Estatística experimental na agropecuária

Alfredo Ribeiro de Freitas

Embrapa

Empresa Brasileira de Pesquisa Agropecuária
Embrapa Pecuária Sudeste
Ministério da Agricultura, Pecuária e Abastecimento

Estatística experimental na agropecuária

Alfredo Ribeiro de Freitas

Embrapa
Brasília, DF
2022

Embrapa Pecuária Sudeste
Rodovia Washington Luiz, Km 234 s/nº,
Fazenda Canchim, Caixa Postal: 339,
CEP: 13560-970, São Carlos, SP
Fone: (16) 3411-5600
www.embrapa.br
www.embrapa.br/fale-conosco/sac

Unidade responsável pelo conteúdo
Embrapa Pecuária Sudeste

Comitê Local de Publicações

Presidente
André Luiz Monteiro Novo

Secretário-executivo
Luiz Francisco Zafalon

Membros
Mara Angélica Pedrochi
Maria Cristina Campanelli Brito
Silvia Helena Piccirillo
Gisele Rosso

Unidade responsável pela edição
Embrapa, Superintendência de Comunicação

Coordenação editorial
Carla Alessandra Timm
Nilda Maria da Cunha Sette

Supervisão editorial
Cristiane Pereira de Assis

Revisão de texto
Everaldo Correia da Silva Filho

Normalização bibliográfica
Márcia Maria Pereira de Souza

Projeto gráfico, editoração eletrônica
e tratamento das ilustrações
Júlio César da Silva Delfino

Capa
Paula Cristina Rodrigues Franco

1ª edição
Publicação digital (2022): PDF

Dados Internacionais de Catalogação na Publicação (CIP)
Embrapa

Freitas, Alfredo Ribeiro de.
Estatística experimental na agropecuária / Alfredo Ribeiro de Freitas. – Brasília, DF : Embrapa, 2022.
PDF (457p.) : il. color.

ISBN 978-65-89957-26-3

1. Estatística agrícola. 2. Análise estatística. 3. Estatística de produção. I. Freitas, Alfredo Ribeiro de. II. Título. III. Embrapa Pecuária Sudeste.

CDD 519.5

Márcia Maria Pereira de Souza (CRB 1/1441)

Autor

Alfredo Ribeiro de Freitas
Engenheiro-agrônomo, doutor em Genética Quantitativa, pesquisador aposentado da Embrapa Pecuária Sudeste, São Carlos, SP

Dedicatória

À minha família: Maria Alice e Matheus, filhos da união de Maria do Carmo e eu. Maria Alice é casada com Helton, e Matheus, com Bruna, que nos presenteou com o maravilhoso neto Cauã.

Dedico também a três pessoas que foram importantes em minha vida profissional. Interagir com elas foi uma das razões para considerar o trabalho com estatística experimental uma diversão e terapia.

Antonio Lourenço Guidoni (in memoriam), que dedicou grande parte de sua vida à estatística experimental, com o objetivo de inovar e aperfeiçoar a qualidade da pesquisa na Embrapa.

João Gilberto Corrêa da Silva: grande conselheiro. Foi o responsável por implantar e organizar as atividades de métodos quantitativos na Embrapa.

Waldomiro Barioni Junior, sempre com os pés no chão e a cabeça nas estrelas. Na área de estatística, é uma pessoa fundamental para conectar pesquisadores dentro da Embrapa e entre ela e as universidades.

Agradecimentos

Aos pesquisadores e demais funcionários da Embrapa Pecuária Sudeste pelo apoio. Quanto à revisão do livro e às valiosas sugestões apresentadas, meu agradecimento, em especial, ao professor Alencariano José da Silva Falcão, professor da Universidade Federal do Tocantins, ao pesquisador Arlei Coldebella (Embrapa Suínos e Aves), à pesquisadora Maria Cristina Neves de Oliveira (Embrapa Soja) e ao pesquisador Waldomiro Barioni Junior (Embrapa Pecuária Sudeste).

Apresentação

A estatística experimental é uma ciência cujo conhecimento é indispensável nas instituições ligadas ao ensino e à pesquisa. Se o propósito da ciência é aumentar o conhecimento e melhorar a compreensão do homem acerca dos fenômenos naturais, a estatística é a ferramenta mais comprometida com a abordagem científica, com o propósito de transformar o cenário da incerteza contido nos dados em conhecimento. Sua teoria pode ser descrita como a "matemática da incerteza". Quando direcionada para a agropecuária, como no presente caso, é uma área fundamental por fornecer ferramentas indispensáveis na análise e interpretação dos dados oriundos de pesquisas, como a interpretação de parâmetros e de fenômenos de natureza biológica, que, além dos tratamentos impostos e controlados pelos pesquisadores, são também influenciados por fatores ambientais e genéticos.

Como sabemos, a quantidade de informação no mundo científico duplica num período de aproximadamente 4 a 5 anos, e o tamanho e a quantidade dos bancos de dados crescem com velocidade ainda maior nas instituições de pesquisas e universidades. Nesse contexto, a estatística experimental se torna ainda mais importante por ser uma potente ferramenta que transforma informações existentes nos dados em resultados que podem ser compreendidos, interpretados e transmitidos ao mundo científico e para a sociedade.

Ao longo dos 14 capítulos deste livro, são abordados tópicos fundamentais da estatística experimental aplicada à pesquisa agropecuária. Em cada capítulo, são discutidas várias aplicações da estatística, na maioria das vezes com dados reais, por meio de exercícios resolvidos e propostos, usando recursos computacionais na análise de dados e solução de problemas usando o software Statistical Analysis System (SAS), robusto pacote estatístico utilizado por grandes corporações.

Rui Machado
Pesquisador da Embrapa Pecuária Sudeste

Prefácio

Neste livro, são abordados tópicos fundamentais da estatística experimental aplicada à agropecuária. Embora as discussões e as aplicações sejam mais direcionadas a estudantes e profissionais ligados às ciências agrárias, o conteúdo do livro é destinado a todos que necessitam familiarizar-se com a aplicação da estatística experimental.

Em cada capítulo, são discutidas várias aplicações da estatística por meio de exercícios resolvidos usando recursos do Statistical Analysis System (SAS) e, na maioria das vezes, utilizando-se dados reais. Devido à grande quantidade de exercícios resolvidos e propostos, o livro é de grande interesse também para uso didático e para candidatos a concursos nos quais são necessários conhecimentos básicos de estatística e suas aplicações. A diversidade dos exercícios resolvidos e propostos procura ilustrar parte da minha experiência como pesquisador da área de estatística e genética na Embrapa.

Outra preocupação na elaboração do livro relacionou-se ao software usado. Foi utilizada a ferramenta SAS, que é, indiscutivelmente, o software mais abrangente e bem documentado em todo o mundo. O SAS é o pacote estatístico mais utilizado pelas grandes corporações e usado por mais de 70 mil empresas em 140 países. Apesar de sua importância, muitos usuários não têm a oportunidade de aprendizagem e de contato com esse software, uma vez que o mundo acadêmico está adotando, principalmente, a linguagem R, que é gratuita. Entretanto, atualmente o SAS está disponibilizando versão estudantil, com acesso on-line gratuito, especificamente para estudos acadêmicos, dissertações, teses, etc.

As informações contidas neste livro resultam das minhas experiências e atividades de pesquisas dentro da Embrapa, um trabalho realizado em equipe, que envolveu pesquisadores da Embrapa, professores e estudantes. Apesar de todo o cuidado na elaboração deste livro, inevitavelmente erros irão existir, pelos quais assumo inteira responsabilidade. Agradeço e sempre serei grato às correspondências e sugestões posteriores com o intuito de melhorias na publicação.

O autor

Sumário

Capítulo 1

Conceitos básicos em estatística

Introdução

Um dos pré-requisitos para iniciar o estudo de estatística é a compreensão de alguns conceitos básicos. Entretanto, antes de utilizarmos esses conceitos, é fundamental um breve histórico, descrito em Salsburg (2009), de como surgiram, e foram assimiladas pela sociedade, as ciências exatas e, dentre elas, a estatística.

Até a metade do século 19, havia muita dificuldade de a população entender os fundamentos da ciência exata, principalmente a estatística. De certa forma, o movimento romântico que aconteceu naquele século foi uma reação ao uso da ciência exata. O romantismo é a arte do sonho e da fantasia que valoriza as forças criativas do indivíduo e da imaginação popular e baseia-se na fé, no sonho, na paixão, na intuição, na saudade. Em oposição, na ciência exata prevalece a razão.

Um exemplo que mostra a dificuldade da sociedade do século 19 quanto à compreensão da matemática era o diálogo entre o imperador Napoleão Bonaparte e Pierre Simon Laplace, nos primeiros anos daquele século. Laplace havia escrito um livro monumental e definitivo, no qual descrevia como calcular as futuras posições de planetas e cometas com base em algumas observações feitas a partir da Terra. “Não encontro menção alguma a Deus em seu tratado, sr. Laplace”, teria questionado Napoleão, ao que Laplace teria respondido: “Eu não tinha necessidade dessa hipótese.”

No entanto, uma prova do potencial da ciência exata que vislumbrou a imaginação popular apareceu na década de 1840. As leis físicas e matemáticas de Newton na aplicação em astronomia foram usadas para prever a existência de mais um planeta − e Netuno foi descoberto no lugar que as leis previram. Dessa forma, a ciência exata chegou ao fim do século 19 mostrando que, por meio de fórmulas matemáticas (como as leis do movimento de Newton e as leis dos gases de Boyle), seria possível descrever a realidade e prever eventos futuros. Tudo de que se necessitava para tal predição era um conjunto completo dessas fórmulas e um grupo de medições a elas associadas, realizadas com suficiente precisão.

Para a cultura popular compreender essa revolução científica, eram necessárias três ideias matemáticas: aleatoriedade, probabilidade e estatística. Aleatoriedade − está associada ao conceito de distribuição probabilística, a qual impõe limitações à aleatoriedade na capacidade de prever eventos futuros aleatórios. Probabilidade − palavra atual para um conceito muito antigo, o qual apareceu em Aristóteles, que afirmou: “É da natureza da probabilidade que coisas improváveis aconteçam.” Como ciência, a probabilidade teve início em 1654 com os matemáticos franceses Blaise Pascal e Pierre de Fermat e foi aplicada a jogos de dados e de cartas. De início, ela envolve a

sensação de alguém a respeito do que se pode esperar. De Moivre conseguiu inserir os métodos de cálculo nessas técnicas, e Jakob Bernoulli foi capaz de estabelecer alguns profundos teoremas fundamentais, chamados “leis dos grandes números”. Apesar da natureza incompleta da teoria da probabilidade, ela se mostrou útil para a ideia, que então se desenvolvia, de distribuição estatística.

Uma distribuição estatística ocorre quando consideramos um problema científico específico. A estatística baseia-se em fórmulas e símbolos matemáticos e tem sua aplicação principal na abordagem da incerteza, principalmente no estudo de dados científicos.

População

É o conjunto de todos os indivíduos ou elementos que representam o universo de determinado problema. Por exemplo, a população humana de um país ou de uma região, o número de peixes de determinada espécie em um rio, o número de aves de uma espécie na fauna brasileira, as árvores doentes de uma floresta, etc.

Conceitualmente, uma população pode ser infinita ou finita, em tamanho; é infinita quando não for possível enumerar ou listar todas as unidades individuais a ela pertencentes, como a população de peixes de determinada espécie em um rio, pois a todo instante novos indivíduos são descobertos e nunca se tem certeza sobre o tamanho da população em determinado tempo.

Uma população é finita quando as unidades são facilmente enumeráveis, como o número de fazendeiros e/ou o número de produtores de leite de um município.

Amostra

Amostra é a parte da população convenientemente escolhida e que apresenta as características dessa população. Uma amostra é aleatória simples quando *n* indivíduos são selecionados a partir de uma população com N indivíduos, tendo todos os elementos da população a mesma probabilidade de pertencerem à amostra. Uma amostra aleatória de tamanho *n* é estratificada quando uma população é dividida em subconjuntos ou subpopulações (estrato) e uma amostra com *n* indivíduos é selecionada de cada estrato. O processo de seleção deve ser tal que cada uma das N unidades amostrais de cada estrato tenha a mesma probabilidade de pertencer à amostra de cada estrato. A amostra final é formada, então, pela união das amostras selecionadas de cada um dos estratos.

Tanto na amostra aleatória simples quanto na amostra aleatória estratificada, a seleção pode ser feita com ou sem reposição.

Vantagens da amostragem

Na maioria das situações, é impossível obter dados e realizar pesquisas com todos os indivíduos da população. Em razão disso, o processo de amostragem frequentemente é a única ferramenta de que dispomos. Além de contribuir para as mais variadas atividades do ser humano, os dados de amostras geralmente são obtidos mais rápidos a um custo barato e, consequentemente, as pesquisas e os resultados também são mais rápidos, o que é importante em todas as áreas.

A seguir são apresentadas algumas ilustrações da aplicação dessa técnica para a solução de problemas práticos. No estudo das populações infinitas e homogêneas, principalmente quando se trata de substâncias, uma pequena amostra pode ser representativa da população.

Para a manutenção da qualidade da água das piscinas, é necessário que a faixa de cloro se situe entre 1,0 mg L^{-1} e 1,5 mg L^{-1}; e a de pH entre 7,4 e 7,6. Diariamente, realiza-se a análise de algumas gotas de água, utilizando-se um kit apropriado. Com isso, é possível verificar a qualidade da água, não importando o tamanho da piscina. Para conhecer a qualidade das águas dos rios, dos lagos, etc., o procedimento é semelhante. Nessas situações, faz-se a coleta de alguns mililitros (mL) de água de alguns pontos, a fim de obter uma amostra homogênea, a qual será enviada ao laboratório para análise. O mesmo procedimento pode ser usado para verificar a qualidade do leite em um tanque resfriador no campo. Colhem-se alguns mililitros (amostra) de leite de um determinado produtor por meio de um kit apropriado, a fim de verificar se o pH do leite está ácido ou básico.

Na agricultura, para se conhecer o nível de fertilidade dos solos e utilizar adubação nos mais variados tipos de cultura, pastagens, etc., geralmente são feitas amostras de solos que são enviadas ao laboratório para análises.

Desvantagens da amostragem

Entre as desvantagens do processo de amostragem, pode-se citar o grau de incerteza tolerável no estudo de uma amostra. Essa incerteza tem origem em duas causas: a) variação natural entre os membros da população, que se refletirá na variabilidade das possíveis amostras (aqui há grande contribuição dos métodos estatísticos); b) definição

inadequada da população ou do processo de amostragem, o que induz a um tipo de erro denominado de viés ou vício, o qual pode ocorrer por subjetividade na escolha da amostra ou pelo fato de a pessoa envolvida no processo ter tido, inconscientemente, preferência na escolha de determinados elementos da amostra em detrimento de outros, o que contraria o processo de aleatorização (todos os elementos devem ter a mesma chance de pertencer à amostra).

Recenseamento e censo

Recenseamento é o estudo científico de um universo de objetos, de indivíduos, etc., observando-se todas suas informações, com o propósito de adquirir conhecimentos quantitativos acerca de características importantes desse universo. Por exemplo, um censo agropecuário é o levantamento de informações sobre estabelecimentos agropecuários, florestais e/ou agrícolas de todos os municípios de um país.

Censo é o exame de todas as unidades ou indivíduos de uma população, de um município, região, estado, que informa diferentes características. Frequentemente, é impossível ou mesmo ineficiente realizar esse processo e, nesse caso, utiliza-se o processo de amostragem.

Embora esses dois conceitos sejam mais utilizados para habitantes de um município, região ou estado, em algumas situações podem ser realizados na agropecuária.

Análise exploratória e inferência estatística

Análise exploratória é o conjunto de procedimentos e técnicas de manejo de dados, com o objetivo de calcular médias e percentagens, bem como de construir gráficos, tabelas, etc., sem envolver grande teorização sobre o assunto. Fornece contribuições valiosas para o gerenciamento e monitoramento da precisão da coleta de dados e avaliação de sua qualidade, além de possibilitar o diagnóstico rápido das atividades, calibrar instrumentos de mensuração, entre outras. Em síntese, a análise exploratória procura padrões e características interessantes nos dados que indiquem possíveis modelos, padrões ou tendências.

Inferência estatística é o emprego de técnicas estatísticas que tem por objetivo a tomada de decisão sobre a população, por meio de evidências fornecidas pela amostra. Essa decisão está associada a um grau de incerteza e, consequentemente, a

uma probabilidade de erro. A inferência estatística é mais complexa do que a análise exploratória, principalmente pelo fato de exigir maior esforço teórico. Toda inferência na estatística está baseada na teoria das probabilidades.

Testes paramétricos e não paramétricos

Testes paramétricos são aqueles em que se pressupõe que as variáveis a serem estudadas são de natureza quantitativa e descritas por meio de parâmetros como média e variância, e que apresentam determinado tipo de distribuição de probabilidade dos erros amostrais, sendo a distribuição normal a mais comum. Permitem detectar diferenças muito sutis entre as variáveis estudadas, porém sua aplicação é mais limitada do que a dos testes não paramétricos, em razão da necessidade de se comprovar que todos os pressupostos estão de fato satisfeitos, como normalidade e independência dos erros, aditividade dos fatores de variação e homogeneidade de variância dos tratamentos. Com todos esses requisitos atendidos, a estatística paramétrica é, em geral, mais poderosa do que a não paramétrica.

Testes não paramétricos são aqueles em que se pressupõe relativamente pouca informação acerca das variáveis envolvidas, não requerem nenhuma distribuição de probabilidade dos erros e as comparações são baseadas no ordenamento (ranking) dos resultados, do mais baixo para o mais alto. Em geral, são menos sensíveis do que os testes paramétricos, mas podem ser aplicados a um conjunto mais amplo de situações.

Estudo observacional e estudo experimental

No estudo observacional, o pesquisador coleta dados e extrai informações, mas faz o possível para não influenciar os eventos. Ainda nesse estudo, os fenômenos surgem espontaneamente, sem nenhum artifício, e independem da vontade do homem, diferindo dos estudos experimentais. Por exemplo, os indivíduos da amostra não são designados aos grupos por processo aleatório, mas já faziam parte do grupo no início da pesquisa. Por exemplo, comparar a produção de leite, por determinado período, de um grupo de vacas com determinada doença com um um grupo de vacas sadias, em que não houve nenhuma intervenção do homem.

Por sua vez, no estudo experimental, o pesquisador deliberadamente influencia os eventos, buscando verificar os efeitos da intervenção. Nesses estudos, o objetivo é estabelecer uma relação de causa e efeito entre variáveis em estudo de forma prática, como descobrir algo desconhecido ou testar uma hipótese.

Estudo transversal e estudo longitudinal

No estudo transversal, o pesquisador coleta os dados de cada indivíduo num único instante no tempo, obtendo-se uma situação momentânea do fenômeno investigado. Esse estudo é relativamente rápido, consome menos recursos e é menos vulnerável às variáveis estranhas; no entanto, só é possível conhecer apenas uma situação pontual do fenômeno e não o fenômeno em toda a sua extensão.

Em um estudo longitudinal, o pesquisador coleta os dados de cada indivíduo ou sujeito em dois ou mais momentos (*follow-up*), havendo acompanhamento do desenrolar do fenômeno considerado. É relativamente lento, consome mais recursos e é mais vulnerável a variáveis estranhas. Propicia, no entanto, condições de avaliar todo o desenrolar do fenômeno em si. Esse estudo é bastante similar à análise de séries temporais. Porém, esta última dispõe de uma única unidade amostral ou indivíduo com muitas observações (geralmente acima de 100) ao longo do tempo. No estudo longitudinal (ou medidas repetidas), as respostas são avaliadas no mesmo animal ou indivíduo, mas geralmente abaixo de 20 observações por indivíduo.

Precisão e acurácia

Precisão se refere à fidedignidade na mensuração feita repetidamente em uma mesma característica. Ela expressa o grau de consistência da característica medida com relação à sua média, ou seja, está diretamente ligada à dispersão das observações, com a dispersão do valor estimado. Quanto mais próximos forem os valores das medidas realizadas, mais precisas são as mensurações. Alta precisão significa pequena ou nenhuma variação nas medidas, enquanto baixa precisão significa grande variação entre as medidas. No entanto, a precisão em um processo de mensuração não significa que a avaliação esteja correta. Por exemplo, se uma balança sistematicamente aumenta o peso em 10%, assim as pesagens sucessivas de um determinado objeto podem ser bastante próximas entre si, mas distantes do valor verdadeiro. Essa distância entre o valor médio das medidas realizadas com o valor verdadeiro é o viés, vício ou tendência.

Acurácia ou exatidão, que envolve tanto efeitos sistemáticos quanto aleatórios, representa o grau de proximidade de uma estimativa com seu parâmetro (ou valor verdadeiro). A acurácia é tomada como sendo o afastamento entre o valor de referência e o valor estimado. Não havendo viés, vício ou tendência, a acurácia se resume à medida de precisão.

Análise univariada e multivariada

Análise univariada refere-se ao uso das técnicas e métodos de estatística para a análise de cada variável separadamente.

Já a análise multivariada refere-se ao uso das técnicas e métodos de estatística para análise conjunta de múltiplas variáveis dependentes e/ou múltiplas variáveis independentes.

Por exemplo, em uma pesquisa com camarão, são avaliadas as características: comprimento total, comprimento do abdômen, perímetro do abdômen, comprimento do cefalotórax, perímetro do cefalotórax e, também, o peso corporal. Se o objetivo ou finalidade da pesquisa é uma análise conjunta dessas características com relação ao sexo (macho e fêmea), trata-se de análise multivariada. Se o objetivo é estudar o efeito de sexo para cada variável separadamente, trata-se de análise univariada.

Variável

Uma variável representa uma descrição ou um nível, atribuindo-se números ou outros símbolos, para cada característica dentro de um conjunto de dados. Geralmente utiliza x_1, x_2 ... x_n para representar os dados de uma amostra de tamanho *n*, e x_1, x_2 ... x_N para representar os dados de uma população de tamanho N.

Variáveis qualitativas e quantitativas

Variáveis qualitativas representam a informação que identifica alguma qualidade, categoria ou característica, não suscetível de medida, mas de classificação (de alocação numa categoria). Podem ser categórica nominal e categórica ordinal.

Variáveis quantitativas ou numéricas representam a informação resultante de características suscetíveis de serem medidas. Podem assumir qualquer valor no conjunto dos números reais, tais como peso, em quilograma (kg), e altura, em centímetro (cm), etc.

Resumindo, são quatro os níveis de medidas de uma variável:

a) Categórica nominal − classifica os dados em classes ou categorias independentes (ex.: macho e fêmea), porém não existe ordenamento entre as categorias ou classes.

b) Categórica ordinal − classifica os dados em classes ou categorias, porém estabelece uma relação de ordem entre as classes ou categorias. Exemplos: pouco importante, importante, muito importante; nível de gravidade de uma doença (1, 2, 3, ...).

c) Numérica discreta − quando envolve uma unidade de medida e o resultado desta é um número inteiro. Exemplos: número de frutos de uma árvore, número de bezerros de uma fazenda nascidos em determinado ano, número de aves em um galpão, etc.

d) Numérica contínua − pode assumir números inteiros e fracionários, geralmente dentro de um intervalo conhecido, porém, pode assumir um conjunto infinito de valores possíveis. Exemplos: altura, em centímetro (cm); peso, em quilograma (kg), etc.

Medidas de tendência central

São estatísticas que fornecem uma tendência central de um conjunto de dados. São a média, a moda e a mediana. Em uma distribuição de probabilidade simétrica, a média aritmética simples, a moda e a mediana são iguais.

A seguir, serão discutidas algumas aplicações dessas três medidas de tendência central − médias (aritmética simples e ponderada, média geométrica simples e ponderada, média harmônica simples e ponderada), a moda e a mediana.

Média

Média aritmética simples

No caso de uma amostra $x_1, x_2, \ldots, x_n$, em que *n* é o tamanho amostral, a média aritmética simples (MA) é dada por:

$$MA = (\sum_{i=1}^{n} x_i)/n$$

Exemplo 1

Se a produtividade de leite diária (y), em quilograma (kg), de sete vacas é: 20 kg, 21 kg, 21 kg, 22 kg, 23 kg, 24 kg, 25 kg, usando-se o software Sistema de Análise Estatística (SAS), o cálculo da média é 22,28 kg.

```
data exemplo1;
input y @@;              /* a variável y é repetida na linha */
datalines;               /* inicia os dados  */
20 21 21 22 23 24 25     /* dados */
;                        /* término dos dados */
proc means mean;         /* calcula a média */
run;                     /* executa o programa */
output                   /* saída dos resultados */
mean
22.28
```

Média aritmética ponderada

Para dados agrupados $x'_1, x'_2, \ldots, x'_m$ com frequências $f_1, f_2, \ldots, f_m$, a média aritmética ponderada (MAP) é calculada por:

$$\text{MAP} = \frac{\sum_{j=1}^{m} x'_j f_j}{\sum_{j=1}^{m} f_j}$$

Exemplo 2

Se a média do salário mensal de trabalhadores de duas categorias for R$ 2.000,00 e R$ 2.500,00, respectivamente, é possível que a média mensal do rendimento dessas duas categorias seja R$ 2.250,00?

A média do rendimento mensal (RM) desses trabalhadores somente será de R$ 2.250,00 se o número de funcionários nas duas categorias for igual. O RM é obtido no programa SAS a seguir considerando-se cinco funcionários (Func).

```
data exemplo2;                /* indica o nome do arquivo */
input func;                   /* fornece a variável "func" com valor de  5 */
rm = (5*2000 + 5*2500)/10;    /* cálculo da média aritmética ponderada */
datalines;                    /* inicia os dados  */
5                             /* valor atribuído a "func"   */
;                             /* término dos dados */
proc print; var rm;           /* imprime a var rm */
run;                          /* executa o programa */
output                        /* saída dos resultados */
rm
2250
```

Média geométrica

A média geométrica (MG) tem aplicações em várias áreas, principalmente de finanças e engenharia. É utilizada somente para números positivos. A média geométrica de uma amostra é sempre menor que a média aritmética.

$$MG = \sqrt[n]{x_1 x_2 \ldots x_n} = (x_1\, x_2 \ldots x_n)^{1/n}$$

Exemplo 3

Sejam x_1 e x_2 números diferentes, inteiros e positivos, provar que a média aritmética (MA) simples desses números é maior do que a média geométrica (MG), ou seja, MA > MG.

Tem-se $MA = (x_1 + x_2)/2$ e $MG = \sqrt{x_1 x_2}$

Essa prova é demonstrada abaixo para $X_1 = 10$ e $X_2 = 15$. Obtém-se MA = 12,50 e MG = 12,24.

```
data exemplo3;
input x1 x2;
MA = (x1+ x2)/2;
MG = (x1*x2)**.5;
datalines;
10 15
;
proc print; var MA MG;
run;
output
obs   MA      MG
1     12.50   12.24
```

Média geométrica ponderada

A média geométrica ponderada (MGP) ocorre quando cada valor x_i tem uma frequência ou peso f_i. A interpretação é a mesma da média geométrica simples. Se todas as frequências são iguais, a média geométrica ponderada é igual à média geométrica.

$$MGP = \sqrt[\sum f_i]{x_1^{f_1} x_2^{f_2} \ldots x_m^{f_m}} = \sqrt[\sum f_i]{\prod_{i=1}^{m} x_i^{f_i}}$$

Média harmônica

Média harmônica simples

A média harmônica simples (MH) é empregada em várias áreas e situações e geralmente está associada com a obtenção de médias de alguma classe ou categoria quando elas possuem número diferente de observações.

$$MH = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

em que *n* é o tamanho amostral.

Isso significa dizer que:

- A média harmônica é a recíproca da média aritmética dos recíprocos dos valores.
- O inverso da média harmônica é a média aritmética dos recíprocos dos valores.

Média harmônica ponderada

A média harmônica ponderada (MHP) ocorre quando cada valor x_j tem uma frequência f_j.

$$MHP = \frac{\sum_{j=1}^{m} f_j}{\sum_{j=1}^{m} \frac{1}{x_j} f_j}$$

Moda

Moda (MO) é o valor ou o intervalo de classe, no caso de dados discretos e dados contínuos, respectivamente, que surge com maior frequência. A moda é especialmente útil para reduzir a informação de um conjunto de dados qualitativos, que se apresenta sob a forma de nomes ou categorias, para os quais não se pode calcular a média nem a mediana, uma vez que não é possível a ordenação dos dados.

$$MO = l_{MO} + \frac{h(f_{MO} - f_{MO-1})}{(2f_{MO} - f_{MO-1} - f_{MO+1})}$$

em que:

l_{MO} = limite inferior da classe modal.
h = amplitude da classe modal.
f_{MO} = frequência absoluta da classe modal.
f_{MO-1} = frequência absoluta da classe anterior à classe modal.
f_{MO+1} = frequência absoluta da classe posterior à classe modal.

Mediana

Para um conjunto de dados $x_1, x_2, ..., x_n$, dispostos em ordem crescente, a mediana (MD), pertencente ou não à amostra, é o valor central se *n* for ímpar, ou a média aritmética dos dois termos centrais, se *n* for par. Para um conjunto de dados com frequência, em ordem crescente e agrupados em classes, a mediana (P_{50}) pode ser calculada pela fórmula a seguir. Para calcular outro percentil basta alterar o termo ($\frac{50n}{100}$).

$$MD = l_i + \left(\frac{\frac{50n}{100} - \Sigma f}{f_d}\right) h$$

em que:

l_i = limite inferior da classe que contém a mediana.

n = número de observações.

Σf = frequências acumuladas até a classe anterior à classe da mediana.

f_d = frequências absolutas da classe da mediana.

h = amplitude de classe.

Relação empírica entre a média, a moda e a mediana

$\overline{X}$ - Moda = 0 (assimetria nula ou distribuição simétrica).

$\overline{X}$ - Moda < 0 (assimetria negativa ou assimetria à esquerda).

$\overline{X}$ - Moda > 0 (assimetria positiva ou assimetria à direita).

Considerando-se os dados fictícios do rendimento de uma cultura, em toneladas por hectare (t ha^{-1}), conforme mostrados na Tabela 1, serão realizados cálculos de medidas de tendência central.

Tabela 1. Distribuição de classes e frequências considerando-se dados fictícios do rendimento de uma determinada cultura.

Classe (t ha^{-1})	Frequência absoluta	Frequência acumulada	Frequência relativa	Frequências relativa acumulada (%)
2,0−2,8	2	2	18,2	18,2
2,8−3,6	3	5	27,3	45,5
3,6−4,4	1	6	9	54,5
4,4−5,2	3	9	27,3	81,8
5,2−6,0	2	11	18,2	100

Exemplo 4

Da Tabela 1, selecionando-se cinco números inteiros (2 t ha^{-1}, 3 t ha^{-1}, 4 t ha^{-1}, 5 t ha^{-1}, 6 t ha^{-1}), na coluna classe, tem-se alguns exemplos de cálculos:

a) Média aritmética ponderada (MAP):

$$\text{MAP} = \left(\frac{2 \text{ x } 2 + 3 \text{ x } 3 + 4 \text{ x } 1 + 5 \text{ x } 3 + 6 \text{ x } 2}{2 + 3 + 1 + 3 + 2}\right) = \frac{44}{11} = 4$$

b) Média geométrica ponderada (MGP):

$$\text{MGP} = \sqrt[11]{2^2 3^3 4^1 5^3 6^2} = (1944000)^{1/11} = 3{,}72$$

c) Média harmônica ponderada (MHP):

$$\text{MHP} = \frac{2 + 3 + 1 + 3 + 2}{\frac{1}{2}2 + \frac{1}{3}3 + \frac{1}{4}\ 1 + \frac{1}{5}3 + \frac{1}{6}2 =} = 3{,}45$$

No programa SAS *exemplo4*, calculam-se os itens (a), (b) e (c), obtendo-se 4,00; 3,72 e 3,45, respectivamente, para a média aritmética ponderada (MAP), média geométrica ponderada (MGP) e média harmônica ponderada (MHP).

```
data exemplo4;
input x1 f1 x2 f2 x3 f3 x4 f4 x5 f5;          /*input dos valores e frequências absolutas */
sf = f1+f2+f3+f4+f5;                          /* soma de frequências absolutas */
map = (x1*f1 + x2*f2 + x3*f3 + x4*f4 + x5*f5)/sf;          /* média aritmética */
mgp = (x1**f1*x2**f2*x3**f3*x4**f4*x5**f5)**(1/sf);        /*média geométrica*/
mhp = sf/(1/x1*f1+1/x2*f2+1/x3*f3+1/x4*f4+1/x5*f5);        /* média harmônica */
datalines;
2 2 3 3 4 1 5 3 6 2
```

```
;
proc print noob; var map mgp mhp; run;          /* imprime map, mgp, mhp */
Output
map     mgp      mhp
4.00    3.72     3.45
```

Exemplo 5

Dada uma progressão geométrica do tipo 3, 6, 12, 24, 48, provar que a média geométrica desses números é o termo central da progressão geométrica.

Inicialmente, atribui-se x para 3 e a série 3, 6, 12, 24, 48 resulta, portanto, em x, 2x, 4x, 8x, 16x, que também pode ser reescrita na forma 2^0x, 2^1x, 2^2x, 2^3x, 2^4x.

A média geométrica dessa série é:

$$MG = \sqrt[5]{2^0 \text{ x } 2^1 \text{ x } 2^2 \text{ x } 2^3 \text{ x } 2^4 \text{ x}} = 5\sqrt{2^{10}X^5} = (2^{10}\text{x}^5)^{1/5} = 2^{10/5}.\ \text{x}^{5/5} = 2^2\text{x} = 4\text{x}$$

Como foi atribuído x = 3, então 4x = 4 x 3 = 12 é o termo central da progressão.

```
data exemplo5;
input x;
mg = (2**10*x**5)** (1/5);
cards;
3 ;
proc print; var mg;run;
output
mg
12
```

Quartis, decis e percentis

São estatísticas ou separatrizes que se calculam a partir de uma amostra de dados ordenados.

Quartis

Os dados ordenados são divididos em quatro partes iguais e cada parte tem 25% dos dados. Na posição 25% tem Q_1; na posição 50% tem Q_2 e na posição 75% tem Q_3.

Para n dados discretos: $Q_1(n/4)$; Q_2 $(2n/4=n/2=$ Md); $Q_3(3n/4)$.

$$Q_1 = l_i + [\frac{(\frac{n}{4} - \Sigma f)}{fq_1}] \, h$$

em que:

l_i = limite inferior da classe que contém Q_1.

n = tamanho da amostra; aqui n/4 é a posição para Q_1 (25%).

Σf = soma das frequências das classes inferiores a Q_1.

fq_1 = frequência da classe que contém Q_1.

h = amplitude da classe que contém Q_1.

Q_2 = mediana.

Para o cálculo do terceiro quartil (Q_3), na fórmula anterior, substitui-se (n/4) por (3n/4), e seguem os demais componentes para a classe que contém Q_3.

Frequentemente se utiliza o coeficiente de variação quartílico (CVQ) para fazer comparações dentro e entre conjuntos de dados. Sua interpretação é similar ao coeficiente de variação.

$$CVQ = \frac{Q_3 - Q_1}{Q_3 + Q_1} \text{ x } 100$$

Exemplo 6

Os dados a seguir se referem ao peso corporal, em gramas (g), de dez camarões machos e dez fêmeas, com aproximadamente 3 meses de idade, de um estoque pronto para despesca, cultivado no Rio Grande do Norte. Segue-se uma interpretação desses dois conjuntos de dados usando-se o coeficiente de dispersão quartílico.

Machos (g): 16,48; 15,26; 14,93; 14,62; 14,41; 13,26; 10,32; 14,03; 15,65 e 13,28.

Q_1 = 13,28; Q_3 =15,26 e CVQ = 6,94%

Fêmeas (g): 9,94; 8,57; 9,90; 14,37; 10,21; 13,65; 10,95; 11,32; 12,76 e 10,89.

Q_1 = 9,94; Q_3 =12,76 e CVQ = 12,42%

Observa-se maior uniformidade dos dados de pesos para camarões machos.

Decis

Os dados ordenados são divididos em décimas partes, e cada parte tem 10% dos dados. Na posição 10% tem-se D_1, na posição 20% tem-se D_2, e assim por diante.

Para *n* dados discretos: D_1 (n/10), D_2 (2n/10), ..., D_9 (9n/10).

Percentis

Os dados ordenados são divididos em centésimas partes, e cada parte tem 1% dos dados. Na posição 1% tem-se P_1, na posição 2% tem-se P_2, e assim por diante.

Para *n* dados discretos: P_1 (1n/100), P_2 (2n/100), ..., P_{99} (99n/100).

Para dados agrupados em classes, a fórmula usada no cálculo de Q_1, citada em Quartis, pode ser usada para calcular decis e percentis, desde que se usem os devidos componentes da fórmula para a classe que contém o decil ou percentil a ser calculado.

Exemplo 7

Considerando-se os dados da Tabela 1, os percentis P_{10}, P_{25}, P_{50} e P_{75} são calculados abaixo e também pelo *exemplo7*.

$$P_{10} = 2{,}0 + \left[\frac{(10x11)/100)-0}{2}\right] x\ 0{,}8 = 2{,}44$$

$$P_{25} = 2{,}8 + \left[\frac{(25x11)/100)-2}{3}\right] x\ 0{,}8 = 3{,}00$$

$$P_{50} = 3{,}6 + \left[\frac{(50x11)/100)-5}{1}\right] x\ 0{,}8 = 4{,}00$$

$$P_{75} = 4{,}4 + \left[\frac{(75x11)/100)-6}{3}\right] x\ 0{,}8 = 5{,}00$$

```
data exemplo7;
/* calcula os percentis: p10, p25, p50 e p75:  */
input n  h;
p10 = 2.0 + (((10*n/100)-0)/2)*h;
p25 = 2.8 + (((25*n/100)-2)/3)*h;
p50 = 3.6 + (((50*n/100)-5)/1)*h;
p75 = 4.4 + (((75*n/100)-6)/3)*h;
datalines;
11  .8
;
proc print noobs; var p10 p25 p50 p75;run;
output
p10    p25      p50      p75
2.44   3.00     4.00     5.00
```

Medidas de dispersão

As medidas de dispersão expressam diferentes formas de quantificação da tendência que os resultados de um experimento aleatório têm de se concentrarem ou não em determinados valores.

Medidas de dispersão absolutas: amplitude total, variância, desvio-padrão, desvio médio e desvio médio absoluto

Amplitude total

É a diferença entre o maior e o menor valor observado de um conjunto de dados. Com intervalos de classe, é a diferença entre o limite superior da última classe e o limite inferior da primeira classe. A amplitude total é importante para caracterizar a abrangência de um estudo. Por exemplo, amplitude de variáveis climáticas, temperatura em um dia, controle de qualidade de um produto industrial ou do resultado de um laboratório, amplitude de renda familiar de uma região, rendimento das culturas agrícolas de uma área, entre outros.

Variância

É a mais usada entre as medidas absolutas de dispersão ou variabilidade em torno da média, sendo extremamente importante na inferência estatística. Na teoria da probabilidade e na estatística, a variância indica o quão distante os valores se encontram do valor esperado. No entanto, tem pouca utilidade como estatística descritiva e de difícil interpretação.

A variância de uma amostra é representada por s^2 e de uma população, por σ^2. A variância de uma amostra de variáveis aleatórias, $X_1, X_2, \ldots, X_n$, é dada por:

$$s^2 = \frac{\Sigma X_i^2 - (\Sigma X_i)^2/n}{n-1} = \frac{\Sigma (X_i - \overline{X})^2}{n-1}$$

A somatória dos quadrados (SQ) dos desvios $(x_i - \overline{x})^2$ é dividida por $(n - 1)$ para a variância amostral (s^2), e por N para a variância populacional (σ^2). A variância e a média aritmética representam os parâmetros da função densidade de probabilidade normal; portanto, a função normal é caracterizada por esses dois parâmetros. Enquanto a média

aritmética é expressa na mesma unidade da medida, a variância é expressa no quadrado da unidade de medida, e disso decorre a dificuldade de se utilizá-la como estatística descritiva. Para um conjunto de dados agrupados, x_1, x_2 , ... x_n, com frequências f_1, f_2... f_n, a variância é calculada por:

$$S^2 = \sum_{i=1}^{n} (X_i - \overline{X})^2 f_i /(n - 1)$$

em que:

f_i = frequência de X_i.

Algumas propriedades da variância

Para qualquer constante k, tem-se:

$v(kx) = k^2 v(x)$

$v(k \pm x) = v(x)$

Essas propriedades são demonstradas no exemplo 8.

Exemplo 8

Considerando-se x_i= 2, 3, 4, 5 e k = 2, o programa *exemplo8* calcula quatro variâncias:

$$v_1 = v(x_i); v_2 = v(2 - x_i); v_3 = v(2 + x_i); v_4 = v(2x_i) = 4v(x_i)$$

As variâncias v_1 a v_3 mostram que $v(k \pm x) = v(x)$, e v_4 mostra que $v(kx) = k^2 v(x)$.

```
data exemplo8;
input x @@;
v1 =x; v2 = 2 − x; v3= 2+ x; v4 = 2* x;
datalines;
2 3 4 5
 ;
proc means var; var v1-v4;
run;
output
variable     variance
v1           1.66
v2           1.66
v3           1.66
v4           6.66
```

Exemplo 9

Dois grupos A e B com 10 pessoas cada foram submetidos a um teste durante 1 hora. A percentagem de erros de cada pessoa durante o teste foi:

A: 1, 2, 1, 2, 1, 2, 3, 3, 1, 2.

B: 1, 2, 2, 2, 1, 1, 2, 2, 1, 1.

Aplicando-se conhecimentos de estatísticas descritivas, qual grupo foi mais eficiente?

Grupo A: $\overline{X}_A = 1{,}80$ e $S^2_A = 0{,}6222$.

Grupo B: $\overline{X}_B = 1{,}50$ e $S^2_B = 0{,}2778$.

Como o grupo B tem menor média de erros do que o grupo A (1,50 versus 1,80), menor variância (0,2778 versus 0,6222) e ainda menor amplitude total (2 − 1 = 1), conclui-se que o grupo B é mais eficiente do que o grupo A. Esses resultados estão descritos também no programa abaixo:

```
data exemplo9;
input a b @@;
cards;
1 1 2 2 1 2 2 2 1 1 2 1 3 2 3 2 1 1 2 1
;
proc means mean var; run;
output
variable   mean          variance
a          1.8000        0.6222
b          1.5000        0.2777
```

Desvio-padrão

O desvio-padrão é a raiz quadrada da variância, $s = \sqrt{s^2}$, e representa uma medida de dispersão ou variabilidade em torno da média, porém, é mais fácil de interpretar do que a variância, pois a unidade é a mesma dos dados. Ele é sempre positivo e tanto maior quanto maior for a variabilidade entre os dados; se $s = 0$, então não existe variabilidade, isto é, os dados são todos iguais.

Relação empírica entre desvio-padrão e amplitude total

Na maioria dos casos, o desvio-padrão se situa entre 1/6 e 1/3 da amplitude total (A_t), ou seja, $1/6A_t < s < 1/3A_t$.

O desvio-padrão amostral *s* é bastante sensível a *outliers,* que são observações que não seguem o mesmo padrão ou são inconsistentes do restante do conjunto de dados a que elas pertencem, distanciando-se significativamente das demais. Assim, uma estimativa robusta de *s* pode ser obtida pela amplitude interquartílica (AI).

$$AI = (Q_3 - Q_1)/1{,}34898$$

Erro-padrão da média

O erro-padrão da média é uma medida de variação de uma média amostral em relação à média da população. Como na prática se trabalha com uma amostra, inicialmente calcula-se a estimativa da variância da média, que é dada por:

$$s^2_{\overline{X}} = \frac{s^2}{n}$$

em que:

s^2 = variância da amostra.

n = tamanho da amostra.

Uma vez obtida a variância da média trabalhando-se com apenas uma amostra, que é uma estimativa da variabilidade das médias que seria obtida, caso tivesse tomado nas mesmas condições todas as amostras possíveis, calcula-se o erro-padrão da média:

$$s_{\overline{X}} = \frac{s}{\sqrt{n}}$$

Na distribuição de probabilidade da curva normal ou gaussiana, a média verdadeira tem a chance de 95% de ficar no intervalo $\overline{X} \pm 1{,}96\, s_{\overline{X}}$.

Desvio médio

É a média do valor absoluto dos desvios em relação à média.

$$D = \sum_{i=1}^{n} |X_i - \overline{X}|/n$$

Desvio médio absoluto

O desvio médio absoluto (DMA) de uma amostra ou conjunto de dados é a diferença absoluta entre cada dado e uma medida de tendência central que pode ser a mediana ou a média. Considerando-se a média, tem-se:

$$DMA = \sum_{i=1}^{n} | x_i - \bar{x} | /n$$

Para dados agrupados com *m* grupos e f_j a frequência do j-ésimo grupo, tem-se:

$$DMA = \sum_{j=1}^{n} | x_i - \bar{x} | f_J / \sum_{j=1}^{m} f_J$$

O desvio médio de cada dado com relação à mediana dá uma ideia dos valores que estão provocando assimetria com viés para a direita ou para a esquerda, e, também, dos valores que mais estão influenciando a medida de assimetria. Para amostras de tamanhos iguais e mesma variável, calculando-se $\sum_{i=1}^{n} |x_i - \text{mediana}|$, o menor valor corresponde à amostra mais homogênea.

No ajuste de modelos, tais como análises de regressão, do tipo linear e não linear, o DMA pode ser utilizado para avaliar a qualidade de ajuste de um modelo e, também, entre modelos, o que significa calcular o DMA para a diferença entre dados observados (y) e estimados ou preditos ($\hat{y}$), ou seja:

$$DMA = \sum_{i=1}^{n} |yi - \hat{y}|/n$$

Exemplo 10

Na Tabela 2, encontra-se o preço mensal da arroba do suíno em duas cidades (A e B). Em qual das duas cidades, o mercado físico do suíno foi mais estável?

Tabela 2. Preço mensal, em R$, da arroba do boi em duas cidades (A e B).

Cidade	Jan.	Fev.	Mar.	Abr.	Maio	Jun.	Jul.	Ago.	Set.	Out.	Nov.	Dez.	Média
Cidade A	238,1	236,8	238,0	239,7	239,0	239,0	239,8	240,0	242,0	245,7	244,7	244,3	240,6
Cidade B	239,4	238,1	238,8	240,7	239,9	240,4	240,8	241,4	242,1	245,9	245,6	245,9	241,6

Solução

Para responder à questão, utilizando-se o DMA, tem-se que a média para a cidade A é 240,6 e para a cidade B é 241,6.

Cidade A: $DMA_A = (|238,1 - 240,6| + |236,8 - 240,6| + ... + |244,3 - 240,6|)/12$

$= 2,5 + 3,8 + ... + 3,7 = 28,7$

Cidade B: $DMA_B = (|239,4 - 241,6| + |238,1 - 241,6| + ... + |245,9 - 241,6|)/12$

$= 2,2 + 3,5 + ... + 4,3 = 26,4$

O DMA foi menor na cidade B, indicando que nesse centro comercial o preço foi mais estável do que na cidade A e, portanto, mais confiável para projeções.

Utilizando-se a rotina SAS, para o exemplo 10, tem-se:

```
data exemplo10;
input cidade_a cidade_b ;
a = abs(cidade_a - 240.6);
b = abs(cidade_b - 241.6);
datalines;
238.1  239.4
236.8  238.1
238.0  238.8
239.7  240.7
239.0  239.9
239.0  240.4
239.8  240.8
240.0  241.4
242.0  242.1
245.7  245.9
244.7  245.6
244.3  245.9
;
proc means sum; var a b;run;
output
variable   sum
a          28.7000000
b          26.4000000
```

Coeficiente de variação

O coeficiente de variação (CV) é a razão entre o desvio-padrão e a média aritmética multiplicada por 100; expressa a variação como a percentagem da média. Sua aplicação é útil para variáveis positivas e tem grande aplicação nos delineamentos experimentais.

$$CV = (s/\overline{X}) \times 100$$

Coeficientes de assimetria e de curtose

Coeficiente de assimetria

É uma medida do afastamento ou viés de quanto a curva de uma distribuição de frequência se afasta da posição simétrica. Geralmente é considerada a distribuição normal. Existem três tipos de coeficientes de assimetria:

- Assimetria positiva: a cauda da curva tem viés para a direita, e a média é maior do que a mediana, que por sua vez é maior do que a moda.
- Assimetria negativa: a cauda da curva tem viés para a esquerda, e a média é menor do que a mediana, que por sua vez é menor do que a moda.
- Simétrica: a curva tem a forma de sino; o valor da assimetria é zero e a média, a mediana e a moda são iguais.

Geralmente é utilizado o coeficiente de assimetria de Pearson (CA), que tem por base as medidas de tendência central:

- CA = (média - moda) / desvio-padrão.

Tem-se a seguinte interpretação:

- CA = 0: a distribuição é simétrica: média = mediana = moda.
- CA > 0: a distribuição é assimétrica positiva: média > mediana > moda.
- CA < 0: a distribuição é assimétrica negativa: média < mediana < moda.

Tem-se as seguintes escalas de assimetria:

- $|CA| < 0{,}15 \Rightarrow$ assimetria pequena.
- $0{,}15 < |CA| < 1 \Rightarrow$ assimetria moderada.
- $|CA| > 1 \Rightarrow$ assimetria elevada.

Coeficiente de curtose

É uma medida do grau de achatamento de uma distribuição de valores em relação a uma distribuição simétrica como a da distribuição normal. Essa medida pode ser de três tipos:

- Mesocúrtica: quando os dados têm distribuição normal.
- Leptocúrtica: distribuição mais pontiaguda do que a normal.
- Platicúrtica: distribuição mais achatada do que a normal.

As medidas acima são obtidas por meio do coeficiente percentílico da curtose (C), que é função de Q_1, Q_3 , P_{10} e P_{90}:

$C = (Q_3 - Q_1)/2(P_{90} - P_{10})$
$C \cong 0{,}263 \rightarrow$ a curva é mesocúrtica.
$C < 0{,}263 \rightarrow$ a curva é leptocúrtica.
$C > 0{,}263 \rightarrow$ a curva é platicúrtica.

Exemplo 11

De duas fazendas independentes, foram obtidos os pesos de 100 bovinos da raça Nelore, com aproximadamente 2 anos de idade. Esses dados foram analisados por meio do procedimento *Means* do SAS, obtendo-se as seguintes estatísticas: número de observações (N), média (*mean*), desvio-padrão (Std Dev), assimetria (*skewness*), curtose (kurtosis), soma de quadrados não corrigida (USS), coeficiente de variação (CV), valor máximo da amostra (Max), terceiro quartil (Q_3), mediana (Med), primeiro quartil (Q_1), valor mínimo da amostra (Min), amplitude (*range*), diferença interquartílica ($Q_3 - Q_1$) e moda (*mode*) (Tabela 3).

Tabela 3. Dados de pesos de 100 bovinos da raça Nelore, com aproximadamente 2 anos de idade, obtidos de duas fazendas (A e B).

Estatística	Fazenda A	Fazenda B
Número de observações (N)	100	100
Média (*mean*)	388,3400	354,7800
Desvio-padrão (Std Dev)	108,0536	99,8336
Assimetria (*skewness*)	0,4893	0,9007
Curtose (kurtosis)	-1,1317	0,3111
Soma de quadrados não corrigida (USS)	16.236.678,0	13.573.592,0
Coeficiente de variação (CV)	27,8245	28,1396
Valor máximoda amostra (Max)	610,0000	669,0000
Terceiro quartil (Q_3)	492,0000	419,0000
Mediana (Med)	358,0000	324,0000
Primeiro quartil (Q_1)	295,5000	283,5000
Valor mínimo da amostra (Min)	237,0000	200,0000
Amplitude (*range*)	373,0000	469,0000
Diferença entre interquartílica ($Q_3 - Q_1$)	196,5000	135,0000
Moda (*mode*)	300,0000	297,0000

Pergunta-se:

a) Qual amostra era mais homogênea? Use as informações da média, da variância e do coeficiente de variação.

Para decidir qual amostra de dados de peso é a mais homogênea, considerando-se três informações − média, variância e coeficiente de variação, deve-se inicialmente analisar as duas médias. Sendo diferentes, como no presente caso ($\overline{X}_A$ = 388,34 kg versus $\overline{X}_B$ = 354,78 kg), a amostra mais homogênea é a que possui o menor coeficiente de variação, no caso é a da fazenda A (CV_A = 27,82% versus CV_B = 28,14%). Caso as duas médias fossem iguais, a amostra mais homogênea então seria a da fazenda B, pois é a que teria a menor variância (σ^2_A= 11.675,58 kg^2 versus σ^2_B= 9.966,74 kg^2).

b) Qual é o intervalo dos pesos de 50% dos animais que ocorre com maior frequência, ou seja, qual é a diferença interquartílica $Q_3 - Q_1$?

O intervalo de pesos de 50% dos animais que ocorre com maior frequência é dado pela diferença interquartílica ($Q_3 - Q_1$). Assim, para a fazenda A, tem-se Q_1 = 295,5 kg e Q_3 = 492,0 kg; logo, o intervalo 295,5 kg ≤ pesos ≤ 492,0 kg contém 50% desses animais; para a fazenda B, tem-se Q_1 = 283,0 kg e Q_3 = 419,0 kg, e o intervalo é dado por 283,0 kg ≤ pesos ≤ 419,0 kg. A diferença interquartílica ($Q_3 - Q_1$) para a fazenda A é 196,50 kg e para a fazenda B é 136,00 kg.

c) Qual é o intervalo de pesos de 25% dos animais inferiores de cada fazenda?

A amplitude de pesos de 25% dos animais inferiores corresponde ao intervalo dado pelo valor mínimo da amostra até o valor de posição 25° que corresponde ao primeiro quartil (Q_1). Assim, para a fazenda A, o valor mínimo é de 237,0 kg e o intervalo é dado por 237,0 kg ≤ pesos ≤ 295,5 kg, ou simplesmente pesos ≤ 295,5 kg; para a fazenda B, o valor mínimo é de 200,0 kg, e o intervalo é dado por 200,0 kg ≤ pesos ≤ 283,5 kg ou pesos ≤ 283,5 kg.

d) Descreva a amplitude dos pesos dos animais de cada fazenda.

A amplitude total de peso dos animais corresponde à diferença dos dois valores extremos da amostra, ou seja, 373,0 kg e 469,0 kg, para a fazenda A e B, respectivamente.

e) Descreva a média de pesos de cada fazenda com o seu respectivo erro-padrão.

A média de pesos e o respectivo erro-padrão são dados por $\overline{x} \pm s_{\overline{x}}$, em que $s_{\overline{x}} = \frac{s}{\sqrt{n}}$. Para a fazenda A, é 388,34 kg ± 10,80 kg e para a fazenda B, 354,78 kg ± 9,98 kg.

f) Qual amostra mais se aproxima da distribuição normal? Por quê?

A distribuição normal se caracteriza por possuir coeficientes de assimetria e de curtose iguais a zero e também média = moda = mediana. Pela análise dos coeficientes de assimetria e de curtose, não é possível decidir (para a fazenda A: 0,4893 e -1,1317; e para a fazenda B: 0,9007 e 0,3111). Entretanto, os valores, em ordem crescente, das três medidas de tendência central, média, moda e mediana, são 300,0 kg, 358,0 kg e 388,3 kg para a fazenda A, e 297,0 kg, 324,50 kg e 354,7 kg para a fazenda B. Verifica-se que os três valores para a última fazenda são mais uniformes, dando à amostra de pesos dessa fazenda uma distribuição mais simétrica.

g) Quais pesos estão influenciando mais a variância? Os menores ou os maiores?

As duas amostras possuem assimetria positiva, indicando que a distribuição de frequências tem viés para a direita, ou seja, a cauda da distribuição é mais alongada para a direita, em relação à situação em que a assimetria é nula. Assim, essas duas amostras possuem uma proporção de dados com valores maiores do que o esperado em uma distribuição normal. Conclui-se, portanto, que esses dados com valores maiores são os que estão influenciando mais a variância.

h) Comente os coeficientes de assimetria e de curtose.

As duas amostras possuem vieses para a direita, sendo o pico da distribuição na fazenda A mais achatado do que o da distribuição normal, caracterizando distribuição mesocúrtica; na fazenda B o pico é mais pontiagudo do que o da normal, caracterizando distribuição platicúrtica.

i) Considerando-se a média e o respectivo erro-padrão das duas amostras, qual delas possui maior erro?

A média e o respectivo erro-padrão da fazenda A são 388,34 ± 10,80 kg e da fazenda B, 354,78 ± 9,98 kg. Uma das maneiras de saber qual dessas médias possui maior erro percentual é dividir o erro-padrão pela respectiva média e multiplicar por 100. No caso da fazenda A é (10,80/388,34) x 100 = 2,78% e da fazenda B, (9,98/354,78) x 100 = 2,81%. Verifica-se que a média da fazenda B possui maior erro.

Exemplo 12

Considerando-se que, nas duas fazendas do exemplo 11, houve nova pesagem dos animais, em quilograma (kg), porém com amostra inferior a 100. Analisando-se os dados por meio do procedimento *Means* do SAS, foram obtidos os resultados apresentados na Tabela 4.

Tabela 4. Dados de peso de bovino da raça Nelore, obtidos de duas fazendas (A e B).

Estatística	Fazenda A	Fazenda B
Numero de observações (N)	40	50
Média (*mean*)	489,30	459,04
Desvio-padrão (Std Dev)	74,68	"5"
Assimetria (*skewness*)	-0,5734	-0,5668
Soma de quadrados não corrigida (USS)	"4"	"4"
Coeficiente de variação (CV)	"2"	22,46
Valor máximo da amostra (Max)	"3"	610,0
Terceiro quartil (Q_3)	545,00	"6"
Mediana (Med)	500,00	492,00
Primeiro quartil (Q_1)	450,00	381,00
Valor mínimo da amotra (Min)	255,00	"7"
Amplitude (*range*)	295,00	355,00
Diferença interquartílica ($Q_3 - Q_1$)	"1"	154,00

Com base na solução do exemplo 11, estimar as células faltantes (1 a 7).

Solução

- Célula "1": a diferença interquartílica é dada por Q_3 - Q_1 ou 545,00 kg - 450,00 kg = 95,0 kg.
- Célula "2": como CV = (100s)/$\bar{x}$, o cálculo do CV é direto, ou seja, CV = (100 x 74,68)/489,30 = 15,26%.
- Célula "3": amplitude total = valor máximo − valor mínimo $\Rightarrow$ (leia-se "implica") valor máximo = amplitude + valor mínimo = 295,00 kg + 255,00 kg = 550,00 kg.
- Célula "4": a soma de quadrados não corrigida (USS) é dada por $\sum_{i=1}^{n} x_i^2$. Com o uso da fórmula da variância $s^2 = [\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2]/(n-1)$ e ainda com o fato de que $\frac{\sum X_i}{n}$ é igual à média que também é conhecida por simples manipulação algébrica, obtém-se $\sum_{i=1}^{n} x_i^2$ = 9.793.306,5940 kg² (fazenda A) e $\sum_{i=1}^{n} x_i^2$ = 10.546.515,64 kg² (fazenda B).
- Célula "5": como CV = (100s)/$\bar{x}$, tem-se s = (CV$\bar{x}$)/100 = (22,46 x 459,04)/100 = 103,10 kg.
- Célula "6": como Q_3 - Q_1 = 154 $\Rightarrow$ Q_3 = Q_1 + 154 kg = 381,00 kg + 154,00 kg = 535,00 kg.

- Célula “7”: amplitude total = valor máximo − valor mínimo ⇒ valor máximo − amplitude = 610,00 kg - 355,00 kg = 255,00 kg.

Exercícios[1]

1) No texto a seguir, existem afirmações incorretas quanto a conceitos de estatística. Reescreva o texto colocando definições corretas e sublinhe ou coloque em negrito onde houve definições incorretas.

Estatísticas descritivas são usadas para descrever as características dos dados. Várias são as técnicas usadas para classificar dados: descrição gráfica, descrição tabular e sumários estatísticos, entre outras. É imprescindível compreender alguns conceitos: parâmetros e estatísticas: uma população é caracterizada por parâmetros, que geralmente são conhecidos, os quais são funções de valores populacionais; as estatísticas, porém, são funções de valores amostrais e sempre representam com precisão os parâmetros; dados qualitativos: representam a informação que identifica alguma qualidade, categoria ou característica, que não é susceptível de classificação, mas de medida; dados quantitativos: representam a informação resultante de características susceptíveis de serem medidas, podendo ser de natureza discreta ou contínua e dividem-se em: variável discreta − que é constituída de partes ou categorias separadas e distintas, e variável contínua − que somente pode assumir um conjunto ordenado de valores inteiros dentro de determinado limite ou intervalo; estudo transversal − permite uma situação momentânea do fenômeno investigado; o estudo é relativamente rápido, consome menos recursos e é menos vulnerável a variáveis estranhas; estudo longitudinal − os dados de cada indivíduo são coletados em dois ou mais momentos, possibilitando acompanhamento do desenrolar do fenômeno considerado. Porém, é mais complicado que o estudo transversal e não é recomendado para a pesquisa.

2) Na Tabela 5 são apresentados os valores da média, da mediana, do erro-padrão da média e do coeficiente de variação de dados de produtividade de matéria seca (PMS), kg ha^{-1} de 92 cultivares de alfafa (*Medicago sativa* L.) avaliados em cinco cortes. Cada estatística foi calculada duas vezes (com e sem *outliers*).

A existência de *outliers* nos cortes 2, 4 e 5 prejudicou a qualidade dos dados? Por quê?

[1] As respostas dos exercícios podem ser consultadas no Apêndice 1.

Tabela 5. Média, mediana, erro-padrão da média e coeficiente de variação de dados de produtividade de matéria seca (em kg ha^{-1}) de 92 cultivares de alfafa avaliadas em cinco cortes.

Corte	Média		Mediana		Erro-padrão da média		Coeficiente de variação	
	Com outlier	Sem outlier	Com outlier	Sem outlier	Com outlier	Sem outlier	Com outlier	Sem outlier
1		2.248,4		2.256,9		31,80		19,2
2 (2)+	2.730,5	2.757,9	2.757,8	2.762,6	35,69	30,30	17,7	14,8
3		2.041,4		2.035,1		29,89		19,8
4 (1)	1.266,6	1.273,1	1.250,8	1.250,9	21,79	20,93	23,3	22,2
5 (1)	1.222,1	1.213,1	1.227,02	1.225,6	23,42	21,73	26,0	24,2

"+" = Número de *outliers* por corte, dentro do parêntese.
Fonte: Freitas et al. (2008).

3) Com relação à Tabela 5, discutir a homogeneidade dos dados de cada corte com relação ao coeficiente de variação.

4) Com base na média e na mediana da Tabela 5, discuta os coeficientes de simetria dos dados de cada corte.

5) A Tabela 6 refere-se à técnica de análises exploratórias aplicadas à nove pesagens de dados de coelhos (P_0 = peso ao nascimento; P_1 a P_8 = pesagens posteriores).

Tabela 6. Análises exploratórias aplicadas a nove pesagens de dados de coelhos (P_0 = peso ao nascimento; P_1 a P_8 = pesagens posteriores).

Coeficiente	Po	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8
Assimetria	0,30	0,69	0,46	0,45	0,82	0,90	0,92	1,00	1,07
Curtose	3,02	4,04	0,70	0,29	0,76	1,40	1,32	1,38	1,63
Locação(*)	1,2 = 3	2,3,1	2,3,1	1,2,3	2,3,1	2,3,1	3,1,2	3,1,2	2,3,1

(*)Locação = indica a ordem de ocorrência (1= média 1; 2 = moda; 3 = mediana).

Interpretar as análises exploratórias com relação à distribuição normal.

Capítulo 2

Tabulação e cálculos

Introdução

O objetivo deste capítulo é mostrar algumas etapas básicas utilizadas na análise de variância (Anova) de um experimento, iniciando-se com a elaboração do arquivo de dados e a combinação de recursos matemáticos e estatísticos usados na análise dos dados. Da matemática, utilizam-se equações, funções, fórmulas quadráticas, derivadas, integrais e cálculo diferencial, matrizes, vetores, distribuição de probabilidades contínuas, modelos não lineares, etc. Grande parte dessas ferramentas são utilizadas por meio de modelos lineares; no entanto, os não lineares são bastante usados, principalmente em curvas de crescimento. Desses recursos, a teoria da probabilidade é fundamental; é utilizada na quantificação da aleatoriedade e incerteza de eventos na natureza. Já a estatística experimental é compreendida como a ciência da coleta, descrição e análise de dados; trabalha com métodos que auxiliam na tomada de decisões diante da incerteza e tem por base a variabilidade ou erro entre as variáveis. Inicia-se com uma amostra de variáveis aleatórias $x_1, x_2, \ldots, x_n$ e, a partir desta, todos os cálculos são realizados.

São ilustrados, neste capítulo, elaboração do arquivo de dados de um experimento; distribuição de frequências, cálculos de somas de quadrados, matrizes, vetores, momentos estatísticos e distribuição de probabilidades discretas e contínuas na Anova, bem como uso de modelos não lineares.

Elaboração do arquivo de dados de um experimento

Uma etapa fundamental de um experimento de campo, de um ensaio em laboratório e de qualquer pesquisa, de modo geral, é a coleta de dados e a organização do arquivo de dados, o qual deve ser bem documentado, de modo que as análises estatísticas possam ser devidamente executadas e que os dados e as informações deste arquivo possam ser recuperados sempre que necessário. Dependendo da instituição e da finalidade da pesquisa, a estrutura de um arquivo de dados pode variar. Na Embrapa, por exemplo, na sua estrutura programática de pesquisa, tem-se o projeto de pesquisa, que se divide em planos de ação, e estes, por sua vez, dividem-se em atividades (AT).

Na documentação do arquivo, é importante mencionar o título da pesquisa ou projeto, o responsável, etc. Sempre que possível, devem estar claramente descritos os seus objetivos, o delineamento experimental, a unidade experimental ou parcela com o respectivo tamanho, a subparcela, quando for o caso.

Na descrição dos tratamentos, deve-se informar exatamente o que foi utilizado, os correspondentes níveis, suas nomenclaturas, etc., evitando-se informações como tratamentos A, B, C, etc. Na descrição das variáveis a serem analisadas, devem-se informar as respectivas unidades e as novas variáveis que serão geradas a partir das variáveis originais. É fundamental informar também local e data da realização do experimento.

A estruturação do arquivo de dados será demonstrada utilizando-se dados de um experimento realizado na Embrapa Pecuária Sudeste, São Carlos, SP, cujo objetivo foi avaliar o efeito de adubação nitrogenada sobre a produtividade de forragem em condição não irrigada. O delineamento experimental foi o de blocos casualizados (DBC), com três repetições e parcelas divididas. Nas parcelas principais, foram distribuídos aleatoriamente 20 tratamentos em esquema fatorial 5 x 4 (cinco espécies de capim: Tanzânia, *Brachiaria*, Marandu, Pojuca e *Coast-cross*, e quatro doses de nitrogênio em cobertura: 0 kg ha^{-1}, 20 kg ha^{-1}, 40 kg ha^{-1} e 60 kg ha^{-1}). Nas subparcelas, foram considerados seis cortes realizados no tempo, totalizando-se 360 observações. A teoria do delineamento de blocos casualizados (DBC) está apresentada no Capítulo 10.

A elaboração do arquivo de dados do experimento em DBC descrito anteriormente será feita com a rotina do Statistical Analysis System (SAS). Nas partes que se referem à documentação e comentários, o procedimento inicia-se com "/*" e termina com "*/"; na sequência tem o arquivo de dados "forragem" colocado após "data" e termina com ";".

```
/* informações de identificação
projeto:          < título do projeto >
coordenador:      < nome do coordenador do projeto >

objetivo:
avaliar o efeito da adubação nitrogenada sobre a produtividade de forragem.

delineamento experimental
blocos casualizados, três repetições e parcelas subdivididas.
tratamento = 20 (organizados em esquema fatorial 5x4)
5 espécies de capim: 1. tanzânia, 2. brachiaria, 3. marandu, 4. pojuca e 5. coast-cross
4 doses de adubo: nitrogênio em cobertura, kg/ha: (0 − 20 - 40 − 60)
parcela = área com 5 m de comprimento x 1 m de largura
subparcela = seis cortes consecutivos realizados mensalmente (1,2,3,4,5,6)

variável avaliada
y = rendimento em tonelada de matéria seca, t/ha
*/
```

```
data forragem;
input bloco capim dose corte y;
x = dose;
x2= x*x; x3 =x2*x; x4 =x3*x; y2 =y*y; y3 =y2*y; y4 =y3*y; xy =x*y;
datalines;
 1     1     0     1     2.9
 1     1    20     1     2.9
...
 3     5    40     6     0.9
 3     5    60     6     1.5
;
```

Distribuição de frequências

Conhecer a distribuição de frequências dos dados é uma das atividades iniciais da análise estatística. Para isso, é organizada uma tabela que mostra a distribuição dos valores das variáveis, após sua ordenação, em ordem crescente. Na Tabela 1 são apresentadas as classes, frequências absolutas, frequências absolutas acumuladas, frequências relativas e frequências relativas acumuladas da variável produtividade de forragem, em tonelada por hectare (t ha^{-1}) de matéria seca, calculadas a seguir.

Para um conjunto de dados com *n* observações em ordem crescente, o número de classes (k) é obtido pela fórmula de *Sturges*:

$k = 1 + 3{,}322\log_{10}(n)$.

Para $n = 360$, $k = 1 + 3{,}322 \times (2{,}55630) = 9{,}49204 \approx 10$ classes.

A amplitude de cada classe (AC) é dada por:

AC = (amplitude dos valores da amostra)/número de classes.

Como o valor mínimo e máximo de *y*, em t ha^{-1}, foi respectivamente 0,50 e 8,8 então foi escolhida AC = 0,8.

Na pesquisa, quando se deseja estudar uma variável ou característica, a primeira preocupação de um pesquisador deve ser a organização dos dados em uma tabela de frequências, como a Tabela 1. A distribuição de frequências dos dados de um fenômeno aleatório ou casual, seguida da elaboração de gráficos, é um recurso poderoso para entender a natureza da variabilidade dos dados. Assim, um interesse imediato é a construção de um histograma ou gráfico de barra, como o representado na Figura 1. Os gráficos são as formas mais eficientes para mostrar a estrutura dos dados e possibilitar refinamentos metodológicos, além de visualizar ajuste de modelos na Anova.

Tabela 1. Classes e frequências da variável produtividade de forragem, em tonelada por hectare (t ha^{-1}) de matéria seca.

Classe (t ha^{-1})	Frequência absoluta	Frequência absoluta acumulada	Frequência relativa	Frequência relativa acumulada
1 (0,4–1,2)	22	22	6,11	6,11
2 (1,2–2,0)	62	84	17,22	23,33
3 (2,0–2,8)	92	176	25,56	48,89
4 (2,8–3,6)	88	264	24,44	73,33
5 (3,6–4,4)	44	308	12,22	85,56
6 (4,4–5,2)	30	338	8,33	93,89
7 (5,2–6,0)	13	351	3,61	97,50
8 (6,0–6,8)	4	355	1,11	98,61
9 (6,8–7,6)	3	358	0,83	99,44
10 (> 7,6)	2	360	0,56	100,00

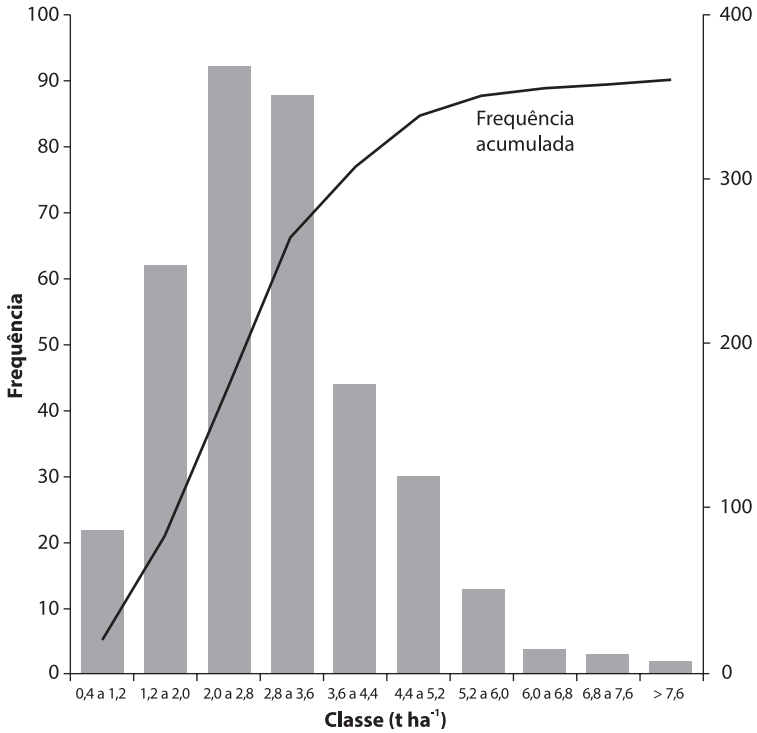


Figura 1. Histograma de frequência absoluta por classe (t ha^{-1}), de produtividade de forragem de matéria seca.

Em síntese, um estudo preliminar ou exploratório possibilita investigar características latentes nos dados e sugerir possíveis modelos para uma análise de variância.

Considerando-se que toda investigação e todo avanço do conhecimento parte da identificação de um problema, questão ou suposição, que é a hipótese científica, a sequência de atividades fica:

- Análise clássica:
  Problema => dados => modelo => análises => conclusões.
- Análise exploratória:
  Problema => dados => análises => modelo => conclusões.

Cálculos de somas de quadrados na Anova

Após um estudo inicial de estatística descritiva, o passo seguinte na análise dos dados de um experimento é realizar os diversos cálculos associado a uma Anova, assunto dos Capítulos de 9 a 11, após escolher um modelo matemático linear padrão ou linear misto de acordo com o delineamento experimental.

A forma mais simples de expressar um modelo na Anova é do tipo y = x´b+ ε, em que a variável dependente, de efeito ou de resposta (y), pode ser escrita como a soma de duas partes, uma envolvendo os efeitos fixos (x´b) que é função linear dos coeficientes independentes, que variam para cada experimento, e a outra que é o erro (ε).

Geralmente a análise de variância dos dados coletados de um experimento é feita a partir do cálculo de somas de quadrados de uma variável e soma de produto envolvendo duas variáveis. Como exemplo, para a análise dos dados do arquivo “data forragem”, se considerarmos a variável dependente *y* e a variável preditora *x* (dose), alguns tipos de somatório efetuados são:

$$\sum_{i=1}^{360} x_i = 0 + 20 + 40 + 60 + ... + 60 = 10.800,0$$

$$\sum_{i=1}^{360} y_i = 4,0 + 4,8 + ... + 1,5 = 1.052,8$$

$$\sum_{i=1}^{360} x_i^2 = 0^2 + 20^2 + 40^2 + 60^2 ... + 60^{2.} = 504.000,0$$

$$\sum_{i=1}^{360} y_i^2 = 4,0^2 + 4,8^2 + ... + 1,5^2 = 3.707,5$$

$$\sum_{i=1}^{360} x_i y_i = 0x4{,}0 + \ldots + 60x1{,}5 = 33.364{,}0$$

A partir de somatórios como estes, para a execução de uma Anova, são calculadas diversas somas de quadrados (SQ), como a SQ total corrigida e a SQ total não corrigida para a variável dependente *y*, que são fundamentais para a Anova, para dividir a variabilidade total nos dados, em componentes devidos a efeitos de tratamentos e a variabilidade devida ao acaso, conceitos que serão estudados nos Capítulos 7 e 8.

Exemplos:

SQ não corrigida = $\sum_{i=1}^{n} y_i^2 = 3.707{,}54$

SQ corrigida = $\sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2 n = 628{,}68$

em que $(\sum_{i=1}^{n} y_i)^2/n = 3.078{,}85$ é o fator de correção.

Para ilustrar melhor a necessidade desses cálculos, tem-se que em uma amostra aleatória com n elementos, $x_1, x_2, \ldots, x_n$, o termo $(x_i - \bar{x})$ representa o desvio ou erro; o termo $\sum_{i=1}^{n} (x_i - \bar{x})$ representa o somatório dos erros ou desvios. Porém, para evitar que o somatório resulte em zero, cada erro ou desvio é elevado ao quadrado, o que resulta na expressão $\sum_{i=1}^{n} (x_i - \bar{x})^2$, que dá origem à soma de quadrados corrigida.

Dando sequência à Anova, são calculadas as somas de quadrados do efeito principal de um fator (*y:* produtividade de forragem) e a interação entre dois fatores (capim e dose). Na Tabela 2, cada célula tem a soma de *y* para a respectiva dose e capim, com os respectivos totais marginais nas linhas e colunas.

Tabela 2. Produtividade de forragem, em tonelada por hectare (t ha^{-1}) de matéria seca. Soma para cada dose e capim.

Capim	Dose 0	Dose 20	Dose 40	Dose 60	Total
Capim 1	55,7	56,9	58,1	64,8	235,5
Capim 2	45,0	50,8	57,3	58,4	211,5
Capim 3	41,4	52,6	54,4	52,6	201,0
Capim 4	56,0	79,6	57,6	69,1	262,3
Capim 5	29,1	31,6	38,2	43,6	142,5
Total	227,2	271,5	265,6	288,5	1.052,8

Tem-se:

a) Somas de quadrados para dose e capim:

$$S.Q.Doses = \frac{1}{90}[227,2^2 + 271,5^2 + 265,6^2 + 288,5^2] - \frac{(1.052,8)^2}{360} =$$

$$S.Q.Doses = \frac{1}{90}[279.107,70] - \frac{(1.052,8)^2}{360} = 3.101,1967 - 3078,8551 = 22,3416$$

$$S.Q.Capim = \frac{1}{72}[235,5^2 + 211,5^2 + 201,0^2 + 262,3^2 + 142,5^2] - \frac{(1.052,8)^2}{360} =$$

$$S.Q.Capim = \frac{1}{72}[229701,0400] - \frac{(1.052,8)^2}{360} = 3190,2922 - 3078,8551 = 111,4371$$

b) Somas de quadrados para a interação dose × capim:

$$S.Q.\ Doses, Capim = \frac{1}{18}[\ 55,7^2 + 56,9^2 + \ldots + 38,2^2 + 43,6^2] - \frac{(1.052,8)^2}{360}$$

$$S.Q.\ Doses, Capim = \frac{1}{18}[58191,38] - \frac{(1.052,8)^2}{360} = 3232,8544 - 3078,8551 = 153,9993$$

S.Q. Interação Doses x Capim = S.Q.Doses, Capim - S.Q.Doses - S.Q.Capim

S.Q. Interação Doses x Capim = 153,9993 - 22,3416 - 111,4371 = 20,2206

Matrizes e vetores na Anova

A matriz é a estrutura matemática mais apropriada para trabalhar com um arquivo de dados. Um arquivo é uma matriz em que cada valor pertence a uma linha e uma coluna, o que significa que ele tem um endereço, e os dados de uma linha ou de uma coluna representam um vetor da matriz. A maioria da programação e cálculos necessários para a estatística experimental são obtidos em termos de álgebra de matriz. Um exemplo típico são as técnicas estatísticas multivariadas de fundamental importância nas pesquisas biológicas em que os dados são coletados nas várias dimensões para um mesmo indivíduo.

Em um modelo linear de análise de variância, as equações normais são do tipo y = xb + e, que, utilizando-se vetores e matrizes, resulta em:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_{10} & x_{11} & \ldots & x_{1p} \\ x_{20} & x_{21} & \ldots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n0} & x_{n1} & \ldots & x_{np} \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

em que:

y = vetor de variáveis de respostas.

x = matriz de informação.

b = vetor de parâmetros.

e = vetor de erros.

n = número de unidades experimentais.

p = número de variáveis envolvidas na análise.

As estimativas dos parâmetros do vetor b são obtidas do sistema de equações:

$$\hat{b} = (x'x)^{-1}x'y$$

O produto matricial $x'x$, por exemplo, envolve vários cálculos:

$$x'x = \begin{bmatrix} n & \sum x_2 & \sum x_{2\ldots} & \sum x_m \\ \sum x_1 & \sum x_1^2 & \sum x_1 x_{2\ldots} & \sum x_1 x_m \\ \sum x_2 & \sum x_2 x_1 & . & \sum x_2 x_m \\ . & . & . & . \\ \sum x_m & \sum x_m x_1 & \sum x_m x_{2\ldots} & \sum x_m^2 \end{bmatrix}$$

Quando a situação experimental fica mais complexa como nos modelos lineares mistos e análises de medidas repetidas no tempo, como os descritos no Capítulo 9, tem-se que trabalhar com sistemas matriciais mais complexos:

$$y = xb + zu + e$$

em que:

$$\text{Var}(u) = G;\ \text{Var}(e) = R;\ \text{Var}(y) = zGz' + R = V.$$

Para estimar os efeitos $\hat{b}$ e aleatórios $\hat{u}$, tem-se:

$$\hat{b} = (x'V^{-1}x)x'V^{-1}y$$

$$\hat{u} = (x'z'V^{-1})(y - x\hat{b})$$

Momentos estatísticos

Para uma amostra $x_1, x_2, \ldots, x_n$, o p-ésimo momento amostral até a quarta ordem (p = 4) centrado na média é dado por:

$$m_p = 1/n \sum_{i=1}^{n} (x_i - \bar{x})^p$$

Até a quarta ordem, são dados por:

$$m_1 = 1/n \sum_{i=1}^{n} (x_i - \bar{x})^1 = \text{média centrada}$$

$$m_2 = 1/n \sum_{i=1}^{n} (x_i - \bar{x})^2 = \text{variância}$$

$$m_3 = 1/n \sum_{i=1}^{n} (x_i - \bar{x})^3$$

$$m_4 = 1/n \sum_{i=1}^{n} (x_i - \bar{x})^4$$

Como $x_1, x_2, \ldots x_n$ são amostras aleatória da população, então a esperança do k-ésimo momento amostral é igual ao k-ésimo momento populacional, ou seja:

$$E(m_p) = \mu_p$$

Os momentos que originam a média centrada e a variância são conhecidos como momentos de baixa ordem. Os momentos de ordem p = 3 e p = 4 originam os coeficientes de assimetria e de curtose (Tabela 3).

Tabela 3. Coeficientes de assimetria e de curtose para uma população e amostra.

Coeficiente	População	Amostra
Coeficiente de assimetria	$\gamma_1 = \mu_3/\mu_2^{3/2}$	$g_1 = m_3/m_2^{3/2}$
Coeficiente de curtose	$\gamma_2 = \mu_4/\mu_2^2 - 3 = \mu_4/\sigma^4 - 3$	$g_2 = m_4/m_2^2 - 3 = m_4/m_2^2 - 3$

Distribuição de probabilidades discretas e contínuas na Anova

As distribuições de probabilidades discretas são descritas no Capítulo 5, e, na experimentação, as mais importantes são a binomial e a Poisson. Entretanto, a maioria das variáveis estudadas na estatística experimental é de natureza aleatória e contínua, cujas distribuições de probabilidade contínua são descritas no Capítulo 6. Para o cálculo dessas distribuições requer-se o uso de uma integral, ferramenta matemática geralmente usada para determinar a área sob uma curva no plano cartesiano (*x*, *y*).

Admitindo-se que *x* seja uma variável aleatória contínua e assume valor no intervalo de *a* até *b* (a < b), então a distribuição de probabilidade para a variável aleatória contínua *x,* denominada de função de densidade de probabilidade (FDP), é dada pela área sob a curva limitada por *a* e *b*, em que f(x) é a função de densidade de probabilidade de *x*:

$$P(a < x < b) = \int_a^b f(x)\, dx$$

Para uma amostra aleatória de dados, x_1, ..., x_n, assumindo qualquer valor dentro do intervalo de $-\infty < x < \infty$, podemos afirmar que a soma das probabilidades desses valores é igual a 1, conforme mostra a expressão:

$$P(-\infty < x < \infty) = \int_{-\infty}^{\infty} f(x)dx = 1$$

Uma FDP de uma variável aleatória contínua pode ser compreendida como a relação matemática que fornece, para cada valor da variável, o somatório das probabilidades de todas as ocorrências até aquele ponto.

Um exemplo é o gráfico da Figura 2, em que no eixo *x* são expressos os valores da variável aleatória, em ordem crescente, da produtividade de forragem em tonelada por hectare (t ha^{-1}) de matéria seca, dos dados apresentados na Tabela 1; e no eixo *y* é expresso o valor da sua função de distribuição, sendo que a área vale 1.

O interesse em localizar variáveis contínuas que se ajustam à distribuição normal é que as informações latentes nos dados podem ser facilmente interpretáveis por meio de uma curva e apenas dois parâmetros: média e variância. Também é comum falar e considerar que as variáveis contínuas têm distribuição normal.

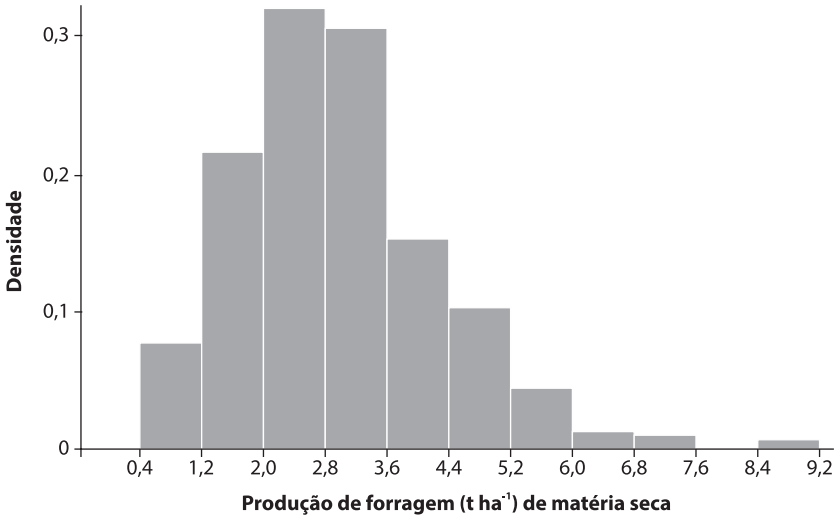


Figura 2. Função de distribuição de probabilidade da produtividade de forragem.

Quando a distribuição f(x) é normal ou gaussiana, ela é perfeitamente caracterizada pela média μ e pela variância σ^2, que são exemplos de momentos de ordem 1 e 2 respectivamente, centrados na média, os quais serão descritos neste capítulo.

$$\mu = E(x) = \int_{-\infty}^{\infty} x\, f(x)dx$$

$$\sigma^2 = E[x - \mu]^2 = \int_{-\infty}^{\infty} (x - \mu)^2\, f(x)dx$$

Utilizando-se o arquivo Data Forragem no início deste capítulo, um resumo das estatísticas descritivas da variável *y* é obtido por meio do procedimento *means* do SAS (Tabela 4).

```
proc means data = forragem;
n min max range sum mean var std stderr cv uss css q1 median q3 qrange kurt skew;
var y;
run;
```

Tabela 4. Estatísticas associadas à variável *y*.

Estatística	y
n = número de observações da amostra	360
min = menor valor da amostra	0,50
max = maior valor da amostra	8,80
range = amplitude total	8,30
sum = soma dos valores da variável	1.053
mean = média	2,92
var = variância	1,75
std = desvio-padrão	1,32
stderr = erro-padrão da média	0,07
cv = coeficiente de variação	45,25
uss = soma de quadrados não corrigida	3.708
css = soma de quadrados corrigida	628,70
q_1 = primeiro quartil	2,00
median = mediana	2,80
q_3 = terceiro quartil	3,66
qrange = amplitude interquartílica (q_3 - q_1)	1,65
kurt = curtose	1,63
skew = assimetria	0,90

Usos de modelos não lineares

Considerando-se um modelo qualquer em que a variável dependente *y* é função de variáveis independentes *x*, geometricamente a derivada representa a inclinação de uma curva de $y = f(x)$ e fisicamente representa a taxa de variação desta. A derivada é uma ferramenta matemática importante na estatística experimental e tem grande aplicação para representar a taxa de crescimento em um modelo da forma $y = f(t)$, em que *t* é o tempo.

A derivada de *y* com relação a *t* é representada por ∂_y/∂_t e, nesse caso, a variação média ou a taxa de crescimento entre dois tempos t_1 e t_2 é dada por:

$$\frac{\partial_y}{\partial_t} = \frac{f(t_2) - f(t_1)}{t_2 - t_1}$$

A taxa de variação de uma função f num ponto t_i qualquer é denotada pela derivada de f em relação a t_i ou $f'(t_i)$. Assim, $f'(t_1)$ é a derivada de f com relação a t para $t = t_1$, em que $f'(t_1)$ é a taxa instantânea da variação de f no ponto t_1.

Todos os cálculos anteriores a esse item foram realizados por meio de modelos lineares. No entanto, na estatística experimental, os modelos não lineares, geralmente utilizados como curvas de crescimento, são de grande importância.

Em um modelo não linear, os dados são modelados por uma função que é uma combinação não linear dos parâmetros. Para tanto, vamos considerar o estudo de curvas de crescimento em que a variável dependente y representa o crescimento dos organismos vivos em função da idade t.

No exemplo, vamos utilizar dois modelos não lineares bastante utilizados na estimativa de crescimento de animais:

$$\text{Brody:} \quad y_t = A(1 - be^{-kt}) + \varepsilon_t$$

$$\text{Von Bertalanffy:} \quad y_t = A(1 - be^{-kt})^3 + \varepsilon_t$$

em que:

y_t = peso estimado do animal na idade t (em dias, semanas, meses, etc.).

A = estimativa do peso limite do animal ou peso assintótico.

b, k = parâmetros que modelam a curva de crescimento.

e = base do logaritmo neperiano ($e = 2{,}71828...$).

ε_t = erro aleatório.

Esses modelos foram ajustados a dados de pesos de bovinos Nelore, criados no bioma Amazônia, do nascimento até 750 dias de idade, de Marinho et al. (2013). Para cada animal foram utilizados nove pares peso-idade. A primeira pesagem feita no nascimento e as demais em intervalos de 3 meses cada. Os modelos estimados ficaram:

$$\text{Brody:} \quad \hat{y}_t = 384{,}60\ (1 - 0{,}9192e^{-0{,}0022t}) + \varepsilon_t$$

$$\text{Von Bertalanffy:} \quad \hat{y}_t = 313{,}41(1 - 0{,}5153e^{-0{,}0045t})^3 + \varepsilon_t$$

Na Figura 3 é apresentada a estimativa de crescimento, em quilograma (kg), de bovinos Nelore e a taxa de crescimento absoluta instantânea (TCI), que é obtida pela derivada da variável dependente (y_t) em relação ao tempo t, em dias, ($\partial y_t / \partial t$).

Para essas duas funções, a expressão $\partial y_t / \partial t$ tem a forma:

Brody: $\partial y_t/\partial t = Abke^{-kt}$

Von Bertalanffy: $\partial y_t/\partial t = 3Abke^{-kt}(1- be^{-kt})^2$

Observa-se que a média da taxa de crescimento do rebanho (Figura 3B), do nascimento até 750 dias de idade, variou, aproximadamente, de 0,80 kg até 0,15 kg dia^{-1}.

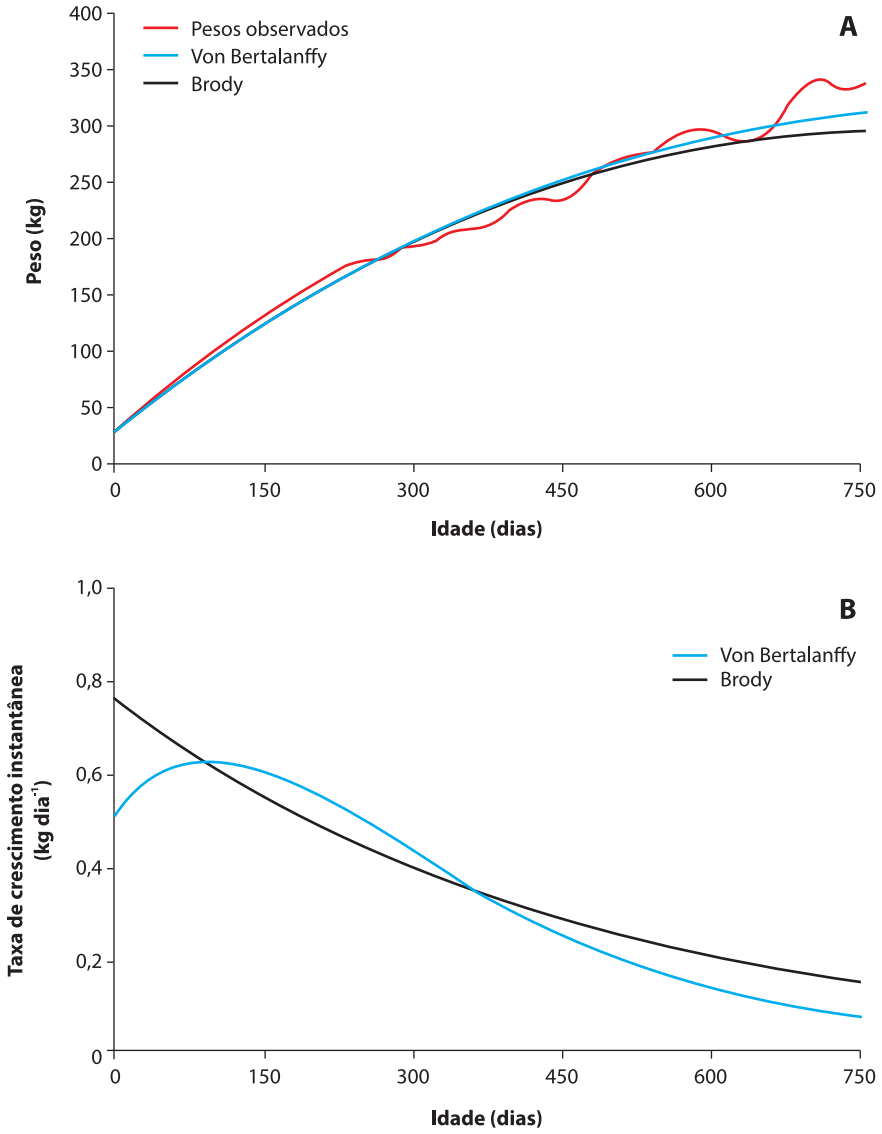


Figura 3. Estimativa de crescimento de bovinos Nelore por meio dos modelos Brody e Von Bertalanffy e pesos observados (A). Taxa de crescimento instantânea (TCI) obtida por esses modelos (B).

Fonte: Marinho et al. (2013).

Desses modelos não lineares, pode ser calculada a taxa de maturidade absoluta (TMA), que é a TCI dividida pelo peso assintótico A:

$$TMA = (\partial y_t/\partial t)A^{-1}$$

Para um intervalo de tempo t_i e t_j (i < j), a média de ganho de peso por dia é calculada por $(y_j - y_i)/(t_j - t_i)$, enquanto o incremento no peso para cada unidade de tempo *t* é calculado por: $A^{-1}(y_j - y_i)/(t_j - t_i)$.

Um interesse nesses estudos é calcular o tempo necessário para o animal atingir 50%, 70%, 90% do peso adulto. Isso equivale a ter, respectivamente, $y_{t/A} = 0,5$; $y_{t/A} = 0,7$; $y_{t/A} = 0,9$. Considerando-se que o valor de A para Brody é 384,6 kg, tem-se que os pesos, em quilograma (kg), nesta ordem, são: 192,3 kg; 269,2 kg e 346,1 kg.

Para saber quais são as idades para que o animal atinja esses pesos, pode-se calcular o tempo *t* pela fórmula.

$$t = -\log_e\left[\frac{1 - \frac{y_t}{A}}{b}\right]k^{-1}$$

Assim, o animal atinge 50%, 70% e 90% do peso adulto com 276, 509 e 1.008 dias, respectivamente. Os dois primeiros resultados podem ser facilmente vistos na Figura 3B; entretanto, como para atingir 90% do peso adulto são necessários 1.008 dias, esse resultado não pode ser visto na Figura 3, pois foram considerados até a idade de 750 dias, fase em que os animais estavam ainda em crescimento.

Exercícios[2]

1) No texto a seguir, existem afirmações incorretas quanto a conceitos de estatística. Reescreva o texto colocando definições corretas e sublinhe ou coloque em negrito onde houve definições incorretas.

Geralmente, os dados coletados de um experimento de campo, de um ensaio em laboratório e de qualquer pesquisa de modo geral dão origem a um arquivo de dados. É importante documentar esse arquivo, para que suas informações possam ser recuperadas sempre que necessário. Independentemente da instituição e/ou finalidade da pesquisa, a documentação é padrão. Independentemente do número de variáveis

[2] As respostas dos exercícios podem ser consultadas no Apêndice 1.

medidas em um experimento, a análise estatística é realizada a partir do cálculo de somas de quadrados de uma única variável e soma de produto envolvendo duas variáveis. O termo $\sum_{i=1}^{n} (x_i - \bar{x})$ representa o somatório dos erros ou desvios; porém, para evitar que o somatório resulte em zero, cada termo é elevado ao quadrado, resultando na expressão da variância ou soma de quadrados corrigida: $\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^{2/n}$, em que $(\sum_{i=1}^{n} x_i)^{2/n}$, é o conhecido fator de correção. Com relação aos momentos estatísticos de ordem p (p = 1, p = 2, p = 3 e p = 4), os mais utilizados são os de segunda e terceira ordens, que originam, respectivamente, a média e a variância, que são conhecidos como momentos de baixa ordem. Geralmente são utilizados os momentos até a quarta ordem. Todos os momentos, até a quarta ordem, são utilizados nos cálculos de coeficientes de curtose e assimetria.

2) A distribuição de frequências (Tabela 5) mostra o peso médio do ovo (g) de galinha. Organizar uma nova tabela considerando-se três categorias:

a) Ovo pequeno, peso < 49,3g.

b) Ovo médio, 49,3g ≤ peso < 55,5g.

c) Ovo grande, peso ≥ 55,5g.

Tabela 5. Distribuição de frequências para peso médio do ovo (g) de galinha.

Classe	Frequência	Frequência relativa (%)	Frequência acumulada (%)
40,1 – 43,1	2	5,0	5,0
43,2 – 46,2	3	7,5	12,5
46,3 – 49,3	8	20,0	32,5
49,4 – 52,4	10	25,0	57,5
52,5 – 55,5	8	20,0	77,5
55,6 – 58,6	5	12,5	90,0
58,7 – 61,7	4	10,0	100,0

3) Com o objetivo de avaliar a produtividade e a qualidade da forragem de capim *Coast-cross*, foi realizado um delineamento experimental em blocos casualizados, com quatro repetições (rep) e parcelas subdivididas. Nas parcelas, foram distribuídos 10 tratamentos organizados em esquema fatorial 2 x 5 [duas fontes de nitrogênio (N): ureia e nitrato de amônio e cinco doses de N: 0 kg ha^{-1}, 25 kg ha^{-1}, 50 kg ha^{-1}, 100 kg ha^{-1}, 200 kg ha^{-1} por corte] em cinco cortes consecutivos (corte) na subparcela.

As variáveis colhidas em cada parcela foram:

Prod = produtividade da área útil da parcela (g).

Gaf = amostragem de material fresco colhido na parcela (g).

Gas = amostra de material seco (g).

Pb = proteína bruta (%).

Fdn = fibra em detergente neutro (%).

Div = digestibilidade in vitro da matéria seca (%).

No_3 = nitrato solúvel na biomassa (mg kg^{-1}).

N = nitrogênio (g kg^{-1}).

Considerando-se apenas as duas primeiras observações ou linhas de dados (Tabela 6), descreva uma rotina SAS mostrando a descrição do experimento e o arquivo de dados.

Tabela 6. Dados de produtividade e de qualidade de forragem de capim *Coast-cross* colhidos nas duas primeiras observações de um experimento.

Rep	Fonte	Dose	Corte	Prod	Gaf	Gas	Pb	Fdn	Div	No_3	N
1	1	0	1	51,0	48,9	13,0	7,4	82,9	58,2	0,0	10,7
1	1	0	2	600,0	448,1	105,2	10,3	86,5	58,2	33,3	15,5
...											

Prod: produtividade da área útil da parcela (g); Gaf: amostragem de material fresco colhido na parcela (g); Gas: amostra de material seco (g); PB: proteína bruta (%); Fdn: fibra em detergente neutro (%); div: digestibilidade in vitro da matéria seca (%); No_3: nitrato solúvel na biomassa (mg kg^{-1}); N: nitrogênio (g kg^{-1}).

4) A seguir está apresentado o ganho de peso médio diário, em grama (y) e o consumo de ração médio diário, em grama (x) para uma amostra aleatória de cinco lotes de aves, no período de 1 a 28 dias.

y (g)	871	772	800	904	843
x (g)	1.415	1.295	1.328	1.510	1.399

A partir desses dados foram elaborados os cálculos a seguir. Alguns estão errados. Identifique os cálculos com erros.

$\sum x = 1.415 + ... + 1.399 = 6.947$

$\sum x^2 = 1.415^2 + ... + 1.399^2 = 9.680.137$

$\sum y = 871 \quad ... + 843 = 4.190$

$\sum y^2 = 871^2 + ... + 843^2 = 3.522.490$

$\sum xy = 871x1.415 + ... + 843 \times 1.399 = 5.839.002$

5) Em um cálculo matricial, tem-se: SQTotal = y'y - $(\sum y_i)^2/n$ = 11.270. Sabendo-se que $\sum y_i^2$ = 3.522.490 e n = 5. Determine o valor de $\sum y_i$.

6) As matrizes abaixo representam informação de 16 frangos de corte em que $GP_{1\text{-}28}$ e $GP_{29\text{-}42}$ indicam, respectivamente, o ganho de peso, em gramas (g), no período de 1 a 28 dias e de 29 a 42 dias de idade.

$$GP_{1\text{-}28} = \begin{bmatrix} 701 & 689 & 675 & 613 \\ 690 & 715 & 669 & 584 \\ 719 & 725 & 699 & 661 \\ 807 & 687 & 673 & 660 \end{bmatrix}; GP_{29\text{-}42} = \begin{bmatrix} 651 & 658 & 650 & 632 \\ 648 & 659 & 644 & 615 \\ 623 & 588 & 607 & 669 \\ 545 & 667 & 618 & 656 \end{bmatrix}$$

Represente, em termos matriciais, o ganho de peso total do frango e também a média do ganho diário, do nascimento a 42 dias de idade.

Capítulo 3

Apresentação gráfica

Introdução

A quantidade de informação no mundo duplica num período de aproximadamente 4 a 5 anos, e o tamanho e a quantidade dos bancos de dados crescem com velocidade ainda maior.

Nesse contexto, os gráficos são ferramentas imprescindíveis no mundo de hoje, pois possibilitam explorar grande quantidade de dados, com rapidez e precisão. Na área científica, os gráficos são eficientes para mostrar a estrutura e o comportamento dos dados, detectar padrões ou tendências, propiciar refinamentos metodológicos, visualizar o ajuste de modelos, resumir e transmitir de forma precisa e ágil as informações latentes em um conjunto de dados. Além disso, a exposição por meio de gráficos é a melhor forma de atrair a atenção do público no mundo científico, em jornais, em revistas, etc.

Neste capítulo, são apresentados conceitos de gráficos uni, bi e tridimensionais, com ampla discussão em seu emprego, utilizando-se dados reais.

Gráficos unidimensionais

Histogramas ou gráficos de barras

Os histogramas ou gráficos de barras, do tipo horizontal e vertical, são os mais simples que existem; eles mostram as distribuições de variáveis do tipo intervalares ou nominais, e, geralmente, são utilizados com o intuito de analisar as projeções em determinado período. As barras devem ter a mesma largura, não justapostas e de intervalo constante. O comprimento das barras deve ser proporcional às frequências de cada classe. Geralmente o eixo *y* representa a frequência relativa de cada classe, enquanto o eixo *x* representa as faixas ou classes de valores.

Com o gráfico de barras, é possível visualizar aproximadamente a forma da distribuição do conjunto de dados, como a localização do valor central, a amplitude dos dados e a frequência de cada classe. Pode-se ainda ter informações da variância, da assimetria, da curtose e da distribuição teórica a que os dados estão associados. Entretanto, a percepção da forma da distribuição pode ser extremamente influenciada pela mudança na largura e na posição das barras.

Como exemplo de gráfico de barras, são apresentadas, na Figura 1, a frequência absoluta e a frequência acumulada da produtividade de forragem, descritas na Tabela 1 do Capítulo 2.

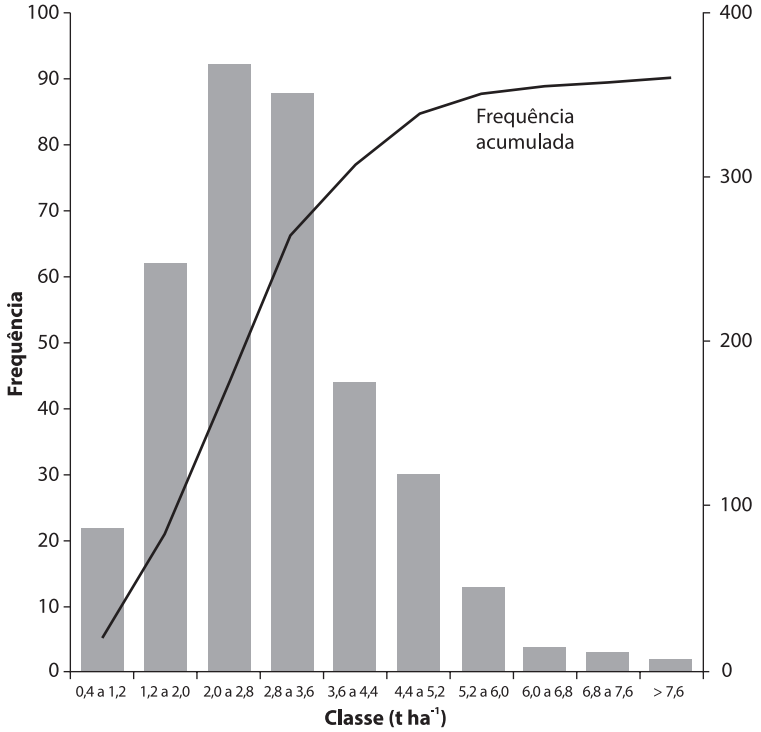


Figura 1. Histograma de frequências absoluta e acumulada (eixo *y*) por tonelada de produtividade de forragem (t ha^{-1}) de matéria seca em cada classe (eixo *x*).

Na Figura 2, é apresentada a distribuição de frequências de dados de pesos de uma amostra de 541.900 bovinos da raça Nelore, fêmeas e machos, controlados pela Associação Brasileira de Criadores de Zebu (ABCZ). Nessa amostra, a amplitude de peso dos animais variava de 150 kg a 400 kg, e a pesagem foi realizada antes do uso de balanças eletrônicas pela ABCZ. A construção do histograma − além de mostrar a distribuição da frequência de animais por categoria de peso, que se assemelha a uma distribuição normal − revela ainda informações valiosas que se encontram latentes, principalmente em um grande conjunto de dados. Nesse estudo, o histograma revela a existência de vícios na pesagem dos animais no campo.

No histograma, visualizam-se três tipos de frequência: os picos maiores, que correspondem às pesagens terminadas em zero, ou seja, pesos do tipo 150 kg, 160 kg,

etc.; os picos intermediários, que correspondem à frequência de pesos com final 5, ou seja, 155 kg, 165 kg, etc.; por último, a frequência compacta, de distribuição contínua dos pesos.

Observa-se que as frequências de pesos localizadas nas dezenas são bastante superiores àquelas com final 5, o que discorda do esperado em uma distribuição normal, ou seja, frequências maiores na média, reduzindo-se a partir desse ponto para os extremos, porém distribuídas simetricamente.

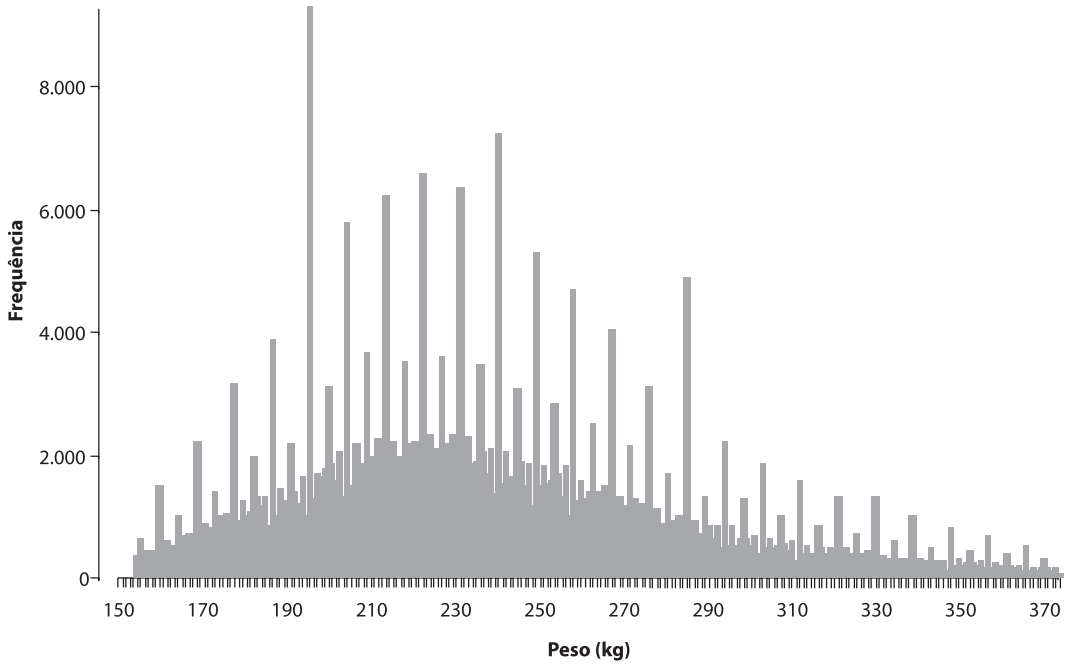


Figura 2. Frequência absoluta de animais da raça Nelore por categoria de peso.
Fonte: Freitas et al. (2000).

Para esse conjunto de dados, a análise gráfica não somente possibilitou detectar problemas no manejo do peso dos animais, o que é importante para o melhoramento genético da raça, mas também propiciou a visualização de informações sobre estatísticas descritivas, tais como *outliers*, assimetria, curtose, medidas de tendência central e heterogeneidade de variâncias.

Diagramas de caixa

Diagramas de caixa (*box plot*) são representações gráficas de um conjunto de dados. São eficientes para comparar distribuições de dados contínuos e para revelar características importantes, como a dispersão dos dados em torno da média, o grau e a direção da assimetria, a existência de heterogeneidade de variâncias, a presença de *outliers* e a assimetria, entre outras.

No diagrama de caixa (Figura 3), que apresenta a produtividade de matéria seca de alfafa da variedade Crioula, em um experimento realizado na Embrapa Pecuária Sudeste, São Carlos, SP, a linha horizontal cheia no meio da caixa indica a mediana, e a linha fina, a média aritmética. O comprimento da caixa corresponde à distância interquartílica (Q_3 - Q_1), em que os limites da parte inferior e da parte superior da caixa indicam, respectivamente, o primeiro (Q_1) e o terceiro (Q_3) quartil, correspondendo aos elementos de posição 25º e 75º da amostra. As caixas estreitas (*whiskers*), acima e abaixo da caixa central, possuem distância não superior a 1,5 vez a distância interquartílica (Q_3 - Q_1), e as marcações individuais além desses limites são potencialmente *outliers*.

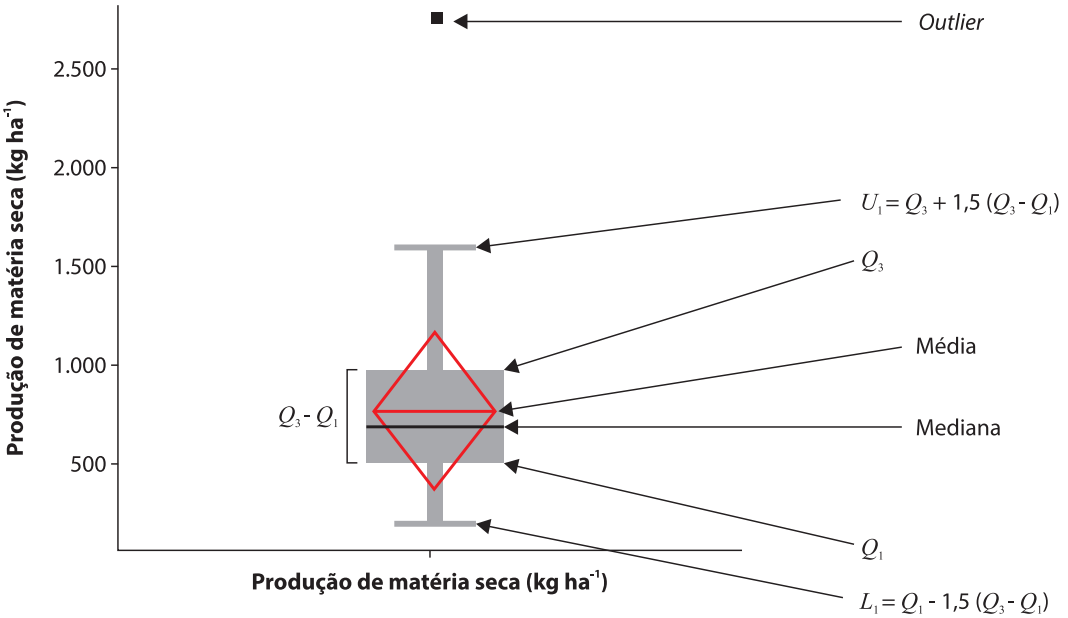


Figura 3. Diagrama de caixa para produtividade de matéria seca de alfafa da variedade Crioula.

Possíveis *outliers* são determinados a partir de pontos calculados no gráfico:

$$L_1 = Q_1 - 1{,}5(Q_3 - Q_1)$$

$$U_1 = Q_3 + 1{,}5(Q_3 - Q_1)$$

L_1 e U_1 são as delimitações internas. Valores menores que L_1 e maiores que U_1 são considerados discrepantes (*outliers*).

Gráficos de pizza

Os gráficos de pizza são circulares e divididos em setores. São utilizados para examinar a contribuição ou participação de cada setor com relação ao todo. Seu uso é bastante comum nas diversas áreas, porém, têm maior participação na análise de dados de indústrias, de bancos e informações financeiras de modo geral. Em termos numéricos, existe uma variável dependente que, na maioria das situações, expressa, em percentagem, o valor do círculo total como 100%. Como variável independente, têm-se as categorias. Cada uma delas está associada a um setor ou fatia e a soma delas representa o fenômeno como um todo. O gráfico de pizza da Figura 4 mostra um exemplo fictício de venda de leite mensal, em tonelada (t), de 12 fazendas (A a L) de uma cooperativa. No gráfico, os 12 setores indicam o nome da fazenda com a respectiva quantidade de leite vendida, e a parte interna do gráfico apresenta as fazendas divididas por cores e ainda o tamanho proporcional à quantidade vendida.

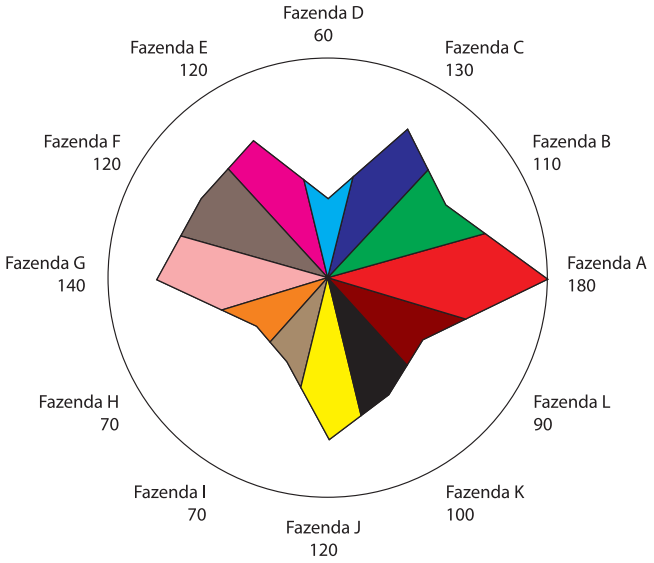


Figura 4. Venda de leite mensal, em tonelada (t), de 12 fazendas (A a L).

Gráficos bidimensionais

Os gráficos bidimensionais mostram o parentesco entre duas variáveis, que geralmente são provenientes de dados organizados em tabelas de dupla entrada ou dados bivariados, em que cada par (x_i, y_i) é representado por um ponto em um sistema de eixos coordenados. Geralmente em situações experimentais, para obter os dados (x_i, y_i) muitas vezes é necessário trabalhar com funções. Por exemplo, em um experimento realizado com pastagens, com o objetivo de estimar a produtividade de matéria seca de capim *Coast-cross*, em quilograma por hectare (kg ha^{-1}), em função de doses crescentes de adubo, foram utilizadas cinco doses de nitrogênio (0 kg ha^{-1}, 25 kg ha^{-1}, 50 kg ha^{-1}, 100 kg ha^{-1} e 200 kg ha^{-1}). Nesse experimento, após analisar os dados, o pesquisador obteve a seguinte equação:

$$y_i = 446{,}83 + 45{,}99496x_i - 0{,}1112x^2_i$$

em que:

y_i = i-ésima produtividade de matéria seca (PMS) da forragem.

x_i = i-ésima dose de adubo.

Com essa função, pode-se estimar a média da produtividade de PMS para qualquer dose de adubo, desde que ela esteja no intervalo de 0 kg a 200 kg. Por exemplo, na dose zero, a média de PMS é 446,83 kg ha^{-1}, enquanto na dose 200 a média da PMS é igual a 5.197,82 kg ha^{-1}.

Histogramas ou gráficos de barras

Histogramas ou gráficos de barras permitem avaliar as distribuições de variáveis do tipo intervalares ou nominais. São gráficos bastante comuns e suas características já foram discutidas nos gráficos unidimensionais no início deste capítulo. Um exemplo é apresentado na Figura 5, que é o resultado de um estudo de seis produtores familiares do município de Muriaé, MG, cuja principal atividade era a produtividade de leite. O estudo fez parte de um projeto de pesquisa e desenvolvimento (P&D) em agricultura familiar, executado pela Embrapa Pecuária Sudeste, São Carlos, SP. Com o intuito de ressaltar o impacto das tecnologias nos estabelecimentos familiares em 1999 e 2000, em relação a 1998, considerou-se esse ano referência = 100%. Conforme se pode visualizar na figura, houve aumento crescente na produtividade de leite em 1999 e 2000, cujos índices variaram de 8% a 69%. Os resultados evidenciaram que, de 1998 a 2000, o

gerenciamento e as tecnologias adotadas pelos produtores familiares, nesse município, refletiram positivamente na produtividade de leite.

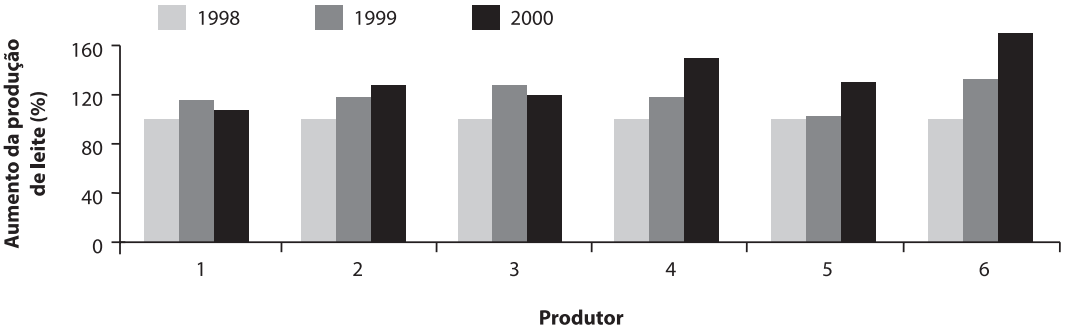


Figura 5. Aumento relativo, em percentagem (%), na produtividade de leite em estabelecimentos familiares, no município de Muriaé, MG, em 1999 e 2000, em relação a 1998.

Diagramas de caixa

Na Figura 6 ilustram-se dois diagramas de caixa que apresentam a produtividade de forragem, em tonelada por hectare (t ha^{-1}), de seis cortes. No primeiro diagrama (Figura 6A), tem-se a produtividade *y* em função de cinco espécies de capim (Tanzânia, *Brachiaria*, Marandu, Pojuca e *Coast-cross*). No segundo diagrama (Figura 6B), é apresentada a produtividade em função de quatro doses de adubo: (0 kg ha^{-1}, 20 kg ha^{-1}, 40 kg ha^{-1} e 60 kg ha^{-1}).

Várias informações experimentais podem ser obtidas a partir desses diagramas, como, por exemplo, que a produtividade de forragem (*y*) é distinta entre os cinco tipos de capins nos seis cortes. Quando se analisa a mediana (linha em negrito do retângulo), verifica-se, na Figura 6A, que a produtividade é crescente na seguinte ordem das forrageiras: 5, 3, 2, 1 e 4. Analisando-se a amplitude interquartílica (Q_3 - Q_1), em que Q_1 é a parte inferior do retângulo e Q_3 é parte superior, verifica-se que a produtividade *y* também difere entre as espécies de capim, ou seja, há uniformidade de produtividade para os capins 1, 2 e 5 e maior variabilidade para os capins 3 e 4.

Analisando-se a Figura 6B, verifica-se que a produtividade (eixo *y*) é crescente à medida que a dose de adubo aumenta. Porém, com base na amplitude interquartílica (Q_3 - Q_1), que é o retângulo, observa-se que a variabilidade da produtividade é praticamente constante nas quatro doses de adubo.

A presença de observações influentes e que são candidatas a *outliers* também podem ser visualizadas no gráfico. Exceto para o capim 5 (Figura 6A) e para a dose de 40 kg ha^{-1} de nitrogênio (Figura 6B), todos os pontos observados têm grande probabilidade de serem *outliers*.

Outra informação relevante que é observada na Figura 6 relaciona-se à assimetria, isto é, a medida da forma da distribuição dos dados com relação à distribuição simétrica (distribuição normal). Ela é representada pela barra acima e abaixo do retângulo, e, em ambos os gráficos, exceto para o capim 2 (Figura 6A), a assimetria é positiva.

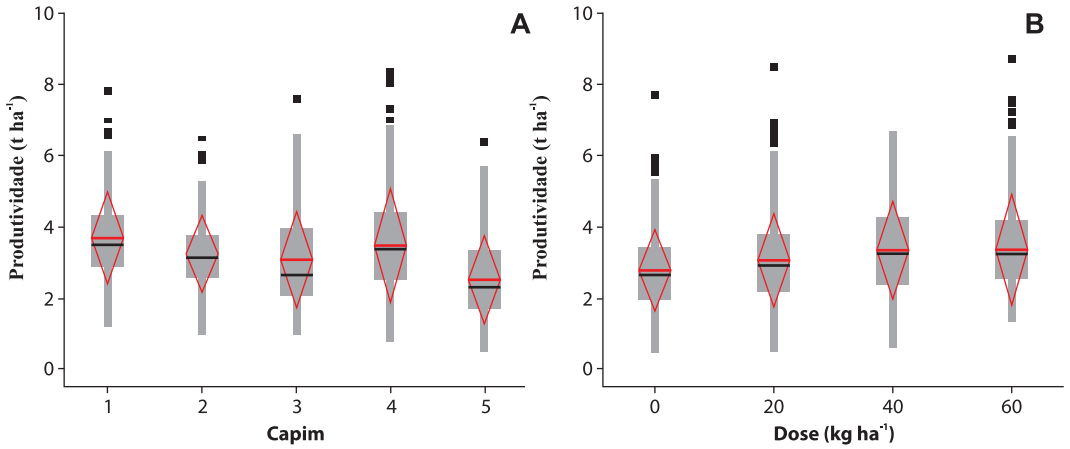


Figura 6. Diagramas de caixa da produtividade de forragem, em tonelada por hectare (t ha^{-1}) no eixo *y*, em função de cinco espécies de capins: Tanzânia-1, *Brachiaria*-2, Marandu-3, Pojuca-4, *Coast-cross*-5 (A) e quatro doses de nitrogênio: 0 kg ha^{-1}, 20 kg ha^{-1}, 40 kg ha^{-1}, 60 kg ha^{-1} (B).

As informações indicadas pelo diagrama de caixa são as mais variadas possíveis. Na Figura 7, são apresentadas respostas bastante diferentes daquelas fornecidas pelos diagramas de caixa da Figura 6. O diagrama da Figura 7A mostra nove pesagens − ao nascimento (P_0) e oito (P_1 a P_8) de bovinos da raça Guzerá, machos e fêmeas. As pesagens foram realizadas em intervalos trimestrais até os 2 anos de idade. À medida que os animais ficam mais velhos, a amplitude dos dados, na parte superior da mediana, aumenta proporcionalmente, caracterizando-se assimetria positiva e com possível presença de *outliers*.

Verifica-se também variância crescente no peso dos animais com o aumento da idade, indicando a existência de heterogeneidade de variâncias, fenômeno conhecido como inflação de variância. Esse comportamento observado da variância crescente

com a idade é de interesse para a área de melhoramento genético no estudo de curvas de crescimento, utilizando-se modelos não lineares e também para o estudo de dados longitudinais ou de medidas repetidas.

O diagrama da Figura 7B mostra dados de produtividade de matéria seca (PMS) de alfafa, obtidos de 20 cortes de um experimento realizado na Embrapa Pecuária Sudeste, São Carlos, SP. Os pontos individuais nos extremos das caixas estreitas são considerados dados discrepantes e candidatos a serem *outliers*. Observa-se que os maiores valores de PMS, pela ordem, foram obtidos nos cortes 2, 1 e 3 e os menores nos cortes 9, 17, 18 e 19. A flutuação da produtividade de PMS ao longo dos cortes não é casuística e sim em razão do efeito de sazonalidade na produção, isto é, períodos de seca (abril a setembro) e de chuvas (outubro a março), que ocorreram durante a realização dos 20 cortes mensais.

Da mesma forma que o gráfico da Figura 7A, os dados de produtividade das forrageiras obtidos de vários cortes ao longo do tempo são de interesse para análises de medidas repetidas, pois, nesse tipo de experimento, a ordem das observações realizadas na parcela ou na unidade experimental é fundamental, sendo as mais próximas mais correlacionadas. Associado ao comportamento do fenômeno biológico em função do tempo de ambos os gráficos, existe uma estrutura de correlação decorrente do fato de os dados avaliados em função do tempo serem correlacionados. Para cada caso, pode-se determinar a estrutura de correlação ou matriz de variância-covariância que descreva a natureza da correlação entre os dados.

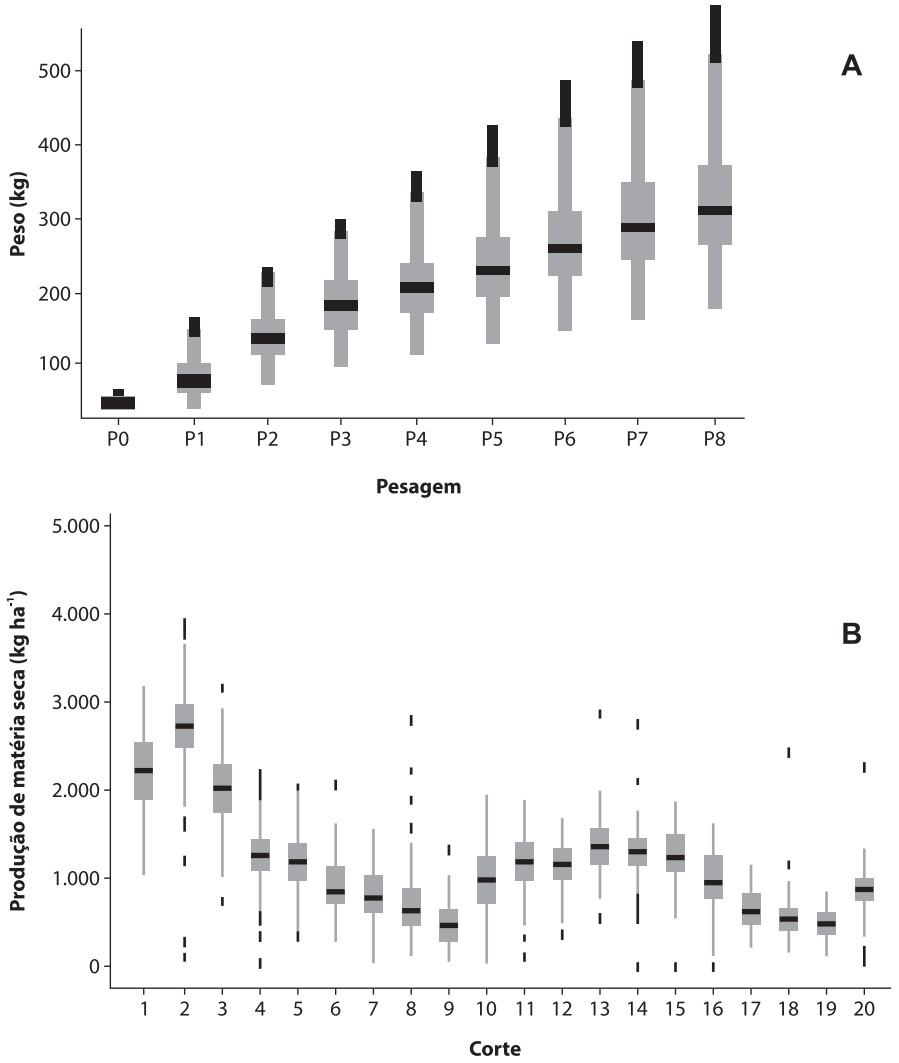


Figura 7. Diagramas de caixa de peso de bovinos Guzerá, em quilograma (kg), no eixo *y*, obtidos de nove pesagens trimestrais, do nascimento (P0) aos 2 anos de idade (P8) (A). Produtividade de matéria seca de alfafa Crioula, em quilograma por hectare (kg ha^{-1}) de 20 cortes mensais (B). Freitas et al. (2005)A e (2007)B.

Gráficos em linha (x-y)

Gráficos em linhas x-y ou simplesmente gráficos de linha geralmente são usados para indicar tendências de dados contínuos sobre o tempo. Em agricultura, normalmente o eixo *y* é representado por variáveis ou características, como peso, percentagem, altura, comprimento, taxa, etc. O objetivo principal é verificar o comportamento dessas variáveis

sobre o eixo *x*, que geralmente é o tempo, que pode ser em horas, dias, semanas, anos, etc. Em muitas situações, várias linhas ou fatores de uma mesma variável dependente precisam ser representados, tais como sexo, doses de adubo, local, raças de animais, e necessitam serem comparados simultaneamente.

Esses gráficos geralmente expressam *y* como função de *x* por meio da função *y* = f(*x*). Como exemplos, pode-se expressar a produtividade de alimento *y* (quilograma por hectare − kg ha^{-1}) de uma cultura em função de níveis de adubação *x*.

Na Figura 8, é apresentada a taxa de crescimento instantânea (TCI), em quilograma por dia (kg dia^{-1}) de bovinos Nelore estimada pelos modelos não lineares: Brody e Von Bertalanffy. A TCI é a taxa de crescimento do indivíduo em um particular tempo. Em valores aproximados, a estimativa da TCI pelo modelo Brody foi de 0,77 kg dia^{-1} logo após o nascimento, e 0,30 kg dia^{-1} aos 750 dias de idade. Para o modelo Von Bertalanffy, a TCI foi de 0,51kg dia^{-1} por ocasião do nascimento, com um máximo de 0,66 kg dia^{-1} aos 100 dias de idade e 0,10 kg dia^{-1} aos 750 dias de idade.

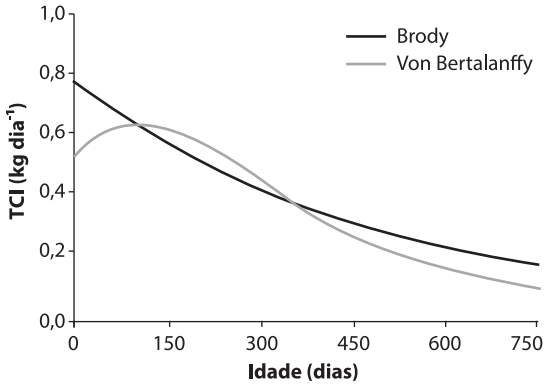


Figura 8. Taxa de crescimento instantânea (TCI), em quilograma por dia (kg dia^{-1}) do nascimento até 750 dias de idade de bovinos Nelore, estimada pelos modelos Brody (**—**) e Von Bertalanffy (—).
Fonte: Marinho et al. (2013).

Gráfico de dispersão (*scatter plot*)

O gráfico de dispersão ou *scatter plot* geralmente mostra um conjunto de pontos e cada um deles é representado por uma coordenada no plano x-y. O objetivo é visualizar todos os dados de uma amostra e, com isso, detectar padrões nos dados, tais como a presença de *outliers*. Uma das grandes contribuições do *scatter plot* é a possibilidade de

detectar o parentesco linear ou não linear entre as variáveis. Essa informação é relevante na pesquisa científica, pois possibilita escolher quais análises e modelos devem ser utilizados, e, com isso, fazer os refinamentos metodológicos necessários.

Nos estudos de regressão linear, por exemplo, há dois tipos de variáveis: uma variável dependente, geralmente no eixo *y*, e outra independente no eixo *x*. Se não existe relação de dependência, então qualquer uma delas pode ser plotada no eixo *x* e no eixo *y*, e, nesse caso, o gráfico mostra o grau de correlação entre elas.

A Figura 9 foi elaborada utilizando-se os mesmos dados da Figura 6, isto é, a produtividade da forragem de cinco espécies de capim. Ela apresenta gráficos de *scatter plot*, que é ideal para verificar a variabilidade entre os dados de cada forrageira e, também, de cada corte, facilitando, com isso, a detecção da presença de dados influentes e de *outliers*. Na Figura 9A, o gráfico permite visualizar que há uniformidade da produtividade para os capins 2, 3 e 5 e maior variabilidade para os capins 1 e 4.

Com relação à Figura 9B, observa-se que a produtividade varia entre os seis cortes e que a maior variabilidade entre os dados ocorre no corte 2, seguido do corte 4. Contudo, esses gráficos não informam a distribuição de dados, o coeficiente de assimetria, o coeficiente de curtose, a média e a mediana.

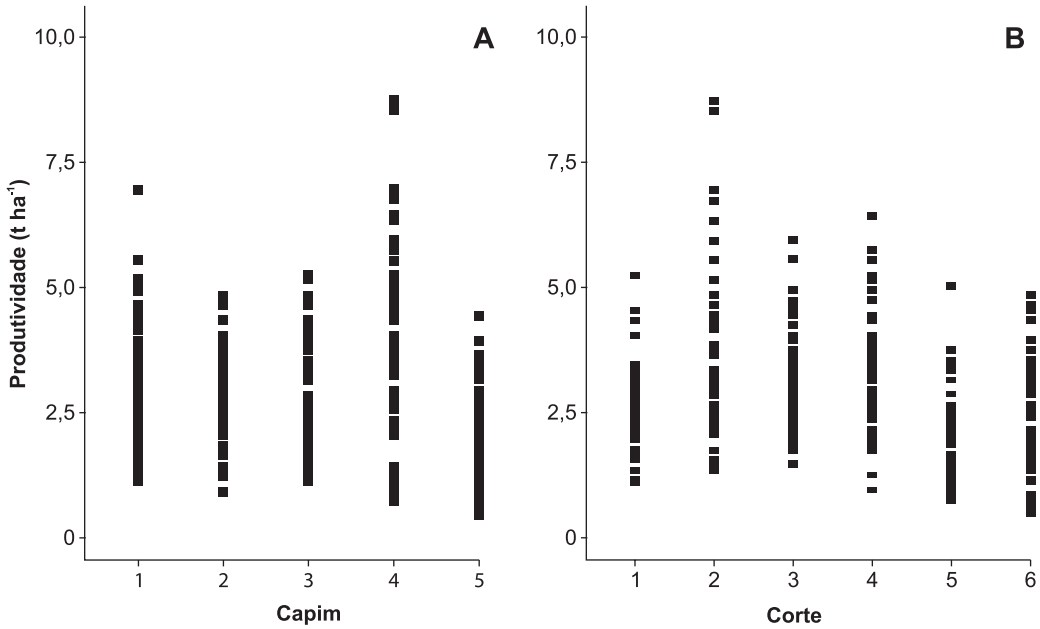


Figura 9. *Scatter plot* para produtividade de forragem (t ha^{-1}) no eixo *y*, de cinco espécies de capins: Tanzânia 1; *Brachiaria*-2; Marandu-3; Pojuca-4; *Coast-cross*-5 (A); e de seis cortes mensais (B).

Uma das grandes contribuições do *scatter plot* é a possibilidade de detectar o parentesco linear ou não linear entre as variáveis, como mostram os gráficos da Figura 10. No gráfico da Figura 10A, têm-se dados de pesagens do nascimento até 750 dias de idade de 17 mil animais da raça Nelore, e, com isso, observa-se uma massa compacta de pontos. Uma constatação imediata é a existência de uma relação linear ou parentesco linear do peso com a idade, isto é, o animal tem um crescimento corporal proporcional à idade. Observando-se mais cuidadosamente, informações valiosas são obtidas; por exemplo, quando os animais ficam mais velhos, principalmente a partir de 600 dias de idade, pois ocorre redução significativa no número de observações, fato que pode estar associado ao descarte de animais, venda, mortes, entre outros fatores.

Verifica-se que a variabilidade entre os dados é crescente com a idade dos animais, o que caracteriza aumento de variância entre pesos à medida que o animal fica mais velho, fenômeno também conhecido como heterocedasticidade de variâncias. Constata-se ainda, na Figura 10A, a existência de pontos discrepantes abaixo da massa compacta, o que indica animais com peso muito baixo em relação à idade. Observa-se também pontos discrepantes acima da massa compacta, os quais indicam animais com peso alto em relação à idade. Tais pontos ou dados discrepantes devem ser avaliados com cuidado antes de qualquer análise estatística, pois são suspeitos de serem observações influentes ou *outliers*.

No gráfico da Figura 10B, a variável percentagem de defeitos do sêmen (Deftot) representa uma somatória dos defeitos totais no esperma de bovinos Nelore, defeitos esses que variam de acordo com o crescimento da circunferência escrotal (CE). O objetivo do gráfico é mostrar o tipo de relacionamento ou parentesco existente entre Deftot e CE.

A linha circular vermelha, conhecida como elipse de predição, delimita uma região para predizer uma nova observação da população ou determinada percentagem da população. Ela mostra que as observações dentro de seu interior têm 90% de probabilidade de ocorrerem, enquanto as observações externas têm apenas 10% em amostras de animais dessa população. Observa-se, no gráfico, que a percentagem de defeitos ficou distribuída na amplitude da CE de 20 cm a 40 cm, com a maior concentração no intervalo de 30 cm a 35 cm. Na CE acima de 35 cm, tem-se estabilidade.

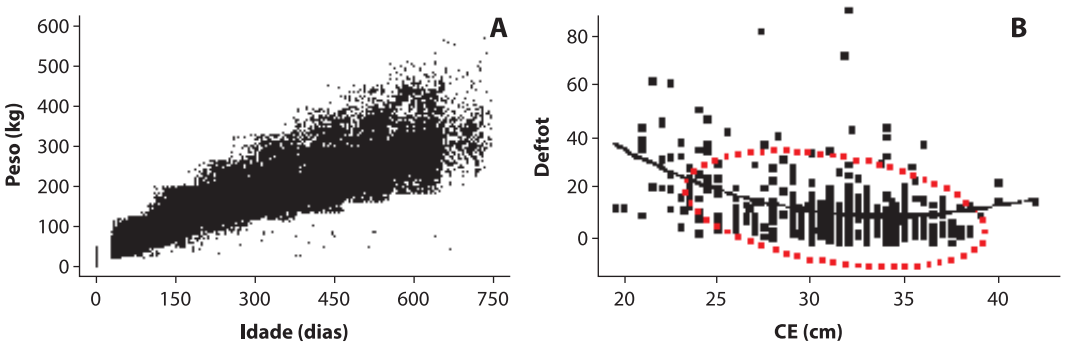


Figura 10. *Scatter plot* para peso corporal, em quilograma (kg), em função da idade (A); percentagem (%) de defeitos do sêmen (Deftot) em função da circunferência escrotal de bovinos Nelore (B).

Matrizes de dispersão (*scatter plot matrix*)

São matrizes de diagrama de dispersão que possibilitam explorar relacionamentos bidimensionais, os quais podem revelar várias informações sobre os dados, tais como dependências, *clusters* e *outliers*.

As variáveis são pareadas em uma matriz de diagrama de dispersão, cuja dimensão é representada pelo número de variáveis. Na Figura 11, tem-se um diagrama de dispersão 3 x 3, representando os teores de matéria seca (MS), de proteína bruta (PB) e de cálcio (CA), obtidos de um experimento de pastagem de capim *Coast-cross* (*Cynodon dactylon* "*Coast-cross*"), realizado na Embrapa Pecuária Sudeste.

Cada célula da diagonal da matriz contém o nome da variável e o seu valor mínimo e máximo: MS variando de 89,9% a 95,65%, PB variando de 6,88% a 22,19% e CA variando de 2,2 g kg^{-1} a 4,6 g kg^{-1}. Para cada diagrama de dispersão, uma variável está representada no eixo *y* e outra no eixo *x*. Na matriz, abaixo da diagonal, os gráficos de dispersão são os mesmos, apenas invertem-se os eixos: eixo *x* por *y* e eixo *y* por *x* (relação simétrica entre as variáveis). Pode-se observar, simultaneamente, as correlações pareadas das três variáveis.

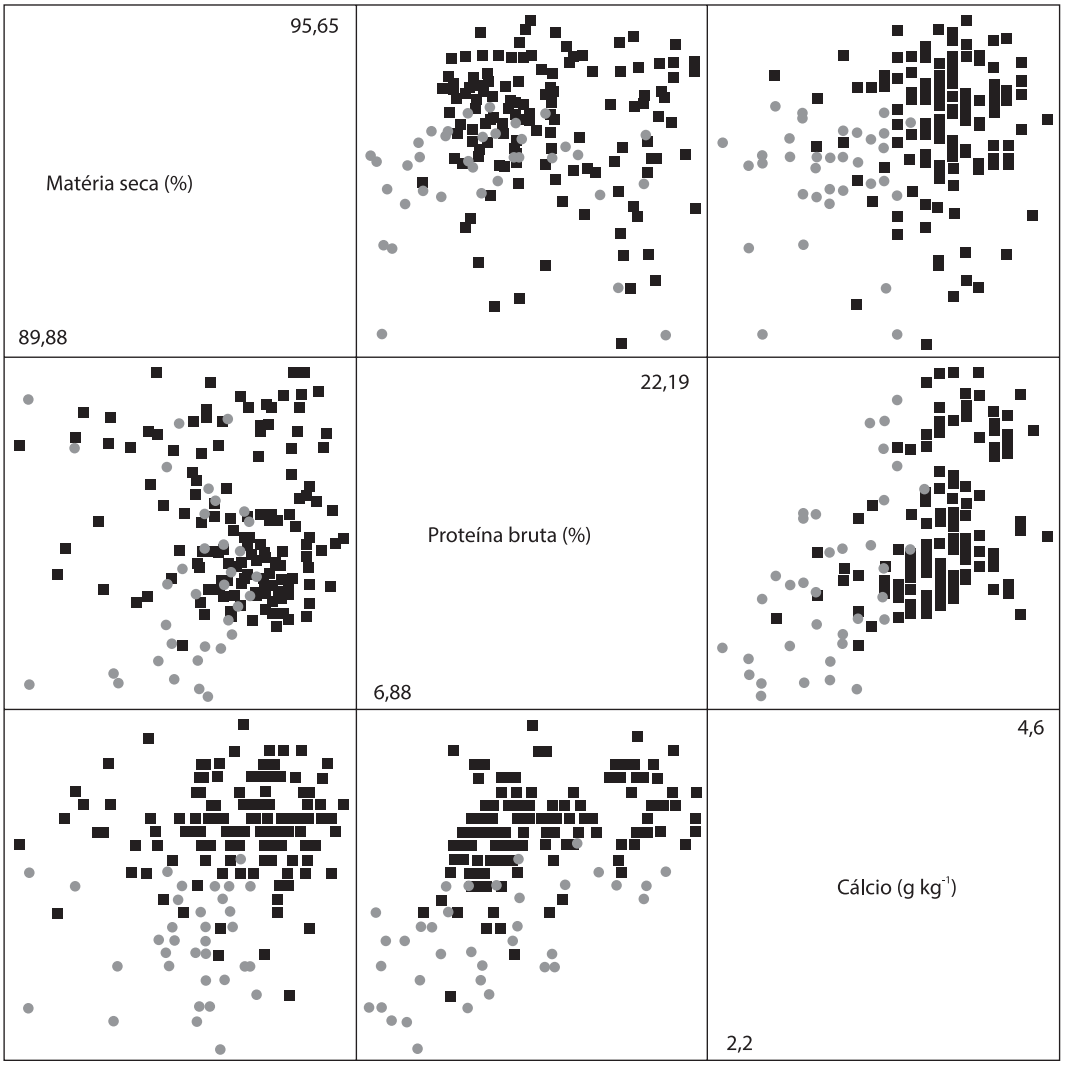


Figura 11. Matrizes de dispersão para matéria seca e proteína bruta em percentagem (%), e de cálcio em grama por quilograma (g kg^{-1}).

Gráficos tridimensionais

São utilizados para examinar o parentesco entre três variáveis (*x*, *y*, *z*), o que permite visualizar a altura, a largura e a profundidade no gráfico. Eles possibilitam analisar um fenômeno mais profundamente, revelando características importantes

da estrutura de um conjunto de dados que não são aparentes em gráficos de duas dimensões (*x*, *y*).

Existem vários tipos de gráficos tridimensionais que são apropriados para as mais diversas situações. Com recursos como rotação de eixos e combinação de cores, as possibilidades de uso desses gráficos aumentam significativamente. Os gráficos tridimensionais podem ser rotacionados como: ajuste de superfície (*surface plot*), de contorno e *scatter plot*. No gráfico de contorno, uma quarta variável pode também ser usada para colorir os contornos de superfície ao longo da direção do eixo *z* no espaço tridimensional.

Gráfico de superfície (*surface plot*)

Mostra o parentesco entre três variáveis por meio de uma superfície contínua. Este gráfico é muito comum nas metodologias de superfície de resposta que são utilizadas em experimentos de adubação, nos quais se deseja avaliar a combinação de adubos e de doses que proporcionam o maior rendimento de uma cultura.

Algebricamente, esse gráfico pode ser representado pela função $z = f(x, y)$, em que $z = 0$ representa a origem, e a superfície é construída para $z > 0$. No caso de experimentos de adubação, *x* e *y* podem representar dois tipos de adubo, cada qual com diferentes níveis de doses, enquanto *z* representa a produtividade.

Na maioria dos estudos de superfície de resposta, a forma do relacionamento entre a variável dependente e as independentes, isto é, $z = f(x, y)$, é desconhecida. Assim, o primeiro passo é encontrar uma relação ótima para essa função. Nos experimentos de adubação, essa relação ótima geralmente é obtida após uma análise de variância dos dados. Por exemplo, em uma situação em que o objetivo do experimento foi realizado de acordo com um esquema fatorial 5^2, cinco doses de fósforo (P), combinadas com cinco doses de nitrogênio (N), após a análise de variância, pode-se encontrar uma relação do tipo $z = (c_1 + c_2N + c_3P + c_4NP + c_5N^2 + c_6P^2 + c_7N^2P^2)$, em que c_1 a c_7 representam constantes, e N, P, NP, N^2, P^2 e N^2P^2 representam combinação das doses dos fatores N e P.

Na Figura 12 mostra-se uma relação da função PMS = f(PB, Altura), em que PMS é a média de produtividade de matéria seca, em quilograma por hectare (kg ha^{-1}), PB é proteína bruta, em percentagem, e Altura é a altura da planta, em centímetro (cm), obtidas de um experimento realizado com alfafa na Embrapa Pecuária Sudeste. A PMS variou de 1.139,78 kg ha^{-1} a 2.161,89 kg ha^{-1}, a PB de 25,46% a 30,51% e a Altura de 33,55 cm a 53,46 cm. De imediato, pode-se observar, na Figura 12, que a PMS tem um

acréscimo praticamente linear com o aumento da altura da planta. Até uma altura de aproximadamente 40 cm, o aumento da PMS varia de acordo com o aumento da percentagem de proteína.

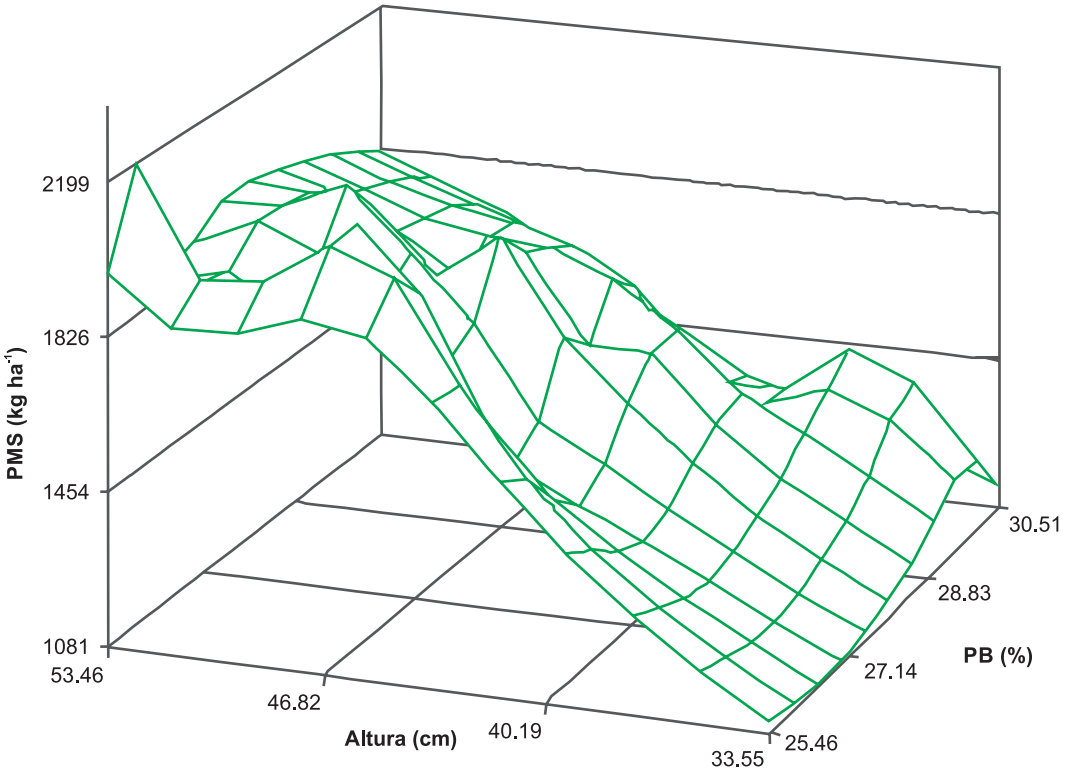


Figura 12. Produtividade de matéria seca (PMS), em quilograma por hectare (kg ha^{-1}) de alfafa em função da altura da planta (Altura), em centímetro (cm) e proteína bruta (PB), em percentagem (%).

Gráficos do tipo contorno

Os gráficos do tipo contorno utilizam linhas para representar níveis de magnitude de uma variável de contorno sob a superfície dos eixos horizontais e verticais. Visualmente é similar a um gráfico de duas dimensões; entretanto, é um gráfico de terceira dimensão, pois examina o parentesco entre três variáveis (*x*, *y*, *z*). Geralmente expressa *z* como função de *x* e *y* por meio da função *z* = f (*y, x*). Como exemplos, pode-se expressar a produtividade *z* (quilograma de alimento por hectare) de uma cultura como função de níveis de adubação *x* (quilograma de adubo por hectare) e do tipo de adubo *y* (tipo de adubo: A, B, C, …). Neste caso, *z* pode ser o resultado de um experimento e

ser expresso por $z = f(x, y) = x^2 + y^2$, sendo a construção similar àquela apresentada no gráfico com superfície (*surface plot*).

As variáveis *x* e *y* representam o plano e a variável *z* representa linhas de curvas de nível, de contorno ou isolinhas, que são traçadas como curvas. As curvas de nível podem ser utilizadas para exibir informações sobre duas variáveis sem se referir a uma superfície. Por exemplo, *z* pode representar a produtividade de uma cultura, enquanto *x* e *y* podem representar variáveis associadas ao tempo, como precipitação e temperatura.

As isolinhas são linhas que unem entre si pontos da superfície que tenham o mesmo valor. A compreensão desse valor depende da superfície que se está estudando. Para uma superfície que representa temperatura, as isolinhas são as isotermas; para superfície de um terreno, elas representam curvas de nível. Dependendo do software e do interesse, as áreas entre as isolinhas podem ser coloridas para tornar o gráfico mais atrativo.

Continuando com o exemplo do rendimento de matéria seca, em tonelada por hectare (t ha^{-1}), na Figura 13 mostram-se, no plano bidimensional, as quatro doses de nitrogênio (eixo *y*: 0 kg ha^{-1}, 20 kg ha^{-1}, 40 kg ha^{-1} e 60 kg ha^{-1}) e cinco espécies de capim (eixo *x*: Tanzânia, *Brachiaria*, Marandu, Pojuca e *Coast-cross*). A legenda à direita é o rendimento de matéria seca, em tonelada por hectare (t ha^{-1}) (eixo *z*). Cada isolinha indica uma determinada produtividade. Por exemplo, a maior produtividade mostrada é representada pelo isolinha 3,9 t ha^{-1}, obtida nas doses de nitrogênio, 40 kg ha^{-1} a 60 kg ha^{-1}; a segunda maior é 3,6 t ha^{-1}, obtida nas doses de nitrogênio 20 kg ha^{-1} a 60 kg ha^{-1}, ambas com o capim 1. A menor produtividade é representada pela isolinha 1,7 t ha^{-1}, com o capim 5 na dose de nitrogênio, 0 kg ha^{-1} a 20 kg ha^{-1}. Observando-se as isolinhas, pode-se verificar que determinadas espécies de capim podem apresentar a mesma produtividade em todos os níveis de adubação.

Em vez de usar linhas, no gráfico de contorno, pode-se usar regiões no plano bidimensional como na Figura 14, que mostra o crescimento da circunferência escrotal (CE), dos 20 cm aos 40 cm, em função da idade, em meses, com características espermáticas de bovinos. Nos eixos *x*, *y* e *z*, são apresentados, respectivamente, a idade dos animais, em meses (Idade), a circunferência escrotal (CE), em centímetro (cm), e os defeitos totais (Deftot), em percentagem (%), das características espermáticas. O objetivo foi avaliar a ocorrência de Deftot em função da idade e do tamanho da CE dos animais. Observa-se que a maior parte do mapa é de cor branca, indicando que, na maioria dos animais, a percentagem de defeitos totais nas características espermáticas foi de aproximadamente 10% (cor branca); em uma percentagem pequena dos animais, a Deftot é de aproximadamente 15% (cor cinza); e há algumas ocorrências de Deftot de 15% a 33% (cor preta).

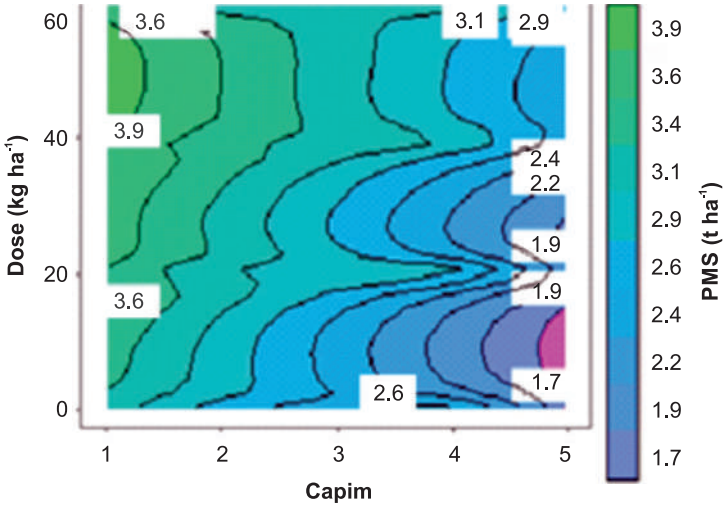


Figura 13. Gráfico do tipo contorno, mostrando a produtividade de matéria seca (PMS), em tonelada por hectare (t ha^{-1}) de cinco espécies de capim (Tanzânia-1; *Brachiaria*-2; Marandu-3; Pojuca-4; *Coast-cross*-5) em função de quatro doses de nitrogênio (0 kg ha^{-1}, 20 kg ha^{-1}, 40 kg ha^{-1} e 60 kg ha^{-1}).

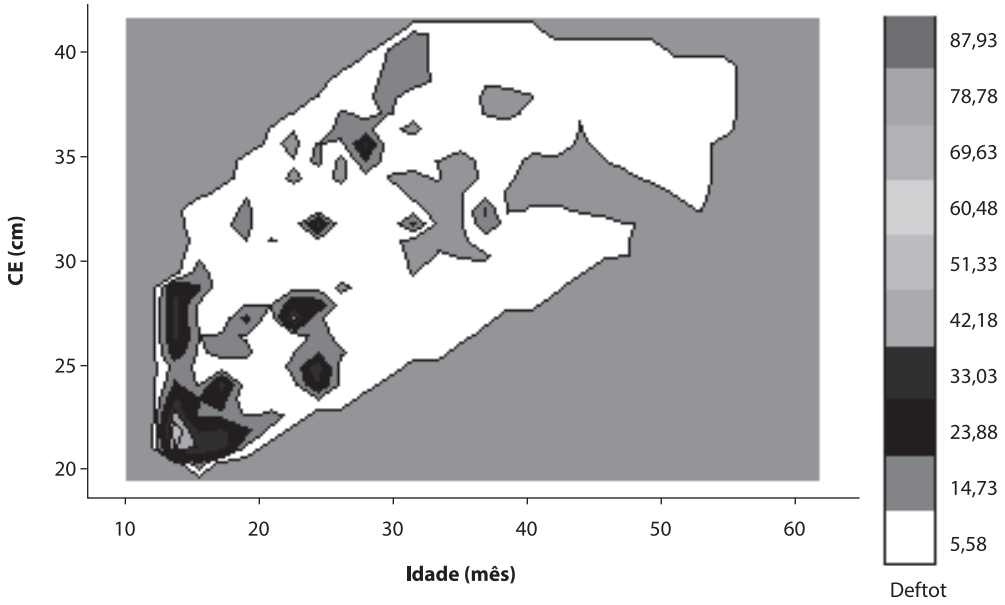


Figura 14. Gráfico de contorno mostrando a associação da circunferência escrotal (CE) com a ocorrência de defeitos totais (Deftot), em características espermáticas de bovinos Nelore até os 60 meses de idade.

Gráfico de dispersão rotacionado

Gráficos rotacionados do tipo *scatter plot* possibilitam examinar dados ou observações de três dimensões ao invés de examinar superfície, e, com isso, é possível obter informações dos dados que não seriam perceptíveis com outros tipos de gráficos ou métodos analíticos. Cada observação representa uma coordenada no plano tridimensional. O gráfico da Figura 15 é construído com os dados de rendimento de matéria seca, em tonelada por hectare (t ha^{-1}) (y), descritos neste Capítulo 3, cujos eixos são:

y: rendimento de matéria seca, em tonelada por hectare (t ha^{-1}).

Corte: seis cortes.

Dose: quatro doses de nitrogênio (0 kg ha^{-1}, 20 kg ha^{-1}, 40 kg ha^{-1}, 60 kg ha^{-1}).

Examinando o eixo *y*, pode-se observar, em valores aproximados, a amplitude do rendimento de matéria seca, em tonelada por hectare (t ha^{-1}), por meio da maior ou menor concentração de dados ou pontos da linha associada a cada combinação de cortes e doses. Na dose zero de nitrogênio (controle), o rendimento de matéria seca, em tonelada por hectare (t ha^{-1}) (y), varia de valores próximos de zero até 4,0 t ha^{-1}; na dose de 20 kg ha^{-1}, o rendimento varia de 2,0 t ha^{-1} a 7,0 t ha^{-1}, principalmente nos cortes 2, 3, 4 e 6; na dose de 40 kg ha^{-1}, o rendimento é semelhante ao observado para a dose de adubo de 20,0 kg ha^{-1}. Para a maior dose de adubo (60 kg ha^{-1}), a produtividade varia de 2,0 t ha^{-1} a 8,0 t ha^{-1}. Embora se observe que, para a dose zero de adubo, ela foi menor; para as demais doses, entretanto, não se verifica, na Figura 15, influência significativa da adubação nitrogenada sobre o rendimento da matéria seca, em tonelada por hectare (t ha^{-1}) (y).

Utilizando-se os mesmos dados da Figura 15, outra alternativa de gráfico do tipo *scatter plot*, sem rotação, é o apresentado na Figura 16. Nesse gráfico, é possível visualizar melhor que o rendimento de matéria seca, em tonelada por hectare, concentra na amplitude de 2 t ha^{-1} a 5 t ha^{-1}. E essa produtividade ocorre em todas as doses de adubo. Nesse gráfico, a altura está associada com a dose de adubo.

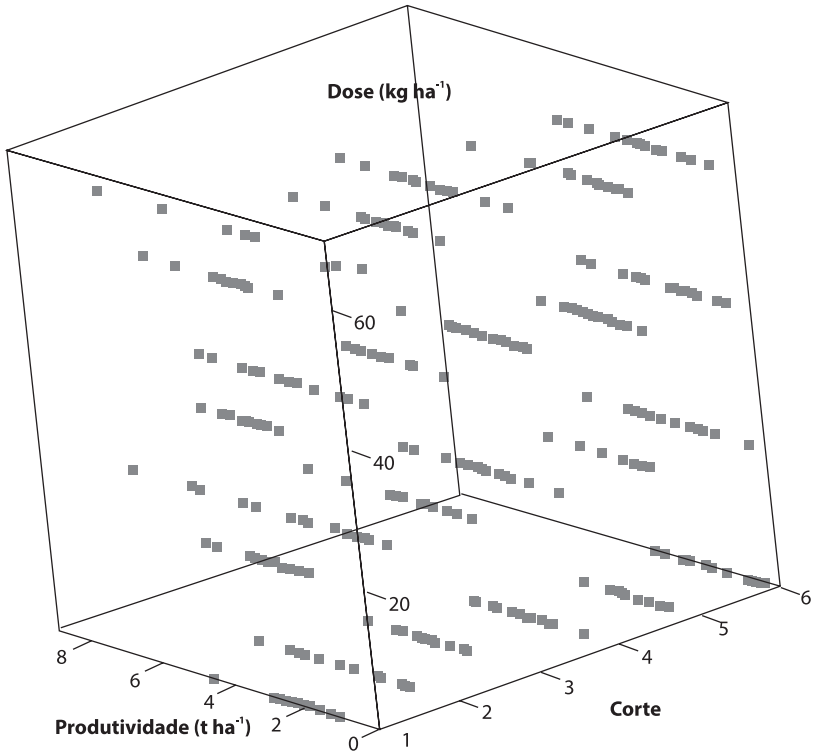


Figura 15. Gráfico rotacionado do tipo *scatter plot* mostrando o rendimento de matéria seca, em tonelada por hectare (t ha^{-1}) (Y), de cinco espécies de capins em função de quatro doses de nitrogênio (0 kg ha^{-1}, 20 kg ha^{-1}, 40 kg ha^{-1} e 60 kg ha^{-1}) e seis cortes mensais (corte: 1 a 6).

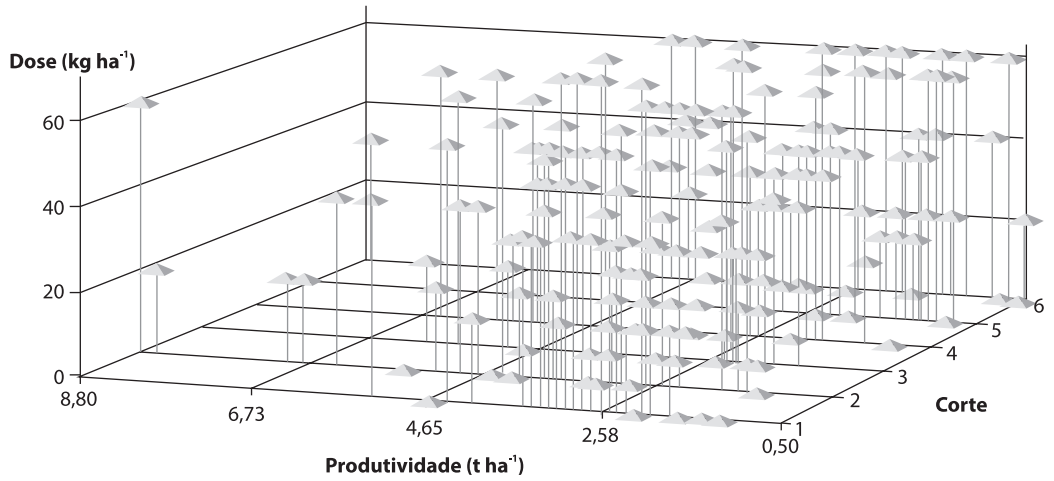


Figura 16. Gráfico rotacionado do tipo *scatter plot*, mostrando a produtividade de forragem, em tonelada por hectare (t ha^{-1}) (y), em função de quatro doses de nitrogênio (0 kg ha^{-1}, 20 kg ha^{-1}, 40 kg ha^{-1} e 60 kg ha^{-1}) e seis cortes mensais (corte).

Exercícios[3]

1) No texto a seguir, existem afirmações incorretas quanto a conceitos de estatística. Reescreva o texto colocando definições corretas e sublinhe ou coloque em negrito onde houve definições incorretas.

Na análise gráfica, uma das ferramentas mais importantes na estatística descritiva é o *box plot*, pois ele fornece: o valor mínimo e o máximo da amostra, o primeiro (Q_1) quartil que representa 25% dos dados, o segundo (Q_2) quartil que equivale à mediana (valor que representa 50% dos dados ordenados), o terceiro quartil (Q_3) ou quartil superior, que representa 75% dos dados ordenados. Outra informação que pode ser obtida do *box plot* é o intervalo interquartílico (Q_3 - Q_1), que inclui 75% dos dados da amostra. Dentre os gráficos, o histograma é o mais apropriado para exibir frequências de classes, e, por isso, é o mais utilizado para representar dados categóricos. O histograma é até mais apropriado para identificar *outliers* do que o próprio *box plot*, mas fornece poucas informações para visualizar simetria e curtose. Um outlier é uma observação que é discrepante dos demais dados e, principalmente, discrepante da média. Uma curiosidade acerca dos *outliers* é que a sua ocorrência em um conjunto de dados é numericamente igual no extremo inferior e extremo superior da amostra, isto é, se uma amostra tem quatro *outliers*, obrigatoriamente, dois estão no extremo inferior e dois estão no extremo superior. Os *outliers*, quando presentes em um conjunto de dados, não causam preocupação pois a sua influência nas estimativas de valores populacionais é mínima.

2) O gráfico de dispersão (Figura 17) mostra a estimativa do peso de bovinos Nelore fêmeas, em função da idade. Sabendo-se que cada ponto representa o peso de um animal, responda as questões a seguir.

a) Pode-se afirmar que o crescimento no período é praticamente linear?

b) Qual é o peso máximo atingido pelo animal?

c) Qual seria, aproximadamente, a média de peso aos 150, aos 300, aos 450 e aos 600 dias de idade?

d) Pode-se afirmar que há maior variabilidade de peso com o avançar da idade do animal?

e) Considerando-se que a média de peso é de 150 kg aos 200 dias de idade e 300 kg aos 550 dias, qual é a média do ganho de peso diário dos animais nesse período?

[3] As respostas dos exercícios podem ser consultadas no Apêndice 1.

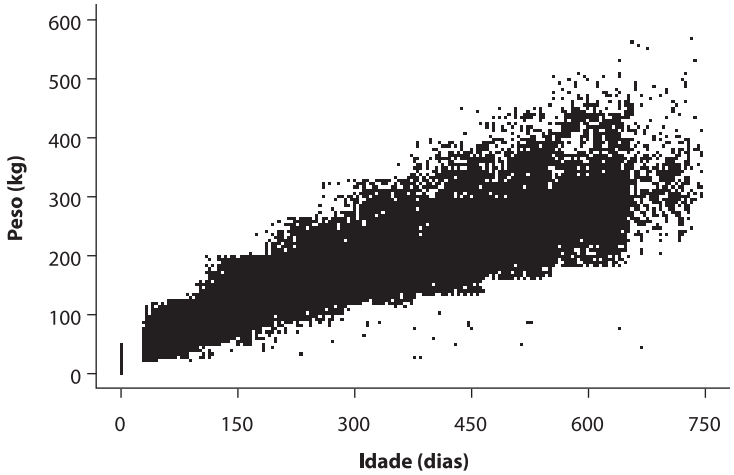


Figura 17. Peso de bovinos Nelore fêmeas em função da idade.

f) Sabendo-se a idade do animal, é possível afirmar qual seria o intervalo de seu peso?

3) Os dados da Tabela 1 referem-se a pesos, em grama, de coelhos machos da raça Nova Zelândia Branca, do nascimento aos 70 dias de idade.

Tabela 1. Pesos, em grama (g), de coelhos machos da raça Nova Zelândia Branca, obtidos de cinco animais, do nascimento aos 70 dias de idade.

Animal	Nascimento	7 dias	15 dias	21 dias	30 dias	45 dias	70 dias
1	51,9	82,5	182,4	303,0	610,6	1.130,0	1.777,0
2	52,6	81,4	171,5	320,0	642,2	1.180,0	1.469,0
3	53,3	78,2	154,2	310,0	680,8	1.120,0	1.859,0
4	48,6	65,0	165,7	290,0	592,7	930,0	1.469,0
5	46,6	68,0	195,2	285,0	562,3	940,0	1.447,0

Com relação a esses dados, indicar o gráfico mais adequado para:

a) Expressar a média de crescimento em cada idade.

b) Mostrar os pesos de cada animal nas várias idades.

c) Mostrar, simultaneamente, o peso, a idade e o animal.

d) Qual dos gráficos anteriores reúne maior quantidade de informação?

4) Na Tabela 2, encontram-se esquematizados os resultados de um experimento em que foram analisadas a produtividade de leite, a gordura, a proteína e sólidos de 20 vacas submetidas a sete controles leiteiros cada. No final do experimento, para cada variável foram obtidos 20 dados, um para cada vaca. Considerando-se essa tabela, qual o gráfico mais apropriado para representar as quatro variáveis?

Tabela 2. Produtividade de leite, gordura, proteína e sólidos de 20 vacas submetidas a sete controles leiteiros cada.

Vaca	Controle leiteiro							Resposta			
	1	2	3	4	5	6	7	Produção de leite	Gordura	Proteína	Sólidos
1	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	X_{16}	X_{17}	L_{11}	G_{11}	P_{11}	S_{11}
2	X_{21}	X_{22}	X_{23}	X_{24}	X_{25}	X_{26}	X_{27}	L_{21}	G_{21}	P_{21}	S_{21}
3	X_{31}	X_{32}	X_{33}	X_{34}	X_{35}	X_{36}	X_{37}	L_{31}	G_{31}	P_{31}	S_{31}
...	...							...	...	...	...
20	$X_{20,1}$	$X_{20,2}$	$X_{20,3}$	$X_{20,4}$	$X_{20,5}$	$X_{20,6}$	$X_{20,7}$	$L_{20,1}$	$G_{20,1}$	$P_{20,1}$	$S_{20,1}$

5) Na Figura 18 é apresentado o diagrama de caixa da produtividade de matéria seca (PMS), em quilograma por hectare (kg ha^{-1}), de 20 cortes mensais de alfafa (Freitas et al., 2011). A marca central em negrito no meio da caixa indica a mediana, elemento de posição 50% ou segundo quartil (Q_2). As partes inferior e superior da caixa correspondem, respectivamente, ao quartil Q_1 e Q_3. As barras estreitas (*whiskers*) acima e abaixo da caixa central possuem distância não superior a 1,5 vez à distância interquartílica (Q_1 - Q_3).

Comente, para cada corte, a dispersão dos dados em torno da média e, também, o grau e a direção da assimetria.

6) Em bovinos de corte Zebu, como os das raças Nelore, Gir, Indubrasil, Tabapuã, Guzerá, controlada pela Associação Brasileira dos Criadores de Zebu (ABCZ), normalmente são realizadas nove pesagens por animal, em intervalos aproximadamente trimestrais, do nascimento até 750 dias de idade. Geralmente, os bancos de dados contêm pesos de milhares de animais pertencentes a várias raças, estados da federação, sexo, etc.

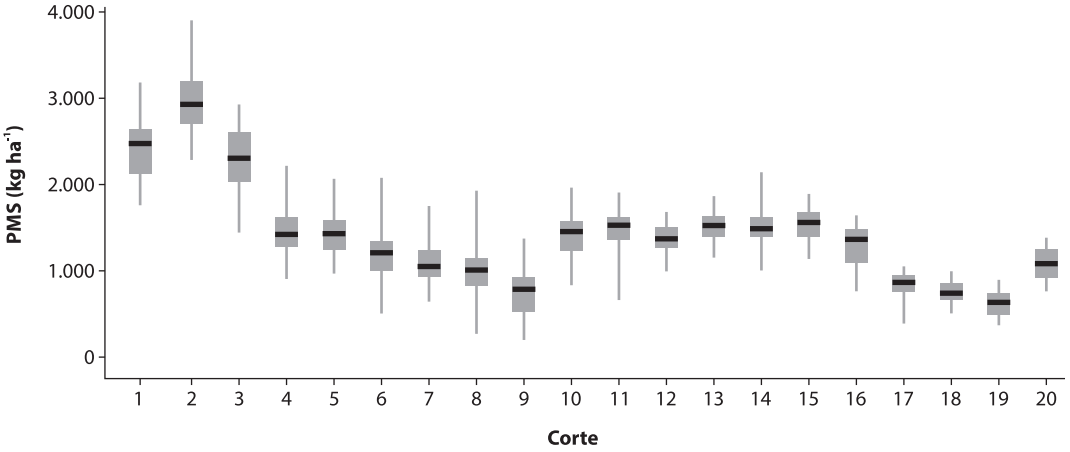


Figura 18. Produtividade de matéria seca (PMS), eixo *y*, em quilograma por hectare (kg ha^{-1}), de 20 cortes mensais de alfafa (eixo *x*).
Fonte: Freitas et al. (2011).

Com base no enunciado, mostre como elaborar gráficos com a relação peso-idade para as seguintes situações:

a) Gráfico peso-idade por raça.

c) Gráfico peso-idade por raça e sexo.

d) Gráfico peso-idade por raça, sexo e ano de nascimento dos animais.

Capítulo 4

Noções básicas de probabilidades

Introdução

A probabilidade teve início em 1654 com os matemáticos franceses Blaise Pascal (1623−1662) e Pierre de Fermat (1601−1665) que, utilizando também análise combinatória, aplicaram essas teorias a jogos de dados e cartas ((Wikipédia, 2019b).

Pascal também associou o estudo da probabilidade com o triângulo aritmético, que mais tarde ficou conhecido como triângulo de Pascal. Em 1657, Christiaan Huygens (1629−1695), matemático, físico e astrônomo holandês, publicou o folheto *De ratiociniis in ludo aleae* (sobre o raciocínio em jogos de dados). Entretanto, o mais antigo volume substancial sobre a teoria das probabilidades é o *Ars conjectandi*, de Jacques Bernoulli ou Jacob Bernoulli (1654−1705), matemático suíço, publicado em 1713, 8 anos após a sua morte (Boyer; Merzbach, 1996).

No período de 1751 a 1765, Leonhard Paul Euler (1707−1783), matemático e físico suíço, e Jean Le Rond d'Alembert (1717−1783), matemático e físico francês, realizaram várias aplicações da probabilidade sobre problema de expectativa de vida, loterias e outros aspectos da ciência social (Boyer; Merzbach, 1996).

Thomas Bayes (1702−1761), matemático e pastor inglês, no século 18, utilizou a probabilidade de forma intuitiva e estabeleceu as bases para a inferência estatística, tornando-se conhecido por ter formulado o famoso teorema de Bayes, que é um corolário do teorema da probabilidade total, uma das interpretações mais populares do conceito de probabilidade (Wikipédia, 2019j).

Apesar de importantes, as fases descritas anteriormente são introdutórias, pois grande avanço na teoria dessa área da matemática se deve ao francês Pierre Simon Laplace (1749−1827), considerado o fundador da teoria das probabilidades. A partir de 1774, ele escreveu muitos artigos sobre o assunto, os quais foram incorporados no *Théorie analytique des probabilitiés*, de 1812. Dois anos mais tarde, em 1814, ele publicou o trabalho *Essai philosophique dês probabilitiés*, considerada a primeira exposição do assunto disponível para o leitor comum.

No início do século passado, a teoria das probabilidades teve grande avanço em razão de vários acontecimentos favoráveis. Em 1901, é fundada a *Biometrika* por Karl Pearson (1857−1936); em 1906−1907, o russo Andrei A. Markov (1856−1922) introduziu as cadeias Markov de probabilidades e as teorias modernas de integração; em 1909, Félix Édouard Justin Émile Borel (1871−1956), matemático francês, contribuiu com *Elements de la théorie des probabilitiés*. No entanto, eventos de maiores vultos aconteceram após a Segunda Guerra Mundial. A matemática teve grande avanço, principalmente, com a

teoria dos conjuntos e a teoria da medida, e a probabilidade foi bastante influenciada por essas duas áreas (Wikipédia, 2019f).

A probabilidade hoje tem aplicação no dia a dia das pessoas e em vários setores da economia do País. Sua aplicação vai muito além dos jogos de loteria, dados, baralhos, etc. Dentre alguns exemplos de aplicações importantes, tem-se a previsão de chuvas, de frio, de ataques de pragas, entre outras, na agricultura. A probabilidade não fornece um resultado preciso de um evento que vai acontecer, porém, é a ferramenta mais indicada para fornecer um prognóstico ou previsão mais segura desse evento. No seu aspecto mais geral, a probabilidade quantifica a incerteza sobre um evento que já ocorreu, que está ocorrendo no presente e que ocorrerá no futuro. Porém, o leitor deve estar sempre consciente de que todas as afirmações, quando baseadas em expectativas futuras, e não em fatos históricos, envolvem riscos e incertezas.

A probabilidade é uma quantificação da incerteza em eventos futuros; porém, baseia-se em eventos passados, muitas vezes de grandes amostras. Está associada a um modelo estatístico e que possui uma distribuição, etc., portanto, tem fundamento científico. O leitor não deve confundir com informações do tipo – o homem das cavernas ou tal espécie animal surgiu na terra há 5 ou 6 milhões de anos ou ainda que, em um determinado evento, fontes de informações diferentes relatam que havia 5 mil pessoas, 7 mil pessoas, 10 mil pessoas, etc. Essas informações, geralmente, não possuem fundamento científico.

Intuitivamente, todos os seres vivos utilizam a probabilidade no dia a dia para tomar pequenas ou grandes decisões. Por exemplo, um pássaro decide escolher qual é o galho de uma árvore que ele decide pousar, ou quando um cachorro decide atravessar um rio utilizando a passarela mais larga, na verdade, esses indivíduos estão fazendo uma escolha de qual caminho ou alternativa ele terá mais probabilidades de ter sucesso.

Neste capítulo, serão discutidas ferramentas que possibilitam tomar decisões e fazer previsões com maior segurança.

Na genômica, por exemplo, que estuda o genoma completo de um organismo, possibilitando estudar a origem, a evolução e o futuro da humanidade, a sua estrutura e mecanismos de herança são fundamentados no uso de modelos probabilísticos. Apenas para ilustrar a complexidade da genômica no ser humano e a sua dependência da teoria da probabilidade, temos que a origem da vida acontece quando uma célula reprodutiva feminina (óvulo) funde com uma célula reprodutiva masculina (espermatozoide). Após essa fusão, por meio de mecanismos de multiplicação e diferenciação celular, o novo indivíduo é formado. No corpo humano, cada célula contém um núcleo e lá dentro estão

46 cromossomos, sendo 23 herdados do pai e 23 da mãe; aqui já começa a importância da probabilidade.

Como sabemos, os cromossomos são estruturas constituídas por proteínas e por molécula de *deoxyribonucleic acid* (DNA), estruturas em forma de hélice dupla, formadas por quatro blocos representados pelas letras G (Guanina), T (Timina), C (Citosina) e A (Adenina), sendo que A pareia com T e C pareia com G. Ora, no genoma humano existe em torno de 3 bilhões de pares dessas bases, que são responsáveis pela constituição de, aproximadamente, 25 mil genes, a unidade fundamental da hereditariedade. A organização desses genes no ser humano e sua transmissão de pais para filhos são responsáveis por formar o caráter e a personalidade de cada pessoa. A compreensão de todo esse mecanismo genético de transmissão e herança, como prever a possibilidade de ocorrer determinada doença, entre outras, tem por base a teoria da probabilidade.

Vale ressaltar ainda que os 3 bilhões de pares de letras ou bases do genoma humano podem ser lidos em apenas 1 dia e que existe também uma quantidade enorme de dados de várias espécies para se fazer inferências sobre a função do DNA. Com isso, cada vez mais há necessidade de conhecimento de áreas como a estatística (teoria da probabilidade), ciência da computação e genética.

Embora a probabilidade tenha aplicação na maioria das atividades das pessoas e em vários setores da economia do País, o caminho para o seu aprendizado básico é bastante simples e não exige conceitos matemáticos complicados. Exemplos com aplicações de dados, baralhos, bolas coloridas dentro de urnas, etc., com conhecimentos básicos sobre a teoria dos conjuntos, são os mecanismos necessários para ensinar probabilidade para qualquer usuário. Neste capítulo, esses itens serão estudados por meio de vários exemplos.

Conceitos

Probabilidade é o ramo da matemática que permite a quantificação da incerteza, possibilitando que as pessoas no seu dia a dia possam fazer previsões com determinada precisão dos fenômenos futuros. A incerteza é a propriedade que os acontecimentos ou eventos possuem de não serem completamente previsíveis, isto é, a incerteza pode ser minimizada, mas nunca eliminada. Na quantificação da incerteza, a probabilidade ou chance ou verossimilhança, que são sinônimos, de um evento ocorrer é representada por um número real entre 0 (nenhuma certeza de que o evento vai ocorrer, ou seja, plena certeza de que o evento não vai ocorrer) e 1 (plena certeza de que o evento vai ocorrer).

A finalidade da probabilidade e, por conseguinte, deste capítulo, é determinar com a maior precisão qual é o valor, de 0 a 1, dos eventos ou fenômenos futuros.

No mundo dos negócios, como na bolsa de valores, a probabilidade é usada no dia a dia para identificar previsões, as quais envolvem riscos ou incertezas. Com isso, os resultados de operações futuras na bolsa de valores podem ser bastante diferentes dos esperados.

Ensaio

É comum definir como ensaio ou experimento na área da probabilidade qualquer ação que fazemos e que, mantendo as mesmas condições, possa ser repetida infinitas vezes. Como neste livro a palavra experimento tem outra conotação, neste capítulo, será adotada a palavra ensaio.

Iniciaremos com ensaios bastante simples − jogar uma moeda sobre uma superfície plana e observar o resultado (cara ou coroa); jogar um dado não viciado e esperar que ocorra o número 6 na face voltada para cima, etc. Em todos os eventos ou experimentos em que os resultados não são previamente conhecidos, como os exemplos da moeda e do dado, tem-se os chamados modelos estocásticos ou modelos probabilísticos, que é o tema deste capítulo.

Existem situações, no entanto, em que os resultados são previamente conhecidos, como a ebulição da água; sabemos que sob condições normais de temperatura e de pressão, ela sempre acontece aos 100 °C; sabemos também que dois mais dois é igual a quatro, etc. Esses fenômenos que são previamente conhecidos são denominados modelos determinísticos.

Espaço amostral

É o conjunto de todos os resultados possíveis de um ensaio aleatório. A palavra aleatório vem do fato de que não sabemos qual será o resultado.

Exemplos de ensaios (ε) e de espaços amostrais (S):

ε_1) Atirar uma moeda duas vezes e observar a face voltada para cima.

Designando-se cara por C e coroa por K, esse ensaio gera o seguinte espaço amostral:

$S_1 = \{CC, CK , KC, KK\}$

ε_2) Atirar uma moeda quatro vezes e observar a face voltada para cima.

$S_2 = \{CCCC, CCCK, CCKC, CKCC, KCCC, CCKK, CKKC, KKCC, CKCK, KCKC, KCCK, CKKK, KCKK, KKCK, KKKC, KKKK\}$.

ε_3) Jogar um dado e observar o número mostrado na face voltada para cima.

$S_3 = \{1, 2, 3, 4, 5, 6\}$

Os três exemplos anteriores são bastante simples. Porém, na maioria das situações, podem ser bem mais complexos.

Um pesquisador conduz uma pesquisa em agricultura com o objetivo de avaliar o efeito de doses de dois adubos (A_1 e A_2) para a produtividade de gramíneas. Para cada adubo, foram usadas cinco doses: 0 kg ha^{-1}, 25 kg ha^{-1}, 50 kg ha^{-1}, 100 kg ha^{-1}, 200 kg ha^{-1}, denominadas, respectivamente, de D_1, D_2, D_3, D_4 e D_5. Nesse caso, temos dois fatores (adubo e dose) que correspondem a 10 tratamentos: A_1D_1, A_1D_2, A_1D_3, A_1D_4, A_1D_5, A_2D_1, A_2D_2, A_2D_3, A_2D_4 e A_2D_5. De acordo com esse enunciado, tem-se:

ε_4) Formular o número de tratamentos.

$S_4 = \{A_1D_1, A_1D_2, A_1D_3, A_1D_4, A_1D_5, A_2D_1, A_2D_2, A_2D_3, A_2D_4 \text{ e } A_2D_5\}$

Evento

É o subconjunto do espaço amostral (S) ou o número de resultados favoráveis de um ensaio (ε). Associados aos ensaios ε_1 a ε_4 e aos espaços amostrais S_1 a S_4, podem-se formular alguns eventos (A):

A_1) Ocorrência de duas caras:

$A_1 = \{CC\}$

A_2) Ocorrência de pelo menos três caras:

$A_2 = \{CCCC, CCCK, CCKC, CKCC, KCCC\}$

A_3) Ocorrência de números pares no lançamento de um dado:

$A_3 = \{2, 4, 6\}$

A_4) Número de tratamentos formulados com o adubo 1:

$A_4 = \{A_1D_1, A_1D_2, A_1D_3, A_1D_4, A_1D_5\}$

A_5) Número de tratamentos formulados com o adubo 2:

$A_5 = \{A_2D_1, A_2D_2, A_2D_3, A_2D_4, A_2D_5\}$

A_6) Número de tratamentos com o adubo 1, cujas doses sejam maiores do que 50 kg ha^{-1}:

$A_6 = \{A_1D_4, A_1D_5\}$

A_7) Número de tratamentos com o adubo 2, cujas doses sejam menores do que 100 kg ha^{-1}:

$A_7 = \{A_2D_1, A_2D_2, A_2D_3\}$

Probabilidade de um evento

A probabilidade de um dado evento P(A) é sempre um número de 0 a 1 ou de 0% a 100%:

P(A) = Nº de casos favoráveis / Nº total de casos possíveis.

As probabilidades dos eventos de A_1 a A_7 são:

$P(A_1) = 1/4$

$P(A_2) = 5/16$

$P(A_3) = 3/6 = 1/2$

$P(A_4) = 5/10 = 1/2$

$P(A_5) = 5/10 = 1/2$

$P(A_6) = 2/10 = 1/5$

$P(A_7) = 3/10$

Eventos dependentes e eventos independentes

Quando se estudam simultaneamente dois eventos, existem duas possibilidades quanto à relação entre as suas probabilidades:

a) Eventos dependentes: quando a ocorrência de um influencia a probabilidade de ocorrência do outro.

b) Eventos independentes: quando a ocorrência de um em nada interfere na ocorrência do outro. Assim, se A e B são eventos independentes, então a probabilidade de que ambos aconteçam ao mesmo tempo é necessariamente igual à probabilidade isolada de um deles ocorrer, multiplicada pela probabilidade isolada do outro, ou seja, em notação matemática: P(A e B) = $P(A \cap B) = P(A)P(B)$.

Leis da probabilidade

a) A probabilidade de um evento A é um número não negativo: $0 \leq P(A) \leq 1$.

b) A probabilidade de um evento certo é 1: $P(S) = 1$. Aqui, S significa o conjunto de todos os resultados possíveis de um ensaio aleatório.

c) A probabilidade de um evento impossível é zero: $P(\phi) = 0$.

d) A probabilidade de um evento complementar $P(A^C)$ é igual a 1 menos a probabilidade de A: $P(A^C) = 1 - P(A)$.

Teoria dos conjuntos

Um conjunto é formado por elementos ou membros que podem ser de qualquer natureza: números, frutas, animais, plantas, pontos, pessoas, etc. O nome de um conjunto é representado por letras maiúsculas A, B, C, ..., e os seus elementos normalmente são representados por letras minúsculas a, b, c, ..., ou números 1, 2, 3, ..., colocados entre chaves.

Exemplos de conjuntos:

A = {1, 2, 3}; B = {3, 4, 13}; C = {a, b, c}. Um conjunto vazio A pode ser representado por A = { } ou $A = \phi$.

As operações ou relações dos eventos em um espaço amostral são representadas pelo diagrama de Venn (Figura 1), que são curvas simples desenhadas sobre um plano de forma a simbolizar, graficamente, os conjuntos e suas propriedades.

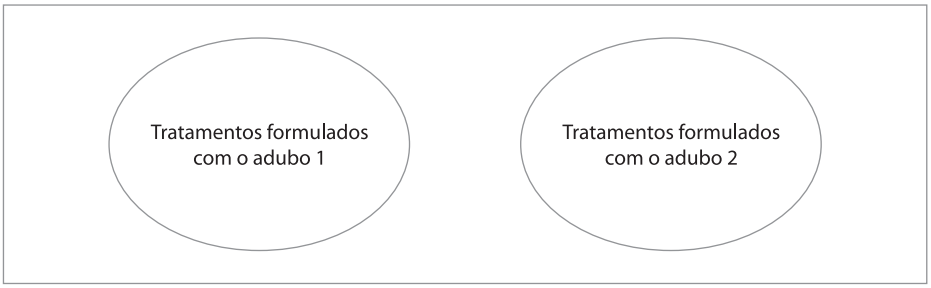


Figura 1. Diagrama de Venn de dois conjuntos.

Conjuntos: união, intersecção e complementação

Nas notações a seguir, utilizaremos os conjuntos A e B, porém, elas são válidas para quaisquer conjuntos.

União de conjuntos

A união de dois conjuntos A e B resulta no conjunto que representa todos os elementos que pertencem ao conjunto A e B (Figura 2).

Notação: $A \cup B = \{x : x \in A \text{ ou } x \in B\}$

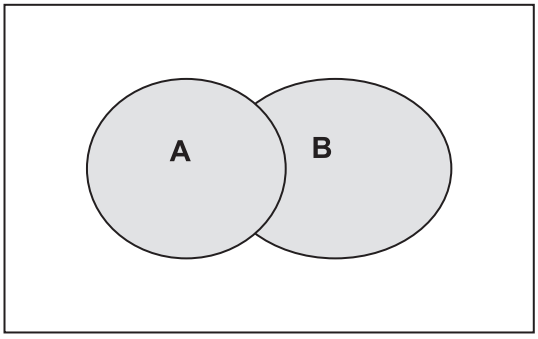


Figura 2. União dos conjuntos A e B (A U B).

Algumas propriedades da união:

$A \cup B = B \cup A$

$A \cup \phi = A$

$A \cup (B \cup C) = (A \cup B) \cup C$

$A \cup B \cup C = A \cup (B \cup C) = (A \cup B) \cup C$

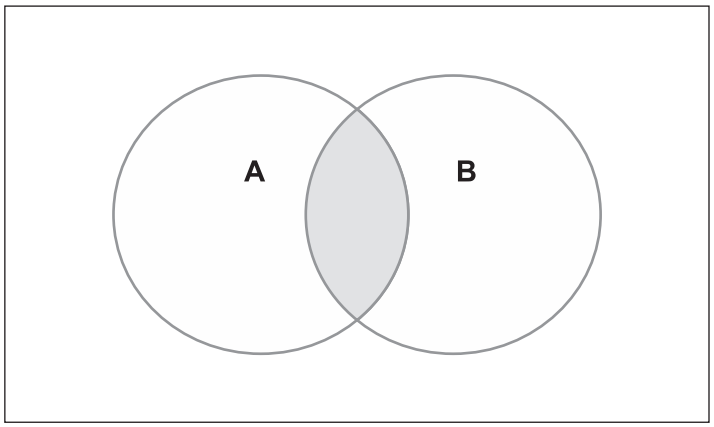
$$A_1 \cup A_2 \cup A_3 \ldots \cup A_n = \bigcup_{i=1}^{n} A_i$$

Aplicação

Seja $A = \{1, 2, 3, 5, 9\}$ e $B = \{3, 4, 13\}$:
$A \cup B = \{1, 2, 3, 4, 5, 9, 13\}$

Interseção de conjuntos

A interseção de dois conjuntos A e B ($A \cap B$) é o conjunto de todos os pontos que pertencem aos conjuntos A e B simultaneamente (Figura 3).
Notação: $A \cap B = \{x / x \in A \text{ e } x \in B\}$



$(A \cap B)$

Figura 3. Interseção dos conjuntos A e B ($A \cap B$).

Algumas propriedades da união e da interseção:

$A \cap \phi = \phi$

$A \cap B = B \cap A$

$A \cap (B \cap C) = (A \cap B) \cap C$

$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$

$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$

$$A_1 \cap A_2 \cap A_3 \ldots \cap A_n = \bigcap_{i=1}^{n} A_i$$

Aplicação:

$A = \{1, 2, 3, 5, 9\}$ e $B = \{3, 4, 13\}$

$A \cap B = \{3\}$

Evento complementar

Evento complementar (A^C) é o Evento complementar de A, representa os pontos do espaço amostral que não estão contidos em A (Figura 4):

$A^C = \{x / x \notin A\}$

$X \in A$: lê-se x pertence a A

$X \notin A$: lê-se x não pertence a A

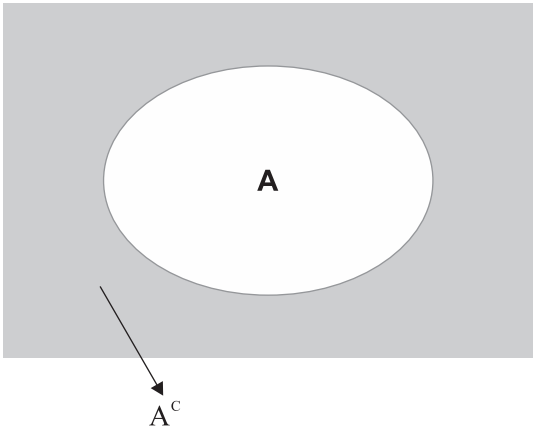


Figura 4. Evento complementar de A (A^C).

O complementar de A U B está representado no diagrama de Venn (Figura 5).

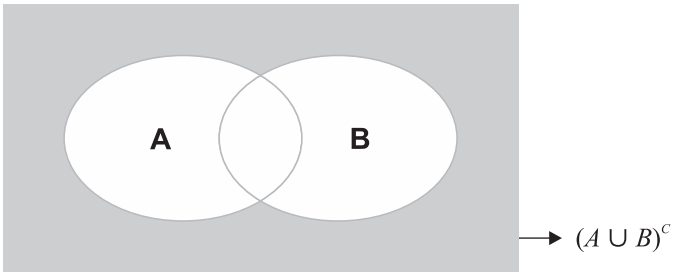


Figura 5. Evento complementar de A união B $(A \cup B)^c$.

Algumas propriedades:

$P(A^C) = 1 - P(A)$
$P(A) + P(A^C) = 1$

$A \cap A^c = \phi$

$(A \cap B)^C = A^C \cup B^C$

$(A \cup B)^C = A^C \cap B^C$

$A^c \cup A = A \cup A^c = S$

$S^C = \phi, \phi^C = S, (A^C)^C = A$

Aplicação:

Seja S = {1, 2, 3, 4, 5, 6, 7, 8, 9, 10} o conjunto fundamental formado por todos os elementos que estão sendo estudados e ainda os conjuntos:

$A = \{1, 2, 3, 4\}$

$B = \{3, 4, 5, 6\}$

$C = \{6, 7, 8, 9\}$

$D = \{1, 2, 3, 5, 9\}$

$E = \{3, 4, 5\}$

Determinar:

a) A^C b) $A \cup B$ c) $A \cup B$ d) $A^C \cup B$ e) $(A^C \cup B^C)$ f) D^C g) E^C

Resposta:

$A^C = \{5, 6, 7, 8, 9, 10\}$

$A \cup B = \{1, 2, 3, 4, 5, 6\}$

$A \cap B = \{3, 4\}$

$A^C \cap B = \{5, 6\}$

$A^C \cap B^C = \{7, 8, 9, 10\}$

$D^C = \{4, 6, 7, 8, 10\}$

$E^C = \{1, 2, 6, 7, 8, 9, 10\}$

Complemento relativo

Dada a diferença entre dois conjuntos A e B, o termo A − B indica o complemento relativo de A com relação a B, ou seja, a parte que está em A e não está em B:

$$A - B = A \cap B^C = \{x : \in A \text{ e } x \notin B\}$$

$$B - A = A^C \cap B = \{x : x \in B \text{ e } x \notin A\}$$

O complemento relativo de A com relação a B (A − B) e complemento relativo de B com relação a A (B − A) estão representados na Figura 6.

Figura 6. Complemento relativo de A − B e B − A.

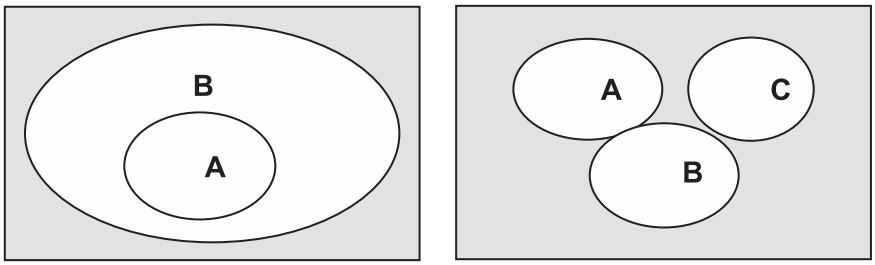
Um conjunto também pode ser elemento de outro conjunto, ou seja, existem conjuntos de conjuntos. Pode-se tomar, como exemplo, o plano, que é um conjunto de retas, e estas, por sua vez, são conjuntos de pontos. Os planos são, portanto, conjuntos de conjuntos. Dessa maneira, pode-se dizer que há dois tipos de relação nesse caso: entre elementos e conjuntos e entre conjuntos. As primeiras são chamadas relações de pertinência e as últimas, de relações de inclusão:

a) Pertinência: se **a** é um elemento do conjunto A, dizemos que **a** pertence ao conjunto A (notação: **a** ∈ A).

b) Inclusão: se A e B são conjuntos, e se todo elemento de A é um elemento de B (todo elemento de A também pertence ao conjunto B), diz-se que A é um subconjunto de B, ou que A está contido em B. De forma análoga, pode-se dizer que B contém A (notação: $A \subset B$ ou $B \supset A$).

Os diagramas da Figura 7 ilustram, respectivamente, $(A \subset B)^c$ e $(A \cup B \cup C)^c$.

A parte hachurada no diagrama à esquerda corresponde a $(AUB)^c = B^c$, pois $A \subset B$.



$(A \subset B)^c$ $(A \cup B \cup C)^c$

Figura 7. Complementar de A contido em B $(A \subset B)^c$ e complementar de A união a B e união a C $(A \cup B \cup C)^c$.

Se $A \subset B, B \subset C, \Rightarrow A \subset C$ ("$\Rightarrow$", implica ou equivale)

Se $A \subset B, B = AU(B - A)$

Se $A = A \Rightarrow A \subset A$

Se $A = B, B = C, \Rightarrow A = C$

$A \Delta B = (A - B) U (B - A)$

Notação: A - B = A menos B e B - A = B menos A

Frequência relativa de um evento

Admita-se que um ensaio ε seja repetido ***n*** vezes; sejam A e B dois eventos associados a ε, com n_A e n_B, o número de vezes que os eventos A e B ocorrem nas *n* repetições, respectivamente. Sejam $f_A = n_A/n$ e $f_B = n_B/n$ as frequências relativas dos eventos A e B.

Uma frequência relativa f_i, nos eventos A e B, tem as seguintes propriedades:

a) $0 \leq f_i \leq 1$.

b) $f_A = 1$; somente A ocorre nas ***n*** repetições.

c) $f_A = 0$; A nunca ocorre nas ***n*** repetições.

Permutação, arranjo e combinação

Permutação

$$_nP_n! = n!$$

em que *n*! é o fatorial de *n*.

O fatorial de n é dado por:

n! = n.(n-1) ... 2.1

1! = 1(o fatorial de 1 é igual a 1, por convenção)

0! = 1 (o fatorial de zero é igual a 1, por convenção)

De quantas maneiras podem permutar as letras abc? Nesse caso, tem-se:

3! = 3.2.1 = 6 possibilidades:

abc, acb, bac, bca, cab, cba.

Arranjo

Quando se tem ***n*** objetos diferentes e se quer permutar ***p*** desses objetos (**p** ≤ **n**), tem-se o arranjo de ***n*** elementos tomados ***p*** a ***p***. Empregando-se o uso do fatorial, tem-se:

$$A_{n,p} = \frac{n!}{(n-p)!}$$

Também se tem:

$$A_{n,n} = \frac{n!}{(n-n)!} = n!/0! = n! = {}_nP_n! = n! \text{ (aqui mostra que } 0! = 1)$$

De quantas maneiras podem permutar as letras *abcd*, duas a duas? Nesse caso, tem-se n = 4 e p = 2.

$$A_{4,2} = \frac{4}{(4 - 2!)} = \frac{4!}{2!} = 4.3.2!/2! = 4 \text{ x } 3 = 12 \text{ possibilidades}$$

Combinação ou análise combinatória

Na combinação, a ordem dos elementos não é levada em conta. Considerando-se o exemplo anterior, ou seja, da combinação das letras *abcd*, duas a duas, tem-se:

$$Cn,p = \frac{n!}{p!(n-p)!} = \frac{4!}{2!(4-2)!} = \frac{4!}{2!2!} = 6 \text{ possibilidades: ab, ac, ad, bc, bd, cd}$$

Os cálculos envolvendo permutação, arranjo e combinação podem ser feitos facilmente por meio de rotinas do SAS. Consideremos a situação em que n = 5, e deseja calcular o fatorial de *n* (n!), o arranjo de *n* tomados 2 a 2 e a análise combinatória de *n* tomados 2 a 2.

```
data;
input x;
fat_x = fact(x);
comb_x = comb(x,2);
permut_x = perm(x,2);
datalines;
5
;
proc print; var x fat_x comb_x permut_x;run;
Output
x   fat_x   comb_x   permut_x
5   120     10       20
```

Teorema da soma e do produto

Teorema da soma

Para quaisquer dois eventos A e B de um espaço amostral, tem-se:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Se os eventos A e B são independentes ou disjuntos, então:

$$P(A \cup B) = P(A) + P(B)$$

Teorema do produto

A probabilidade de ocorrer, simultaneamente, dois eventos A e B que é representada por $P(A \cap B)$, do mesmo espaço amostral, é dada pelo produto das probabilidades individuais:

$$P(A \cap B) = P(A)\,P(B)$$

Algumas propriedades da soma e do produto:

$$P(A \cup B) \leq P(A) + P(B)$$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$P(A \cup B) = P(A) + P(A^C \cap B)$$

$$P(A^C \cap B) = P(B) - P(A \cap B)$$

Se $A \subset B$, então $P(B \backslash A) = P(B) - P(A)$

Se $A \subset B$, então $P(A) \leq P(B)$

Se $A_1, A_2, \ldots A_n$ são conjuntos pareados disjuntos, então $P(A_1 \cup A_2 \cup A_3 \ldots \cup A_n) = P(\bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i)$

$$P(A \cap B) = P(A)\,P(B)$$

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(B \cap C) - P(A \cap C) + P(A \cap B \cap C)$$

Aplicação

O espaço amostral gerado no lançamento de dois dados (Tabela 1) é representado por:

Tabela 1. Espaço amostral no lançamento de dois dados.

X\Y	1	2	3	4	5	6
1	1,1	1,2	1,3	1,4	1,5	1,6
2	2,1	2,2	2,3	2,4	2,5	2,6
3	3,1	3,2	3,3	3,4	3,5	3,6
4	4,1	4,2	4,3	4,4	4,5	4,6
5	5,1	5,2	5,3	5,4	5,5	5,6
6	6,1	6,2	6,3	6,4	6,5	6,6

Tem-se S = {(1,1); (1,2); ...; (6,6)} = 36 resultados.

Sejam os eventos:

ε_1) Resultar soma 5:ε_1 = {(1,4); (4,1); (3,2) e (2,3)} = 4 resultados possíveis:

P (ε_1) = P(1,4) + P(2,3) + P(3,2) + P(4,1)= 4/36 = 1/9

ε_2) Resultar soma $\geq$ 6:

P (ε_2) = 26/36 = 13/18 = 0,72

ε_3) Resultar soma $\geq$ 13:

P (ε_3) = 0/36 = 0 (evento impossível, pois não existem pares (i,j), em que i+j $\geq$13).

ε_4) Resultar soma $\geq$ 2:

P (ε_4) = 36/36 = 1(evento certo, pois todos os pares (i,j) têm soma i+j $\geq$ 2).

ε_5) Obter um par de 6 (eventos independentes):

P(A) = 1/6; P(B) = 1/6

P(A) = 1/6; P(B) = 1/6; P(A$\cap$ B) = P(A) P(B) =1/6.1/6 = 1/36

ε_6) Probabilidade da soma dos pontos dos dois dados Z = X + Y (Tabela 2)

Tabela 2. Somas e respectivas probabilidades dos pontos do lançamento dos dois dados.

Z	2	3	4	5	6	7	8	9	10	11	12
P(Z)	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

Aplicação

Têm-se dois baralhos de 52 cartas e extrai-se uma carta de cada um. Qual a probabilidade de que pelo menos uma delas seja rei de paus? Em um baralho, existe apenas um rei de paus, e, portanto, nos dois baralhos existem dois reis de paus:

R_1: extrair um rei de paus do 1º baralho.

R_2: extrair um rei de paus do 2º baralho.

$P(R_1 \cup R_2) = P(R_1) + P(R_2) - P(R_1 \cap R_2)$

$P(R_1 \cup R_2) = 1/52 + 1/52 - (1/52)^2$

Uma urna possui cinco bolas vermelhas (v_1), quatro brancas (b_1) e seis pretas (p_1); uma segunda urna possui seis bolas vermelhas(v_2), cinco brancas (b_2) e sete pretas(p_2). Extraindo-se uma bola de cada urna, qual a probabilidade de que ambas sejam da mesma cor?

$P(b_1) = 4/15$ e $P(b_2) = 5/18$; $P(b_1 \cap b_2) = P(b_1)\,P(b_2) = (4/15) \times (5/18)$

$P(v_1) = 5/15$ e $P(v_2) = 6/18$; $P(v_1 \cap v_2) = P(v_1)\,P(v_2) = (5/15) \times (6/18)$

$P(p_1) = 6/15$ e $P(p_2) = 7/18$; $P(p_1 \cap p_2) = P(p_1)\,P(p_2) = (6/15) \times (7/18)$

$P((b_1 \cap b_2) \cup (v_1 \cap v_2) \cup (p_1 \cap p_2)) = P(b_1)P(b_2) + P(v_1)P(v_2) + P(p_1)P(p_2) =$

$= 2/27 + 1/9 + 7/45 = 0{,}34$

Probabilidade condicional

Se dois eventos A e B estão associados ao ensaio ε, denota-se por P(B|A) a probabilidade condicionada do evento B, dado que o evento A tenha ocorrido. Sempre que se calcula P(B|A), calcula-se P(B), em relação ao espaço amostral reduzido A. No diagrama de Venn (Figura 8), a probabilidade correspondente ao evento A é ilustrada pela área da elipse pontilhada, enquanto a probabilidade do evento B é dada pela área da elipse contínua. A probabilidade de ocorrências simultâneas é dada pela área da interseção entre as duas elipses (A⋂B). Assim, a probabilidade de A, dado que B tenha ocorrido P(A|B), é calculada pela razão da intersecção entre A e B dividida pela área de B, ou em notação matemática: $P(A|B) = P(A \cap B)/P(B)$. Naturalmente, é possível usar o mesmo raciocínio para calcular a probabilidade de B dado que A tenha ocorrido, isto é, $P(B|A) = P(A \cap B)/P(A)$.

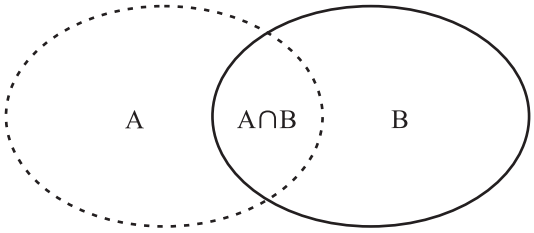


Figura 8. Diagrama de Venn ilustrando probabilidade condicional.

Algumas propriedades da probabilidade condicional:

$P(A \cap B) = P(A)\ P(B|A)$, com intersecção

$P(A \cap B) = P(A)\ P(B)$, sem intersecção

$P(B|A) = P(A \cap B) / P(A)$

$P(A|B) = P(A)$, quando A e B são independentes

$P(B|A) = P(B)$, quando A e B são independentes

$P(A \cap B) = P(B)P(A|B)$

$P(A \cap B \cap C) = P(C|A \cap B)P(B|A)P(A)$

Aplicação

Considerando-se o lançamento de dois dados, o espaço amostral é representado por 36 pontos (x, y), igualmente possíveis. Sejam os eventos:

$A = \{(x_1, x_2) | \ x_1 + x_2 = 10\} \Rightarrow A = \{(5,5); (4,6); (6,4)\}$

$B = \{(x_1, x_2) | \ x_1 = x_2\} \Rightarrow B = \{(1,1); (2,2); (3,3); (4,4); (5,5); (6,6)\}$

Assim, $P(A) = 3/36$; $P(B) = 6/36$ e $P(B/A) = 1/3$, pois o espaço amostral é representado por A (três resultados) e apenas um deles (5,5) pertence também a B.

Aplicação:

Em um baralho com 52 cartas, tem-se quatro naipes (ouros, copas, espadas e paus); e cada naipe tem 13 cartas (C1 a C10, mais três cartas de figuras: valete = CV, dama = CD e rei = CR). Obtêm-se os eventos de E_1 a E_4 de acordo com o seguinte esquema:

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	CV	CD	CR
Ouro	E_2	E_2	E_2	E_2	E_2		E_3						
Copas					E_1		E_3						
Espada							E_3				E_4	E_4	E_4
Paus							E_3						

E_1) Saiu Copas, qual a probabilidade de ser 5?

$P(5/Copas) = P(5 \cap Copas)/ P(Copas) = (1/52) / (13/52) = 1/13$

E_2) Saiu Ouro, qual a probabilidade de ser ≤ 5?

$P(\leq 5/Ouros) = P(\leq 5 \cap Ouros)/ P(Ouros) = (5/52) / (13/52) = 5/13$

E_3) Qual a probabilidade de sair 7, independentemente do naipe?

$P(C7) = P(Ouro \cap C7) + P(Copas \cap C7) + P(Espada \cap C7) + P(Paus \cap C7) = 4/52$

E_4) Saiu Figura, qual a probabilidade de ser Espada?

$P(Figura /Espada) = P(Figura \cap Espada)/P(Espada) = (3/52) / (13/52) = 3/13$

Em um experimento de campo, o objetivo é verificar o efeito de cinco doses de adubo na produtividade de forragem tropical. As doses do adubo utilizadas são: 0 kg ha^{-1}, 50 kg ha^{-1}, 100 kg ha^{-1}, 150 kg ha^{-1} e 200 kg ha^{-1}, denominadas, respectivamente, de D_1, D_2, D_3, D_4 e D_5. Inicialmente, é escolhida uma faixa de solo mais homogênea possível em termos de fertilidade para conduzir o experimento. Entretanto, verifica-se, por meio de análise, que mesmo escolhendo o solo mais homogêneo possível, há um gradiente crescente de fertilidade, que varia de 1 (solo menos fértil) até 5 (solo mais fértil).

Aplicação

Seguindo um dos princípios básicos da experimentação, que é a casualização, as cinco doses (D_1, D_2, D_3, D_4 e D_5) são sorteadas e distribuídas ao acaso nas cinco áreas (parcelas). O objetivo é dar a mesma oportunidade para que qualquer das doses possa ser alocada em qualquer das parcelas, evitando-se, com isso, que doses altas de adubo sejam alocadas, preferencialmente, em áreas mais férteis e vice-versa.

Se as doses de adubo sorteadas são distribuídas, pela ordem, nas parcelas 1, 2, 3, 4 e 5, qual é a probabilidade de as doses mais altas de adubo serem distribuídas nas áreas mais férteis, ou seja, as doses D_1, D_2, D_3, D_4 e D_5 serem distribuídas nas parcelas 1, 2, 3, 4 e 5, respectivamente?

Para responder essa questão, cinco eventos (E_i) são formulados:

- Probabilidade de um evento P(E):

E_1) Sair a D_5 no primeiro sorteio:

$P(E_1) = 1/5$

E_2) Sair a D_4 no segundo sorteio:

$P(E_2) = 1/4$

E_3) Sair a D_3 no terceiro sorteio:

$P(E_3) = 1/3$

E_4) Sair a D_2 no quarto sorteio:

$P(E_4) = 1/2$

E_5) Sair a D_1 no quinto sorteio:

$P(E_5) = 1$

Solução

Como os eventos são independentes, a probabilidade de as doses mais altas de adubo serem distribuídas nas áreas mais férteis é dada por:

$$P(E_1 \cap E_2 \cap E_3 \cap E_4 \cap E_5) = P(E_1)\, P(E_2)\, P(E_3)\, P(E_4)\, P(E_5) = (1/5)(1/4)(1/3)(1/2)(1) = 1/120$$

Aplicação

Uma região apresenta, durante os sete dias da semana, situações bastante variadas com relação às condições climáticas, cuja probabilidade de ocorrência dessas condições, em cada dia, está dentro do parêntese:

Dia 1 → calor, ensolarado, temperatura de 18 °C a 35 °C (0,20).

Dia 2 → calor, nublado, vento, temperatura de 18 °C a 28 °C (0,10).

Dia 3 → calor, nublado, chuvas abundantes, temperatura de 18 °C a 30 °C (0,15).

Dia 4 → calor, nublado, temperatura de 18 °C a 28 °C (0,20).

Dia 5 → frio, vento, nublado, chuvas esparsas, temperatura de 15 °C a 30 °C (0,10).

Dia 6 → frio, vento, ensolarado e temperatura de 15 °C a 25 °C (0,20).

Dia 7 → frio, ensolarado e temperatura de 13 °C a 25 °C (0,05).

Eventos e respectivas probabilidades:

E_1) Encontrar temperatura acima de 25 °C:

$P(E_1) = 5/7$, pois essa situação ocorre nos dias de 1 a 5.

E_2) Encontrar dia frio:

$P(E_2) = 3/7$, pois ocorre nos dias 5, 6 e 7.

E_3) Encontrar dia com calor:

$P(E_3) = 4/7$, pois ocorre nos dias de 1 a 4.

E_4) Encontrar dia ensolarado:

$P(E_4) = 3/7$, pois ocorre nos dias 1, 6 e 7.

E_5) Encontrar dia com tempo nublado ou com chuvas:

- Existem quatro dias em que ocorre tempo nublado (2, 3, 4, 5) com probabilidade 4/7; dois dias em que ocorrem chuvas (3, 5) e dois dias em que ocorrem tempo nublado e chuva, simultaneamente, (3 e 5), ambos com probabilidade de 2/7. Logo,

$P(E_5)$ = P(templo nublado) + P(templo chuvoso) - P(chuva $\cap$ nublado) =

$P(E_5) = 4/7 + 2/7 - 2/7 = 4/7$

E_6) Encontrar dia com tempo nublado **e** com chuvas:

$P(E_6) = 2/7 = 0{,}28$

Teorema de Bayes

Em 1762, o reverendo Thomas Bayes (Wikipédia, 2019j) demonstrou um procedimento bastante importante para se calcular a probabilidade de um evento, dado que outro evento tenha ocorrido. A fórmula simplificada da aplicação desse teorema é dada por:

$$P\,(A/B) = \frac{P(B/A)P(A)}{P(B)}$$

$$P\,(B/A) = \frac{P(A/B)P(B)}{P(A)}$$

em que:

P(A) e P(B) são probabilidades a priori de A e B.

P(A/B) e P(B/A) são probabilidades a posteriori de A e B, indicando, respectivamente, a probabilidade condicional de A dado B e a probabilidade condicional de B dado A.

O uso do teorema de Bayes nos mostra que a probabilidade de um evento ou informações futuras que estamos calculando depende de informações atuais acrescidas às informações já existentes. Observando-se as fórmulas acima, P(A) e P(B) indicam as informações já existentes ou o que pensávamos sobre o evento (probabilidade a priori); P(B/A) indica as informações atuais do evento (verossimilhança), enquanto P(A/B) indica as informações futuras (probabilidade a posteriori).

Assim, é importante definirmos aqui o princípio da verossimilhança:

- Sabemos que os parâmetros θ que caracterizam uma população geralmente são desconhecidos e, portanto, na prática, trabalhamos com uma amostra e acreditamos que ela possui informação suficiente para obter esses parâmetros. O princípio da verossimilhança está baseado na ideia de que diferentes populações geram amostras diferentes e que é mais provável que determinada amostra venha de determinada população do que de outra. Isto é, há condição de saber de qual população uma amostra pertence.

Assim, se uma variável aleatória x tiver uma distribuição de probabilidade f(x) caracterizada por K parâmetros θ e se observamos uma amostra x_1, x_2, ..., x_n, então os estimadores de máxima probabilidade de θ serão aqueles valores dos parâmetros que geram, mais frequentemente, a amostra observada. Em outras palavras, os estimadores de máxima probabilidade de θ serão aqueles valores para os quais a probabilidade (ou densidade de probabilidade) de determinado conjunto de valores amostrais está no máximo, isto é, para acharmos os estimadores de máxima probabilidade dos parâmetros de θ, temos de achar aqueles valores que maximizam $f(x_1, x_2, ..., x_n)$.

Para ilustrar o teorema de Bayes consideremos o diagrama de Venn (Figura 9) a seguir com três eventos independentes: B_1, B_2 e B_3, com cada B_i representando uma divisão do espaço amostral total S. No diagrama, tem-se $P(B_i) > 0$ e o evento A, com $P(A) > 0$, que também corresponde ao espaço amostral total S. Tudo o que será discutido a seguir é válido para n eventos independentes B_1, B_2,..., B_n.

Inicialmente, é fundamental compreender o conceito de probabilidade total.

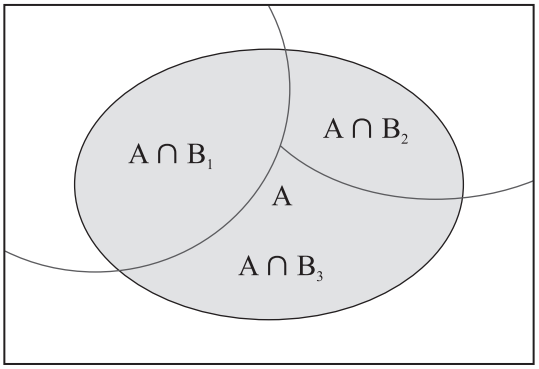


Figura 9. Teorema de Bayes com três eventos independentes: B_1, B_2 e B_3.

Probabilidade total

Considerando-se os três eventos independentes B_1, B_2, B_3 da Figura 9 representando a partição do espaço total S, tem-se:

$B_i \cap B_j = \varnothing$, para qualquer $i \neq j$

$\bigcup_{i=1}^{3} B_i = B_1 \cap B_2 \cap B_3 = S$

$P(B_i) > 0$, para todo i

Como os $A \cap B_i$ são mutuamente excludentes, dois a dois, pode-se aplicar a propriedade da adição de eventos:

$$P(A) = P(A \cap B_1) + P(A \cap B_2) + P(A \cap B_3)$$

$$P(B) = P(B_1) + P(B_2) + P(B_3)$$

Mas, $P(A \cap B_i) = P(B_i/A)\, P(A) = P(A/B_i)\, P(B)$

Assim, tem-se:

$$P(A) = P(B_1/A)\, P(A) + P(B_2/A)\, P(A) + P(B_3/A)\, P(A)$$

Ou

$$P(A) = P(A/B_1)\, P(B_1) + P(A/B_2)\, P(B_2) + P(A/B_3)\, P(B_3)$$

Considerando-se o diagrama anterior com os três eventos independentes B_1, B_2 e B_3, a expressão do teorema de Bayes fica:

$$P(B_i/A) = \frac{P(A/B_i)P(B_i)}{P(A/B_1)P(B_1)+P(A/B_2)\,P(B_2)+P(A/B_3)P(B_3)}$$

Aplicação

Uma cooperativa vende queijos produzidos em três municípios: B_1, B_2 e B_3. O município B_1 é responsável por 45% da produção: o município B_2 é responsável por 30%; e o município B_3 por 25%. Cada um dos municípios, no entanto, produz uma percentagem de queijos que não atende aos controles sanitários. Tais produtos são considerados defeituosos e correspondem a 1,2%, 1,0% e 1,5%, respectivamente, dos totais produzidos nos municípios B_1, B_2 e B_3. Na cooperativa, é feito um controle de qualidade da produção combinada de queijos dos três municípios.

a) Qual é a probabilidade de encontrar um produto defeituoso durante a inspeção de qualidade?

Solução

Sejam os eventos:

A = {o produto é defeituoso}

B_i = {o produto pertence ao município i}

De acordo com o enunciado, tem-se:

$P(B_1) = 0,45$

$P(B_2) = 0,30$

$P(B_3) = 0,25$

$P(A/B_1) = 0,012$

$P(A/B_2) = 0,010$

$P(A/B_3) = 0,015$

Pela lei da probabilidade total, podemos calcular qual é a probabilidade de encontrar um produto defeituoso na cooperativa, independente de qual município ele é produzido:

$$P(A) = P(A/B_1)\,P(B_1) + P(A/B_2)\,P(B_2) + P(A/B_3)\,P(B_3) =$$

$$P(A) = 0,012 \times 0,45 + 0,010 \times 0,30 + 0,015 \times 0,25 = 0,01215$$

Pela lei da probabilidade total, temos que 1,22% dos produtos apresentam defeitos.

b) Se durante a inspeção encontramos um produto defeituoso, qual é a probabilidade de que ele tenha sido produzido no município B_3?

$P(B_3/A) = P(A/B_3)P(B_3)/P(A) = 0,015 \times 0,25/0,01215 = 0,308$

Exercícios[4]

1) No texto a seguir, existem afirmações incorretas quanto a conceitos de estatística. Reescreva o texto colocando definições corretas e sublinhe ou coloque em negrito onde houve definições incorretas.

Distribuição de probabilidade − descreve os valores e probabilidades que uma variável aleatória discreta ou contínua pode assumir. Os valores cobrem todos os resultados possíveis de um evento, enquanto a probabilidade total precisa somar exatamente 1 ou 100%. Três leis fundamentais da probabilidade são: a) a probabilidade de um evento A é um número real, isto é, -1 < P(A) < 1; b) a probabilidade de um evento certo é 1; c) a probabilidade de um evento impossível é zero: $P(\phi) = 0$. O conceito de fatorial, isto é, o fatorial de n elementos é dado por n! = n. (n-1) ... 2.1.0; o arranjo de n elementos tomados p a p é dado por: $A_{n,p} = n!/(n-p)!$. Já na análise combinatória, tem-se que a combinação de n elementos tomados p a p é dada por $\frac{n!}{p!(n-p)!}$. A união de dois conjuntos A e B resulta em um novo conjunto, o qual contém todos os elementos que pertencem a esses dois conjuntos. Assim, se um conjunto A tem quatro elementos e um conjunto B tem cinco elementos, obrigatoriamente o conjunto resultante de A U B terá nove elementos. Um modelo estocástico ou modelo probabilístico é aquele que contém variáveis aleatórias, e os resultados não são previamente conhecidos; entretanto, nos modelos determinísticos, os resultados podem ser previamente conhecidos ou não.

2) O espaço amostral (Ω), definido em cada experimento abaixo, está correto?

E_1: atirar uma moeda quatro vezes e observar a face voltada para cima (c = cara, k = coroa).

E_2: jogar um dado e observar o número mostrado na face voltada para cima.

E_3: escolher três peças, ao acaso, da produção diária de uma linha de montagem e classificá-las como defeituosas (D) e não defeituosas (N), de acordo com a ordem.

[4] As respostas dos exercícios podem ser consultadas no Apêndice 1.

E4: observar o sexo (F = fêmea; M = macho) quanto ao nascimento de duas crianças (não gêmeas) em uma família.

3) Com relação ao exercício anterior, calcular as probabilidades dos eventos:

E_1: ocorrência de pelo menos duas caras.

E_2: ocorrência de face par no lançamento do dado.

E_3: Pelo menos duas peças não defeituosas.

4) Dados os conjuntos:

$A = \{x/\ 0 \leq x \leq 1\}$

$B = \{x/\ ½ \leq x \leq ¾\}$

$C = \{x/\ x^2 + 1 = 0\}$

$D = \{x/\ ½ \leq x < 2\}$

Determinar:

a) $A \cup B$; b) $(A \cap B)$; c) $(B \cup C)$; d) $B \cap D$.

5) Uma região apresenta, durante os 7 dias da semana, condições bastante variadas com relação às condições meteorológicas, cuja probabilidade de ocorrência em cada dia está dentro do parêntese:

Dia 1 → calor, tempo limpo, temperatura de 18 °C a 35 °C (0,15).

Dia 2 → calor, nublado, vento, temperatura de 18 °C a 28 °C (0,10).

Dia 3 → calor, nublado, chuvas abundantes, temperatura de 18 °C a 30 °C (0,15).

Dia 4 → calor, nublado, temperatura de 18 °C a 28 °C (0,20).

Dia 5 → frio, vento, nublado, chuvas esparsas, temperatura de 18 °C a 30 °C (0,10).

Dia 6 → frio, vento, tempo limpo e temperatura 18 °C a 25 °C (0,20).

Dia 7 → frio, tempo limpo e temperatura 13 °C a 25 °C (0,10).

Determine a probabilidade dos eventos:

E_1: encontrar temperatura acima de 25 °C, nos dias de 1 a 5.

E_2: encontrar dia frio, nos dias de 5 a 7;

E_3: encontrar pelo menos um dia com calor, nos dias de 1 a 4.

Capítulo 5

Distribuições discretas

Introdução

O estudo das distribuições discretas iniciou-se no século 17 (Boyer; Merzbach, 1996), e uma das primeiras contribuições foi dada pelo francês Blaise Pascal (1623−1662), que propôs o arranjamento triangular dos coeficientes binomiais e associou o estudo da probabilidade com o triângulo aritmético, que, mais tarde, ficou conhecido como triângulo de Pascal. Outra contribuição importante para as distribuições discretas que foi atribuída ao suíço Jacob Bernoulli (1654−1705) é a prova de Bernoulli ou experimento Bernoulli, cujo resultado é aleatório e pode ser qualquer de duas possibilidades: sucesso ou falha (Wikipédia, 2019e).

As distribuições discretas mais importantes são a distribuição binomial e a distribuição de Poisson. A notação e a expansão do coeficiente binomial $\binom{n}{k}$, de grande interesse nos estudos das distribuições discretas, foram desenvolvidas por Andreas Freiherr von Ettingshausen (1796−1878). A distribuição de Poisson, também conhecida como lei de Poisson ou lei dos eventos raros, uma vez que está associada a uma amostra grande e probabilidade muita pequena, foi desenvolvida em 1837 e atribuída ao francês Siméon-Denis Poisson (Wikipédia, 2019h).

As aplicações das distribuições discretas são importantes para modelar dados de contagens na experimentação científica, pois várias características ou variáveis não são susceptíveis de medida, mas são classificadas em classes ou categorias, que muitas vezes identificam alguma qualidade, e, por isso, também denominadas de variáveis ou dados qualitativos.

Como exemplos de dados qualitativos na experimentação científica, podem-se citar as pesquisas de campo envolvendo aplicação de questionários, com variáveis dicotômicas de respostas sim ou não, presença ou ausência e, também, respostas de múltipla escolha, dando origem a variáveis categóricas. Nesse último exemplo, têm-se as perguntas com escores e/ou notas variando de 1 a 5 − de 1 (insatisfeito) até 5 (totalmente satisfeito), e escalas de concordância que avaliam a pergunta por meio de graus de aprovação/reprovação (concordo totalmente até discordo totalmente).

Na indústria, a classificação de peças defeituosas, após determinado tempo de uso, pode ser do tipo: 0 − peça defeituosa, 1− peça defeituosa, 2 − peças defeituosas, 2 − ou mais peças defeituosas, entre outras.

Na área da agricultura, grande parte dos dados dos experimentos são avaliados por meio de categorias, classes, notas ou escores. Por exemplo, o sexo dos animais (macho, fêmea), os escores de precocidade de terminação da carcaça (notas de 1 a 5), o grau de severidade de doenças de plantas, variando de nota zero (planta sadia) até nota

5 (planta severamente atacada), o comportamento da produtividade de uma cultivar – 1 (excelente), 2 (ótimo), 3 (bom), 4 (regular), 5 (ruim).

Na área da genética, inúmeros são os exemplos de dados categóricos que são analisados por meio de distribuições discretas: o número de progênies classificado segundo diferentes fenótipos e genótipos, ou classificado segundo as diversas gerações de cruzamento e de retrocruzamento. No entanto, para o leitor compreender o potencial da utilização das distribuições discretas na genética, é fundamental conhecer alguns conceitos e, também, alguns princípios utilizados nessa área, principalmente os da reprodução e das leis da hereditariedade. A seguir, serão descritas algumas informações que facilitarão a compreensão dos princípios fundamentais dessa área e, como consequência, do uso das contribuições das variáveis discretas. Serão apresentados conceitos, definições e exercícios das distribuições discretas: binomial, binomial negativa, Poisson, geométrica, hipergeométrica e multinomial. O leitor compreenderá o emprego dessas distribuições em aplicações isoladas e como ferramentas na análise de variância.

Base genética da hereditariedade

O indivíduo (animal ou planta), para a sua existência, depende de dois fatores ou fenômenos distintos: a genética, que durante o processo da reprodução é transmitida a ele pelos seus parentais, e o ambiente, que é representado pelas demais condições que o influencia durante toda a sua vida a partir de seu nascimento.

Para simplificar, consideremos que o material genético, que é transmitido pelos pais aos filhos, está contido em uma substância chamada ácido desoxirribonucleico (DNA), que está apresentada na Figura 1. O DNA guarda as informações que serão transmitidas por meio de sequências ou blocos de bases de nucleotídeos: adenina, citosina, timina, guanina, cuja ordem dessas bases forma o dicionário dos genes, isto é, forma letras e palavras, que representam instruções, e cada instrução equivale a um gene.

Um gene pode ser definido como um segmento do DNA que participa da formação do caráter ou característica do indivíduo e é conhecido como a unidade fundamental da hereditariedade. O gene está situado em uma posição específica do cromossomo, estrutura que é observada durante as divisões celulares e localiza-se no núcleo da célula. O cromossomo representa uma longa sequência de DNA, em que se localizam os genes; o número de cromossomos é variável de espécie para espécie; por exemplo, 46 no homem, 60 nos bovinos, 40 nos suínos e 24 no arroz; porém, dentro da espécie, o número de cromossomos é constante em todas as células.

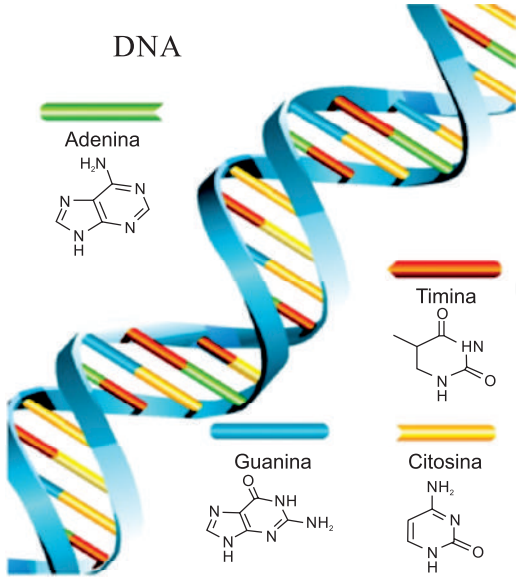


Figura 1. Estrutura do DNA.
Fonte: Biologia (2021).

Dentro das células do indivíduo, os cromossomos ocorrem aos pares, sendo, por isso, chamados de cromossomos homólogos, os quais apresentam a mesma morfologia e são portadores dos mesmos genes. Situados em um mesmo local ou segmento dos cromossomos homólogos, existem os alelos que são formas alternativas de um gene. Os alelos possuem o mesmo número de bases de nucleotídeos. Existe também o alelo múltiplo, isto é, quando um gene é representado por mais de dois alelos.

Durante o processo de fertilização ou fecundação, existe a meiose − processo de divisão celular responsável por reduzir à metade o número de cromossomos das células sexuais, tanto do indivíduo masculino quanto do feminino. Cada metade do número de cromossomos que é transmitida pelo pai e pela mãe ao descendente, durante a fertilização ou fecundação, é chamada de gameta, que contém todos os genes do indivíduo. Aqui surge um conceito muito importante que é o genoma: conjunto haploide de cromossomo que contém todos os genes de uma espécie.

Durante a fecundação, os gametas se unem para formar a célula-ovo ou zigoto, que, por multiplicação e diferenciação, irá formar o novo indivíduo. No homem, o zigoto corresponde à fusão da célula reprodutiva feminina − o óvulo, à célula reprodutiva masculina − o espermatozoide. O zigoto ou ovo contém metade do número de cromossomos do pai e metade do número de cromossomos da mãe, garantindo que o número de cromossomos da espécie permaneça constante de geração a geração.

Existem outros conceitos que são fundamentais:

- Indivíduo homozigoto e indivíduo heterozigoto − Um indivíduo homozigoto é aquele que apresenta alelos iguais de um mesmo gene ou, ainda, que os alelos presentes em um *locus* (local) genético são idênticos. Um indivíduo heterozigoto é o que tem dois alelos diferentes do mesmo gene. Por exemplo, os indivíduos que apresentam os alelos AA ou aa são homozigotos, e aqueles com alelos Aa são heterozigotos.
- Alelo dominante e alelo recessivo − Alelo dominante é aquele que manifesta seu caráter hereditário estando em dose simples ou dupla. Um alelo recessivo é aquele que só manifesta seu caráter hereditário se estiver em dose dupla e na ausência do alelo dominante. Por exemplo, em cães da raça labrador, a cor da pelagem preta e marrom é controlada geneticamente pelo gene B; quando o animal possui os alelos BB e Bb, a cor da pelagem é preta, isto é, o alelo B é dominante; o animal tem pelagem marrom somente quando tem os alelos bb.

Aqui surge também o conceito de genótipo, que é a constituição genética de um indivíduo, ou seja, o indivíduo tem genótipo AA ou Aa ou aa. Grande parte das características dos seres vivos, tais como o peso, é influenciada por grande número de alelos, cada um contribuindo com pequena fração para a característica.

É importante também exemplificar dois conceitos envolvidos no modo de reprodução dos indivíduos em vegetal:

- Autofecundação − Modo de reprodução em que a fecundação se dá entre gametas masculinos e femininos oriundos do mesmo indivíduo, muito comum nos vegetais. O milho, por exemplo, possui flores masculinas (pendão) e femininas (cabelo da espiga) na mesma planta. Para a fecundação, é necessário, portanto, que o grão de pólen do pendão caia no cabelo da espiga, sendo que cada cabelo fecundado vai formar um grão.
- Cruzamento parental − Cruzamento entre dois indivíduos macho e fêmea, geralmente cada um deles pertencente a uma linhagem ou raça pura para um determinado caráter: $P_1 \times P_{2...}$ ou...♂ x ♀.
- Geração filial ou progênies − F_1: primeira geração filial: indivíduos descendentes do cruzamento parental; F_2: segunda geração filial: indivíduos provenientes da autofecundação de indivíduos da geração F_1.

- Retrocruzamento – Cruzamento de um indivíduo F_1 com qualquer um dos seus genitores.

Na Tabela 1 é apresentada a frequência esperada de genótipos com alelos BB, Bb e bb, com as sucessivas autofecundações de um indivíduo de genótipo Bb da geração F_1.

Resumidamente, no caso das distribuições discretas, o espaço amostral é composto de valores susceptíveis de contagem que podem ser classificados em classes ou categorias, sendo que a cada valor ou elemento que corresponde a uma variável aleatória discreta pode associar uma probabilidade.

Tabela 1. Frequências genotípicas esperadas com as sucessivas autofecundações de um indivíduo de genótipo Bb da geração F_1.

Geração	Frequência genotípica		
	BB	Bb	bb
F_1	0	1	0
F_2	¼	½	¼
F_3	3/8	2/8	3/8
F_4	7/16	2/16	7/16
...	...	...	...
F_G	$[1 - (1/2)^{g-1}]/2$	$(1/2)^{g-1}$	$[1 - (1/2)^{g-1}]/2$

Fonte: Ramalho et al. (2005).

Se x é uma variável aleatória discreta com valores $x_1, x_2, ..., x_n$ e com probabilidades associadas $f(x_1)$, $f(x_2)$, ..., $f(x_n)$, então, a função ou distribuição de probabilidades de x é formada com os pares a seguir em que x_i é um evento e $f(x_i)$ é a probabilidade desse evento:

$x_1 \quad f(x_1)$

$x_2 \quad f(x_2)$

...

$x_n \quad f(x_n)$

Se $x_1 < x_2 ... < x_n$, então, a probabilidade cumulativa até um valor k (x_k) é dada por:

$$f(x_k) = f(x_1) + f(x_2) + ... + f(x_k) = \sum_{i=1}^{k} f(x_i)$$

Como exemplo, admitindo-se que o progresso de uma doença em plantas é classificado segundo a atribuição de notas ou escores variando de 0 a 5:

0 − Sem sintomas.

1 − Pequenas manchas ou pontos nas folhas.

2 − Pequenos pontos marrons com poucas conexões entre si.

3 − Lesões de tamanho intermediário.

4 − Poucas lesões grandes.

5 − Muitas lesões grandes.

Na Tabela 2, é apresentado o resumo de 100 plantas classificadas segundo notas ou escores variando de 0 a 5.

Tabela 2. Variável aleatória discreta, x_1, x_2, ..., x_6, representando notas ou escores de 0 a 5 de doenças em plantas com respectivas probabilidades $f(x_1)$, $f(x_2)$, ..., $f(x_6)$, de um espaço amostral consistindo de n = 100 plantas.

X	Elemento do espaço amostral	f(x)
0	30	30/100 = 0,30
1	20	20/100 = 0,20
2	15	15/100 = 0,15
3	15	15/100 = 0,15
4	13	13/100 = 0,13
5	7	7/100 = 0,07

Na análise de dados qualitativos, uma vez formulada a hipótese ou definido o objetivo de estudo, um interesse imediato é identificar a distribuição dos dados. Se o objetivo é calcular a probabilidade de os futuros filhos de determinada vaca serem todos machos ou todos fêmeas, vale lembrar que o sexo do bezerro em cada parto é uma variável aleatória discreta, cuja ocorrência, macho ou fêmea, é influenciada por uma série de mecanismos probabilísticos discretos. Conhecer esses mecanismos probabilísticos significa estudar as distribuições de probabilidades discretas. Neste capítulo, serão abordadas as seguintes distribuições:

- Distribuição binomial.
- Distribuição binomial e regressão logística.

- Distribuição binomial negativa.
- Distribuição de Poisson.
- Distribuição geométrica.
- Distribuição hipergeométrica.
- Distribuição multinomial.

Distribuição binomial

É uma das distribuições mais utilizadas para análise de dados discretos que incluem duas categorias, tais como: o número de insetos vivos e mortos após a aplicação de diferentes dosagens de um inseticida; o número de plantas doentes e sadias; o sexo dos animais (macho e fêmea); os animais castrados e não castrados, o número de locos em homozigose ou heterozigose, etc.

Na distribuição binomial, em uma amostra de tamanho *n*, cada tentativa ou prova é independente entre si e possui apenas dois resultados: sucesso com probabilidade *p* ou falha com probabilidade *q* = 1 - *p*. Cada prova ou tentativa, pelo fato de apresentar dois resultados, é também conhecida como prova de Bernoulli.

Para *n* provas independentes com desejo de *k* sucessos (*k* = 0, 1, ..., *n*), a variável aleatória discreta *x* = 0, 1, ..., *n* tem distribuição binomial com parâmetros *n* e *p* (0 < *p* < 1) dada pela função de densidade de probabilidade (FDP):

$$P(X=k)=\binom{n}{k}p^k(1-p)^{n-k}=\frac{n!}{k!\,(n-k)!}p^k(1-p)^{n-k}$$

Parâmetros e estatísticas descritivas da distribuição:

Média → np.

Variância → npq.

Aplicação

Admitindo-se que frangos para o abate sejam avaliados nas granjas quanto ao peso aos 40 dias de idade, e considerando que, nesta idade, apenas 40% dos frangos devem apresentar peso corporal inferior a 2,2 kg, deseja-se calcular, em uma granja,

a probabilidade de *x* (número de frangos com peso inferior a 2,2 kg) em uma amostra aleatória de dez frangos. Nesse problema, tem-se: n = 10, p = 0,4 e q = 1 − p = 0,6.

A seguir, são apresentados os resultados da probabilidade de *k* sucessos (k = 0, 1, ..., 10), obtidos da FDP da distribuição binomial.

$$P(x=0)=\binom{10}{0}(0,4)^{0}(0,6)^{10-0}=0,00605 \qquad P(x=6)=\binom{10}{6}(0,4)^{6}(0,6)^{10-6}=0,11148$$

$$P(x=1)=\binom{10}{1}(0,4)^{1}(0,6)^{10-1}=0,04031 \qquad P(x=7)=\binom{10}{7}(0,4)^{7}(0,6)^{10-7}=0,04247$$

$$P(x=2)=\binom{10}{2}(0,4)^{2}(0,6)^{10-2}=0,12093 \qquad P(x=8)=\binom{10}{8}(0,4)^{8}(0,6)^{10-8}=0,01062$$

$$P(x=3)=\binom{10}{3}(0,4)^{3}(0,6)^{10-3}=0,21499 \qquad P(x=9)=\binom{10}{9}(0,4)^{9}(0,6)^{10-9}=0,00157$$

$$P(x=4)=\binom{10}{4}(0,4)^{4}(0,6)^{10-4}=0,25082 \qquad P(x=10)=\binom{10}{10}(0,4)^{10}(0,6)^{10-10}=0,00010$$

$$P(x=5)=\binom{10}{5}(0,4)^{5}(0,6)^{10-5}=0,20066$$

Podem-se determinar vários eventos (E) ou provas na amostra de dez frangos:

E_1: nenhum frango possui peso inferior a 2,2 kg.

P (x = 0) = 0,00605 ≈ 0,6%

E_2: um frango tem peso inferior a 2,2 kg.

P (x = 1) = 0,04031 ≈ 4,0%

E_3: dois frangos têm peso inferior a 2,2 kg.

P (x = 2) = 0,12093 ≈ 12,1%

...

E_{10}: dez frangos têm peso inferior a 2,2 kg.

P (*x* = 10) = 0,00010≈ 0,0%

A soma anterior das probabilidades das dez provas é igual a 1. Porém, o cálculo das probabilidades individuais P(*x* = 0), ..., P(*x* = 10), e também as probabilidades acumuladas P(*x* = 0) + P(*x* = 1) + ... + P(*x* = 10) são facilmente obtidas utilizando-se funções do programa SAS.

As estimativas dos parâmetros e estatísticas descritivas associadas a esse experimento são:

Média = np = 4,0000

Variância = npq = 2,4000

Desvio-padrão = $\sqrt{npq}$ = 1,5492

A função pdf('binomial', *x, p, n*) e a função cdf('binomial', *x, p, n*), com parâmetro *n* e *p*, retorna, respectivamente, a probabilidade individual e a probabilidade acumulada para a variável aleatória discreta *x*. O uso dessas duas funções é ilustrado no programa do Sistema de Análise Estatística (SAS) a seguir:

```
data binomial;
p = 0.4; n = 10;
do x = 0 to 10 by 1;
prob = pdf('binomial', x, p, n);
probcum = cdf('binomial', x, p, n);
output binomial; end;
proc print noobs; var x prob probcum; run;
```

output

x	prob	probcum
0	0,00605	0,00605
1	0,04031	0,04636
2	0,12093	0,16729
3	0,21499	0,38228
4	0,25082	0,63310
5	0,20066	0,83376
6	0,11148	0,94524
7	0,04247	0,98771
8	0,01062	0,99832
9	0,00157	0,99990
10	0,00010	1,00000

O termo $\frac{n!}{k!(n-k)!}$ da distribuição binomial indica o número de sucesso ou falha que ocorre. Na amostra anterior de dez frangos, admitindo-se a existência de apenas um frango com peso inferior a 2,2 kg (sucesso), as possibilidades desse evento ocorrer são calculadas por: $\frac{10!}{1!(10-1)!} = 10$. O espaço amostral dessas dez possibilidades, para peso inferior a 2,2 kg (▲) e superior a 2,2 kg (Δ), fica:

▲ Δ Δ Δ Δ Δ Δ Δ Δ Δ
Δ ▲ Δ Δ Δ Δ Δ Δ Δ Δ
Δ Δ ▲ Δ Δ Δ Δ Δ Δ Δ
Δ Δ Δ ▲ Δ Δ Δ Δ Δ Δ
Δ Δ Δ Δ ▲ Δ Δ Δ Δ Δ
Δ Δ Δ Δ Δ ▲ Δ Δ Δ Δ
Δ Δ Δ Δ Δ Δ ▲ Δ Δ Δ
Δ Δ Δ Δ Δ Δ Δ ▲ Δ Δ
Δ Δ Δ Δ Δ Δ Δ Δ ▲ Δ
Δ Δ Δ Δ Δ Δ Δ Δ Δ ▲

O termo $\frac{n!}{k!(n-k)!}$ é também usado para calcular os coeficientes usados na expansão binomial de $p^k(1 - p)^{n-k} = p^k q^{n-k}$. Para n = 10, constrói-se o triângulo de Pascal (Tabela 3) com 11 linhas (n = 0, 1, ..., 10), e, em cada linha, o termo k (k = 0, 1, ..., n) indica o número de coeficientes.

Exemplos:

Linha 0: $n = 0$; $k = 0$ 0!/[0!(0-0)!] = 1

Linha 1: $n = 1$; $k = 0, 1$ 1!/[0!(1-0)!] = 1; 1!/[1!(1-1)!] = 1

Linha 2: $n = 2$; $k = 0, 1, 2$ 2!/[0!(2-0)!] = 1; 2!/[1!(2-1)!] = 2; 2!/[2!(2-2)!] = 1

Tabela 3. Triângulo de Pascal.

Linha (n)	Coeficiente
0	1
1	1 1
2	1 2 1
3	1 3 3 1
4	1 4 6 4 1
5	1 5 10 10 5 1
6	1 6 15 20 15 6 1
7	1 7 21 35 35 21 7 1
8	1 8 28 56 70 56 28 8 1
9	1 9 36 84 126 126 84 36 9 1
10	1 10 45 120 210 252 210 120 45 10 1

Na distribuição binomial, o tipo de simetria depende do valor da probabilidade *p*; se *p* = 0,5, a distribuição é simétrica; se *p* > 0,5, a distribuição é viesada para a direita e, se *p* < 0,5, é viesada para a esquerda. A distribuição binomial tende a se tornar cada vez mais simétrica à medida que o tamanho da amostra *n* aumenta. Para *n* grande e *p* não muito próximo de zero, essa distribuição tem comportamento igual à distribuição normal. Esse fato é consequência do conhecido teorema do limite central. Na Figura 2, são apresentados gráficos da função de distribuição binomial para três tamanhos amostrais (*n* = 10, *n* = 30 e *n* = 50); para *n* = 30 e *n* = 50, os gráficos se assemelham à distribuição normal.

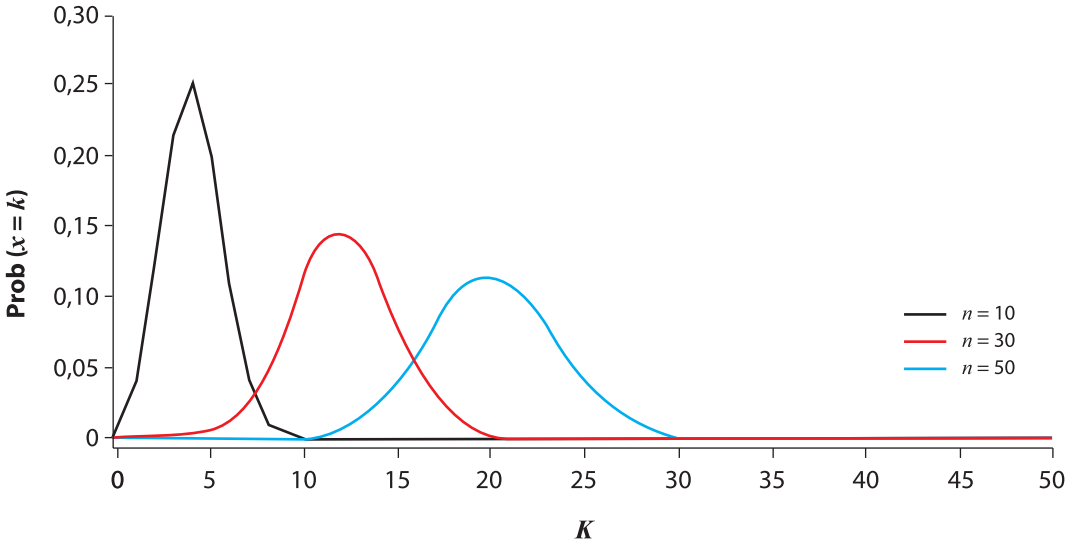


Figura 2. Probabilidades da função de distribuição binomial para *n* = 10, *n* = 30 e *n* = 50.

Na Figura 3, é apresentado o gráfico da função de distribuição binomial acumulada − cdf('binomial', *x, p, n*) e o gráfico da distribuição normal acumulada − cdf('normal', k, μ, σ) para uma amostra *n* = 50. Essas funções retornam a probabilidade acumulada até o maior nível da variável x_i (i = 1, 2, ..., 50). Verifica-se boa concordância entre as duas distribuições para *n* maior que 25. Para fazer corresponder essas duas funções, será usado o fato de que a média (μ) da distribuição normal é igual a *np* = 50 x 0,4 = 20 e $\sigma = \sqrt{npq} = \sqrt{50x0,4x0,6} = 3,4641$. Os dados utilizados para elaboração da Figura 3 são obtidos por meio do programa:

```
data figura3;
do x        = 1 to 50 by 1;
binomcdf = cdf('binomial', floor(i), 0.4, 50);
```

```
normcdf  = cdf('normal', i, 20, 3.4641);
output figura2;
end;
proc print noobs;var x binomcdf normcdf; run;
```

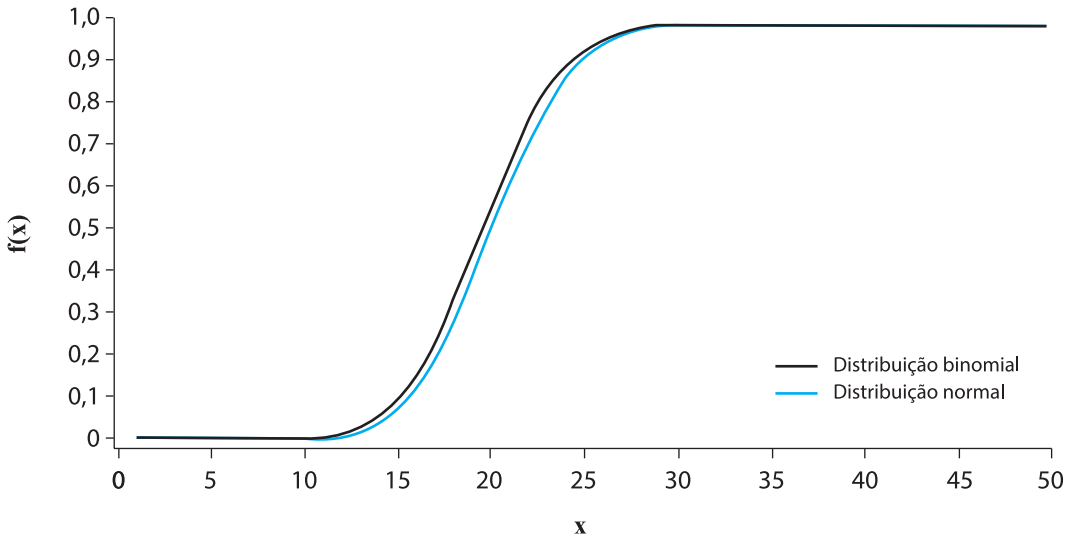


Figura 3. Função densidade de probabilidade acumulada, f(x), da distribuição binomial e normal para *x* até 50.

A distribuição binomial tem grande aplicação também na genética. Um exemplo é o cálculo de frequências genotípicas esperadas em sucessivas autofecundações de um indivíduo heterozigótico, em que o número de locos em homozigose e em heterozigose esperado em cada autofecundação pode ser considerado como sucesso e fracasso, respectivamente. Conforme apresentado na Tabela 1, em *g* gerações de autofecundações, a probabilidade de o loco estar em homozigose é dada por $p = 1- (1/2)^{g-1}$, e a probabilidade do loco de estar em heterozigose é $q = 1- p = (1/2)^{g-1}$. Para *n* locos envolvidos, a probabilidade $k = 0, 1, ..., n$ de locos em homozigose é dada pela FDP:

$$P(x = k) = \binom{n}{k} p^k(1 - p)^{n-k} = \frac{n!}{k!(n - k)!} p^k(1 - p)^{n-k}$$

Aplicação

Seja a autofecundação de um indivíduo até a geração seis, admitindo-se que na geração F_1 ele tem o genótipo AaBbCc. Nesse caso, o número de locos ou de genes envolvidos é três, e o número de locos em homozigose ($k = 0, 1, 2, 3$) é dado por:

- k = 0 loco em homozigose ou os três em heterozigose: AaBbCc.
- k = 1 loco em homozigose ou dois em heterozigose: AABbCc ou AaBBCc ou AaBbCC.
- k = 2 locos em homozigose ou um em heterozigose: AABBCc ou AABbCC ou AaBBCC.
- k = 3 locos em homozigose ou zero em heterozigose: AABBCC.

A frequência esperada de locos em homozigose (AA ou BB ou CC) e em heterozigose (Aa ou Bb ou Cc) na g-ésima geração de autofecundação é obtida, respectivamente, por: $p = 1 - (1/2)^{g-1}$ e $q = 1- p = (1/2)^{g-1}$. Quando se tem um gene, por exemplo o gene A de um indivíduo heterozigótico, o número de gametas diferentes em F_1 é dois (AA, Aa); na geração F_2, o número de gametas possíveis é quatro (AA, Aa, Aa, aa); o número de classes de genótipos diferentes é três (AA, Aa, aa); e o número de genótipos homozigotos é dois (AA, aa). Generalizando, para n genes, o número de gametas diferentes em F_1 será 2^n; e na geração F_2, o número de gametas possíveis será 4^n; o número de classes de genótipos diferentes será 3^n e o número de genótipos homozigotos será 2^n.

Na Tabela 4 é apresentada a frequência esperada de locos em homozigose e em heterozigose até a sexta geração (g = 6) de um indivíduo F_1 com genótipo AaBbCc.

Tabela 4. Frequências esperadas de locos em homozigose e em heterozigose de um indivíduo F_1 com genótipo AaBbCc, autofecundado até a geração seis.

Geração	Homozigose (p)	Heterozigose ($q = 1- p$)
F_1	0,00000	1,00000
F_2	0,50000	0,50000
F_3	0,75000	0,25000
F_4	0,87500	0,12500
F_5	0,93750	0,06250
F_6	0,96875	0,03125

Com as frequências esperadas de locos em homozigose e em heterozigose calculadas até a geração seis, conforme descritas na Tabela 4, a probabilidade do loco estar em homozigose pode ser calculada pela expansão do binômio de Newton $(p+q)^n$, para n = 6, que corresponde à expressão a seguir, em que os coeficientes (1, 6, 15, 20, 15, 6, 1) é calculado por $\frac{n!}{k!(n-k)!}$ e o produto pq pelo desenvolvimento de $p^k(1- p)^{n-k}$:

$$1p^6q^0 + 6p^5q^1 + 15p^4q^2 + 20p^3q^3 + 15p^2q^4 + 6p^1q^5 + 1p^0q^6$$

Para o desenvolvimento dessa expressão, deve-se observar três regras:

a) Todos os termos devem apresentar o produto pq.

b) A soma dos expoentes do produto pq é constante e deve sempre somar n.

c) Os coeficientes da equação correspondem à linha seis do triângulo de Pascal (Tabela 3).

Seja a situação em que se deseja determinar o sexo de nascimento de bovinos em que a probabilidade da progênie ter sexo feminino é $p = 0,5$, e a probabilidade de ter sexo masculino é $q = 0,5$; nesse caso $p = q$. Designando-se o número de fêmeas e de machos por F e M, respectivamente, todas as possibilidades de sexo na progênie são calculadas por $n!/(F!M!)p^Fq^M$, em que os coeficientes da expansão do binômio de Newton são obtidos por $n!/(F!M!)$, e o produto pq pelo desenvolvimento de p^Fq^M. As possibilidades de nascimento de machos e fêmeas, considerando-se dez vacas ($n = 10$) são:

Prob (0 fêmea, 10 machos) = $10!/(0!10!)\ (0,5)^0\ (0,5)^{10} = 0,00098$

Prob (1 fêmeas, 9 machos) = $10!/(1!9!)\ (0,5)^1\ (0,5)^9 = 0,00977$

Prob (2 fêmeas, 8 machos) = $10!/(2!8!)\ (0,5)^2\ (0,5)^8 = 0,04395$

Prob (3 fêmeas, 7 machos) = $10!/(3!7!)\ (0,5)^3\ (0,5)^7 = 0,11719$

Prob (4 fêmeas, 6 machos) = $10!/(4!6!)\ (0,5)^4\ (0,5)^6 = 0,20508$

Prob (5 fêmeas, 5 machos) = $10!/(5!5!)\ (0,5)^5\ (0,5)^5 = 0,24609$

Prob (6 fêmeas, 4 machos) = $10!/(6!4!)\ (0,5)^6\ (0,5)^4 = 0,20508$

Prob (7 fêmeas, 3 machos) = $10!/(7!3!)\ (0,5)^7\ (0,5)^3 = 0,11719$

Prob (8 fêmeas, 2 machos) = $10!/(8!2!)\ (0,5)^8\ (0,5)^2 = 0,04395$

Prob (9 fêmeas, 1 macho) = $10!/(9!1!)\ (0,5)^9\ (0,5)^1 = 0,00977$

Prob (10 fêmeas, 0 macho) = $10!/(10!0!)\ (0,5)^{10}(0,5)^0 = 0,00098$

Regressão logística

Muitas vezes, há necessidade de estudo de variáveis binárias obtidas de situações experimentais; e, nesse caso, a regressão logística será usada para estimar e testar a influência de variáveis regressoras sobre a variável resposta binária. A regressão logística é muito usada na área médica, e várias são as situações em que variáveis explanatórias são avaliadas nos indivíduos em vários tempos ou ocasiões. As variáveis respostas são do tipo binária (0: menor que determinado nível de uma dose; 1: maior que determinado nível de uma dose), parentesco entre medidas químicas do sangue e a condição de diabete (0 = normal; 1 = diabete química).

Na agricultura, a aplicação da regressão logística e transformação *logit* é muito comum. Várias são as situações em que um indivíduo ou unidade experimental pode assumir um de dois valores: sucesso (y = 1) ou não sucesso (y = 0); ocorrência de doença (y = 0: doença presente; y = 1: doença ausente); prenhez de uma fêmea (y = 0: prenhez negativa; y = 1: prenhez positiva); presença de um parasita no animal (y = 0: presente; y = 1: ausente).

Como nos casos clássicos de regressão linear, pode-se fazer um estudo para investigar o parentesco entre variáveis dependentes binárias e variáveis explanatórias. Nessas situações, pode-se realizar a análise envolvendo função do tipo *Probit* ou *logit*, como no exemplo a seguir, com o propósito de modelar a proporção ou probabilidade p de indivíduos que respondem ao estímulo y = 1 (sucesso).

$$\text{logit}(p_i) = \log(p_i/1 - p_i) = \alpha + \beta x_i$$

em que:

p_i = Prob ($y_i = 1 | x_i$) é a probabilidade da resposta y = 1 (sucesso) ser modelada.
α = intercepto.
β = vetor de parâmetros angular.
x_i = vetor de variáveis explanatórias.
log = logaritmo decimal.

Os leitores interessados em maiores detalhes sobre essas análises podem consultar Allison (1999).

Distribuição binomial negativa

A distribuição binomial negativa, também conhecida por distribuição de Pascal, é uma extensão da distribuição binomial. Para uma variável aleatória X e parâmetros *k* e *p*, sua FDP é dada por:

$$P(X = x) = \binom{x-1}{k-1} p^k (1-p)^{x-k} \; ; \; x = k, k+1, k+2, \ldots,$$

em que:

k = número de sucessos desejado.

x = número de ensaios a serem realizados até ocorrer *k* sucessos.

(1 − p) = probabilidade de fracasso em cada ensaio.

A variável *x* pode ser interpretada como o número de tentativas necessárias para ocorrer o primeiro sucesso, mais o número de tentativas para ocorrer o segundo sucesso, etc., mais o número de tentativas para ocorrer o *k*-ésimo sucesso. Quando ocorre o k-ésimo sucesso na x-ésima tentativa, já houve (k - 1) sucessos e (x - 1) falhas, e o número de possibilidades como isso ocorre é dado pela expansão do termo $\binom{x-1}{k-1}$.

Situações ou experimentos em que se deseja um número previamente determinado de *k* sucessos são comuns. Na pesquisa social, pessoas são entrevistadas até encontrar um número de pessoas que atendam a determinados requisitos (sucessos). Na agricultura, são frequentes as situações em que a distribuição binomial negativa pode ser aplicada, e o objetivo é obter sucessos. Por exemplo, quando se realiza amostragem em uma pastagem, em mata, em água de rios, etc., o interesse é determinar a ocorrência de determinada praga ou inseto, um indivíduo contaminado, um indivíduo com doença, ou outra característica qualquer.

Parâmetros e estatísticas descritivas da distribuição:

Média → k/p.

Variância → kq/p^2.

Aplicação

Um pesquisador verifica que determinada espécie de pássaro está em extinção em uma floresta. De acordo com estudos, de cada mil nascimentos, nasce apenas um exemplar da espécie do pássaro que está em extinção. Admitindo-se que o pesquisador analisa amostragens independentes de mil nascimentos cada, pergunta-se:

a) Qual é o número médio de pássaros que necessitam ser amostrados ao nascimento para que três aves em extinção da referida espécie sejam identificadas?

Tem-se:

$x = x1 + x2 + x3$

$k = 3$ (número de sucessos)

Como a probabilidade é p = 1/1.000 = 0,001 e μ = k/p = 3/0,001 = 3.000, então necessitam ser amostrados 3 mil pássaros.

b) Qual é a probabilidade de que três pássaros em extinção sejam encontrados em cinco amostras de mil cada?

Admitindo-se que o primeiro sucesso ocorra na terceira amostra, o segundo na quarta amostra e o último na quinta, tem-se:

$$P(X=x)=\binom{x-1}{k-1}p^{k}(1-p)^{x-k}$$

$$P(x=3)+P(x=4)+P(x=5)=\sum_{x=3}^{5}\binom{x-1}{3-1}(0{,}001)^{3}(1-0{,}001)^{x-1}$$

$$=\binom{2}{3-1}(0{,}001)^{3}(1-0{,}001)^{0}+\binom{3}{3-1}(0{,}001)^{3}(1-0{,}001)^{1}+\binom{4}{3-1}(0{,}001)^{3}(1-0{,}001)^{2}=9{,}97\text{E-}9$$

A probabilidade de que três pássaros em extinção sejam encontrados, isto é, um na terceira, um na quarta e um na quinta amostra de tamanho mil é 9,97E-9 ou 0,00000000997. Esse valor pode ser obtido por meio da função *pdf('negbinomial', m, p, k)*; para cada ensaio, m (m ≥ 0) indica o número de falhas, *p* é a probabilidade de sucesso e *k* é o número de sucessos. O programa SAS, a seguir, ilustra essa aplicação.

```
data;
input m p k;
valor = pdf('negbinomial',m,p,k);
datalines;
3  .001  3
;
proc print; var valor;
run;
output
obs      valor
1        9.97e-9
```

Distribuição de Poisson

É uma distribuição de probabilidade discreta (p) de eventos raros e independentes um do outro, que ocorrem em dado espaço, tempo, volume ou qualquer outra dimensão. Por exemplo, a distribuição do número de ganhadores em loteria de um estado, número de e-mails de determinado assunto que é recebido, número de chamadas de telefone, número de casos de Aids em uma população, número de acidentes mensais em uma rodovia, número de assaltos mensais em uma cidade, número de mutações em uma sequência de DNA de um cromossomo. A distribuição de Poisson é útil em várias áreas da pesquisa. Sua origem data de 1837 e foi atribuída ao francês Siméon-Denis Poisson (1781−1840), conforme Simeon-Denis Poisson (Wikipédia, 2019h).

Essa distribuição tem estreita ligação com a distribuição binomial, e pode substituir esta quando o número de eventos *n* é muito grande e *p* é muito pequeno ($p < 0{,}1$ e $n > 50$). Uma situação típica é o que acontece no caso de uma doença rara em uma população, pois a probabilidade de qualquer pessoa contrair a doença é muito pequena, porém, muitas pessoas da população estão em risco, o que implica *n* grande. Diferentemente da distribuição binomial, as probabilidades da distribuição de Poisson não são conhecidas previamente. A função densidade de probabilidade (FDP) da distribuição binomial é dada por:

$$\frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$$

Assim, para uma variável aleatória discreta, x = 0, 1, ..., n, com *n* tendendo ao infinito e *p* a zero ($n \to \infty$ e $p \to 0$), fazendo $\lambda = np$, obtém-se a distribuição de Poisson a seguir que é um caso particular da distribuição binomial:

$$P(x = k) = (\lambda^k e^{-\lambda})/ k!$$

em que:

e = base do logaritmo natural (*e* = 2.71828...).

k = número de ocorrências de um evento (k = 0, 1, ...).

k! = fatorial de k.

λ = média de eventos em dado intervalo de tempo.

Quando λ aumenta, ambos, média e variância também aumentam, e a distribuição torna mais simétrica.

Parâmetros e estatísticas da distribuição de Poisson:

Média → $\lambda = np$.

Variância → $\lambda = np$.

Aplicação

De cada lote de 300 peças fabricadas em uma indústria uma é defeituosa. Qual é a probabilidade de que, na produção de 600 peças, exista:

a) 0 peça defeituosa.

b) 1 peça defeituosa.

c) 2 peças defeituosas.

d) 2 ou mais peças defeituosas.

A probabilidade de uma peça ser defeituosa em um lote de 300 unidades é $p = 1/300$. Uma vez que a distribuição de Poisson deriva da distribuição binomial e até pode ser compreendida como um caso limite desta quando $p < 0,1$ e $n > 50$, a média é obtida por $\lambda = np = 600x1/300 = 2$.

As probabilidades obtidas da FDP da distribuição de *Poisson* para $k = 0, 1, 2$, são:

$P(x = 0) = (e^{-2}2^0)/0! = 0,1353$

$P(x = 1) = (e^{-2}2^1)/1! = 0,2707$

$P(x = 2) = (e^{-2}2^2)/2! = 0,2707$

Como as probabilidades têm que somar 1, a maneira mais fácil de calcular a probabilidade de obter duas ou mais peças defeituosas é efetuá-la por diferença:

$$P(x > 2) = 1 - (P(x = 0) + P(x = 1) + P(x = 2)) = 1 - (0,1353 + 0,2707 + 0,2707) = 0,3233$$

A função pdf('poisson', *k*, *m*) e a função cdf('poisson', *k, m*), com parâmetro *k* e *m* (média), retorna, respectivamente, a probabilidade individual e a probabilidade acumulada para a variável aleatória $x = k$ ($k = 0, 1, 2, ...$). No programa SAS, a seguir, são calculadas essas probabilidades para $k = 1, 2, ..., 10$, e $m = 2$.

```
data poisson;
do k = 0 to 10 by 1;
prob = pdf('poisson',k,2);
probcum = cdf('poisson',k,2);
output poisson;
end;
proc print noobs; var k prob probcum; run;
output
k     prob     probcum
 0   0,13534   0,13534
 1   0,27067   0,40601
 2   0,27067   0,67668
 3   0,18045   0,85712
 4   0,09022   0,94735
 5   0,03609   0,98344
 6   0,01203   0,99547
 7   0,00344   0,99890
 8   0,00086   0,99976
 9   0,00019   0,99995
10   0,00004   0,99999
```

Como pode ser observado, apenas os dez primeiros valores de *k* de uma amostra de 600 já obtêm somatória de probabilidades igual a 1. As probabilidades individual e acumulada são apresentadas na Figura 4.

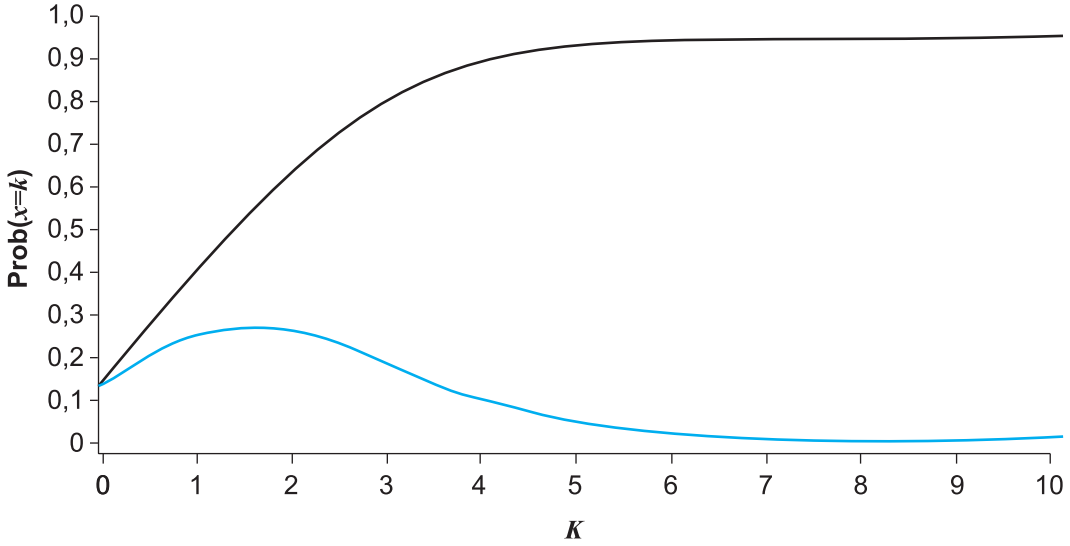


Figura 4. Função densidade de probabilidade acumulada (—) e individual (—) da distribuição de Poisson.

Distribuição geométrica

Esta distribuição pode ser interpretada como duas funções de distribuição de probabilidade discretas:

a) Do número de provas necessárias para ocorrer o primeiro sucesso, representada pelo conjunto ($x = 1, 2, 3, ...,$), então, a FDP da distribuição geométrica com parâmetro *p* é dada por:

$$P(x = k) = (1 - p)^{k-1}p; \quad k = 1, 2, 3, \ ...$$

em que:

k = número de ensaios realizados até ocorrer o primeiro sucesso.

p = probabilidade de sucesso em cada ensaio.

($1 - p$) = probabilidade de fracasso em cada ensaio.

Parâmetros e estatísticas descritivas da distribuição:

Média → $1/p$.

Variância → q/p^2.

b) Do número de falhas $y = x - 1$, antes de ocorrer o primeiro sucesso, representada pelo conjunto (y = 0, 1, 2, ...,), então, a FDP da distribuição geométrica com parâmetro p é dada por:

$$P(y = k) = (1 - p)^k p; \quad k = 0, 1, 2, 3, ...$$

Parâmetros e estatísticas descritivas da distribuição:

Média → q/p.

Variância → q/p^2.

A distribuição geométrica é útil em várias situações. Por exemplo, admita-se que X (x = 1, 2, ..., n) pessoas candidatas ao preenchimento de uma vaga em uma empresa são entrevistadas. Se a primeira pessoa entrevistada é selecionada, então x = 1; se a primeira é rejeitada, mas a segunda é aceita, então x = 2, e assim por diante. Quando $x = n$, significa que as n - 1 pessoas anteriormente entrevistadas foram reprovadas, e o sucesso ocorreu na n-ésima tentativa. A probabilidade de falhar na primeira tentativa é (1 - p); a probabilidade de falhar nas duas primeiras tentativas é (1 - p)(1 - p); e a probabilidade de falhar nas n - 1 primeiras tentativas é $(1 - p)^{(n-1)}$.

Aplicação

Suponha que a probabilidade de um medicamento ser ineficiente no tratamento de determinada doença em bovinos é 0,2. Admitindo-se que todos os animais de um rebanho (obedecendo à ordem crescente do número do animal) são tratados com esse medicamento, qual seria a probabilidade do primeiro resultado ineficiente ocorrer no décimo animal tratado? Nesse exemplo, o sucesso (p = 0,2) é identificar o primeiro animal em que o tratamento não foi eficaz.

$$P(x = 10) = (1 - 0{,}2)^{10-1} \times 0{,}2 = 0{,}02684$$

Esse valor também pode ser obtido por meio da função pdf('geometric', m, p) do SAS, que depende do número de falhas m ($m \geq 0$) e da probabilidade de sucesso p ($0 \leq p \leq 1$). Para o primeiro sucesso ocorrer no décimo animal, significa que houve nove falhas (m = 9). O programa, a seguir, ilustra o cálculo para p = 0,2 e m = 9.

```
data geometric;
input m p;
valor = pdf('geometric',m,p);
cards;
9   0.2
;
proc print; run;
output
 m    p      valor
 9   0.2    0.02684
```

Distribuição hipergeométrica

Uma variável aleatória x segue distribuição hipergeométrica com parâmetros N, *m*, *n*, se a sua FDP é dada por:

$$P(x = k) = \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}}, k = 0, 1, 2, ...$$

Uma maneira fácil de compreender a distribuição hipergeométrica é por meio do estudo clássico de bolas em uma urna. Suponha que em uma urna existem N bolas, divididas em bolas brancas e bolas pretas. Definindo como sucesso (*m*) a retirada de bolas brancas e como insucesso a retirada de bolas pretas, então N − *m* corresponde ao número de bolas pretas na urna.

Assumindo-se que existam 20 bolas brancas e 80 bolas pretas na urna; então, retiram-se 10 bolas, uma a uma, sem reposição. Qual é a probabilidade de que, das dez bolas retiradas, quatro sejam brancas? Embora o evento seja do tipo sucesso/falha, ele não pode ser modelado como binomial, uma vez que, em cada retirada, a probabilidade de obter uma bola branca é alterada. Na Tabela 5 ilustra-se a composição das bolas na urna, possibilitando compreender melhor os parâmetros da distribuição hipergeométrica.

Tabela 5. Exemplos de distribuição hipergeométrica.

	Retirada	Não retirada	Total
Sucesso	k = 4	m - k = 20-4 = 16	m = 20
Falha	n - k = 10 - 4 = 6	N + k - n - m = 100 + 4 - 10 - 20 = 74	N - m = 100 - 20 = 80
Total	n = 10	N - n = 100 - 10 = 90	N = 100

N = população que se divide em dois conjuntos (m, N - m).

N - m = número de falhas presentes na população (bolas pretas).

m = total de sucessos presentes na população ou bolas brancas

(k = retirada; m - k = não retirada).

$\binom{N}{n}$ = número de amostras possíveis de tamanho n.

$\binom{m}{k}$ = número de possibilidades de obter k sucessos.

n - k = número de falhas (retirada).

N + k − n - m = número de falhas (não retirada).

$\binom{N-m}{n-k}$ = número de possibilidades de falha.

$0 \leq m \leq N;\ 0 \leq n \leq N;\ 0 \leq k \leq n;\ 0 \leq k \leq m$

Parâmetros e estatísticas descritivas da distribuição:

$\text{Média} \rightarrow nm/N.$

$\text{Variância} \rightarrow \frac{nm}{N}(1-\frac{m}{N})(\frac{N-n}{N-1})$

Aplicação

Do problema anterior, tem-se:

$$P(x = k) = \mathrm{f}(n, m, N, k) = \frac{\binom{m}{k}\binom{N\text{-}m}{n\text{-}k}}{\binom{N}{n}}$$

$$P(x = 4) = \mathrm{f}(10, 20, 100, 4) = \frac{\binom{20}{4}\binom{80}{6}}{\binom{100}{10}} = 0{,}084107$$

Como o termo $\binom{n}{p}$ significa o arranjo de n elementos combinados *p* a *p*, os cálculos, acima, podem ser obtidos por meio da função comb(N,P) do SAS, conforme programa a seguir, considerando-se n1 = N e n2 = n.

```
data;
input n1 m k n2;
prob2 = (comb(m,k)*comb(n1-m,n2-k))/comb(n1,n2);
datalines;
100  20  4  10
;
proc print; var prob2;run;
output
prob2
0.0841
```

Outro exemplo de aplicação da distribuição hipergeométrica é na prova de Fisher, uma estatística não paramétrica, que é aplicada em análises de dados discretos (nominais ou ordinais) de amostras pequenas que são organizadas em Tabelas 2 x 2. Na Tabela 6, as frequências se enquadram em uma de duas classes mutuamente exclusivas em que os totais marginais são fixos, e os grupos 1 e 2 se referem a quaisquer dois grupos ou tratamentos independentes. A prova de Fisher é recomendável para N < 30 e quando nenhum dos totais marginais é maior que 15. Essa prova avalia se os dois grupos diferem significativamente na proporção de sinais “mais” e “menos” atribuídos a cada um.

Tabela 6. Organização de tabela para aplicação da prova de Fisher.

	-	+	Total
Grupo 1	a	b	a + b
Grupo 2	c	d	c + d
Total	a + c	b + d	n

A prova de Fisher é calculada por:

$$P = \frac{\binom{a+c}{a}\binom{b+d}{b}}{\binom{n}{a+b}} = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{n!a!b!c!d!}$$

Com os dados da Tabela 7 tem-se uma aplicação da prova de Fisher por meio do programa SAS.

Tabela 7. Dados para aplicação da prova de Fisher.

	-	+	Total
Grupo 1	10	2	12
Grupo 2	4	8	12
Total	14	10	24

```
data;
input a b c d n;
prob = comb(a+c,a)*comb(b+d,b)/comb(n,a+b);
datalines;
10  2  4  8 24
;
proc print; var prob;run;
output
prob
0.0166
```

No exemplo anterior, calculou-se a probabilidade de que a proporção de sinais "-" e "+" é igual nos dois grupos. Se os dois grupos fossem iguais, a probabilidade seria próxima de meio. Como a probabilidade calculada foi muito baixa (p = 0,0166), existe forte evidência de que há diferença entre os dois grupos.

Distribuição multinomial

A distribuição multinomial é uma generalização da distribuição binomial, sendo a diferença o fato de que mais de dois resultados são possíveis em cada prova.

Seja $x = (x_1, x_2, ..., x_k)$ uma variável aleatória discreta de dimensão k e n o número de indivíduos total classificados nestas k classes, a probabilidade de que $n_1, n_2, ..., n_k$ indivíduos sejam classificados, respectivamente, nas k classes ou categorias é obtida por:

$$P = (x_1 = n_1, x_2 = n_2, ..., x_k = n_k) = \frac{n!}{n_1!n_2!...n_k!} p^{n_1} p^{n_2} ... p^{nk}$$

em que:

$n_1 + n_2 + ... + n_k = n$, total de indivíduos.

$p_1 + p_2 + ... + p_k = 1$= soma das probabilidades ($p_i \geq 0$).

Parâmetros e estatísticas descritivas da distribuição:

Média → $E(x_i) = np_i$.

Variância → $V(x_i) = np_i(1- p_i)$.

Na distribuição multinomial para qualquer par de frequência x_i e x_j, com probabilidade p_i e p_j, tem-se:

covariância: $cov(x_i, x_j) = -np_ip_j (i \neq j)$

correlação: $corr(ni,nj) = -p_ip_j/[(1-p_i)(1-p_j)]^{-(0,5)}$

Na agricultura, vários são os exemplos nos quais temos interesse em estudar a classificação das variáveis em classes ou categorias. Por exemplo, em doenças de plantas e animais, é comum atribuir notas ou escores do tipo: 1 − excelente; 2 − ótimo; 3 − bom; 4 − regular; 5 − ruim; ou 0 − susceptível; 1 − baixa resistência; 2 − moderadamente resistente; 3 − altamente resistente.

Aplicação

O peso de bezerros, ao nascimento, é influenciado pela mãe, raça, sexo, ano e mês de nascimento, e pode ser considerado como uma forma indireta de avaliar o crescimento pré-natal e, com isso, avaliar a qualidade da mãe, tais como o seu estado nutricional durante a gestação, o peso e o tamanho.

No rebanho bovino Nelore, em condições normais, o peso do bezerro, ao nascimento, pode ser considerado uma variável aleatória real x com média de 30,0 kg

e distribuída no intervalo de 25 kg a 40 kg. Se dividir os resultados esperados em três categorias de pesos, ao nascimento, (Cat 1 a Cat 3), com as três probabilidades abaixo, tem-se:

$$\text{Cat 1} = \{x < 29{,}0\};\ P1 = P(\text{Cat 1}) = 0{,}10$$

$$\text{Cat 2} = \{29{,}0 \leq x < 31{,}0\};\ P2 = P(\text{Cat 2}) = 0{,}80$$

$$\text{Cat 3} = \{x \geq 31{,}0\};\ P3 = P(\text{Cat 3}) = 0{,}10$$

Considerando-se o nascimento, ao acaso, de 100 bezerros Nelore, qual seria a probabilidade (p) de todos os bezerros terem pesos dentro da categoria 2 (29,0 ≤ x < 31,0)?

A probabilidade seria:

$$p = (x_1 = 0, x_2 = 100, x_3 = 0) = \frac{100!}{0!100!0!}\ (0{,}10)^0(0{,}80)^{100}(0{,}10)^0 = 2{,}0370\text{E-}10$$

Exercícios[5]

1) No texto a seguir, existem afirmações incorretas quanto a conceitos de estatística. Reescreva o texto colocando definições corretas e sublinhe ou coloque em negrito onde houve definições incorretas.

As distribuições discretas são importantes para modelar dados de contagens que não são susceptíveis de medida, porém são classificados em classes ou categorias; entretanto, estas distribuições são bastante utilizadas para modelar dados de natureza contínua. Das distribuições discretas, sem dúvida, a mais importante é a de Poisson, também conhecida como lei de Poisson ou lei dos eventos raros, uma vez que está associada a uma amostra pequena e com probabilidade também pequena. Essa distribuição pode ser derivada como um caso limite da distribuição binomial. Diferentemente da distribuição binomial, as probabilidades da distribuição de Poisson são conhecidas previamente. Na distribuição binomial, em uma amostra de tamanho *n*, cada tentativa ou prova é dependente entre si e pode apresentar dois ou mais resultados. A distribuição binomial que é utilizada para análise de dados discretos e que incluem duas ou mais categorias é uma das mais utilizadas. A distribuição binomial tende a se tornar cada vez mais simétrica à medida que o tamanho da amostra *n* aumenta. Verifica-se boa concordância entre ela com a distribuição normal para *n* maior que 10.

[5] As respostas dos exercícios podem ser consultadas no Apêndice 1.

2) Considerando-se que frangos para o abate sejam avaliados nas granjas quanto ao peso aos 40 dias de idade; admitindo-se que, nessa idade, cerca de 60% dos frangos devem apresentar peso corporal superior a 2,2 kg, calcule as três probabilidades, abaixo, em uma amostra aleatória de dez frangos retirados de uma granja.

Considere-se $n = 10$, $p = 0,6$ e $q = 1 - p = 0,4$:

a) Seis frangos possuem peso superior a 2,2 kg.

b) Sete frangos possuem peso superior a 2,2 kg.

c) Os dez frangos têm peso superior a 2,2 kg.

3) Considerando-se a situação da granja descrita na questão 2, calcular as três probabilidades abaixo, em uma amostra aleatória de 10 frangos retirados da granja.

a) Um frango possui peso inferior a 2,2 kg.

b) Dois frangos possuem peso inferior a 2,2 kg.

c) Três frangos possuem peso inferior a 2,2 kg.

4) Se em uma indústria de cada lote de 100 peças fabricadas uma é defeituosa, qual é a probabilidade de que na produção de 200 peças, exista:

a) 0 peça defeituosa.

b) 1 peça defeituosa.

c) 2 ou mais peças defeituosas.

5) Dois grupos de suínos, machos e fêmeas, são submetidos a um tratamento para a cura de uma doença, cujos resultados estão na Tabela 8. Verificar pelo teste de Fisher se o grupo de fêmeas apresentou resultado significativamente melhor do que o grupo de machos.

Tabela 8. Resultado de um tratamento para dois grupos de suínos, machos e fêmeas.

Resultado	Suínos machos	Suínos fêmeas	Total
Teste negativo	7	8	15
Teste positivo	5	1	6
Total	12	9	21

Capítulo 6

Distribuições de probabilidades contínuas

Introdução

No capítulo anterior foram discutidas as distribuições de probabilidade de dados discretos. Entretanto, na natureza e, principalmente, na agricultura, a maioria dos fenômenos se baseia em interpretação de dados contínuos.

Apenas para ilustrar, a maioria dos dados estudados nos seres vivos está incluída nas seguintes categorias de medida: comprimento, largura, altura, peso e percentagem, que geralmente são avaliadas no tempo e no espaço e são de natureza contínua. Alguns exemplos da variável peso são: peso dos animais, do nascimento até a idade adulta, o peso dos frutos por árvore e por hectare; o peso da produção das forragens por vaso, por talhão, por parcela experimental, etc.

Se examinar o peso ao nascimento de animais de uma espécie e ordenar os dados em ordem crescente, observa-se que a variação de peso de um animal para o outro é contínua. Esse comportamento acontece em praticamente todas as características avaliadas nos seres vivos. Diferentemente dos dados discretos, nos dados contínuos não é possível calcular a probabilidade de ocorrer determinado valor, como no evento sair cara ou coroa no lançamento de uma moeda, ou obter um número de 1 a 6 no lançamento de um dado.

Por exemplo, se um bezerro tem peso ao nascer de 35,3 kg, é impossível calcular a probabilidade desse evento, pois ele não é pontual como no caso dos dados discretos. Nos dados contínuos, a variável associada a um evento é a variável aleatória contínua x. Admitindo-se que *x* assuma um valor no intervalo de *a* a *b* (a < b), então a área sob a curva limitada por *a* e *b*, da função de densidade de probabilidade de *x*, é dada por f(x):

$$P\,(a < x < b) = \int_a^b f(x)dx$$

Para uma amostra $x_1,\ldots, x_n$, em que a variável *x* assume qualquer valor dentro do intervalo ($-\infty < x < \infty$), pode-se afirmar que a soma das probabilidades desses valores é igual a 1:

$$P\,(-\infty < x < +\infty) = \int_{-\infty}^{\infty} f(x)dx = 1$$

Uma função de distribuição de probabilidade de uma variável aleatória contínua *x* pode ser compreendida como a relação matemática que fornece, para cada valor da variável, o somatório das probabilidades de todas as ocorrências até aquele ponto. Isso pode ser demonstrado num gráfico cartesiano x-y, em que o eixo *x* (horizontal) expressa os valores da variável aleatória, em ordem crescente, e o eixo *y* (vertical), o valor da função de distribuição (Figura 1).

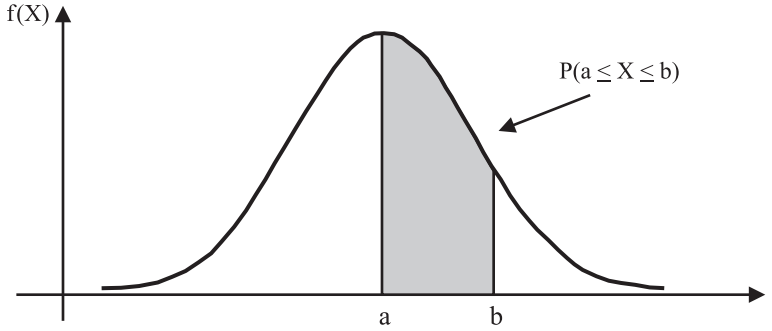


Figura 1. Função de distribuição de probabilidade f(x) da normal.

Neste capítulo, serão estudadas as distribuições contínuas: normal ou gaussiana, normal reduzida, qui-quadrado, F e t de Student. Entre as distribuições de probabilidades contínuas, a mais importante é a distribuição normal ou gaussiana. Quando uma distribuição de probabilidade contínua, como a normal, se ajusta adequadamente a um conjunto de dados, estes podem ser facilmente interpretáveis por meio de uma curva e apenas dois parâmetros: média e variância.

As outras distribuições contínuas citadas, de alguma forma, estão relacionadas com a distribuição normal e, muitas vezes, com o propósito de preencher algumas limitações desta, principalmente quando o tamanho da amostra é pequeno. À medida que o tamanho amostral cresce, quase todas as distribuições apresentam formato semelhante ao da normal. Essa propriedade, na verdade, tem por base o teorema central do limite ou teorema do limite, isto é, quando o tamanho da amostra aumenta, a distribuição amostral da sua média aproxima-se cada vez mais de uma distribuição normal.

Amostras grandes e o teorema do limite central estão associados à lei dos grandes números, a qual tem bastante aplicação na agricultura e em várias outras áreas. Resumidamente, essa lei afirma que, se um evento ou experimento é repetido várias vezes, em ocasiões independentes, a média da probabilidade ou da proporção observada de um dado fenômeno se aproximará do verdadeiro valor. Por exemplo, se não sabemos qual é a probabilidade de chover, ou do número de horas de frio que ocorre em determinado mês numa região, ou ainda a proporção de animais de determinado rebanho ou população que venha a contrair determinada doença, é possível aproximar da probabilidade ou proporção verdadeira trabalhando com um número de observações suficientemente grande.

A distribuição de probabilidade de qui-quadrado, que é uma distribuição de uma soma de quadrados de variáveis aleatórias normal padronizada e independentes, foi inicialmente estudada por Sir Ronald Aylmer Fisher (1890−1962) e por Karl Pearson em 1900 (1857−1936). Esses pesquisadores dedicaram grande parte de suas vidas aplicando essa distribuição e outros métodos estatísticos a problemas biológicos e genéticos. Essas informações têm por base os documentos de Sir Ronald Aylmer Fisher (Wikipédia, 2019i) e Karl Pearson (Wikipédia, 2019f).

A distribuição F, que tem grande aplicação na análise de variância, também ficou conhecida por meio dos trabalhos de Fisher. Já a distribuição t de Student teve origem em uma publicação de 1908 de William Sealy Gosset; porém o conhecido teste t, usado em análises de variância, também ficou conhecido por meio dos trabalhos de R. A. Fisher. Essas informações estão descritas em Zabell (2008). Finalmente, a distribuição exponencial tem grande aplicação para calcular a probabilidade de ocorrência de eventos extremos em dados com forte assimetria.

Essas distribuições são apresentadas a seguir.

Distribuição normal ou gaussiana

A distribuição normal ou gaussiana f(x) de uma variável aleatória contínua *x* tem a sua função de densidade de probabilidade (FDP) dada por:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} dx$$

em que:

x = variável aleatória ($-\infty < x < \infty$).

μ = média de x ou esperança de x.

σ^2 = variância de x.

σ = desvio-padrão de x, $\sigma > 0$.

e = 2,71828... é a base do logaritmo neperiano.

π = 3,1416 (valor de pi).

A FPD da distribuição normal f(x) é caracterizada por dois parâmetros μ e σ^2, de modo que, para obter a probabilidade da ocorrência de uma variável aleatória *x*, depende somente da média e da variância da amostra, uma vez que ***e*** e π são constantes.

A variação de μ não afeta a aparência da curva, simplesmente determina o ponto central; a magnitude de σ^2 é o que determina o aspecto e formato da curva, fazendo com que ela se torne mais ou menos achatada. Cerca de 68,26%, 95,44%, 99,73% e 99,99% da distribuição está compreendida, respectivamente, no intervalo de μ±1σ, μ±2σ, μ±3σ e μ±4σ (Figura 2).

A distribuição normal é a mais utilizada na estatística experimental; entretanto, não é comum encontrar um conjunto de dados que atenda com precisão todas as suas propriedades. O que ocorre é que, em condições experimentais, quando não obedece a alguns princípios (repetição, controle do material experimental e aleatorização), os dados podem apresentar problemas de valores discrepantes (*outliers*), de assimetria e de curtose. Essas anomalias, além de afastar a distribuição dos dados de uma normal exata, podem omitir resultados importantes e gerar viés nas estimativas dos parâmetros.

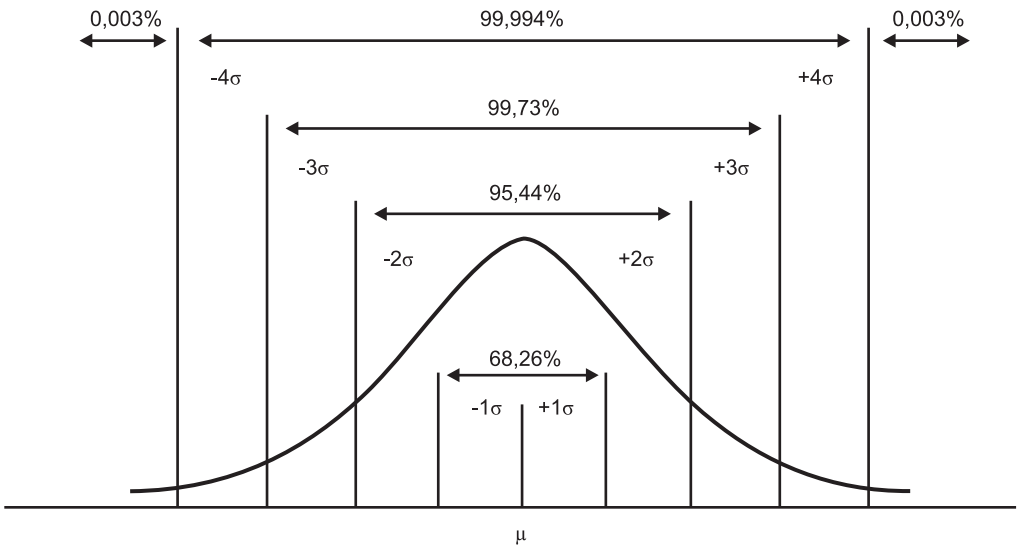


Figura 2. Algumas áreas da distribuição normal.
Fonte: Portal Action (2019).

Na função de distribuição de probabilidade normal, a curva é simétrica em torno da média μ. Esse fato implica que a média, a moda e a mediana são iguais; a altura máxima da curva é atingida no ponto x = μ, ponto que divide a curva em duas áreas exatamente iguais:

a) A curva tem domínio de -∞ a +∞, portanto é ilimitada.

b) Existem dois pontos de inflexão (PI): $x = \mu - \sigma$ e $x_2 = \mu + \sigma$, isto é, pontos em que a curva torna mais achatada e muda de concavidade, significando que o desvio-padrão σ mede a distância do centro da distribuição ($x = \mu$) ao ponto de inflexão.

c) f(x) tem valor máximo no ponto $x = \mu$ e sua ordenada vale $\frac{1}{\sigma\sqrt{2\pi}}$.

d) A área sob a curva é igual à unidade, ou seja, $\int_{-\infty}^{\infty} f(x)dx = 1$; quando x tende a $-\infty$, f(x) tende a zero, e quando x tende a $+\infty$, f(x) tende a 1.

e) Se uma variável aleatória contínua *x* tem distribuição normal com média μ e variância σ^2, a notação fica: $x \sim N(\mu, \sigma^2)$.

f) 95% da área da curva está dentro do intervalo $[\mu - 1{,}96\sigma; \mu + 1{,}96\sigma]$.

Para ilustrar as propriedades da distribuição normal a partir de uma amostra obtida experimentalmente, foram utilizados os dados da produtividade de matéria seca (x) de forragem, em tonelada por hectare (t ha^{-1}), do experimento de adubação nitrogenada, descrito no Capítulo 2.

O arquivo de dados contém 360 observações oriundas da combinação dos fatores experimentais: 3 blocos, 5 capins, 4 doses e 6 cortes (3 x 5 x 4 x 6 =360). A variável *x* variou de zero a 6,88 t ha^{-1}, com estimativas da média de 2,92 kg ha^{-1} e do desvio-padrão de 1,32 kg ha^{-1}.

Na Figura 3 mostra-se a função de densidade de probabilidade f(x) da distribuição normal e acumulada para dados de produtividade de matéria seca, que foi elaborada pela rotina SAS, considerando-se a função *pdf('normal', x,* μ, σ*)* e *cdf('normal', x,* μ, σ*)* em que *x* é a produtividade de matéria seca, μ é a média e σ, o desvio-padrão. A Figura 3 foi elaborada considerando-se que 99% da área da curva está dentro do intervalo $[\mu - 3\sigma; \mu + 3\sigma] = [-1{,}04; 6{,}88]$. O programa do Sistema de Análise Estatística (SAS) usado está descrito a seguir:

```
proc options = reset;
data normal;
do y = -1.04  to 6.88  by 0.05;
prob = pdf('normal', y, 2.9244, 1.3233);
probcum = cdf('normal', y, 2.9244, 1.3233);
output normal; end;
goptions reset = global inborder cback = white ftitle=swissb ftext=swiss htitle=2 htext =
2 csymbol = black;
symbol1 interpol = join  value = dot  height = 1 color = black;
symbol2 interpol = join  value = height=1 color = black;
```

```
axis1 label = (‘x’)   order=(-1 to 7  by 1);
axis2 label = (‘f(x)’)order=( 0 to 1 by .1);
run;
proc gplot;
      plot prob*y probcum*y/ frame
            haxis = axis1  vaxis = axis2 overlay;  run;
```

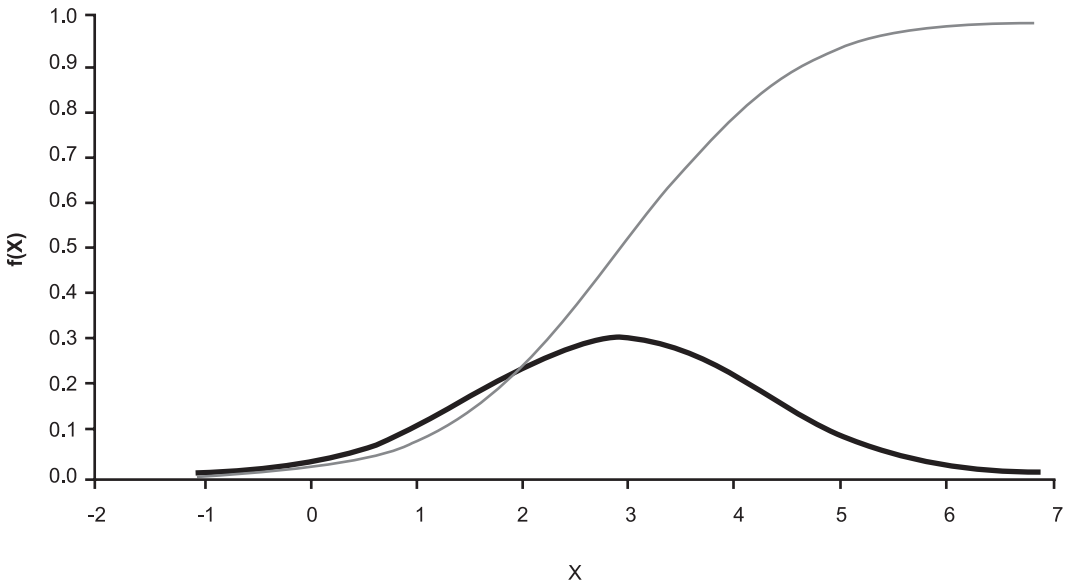


Figura 3. Função de densidade de probabilidade f(x) da distribuição normal individual (**—**) e acumulada (—) para dados de produtividade de matéria seca (*x*) de forragem, em tonelada por hectare (t ha^{-1}).

Considerando-se os dados de produtividade de matéria seca, o valor máximo no ponto x = μ é dado por:

$$\frac{1}{\sigma\sqrt{2\pi}} = \frac{1}{1{,}32\sqrt{2x3{,}1416}} = 30$$

Os dois pontos de inflexão são:

$$x_1 = \mu - \sigma = 2{,}92 - 1{,}32 = 1{,}60$$

$$x_2 = \mu + \sigma = 2{,}92 + 1{,}32 = 4{,}24$$

Como a Figura 3 foi elaborada considerando-se as estimativas da média e da variância da amostra, ela omite certas anomalias nos dados originais, pois, no presente

estudo, o coeficiente de simetria foi de 0,90, indicando que a distribuição dos dados é viesada para a direita; o coeficiente de curtose foi 1,63, significando que o pico da curva é mais pontiagudo do que se espera para uma curva normal. Isso também é comprovado pelas medidas de tendência central em que a média de 2,92 t ha^{-1} de matéria seca foi levemente superior à mediana (2,80 t ha^{-1}).

Distribuição normal reduzida

A distribuição normal, discutida acima, não permite a comparação de diferentes conjuntos de dados com médias e variâncias específicas e muitas vezes dados com unidades de medidas diferentes.

Para solucionar essas limitações da distribuição normal, tem-se a distribuição normal reduzida ou padronizada em que a variável *x* é substituída pela variável aleatória $z = (x - \mu)/\sigma$. Essa distribuição independe de unidades de medida e de parâmetros específicos de cada amostra, como a média e a variância. Para qualquer amostra em particular, pode-se construir um gráfico em que os valores da variável são transformados em probabilidades de 0 a 1 ou de 0 a 100%. Com isso, pode-se comparar a eficiência de experimentos diferentes e, também, determinar a importância ou desempenho de um particular indivíduo em diversas situações.

A FDP f(z) da normal reduzida que também é conhecida por $\Phi(\mathbf{z})$ é dada por:

$$\Phi(z) = 1/\sqrt{2\pi}\int_{-\infty}^{+\infty} e^{-z^2/2}\, dz$$

Propriedades:

a) Notação: se $x \sim N(\mu, \sigma^2)$, então $z = (x - \mu)/\sigma$ tem $\sim N(0, 1)$.

b) Média = moda = mediana = 0.

c) Existem dois pontos de inflexão: z = +1 e z = -1, sendo que a área entre eles é 0,6826.

d) A curva tem forma de sino e a sua área vale 1.

e) A função tem valor máximo no ponto $z = \mu$ e sua ordenada vale $1/\sqrt{(2\pi)} = 0,3989$.

f) A função $\Phi(\mathbf{z})$ é definida para qualquer *z* entre $-\infty$ e $+\infty$, porém $\Phi(\mathbf{z})$ tende a zero quando *z* aproxima de -4, e tende a 1 quando *z* aproxima de +4.

Por exemplo, o uso da normal reduzida N(0,1) possibilita comparar o desempenho de um indivíduo em dois experimentos distintos, necessitando apenas da respectiva

média e do desvio-padrão de cada indivíduo, respectivamente. É possível também comparar a eficiência de uma empresa no Brasil e em outro país.

Aplicação

Um candidato fez exame em dois concursos; no primeiro ele obteve 80 pontos em uma prova cuja média de pontos foi igual a 60 e o desvio-padrão 10; no segundo concurso, em que a média de pontos foi 50 e o desvio-padrão 9, ele obteve 72 pontos. Em qual concurso ele foi mais eficiente?

No primeiro concurso, tem-se z_1= (80 - 60)/10 = 2,0, e, no segundo concurso, tem-se z_2 = (72 - 50)/9 = 2,4. As probabilidades (prob) de z_1 e z_2 são obtidas por meio das integrais abaixo, utilizando o programa SAS. Conclui-se que o candidato, no segundo concurso, teve melhor desempenho, pois teve eficiência de 99,2% (prob de z_1 = 0,9918), enquanto, no primeiro, sua eficiência foi 97,7% (prob de z_2 = 0,9772).

$$\Phi(z_1) = 1/\sqrt{2\pi}\int_{-\infty}^{z_1} e^{-z^2/2}dz \text{ e } \Phi(z_2) = 1/\sqrt{2\pi}\int_{-\infty}^{z_2} e^{-z^2/2}\,dz$$

```
data;
z1 = 2; z2 = 2.4;
prob_z1 = probnorm(z1);
prob_z2 = probnorm(z2);
proc print;
var prob_z1 prob_z2;
run;
output
  prob_z1    prob_z2
 0.9772      0.9918
```

Outra propriedade da distribuição normal reduzida é:

- Se $x \sim N(\mu, \sigma^2)$, então $P(a \leq x \leq b) = P[(a - \mu)/\sigma \leq z \leq (b - \mu)/\sigma]$

Essa propriedade mostra que a probabilidade de variável aleatória *x* e da variável aleatória *z*, dentro do intervalo [a; b], é equivalente. Isso é facilmente calculado com os recursos das funções Φ(a) e Φ(b).

$$\Phi(a) = 1/\sqrt{2\pi}\int_{-\infty}^{a} e^{-z^2/2}\,dz$$

$$\Phi(b) = 1/\sqrt{2\pi}\int_{-\infty}^{b} e^{-z^2/2}\,dz$$

$$P(a \leq z \leq b) = \Phi(b) - \Phi(a)$$

$$P(-z) = 1 - \Phi(z)$$

Aplicação

Se estamos interessados na probabilidade de *z* estar compreendido no intervalo entre 1 e 2, esta pode facilmente ser obtida pela função SAS *probnorm(x)*, a qual retorna a probabilidade de uma observação da distribuição normal padrão ser menor ou igual a *x*. Verifica-se que a probabilidade de *z* estar compreendido no intervalo entre 1 e 2 é:

$$P(1 \leq z \leq 2) = 1/\sqrt{2\pi}\int_1^2 e^{-z^2/2}dz = \Phi(2) - \Phi(1) = 0{,}9772 - 0{,}8413 = 0{,}1359$$

```
data prob;
a = 1; b = 2;
prob1 = probnorm(1);
prob2 = probnorm(2);
prob12 = prob2-prob1;
proc print;run;
output
prob1       = 0,8413
prob2       = 0,9772
prob12      = 0,1359
```

Como visto, pode-se construir a curva de distribuição normal de qualquer conjunto de dados, conhecendo-se a média e a variância. Tudo isso se resume no cálculo da probabilidade da variável aleatória *z*, em várias situações. Essa distribuição está tabelada em vários livros-textos e pode facilmente ser obtida por meio de rotina SAS, como demonstrado. Por exemplo, quando *z* pertence ao intervalo *a* e *b* e queremos saber a área da curva dentro desse intervalo, usamos *probnorm(b) - probnorm(a).* A probabilidade de *z* estar no intervalo de -∞ até a (-∞; a) é calculada por *probnorm(a),* e a probabilidade de *z* estar no intervalo de *b* até +∞ (b; +∞) é dada por: *1 - probnorm(b)*.

Aplicação

Na rotina SAS abaixo, é calculada a probabilidade de *x* estar no intervalo delimitado por z = -1,96 e z = +1.96. Como resultado, tem o limite inferior de 2,5% e o superior de 97,5%, e a área compreendida nesse intervalo é de 95,0%.

```
data prob;
ls = probnorm(1.96);
li = probnorm(-1.96);
prob = ls - li;
proc print;run;
output
li          ls          prob
0.0249   0.9750      0.9500
```

A produção de carcaças de suínos com menor espessura de toucinho (ET) e maior teor de carne magra é a preferida pelas indústrias frigoríficas, as quais geralmente bonificam os produtores de suínos que fornecem matéria-prima de melhor qualidade.

Se um frigorífico decide bonificar produtores para suínos que apresentem ET no máximo de 2,00 cm, qual é a percentagem de suínos que estará apta para atender a essa exigência em uma amostra, cuja média de ET é 1,90 cm e o desvio-padrão é 0,15?

Solução

O primeiro passo é calcular *z* por meio da relação $z = (x - \bar{x})/s$, a qual transforma uma resposta biológica com distribuição normal em uma distribuição padrão *z* que têm ~ N(0, 1).

$$z = (2{,}00 - 1{,}90)/0{,}15 = 0{,}67$$

Consultando-se a Tabela 1.1 (Anexo 1), verifica-se que a área correspondente de um valor inferior a 0,67 é 0,7486. Assim, a probabilidade de encontrar suínos na amostra examinada, no abate, que possuam no máximo 2,00 cm de ET é 74,86%.

A probabilidade de *z* situar dentro do intervalo de -∞ até 0,67 pode ser obtida também usando-se a rotina SAS a seguir:

```
data prob;
prob = probnorm(0.67);
proc print; run;
output
prob
0.7486
```

Em diversas áreas do conhecimento, várias estatísticas são calculadas utilizando-se recursos da distribuição normal padronizada para situações de grandes amostras (n > 30). Um exemplo são as técnicas não paramétricas, que são bastante utilizadas principalmente nas ciências do comportamento. Essas estatísticas são conhecidas como de distribuição livre, pois, ao contrário dos métodos paramétricos, elas não especificam condições sobre os parâmetros da população da qual se extraiu a amostra e, também, são pouco exigentes quanto à mensuração dos dados. As técnicas não paramétricas exigem apenas classificação nominal ou em forma de postos, do tipo, maior do que, menor do que, etc.

A seguir, são apresentados alguns exemplos de técnicas estatísticas não paramétricas, com as respectivas médias e desvios-padrão que, em situações de

grandes amostras, utilizam a distribuição normal padronizada $z = (x - \mu)/\sigma$. Nas técnicas a seguir, *n* é o número de casos; *p* é a probabilidade da variável aleatória *z*, q = 1 - p; n_1 e n_2 indicam, respectivamente, número de casos no grupo menor e no grupo maior.

Tabela 1. algumas métodos estatísticos não paramétricas.

Método	Média: μ	Desvio-padrão: σ
Teste Binomial	np	$(npq)^{1/2}$
Prova dos Sinais	np = n/2	$(npq)^{1/2} = (n)^{1/2} \times 0,5$
Prova de Wilcoxon	n(n+1)/4	$[(n(n+1))(2n+1))/24]^{1/2}$
Prova de Mann-Whitney	$n_1n_2/2$	$[(n_1n_2/(n_1 + n_2 + 1))/12]^{1/2}$
Prova de aleatorização	0	$\sum_{i=1}^{n} d_i^2$ (d_i = escores)

Distribuição de qui-quadrado

Se *v* é a soma do quadrado de *v* variáveis aleatórias, z_1, z_2, ..., z_v, de distribuição normal reduzida, $z_i \sim N(0, 1)$, então *v* tem distribuição de qui-quadrado (χ^2) com v graus de liberdade dada por:

$$v = \sum_{i=1}^{v} z_i^2 \sim \chi_v^2$$

A FDP da distribuição de χ^2 é dada por:

$$f(v) = \frac{1}{2^{n/2}\Gamma(n/2)} v^{n/2-1}e^{-v/2},\ v > 0$$

em que:

$\Gamma(n/2)$ = função gama com argumento (n/2) e Γ(n) = (n-1)!

O formato da distribuição de χ^2 está diretamente associado com seu grau de liberdade (GL); à medida que o tamanho da amostra cresce, a distribuição apresenta formato semelhante ao da normal. Isso pode ser visto na Figura 4.

A seguir, tem-se a rotina SAS utilizada para a construção da Figura 4, para graus de liberdade = 2, 3 e 20.

```
proc options=reset notitle;
data chisq;
keep v y gl;
 do gl = 2, 3, 20;
 k1 = gl/2;
 k2 = 1/((2**k1)*gamma(k1));
      do v =0 to 20 by 0.001;
      h1 = v**(k1-1);
         h2 = exp(-v/2);
         y = k2*h1*h2;
         output;
         end;
  end;
proc gplot data = chisq;
plot y*v = gl/
vaxis = axis1 haxis = axis2;
symbol1 interpol = join    value = none line=20 height=2 color = lack;
symbol2 interpol = join    value = none height=2  color = red;
symbol3 interpol = join    value = dot height=0 color = blue; run;
```

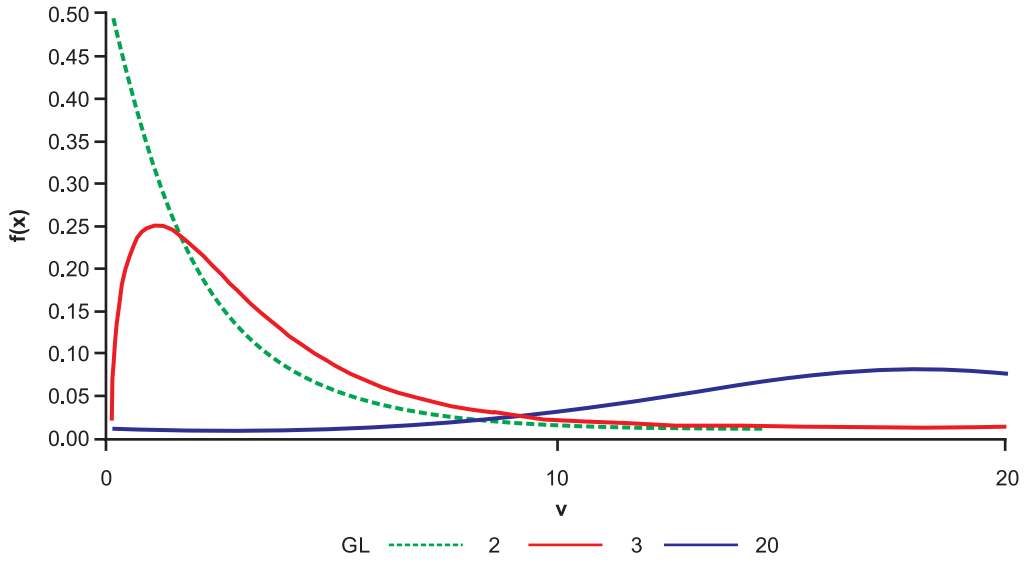


Figura 4. Densidade de probabilidade f(x) da distribuição de qui-quadrado com graus de liberdade (GL) = 2, 3 e 20.

Os valores da variável aleatória *v* com distribuição de χ^2 são transformados em probabilidades de 0 a 1 ou de 0 a 100%. A partir dessa distribuição, tem-se o teste de χ^2 ou estatística de χ^2 que, associada aos respectivos graus de liberdade, tem bastante aplicações na análise de dados categóricos.

A análise de dados categóricos é comum em diversas áreas do conhecimento. O primeiro passo é organizar uma tabela de contingência, com I linhas e J colunas. As linhas e colunas são representadas por variáveis mensuradas no nível nominal e ordinal, tais como raça, sexo, e cada célula n_{ij} contém as frequências de cada combinação das linhas e colunas. Uma das aplicações mais comum da estatística de χ^2 é testar valores de frequências observadas versus valores preditos, em que estes últimos são calculados sob a suposição de que não há associação entre as duas variáveis (linhas e colunas) ou que a diferença entre elas é em razão do acaso.

Na Tabela 2, é apresentado um exemplo de tabela de contingência em que 13 animais com determinada doença foram divididos em dois grupos: vacinados, com oito animais, e não vacinados ou controle, com cinco animais. Após certo período de avaliação, observou-se a resposta dos animais à vacinação.

Tabela 2. Tabela de contingência.

Vacinação	Recuperação		Total
	Sim	Não	
Vacinados	6 (n_{11})	2 (n_{12})	8 (n_{11+} $n_{12=}$ $n_{1.}$)
Controle	1 (n_{21})	4 (n_{22})	5 (n_{21+} $n_{22=}$ $n_{2.}$)
Total	7 (n_{11+} $n_{21=}$ $n_{.1}$)	6 (n_{12+} $n_{22=}$ $n_{.2}$)	13 ($n_{..}$)

A aplicação dessa distribuição está demonstrada no Capítulo 13.

Distribuição F

A distribuição de probabilidade F com n_1 e n_2 graus de liberdade, também conhecida por distribuição F de Snedecor, deriva da razão de duas variáveis (v_1, v_2), tendo ambas a distribuição de qui-quadrado, com n_1 e n_2 graus de liberdade.

$$u = \frac{v_1/n_1}{v_2/n_2} \sim F_{n_1,n_2}$$

Uma variável aleatória contínua *x* tem distribuição F de Snedecor com n_1 graus de liberdade no numerador e n_2 graus de liberdade no denominador se sua função densidade de probabilidade f(x) é definida por:

$$f(x) = \frac{\Gamma(\frac{n_1 + n_2}{2})(n_1/n_2)^{n_1/2} x^{n_2/2-1}}{\Gamma(\frac{n_1}{2})\Gamma(\frac{n_2}{2})[(n_1/n_2)x+1]^{(n_1+n_2)/2}} \quad x \in [0,\infty]$$

A estatística ou teste F é fundamental em análises de variância, principalmente para testar a razão entre a variância do quadrado médio de tratamentos (QMT) e a variância residual (QMR). Comumente, tem-se a notação $F_{n1,n2}$, que é a estatística F obtida da razão entre QMT/QMR, com n_1 e n_2 graus de liberdade para QMT e QMR, respectivamente. O uso da estatística F em análises de variância será demonstrado nos Capítulos 7 a 11.

$$F_{n1,n2} = \frac{QMT}{QMR}$$

Os valores de F para uso em análises de variância encontram tabelados na maioria dos livros de estatística, mas podem ser obtidos facilmente por meio de funções SAS, as quais são associadas a: x, n_1, n_2 e nc, em que:

x = variável aleatória numérica ($x \geq 0$).

n_1 = grau de liberdade do numerador.

n_2 = grau de liberdade do denominador.

<nc> = parâmetro de não centralidade, em que a notação < > indica que nc é optativo e, se não especificado, vale zero.

As funções são:

- *probf(x, n_1, n_2, <nc>)*

Retorna a probabilidade que uma observação de distribuição F é menor ou igual a *x*.

- *p = 1- probf(x, n_1, n_2, <nc>)*

Retorna o nível de significância para o teste F tabelado.

- *finv(p, n_1, n_2, <nc>)*

Calcula o p-ésimo quantil de uma distribuição F.

Na rotina SAS a seguir, é apresentado um exemplo dessas funções. Tem-se um valor de F tabelado com 5 graus de liberdade no numerador e 10 no denominador ($F_{5,10}$ = 3,3258). Usando-se esses valores na função probf(3.3258, 5, 10), ela retorna o valor 0,9500 que corresponde a 0,95 ou 95,0%, que é a probabilidade de obter na FDP um valor menor ou igual a 3,3258.

```
data;
a = probf(3.3258, 5, 10);
b = 1- probf(3.3258, 5, 10);
c = finv(0.95, 5, 10);
proc print; var a b c; run;
output
a          b          c
0.9500  0.0500   3.3258
```

Distribuição t de Student

A distribuição de probabilidade t foi desenvolvida em 1908 pelo inglês William Sealy Gosset (1876–1937), mais conhecido pelo pseudônimo de Student. Essa distribuição é caracterizada pelo parâmetro *v* (graus de liberdade) que define e caracteriza a sua forma; e quanto maior for esse parâmetro, mais a distribuição se aproxima da normal. A distribuição de t tem muitas similaridades com a normal, podendo ser interpretada como uma limitação desta (Wikipédia, 2019L).

Como já discutido nas distribuições z, χ^2 e F, a FDP de t fornece os valores de probabilidades de 0 a 1 ou de 0 a 100%. Dessa distribuição deriva também o conhecido teste t de Student que tem várias aplicações na estatística experimental. Uma suposição requerida para aplicação do teste t é que as variáveis envolvidas sejam intervalares ou de razão e que tenham distribuição normal. O teste t permite testar se um resultado é ou não estatisticamente significativo.

A distribuição t de Student com v graus de liberdade é dada por:

a) $t_v = z/\sqrt{\chi^2_v/v}$

em que:

z = variável aleatória de distribuição normal, com média zero e variância 1.

χ^2_v = distribuição de qui-quadrado (χ^2) com v graus de liberdade.

b) Para uma amostra de n variáveis aleatórias, $x_1, ..., x_n$, com média $\bar{x}$ e variância.

$$s^2_n = \sum_{i=1}^{n} (x_i - \bar{x})^2/(n-1)$$

Pode ser deduzido que:

$$v = (n-1)\frac{s^2}{\sigma^2} \sim \chi^2_{n-1}$$

Pode ser demonstrado ainda que:

$z = (\bar{x}-\mu)\sqrt{n/\sigma}$ tem distribuição normal com média zero e variância 1, considerando que $\bar{x}$ é normalmente distribuída com média μ e variância $\sigma/\sqrt{n}$.

Finalmente, pode-se mostrar que a distribuição de v e z são independentes e que a quantidade $t = (\bar{x} - \mu_0)\sqrt{n/s}$ tem distribuição de Student com n graus de liberdade.

A FDP de t é dada por:

$$f(t) = \frac{(\Gamma[(v+1)/2]}{\Gamma(v/2)\sqrt{\pi v}}(1+\frac{t^2}{v})^{-(v+1)/2}, \; -\infty < t < \infty$$

em que:

f(t) = distribuição t de Student com média zero e variância n/(n-2).

Γ ( ) = função gama.

v = graus de liberdade.

π = 3,1416.

A comparação da distribuição normal e da distribuição de t com um grau de liberdade é apresentada na Figura 5. Observa-se que a distribuição de t apresenta cauda mais longa que a normal, de modo que uma simulação de dados por meio da distribuição de t pode gerar valores mais extremos que uma simulação por meio da normal. Na medida em que aumenta o tamanho da amostra, a distribuição t se aproxima da distribuição normal, que para n > 30, as diferenças são pequenas.

A Figura 5 foi construída pela rotina SAS a seguir:

```
data normal;
do x = -4 to 4 by .1;
pi = 3.1416;
normal_01 = sqrt(2*pi)**(-1)*exp(-1*( (x*x) /2));
dfval =1;
```

```
t_1gl=gamma((dfval+1)/2)/(sqrt(dfval*pi)*gamma(dfval/2))*(1+(x*x)/dfval)**(-
(dfval+1)/2);
output;
 end;
run;
legend1 label = none position = (top inside right) frame down=2 value =    ("n(0,1)" tick
= 2 "t com 1 gl");
axis1 label = (angle=90 "f(x)") minor = none order = (0 to .4 by .1);
axis2 minor = none order = (-4 to  4 by 2);
symbol1 i = j  v = none  l = 1   c = black w = 5;
symbol2 i = j  v = none  l = 21 c = black w = 5;
proc gplot data = normal;
    plot (normal_01 t_1gl) * x / overlay legend=legend1 vaxis = axis1 haxis = haxis2;
run; quit;
```

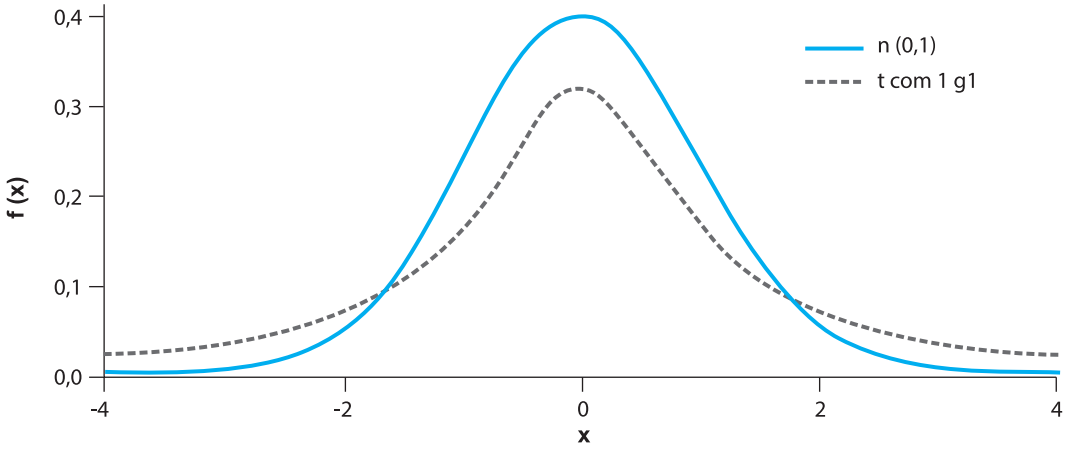


Figura 5. Comparação de duas distribuições de probabilidade, f(x): distribuição normal padrão (▬) e distribuição t com um grau de liberdade (▬ ▬).

A seguir, são apresentadas algumas funções SAS utilizadas na FDP de t:

probt(x,df)

Calcula a probabilidade (0 < p < 1) de uma variável aleatória com distribuição t ser menor ou igual a x, ou seja, calcula a área até a variável x. Para um valor de t de 1,85 com nove graus de liberdade, a função *z = probt(1.85, 9)* retorna p = 0,95132.

tinv(p,df)

Calcula o valor de *t* com probabilidade (0 < p < 1) e grau de liberdade df. Para uma probabilidade p = 0,95 e 9 graus de liberdade, a função t = tinv(0.95, 9) retorna t = 1,85.

Existem várias estatísticas ou quantidades que são distribuídas como uma distribuição t de Student e que são denominadas teste t ou estatística t, o que significa que podemos fazer teste de hipóteses e construir intervalos de confiança dessas estatísticas usando as informações da FDP de t. Essas estatísticas são utilizadas na prática em situações, tais como uma amostra independente, duas amostras independentes, duas amostras de tamanhos iguais e pareadas.

Distribuição exponencial

Uma variável aleatória contínua *x* tem distribuição exponencial com parâmetro $\lambda > 0$, se a sua FDP f(x) é dada por:

$f(x) = \lambda e^{-\lambda x}, (x \geq 0)$

$f(x) = 0, (x < 0)$

em que:

x = variável aleatória contínua.

λ = parâmetro que determina a taxa com que um evento ocorre na unidade de tempo ($\lambda > 0$).

e = 2,71828... (base do logaritmo neperiano).

Parâmetros e estatísticas descritivas da distribuição:

Média: $\mu = 1/\lambda$.

Variância: $\sigma^2 = 1/\lambda^2$.

Essa distribuição é de fundamental importância em vários tipos de fenômenos e por fazer parte de uma grande quantidade de modelos. Várias são as situações na vida real em que a ocorrência de um evento pode ser modelada por uma função exponencial. Ela tem sido usada extensivamente como modelo para o tempo de vida de certos produtos e materiais e faz parte de vários modelos que descrevem o crescimento dos seres vivos. Possui afinidade com a distribuição de Poisson, enquanto nesta tem a análise de falhas por intervalo, em eventos discretos; na exponencial tem a análise de intervalo por falha, em eventos contínuos.

Matematicamente, o termo e^x é definido por:

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} = \frac{x^0}{1!} + \frac{x^1}{2!} + \ldots = \lim_{n \to \infty} (1 + \frac{x}{n})^n$$

Com a FDP anterior, a probabilidade acumulada de zero até *x* é dada por:

$$P(X \leq x) = \int_0^x \lambda e^{-\lambda x}\, dx = 1 - e^{-\lambda x}$$

A probabilidade complementar ou probabilidade acumulada para valores maiores que x é:

$$P(X > x) = \int_x^\infty \lambda e^{-\lambda x} dx = e^{-\lambda x}$$

Na Figura 6 é apresentada a função de distribuição $f(x) = \lambda e^{-\lambda x}$ para uma variável aleatória *x*. Na Figura 7 é apresentada a função de distribuição acumulada da distribuição exponencial $P(X \leq x) = 1 - e^{-\lambda x}$. Para ambas as distribuições, foram usados diferentes valores para o parâmetro λ.

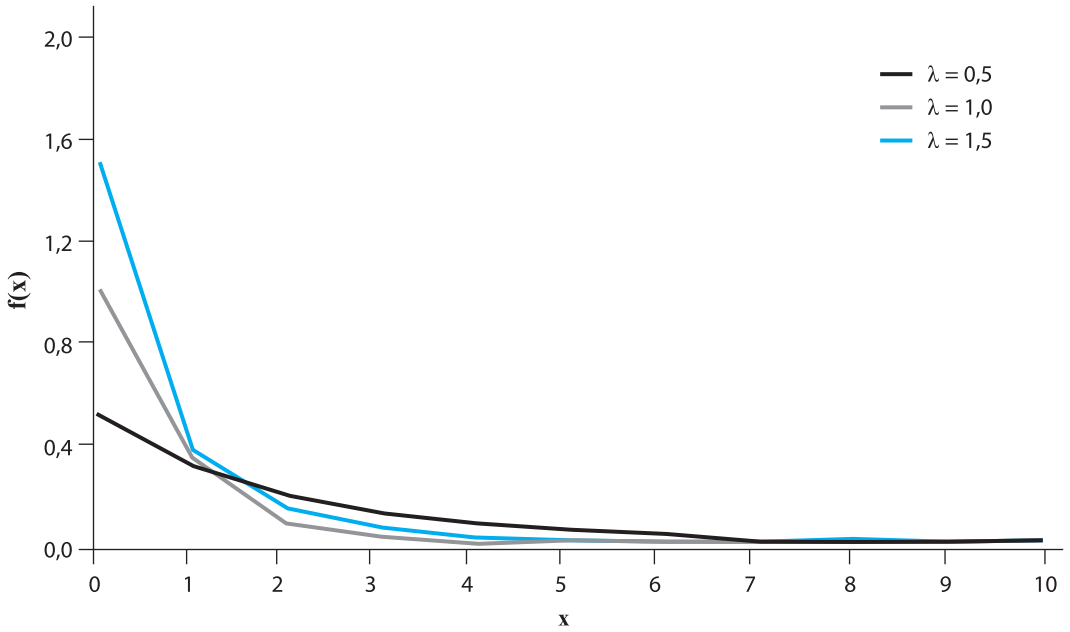


Figura 6. Função densidade de probabilidade da distribuição exponencial f(x) da variável aleatória *x* para: λ = 0,5 (▬); λ = 1,0 (▬) e λ = 1,5 (▬).

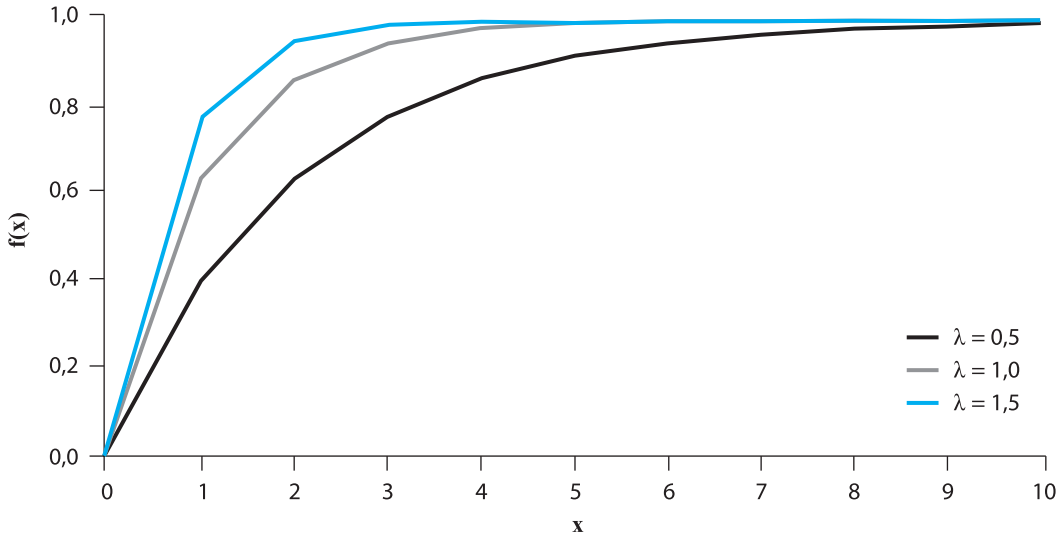


Figura 7. Função de distribuição acumulada da distribuição exponencial f(x) para a variável aleatória *x* para: λ = 0,5 (▬); λ = 1,0 (▬) e λ = 1,5 (▬).

A função *cdf(´exponential´,x, λ)*, no SAS, retorna a probabilidade que uma observação de distribuição exponencial com parâmetro de escala λ ser menor ou igual a *x*.

Aplicação

Uma ferramenta produzida por uma indústria apresenta uma vida média de 8 mil horas. Considerando-se o comportamento segundo a distribuição exponencial, qual a probabilidade dessa ferramenta durar mais de 9 mil horas?

A vida média é o parâmetro μ, sendo λ, o inverso de μ, então, λ = 1/8.000. Usa-se a função de distribuição acumulada para determinar a probabilidade até 9 mil horas.

$$P(x \leq 9.000) = 1 - e^{-\lambda x} = 1 - e^{-(1/8.000)\,9.000} \approx 0{,}67535$$

Em seguida, subtrai-se de 1 para a probabilidade acima de 9 mil horas.

$$P(x > 9.000) = 1 - P(x \leq 9.000) = 1 - (1 - e^{-(1/8.000)\,9.000}) \approx 0{,}32465$$

Finalmente, a probabilidade de a ferramenta durar mais de 9 mil horas é, aproximadamente, 32,5%, valor que também pode ser obtido por:

$$P(x > 9.000) = e^{-\lambda x} = e^{-(1/8.000)\,9.000} \approx 0{,}32465$$

Utilizando-se a função *cdf (´exponential´, x, λ)* por meio do SAS, tem-se:

```
data teste;
input x lambda;
prob_9000 = cdf('exponential', x, lambda); /*Prob: durar ≤ 9000 horas*/
valor = 1 - prob_9000; /*Prob: durar > 9000 horas*/
cards;
9000 8000
;
proc print; var prob_9000  valor;
run;
output
prob_9000  valor
0.6753        0.3246
```

Função exponencial em estudos de crescimento populacional

Quando estudada na forma $f(x) = k_0 \times a^t$, a função exponencial é importante nos estudos de crescimento populacional. Nessa forma, k_0 é a quantidade inicial, quando o tempo t é igual a zero (t = 0), e a é o valor pelo qual f(x) varia quando t aumenta de 1 (base a). Para a > 1, tem-se crescimento exponencial, e, para 0 < a < 1, a função é decrescente ou tem decaimento exponencial, como a taxa de eliminação de uma droga pela corrente sanguínea. Um paciente recebe um medicamento, o qual é eliminado pelo corpo, em miligrama (mg), a uma velocidade de tempo (t), em horas (Tabela 3).

Tabela 3. Velocidade de eliminação de uma droga pela corrente sanguínea em um paciente.

Tempo (hora)	0	1	2	3	4	5
Perda (mg)	240,0	140,0	78,0	45,1	28,0	16,2

Um outro tipo de aplicação é por meio da fórmula abaixo:

$$y = y_o.e^{rx}$$

em que:

y = grau de severidade da doença.

y_0 = quantidade inicial da doença.

r = taxa de progresso desta doença.

x = variável aleatória contínua.

Função exponencial em estudos de curvas de crescimento

O termo exponencial "e^x" faz parte de vários modelos não lineares que tem aplicações importantes no estudo de curvas de crescimento de animais e de plantas. Como exemplo, tem-se o modelo Von Bertalanffy dado por:

$$f(x) = A(1 - be^{-kx})^3 + \varepsilon_x$$

em que:

f(*x*) = peso do indivíduo na idade *x,* que é uma variável aleatória.

ε_x = erro aleatório associado a cada peso.

A = estimativa do peso assintótico ou peso limite do indivíduo, quando $x \to \infty$.

k = índice de maturidade do indivíduo.

b = constante, sem interpretação biológica.

Maiores detalhes são descritos em Freitas (2005).

A taxa de crescimento instantânea (TCI), é derivada de f(*x*) em função de x: $\partial f(x)/\partial x$. Estima-se o incremento no peso para cada unidade *x*:

$$TCI = 3Abke^{-kx}(1-be^{-kx})^2$$

A taxa de crescimento instantânea relativa (Tcir) estima a taxa de crescimento instantânea em relação ao peso do indivíduo na particular variável aleatória *x*.

$$Tcir = 3f(x)bke^{-kx}/(1-be^{-kx})$$

A taxa de maturidade absoluta (TMA) é a razão da TCI em relação ao peso assintótico A (TCI/A). Estima-se a taxa de crescimento instantânea em relação ao tamanho máximo que o indivíduo possa atingir (A). A TMA representa a taxa de troca, na escala de 0 a 1 ou na escala de 0 a 100%, em relação ao peso global ou ao peso adulto do indivíduo.

$$TMA = 3bke^{-kx}(1-be^{-kx})^2$$

Aplicação

Dados de peso-idade de 29.221 fêmeas da raça Nelore, nascidas entre 1976 e 2006, em diversas fazendas da região Norte do Brasil, com pesagens do nascimento até 750 dias de idade, foram fornecidos pela Associação Brasileira de Criadores de Zebu (ABCZ). Foram realizadas nove pesagens por animal em intervalos, aproximadamente,

trimestrais. Os valores estimados pelos modelos de Brody: $f(x)=A(1-be^{-kx}) + \varepsilon_x$ e Von Bertalanffy: $f(x) = A(1 - be^{-kx})^3$ para TCI, Tcir e TMA são apresentados, respectivamente, nas Figuras 8, 9 e 10.

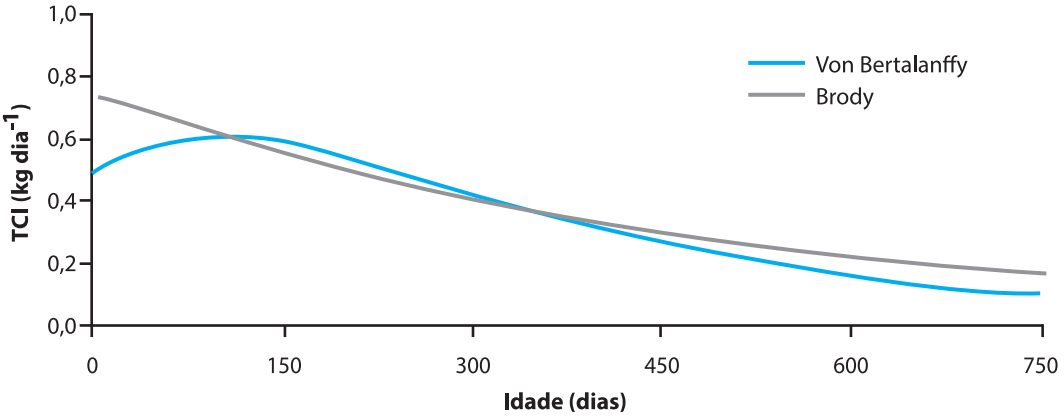


Figura 8. Taxa de crescimento instantânea (TCI), em quilograma por dia (kg dia^{-1}), obtida dos modelos Brody (▬): $f(x)=A(1-be^{-kx}) + \varepsilon_x$ e Von Bertalanffy (▬): $f(x) = A(1 - be^{-kx})^3 + \varepsilon_x$.
Fonte: Marinho et al. (2013).

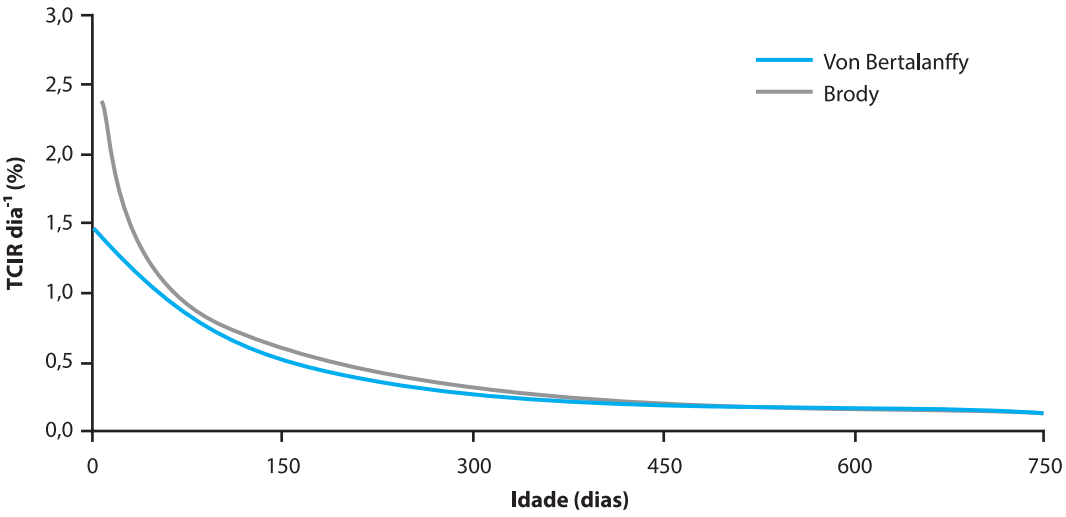


Figura 9. Percentagem da taxa de crescimento instantânea relativa (TCIR) por dia, obtida dos modelos Brody (▬): $y_x =A(1-be^{-kx}) + \varepsilon_x$ e Von Bertalanffy (▬): $y_x=A(1-be^{-kx})^3 + \varepsilon_x$.
Fonte: Marinho et al. (2013).

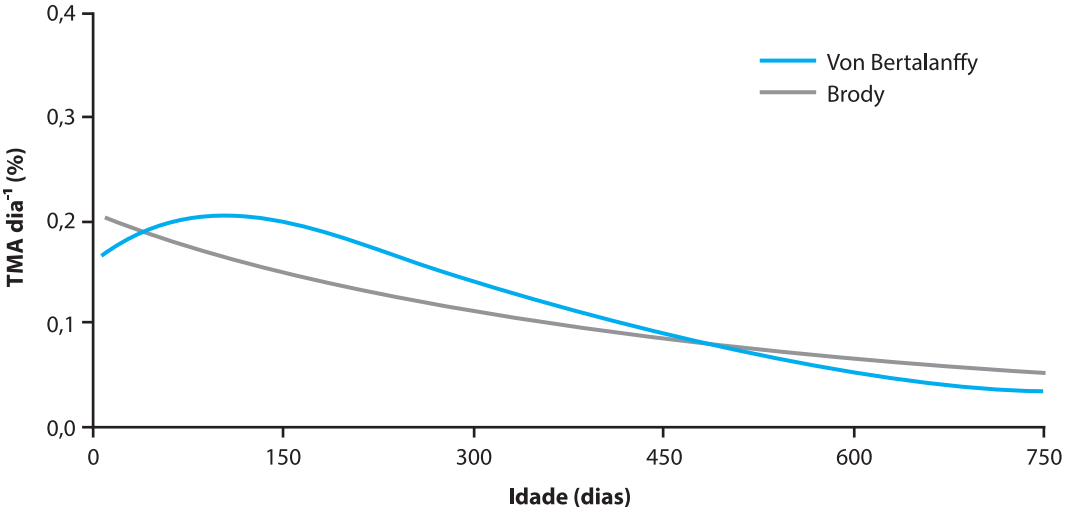


Figura 10. Percentagem da taxa de maturidade absoluta (TMA) por dia obtidas dos modelos Brody (▬): $y_x = A(1-be^{-kx}) + \varepsilon_x$ e Von Bertalanffy (▬): $y_x = A(1-be^{-kx})^3 + \varepsilon_x$.
Fonte: Marinho et al. (2013).

Em síntese, distribuição normal é considerada a mais importante pelo fato de representar quase todos os fenômenos experimentais, cujos dados devem ser quantitativos resultantes de características suscetíveis de serem medidas. A distribuição normal reduzida possui propriedades que possibilitam comparar resultados de experimentos realizados em diferentes condições, como estudo em diferentes épocas e também avaliar o desempenho de um indivíduo ou de um equipamento dentro de um experimento.

As distribuições t e F proporcionam as estatísticas t e F que geralmente são utilizadas após uma análise de variância (Anova) ter sido realizada para tomada de decisão.

A distribuição de χ^2 é utilizada em tabelas de contingência com dados agrupados em frequências, mas tem aplicação também para dados contínuos em testes de hipóteses e intervalos de confiança. Finalmente, a distribuição exponencial tem aplicação em várias áreas de estudo. O termo exponencial "e^x" faz parte de vários modelos não lineares que têm aplicações importantes no estudo de curvas de crescimento de animais e de plantas.

Exercícios[6]

1) No texto a seguir, existem afirmações incorretas quanto a conceitos de estatística. Reescreva o texto colocando definições corretas e sublinhe ou coloque em negrito onde houve definições incorretas.

A distribuição normal tem grande aplicação na estatística, principalmente para dados discretos. Com relação a essa distribuição, é correto afirmar: a média, a moda e a mediana são iguais; a curva é simétrica em torno do desvio-padrão σ; os coeficientes de assimetria são indicativos se a curva de distribuição é viesada para esquerda ou para a direita, enquanto os coeficientes de curtose indicam se a curva é mais pontiaguda ou mais achatada que a normal. Para uma amostra ou população $x_1,..., x_n$, a distribuição normal é compreendida no intervalo de $0 < x < \infty$, podendo afirmar que a soma das probabilidades desses valores é igual a 1, conforme mostra a expressão $P(0 < x < +\infty) = \int_0^{\infty} f(x)dx = 1$. A variância é o parâmetro mais apropriado do que o coeficiente de variação para comparar a dispersão ou homogeneidade de duas amostras. Por meio do desvio-padrão (raiz quadrada da variância), é possível conhecer a dispersão dos dados de uma amostra e, também, determinar o erro-padrão da média. Na distribuição normal, 95% da área da curva está dentro do intervalo [μ - 3σ; μ + 3σ]. Com relação à distribuição normal reduzida, algumas propriedades e notação são: a) se $x \sim N(\mu, \sigma^2)$, então $z = (x - \mu)/\sigma$ tem ~ N(0, 1); b) existem dois pontos de inflexão: z = +2 e z = -2, sendo que a área entre eles é 0,6826; c) a curva tem forma de sino e a sua área vale 1; d) a função tem valor máximo no ponto z = μ e sua ordenada vale $1/\sqrt{(2\pi)} = 0,3989$; e) ela é definida para qualquer *z* entre -∞ e +∞, porém a curva da FDP tende a zero quando *z* tende a ± 2 ($z \to \pm 2$). Se *v* é a soma do quadrado de *v* variáveis aleatórias $z_1, z_2, ..., z_v$ de distribuição normal reduzida, então $z_i \sim N(0, 1)$ e *v* tem distribuição de qui-qradrado (χ^2) com *k* graus de liberdade, sendo $v = \sum_{i=1}^{k} Z_i^2 \sim X_k^2$.

2) Sabendo-se que a probabilidade da variável z estar compreendida no intervalo [a,b] da distribuição normal reduzida é dada por: $P(a \leq z \leq b) = 1/\sqrt{2\pi}\int_a^b e^{-z^2/2} dz = \Phi(b) - \Phi(a)$, e, ainda que a função *probnorm(z)* retorna o valor da área acumulada até o valor z, elabore uma rotina SAS para provar que essa função tende a zero quando z aproxima de -4, e tende a 1 quando z aproxima de +4.

3) Em um experimento os valores da soma de quadrado do resíduo (SQR) e soma de quadrado de tratamentos (SQT), com 5 e 20 graus de liberdade, foi respectivamente: 20,52 e 12,24.

[6] As respostas dos exercícios podem ser consultadas no Apêndice 1.

Calcule o valor de $F_{5,20}$ e com o resultado utilize as funções *probf* ($f_{5,20,5,\ 20)}$ e *finv (0.95, 5, 20)* em uma rotina SAS, e interprete os resultados.

4) Na rotina SAS a seguir, explicar o significado das variáveis li e ls.

```
title;
data peso;
input peso @@;
cards;
91.9  97.8  111.4  122.3  105.4  95.0  103.8  99.6  96.6  119.3  104.8  101.7
;
proc means data = peso;
output out = saida  mean = media  stddev = s;
data peso1i;
set saida;
t = tinv(.95, 11);
li = media - t*(s/sqrt(12));
ls = media + t*(s/sqrt(12));
keep media s t li ls;
proc print; var media li ls;run; output
media       li        ls
104.133   99.2615   109.005
```

5) Se um equipamento produzido por uma indústria apresenta uma vida média de 400 dias. Considerando a distribuição exponencial, qual a probabilidade deste equipamento durar mais de 420 dias?

Capítulo 7

Hipóteses científicas e testes de hipóteses

Introdução

O método científico pode ser definido como um conjunto de técnicas para investigar os fenômenos naturais, possibilitando atualizar e corrigir conhecimentos anteriores e adquirir novos conhecimentos. A ciência pode ser entendida como um conjunto de técnicas teóricas, empíricas e práticas acerca do mundo natural que são produzidas por cientistas e pesquisadores usando o método científico. O conhecimento científico e outras classes de conhecimento sofreram grandes transformações durante a revolução científica nos séculos 16 e 17. A filosofia realista, por exemplo, começa com Aristóteles e sua metafísica e culmina com Tomás de Aquino, em que o conhecimento antecede o pensamento, isto é, "as coisas são a medida do nosso conhecer".

O inglês Francis Bacon (1561 a 1626), conforme Francis Bacon (Wikipédia, 2019c), mostrou a importância da experimentação e do empirismo para a aquisição dos conhecimentos científicos, enquanto o francês René Descartes (1596−1650), que se dedicou à metodologia científica e ao empirismo, é considerado o fundador da ciência moderna (Boyer; Merzbach, 1996). O empirismo pode ser definido como uma filosofia ou movimento que acredita nas experiências como as formadoras das ideias, isto é, todo o processo do conhecer, do saber e do agir é aprendido pela experiência, por meio de tentativa e erro. O italiano Galileu Galilei (1564−1642), que se dedicou ao método científico e ao experimentalismo, é um dos responsáveis pelo estabelecimento das bases do pensamento científico moderno e do método experimental (Wikipédia, 2019d). Embora seja considerado por muitos como o pai da física moderna, o seu princípio racional era matemático, e é atribuída a ele a frase "a matemática pode ajudar a compreender a natureza do mundo".

Em resumo, sabemos que a ciência é sempre passível de aperfeiçoamento. Mas, devemos sempre conhecê-la e acompanhá-la, pois ela conduz à verdade.

Neste capítulo, são apresentados conceitos fundamentais da estatística experimental, tais como: método científico, hipótese científica, experimento, princípios básicos da experimentação, hipótese estatística e teste de uma hipótese estatística, erro experimental, instalação, condução e análise de um experimento e estatísticas associadas a uma análise de variância.

Hipótese científica

A estatística trabalha com métodos que auxiliam na tomada de decisões diante da incerteza. Cada decisão requer no mínimo que uma hipótese seja testada. Toda investigação e todo avanço do conhecimento parte da identificação de um problema, questão ou suposição, que é a hipótese científica.

As hipóteses são predições que, se um conjunto de suposições é verdade, então certos parâmetros terão valores especificados. Admitamos que na pecuária exista uma vacina A tradicionalmente utilizada para a cura de uma doença em bovinos e que a questão ou hipótese científica seria o desenvolvimento por um laboratório de uma vacina alternativa B que seja mais eficiente e mais econômica do que a vacina A.

A hipótese científica então seria:

- É possível desenvolver uma vacina B que seja mais eficiente e mais econômica do que a vacina tradicional A para o tratamento da doença X em bovinos.

A palavra "é possível" é necessária na formulação de uma hipótese científica, pois se trata de uma suposição e partimos sempre da hipótese de nulidade (H_0), ou seja, de que as duas vacinas A e B são igualmente eficientes. A partir dos resultados de um experimento, temos evidências e conhecimentos para a aceitação e ou rejeição da hipótese Ho em favor da hipótese alternativa (H_a). Dentro do exemplo, conclui-se que a vacina B é mais eficiente do que a vacina A.

A hipótese normalmente é colocada no item Material e Método de um projeto de pesquisa, que geralmente possui os itens abaixo:

a) Título
b) Resumo
c) Introdução
d) Revisão de Literatura
e) Objetivo
f) Hipótese Científica
g) Material e Método
h) Cronograma de Trabalho
i) Estratégia de Ação
j) Equipe Técnica
k) Referências Bibliográficas

O experimento

O experimento é a etapa mais importante do método científico e tem a finalidade de responder questões, investigar problemas, testar teorias e, principalmente, testar hipóteses, oriundas de um planejamento experimental adequado, e que atendam a alguns princípios básicos avaliando e comparando efeitos de tratamentos. As conclusões de um experimento podem apresentar argumentos que levam à aceitação e ou rejeição de uma hipótese ou de uma teoria. Conforme afirmação do filósofo Karl Popper (Popper, 2005), qualquer hipótese pode ser falsificada. Um experimento não prova uma hipótese, ele apenas pode adicionar conhecimentos para mantê-la.

Vários itens são importantes por ocasião da instalação de um experimento: escolha do local, escolha do material experimental, delineamento experimental, número de repetições, número de tratamentos, número de amostras a serem obtidas por unidade experimental, variáveis respostas a serem avaliadas, acompanhamento do experimento, planilha para a coleta dos dados, tipo de análise, etc. O experimento é um método de pesquisa explicativa em que o pesquisador interfere na amostra por meio da imposição de tratamentos.

A escolha dos tratamentos e dos fatores com os respectivos níveis e a inclusão de um tratamento-controle são itens fundamentais de um experimento. Quando se trata de um fator quantitativo como doses de adubo, o tratamento-controle geralmente é a dose zero, ou seja, o não uso de adubo. Na área médica, mais comumente na psiquiatria, um tratamento correspondente ao controle é o placebo, que consiste em pílula de farinha ou açúcar, ou mesmo remédios verdadeiros, porém, em dose pequena ou insuficiente para provocar o efeito esperado. O efeito placebo reflete um dos mais extraordinários poderes da mente humana e que ainda são mal compreendidos e interpretados. Ele mostra que o cérebro é capaz de produzir reações cuja capacidade de cura é comparável ao de drogas poderosas. A descrição do experimento é feita no item Material e Método, e deve ser detalhada de forma que um pesquisador possa repeti-lo.

Os elementos básicos de um trabalho científico são:

a) Título
b) Resumo
c) Abstract
d) Introdução
e) Material e Método
f) Resultados e Discussão
g) Conclusões
h) Referências Bibliográficas

Princípios básicos da experimentação

Controle local

É o processo de avaliar o local ou ambiente onde será instalado o experimento e escolher o desenho experimental mais adequado. Se o ambiente é totalmente homogêneo, é recomendado o delineamento inteiramente casualizado (DIC), por ser mais simples, tanto para instalação quanto para análise dos dados, e geralmente mais barato.

Repetição

É o número de parcelas ou unidades experimentais de um tratamento. É necessária para estimar o erro experimental e determinar a precisão com que estimamos um parâmetro. Conhecendo-se antecipadamente o tipo de variabilidade dos dados que serão obtidos de um experimento, pode-se estimar o número mais conveniente de repetições a ser utilizado de modo a reduzir as estimativas dos erros-padrão e trabalhar com um grau de precisão desejada. Deve-se ter em mente que a variação é maior para material biológico do que para material inanimado. Na ciência física, por exemplo, os processos são mais determinísticos, e o erro é principalmente de medida.

Casualização

É o processo de distribuição aleatória dos tratamentos às parcelas ou unidades experimentais. Possibilita que todas as parcelas tenham a mesma chance de receber qualquer um dos tratamentos, e, com isso, evita que um tratamento seja favorecido por ser distribuído em ambiente mais apropriado. A casualização garante a independência dos erros, uma das suposições mais importantes da experimentação. A sua falta implica ajustamentos complicados e introduz vícios na comparação de tratamentos.

Erro experimental

Uma suposição é que os erros sejam não correlacionados (independentes) e com variâncias homogêneas entre os fatores ou tratamentos. O erro é inerente ao material experimental, mas tem como controlá-lo ou minimizá-lo. A variação tende mascarar os efeitos dos tratamentos levando a interpretações erradas dos resultados e do julgamento

acerca dos melhores tratamentos. O teste de Bartlett é geralmente usado para testar igualdade de variâncias e tem grande poder quando os dados têm distribuição normal.

Desenho experimental

Processo de organizar um experimento de maneira que os dados coletados possibilitem responder as questões de interesse. É a maneira que os tratamentos são distribuídos às unidades experimentais.

Organização dos tratamentos

Geralmente os tratamentos são organizados em fatores e níveis. Por exemplo, sexo é um fator e possui dois níveis: macho e fêmea; caso esteja planejando um ensaio de ganho de peso com três raças de bovinos e dois sexos, então raça e sexo são fatores, com três e dois níveis, respectivamente.

Um dos grandes interesses em utilizar tratamentos organizados em esquema fatorial é obter resultados mais abrangentes e conclusivos e, também, avaliar a eficiência de um tratamento nos vários níveis do outro fator, o que é chamado de interação. Uma desvantagem desses delineamentos é que facilmente cresce o número de tratamentos e daí a dificuldade em conseguir um ambiente ou local com condições homogêneas para distribuir todos os tratamentos. O mais comum é planejar um experimento com os tratamentos organizados com no máximo três fatores. Uma quantidade maior de fatores dificulta a sua condução, tornando-o mais caro, e a análise estatística mais complicada. Quando o efeito de interações de alta ordem é não significativo, este pode ser adicionado ao resíduo aumentando-se o número de graus de liberdade.

No cenário atual, com possíveis alterações do clima, com elevação da temperatura e mudanças na distribuição de chuvas, pode ocorrer aumento da incidência de pragas e de doenças na agricultura, com interferências na experimentação. Com isso, na estatística experimental, sempre devem ocorrer refinamentos de metodologias de estatísticas e alterações dos fatores e níveis dos tratamentos. Alguns exemplos de fatores de experimentos em agricultura são apresentados na Tabela 1.

Tabela 1. Exemplos de fatores de experimentos em agricultura.

Animal	Vegetal
Raça/grupo genético	Variedade
Mês e ano de nascimento	Semana, mês e ano de plantio
Sexo	Época de plantio (chuvosa; seca)
Local de nascimento	Local de plantio
Manejo	Característica morfológica
Categoria de idade dos animais	Fontes e doses de adubo
Posição de uma instalação	Estádio de crescimento
Luminosidade	Manejo de corte
Ventilação	Profundidades de aração
	Tipos de inseticida

Hipótese estatística

É uma suposição feita acerca de um parâmetro ou de uma característica da população, a qual normalmente é formulada junto com o experimento. Se o objetivo é a avaliação das duas vacinas A e B, tem-se:

Hipótese nula (H_0):

H_0: a vacina B é tão eficiente quanto a vacina A.

Duas hipóteses unilaterais:

H_{a1}: a vacina B é menos eficiente que a vacina A.

H_{a2}: a vacina B é mais eficiente que a vacina A.

Hipótese bilateral que inclui as duas hipóteses unilaterais:

H_{a3}: a vacina B é diferente da vacina A.

O objetivo é rejeitar a hipótese H_0 em favor de hipóteses alternativas (H_a).

Generalizando, no caso de k tratamentos, tem-se:

H_0: $t_1 = t_2 = ... = t_k$ versus H_a: pelo menos dois tratamentos diferem entre si.

Teste de uma hipótese estatística

É utilizado para tomar decisões sobre a rejeição ou não da hipótese H_0 por meio de um teste estatístico, sendo α o nível de significância do teste, que geralmente é usado ao nível de 5% de probabilidade (α = 0,05).

As contribuições da estatística experimental concentram-se no fornecimento de informações, metodologias e ferramentas para inferir sobre uma população que é desconhecida, trabalhando-se com dados de uma amostra, obtida de um experimento, ensaio, etc. Portanto, o tamanho da amostra é imprescindível, pois os resultados dela serão utilizados para inferir sobre a população.

Na Figura 1, no gráfico superior, visualiza-se o valor de α à direita do gráfico, (α = 0,05), para rejeitar uma hipótese de nulidade H_0 unilateral em favor da hipótese alternativa H_a.

Exemplo:

H_{a1}: a vacina B é menos eficiente que a vacina A.

Ou

H_{a2}: a vacina B é mais eficiente que a vacina A.

No gráfico inferior, visualiza-se o valor de α/2 = 0,025 em cada extremo da curva, para rejeitar uma hipótese H_0 bilateral em favor da hipótese alternativa H_a.

Exemplo:

H_{a3}: a vacina B é diferente da vacina A.

- Região de confiança (1- α) é a região do gráfico que indica alta probabilidade de afirmar que H_0 é verdadeira, geralmente 95%, o que significa que α ou o erro é 5%. A não rejeição de H_0 não implica o fato de que ela seja verdadeira. Apenas nos mostra que, com 95% de confiança, não temos evidência suficiente para considerá-la como falsa.
- Região crítica é a região do gráfico em que, muitas vezes, tomamos decisão errada. Rejeitar H_0 quando de fato ela é verdadeira, comete-se o erro do tipo I (α). Não rejeitar H_0 quando ela não é verdadeira, comete-se um erro denominado de erro do tipo II (β).

Resumindo:

- Erro do tipo I (α) → prob (erro do tipo I) = α.
- Erro do tipo II (β) → prob (erro do tipo II) = β.

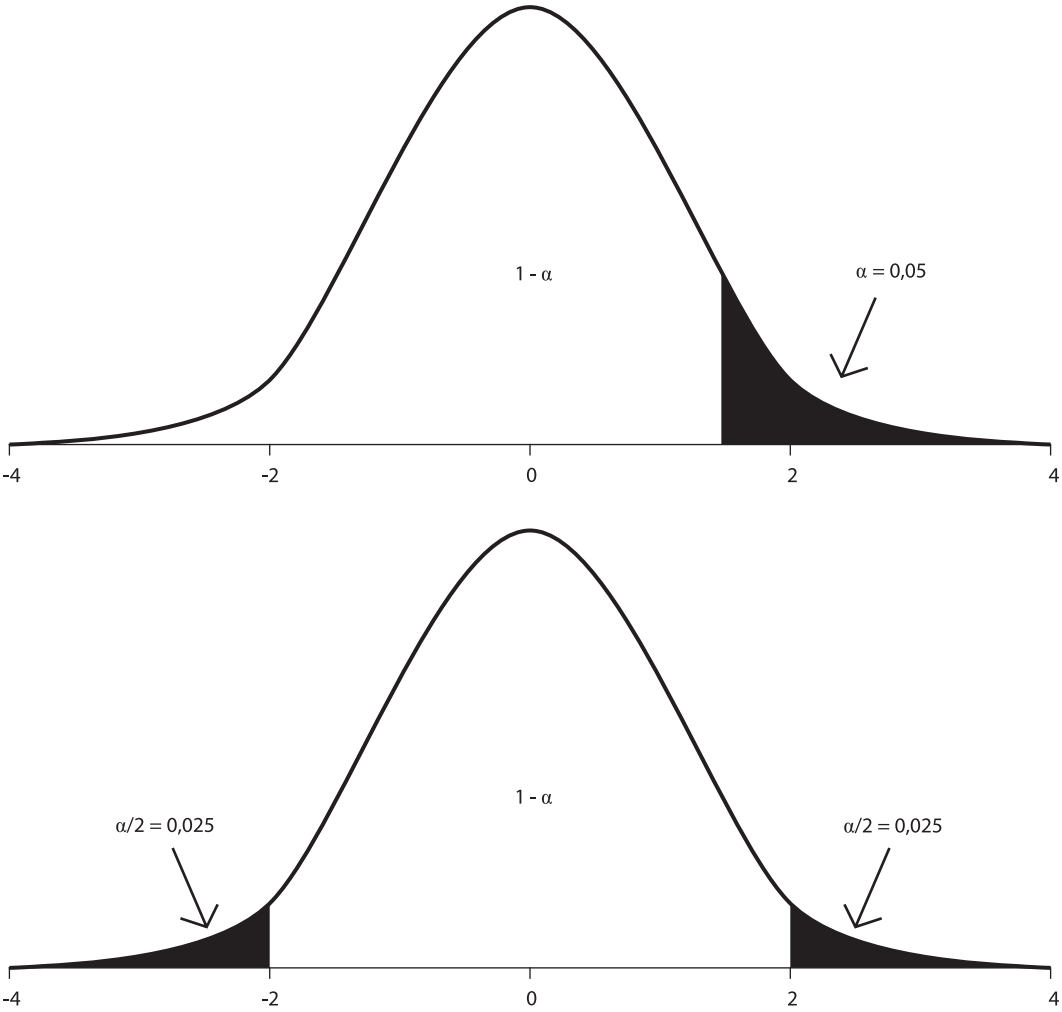


Figura 1. Região de confiança (1 - α) e região crítica (α = 0,05; α/2=0,025).

Na Tabela 2, são apresentadas as alternativas de decisões para aceitar e/ou rejeitar uma hipótese.

Tabela 2. Alternativas de decisões para aceitar e/ou rejeitar uma hipótese.

	Hipótese nula: (H_0) verdadeira	Hipótese alternativa: (H_a) verdadeira
Aceita H_0	Decisão correta	Decisão errada: erro tipo II (β)
Rejeita H_0	Decisão errada: erro tipo I (α)	Decisão correta

Os testes de hipóteses são construídos com base em resultados experimentais obtidos de amostras aleatórias de uma população e, portanto, estão sujeitos a flutuações amostrais. Em razão disso, pode-se ter uma amostra que não represente bem a população, como a presença de observações extremas (*outliers*), o que leva a conclusões que não correspondem à realidade. Uma alternativa experimental é aumentar o tamanho amostral.

O objetivo de comparar duas hipóteses é obter inferências da população com base em evidências experimentais, que são sujeitas a erros. Em uma amostra de tamanho n: $x_1, x_2, \ldots, x_n$, o termo $(x_i - \bar{x})$ representa o erro ou desvio, que é a base da estatística experimental. Por meio de um teste estatístico, procura-se condições que garantam que os resultados de um experimento possam ser generalizados; isto é, usa o raciocínio indutivo: do particular para o geral (inferência estatística).

Como as variáveis de uma amostra são de caráter puramente aleatório, para que um teste de hipótese seja realizado, é necessário assumir uma distribuição de probabilidade para a variável ou variáveis envolvidas, e depois conhecer a média e a variância dessa distribuição.

As distribuições de probabilidades contínuas, geralmente utilizadas na análise dos dados de um experimento, são: normal padronizada (z), qui-quadrado (χ^2), F e t. Elas fornecem os respectivos testes associados ao valor crítico α e graus de liberdade ν, ν_1 e ν_2, conforme notação a seguir:

α = nível de valor crítico de uma distribuição de probabilidade.

z_α ou $z(\alpha)$ = percentil da distribuição normal padrão.

$\chi^2(\alpha,\nu)$ = percentil da distribuição de qui-quadrado com ν graus de liberdade.

$F(\alpha,\nu_1,\nu_2)$ = percentil da distribuição F com ν_1 e ν_2 graus de liberdade.

$t(\alpha,\nu)$ = percentil da distribuição t com ν graus de liberdade.

Distribuições z, χ^2, F, t e correspondentes testes estatísticos

Das distribuições z, χ^2, F, t são produzidos testes que são utilizados em análises estatísticas.

Teste z

É aplicado a dados ou proporções para grandes amostras e com a suposição de normalidade atendida. Quando se trabalha com grandes amostras, a normalidade é garantida pelo teorema do limite central: "Se de uma população retira-se n amostras de tamanho grande ($n > 30$ observações), e calcula-se as médias de cada uma delas ($\bar{x}_1,...,\bar{x}_n$), a distribuição dessas médias é aproximadamente normal com média μ e desvio-padrão $\sigma/\sqrt{n}$. Isso acontece mesmo quando os dados originais não possuem distribuição normal."

Aplicação do teste z para uma amostra e para duas amostras

Nesses testes, há interesse de testar a hipótese nula (H_0) versus hipóteses alternativas (H_1), considerando-se um valor hipotético Δ_0 que pode ser $\Delta_0 = 0$ ou qualquer outro valor, dependendo do experimento.

Uma amostra de tamanho n

H_0: H_0: $\mu = \Delta_0$

versus

H_{a1}: $\mu > \Delta_0$, ou

H_{a1}: $\mu < \Delta_0$, ou : $\mu \neq \Delta_0$

A fórmula geral para o cálculo da estatística z é:

$$z = \frac{\bar{x} - \Delta_0}{\sigma} \sqrt{n}$$

Exemplo considerando-se dados de proporções:

Um fazendeiro afirma que 90% de suas vacas produzem acima de 20 kg de leite por dia. Em uma amostra de 300 vacas dessa fazenda, constatou-se que 260 delas tinham produtividade maior que 20 kg dia^{-1} e 40, inferior a 20 kg dia^{-1}. Considerando $\alpha = 0,05$ a hipótese do fazendeiro é confirmada?

A proporção de vacas na amostra com produtividade de leite acima de 20 kg dia^{-1} é:

$$\hat{p} = \frac{260}{300} = 0,87$$

Com essa proporção, testa-se a hipótese:

H_0: p = 0,90 versus H_1: p < 0,90

Em uma amostra de dados de proporções, o desvio-padrão é dado por: $s = \sqrt{\frac{pq}{n}}$.

Assim, a estatística z é calculada por:

$$Z = \frac{\hat{p} - \Delta_0}{\sqrt{(pq)/n}} = \frac{0,87 - 0,90}{\sqrt{(0,87 \times 0,13)/300}} = -1,54$$

Consultando a Tabela 1.1 (Anexo 1), para a área sob a curva normal de 0 a z, encontra-se *z* = 0,9382. Assim, α = 1 − z = 1 - 0,9382 = 0,0618. Como o valor de α é maior do que 5%, o valor de *z* = -1,54 está dentro da região de confiança da curva. Portanto, a hipótese de nulidade não é rejeitada; e conclui-se que 90% das vacas do fazendeiro têm produtividade igual a 20 kg dia^{-1}, mas não superior.

Duas amostras de tamanhos n_1 e n_2

H_0: $\mu_1 - \mu_2 = \Delta_0$

versus uma das hipóteses

H_{a1}: $\mu_1 - \mu_2 > \Delta_0$

H_{a2}: $\mu_1 - \mu_2 < \Delta_0$

H_{a3}: $\mu_1 - \mu_2 \neq \Delta_0$

A fórmula geral para o cálculo da estatística z é:

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta_0}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

Como não se conhecem as variâncias populacionais (σ_1^2 e σ_2^2), mas tendo-se $n \geq 30$, usam-se as estimativas amostrais (S_1^2 e S_2^2).

$$z = \frac{(\bar{X}_1 - \bar{X}_2) - \Delta_0}{\sqrt{S_1^2/n_1 + S_2^2/n_2}}.$$

Para dados de proporções, a estatística Z é calculada por:

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - \Delta_0}{\sqrt{[\hat{p}_1 - (1 - \hat{p}_1)]/n_1 [\hat{p}_2 - (1 - \hat{p}_2)]/n_2}}$$

Os requisitos abaixo devem ser atendidos:

$$\hat{p}_1 = \frac{x_1}{n_1}\ ;\ \hat{p}_2 = \frac{x_2}{n_2}\ ;\ n_1p_1 > 5;\ n_1(1-p_1) > 5;\ n_2p_2 > 5 \text{ e } n_2(1-p_2) > 5$$

Para a hipótese Δ_0: $p_1 = p_2$, as proporções são combinadas, $\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$.

$$z_{calc} = \frac{(\hat{p}_2 - \hat{p}_2)}{\sqrt{\hat{p} - (1-\hat{p})(1/n_1 + 1/n_2)}} \sim N(0,1)$$

Hipóteses:

Unilateral (à direita): $\text{Prob}\,[z_{calc} > z_\alpha] = \alpha$.

Unilateral (à esquerda): $\text{Prob}\,[z_{calc} < -z_\alpha] = \alpha$.

Bilateral: $\text{Prob}\,[z_{calc} < -z_{\alpha/2}] = \text{Prob}\,[z_{calc} > z_{\alpha/2}] = \alpha/2$.

Aplicação

Um grupo de 800 animais foi submetido a dois esquemas de vacinação: 400 animais receberam uma vacina nova (N) e 400 receberam uma vacina tradicional (T). Após um tempo de observação, foram obtidos os resultados, conforme apresentado na Tabela 3.

Tabela 3. Comparação de uma vacina nova (N) versus a vacina tradicional (**T**).

Vacina	+	–	Total
Vacina N	340	60	400
Vacina T	300	100	400
Total	640	160	800

O objetivo é verificar se a vacina nova (N) é superior à vacina tradicional (T), o que equivale a testar a hipótese H_0: $p_1 = p_2$ versus H_a: $p_1 > p_2$, em que $\hat{p}_1 = 340/400 = 0{,}85$ e $\hat{p}_2 = 300/400 = 0{,}75$. Isso indica, respectivamente, a proporção de animais que responderam positivamente às vacinas N e T.

Como as duas proporções são semelhantes, o primeiro passo é obter o valor de *z* calculado (z_{calc}), considerando $\hat{p} = 0{,}80$ e $\hat{q} = 0{,}20$.

$$Z_{cal} = (\hat{p}_1 - \hat{p}_2)/\sqrt{\hat{p} - (1-\hat{p})(1/n_1 + 1/n_2)}$$

$$Z_{cal} = (0{,}85 - 0{,}75)/\sqrt{0{,}80 - (1-0{,}80)(1/400 + 1/400)} = 0{,}1119$$

Como o teste é unilateral à direita, consultando-se a Tabela 1.1 (Anexo 1), para se ter uma área de 0,95, o valor de *z* tabelado para $\alpha = 0,05$ é 1,65. A hipótese de nulidade somente é rejeitada para $\alpha = 0,05$ se Prob $[z_{calc} > 1,65]$. Como $z_{calc} = 0,1119$, não se rejeita a hipótese de nulidade, e conclui-se que a vacina nova (*N*) não é superior à vacina tradicional.

Teste de χ^2

A seguir, serão apresentadas algumas estatísticas estudadas por meio do teste de χ^2.

Teste de tendência

O objetivo é detectar se há tendência positiva ou negativa entre as variáveis respostas de um experimento de acordo com o tratamento. Seja a situação em que se deseja testar doses crescentes de um produto para controle de carrapatos em bovinos. Para isso, um conjunto de animais são infestados artificialmente. Após certo tempo, doses crescentes do produto são aplicadas a esses animais. O interesse é verificar se a mortalidade dos parasitas aumenta na mesma proporção em que a dose do produto aumenta. Esse tipo de pesquisa é muito comum nos laboratórios médicos, onde se desejam testar doses crescentes de um medicamento para controlar, por exemplo, dor de pacientes.

Teste de aderência

O objetivo é testar se uma amostra de dados nominais difere de uma distribuição hipotética. É comumente usado em genética, nos cruzamentos entre plantas e entre animais em que as classes fenotípicas das progênies seguem proporção esperada de acordo com leis mendelianas, tais como as proporções: 1:2:1; 1:3:3:1; 1:4:6:4:1.

Teste de concordância

O objetivo é verificar se há concordância entre as variáveis de classificação das linhas e as variáveis de classificação das colunas. Por exemplo, animais são classificados em três categorias (regular, bom e ótimo) quanto à ocorrência de determinada doença. Dois veterinários vão examinar esses animais; e o objetivo é verificar se há concordância ou não entre eles.

Teste de independência

O objetivo é testar se há independência entre as frequências observadas e esperadas das linhas e colunas.

Aplicação

Um grupo de 18 indivíduos com determinada doença foi dividido em dois subgrupos: tratados e não tratados, respectivamente, com 10 e 8 indivíduos cada. Dois meses após, período suficiente para a droga provocar o seu efeito, foram obtidos os resultados apresentados na Tabela 4. O interesse foi verificar a eficiência do tratamento.

Tabela 4. Eficiência de uma droga comparando indivíduos tratados e indivíduos-controles.

Tratamento	Recuperação		Total
	Sim	Não	
Tratados	6 (4,44)	4 (5,55)	10
Controle	2 (3,56)	6 (4,44)[1]	8
Total	8	10	18

[1] Valores observados e esperados, respectivamente.

A partir dos dados apresentados na Tabela 4, que representa uma tabela de contingência 2×2, calcula-se a estatística de χ^2:

$$\chi^2_v = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

em que:

χ^2_v = valor de χ^2 com v graus de liberdade, em que $v = (I - 1) \times (J - 1)$.

o_{ij} = número de casos observados na linha i da coluna j.

e_{ij} = número de casos esperados sob a hipótese de que não há diferença na proporção de animais recuperados e não recuperados dos grupos tratados e controle. Na Tabela 4 os valores esperados estão dentro do parêntese:

$e_{11} = (8 \times 10)/18 = 4{,}44$; $e_{12} = (10 \times 10)/18 = 5{,}55$; $e_{21} = (8 \times 8)/18 = 3{,}56$; $e_{22} = (10 \times 8)/18 = 4{,}44$.

$\sum_{i=1}^{I}\sum_{j=1}^{J}$ = somatório sobre todas as células das I linhas e as J colunas.

$$\chi_1^2 = \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} = \frac{(6-4{,}44)^2}{4{,}44} + \frac{(4-5{,}55)^2}{5{,}55} + \frac{(2-3{,}56)^2}{3{,}56} + \frac{(6-4{,}44)^2}{4{,}44} = 2{,}2050$$

Esse valor pode ser obtido por meio da rotina do Statistical Analysis System (SAS).

```
data doenca;
      input trata resposta freq;
      datalines;
1  0  6
1  1  2
0  0  4
0  1  6
;
 proc freq data = doenca order = data;
    weight freq;
    tables trata*resposta / chisq;
     run;
output
Statistic                DF      Value     Prob
 Chi-Square              1       2.2050    0.1376
```

Na saída (*output*), tem-se $\chi_1^2 = 2{,}2050$, com um nível de probabilidade ($\alpha = 0{,}1376$).

A função *probchi(x,df)*, a seguir, retorna os valores anteriores e também a área correspondente da distribuição de χ^2 para (1 - α) que é de 86,24%.

```
data;
input x gl;
area = probchi(x, gl);
alfa = 1- area;
datalines;
2.2050  1
;
proc print;run;
output
 x          gl     area         alfa
2.2050     1    0.8624     0.13756
```

Assim, o primeiro passo em tabelas de contingência é calcular a estatística de χ^2 juntamente com o número de graus de liberdade (linhas -1) x (colunas -1) e um nível

α, geralmente α = 0,05. Nesse exemplo, com o valor de α = 0,13756, conclui-se que o tratamento não foi eficiente. O teste de χ^2 é usado em inferência estatística, teste de hipóteses e construção de intervalos de confiança.

Teste F

O teste F tem grande aplicação nas análises de variância de um experimento. É obtido pela razão entre a variância devida a tratamentos (QMT) e a variância residual (QMR), respectivamente, com n_1 e n_2 graus de liberdade. Em 1924, Fisher apresentou a fundamentação básica desse teste e sua posterior formulação se deve à Snedecor.

Exemplo:

$$F_{n_1, n_2} = \frac{QMT}{QMR}$$

Com

$$F_{5,20} = \frac{157,05}{49,12} = 3,1972$$

Consultando a distribuição de F na Tabela 1.4 (Anexo 1) para α = 0,05, verifica-se que o valor tabelado para $F_{5,20}$ é 2,71. Utilizando a rotina SAS a seguir, verifica-se que o valor exato de α para o teste F calculado ($F_{5,20}$ = 3,1972) é α = 0,0279, com correspondente área de 0,9721.

```
data;
input x n1 n2;
area = probf(x, n1, n2);
alfa = 1- area;
datalines;
3.1972  5  20
;
proc print;run;
output
x          n1    n2     area        alfa
3.1972     5     20     0.9721      0.0279
```

Teste t

É utilizado para amostras pequenas (n < 30), com suposição de normalidade atendida e variâncias estimadas; o teste t é aplicado em situações nas quais o uso do teste z leva a resultados incorretos.

Teste t para uma amostra

Testa a hipótese que uma média amostral ($\overline{x}$) é igual à média de uma população hipotética μ_0. Tem-se que calcular a média e o desvio-padrão (s) da amostra.

$$t = \frac{\overline{x} - \mu_o}{s \div \sqrt{n}}$$

Teste t para duas amostras dependentes

É também conhecido por teste t pareado. É utilizado em situações experimentais quando uma característica ou indivíduo é avaliado antes e após receber um tratamento. Nesse caso, o objetivo é testar se a diferença entre as duas respostas tem valor zero.

$$t = \frac{\overline{x}_1 - \overline{x}_2}{s(\overline{x}_1 - \overline{x}_2)} = s(\overline{x}_1 - \overline{x}_2) = \sqrt{\frac{s_1^2 + s_2^2}{n}}$$

Teste t para duas amostras independentes

É o caso de duas amostras independentes obtidas, aleatoriamente, de duas populações independentes a serem comparadas. Várias são as situações considerando as amostras e as variâncias:

- Tamanhos amostrais iguais ($n_1 = n_2 = n$) e variâncias iguais ($s_1^2 = s_2^2 = s^2$):

$$t = \frac{\overline{x}_1 - \overline{x}_2}{s_{x1x2}\sqrt{\frac{2}{n}}}$$

em que:

$$s_{x1x2} = \sqrt{\frac{1}{2}(s_1^2 + s_2^2)}$$

- Tamanhos amostrais diferentes ($n_1 \neq n_2$) e variâncias iguais ($s_1^2 = s_2^2 = s^2$):

$$t = \frac{\overline{x}_1 - \overline{x}_2}{s_{x1x2}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

em que:

$$s_{x1x2} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_{x1x2}}{n_1 + n_2 - 2}}$$

- Tamanhos amostrais diferentes ($n_1 \neq n_2$) e variâncias diferentes ($s_1^2 \neq s_2^2$):

$$t = \frac{\overline{X}_1 - \overline{X}_2}{S_{\overline{X}1-\overline{X}2}}$$

em que:

$$S_{\overline{X}1-\overline{X}2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Aplicação

A média de peso ao nascer de bezerros de bovinos da raça Nelore de um município é μ_0 = 30,7 kg, ou seja, é a média de uma população hipotética. Analisando-se pesos ao nascer de uma amostra de 20 bezerros, a média foi 30,0 kg e o desvio-padrão, 1,05 kg. Essa média difere ou não da população hipotética?

$$t_{19} = \frac{30,0 - 30,7}{1,05 \div \sqrt{20}} = -2,9814$$

A função *probt(x,gl,nc),* na rotina SAS a seguir, retorna a probabilidade da variável x com distribuição t de Student, com graus de liberdade (GL) e parâmetro de não centralidade nc (opcional); se não for especificado, nc = 0.

```
data testet;
input x gl nc;
area = probt(x,gl,nc);
alfa = 1-area;
datalines;
-2.9814  19   0
;
proc print;var area alfa; run;
Output
area          alfa
0.0038     0.9962
```

Conclusão: para o teste t com 19 graus de liberdade, tem-se que módulo de t é 2,9814, |t= 2,9814|. Assim, rejeita-se a hipótese de nulidade e conclui-se que a média de peso de 30,0 kg é significativamente inferior à média da população hipotética (μ0 = 30,7 kg).

Testes associados com a análise de variância

Exemplo de análise de variância e uso do teste F

Após a coleta de dados de um experimento, a próxima atividade é realizar a análise de variância (Anova), e a principal decisão baseia-se no uso do teste F que indica se há ou não significância entre os efeitos de tratamentos.

Considera-se, como exemplo, o quadro de análise de variância (Tabela 5), da análise do rendimento, em tonelada por hectare (t ha^{-1}) de matéria seca, de um experimento em que foram avaliadas cinco espécies de capim (Tanzânia, *Brachiaria*, Marandu, Pojuca e *Coast-cross*) utilizadas como forrageiras na alimentação de bovinos e quatro doses de adubação nitrogenada (0 kg ha^{-1}, 20 kg ha^{-1}, 40 kg ha^{-1} e 60 kg ha^{-1}). Após o cálculo do teste F, verificou-se que houve significância estatística entre os capins e, também, entre as doses, o que significa que a adubação nitrogenada influenciou a produtividade, e os capins se comportaram diferentemente.

Tabela 5. Análise de variância.

Fator de variação	Grau de liberdade	Soma de quadrados	Quadrado médio	F	
Capim – C	4	18,1823	4,5455	13,42	(<,0001)
Dose – D	3	4,0525	1,3508	3,99	0,0141
Interação C x D	12	5,2083	0,4340	1,28	0,2668
Resíduo	40	13,5467	0,3386		
Total	59	40,9898			

Contrastes ortogonais

Um contraste ortogonal (C) é uma combinação linear de *k* médias com *r* repetições cada e cuja soma dos coeficientes é zero.

$$C = \sum_{i=1}^{k} c_i\mu_i = c_1\mu_1 + c_2\mu_2 + \ldots + c_k\mu_k \, (c_1 + c_2 + \ldots + c_k = 0)$$

A variância de um contraste é definida por:

$$Var(C) = QMR \sum_{i=1}^{k} \frac{C_i^2}{r_i}$$

Para um conjunto de *k* tratamentos, tem-se k - 1 contrastes ortogonais com um grau de liberdade cada, os quais são independentes entre si, o que significa que o produto dos coeficientes correspondentes de dois contrastes é zero. Em uma Anova, os contrastes devem ser planejados a priori, isto é, por ocasião da instalação do experimento.

Os contrastes ortogonais são importantes para testar hipóteses complexas em uma Anova ou regressão múltipla. Em uma Anova, o uso de contrastes ortogonais não requer que o efeito de tratamentos seja significativo pelo teste F. Entretanto, se F é significativo, pelo menos um contraste também será.

Na Tabela 6, com as médias e respectivos totais dos cinco tratamentos, apresentam-se os coeficientes dos quatro contrates ortogonais. Outros conjuntos de contrastes ortogonais podem ser formulados de acordo com o interesse.

Tabela 6. Médias e contrastes.

Capim	Média	Total (12 repetições)	Contraste			
			C_1	C_2	C_3	C_4
1	3,75	45,00	-1	0	-1	-1
2	2,99	35,88	+1	0	-1	-1
3	2,47	29,64	0	-1	+1	-1
4	2,31	27,72	0	+1	+1	-1
5	2,30	27,60	0	0	0	+4

A partir dos contrastes apresentados na Tabela 6, tem-se:

a) Estimativa dos contrastes:

$\hat{C}_1 = -1 \times 45{,}00 + 1 \times 35{,}88 + 0 \times 29{,}64 + 0 \times 27{,}72 + 0 \times 27{,}60 = -9{,}12$
$\hat{C}_2 = 0 \times 45{,}00 + 0 \times 35{,}88 - 1 \times 29{,}64 + 1 \times 27{,}72 + 0 \times 27{,}60 = -1{,}92$
$\hat{C}_3 = -1 \times 45{,}00 - 1 \times 35{,}88 + 1 \times 29{,}64 + 1 \times 27{,}72 + 0 \times 27{,}60 = -23{,}52$
$\hat{C}_4 = -1 \times 45{,}00 - 1 \times 35{,}88 - 1 \times 29{,}64 - 1 \times 27{,}72 + 4 \times 27{,}60 = -27{,}84$

b) Variância das estimativas dos contrastes: $\text{Var}(\hat{C}) = \text{QMR} \sum_{i=1}^{k} \frac{C_i^2}{r_i}$

$\text{Var}(\hat{C}_1) = 0{,}3386 \times [(-1)^2 + 1^2 + 0^2 + 0^2 + 0^2)]/12 = 0{,}3386 \times 2/12 = 0{,}0564$
$\text{Var}(\hat{C}_2) = 0{,}3386 \times [(0^2 + 0^2 + (-1)^2 + 1^2 + 0^2)]/12 = 0{,}3386 \times 2/12 = 0{,}0564$
$\text{Var}(\hat{C}_3) = 0{,}3386 \times [(-1)^2 + (-1)^2 + 1^2 + 1^2 + 0^2)]/12 = 0{,}3386 \times 4/12 = 0{,}1129$
$\text{Var}(\hat{C}_4) = 0{,}3386 \times [(-1)^2 + (-1)^2 + (-1)^2 + (-1)^2 + 4^2)]/12 = 0{,}3386 \times 20/12 = 0{,}5643$

c) Soma de quadrados dos contrastes: $SQC_i = \frac{(\sum_{i=1}^{k} c_i T_i)^2}{r(\sum_{i=1}^{k} C_i^2)}$:

$$SQ\,\hat{C}_1 = \frac{[(-1x45{,}00 + 1x35{,}88 + 0x29{,}64 + 0x27{,}72 + 0x27{,}60)^2}{12[(-1)^2 + (1)^2]} = \frac{(-1{,}92)^2}{12x2} = 3{,}46566$$

$$SQ\,\hat{C}_2 = \frac{[(0x45{,}00 + 0x35{,}88 + 1x29{,}64 + 1x27{,}72 + 0x27{,}60)^2}{12[(-1)^2 + (1)^2]} = \frac{(-1{,}92)^2}{12x2} = 0{,}1536$$

$$SQ\,\hat{C}_3 = \frac{[(-1x45{,}00 + 1x35{,}88 + 1x29{,}64 + 1x27{,}72 + 0x27{,}60)^2}{12[(-1)^2 + (-1)^2 + 1^2 + 1^2]} = \frac{(-23{,}52)^2}{12x4} = 11{,}5248$$

$$SQ\,\hat{C}_4 = \frac{[(-1x45{,}00 - 1x35{,}88 - 1x29{,}64 - 1x27{,}72 + 4x27{,}60)^2}{12[(-1)^2 + (-1)^2 + (-1)^2 + (-1)^2 + (4)^2]} = \frac{(-27{,}84)^2}{12x20} = 3{,}2294$$

d) Produto dos coeficientes entre dois contrastes:

$\hat{C}_1$ x $\hat{C}_2$ = -1 x 0 + 1 x 0 + 0 x-1 + 0 x 1 + 0 x 0 = 0
$\hat{C}_1$ x $\hat{C}_3$ = -1 x-1 + 1 x-1 + 0 x 1 + 0 x 1 + 0 x 0 = 0
$\hat{C}_1$ x $\hat{C}_4$ = -1 x-1 + 1 x-1 + 0 x-1 + 0 x-1 + 0 x 4 = 0
$\hat{C}_2$ x $\hat{C}_3$ = 0 x-1 + 0 x-1 + -1x 1 + 1 x 1 + 0 x 0 = 0
$\hat{C}_2$ x $\hat{C}_4$ = 0 x-1 + 0 x-1 + -1 x-1 +1 x-1 + 0 x 4 = 0
$\hat{C}_3$ x $\hat{C}_4$ = -1 x-1 + -1 x-1 + 1 x-1 + 1 x-1 + 0 x 4 = 0

Para a significância de cada contraste (Tabela 7), para α = 0,05, pode ser consultada, na Tabela 1.4 (Anexo1), a distribuição F com 1 GL no numerador (GLN) e 40 GL no denominador (GLD). Observa-se que $F_{1,40}$ = 4,08, indicando que apenas o contraste 2 não é significativo. No entanto, o nível de significância exato pode ser obtido por meio da rotina SAS.

```
data;
input contraste gln gld f;
area = probf(f, gln, gld);
alfa = 1- area;
datalines;
1 1 40   10.23
2 1 40    0.45
3 1 40   34.04
4 1 40    9.54
;
```

```
proc print; var contraste  alfa;run;
output
contraste      alfa
   1        0.00270
   2        0.50619
   3        0.00000
   4        0.00365
```

Tabela 7. Análise de variância dos contrastes ortogonais.

Fator de variação	Grau de liberdade	Soma de quadrados	Quadrado médio	F
Capim – C	4	18,1771	4,5455	13,42 (<,0001)
Contraste 1	1	3,4656	3,4656	10,23 (<,0027)
Contraste 2	1	0,1536	0,1536	0,45 (>,0500)
Contraste 3	1	11,5248	11,5248	34,04 (<,0001)
Contraste 4	1	3,2294	3,2294	9,54 (<,0036)
Resíduo	40	13,5467	0,3386	
Total	59	40,9898		

No caso das quatro doses de adubação nitrogenada, como os níveis são equidistantes (0 kg ha^{-1}, 20 kg ha^{-1}, 40 kg ha^{-1} e 60 kg ha^{-1}), podem-se formular três contrastes ortogonais (Tabela 8). Como tem três graus de liberdade e os tratamentos são doses de adubo, o interesse é desdobrar esses três GL em componente linear, quadrático e cúbico.

Tabela 8. Três contrastes ortogonais.

Dose (4 níveis)	Média	Total	Contraste		
			Linear – L	Quadrático – Q	Cúbico – C
0	2,34	35,20	-3	1	-1
20	2,80	42,10	-1	-1	3
40	2,87	43,00	1	-1	-3
60	3,05	45,80	3	1	1
Produto entre contrastes LxQ = -3 x 1 + -1 x-1 + 1 x-1 + 3 x 1 = 0 LxC = -3 x-1 + -1 x 3 + 1 x-3 + 3 x 1 = 0 QxC = 1 x-1 + -1 x 3 + -1 x-3 + 1 x 1 = 0					

Seguindo-se o mesmo raciocínio dos cálculos executados para capim, calculam-se as somas de quadrados (SQ) para os três contrastes (Tabela 9). Dividindo-se cada SQ do contraste por QMR, obtém-se o correspondente valor de F com graus de liberdade 1 no numerador (GLN) e 40 no denominador (GLD). Na rotina SAS a seguir, calculam-se as correspondentes probabilidades do erro tipo I. Observa-se efeito linear significativo para as dosagens.

```
data;
input contraste gln gld f;
area = probf(f, gln, gld);
alfa = 1- area;
datalines;
1  1  40   10.5257
2  1  40    0.8275
3  1  40    0.6143
;
proc print; var contraste  alfa;run;
output
contraste      alfa
   1         0.0024
   2         0.3684
   3         0.4378
```

Tabela 9. Análise de variância com três contrastes ortogonais.

Fator de variação	Grau de liberdade	Soma de quadrados	Quadrado médio	F
Dose	3	4,0522	1,3507	3,9900 (<,0141)
Linear	1	3,5640		10,5257 (<,0024)
Quadrático	1	0,2802		0,8275 (>,3684)
Cúbico	1	0,2080		0,6143 (>,4378)
Resíduo	40	13,5467	0,3386	
Total	59	40,9898		

Comparações pareadas

A significância do teste F em uma Anova informa que os tratamentos foram significativos, mas não se sabe quais médias diferem entre si. Uma alternativa é localizar essas diferenças por meio de contrastes ortogonais. Porém, o mais comum é realizar comparações pareadas entre todas as médias dos tratamentos; as *k* médias de

um experimento são colocadas em ordem decrescente, e o número de comparações pareadas é k(k - 1)/2. Os testes utilizados nas comparações pareadas são: teste t, teste de Student Newman-Keuls (SNK), teste de Tukey, teste de Scheffé, teste de Duncan, teste de Dunnett e teste de Bonferroni (Hicks, 1973; Sampaio, 1998).

A estimativa de um contraste entre duas médias é dada por: $\hat{C}_{ij} = \overline{y}_i - \overline{y}_j$.

Se as médias têm repetições iguais, a estimativa da variância do contraste é dada por $Var(\hat{C}_{ij}) = \sqrt{2QMR/r}$; caso contrário, fica $Var(\hat{C}_{ij}) = \sqrt{\frac{QMR}{2}\left(\frac{1}{ri} + \frac{1}{rj}\right)}$.

Em seguida, calcula-se a diferença mínima significativa (DMS), que é a diferença em que duas médias diferem, significativamente, entre si a um nível de probabilidade α. A DMS é construída com a $Var(\hat{C}_{ij})$ e o valor tabelado de um teste de comparação de médias a um nível α. A DMS será demonstrada nos diversos testes utilizados nas comparações pareadas a seguir.

Teste t

O teste t é o mais utilizado para testar contrastes ortogonais, de preferência que eles sejam escolhidos previamente. Porém, nas comparações pareadas, alguns cuidados devem ser tomados. Colocando-se as k(k - 1)/2 comparações pareadas de um experimento em ordem decrescente, utilizando-se o teste t, apenas na comparação envolvendo as duas médias maiores, tem-se erro do tipo I correto (α = 0,05). Nas demais comparações a probabilidade do erro do tipo I é dada por $\alpha = 1 - (1 - 0{,}05)^k$, em que *k* é o número de médias incluídas no contraste. Se o contraste é entre a primeira e a terceira média, então k = 3, e assim por diante.

Considerando-se o quadro de análise de variância (Tabela 5), e as médias dos cinco tratamentos (Tabela 6), o número de comparações pareadas que pode realizar é dado por 5(4)/2 = 10.

Dessas dez comparações pareadas, em sete delas houve significância estatística ($\overline{y}_i \neq \overline{y}_j$), isto é, $\hat{C}_{ij} > DMS$. No entanto, apenas na primeira comparação, o teste de hipótese tem erro do tipo I correto (α = 0,05). Para as demais comparações, esse erro vai aumentando, de modo que o maior erro do tipo I é na comparação envolvendo as duas médias mais extremas.

Comparação $\hat{C}_{ij} = \bar{y}_i - \bar{y}_j$	Valor de α
1 – 2 = 0,76	0,0500
1 – 3 = 1,28	0,0975
1 – 4 = 1,44	0,1426
1 – 5 = 1,45	0,1854
2 – 3 = 0,52	0,2262
2 – 4 = 0,68	0,2649
2 – 5 = 0,69	0,3016
3 – 4 = 0,16	0,3366
3 – 5 = 0,17	0,3697
4 – 5 = 0,01	0,4013

Cálculo da DMS:

$$DMS(t) = t\sqrt{\frac{2QMR}{r}} = 1,6839 = \sqrt{\frac{2x0,3386}{12}} = 0,3982$$

em que:

QMR = quadrado médio do resíduo = 0,3386.

r = número de repetições por tratamento = 12.

t = 1,6839, valor tabelado para α = 0,05.

GL do resíduo = 40 (Tabela 1.3, do Anexo 1).

Teste de Student Newman-Keuls (SNK)

Foi proposto por Newman (1939), conforme citação de Sampaio (1998), com o propósito de corrigir as distorções do teste t para mais de dois tratamentos. Colocando-se as *k* médias em ordem crescente ou decrescente, é feita uma comparação entre a maior e a menor. Se a diferença for significativa, então continua com as comparações entre a maior e a menor média do grupo restante. Uma diferença entre duas médias *i* e *j* é significativa se:

$$(\bar{y}_i - \bar{y}_j)/\sqrt{Var(\hat{C}_{ij})} \geq q$$

em que:

q = é um valor tabelado em função de α, dos graus de liberdade do erro e do número de médias incluídas na comparação (p + 2; sendo p o número de médias entre as duas médias a serem comparadas).

Por exemplo, se existe um conjunto de cinco médias ordenadas, então, na comparação entre a maior e a menor delas, existem três médias intermediárias, e *p* é igual a três.

Os valores de *q* diminuem com o aumento do número de graus de liberdade do resíduo (GLR), mas aumentam com a distância entre as médias, o que possibilita corrigir erros do tipo I. Para um conjunto pequeno de tratamentos, o teste SNK é mais poderoso e menos conservativo do que o teste de Tukey (que será discutido adiante).

Cálculo da DMS:

$$DMS_{(SNK)} = q\sqrt{\frac{QMR}{r}} = 4{,}04\sqrt{\frac{0{,}3386}{12}} = 0{,}6786$$

em que:

QMR = quadrado médio do resíduo = 0,3386.

r = número de repetições por tratamento = 12.

q = 4,04, valor tabelado para α = 0,05; GL do resíduo = 40 e *p* = 5 (Tabela 1.5, Anexo 1).

Teste de Tukey

Este teste, proposto por Tukey, em 1953, é bastante rigoroso e controla muito bem o erro do tipo I. Ele corrige as limitações do teste t quando o número de comparações pareadas é maior que 1. No entanto, permite o aparecimento do erro tipo II, ou seja, considera duas médias iguais, quando na verdade elas são diferentes. Dependendo do número de comparações pareadas, esse teste pode até prejudicar a eficiência de um tratamento, isto é, considerar que ele não difere, quando, na verdade, houve significância. Quando os tamanhos amostrais dos grupos são iguais, o teste de Tukey é exato, isto é, para as *k* médias de um experimento, em que se tem k(k - 1)/2 comparações duas a duas, em que a taxa do erro é exatamente α, e a do intervalo de confiança é (1 - α). Diferentemente do teste de SNK, todas as comparações pareadas são feitas, podendo também estimar contrastes entre elas.

Considerando-se a Tabela 7, o valor da DMS do teste de Tukey é obtido por:

$$DMS_{(Tukey)} = q\sqrt{\frac{QMR}{r}} = 4{,}10\sqrt{\frac{0{,}3386}{12}} = 0{,}6786$$

em que:

QMR = quadrado médio do resíduo = 0,3386.

r = número de repetições por tratamento = 12.

q = 4,10, valor tabelado para α = 0,05; GL do resíduo = 40 e número de médias a serem comparadas (p = 5) (Tabela 1.6, Anexo 1).

Teste de Scheffé

Este teste foi proposto por Henry Scheffé em 1953, conforme publicação (Scheffé, 1969), com o objetivo de comparar qualquer contraste entre médias e permitir diferente número de repetições por tratamento. O teste é um pouco mais rigoroso que o de Tukey, merecendo, portanto, os mesmos comentários com relação ao perigoso aumento do erro do tipo II. Quando há muitos contrastes, o teste de Scheffé é mais poderoso do que Bonferroni, que será discutido adiante.

Cálculo da DMS:

$$DMS_{(Scheffé)} = \sqrt{glnxF(gln.gld)xVar(\hat{C}_{ij})} = \sqrt{4x2,61 \text{ x } \frac{0,3386}{12}} = 0,5427$$

em que:

$F_{4,40}$ = 2,61, valor tabelado para α = 0,05, graus de liberdade do numerador = 4 e graus de liberdade do denominador = 40 (Tabela 1.4, Anexo 1).

Teste de Duncan

Introduzido por Duncan (1955), este teste indica resultados significativos em situações em que o teste de Tukey não permite obter. Da mesma forma que o teste de Tukey, os resultados são exatos quando todos os tratamentos têm o mesmo número de repetições. Seguindo a estrutura do SNK, coloca as k médias em ordem decrescente e faz uma comparação entre a maior e a menor média. Somente se essa diferença for significativa, é que outras comparações no intervalo são feitas. Comparado ao SNK, o teste de Duncan tem o mesmo rigor no controle do erro tipo I, mas, na comparação de médias mais afastadas, cria uma oportunidade para o aparecimento do erro tipo II, uma vez que o maior valor da diferença mínima significativa se equipara com aquela do teste Tukey. O teste de Duncan é, porém, menos conservador, isto é, proporciona diferenças significativas com mais facilidade, conduzindo a afirmativas erradas com maior frequência.

Cálculo da DMS:

$$DMS_{(Duncan)} = q\sqrt{\frac{QMR}{r}} = 3,171\ \sqrt{\frac{0,3386}{12}} = 0,5327$$

em que:

QMR = quadrado médio do resíduo = 0,3386.

r = número de repetições por tratamento = 12.

q = 3,171, valor tabelado para α = 0,05; grau de liberdade do resíduo = 40 e p = 5 (Tabela 1.7, Anexo 1).

Teste de Dunnett

O teste de Dunnett é utilizado para comparar as médias de cada tratamento com o tratamento-controle; de *k* tratamentos, tem-se (k - 1) comparações. Dos cinco capins analisados, considerando o capim 1, como controle, tem-se:

Controle	Grupo tratado			
Capim 1	Capim 2	Capim 3	Capim 4	Capim 5

Cálculo da DMS:

$$DMS_{(Duncan)} = q\sqrt{\frac{2QMR}{r}} = 2{,}54\sqrt{\frac{2x0{,}3386}{12}} = 0{,}6034$$

em que:

QMR = quadrado médio do resíduo = 0,3386.

r = número de repetições por tratamento = 12.

q = 2,54, valor tabelado para α = 0,05; GL do resíduo = 40 e número de tratamentos (k = 5) (Tabela 1.8, Anexo 1).

Teste de Bonferroni

Esse teste é um aperfeiçoamento do teste t, e, às vezes, é chamado de teste t de Bonferroni; é conservativo quando o número de médias é grande.

A mesma preocupação discutida no teste t sobre a probabilidade de obter resultados falsos positivos (erro tipo I) vale para o teste de Bonferroni.

Para reduzir a probabilidade de obter resultados falsos positivos (erro tipo I), foi criada a correção de Bonferroni. Para *k* possíveis combinações de médias, o valor de α é corrigido para α/k.

No caso de cinco tratamentos, tem-se k(k - 1)/2 = 10 comparações pareadas, duas a duas. Para um erro experimental de α = 0,05, precisa usar 0,05/10 = 0,005, como nível de significância para cada teste, ou seja, 0,5%.

O erro do tipo 1 é igual a $1 - (1 - 0,005)^{10} = 0,0488$, que é menos que 0,05.

Cálculo da DMS:

$$DMS_{(Bonferroni)} = t\sqrt{\frac{2xQMR}{r}} = 2,7045\sqrt{\frac{2x0,3386}{12}} = 0,6425$$

em que:

QMR = quadrado médio do resíduo = 0,3386.

r = número de repetições por tratamento = 12.

t = 2,7045, para valor tabelado para α = 0,005 e GL do resíduo = 40 (Tabela 1.3, Anexo 1).

Na Tabela 10, resume-se a significância para os sete testes discutidos anteriormente.

Tabela 10. Testes de comparações pareadas e suas probabilidades.

i,j	dif	Teste t	SNK	Tukey	Scheffé	Duncan	Dunett	Bonferroni
1,2	0,76	0,0028	(*)	0,0218	0,0541	(*)	0,0099	0,0275
1,3	1,28	< ,0001	(*)	< ,0001	0,0002	(*)	< ,0001	< ,0001
1,4	1,43	< ,0001	(*)	< ,0001	<,0001	(*)	< ,0001	< ,0001
1,5	1,44	< ,0001	(*)	< ,0001	<,0001	(*)	< ,0001	< ,0001
2,3	0,52	0,0356	(*)	0,2101	0,3332	(*)		0,3562
2,4	0,68	0,0070	(*)	0,0518	0,1103	(*)		0,0704
2,5	0,68	0,0064	(*)	0,0477	0,1031	(*)		0,0642
3,4	0,16	0,5090	ns	0,9624	0,9779	ns		1,0000
3,5	0,17	0,4870	ns	0,9550	0,9734	ns		1,0000
4,5	0,01	0,9722	ns	1,0000	1,0000	ns		1,0000

(*) = significativo; ns = não significativo.

Em resumo, teste t deve ser utilizado para testar contrastes ortogonais que foram escolhidos previamente, cujo número não exceda os graus de liberdade de tratamentos; não é recomendável para comparações pareadas. Se o pesquisador quer ter alta chance de rejeitar a hipótese de que as médias são iguais, pode optar pelo teste Duncan e SNK.

Se o interesse é rejeitar a hipótese de que as médias são iguais com muita confiança, deve optar pelo teste de Tukey.

Quando deseja comparar todos os tratamentos com um tratamento-controle ou referência, o teste adequado é o de Dunnett. Já o teste de Scheffé é aconselhável para testar contrastes mais complicados; para comparações pareadas, ele somente detecta significância se esta já foi comprovada pelo teste F global, portanto, bastante rigoroso. O teste de Bonferroni, com a correção, também é um teste que pode ser usado, porém é bastante rigoroso, e mais recomendado para testar contrastes mais complexos.

Erro-padrão da média, coeficiente de variação e coeficiente de determinação

Erro-padrão da média

Corresponde ao desvio-padrão (s) de uma população de *n* médias amostrais dividido pela raiz quadrada do tamanho amostral ($s_{\overline{x}} = \frac{s}{\sqrt{n}}$). Na análise de variância, o $s_{\overline{x}}$ e uma média de tratamento com *r* repetições é calculado pela fórmula a seguir, em que QMR é o quadro médio do resíduo. Se as médias dos tratamentos têm repetições iguais, o valor é calculado apenas uma vez; caso contrário, *r* deve ser alterado para cada média.

$$s_{\overline{x}} = \sqrt{\frac{QMR}{r}}$$

Coeficiente de variação

É a razão entre o desvio-padrão (s) e a média aritmética $(\overline{x})$ multiplicada por 100 que expressa a variação como percentagem da média. Em uma análise, o coeficiente de variação (CV) é calculado por:

$$CV = \sqrt{\frac{QMR}{\overline{x}}} = 100$$

em que:

$\overline{x}$ = média geral do experimento.

QMR = Quadrado médio do resíduo.

Teoricamente, o CV varia de zero (quando todos os dados são iguais), o que é praticamente impossível, até valores bem superiores a 100%. De modo geral, na

agricultura, em experimentos bem planejados, bem conduzidos e com variáveis estáveis, o CV geralmente não ultrapassa a 30%.

Não existe um conceito sobre os valores do CV; entretanto, na experimentação agrícola, de 0,0 a 10,0%, o CV é considerado baixo, de 10,0% a 20,0%, médio a alto, de 20,0% a 30,0% é alto. Em experimentos com herbicidas na contagem de plantas daninhas em parcelas, ele pode atingir valores muito altos, às vezes bem superiores a 100,0%. Quando dois experimentos são semelhantes e instalados próximos, valores diferentes de CV podem refletir falhas na condução do experimento, na coleta dos dados e na análise.

Coeficiente de determinação

É a proporção da variação que é explicada pelo ajuste de um modelo em relação à variação total. Para calcular o coeficiente de determinação (R^2) em uma análise de variância, no numerador, colocam-se as somas de quadrados (SQ) de todos os efeitos ajustados, com exceção da SQ do erro, e dividem-se pela SQ total. Utilizando-se os resultados da Anova da Tabela 5, tem-se:

$$R^2 = \frac{\text{variação explicada}}{\text{variação total}}$$

$$R^2 = \frac{\text{SQ(capim + dose + interação capim x dose)}}{\text{variação total}} = \frac{(18{,}1823+4{,}0525+5{,}2083)}{40{,}9898} = \frac{27{,}3838}{40{,}9898} = 0{,}66$$

Quanto mais próximo de 1,0 é o valor de R^2, melhor é a qualidade dos dados e mais eficiente é a análise estatística realizada.

Intervalos de confiança

Na estatística, o termo inferência refere-se ao conjunto de métodos utilizados para obter medidas representativas da população a partir da amostra. Assim, um estimador obtido a partir de uma amostra precisa de algumas características para representar o parâmetro populacional, as quais são:

a) Consistência: um estimador consistente é aquele que tende para o valor verdadeiro do parâmetro quando o tamanho da amostra cresce.

b) Acurácia: correlação entre o valor verdadeiro e o valor estimado (predito).

c) Precisão: grau de repetitividade de um resultado.

d) Viés ou vício: é a diferença entre a medida ou a média de uma característica em relação ao seu verdadeiro valor.

e) Eficiência: um estimador eficiente possui precisão ou acurácia e exatidão ao mesmo tempo.

Como as estimativas de um parâmetro são obtidas de amostras sujeitas a flutuações, para calcular um intervalo de confiança (IC), primeiramente define-se o nível de significância alfa e delimita os limites de confiança: limite inferior (LI) e superior (LS), dentro do qual há probabilidade 1- alfa de encontrar o parâmetro.

Segue-se uma interpretação do IC:

"Se 100 amostras similares são obtidas de uma população e para cada uma calcula-se um intervalo de confiança com 95% de probabilidade, nós estamos confiantes de que 95% dos intervalos calculados incluem o verdadeiro parâmetro populacional".

Uma interpretação mais usual dentro da experimentação, quando se analisa um experimento, é afirmar que temos 95% de certeza de que o intervalo de confiança inclui o verdadeiro parâmetro populacional. Portanto, são a variação dos dados de uma amostra e o grau de confiança que queremos (95%, 99%, por exemplo), que delimitam a amplitude do IC. A largura do IC dá uma ideia acerca da incerteza que se tem do parâmetro populacional. Um IC largo sugere que mais dados devem ser coletados para reduzir a incerteza acerca do parâmetro. Uma vez que fornece o intervalo que incluirá o parâmetro populacional de interesse, o IC é mais informativo do que os testes de hipóteses.

Para uma média amostral $\overline{x}$ em que $s(\overline{x})$ é o erro-padrão da média, *z* é a normal reduzida e *t* é o valor tabelado de com n - 1 graus de liberdade, um intervalo com 100(1 - α) de confiança para a média populacional ou parâmetro μ é dado por:

- Variância conhecida: $\overline{x} - s(\overline{x})z_{\alpha/2} \leq \mu \leq \overline{x} + s(\overline{x})z_{\alpha/2}$.
- Variância desconhecida: $\overline{x} - s(\overline{x})t_{\alpha/2;\, n-1} \leq \mu \leq \overline{x} + s(\overline{x})t_{\alpha/2;\, n-1}$.

O IC para a diferença entre duas médias populacionais é (μ1 - μ2). Para esse exemplo, considera-se a Tabela 11. O IC será construído para a diferença entre as duas primeiras médias.

Tabela 11. Médias e total de 12 repetições para a produção, em toneladas por hectare (t ha^{-1}), de cinco capins.

Capim	Média	Total (12 repetições)
1	3,75	45,00
2	2,99	35,88
3	2,47	29,64
4	2,31	27,72
5	2,30	27,60

Teste de hipótese bilateral:

$H_0: \mu_1 = \mu_2$ versus $H_a: \mu_1 \neq \mu_2$

ou:

$H_0: \mu_1 - \mu_2 = 0$ versus $H_a: \mu_1 - \mu_2 \neq 0$

Estimativa do contraste:

$$\hat{y}_1 = \overline{x}_1 - \overline{x}_2 = 3{,}75 - 2{,}99 = 0{,}76$$

Variância da estimativa do contraste:

$$\text{Var}(\hat{y}_1) = \text{QMR} \sum_{i=1}^{2} \frac{C_i^2}{r_i} = 0{,}3386 \times \frac{2}{12} = 0{,}0564$$

Erro-padrão da estimativa do contraste:

$$s(\hat{Y}_1) = \sqrt{0{,}0564} = 0{,}2375$$

Cálculo do intervalo de confiança:

$[y_1 - zs(y_1); y_1 + zs(y_1)]$

$[y_1 - ts(y_1); y_1 + ts(y_1)]$

O valor de *z* e *t* tabelado depende do grau de certeza que queremos (95%, 99%, por exemplo) para o intervalo de confiança. Para 95% de segurança, z = 1,96, valor obtido da Tabela 1.1 (Anexo 1), tem-se:

$[0{,}76 - 1{,}96 \times 0{,}2375; 0{,}76 + 1{,}96 \times 0{,}2375]$

$[0{,}2945; 1{,}2255]$

Diagnóstico de adequabilidade

Além das estatísticas anteriores, são recomendados teste de normalidade e de homogeneidade de variâncias dos dados e teste de normalidade dos erros.

Teste de normalidade dos dados

Alguns testes testam a hipótese nula de que os dados em estudo correspondem a uma amostra aleatória proveniente de uma distribuição normal.

- Shapiro-Wilks (S-W): apropriado para tamanho amostral menor ou igual a 2 mil. Calcula-se a estatística W ($0 < W \leq 1$) e sua probabilidade ($0 \leq Prob \leq 1$); se W é igual a 1, os dados ajustam perfeitamente à distribuição normal, enquanto valores pequenos de W são evidências de desvios da normalidade.
- Kolmogorov-Smirnov (K-S): avalia a discrepância entre a distribuição empírica Fn(y) e a distribuição normal considerada referência F(y). É apropriado para amostras grandes e testa a hipótese:

$$H_o: Fn(y) = F(y) \text{ versus } H_a: Fn(y) \neq F(y)$$

Esse teste é mais sensível em pontos próximos da mediana da distribuição do que nas caudas.

Aplicação

Na Tabela 12, os testes de Shapiro-Wilks e Kolmogorov-Smirnov são aplicados a dados de produtividade, em quilograma por hectare (kg ha^{-1}), de matéria seca, de 5 cortes de 92 cultivares de alfafa.

Tabela 12. Testes de normalidade: Shapiro-Wilks e Kolmogorov-Smirnov.

Corte	Shapiro-Wilks (S-W)	Kolmogorov-Smirnov (K-S)
1	0,3683	>0,1500
2	<0,0001	<0,0100
3	0,9148	>0,1500
4	0,0032	<0,0100
5	0,0001	>0,1500

Pelos dois testes, houve concordância de que os dados do corte três são os que mais se aproximam de uma distribuição normal. Os testes são concordantes quanto ao afastamento da normalidade dos cortes dois e quatro.

Teste de homogeneidade de variâncias

Pelo procedimento do modelo linear generalizado (GLM, do inglês *General linear model*) com a opção *hovtest*, usa-se o teste de Levene para testar a homogeneidade de variâncias de um grupo de tratamentos. Pode-se usar também a opção *welch* em uma Anova quando as variâncias não são assumidas homogêneas. O teste Welch é robusto para esta situação.

Considerando-se os dados de capins e doses discutidos neste capítulo, no programa SAS a seguir, com a opção "/ *hovtest welch*", calcula-se o teste de Levene para testar a homogeneidade de variâncias e realiza uma Anova robusta, mesmo sob a suposição de não homogeneidade de variâncias para os cinco capins.

```
proc glm data = capim;
  class bloco capim dose;
  model y = capim ;
means capim / hovtest welch;
run;
```

Pelo teste de Levene, rejeitou-se a hipótese de homogeneidade de variâncias entre os cinco capins (Prob > F = 0,0002). No entanto, a hipótese de efeito significativo entre capins foi confirmada pelo teste de Welch (Prob > F = 0,0001).

```
Source                  DF     Type III SS     Mean Square    F Value    Prob > F
capim                    4     18.18233333      4.54558333      10.96      <.0001
Levene's test for homogeneity of y variance
Anova of squared deviations from group means
                        sum of       mean
Source      DF         squares      square    F Value    Prob > F
capim        4          4.3871      1.0968       6.82      0.0002
error       55          8.8422      0.1608
welch's anova for y
source          DF     F value    Prob > F
capim       4.0000        8.62      0.0001
```

Teste de normalidade dos erros

É uma das exigências da Anova. Entretanto, duas propriedades estatísticas garantem que pequenos desvios de normalidade dos erros não afetam os resultados da análise de variância.

a) De acordo com o teorema do limite central, se de uma população retirarmos n amostras de tamanho grande (n > 30) e calcularmos as médias de cada uma delas, a distribuição destas médias é aproximadamente normal.

b) Como citado por Gill (1978), o teste F usado na Anova é robusto, e as probabilidades de erro do tipo I e do tipo II são pouco afetadas por moderados desvios de normalidade dos erros.

A independência dos erros é também uma exigência da Anova, mas essa suposição é atendida pela casualização dos tratamentos nas unidades experimentais. Usando o programa a seguir para os dados citados no item Testes Associados com a análise de variância (Anova), obtém-se a Figura 2 com a variável dependente no eixo *y* e os quantis normais no eixo *x*. Quanto mais próximo da reta é o gráfico, mais os erros se ajustam a uma distribuição normal.

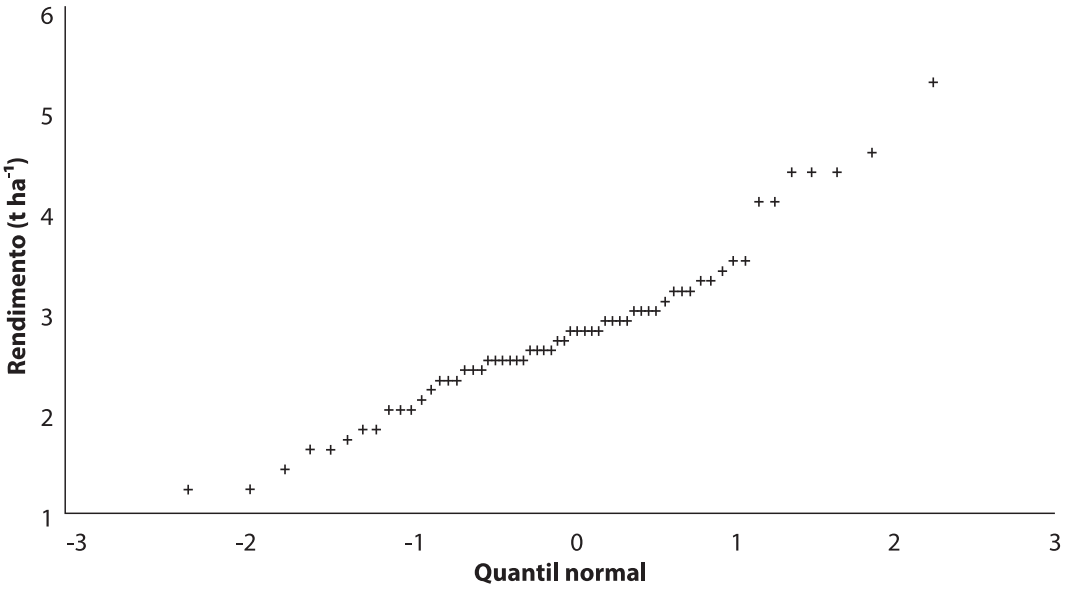


Figura 2. Rendimento em tonelada por hectare (t ha^{-1}) de alfafa (eixo *y*) e quantis normais no eixo *x*.

```
data capim;
label y = ' rendimento, t/ha';
run;
   symbol v=plus;
   title 'normal quantile-quantile plot for hole distance';
   proc univariate data=cap7 noprint;
      qqplot y;
   run;
```

Em resumo, o roteiro básico para a experimentação e Anova começa com a formulação da hipótese científica que geralmente é colocada no item Material e Método de um projeto de pesquisa.

Em seguida, tem-se que pensar no experimento, etapa mais importante do método científico, o qual tem a finalidade de investigar problemas, testar teorias e, principalmente, testar hipóteses. Três princípios básicos envolvem a experimentação: a) controle local − processo de avaliar o local ou ambiente onde será instalado o experimento e escolher o desenho experimental mais adequado; b) repetição − o número de parcelas ou unidades experimentais em que um tratamento será aplicado; é necessário para estimar o erro experimental e determinar a precisão com que estimamos um parâmetro; c) casualização − processo de distribuição aleatória dos tratamentos às parcelas ou unidades experimentais, que garante a independência dos erros, umas das suposições mais importantes da experimentação.

Com a instalação, condução e conclusão do experimento no campo, tem-se que avaliar a hipótese por meio de um teste estatístico, para tomar decisões sobre a rejeição ou não da hipótese H_0, sendo α o nível de significância do teste.

Na formulação dos testes estatísticos utilizados na Anova para testar a hipótese estatística, é fundamental a participação das distribuições de probabilidades, tais como a normal padronizada (z), qui-quadrado (χ^2), F e t, as quais fornecem os respectivos testes associados ao valor crítico α e graus de liberdade.

Havendo significância do teste F na Anova, surge a necessidade de três estatísticas: a) contrastes ortogonais − de preferência que sejam escolhidos previamente; b) comparações pareadas − comparações entre todos os pares de médias; c) comparação de todos os tratamentos com o controle ou tratamento referência. O teste t é o indicado para testar contrastes ortogonais. Já o teste de Scheffé é aconselhável para testar contrastes mais complicados, desde que uma significância global dos tratamentos tenha sido detectada pelo teste F.

Para as comparações pareadas, caso o interesse seja detectar significância entre as médias de tratamentos, sem preocupar muito com a eficiência ou confiança do teste, pode-se optar pelo teste Duncan e SNK. Caso o interesse é rejeitar a hipótese de que as médias são iguais com muita confiança, o recomendável é o teste de Tukey. Para comparar todos os tratamentos com um tratamento-controle, deve-se utilizar o teste de Dunnett.

Finalmente, várias estatísticas são fundamentais para indicar a precisão com que um experimento é conduzido e as conclusões obtidas. São elas, o erro-padrão da média, o coeficiente de variação (CV), o coeficiente de determinação (R^2), os intervalos de confiança. Além dessas estatísticas, recomenda-se também testar a homogeneidade de variâncias entre tratamentos e a normalidade dos erros.

Exercícios[7]

1) No texto a seguir, existem afirmações incorretas quanto a conceitos de estatística. Reescreva o texto colocando definições corretas e sublinhe ou coloque em negrito onde houve definições incorretas.

O empirismo é caracterizado pelo conhecimento científico e acredita nas experiências e na intuição como formadoras das ideias. O inglês Francis Bacon (1561−1626), que mostrou a importância da experimentação e do empirismo para a aquisição dos conhecimentos científicos, é considerado o fundador da ciência moderna, sendo atribuída a ele a frase “a matemática pode ajudar a compreender a natureza do mundo”. O italiano Galileu Galilei (1564−1642) é um dos responsáveis pelo estabelecimento das bases do pensamento científico moderno e o método experimental; o seu princípio racional era a matemática. Quanto ao experimento ou teste que é conduzido usando o método científico, pode-se afirmar que ele é usado para responder uma questão ou investigar um problema, testar teorias e, principalmente, testar hipóteses. Conforme afirmação do filósofo Popper (1959) qualquer hipótese é falsificada. Porém, as conclusões de um experimento sempre apresentam argumentos que permitem provar uma hipótese. Uma hipótese estatística é uma suposição feita acerca de um parâmetro ou de uma característica da população, a qual pode ser aceita ou rejeitada por meio de uma amostra aleatória retirada da população. Para formular uma hipótese estatística, sempre iniciamos com a hipótese alternativa (H_a), sendo que o objetivo é sempre rejeitar a hipótese de nulidade (H_0) em favor de hipóteses alternativas (H_a).

[7] As respostas dos exercícios podem ser consultadas no Apêndice 1.

2) Em um experimento o valor da estatística ou teste de χ^2 calculado com um grau de liberdade foi $\chi^2_1 = 7,87$. Os valores tabelados são P ($\alpha = 0,05$) = 3,84; P ($\alpha = 0,01$) = 6,64 e P($\alpha = 0,001$) = 10,83. Quais são as conclusões?

3) Formule uma hipótese científica e uma hipótese estatística em que a aplicação de doses de adubo nitrogenado (0 kg ha^{-1}, 20, 40, 60, 80 e 100 kg ha^{-1}) aumenta linearmente a produtividade de forragem, em kg de matéria seca ha^{-1}.

4) Material e Método ou Metodologia, é um item obrigatório na publicação de um artigo (*paper*) em revista científica, projeto de pesquisa, etc. Quais são os itens imprescindíveis que devem ser colocados?

5) Em uma publicação científica está escrito que 90% das vacas de uma região produzem acima de 25 kg de leite por dia. Se em uma amostra de 200 vacas dessa região constatou-se que 184 delas tinham produtividade acima de 25 kg dia^{-1} e 16 tinham produção inferior a 25 kg dia^{-1}. A hipótese é confirmada?

Capítulo 8

Modelo linear geral versus modelo misto

Introdução

Os modelos lineares gerais no Sistema de Análise Estatística (SAS) são ajustados pelo método dos quadrados mínimos pelo procedimento do modelo linear geral (GLM). Vários tipos de análises são executados: regressões, análises de variâncias univariadas e multivariadas, e, em situações especiais, podem-se fazer análises de medidas repetidas (MR).

Os modelos lineares mistos podem ser compreendidos como uma generalização dos modelos lineares padrão, e, no SAS, são executados pelo procedimento modelo linear misto (Mixed). Ele permite modelar as estruturas de variâncias e covariâncias dos erros.

A necessidade de modelar as variâncias e as covariâncias a partir do erro, nos modelos mistos, surge basicamente de duas situações:

a) As unidades experimentais sobre as quais os dados são avaliados podem ser agrupadas dentro de *clusters*, e os dados dentro deles são correlacionados.

b) Nas análises de medidas repetidas no tempo ou no espaço, várias medidas são avaliadas na mesma unidade experimental, e elas são correlacionadas.

As análises de medidas repetidas em estudos de crescimento de animal e vegetal, por exemplo, em que as medidas de pesos são feitas periodicamente nas unidades experimentais ou indivíduos, têm grande aplicação na pesquisa agropecuária. Essas análises representam grande superioridade do procedimento Mixed em relação ao GLM, pois o Mixed possibilita realizar análise de perfis cujo objetivo é estudar o efeito dos tratamentos nas diversas ocasiões de avaliação, podendo responder a questões do tipo:

a) Se os perfis médios de resposta dos diferentes tratamentos são paralelos, a interação entre tratamento e tempo é nula.

b) Se os perfis médios de resposta dos diferentes tratamentos são coincidentes, o efeito de tratamento é nulo.

c) Se os perfis médios de resposta dos diferentes tratamentos são horizontais, o efeito do tempo é nulo.

Dados utilizados nas aplicações do GLM e do Mixed

Para demonstrações de análises com os procedimentos do GLM e do Mixed, utilizaremos dados de produtividade de matéria seca (PMS) de 20 cortes sequenciais de avaliação de 92 genótipos de alfafa (*Medicago sativa* L.), distribuídos em delineamento de blocos completos casualizados e duas repetições. O experimento foi realizado em condições de campo na Embrapa Pecuária Sudeste, São Carlos, SP. As análises apresentadas, neste capítulo, com o GLM e o Mixed, utilizam a estrutura de dados descrita na Tabela 1.

O modelo linear associado aos dados da Tabela 1 é:

$$y_{ijk} = \mu + \alpha_i + \delta_{ij} + t_k + (\alpha t)_{ik} + \varepsilon_{ijk}$$

em que:

y_{ijk} = valor observado da PMS no corte *k*, na unidade experimental ou indivíduo *j* e no tratamento *i*.

μ = efeito médio global.

α_i, t_k, $(\alpha t)_{ik}$ = respectivamente, efeito fixo do tratamento *i*; efeito fixo do corte *k* e da interação de tratamento e corte.

δ_{ij} = efeito aleatório da unidade experimental *j* no tratamento *i*.

ε_{ijk} = efeito aleatório associado à PMS no corte *k*, na unidade experimental *j* e no tratamento *i*.

Tabela 1. Estrutura de dados em medidas repetidas.

Tratamento, bloco	Indivíduo	Avaliação no indivíduo (Corte de 1 a 20)
1,1	1	$y_{1,1,1}$ $y_{1,1,2}$ ... $y_{1,1,19}$ $y_{1,1,20}$
1,1	2	$y_{1,1,1}$ $y_{1,1,2}$ ... $y_{1,1,19}$ $y_{1,1,20}$
...	...	...
92,1	92	$y_{92,1,1}$ $y_{92,1,2}$ ... $y_{92,1,19}$ $y_{92,1,20}$
1,2	93	$y_{1,2,1}$ $y_{1,2,2}$ ... $y_{1,2,19}$ $y_{1,2,20}$
2,2	94	$y_{2,2,1}$ $y_{2,2,2}$ ... $y_{2,2,19}$ $y_{2,2,20}$
...	...	...
92,2	184	$Y_{92,2,1}$ $y_{92,2,2}$ ... $y_{92,2,19}$ $y_{92,2,20}$

Fonte: Freitas et al. (2011).

Estatísticas associadas ao GLM

A análise univariada inicia-se com o modelo matricial:

$$y_{nx1} = X_{nxp}\, b_{px1} + e_{nx1}$$

em que:

y_{nx1} = vetor de valores dependentes.

$E(y) = Xb;\ Var(y) = V(e) = R = \sigma^2 I_n$

σ^2 = quadrado médio do erro.

I_n = matriz de identidade de ordem n.

X = matriz de especificação.

b = vetor que contém os efeitos fixos.

e = vetor que contém os erros aleatórios, E(e) = 0.

O GLM exige que os erros e_{ijk} sejam independentes e identicamente distribuídos com média zero e com distribuição pelo menos aproximada da normal, ou seja, $V(e) = R = \sigma^2 I_n$. Isso pressupõe variância constante na diagonal principal e correlação nula para os elementos fora da diagonal.

Estimativa dos efeitos fixos

$$\hat{b} = (X'X)^{-1}X'Y.$$

Teste de uma hipótese linear dos parâmetros:

H_0: $L\beta = 0$ versus H_a: $L\beta \neq 0$ (L = matriz de hipótese).

Uma necessidade fundamental é ter uma combinação linear dos parâmetros Lβ que seja estimável.

Cálculo da soma de quadrados (SQ) para H_0: $L\beta = 0$:

É estimada por $Lb'(L(X'X)^{-}L')^{-1}Lb$.

Cálculo do erro-padrão (EP):

$$EP = \sqrt{\sigma^2 L(X'X)^{-}L'}$$

Somas de quadrados na análise de variância univariada

Existem quatro tipos de somas de quadrados **(SQ)**: tipo I, tipo II, tipo III e tipo IV.

Nos cálculos, são utilizados a matriz de hipótese L e os vetores L_1, L_2, etc. nas funções estimáveis para construir os coeficientes dos efeitos nos modelos nos diversos testes de uma hipótese linear dos parâmetros:

$$H_0: L\beta = 0 \text{ versus } H_a: L\beta \neq 0$$

Somas de quadrados do tipo I

Os modelos são dependentes da ordem dos efeitos; cada efeito é ajustado somente para os efeitos precedentes no modelo. A SQ do tipo I é também chamada de SQ sequencial, pois, para cada efeito adicionado ao modelo, pode-se ver o incremento na SQ do erro. É usada principalmente em modelos polinomiais em que queremos saber a contribuição de um termo no modelo quando ele é ortogonal ao efeito precedente. Para um modelo do tipo $E(Y) = X_1B_1 + X_2B_2 + X_3B_3$, as SQs do tipo I são apresentadas na Tabela 2, em que:

- $R(B_1) \rightarrow$ ajusta para B_1.
- $R(B_2|B_1) \rightarrow$ ajusta para B_2 dado que B_1 já foi ajustado.
- $R(B_3|B_2,B_1) \rightarrow$ ajusta para B_3 dado que B_1 e B_2 já foram ajustados.

A notação "R" significa uma diferença da SQ entre dois modelos. Por exemplo, a notação $R(B_2|B_1)$ significa uma diferença entre a SQ do modelo contendo os efeitos B_1 e B_2 e a SQ do modelo contendo apenas o efeito B_1.

Tabela 2. Somas de quadrados (SQs) do tipo I.

Efeito	SQs tipo I
B_1	$R(B_1)$
B_2	$R(B_2\|B_1)$
B_3	$R(B_3\|B_1,B_2)$

Somas de quadrados do tipo II ou somas de quadrados do tipo parcial

Considerando-se o modelo $E(Y) = X_1B_1 + X_2B_2 + X_3B_3$, a SQ do tipo II ajusta cada efeito, levando-se em conta que todos os outros já foram ajustados (Tabela 3). Ela possibilita ver a redução na SQ do erro em razão da adição de um efeito dado que os outros já foram ajustados. Por exemplo, a notação $R(B_1|B_2, B_3)$ significa o ajuste para o efeito B_1 que é realizado após os ajustes de B_2 e B_3.

Tabela 3. Somas de quadrados (SQs) do tipo II.

Efeito	SQs tipo II
B_1	$R(B_1\|B_2, B_3)$
B_2	$R(B_2\|B_1, B_3)$
B_3	$R(B_3\|B_1, B_2)$

Somas de quadrados do tipo III e do tipo IV

São conhecidas como SQs parciais e são as mais usadas em uma Anova. Para dados balanceados, estas duas SQs são iguais. A SQ do tipo III é calculada por:

$$L\beta = Lb'(L(X'X)- L')^{-1} Lb = 0$$

Seguem algumas propriedades das SQs:

a) As hipóteses para um efeito não envolvem parâmetros de outros efeitos, exceto se eles são necessários na sua estimativa.

b) As hipóteses a serem testadas são invariantes ao ordenamento dos efeitos no modelo.

c) Mesmo tendo dados perdidos, as hipóteses são as mesmas que são testadas considerando dados balanceados.

d) Em uma Anova, considerando um experimento fatorial e uma interação de dois fatores (A, B), com dois níveis cada (2 × 2), as hipóteses do tipo III para uma interação A x B são mostradas na Tabela 4.

Considerando-se a interação A × B, com dois níveis cada (2 × 2), as hipóteses do tipo III somente para o efeito A são apresentadas na Tabela 5.

Tabela 4. Hipóteses do tipo III para uma interação A × B.

Efeito	Forma geral	L1 = L2 = L4+0
μ	L1	0
A1	L2	0
A2	L1 - L2	0
B1	L4	0
B2	L1 - L4	0
AB11	L6	L6
AB12	L2 - L6	-L6
AB21	L4 - L6	-L6
AB22	L1- L2 - L4 + L6	L6

Tabela 5. Hipótese do tipo III para o efeito *A*.

Efeito	Forma geral	L1 = L4 = 0	L6 = K x L2	K = 0,5
μ	L1	0	0	0
A1	L2	L2	L2	L2
A2	L1 - L2	-L2	-L2	-L2
B1	L4	0	0	0
B2	L1 - L4	0	0	0
AB11	L6	L6	K*L2	0,5*L2
AB12	L2 -L6	L2 - L6	(1 - K)*L2	0,5*L2
AB21	L4 - L6	-L6	-K*L2	-0,5*L2
AB22	L1- L2 - L4 + L6	-L2 + L6	-(1- K)*L2	-0,5*L2

No GLM, as quatro somas de quadrados (I, II, III e IV) podem ser calculadas pela rotina SAS a seguir, na opção "*model y=a b a*b / e e1 e2 e3 e4;",* em que a opção *"e"* dá um resumo e indica a ordem dos parâmetros em contrastes.

```
proc glm;
    class a b;
    model y=a b a*b / e e1 e2 e3 e4;
  run;
```

Especificação dos efeitos no modelo matemático

Usando-se notação "|" e operador "@", seguem alguns exemplos de como colocar efeitos no procedimento: model y= " ":

"a | c(b)"	equivale a	a c(b) a*c(b)
"a(b) | c(b)"	equivale a	a(b) c(b) a*c(b)
"a(b) | b(d e)"	equivale a	a(b) b(d e)
"a | b(a) | c"	equivale a	a b(a) c a*c b*c(a)
"a | b(a) | c@2"	equivale a	a b(a) c a*c
"a | b | c"	equivale a	a b a*b c a*c b*c a*b*c

Aplicação do procedimento GLM

Uma aplicação típica do modelo linear padrão é a dos delineamentos de blocos casualizados com parcelas subdivididas (*split-plot*), um dos mais utilizados na agricultura. No *split-plot*, são utilizados dois níveis de tratamentos ou fatores experimentais, geralmente denominados de A e B, os quais são aleatoriamente atribuídos às parcelas principais e subparcelas, respectivamente. Geralmente, esse tipo de análise é apropriado quando a característica é avaliada apenas uma vez na unidade experimental.

Para análises MR com estrutura de erros mais complexas do que $R = \sigma^2 I$, o GLM proporciona resultados incorretos e limitados, pois produz estimativas inadequadas dos erros-padrão para a maioria delas. Para que o uso do teste F seja correto, proporcionando erro do tipo I exato para teste de todas as hipóteses, é necessário que os erros e_{ijk} atendam às suposições de independência, normalidade e, ainda, homogeneidade de variâncias. Para essa última suposição, a matriz de variâncias e covariâncias R precisa atender à condição de circularidade e esfericidade, ou seja, as variâncias da diferença entre quaisquer pares de medidas dentro da unidade experimental ou indivíduo são iguais. Caso isso não ocorra, há violação da esfericidade. Uma matriz com essa estrutura é a Huynh-Feldt (HF). A matriz simetria composta (CS), que é também um caso particular da matriz HF, atende a esta condição, e ela é a preferida por ter número de parâmetros menor do que a HF. Quando a matriz de variâncias e covariâncias não atende a esta condição, o GLM oferece dois critérios, em análises univariadas, para ajustar o teste F: Epsilon de Greenhouse-Geisser (GG) e Epsilon de Huynh-Feldt (HF).

Exemplificando-se as aplicações do GLM, quando os erros e_{ijk} são independentes, identicamente distribuídos, e com V(e) = R = σ^2I, a Anova apresentada na Tabela 6 é realizada pelo programa 1.

```
/* programa1 */
proc glm;
class b t c; "b = blocos, t = tratamentos, c = cortes";
 model pms = t  b(t) c t*c;
test h = t  e = b(t); "testa o efeito fixo t usando b(t) como erro)";
run;
```

Tabela 6. Análise de variância.

Fator de variação	Grau de liberdade
Blocos (b)	1
Tratamentos (t)	91
Resíduo_a = interação b x t	91
(Parcelas)	(183)
Cortes (c)	19
Interação t x c	1.729
Resíduo_b = interação b x c + interação b x t x c	1.748
Total	3.679

Fonte: Freitas et al. (2011).

Na Tabela 7, os testes F para tratamentos, cortes e interação tratamentos × cortes são válidos e indicam que os efeitos são altamente significativos.

Tabela 7. Resultados parciais do programa 1.

Efeito	GL do numerador	Prob > F
Tratamentos (t)	91	0,0028
Cortes (c)	19	< 0,0001
Interação t x c	1.728	< 0,0005

As médias obtidas por quadrados mínimos, intervalos de confiança e testes de hipóteses são calculadas no programa 2.

```
/* programa2 */
proc glm;
class b t  c ;
 model pms =t  b(t)  c t*c/ cli clm;
test h = t e = b(t);
lsmeans t; run;
```

No programa 2 acima, as opções Cli e Clm calculam o limite inferior e superior do intervalo de confiança, com 95% de probabilidade para cada observação. Na opção *lsmeans*, o programa calcula as médias por quadrados mínimos. Os resultados são apresentados na Tabela 8.

Tabela 8. Resultados parciais do programa 2.

Observação	Observado	Predito	Resíduo	Intervalo de confiança	
				Inferior	Superior
1	2.232,69	2.223,75	8,94	1.939,72	2.507,78
2	2.627,06	2.571,58	55,48	2.287,55	2.855,61
...	...	...	...	...	...

Tratamento	Média ± erro-padrão
1	2.563,4 + 62,5
2	3.182,5 + 62,5
3	2.616,7 + 62,5
...	

A opção *random*, no programa 3 apresentado, calcula a esperança dos quadrados médios, E(QM), para os efeitos do modelo. A opção *test* em "*random b(t)/test;*" produz testes de hipótese dos efeitos usando os erros determinados na esperança do quadrado médio, E(QM), ou seja, os mesmos resultados proporcionados pelo comando Test H = tratamentos E = blocos (tratamentos). Na Tabela 9, são apresentados os resultados da Anova realizada por meio do programa 3, com a letra "Q" associada aos efeitos fixos na E(QM).

```
/* programa3*/
proc glm;
class b t  c ;
 model pms = t  b(t)  c t*c;
random b(t)/test; run;
```

Tabela 9. Anova realizada por meio do programa 3.

Causa de variação	Grau de liberdade	Quadrado médio	Esperança do (quadrado médio)	F	Prob > F
Tratamentos (t)	91	1.191.471	σ^2+ 19,848$\sigma^2_{B(T)}$+ Q(t,t*c)	1,80	0,0028
Blocos(tratamentos): b(t)	92	663.529	σ^2+19,848$\sigma^2_{B(T)}$	16,61	<0,0001
Cortes: c	19	62.917.117	σ^2+ Q(c,t*corte)	1.575,10	<0,0001
Interação t x c	1.728	46.761	σ^2+ Q(t*c)	1,17	0,0005
Resíduo	1.734	39.945	σ^2		

Observando-se a coluna Esperança do quadrado médio na Tabela 9, verifica-se que o denominador apropriado para testar tratamentos (t) é Blocos(Tratamentos): b(t), enquanto o efeito de cortes (c) e da interação tratamentos × cortes (*t* x *c*) são testados com o resíduo σ^2. Os efeitos de t, além dos componentes fixos, contêm uma fração em razão dos dois fatores aleatórios (σ^2, $\sigma^2_{b(t)}$). Quanto ao efeito de cortes e da interação *t* x *c*, além dos componentes fixos, eles contêm apenas a fração devida à variância residual.

As estimativas dos componentes de variância residual σ^2 e de bloco dentro de tratamentos $\sigma^2_{b(t)}$ foram, respectivamente, 39.945 e 31.418, indicando que, da variabilidade total, cerca de 56,0% foi devida ao erro aleatório (σ^2), e 44,0% por causa do erro de blocos dentro de tratamentos ($\sigma^2_{b(t)}$) ou entre unidades experimentais. A correlação intraclasse ou proporção da variância total que ocorre entre as unidades experimentais foi estimada por $\sigma^2_{b(t)}/(\sigma^2 + \sigma^2_{b(t)})$ = 31.418/(39.945 + 31.418) = 0,44.

No programa 4 a seguir, a opção *contrast* permite estimar contrastes ortogonais entre médias e testar a hipótese de nulidade pelo teste t. No exemplo, foi estimado o contraste entre as médias dos cortes 1 e 20. Ao nível de p ≤ 0,0001, rejeitou-se a hipótese de nulidade H_0, da igualdade entre esses dois cortes (Tabela 10).

```
/* programa4*/
proc glm;
class b t c ;
 model pms = t b(t) c t*c;
contrast 'corte 1 vs corte 20' corte 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 -1; test h=t e=b(t);
run;
```

Tabela 10. Resultados parciais do programa 4.

Contraste	Grau de liberdade	Soma de quadrados	Quadrado médio	F	Prob > F
Corte 1 versus corte 20	1	169.698.838,4	169.698.838,4	4.248,29	<0,0001

Uso do GLM na análise multivariada

Para *p* variáveis dependentes, *k* parâmetros e *n* observações e para cada variável dependente *y*, que pode ser testada separadamente, tem-se modelo matricial:

$$y_{nxp} = X_{nxk}\, b_{kxp} + e_{nxp}$$

O vetor de erros tem: vec(e) ~ N (0, $I_n \otimes \sum_{pxp}$), em que ⊗ = produto de *Kronecker*.

A matriz $\sum_{pxp}$ é estimada por:

$$[(e'e)/(n - r)] = [((y - Xb)'(y - Xb))/(n - r)]$$

em que:

$b = (X'X)^{-}X'y$.

r = posto da matrix X.

e = matriz de resíduos.

Teste de hipótese linear na análise multivariada

Para o vetor de parâmetros β, matriz de hipótese linear L e matriz de identidade *p* x *p* (M), a fórmula geral é:

H_0: LβM = 0 versus H_a: LβM ≠ 0

Quando há muitas variáveis e amostras, existem quatro testes comumente usados para analisar a hipótese de nulidade (H_0) de que todas as amostras vêm de populações com o mesmo vetor médio:

Lambda (λ) de Wilks = det(E)/det(H+E).

Traço de Pillai = traço$(H(H + E)^{-1})$.

Traço de Lawley Hotelling = traço$(E^{-1}H)$.

Raiz máxima de Roy = λ = maior valor de $(E^{-1}H)$.

det = determinante.

O numerador e o denominador, no cálculo do teste F, são dados por:

$H = M'(Lb)'(L(X'X)^{-}L')^{-1}(Lb)M$

$E = M'(y'y - b')(X'X)b)M$

Embora o traço de Pillai seja o método que apresenta maior robustez, o teste mais utilizado é o Lambda de Wilks.

Para aplicação de análises multivariadas pelo GLM, as colunas do arquivo de dados descrito na Tabela 1 foram organizadas na seguinte ordem: Blocos Tratamentos Corte1, Corte2, Corte…, Corte20. Foram comparados os tratamentos ou variedades dentro de cada corte, porém, nessa análise, não é possível comparações entre cortes e nem verificar tendências dos resultados das variedades ao longo dos cortes, não sendo, portanto, consideradas análises de medidas repetidas.

Em cada análise, o erro, para testar o efeito de tratamentos, foi Bloco (tratamentos). Na Tabela 11, são apresentados resultados parciais (corte 1 e corte 2). De acordo com o teste F, os tratamentos não diferiram entre si no corte 1 ($p > 0,05$); porém, no corte 2, houve diferença significativa no nível $p = 0,0050$. Nas análises multivariadas pelo GLM, como as executadas pelo programa 5, são requeridas as suposições de normalidade multivariada, independência e homogeneidade de matrizes de covariâncias.

```
/* programa5*/
proc glm;
class b t  c ;
 model c1-c20 = t/ss3; run;
```

Tabela 11. Resultados parciais do programa 5.

Contraste	Grau de liberdade	Soma de quadrados	Quadrado médio	F	Prob > F
Corte – 1	91	16.984.699,71	186.645,05	1,01	0,4889
Corte – 2	91	20.489.238,10	225.156,46	1,72	0,0050
...	...	...	...	...	...

No programa 6, com a opção Manova, realiza-se a análise multivariada para testar a hipótese global:

H_0 = os tratamentos não diferem entre si

versus

H_a = os tratamentos diferem entre si.

A comparação entre os tratamentos é feita de forma ponderada para todos os cortes. Nos resultados parciais do programa 6, apresentados na Tabela 12, os quatro testes (Pillai, Hotelling, Wilks e Roy) são funções de autovalores de $E^{-1}H$ ou $(E+H)^{-1}H$,

em que E é a matriz de somas de quadrados de blocos(tratamentos) e H, a matriz de somas de quadrados do tipo III para tratamentos. Essas estatísticas multivariadas são convertidas em valores de probabilidades do teste F. Pelo teste λ de Wilks, rejeita-se a hipótese de nulidade e conclui-se que, independentemente de cortes, os tratamentos diferem entre si.

```
/* programa6 */
proc glm;
class b t c;
 model corte1-corte20 = t b(t) /nouni;
manova h= t e = b(t);run;
```

Tabela 12. Resultados parciais do programa 6.

Estatística	Valor da estatística	Valor de F	Grau de liberdade do numerador	Grau de liberdade do denominador	Prob > F
λ de Wilks	0,00000001	1,15	1.800	1.318	0,0030
Traço de Pillai	10,99583174	1,07	1.800	1.580	0,0775
Traço de Hotelling-Lawley	39,72475770	1,28	1.800	727	<0,0001
Raiz máxima de Roy	9,94051656	8,73	90	79	<0,0001

No programa 7, as variáveis corte 1 a corte 20 foram renomeadas para c_1 a c_{20}. A opção *"m"*, após *manova,* produz uma matriz de transformação M contendo contrastes ortonormais representados por diferenças sucessivas entre as variáveis dependentes. Quando as medidas repetidas representam doses de algum fator, utilizando-se a opção *"m"* é possível transformar as variáveis dependentes em componentes de polinômios ortogonais e ajustar regressões polinomiais do tipo linear, quadrática, etc. São geradas 19 novas variáveis, e a opção *prefix = t* identifica essas variáveis por t_1, t_2 ,..., t_{19}.

```
/* programa7*/
proc glm; class t b; model c1-c20 = t b(t) /nouni;
manova h = t e = b(t)
m=c1-c2, c2-c3, c3-c4, c4-c5, c5-c6, c6-c7, c7-c8, c8-c9, c9-c10, c10-c11, c11-c12,
c12-c13, c13-c14, c14-c15, c15-c16, c16-c17, c17-c18, c18-c19, c19-c20 prefix=t;
run;
```

As probabilidades do teste F para o teste de hipótese de nulidade do efeito de tratamentos, com relação às variáveis definidas pela matriz de transformação M, são apresentadas na Tabela 13. Os valores de F para os testes Pillai, Hotelling, Wilks e Roy são

levemente diferentes dos apresentados no programa 6, pois as variáveis transformadas são diferentes das originais. Porém, observa-se que a probabilidade de F para o traço de Pillai indica não significância (p > 0,05) dos tratamentos com relação às combinações lineares dos cortes, conforme geradas pela matriz M. Considerando-se os outros três testes e, principalmente, o λ de Wilks, que é o mais usado, rejeita-se a hipótese de nulidade.

Tabela 13. Resultados parciais do programa 7.

Estatística	Valor	Valor de F	Grau de liberdade do numerador	Grau de liberdade do denominador	Prob > F
λ de Wilks	0,00000002	1,16	1.710	1.266	0,0027
Traço de Pillai	10,49541661	1,08	1.710	1.501	0,0553
Traço de Hotelling-Lawley	36,24451099	1,25	1.710	718	<0,0002
Raiz máxima de Roy	7,86619310	6,90	90	79	<0,0001

Para realizar análises de MR e testar hipóteses de fatores dentro de indivíduos e as interações envolvendo as MR com os demais efeitos fixos, utiliza-se a opção Repeated (programa 8). Essas análises são robustas no que se refere às suposições de normalidade multivariada e independência, mas requerem que a condição de esfericidade seja atendida. Para obter a decomposição ortogonal, os dados de PMS dos 20 cortes são transformados por meio dos seguintes contrastes ortogonais: corte.1, corte.2, …, corte.19, em que corte.1 = corte1 - (corte2 + … + corte20)/19; corte.2 = corte2 - (corte3 + … + corte20)/18, e assim por diante. Isto é, cada nível do fator dentro de indivíduo é comparado em relação à média global dos demais níveis.

Na Tabela 14 são apresentados resultados parciais do programa 8 que mostra o teste de Mauchly, que testa se a esfericidade é violada ou não. Para o nível α = 0,05, caso tenha $p \leq \alpha$, a esfericidade não pode ser assumida; se p > α, não podemos rejeitar a condição de esfericidade. O teste de Mauchly foi $1{,}6659 \times 10^{-6}$, para componentes ortogonais, e a aceitação ou rejeição da condição de esfericidade é avaliada pelo teste de qui-quadrado (χ^2). Uma vez que $\chi^2_{189} = 964{,}3885$ ($p \leq 0{,}0001$), rejeita-se a suposição de esfericidade. Nesse caso, o teste F pode aumentar o erro tipo I e precisa ser interpretado com cuidado. Para prosseguir com a análise pelo GLM, recomendam-se dois fatores de correção para ajustar os graus de liberdade do teste *F*: Epsilon de Greenhouse-Geisser (GG) e Epsilon de Huynh-Feldt (HF). O HF é o mais indicado, uma vez que o GG dificulta a detecção da diferença entre médias de tratamentos quando elas existem.

Ambos os fatores de correção GG e HF variam de 0 a 1, e os valores menores indicam maior afastamento da esfericidade.

```
/* programa8 */
proc glm; class t; model c1-c20 = t /nouni;
repeated c helmert / summary; run;
```

Tabela 14. Resultados parciais do programa 8.

Variável	Teste de Mauchly		Qui-quadrado	Prob > χ^2
	GL	Critério		
Variáveis transformadas	189	$1{,}1238 \times 10^{-9}$	1.493,6125	<0,0001
Componentes ortogonais	189	$1{,}6659 \times 10^{-6}$	964,3885	<0,0001

Na Tabela 15, as duas últimas colunas apresentam o valor do teste F, ajustado para GG e HF, para testar o efeito de cortes e de tratamentos versus cortes. Observa-se que a correção proporcionou o mesmo valor para a probabilidade de F para o efeito de corte; no entanto, a interação tratamentos versus cortes foi não significativa (p > 0,0953) para a correção GG e significativa para HF (p ≤ 0,0232). Como HF é preferido em relação a GG, conclui-se que ambos os efeitos da Anova da Tabela 15 são significativos.

Tabela 15. Análise de variância – resultados parciais do programa 8.

Fonte	GL	SQ tipo III	Quadrado médio	F	Prob > F	Adj Prob > F	
						GG	HF
Cortes: C	19	1.067.719.734	56.195.775	1.425,38	<0,0001	<0,0001	<0,0001
C x tratamentos	1.710	76.179.409	44.549	1,13	0,0074	0,0953	0,0232
Erro	1.501	59.177.066	39.425				

GL: Grau de liberdade; SQ tipo III: Soma de quadrado tipo III; Prob > F: probabilidade de F; GG: Greenhouse-Geisser; HF: Huynh-Feldt.

Na Tabela 16 são apresentados os quatro testes de análise multivariada: Pillai, Hotelling, Wilks e Roy para o efeito de cortes e da interação tratamentos × cortes. Com exceção do teste de traço de Pillai, ambos os efeitos foram significativos (p ≤ 0,0001), concordando com os resultados da análise univariada da Tabela 15.

Tabela 16. Análise de variância – resultados parciais do programa 8.

Estatística	Valor da estatística	F	Grau de liberdade do numerador	Grau de liberdade do denominador	Prob > F
Cortes					
Lambda de Wilks	0,00541025	590,20	19	61	<0,0001
Traço de Pillai	0,99458975	590,20	19	61	<0,0001
Traço de Hotelling-Lawley	183,83430833	590,20	19	61	<0,0001
Raiz máxima de Roy	183,83430833	590,20	19	61	<0,0001
Interação tratamentos x cortes					
Lambda de Wilks	0,00000002	1,16	1.710	1.266	<0,0027
Traço de Pillai	10,49541661	1,08	1.710	1.501	<0,0553
Traço de Hotelling-Lawley	36,24451099	1,25	1.710	61	<0,0002
Raiz máxima de Roy	7,86619310	6,90	1.710	61	<0,0001

Os exemplos apresentados na Tabela 6 até Tabela 16 mostram que o procedimento GLM é bastante versátil. Entretanto, os modelos lineares mistos, por meio do procedimento Mixed, descrito a seguir, são mais poderosos e têm grande aplicabilidade em diversas áreas da pesquisa.

Aplicação do procedimento Mixed

Modelo linear misto na forma matricial

$$y_{nx1} = x_{nxp}\, b_{px1} + z_{nxq}\, u_{qx1} + e_{nx1}$$

em que:

n, p, q = número de observações, de efeitos fixos e de efeitos aleatórios, respectivamente.

x, z = matrizes de incidência ou matrizes de desenho.

y = vetor de valores observados.

b = vetor desconhecido de parâmetros de efeitos fixos.

u = vetor desconhecido de parâmetros de efeitos aleatórios entre unidades; experimentais (indivíduos).

e = vetor de erros.

E(u) = 0; Var(u) = G.

E(e) = 0; V(e) = R.

Var(y) = V(zu + e) = V = zGz'+R.

Quando $R = \sigma^2 I_n$ e z = 0, o modelo misto reduz ao modelo linear padrão.

Para quaisquer duas respostas nos tempos k e k´ (k ≠ k´), avaliadas na mesma parcela:

$$Cov(y_{ijk}, y_{ijk'}) = Cov(\delta_{ij} + e_{ijk}, \delta_{ij} + e_{ijk'}) = Var(\delta_{ij}) + Cov(e_{ijk}, e_{ijk'})$$

Para quaisquer duas respostas nos tempos e entre parcelas diferentes, a covariância é zero.

Estimativas de *b* e *u*

$$\hat{b} = (X'\hat{V}^{-1}X)^{-1} X'\hat{V}^{-1} y$$

$$\hat{V}^{-1} y$$

$$\hat{u} = (\hat{G}Z'\hat{V}^{-1}(y - X\hat{b})$$

Cálculo de médias por quadrado mínimo:

São calculadas por L´b.

Cálculo do erro-padrão (EP):

$$EP = \sqrt{L(X'V^{-1}X)^{-}L'}$$

em que:

L = matriz de hipótese.

V = matriz de variâncias e covariâncias.

Estimação de parâmetros por máxima verossimilhança e restrita

Tem-se por suposição que os efeitos aleatórios (*u*, *e*) têm distribuição normal. Uma propriedade importante desses métodos é o fato de que, se as avaliações realizadas em uma unidade experimental ou indivíduo não estiverem completas, as informações restantes serão consideradas para a análise. Associados aos métodos de

máxima verossimilhança (ML) e ML restrita (REML), são calculadas as funções objetivo ou funções de logaritmo que maximizam os parâmetros desconhecidos.

Funções de logaritmo da ML:

$l = -0,5\log|V| - 0,5r'V^{-1}r - 0,5n\log(2\pi)$

Funções de logaritmo da REML:

$l_R = -0,5\log|V| - 0,5\log|X'V^{-1}X| - 0,5r'V^{-1}r - 0,5(n - p)\log(2\pi)$

$r = y - X(X'V^{-1}X)^{-}X'V^{-1}y$

p = posto de X

Em uma rotina SAS, ML e REML são usados por:

"proc mixed method = ml;"

"proc mixed method = reml;"

Em termos computacionais no processo de estimação de parâmetros pelo procedimento Mixed, é utilizado o algoritmo de Newton-Raphson (Lindstrom; Bates, 1988). Fornece um valor inicial ao parâmetro, e, daí por diante, por um processo iterativo, obtém as estimativas dos parâmetros maximizadas. Ele é bastante eficiente e permite obter a convergência em poucas iterações.

É recomendável sempre trabalhar com amostras grandes para se obter uma matriz de variância-covariância assintótica de G e R e, portanto, testes de intervalos de confiança com base na normalidade assintótica podem ser obtidos. Quando se utilizam amostras pequenas, no entanto, as estimativas não são confiáveis, especialmente as de variâncias e covariâncias, pois a distribuição delas tende a ser viesada para a direita.

Critérios de convergência

No procedimento Mixed "*proc mixed;*" o padrão (*default*) do SAS para estimativas de parâmetros é o REML, e para convergência é o critério de convergência Hessiana "*convh*". É fornecido um valor inicial para o parâmetro de 0,00000001 (1E-8), significando que a convergência se estabiliza quando o valor da função de máxima verossimilhança entre duas iterações consecutivas for menor ou igual a 0,00000001.

Seleção de estrutura de covariâncias

Quando as distribuições dos dados tendem à normalidade, o procedimento Mixed disponibiliza cerca de 40 tipos de matriz R. A seguir, são apresentadas as estruturas de covariâncias mais utilizadas nas análises de medidas repetidas e alguns parâmetros para a escolha da mais adequada:

a) Componentes de variância (VC): usar essa estrutura quando as medidas repetidas possuem variâncias iguais e não estão correlacionadas.

b) Simetria composta (CS): usar essa estrutura quando as medidas repetidas são igualmente correlacionadas e têm variâncias e covariâncias constantes.

c) Huynh - Feldt (HF): usar essa estrutura quando as medidas repetidas podem ter variâncias diferentes, isto é, heterogeneidade ao longo da diagonal principal. A estrutura de HF é mais flexível do que a CS, e contém apenas mais alguns parâmetros. Ela é similar à simetria composta heterogênea (CSH), tem o mesmo número de parâmetros e heterogeneidade ao longo da diagonal principal.

d) AR(1) e Arma(1,1): usar essas estruturas quando as medidas repetidas (MR) podem ter variâncias iguais, mas a correlação entre as MR pode ser proporcional à distância entre elas, situação que ocorre geralmente com dados de organismos vivos. Nesse caso, quanto mais distantes são as medidas repetidas, menor é a correlação entre elas.

Como selecionar uma estrutura de variância-covariância R mais adequada

a) Comparando uma particular estrutura de covariância com $R = \sigma^2 I_n$.

Teste de razão de verossimilhança de distribuição de χ^2, que compara a estrutura de covariância R, em particular, com a estrutura clássica ($R = \sigma^2 I_n$).

$$H_0: R = \sigma^2 I_n \text{ versus } H_a: R \neq \sigma^2 I_n.$$

Se a hipótese H_0 não for rejeitada, a Anova pode ser feita pelo método de quadrados mínimos ordinários. Esse teste é padrão do SAS (*default*).

b) Comparando duas estruturas (R_1 e R_2), ambas diferentes de $\sigma^2 I_n$, pelo teste de razão de verossimilhança.

Geralmente são usados dois critérios (Bozdogan, 1987): *Akaike's Information Criterion* (AIC) e *Bayesian Information Criterion* (BIC). Para *p* parâmetros e *n* observações da amostra:

$$AIC = -2\log(L) + 2p$$

$$BIC = -2L + p\log(n)$$

L = Teste de razão de verossimilhança (LRT: *Likelihood ratio test*)

$$L = -2\log\left(\frac{\text{Verossimilhança do modelo nulo}}{\text{Verossimilhança do modelo nulo alternativo}}\right)$$

Quando várias estruturas de *R* são ajustadas, é comum escolher a de menor valor para AIC como a mais adequada. No entanto, em algumas situações matrizes com valores de AIC maiores podem apresentar menor número de parâmetros e serem de interesse. Assim, é interessante comparar as matrizes duas a duas e escolher a que apresenta significativamente o melhor ajuste. A partir do teste de razão de verossimilhança restrito, constrói um teste de χ^2 com graus de liberdade igual à diferença do número de parâmetros das duas matrizes a serem testadas.

No programa 9, foram testadas as estruturas de covariâncias: simetria composta (CS), não estruturada (UN) e Huynh-Feldt (HF). Essas estruturas estão apresentadas na Figura 1.

UN

$$\begin{bmatrix} \sigma_1^2 & \sigma_{21} & \sigma_{31} & \sigma_{41} \\ \sigma_{21} & \sigma_2^2 & \sigma_{32} & \sigma_{42} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 & \sigma_{43} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_4^2 \end{bmatrix}$$

H-F

$$\begin{bmatrix} \sigma_1^2 & \frac{\sigma_1^2+\sigma_2^2}{2}-\lambda & \frac{\sigma_1^2+\sigma_3^2}{2}-\lambda \\ \frac{\sigma_2^2+\sigma_1^2}{2}-\lambda & \sigma_2^2 & \frac{\sigma_2^2+\sigma_3^2}{2}-\lambda \\ \frac{\sigma_3^2+\sigma_1^2}{2}-\lambda & \frac{\sigma_3^2+\sigma_2^2}{2}-\lambda & \sigma_3^2 \end{bmatrix}$$

CS

$$\begin{bmatrix} \sigma^2+\sigma_1 & \sigma_1 & \sigma_1 & \sigma_1 \\ \sigma_1 & \sigma^2+\sigma_1 & \sigma_1 & \sigma_1 \\ \sigma_1 & \sigma_1 & \sigma^2+\sigma_1 & \sigma_1 \\ \sigma_1 & \sigma_1 & \sigma_1 & \sigma^2+\sigma_1 \end{bmatrix}$$

Figura 1. Estruturas de variâncias e covariâncias não estruturada (UN), Huynh-Feldt (HF) e simetria composta (CS).

Assim, o primeiro objetivo no uso do procedimento Mixed, em uma análise de variância, é modelar os erros entre as avaliações da unidade experimental, o que é feito pelo programa 9, utilizando a estrutura de dados descrita na Tabela 1, que tem como base a publicação de Freitas et al. (2011). Com os resultados parciais dessa análise (Tabela 17), foi realizada a comparação das estruturas de covariâncias duas a duas, pelo teste de χ^2.

```
/* programa9*/
proc mixed data = alfafa  method = reml;
class bloco tratamento corte ;
model pms = tratamento corte tratamento*corte;
repeated corte /sub = bloco(tratamento) type = cs;
/* type = un; type = hf */
run;
```

Notação:

"repeated corte": indica que corte é considerado medidas repetidas.

"sub = bloco (tratamento)": a unidade experimental é representada por bloco (tratamento).

"type = cs": indica que a estrutura de covariância CS está sendo testada.

Tabela 17. Resultados parciais do programa 9.

Estrutura[1]	Grau de liberdade	-2 Res Log Likelihood	AIC[2]
CS	1	14.128,1	14.132,1
HF	20	14.416,4	14.458,4
UN	209	14.110,3	14.850,3

[1] Simetria composta (CS), Huynh-Feldt (HF) e estruturas de variâncias e covariâncias não estruturada (UN); [2] *Akaike's Information Criterion*.

Comparação pareada das estruturas pelo teste de razão de verossimilhança

CS versus HF $\chi^2_{19} = |14128,1 - 14416,4| = 288,3$ ($p < 0,001$)

CS versus UN $\chi^2_{208} = |14128,1 - 14110,3| = 17,8$ ($p > 0,05$)

HF versus UN $\chi^2_{189} = |14416,4 - 14110,4| = 306,1$ ($p < 0,001$)

As estruturas CS, HF e UN, nessa ordem, foram as que apresentaram os menores valores de AIC. Fazendo-se as comparações pareadas pelo teste de razão de verossimilhança, a estrutura CS não diferiu significativamente de UN ($p > 0,05$). No

entanto, nessa análise, a CS é a escolhida por atender à condição de esfericidade (as variâncias da diferença entre pares de erros são todas iguais), bem como possuir apenas dois parâmetros e menor valor de AIC.

Uma condição suficiente para que o teste F da análise de variância usual, em análises de medidas repetidas, seja válido, isto é, para avaliar tempo e a interação tratamento × tempo, é que a matriz de covariâncias seja do tipo simétrica composta heterogênea (CSH). Uma matriz que atende a essa condição é a HF, o que significa que as variâncias da diferença entre pares de erros sejam todas iguais.

Uma vez que a matriz HF não foi escolhida para realizar análises de medidas repetidas pelo Mixed, outras estruturas de variâncias são avaliadas. Como dados biológicos avaliados ao longo do tempo geralmente apresentam efeito de sazonalidade na produção, duas estruturas de variâncias são de interesse:

a) Ar(1): autorregressiva de primeira ordem. Contém σ^2 e um parâmetro autorregressivo (ρ), sendo a tendência ao longo do tempo dada por $_{\sigma}{}^{2}{}_{d}+\sigma^2{}_{p}\rho^{lag}$.

b) Arma(1,1): autorregressiva de primeira ordem de média móvel. Além de σ^2, ρ, contém γ, parâmetro que modela um componente de média móvel, sendo a tendência ao longo do tempo dada por $_{\sigma}{}^{2}{}_{d}+\sigma^2{}_{p}\gamma\rho^{lag}$.

Na Arma(1,1), autorregressiva de primeira ordem de média móvel, significa que, em uma previsão futura, considera a média das observações passadas como recentes; o termo média móvel indica que, quando a medida da próxima observação de uma série se torna disponível, a média das observações é recalculada com a inclusão desse novo valor e elimina-se a mais antiga.

As duas estruturas AR(1), Arma(1,1), que estão apresentadas na Figura 2 e também CS são calculadas no programa 10, cujos resultados parciais são apresentados na Tabela 18. Observa-se que a Arma(1,1), AR(1) e CS, nessa ordem, apresentaram os menores valores de AIC. Como AR(1) e CS têm dois parâmetros, não é necessário utilizar o teste de χ^2 para compará-las. Nesse caso, a mais adequada é a AR(1), pois tem o menor valor de AIC. Para comparar AR(1) e Arma(1,1), tem-se $\chi^2 = |13.957,5 - 14.123,8| = 166,3$ com grau de liberdade = 1 (2 - 1). Consultando-se $\chi^2_1 = 166,3$, verificou-se que o valor de *p* é menor que 0,001, indicando que a matriz Arma(1,1) foi a que melhor modelou a correlação entre as medidas repetidas.

Foram acrescentadas ainda, no Programa 10, as opções Asycorr, Asycov após Proc Mixed, que produzem, respectivamente, a matriz de correlação assintótica e a matriz de covariância assintótica das estimativas dos três parâmetros da Arma(1,1): σ^2, ρ e γ, cuja finalidade é mostrar o parentesco existente entre eles (Tabela 19).

$$\sigma^2\begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}$$

Autorregressiva de primeira ordem – AR(1)

$$\sigma^2\begin{bmatrix} 1 & \gamma & \gamma\rho & \gamma\rho^2 \\ \gamma & 1 & \gamma & \gamma\rho \\ \gamma\rho & \gamma & 1 & \gamma \\ \gamma\rho^2 & \gamma\rho & \gamma & 1 \end{bmatrix}$$

Autorregressiva de primeira ordem de média móvel – ARMA(1,1)

Figura 2. Estruturas de variâncias e covariâncias de AR(1): autorregressiva de primeira ordem e ARMA(1,1): autorregressiva de primeira ordem de média móvel.

```
/* programa10*/
proc mixed asycorr asycov;
class bloco tratamento  corte ;
model pms = tratamento corte tratamento*corte;
repeated corte / sub = bloco(tratamento) type = arma(1,1);run;
```

Os resultados parciais do programa 10 são apresentados nas Tabelas 18 e 19.

Tabela 18. Resultados parciais do programa 10.

Estrutura de covariância	Grau de liberdade	-2 Res Log Likelihood	Critério de AIC
Ar(1)	1	14.123,8	14.127,8
Arma(1,1)	2	13.957,5	13.963,5
Cs	1	14.128,1	14.132,1

Tabela 19. Resultados parciais do programa 10.

Variável	Matriz de covariância assintótica		
	ρ	γ	σ^2
ρ	0,000091	0,000146	35,9798
γ	0,000146	0,000789	150,3700
σ^2	35,9798	150,3700	36061142
Matriz de correlação assintótica			
	ρ	γ	σ^2
ρ	1,0000	0,5445	0,6272
γ	0,5445	1,0000	0,8912
σ^2	0,6272	0,8912	1,0000

σ^2, ρ e γ: são três parâmetros da Arma(1,1).

Procedimento GLM versus Mixed

Na comparação entre os procedimentos GLM e Mixed, para facilitar a apresentação gráfica, foram analisados dados de PMS de apenas cinco cultivares de alfafa (Bárbara, Crioula, P30, P5715 e LEN 4), que foram as mais produtivas entre as 92 estudadas.

A divergência fundamental entre os procedimentos GLM e Mixed está no erro associado a cada indivíduo na análise de MR. Essa diferença reflete no cálculo dos erros-padrão (EP) das médias e demais cálculos derivados desses, tais como: testes de hipóteses e intervalos de confiança.

No GLM e Mixed, o EP é obtido, respectivamente, da raiz quadrada de $\sigma^2L(X'X)^{-}L'$ e $L(X'V^{-1}X)^{-}L'$, em que X é a matriz de especificação, L, a matriz de hipótese, σ^2, o quadrado médio residual, V, a matriz de variâncias e covariâncias, ou seja, o EP no GLM é função do σ^2, enquanto no Mixed é função de V.

Nos programas 11 (GLM) e 12 (Mixed), o comando:

"lsmeans t c /stderr pdiff = all adjust = tukey cl;run;".

Calculam-se, pela ordem, as médias obtidas por quadrados mínimos para *t* (variedades) e cortes (c); erros-padrão das médias "*stderr*", teste de Tukey nas múltiplas comparações com respectivas probabilidades "*pdiff = all adjust = tukey*" e com respectivos intervalos de confiança "*cl*". Os resultados apresentados, nas Tabelas 20 e 21 e na Figura 3, ilustram as diferenças entre os dois procedimentos (GLM e Mixed).

```
/* programa11*/
proc glm;
class b t c ;
model pms = t b(t) c t*c;
lsmeans t c /stderr pdiff = all adjust = tukey cl;run;

/* programa12*/
proc mixed data = alfafa method = reml;
class b t c ; model pms = t c t*c;
repeated c /sub = b(t) type = arma(1,1) ;
lsmeans t c /stderr pdiff = all adjust = tukey cl;run;
```

Os valores de Pr > F para efeitos de tratamentos, cortes e interação tratamentos versus cortes indicam significância apenas para o efeito de cortes (Tabela 20). Como os dados eram balanceados, os dois métodos são concordantes quanto às estimativas de efeitos fixos. No entanto, observou-se que os erros-padrão, obtidos por máxima verossimilhança restrita (Mixed), foram maiores do que os obtidos por quadrados

mínimos (GLM), diferença essa que pode ser observada nas médias e erros-padrão obtidos por quadrados mínimos (Tabela 21) e, também, nos gráficos com intervalos de confiança com 95% de probabilidade, pois os cálculos destes são feitos com base nos erros-padrão (Figura 3).

Tabela 20. Resultados parciais dos programas 11 e 12 "Prob > F", para os efeitos principais".

Efeito	Prob > F	
	GLM	Mixed
Tratamento	0,8524	0,8932
Corte	<0,0001	<0,0001
Tratamento*Corte	0,9662	0,9563

Tabela 21. Resultados parciais dos programas 11 e 12 (médias ± erro-padrão).

Tratamento	GLM	Mixed
	Variedade de alfafa	
Len 4	1.653,1 ± 31,2 a	1.653,1 ± 155,3 a
Bárbara	1.519,5 ± 31,2 bc	1.519,5 ± 155,3 a
Crioula	1.532,8 ± 32,0 bc	1.534,3 ± 155,5 a
P 30	1.565,0 ± 32,0 ab	1.565,5 ± 155,5 a
P 5715	1.434,2 ± 31,2 c	1.434,2 ± 155,3 a
	Cortes de alfafa	
1	2.563,4 ± 62,5 b	2.563,5 ± 92,1 b
2	3.182,5 ± 62,5 a	3.182,5 ± 92,1 a
3	2.616,7 ± 62,5 b	2.616,7 ± 92,1 b
4	1.748,9 ± 62,5 c	1.748,9 ± 92,1 c
5	1.584,3 ± 68,8 cde	1.586,5 ± 95,7 cdef
6	1.342,7 ± 62,5 df	1.342,7 ± 92,1 def
7	1.296,1 ± 62,5 bef	1.296,1 ± 92,1 efg
8	1.269,2 ± 68,7 ef	1.273,5 ± 95,7 fg
9	928,6 ± 62,5 gh	928,5 ± 92,1 hi
10	1.644,5 ± 62,5 cd	1.644,5 ± 92,1 cd
11	1.557,9 ± 62,5 cde	1.557,9 ± 92,1 cdef
12	1.406,7 ± 62,5 def	1.406,7 ± 92,1 defg
13	1.578,8 ± 62,5 cde	1.578,8 ± 92,1 cdef
14	1.565,9 ± 62,5 cde	1.565,9 ± 92,1 cdef
15	1.613,2 ± 62,5 cd	1.613,2 ± 92,1 cde
16	1.395,5 ± 62,5 def	1.395,5 ± 92,1 defg
17	866,4 ± 68,7 gh	894,5 ± 92,1 hi
18	819,1 ± 62,5 h	819,1 ± 92,1 i
19	652,2 ± 62,5 h	652,2 ± 92,1 i
20	1.159,5 ± 62,5 fg	1.159,5 ± 92,1 gh

Médias seguidas de letras iguais na coluna não diferem entre si, pelo teste de Tukey, a 5% de probabilidade.

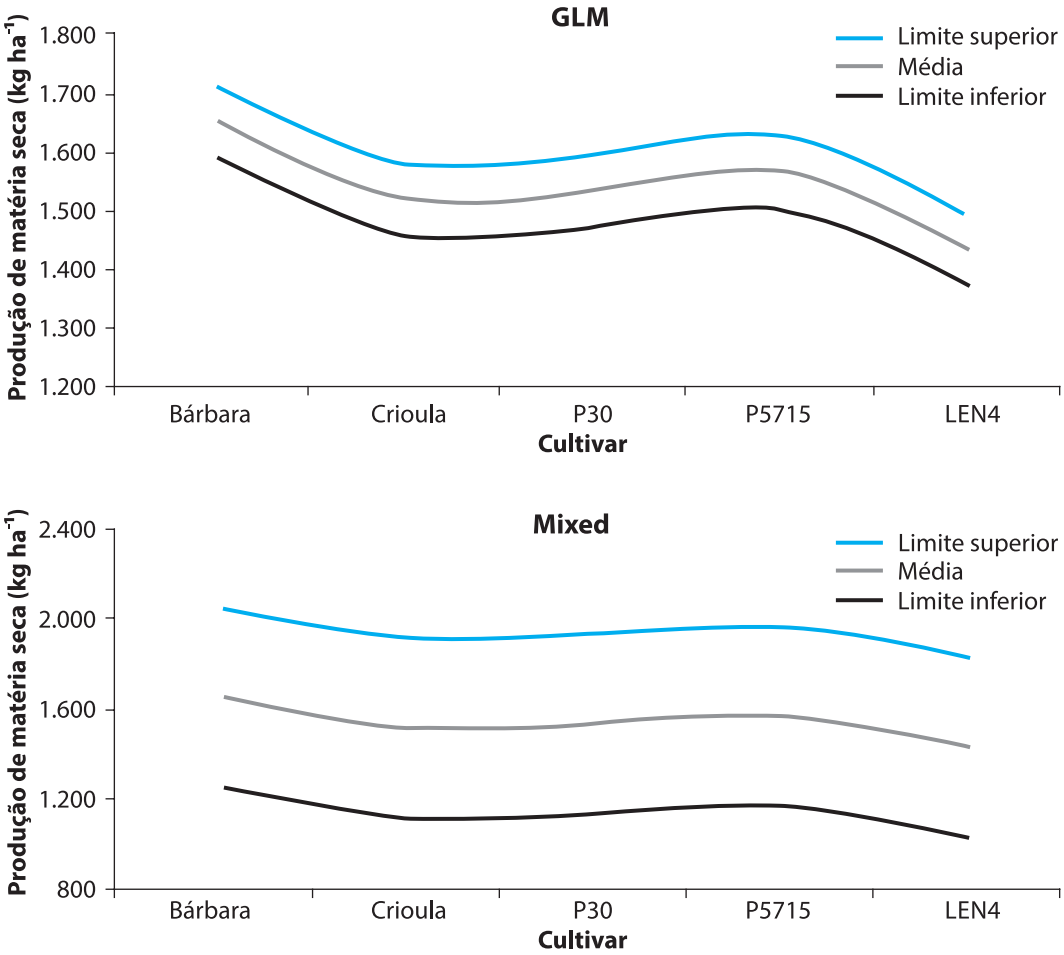


Figura 3. Produtividade de matéria seca, em quilograma por hectare (kg ha^{-1}) de cinco cultivares de alfafa e intervalos de confiança com 95% de probabilidade.

Análise univariada versus análise multivariada

Com exceção de alguns exemplos do GLM, os assuntos discutidos, neste capítulo e nos seguintes, referem-se à análise de variância univariada. No entanto, apresentaremos breve descrição da importância da análise de variância multivariada.

Ao realizar um experimento ou levantamento, em geral, diversas variáveis são avaliadas simultaneamente na mesma unidade experimental ou amostral, e é comum elas apresentarem uma estrutura natural de dependência linear ou de correlação. A análise de cada uma dessas variáveis, isoladamente, poderá não ser adequada para interpretar o fenômeno como ele realmente se apresenta, pois perdem-se valiosas informações ao desconsiderar as correlações entre as variáveis envolvidas.

O que ocorre em diversas situações é a existência de variáveis que atuam camufladamente e precisam também serem consideradas. Sabe-se da teoria da correlação que duas variáveis podem não ter nenhuma relação aparente quando consideradas isoladamente. Por exemplo, elas podem ter uma correlação r = 0,2 inicialmente, e passar para 0,7 ou 0,8 na presença de uma terceira variável no cálculo de correlação condicionada.

Um procedimento interessante na análise de dados é começar pela abordagem num enfoque multivariado, para então chegar às análises univariadas. Não como premissa, mas, sim, pelas consequências dos resultados encontrados após uma exploração mais elaborada dos dados. Um outro ponto fundamental da análise multivariada é que, em algumas situações, pode-se aceitar uma hipótese em uma análise univariada de um experimento, e ela pode ser rejeitada pela análise multivariada, já que esta pondera as informações de forma ótima.

As técnicas de análise multivariada na experimentação são utilizadas basicamente para:

a) Reduzir dimensões de conjunto de variáveis − Análise de componentes principais.

b) Formar grupos de indivíduos − Análise de agrupamentos.

c) Classificar indivíduos em grupos conhecidos − Análise discriminante.

d) Relacionar grupos de variáveis − Correlações canônicas.

e) Comparar grupos de indivíduos − Análise de variância multivariada e análise de regressão.

As análises de componentes principais, por exemplo, são técnicas de análises multivariadas que possibilitam interpretar os dados em uma dimensão reduzida, o que facilita bastante, principalmente quando existem muitas variáveis em estudo. Por meio dessa técnica, são geradas novas variáveis denominadas de componentes principais (CP), independentes entre si, e que são combinações lineares daquelas. Geralmente, a variabilidade ou as informações contidas em um conjunto grande de variáveis, correlacionadas entre si, são resumidas em dois ou três CP, os quais preservam a maior parte das informações contidas nos dados originais e podem ser utilizados como novas variáveis ou como índices.

Na análise de variância multivariada, são requeridas as suposições de normalidade multivariada, independência e homogeneidade de matrizes de covariâncias entre tratamentos. Testa-se a hipótese global:

H_0 = os tratamentos não diferem entre si

versus

H_a = os tratamentos diferem entre si.

A comparação entre os tratamentos é feita com a ponderação de todos os indivíduos pertencentes ao experimento.

Essas exigências para testar uma hipótese mostram que nem sempre um experimento planejado para ser analisado por métodos univariados poderá ser analisado por técnicas multivariadas, pois essas também têm suas limitações e certas pressuposições que precisam ser atendidas. Nesse momento, a estatística univariada é importante, principalmente na análise exploratória de dados para conhecer as variáveis em estudo.

Uma razão principal que limita o uso indiscriminado da análise multivariada é que esta requer, no mínimo, conhecimento mais avançado sobre a teoria de matrizes, além de exigir recursos mais sofisticados dos softwares, mesmo usando-se a ferramenta SAS.

Como exemplo da decomposição matricial na análise multivariada, o modelo geral é dado por:

$$Y_{nxp} = X_{nxk}\beta_{kxp} + \varepsilon_{nxp}$$

em que:

Y = matriz das observações ou de variáveis de respostas, de dimensão n x p.

n = número de unidades experimentais.

p = número de variáveis envolvidas na análise.

De modo que Y é dada por:

$$Y = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1p} \\ y_{21} & y_{22} & \cdots & y_{2p} \\ \vdots & \vdots & & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{np} \end{bmatrix}$$

- Se p = 1 → Análise univariada.
- Se p = 2 → Análise bivariada e, assim, sucessivamente.

X = matriz de incidência ou de delineamento, de dimensão n x k.

β = número de parâmetros, que é uma função do delineamento e do número de tratamentos experimentais, de dimensão k x p.

ε = matriz de resíduos, de dimensão n x p.

Exercícios[8]

1) No texto a seguir, existem afirmações incorretas quanto a conceitos de estatística. Reescreva o texto colocando definições corretas e sublinhe ou coloque em negrito onde houve definições incorretas.

A divergência fundamental entre os procedimentos General Linear Model (GLM) e Mixed Model (Mixed) está no erro do modelo matemático. Em razão disso, a aplicação deles são bastante divergentes. O GLM é apropriado para ajustar modelos lineares gerais pelo método dos quadrados mínimos e, também, por máxima verossimilhança. Permite executar vários tipos de análises: regressões, análises de variâncias univariadas e multivariadas, etc. Entretanto, não é possível realizar pelo GLM análises de medidas repetidas (MR). O modelo linear padrão matricial é do tipo $y_{nx1} = X_{nxp}\ b_{px1} + \varepsilon_{nx1,}$ em que y é o vetor de valores dependentes, sendo que E(y) = Xb; Var(y) = V(ε) = σ^2I em que σ^2 é o quadrado médio do erro; X é a matriz de especificação; b contém os efeitos fixos; ε é o vetor que contém os erros aleatórios. O GLM, exige apenas que erros e_{ijk} sejam independentes, identicamente distribuídos com média zero; porém, não há necessidade de a distribuição dos erros se ajustar a uma normal, nem mesmo ter uma distribuição aproximada da normal. O procedimento Mixed, todavia, possibilita o ajuste de grande variedade de modelos lineares e não lineares a dados que exibem correlação e variabilidade não constante; possibilita modelar não somente as médias, mas também variâncias e covariâncias. As suposições básicas que são requeridas dos dados é que eles sejam normalmente distribuídos e que as variâncias e covariâncias dos dados devem exibir estrutura dentro daquelas disponíveis no Proc Mixed. O Proc Mixed ajusta essas estruturas por meio do método de quadrados mínimos e máxima verossimilhança (ML).

2) Se a matriz de variância e covariância da Tabela 22 é do tipo Huynh-Feldt (HF), determine o valor de λ, sabendo que a covariância situada na i-ésima linha e j-ésima coluna é dada por $\frac{\sigma_i^2 + \sigma_j^2}{2} - \lambda$ e na diagonal principal tem $\sigma_1^2, ..., \sigma_7^2$.

[8] As respostas dos exercícios podem ser consultadas no Apêndice 1.

Tabela 22. Matriz de variância e covariância.

	C_1	C_2	C_3	C_4	C_5	C_6	C_7
C_1	6,726	2,032	3,096	0,149	3,780	4,854	2,865
C_2		7,878	3,673	0,725	4,356	5,430	3,441
C_3			10,007	1,790	5,421	6,495	4,506
C_4				4,113	2,474	3,547	1,558
C_5					11,375	7,179	5,190
C_6						13,523	6,263
C_7							9,545

3) A Tabela 23 contém as médias e erros-padrão de cinco cortes de alfafa, obtidos por meio dos procedimentos GLM e Mixed do SAS. Pelo Mixed, duas estruturas de variâncias e covariâncias foram ajustadas: Ar(1) e Arma(1,1).

Tabela 23. Médias e erros-padrão obtidos por meio do GLM e Mixed do SAS.

Corte	GLM	Mixed	
		Ar(1)	Arma(1,1)
1	2.436,1 ± 32,3	2.436,1 ± 40,7	2.436,1 ± 41,1
2	2.945,0 ± 32,3	2.945,0 ± 40,7	2.945,0 ± 41,1
3	2.321,7 ± 32,3	2.321,7 ± 40,7	2.321,7 ± 41,1
4	1.471,1 ± 32,3	1.471,1 ± 40,7	1.471,1 ± 41,1
5	1.436,2 ± 32,1	1.436,2 ± 41,3	1.436,2 ± 41,1

Ar(1): autorregressiva de primeira ordem; Arma(1,1): autorregressiva com média móvel.

a) Com base nos resultados, comente se os dados eram ou não balanceados.

b) Comente a razão da diferença do erro-padrão da média.

4) A Tabela 24 contém o ajuste de cinco estruturas de covariâncias e estatísticas. Selecione, pelos dois critérios ($-2L_R$, AIC), a estrutura mais adequada.

Tabela 24. Resultados do ajuste de cinco estruturas de covariâncias, com respectivos número de parâmetros e estatísticas associadas: razão de máxima verossimilhança restrita ($-2L_R$) e *Akaike's Information Criterion* (AIC).

Estatística	Cs (2)	Csh (136)	HF (21)	Ar(1) (2)	Arma(1,1) (3)
$-2L_R$	5.676,0	5.556,9	5.690,7	5.681,1	5.625,3
AIC	5.680,0	5.598,9	5.732,7	5.685,1	5.631,3

Cs: simetria composta; Csh: simetria composta heterogênea; HF: *Huynh –Feldt*; Ar(1): autorregressiva de primeira ordem; Arma(1,1): autorregressiva de primeira ordem com média móvel.
Obs.: o número de parâmetros está dentro do parêntese.

5) A Tabela 25 contém resultados de estruturas de variâncias e covariâncias e comparações pareadas pelo teste de razão de verossimilhança.

Tabela 25. Estruturas de variâncias e covariâncias com respectivos parâmetros, valores da razão de máxima verossimilhança restrita ($-2L_R$) e *Akaike's Information Criterion* (AIC) e análise de variância do tipo III (Pr > F).

Estrutura[(1)]	Nº de parâmetros	$-2L_R$	AIC	Prob > F		
				Tratamento	Controle	Tratamento x controle
Un	28	725,9	825,9	0,1779	<,0001	0,7494
HF	8	751,6	811,6	0,2507	<,0001	0,9151
Arma(1,1)	3	767,0	817,0	0,1758	<,0001	0,9131
Cs	2	769,5	817,5	0,1808	<,0001	0,9115
Ar(1)	2	773,7	821,7	0,0921	0,0002	0,9093
Vc	1	801,9	847,9	0,0116	<,0001	0,9820

[(1)] Un: não estruturada; HF: Huynh-Feldt; Arma(1,1): autorregressiva de primeira ordem com média móvel; Cs: simetria composta; Ar(1): autorregressiva de primeira ordem e Vc: componente de variância.

A seguir estão as comparações da Un com as demais estruturas, pelo critério da máxima verossimilhança restrita ($-2L_R$). O nível de significância foi verificado consultando-se os valores de qui-quadrado para $\alpha = 0,05$ (Tabela 2.1, Anexo 1): $\chi^2_{20} = 31,41$; $\chi^2_{25} = 37,65$; $\chi^2_{26} = 35,56$; $\chi^2_{26} = 35,56$; $\chi^2_{27} = 40,11$.

Corrigir as comparações que não estão corretas.

Un versus HF $\rightarrow |725,9 - 751,6| \rightarrow \chi^2_{10} = 25,7$ ns.
Un versus Arma (1,1) $\rightarrow |725,9 - 767,0| \rightarrow \chi^2_{20} = 41,1$ (P <0,05).
Un versus CS $\rightarrow |725,9 - 801,9| \rightarrow \chi^2_{26} = 76,0$ (P <0,05).
UN versus Ar1 $\rightarrow |725,9 - 773,7| \rightarrow \chi^2_{26} = 40,8$ (P <0,05).
Un versus Vc $\rightarrow |725,9 - 801,9| \rightarrow \chi^2_{27} = 76,0$ (P <0,05).

6) Em um experimento em blocos completo casualizados os tratamentos principais (A) foram distribuídos nas parcelas principais. Como tratamentos secundários (B) das subparcelas foram consideradas as avaliações mensais realizadas na parcela principal ou indivíduo ao longo do tempo. Pergunta-se: porque este experimento não pode ser considerado um delineamento em blocos completo com parcela dividida (*split-plot*)?

Capítulo 9

Delineamento inteiramente casualizado

Introdução

Quando se consegue um ambiente homogêneo onde se podem instalar todos os tratamentos, o delineamento inteiramente casualizado (DIC) é um dos mais utilizados e eficientes. O DIC apresenta ainda as seguintes vantagens: possui maior flexibilidade, o número de repetições pode variar entre tratamentos, a análise estatística é simples, o número de graus de liberdade associado ao erro geralmente é grande, a parcela perdida não é problema, o modelo requer poucas suposições e o custo na condução do experimento é um dos menores. No entanto, tem como desvantagem sua ineficiência quando não se consegue um ambiente homogêneo, pois as parcelas são heterogêneas e não possibilitam a obtenção de resultados abrangentes.

Neste capítulo, são apresentados os vários conceitos, tipos, croquis de campo, etc. do DIC, que é o mais simples, mais econômico e um dos mais utilizados na experimentação. O leitor já teve a oportunidade de saber como utilizar os fundamentos teóricos dos Capítulos 7 e 8, principalmente o modelo linear generalizado (GLM) e o Sistema de Análise Estatística (SAS).

Inteiramente casualizado tradicional

Seja um experimento em que se deseja testar quatro variedades (A, B, C, D) de uma forrageira, com cinco repetições (1 a 5) cada, totalizando 20 parcelas. Para a instalação do experimento, de modo a garantir a aleatoriedade dessas parcelas no campo, uma das maneiras é sortear o tratamento juntamente com a repetição. Por exemplo, saiu o tratamento B e a repetição 1, então, tem-se a parcela B1; no segundo sorteio saiu o tratamento D e a repetição 1, tendo-se a parcela D1, e assim por diante. Desde que o ambiente seja homogêneo, não importa o formato do experimento no campo. Na Figura 1 apresentam-se dois croquis.

Croqui 1

B1	D1	D4	D5	A5
C2	C1	B5	B3	D3
A3	A1	C3	C5	C4
D2	B2	A2	A4	B4

Croqui 2

B1	D1	D4	A4	A5	B2	C2	A1	D2	B5
C1	A2	C4	B3	D3	C3	D5	B4	C5	A3

Figura 1. Dois croquis com delineamento inteiramente casualizado (DIC).

Modelo matemático

Em que:

$y_{ij} = \mu + a_i + \varepsilon_{ij}$ (i = 1, 2, ..., 4; j = 1, ..., 5).

y_{ij} = valor observado na parcela.

μ = efeito médio global.

a_i = efeito do tratamento i.

ε_{ij} = erro aleatório associado ao tratamento i e repetição j.

Quanto aos erros ε_{ij} (i = 1, 2, ..., 4; j = 1, ..., 5), supõe-se que eles têm média zero, se ajustam a uma distribuição normal, são independentes e identicamente distribuídos com variância σ^2, daí a notação $\varepsilon_{ij} \sim NIID\ (0, \sigma^2)$.

Na Tabela 1, é apresentado o quadro de análise de variância.

Tabela 1. Análise de variância.

Fator de variação	Grau de liberdade	Soma de quadrados	Quadrado médio	F
Tratamentos – t	(i -1)	SQT	QMT = SQT/(i - 1)	QMT/QME
Erro	(j - 1)(i - 1)	SQE	QME = SQE/(j - 1)(i - 1)	
Total	ij -1	SQ total	SQ total	

Aplicação

Os dados da Tabela 2 são fictícios e referem-se a um experimento com cinco tratamentos (trat) e sete repetições (rep), cujo objetivo foi avaliar a produtividade de feijão, em grama (g) por parcela (*y*).

Tabela 2. Produtividade de feijão, em grama (g) por parcela (*y*), de um experimento inteiramente casualizado (dados fictícios).

Repetição	Tratamento					Total
	1 (Controle)	2	3	4	5	
1	85	70	85	55	85	295
2	45	67	42	48	41	243
3	80	47	49	71	49	296
4	62	84	64	80	85	375
5	60	77	67	58	42	304
6	57	90	51	77	52	327
7	25	65	22	59	39	210
Total	414	500	380	448	393	2.050

Código SAS

Arquivo de dados e análise de variância (Anova) pelo procedimento GLM do SAS. Para o teste de comparações múltiplas das médias pareadas, foi usado o teste de Tukey, e para comparar as médias de todos os tratamentos com o controle, foi usado o teste de Dunnett.

```
data a;
input rep trat  y @@;
cards;
1 1 85   2 1 45   3 1 80   4 1 62   5 1 60   6 1 57   7 1 25
1 2 70   2 2 67   3 2 47   4 2 84   5 2 77   6 2 90   7 2 65
1 3 85   2 3 42   3 3 49   4 3 64   5 3 67   6 3 51   7 3 22
1 4 55   2 4 48   3 4 71   4 4 80   5 4 58   6 4 77   7 4 59
1 5 85   2 5 41   3 5 49   4 5 85   5 5 42   6 5 52   7 5 39
;
proc glm data = a; class trat;
model y = trat/ss3;
means trat/ tukey dunnett ('1'); run;
```

A análise de variância da Tabela 3 mostrou que o efeito de tratamentos não foi significativo (P > 0,05). Embora as médias tenham sido bastante diferentes, uma explicação para a não significância é a grande variabilidade entre os dados, pois o coeficiente de variação foi alto (29,07%) e o coeficiente de determinação muito baixo (12,34%).

Tabela 3. Análise de variância.

Fatores de variação	Grau de liberdade	Soma de quadrados	Quadrado médio	F	Prob > F
Tratamentos (trat)	4	1.329,14	332,28	1,06	0,3951
Erro	30	9.436,86	314,56		
Total	34	10.766,00			

Sqtotal = $85^2 + 70^2 + ... + 39^2 - (2050)^2/35 = 10766,00$

Sqtrat = $(414^2 + 500^2 + 380^2 + 448^2 + 393^2)/7 - (2050)^2/35 = 1329,14$

Sqres = SQ Total − SQtrat = 9436,86

Qmtrat = SQtrat/GL trat = 332,28

Qmres = SQErro/GL RES= $s^2 = 314,56$

F = Qmtrat/Qmres = 1,06

CV = (s/61,00) x 100 = 29,07

$R^2 = 1329,14/10766,00 = 0,12$

Médias de tratamentos em ordem decrescente:

2	71,43 a
4	64,00 a
1	59,14 a
5	56,14 a
3	54,27 a

Média geral: 61,00.

“a” = não significância (p > 0,05)

Diferença mínima significativa (DMS) pelo teste de Tukey e teste de Dunnett:

$$DMS_{(Tukey)} = 4,1021\sqrt{\frac{QMR}{7}} = 27,4980$$

$$DMS_{(Dunnett)} = 2,5782\sqrt{\frac{2QMR}{7}} = 24,4410$$

Valor tabelado do teste de Tukey com graus de liberdade 5 e 30 e α = 0,05:

$q_{5;30(0,05)} = 4,1021$

Valor tabelado do teste de Dunnett com graus de liberdade 4 e 30 e α = 0,05:

$D_{4;30(0,05)} = 2,5782$

Todos os intervalos de confiança (IC) com 95% de probabilidade da diferença entre duas médias incluem o zero, indicando não significância.

Teste de Dunnett

Compara os tratamentos 2, 3, 4 e 5 com a testemunha (1).

α = 0,05

DMS = 24,441

Valor crítico = 2,5782

Comparações	Diferenças	IC 95%	
2 - 1	12,27	[-12,16	36,73]
4 - 1	4,86	[-19,58	29,30]
5 - 1	-3,00	[-27,44	21,44]
3 - 1	-4,86	[-29,30	19,58]

Teste de Tukey

Faz comparações pareadas entre todos os tratamentos:

α = 0,05

DMS = 27,4980

Valor crítico = 4,1021

Comparações	diferenças	IC 95%
2 - 4	7,429	[-20,07 34,93]
2 - 1	12,286	[-15,21 39,78]
2 - 5	15,286	[-12,21 42,78]
2 - 3	17,143	[-10,35 44,64]
4 - 2	-7,429	[-34,93 20,07]
4 - 1	4,857	[-22,64 32,35]
4 - 5	7,857	[-19,64 35,35]
4 - 3	9,714	[-17,78 37,21]
1 - 2	-12,286	[-39,78 15,21]
1 - 4	-4,857	[-32,35 22,64]
1 - 5	3,000	[-24,50 30,50]
1 - 3	4,857	[-22,64 32,36]
5 - 2	-15,286	[-42,78 12,21]
5 - 4	-7,857	[-35,36 19,64]
5 - 1	-3,000	[-30,50 24,50]
5 - 3	1,857	[-25,61 29,35]
3 - 2	-17,143	[-44,64 10,35]
3 - 4	-9,714	[-37,21 17,78]
3 - 1	-4,857	[-32,35 22,64]
3 - 5	-1,857	[-29,36 25,64]

Observação: todos os intervalos de confiança incluem zero, indicando não significância entre os tratamentos.

Delineamento inteiramente casualizado com os tratamentos organizados em esquemas fatoriais

Nos experimentos com animais, é comum a necessidade de tratamentos organizados em fatores, tais como rações, grupo genético ou raças, sexo, e ainda envolvendo variação espacial (local, fazenda, etc.) e temporal (ano de nascimento, estação, mês, ano, etc.). Em experimentos com plantas forrageiras, geralmente o interesse é avaliar variedades, efeito de doses de adubo, em quilograma por hectare

(kg ha^{-1}), fontes desses adubos e, ainda, incluir efeitos temporais (mês, estação, ano, cortes, etc.).

Um dos grandes interesses por esses delineamentos é a obtenção de resultados mais conclusivos, como verificar a eficiência de um tratamento nos vários níveis do outro fator, que é chamado de interação. Naturalmente, o mais comum é planejar um experimento com os tratamentos organizados com, no máximo, três fatores. Uma desvantagem desses delineamentos é o crescimento do número de tratamentos, pois existe a dificuldade para conseguir um ambiente ou local com condições homogêneas que permita distribuir todos eles. Além disso, há aumento de custo.

Aplicação com dois fatores

Admitamos que o interesse seja avaliar a produtividade de matéria seca (MS) de forrageira utilizada para alimentação de bovinos. São avaliados 16 tratamentos, organizados em esquema fatorial 4 x 4 (quatro variedades e quatro tipos de adubação), com três repetições para cada tratamento, resultando em 48 parcelas.

Modelo matemático

$$y_{ijk} = \mu + v_i + a_j + (va)_{ij} + \varepsilon_{ijk}$$

$$(i = 1, 2, ..., 4; j = 1, ..., 4; k = 1,2,3)$$

em que:

y_{ijk} = valor observado na parcela.

μ = efeito médio global.

v_i = efeito fixo da variedade i.

a_i = efeito fixo da adubação j.

$(va)_{ij}$ = efeito fixo da interação entre variedade i e adubação j.

ε_{ijk} = erro aleatório associado a cada parcela; $\varepsilon_{ijk} \sim$ NIID $(0, \sigma^2)$.

A análise de variância para um experimento com delineamento inteiramente casualizado com dois fatores é apresentada na Tabela 4.

Tabela 4. Análise de variância de um delineamento inteiramente casualizado (DIC) com dois fatores.

Fator de variação	Grau de liberdade	Soma de quadrados	Quadrado médio	F
Variedade – v	v - 1 = 3	SQT_v	QMTv = SQTv/(v - 1)	QMTv/QME
Adubação – a	a - 1 = 3	SQT_a	QMTa = SQTa/(a - 1)	QMTa/QME
Interação v x a	(v - 1)(a - 1) = 9	SQT_{vxa}	QMTvxa= SQT_{vxa}/(v - 1)(a - 1)	QMTvxa /QME
Erro	va(k - 1) = 32	SQE	QME = SQE/ va(k - 1)	
Total	vak - 1 = 47	SQ total		

Código SAS

Análise de variância pelo procedimento GLM do SAS e comparações pareadas entre tratamentos pelo teste de Tukey, considerando-se que os dados estejam no arquivo “Fatorial”:

```
proc glm data = fatorial; class variedade adubo;
model ms = variedade adubo variedade*adubo /ss3;
means variedade / hovtest;
lsmeans variedade adubo variedade*adubo / adjust=tukey;
run;
/*
ss3 = calcula a soma de quadrados do tipo 3.
hovtest = testa homogeneidade de variâncias entre variedades pelo teste de levene.
tukey = realiza comparações pareadas entre efeitos fixos e interações pelo teste de tukey.
*/
```

Aplicação com três fatores

Quando os tratamentos são organizados em três fatores, além do número de tratamentos ser grande, se a interação tripla for significativa, geralmente dificulta os testes de comparações múltiplas de médias e a interpretação. Os recursos gráficos podem ser bastante úteis nessa situação.

Em um experimento cujo objetivo foi avaliar a proteína bruta (PB), em percentagem (%), de capim *Coast-cross*, os tratamentos foram organizados em esquema fatorial 2 x 5 x 5 (duas fontes de nitrogênio (N): ureia e nitrato de amônia; cinco doses de adubo nitrogenado: 0, 25 kg ha^{-1}, 50 kg ha^{-1}, 100 kg ha^{-1} e 200 kg ha^{-1}, em cinco cortes consecutivos). Em um experimento desse tipo, os cinco cortes são considerados

como medidas repetidas. No entanto, o objetivo aqui é mostrar como o uso adequado de uma figura pode facilitar a interpretação de uma interação tripla significativa. Na Figura 2, as duas linhas indicam as duas fontes de nitrogênio - N (ureia e nitrato de amônia), enquanto o comportamento delas, como a inclinação, mostra o efeito das cinco doses de N. Finalmente, cada gráfico mostra o efeito dos cortes. Para representar uma interação tripla como na Figura 2 a seguir, é necessário que a escala dos eixos (x, y) seja a mesma.

Observa-se que a PB tem melhor desempenho, de forma geral, na presença da fonte de nitrato de amônio. Resta saber o desempenho em relação ao custo ou viabilidade econômica.

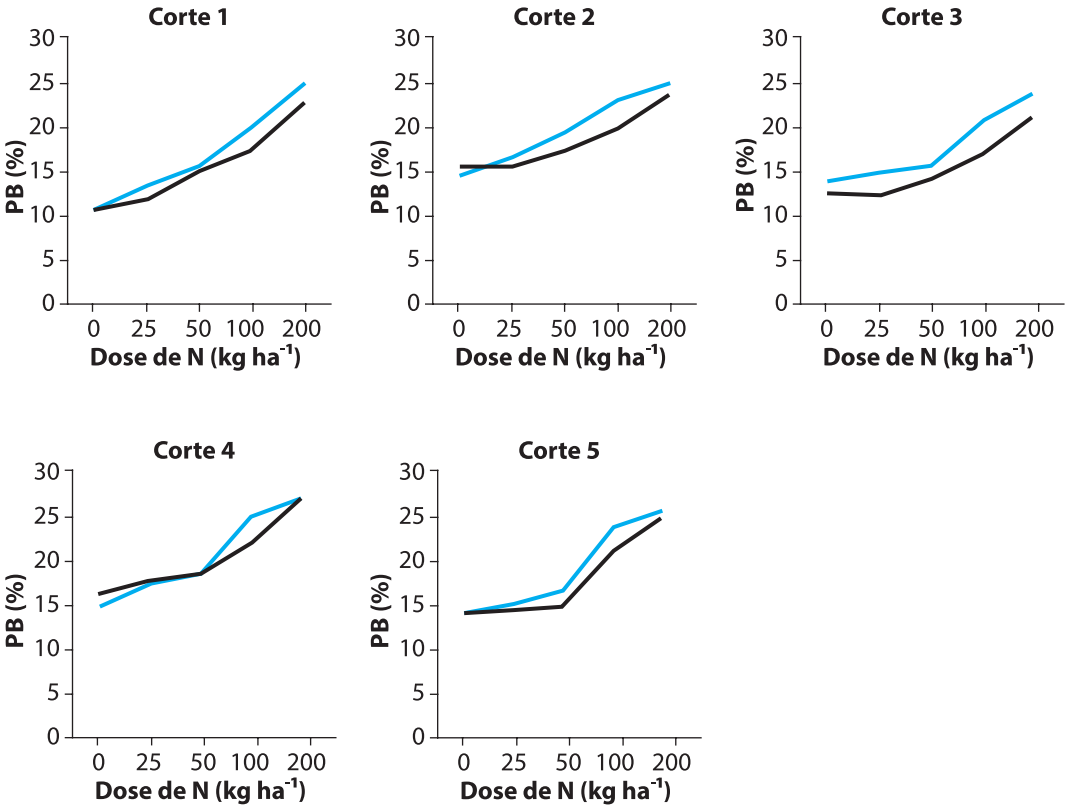


Figura 2. Médias obtidas por quadrados mínimos em função de cinco doses de nitrogênio (N), 0, 25 kg ha^{-1}, 50 kg ha^{-1}, 100 kg ha^{-1}, 200 kg ha^{-1}, para proteína bruta (PB), considerando-se duas fontes de N, ureia (▬)e nitrato de amônio (▬), em cinco cortes no período de 1999 a 2000.

Fonte: Corrêa et al. (2007).

Delineamento inteiramente casualizado com subamostragem

Em várias situações experimentais, não há necessidade e condições de avaliar toda a parcela, mas apenas amostras desta, já que as unidades de observação são subamostras das parcelas. Nesse caso, como existe variabilidade entre as subamostras, elas devem fazer parte da análise de variância, e não se deve calcular a média. Quanto maior o número de subamostras, maior é a precisão do experimento. Em muitas situações, como em laboratórios, é comum o pesquisador fazer duas ou três avaliações de uma mesma medida e calcular a média. Isso acontece, porque, na verdade, trata-se de aferições de medida, mas o fenômeno é o mesmo. Na área médica, é comum o paciente ter que repetir duas ou três vezes o exame de sangue, urina e fezes, para se obter o resultado preciso.

Alguns exemplos de subamostras:

a) Folhas dentro de uma árvore.

b) Frutas dentro de uma caixa.

c) Plantas dentro de touceira.

d) Plantas dentro de uma parcela.

e) Amostras de solos dentro de uma parcela.

Aplicação

Em um experimento, o objetivo foi avaliar a percentagem de germinação de sementes. Foram utilizados quatro substratos (tratamentos) e cinco repetições. Dentro de cada parcela, foram utilizadas cinco subamostras de 20 sementes cada.

Modelo matemático

$$y_{ijk} = \mu + t_i + \varepsilon_{j(i)} + \varepsilon_{k(ij)}$$

$$(i = 1, 2, ..., 4;\ j = 1, ..., 5;\ k = 1, ..., 5)$$

em que:

y_{ijk} = valor observado na amostra.

μ = efeito médio global.

t_i = efeito fixo do tratamento i.

$\varepsilon_{j(i)}$ = efeito aleatório entre parcelas.

$\varepsilon_{k(ij)}$ = erro amostral; $\varepsilon_{k(ij)} \sim$ NIID $(0, \sigma^2)$.

A análise de variância é apresentada na Tabela 5.

Tabela 5. Análise de variância de um delineamento inteiramente casualizado (DIC) com subamostragem.

Fator de variação	Grau de liberdade	Soma de quadrados	Quadrado Médio	F
Tratamentos – t	t - 1 = 3	SQT	QMT = SQT/(t -1)	QMT/QMR
Resíduo	t(r - 1) = 16	SQR	QMR = SQR/ t(r - 1)	
(Parcelas)	(tr - 1) = (19)			
Erro amostral – k	tr(k - 1) = 80			
Total	trk -1 = 99	SQ total		

Código SAS

Código SAS para trat = tratamentos; rep = repetições;

```
proc glm;
  class trat rep;
  model y = trat rep(trat);
  test h = trat  e = rep(trat);
run;
```

Delineamento inteiramente casualizado com a unidade experimental avaliada no tempo

Aplicação

Neste experimento, cada unidade experimental ou indivíduo é avaliado ao longo do tempo. Como exemplo, serão apresentados os resultados de um experimento realizado na Embrapa Pecuária Sudeste (Freitas et al., 2012), São Carlos, SP, em que foram comparados três tratamentos com oito vacas holandesas cada.

A produtividade de leite (*y*) de cada vaca foi avaliada por sete controles a cada 14 dias. Em experimentos com animais, as respostas avaliadas no mesmo indivíduo são correlacionadas, em razão de uma contribuição comum do animal (genética), e as avaliações em animais diferentes são independentes. Inicialmente, houve um período pré-experimental de 7 dias, e essa produtividade foi analisada como covariável (pre).

As informações do experimento e da estrutura dos dados estão no arquivo SAS.

```
/* Informações de identificação
Objetivo

Avaliar a produtividade de leite de vacas holandesas durante sete controles com intervalo
de 14 dias cada.
Delineamento experimental
Inteiramente casualizado com três tratamentos e oito vacas cada. A parcela foi
representada pela vaca, sendo a produtividade de leite (y) avaliada durante sete controles
e os dados analisados por meio de medidas repetidas.

Descrição das variáveis do programa SAS - input
TRAT = Tratamentos

1.  confinamento em área de descanso com silagem de milho + 11,0 kg de concentrado;
2.  pastejo restrito de alfafa + silagem de milho + 11,0 kg de concentrado;
3.  pastejo de alfafa à vontade + silagem de milho + 8,0 kg de concentrado.

NVACA = número de identificação da vaca;
CONTROLE = produtividade de leite, kg/vaca/dia avaliada no tempo (1 a 7);
Pre = produtividade avaliada no período pré-experimental de 7 dias (covariável)
Os comandos após "data;" transformam um arquivo original com dez colunas
(tratamento vaca controle1 - controle8) em um arquivo final com quatro variáveis
(tratamento vaca controle y). */
data;
input tratamento vaca controle1- controle8;
pre = controle1;
controle = 1; y = controle2;output;
controle = 2; y = controle3;output;
controle = 3; y = controle4;output;
controle = 4; y = controle5;output;
controle = 5; y = controle6;output;
controle = 6; y = controle7;output;
controle = 7; y = controle8;output;
/*drop controle1- controle8;*/
datalines;
```

```
1  343  33.0  34.2  28.0  30.8  31.6 30.0  28.6  27.2
1  365  33.8  34.6  31.4  31.0  28.4 31.0  32.2  32.2
...
3  548  27.4  23.8  29.2  24.6  24.4  23.9  17.2  24.4
;
```

Modelo matemático

$$y_{ijk} = \mu + \alpha_i + \delta_{ij} + t_k + (\alpha t)_{ik} + \varepsilon_{ijk}$$

em que:

y_{ijk} = valor observado na vaca ou indivíduo.

μ = efeito médio geral.

$\alpha_i, t_k, (\alpha t)_{ik}$ = efeito fixo do tratamento i; controle k e interação entre esses dois efeitos.

δ_{ij} = efeito aleatório entre vacas.

ε_{ijk} = erro aleatório entre avaliações no indivíduo;$Var(\varepsilon_{ijk}) = R$.

A análise de variância referente a esse modelo matemático está apresentada na Tabela 6.

Tabela 6. Análise de variância de um delineamento inteiramente casualizado (DIC) com a unidade experimental avaliada no tempo.

Fator de variação	Grau de liberdade	Soma de quadrados	Quadrado Médio	F
Tratamentos – a	a - 1 = 2	SQT	QMT = SQT/(a - 1)	QMT/QMR_a
Vaca (j) = Erro a	a(j - 1) = 21	SQR_a	$QMR_a = SQT_a/a(j - 1)$	-
Controles – c	(c - 1) = 6	SQC	QMC = SQC/(c - 1)	QMC/QME_b
Interação a x c	(a -1)(c - 1) = 12	SQI	QMI = SQI/(a - 1)(c - 1)	QMI/QME_b
Erro b	a(c-1)(j - 1) = 126	SQE_b	$QME_b = SQE_b/a(c-1)(j - 1)$	
Total	ajc - 1 = 167	SQ total		

Código SAS

```
proc mixed covtest asycov;
class tratamento vaca controle;
model y = tratamento controle tratamento *controle pre / s;
```

```
repeated controle / type = hf subject = vaca r rcorr;
lsmeans tratamento controle tratamento*controle/pdiff = all adjust = tukey;
run;
```

Esse programa foi executado sete vezes, uma vez para cada estrutura de variância e covariância da Tabela 7.

Tabela 7. Variâncias e covariâncias com respectivos parâmetros e elementos (i = linha e j = coluna).

Estrutura	Definição	Parâmetro[1]	Elemento i,j
Ar(1)	Autorregressiva de primeira ordem	2	$\sigma^2\rho^{\lvert i-j\rvert}$
Arma(1,1)	Autorregressiva de primeira ordem de média móvel	3	$\sigma^2[\gamma\rho^{\lvert i-j\rvert-1}\mathbf{1}(i \neq j) + \mathbf{1}(i = j)]$
Cs	Simetria composta	2	$\sigma^2 + \sigma_i\ \mathbf{1}(i = j)$
Csh	CS heterogênea	k + 1	$\sigma_i\sigma_j[\rho\mathbf{1}(i \neq j) + \mathbf{1}(i = j)]$
HF	Huynh-Feldt	k + 1	$(\sigma^2_i + \sigma^2_j)/2 + \lambda\mathbf{1}(i \neq j)$
Vc	Componente de variância	q	σ^2_1
Un	Não estruturada	k(k + 1)/2	σ_{ij}

[1] k = número de controles = 7; q = 1= número de fatores.

As estruturas de variâncias e covariâncias com as informações: número de parâmetros, logaritmo da função de verossimilhança restrita (−2L), *Akaike's Information Criterion* (AIC) e análise de variância do tipo III (Pr > F) estão na Tabela 8.

Tabela 8. Estruturas de variâncias e covariâncias, número de parâmetros, logaritmo da função de verossimilhança restrita (−2L), *Akaike's Information Criterion* (AIC) e análise de variância do tipo III (Pr > F).

Estrutura[1]	N⁰ de parâmetros	-2L	AIC	Pr > F Tratamento-controle		Tratamento x controle
HF	8	725,7	741,7	0,4047	<,0001	0,9546
Csh	8	732,9	748,9	0,3956	<,0001	0,9238
Ar(1)	2	745,7	749,7	0,2653	<,0011	0,9641
Cs	2	740,9	749,9	0,4018	<,0001	0,9546
Un	28	703,1	759,1	0,3394	<,0003	0,8249
Vc	1	768,9	770,9	0,0679	<,0003	0,9929

[1]HF: Huynh-Feldt; Csh: simetria composta heterogênea; Ar(1): autoregressiva de primeira ordem; Cs: simetria composta; Un: não estruturada; Vc: componente de variância.

De acordo com o valor de AIC na Tabela 8, a estrutura de covariância escolhida de imediato seria a H-F, pois foi a que apresentou o menor valor de AIC, cuja ordem foi Huynh-Feldt (H-F), simetria composta heterogênea (Csh), autorregressiva de primeira ordem; (Ar(1)), simetria composta (Cs), não estruturada (Un) e componente de variância (Vc). No entanto, o número de parâmetros de cada uma é importante na escolha. Porém, é interessante verificar se ela difere estatisticamente das demais.

Para comparar duas matrizes R_i e R_j, faz-se a diferença de "-2L" entre essas matrizes, que, sob a hipótese de nulidade, o valor obtido tem distribuição de qui-quadrado (χ^2), com graus de liberdade igual à diferença do número de parâmetros entre elas. A comparação pareada entre a estrutura H-F com as demais é feita a seguir:

$$\text{HF versus Csh} \rightarrow |725{,}7 - 732{,}9| \rightarrow \chi^2_0 = 7{,}2 \ (p \leq 0{,}05).$$
$$\text{HF versus Ar(1)} \rightarrow |725{,}7 - 745{,}7| \rightarrow \chi^2_6 = 20{,}0 \ (p \leq 0{,}01).$$
$$\text{HF versus Cs} \rightarrow |725{,}7 - 740{,}9| \rightarrow \chi^2_6 = 15{,}2 \ (p \leq 0{,}05).$$
$$\text{HF versus Un} \rightarrow |725{,}7 - 703{,}1| \rightarrow \chi^2_{20} = 22{,}6 \ (p > 0{,}05).$$
$$\text{HF versus Vc} \rightarrow |725{,}7 - 768{,}9| \rightarrow \chi^2_7 = 43{,}2 \ (p \leq 0{,}001).$$

De acordo com essa avaliação, a estrutura H-F não diferiu estatisticamente de Un ($p > 0{,}05$); porém, pelo fato de Un ter número excessivo de parâmetros (28), e estrutura totalmente heterogênea para as variâncias e covariâncias, a escolhida é a H-F, por várias razões. Possui apenas oito parâmetros; e, embora a estrutura seja heterogênea, a interpretação é mais simples do que a Un, pois as covariâncias representam médias aritméticas das respectivas variâncias; por último, a H-F possui menor valor de AIC do que Un (741,7 versus 759,1).

Na análise de variância realizada com a estrutura H-F (Tabela 8), houve redução da produtividade de leite ($p < 0{,}0001$) ao longo dos sete controles; porém, não houve diferença significativa ($p > 0{,}05$) global entre tratamentos nem entre tratamentos dentro de controles ($p > 0{,}05$). Resultados semelhantes foram obtidos com as estruturas H-F, Cs, Ar(1), Csh e Un. Entretanto, os efeitos de tratamentos foram quase significativos para VC ($p = 0{,}0679$), o que comprova a inadequabilidade dessa estrutura no presente estudo. As médias e erros-padrão obtidos dessa análise para tratamentos, controles e interação tratamentos × controles estão na Tabela 9.

Tabela 9. Médias e erros-padrão obtidos por quadrados mínimos e testes de hipóteses obtidos para os seguintes fatores: tratamentos, controles e interação tratamentos × controles.

Fator							
Tratamento (T)* (P = 0,3013)							
1	2	3					
29,6 + 0,7a	31,1 + 0,7a	30,1 + 0,7a					
Controle (C)** (p < ,0001)							
	1	2	3	4	5	6	7
	31,8 ± 0,5a	31,5 ± 0,5ab	30,5 ± 0,6bc	29,8 ± 0,4c	30,1 ± 0,7c	29,7 ± 0,7c	28,1 ± 0,6d
Interação tratamento (T) × controle (C) (P = 0,9455)							
C x T	**1**	**2**	**3**	**4**	**5**	**6**	**7**
1	31,5 ± 0,9a	30,5 ± 0,9a	29,8 ± 1,1a	29,2 ± 0,7a	29,6 ± 1,2a	29,2 ± 1,3a	27,0 ± 1,1a
2	31,8 ± 0,9a	32,3 ± 1,0a	31,2 ± 1,1a	31,4 ± 0,7a	30,8 ± 1,2a	30,8 ± 1,3a	28,8 ± 1,1a
3	32,1 ± 0,9a	31,8 ± 0,9a	30,5 ± 1,1a	29,0 ± 0,7a	29,9 ± 1,2a	29,0 ± 1,3a	28,5 ± 1,1a

" * " e "**" Sigificativo a 5% e a 1% de probabilidade, respectivamente, pelo teste F.
Médias seguidas por letras iguais, na linha, não diferem entre si, pelo teste de Tukey, a 5% de probabilidade.

Nas Tabelas 10 e 11, estão apresentadas, respectivamente, a matriz de covariância e de correlação estimada pelo teste da razão de verossimilhança, considerando a hipótese nula de que matriz de variância e covariância dos erros (R) é igual a HF para os sete controles leiteiros. Essas matrizes são obtidas para um indivíduo, porém, todos são assumidos ter as mesmas estruturas. Observou-se heterogeneidade de covariância e, consequentemente, heterogeneidade de correlação entre os controles; a menor correlação (0,028) foi entre os controles 1 e 4, e a maior (0,578) entre os controles 5 e 6.

Tabela 10. Matriz de covariância estimada na análise de sete controles leiteiros (C_1 a C_7) de vacas da raça holandesa, como medidas repetidas, em que a estrutura de variância e covariância ajustada entre controles dentro de vaca é a Huynh-Feldt.

	C_1	C_2	C_3	C_4	C_5	C_6	C_7
C_1	6,726	2,032	3,096	0,149	3,780	4,854	2,865
C_2		7,878	3,673	0,725	4,356	5,430	3,441
C_3			10,007	1,790	5,421	6,495	4,506
C_4				4,113	2,474	3,547	1,558
C_5					11,375	7,179	5,190
C_6						13,523	6,263
C_7							9,545

Tabela 11. Matriz de correlação estimada na análise de sete controles leiteiros (C_1 a C_7) de vacas da raça holandesa, como medidas repetidas, em que a estrutura de variância e covariância ajustada entre controles dentro de vaca é a Huynh-Feldt.

	C_1	C_2	C_3	C_4	C_5	C_6	C_7
C_1	1,000	0,279	0,377	0,028	0,432	0,509	0,357
C_2		1,000	0,413	0,127	0,460	0,526	0,396
C_3			1,000	0,279	0,508	0,558	0,461
C_4				1,000	0,361	0,475	0,248
C_5					1,000	0,578	0,498
C_6						1,000	0,551
C_7							1,000

Exercícios[9]

1) No texto a seguir, existem afirmações incorretas quanto a conceitos de estatística. Reescreva o texto colocando definições corretas e sublinhe ou coloque em negrito onde houve definições incorretas.

Quando se consegue um ambiente homogêneo onde pode instalar todos os tratamentos, o delineamento inteiramente (IC) é um dos mais utilizados e eficientes. Esse delineamento tem várias vantagens: não há limite para o número de tratamentos que se quer avaliar, tem maior flexibilidade, o número de repetições pode variar entre tratamentos, a análise estatística é simples, o número de graus de liberdade associado ao erro geralmente é grande, perda de parcela não é problema, o modelo requer poucas suposições, possibilita os resultados mais abrangentes e o custo na condução do experimento é um dos menores. Dentre as desvantagens do IC tem-se a sua ineficiência quando o ambiente não é homogêneo e o número mínimo de parcelas requerido é 20, independente do número de tratamentos. O modelo matemático é do tipo $y_{ik} = \mu + t_i + r_j + \varepsilon_{ik}$, em que y_{ik} é o valor observado no tratamento i (t_i) da repetição j (r_j); μ é o efeito médio global; e ε_{ik} é o erro aleatório, o qual não requer nenhuma suposição.

2) Um experimento com delineamento inteiramente casualizado foi realizado considerando nove tratamentos organizados em esquema fatorial 3 x 3 [três grupos genéticos (G): Canchim, ½ Canchim + ½ Nelore e Nelore, e três níveis de proteína bruta (P): 6, 10 e 13%], com oito repetições (rep) cada, sendo a unidade experimental a novilha e o peso corporal (y) a variável resposta.

[9] As respostas dos exercícios podem ser consultadas no Apêndice 1.

Descreva o modelo matemático, o quadro de análise de variância e um programa SAS para análise de dados.

3) Os dados da Figura 3 mostram resultados de fibra em detergente neutro, % (FDN) de um trabalho de avaliação de pastagem de capim *Coast-cross* – *Cynodon dactylon* cv. *Coast-cross,* realizado na Embrapa Pecuária Sudeste, São Carlos, SP (Correa et al. 2007). Nas parcelas foram aplicados 10 tratamentos em esquema fatorial 2×5 (duas fontes de N: ureia e nitrato de amônio; e cinco doses de N por corte: 0 kg ha^{-1}, 25 kg ha^{-1}, 50 kg ha^{-1}, 100 kg ha^{-1}, 200 kg ha^{-1}); nas subparcelas, foram considerados cinco cortes consecutivos. Interpretar os resultados analisando a Figura 3.

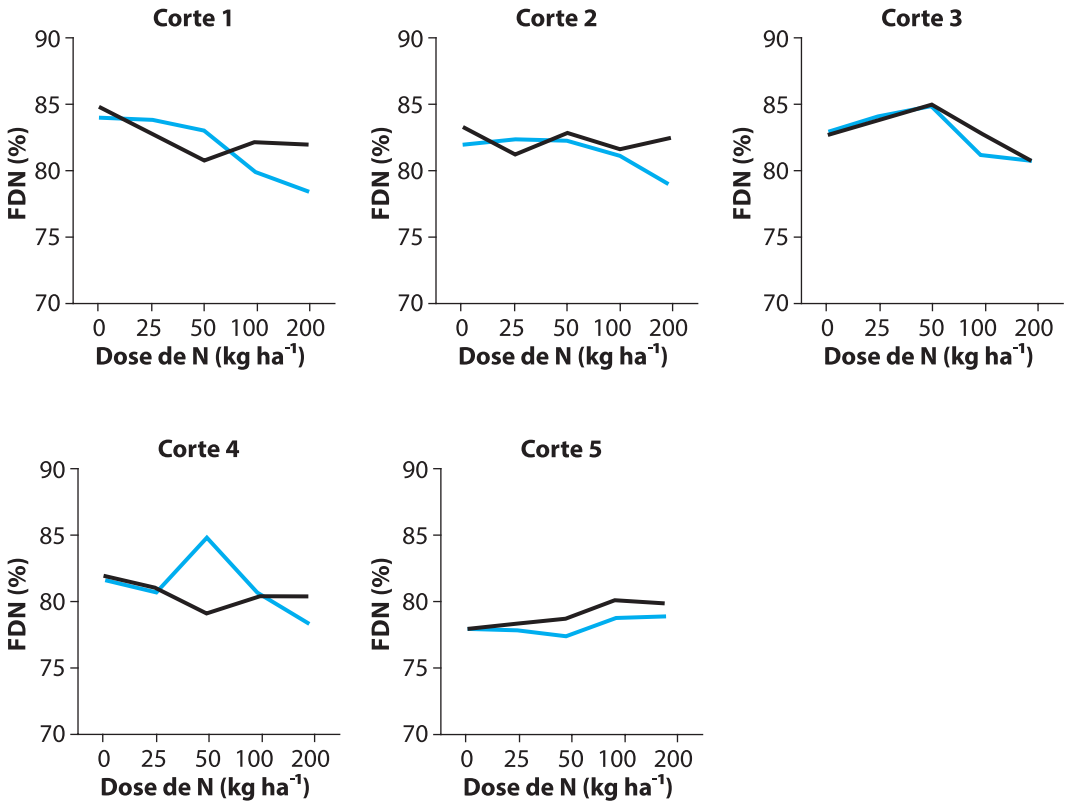


Figura 3. Médias obtidas por quadrados mínimos em função de duas fontes de N: ureia (▬) e nitrato de amônio (▬) com cinco doses de nitrogênio cada: N (0 kg ha^{-1}, 25 kg ha^{-1}, 50 kg ha^{-1}, 100 kg ha^{-1} e 200 kg ha^{-1}) para fibra em detergente neutro,%, (FDN) e cinco cortes no ano de 1999−2000.

Fonte: (Corrêa et al., 2007).

4) Com relação ao delineamento inteiramente casualizado (DIC), assinale a alternativa verdadeira:

( ) Para experimentos em condições de campo o DIC é mais complicado do que os outros delineamentos devido à dificuldade de obter um ambiente homogêneo.

( ) Geralmente, é mais fácil obter animais de pequeno porte (aves e suínos) mais homogêneos do que animais de grande porte, como bovinos.

( ) Em geral, é mais fácil obter ambientes homogêneos ou uniformes em laboratórios e casas de vegetação do que em condições de campo.

( ) Estudar uma interação tripla em análise de variância possibilita resultados mais conclusivos, porém, paga-se uma penalidade devido à redução dos raus de liberdade do resíduo.

( ) Todas as respostas acima são verdadeiras.

5) Num experimento inteiramente casualizado, de competição de variedades de batata, foram utilizados 5 tratamentos (cultivares) e 6 repetições, cuja análise de variância está apresentada na Tabela 12. Interprete essa análise.

Tabela 12. Análise de variância do delineamento inteiramente casualizado.

Fatores de variação	Grau de liberdade	Soma de quadrado	Quadrado médio	F
Tratamentos – t	4	2.200,95	550,24	31,23
Erro	25	440,52	17,62	

Capítulo 10

Blocos casualizados

Introdução

O delineamento de blocos completos casualizados (DBC) é um dos mais utilizados em agricultura. Enquanto no tradicional delineamento inteiramente casualizado (DIC), a preocupação é encontrar um ambiente homogêneo para distribuir todos os tratamentos do experimento, no DBC a preocupação é escolher um ambiente homogêneo para cada bloco, por isso a sua grande popularidade.

Nos DBC consideram-se os três princípios básicos da experimentação: repetição, casualização e controle local. Dentro de cada bloco ou repetição, as condições devem ser as mais homogêneas possíveis, porém pode haver grande variação entre os blocos. Os tratamentos são distribuídos aleatoriamente dentro de cada bloco.

Nos experimentos com animais, por exemplo, o bloco pode ser um local como galpão, tanque, aquário, baia, etc., desde que as características ambientais sejam homogêneas. Em experimentos com grandes animais, como na produtividade de leite em bovinos, os blocos podem ser formados por animais que possuam condições semelhantes quanto à ordem e data de parto, ordem de parição, sexo, produtividade de leite, leitegada, peso dos animais, etc. Existem situações particulares em que o próprio animal pode ser o bloco.

Quando se trabalha com culturas agronômicas em que geralmente se avalia a produção, os exemplos mais comuns de blocos se referem às características do solo, tais como topografia, gradiente de fertilidade e de inclinação. Para experimentos com culturas anuais, essas exigências geralmente são maiores do que para culturas perenes.

Em termos de eficiência, quando comparado com o DIC, considerando-se o mesmo conjunto de tratamentos, o DBC é menos eficiente, pois, neste caso, parte dos graus de liberdade do erro experimental é atribuída ao efeito de blocos. Contudo, como a preocupação é a escolha de ambiente homogêneo apenas para instalar uma repetição dos tratamentos (bloco), a sua versatilidade é muito grande. Cada bloco recebe por sorteio todos os tratamentos e, quando não é possível distribuir todos os tratamentos num mesmo bloco, têm-se os delineamentos de blocos incompletos (DBI).

Para cada delineamento apresentado, serão discutidas as suas características e apresentados o modelo matemático, o quadro de análise de variância (Anova), com a ilustração de exemplo pelo procedimento do modelo linear generalizado (*generalized linear model*, em inglês – GLM).

Blocos casualizados completos

Corresponde ao tradicional blocos casualizados em que os tratamentos principais, denominados de fator A, são distribuídos no bloco ou repetição. Seja um experimento com quatro tratamentos principais (a_1, a_2, a_3, a_4) e cinco blocos (r_1, r_2, r_3, r_4, r_5), totalizando 20 parcelas. Para a instalação do experimento, de modo a garantir a aleatoriedade das parcelas no campo e a independência dos erros entre elas, os tratamentos são sorteados ao acaso dentro de cada bloco, conforme croqui de campo apresentado na Figura 1.

Bloco 1	Bloco 2	Bloco 3	Bloco 4	Bloco 5
a_2	a_4	a_1	a_4	a_1
a_3	a_3	a_3	a_3	a_4
a_1	a_1	a_2	a_2	a_3
a_4	a_2	a_4	a_1	a_2

Figura 1. Croqui de campo de blocos casualizados mostrando-se quatro tratamentos principais (a_1, a_2, a_3, a_4) distribuídos aleatoriamente em cinco blocos.

Modelo matemático

$$y_{ij} = \mu + a_i + r_j + \varepsilon_{ij}$$

$$(i = 1, 2, ..., 4; j = 1, ..., 5)$$

em que:

y_{ij} = valor observado na parcela.

μ = média global.

a_i = efeito fixo do tratamento i.

r_j = efeito do bloco j.

ε_{ij} = erro aleatório associado a cada parcela; $\varepsilon_{ij} \sim$ NIID $(0, \sigma^2)$.

A análise de variância está apresentada na Tabela 1.

Tabela 1. Análise de variância de blocos casualizados.

Fator de variação	Grau de liberdade		Soma de quadrados	Quadrado médio	F
Tratamentos – a	a - 1	3	SQT	QMT = SQT/(a - 1)	QMT/QME
Blocos – r	r - 1	4	SQB	QMB = SQB/(r - 1)	QMB/QME
Erro	(r -1)(a - 1)	12	SQE	QME = SQE/(r - 1)(a - 1)	
Total	ra - 1	19	SQ total	SQ total	

A seguir, há um exemplo de um arquivo de dados fictícios "dbc" com análise de variância pelo procedimento GLM do Statistical Analysis System (SAS) para o desenho experimental da Figura 1, em que *y* é o valor observado.

```
data dbc; input bloco A  y;
datalines;
1 1 22.0
1 2 23.5
...
5 4 25.1
 ;
proc glm;
 class bloco A ;
 model y = bloco A; random bloco;
/* random bloco, calcula a esperança dos quadrados médios dos efeitos de blocos
   e tratamentos */
lsmeans A/stderr pdiff = all adjust = tukey;
/*testa todas as comparações pareadas das medias dos tratamentos A pelo teste de tukey
*/;run;
```

Blocos casualizados com parcela dividida (*split-plot*)

Neste delineamento, dois grupos de tratamentos são avaliados; os tratamentos principais (A) são distribuídos nas parcelas, como no DBC, e os tratamentos secundários (B) são distribuídos nas subparcelas. Seja a situação com quatro tratamentos A (a_1, a_2, a_3, a_4), em que cada parcela seja dividida em três partes iguais para receber um tratamento secundário B com três níveis (b_1, b_2, b_3). Após o sorteio dos tratamentos principais (A) nas parcelas, sorteiam-se os tratamentos secundários (B) dentro de cada subparcela.

Cite-se como exemplo na área agronômica a situação em que cinco blocos são utilizados e o objetivo principal é determinar a influência de quatro doses de nitrogênio (0 kg ha^{-1}, 20 kg ha^{-1}, 40 kg ha^{-1} e 60 kg ha^{-1}) e, como interesse secundário, verificar a resposta dessa adubação na produtividade (y), em kg ha^{-1}, de três variedades de capim. Assim, os tratamentos atribuídos às parcelas são as quatro doses de nitrogênio e, na subparcela, as três variedades, ou seja:

a) Tratamentos principais (a_1, a_2, a_3, a_4) → sorteio em faixa vertical.

b) Tratamentos secundários (b_1, b_2, b_3) → sorteio dentro de cada nível do fator A.

O croqui de campo está apresentado na Figura 2.

Bloco 1				Bloco 2				Bloco 3				Bloco 4				Bloco 5			
a_1	a_3	a_2	a_4	a_3	a_1	a_4	a_2	a_4	a_2	a_1	a_3	a_1	a_3	a_2	a_4	a_4	a_3	a_1	a_2
b_3	b_1	b_1	b_2	b_1	b_3	b_1	b_2	b_3	b_1	b_1	b_2	b_1	b_2	b_1	b_3	b_2	b_3	b_1	b_1
b_2	b_3	b_2	b_3	b_3	b_2	b_2	b_3	b_2	b_2	b_3	b_3	b_2	b_3	b_3	b_2	b_3	b_2	b_3	b_2
b_1	b_2	b_3	b_1	b_2	b_1	b_3	b_1	b_1	b_3	b_2	b_1	b_3	b_1	b_2	b_1	b_1	b_1	b_2	b_3

Figura 2. Croqui de campo para blocos completo com parcela dividida (*split-plot*). Os quatro tratamentos principais (a_1, a_2, a_3, a_4) são distribuídos aleatoriamente em cada um dos cinco blocos e correspondem às parcelas. Os três tratamentos secundários (b_1, b_2, b_3) são distribuídos aleatoriamente em cada parcela e correspondem às subparcelas.

Modelo matemático

$$y_{ijk} = \mu + a_i + r_j + (ar)_{ij} + b_k + (ab)_{jk} + \varepsilon_{ijk}$$

$$(i = 1, 2, ..., 4;\ j = 1, ..., 5;\ k = 1, ..., 3)$$

em que:

y_{ijk} = valor observado na subparcela.

μ = média global.

a_i = efeito do tratamento i atribuído à parcela.

r_j = efeito do bloco j.

$(ar)_{ij}$ = erro aleatório entre parcelas.

b_k = efeito do tratamento k atribuído à subparcela.

$(ab)_{jk}$ = efeito da interação entre tratamentos a e b.

ε_{ijk} = erro aleatório entre subparcelas; $\varepsilon_{ijk} \sim NIID\ (0, \sigma^2)$.

O resumo da análise de variância está apresentado na Tabela 2.

Tabela 2. Análise de variância de blocos casualizados com parcelas divididas (*split-plot*).

Fator de variação	Grau de liberdade		Soma de quadrados	Quadrado médio	F
Blocos – r	r - 1	4	SQB	QMB = SQB/(r - 1)	QMB/QME_a
Doses – a	a - 1	3	SQTa	QMTa = SQTa/(a - 1)	$QMTa/QME_a$
Erro a	(r - 1)(a - 1)	12	SQE_a	$QME_a = SQE_a/(r - 1)(a - 1)$	
Variedades – b	(b - 1)	2	SQTb	QMTb = SQTb/(b - 1)	$QMTb/QME_b$
Interação a*b	(a - 1)(b - 1)	6	SQI_{axb}	$QMI = SQI_{axb}/(a - 1)(b -1)$	QMI/QME_b
Erro b	a(r - 1)(b - 1)	32	SQE_b	$QME_b = SQE_b/a(r - 1)(b -1)$	
Total	abr -1	59	SQ total		

A seguir, há exemplo de um arquivo de dados fictícios *split-plot* com análise de variância pelo procedimento GLM do SAS para o desenho experimental da Figura 2, em que *y* é o valor observado.

```
data splitplot;
 input bloco a b y;
datalines;
1  1   1 55.0
1  1   2 51.5
1  1   3 52.3
...
5  4   3 53.1
;
proc glm;
class bloco a  b;
 model y = bloco a bloco*a b a*b;
 test h = a  e = bloco*a;
lsmeans a*b / pdiff = all adjust = tukey;
/*testa todas as comparações pareadas da interação (a*b) pelo teste de tukey */.
lsmeans a*b/ slice = b;
/* testa os tratamentos principais a dentro de cada nível do tratamento da subparcela b*/
 run;
```

Blocos casualizados em faixas (*strip block design*)

Os delineamentos de blocos em faixas (*strip block design*) têm semelhança com os *split-plot*, porém a distribuição dos tratamentos principais (a) nas parcelas e dos tratamentos secundários (b) nas subparcelas recebe uma casualização diferente da realizada nos blocos com parcela dividida (*split-plot*). Cite-se como exemplo de utilização desse delineamento na agricultura a situação em que dois grupos de tratamentos ou fatores (a e b) exigem grande área para serem instalados. Geralmente esses tratamentos são derivados da combinação de práticas culturais, tais como: preparo do solo (aração, gradeação), plantio (adubação, método, época e profundidade de semeadura, variedade), manejo (adubação em cobertura, irrigação), tratamento fitossanitário (aplicação de agrotóxicos).

Imagine a situação em que quatro espaçamentos e três tipos de aração são dois fatores que precisam ser instalados no mesmo experimento para avaliar o rendimento de uma cultura (y), em kg ha^{-1}. Com esse planejamento, a área experimental de cada bloco será dividida inicialmente em quatro parcelas, denominadas de faixas verticais, onde serão distribuídos os quatro níveis do fator a (a_1, a_2, a_3, a_4). A seguir a área do bloco será dividida em três faixas horizontais, e, em cada uma, será sorteado um dos níveis do fator b (b_1, b_2, b_3), de modo que 12 unidades experimentais (4 x 3) serão formadas no bloco. Na Figura 3 está o sorteio dos três níveis do fator b (b_1, b_2, b_3) em cada bloco de acordo com as cores.

Nos experimentos em faixas, há diminuição da precisão de avaliação dos efeitos principais **a** e **b**, devido à restrição na casualização. No entanto, esses experimentos possibilitam grande precisão na avaliação da interação, que é resultante de cada combinação das faixas horizontais e verticais.

	Bloco 1					Bloco 2					Bloco 3					Bloco 4					Bloco 5				
	a_1	a_3	a_4	a_2		a_3	a_2	a_4	a_1		a_4	a_1	a_2	a_3		a_2	a_3	a_4	a_1		a_1	a_3	a_4	a_2
b_3	b_3	b_3	b_3	b_3	b_2	b_2	b_2	b_2	b_2	b_1	b_1	b_1	b_1	b_1	b_2	b_2	b_2	b_2	b_2	b_1	b_1	b_1	b_1	b_1
b_1	b_1	b_1	b_1	b_1	b_3	b_3	b_3	b_3	b_3	b_2	b_2	b_2	b_2	b_2	b_3	b_3	b_3	b_3	b_3	b_2	b_2	b_2	b_2	b_2
b_2	b_2	b_2	b_2	b_2	b_1	b_1	b_1	b_1	b_1	b_3	b_3	b_3	b_3	b_3	b_1	b_1	b_1	b_1	b_1	b_3	b_3	b_3	b_3	b_3

Figura 3. Croqui de campo para blocos completos em faixas (*strip block design*). O bloco é dividido em quatro faixas verticais e os tratamentos principais (a_1, a_2, a_3, a_4) são sorteados. A seguir cada bloco é dividido em três faixas horizontais e os tratamentos secundários (b_1, b_2, b_3) sorteados.

Modelo matemático

$$y_{ikj} = \mu + a_i + r_j + (ar)_{ij} + b_k + (rb)_{jk} + (ab)_{ik} + \varepsilon_{ijk}$$

$$(i = 1, 2, ..., 4; j = 1, ..., 5; k = 1, ..., 3)$$

em que:

y_{ijk} = valor observado na subparcela.

μ = média global.

a_i = efeito do tratamento i atribuído à parcela.

r_j = efeito do bloco j.

$(ar)_{ij}$ = erro aleatório entre parcelas.

b_k = efeito do tratamento k atribuído à subparcela.

$(rb)_{jk}$ = erro aleatório devido ao tratamento b.

$(ab)_{ik}$ = efeito da interação entre tratamentos a e b.

ε_{ijk} = erro aleatório entre subparcelas; $\varepsilon_{ijk} \sim$ NIID $(0, \sigma^2)$.

O resumo da análise de variância está apresentado na Tabela 3.

Tabela 3. Análise de variância de blocos completo em faixas (*strip block design*).

Fator de variação	Grau de liberdade		Soma de quadrados	Quadrado médio	F
Trat a	a - 1	3	SQT_a	$QMT_a = SQT_a/(a - 1)$	QMT_a/QME_a
Blocos – r	r -1	4	SQB	QMB = SQB/(r - 1)	QMB/QME_a
Erro a	(a - 1)(r - 1)	12	SQE_a	$QME_a = SQE_a/(r - 1)(a - 1)$	
Trat b	(b -1)	2	SQT_b	$QMT_b = SQT_b/(b - 1)$	QMT_b/QME_b
Erro b	(r - 1)(b - 1)	8	SQE_b	$QME_b = SQE_b/(r - 1)(b - 1)$	
Interação a x b	(a - 1)(b - 1)	6	SQI_{axb}	$QMI_{axb} = SQI_{axb}/(a - 1)(b - 1)$	QMI_{axb}/QME_c
Erro c	(a - 1)(r - 1)(b - 1)	24	SQE_c	$QME_c = SQE_c/(r - 1)(a - 1)(b - 1)$	
Total	abr - 1	59	SQ total		

A seguir, encontra-se exemplo de um arquivo de dados fictícios, para blocos completo em faixas (*strip block design*), com análise de variância pelo procedimento GLM do SAS para o desenho experimental da Figura 3, em que *y* é o valor observado.

```
data faixa; input bloco a b y;
datalines;
1   1 55.0
1   2 51.5
1   3 52.3
...
5 3 53.1
 ;
proc glm;
class bloco a b;
model y = bloco a bloco*a b bloco*b;
test h = a  e = bloco*a;
test h = b  e = bloco*b;
run;
```

Blocos casualizados em diferentes locais

Os experimentos citados anteriormente muitas vezes são repetidos em diversos locais, distantes um do outro e com condições ambientais diferentes. Este delineamento é bastante utilizado em agricultura quando há várias variedades de uma cultura, como cana-de-açúcar, e se deseja identificar quais as que produzem mais em determinado ambiente. Esse tipo de delineamento tem grande afinidade com a área de melhoramento genético, uma vez que explica a interação genótipo-ambiente.

Cite-se a situação em que queremos testar a produtividade de cinco variedades de cana-de-açúcar e ver o comportamento delas em cinco diferentes locais e em cada local temos cinco blocos.

O croqui de campo está apresentado na Figura 4.

Bloco	Local 1	Local 2	Local 3	Local 4	Local 5
1	$v_1 \mid v_3 \mid v_5 \mid v_2 \mid v_4$	$v_2 \mid v_1 \mid v_5 \mid v_4 \mid v_3$	$v_4 \mid v_3 \mid v_5 \mid v_2 \mid v_1$	$v_2 \mid v_1 \mid v_5 \mid v_4 \mid v_3$	$v_1 \mid v_3 \mid v_5 \mid v_2 \mid v_4$
2	$v_4 \mid v_2 \mid v_5 \mid v_1 \mid v_3$	$v_2 \mid v_1 \mid v_5 \mid v_4 \mid v_3$	$v_5 \mid v_4 \mid v_2 \mid v_3 \mid v_1$	$v_4 \mid v_2 \mid v_5 \mid v_1 \mid v_3$	$v_5 \mid v_4 \mid v_2 \mid v_3 \mid v_1$
3	$v_3 \mid v_1 \mid v_4 \mid v_2 \mid v_5$	$v_5 \mid v_4 \mid v_2 \mid v_3 \mid v_1$	$v_1 \mid v_3 \mid v_2 \mid v_5 \mid v_4$	$v_2 \mid v_3 \mid v_5 \mid v_1 \mid v_4$	$v_2 \mid v_1 \mid v_5 \mid v_4 \mid v_3$
4	$v_2 \mid v_1 \mid v_5 \mid v_4 \mid v_3$	$v_1 \mid v_3 \mid v_5 \mid v_2 \mid v_4$	$v_2 \mid v_1 \mid v_5 \mid v_4 \mid v_3$	$v_3 \mid v_1 \mid v_5 \mid v_2 \mid v_4$	$v_4 \mid v_2 \mid v_5 \mid v_1 \mid v_3$
5	$v_5 \mid v_4 \mid v_2 \mid v_3 \mid v_1$	$v_1 \mid v_3 \mid v_5 \mid v_2 \mid v_4$	$v_4 \mid v_2 \mid v_5 \mid v_1 \mid v_3$	$v_5 \mid v_4 \mid v_2 \mid v_3 \mid v_1$	$v_1 \mid v_3 \mid v_5 \mid v_2 \mid v_4$

Figura 4. Croqui de campo para blocos completos em cinco locais – l (l_1, l_2, l_3, l_4, l_5). Em cada local, é instalado um experimento em blocos casualizados, isto é, são sorteadas aleatoriamente as cinco variedades (v_1, v_2, v_3, v_4, v_5) dentro de cada bloco.

Modelo matemático

$$y_{ijk} = \mu + l_i + (lr)_{j(i)} + v_k + (lv)_{ik} + \varepsilon_{ijk}$$

$$(i = 1, 2, ..., 5;\ j = 1, ..., 5;\ k = 1, ..., 5)$$

em que:

y_{ijk} = valor observado na parcela.

μ = média global.

l_i = efeito do i-ésimo local.

$(lr)_{j(i)}$ = efeito do j-ésimo bloco dentro i-ésimo local.

v_k = efeito da k-ésima variedade.

$(lv)_{ik}$ = efeito da interação da k-ésima variedade com i-ésimo local.

ε_{ijk} = erro aleatório associado ao valor observado na parcela y_{ijk}; $\varepsilon_{ijk} \sim$ IID $(0, \sigma^2)$.

O resumo da análise de variância está apresentado na Tabela 4.

Tabela 4. Análise de variância de blocos completo em diferentes locais.

Fator de variação	Grau de liberdade	Soma de quadrados	Quadrado médio	F
Locais – l	l – 1	SQL	QML = SQL/(l - 1)	QML/QME_a
Blocos(l) = erro a	l(r - 1)	SQE_a	$QME_a = SQE_a /l(r - 1)$	
Variedades – v	v – 1	SQV	QMV = SQV/(v - 1)	QMV/QMR_b
Interação l x v	(l - 1)(v - 1)	SQI_{lxv}	$QMI_{lxv} = SQI_{lxv}/[(l - 1)(v - 1)]$	QMI_{lxv} /QMR_b
Erro b	l(r - 1)(v - 1)	SQR_b	$QMR_b = SQR_b/l(r - 1)(v - 1)$	
Total	lvr - 1	SQ total		

A seguir há exemplo de um arquivo de dados fictícios, para blocos completo em diferentes locais, com análise de variância pelo procedimento GLM do SAS para o desenho experimental da Figura 4, em que *y* é o valor observado.

```
data local;
input l bloco v y;
datalines;
1 1 1  55.0
1 1 2  51.5
...
5 5 5 53.1
;
proc glm; class l bloco v
model y = l bloco(l) v l*v;
 test h = local e = bloco(l); run;
```

Blocos completos com medidas repetidas no tempo

O croqui de campo corresponde ao tradicional delineamento em DBC, pois as avaliações no tempo são realizadas na mesma parcela experimental, denominada de indivíduo, e os dados são analisados como medidas repetidas. Muitas vezes, esses experimentos são analisados erradamente como blocos completos casualizados com parcela dividida (*split-plot*). A diferença fundamental é que nos *split-plot* os tratamentos secundários (B) são distribuídos ao acaso nas subparcelas, enquanto, no presente caso, as avaliações são realizadas ao longo do tempo e não podem ser sorteadas.

Como exemplo, consideremos um experimento realizado na Embrapa Pecuária Sudeste, São Carlos, SP, onde foi analisada a produtividade de matéria seca (PMS) de alfafa (*Medicago sativa* L.), em kg ha^{-1}. Foi utilizado o delineamento em DBC, com duas repetições, e os tratamentos aplicados às parcelas principais consistiram da avaliação de 92 genótipos de alfafa. Em cada parcela, foram realizados 20 cortes mensais consecutivos representando as medidas repetidas no tempo. Para facilitar a especificação do modelo matemático e das demais propriedades estatísticas, organizou-se a estrutura dos dados do experimento no campo conforme a Tabela 5, onde a parcela ou unidade experimental, que representa cada combinação de tratamento e bloco no qual foram realizados os cortes, foi denominada de indivíduo.

Tabela 5. Estrutura do experimento de alfafa (*Medicago sativa* L.).

Tratamento,Bloco	Indivíduo	Avaliações no indivíduo
1,1	1	$y_{1,1,1}$ $y_{1,1,2}$ ... $y_{1,1,19}$ $y_{1,1,20}$
2,1	2	$y_{2,1,1}$ $y_{2,1,2}$ ... $y_{2,1,19}$ $y_{2,1,20}$
...	...	...
92,1	92	$y_{92,1,1}$ $y_{92,1,2}$ ... $y_{92,1,19}$ $y_{92,1,20}$
1,2	93	$y_{1,2,1}$ $y_{1,2,2}$ ... $y_{1,2,19}$ $y_{1,2,20}$
2,2	94	$y_{2,2,1}$ $y_{2,2,2}$ ... $y_{2,2,19}$ $y_{2,2,20}$
...	...	...
92,2	184	$y_{92,2,1}$ $y_{92,2,2}$ ... $y_{92,2,19}$ $y_{92,2,20}$

Modelo matemático

$$y_{ijk} = \mu + t_i + \delta_{ij} + c_k + (tc)_{ik} + \varepsilon_{ijk}$$

$$(i = 1, 2,\ldots,92; j = 1, 2; k = 1, 2,\ldots,20)$$

em que:

y_{ijk} = valor observado da PMS no corte *k*, no indivíduo *j* e no tratamento *i*.

μ = média global.

t_i = efeito fixo do tratamento i.

δ_{ij} = efeito aleatório do indivíduo j no tratamento i.

c_k = efeito fixo do corte k.

$(tc)_{ik}$ = efeito fixo da interação de tratamento e corte.

ε_{ijk} = erro aleatório do corte k, no indivíduo j e no tratamento i; $\varepsilon_{ijk} \sim IID\ (0, \sigma^2)$.

Na forma matricial, o modelo matemático é descrito por y = Xb + Zu + e, em que o vetor *b* contém os efeitos fixos μ, t_i, c_k e $(tc)_{ik}$; o vetor *u* contém o efeito aleatório de indivíduos (δ_{ij}); o vetor *e* contém os erros associados às avaliações dentro de indivíduos (V(e) = R). Admitindo-se que os três vetores: *y*, *u*, *e* têm distribuição normal, em termos de: esperanças (E), variâncias (Var) e covariâncias (Cov), têm-se:

E(u) = 0; Var(u) = G; E(e) = 0; V(e) = R; Var(y) = V(Zu + e) = ZGZ'+R

O termo δ_{ij} equivale a Blocos + (Interação Tratamentos × blocos). Eles refletem a variação entre os tratamentos para um determinado tempo fixo e são independentes entre si, ou seja, cada parcela com a identificação de um tratamento e um bloco é independente da outra. Por sua vez, os ε_{ijk} na mesma unidade experimental são correlacionados e refletem a variação dentro da unidade experimental que é dada por $V(\varepsilon_{ijk}) = R$. A estrutura de variância e covariância R pode assumir um formato diferente, dependendo do experimento.

O quadro de análise de variância do delineamento de blocos completos com medidas repetidas no tempo, conforme descrito, está apresentado na Tabela 6.

Tabela 6. Análise de variância de blocos completos com medidas repetidas no tempo.

Fator de variação	Grau de liberdade		Soma de quadrados	Quadrado médio	F
Tratamentos – a	a - 1	91	SQT	QMT = SQT/(a - 1)	QMT/QME_a
Blocos(a): erro a	a(r - 1)	92	SQE_a	$QME_a = SQE_a /a(r - 1)$	
Corte – c	(c - 1)	19	SQC	QMC = SQC/(c - 1)	QMC/QMEb
Interação (a x c)	(a -1)(c - 1)	1729	SQI	QMI = SQI/(a - 1)(c -1)	QMI/QME_b
Erro b	a(r - 1)(c - 1)	1748	SQE_b	$QME_b = SQE_b/a(r - 1)(c -1)$	
Total	arc - 1	3679	SQ total	SQ total	

A seguir, há exemplo de análise de variância pelo procedimento Mixed do SAS da Tabela 5 e considerando-se arquivo de dados fictícios “repetida”, em que *y* é o valor observado.

```
data repetida; input bloco a c y;
datalines;
1 1 1  55.0
1 1 2  51.5
...
92 20 53.1
;
proc mixed;
class bloco a c;
model y=bloco a c a*c /sub=bloco(a) type= cs;
/ “type = cs” indica que a estrutura de variância e covariância é a simetria composta */
run;
```

Blocos casualizados com subamostragem

Este desenho é também chamado de desenho aninhado ou desenho hierárquico. Na verdade, é uma extensão do tradicional blocos ao caso (DBC). A diferença é que, no DBC, todo o material da parcela é utilizado para análise, enquanto, neste delineamento, de cada unidade experimental ou parcela, são retiradas b unidades amostrais que correspondem às unidades observacionais.

Considere a situação em que uma parcela de um experimento é composta por determinado número de plantas, sendo que, para análise, é suficiente verificar apenas algumas plantas retiradas ao acaso.

Cite-se um exemplo em que o interesse é instalar um experimento com cinco blocos (r_1 a r_5), quatro tratamentos principais (a_1, a_2, a_3, a_4) sorteados nas parcelas e três amostras (b_1, b_2, b_3) retiradas de cada parcela. O croqui de campo é semelhante ao da Figura 2.

Modelo matemático

$$y_{ijk} = \mu + a_i + r_j + \varepsilon_{j(i)} + \varepsilon_{k(ij)}$$

$$(i = 1, 2, ..., 4; j = 1, ..., 5; k = 1, 2, 3)$$

em que:

y_{ijk} = valor observado na amostra.

μ = média global.

a_i = efeito fixo do tratamento i.

r_j = efeito do bloco j.

$\varepsilon_{j(i)}$ = erro aleatório associado a cada parcela; $\varepsilon_{j(i)} \sim$ NIID $(0, \sigma^2)$.

$\varepsilon_{k(ij)}$ = erro aleatório associado a cada amostra $\varepsilon_{k(ij)} \sim$ NIID $(0, \sigma^2)$.

A análise de variância do delineamento de blocos casualizados com subamostragem está apresentada na Tabela 7.

Tabela 7. Análise de variância de blocos casualizados com subamostragem.

Fator de variação	Grau de liberdade		Soma de quadrados	Quadrado médio	F
Tratamentos – a	a - 1	3	SQT	QMT = SQT/(a - 1)	QMT/QME_a
Blocos – r	r - 1	4	SQB	QMB = SQB/(r - 1)	QMB/QME_a
Erro a	(r - 1)(a - 1)	12	SQE_a	$QME_a = SQE_a/(r - 1)(a -1)$	QME_a/QME_b
Erro b	ar(b -1)	40	SQEb	$QME_b = SQE_n/ar(b -1)$	
Total	rab -1	59	SQ total		

A seguir, há exemplo de um arquivo de dados fictícios, de blocos casualizados com subamostragem e análise de variância pelo procedimento GLM do SAS da Tabela 7, em que *y* é o valor observado.

```
data amostra;
input bloco a b y;
datalines;
1  1   1 55.0
1  1   2 51.5
1  1   3 52.3
...
5  4   3 53.1
;
proc glm;
class bloco a b;
model y = bloco a bloco*a;
test h = a bloco e= bloco*a;
run;
```

Blocos casualizados com k repetições por bloco

Ocorre quando o número de tratamentos é pequeno e, para atingir os requisitos da análise de variância, como no mínimo dez graus de liberdade para o resíduo, são necessários muitos blocos. A alternativa, portanto, é utilizar poucos blocos e usar repetições do tratamento dentro do bloco. Por exemplo, caso se queira testar apenas três tratamentos, seriam necessários pelo menos seis blocos. A alternativa para essa

situação é utilizar três tratamentos, três blocos e três repetições por bloco, conforme croqui de campo apresentado na Figura 5.

Tratamento	Bloco 1	Bloco 2	Bloco 3
1	r_1	r_2	r_3
	r_3	r_3	r_1
	r_2	r_1	r_2
2	r_3	r_1	r_2
	r_1	r_3	r_1
	r_2	r_2	r_3
3	r_3	r_3	r_1
	r_2	r_2	r_2
	r_1	r_1	r_3

Figura 5. Croqui de campo para blocos casualizados de três tratamentos, três blocos e três repetições dos tratamentos dentro de cada bloco.

Modelo matemático

$$y_{ijk} = \mu + t_i + b_j + (tb)_{ij} + \varepsilon_{ijk}$$

$$(i = 1, 2, 3; j = 1, 2, 3; k = 1, 2, 3)$$

em que:

y_{ijk} = valor observado na parcela.

μ = efeito médio global.

t_i = efeito do tratamento i.

b_j = efeito do bloco j.

$(tb)_{ij}$ = efeito da interação entre tratamento e bloco.

ε_{ijk} = erro aleatório associado a cada parcela; $\varepsilon_{ijk} \sim NIID\ (0, \sigma^2)$.

A análise está apresentada na Tabela 8.

Tabela 8. Análise de variância de três tratamentos (a), três blocos (b) e três repetições (r) dos tratamentos dentro de cada bloco.

Fator de variação	Grau de liberdade	Soma de quadrados	Quadrado médio	F
Tratamentos – a	a - 1 = 2	SQT	QMT = SQT/(a - 1)	QMT/QME
Blocos – b	b - 1 = 2	SQB	QMB = SQB/(b - 1)	QMB/QME
Interação a x b	(a - 1)(b - 1) = 4	SQI	QMI = SQI/(b - 1)(a - 1)	QMI/QME
Erro	ab(k - 1) = 18	SQE	QME = SQE/ba(k - 1)	
Total	abk - 1 = 26	SQ total		

Um modelo de análise de variância pelo procedimento GLM do SAS, conforme Tabela 8, para blocos casualizados, considerando três tratamentos (a), três blocos e três repetições (r) dentro do bloco e dados fictícios *y* para o arquivo "repete" é apresentado a seguir:

```
data repete;
input bloco a r y ;
datalines;
1   1 1  32.0
1   1 2  31.2
1  1 3 30.3
...
3  3 3 29.1
 ;
proc glm;
class bloco a r;
model y = bloco a bloco*a bloco*a(r);
test h = a bloco e= bloco*a;
/* testa bloco e a com a interação bloco x a como resíduo */
test h = bloco*a  e = bloco*a(e);
/* testa a interação bloco x a com o resíduo bloco*a(r)  */
run;
```

Blocos incompletos balanceados

Blocos incompletos balanceados (BIB) são delineamentos em que o número de parcelas dentro do bloco é menor que o número de tratamentos a ser testado. Portanto, apenas parte dos tratamentos é avaliada no bloco. No BIB cada par de tratamentos

ocorre o mesmo número de vezes no experimento. Se existem *k* tratamentos a serem instalados e em cada bloco tem ***a*** tratamentos (a < k), então o número de blocos necessários para instalar todos os *k* tratamentos é dado pela fórmula:

$$\binom{k}{a} = \frac{k!}{a!(k-a)!}$$

Por exemplo, se existem cinco tratamentos (a_1, a_2, a_3, a_4, a_5) a serem testados e somente pode-se distribuir três tratamentos por bloco, são necessários dez blocos.

$$\binom{5}{3} = \frac{5!}{3!(5-3)!} = (5.4)/2! = 10 \text{ blocos}$$

Na Figura 6 é apresentado o croqui de campo para esse exemplo e com dados fictícios em cada parcela.

Bloco	Tratamento		
1	$a_1 = 22,5$	$a_3 = 21,2$	$a_4 = 25,5$
2	$a_2 = 21,2$	$a_4 = 22,4$	$a_5 = 22,2$
3	$a_2 = 23,0$	$a_3 = 20,5$	$a_5 = 20,2$
4	$a_3 = 22,8$	$a_4 = 22,2$	$a_5 = 23,5$
5	$a_1 = 20,5$	$a_2 = 19,5$	$a_3 = 19,1$
6	$a_1 = 22,9$	$a_2 = 22,5$	$a_4 = 21,5$
7	$a_1 = 24,0$	$a_4 = 24,5$	$a_5 = 23,5$
8	$a_2 = 22,6$	$a_3 = 21,7$	$a_4 = 20,5$
9	$a_1 = 19,8$	$a_2 = 22,4$	$a_5 = 21,5$
10	$a_1 = 23,5$	$a_3 = 22,5$	$a_5 = 24,5$

Figura 6. Croqui de campo para blocos incompletos balanceados com cinco tratamentos, dez blocos, três tratamentos em cada bloco. Cada par de tratamentos ocorre em três blocos.

Modelo matemático

$$y_{ij} = \mu + a_i + r_j + \varepsilon_{ij}$$

$$(i = 1, 2,... 5; j = 1, 2, ...10)$$

em que:

y_{ij} = valor observado na parcela.

μ = efeito médio global.

a_i = efeito fixo do tratamento i.

r_j = efeito do bloco j.

ε_{ij} = erro associado à parcela $y_{ij;}$ $\varepsilon_{ij} \sim$ NIID $(0, \sigma^2)$.

Várias variáveis estão associadas com o BIB:

k = número de tratamentos (k = 5).

b = número de blocos (b = 10).

r = número de repetições de um tratamento ou o número de vezes que um tratamento ocorre no experimento (r = 6).

a = número de tratamentos por bloco (a = 3), ou seja, a < k.

n = número de observações (n = ba = kr = 30).

λ = número de vezes que cada par de tratamento ocorre no mesmo bloco (λ = 3). É obtido por: $\lambda = r(a - 1)/(k - 1) = 6(3 - 1)/(5 - 1) = 3$.

Por exemplo, o par (a_1, a_3) ocorre junto nos blocos 1, 5, 10. A seguir estão os blocos em que cada par de tratamentos aparece:

Par de tratamento	Bloco
a_1, a_2	5, 6, 9
a_1, a_3	1, 5, 10
a_1, a_4	1, 6, 7
a_1, a_5	7, 9, 10
a_2, a_3	3, 5, 8
a_2, a_4	2, 6, 8
a_2, a_5	2, 3, 9
a_3, a_4	1, 4, 8
a_3, a_5	3, 4, 10
a_4, a_5	2, 4, 7

De modo geral, a análise dos experimentos em blocos incompletos é bem mais complexa do que a dos blocos completos. Embora, do ponto de vista experimental, não

haja interesse em comparar blocos, dependendo do tipo do experimento há necessidade de três tipos de análises:

a) Análise intrabloco – considera-se bloco como o efeito fixo. São feitas comparações entre as parcelas dentro do mesmo bloco, as quais são usadas nas estimativas de efeito de tratamentos que são calculadas dentro de cada bloco.

b) Análise interbloco – considera o efeito de bloco como aleatório.

c) Análise com recuperação da informação interbloco – além das comparações realizadas dentro de cada bloco, ela considera as comparações entre os blocos para a estimação dos efeitos de tratamentos. Uma revisão abrangente sobre esse tópico é encontrada em Trevisol e Cosme (2013).

A análise de variância de blocos incompletos balanceados está apresentada na Tabela 9, em que a soma de quadrados para tratamentos ajustados (SQT_{aj}) é dada. Em outras situações experimentais, podem-se incluir na análise:

a) SQ tratamentos ajustados (SQT_{aj}) e não ajustados (SQT).

b) SQ blocos ajustados (SQB_{aj}) e não ajustados (SQB).

Tabela 9. Análise de variância de blocos incompletos balanceados com *k* repetições por bloco.

Fator de variação	Grau de liberdade	Soma de quadrados	Quadrado médio	F
Tratamentos – k	k - 1	SQT_{aj}	$QMT_{aj} = SQT/(k - 1)$	QMT_{aj}/QME
Blocos – b	b - 1	SQB	QMB = SQB/(b - 1)	QMB/QME
Erro	k(b - 1) - b + 1	SQE	QME = SQE/[k(b - 1) - b + 1]	
Total	bk – 1	SQ total		

A seguir encontra-se a rotina SAS para análise de variância pelo procedimento GLM do SAS para o BIB descrito na Figura 6:

```
data bib;
input bloco a y @@;
/*a = tratamentos; y = variável resposta */
datalines;
1 1 22.5 1 5 20.5 1 6 22.9 1 7 24.0 1 9 19.8 1 10 23.5
2 2 21.2 2 3 23.0 2 8 22.6 2 5 19.5 2 6 22.5 2 9 22.4
```

```
...
5  2    22.2    5  3  20.2   5  4   23.5   5   7   23.5  5    9    21.5   5  10   24.59
;
proc glm;
class bloco a;
model y = bloco  a;
lsmeans a/pdiff=all adjust=tukey cl stderr;
/*realiza todas comparações pareadas das medias de tratamentos  pelo teste de tukey*/
contrast '1 vs 2' a 1 -1 0 0 0;
estimate '1 vs 2' a 1 -1 0 0 0;
/* calcula intervalo de confiança entre tratamento 1 e 2*/;
run;
```

Exercícios[10]

1) No texto a seguir, existem afirmações incorretas quanto a conceitos de estatística. Reescreva o texto colocando definições corretas e sublinhe ou coloque em negrito onde houve definições incorretas.

No delineamento tradicional em blocos casualizados (BC), que é um dos mais utilizados em agricultura, uma exigência é que o ambiente seja o mais homogêneo possível dentro dos blocos e também entre os blocos. Com essa condição, não há necessidade e nem é recomendável sortear os tratamentos dentro do bloco, pois a aleatoriedade das parcelas ou unidades experimentais já está garantida e consequentemente a independência entre os erros. No DBC, mesmo que os blocos sejam grandes, os tratamentos distribuídos às parcelas não podem ser organizados na forma fatorial e somente de um fator. Outra vantagem do delineamento em DBC é que qualquer que seja a natureza do bloco, ele sempre será mais eficiente do que o delineamento inteiramente casualizado. No delineamento em DBC com parcela dividida (*split-plot*), que somente é utilizado em agricultura, a parcela principal é dividida em subunidades chamada de subparcela. No *split-plot*, na parcela é distribuído o tratamento principal A, enquanto a subparcela recebe o tratamento secundário B. Na análise de variância deste delineamento, o quadrado médio residual, formado pela interação tratamentos × blocos (erro a) é utilizado para testar ambos os tratamentos (A e B).

2) Na Tabela 10 tem-se a descrição dos resultados de um experimento.

[10] As respostas dos exercícios podem ser consultadas no Apêndice 1.

Tabela 10. Estimativas da média e dos erros-padrão da produtividade de matéria seca, em kg ha^{-1}, de capim "*coast-cross*" (*Cynodon dactylon* L.) obtidas por quadrados mínimos, considerando-se os efeitos principais de duas fontes de N (1 − ureia; e 2 − nitrato de amônio), cinco doses de N (0 kg ha^{-1}, 25 kg ha^{-1}, 50 kg ha^{-1}, 100 kg ha^{-1} e 200 kg ha^{-1}), cinco cortes (1 a 5) e dois anos agrícolas (1998−1999, 1999−2000).

Efeito	Nível	Ano: 1998−1999	Ano: 1999−2000
		$\overline{X} \pm s(\overline{X})$	$\overline{X} \pm s(\overline{X})$
Fontes	1	2.147,9 ± 34,93	1.704,9 ± 35,55
	2	2.511,6 ± 34,93	1.931,7 ± 35,55
Doses	0	810,1 ± 55,24	520,9 ± 56,22
	25	1.606,2 ± 55,24	1.162,7 ± 56,22
	50	2.426,9 ± 55,24	1.932,3 ± 56,22
	100	3.193,1 ± 55,24	2.630,9 ± 56,22
	200	3.612,4 ± 55,24	2.844,7 ± 56,22
Cortes	1	913,1 ± 57,67	1.329,3 ± 45,68
	2	2.796,5 ± 43,71	1.769,9 ± 45,68
	3	1.688,8 ± 47,68	2.347,0 ± 45,68
	4	4.070,9 ± 53,29	1.411,8 ± 45,68
	5	2.179,4 ± 43,02	2.233,4 ± 45,68

($P < 0,0001$) entre níveis dentro de efeitos principais

Fonte: Corrêa et al. (2007).

Faça uma interpretação dos resultados.

3) Em um experimento de blocos casualizados, com parcelas divididas (*split-plot*), tem-se cinco repetições, os tratamentos na parcela principal são quatro doses de nitrogênio (0 kg ha^{-1}, 20 kg ha^{-1}, 40 kg ha^{-1} e 60 kg ha^{-1}) e na subparcela são avaliadas três variedades de alfafa (*Medicago sativa*). O resumo da análise de variância está apresentado na Tabela 11.

Tabela 11. Análise de variância de blocos casualizados com parcelas divididas (*split-plot*).

Fator de variação	Grau de liberdade	
Doses – a	a - 1	3
Blocos – r	r - 1	4
Erro a	(r - 1)(a - 1)	12
Variedades – b	(b - 1)	2
Interação a*b	(a - 1)(b - 1)	6
Erro b	a(r - 1)(b - 1)	31
Total	abr -1	58

Assinale a alternativa correta:

( ) Houve perda de uma parcela.

( ) Houve perda de uma subparcela.

( ) Houve perda de duas subparcelas.

( ) Não houve perda de parcela e nem de subparcela.

4) Num delineamento de blocos incompletos balanceados (BIB), deseja instalar seis tratamentos e em cada bloco só é possível colocar cinco. Quantos blocos serão necessários para instalar os seis tratamentos?

5) Considerando-se a questão 4, descreva como ficaria o croqui de campo.

Capítulo 11

Delineamento em quadrado latino, greco-latino e Youden

Introdução

Os delineamentos em quadrado latino, quadrado greco-latino e quadrado de Youden são bastante úteis em várias situações experimentais e, principalmente, na agricultura. O termo "quadrado latino" é atribuído à Leonhard Euler (1707−1783), o qual utilizou caracteres latinos como símbolos dos desenhos experimentais. Nesse desenho, as unidades experimentais ou parcelas em que os tratamentos são aplicados possuem dois controles locais (linhas e colunas), isto é, há dois tipos de blocos. Isso significa que, dentro de linhas e dentro de colunas, as condições precisam ser as mais homogêneas possíveis (Wikipédia, 2019g).

Em geral, em cada linha é alocado um fator, como pessoa, animal, máquina, etc., que recebem os tratamentos; e cada coluna representa o tempo, ocasião ou período em que os tratamentos são avaliados. Entretanto, dependendo das necessidades experimentais, os mais diferentes fatores, tais como pessoa, período do ano, local, máquina, podem ser distribuídos nas linhas e colunas.

Em agricultura, como exemplos de dois controles locais, podem-se citar: diferenças entre leitegadas e variabilidade de peso inicial; luz e temperatura em casa de vegetação; gradientes de fertilidade do solo no sentido vertical e horizontal; provadores e ordem de degustação em avaliações sensoriais, etc.

Em geral, o fator distribuído na linha representa o indivíduo ou sujeito que receberá o tratamento, enquanto cada coluna representa o período em que o tratamento é avaliado. É importante salientar que o delineamento é apropriado para situações em que as respostas são de fluxo continuado, isto é, nas quais o mesmo indivíduo (pessoa, animal, máquina, etc.) continua nos diversos períodos. Esse delineamento não pode ser utilizado onde o animal precisa ser sacrificado, onde a máquina precisa ser descartada, etc., após a sua avaliação. É preciso observar que o período entre um tratamento e o subsequente deve ser cuidadosamente planejado, de modo que não exista efeito residual do tratamento anterior.

Neste capítulo, são apresentados conceitos, tipos, croquis de campo, modelos matemáticos, etc., do delineamento em quadrado latino, quadrado greco-latino e quadrado de Youden. Esses delineamentos são importantes quando se deseja instalar experimentos em situações de restrição de animais, equipamentos, etc. A pouca disponibilidade desses insumos é compensada pela repetição no tempo, de modo que o experimento pode ser perfeitamente realizado.

Quadrado latino

O desenho em quadrado latino (QL) possui número igual de linhas, de colunas e de tratamentos, com os *t* tratamentos distribuídos em *t* linhas e *t* colunas, surgindo daí a denominação QL t x t, por exemplo, QL do tipo 3 x 3, 4 x 4, 5 x 5, 6 x 6, 7 x 7, 8 x 8, 9 x 9, os quais atendem à maioria das necessidades do pesquisador.

O duplo controle experimental (linhas e colunas) requerido pelo quadrado latino acarreta decréscimo no número de graus de liberdade (GL) para o resíduo na análise de variância, o que é indesejável. Para atender a esse requisito, é importante escolher um desenho do tamanho 5 x 5 ou maior. Entretanto, se as necessidades do pesquisador são atendidas por um desenho de tamanho 3 x 3 ou 4 x 4, algumas modificações experimentais precisam ser incorporadas, tais como, aumentar o número de indivíduos (linhas), considerar três QL 3 x 3 ou dois QL 4 x 4, além de outras modificações no desenho de campo.

No delineamento em QL, cada tratamento é distribuído aleatoriamente uma vez na linha e uma vez na coluna. A maioria dos experimentos pode ser instalada em delineamentos de tamanho 5 x 5 a 9 x 9. Na Figura 1 a seguir são apresentados alguns croquis de desenhos no campo de QL de tamanho 3 x 3 até 9 x 9, com os respectivos sorteios da distribuição dos tratamentos (A até I) nas parcelas, porém o pesquisador poderá optar por novas alternativas de aleatorização dos tratamentos e também escolher QL maiores.

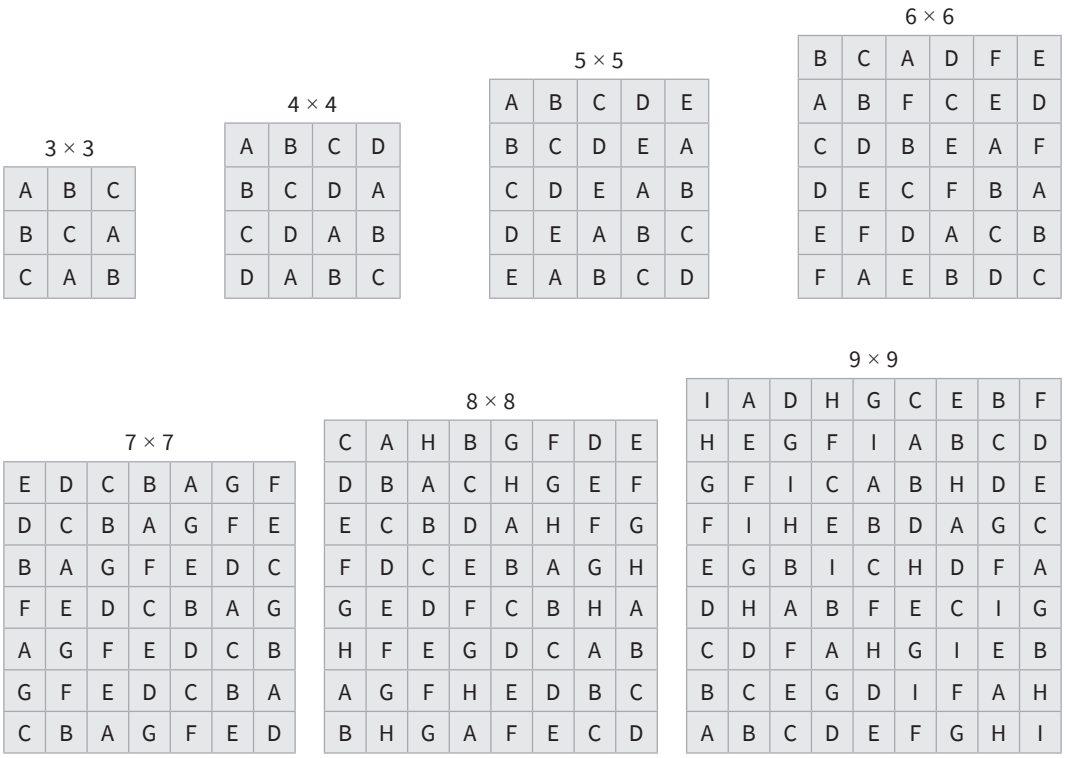


Figura 1. Croquis de campo – delineamento quadrado latino.

Análise de variância

Modelo matemático

$$y_{ijk} = \mu + \alpha_i + \tau_j + \beta_k + \varepsilon_{ijk}$$

$$(i = 1,..., t; j = 1,..., t; k = 1,..., t)$$

em que:

y_{ijk} = variável resposta.

μ = efeito médio global.

α_i = i-ésimo efeito do bloco 1 (efeito de linha).

τ_j = j-ésimo efeito do tratamento.

β_k = k-ésimo efeito do bloco 2 (efeito coluna).

ε_{ijk} = erro aleatório ~ N (0, σ^2) (normalidade, independência, variância constante).

A análise de variância para o delineamento em quadrado latino é apresentada na Tabela 1.

Tabela 1. Análise de variância para o delineamento em quadrado latino.

Fator de variação	Grau de liberdade	Soma de quadrados	Quadrado médio	F
Linhas – l	t - 1	SQL	QML = SQL/(t - 1)	QML/QME
Colunas – c	t - 1	SQC	QMC = SQC/(t - 1)	QMC/QME
Tratamentos – t	t - 1	SQT	QMT = SQT/(t - 1)	QMT/QME
Erro	(t - 1) (t - 2)	SQE	QME = SQE/(t - 1)(t -2)	
Total	t^2 - 1	SQTo		

Aplicação

Na Tabela 2, é apresentado um croqui de campo de um delineamento em QL 5 x 5 em que cinco diferentes animais, na linha, foram avaliados em cinco períodos, na coluna. Em cada coluna, o animal recebe um dos cinco tratamentos (A a E). A resposta por animal em cada período e tratamento são dados fictícios.

Tabela 2. Croqui de campo – delineamento quadrado latino 5 x 5.

Animal	Período				
	1	2	3	4	5
1	B = 164	D = 161	A = 149	C = 157	E = 164
2	E = 168	B = 171	D = 162	A = 152	C = 152
3	C = 153	E = 152	B = 172	D = 163	A = 153
4	D = 160	A = 148	C = 159	E = 160	B = 171
5	A = 145	C = 155	E = 145	B = 169	D = 153

A análise de variância foi realizada pelo procedimento modelo linear geral (GLM). A opção "*lsmeans trt*" estimou as médias dos tratamentos por quadrados mínimos; em seguida, foi utilizado o teste de Tukey (*adjust = tukey*) para comparação pareada entre todas as médias dos tratamentos (*pdiff = all*) e finalmente foi calculado o erro-padrão da média (*stderr*).

```
data  ql5x5;
input animal periodo trat $ y @@;
cards;
1  1 B 164   1 2  D  161  1  3  A  149  1 4   C 157  1  5  E  164
2  1 E 168   2 2  B  171   2  3  D  162  2 4  A 152  2  5  C  152
3  1 C 153   3 2  E  152   3  3  B  172  3 4  D 163  3  5  A  153
4  1 D 160   4 2  A 148   4  3  C   159 4  4  E 160  4  5  B  171
5  1 A 145   5 2  C 155   5  3  E   145  5 4  B 169  5  5  D  153
;
proc glm;
class animal periodo trat;
model y = trat animal periodo;
means trat / lines tukey;
/* medias sem erro-padrão e teste de tukey*/
lsmeans trat / adjust = tukey pdiff = all stderr;
/* medias com erro-padrão e teste de tukey*/
run;
```

Os resultados da análise de variância na Tabela 3 indicam que os tratamentos foram altamente significativos (P < 0,0009), o que pode ser visto nas comparações pareadas pelo teste de Tukey na Tabela 4.

Tabela 3. Análise de variância.

Fator de variação	Grau de liberdade	Soma de Quadrados	Quadrado médio	F	Pr > F
Linhas	4	167,84	41,96	1,54	0,2539
Colunas	4	27,04	6,76	0,25	0,9058
Tratamentos	4	1.072,64	268,16	9,81	0,0009
Erro	12	327,92	27,33		
Total	24	1.595,44			

Tabela 4. Médias e erros-padrão estimados por quadrados mínimos.

Tratamento	Média ± erro-padrão
A	149,40 ± 2,33b
B	169,40 ± 2,33a
C	155,20 ± 2,33b
D	159,80 ± 2,33ab
E	157,80 ± 2,33b

Letras diferentes na coluna indicam significância estatística (P < 0,05) pelo teste de Tukey.

Quadrado latino com desbalanceamento nos dados

Como o mesmo indivíduo é avaliado em vários períodos, para este tipo de delineamento, somente é tolerável a perda de parcela (PP) no último período, pois se a PP ocorrer no meio do experimento, devido à morte de um animal, por exemplo, o experimento seria inviabilizado. Caso haja PP no último período, a análise de variância pode ser efetuada sem problemas pelo procedimento GLM. Dos quatro tipos de Somas de Quadrados (SQ) efetuados por este procedimento (I, II, III, IV), deve-se utilizar a SQ do tipo III para todos os tipos de hipóteses.

Outros recursos do delineamento em quadrado latino

Seja a situação em que o pesquisador deseja trabalhar com um QL 4 x 4. Nesse caso, ele teria apenas seis graus de liberdade para o resíduo e não atenderia a umas das principais exigências da análise de variância, isto é, ter pelo menos 10 GL para o resíduo. Duas alternativas para atender a esse requisito são sugeridas nos croquis de campo da Figura 2, em que se considera animais nas linhas e períodos de avaliação (P1 a P8) nas colunas.

a) Utilizar simultaneamente dois QL 4 x 4, sendo que para isso bastaria ter oito animais.

b) Caso tenha disponibilidade de apenas quatro animais, somente poderia instalar um QL 4 x 4. Nesse caso, o segundo QL poderia ser instalado em outra época, sendo que os mesmos animais poderiam ser utilizados. O mesmo raciocínio pode ser utilizado para outros QL.

a)

Animal	P1	P2	P3	P4
1	D	A	B	C
2	C	D	A	B
3	B	C	D	A
4	A	B	C	D

Animal	P1	P2	P3	P4
5	D	A	B	C
6	A	B	C	D
7	B	C	D	A
8	C	D	A	B

b)

Animal	P1	P2	P3	P4
1	D	A	B	C
2	C	D	A	B
3	B	C	D	A
4	A	B	C	D

Animal	P5	P6	P7	P8
1	D	A	B	C
2	A	B	C	D
3	B	C	D	A
4	C	D	A	B

Figura 2. Croqui de campo de delineamento quadrado latino 4 x 4.

As análises de variância dos dois exemplos de delineamento quadrado latino 4 x 4 estão na Tabela 5.

Tabela 5. Análise de variância de delineamento em quadrado latino 4 x 4.

a)		b)	
FV	**GL**	**FV**	**GL**
Animais	7	Animais	3
Períodos	3	Períodos (QL)	6
Tratamentos	3	Tratamentos	3
Erro	18	Erro	19
Total	31	**Total**	31

FV: fator de variação; GL: grau de liberdade.

Quadrado greco-latino

É o tipo de delineamento que equivale a dois experimentos de quadrado latino (QL) realizados simultaneamente ou superpostos. Considerando-se o mesmo exemplo do QL 5 x 5 da Tabela 2, em que cinco diferentes animais, na linha, são avaliados em

cinco períodos, na coluna, a diferença é que o quadrado greco-latino (Tabela 6) tem mais um controle local, que é representado pelas letras gregas (α, β, γ, δ, ϵ).

Cada tratamento (A, ... ,E) aparece uma única vez em cada coordenada (linha, coluna) junto com uma letra grega. As respostas são dados fictícios e a análise de variância, pela rotina SAS, é apresentada na Tabela 7.

Tabela 6. Esquema de delineamento quadrado greco-latino 5 x 5.

Linha	Coluna				
	1	2	3	4	5
1	Aα = 15	Bϵ = 16	Cβ = 18	Dγ = 23	Eδ = 16
2	Bβ = 17	Cγ = 20	Dδ = 20	Eα = 15	Aϵ = 15
3	Cδ = 19	Dα = 21	Eϵ = 14	Aβ = 13	Bγ = 16
4	Dϵ = 22	Eβ = 15	Aγ = 14	Bδ = 17	Cβ = 17
5	Eγ = 14	Aδ = 16	Bα = 18	Cϵ = 19	Dα = 21

Modelo matemático considerando um quadrado greco-latino 5 x 5

$$y_{ijkl} = \mu + \alpha_i + \tau_j + \beta_k + \omega_l + \varepsilon_{ijkl}$$

$$(i = 1,..., t; j = 1,..., t; k = 1,..., t; l = 1,..., t)$$

em que:

y_{ijkl} = variável resposta.

μ = efeito médio global.

α_i = i-ésimo efeito do bloco 1 (efeito de linha).

τ_j = j-ésimo efeito do tratamento.

β_k = k-ésimo efeito do bloco 2 (efeito coluna).

ω_l = l-ésimo efeito do bloco 3 (efeito letra grega).

ε_{ijkl} = erro aleatório ~ N (0, σ^2) (normalidade, independência, variância constante).

Código SAS

```
data;
input animal periodo trat $  grega $ y @@;
cards;
1  1 a  α   15   1  2  b  ϵ  16   1  3  c  β  18   1  4  d  γ  23  1  5  e  δ  16
...
5  1 e  γ   14   5  2  a  δ  16   5  3  b  α  18   5  4  c  ϵ  19  5  5 d  α  21
;
proc glm;
class animal periodo trat grega;
model y = animal period trat grega;
means trat / lines tukey;
/* medias sem erro-padrão e teste de tukey*/
lsmeans trat / adjust = tukey pdiff = all stderr;
/* medias com erro-padrão e teste de tukey*/
run;
```

Tabela 7. Quadrado greco-latino 5 x 5 – Análise de variância.

Fatores de variação	Graus de liberdade	Somas de quadrados	Quadrado médio	F
Linhas	t - 1 = 4	SQL	QML = SQL/(t - 1)	QMT/QME
Tratamentos	t - 1 = 4	SQT	QMT = SQT/(t - 1)	
Colunas	t - 1 = 4	SQC	QMC = SQC/(t - 1)	
Letras gregas	t - 1 = 4	SQG	QMG = SQG/(t - 1)	
Erro	(t - 1)(t - 3) = 8	SQE	QME = SQR/(t - 1)(t - 3)	
Total	t^2 - 1 24			

Quadrado de Youden

O quadrado de Youden é considerado um QL incompleto em que o número de linhas (l) é igual ao número de tratamentos (l = t) e o número de colunas (c) é menor. Cada tratamento ocorre somente uma vez em cada coluna e eles são balanceados através das linhas.

Nos anos 1930, esse delineamento foi introduzido por Willian John Youden (Wikipédia, 2019a) para experimentação em plantas. Porém, ele pode ser utilizado em várias áreas da pesquisa. Na experimentação animal, uma situação típica são os experimentos de teste de rações de vacas em lactação onde o principal interesse é

avaliar a produtividade de leite. O período de lactação de uma vaca tem duração de 305 dias, porém a produtividade de leite é bastante instável com coeficiente de variação muito alto. O período de maior estabilidade inicia após o pico de lactação (30 a 60 dias após o parto) até a metade da gestação seguinte, que ocorre de três a cinco meses.

Considerando-se que, em um teste de avaliação de ração, a adaptação do animal requer de uma a duas semanas, e admitindo-se esse mesmo período para avaliar o efeito da ração, então haveria condições de avaliar, no máximo, quatro rações por vaca. Como a disponibilidade de vacas geralmente não é problema, o quadrado de Youden seria uma opção excelente, pois a limitação é o número de colunas, não o número de linhas. Na Figura 3, é apresentado o croqui de campo com cinco vacas (linhas), cinco tratamentos (A, B, C, D, E) e quatro períodos para esse delineamento (colunas). A análise de variância encontra-se na Tabela 8.

Vaca	Período			
	1	2	3	4
1	D	E	A	C
2	E	A	B	D
3	B	C	D	A
4	A	B	C	E
5	C	D	E	B

Figura 3. Croquis de campo de um delineamento quadrado de Youden.

No quadrado de Youden, cada par de tratamentos ocorre juntos λ vezes:

$$\lambda = \tau(c - 1)/(t - 1)$$

em que:

τ = número de vezes que o tratamento ocorre no experimento.

c = número de colunas.

t = número de tratamentos.

No croqui da Figura 3, tem-se:

$$\lambda = 4(4 - 1)/(5 - 1) = 3$$

Cada par de tratamentos ocorre três vezes em cada vaca ou linha, o que pode ser visto a seguir:

AB (2, 3, 4), AC (1, 3, 4), AD (1, 2, 3), BC (3, 4, 5), BD (2, 3, 5), BE (2, 4, 5), CD (1, 3, 5), DE (1, 2, 5).

Análise de variância

Modelo matemático

$$y_{ijk} = \mu + t_i + l_j + c_k + \varepsilon_{ijk}$$

em que:

y_{ijk} = variável resposta.

μ = efeito médio global.

t_i, l_j, c_k = i-ésimo, j-ésimo e k-ésimo efeito, respectivamente, de tratamento, linha e de coluna.

ε_{ijk} = erro aleatório.

Tabela 8. Análise variância de um quadrado de Youden 5 × 4.

Fator de variação	Grau de liberdade
Tratamentos	4
Linhas	4
Colunas	3
Erro	8
Total	19

Exercícios[11]

1) No texto a seguir, existem afirmações incorretas quanto a conceitos de estatística. Reescreva o texto colocando definições corretas e sublinhe ou coloque em negrito onde houve definições incorretas.

O delineamento em quadrado latino (QL) é bastante útil em várias áreas, tais como medicina, psicologia, engenharia, agricultura. Em geral, o fator distribuído na linha é o

[11] As respostas dos exercícios podem ser consultadas no Apêndice 1.

sujeito que receberá o tratamento, enquanto as colunas representam o tempo, ocasião, período em que os tratamentos são avaliados. Nesse desenho, as unidades experimentais ou parcelas onde os tratamentos são aplicados possuem dois controles locais – linhas e colunas. Porém, dentro das linhas, há maior rigor no controle experimental do que dentro das colunas. Dentre as desvantagens dos delineamentos em QL, tem-se o fato de que o tipo de sorteio de um experimento é único e também o fato de que não se pode calcular efeito de interação entre linhas e colunas e tratamentos. O desenho em quadrado latino possui número igual de linhas, de colunas e de tratamentos, com os *t* tratamentos distribuídos em *t* linhas e *t* colunas, surgindo daí a denominação QL t x t, por exemplo, QL do tipo 3 x 3, 4 x 4, 5 x 5, 6 x 6, 7 x 7, 8 x 8, 9 x 9. Desde que se utilize corretamente o sorteio em um experimento em um QL, qualquer tipo de tamanho, desde QI 3 x 3 até QL 9 × 9 atende, sem restrição, as necessidades do pesquisador.

2) Se um pesquisador está utilizando um delineamento em QL 5 x 5, em que cinco diferentes animais na linha são avaliados em cinco períodos, e em cada período o animal recebe um dos cinco tratamentos (A a E). Como fica a situação se um animal morre no quarto período?

3) Um pesquisador tem apenas quatro animais e deseja instalar três experimentos QL 4 x 4, isto é, ele vai realizar o experimento em três épocas. Como fica o quadro de análise de variância?

4) Uma análise de variância de experimento QL proporcionou os resultados apresentados na Tabela 9.

Tabela 9. Análise de variância de um delineamento quadrado latino.

Fator de variação	Grau de liberdade
Linhas: 1	5
Colunas: c	5
Tratamentos: t	5
Erro	19
Total	34

Assinale a alternativa correta para o delineamento que proporcionou a análise de variância da Tabela 9:

( ) QL 6 x 6.

( ) QL 6 x 6 com perda de uma parcela.

( ) QL 5 x 5.

( ) Quadrado de Youden 6 x 5.

( ) Nenhuma das anteriores.

5) A análise de variância da Tabela 10 apresenta erros de cálculos para Pr > F. Se a função *probf(x, n_1, n_2),* em que para F calculado (x), com grau de liberdade do numerador (n_1) e grau de liberdade do denominador (n_2), retorna a probabilidade exata de F, determine esses valores.

Tabela 10. Análise de variância.

Fator de variação	Grau de liberdade	Soma de quadrados	Quadrado médio	F	Pr > F
Linhas	4	167,84	41,96	1,54	0,2439
Colunas	4	27,04	6,76	0,25	0,8758
Tratamentos	4	1.072,64	268,16	9,81	0,0050
Erro	12	327,92	27,33		
Total	24	1.595,44			

Capítulo 12

Correlação, regressão linear e covariância

Introdução

Correlação, regressão linear e covariância são técnicas estatísticas bastante utilizadas em aplicações práticas, tais como na interpretação de resultados experimentais.

A correlação estuda a dependência entre duas quantidades, como o parentesco estatístico entre duas variáveis, sem, no entanto, diferenciar variáveis dependentes e independentes.

A regressão linear, simples ou múltipla, em estatística é uma equação para estimar o valor condicional ou esperado de uma variável *y* em função de variáveis *x*. É de fácil aplicação e os modelos são frequentemente ajustados usando a abordagem dos mínimos quadrados.

A covariância ou variância conjunta é uma medida da variabilidade conjunta de duas variáveis aleatórias. A correlação e a covariância indicam se as variáveis são positivamente ou negativamente correlacionadas. Em qualquer tipo de estudo, os valores obtidos e significâncias estatísticas dessas técnicas são dependentes do tamanho amostral.

Neste capítulo, são apresentadas as estatísticas − correlação, regressão linear e covariância − no estudo da relação linear entre duas variáveis. Essas estatísticas estão diretamente associadas aos Capítulos 7 a 11 do livro.

Correlação

Quando se deseja saber apenas o parentesco da relação linear entre duas variáveis aleatórias *x* e *y* obtidas de uma amostra de tamanho *n*, principalmente quando elas são medidas em um mesmo indivíduo, a estimativa do coeficiente de correlação r ($-1 \leq r \leq 1$) é importante. Nesse caso, as duas variáveis precisam estar apenas associadas, sem uma relação entre causa e efeito e não há preocupação com as unidades de medida entre elas. Existem vários tipos de correlação, mas, neste capítulo, utilizaremos o coeficiente de correlação de Pearson (r), que é utilizado quando as duas variáveis são contínuas e intervalares ou de razão. É estimado por:

$$r = \frac{\sum_{1=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{1=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{1=1}^{n}(y_i - \overline{y})^2}}$$

Interpretação:

a) r = -1 → há correlação negativa perfeita entre *x* e *y*.

b) r = 1→ há correlação positiva perfeita entre *x* e *y*.

c) r = 0 → *x* e *y* são independentes entre si.

d) -1 < r < 0 → correlação parcialmente negativa.

e) 0 < r < 1 → correlação parcialmente positiva.

Testes de hipóteses do coeficiente de correlação r

Como o coeficiente de correlação *r* é calculado com base em "n" pares de dados de amostras aleatórias, ele é apenas uma estimativa do coeficiente de correlação populacional ρ, que é obtido se tivéssemos todos os pares *x*, *y* da população. Nesse caso, uma estimativa de r ≠ 0 não é garantia de que ρ ≠ 0, e estamos diante de um problema de inferência. Assim, resolve-se o problema aplicando-se um teste de hipótese para verificar se o valor de *r* é coerente com o tamanho da amostra *n*, a um nível de significância α.

Para isso, testa-se a hipótese de nulidade H_0: ρ = 0 versus três possíveis hipóteses alternativas (Ha):

Ha: ρ > 0 (teste unilateral à direita).

Ha: ρ < 0 (teste unilateral à esquerda).

Ha: ρ ≠ 0 (teste bilateral).

Para aceitar e ou rejeitar uma das hipóteses acima para uma amostra de tamanho *n*, calcula-se o teste t (t_c) abaixo, que corresponde a uma distribuição "t" de Student com n-2 graus de liberdade.

$$t_c = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{r}{s_r}$$

em que: $s_r = \frac{\sqrt{n-2}}{\sqrt{1-r^2}}$ é o erro-padrão do coeficiente de correlação *r*.

Passos para realizar o teste de hipótese:

a) Calcular t_c.

b) Para uma amostra de *n* dados, consultar o valor de *t* na tabela t de Student (Tabela 3) com n - 2 graus de liberdade (GL) e um valor crítico de α para teste unilateral e α/2 para bilateral; geralmente é escolhido α = 0,05.

c) Para teste unilateral à direita, rejeita-se H_0 se $t_{c>} t_{(gl,\alpha)}$.

d) Para teste unilateral à esquerda, rejeita-se H_0 se $t_{c<t(gl,\alpha)}$.

e) Para teste bilateral, rejeita-se Ho se $t_{c<-t(gl,\alpha/2)}$ ou $t_{c>} t_{(gl,\alpha/2)}$.

Intervalo de confiança para *r*

Como o coeficiente de correlação *r* é calculado com base em n pares de dados de uma amostra aleatória, ele é uma estimativa do coeficiente de correlação populacional ρ entre as variáveis aleatórias *x* e *y*, com médias esperadas μ_x e μ_y e desvios-padrão σ_x e $\sigma_{Y,}$ respectivamente.

$$\rho = \frac{\sum_{l=l}^{n}(x_i - \mu_x)(y_i - \mu_y)}{\sigma_X \sigma_Y}$$

Assim, ao invés de realizar testes de hipóteses para *r*, a alternativa mais eficiente é calcular o intervalo de confiança (IC). Como o valor de *r* é obtido de uma amostra de n pares de dados, geralmente a sua distribuição não é normal. Para garantir que *r* seja normalmente distribuído, utiliza-se a transformação *z* proposta por Fisher (1921):

$z = 1/2\ln\left[\frac{1+r}{1-r}\right]$, com desvio-padrão $s_z = 1/\sqrt{n-3}$, sendo *n* o tamanho amostral.

Com essa transformação, o coeficiente de correlação *r* tem distribuição normal ou aproximadamente normal. Inicialmente calcula-se o IC em termos de *z*, cujo limite inferior e superior para um nível α de probabilidade fica:

- Limite inferior (LI): $z - z_{(1-\alpha/2)}1/\sqrt{n-3}$
- Limite superior (LS): $z + z_{(1-\alpha/2)}1/\sqrt{n-3}$

Esses cálculos do LI e LS estão em termos de *z*; porém é necessário fazer a conversão de *z* para *r*.

Na rotina do Sistema de Análise Estatística (do inglês Statistical Analysis System – SAS), a seguir, em que se considera o arquivo de dados fictícios "corre" com cinco variáveis (y, x_1, x_2, x_3, x_4), são calculadas as estatísticas (Tabela 1): a) correlação de *y* com cada variável *x* sem a transformação *z* de Fisher; b) transformação *z* de Fisher; c) ajuste de

r após a transformação de Fisher; d) correlação *r* obtida após a transformação *z* de Fisher (r_{aju}); e) intervalo de confiança com 95% de probabilidade de *r* após a transformação *z* de Fisher; f) teste de hipóteses: H_0: ρ = 0 versus H_a: ρ ≠ 0.

```
data corre;
input y x1 x2 x3 x4;
cards;
10.0   2.0  5.0  8.0   3.0
10.5   2.2  4.5  7.0   3.0
11.0   2.4  4.3  8.5   3.9
12.7   3.0  4.0  3.0   3.7
12.9   3.4  3.9  12.5  4.0
13.0   3.8  3.2  10.5  4.2
13.8   4.2  3.0  0.0   4.3
14.0   4.7  2.8  7.0   5.0
14.6   5.0  2.6  8.9   4.9
15.2   5.3  2.0  9.0   5.7;
proc corr data = corre nosimple fisher;
/* nosimple = não imprime estatisticas descritivas */
/* Fisher = calcula transformação de Fisher */
 var y; with x1 - x4;
 run;
```

Tabela 1. Coeficiente de correlação de *y* com cada variável X (r), *z* de Fisher, ajuste de *r* após transformação de Fisher, correlação ajustada (r_{aju}), intervalo de confiança (IC) e testes de hipóteses: H_0: ρ = 0 versus H_a: ρ ≠ 0.

	r	z de Fisher	Ajuste de r	r_{aju}	IC com 95% de probabilidade		Teste de hipóteses
Y com X_1	0,9768	2,2222	0,0610	0,9738	0,8766	0,9946	<0,0001
Y com X_2	-0,9791	-2,2746	-0,0699	-0,9760	-0,9957	-0,8688	<0,0001
Y com X_3	-0,0985	-0,0988	0,0062	-0,0924	-0,7128	0,6090	0,8086
Y com X_4	0,3051	0,3152	0,0191	0,2878	-0,4652	0,7991	0,4401

Observa-se correlação positiva perfeita entre *y* e x_1; correlação negativa perfeita entre *y* e x_2, correlação praticamente nula entre *y* e x_3 e correlação parcialmente positiva entre *y* e x_4. Quanto ao teste de hipótese: H_0: ρ = 0 versus H_a: ρ ≠ 0 para as correlações *y* e x_3 e y e x_4, não rejeitou H_0 a nível de 5% de probabilidade, indicando que não é possível afirmar que elas diferem de zero.

Regressão linear simples

Técnica estatística que tem por objetivo estimar y_i em função de x_i, ou seja, quando existe uma relação linear entre *x* e *y*. Por exemplo, quando se quer saber a relação dose-resposta de uma determinada droga para que se possa prever a reação fisiológica (y_i) de um indivíduo **i** em função do uso de uma determinada quantidade de medicamento (x_i). Na regressão linear, há uma relação de causa e efeito e, para uma amostra de *n* pares de observações (y_i, x_i), o interesse é estimar a equação abaixo, geralmente pelo método dos quadrados mínimos.

$$y_i = a + bx_i + e_i$$

em que:

y_i = variável dependente ou resposta, aleatória.

a = parâmetro, intercepto ou coeficiente linear a ser estimado.

b = parâmetro, coeficiente angular a ser estimado.

x_i = variável preditora, aleatória.

e_i = erro aleatório.

Estimação dos parâmetros por quadrados mínimos

Significa determinar os valores dos parâmetros (a, b) que minimizam a soma dos quadrados dos resíduos (SQR) dada por:

$$SQR = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

As estimavas de *a* e *b* são obtidas pela diferenciação desta equação que é igualada a zero.

$$\partial(\sum_{i=1}^{n} e_i^2) / \partial b = \partial[\sum_{i=1}^{n} (y_i - a - bx_i)^2] / \partial b, \text{ que resulta em: } \hat{b} = \frac{\sum_{i=1}^{n} xy}{\sum_{i=1}^{n} x^2}$$

$$\partial(\sum_{i=1}^{n} e_i^2) / \partial a = \partial[\sum_{i=1}^{n} (y_i - a - bx_i)^2] / \partial a, \text{ que resulta em: } \hat{a} = \bar{y} - \hat{b}\bar{x}$$

Em notação matricial, a regressão linear é escrita como $y = x\beta + e$, em que:

y = vetor n x 1 de valores a serem estimados.

x = matriz de desenho n x 2 (n observações nas linhas e duas regressoras nas colunas).

β = vetor 2 x 1 de parâmetros desconhecidos.

e = vetor n x 1 de erros aleatórios.

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1\,x_1 \\ 1\,x_2 \\ \vdots \\ 1\,x_n \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

Cálculos da análise de variância (Tabela 2):

SQ Regressão = $\hat{b}'\, x'y - (\sum y)^2/n$.

QM Regressão = SQ Regressão/1.

SQ Resíduo = $y'y - \hat{b}'\, x'y$.

QM Resíduo = SQ Resíduo/(n-2).

Valor de F = QM Regressão/QM Resíduo.

SQ Total = $y'y - (\sum y)^2/n$.

Notação:

$\hat{b}'$ = vetor b estimado e transposto.

y' = vetor y transposto.

x' = matriz x transposta.

Considerando-se o arquivo de dados "*corre*" e o programa SAS a seguir, os diversos cálculos de uma análise de variância acima são apresentados na Tabela 2, mostrando que a regressão linear é altamente significativa (< 0,0001).

```
proc reg data=corre;
    model y = x1;
   run;
```

Tabela 2. Análise de variância da regressão linear.

Causa de variação	Grau de liberdade	Soma de quadrados	Quadrado médio	Valor de F	Pr > F
Regressão linear	1	26,5262	26,5262	145,52	<0,0001
Resíduo	n - 2 = 7	1,2760	0,1823		0,4246
Total	n - 1 = 8	27,8022			

Testes de hipóteses dos parâmetros

Duas hipóteses são de interesse:

H_0: a = 0 versus H_a: a ≠ 0.

H_0: b = 0 versus H_a: b ≠ 0.

Para cada hipótese, calcula-se o teste t (t_c) abaixo, que corresponde a uma distribuição "t" de Student com n - 2 graus de liberdade (gl). O valor de *t* é consultado na Tabela 1.3 (Anexo 1) para α/2; geralmente é escolhido α = 0,05. As conclusões são:

$$t_c = \frac{\hat{a} - 0}{\sqrt{Var(\hat{a})}}$$

em que: $\sqrt{Var\,(\hat{a})} = \sqrt{QMResiduo\,/[1/n+x^2/\sum_{i=1}^{n}(x_i - \bar{x})^2]}$

Decisão: se $|tc| > t_{(\alpha/2,\,gl)}$, rejeita-se H_0: a = 0 versus H_a: a ≠ 0.

$$t_c = \frac{\hat{b} - 0}{\sqrt{Var(\hat{b})}}$$

em que: $\sqrt{Var(\hat{b})} = \sqrt{QMResiduo\,/\sum_{i=1}^{n}(X_i - \bar{X})^2}$

Decisão: se $|tc| \geq t_{(\alpha/2,\,gl)}$, rejeita-se H_0: b = 0 versus H_a: b ≠ 0.

Na Tabela 3, tem-se a regressão linear estimada $\hat{y}_i = 7{,}5891+1{,}4409x_i$ e os testes de hipóteses dos parâmetros calculados pelo programa SAS.

```
proc reg data = corre;
model y = x1; run;
```

O valor de *t* para â e $\hat{b}$, são obtidos por:

$$t_c = \frac{\hat{a} - 0}{\sqrt{Var(\hat{a})}} = 16{,}85$$

$$\hat{b} - 0$$

$$t_c = \frac{}{\sqrt{Var(\hat{b})}} = 12{,}06$$

Para ambas as hipóteses, rejeitou-se H_o.

Tabela 3. Modelo ajustado para a regressão linear.

Variável	Grau de liberdade	Estimativa	Erro-padrão	Valor de t	Pr > | t |
Intercepto	1	7,5891	0,45044	16,85	<0,0001
x1	1	1,4409	0,11945	12,06	<0,0001

Regressão linear e a análise de variância de experimentos

Frequentemente, em um experimento, os tratamentos são representados por níveis quantitativos ou doses crescentes e tem-se interesse de verificar o quanto da variabilidade desses tratamentos pode ser explicado por uma regressão linear.

Seja um experimento em que se avalia o número de animais em pastejo e os tratamentos são quatro lotações (0,5; 1,0; 1,5 e 2,0) de animais por hectare com seis repetições (piquetes) e a variável resposta é o ganho médio mensal por animal (quilograma, kg) (Tabela 4).

Tabela 4. Ganho médio mensal por animal, em quilograma (kg), de acordo a lotação (animal/hectare) em seis repetições (piquetes).

Repetição	Lotação (animal/ha)			
	0,5	1,0	1,5	2,0
1	10,5	12,6	8,1	6,5
2	9,4	7,9	6,7	5,8
3	11,0	10,4	-	4,3
4	8,3	9,3	7,3	7,0
5	15,0	7,2	7,0	3,9
6	12,7	8,9	6,1	4,6

Fonte: Sampaio (1998).

A análise de variância do experimento da Tabela 4 pelo procedimento GLM do SAS, a seguir, para o arquivo de dados (lotacao), tratamentos (trat) e repetições (rep) é apresentada na Tabela 5.

```
data lotacao;
input rep trat y @@;
cards;
1 0.5 10.5  1 1.0 12.6  1 1.5 8.1  1 2.0 6.5
2 0.5  9.4  2 1.0  7.9  2 1.5 6.7  2 2.0 5.8
3 0.5 11.0  3 1.0 10.4  3 1.5   .  3 2.0 4.3
4 0.5  8.3  4 1.0  9.3  4 1.5 7.3  4  2.0 7.0
5 0.5 15.0  5 1.0  7.2  5 1.5 7.0  5  2.0 3.9
6 0.5 12.7  6 1.0  8.9  6 1.5 6.1  6 2.0 4.6
;
proc glm data = lotacao;
     class trat;
     model y = trat/ss3;
    means trat;
     contrast 'regressão linear' trat -3 -1 1 3;
  run;
```

Tabela 5. Análise de variância utilizando dados da Tabela 4.

Fator de variação	Grau de liberdade	Soma de quadrados	Quadrado médio	Valor de F	Pr>F
Tratamentos	3	115,9227	38,6409	12,7351	<0,0001
Regressão linear	1	115,7819	115,7819	38,1600	<0,0001
Resíduo	19	57,6503	3,0342		
Total	22	173,5730			

Na Tabela 5, observa-se que a SQ para tratamentos é praticamente igual à SQ do contraste com 1 grau de liberdade (SQ da regressão linear), o que significa que as quatro lotações (0,5; 1,0; 1,5 e 2,0) de animais por hectare influenciam linearmente o ganho médio mensal, em quilograma, dos animais em pastejo.

Na Tabela 6 são apresentadas as quatro médias dos tratamentos obtidas pelo procedimento "*means trat*" da rotina SAS anterior e as estimadas por meio da regressão $\hat{y}$ = 13,1667 - 3,9415trat, obtida da rotina SAS.

```
proc reg data = lotacao;
  model y = trat;
  run;
```

Os resultados mostram que o aumento do número de animais em pastejo por hectare diminui linearmente o ganho médio mensal por animal. Embora neste capítulo sejam discutidas apenas técnicas que estudam a relação linear entre variáveis, caso o contraste linear não fosse suficiente para explicar os efeitos de tratamentos, poderiam se estimar contrastes quadrático e cúbico.

Tabela 6. Médias e erros-padrão de ganho de peso médio mensal, em quilograma (kg), de quatro lotações (0,5; 1,0; 1,5 e 2,0) de animais por hectare obtidos pela Anova e estimados por regressão linear.

Obtidas pela Anova	Estimadas pela regressão linear
11,15 ± 2,40	11,19
9,38 ± 1,93	9,22
7,04 ± 0,74	7,25
5,35 ± 1,27	5,28

Regressão linear múltipla

Regressão linear múltipla ou simplesmente regressão múltipla é uma técnica estatística que tem por objetivo determinar o parentesco linear entre uma variável resposta *y* em função de *k* variáveis explicativas ou regressoras ($x_1, \ldots, x_k$) e p parâmetros (p = k+1) desconhecidos ou coeficientes de regressão ($b_0, b_1, \ldots, b_k$) por meio da equação:

$$y = b_0 + b_1x_1 + b_2x_2 + \ldots + b_kx_k + e$$

Para uma amostra de *n* valores, tem-se a equação:

$$y_i = b_0 + b_1x_{1i} + b_2x_{2i} + \ldots + b_kx_{ki} + e_i \, (i = 1, 2, \ldots, n)$$

em que:

y_i = variável dependente ou resposta.

x_i = variável explanatória.

b_0 = intercepto (constante).

b_k = coeficiente angular de variável explanatória.

e_i = erro aleatório suposto independente e identicamente distribuído.

A análise de variância de uma regressão linear múltipla é apresentada na Tabela 7.

Tabela 7. Análise de variância de regressão linear múltipla.

Causa de variação	Grau de liberdade	Soma de quadrados	Quadrado médio	F calculado (F_{calc})
Regressão	k	SQ_{reg}	$QM_{Reg} = SQ_{reg}/k$	QM_{Reg}/ QMR
Resíduo	n - k -1	SQR	QMR = SQR/(n - k - 1)	
Total	n - 1	SQ_{total}		

Construindo-se a tabela da Anova a partir do resultado de uma análise, se (F_{calc}) > F Tabelado $F_{(k; n-k-1)}$ para α = 0,05 a regressão é significativa a um nível de confiança de 95%.

Cálculo das somas de quadrados do resíduo (SQR), total (SQ_{total}) e regressão (SQ_{reg}) da Tabela 7:

$$SQR = \sum_{i=1}^{n} (y_i - b_0 - b_1x_{1i} - b_2x_{2i} - ... - b_kx_{ki})^2$$

$$SQ_{total} = y'y - (\sum_{i=1}^{n} y_i)^2/n$$

$$SQ_{reg} = SQ_{total} - SQR$$

Coeficiente de determinação (R^2)

O coeficiente de determinação (R^2) que mostra a proporção dos valores observados que é explicada pelo modelo estimado é calculado por:

$$R^2 = SQ_{reg} / SQ_{total} = 1 - (SQR/SQ_{total})$$

Teste de hipótese dos parâmetros

A hipótese de interesse é testar cada parâmetro individualmente:

Ho: $b_j = 0$ versus Ha: $b_j \neq 0$ (j = 0, ... , k)

Para cada hipótese, calcula-se o teste t (t_c) que corresponde a uma distribuição "t" de Student com n - k - 1 graus de liberdade (gl). O valor de t é consultado na Tabela 1.3 (Anexo 1) para α/2; geralmente é escolhido α = 0,05.

$$tc = \frac{\hat{b}_j}{\sqrt{Var(\hat{b}_j)}}$$

$Var(\hat{b}_j) = S^2_{cjj}$ onde $S^2 = \sum_{i=1}^{n} (y_i - \hat{y})^2/(n-k-1)$ e c_{jj} é o j-ésimo elemento da diagonal de $(x'x)^{-1}$.

Rejeita-se a hipótese nula se $| t | > t_{\alpha/2}$.

A regressão linear múltipla tem aplicações em diversas áreas. Na pesquisa agropecuária, no melhoramento genético de plantas, por exemplo, é comum avaliar variáveis como produtividade de grãos, produção de matéria seca, comprimento da vagem, altura da planta, etc. Como a variável resposta geralmente é a produção, o interesse é escolher quais variáveis preditoras ajustam o melhor modelo. Cinco métodos são mais comumente usados.

a) Minimização de R^2 (*minr*) − inicia com cada modelo tendo somente uma variável preditora e ajusta do menor para o maior R^2. Em seguida ajusta do menor para o maior valor de R^2 para modelos com duas variáveis preditoras e assim por diante.

b) Maximização de R^2 (*maxr*) − ajusta o modelo com o maior R^2 para uma variável preditora. Em seguida ajusta-se o modelo com o maior R^2 para duas variáveis preditoras e assim por diante.

c) *Forward selection* − começa somente com o intercepto no modelo "b_o" e fixa-se um valor máximo para α (ex.: α =0,15) para uma variável ser adicionada no modelo (*slentry* = α). Para cada variável independente que é adicionada no modelo, calcula-se a estatística F para indicar a contribuição no modelo com a inclusão da variável. O Forward então adiciona a variável que tem a maior estatística F no modelo. O processo se repete para as variáveis que ainda estão fora do modelo. As variáveis são adicionadas uma a uma no modelo até que nenhuma variável remanescente produza um F significativo. Uma vez que a variável esteja no modelo, ela permanece.

d) *Backward elimination* − inicia por calcular F para um modelo contendo todas as variáveis preditoras. A partir daí as variáveis são retiradas do modelo uma

a uma até que todas as variáveis remanescentes no modelo produzam um F significativo de acordo com um valor de α pré-estabelecido para a variável permanecer no modelo (slstay= α), por exemplo, α = 0,10. Em cada etapa, a variável que mostra ter a menor contribuição no modelo é retirada.

e) *Stepwise regression* − é similar ao Forward selection, inicia somente com o intercepto do modelo "b_o", porém, com o avançar da técnica, as variáveis que já estão no modelo podem não permanecer lá.

As variáveis são adicionadas uma a uma no modelo, e a estatística F precisa ser significativa para o valor de α para a variável entrar no modelo (s*lentry* = α). Após a adição de cada variável, faz-se uma checagem de todas as variáveis já incluídas e retira-se a variável que não produz um F significativo de acordo com o valor de α para ela permanecer no modelo (*slstay* = α). O processo continua e somente termina quando nenhuma das variáveis fora do modelo produz um F significativo para um valor de α para a variável entrar no modelo (*slentry* = α).

Na rotina SAS a seguir, o arquivo *multi* contém dados de camarões com aproximadamente três meses de idade. Foram medidas cinco variáveis, em centímetro (cm): comprimento total (x1), comprimento do abdômen (x2), perímetro do abdômen (x3), comprimento do cefalotórax (x4) e perímetro do cefalotórax (x5) e peso corporal, em gramas (g) (y). A rotina contém informações de como utilizar as cinco metodologias descritas acima, mas ela foi executada para este arquivo somente para o procedimento *stepwise regression*

*model y= x1 x2 x3 x4 x5 / selection= stepwise slentry=**0.20** slstay=**0.10** details;*

em que:

"*details*": produz sumário estatístico para cada etapa.

"*slentry* = α": critério para a variável entrar no modelo. Ela precisa ser significativa ao nível p = 0,20 pelo teste F.

"*slstay* = α": critério para a variável permanecer no modelo. Ela precisa ser significativa ao nível p=0,10 pelo teste F.

```
data multi;
input x1-x5 y;
cards;
13.0  8.7  3.9  3.5  5.4  16.48
12.7  8.3  4.4  3.3  4.8  15.26
12.4  8.1  4.3  3.2  5.0  14.93
12.5  8.3  4.2  3.3  4.6  14.62
12.6  7.7  4.2  3.3  5.0  14.41
12.1  7.5  4.2  3.1  4.5  13.26
10.9  7.0  3.7  3.0  3.9  10.32
12.0  7.5  4.2  3.1  4.6  14.03
12.5  8.4  4.5  3.3  4.7  15.65
11.5  8.0  4.4  2.9  4.6  13.28
11.1  6.6  3.8  2.8  4.4   9.94
10.4  6.8  3.5  2.8  4.0   8.57
11.2  7.5  3.8  2.8  4.0   9.90
12.5  8.0  4.4  3.2  4.7  14.37
10.9  7.0  3.8  2.8  4.4  10.21
12.5  7.9  4.2  3.1  4.6  13.65
11.4  7.2  3.9  2.8  4.2  10.95
11.6  7.6  4.2  3.2  4.4  11.32
12.0  8.3  4.0  3.1  4.5  12.76
11.7  7.7  4.1  3.4  4.0  10.89;
proc reg data = multi;
    model y = x1 x2 x3 x4 x5 / selection = minr detalis;
    model y = x1 x2 x3 x4 x5 / selection = maxr details;
    model y = x1 x2 x3 x4 x5 / selection = backward slstay = 0.10 details ;
    model y = x1 x2 x3 x4 x5 / selection = forward slentry = 0.20 details;
    model y = x1 x2 x3 x4 x5 / selection = stepwise slentry = 0.20 slstay = 0.10 details ;
  run;
```

Na análise pelo *forward selection* (Tabela 8) o procedimento inicia somente com o intercepto "b_o"; na etapa 1 entra x1; na etapa 2 entra x5; na etapa 3 entra x3; na etapa 4 entra x2. Porém, x2 tem significância pelo teste F de p = 0,1174, não atendendo ao critério de permanência (*slstay = 0,10*). Na etapa 5, que é a última, a variável x2 é removida e o modelo final é ajustado. Naturalmente, quanto menor o valor de α, maior é o rigor no modelo final.

Tabela 8. Etapas da Anova do *Stepwise regression* na análise de dados de camarões do arquivo Multi.

Variável	Estimativa	Erro-padrão	SQ tipo II	Valor de F	Pr > F
Etapa 1 → Variável X1 entrando no modelo R^2 = 0,8978					
Intercepto	-22,8167	2,8320	36,7784	64,91	<,0001
X1	2,9942	0,2380	89,6328	158,19	<,0001
Etapa 2 → Variável X5 entrando no modelo R^2 = 0,9257					
Intercepto	-21,6997	2,5237	32,2451	73,93	<,0001
X1	2,2209	0,3705	15,6679	35,92	<,0001
X5	1,7865	0,7071	2,7840	6,38	0,0217
Etapa 3 → Variável X3 entrando no modelo R^2 = 0,9424					
Intercepto	-22,8980	2,3564	33,9073	94,42	<,0001
X1	1,6779	0,4201	5,7287	15,95	<,0010
X3	1,6201	0,7514	1,6692	4,65	0,0466
X5	2,0142	0,6502	3,4457	9,60	0,0069
Etapa 4 → Variável X2 entrando no modelo R^2 = 0,9514					
Intercepto	-22,4826	2,2505	32,2851	99,80	<,0001
X1	1,2138	0,4868	2,0108	6,22	0,0248
X2	0,7292	0,4388	0,8930	2,76	0,1174
X3	1,4601	0,7197	1,3316	4,12	0,0606
X5	2,0433	0,6174	3,5431	10,95	0,0048
Etapa 5 → Variável X2 removida do modelo R^2 = 0,9424					
Intercepto	-22,8980	2,3564	33,9073	94,42	<,0001
X1	1,6779	0,4201	5,7287	15,95	0,0010
X3	1,6201	0,7514	1,6692	4,65	0,0466
X5	2,0142	0,6502	3,4457	9,60	0,0069

Covariância

A covariância (Cov) é uma medida de associação linear entre duas variáveis aleatórias que combina algumas características da regressão e da análise de variância. Se *x* e *y* são duas variáveis aleatórias, a covariância entre elas é definida por:

$$\text{Cov}\,(x,y) = \delta_{xy} = E[x - E(x)][y - E(y)] = \Sigma xy/(n-1)$$

É o valor médio do produto dos desvios de *x* e *y* em relação às suas respectivas médias.

Se *x* e *y* são independentes, a Cov(x,y) = 0.

Como expressão matemática, a covariância é parte integrante do cálculo da correlação e também da regressão linear. Ela tem grande importância também na análise de variância (Anova), na qual recebe o nome de covariável.

Participação da covariância na análise de variância

Nos experimentos planejados, embora tenhamos controle de todas as situações, acontece que, em determinadas situações, temos variáveis que fazem parte dos experimentos, interferem nos resultados, mas não temos condições de controlá-las.

Seja um experimento com vacas leiteiras cujo objetivo é avaliar tratamentos com respeito à alimentação dos animais e verificar como esses tratamentos interferem na produção de leite. Nessas situações, mesmo selecionando as vacas quanto à idade, número de partos, produção de leite, etc., verifica-se que os animais diferem quanto à produção inicial de leite. Se não controlar essa variável, os tratamentos serão prejudicados, pois vacas com maior produção de leite inicial irão beneficiar os tratamentos e vice-versa.

Quando se inclui a covariável na Anova, corrige-se o seu efeito prejudicial e garante-se que todos os tratamentos serão avaliados nas mesmas condições. Isto é, garante-se que as vacas iniciem o experimento com a mesma produção de leite.

O programa SAS a seguir foi utilizado no Capítulo 9 no item delineamento inteiramente casualizado (DIC) com a Unidade Experimental Avaliada no Tempo, em que foram analisados dados de produção de leite de vaca durante sete controles com intervalos de 14 dias cada. No experimento, houve um período pré-experimental de 7 dias e essa produção foi analisada como covariável (pre). Em uma rotina SAS para análise de variância, a covariável não é declarada na opção “class” e é o último efeito colocado no modelo (*model*).

```
proc mixed covtest asycov;
class tratamento vaca controle;
model y = tratamento controle tratamento*controle pre / s;
/* Pre = produção avaliada no período pré experimental de 7 dias (covariável)  */
repeated controle / type = hf subject = vaca r rcorr;
lsmeans tratamento controle tratamento*controle/pdiff = all adjust = tukey;
run;
```

Quando se realiza uma análise de variância de um experimento em que estamos avaliando tratamentos (trat) e *x* é a covariável, as variáveis respostas são corrigidas pela covariável de acordo com o modelo abaixo.

$$y_{ij} = \mu + b(x_{ij} - \bar{x}) + trat_j + \varepsilon_{ij}$$

em que:

y_{ij} = variável resposta associada à covariável *i* e tratamento *j*.

μ = efeito médio global.

b = coeficiente de regressão linear entre *y* e *x*.

$\bar{x}$ = média dos valores de *x*.

x_{ij} = covariável associada à parcela *i* e tratamento *j*.

$trat_j$ = tratamento *j*.

ε_{ij} = erro aleatório.

Exercícios[12]

1) No texto a seguir, existem afirmações incorretas quanto a conceitos de estatística. Reescreva o texto colocando definições corretas e sublinhe ou coloque em negrito onde houve definições incorretas.

Regressão linear, covariância e correlação são técnicas estatísticas utilizadas no estudo entre duas variáveis *x* e *y*, que necessariamente devem ser contínuas. A regressão linear tem por objetivo estimar y_i em função de x_i em uma relação de *n* pares de variáveis aleatórias (y_i, x_i, i = 1, 2, ... , n), porém o modelo estimado é mais importante quando não existe uma relação linear entre *x* e *y*. Na covariância é determinado o parentesco entre estas duas variáveis, isto é, como elas se covariam, enquanto na regressão linear e correlação, é estudada a relação entre *x* e *y*; porém, em ambas as técnicas, devem existir uma relação de causa e efeito entre elas. Uma vantagem da aplicação dessas três técnicas estatísticas é que, em qualquer situação, o valor obtido e sua significância estatística independem do tamanho amostral. Quando se deseja saber apenas o parentesco da relação linear entre duas variáveis aleatórias *x* e *y* obtidas de uma amostra de tamanho *n*, principalmente quando elas são medidas em um mesmo indivíduo, o coeficiente de correlação *r* (-1 < r < 1) é suficiente.

[12] As respostas dos exercícios podem ser consultadas no Apêndice 1.

2) Na Figura 1 a seguir é apresentada a produção de matéria seca, g vaso^{-1}, em função do teor de fósforo, g kg^{-1}. Interprete este resultado.

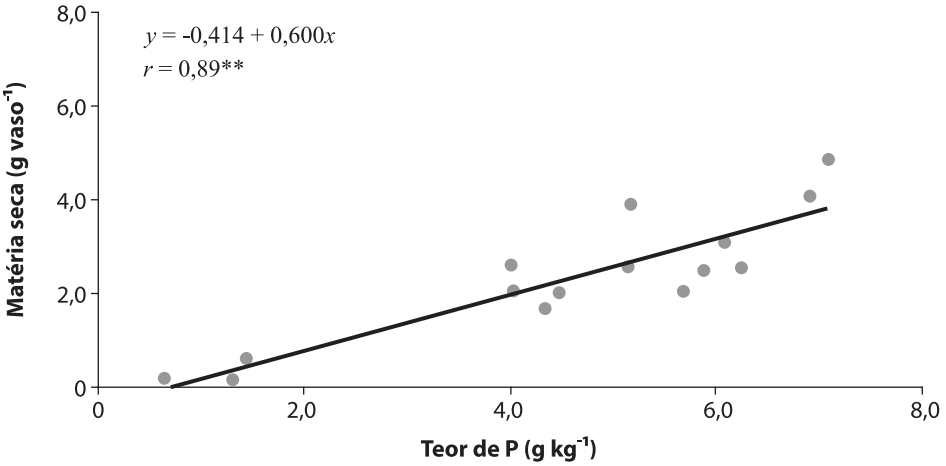


Figura 1. Produção de matéria seca, g vaso^{-1}, em função do teor de fósforo, g kg^{-1}.
Fonte: Moreira et al. (2008).

3) No quadro de análise de variância para regressão linear da Tabela 9, a regressão linear estimada foi: $\hat{y}_i = 56+0,1x_i$, em que y é a vida, em horas, de um material elétrico cuja duração é afetada pela temperatura (x). No experimento foram utilizadas cinco temperaturas (0 °C, 25 °C, 50 °C, 75 °C e 100 °C).

Tabela 9. Análise de variância para regressão linear.

Fator de variação	Grau de liberdade	Soma de quadrados	Quadrado médio
Entre temperaturas	4	660,00	-
Devido à regressão linear	1	187,50	187,50
Desvio da regressão linear	3	472,50	157,50
Erro	10	70,00	7,00
Total	14	730,00	

Fonte: Hicks (1973).

Determine os cinco valores estimados e os cinco resíduos, sabendo-se que a correspondente média para cada temperatura, respectivamente, foi: 50,0 °C; 60,0 °C; 70,0 °C; 65,0 °C e 60,0 °C.

Solução:

```
data;
input x media;
y = 56 + 0.1*x;
res = y - media;
cards;
0    50.0
25   60.0
50   70.0
75   65.0
100  60.0
;
proc print; var x y media res;run;
```

Os cinco valores estimados (ŷ) e os cinco resíduos estão abaixo.

x	ŷ	Média	Resíduo
0	56,0	50	6,0
25	58,5	60	-1,5
50	61,0	70	-9,0
75	63,5	65	-1,5
00	66,0	60	6,0

4) Com relação ao exercício 3, a regressão linear foi suficiente para estimar os resultados dos tratamentos ou necessita ajustar novos modelos?

5) Na Tabela 10, corrigir a interpretação quando necessária.

Tabela 10. Coeficientes de correlação e interpretação.

Coeficiente de correlação	Interpretação
r = -1	Razoavelmente negativa
-0,2 ≤ r < 0,1	Significativamente negativa
0,2 ≤ r ≤ 0,4	Razoavelmente positiva
0,7 ≤ r ≤ 0,9	Altamente positiva
-0,3 ≤ r ≤ -0,1	Altamente negativa
r = 0	Correlação nula
r = 1	Correlação positiva perfeita

6) Na Figura 2 tem-se, no eixo-y, a taxa de maturidade absoluta, em percentagem, - TMA por dia em função da idade, em dias. A TMA é uma variável obtida quando

se utiliza modelos não lineares para estimar o crescimento de animais. Na Figura 2 é apresentada a TMA obtida do ajuste de dois modelos não lineares utilizados na estimativa de crescimento de vacas Nelore, do nascimento até 750 dias de idade, criadas na região Norte do Brasil. Caso você tivesse o arquivo de dados original acrescido da variável TMA por dia, quais contribuições a correlação e a regressão linear poderiam dar a este estudo?

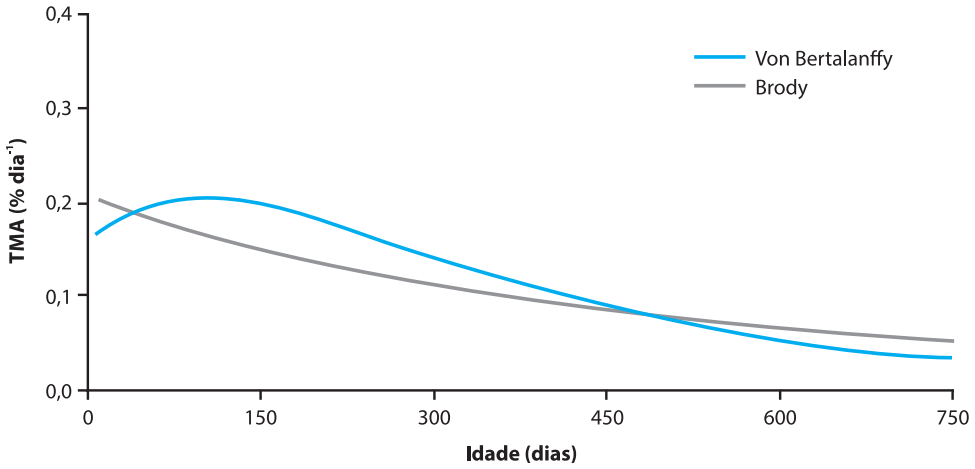


Figura 2. Percentual de maturidade absoluta por dia [TMA (% dia^{-1})] em função da idade em dias obtidas de dois modelos não lineares: Brody (—) e Von Bertalanffy (—).

Fonte: Marinho et al. (2013).

Capítulo 13

Dados categóricos

Introdução

Dados categorizados representam categorias ou características que não são susceptíveis de medida, mas, sim, de classificação. A análise de dados categorizados organizados em tabelas de contingência é comum nas diversas áreas do conhecimento, tais como agricultura, indústria, laboratórios, pesquisas de opinião pública, medicina, ensino, áreas sociais, etc. Na aplicação de questionários, por exemplo, as perguntas geralmente são fechadas e possibilitam respostas do tipo sim ou não; múltipla escolha, respostas do tipo escalas de avaliação; escores e/ou notas, que representam dados categorizados, os quais são organizados em tabelas de contingências e analisados pelo teste de qui-quadrado de Pearson (χ^2) e testes correlatos. Alguns exemplos de aplicações do teste de χ^2 são descritos a seguir:

a) Teste de tendência − detecta se há tendência positiva ou negativa entre as variáveis respostas de um experimento.

b) Teste de aderência − testa se uma amostra de dados nominais difere de uma distribuição hipotética; como, por exemplo, avaliando a normalidade dos dados.

c) Teste de independência − testa se há independência entre as frequências observadas e esperadas das linhas e colunas.

d) Teste de concordância − verifica se há concordância entre as variáveis de classificação das linhas e variáveis de classificação das colunas.

Neste capítulo, são apresentados vários exemplos de tabelas de contingências e aplicações do teste de qui-quadrado de Pearson (χ^2) e correlatos. O leitor, além de se familiarizar com esses assuntos, terá a oportunidade de saber como utilizar os fundamentos teóricos, principalmente do Capítulo 5.

Tabelas de contingência

São tabelas de classificação, r x c, organizadas com dados categóricos consistindo de *r* linhas e *c* colunas (i = 1, 2, ..., r; j = 1, 2, ..., c). As linhas e colunas correspondem a grupos, tais como, raça, sexo, etc. Cada combinação r × c representa uma célula. Um exemplo de tabela r × c é apresentado na Tabela 1. Serão discutidas a importância e estatísticas específicas para quatro tipos de tabelas de contingência: 1 × c, 2 × 2, 2 × k e r × c.

Tabela 1. Contingência r x c.

Linha	Coluna			
	1	...	c	Total
1	n_{11}	...	n_{1j}	$n_{1.}$
2	n_{21}	...	n_{2j}	$n_{2.}$
...	...	...	...	...
r	n_{i1}	...	n_{ij}	$n_{i.}$
Total	$n_{.1}$	...	$n_{.j}$	n..

Um dos principais objetivos em uma tabela de contingência é calcular a estatística de χ^2 de Pearson para um teste de independência:

$$\chi^2_{gl} = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

em que:

χ^2_{gl} = χ^2 de Pearson com graus de liberdade gl = (r - 1)(c - 1).

o_{ij} = frequência observada em cada célula.

e_{ij} = frequência esperada; $e_{ij} = n_{i.}\, n_{.j}/n..$

A estatística de χ^2 de Pearson testa a hipótese:

a) Hipótese nula H_0: As linhas e colunas são homogêneas entre si.

versus

b) Hipótese alternativa H_a: As linhas e colunas são independentes entre si.

ou

c) H_0: As frequências observadas não diferem das frequências esperadas.

d) Ha: As frequências observadas são diferentes das frequências esperadas.

Algumas recomendações para uso do teste de χ^2

O teste de χ^2 deve ser usado quando ambos, o número de observações de cada célula (n_{ij}) e a menor frequência esperada (e_{ij}), forem maiores ou iguais a 5 ($n_{ij} \geq 5$, $e_{ij} \geq 5$).

Correção de Yates

Quando n < 40 ou uma das células nij for menor que 5, utiliza-se esta correção, a qual ajusta a fórmula do teste de χ^2 por subtrair 0,5 da diferença entre cada valor observado e seu valor esperado. Essa correção reduz o valor de χ^2 obtido e aumenta o valor de *p*; com isso evita a superestimação da significância estatística para frequências pequenas.

$$\chi^2_{gl} = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(|n_{ij} - e_{ij}| - 0{,}5)^2}{e_{ij}}$$

Atendendo a essas recomendações, o teste de χ^2 é apropriado para quase todos os tipos de análises de dados categóricos e pode detectar qualquer tipo de independência e ou associação entre variáveis, porém é menos poderoso para detectar tendência entre linhas e colunas como uma tendência linear para variáveis ordinais.

Tabela de contingência 1 x c

É uma tabela de classificação simples l x c (r = 1; j = 1, 2, ..., c) que é apropriada para utilizar estatísticas para proporções binomiais, que podem ser proporções iguais e outras proporções especificadas (Tabela 2).

Tabela 2. Contingência 1 × c.

	Coluna			
Linha	**1**	**...**	**C**	**Total**
1	n_{11}	...	n_{1j}	$n_{1.} = n..$

A estatística de χ^2 com c - 1 grau de liberdade é calculada por:

$$\chi^2_{c-1} = \sum_{i=1}^{c} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$$

Proporções binomiais (p) em cada célula:

- Estimativa: $\hat{p} = n_{ij}/n$
- Erro-padrão $s(\hat{p}) = \sqrt{\hat{p}(1 - \hat{p})/n}$
- Intervalo de confiança: $\hat{p} \pm z_{\alpha/2}.\, s(\hat{p}) + (1/2n)$
- $z_{\alpha/2}$ é o 100(1 - α/2) percentil da distribuição normal padrão $z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$
- p_0 = proporção teórica ou especificada a priori.

Aplicação

Em um cruzamento de plantas, na geração F_1 foram obtidas quatro categorias com a seguinte distribuição: 119, 45, 38, 14, para as categorias de 1 a 4, respectivamente (Tabela 3). Se, na geração F_1 era esperada uma proporção de 9:3:3:1 de acordo com a Lei de Mendel, os resultados experimentais são compatíveis com essa teoria aos níveis de 5% e 1% de significância? O que se espera aqui é a não rejeição da hipótese de nulidade.

Tabela 3. Contingência 1 × c: proporções binomiais.

	Categoria 1	Categoria 2	Categoria 3	Categoria 4	Total
Observado	119 (e_{11})*	45(e_{12})	38(e_{13})	14(e_{14})	n = 216

* = frequência esperada.

Cálculo das frequências esperadas:

$$e_{11} = (216/16)x9 = 121{,}5;\ e_{12} = e_{13} = (216/16)x3 = 40{,}5;\ e_{14} = (216/16)x1 = 13{,}5$$

Como há quatro categorias, a estatística de χ^2 com três graus de liberdade é calculada por:

$$\chi^2_3 = \sum_{i=1}^{4} \frac{(119-121,5)^2}{121,5} + \frac{(45-40,5)^2}{40,5} + \frac{(38-40,5)^2}{40,5} + \frac{(14-13,5)^2}{13,5} = 0,7243$$

A estatística de χ^2 pode ser calculada por meio da rotina do Sistema de Análise Estatística (do inglês, Statistical Analysis System − SAS) a seguir com a opção "*testp*", em que as proporções esperadas 9:3:3:1, devem ser transformadas em percentagem resultando em soma de 1. O valor (pr > chisq = 0,8675), indica que χ^2_3 = 0,7243 é não significativo e a hipótese de nulidade não é rejeitada. Portanto, o número de plantas: 119, 45, 38 e 14 na geração F_1, das quatro categorias, está de acordo com a proporção de 9:3:3:1, ou seja, as frequências observadas não diferem das esperadas. Essa conclusão também pode ser obtida consultando a Tabela 2.1 (Anexo 1) para χ^2_3 = 0,7243.

```
data;
input categoria count;
datalines; /* frequências observadas nas categorias de 1 a 4 */
1 119
2   45
3   38
4   14
;
proc freq order = data; weight count;
     tables categoria / nocum testp=(.5625 .1875 .1875 .0625);   /* soma = 1 */
  run;
output
chi-square        0.7243
df                3
pr > chisq        0.8675
```

Tabelas de contingência 2 x 2

Em várias áreas das pesquisas, os dados obtidos são de natureza binária. Por exemplo, na área médica: indivíduos doentes e não doentes, vacinados e não vacinados, expostos e não expostos a um fator de risco, etc. Nas pesquisas com animais: macho-fêmea, morto-vivo, prenhez-não prenhez, fértil-não fértil, etc. Dependendo do experimento e do objetivo do pesquisador, certos tipos de dados tornam-se binários. Por exemplo, frequência de animais sadios e doentes pode ser classificada por fatores que os tornam dicotômicos, tais como, raça, sexo, estações do ano, local, etc.

As tabelas de contingência 2 x 2, além de fácil construção, fornecem informações bastante eficientes; porém precisam ser analisadas com cuidado. Muitas vezes o que se deseja não é verificar se há independência entre efeitos de linhas e de colunas e sim associação. Existe associação entre duas variáveis quando o comportamento da primeira é diferente, dependendo do nível da segunda variável. Muitas vezes utiliza-se o χ^2 de Pearson, porém os resultados fornecidos por essa estatística são inadequados ou insuficientes em muitas situações. Por exemplo, quando se têm proporções altamente correlacionadas, como as observações tomadas sobre o mesmo indivíduo; ou ainda quando esse é submetido a dois tipos de tratamentos, caracterizando situações "antes" e "após". Aplicações importantes dessas tabelas são feitas na área animal como em epidemiologia e estudos de prevalência (ex.: frequência ou número de indivíduos doentes em dado tempo) e incidência (ex.: frequência ou número de indivíduos que tornaram doentes).

Para determinar a probabilidade de um animal adquirir determinada doença, é necessário conhecer as variáveis ou fatores de risco que estão diretamente associadas a ela. Na associação entre doenças e fatores de risco, alguns estudos utilizados são denominados de retrospectivo, prospectivos, seccional cruzada ("*cross-sectional*"), etc., que são importantes para planejar racionalmente o controle sanitário do rebanho.

Estudos observacionais

O objetivo é avaliar se existe associação entre um determinado fator e a ocorrência de um desfecho, sem, no entanto, intervir diretamente na relação analisada. Os indivíduos que fazem parte do estudo não são distribuídos aleatoriamente aos tratamentos, mas já estavam classificados nos respectivos grupos, no início da pesquisa; não se tratando de um experimento. Portanto, não é possível controlar as condições de exposição e nem determinar quais os indivíduos que estão expostos e não expostos.

Por exemplo, se o interesse é estudar algum fator de um grupo de alcoólatras, não há a possibilidade de induzir um grupo de indivíduos a se tornar alcoólatra; o estudo inicia-se com grupo que já é alcoólatra e outro de não alcoólatras como grupo-controle, portanto, o estudo é observacional. Nos seres humanos, os estudos observacionais são fundamentais para estudar efeitos colaterais das diferentes terapias e suas contra indicações. Em algumas situações, estudos semelhantes são realizados com animais.

Os dados desse tipo de estudo, entretanto, são mais complicados de serem analisados do que aqueles coletados de um desenho experimental. Pode ser descritivo quando o pesquisador apenas observa, ou analítico, quando ele testa hipótese, estabelece correlações e faz inferências.

Estudos retrospectivos

São estudos observacionais cujo evento a ser pesquisado já ocorreu. Os indivíduos são acompanhados do "efeito" para a "causa". Para isso, recuperam-se fichas e outros tipos de dados dos indivíduos que estão sendo analisados, então os acontecimentos que ocorreram no passado são interpretados. Como o evento já ocorreu, esse tipo de estudo tem algumas deficiências devido a erros de confundimentos e vícios que são mais comuns do que em estudos prospectivos.

Como exemplo, considera-se a contagem ou frequência do número de indivíduos segundo o fator de risco (presente, ausente) que ocorreu no passado e o resultado ou doença atual (presente, ausente) apresentados na Tabela 4.

Tabela 4. Fatores de riscos associados a resultado ou doença.

Fator de risco (passado)	Doença (atual)		Total
	Presente (+)	Ausente (-)	
Presente (+)	$n_{11} = 25$	$n_{12} = 9$	$n_{1.} = 34$
Ausente (-)	$n_{21} = 7$	$n_{22} = 17$	$n_{2.} = 24$
Total	$n_{.1} = 32$	$n_{.2} = 26$	$n.. = 58$

Várias estatísticas são utilizadas para interpretar dados de tabelas 2 x 2 como a Tabela 4.

Sensibilidade (s)

Probabilidade de o fator de risco estar presente, dado que o indivíduo testado realmente tem a doença (caso):

$$s = n_{11}/n_{.1} = 25/32 = 0,78$$

Especificidade (e)

Probabilidade de o fator de risco estar ausente, dado que o indivíduo testado realmente não tem a doença (não caso):

$$e = n_{22}/ n_{.2} = 17/26 = 0,65$$

Valor preditivo positivo (p+)

Probabilidade de o indivíduo com fator de risco presente estar realmente doente:

$p+ = n_{11}/n_{1.} = 25/34 = 0,74$

Valor preditivo negativo (p-)

Probabilidade de o indivíduo com fator de risco ausente estar realmente sadio:

$p- = n_{22}/n_{2.} = 17/24 = 0,71$

Prevalência (p)

É a proporção de indivíduos doentes na população total avaliada:

$p = n_1/n = 32/58 = 0,55$

Acurácia (a)

Probabilidade de um método de análise acertar o diagnóstico:

$a = (n_{11} + n_{22})/n.. = (25 + 17)/58 = 0,72$

n_{11}: expostos (possuem a doença).

n_{22}: não expostos (não possuem a doença).

n.. = total de indivíduos.

Coeficiente Phi (Φ)

É uma medida de associação entre duas variáveis binárias:

$$\Phi = \frac{n_{11}n_{22} - n_{12}n_{21}}{\sqrt{n_{1.}n_{2.}n_{.1}n_{.2}}} ; (0 \leq \Phi \leq 1)$$

O coeficiente Φ varia de 0 a 1 e pode ser descrito como um caso particular de *r* de Pearson, tendo a seguinte interpretação:

a) -1,0 a -0,7: forte associação negativa.

b) -0,7 a -0,3: fraca associação negativa.

c) -0,3 a 0,3: fraca ou nenhuma associação.

d) 0,3 a 0,7: associação positiva fraca.

e) 0,7 a 1,0: forte associação positiva.

O coeficiente Φ pode ser usado para aceitar e ou rejeitar a hipótese de nulidade. Nesse caso, há uma relação entre Φ e o valor de χ^2:

$$\Phi^2 = \chi^2/n \text{ ou } \chi^2 = n\Phi^2.$$

Para interpretar Φ, precisa convertê-lo para o valor de χ^2 ($\chi^2 = n\Phi^2$). Escolhe-se o nível α que quer utilizar e consulta-se o valor crítico de χ^2 da Tabela 1.2 (Anexo 1) com grau de liberdade 1. Para $\alpha = 0,05$, tem-se $\chi^2_1 = 3,84$. Então se $n\Phi^2 < 3,84$, não se rejeita a hipótese de nulidade.

Estudo de Coorte (*Cohort Study*)

O objetivo é observar dois grupos de indivíduos por um período de tempo, isto é, expostos (possuem) e não expostos (não possuem) a fatores de riscos ou características e verificar o desenvolvimento de enfermidades ou doenças. O interesse principal neste estudo é a identificação dos efeitos da exposição na incidência de evento de interesse ou doença. Compreende dois estudos.

Estudos retrospectivos ou não concorrentes (coorte histórica)

É necessário que o pesquisador tenha acesso às informações sobre a exposição ao fator de risco no passado, ou seja, no início da investigação. Só assim é possível acompanhar os indivíduos selecionados. Em algumas situações, os dados podem ser incompletos e a análise estatística pode ficar limitada. No entanto, apresenta algumas vantagens: a) não há possibilidade de vícios, isto é, de escolher ou associar doenças a determinados indivíduos; b) tanto os indivíduos que desenvolveram a doença quanto os que não desenvolveram provêm da mesma população; c) como requer

apenas recuperação de fichas e de dados, o trabalho é somente de análise estatística e interpretação dos resultados, sendo, portanto, um experimento barato e rápido; d) são importantes para estudar doenças raras, pois os indivíduos já apresentaram a doença.

Estudos prospectivos ou concorrentes

Neste estudo, os dois grupos de indivíduos expostos e não expostos a um fator de risco, são selecionados no início do estudo e acompanhados por um período de tempo, então, verifica-se a presença ou ausência de doença no futuro (Tabela 5). Em algumas situações, no entanto, será impossível determinar o início exato da exposição. Como o evento a ser pesquisado irá ocorrer, os indivíduos são acompanhados da "causa" para o "efeito".

Os estudos prospectivos são do tipo longitudinais e observacionais, em que cada observação é analisada ao longo do tempo. São úteis para verificar se um grupo de indivíduos selecionados, ao serem expostos a um fator de risco, desenvolvem uma doença com maior ou menor probabilidade do que um grupo de indivíduos com as mesmas características, mas que não apresentam a doença.

Por exemplo, deseja estudar a diferença de indivíduos fumantes e não fumantes quanto à probabilidade de desenvolverem câncer. É um estudo mais preciso uma vez que o pesquisador tem mais controle do experimento − ele escolhe as variáveis que vai medir, como vai fazer isso e como apresentará as respostas. Como desvantagens, é um experimento mais caro, mais longo e, quanto mais rara é a doença, maior é o tamanho da amostra necessária.

Tabela 5. Contingência para estudos prospectivos.

Fator de risco	Doença no futuro		
	Presente	Ausente	Total
Expostos	n_{11}	n_{12}	$n_{1.}$
Não expostos	n_{21}	n_{22}	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	$n_{..}$

Risco relativo em estudo de coorte

O Risco Relativo (RR) é a probabilidade de um indivíduo do grupo exposto desenvolver a doença em relação ao indivíduo do grupo não exposto. Observando os dados da Tabela 4, tem-se:

- Incidência de doenças nos indivíduos expostos: I_E.

$$I_E = n_{11/}n_{1.} = 25/34 = 0,74$$

- Incidência de doenças nos indivíduos não expostos: I_N.

$$I_N = n_{21/}n_{2.} = 7/24 = 0,29$$

Cálculo do RR:

$$RR = \frac{n_{11}/n_{1.}}{n_{21}/n_{2.}} = 2,52$$

Interpretação:

- RR = 1: o risco dos indivíduos expostos e dos não expostos é o mesmo.
- RR > 1: a exposição dos indivíduos é fator de risco.
- RR < 1: a exposição dos indivíduos não é fator de risco.

Como RR > 1, existe associação entre o fator de risco e o desenvolvimento futuro da doença. Geralmente RR é importante quando é maior que 1,5. No exemplo, a probabilidade de incidência de doença nos indivíduos expostos ao fator de risco é 2,52 vezes maior quando comparada aos indivíduos não expostos ao fator de risco.

Razão de prevalência

A razão de prevalência (RP) estima a proporção de doentes entre os indivíduos expostos (P_E) em relação aos indivíduos não expostos ($P_{Não}$). É uma forma de avaliar o número de vezes que o risco de ficar doente entre os expostos é maior que o risco de ficar doente entre os não expostos.

$$RP = \frac{P_E}{P_{Não}} = \frac{n_{11}/n_{.1}}{n_{12}/n_{.2}} = \frac{25/32}{9/26} = 2,26$$

O risco de os indivíduos expostos ao fator de risco ficarem doentes é 2,26 vezes maior em relação aos não expostos.

Odds Ratios

Odds Ratios (*OR*) significa razão de chances e é tipicamente estimado de um estudo caso-controle (*case-control study*), em que os participantes são selecionados com base na presença/ausência da doença. É a razão entre número de eventos observados e os não observados.

De acordo com a Tabela 4, tem-se:

a) Probabilidade de doença no grupo de indivíduos expostos = n_{11}/n_1.

b) Probabilidade de doença no grupo de indivíduos não expostos = n_{21}/n_2.

Chance é a probabilidade de ocorrência de um evento dividida pela probabilidade da não ocorrência deste evento.

Chance = probabilidade de adoecer / (1- probabilidade de adoecer).

Chance de doença no grupo de indivíduos expostos:

$$\frac{n_{11}/n_{1.}}{1 - n_{11}/n_{1.}} = \frac{25/34}{1\text{-}25/34} = 2{,}78$$

Chance de doença no grupo de indivíduos não expostos:

$$\frac{n_{21}/n_{2.}}{1 - n_{21}/n_{2.}} = \frac{7/24}{1\text{-}7/24} = 0{,}41$$

Dividindo-se estas duas chances tem a razão de chances (OR: *odds ratios*):

$$OR = \frac{2{,}78}{0{,}41} = 6{,}78$$

O intervalo de confiança (IC) de OR com 95% de confiança é dado por:

$$e^{(\log(OR) \pm [1{,}96.EP(\log(OR))])}$$

O erro-padrão (EP) para log(OR) é:

$$EP\,(\log(OR)) = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{21}} + \frac{1}{n_{12}} + \frac{1}{n_{22}}}$$

IC: [2,1054; 21,6087].

A chance de os indivíduos desenvolverem a doença no grupo expostos é 6,78 vezes comparado ao grupo não exposto. Com 95% de confiança o verdadeiro *odds ratios* situa no intervalo de [2,1054; 21,6087].

Caso-controle (*case-control*) em estudos retrospectivos

É um tipo de estudo observacional em que o pesquisador parte da doença já instalada e por isso é um estudo retrospectivo. O grupo de indivíduos com a doença recebe o nome de caso (*case*) e o grupo de indivíduos sem a doença que é o padrão para comparação é o controle (*control*). Não é possível saber quais são os casos novos e também não apresenta a real prevalência do desfecho da população, por isso não é correto calcular a incidência e verificar o risco relativo nesse tipo de estudo.

Se um pesquisador tem por objetivo avaliar a influência de fatores de risco no câncer de mama, em um hospital ele seleciona fichas de pacientes com determinadas características e que tiveram câncer (Casos). Da mesma forma ele seleciona fichas de um grupo de indivíduos contemporâneos e com características semelhantes do grupo que apresentou câncer, mas que não apresentaram a doença (grupo-controle).

Corte transversal (*cross section*) – estudos transversais ou estudos de prevalência

O pesquisador verifica a associação entre a exposição e o desfecho em apenas um instante na linha do tempo e, assim, é possível avaliar a prevalência da doença e utilizar a razão de prevalência para avaliar a força da associação entre a exposição e a doença, por meio, por exemplo, do uso de fotografias. São razoavelmente baratos, pois são estudos comuns. Por exemplo, pode-se verificar se homens respondem melhor à determinada cirurgia do que as mulheres. Os dados são coletados de uma amostra de indivíduos em um espaço de tempo curto.

Teste de McNemar

O teste de McNemar é apropriado a tabelas de contingência 2 x 2, em que uma amostra de *n* indivíduos é submetida a uma situação antes e após, em que cada paciente serve como seu próprio controle; quando se deseja testar a diferença entre proporções pareadas ou quando os indivíduos avaliados em dois testes (teste 1, teste 2) como na Tabela 6. Suponhamos que o pesquisador deseja determinar se uma droga tem efeito sobre uma particular doença. As contagens de indivíduos são dadas em uma tabela com o diagnóstico (doença: presente ou ausente) antes do tratamento ser dado (linhas), e o diagnóstico após tratamentos (colunas). O teste requer que os mesmos indivíduos sejam incluídos antes e após avaliações (*matched pairs*).

Tabela 6. Tabela de contingência para aplicação do teste de McNemar.

Antes	Após		Total
	Teste 2 (+)	Teste 2 (-)	
Teste 1 (+)	n_{11}	n_{12}	$n_{1.}$
Teste 1 (-)	n_{21}	n_{22}	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	$n_{..}$

A primeira observação a ser feita é a hipótese nula para homogeneidade marginal, significando que as probabilidades marginais para cada resultado são as mesmas, ou seja:

$$n_{1.} = n_{.1} \text{ e } n_{2.} = n_{.2}$$

O teste de qui-quadrado com 1 g.l é dado por:

$$\chi^2_1 = (n_{12} - n_{21})^2 / (n_{12} + n_{21})$$

Com correção de continuidade, pode ser reescrito para:

$$\chi^2_1 = (|n_{12} - n_{21}| - 0{,}5)^2/(n_{12} + n_{21}) \text{ ou } \chi^2_1 = (|n_{12} - n_{21}| - 1)^2/(n_{12} + n_{21})$$

Se o teste de χ^2 é significativo, a homogeneidade marginal é rejeitada. Se n_{12} ou n_{21} são pequenos ($n_{12} + n_{21} < 10$), deve-se usar o teste exato de Fisher.

Teste exato de Fisher

É usado quando as amostras de uma tabela de frequência 2 x 2 são pequenas, ou seja, para n < 20 ou ainda quando 20 < n < 40 e a menor frequência esperada for menor que 5. A hipótese nula é que não existe associação entre a variável da linha e da coluna, isto é, hipótese de independência.

Por exemplo, em um experimento, 13 suínos com determinada doença foram divididos em dois grupos: vacinados (8) e não vacinados ou controle (5). Após certo período de observação, a resposta dos animais se apresentou conforme a Tabela 7. O interesse é saber se a vacina foi eficiente na recuperação dos animais. Uma vez que a maioria das células tem frequências esperadas menores do que 5, o teste de qui-quadrado não é válido e o recomendado é o teste exato de Fisher.

Tabela 7. Fatores independentes em uma amostra pequena.

Vacinação	Recuperação		Total
	Sim	Não	
Vacinados	$n_{11} = 6$	$n_{12} = 2$	$n_{1.} = 8$
Controle	$n_{21} = 1$	$n_{22} = 4$	$n_{2.} = 5$
Total	$n_{.1} = 7$	$n_{.2} = 6$	$n_{..} = 13$

O interesse é verificar se a proporção de animais recuperados no grupo tratado ($\hat{p}_{11} = 6/8 = 0,75$) é superior à proporção de animais recuperados no grupo-controle = ($\hat{p}_{21} = 1/5 = 0,20$).

A hipótese a ser testada é:

- Ho: Vacinados = controle.

versus

- Ha: Vacinados ≠ controle.

No teste exato de Fisher, o valor de probabilidade *p* é calculado diretamente por meio da fórmula:

$$p = n_{1.}!\, n_{2.}!\, n_{.1}!\, n_{.2}!/(n_{11}!\, n_{12}!\, n_{21}! n_{22}! n!) = 0,082$$

Como o valor $\hat{p} = 0,082$ não é significativo ao nível de 5% de probabilidade, não se rejeita a hipótese de nulidade, o que indica que a amostra não forneceu evidência de

que o grupo de animais tratado difira do grupo controle, apesar da superioridade do grupo tratada (0,75 versus 0,20).

Essa probabilidade pode ser obtida por meio da rotina SAS:

```
data;
input n11 n12 n21 n22;
n = n11+ n12+ n21+ n22;
p = fact(n11+n12)*fact(n21+n22)*fact(n11+n21)*fact(n12+n22)/
    (fact(n11)*fact(n12)*fact(n21)*fact(n22)* fact(n));
cards;
6 2 1 4
;
proc print; var p;run;
output
    p
 0.0816
```

A alternativa é utilizar a rotina do SAS a seguir:

```
data;
input tratados recuperados valor;
datalines;
1 1 6
1 2 2
2 1 1
2 2 4
;
title fisher;
proc freq;
weight valor;
tables tratados*recuperados /nopercent nocol norow ;
exact fisher;
run;
output
p
0.0816
```

Além do valor de probabilidade *p*, outros resultados são calculados por esse programa, mas não são discutidos aqui. Como o valor de *p* calculado é superior a 5% de probabilidade, se aceita a hipótese nula, ou seja, a vacinação não foi eficiente para a recuperação dos animais. Esse fato pode ser devido ao tamanho amostral. Se

essa situação se repetisse em um rebanho cinco vezes maior, como demonstrado na Tabela 8, qual seria a conclusão?

Tabela 8. Fatores independentes em uma amostra grande.

Vacinação	Recuperação		Total
	Sim	Não	
Vacinados	30	10	40
Controle	5	20	25
Total	35	30	65

A Tabela 8 não é adequada para o teste exato de Fisher, pois N é maior que 20. No entanto, ao repetir a rotina SAS para esse teste, tem-se: $\hat{p} = 0{,}000014967$.

Esse resultado é altamente significativo, indicando que, no subgrupo de animais tratados, a recuperação é maior do que no subgrupo de animais controle, conclusão oposta à situação inicial. Essas análises mostram a grande importância da representatividade da amostra nos testes de hipóteses.

Tabelas de contingência 2 x k

Teste de tendência de Cochran-Armitage

É apropriado para tabela de contingência 2 x k, cujo objetivo é avaliar a associação entre uma variável resposta com dois níveis e uma variável explanatória ordinal com *k* categorias ordenadas. O teste de tendência de Cochran-Armitage é sensível à linearidade entre as variáveis de tratamento (linhas) e as variáveis respostas (colunas). Com isso detecta tendência positiva ou negativa entre elas por meio de medidas de associação (correlação de Pearson e de Spearman) e de tendência (estatística *z*). Dependendo do valor de *z*, a tendência linear das proporções com o tempo é crescente ou decrescente.

As perguntas normalmente de interesse são:

a) Existe parentesco entre as variáveis das linhas e das colunas?

b) Se parentesco existe qual é a sua magnitude e quão forte ele é?

c) Qual é a direção e sentido do parentesco?

d) É o parentesco devido a alguma variável interferindo no modelo?

Pelo teste de tendência de Cochran-Armitage é testada a seguinte hipótese:

a) Ho: não há tendência linear nas proporções segundo as escalas numéricas.

versus

b) Ha: há tendência linear nas proporções segundo as escalas numéricas.

Aplicação

Uma aplicação bastante comum é verificar a ocorrência de doença ou não doença em dado período de acordo com doses crescentes de um medicamento. Por exemplo, dados hipotéticos de um ensaio clínico usando seis doses (0 a 5) de um medicamento são utilizadas para controlar a dor de indivíduos ou pacientes. Cada indivíduo recebe aleatoriamente o tratamento-controle (dose zero) ou uma das cinco drogas e uma resposta é avaliada (adversa = sim ou adversa = não, ”não”). A frequência de valores indica o número de indivíduos para cada combinação dose × resposta (Tabela 9).

Tabela 9. Frequência de valores para cada combinação dose × resposta.

Resposta	Dose					
	0	1	2	3	4	5
Adversa = ´sim`	8	7	9	14	22	23
Adversa = ´não`	22	23	21	16	8	7

Os resultados da análise por meio da rotina SAS dos dados da Tabela 9 estão apresentados também na Tabela 10. A opção “*trend*” testa a tendência através dos valores ordinais da variável dose, da variável resposta (adversa = sim ou adversa = não, ”não”) e da contagem (frequência) por meio do teste de tendência de Cochran-Armitage.

```
data; input dose resposta  $ contagem @@;
datalines;
    0 nao 22   0 sim  8
    1 nao 23   1 sim  7
    2 nao 21   2 sim  9
    3 nao 16   3 sim 14
    4 nao  8   4 sim 22
    5 nao  7   5 sim 23
;
proc freq;
      weight contagem;
```

```
    tables dose*resposta / nofreq nopercent nocol trend;
 run;
output
cochran-armitage trend test
statistic (z) -5.4720
one-sided pr < z <.0001
two-sided pr > |z| <.0001
```

Tabela 10. Frequência de resposta (sim, não) em função de doses (0, 1, 2, 3, 4, 5).

Dose	Frequência de resposta	
	Não	Sim
0	73,33	26,67
1	76,67	23,33
2	70,00	30,00
3	53,33	46,67
4	26,67	73,33
5	23,33	76,67

Interpretação

As proporções na Tabela 10 variaram de 23,33% a 76,67%, ou seja, tendências crescentes e decrescentes com o incremento da dosagem para a proporção de respostas 'Sim' e 'Não', respectivamente. Essas tendências, embora não sejam totalmente lineares, são suportadas pelo teste de Cochran-Armitage, pois a probabilidade (Pr < Z, < 0,0001) indica que o nível de resposta da coluna 1 (Adversa = ´não`) diminui significativamente quando a dose aumenta ou, equivalentemente, a probabilidade para o nível de resposta da coluna 2 (resposta = sim) aumenta quando a dose aumenta. Da mesma forma, o valor bilateral (Pr > |z|, <0,0001) indica efeito adverso ou significativo da droga sobre os pacientes, porém não informando qual o direcionamento. O teste bilateral, que testa simultaneamente a hipótese que a resposta aumenta ou reduz, é apropriado quando se deseja verificar se o efeito da dose é progressivo, porém sem preocupar com a direção.

Teste de Cochran-Mantel-Haenszel

O teste de Cochran-Mantel-Haenszel (CMH) testa a associação entre variáveis das linhas e colunas em tabelas 2 x 2 repetidas. São dados que contêm três variáveis

nominais, sendo duas nas tabelas 2 x 2 e uma terceira (estrato) que identifica repetição (de locais, de tempos, de estudos, etc.). Tipicamente, em cada tabela, nas linhas estão os tratamentos (controle, sexo, droga, etc.) e nas colunas a variável resposta (sadios, doentes, etc.).

A hipótese a ser testada é:

a) Ho: as respostas em cada estrato são independentes de tratamentos.

versus

b) Ha: as respostas em cada estrato são dependentes de tratamentos.

Como exemplo, na Tabela 11, a tabela 2 x 2 é a relação entre exposição (linhas) e doenças (colunas) segundo a raça humana − caucasianos e negros (Pereira, 2004), que é a terceira variável.

Tabela 11. Dados de exposição (linhas) e doenças (colunas) de acordo com raças (caucasianos e negros).

Exposição	Caucasianos		Negros	
	Doentes	Sadios	Doentes	Sadios
Exposto	70	30	8	17
Não exposto	20	5	42	58

Na rotina SAS a seguir, o valor de (P < 0,0001) para "*General Association*" indica forte associação entre exposição e resposta (doentes e sadios), após o ajuste para raça (caucasianos e negros). Portanto, rejeita-se a hipótese nula de que a proporção de exposição é a mesma para caucásio e negros.

```
data;
input raca $ exposicao $ resposta $ count @@;
datalines;
caucasia exposto  doente  70  caucasia exposto  sadio  30
caucasia nao      doente  20  caucasia nao      sadio   5
negros   exposto  doente   8  negros   exposto  sadio  17
negros   nao      doente  42  negros   não      sadio  58
;
proc freq;
    weight count;
    tables raca*exposicao*resposta/ cmh nofreq norow nocol;
  run;
```

output

alternative hypothesis	*df*	*value*	*prob*
general association	*2*	*91.9532*	*<.0001*

Tabelas de contingência r x c

São tabelas de classificação que contêm *r* linhas e c colunas (i = 1, 2, ..., r; j = 1, 2, ..., c). Quando a hipótese nula é rejeitada, o objetivo é avaliar a natureza do parentesco e a correlação entre as variáveis. Várias medidas de associação podem ser utilizadas. Além da estatística de χ^2 várias outras que descrevem a associação entre duas variáveis são também usadas nestas tabelas.

Coeficiente V de Cramer

É uma medida de associação que também varia de 0 quando não há parentesco entre duas variáveis, até o valor máximo de 1.

$$V = \frac{ad - bc}{(a+b)(c+d)(a+c)(b+d)} = \sqrt{\frac{\Phi^2}{q}} = \sqrt{\frac{\chi^{2/n\ (\text{qui-quadrado dividido por n})}}{q}}$$

em que:

q = menor valor (r-1; c-1).

n = total de observações.

χ^2= qui-quadrado de Pearson.

- Para tabelas 2 x 2: $-1 \leq V \leq 1$.
- Para as demais tabelas: $0 \leq V \leq 1$.

Coeficiente de contingência de Pearson

É uma medida do grau de parentesco entre duas variáveis. O valor do coeficiente de contingência de Pearson de (C) é ajustado para diferentes tamanhos amostrais.

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

C é sempre menor do que 1 e o valor máximo que pode atingir depende do número mínimo (m) de linhas ou colunas da tabela de contingência.

$$0 \leq C \leq \sqrt{\frac{m-1}{m}}$$

Coeficiente kappa de Cohen

É uma medida estatística de concordância geralmente usada entre avaliadores. Os dados são organizados em uma tabela de contingência r x c em que os níveis das colunas representam as observações de um avaliador e os níveis das linhas representam as observações do outro avaliador.

É uma medida mais robusta do que o simples cálculo de percentagem, pois leva em conta o cálculo de probabilidade. As estatísticas abaixo são demonstradas por meio da Tabela 12.

$$\kappa = \frac{P_0 - P_e}{1 - P_e}$$

Probabilidades:

- P_0 = probabilidade calculada com os dados observados.

$$P_0 = \sum_{i=1}^{r} \frac{n_{ii}}{n}$$

- P_e = probabilidade calculada com as frequências esperadas e representa a concordância entre os avaliadores.

$$P_e = \sum_{i=1}^{r} \frac{n_{i.}\, n_{.i}}{n^2}$$

Interpretação do coeficiente κ:

κ < 0 não concordância.

0 < κ ≤ 0,20 concordância fraca.

0,20 < κ ≤ 0,40 concordância razoável.

0,40 < κ ≤ 0,60 concordância boa.

0,60 < κ ≤ 0,80 concordância excelente.

0,80 < κ ≤ 1,00 concordância perfeita.

Aplicação

Suponhamos que dois juízes (1 e 2) estão avaliando um grupo de 44 touros nas pistas de exposições agropecuárias para finalidade de premiação. Cada touro então é avaliado pelos dois juízes que dizem "sim" ou "não" para ganhar prêmio (Tabela 12).

Tabela 12. Frequência de concordância e ou discordância.

Juiz 1	Juiz 2	
	Sim	Não
Sim	18	5
Não	6	15

A percentagem de concordância entre os dois avaliadores é:

- P_0 = (18+15)/44 = 0,75.

Cálculo de P_e:

- O juiz 1 diz "sim" para 23 touros e "não" para 21.
- O juiz 2 diz "sim" para 24 touros e "não" para 20.
- Probabilidade de ambos juízes dizerem "sim":(23/44).(24/44) = 0,52x0,54 = 0,2851.
- Probabilidade de ambos juízes dizerem "não":(21/44).(20/44) = 0,47x0,45 = 0,2169.
- Probabilidade global de concordância: 0,2851 + 0,2169 = 0,502.

O coeficiente κ é calculado por:

$$\kappa = \frac{0,75 - 0,502}{1 - 0,502} = 0,248/0,498 = 0,4979$$

```
data;
input juiz1 $ juiz2 $ freq;
cards;
sim sim 18
sim nao  5
nao sim  6
nao nao 15
```

```
;
proc freq order = data;
    weight freq;
    tables juiz1*juiz2/agree noprint ;
    test kappa;
run;
output

simple kappa coefficient
kappa                    0.4979
95% lower conf limit     0.2414
95% upper conf limit     0.7544

test of h0: kappa = 0

ase under h0             0.1506
z                        3.3063
one-sided pr >  z        0.0005
two-sided pr > |z|       0.0009
```

O resultado mostra boa concordância entre os dois juízes na avaliação dos touros. O intervalo de confiança (IC) com 95% de probabilidade para o coeficiente κ foi (0,2414 < κ < 0,7544). Naturalmente, quanto maior o tamanho da amostra, menor a amplitude do IC e mais eficiente ele fica. Nesse exemplo, simulamos quatro tamanhos amostrais diferentes, porém mantendo-se as mesmas proporções. Os valores da tabela foram multiplicados por 2, 4, 10 e 100. A seguir são apresentados os valores dos IC, sendo que o último praticamente identificou o valor populacional.

(x2) → IC = (0,3165 < κ < 0,6798).

(x4) → IC = (0,3697 < κ < 0,6262).

(x10) → IC = (0,4168 < κ < 0,5790).

(x100) → IC = (0,4723 < κ < 0,5236).

A aplicação da tabela de contingência r x c será exemplificada com dados da cadeia produtiva e de processamento do couro de bovinos. Entre os vários estágios deste estudo, está a pele salgada, também denominada de couro verde ou cru; o couro curtido úmido ou parcialmente processado (*wet-blue*) e o semiacabado (*crust*). Durante o processamento, ocorrem perdas de peles ainda no pasto, decorrentes de marcas e cicatrizes causadas na pele por agentes, como os ectoparasitas (berne, carrapatos, sarnas e micoses), e por problemas de manejo no uso indevido de arame farpado, marca

a fogo e local inadequado. É de interesse avaliar como as perdas de peles nos estados produtores ocorrem.

Na Tabela 13 são apresentadas frequências de peles com defeitos em função dos agentes e estados da Federação de pesquisas realizadas na Embrapa Pecuária Sudeste, São Carlos, SP.

Tabela 13. Tabela de contingência r × c. Os dados representam frequências de peles com defeitos em função dos agentes e estados da Federação.

Agente	Estado							Total
	BA	MG	MS	MT	PA	RS	SP	
Berne aberto	17	20	30	15	12	2	2	98
Berne curado	23	19	17	15	13	4	15	106
Carrapato	90	20	30	15	12	2	2	171
Risco aberto	22	19	25	15	13	4	3	101
Risco cicatrizado	122	110	123	117	114	125	132	843
Dermatite por sarna	110	98	115	111	118	109	112	773
Marca a fogo	220	215	203	199	213	201	198	1.449
Total	604	501	543	487	495	447	464	3.541

O interesse no uso do teste de χ^2 é testar a hipótese:

a) Ho: a ocorrência de perdas de peles é homogênea entre agentes e estados da Federação.

versus

b) Ha: a ocorrência de perdas de peles é independente entre agentes e estados da Federação.

```
data;
input estado $ agente freq @@;
cards;
ba 1 17 mg 1 20 ms 1 30 mt 1 15 pa 1 12 rs 1 2 sp 1 2
ba 2 23 mg 2 19 ms 2 17 mt 2 15 pa 2 13 rs 2 4 sp 2 15
ba 3 90 mg 3 20 ms 3 30 mt 3 15 pa 3 12 rs 3 2 sp 3 2
ba 4 22 mg 4 19 ms 4 25 mt 4 15 pa 4 13 rs 4 4 sp 4 3
ba 5 122 mg 5 110 ms 5 123 mt 5 117 pa 5 114 rs 5 125 sp 5 132
ba 6 110 mg 6 98 ms 6 115 mt 6 111 pa 6 118 rs 6 109 sp 6 112
ba 7 220 mg 7 215 ms 7 203 mt 7 199 pa 7 213 rs 7 201 sp 7 198
;
```

```
proc freq order = data;
     weight freq;
     tables estado          / chisq nofreq norow nocol;
     tables agente          / chisq nofreq norow nocol;
     tables estado*agente /nofreq norow nocol chisq;
    run;
output
statistic                      df     value       prob
entre estados
chi-square                      6      33.0624    <.0001
entre agentes
chi-square                      6    3314.811     <.0001
estados x agentes
chi-square                     36     269.7419    <.0001
mantel-haenszel chi-square      1      75.9431    <.0001
 phi coefficient                       0.2760
 contingency coefficient               0.2661
 cramer's v                            0.1127
```

As frequências e percentagens de peles com defeitos variam entre os estados. Na Tabela 14, é apresentada a relação do pior para o melhor estado e também da maior para a menor frequência quanto aos agentes causadores de defeitos nas peles de bovinos.

Tabela 14. Frequências e percentagens de peles com defeitos na tabela de contingência r × c.

Estado	Frequência	(%)	Agente	Frequência	(%)
BA	604	17,06	Marca a fogo	1.449	40,92
MS	543	15,33	Risco cicatrizado	843	23,81
MG	501	14,15	Dermatite por sarna	773	21,83
PA	495	13,98	Carrapato	171	4,83
MT	487	13,75	Berne curado	106	2,99
SP	464	13,10	Risco aberto	101	2,85
RS	447	12,62	Berne aberto	98	2,77

A estatística de qui-quadrado de Pearson (χ^2) indica associação significativa entre estados da Federação ($p < 0,0001$), devido, principalmente, ao estado da Bahia, seguido de Mato Grosso do Sul e entre agentes ($p < 0,0001$), devido, principalmente, à marca a fogo, risco cicatrizado e dermatite por sarna.

Exercícios[13]

1) No texto a seguir, existem afirmações incorretas quanto a conceitos de estatística. Reescreva o texto colocando definições corretas e sublinhe ou coloque em negrito onde houve definições incorretas.

Dados categorizados representam categorias ou características, que são susceptíveis de medida, porém, não de classificação. A análise de dados categorizados organizados em tabelas de contingência é comum nas mais diversas áreas do conhecimento. Um exemplo são os questionários que são utilizados para as mais diversas finalidades, e as perguntas fechadas geralmente possibilitam respostas do tipo sim ou não; múltipla escolha, com respostas do tipo escalas de avaliação, perguntas com escores e ou notas representam dados categorizados. Os dados categorizados são organizados em tabelas de contingências e analisados pelo teste de qui-quadrado de Pearson (χ^2) e testes correlatos. O teste de χ^2 deve ser usado quando, ambos, o número de observações em cada célula (n_{ij}) e a menor frequência esperada (e_{ij}) forem maiores ou iguais a 10. Quando $n < 40$ ou uma das células n_{ij} for menor que 10, utiliza-se a correção de Yates, a qual ajusta a fórmula do teste de χ^2 por subtrair 0,5 da diferença entre cada valor observado e seu valor esperado. Essa correção aumenta o valor de χ^2 obtido e também do valor de p, evitando a superestimação da significância estatística para dados pequenos. Existem vários testes, destacando-se a estatística de Cochran-Mantel-Haenszel que testa a hipótese Ho: não existe associação entre a as variáveis das linhas e as variáveis das colunas. versus Ha: há uma associação não linear entre as variáveis das linhas e as variáveis das colunas. Já o teste de McNemar é apropriado para tabelas de contingência 2 x 2 quando os dois fatores de classificação são dependentes, por exemplo quando se deseja testar a diferença entre proporções pareadas. Por exemplo, em estudos nos quais cada indivíduo serve como seu próprio controle ou em estudos que envolvem situações antes e após.

2) Em um experimento com 75 vacas leiteiras, após a ordenha realizou-se tratamento para infecção de mastite. Duas tetas do lado esquerdo de cada vaca foram tratadas e as duas do lado direito foram usadas como controle (Tabela 15). Cerca de 18 vacas desenvolveram mastite no lado-controle, sendo que nove e outras duas desenvolveram do lado tratado.

[13] As respostas dos exercícios podem ser consultadas no Apêndice 1.

Tabela 15. Tabela de contingência 2 × 2 – cada paciente serve como seu próprio controle.

Tratado	Controle		Total
	+	-	
+	n_{11} = 9	n_{12} = 2	$n_{1.}$ =11
-	n_{21} = 9	n_{22} = 55	$n_{2.}$ = 64
Total	$n_{.1}$ = 18	$n_{.2}$ = 57	$n_{..}$ = 75

Verifique se o tratamento foi ou não eficiente utilizando o teste de McNemar.

3) Em uma fazenda num período de 4 semanas a frequência de nascimento de bezerros segundo os dias da semana se distribuiu conforme a Tabela 16.

Tabela 16. Tabela de contingência 1 × c (frequência de bezerros nos dias da semana).

Dias	Dom.	Seg.	Ter.	Qua.	Qui.	Sex.	Sab.
Bezerros	35	28	30	29	31	27	32

Considerando-se o teste de qui-quadrado para proporções binomiais, é possível afirmar que o nascimento de bezerros é constante e com uma frequência de 30 diariamente?

4) A frequência de abscessos hepáticos em 658 caprinos jovens e adultos se distribui conforme a tabela de contingência apresentada na Tabela 17.

Tabela 17. Contingência 2 × 2.

Faixa etária	Abscesso hepático		Total
	ausente	Presente	
Jovens (0 – 12 meses)	6	384	390
adultos (>12 meses)	11	257	268
	17	641	658

Fonte: Rosa et al. (1989).

A ausência de abscessos hepáticos é idêntica nas duas faixas etárias?

5) Considerando-se os dados da Tabela 18, calcule a probabilidade p do teste exato de Fisher e discuta os resultados.

Tabela 18. Contingência 2 × 2 com variáveis nominais nas linhas e colunas.

Vacinação	Recuperação		Total
	Sim	Não	
Vacinados	$n_{11} = 5$	$n_{12} = 4$	$n_{1.} = 9$
Controle	$n_{21} = 5$	$n_{22} = 20$	$n_{2.} = 25$
Total	$n_{.1} = 10$	$n_{.2} = 24$	$n_{..} = 34$

Capítulo 14

Recursos do Sistema de Análise Estatística (SAS)

Introdução

O Sistema de Análise Estatística (do inglês, Statistical Analysis System − SAS) é usado por mais de 80 mil sites entre as maiores empresas de acordo com a classificação *Fortune Global 500, em 2018*. A maioria dos clientes opta por continuar usando o SAS a cada ano. Segundo Jim Goodnight, o principal executivo do SAS, a razão do crescimento são os grandes funcionários que criam um ótimo software e serviço (SAS Institute, 2017).

O SAS, conhecido como "sistema de análise estatística", iniciou na Universidade do Estado da Carolina do Norte, EUA, como um projeto para analisar dados de pesquisas agrícolas. Como a demanda por esse tipo de software cresceu, em 1976, foi fundado o Instituto SAS, com o objetivo de ajudar todos os tipos de clientes, tais como empresas farmacêuticas, bancos e até entidades acadêmicas e governamentais.

O SAS − software e empresa, prosperou ao longo das décadas seguintes. O desenvolvimento do software atingiu novos patamares na indústria por ser capaz de rodar em todas as plataformas, usando a arquitetura de vários fornecedores para o qual é conhecido hoje. Embora o âmbito da empresa tenha se espalhado por todo o mundo, a cultura corporativa encorajadora e inovadora permaneceu a mesma.

Algumas pessoas veem os dados como fatos e números, porém eles são muito mais do que isso. Eles representam a alma do seu negócio, eles contam a história da organização e estão tentando dizer algo.

O SAS é uma poderosa ferramenta de análise estatística e uma linguagem de programação da quarta geração que possibilita realizar atividades, tais como pesquisa operacional, análise matemática e estatística, controle de qualidade, planejamento, previsão, suporte, tomada de decisões mercadológicas, gerar relatórios e gráficos. Permite o acesso a diferentes bases de dados, está disponível para uso em vários ambientes e possui vários outros recursos. O SAS processa pequenos e grandes volumes de dados, dependendo da máquina local ou remota.

Neste capítulo, são apresentadas algumas informações sobre o sistema Statistical Analysis System (SAS) for Windows (SAS Institute, 2017), as quais estão associadas com o conteúdo deste livro.

Transformar e dar sentido à mensagem

Como líder em análise de negócios de software e serviços, o SAS ajuda a transformar dados em visões que dão uma nova perspectiva ao negócio, auxiliando na tomada de decisão. Pode-se identificar o que está funcionando, corrigir o que não está e ainda descobrir novas oportunidades.

Missão e valores

O SAS oferece soluções comprovadas que impulsionam a inovação e melhora o desempenho. O SAS ajuda as organizações a transformar grandes quantidades de dados em conhecimento que elas podem usar.

Estruturas associadas a um arquivo

Áreas ou comandos

Um arquivo em SAS permite armazenar dados no formato American Standard Code for Information Interchange (ASCII). Esse arquivo contém tanto os valores das observações (dados) quanto às informações descritivas (variáveis).

A seguir será apresentada a estrutura de um arquivo criado no sistema SAS e os comandos pertinentes.

a) Área de edição do SAS − usada para a programação em SAS propriamente dita. Todos os arquivos que são salvos nessa área possuem extensão ".SAS" e poderão ser abertos no SAS ou em formato bloco de notas (formato ASCII) ou ainda podem ser visualizados em arquivo texto, por exemplo extensão ".doc".

b) Área explorer do SAS − nessa área, será possível visualizar os conjuntos de dados ou arquivos (*datasets*) criados de maneira rápida.

c) Área de log do SAS − mostra os resultados das execuções da programação e também se a execução foi feita com sucesso ou não (mensagens de erro).

d) Área de output do SAS − mostra os resultados (saídas) da execução de programas SAS.

A seguir, é apresentada a estrutura básica de um arquivo <*data ferrugem;*> de um experimento:

/ ===== Início dos Comentários ==========
informações de identificação
título do projeto: < título do projeto >
res. pelo projeto:<nome do responsável pela atividade>
instituição:< nome da instituição >
data: local:
objetivo:
avaliar a produção e a qualidade da forragem do capim-coastcross.
delineamento experimental
blocos casualizados, 4 repetições em parcelas subdivididas. nas parcelas foram distribuídos 10 tratamentos em esquema fatorial 2 x 5 (duas fontes de nitrogênio: ureia e nitrato de amônio e cinco doses: 0, 25, 50, 100, 200 kg/ha/corte de nitrogênio, em cinco cortes consecutivos - subparcela). A parcela é constituída por área com 5m de comprimento x 1m de largura e a subparcela são os cinco cortes consecutivos do capim-coastcross:
corte 1: 10/12/98
corte 2: 04/01/99
corte 3: 03/02/99
corte 4: 04/03/99
corte 5: 05/04/99

descrição das variáveis – input:

rep = blocos: 4 (1,2,3,4)
fonte = fontes de adubo: 2 (1- ureia; 2 - nitrato de amônio)
dose = doses de adubo: 5 (0, 25, 50, 100, 200)
corte = cortes da forrageira (1,2,3,4,5)

variáveis respostas:

prod = produção da área útil da parcela, g
area = área útil da parcela, m2
gaf = amostragem de material fresco colhido na parcela, g
gas = amostra de material seco, g
pb = proteína bruta, %

variáveis geradas:

ms = materia seca,%
ms = gas/gaf (g/g)
pms = produção de materia seca, kg/ha
pms = ms*prod*(10/area)

*==========fim dos comentários ========== */*

```
data ferrugem;
input rep fonte dose corte prod area gaf gas pb;
/* espaço destinado à manipulação do arquivo: geração de variáveis, etc */
ms = (gas/gaf);
pms = ms*prod*(10/area);
datalines; /* aqui inicia as linhas de dados */
1   1     0    1    51.0   6     48.9     13.0     7.44

4   2  200    5  750.0   6    550.1   106.2   11.27
; /* fim dos dados */
/* espaço destinado a procedimentos e comandos para as diversas análises */
run; /* executa o  programa */
```

Comandos pertinentes a um arquivo

Comentário

"/* " → início de um comentário " */ " → fim

Exemplo: /* Informações de identificação */

Variáveis associadas a um arquivo

Variáveis são quantidades ou características sendo medidas. Existem dois tipos de variáveis: numérica, que são formadas por dígitos numéricos de 0 a 9, e não numérica (caracteres), que são formadas por caracteres alfabéticos, dígitos numéricos de 0 a 9 e ainda outros caracteres especiais. Cada variável é identificada por um nome e representa as colunas em um arquivo de dados e cada linha de um arquivo de dados representa as observações. As variáveis são criadas dentro do procedimento *Data* e após *Input*.

Exemplo de arquivo com variáveis numéricas

O arquivo A contém as variáveis *x*, *y*, *z* e alguns operadores aritméticos.

```
data a;                          /* cria o arquivo a */
input x y z;                     /* cria as variáveis x, y, z */
w = x + 5;                       /* adição */
m = (x + z)/2;                   /* adição e divisão */
x1   = x**2;                     /* potência: x1 é o quadrado de x */
if x < 5 then x = .;             /* criando dado perdido para x menor que 5  */
if (40 <= z < 45)   then classe =1; /* criando a variável classe */
datalines;
30 10 40
31 12 50
...
;
run;
```

Exemplo de arquivo com variáveis não numéricas (caracteres)

O arquivo *b_1* contém quatro variáveis associadas a um bovino, cujas explicações estão a seguir:

```
data b_1;
```

É criado o arquivo usando-se underscore ou traço abaixo (_).

```
input numero d_nasc $ @13 raca $ @20 fazenda $9.  peso ;
```

No comando *input* o símbolo "$" após < *d_nasc* > indica que essa variável é de caracteres, enquanto o símbolo *"$9."* após < fazenda > indica que esta variável de caractere é lida em 9 colunas do registro. Os símbolos *@13* e *@20* indicam que a variável "raca" inicia na coluna 13 e "fazenda" inicia na coluna 20.

```
data b_2; set b_1 (drop=raca);
```

No comando anterior é criado o arquivo < *b_2* >, utilizando-se os dados do arquivo < *b_1* >, sem a variável "raca", por meio do comando < set b_1 (drop=raca); >

O mesmo resultado é obtido no comando abaixo na criação do arquivo < *b_3* > utilizando-se os dados do arquivo < *b_1* > em que < *keep* > indica as variáveis do arquivo < *b_1* > que serão mantidas:

```
data b_3; set b_1 (keep=numero d_nasc  fazenda peso);
```

O comando < label variável = “nome da variável”> é usado para dar nome às variáveis nos relatórios de saída (output), gráficos, tabelas, etc.

```
data b_1;
input  numero d_nasc $ @13 raca $ @20 fazenda $9. peso ;
    label raca      = ‘raça do animal’;
    label fazenda = ‘fazenda de nascimento do animal’;
datalines;
200 27/06/19 nelore saojoao 101
500 17/12/19 zebu valinhos    95
...
;
data b_2; set b_1 (drop=raca);
data b_3; set b_1 (keep=numero d_nasc fazenda peso);
 run;
```

Comando

Cada comando termina com “ **;** ” e pode ter vários comandos em uma mesma linha.

```
data < nome >;
```

<nome> = nome do arquivo, que deve possuir de 1 a 32 caracteres, começar com uma letra (A até Z) e depois é livre para letras e números, podendo usar *underscore* ou traço abaixo (_). Caracteres especiais, tais como, $, %, #, &, @, etc., não podem fazer parte do nome do arquivo.

Exemplo:

```
data forragem;
```

Input

Após o input são colocados nome de variáveis, etc., que representam a identificação das colunas do arquivo de dados.

Exemplo onde < fonte > é uma variável de caractere e as demais são numéricas:

input rep fonte $ dose corte prod area gaf gas pb;

Input colunado

Especifica onde encontrar os valores (dados) correspondentes com a posição da coluna (variável).

input animal $ 1-3

A variável animal é alfanumérica e ocupa as colunas: 1 a 3. Para distinguir no SAS uma variável alfanumérica seu nome deve ser seguido do sinal $ (dólar).

input nome $10. #2 id 3-4;

A variável < nome > é lida nas primeiras 10 colunas do registro e move o apontador para o segundo registro ou segunda linha para ler < id > nas colunas 3 e 4.

input x1 x2 x3 @@;

Repete as variáveis x1, x2 e x3 na mesma linha até utilizar toda a extensão desta.

input @23 numero 4. +5 peso 4.;

Move o apontador para a coluna 23 para ler a variável “numero”, que ocupa as colunas de 23 a 26. O apontador avança 5 colunas para ler a variável peso que ocupa as colunas de 32 a 35.

input numero raça $ idade sexo $ / peso 2-5;

As variáveis numero, raça, idade e sexo são lidas no primeiro registro, sendo raça e sexo variáveis alfanuméricas. A variável peso é lida no segundo registro nas colunas de 2 a 5.

Linha

Geralmente cada linha corresponde a uma observação ou registro.

Exemplo:

1 1 0 1 51.0 6 48.9 13.0 7.44

Coluna

datalines; ou cards;

Após esses comandos, são colocados os dados. Logo abaixo da última linha de dados têm ";".

Comando set

O comando *set* é usado para criar um arquivo SAS a partir de um arquivo já existente. No exemplo abaixo, é criado o arquivo a2 a partir do arquivo a1;

Exemplo:

```
data a2;
      set a1;
run;
```

Operadores aritméticos

São recursos para criar e fazer alterações e ou modificações com as variáveis e dados. São geralmente usados após o *input*. Alguns exemplos são apresentados na Tabela 1.

Tabela 1. Exemplos de operadores aritméticos.

Símbolo		Significado	Exemplo	SAS
**		exponenciação	$y = x^4$	y =x**4;
*		multiplicação	y = a.b	y=a*b;
/		divisão	y=a/b	y=a/b;
-		subtração	y=a-b	y=a-b;
+		adição	y=a+b	y=a+b;
gt	>	maior que	a > b	a > b; a gt b;
lt	<	menor que	a < b	a < b; a lt b;
eq	=	igual a	a = b	a = b; a eq b;
le	<=	menor ou igual a	a <= b	a <= b; a le b;
ge	>=	maior ou igual a	a >= b	a >= b; a ge b;
ne	~=	diferente	a ≠ b	a ~= b; a ne b;

Funções de data e tempo

O programa a seguir calcula informações relativas a datas considerando uma mistura de formatos. São geralmente usados entre os comandos "*input*" e "*datalines*" ou "*input*" e "*cards*".

```
data dates;
      informat date1 date2 date9.;
      input date1 date2;
      format date1 date2 date9.;
/* informações extraída de uma data ou coluna: ex: 01feb2005*/
      ano        = year(date1);                    /*  ano = 2005  */
      quadri     = qtr(date1);                  /*  quadrimestre = 1 */
      mes        = month(date1);                        /*  mes = 2 */
      semana     = week(date1);                      /*  semana = 5 */
      dia_mes = day(date1);                      /* dia do mes = 1 */
      dia_semana = weekday(date1);          /*  dia da semana = 3 ´terça` */
/* informações entre duas datas */
      semana_2   = intck('week', date1, date2);          /*  semanas */
      mes_2      = intck('month', date1, date2);           /*  meses */
      anos_2     = intck('year', date1, date2);             /*  anos  */
      quadri_2   = intck('qtr',date1,date2);          /*  quadrimestres  */
      ano_2      = yrdif(date1, date2, 'actual');           /*  anos  */
/* informações posteriores a uma data: exemplo "01feb2005"  */
      next_month =intnx('month',date1,1);            /*  mes seguinte  */
next_year    =intnx('year',date1,1);              /*  ano seguinte   */
```

```
next_qtr      =intnx('qtr',date1,1);              /*  quadrimestre seguinte   */
      six_month  =intnx('month',date1,6);          /*  seis meses após   */
      format next:six_month date9.;
/* número de anos considerando uma data fixa */
       anos1  = ('14may2014'd - date1)/365.25;        /* valor com decimal  */
       anos2  = int(anos1);                              /* valor inteiro   */
       format date1 mmddyy10.;
/* informações de uma data fixa: ex: '01jan2000:5:15:30'*/
       dt='01jan2000:5:15:30'dt;
       hora       = hour(dt);                      /*  retorna a hora:       5  */
       minuto     = minute(dt);                    /*  retorna minutos:    15 */
       segundo    = second(dt);                    /*  retorna segundos:   30 */
       format dt datetime.;
      datalines;
      01feb2005  16may2014
run;
proc print data = dates heading=h; id date1 date2;
run;
```

Funções que retornam estatísticas de uma amostra: $x_1, x_2, ..., x_n$

São geralmente usadas entre os comandos "*input*" e "*datalines*" ou "*input*" e "*cards*". Na Figura 1 tem-se um exemplo de output (saída dos resultados do programa SAS).

```
/* funções que retornam algumas estatísticas descritivas para uma amostra: x1, x2, ..., xn.
No exemplo: x1= 10, x2 = 11, x3 = 12, x4 = 13, x5 = 14 */

data;
input x1-x5;
css = css(of x1-x5);                    /* soma de quadrados  corrigida */
cv = cv(of x1-x5);                      /* coeficiente de variação */
kurtosis = kurtosis(of x1-x5);          /* curtose */
max = max(of x1-x5);                    /* valor máximo */
min = min(of x1-x5);                    /* valor mínimo */
mean = mean(of x1-x5);                  /* media */
n = n(of x1-x5);                        /* número de valores não perdidos */
nmiss = nmiss(of x1-x5);                /* número de valores perdidos */
range = range(of x1-x5);                /* amplitude dos valores */
```

```
skewness = skewness(of x1-x5);                    /* assimetria */
std = std(of  x1-x5);                             /* desvio padrão */
stderr = stderr(of  x1-x5);                       /* erro padrão da média */
sum = sum(of  x1-x5);                             /* soma  */
uss = uss(of  x1-x5);                  /* soma de quadrados não corrigida */
var = var(of  x1-x5);                             /* variância */
abs = abs(-10);                                   /* valor absoluto */
round = round(10.222);                            /* valor inteiro */
sqrt = sqrt(10);                                  /* raiz quadrada */
log10 = log10(10);                                /* logaritmo de base 10 */
log = log(10);                                    /* logaritmo natural de base  e */
exp = exp(10);                                    / * valor exponencial: ex. e10 */
log2 = log2(10);                                  /* logaritmo de base 2 */
a1 = smallest(1, of x1-x5);                       /* menor valor  */
cards;
10 11 12 13 14
;
proc print;run;
```

x1	x2	x3	x4	x5	css	cv	kurtosis	max	min	mean	nmiss	range	skewness	std	stderr	sum	uss	var	abs	round	sqrt	log10	log	exp	log2	a1
0	11	12	13	14	10	13.1762	-1.2	14	10	12	5	0	4	0	1.58114	0.70711	60	730	2.5	10	10	3.16228	1	2.30259	22026.47	3.32193	10

Figura 1. Output (saída dos resultados do programa SAS.

Funções que retornam probabilidades cumulativas de uma distribuição

Observação: “nc” é o parâmetro de não centralidade e se não for especificado é considerado zero.

Distribuição binomial

probbnml(p,n,m).

p = probabilidade de sucessos ($0 \leq p \leq 1$).

n = número de experimentos ou tentativas independentes de Bernoulli ($n > 0$).

m = número de sucessos ($m = 0, 1, ..., n$).

Distribuição hypergeométrica

probhypr(N,k,n,x).

N = tamanho da população (N = 1, 2, ...).

k = número de itens da categoria de interesse na população (k = 0, 1, ..., n).

n = tamanho amostral (n = 1, 2, ..., n).

x = variável aleatória que indica o número de itens da categoria de interesse na amostra.

Distribuição de Poisson

poisson(m,n).

m = número de sucessos ($m \geq 0$).

n = número de sucessos de experimentos ou tentativas ($n \geq 0$).

Distribuição de qui-quadrado

probchi(x,gl, nc).

x = variável aleatória ($x \geq 0$).

gl = graus de liberdade ($x > 0$).

Distribuição F de Fisher-Snedecor

probf(x,gln,gld,nc).

x = variável aleatória; graus de liberdade do numerador (*gln*) e denominador (*gld*).

Distribuição t de Student

probt(t,gl,nc).

t = variável aleatória numérica.

gl = graus de liberdade.

Distribuição normal padrão

probnorm(x).

x = variável aleatória.

Exemplo:

```
/* funções que retornam probabilidades cumulativas de uma distribuição */
data _null;
binomial = betainv(0.5,20,6);                    /*p= 0.5; n=20; m =6) */
hyper = probhypr(100,30,10,2);                   /* n=100; k=30; n=10; x=2) */
poisson = poisson(1,2);                          /* m = 1; n = 2 */
qui = probchi(5.50,20, 0);                       /* x  = 5.50; gl = 20 ; nc = 0 */
f = probf(4.73,10,5,0);                          /* f = 4.73; gln=10; gld = 5; nc = 0 */
t = probt(2.40,9,0);                             /* t = 2.40; gl = 9; nc = 0 */
normal = probnorm(1.96);                         /* x = 1.96 */
title " ==== resultados das distribuições ==== ";
put  binomial = hyper = poisson = qui = f = t = normal =;
run;
```

Log do SAS: resultado das distribuições descritas no programa SAS acima.

```
binomial = 0.7762090862 hyper = 0.3728565944 poisson = 0.9196986029 qui
=0.0005762219 f = 0.9498928087 t = 0.9800510567 normal = 0.9750021049
```

Funções que retornam o percentil de uma distribuição

Observação: "nc" é o parâmetro de não centralidade e se não for especificado é considerado zero.

be*tainv (p,a,b)

Retorna o p-ésimo percentil da distribuição beta com parâmetros: média (a) e variância (b).

cinv (p,gl<,nc>).

Retorna o p-ésimo percentil da distribuição de qui-quadrado com graus de liberdade gl.

finv (p,gln,gld,nc).

Retorna o p-ésimo percentil da distribuição F com graus de liberdade gln (numerador) e gld (denominador).

gaminv(p,a)

Retorna o p-ésimo percentil da distribuição Gamma.

probit(p)

Retorna o p-ésimo percentil da distribuição normal padrão.

tinv(p,gl<,nc>)

Retorna o p-ésimo percentil da distribuição t com gl graus de liberdade.

Exemplo:

```
/*
argumento
p = probabilidade do erro do tipo 1 (α =0,05)
a = média; b = variância
nc = parâmetro de não centralidade (nc=0)
gl = graus de liberdade; gln = gl do numerador; gld = gl do denominador
*/
data;
input  a b p gl  gln  gld  nc;
beta      = betainv(p,a,b);
cinv      = cinv(p,gl,nc);
finv      = finv(p,gln,gld,nc);
gama      = gaminv(p,a);
probit    = probit(p);
tinv      = tinv(p,gl,nc);
datalines;
2 3 0.95 20  15 30  0
;
proc print;
var beta cinv finv gama probit tinv;
run;
output
 beta          cinv       finv        gama       probit      tinv
 0.75140      31.4104   2.01480   4.74386   1.64485   1.72472
```

Estruturas dos programas

Os programas SAS são divididos por dois passos:

< *data step* >: Etapas da criação de um arquivo de trabalho.

< *proc step* >: Etapas da chamada de procedimentos para execução.

Em que *proc* é abreviatura para *procedure*, que significa procedimento, e *step,* etapas.

Enfim, no sistema SAS a combinação de *data Steps* e *Proc Steps* gera um programa no SAS. Nos exemplos a seguir, serão apresentados alguns procedimentos dentro da estrutura de um programa SAS.

proc sort

Esse procedimento ordena uma ou mais variáveis, numérica ou alfanumérica, de um arquivo, em ordem crescente ou decrescente.

Exemplo: os procedimentos abaixo a partir do arquivo Ferrugem cria o arquivo Ferrugem1 com a variável “ms” classificada em ordem crescente. Usando a opção <*descending*>, cria o arquivo ferrugem1 com a variável “ms” classificada em ordem decrescente.

```
proc sort data = ferrugem out = ferrugem1;
by <descending> ms;
 run;
```

proc means

O procedimento Means do SAS calcula várias estatísticas de uma amostra de dados: $x_1, x_2, ..., x_n$, como demonstrado na Tabela 2.

Tabela 2. Estatísticas associadas a uma amostra de dados: $x_1, x_2, ..., x_n$.

Estatística	O que retorna
N	Número de observações usadas
Nmiss	Número de observações com valores perdidos
Mean	Média aritmética
Min	Valor mínimo
Max	Valor máximo
Range	Amplitude
Sum	Soma dos dados
Var	Variância
Std	Desvio-padrão
Stderr	Erro-padrão da média
Css	Soma de quadrados corrigida
Uss	Soma de quadrados não corrigida
Skewness	Simetria da distribuição dos dados
Kurtosis	Curtose ou pico da distribuição dos dados
Median	Mediana ou percentil 50% ou segundo quartil
p25	q1, percentil 25% ou primeiro quartil
p50	Percentil 50% ou segundo quartil
p75	Percentil 75% ou terceiro quartil
p1, p5, p10, p90, p95, p99	Percentis, 1%, 5%, 10%, 90%, 95%, 99%
Qrange	Amplitude do quartil

```
proc means data= <arquivo>
n nmiss mean min max range sum var std stderr css uss skewness kurtosis median p1 p5
p10 p25 p50 p75 p90  p95 p99 qrange;
run;
```

proc freq.

O procedimento *freq* produz tabelas de frequência simples e tabelas de contingência, r x c, organizada com dados categóricos consistindo de *r* linhas e *c* colunas (i = 1, 2, ... , r; j = 1, 2, ... , c), em que cada combinação i,j de r x c representa uma célula. Esse procedimento calcula várias estatísticas dentro de cada célula e através das células (linhas e colunas). Abaixo são mostrados alguns tipos de usos do *proc freq* e maiores detalhes de aplicações podem ser vistos no Capítulo 13.

proc freq < options > ;
by variables;
exact statistic-options < / computation-options > ;
output < out = SAS-data-set > options ;
tables requests < / options > ;
test options ;
weight variable < / option >;

proc freq < options > ;

< options >
< data = >

Especifica o arquivo de dados.

< compress >

Inicia a próxima tabela simples na página atual.

< nlevels >

Mostra o número de níveis para todas as variáveis da tabela.

< noprint >

Suprime toda a saída.

< order = >

Especifica a ordem para listar os valores das variáveis.

< page >

Mostra uma tabela por página.

order = < data | formatted | freq | internal > < data >

Ordena os valores de acordo com a ordem do Input.

< formatted >

Ordena os dados de acordo com os valores formatados.

< freq >

Ordena os valores de acordo com a frequência de contagem em ordem descendente.

< internal >

by variáveis;

Para realizar análise separada dentro de cada variável (por grupo), que geralmente é de classificação, é importante que essas variáveis tenham sido ordenadas antes pelo procedimento *proc sort* do SAS.

exact statistic-options < / computation-options >;

O *proc freq* calcula testes exatos com algarismos rápidos e eficientes. Esses testes são apropriados quando o arquivo é pequeno, assimétricos, de afastamento de curtose e de simetria. Cálculos de testes exatos requerem grande capacidade de memória e tempo. Testes exatos podem ser requeridos com uma das opções abaixo:

< agree >

Teste de McNemar para tabelas 2 x 2, coeficiente de Kappa simples e coeficiente de Kappa ponderado.

< binomial >

Teste de proporção binomial para tabelas de uma entrada (*one-way tables*).

< chisq >

Calcula teste de qui-quadrado (χ^2) para tabelas de uma entrada (*one-way table*), χ^2 de Pearson, χ^2 para razão de verossimilhança e teste de χ^2 de Mantel-Haenszel para tabelas 2 x 2.

< comor >

Calcula IC para razão de probabilidade para tabelas hx2x2.

< Fisher >

Teste exato de Fisher.

< JT >

Teste de Jonckheere-Terpstra.

< kappa >

Teste de coeficiente de Kappa simples.

< lrchi >

Teste de χ^2 de razão de verossimilhança.

< mcnem >

Teste de McNemar.

< measures >

Teste de correlação de Pearson, de Spearman e IC para razão de probabilidade para tabelas 2 x 2.

< mhchi >

Teste de χ^2 de Mantel-Haenszel.

< or >

Calcula IC para tabelas de razão de probabilidades 2 x 2.

< pchi >

Calcula teste de χ^2 de Pearson.

< pcorr >

Calcula teste para coeficiente de correlação de Pearson.

< scorr >

Calcula teste para coeficiente de correlação de Spearman.

< trend >

Teste de Cochran-Armitage.

< wtkap >

Teste de coeficiente de Kappa ponderado.

< / computation-options >

Das opções, a mais comumente usada é:

< alpha = α >

Especifica o valor de α para construir limites de confiança 100 (1 - α) de estimativas obtidas por meio de Monte Carlo. O *default* (usual) é α = 0,01. O limite de confiança é dado por 100(1- α)%.

< mc >

Requer estimação de Monte Carlo para os valores de p exatos ao invés de utilizar aproximações assintóticas que requerem muito tempo e muita memória.

output < out=SAS-data-set > options;

Cria um arquivo contendo estatísticas calculadas pelo *proc freq*, as quais o usuário seleciona em < *options* >. A lista de estatísticas é grande.

Por exemplo:
agree, binomial, chisq, cmh, fisher, exact, kappa.

tables requests < / options > ;

< requests >

O usuário tem várias opções para construir tabelas de contingência (Tabela 3).

Tabela 3. Opções para construir tabelas de contingência.

Sintaxe	Equivalência
tables A*(B C);	tables A*B A*C;
tables (A B)*(C D);	tables A*C B*C A*D B*D;
tables (A B C)* D;	tables A*D B*D C*D;
tables A - - D;	tables A B C D;
tables (A - - C)*D;	tables A*D B*D C*D;

Não usando < options >

O *proc freq* produz para tabelas simples apenas frequência, percentagem, frequência e percentagem acumulada para cada valor da variável. Para tabelas de dupla entrada, calcula em cada célula a frequência, percentagem, percentagem na linha e percentagem na coluna para os valores de cada variável.

Com a opção < options >

Produz vários testes. Por exemplo:
agree, all, binomial, chisq, cmh, fisher, exact, kappa test options, etc.

weight variable < / option > ;

O *weight* especifica que a variável numérica tem um valor que representa a frequência de dados da observação, célula ou categoria.

Exemplo: dois juízes (1 e 2) avaliaram um grupo de 44 touros nas pistas de exposições agropecuárias para finalidade de premiação. Cada touro então é avaliado pelos dois juízes que dizem "sim" ou "não" para ganhar prêmio (Tabela 4).

Tabela 4. Tabela de contingência 2 x 2 da avaliação de touros por dois juízes.

Juiz 1	Juiz 2	
	Sim	Não
Sim	18	5
Não	6	15

No programa a seguir, a variável *weight* representa a frequência de cada observação da variável resposta (resp).

```
data;
input juiz1 $ juiz2 $ resp;
cards;
sim sim 18
sim nao  5
nao sim  6
nao nao 15
;
proc print;
proc freq;
title "teste exato de fisher e teste de kappa";
tables juiz1*juiz2/fisher kappa; /* gera a tabela 1 com as respectivas estatísticas do teste
de fisher e de kappa*/
weight resp;
run;
```

proc glm

O procedimento *glm* do SAS é utilizado para ajustar modelos lineares gerais pelo método dos quadrados mínimos. Entre os métodos estatísticos disponíveis no *proc glm*, têm-se: regressão, análise de variância, análise de covariância, análise de variância multivariada e correlação parcial.

O modelo linear geral em análise univariada, na forma matricial é:

$$y_{nx1} = X_{nxp}\, b_{px1} + e_{nx1}$$

em que:

y_{nx1} = vetor de valores dependentes.

$E(y) = Xb$; $Var(y) = V(e) = R = \sigma^2 I_n$.

σ^2 = quadrado médio do erro.

I_n = matriz de identidade de ordem *n*.

X = matriz de especificação.

b = vetor que contém os efeitos fixos.

e = vetor que contém os erros aleatórios, E(e) = 0.

O glm exige que os erros e_{ijk} sejam independentes e identicamente distribuídos com média zero e com distribuição pelo menos aproximada da normal, ou seja, $V(e) = R = \sigma^2 I_n$. Isso pressupõe variância constante na diagonal principal e correlação nula para os elementos fora da diagonal.

Os seguintes comandos abaixo são disponíveis no *proc glm*.

proc glm < opções > ;
class variáveis < / opção > ;
model variáveis dependentes= variáveis independentes < / opções > ;
absorb variáveis;
by variáveis ;
freq variável;
id variáveis ;
weight variável;
contrast 'nomes dos valores dos efeitos' < . valores dos efeitos > < / opções > ;
estimate 'nomes dos valores dos efeitos' < . valores dos efeitos > < / opções > ;
lsmeans efeitos < / opções > ;

manova < testes-opções >< / detalhes- opções > ;
means efeitos < / opções > ;
output < out=arquivo SAS >
keyword=names < ... keyword=names > < / opção > ;
random efeitos < / opções > ;
repeated fator de especificação < / opções > ;
test < h= efeitos > e= efeitos < / opções > ;

proc glm < opções >;

< alpha=p >

Especifica o limite de confiança para *p* para o intervalo de confiança 100(1-*p*)%; o *default* é 0,05.

< data = nome do arquivo >

Indica o nome do arquivo a ser usado. Se omitido, usa-se o último.

< manova >

Nas variáveis dependentes que serão usadas em análises multivariadas, as observações com valores perdidos serão eliminadas da análise.

< noprint >

Não imprime as observações e comentários que normalmente são impressos juntos com os resultados das análises.

< order=data | formatted | freq | internal >

Especifica o tipo de ordem das variáveis de classificação.

< data >

Mantém a ordem dos dados que aparece no arquivo.

< formatted >

Mantém valores formatados externamente, exceto para variáveis numéricas não explicitamente formatadas. Estas mantêm seus valores formatados internamente.

< freq >

Variáveis de classificação com maiores frequências em seus níveis são consideradas primeiro.

< outstat = nome do arquivo >

Cria e dá nome a um arquivo que contém estatísticas calculadas.

class variáveis;

No comando *class* são colocadas as variáveis de classificação que são utilizadas na análise (variáveis definidas como fonte de variação ou causa de variação de uma análise de variância), podendo ser numérica ou não numérica. Geralmente essas variáveis estão associadas com o delineamento experimental e indicam a coordenada de uma unidade experimental, de um indivíduo e de uma observação no arquivo de dados.

model variáveis dependentes = variáveis independentes < / opções > ;

< / opções >

< intercept >

Produz testes para o intercepto de um modelo.

< noint >

Omite o intercepto de um modelo.

< solution >

Produz solução para as estimativas de parâmetros das equações normais.

< E >

Mostra a forma geral de todas as funções estimáveis.

< E1, E2, E3, E4 >

Mostra, respectivamente, as somas de quadrados (SQ) do tipo I, II, III e IV.

< cli >

Produz IC para os valores preditos de cada observação.

< clm >

Produz IC para a média dos valores preditos de cada observação.

< clparm >

Produz IC para as estimativas de parâmetros e para todos os resultados envolvendo estas estimativas.

< nouni >

Suprime os resultados de estatísticas univariadas.

< SS1 SS2 SS3 SS4 >

Quando está trabalhando com análise multivariada produz SQ para as funções estimáveis do tipo I, II, III e IV, respectivamente.

absorb < variável;

É uma técnica usada para reduzir recursos computacionais. Ela permite absorver uma variável na análise.

by variáveis;

Esse comando é usado quando se deseja análise separada dentro de cada variável, que geralmente é de classificação, como raça, sexo, local, etc., e que já foi definida no Input. É importante que essas variáveis tenham sido ordenadas antes.

freq variável;

Indica uma coluna no arquivo de dados que será usada apenas como frequência.

id variáveis;

Para cada variável identificada em " id ", são calculados valores observados, predictos e residual. Essas variáveis são incluídas em dois arquivos: < *outp* = arquivo 1 > que inclui os efeitos aleatórios predictos e < *outpm* = arquivo 2 > que inclui somente os efeitos fixos.

weight variável;

Quando *weight* é atribuído a uma variável, esse fator de ponderação ou peso participa de alguns cálculos como a soma de quadrado residual (SQR) que é calculada por $\sum_i w_i(y_i - \hat{y}_i)^2$, em que w_i é o valor da variável especificada. Também nas comparações múltiplas pelo *glm*, quando especificamos *weight* a variância da diferença entre médias de grupos i e j é calculada como:

QMR$(1/n_i + 1/n_j)$, em que n_i e n_j é o tamanho do grupo i e j, respectivamente.

contrast 'nomes dos valores dos efeitos' < valores dos efeitos > < / opções > ;

Este comando possibilita calcular contrastes de interesse e realizar testes de hipóteses. Nas análises univariadas permite testar a hipótese $l\beta = 0$ e nas análises multivariadas, testa a hipótese $l\beta m = 0$. Exemplos de contrastes são apresentados nos Capítulos 7 e 8.

estimate 'nomes dos valores dos efeitos' < valores dos efeitos > < / opções > ;

Permite estimar funções lineares dos parâmetros pela multiplicação do vetor *l* pelo vetor de estimativas dos parâmetros β, gerando $l\beta$.

lsmeans < efeitos fixos > < / opções >;

Quando se realiza uma análise de variância com a opção *lsmeans*, calculam-se as médias obtidas por quadrados mínimos a partir dos < efeitos fixos > que são colocados na opção *class.*

No item < opções >, realiza testes de hipóteses entre as médias por meio dos testes a seguir.

</ opções >
adjust = bonferroni.
adjust = dunnett.
adjust = scheffe.
adjust = sidak.
adjust = smm | gt2.
adjust = tukey.
< pdiff = all > Requer comparações pareadas dos testes acima.

A opção "*stderr*" produz erros-padrão das médias obtidas por quadrados mínimos.

Manova < opções de testes > < / detalhes-opções > ;

Com a opção *Manova* realiza análise multivariada para testar hipótese global:

- Ho = os tratamentos não diferem entre si versus Ha = os tratamentos diferem entre si. Nesta hipótese global, o conjunto de variáveis dependentes ou variáveis resposta é analisado de uma só vez, sendo que qualquer observação com valor perdido é excluída da análise. Esta opção é usada para duas ou mais variáveis resposta (y_1, y_2, ...).

< opções de testes >

Várias opções podem ser utilizadas em *Manova* para definir quais testes serão utilizados:

A hipótese multivariada do modelo linear geral é:

$$l\beta m = 0 \text{ versus } l\beta m \neq 0$$

em que: l corresponde ao tipo de teste usado (I, II, III ou IV); m é a matriz identidade p x p e β é a matriz de parâmetros.

Os testes usados para testar a hipótese multivariada do modelo linear geral assim como o cálculo do teste F utilizado são descritos no Capítulo 8.

O numerador e o denominador no cálculo do teste F é dado por:

$H = m'(lb)'(l(x'x)^{-}l')^{-1}(lb)m.$
$E = m'(y'y - b')(x'x)b)m.$
h=effects | intercept | _all_ .
< effects >

São colocados os efeitos os quais serão utilizados na análise e na formação das matrizes de hipóteses.
< intercept >

É usado para produzir testes para o intercepto.
< _all_ >

É usado para produzir testes para todos os efeitos listados no modelo.

< / detalhes-opções >

< canonical >

Realiza uma análise canônica das matrizes H e E, transformada pela matriz M se especificada.

< etype=n >

Especifica o tipo de teste (Tipo I, II, III ou IV) para a matriz E.

< htype=n >

Especifica o tipo de teste (Tipo I, II, III ou IV) para a matriz H.

< mstat=fapprox >

Especifica que a análise multivariada será avaliada usando a aproximação usual do teste F.

Raiz máxima de Roy = λ = maior valor de ($e^{-1}h$), o cálculo não é exato.

< mstat=exact >

Especifica que a análise multivariada será avaliada usando cálculos exatos para os testes.

Os valores de probabilidade p são calculados exatos para os três testes:

Lambda de Wilks = det(e)/det(h+e)

Traço de Lawley Hotelling = traço($e^{-1}h$)

Raiz máxima de Roy = λ = maior valor de ($e^{-1}h$)

O cálculo exato para este teste proporciona significante melhoria na análise. Usando o método *default* (*mstat=fapprox*), o erro do tipo I para aceitar e ou rejeitar a hipótese não é exato.

No caso do quarto teste (IV), o cálculo é mais preciso:

Traço de Pillai = traço($h(h + e)^{-1}$).

< orth >

Requer uma transformação ortonormal para a matriz *m* antes da análise (ver programa 7 no Capítulo 8).

< printe >

Calcula a soma de quadrados de produtos cruzados da matriz E.

means efeitos < / opções > ;

< efeitos >

Médias aritméticas e desvios-padrão são calculados para cada efeito ou variável de classificação.

< / opções >

Em < / opções >, pode colocar um ou mais dos testes abaixo para realizar comparações múltiplas entre as médias de cada efeito.

T
duncan
snk

regwq
tukey
bon
sidak
scheffe
Gabriel

< out=SAS-data-set >

Possibilita criar um novo arquivo, o qual contém todas as variáveis do arquivo original e ainda as novas variáveis e outras estatísticas calculadas com a análise de variância. Este arquivo facilita o usuário calcular novas estatísticas e principalmente construir gráficos.

< ...keyword=names >

Especifica as estatísticas a serem incluídas no *SAS-data-set* e fornece nomes para essas variáveis.

Alguns exemplos de < names >

< lcl >

Intervalo de confiança (IC) para uma predição individual.

< lclm >

Limite inferior do IC para um valor esperado (média).

< predicted | p >

Valores predictos.

< ucl >

Limite superior do IC para uma predição individual.

< uclm >

Limite superior do IC para o valor esperado (média) do valor predito.

random efeitos < / opções > ;

< efeitos >

Quando alguns efeitos especificados em < *class* > são aleatórios, podemos especificar esses efeitos após <*random*>, para calcular valores esperados dos quadrados médios de uma análise de variância, contrastes e também para realizar testes de efeitos aleatórios na análise de variância. Ver exemplo de aplicação no Capítulo 8.

< opções >
< q >

Calcula formas quadráticas dos efeitos fixos que aparecem nas esperanças dos quadrados médios de uma análise de variância – E(QM).

Em uma análise do tipo:

model Y=A B(A) C A*C,
random B(A),

com B(A) declarado como aleatório, a esperança do quadrado médio de cada efeito é apresentada como:

Var (erro) + constante x Var(B(A)) + Q(A, C, A*C).

Os valores verdadeiros da forma quadrática da matriz Q podem ser apresentados usando a opção Q.

Um exemplo de E(QM) foi produzido pelo programa 3, no Capítulo 8.

< test >

Calcula teste de hipótese para cada efeito especificado no modelo.

repeated fator de especificação < / opções > ;

Utilizado em análises de variância de medidas repetidas (MR). Exemplos são apresentados no Capítulo 8.

test < h= efeitos > e= efeitos < / opções > ;

Utilizado para especificar o teste de um efeito na análise de variância com respectivo quadrado médio residual. O exemplo abaixo foi obtido de um experimento em blocos casualizados, descrito em Freitas et al. (2011), em que se avaliou a produção de matéria seca e os tratamentos (trat) correspondem a genótipos de alfafa e o "tempo", a cortes sequenciais. Para testar o efeito de tratamento "trat", o erro foi "blocos(trat)".

```
proc glm;
 class blocos trat  tempo;
 model PMS = trat blocos(trat) tempo trat*tempo;
test h = trat  e = blocos(trat);
run;
```

proc mixed

O procedimento *mixed* é uma generalização do modelo linear padrão usado no procedimento *glm.* Porém, ele permite não somente modelar as médias dos dados, mas também modelar as variâncias e covariâncias.

Os seguintes comandos abaixo são disponíveis no *proc mixed*.

proc mixed *< opções > ;*
by *variáveis;*
class *variáveis;*
id *variáveis;*
model *dependente = < efeitos fixos > < / opções >;*
random *efeitos aleatórios < / opções >;*
repeated *< efeitos repetidos > < / opções >;*
parms *(lista de valores) < / opções >;*
prior *< distribuição > < / opções >;*
contrast *'label' <valores de efeitos fixos> <|valores de efeitos aleatórios >, ... </opções >;*
estimate *'label' <valores de efeitos fixos> < | valores de efeitos aleatórios >< / opções >;*
lsmeans *efeitos fixos < / opções >;*
weight *variável;*

proc mixed < opções > ;

< opções >

Uma das primeiras opções refere-se às estimativas de parâmetros. No procedimento *mixed,* elas são obtidas iterativamente por três critérios (*proc mixed convf; proc mixed convg e proc mixed convh),* conforme demonstrado a seguir. Em todos os critérios, o número de tolerância *n default* do SAS é 1E-8, significando que a convergência se estabiliza quando o valor da função de máxima verossimilhança entre duas iterações consecutivas for menor ou igual a 0,00000001.

< proc mixed convf = n;>

Requer o critério de convergência da função relativa com número de tolerância *n*.

$$\frac{|f_k - f_{k-1}|}{|f_k|} < n$$

f_k = valor da função objetivo na iteração *k*. Para evitar divisão por $|f_k|$, deve-se usar a opção *absolute*.

< proc mixed convg = n; >

Requer o critério de convergência relativa gradiente com número de tolerância *n*.

$$\frac{\max_j |g_{jk}|}{|f_k|} \leq n$$

em que: f_k é o valor da função objetivo na iteração *k*, g_{jk} é o j-ésimo elemento do gradiente (primeira derivada) da função objetivo, ambos na iteração *k*.

< proc mixed convh = n; >

Requer o critério de convergência Hessiana relativa com número de tolerância *n*.

$$\frac{g_{k'} H_k^{-1} g_{k'}}{|f_k|} \leq n$$

em que: f_k é o valor da função objetivo na iteração *k*, g_k é o gradiente (primeira derivada) da função objetivo, e H_k é a Hessiana (segunda derivada) da função objetivo, todos na iteração *k*.

< proc mixed alpha = 0.05;>

Valor de α para construir limites de confiança das estimativas dos parâmetros de covariância; α deve se situar entre 0 e 1; o *default* é 0,05.

< proc mixed info; >

É uma opção *default*. Ela fornece tabelas contendo "*Model Information", "Dimensions*", and "*Number of Observations*": se não quiser, usa a opção "*Noinfo"*.

< proc mixed lognote; >

Informa no *SAS log* o status dos cálculos.

< proc mixed maxfunc = valor >

Em < valor > é especificado o número máximo de avaliações por verossimilhança.

< proc mixed maxiter= valor >

Em < valor > é especificado o número máximo de iterações. O *default* é 50.

< proc mixed method = reml; >

Especifica o método de estimação para os parâmetros de covariância. O *default* é reml (*restricted maximum likelihood*). Pode usar também:

ml.
mivque0.
type1.
type2.
type3.

< proc mixed covtest; >

Produz erros-padrão assintóticos e testes Z de Wald para estimativas de parâmetros de covariância. Para isso, requer grande tamanho amostral (pelo menos 400 indivíduos são recomendados).

< proc mixed asycorr; >

Produz a matriz de correlação assintótica das estimativas dos parâmetros de covariância. O *default* é a inversa da matriz de informação de Fisher que é igual a $2H^{-1}$, em que H é a matriz Hessiana (segunda derivada) da função objetivo. Para isso, requer grande tamanho amostral (pelo menos 400 indivíduos são recomendados).

< proc mixed asycov; >

Produz a matriz de covariância assintótica das estimativas dos parâmetros de covariância.

< proc mixed itdetails; >

Mostra os valores e parâmetros a cada iteração e facilita a gravação de notas no *SAS log* com relação à informação do tipo "*infinite likelihood*" e "*singularities*" durante iteração por Newton-Raphson.

< proc mixed cl; >

Calcula intervalo de confiança das estimativas dos parâmetros de covariância.

< proc mixed ic; >

Mostra a tabela dos vários critérios de informação:

$AIC = -2l + 2d$

$AICC = -2l + 2dn/(n-d-1)$

$HQIC = -2l + 2d\ loglog(n)$

$BIC = -2l + d\log n$

$CAIC = -2l + d(\log n + 1)$

Em que:

l = valor máximo do log da função de verossimilhança.

d = dimensão do modelo.

n = número de observações.

by variáveis;

Este comando é usado quando se deseja análise separada dentro de cada variável, que geralmente é de classificação, como raça, sexo, local, etc., e que foi colocada no Input. É importante que essas variáveis tenham sido ordenadas antes.

class variáveis;

No comando *class* são colocadas as variáveis de classificação que são utilizadas na análise, podendo ser numérica ou não numérica. Geralmente essas variáveis estão associadas com o delineamento experimental e indicam a coordenada de uma unidade experimental, de um indivíduo e de uma observação no arquivo de dados.

id variáveis;

Para cada variável identificada em "*id*" são calculados valores observados, predictos e residual. Essas variáveis são incluídas em dois arquivos: < outp = arquivo 1 >, que inclui os efeitos aleatórios predictos; e < *outpm* = arquivo 2 >, que inclui somente os efeitos fixos.

model dependente = < efeitos fixos > < / opções > ;

" dependente ">

Aqui é colocado o nome da variável dependente ou variável resposta que será analisada.

< efeitos fixos >

Nesse item são colocados os efeitos na ordem em que aparecem em uma análise de variância.

< / opções >

Várias estatísticas são calculadas em < / opções >.

Alguns exemplos:

< alpha=number >

Requer que um intervalo de confiança do tipo t seja calculado para cada parâmetro de efeito fixo. O valor do número (*number*) precisa estar entre 0 e 1. O usual (*default*) é 0.05.

< alphap=number >

Requer que um intervalo de confiança do tipo t seja calculado para valores predictos. O valor do número (*number*) precisa estar entre 0 e 1. O usual (*default*) é 0.05.

E

Requer que as matrizes de coeficientes *l* do tipo I, tipo II e tipo III, simultaneamente, sejam calculadas para todos os efeitos fixos especificados. Uma matriz *l* específica para tipo I ou tipo II ou tipo III é calculada desde que o E seja substituído por E1 ou E2 ou E3, respectivamente.

< chisq >

Requer que um teste de χ^2 seja calculado para todos os efeitos especificados em adição ao teste F. O usual (*default*) é o tipo III, mas pode ser escolhido testes do tipo I e do tipo II usando a opção htype = E1 ou htype = E2.

< cl >

Requer que limites de confiança usando a estatística *t* sejam construídos para cada das estimativas dos parâmetros de efeitos fixos. O nível de confiança usual (*default*) é 0,95.

< corrb >

Produz uma matriz de correlação aproximada das estimativas de parâmetros dos efeitos fixos.

< solution >

Produz uma solução para os parâmetros de efeitos fixos, representada pela estimativa do vetor β ($\hat{\beta}$), cujos erros-padrão são estimados por $l\,(X'V^{-1}X)^{-}l'$, em que *l* é a matriz de hipótese e V é a matriz de variâncias e covariâncias.

< ddf = Lista de valores >

Possibilita especificar os seus próprios graus de liberdade (gl) do denominador para os efeitos fixos. Os graus de liberdade devem ser colocados, separados por vírgula, na ordem que os efeitos fixos aparecem no procedimento "*Tests of Fixed Effects*" produzida pelo SAS. Se não quiser colocar grau de liberdade para determinado efeito, utilize valor perdido " . " no lugar da vírgula.

Considerando-se uma situação fictícia de um delineamento em blocos casualizados onde se analisa a produtividade de matéria seca (PMS) coletadas de 30 genótipos de alfafa (*Medicago sativa* L.), que são os tratamentos (trat) com 20 cortes no tempo. O programa abaixo mostra como se declara graus de liberdade (gl) do denominador no procedimento mixed do SAS. No exemplo mostra o uso de gl de 25, 567 e 540 para tratamento (trat), corte e a interação trat × corte.

```
proc mixed;
class bloco trat corte;
model pms = trat corte trat*corte/ddf = 25,567,540;
repeated corte/sub = bloco(trat) type = arma(1,1);
run;
```

random efeitos aleatórios < / opções > ;

A declaração *random* define os efeitos aleatórios que constituem o vetor de parâmetros desconhecidos *u* do modelo misto, que podem ser variáveis de classificação ou contínuas. Ela pode ser usada para especificar modelos de componentes de variâncias tradicionais e os coeficientes aleatórios. Várias opções *random* simultaneamente são possíveis. No modelo misto, o propósito da opção *random* é definir a matriz *Z,* os efeitos aleatórios do vetor *u* e a estrutura de G = Var(u) que é definida por *type* = *option*. Pode-se especificar o *intercept* (ou *int*) como um efeito aleatório para indicar o intercepto. Por *default* o *proc mixed* não inclui o intercepto na declaração *random*.

repeated < efeito repetido > < /subject=< indivíduo > type = < estrutura > rcorr vcorr;

< efeito repetido >

Indica o efeito que é avaliado no tempo em cada indivíduo.

< subject = opção >

Define os indivíduos e cria blocos de diagonalidade da estrutura R definida após "Type".

< type = estrutura de covariância >

Especifica a estrutura da matriz de variância e covariância R a ser testada.

< rcorr >

Calcula a matriz de correlação entre as medidas repetidas.

< vcorr >

Calcula a matriz de covariância entre as medidas repetidas.

Exemplo do comando Repeated:

```
proc mixed data = alfafa  method = reml;
class bloco trat corte ;
model pms = trat corte trat*corte;
repeated corte /sub = bloco(trat) type = cs rcorr; run;
```

No programa acima após '*type* =' foi usada a estrutura de matriz de variância e covariância cs (*compound symmetry*). As demais estruturas são descritas na Tabela 5.

Tabela 5. Estruturas de variâncias e covariâncias.

Estrutura	Descrição	Parâmetros*
ante(1)	*ante-dependence*	2t-1
ar(1)	*autoregressive(1)*	2
arh(1)	*heterogeneous r(1)*	t+1
arma(1,1)	*arma(1,1)*	3
Cs	*compound symmetry*	2
Csh	*heterogeneous cs*	t+1
fa(q)	*factor analytic*	[q/2](2t -q + 1) + t
fa0(q)	*no diagonal fa*	[q/2](2t -q + 1)
fa1(q)	*equal diagonal fa*	[q/2](2t -q + 1) + 1
Hf	*huynh-feldt*	t+1
lin(q)	*general linear*	q
Toep	*Toeplitz*	t

Continua...

Tabela 5. Continuação.

Estrutura	Descrição	Parâmetros*
toep(q)	*banded toeplitz*	q
Toeph	*heterogeneous toep*	2t-1
toeph(q)	*banded hetero toep*	t+q-1
Un	*unstructured*	t(t+1)/2
un(q)	*Banded*	[q/2](2t-q+1)
Unr	*unstructured corrs*	t(t+1)/2
unr(q)	*banded correlations*	[q/2](2t-q+1)
un@ar(1)	*direct product ar(1)*	$t_1(t_1+1)/2 + 1$
un@cs	*direct product cs*	$t_1(t_1+1)/2 + 1$
un@un	*direct product un*	$t_2(t_2+1)/2 - 1 + t_2(t_2+1)/2 - 1$
Vc	*variance components*	q

Parâmetros* = número de parâmetros de covariâncias na matriz.
"t" = dimensão global da matriz de covariância.
"q" = número de fatores.
t_1 e t_2 = em um produto direto de duas matrizes, referem-se à primeira e à segunda estrutura.

parms < lista de valores > ... < / opções > ;

Especifica valores iniciais para os parâmetros de covariância.

< lista de valores >

Pode ser um valor (m) ou vários valores (m_1, m_2, ... , m_n).

No exemplo a seguir três componentes de variâncias são conhecidos (60, 20 e 6). A opção < noiter > evita quaisquer iterações pelo método de Newton-Raphson, de modo que os resultados são baseados nestes componentes de variâncias.

```
proc mixed data = sp noprofile;
    class block a b;
    model y = a b a*b;
    random block a*block;
    parms (60) (20) (6) / noiter;
  run;
```

prior < distribuição > < / opções >;

Utilizado quando se realiza análise bayesiana, tema não abordado neste livro.

contrast 'label' < fixed-effect values ... > <random-effect values ... > , ... < options > ;

Possibilita realizar testes de hipóteses de interesse particular do usuário. Contrastes possibilitam testar hipótese do tipo:

$l'\phi = 0 \text{ onde } l' = K'M \text{ e } \Phi' = b'u'$

em que:

l = matriz de hipótese; a escolha de *l* é o que determina o nosso interesse particular.

M = matriz que determina o espaço de inferência. Se M = 0, a inferência é a população inteira de onde os efeitos aleatórios foram amostrados. Quando os elementos de M são diferentes de zero, nossa inferência são os níveis observados dos efeitos aleatórios.

O cálculo de um contraste inicia com inferências estatísticas obtidas por testar a hipótese:

$$H: l \begin{bmatrix} \hat{b} \\ \hat{u} \end{bmatrix} = 0$$

As estimativas de *b* e *u* são obtidas por:

$$\hat{b} = (X'\hat{V}^{-1}X)^{-1} X'\hat{V}^{-1}y$$

$$\hat{u} = \hat{G}Z'\hat{V}^{-1}(y - X\hat{b})$$

Quando *l* consiste de apenas uma linha e normalidade assumida para *u* e *e*, pode ser calculado o teste t, o qual em geral tem distribuição normal aproximada e o grau de liberdade precisa ser estimado utilizando os vários recursos em <DDFM =>:

$$t = \frac{l\begin{bmatrix} \hat{b} \\ \hat{u} \end{bmatrix}}{\sqrt{l\hat{C}l'}}$$

Aqui C é uma estimativa da inversa generalizada da matriz de coeficientes das equações de modelos mistos.

O intervalo de confiança é calculado por:

$$l\begin{bmatrix}\hat{b}\\ \hat{u}\end{bmatrix} + t_{v.\alpha/2}\sqrt{l\hat{C}l'} = 0\ ;\ t_{v.\alpha/2} = (1-\alpha/2)100\% \quad \text{percentil da distribuição t.}$$

Se o posto de $l > 1$, constrói uma distribuição F aproximada, cujo grau de liberdade do numerador é o posto(l) e o grau de liberdade do denominador é υ. As estatísticas t e F são utilizadas para fazer inferências sobre os efeitos fixos.

$$F = \frac{\begin{bmatrix}\hat{b}\\ \hat{u}\end{bmatrix}'(l'\hat{C}l)^{-1}l\begin{bmatrix}\hat{b}\\ \hat{u}\end{bmatrix}}{\text{posto}(l)}$$

Erro-padrão da média (EP).

$$EP = \sqrt{l(X'V^{-1}X)^{-}l'}$$

Exemplo: consideremos um exemplo de contraste em um experimento em blocos casualizados, quatro blocos, com parcela dividida (*split-plot*) onde o tratamento principal (A) tem três níveis e o tratamento distribuído na subparcela (B) tem dois níveis. A análise de variância e a estimativa de contraste pelo procedimento Mixed fica:

```
proc mixed;
  class a b bloco;
  model y = a b a*b;
  random bloco a*bloco;
run;
contrast 'a amplo'
        a  1 -1 0    a*b  0.5 0.5 -0.5 -0.5  0     0,
        a  1 0 -1    a*b  0.5 0.5  0     0    -0.5 -0.5 / df = 6;
```

No exemplo l' tem duas linhas e o grau de liberdade do numerador é 2. De acordo com a análise, o grau de liberdade do denominador é 9, mas por opção ele foi alterado para 6.

estimate 'label' < valores de efeitos fixos ... > < | valores de efeitos aleatórios ... > < / opções >;

É idêntico ao < contrast >, exceto que somente uma linha da matriz *l* é usada.

lsmeans efeitos fixos < / opções >;

Quando se realiza uma análise de variância, o *lsmeans* calcula médias obtidas por quadrados mínimos dos <efeitos fixos> que são colocados na opção *class* e em "< / opções >" são declarados testes para realizar testes de hipóteses entre as médias; a opção < pdiff = all > realiza comparações por meio do teste escolhido.

</ opções >
adjust = bon
adjust = dunnett
adjust = scheffe
adjust = sidak
adjust = simulate<(simoptions)>
adjust = smm | gt2
adjust = tukey

< pdiff = all >

Realiza comparações pareadas.

< cl >

Calcula intervalo de confiança para cada média.

< slice = >

Em uma interação significativa permite realizar o desdobramento, isto é, testar o efeito de um tratamento dentro de cada nível do outro. Se, em uma análise de variância, a interação A*B é significativa, para testar o efeito de A dentro de cada nível de B, temos: lsmeans A*B / slice=B.

< df = valor >

Especifica o grau de liberdade para o teste t e intervalo de confiança.

weight variável;

Se no *proc mixed* não usar *repeated*, a opção *weight* é utilizada como no glm. Neste caso, as matrizes X´X e Z´Z são substituídas por X´WX e Z´WZ, onde W é a matriz de pesos na diagonal. Se usar *repeated* então a matriz R é substituída por *l*R*l*, onde *l* é uma matriz diagonal com elementos $W^{-1/2}$. Observações onde os pesos são perdidos ou são negativos não são incluídas na análise.

Recursos disponíveis no *insight*

O *insight* é um módulo para explorar e analisar dados na estrutura *click-point*. Fornece um conjunto de ferramentas exploratórias e analítica que possibilita alto nível de interatividade quando comparado com outros módulos. Com o *insight* é possível explorar vários tipos de gráficos de uma, duas e três dimensões, fixos ou em rotação, com uso de cores para destaque de informações específicas, além de ser praticamente ilimitado quanto ao tamanho do conjunto de dados.

Existem recursos para explorar distribuições univariadas e multivariadas, realizar análises de estatísticas descritivas de um conjunto de dados, análise de variância, análises de vários tipos de regressão, correlações, componentes principais, etc.

Etapas para utilizar o insight

1) Acessar a página principal do SAS (Figura 2).
2) Na janela principal, no topo a esquerda, aparece na barra de ferramentas os Ícones:

FILE EDIT VIEW TOOLS RUN SOLUTION WINDOW HELP

3) Clicar em < Solution >.
4) Clicar em < Analysis >.
5) Clicar em < Interactive Data Analysis >.

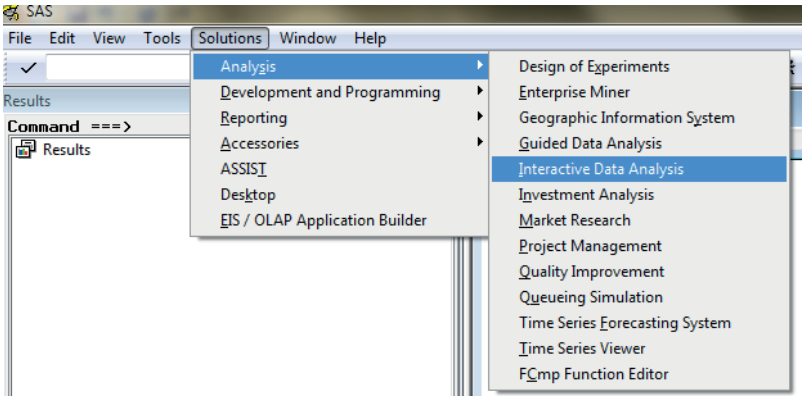


Figura 2. Etapas para utilizar o *insight.*

Em *Library* clicar *Work* e selecionar o arquivo de dados, disponível na área de trabalho, que já deve ter sido executado pelo SAS no comando <*Run*> .

Digite <*Open*> para ter acesso aos dados do arquivo.

6) Clicar em <Analyze> e aparece os módulos para exploração e análise de dados:

Histogram/Bar Chart (Y)
Box Plot/Mosaic Plot (Y)
Line Plot (XY)
Scatter Plot (Y X)
Contour Plot (Z Y X)
Rotating Plot (Z Y X)
Distribution (Y)
Fit (YX)
Multivariate (YX)

Neste momento é só escolher o módulo e ir explorando os dados de forma interativa (no sistema "*click-point*").

Referências

ALLISON, P. D. **Logistic regression using the SAS System**: theory and application. Cary: SAS Institute, 1999.

BIOLOGIA do DNA. Disponível em: https://www.shutterstock.com/pt/image-vector/dna-biology-98546189. Acesso em: 21 set. 2021.

BOYER, C. B.; MERZBACH, U. C. **História da matemática**. 2. ed. São Paulo: Edgard Blucher, 1996. 496 p.

BOZDOGAN, H. Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions. **Psychometrika**, v. 52, n. 3, p. 345-370, 1987. DOI: 10.1007/BF02294361.

CORRÊA, L. de A.; CANTARELLA, H.; PRIMAVESI, A. C.; PRIMAVESI, O.; FREITAS, A. R.; SILVA, A. G. Efeito de fontes e doses de nitrogênio na produção e qualidade da forragem de capim-coastcross. **Revista Brasileira de Zootecnia**, v. 36, n. 4, p. 763-772, 2007.

DUNCAN, D. B. Multiple range and multiple F tests. **Biometrics**, v. 11, n. 1, p. 1-42, 1995. DOI: 10.2307/3001478. Acesso em: 3 abr. 2020.

FISHER, R. A. On the ´probable error` of a coefficient of correlation deduced from a small sample. **Metron**, v. 1, p. 3-32, 1921.

FREITAS, A. R. Curvas de crescimento na produção animal. **Revista Brasileira de Zootecnia**, v. 34, n. 3, p. 786-795, 2005.

FREITAS, A. R.; BARIONI JUNIOR, W.; FERREIRA, R. de P.; CRUZ, C. D.; MOREIRA, A.; VILELA, D. Técnicas de análises exploratórias em dados de cultivares de alfafa. **Revista Brasileira de Zootecnia**, v. 37, n. 9, p. 1531-1536, 2008. DOI: 10.1590/S1516-35982008000900003.

FREITAS, A. R.; DESTEFANI, C. R.; FERREIRA, R. P.; MOREIRA, A. Distribuição de dados de 20 cortes de rendimentos de matéria seca de 92 acessos de alfafa, aplicando análise de medidas repetidas. In: REUNION ASOCIACION LATINOAMERICANA DE PRODUCCION ANIMAL, 20.; REUNION ASOCIACION PERUANA DE PRODUCCION ANIMAL, 30.; CONGRESSO INTERNACIONAL DE GANDEIRA DE DOBLE PROPOSTO, 5., 2007, Cuzco. **Annales** [...] Cuzco: Alpa, 2007. 1 CD-ROM.

FREITAS, A. R; FERREIRA, R. P.; MOREIRA. A. Análises de dados de medidas repetidas por meio do modelo linear geral e do modelo misto. **Revista de Ciências Agrárias**, v. 54, n. 3, p. 214-224, 2011.

FREITAS, A. R.; PERES NETTO, D.; MOREIRA, A.; FERREIRA, R. P.; RODRIGUES, A. Avaliação de controle leiteiro de bovinos usando análises de medidas repetidas. **Revista de Ciências Agrárias**, v. 55, n. 1, p. 5 -10, jan./mar. 2012.

FREITAS, A. R.; PRESOTTI, C. V.; TORAL, F. L. Alternativas de análises em dados de medidas de bovinos de corte. **Revista Brasileira de Zootecnia**, v. 34, n. 6, p. 2233-2244, 2005. Suplemento.

FREITAS, A. R.; RODRIGUES, A. de A.; FERREIRA, R. de P.; MOREIRA, A. Uso do modelo linear e do modelo misto na análise de dados de forragicultura. In: CONGRESSO BRASILEIRO DE ZOOTECNIA, 19., 2009, Águas de Lindóia. **Anais** [...] Águas de Lindóia: FZEA/USP; Uberaba: ABZ, 2009. 1 CD-ROM.

FREITAS, A. R.; SILVA, L. O. C.; JOSAHKIAN, L. A.; ALENCAR, M. M. A qualidade de pesagem de bovinos da raça Nelore In: REUNIÃO ANUAL DA SOCIEDADE BRASILEIRA DE ZOOTECNIA, 37., 2000, Viçosa, **Anais** [...] Viçosa: SBZ, 2000. 1 CD-ROM.

GILL, J. L. **Design and analysis of experiments in the animal and medical science**. Ames: Iowa State University, 1978. 410 p.

HICKS, C. R. **Fundamental concepts in the design of experiments**. New York: Holt, Rinehart and Winston, 1973. 349 p.

LINDSTROM, M. J.; BATES, D. M. Newton-Raphson and EM algorithms for linear mixed-Effects models for repeated-measures data. **Journal of the American Statistical Association**, v. 83, n. 404, p. 1014-1022, 1988.

MARINHO, K. N. S.; FREITAS, A. R.; FALCÃO, A. J. S.; DIAS, F. E. F. Nonlinear models for fitting growth curves of Nellore cows reared in the Amazon Biome. **Revista Brasileira de Zootecnia** v. 42, n. 9, p. 645-650, 2013.

MOREIRA, A.; HEINRICHS, R.; FREITAS, A R. Relação fósforo e magnésio na fertilidade do solo, no estado nutricional e na produção da alfafa. **Revista Brasileira de Zootecnia**, v. 37, n. 6, p. 984-989, 2008.

PEREIRA, J. C. R. **Análise de dados qualitativos: Estratégias metodológicas para as ciências da saúde, humanas e sociais**. 3. ed. São Paulo: Ed. Edusp, 2004. 157 p.

POPPER, K. **The logic of scientific discovery**. London: Routledge Classics, 2005. Disponível em: http://strangebeautiful.com/other-texts/popper-logic-scientific-discovery.pdf. Acesso em: 20 fev. 2017.

PORTAL ACTION. **6.2-Distribuição de frequências**. Disponível em: https://xdocs.com.br/doc/distribuiao-normal-padronizada-probabilidades-portal-action-jozmmeq6jlnz. Acesso em: 28 out. 2019.

RAMALHO, M. A. P.; FERREIRA, D. F.; OLIVEIRA, A. C. **Experimentação em genética e melhoramento de plantas**. 2. ed. Lavras: Ed. Ufla, 2005. 322 p.

ROSA, J. S.; JOHNSON, E. H.; ALVES, F. S. F.; SANTOS, L. de F. L. Ocorrência de abscesso hepático em caprinos. **Pesquisa Agropecuária Brasileira**, v. 24, n. 1, p. 63-68, jan. 1989.

SALSBURG, D. **Uma senhora toma chá**...: Como a estatística revolucionou a ciência no século XX. Rio de Janeiro: Jorge Zahar, 2009. 288 p.

SAMPAIO, I. B. M. **Estatística aplicada à experimentação animal**. 3. ed. Belo Horizonte: Fundação de Ensino e Pesquisa em Medicina Veterinária e Zootecnia, 1998. 264 p.

SAS INSTITUTE. **About SAS**. Disponível em: http://www.sas.com. Acesso em: 17 ago. 2017.

SCHEFFÉ, H. A method for judging all contrasts in the analysis of variance. **Biometrika**, v. 56, n. 1, p. 229, Mar. 1969. Disponível em: https://doi.org/10.1093/biomet/56.1.229. Acesso em: 27 out. 2021.

TREVISOL, A. M.; COSME, K. F. S. G. **Análise de experimentos em blocos incompletos**. 2013. 40 f. Monografia (Bacharelado em Estatística) – Universidade de Brasília, Brasília, DF. Disponível em: http://bdm.unb.br/handle/10483/5718. Acesso em: 19 mar. 2019.

WIKIPÉDIA. **Blaise Pascal**. 2019b. Disponível em: https://pt.wikipedia.org/wiki/Blaise_Pascal. Acesso em: 24 jul. 2019.

WIKIPÉDIA. **Francis Bacon**. 2019c. Disponível em: https://pt.wikipedia.org/wiki/Francis_Bacon. Acesso em: 5 set. 2019.

WIKIPÉDIA. **Galileu Galilei**. 2019d. Disponível em: https://pt.wikipedia.org/wiki/Galileu_Galilei. Acesso em: 24 jul. 2019.

WIKIPÉDIA. **Jakob Bernoulli**. 2019e. Disponível em: https://pt.wikipedia.org/wiki/Jakob_Bernoulli. Acesso em: 24 jul. 2019.

WIKIPÉDIA. **Karl Pearson**. 2019f. Disponível em: https://en.wikipedia.org/wiki/Karl_Pearson. Acesso em: 18 set. 2019.

WIKIPÉDIA. **Leonhard Euler**. 2019g. Disponível em: https://pt.wikipedia.org/wiki/Quadrado_latino. Acesso em: 22 nov. 2019.

WIKIPÉDIA. **Simeon-Denis Poisson**. 2019h. Disponível em: https://pt.wikipedia.org/wiki/Simeon_Denis_Poisson. Acesso em: 7 nov. 2019.

WIKIPÉDIA. **Sir Ronald Aylmer Fisher**. 2019i. Disponível em: http://en.wikipedia.org/wiki/Ronald_Fisher. Acesso em: 18 set. 2019.

WIKIPÉDIA. **Thomas Bayes**. 2019j. Disponível em: https://en.wikipedia.org/wiki/Thomas_Bayes. Acesso em: 26 out. 2021.

WIKIPÉDIA. **William J. Youden**. 2019a. Disponível em: https://en.wikipedia.org/wiki/William_J._Youden. Acesso em: 15 set. 2019.

WIKIPÉDIA. **William Sealy Gosset**. 2019L. Disponível em: poissonPYUOhttps://pt.wikipedia.org/wiki/William_Sealy_Gosset. Acesso em: 18 set. 2019.

ZABELL, S. L. On Student's 1908 Article "The Probable Error of a Mean". **Journal of the American Statistical Association**, v. 103, n. 481, 2008. DOI: 10.1198/016214508000000030.

Anexo 1

Tabelas estatísticas

Tabela 1.1. Distribuição normal padronizada Z (área sob a curva normal de 0 a Z).

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	,5000	,5040	,5080	,5120	,5160	,5199	,5239	,5279	,5319	,5359
0,1	,5398	,5438	,5478	,5517	,5557	,5596	,5636	,5675	,5714	,5754
0,2	,5792	,5832	,5871	,5909	,5948	,5987	,6026	,6064	,6103	,6141
0,3	,6179	,6217	,6255	,6293	,6331	,6368	,6406	,6443	,6480	,6517
0,4	,6554	,6591	,6628	,6664	,6700	,6736	,6772	,6808	,6844	,6879
0,5	,6914	,6950	,6985	,7019	,7054	,7088	,7123	,7157	,7190	,7224
0,6	,7257	,7291	,7324	,7356	,7389	,7421	,7454	,7486	,7518	,7549
0,7	,7580	,7612	,7642	,7673	,7703	,7734	,7764	,7794	,7823	,7852
0,8	,7881	,7910	,7939	,7967	,7996	,8023	,8051	,8078	,8106	,8133
0,9	,8159	,8186	,8212	,8238	,8264	,8289	,8315	,8340	,8365	,8389
1,0	,8413	,8438	,8461	,8485	,8508	,8531	,8554	,8577	,8599	,8621
1,1	,8643	,8665	,8686	,8708	,8729	,8749	,8770	,8790	,8810	,8830
1,2	,8849	,8869	,8888	,8906	,8925	,8944	,8962	,8979	,8997	,9015
1,3	,9032	,9049	,9066	,9082	,9099	,9115	,9131	,9147	,9162	,9177
1,4	,9192	,9207	,9222	,9236	,9251	,9265	,9278	,9292	,9305	,9319
1,5	,9332	,9345	,9357	,9370	,9382	,9394	,9406	,9418	,9429	,9441
1,6	,9452	,9463	,9474	,9484	,9495	,9505	,9515	,9525	,9535	,9545
1,7	,9554	,9564	,9573	,9582	,9591	,9599	,9608	,9616	,9625	,9633
1,8	,9641	,9648	,9656	,9664	,9671	,9678	,9685	,9693	,9699	,9706
1,9	,9713	,9719	,9726	,9732	,9738	,9744	,9750	,9756	,9762	,9767
2,0	,9772	,9778	,9783	,9788	,9793	,9798	,9803	,9808	,9812	,9816
2,1	,9821	,9826	,9830	,9834	,9838	,9842	,9846	,9850	,9854	,9857
2,2	,9861	,9864	,9868	,9871	,9874	,9878	,9881	,9884	,9887	,9890
2,3	,9893	,9895	,9898	,9901	,9904	,9906	,9909	,9911	,9913	,9916
2,4	,9918	,9920	,9922	,9924	,9927	,9929	,9930	,9932	,9934	,9936
2,5	,9938	,9940	,9941	,9943	,9945	,9946	,9948	,9949	,9951	,9952
2,6	,9953	,9955	,9956	,9957	,9958	,9960	,9961	,9962	,9963	,9964
2,7	,9965	,9966	,9967	,9968	,9969	,9970	,9971	,9972	,9973	,9974
2,8	,9974	,9975	,9976	,9977	,9977	,9978	,9979	,9979	,9980	,9981
2,9	,9981	,9982	,9982	,9983	,9984	,9984	,9985	,9985	,9986	,9986
3,0	,9986	,9987	,9987	,9988	,9988	,9989	,9989	,9989	,9989	,9990
3,1	,9990	,9991	,9991	,9991	,9992	,9992	,9992	,9992	,9993	,9993
3,2	,9993	,9993	,9994	,9994	,9994	,9994	,9994	,9995	,9995	,9995
3,3	,9995	,9995	,9995	,9996	,9996	,9996	,9996	,9996	,9996	,9996
3,4	,9997	,9997	,9997	,9997	,9997	,9997	,9997	,9997	,9998	,9998
3,5	,9998	,9998	,9998	,9998	,9998	,9998	,9998	,9998	,9998	,9998
3,6	,9998	,9998	,9998	,9999	,9999	,9999	,9999	,9999	,9999	,9999
3,7	,9999	,9999	,9999	,9999	,9999	,9999	,9999	,9999	,9999	,9999
3,8	,9999	,9999	,9999	,9999	,9999	,9999	,9999	,9999	,9999	,9999
3,9	,9999	,9999	,9999	,9999	,9999	,9999	,9999	,9999	,9999	,9999

Tabela 1.2. Distribuição de qui-quadrado de acordo com o nível de probabilidade α ($\chi^2\alpha$).

GLR \ α	χ^2.995	χ^2.990	χ^2.975	χ^2.950	χ^2.000	χ^2.100	χ^2.050	χ^2.025	χ^2.010	χ^2.005
1	0,000	0,000	0,001	0,004	0,016	2,706	3,841	5,024	6,635	7,879
2	0,010	0,020	0,051	0,103	0,211	4,605	5,991	7,378	9,210	10,597
3	0,072	0,115	0,216	0,352	0,584	6,251	7,815	9,348	11,345	12,838
4	0,207	0,297	0,484	0,711	1,064	7,779	9,488	11,143	13,277	14,860
5	0,412	0,554	0,831	1,145	1,610	9,236	11,070	12,833	15,086	16,750
6	0,676	0,872	1,237	1,635	2,204	10,645	12,592	14,449	16,812	18,548
7	0,989	1,239	1,690	2,167	2,833	12,017	14,067	16,013	18,475	20,278
8	1,344	1,646	2,180	2,733	3,490	13,362	15,507	17,535	20,090	21,955
9	1,735	2,088	2,700	3,325	4,168	14,684	16,919	19,023	21,666	23,589
10	2,156	2,558	3,247	3,940	4,865	15,987	18,307	20,483	23,209	25,188
11	2,603	3,053	3,816	4,575	5,578	17,275	19,675	21,920	24,725	26,757
12	3,074	3,571	4,404	5,226	6,304	18,549	21,026	23,337	26,217	28,300
13	3,565	4,107	5,009	5,892	7,042	19,812	22,362	24,736	27,688	29,819
14	4,075	4,660	5,629	6,571	7,790	21,064	23,685	26,119	29,141	31,319
15	4,601	5,229	6,262	7,261	8,547	22,307	24,996	27,488	30,578	32,801
16	5,142	5,812	6,908	7,962	9,312	23,542	26,296	28,845	32,000	34,267
17	5,697	6,408	7,564	8,672	10,085	24,769	27,587	30,191	33,409	35,718
18	6,265	7,015	8,231	9,390	10,865	25,989	28,869	31,526	34,805	37,156
19	6,844	7,633	8,907	10,117	11,651	27,204	30,144	32,852	36,191	38,582
20	7,434	8,260	9,591	10,851	12,443	28,412	31,410	34,170	37,566	39,997
21	8,034	8,897	10,283	11,591	13,240	29,615	32,671	35,479	38,932	41,401
22	8,643	9,542	10,982	12,338	14,041	30,813	33,924	36,781	40,289	42,796
23	9,260	10,196	11,689	13,091	14,848	32,007	35,172	38,076	41,638	44,181
24	9,886	10,856	12,401	13,848	15,659	33,196	36,415	39,364	42,980	45,559
25	10,520	11,524	13,120	14,611	16,473	34,382	37,652	40,646	44,314	46,928
26	11,160	12,198	13,844	15,379	17,292	35,563	38,885	41,923	45,642	48,290
27	11,808	12,879	14,573	16,151	18,114	36,741	40,113	43,195	46,963	49,645
28	12,461	13,565	15,308	16,928	18,939	37,916	41,337	44,461	48,278	50,993
29	13,121	14,256	16,047	17,708	19,768	39,087	42,557	45,722	49,588	52,336
30	13,787	14,953	16,791	18,493	20,599	40,256	43,773	46,979	50,892	53,672
40	20,707	22,164	24,433	26,509	29,051	51,805	55,758	59,342	63,691	66,766
50	27,991	29,707	32,357	34,764	37,689	63,167	67,505	71,420	76,154	79,490
60	35,534	37,485	40,482	43,188	46,459	74,397	79,082	83,298	88,379	91,952
70	43,275	45,442	48,758	51,739	55,329	85,527	90,531	95,023	100,425	104,215
80	51,172	53,540	57,153	60,391	64,278	96,578	101,879	106,629	112,329	116,321
90	59,196	61,754	65,647	69,126	73,291	107,565	113,145	118,136	124,116	128,299
100	67,328	70,065	74,222	77,929	82,358	118,498	124,342	129,561	135,807	140,169

Tabela 1.3. Valores críticos da distribuição t de Student (α =unilateral; α/2= bilateral). GL = grau de liberdade.

α / GL	0,05	0,025	0,01	0,005	0,0025	0,001
1	6,3138	12,7065	31,8193	63,6551	127,3447	318,4930
2	2,9200	4, 3026	6,9646	9,9247	14,0887	22,3276
3	2,3534	3,1824	4,5407	5,8408	7,4534	10,2145
4	2,1319	2,7764	3,7470	4,6041	5,5976	7,1732
5	2,0150	2,5706	3,3650	4,0322	4,7734	5,8934
6	1,9432	2,4469	3,1426	3,7074	4,3168	5,2076
7	1,8946	2,3646	2,9980	3,4995	4,0294	4,7852
8	1,8595	2,3060	2,8965	3,3554	3,8325	4,5008
9	1,8331	2,2621	2,8214	3,2498	3,6896	4,2969
10	1,8124	2,2282	2,7638	3,1693	3,5814	4,1437
11	1,7959	2,2010	2,7181	3,1058	3,4966	4,0247
12	1,7823	2,1788	2,6810	3,0545	3,4284	3,9296
13	1,7709	2,1604	2,6503	3,0123	3,3725	3,8520
14	1,7613	2,1448	2,6245	2,9768	3,3257	3,7874
15	1,7530	2,1314	2,6025	2,9467	3,2860	3,7328
16	1,7459	2,1199	2,5835	2,9208	3,2520	3, 6861
17	1,7396	2,1098	2,5669	2,8983	3,2224	3,6458
18	1,7341	2,1009	2,5524	2,8784	3,1966	3,6105
19	1,7291	2,0930	2,5395	2,8609	3,1737	3,5794
20	1,7247	2,0860	2,5280	2,8454	3,1534	3,5518
21	1,7207	2,0796	2,5176	2,8314	3,1352	3,5272
22	1,7172	2,0739	2,5083	2,8188	3,1188	3,5050
23	1,7139	2,0686	2,4998	2,8073	3,1040	3,4850
24	1,7109	2,0639	2,4922	2,7970	3,0905	3,4668
25	1,7081	2,0596	2,4851	2,7874	3,0782	3,4502
26	1,7056	2,0555	2,4786	2,7787	3,0669	3,4350
27	1,7033	2,0518	2,4727	2,7707	3,0565	3,4211
28	1,7011	2,0484	2,4671	2,7633	3,0469	3,4082
29	1,6991	2,0452	2,4620	2,7564	3,0380	3,3962
30	1,6973	2,0423	2,4572	2,7500	3,0298	3,3852
40	1,6839	2,0211	2,4233	2,7045	2,9712	3,3069
50	1,6759	2,0086	2,4033	2,6778	2,9370	3,2614
60	1,6706	2,0003	2,3901	2,6603	2,9146	3,2317
70	1,6669	1,9944	2,3808	2,6479	2,8987	3,2108
80	1,6641	1,9901	2,3739	2,6387	2,8870	3,1953
90	1,6620	1,9867	2,3685	2,6316	2,8779	3,1833
100	1, 6602	1,9840	2,3642	2,6259	2,8706	3,1738

Tabela 1.4. Distribuição dos valores de F de acordo com grau de liberdade do numerador (GLN) e do denominador (GLD) para α = 0,05.

GLD \ GLN	1	2	3	4	5	6	7	8	9	10	20
1	161,4	199,3	215,7	224,6	230,2	234,0	237,0	238,9	241,0	242,0	248,0
2	18,51	19,00	19,16	19,25	19,30	19,33	19,4	19,37	19,38	19,40	19,4
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,66
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,80
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,56
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	3,87
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,44
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,15
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	2,94
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,77
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85	2,65
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75	2,54
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67	2,46
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60	2,39
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54	2,33
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,28
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45	2,23
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41	2,19
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38	2,16
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,12
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30	2,07
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25	2,03
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22	1,99
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19	1,96
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16	1,93
35	4,12	3,27	2,87	2,64	2,49	2,37	2,29	2,22	2,16	2,11	1,88
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08	1,84
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99	1,75
80	3,96	3,11	2,72	2,49	2,33	2,21	2,13	2,06	2,00	1,95	1,70
100	3,94	3,09	2,70	2,46	2,31	2,19	2,10	2,03	1,97	1,93	1,68
∞	1,04	3,00	2,61	2,37	2,21	2,10	2,01	1,94	1,88	1,83	1,57

Tabela 1.5. Valores tabelados de Student Newman Keuls (SNK) para α = 0,05, de acordo com a distância entre as médias em ordem decrescente (p) e graus de liberdade do resíduo (GLR).

P / GLR	2	3	4	5	6	7	8	9	10	11
1	17,97	26,98	32,82	37,00	40,41	43,12	45,40	47,36	49,07	50,59
2	6,09	8,33	9,80	10,88	11,74	12,44	13,03	13,54	13,99	14,39
3	4,50	5,91	6,83	7,50	8,04	8,48	8,85	9,18	9,46	9,72
4	3,93	5,04	5,76	6,29	6,71	7,05	7,35	7,60	7,83	8,72
5	3,64	4,60	5,22	5,67	6,03	6,33	6,58	6,80	6,99	7,17
6	3,46	4,34	4,90	5,31	5,63	5,89	6,12	6,32	6,49	6,65
7	3,34	4,16	4,68	5,06	5,36	5,61	5,82	6,00	6,16	6,30
8	3,26	4,04	4,53	4,89	5,17	5,40	5,60	5,77	5,92	6,05
9	3,20	3,95	4,42	4,76	5,02	5,24	5,43	5,60	5,74	5,87
10	3,15	3,88	4,33	4,65	4,91	5,12	5,30	5,46	5,60	5,72
11	3,11	3,82	4,26	4,57	4,82	5,03	5,20	5,35	5,49	5,61
12	3,08	3,77	4,20	4,51	4,75	4,95	5,12	5,27	5,40	5,51
13	3,06	3,73	4,15	4,45	4,69	4,88	5,05	5,19	5,32	5,43
14	3,03	3,70	4,11	4,41	4,64	4,83	4,99	5,13	5,25	5,36
15	3,01	3,67	4,08	4,37	4,60	4,78	4,94	5,08	5,20	5,31
16	3,00	3,65	4,05	4,33	4,56	4,74	4,90	5,03	5,15	5,26
17	2,98	3,63	4,02	4,30	4,52	4,71	4,86	4,99	5,11	5,21
18	2,97	3,61	4,00	4,28	4,49	4,67	4,82	4,96	5,07	5,17
19	2,96	3,59	3,98	4,25	4,47	4,65	4,79	4,92	5,04	5,14
20	2,95	3,58	3,96	4,23	4,45	4,62	4,77	4,90	5,01	5,11
24	2,92	3,53	3,90	4,17	4,37	4,54	4,68	4,81	4,92	5,01
30	2,89	3,49	3,84	4,10	4,30	4,46	4,60	4,72	4,83	4,92
40	2,86	3,44	3,79	4,04	4,23	4,39	4,52	4,63	4,74	4,82
60	2,83	3,40	3,74	3,98	4,16	4,31	4,44	4,55	4,65	4,73
∞	2,80	3,36	3,69	3,92	4,10	4,24	4,36	4,48	4,56	4,64

Tabela 1.6. Valores de q tabelado para o teste de Tukey para α = 0,05, de acordo com número de médias a serem comparadas (p) e graus de liberdade do resíduo (GLR).

P / GLR	2	3	4	5	6	7	8	9	10	20
1	18,0	27,0	32,8	37,1	40,4	43,1	45,4	47,4	49,1	59,60
2	6,08	8,33	9,80	10,88	11,73	12,43	13,03	3,54	13,99	16,77
3	4,50	5,91	6,82	7,50	8,04	8,48	8,85	9,18	9,46	11,24
4	3,93	5,04	5,76	6,29	6,71	7,05	7,35	7,60	7,83	9,24
5	3,64	4,60	5,22	5,67	6,03	6,33	6,58	6,80	6,99	8,13
6	3,46	4,34	4,90	5,30	5,63	5,90	6,12	6,32	6,49	7,59
7	3,34	4,16	4,68	5,06	5,36	5,61	5,82	6,00	6,16	7,16
8	3,26	4,04	4,53	4,89	5,17	5,40	5,60	5,77	5,92	6,87
9	3,20	3,95	4,41	4,76	5,02	5,24	5,43	5,59	5,74	6,63
10	3,15	3,88	4,33	4,65	4,91	5,12	5,30	5,46	5,60	6,47
11	3,11	3,82	4,26	4,57	4,82	5,03	5,20	5,35	5,49	6,33
12	3,08	3,77	4,20	4,51	4,75	4,95	5,12	5,27	5,39	6,21
13	3,06	3,73	4,15	4,45	4,69	4,88	5,05	5,19	5,32	6,11
14	3,03	3,70	4,11	4,41	4,64	4,83	4,99	5,13	5,25	6,03
15	3,01	3,67	4,08	4,37	4,59	4,78	4,94	5,08	5,20	5,96
16	3,00	3,65	4,05	4,33	4,56	4,74	4,90	5,03	5,15	5,90
17	2,98	3,63	4,02	4,30	4,52	4,70	4,86	4,99	5,11	5,84
18	2,97	3,61	4,00	4,28	4,49	4,67	4,82	4,96	5,07	5,79
19	2,96	3,59	3,98	4,25	4,47	4,65	4,79	4,92	5,04	5,75
20	2,95	3,58	3,96	4,23	4,45	4,62	4,77	4,90	5,01	5,71
30	2,92	3,53	3,90	4,17	4,37	4,54	4,68	4,81	4,92	0,00
40	2,89	3,49	3,85	4,10	4,30	4,46	4,60	4,72	4,82	0,00
50	2,86	3,44	3,79	4,04	4,23	4,39	4,52	4,63	4,73	0,00
60	2,83	3,40	3,74	3,98	4,16	4,31	4,44	4,55	4,65	0,00
120	2,80	3,36	3,68	3,92	4,10	4,24	4,36	4,47	4,56	0,00
∞	2,77	3,31	3,63	3,86	4,03	4,17	4,29	4,39	4,47	0,00

Tabela 1.7. Valores tabelados de q_i para o teste de Duncan, de acordo com a distância entre as médias em ordem decrescente (p) e graus de liberdade do resíduo (GLR).

P / GLR	2	3	4	5	6	7	8	9	10	20
1	17,969	17,969	17,969	17,969	17,969	17,969	17,969	17,969	17,969	17,969
2	6,085	6,085	6,085	6,085	6,085	6,085	6,085	6,085	6,085	6,085
3	4,501	4,516	4,516	4,516	4,516	4,516	4,516	4,516	4,516	4,516
4	3,926	4,013	4,033	4,033	4,033	4,033	4,033	4,033	4,033	4,033
5	3,635	3,749	3,796	3,814	3,814	3,814	3,814	3,814	3,814	3,814
6	3,460	3,586	3,649	3,680	3,694	3,697	3,697	3,697	3,697	3,697
7	3,344	3,477	3,548	3,588	3,611	3,622	3,625	3,625	3,625	3,625
8	3,261	3,398	3,475	3,521	3,549	3,566	3,575	3,579	3,579	3,579
9	3,199	3,339	3,420	3,470	3,502	3,523	3,536	3,544	3,547	3,547
10	3,151	3,293	3,376	3,430	3,465	3,489	3,505	3,516	3,522	3,525
11	3,113	3,256	3,341	3,397	3,435	3,462	3,480	3,493	3,501	3,510
12	3,081	3,225	3,312	3,370	3,410	3,439	3,459	3,474	3,484	3,498
13	3,055	3,200	3,288	3,348	3,389	3,419	3,441	3,458	3,470	3,490
14	3,033	3,178	3,268	3,328	3,371	3,403	3,426	3,444	3,457	3,484
15	3,014	3,160	3,250	3,312	3,356	3,389	3,413	3,432	3,446	3,480
16	2,998	3,144	3,235	3,297	3,343	3,376	3,402	3,422	3,437	3,477
17	2,984	3,130	3,222	3,285	3,331	3,365	3,392	3,412	3,429	3,475
18	2,971	3,117	3,210	3,274	3,320	3,356	3,383	3,404	3,421	3,474
19	2,960	3,106	3,199	3,264	3,311	3,347	3,375	3,397	3,415	3,474
20	2,950	3,097	3,190	3,255	3,303	3,339	3,368	3,390	3,409	3,473
22	2,933	3,080	3,173	3,239	3,288	3,326	3,355	3,379	3,398	3,472
24	2,919	3,066	3,160	3,226	3,276	3,315	3,345	3,370	3,390	3,472
26	2,907	3,054	3,149	3,216	3,266	3,305	3,336	3,362	3,382	3,471
28	2,897	3,044	3,139	3,206	3,257	3,297	3,329	3,355	3,376	3,470
30	2,888	3,035	3,131	3,199	3,250	3,290	3,322	3,349	3,371	3,470
40	2,858	3,005	3,102	3,171	3,224	3,266	3,300	3,328	3,352	3,469
60	2,829	2,976	3,073	3,143	3,198	3,241	3,277	3,307	3,333	3,468
80	2,814	2,961	3,059	3,130	3,185	3,229	3,266	3,297	3,323	3,467
∞	2,772	2,918	3,017	3,089	3,146	3,193	3,232	3,265	3,294	3,466

Tabela 1.8. Valores tabelados para o teste de Dunnett (α=0,05) em função do grau de liberdade do resíduo (GLR) e k tratamentos, incluindo o controle.

GLR \ k	2	3	4	5	6	7	8	9	10
5	2,57	3,03	3,29	3,48	3,62	3,73	3,82	3,90	3,97
6	2,45	2,86	3,10	3,26	3,39	3,49	3,57	3,64	3,71
7	2,36	2,75	2,97	3,12	3,24	3,33	3,41	3,47	3,53
8	2,31	2,67	2,88	3,02	3,13	3,22	3,29	3,35	3,41
9	2,26	2,61	2,81	2,95	3,05	3,14	3,20	3,26	3,32
10	2,23	2,57	2,76	2,89	2,99	3,07	3,14	3,19	3,24
11	2,20	2,53	2,72	2,84	2,94	3,02	3,08	3,14	3,19
12	2,18	2,50	2,68	2,81	2,90	2,98	3,04	3,09	3,14
13	2,16	2,48	2,65	2,78	2,87	2,94	3,00	3,06	3,10
14	2,14	2,46	2,63	2,75	2,84	2,91	2,97	3,02	3,07
15	2,13	2,44	2,61	2,73	2,82	2,89	2,95	3,00	3,04
16	2,12	2,42	2,59	2,71	2,80	2,87	2,92	2,97	3,02
17	2,11	2,41	2,58	2,69	2,78	2,85	2,90	2,95	3,00
18	2,10	2,40	2,56	2,68	2,76	2,83	2,89	2,94	2,98
19	2,09	2,39	2,55	2,66	2,75	2,81	2,87	2,92	2,96
20	2,09	2,38	2,54	2,65	2,73	2,80	2,86	2,90	2,95
24	2,06	2,35	2,51	2,61	2,70	2,76	2,81	2,86	2,90
30	2,04	2,32	2,47	2,58	2,66	2,72	2,77	2,82	2,86
40	2,02	2,29	2,44	2,54	2,62	2,68	2,73	2,77	2,81
60	2,00	2,27	2,41	2,51	2,58	2,64	2,69	2,73	2,77

Tabela 1.9. Tabela de Bonferroni em função do número de comparações simultâneas (p) e graus de liberdade do resíduo (GLR) para α = 0,05.

GLR \ p	1	2	3	4	5	6	7	8	9	10
1	12,71	25,45	38,19	50,92	63,66	76,39	89,12	101,9	114,6	127,3
2	4,303	6,205	7,649	8,860	9,925	10,89	11,77	12,59	13,36	14,09
3	3,182	4,177	4,857	5,392	5,841	6,232	6,580	6,895	7,185	7,453
4	2,776	3,495	3,961	4,315	4,604	4,851	5,068	5,261	5,437	5,598
5	2,571	3,163	3,534	3,810	4,032	4,219	4,382	4,526	4,655	4,773
6	2,447	2,969	3,287	3,521	3,707	3,863	3,997	4,115	4,221	4,317
7	2,365	2,841	3,128	3,335	3,499	3,636	3,753	3,855	3,947	4,029
8	2,306	2,752	3,016	3,206	3,355	3,479	3,584	3,677	3,759	3,833
9	2,262	2,685	2,933	3,111	3,250	3,364	3,462	3,547	3,622	3,690
10	2,228	2,634	2,870	3,038	3,169	3,277	3,368	3,448	3,518	3,581

11	2,201	2,593	2,820	2,981	3,106	3,208	3,295	3,370	3,437	3,497
12	2,179	2,560	2,779	2,934	3,055	3,153	3,236	3,308	3,371	3,428
13	2,160	2,533	2,746	2,896	3,012	3,107	3,187	3,256	3,318	3,372
14	2,145	2,510	2,718	2,864	2,977	3,069	3,146	3,214	3,273	3,326
15	2,131	2,490	2,694	2,837	2,947	3,036	3,112	3,177	3,235	3,286
16	2,120	2,473	2,673	2,813	2,921	3,008	3,082	3,146	3,202	3,252
17	2,110	2,458	2,655	2,793	2,898	2,984	3,056	3,119	3,173	3,222
18	2,101	2,445	2,639	2,775	2,878	2,963	3,034	3,095	3,149	3,197
19	2,093	2,433	2,625	2,759	2,861	2,944	3,014	3,074	3,127	3,174
20	2,086	2,423	2,613	2,744	2,845	2,927	2,996	3,055	3,107	3,153
21	2,080	2,414	2,601	2,732	2,831	2,912	2,980	3,038	3,090	3,135
22	2,074	2,405	2,591	2,720	2,819	2,899	2,965	3,023	3,074	3,119
23	2,069	2,398	2,582	2,710	2,807	2,886	2,952	3,009	3,059	3,104
24	2,064	2,391	2,574	2,700	2,797	2,875	2,941	2,997	3,046	3,091
25	2,060	2,385	2,566	2,692	2,787	2,865	2,930	2,986	3,035	3,078
26	2,056	2,379	2,559	2,684	2,779	2,856	2,920	2,975	3,024	3,067
27	2,052	2,373	2,552	2,676	2,771	2,847	2,911	2,966	3,014	3,057
28	2,048	2,368	2,546	2,669	2,763	2,839	2,902	2,957	3,004	3,047
29	2,045	2,364	2,541	2,663	2,756	2,832	2,894	2,949	2,996	3,038
30	2,042	2,360	2,536	2,657	2,750	2,825	2,887	2,941	2,988	3,030
40	2,021	2,329	2,499	2,616	2,704	2,776	2,836	2,887	2,931	2,971
50	2,009	2,311	2,477	2,591	2,678	2,747	2,805	2,855	2,898	2,937
60	2,000	2,299	2,463	2,575	2,660	2,729	2,785	2,834	2,877	2,915
70	1,994	2,291	2,453	2,564	2,648	2,715	2,771	2,820	2,862	2,899
80	1,990	2,284	2,445	2,555	2,639	2,705	2,761	2,809	2,850	2,887
90	1,987	2,280	2,440	2,549	2,632	2,698	2,753	2,800	2,841	2,878
100	1,984	2,276	2,435	2,544	2,626	2,692	2,747	2,793	2,834	2,871
1000	1,962	2,245	2,398	2,502	2,581	2,644	2,696	2,740	2,779	2,813

Apêndice 1

Respostas dos exercícios

Capítulo 1

Resposta do exercício 1

Estatísticas descritivas são usadas para descrever as características dos dados. Várias são as técnicas usadas para classificar dados: descrição gráfica, descrição tabular e sumários estatísticos, entre outras. É imprescindível compreender alguns conceitos: parâmetros e estatísticas: uma população é caracterizada por parâmetros, que geralmente são desconhecidos, os quais são funções de valores populacionais. **Na inferência estatística, uma amostra é escolhida para representar a população em uma análise estatística;** as estatísticas, porém, são funções de valores amostrais e nem sempre representam com precisão os parâmetros; dados qualitativos: representam a informação que identifica alguma qualidade, categoria ou característica, que é susceptível de classificação, mas **não** de medida; dados quantitativos: representam a informação resultante de características susceptíveis de serem medidas, podendo ser de natureza discreta ou contínua e dividem em: variável discreta − que é constituída de partes ou categorias separadas e distintas, e variável contínua − que **pode assumir um conjunto ordenado de valores inteiros e fracionários** dentro de determinado limite ou intervalo; estudo transversal − permite uma situação momentânea do fenômeno investigado; o estudo é relativamente rápido, consome menos recursos e é menos vulnerável a variáveis estranhas; estudo longitudinal − os dados de cada indivíduo são coletados em dois ou mais momentos, possibilitando acompanhamento do desenrolar do fenômeno considerado. **Geralmente** é mais complicado que o estudo transversal; **porém, é bastante utilizado e** recomendado para a pesquisa.

Resposta do exercício 2

Nos dados originais a média foi maior que a mediana no corte 4 e caracterizou-se simetria positiva e foi menor que a mediana no corte 2 e 5, caracterizando simetria negativa. O mais importante é que nos cortes 2, 4 e 5, a presença de *outliers* proporcionou maior coeficiente de variação. Portanto, a presença de *outliers,* como geralmente acontece, prejudicou a qualidade dos dados.

Resposta do exercício 3

Como se sabe o coeficiente de variação (CV) é sensível à *outliers*. Entretanto, nesse estudo, considerando os dados sem *outliers*, os CV foram relativamente altos nos

cinco cortes e a ordem decrescente dos valores do CV nos cinco cortes foi: 2, 1, 3, 4 e 5. Nos três cortes em que houve a ocorrência de *outliers* (2, 4 e 5), o acréscimo no valor do CV foi de 19,5% no corte 2, 4,9% no corte 4 e 7,4% no corte 5.

Resposta do exercício 4

Com a eliminação dos *outliers* nos cortes 1, 2 e 5, a média foi menor que a mediana, indicando que a cauda da curva da distribuição dos dados é viesada à esquerda. Nos cortes 3 e 4, a média foi menor que a mediana, indicando que a curva da distribuição dos dados é viesada para a direita.

Resposta do exercício 5

Inicialmente, tem-se que considerar as escalas de assimetria dos coeficientes de assimetria e de curtose.

Coeficiente de assimetria (CA):

$|CA| < 0,15$: assimetria pequena.

$0,15 < |CA| < 1$: assimetria moderada.

$|CA| > 1$: assimetria elevada.

Coeficiente de curtose (C):

$C \cong 0,263$: mesocúrtica: quando os dados têm distribuição normal.

$C < 0,263$: leptocúrtica: distribuição mais pontiaguda do que a normal.

$C > 0,263$: platicúrtica: distribuição mais achatada do que a normal.

Em uma distribuição normal, tem-se média = moda = mediana e isso não ocorreu em nenhuma das pesagens. Entretanto, observa-se tendência de acréscimo da assimetria e da curtose de P_3 a P_8, indicando maior afastamento dos dados em relação à distribuição normal. Analisando os coeficientes de curtose (C), verifica-se que todas as pesagens tiveram distribuição mais achatada do que a normal ($C > 0,263$: platicúrtica).

Nas pesagens P_1, P_2, P_4, P_5 e P_8, observou-se assimetria positiva (Moda < mediana < média).

Capítulo 2

Resposta do exercício 1

Geralmente, os dados coletados de um experimento de campo, de um ensaio em laboratório e de qualquer pesquisa de modo geral, dão origem a um arquivo de dados. É importante documentar esse arquivo, para que suas informações possam ser recuperadas sempre que necessário. **Dependendo da instituição e da finalidade da pesquisa, a estrutura de um arquivo de dados pode variar.** Independentemente do número de variáveis medidas em um experimento, **geralmente a análise estatística inicia com uma amostra de variáveis aleatórias x_1, x_2, ... , x_n e a partir desta todos os cálculos são realizados.** O termo $\sum_{i=1}^{n} (x_i - \bar{x})$ representa o somatório dos erros ou desvios; porém, para evitar que o somatório resulte em zero, cada termo é elevado ao quadrado, resultando na expressão da variância ou soma de quadrados corrigida: $\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} (x_i)^2/n$, em que $(\sum_{i=1}^{n} (x_i)^2/n$ é o conhecido fator de correção. Com relação aos momentos estatísticos de ordem p (p = 1, p = 2, p = 3 e p = 4), os mais utilizados são **os de primeira e segunda**, que originam, respectivamente, a média e a variância, que são conhecidos como momentos de baixa ordem. **Os momentos, da terceira até a quarta ordem,** são utilizados nos cálculos de coeficientes de curtose e assimetria.

Resposta do exercício 2

Classe	Frequência	Frequência relativa (%)	Frequência acumulada
Peso < 49,3	5	12,5	12,5
49,3g ≤ peso < 55,5g	18	45,0	70,0
peso ≥ 55,5g	17	42,5	100,0

Resposta do exercício 3

/ informações de identificação*
delineamento experimental
blocos casualizados, quatro repetições e parcelas subdivididas.
repetições: quatro (1, 2, 3, 4)
tratamento nas parcelas: 10 (organizados em esquema fatorial 2x5)
duas fontes de nitrogênio de nitrogênio (n): ureia e nitrato de amônio
cinco doses de adubo: nitrogênio, kg/ha: (0, 25, 50, 100, 200)
subparcela: cinco cortes consecutivos (1 a 5)

```
variáveis avaliadas
prod = produtividade da área útil da parcela, g
gaf = amostragem de material fresco colhido na parcela, g
gas = amostra de material seco, g
pb = proteína bruta, %
fdn = fibra em detergente neutro, % (fdn)
div = digestibilidade in vitro da matéria seca, %
no3 = nitrato solúvel na biomassa, mg/kg
n = nitrogênio, g/kg(n)
*/
data forragem;
input rep fonte dose corte prod gaf gas  pb fdn div no3 n;
datalines;
1  1  0  1    51.0    48.9   13.0   7.4   82.9  58.2    0.0  10.7
1  1  0  2  600.0   448.1   105.2  10.3  86.5  58.2  33.3  15.5
...
;
```

Resposta do exercício 4

$\sum x^2$ está errado. O valor correto é: $\sum x^2 = 1.415^2 + ... + 1.399^2 = 9.680.135$.

Resposta do exercício 5

Como $y'y = \sum y_i^2 = 3522490$, tem-se:

$3522490 - (\sum y_i)^2/5 = 11.270$

$- (\sum y_i)^2/5 = 11270 - 3.522.490$

$(\sum y_i)^2/5 = 3522490 - 11270 = 3.511.220$

E, finalmente, tem-se: $\sum Y_i = 4.190$.

Resposta do exercício 6

O ganho de peso total do frango, do nascimento até 42 dias de idade (GP_{1-42}) é obtido da soma de $GP_{1-28} + GP_{29-42}$:

$$GP_{1\text{-}42} = GP_{1\text{-}28} + GP_{29\text{-}42} =$$

$$GP_{1\text{-}42} = \begin{bmatrix} 1352 & 1347 & 1325 & 1245 \\ 1338 & 1374 & 1313 & 1199 \\ 1342 & 1313 & 1306 & 1330 \\ 1352 & 1294 & 1291 & 1306 \end{bmatrix}$$

A média do ganho diário, do nascimento a 42 dias de idade ($GDM_{1\text{-}42}$), é obtida por: $GP_{1\text{-}42}/42$.

$$GDM_{1\text{-}42} = \begin{bmatrix} 32{,}19 & 32{,}07 & 31{,}55 & 29{,}64 \\ 31{,}86 & 32{,}71 & 31{,}26 & 28{,}54 \\ 31{,}95 & 31{,}26 & 31{,}09 & 31{,}66 \\ 32{,}19 & 30{,}80 & 30{,}74 & 31{,}09 \end{bmatrix}$$

Capítulo 3

Resposta do exercício 1

Na análise gráfica, uma das ferramentas mais importantes na estatística descritiva é o *box plot*, pois ele fornece: o valor mínimo e o máximo da amostra, o primeiro (Q_1) quartil que representa 25% dos dados, o segundo (Q_2) quartil que equivale à mediana (valor que representa 50% dos dados ordenados), o terceiro quartil (Q_3) ou quartil superior, que representa 75% dos dados ordenados. Outra informação que pode ser obtida do *box plot* é o intervalo interquartílico (Q_3 - Q_1), que inclui **50%** dos dados da amostra. Dentre os gráficos, o histograma é o mais apropriado para exibir frequências de classes, e por isso, é o mais utilizado para representar dados categóricos. **O histograma não permite visualizar *outliers*, mas, além da amplitude dos dados, frequência de cada classe, ele fornece informações sobre variância, assimetria, curtose, sendo possível visualizar aproximadamente a forma da distribuição do conjunto de dados.** Um *outlie*r é uma observação que é discrepante dos demais dados e, principalmente, discrepante da média. **É um dado fora da curva normal. Ele pode causar anomalias nos resultados obtidos de uma análise estatística. Um conjunto de dados pode apresentar um ou vários *outliers*. Eles podem ser valores extremamente pequenos ou extremamente grandes em uma amostra.**

Resposta do exercício 2

As respostas das perguntas de a até f são com base na visualização do Figura 17 (Capítulo 3) e podem variar de pessoa para pessoa.

a) Sim. Apresenta comportamento praticamente linear.

b) Em torno de 580 kg.

c) 150 dias: 125 kg; 300 dias: 200 kg; 450 dias: 290 kg e 600 dias: 310 kg.

d) Sim. A nuvem de pontos fica mais larga. É comum nos estudos de curvas de crescimento em bovinos ocorrer aumento da variância dos pesos com a idade.

e) Como no intervalo de 200 a 550 dias, o numero de dias é 350 e nesse intervalo o ganho de peso foi de 150 kg (300 kg − 150 kg), então a média de ganho diário de peso é obtida por: (150 kg/350 dias) = 0,428 kg dia^{-1} ou 428 g dia^{-1}.

f) É possível visualizar um valor aproximado.

Resposta do exercício 3

a) Gráfico em linha x - y.

b) Gráficos de dispersão (*scatter plot*) como o da Figura 9 (Capítulo 3).

c) Gráficos de superfície (*surface plot*) como o da Figura 12 (Capítulo 3).

d) O Gráfico de superfície (*surface plot*).

Resposta do exercício 4

Matrizes de dispersão (*scatter plot matrix*). São matrizes de diagrama de dispersão que possibilitam explorar relacionamentos bidimensionais, os quais podem revelar várias informações sobre os dados, tais como dependências, *clusters* e *outliers*.

Resposta do exercício 5

No diagrama de caixa da Figura 18 (Capítulo 3), a linha central da caixa é a média. Quando a linha central está exatamente no meio da caixa a simetria é zero. Se ela está acima, a simetria é viesada à esquerda; se a linha central está abaixo, a simetria é viesada à direita. Assim, observa-se que a simetria é viesada à esquerda para os cortes: 1, 5, 6, 8, 9, 10, 11, 15 16 e 19. É viesada à direita para os cortes 3, 4, 7, 12, 14 e 18. Tem simetria zero para os cortes 2, 13, 17 e 20.

Resposta do exercício 6

a) Gráfico peso-idade por raça: fazer um gráfico linha x−y, com cinco linhas, cada uma representando uma raça.

b) Fazer dois gráficos − um gráfico com cinco linhas representando raças para fêmeas e outro gráfico com cinco linhas representando raças para machos.

c) Gráfico peso-idade por raça e sexo: fazer dois gráficos, um gráfico com cinco linhas representando raças para fêmeas e outro gráfico com cinco linhas representando raças para machos.

d) Gráfico peso-idade por raça, sexo e ano de nascimento dos animais: fazer dois tipos de gráficos: com cinco linhas representando raças para fêmeas e gráfico com cinco linhas representando raças para machos. Repetir esses dois tipos de gráficos para anos de nascimento.

Capítulo 4

Resposta do exercício 1

Distribuição de probabilidade − descreve os valores e probabilidades que uma variável aleatória discreta ou contínua pode assumir. Os valores cobrem todos os resultados possíveis de um evento, enquanto que a probabilidade total precisa somar exatamente 1 ou 100%. Três leis fundamentais da probabilidade são: a) a probabilidade de um evento A é um número real, **não negativo, 0 ≤ P(A) ≤ 1**; b) a probabilidade de um evento certo é 1; c) a probabilidade de um evento impossível é zero: $P(\phi) = 0$. O conceito de fatorial, isto é, o fatorial de n elementos é dado por n! = n. (n-1) ... 2.1; o arranjo de n elementos tomados p a p é dado por: $A_{n,p}$ = n!/(n-p)!. Já na análise combinatória, tem-se que a combinação de n elementos tomados p a p, é dada por $\frac{n!}{p!(n-p)!}$. A união de dois conjuntos A e B resulta em um novo conjunto que contém todos os elementos que pertencem a esses dois conjuntos. Assim, se um conjunto A tem quatro elementos e um conjunto B tem cinco elementos, **não necessariamente** o conjunto resultante de A U B terá nove elementos, **pois eles podem ter números em comum**. Um modelo estocástico ou modelo probabilístico é aquele **cujos resultados não são previamente conhecidos**; entretanto, nos modelos determinísticos, **como não envolve variáveis aleatórias**, os resultados **são sempre os mesmos.**

Resposta do exercício 2

Evento 1: atirar uma moeda quatro vezes e observar a face voltada para cima (c = cara, k = coroa).

$$\Omega = \{(c,c,c,c), (k,k,k,k), (c,c,c,k), (c,c,k,c), (c,k,c,c), (k,c,c,c), (c,c,k,k), (k,k,c,c), (c,k,c,k), (k,c,k,c), (c,k,k,c), (k,c,c,k), (k,k,k,c), (k,k,c,k), (k,c,k,k), (c,k,k,k)\}$$

= 16 possibilidades

Evento 2: jogar um dado e observar o número mostrado na face voltada para cima.

$$\Omega = \{1,2,3,4,5,6\}$$

Evento 3: escolher três peças, ao acaso, da produção diária de uma linha de montagem e classificá-las como defeituosas (D) e não defeituosas (N), de acordo com a ordem.

$$\Omega = \{(D,D,D), (N,N,N,), (D,N,N), (N,D,N), (N,N,D), (D,D,N), (D,N,D), (N,D,D)\}$$

= 8 possibilidades

Evento 4: observar o sexo (F = fêmea; M = macho) quanto ao nascimento de duas crianças (não gêmeas) em uma família.

$$\Omega = \{(M,M) , (F,F) , (M,F) , (F,M)\} = 4 \text{ possibilidades}$$

Resposta do exercício 3

Com relação ao exercício anterior, calcular as probabilidades dos eventos:

Evento 1: ocorrência de pelo menos duas caras: p (E_1) = 11/16.

Evento 2: ocorrência de face par no lançamento do dado: p (E_2) = 3/6.

Evento 3: pelo menos duas peças não-defeituosas: p (E_3) = 4/8.

Resposta do exercício 4

a) $A \cup B = \{x/\ 0 \leq x \leq 1\}$

b) $(A \cap B) = \{x/ \frac{1}{2} \leq x \leq \frac{3}{4}\}$

c) $(B \cup C) = \{x/ \frac{1}{2} \leq x \leq \frac{3}{4}\}$

d) $B \cap D = \{x/ \frac{1}{2} \leq x \leq \frac{3}{4}\}$

Resposta do exercício 5

E_1: encontrar temperatura acima de 25 °C, nos dias de 1 a 5

$P(E_1) = 0,15 + 0,10 + 0,15 + 0,20 + 0,10 = 0,70$

$P(E_1) = 0,70$, pois esta situação ocorre nos dias de 1 a 5

E_2: encontrar dia frio, nos dias de 5 a 7

$P(E_2) = 0,10 + 0,20 + 0,10 = 0,40$

$P(E_2) = 0,40$, pois esta situação ocorre nos dias de 5 a 7.

E_3: encontrar pelo menos um dia com calor, nos dias de 1 a 4

$P(E_3) = 0,15 + 0,10 + 0,15 + 0,20 = 0,60$

$P(E_3) = 0,60$, pois esta situação ocorre nos dias de 1 a 4

Capítulo 5

Resposta do exercício 1

As distribuições discretas são importantes para modelar dados de contagens que não são susceptíveis de medida, porém são classificados em classes ou categorias; entretanto, estas distribuições são bastante utilizadas para modelar dados de **natureza discreta,** também denominadas de variáveis ou dados qualitativos. Das distribuições discretas, sem dúvida, a mais importante é a de Poisson, também conhecida como lei de Poisson ou lei dos eventos raros, uma vez que está associada a uma amostra pequena e com probabilidade também pequena. Essa distribuição pode ser derivada como um caso limite da distribuição binomial. **Pode substituir essa quando o número de eventos *n* é muito grande e *p* é muito pequeno (p < 0,1 e n > 50)**. Diferentemente da distribuição binomial, as probabilidades da distribuição de Poisson não são conhecidas previamente. Na distribuição binomial, em uma amostra de tamanho *n*, cada tentativa ou prova é **independente entre si e possui apenas dois resultados: sucesso com probabilidade p ou falha com probabilidade q = 1 - p.** A distribuição binomial que é utilizada para análise de dados discretos e que incluem **duas categorias** é uma das mais utilizadas. A distribuição binomial tende a se tornar cada vez mais simétrica à medida que o tamanho da amostra *n* aumenta. Verifica-se boa concordância entre ela com a distribuição normal para n maior que **30.**

Resposta do exercício 2

A solução é pela distribuição binomial. Considerando-se que 60% dos frangos devem apresentar peso corporal superior a 2,2 kg, então, para uma amostra aleatória de 10 frangos retirados da granja, tem-se:

1) n = 10, p = 0,6 e q = 1 − p = 0,4. Considerando essas informações, vamos calcular dez sucessos (k = 0, 1,..., 10), que é obtido dada pela função de densidade de probabilidade (FDP):

$$P(x = k) = \binom{n}{k} p^k (1-p)^{n-k} = \frac{n!}{k!(n-k)!} p^k (1 - p)^{n-k}$$

A seguir, são apresentados os resultados da probabilidade de k sucessos (k = 0, 1,..., 10).

$P(x=0)=\binom{10}{0}(0,6)^0(0,4)^{10-0}=0,00010$	$P(x=6)=\binom{10}{6}(0,6)^6(0,4)^{10-6}=0,25082$
$P(x=1)=\binom{10}{1}(0,6)^1(0,4)^{10-1}=0,00157$	$P(x=7)=\binom{10}{7}(0,6)^7(0,4)^{10-7}=0,21499$
$P(x=2)=\binom{10}{2}(0,6)^2(0,4)^{10-2}=0,01062$	$P(x=8)=\binom{10}{8}(0,6)^8(0,4)^{10-8}=0,12093$
$P(x=3)=\binom{10}{3}(0,6)^3(0,4)^{10-3}=0,04247$	$P(x=9)=\binom{10}{9}(0,6)^9(0,4)^{10-9}=0,04031$
$P(x=4)=\binom{10}{4}(0,6)^4(0,4)^{10-4}=0,11148$	$P(x=10)=\binom{10}{10}(0,6)^{10}(0,4)^{10-10}=0,00605$
$P(x=5)=\binom{10}{5}(0,4)^5(0,6)^{10-5}=0,20066$	

Finalmente, as três probabilidades são:

Seis frangos possuem peso superior a 2,2 kg:

$$P(a) = 0,25082 \approx 25,1\%$$

Sete frangos possuem peso superior a 2,2 kg:

$$P(b) = 0,21499 \approx 21,5\%$$

Os dez frangos possuem peso superior a 2,2 kg:

$$P(c) = 0,00605 \approx 0,6\%$$

As dez probabilidades anteriores e também a probabilidade acumulada são obtidas também pela rotina SAS:

```
data binomial;
p = 0.6; n = 10;
do x = 0 to 10 by 1;
prob = pdf('binomial', x, p, n);
probcum = cdf('binomial', x, p, n);
output binomial; end;
proc print noobs; var x prob probcum; run;
```

output

x	*prob*	*probcum*
0	*0.00010*	*0.00010*
1	*0.00157*	*0.00168*
2	*0.01062*	*0.01229*
3	*0.04247*	*0.05476*
4	*0.11148*	*0.16624*
5	*0.20066*	*0.36690*
6	*0.25082*	*0.61772*
7	*0.21499*	*0.83271*
8	*0.12093*	*0.95364*
9	*0.04031*	*0.99395*
10	*0.00605*	*1.00000*

Resposta do exercício 3

Para essa situação tem-se: n = 10, p = 0,4 e q = 1 − p = 0,6.

A seguir, são apresentados os resultados da probabilidade de *k* sucessos (k = 0, 1, 2, 3):

$$P(x = 0) = \binom{10}{0} (0,4)^0(0,6)^{10-0} = 0,00605$$

$$P(x = 1) = \binom{10}{1} (0,4)^1(0,6)^{10-1} = 0,04031$$

$$P(x = 2) = \binom{10}{2} (0,4)^2(0,6)^{10-2} = 0,12093$$

$$P(x = 3) = \binom{10}{3} (0,4)^3(0,6)^{10-3} = 0,21499$$

Finalmente, as três probabilidades são:

a) Um frango possui peso inferior a 2,2 kg:

$P(a) = 0,04031 \approx 4,0\%$

b) Dois frangos possuem peso inferior a 2,2 kg:

$P(b) = 0,12093 \approx 12,1\%$

c) Três frangos possuem peso inferior a 2,2 kg:

$P(c) = 0,21499 \approx 21,5\%$

Resposta do exercício 4

A probabilidade de uma peça ser defeituosa em um lote de 100 unidades é p = 1/100 = 0,01. Para p < 0,1 e n > 50, tem-se uma situação típica da distribuição de Poisson, cuja FDP é:

$$P(X = k) = (\lambda^k e^{-\lambda})/ k!$$

A média é obtida por $\lambda = np = 200 \times 0,01 = 2$.

Para k = 0, 1, as probabilidades são:

$$P(X = 0) = (2^0 e^{-2})/ 0! = 0,1353$$

$$P(X = 1) = (2^1 e^{-2})/ 1! = 0,2707$$

Como as probabilidades têm que somar 1, a maneira mais fácil de calcular a probabilidade de obter duas ou mais peças defeituosas é por diferença:

$$P(x > 1) = 1- (P(x = 0) + P(x = 1)) = 1- (0,1353 + 0,2707) = 0,594$$

Finalmente, tem-se as três probabilidades:

0 peça defeituosa:

$$P(a) = 0,1353 \approx 13,5\%$$

1 peça defeituosa:

$$P(b) = 0,2707 \approx 27,1\%$$

2 ou mais peças defeituosas:

$$P(c) = 0,5940 \approx 59,4\%$$

Resposta do exercício 5

O Teste de Fisher é aplicado em análises de dados discretos (nominais ou ordinais), de amostras pequenas que são organizadas em Tabelas 2 x 2 como na tabela abaixo.

Amostra	-	+	Total
Grupo 1	a	b	a + b
Grupo 2	c	d	c + d
Total	a + c	b + d	n

Esse teste é recomendável para N ≤ 30 e quando nenhum dos totais marginais é maior que 15.

O teste de Fisher é calculado por:

$$P = \frac{\binom{a+c}{a}\binom{b+d}{b}}{\binom{n}{a+b}} = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{n!a!b!c!d!}$$

Utilizando-se os dados da Tabela 8 a aplicação do teste de Fisher e feita por meio do programa SAS.

```
data;
input a b c d n;
prob = (comb(a+c,a)*comb(b+d,b))/comb(n,a+b);
datalines;
7  8  5  1 21
;
proc print; var prob;run;
output
0.13136
```

Como a probabilidade calculada (p = 0,13136) é maior do que p = 0,05, significa que não há diferença estatística quanto à resposta de suínos machos e fêmeas quanto ao tratamento.

Capítulo 6

Resposta do exercício 1

A distribuição normal tem grande aplicação na estatística, **para dados contínuos**. Com relação a esta distribuição é correto afirmar: a média, a moda e a mediana são iguais; **o desvio-padrão σ mede a distância do centro da distribuição (x = μ) em relação aos dois ponto de inflexão: x1 = μ - σ e x2 = μ + σ**; os coeficientes de assimetria são indicativos se a curva de distribuição é viezada para esquerda ou para a direita, enquanto os coeficientes de curtose indicam se a curva é mais pontiaguda ou mais achatada que a

normal. **Para uma amostra x_1,..., x_n**, a distribuição normal é compreendida no intervalo de $-\infty < x < \infty$), podendo afirmar que a soma das probabilidades destes valores é igual a 1, conforme mostra a expressão $P(-\infty < x < +\infty) = \int_{-\infty}^{\infty} f(x)dx = 1$. **O coeficiente de variação é a estatística mais apropriada para comparar a dispersão ou homogeneidade de duas amostras**. Por meio do desvio- padrão (raiz quadrada da variância), é possível conhecer a dispersão dos dados de uma amostra e também determinar o erro-padrão da média. Na distribuição normal, 95% da área da curva está dentro do intervalo [μ - **1,96**σ; μ + **1,96**σ]. Com relação à distribuição normal reduzida, algumas propriedades e notação são: a) se $x \sim N(\mu, \sigma^2)$, então $z = (x - \mu)/\sigma$ tem ~ N(0, 1); b) existem dois pontos de inflexão: z = **-1** e z = **+1**, sendo que a área entre eles é 0,6826; c) a curva tem forma de sino e a sua área vale 1; d) a função tem valor máximo no ponto $z = \mu$ e sua ordenada vale $1/\sqrt{(2\pi)}$ = 0,3989; e) ela é definida para qualquer z entre -∞ e +∞; **porém, a curva da fdp tende a zero quando z se aproxima de -4 e tende a 1 quando z se aproxima de +4**. Se v é a soma do quadrado de v variáveis aleatórias z_1, z_2, ..., z_v de distribuição normal reduzida, então $z_i \sim N(0, 1)$ e v tem distribuição de qui-qradrado (χ^2) com **v** graus de liberdade dada, sendo $v = \sum_{i=1}^{v} \mathbf{z_i^2} \sim \boldsymbol{\chi_v^2}$ **.**

Resposta do exercício 2

Como pode ver pelo resultado (*output*) da rotina SAS abaixo, a função tende a zero quando z aproxima de -4, pois *probnorm(-4)* = 0,000031671, e tende a 1 quando z aproxima de +4, pois *probnorm(4)* = 0,9999.

```
data prob;
area1 = probnorm(-4);
area2 = probnorm(4);
proc print; var area1 area2; run;
Output area1      area2
.000031671 0.9999
```

Resposta do exercício 3

Para calcular $F_{5,20}$, dividimos a soma de quadrado de tratamentos (SQT) e soma de quadrado do resíduo (SQR), pelos respectivos graus de liberdade para obter o quadrado médio do tratamentos (QMT) e quadrado médio do resíduo (QMR).

$$F_{5,20} = \frac{QMT}{QMR} = \frac{20,52/5}{12,14/20} = \frac{4,104}{0,612} = 6,71$$

Esses valores são colocados na rotina SAS a seguir:

```
data;
x = probf(6.71,5, 20);
alfa = 1 − x;
proc print; var x alfa;run;
output
x          alfa
0.99920  0.0008
```

Colocando o valor $F_{5,20}$ = 6,71, que é valor de F tabelado com 5 graus de liberdade no numerador e 20 no denominador, a função probf(6.71,5, 20) retorna o valor de 0,99920 que corresponde a 99,92%. Conclui-se, portanto, que o teste F para tratamentos foi altamente significativo (p < 0,0008), ou seja, com um grau de certeza de 99,92% e um erro α = 0,0008, que é a diferença (1 − 0,9992).

Resposta do exercício 4

Na saída (*output*) da rotina tem:

media li ls

104.133 99.261 109.005

li e *ls* representam, respectivamente, o limite inferior e o limite superior do intervalo de confiança da média (104,133), calculado com 95% de probabilidade:

[99,261; 109,005]

Resposta do exercício 5

```
data;
input x lambda;
prob420 = cdf('exponential', x, lambda); /*Prob de durar até 420 dias*/
valor = 1 - prob420; /*Prob de durar mais que 420 dias */
cards;
420 400
;
proc print; var prob420  valor;
run;
output
prob420    valor
0.65006    0.34994
```

Explicação:

1) μ é o parâmetro que determina a vida média do equipamento, (μ = 400).
2) λ é o parâmetro que determina a taxa com que um evento ocorre na unidade de tempo; é o inverso de μ (λ = 1/400).
3) x é a variável aleatória contínua (x > 0), que assume o valor de 420.

A probabilidade de durar até 420 dias é calculada pela função de distribuição acumulada:

$$P(x \leq 420) = 1 - e^{-\lambda x} = 1 - e^{-(1/400)\,420} \approx 0.65006$$

Na rotina SAS, P(x ≤ 420) é obtido por *prob420 = cdf('exponential', x, lambda);*

A probabilidade de durar mais de 420 dias é obtida subtraindo de 1, a probabilidade de durar até 420 dias.

$$P(x > 420) = 1 - P(x \leq 420) = 1 - (1 - e^{-\lambda x} = 1 - e^{-(1/400)\,420}) \approx 0{,}34994$$

Na rotina SAS é obtida por *valor = 1 - prob420;.*

Finalmente, a probabilidade deste equipamento durar mais de 420 dias é aproximadamente, 35,0%.

Capítulo 7

Resposta do exercício 1

O empirismo é caracterizado pelo conhecimento científico e acredita nas experiências e na intuição como formadoras das ideias. O inglês Francis Bacon (1561-1626), **mostrou a importância da experimentação e do empirismo para a aquisição dos conhecimentos científicos. Porém, é o francês René Descartes (1596−1650),** que é considerado o fundador da ciência moderna. O italiano Galileu Galilei (1564 -1642) é um dos responsáveis pelo estabelecimento das bases do pensamento científico moderno e o método experimental; o seu princípio racional era a matemática. **É atribuída a ele a frase "a matemática pode ajudar a compreender a natureza do mundo"**. Quanto ao experimento ou teste que é conduzido usando o método científico, pode-se afirmar que ele é usado para responder uma questão ou investigar um problema, testar teorias e, principalmente, testar hipóteses. Conforme afirmação do filósofo Popper (1959)

qualquer hipótese é falsificada. **As conclusões de um experimento podem apresentar argumentos** que levam à aceitação e ou rejeição de uma hipótese**. Um experimento não prova uma hipótese, ele apenas pode adicionar conhecimentos para mantê-la.** Uma hipótese estatística é uma suposição feita acerca de um parâmetro ou de uma característica da população, a qual pode ser aceita ou rejeitada por meio de **dados observados de um determinado experimento**. Para formular uma hipótese estatística, sempre iniciamos com a **hipótese nula (H_0)**, sendo que o objetivo é sempre rejeitar a hipótese nula (H_0) em favor de hipóteses alternativas (H_a).

Resposta do exercício 2

O teste de χ^2 é usado em inferência estatística, teste de hipóteses e construção de intervalos de confiança. Nesse e estudo, o valor do teste de χ^2 calculado com um grau de liberdade é $\chi^2_1 = 7,87$, é maior do que o valor tabelado, P ($\alpha = 0,01$) = 6,64. Conclui-se que o teste de χ^2_1 foi significativo ($P < 0,01$) e a hipótese de nula foi rejeitada.

Resposta do exercício 3

É importante que a hipótese formulada seja testável e que possa ser aprovada ou rejeitada com o resultado da experimentação. Seguem duas sugestões de hipóteses:

1) A aplicação de doses de adubo nitrogenado de até 100 kg ha^{-1} pode aumentar linearmente a produtividade de matéria seca de forragem, em kg de matéria seca ha^{-1}.

ou

2) A produtividade de matéria seca de forragem, em kg de matéria seca ha^{-1}, pode aumentar linearmente com a aplicação de doses de adubo nitrogenado de até 100 kg ha^{-1}.

Resposta do exercício 4

- O Material e Método ou Metodologia devem ser descritos de tal forma que um leitor possa ter todas as informações para repetir o experimento, caso deseje.
- Delineamento experimental: citar o modelo matemático, tamanho e forma de parcelas, subparcelas, variáveis que serão avaliadas com respectivas unidades – cm, kg, etc.

- Software que será usado para análise dos dados.
- Experimento: citar o local, data de instalação, período de inicio até o término.
- Nunca cite o nome comercial de um equipamento que será usado e sim a sua característica. Deve-se evitar em uma pesquisa científica qualquer tipo de material, equipamento, etc., que indique propaganda.

Resposta do exercício 5

O objetivo é rejeitar a hipótese de nulidade H_0 em favor da hipótese alternativa H_a.

H_0: p = 0,90 versus H_a: p > 0,90.

A etapa seguinte é calcular a estatística z:

$$z = \frac{\hat{p} - \Delta_0}{\sqrt{(\hat{p}\hat{q})/n}}$$

em que:

$\sqrt{\frac{\hat{p}\hat{q}}{n}}$ = estimativa do desvio-padrão da amostra em dados de proporções:

$\hat{p}$ = proporção de vacas na amostra com produtividade de leite acima de 20 kg dia^{-1}.

$$\hat{p} = \frac{184}{200} = 0,92;\ \hat{q} = 1 - \hat{p} = 0,08.$$

n = tamanho da amostra (n = 200).

Δ_0 = é a proporção hipotética (Δ_0 = 0,90).

$$z = \frac{\hat{p} - \Delta_0}{\sqrt{(\hat{p}\hat{q})/n}} = \frac{0,92 - 0,90}{\sqrt{(0,92x0,08)/200}} = 1,0426\ .$$

Consultando a Tabela 1.1 do Anexo 1, para a área sob a curva normal de 0 a 1,0426, encontra-se *z* = 0,8508. Assim, α = 1 - z = 1 - 0,8508 = 0,1492. Como geralmente aceitamos e ou rejeitamos uma hipótese com no máximo de 5% de erro (α = 0,05), nesse estudo a hipótese de nulidade não é rejeitada. Conclui-se que 90% das vacas da região de estudo não têm produtividade acima de 25 kg de leite por dia.

Capítulo 8

Resposta do exercício 1

A divergência fundamental entre os procedimentos General Linear Model (GLM) e Mixed Model (Mixed) está no erro do modelo matemático. Em razão disso, a aplicação deles são bastante divergentes. **O GLM é apropriado para ajustar modelos lineares gerais pelo método dos quadrados mínimos.** Permite executar vários tipos de análises: regressões, análises de variâncias univariadas e multivariadas, **e, em situações especiais, podem-se fazer análises de medidas repetidas (MR)**. O modelo linear padrão é do tipo $y_{nx1} = X_{nxp}\ b_{px1} + \varepsilon_{nx1,}$ em que y é o vetor de valores dependentes, sendo que E(y) = Xb; $Var(y) = V(\varepsilon) = \sigma^2 I$ em que σ^2 é o quadrado médio do erro; X é a matriz de especificação; b contém os efeitos fixos; ε é o vetor que contém os erros aleatórios. O GLM, exige que erros e_{ijk} sejam independentes, identicamente distribuídos com média zero **e com distribuição pelo menos aproximada da normal**. O procedimento Mixed, todavia, possibilita o ajuste de grande variedade de modelos lineares e não lineares a dados **que podem ser correlacionados e com variabilidade constante ou não constante, pois ele modela as variâncias e covariâncias**. As suposições básicas que são requeridas dos dados é que eles sejam normalmente distribuídos e que as variâncias e covariâncias dos dados devem exibir estrutura dentro daquelas disponíveis no Proc Mixed. O Proc Mixed ajusta essas estruturas por meio do método de quadrados mínimos, máxima verossimilhança (ML), **máxima verossimilhança restrita (REML), também conhecida como máxima verossimilhança residual.**

Resposta do exercício 2

Considerando as duas primeiras variâncias da diagonal principal e a covariância entre elas, tem-se:

$$\sigma_1^2 = 6{,}726,\ \sigma_2^2 = 7{,}878 \text{ e } (\sigma_1^2 + \sigma_2^2)/2 - \lambda = 2{,}032$$

Substituindo a covariância com as duas variâncias, tem-se:

$$(6{,}726 + 7{,}878)/2 - \lambda = 2{,}032$$

$$7{,}302 - \lambda = 2{,}032$$

$$-\lambda = 2{,}032 - 7{,}302$$

$$\lambda = 7{,}302 - 2{,}032 = 5{,}27$$

Observação: para que o teste F seja correto no teste de todas as hipóteses, é necessário que os erros e_{ijk} atendam às suposições de independência, normalidade e homogeneidade de variâncias. Para essa última suposição, a matriz de variâncias e covariâncias R precisa atender à condição de circularidade e esfericidade, ou seja, as variâncias da diferença entre quaisquer pares de medidas dentro da unidade experimental são iguais. Uma matriz com essa estrutura é a Huynh-Feldt (HF).

Resposta do exercício 3

Como as médias são iguais, os dados eram balanceados, pois nessa condição os dois métodos (GLM e Mixed) são concordantes quanto às estimativas de efeitos fixos.

Observa-se que os erros-padrão (EP) das médias obtidos por máxima verossimilhança restrita (Mixed) são maiores do que os obtidos por quadrados mínimos (GLM). Isso é explicado pelo fato de que os EP desses dois procedimentos são calculados diferentemente:

$$\text{Glm: } EP = \sqrt{\sigma^2 L(X'X)^{-}L'} \text{ ; Mixed: } EP = \sqrt{L(X'V^{-1}X)^{-}L'}$$

em que:

σ^2 = quadrado médio residual.

X = matriz de especificação.

L = matriz de hipótese.

V = matriz de variâncias e covariâncias.

A diferença reflete no cálculo dos erros-padrão das médias e demais cálculos derivados desses, tais como testes de hipóteses e intervalos de confiança.

Resposta do exercício 4

Quando várias estruturas de R são ajustadas, é comum escolher a de menor valor para AIC. Porém, é de interesse também escolher a que apresenta o menor número de parâmetros. Na Tabela 24 a matriz Csh tem o menor valor para AIC (AIC = 5598); porém, tem o maior número de parâmetros (136). Já a matriz Arma(1,1) tem o segundo menor valor de AIC e apenas três parâmetros. Portanto, é a escolhida.

Resposta do exercício 5

Abaixo estão as comparações corretas e onde tinham erros estão em negrito.

Un versus HF $\rightarrow |725,9 - 751,6| \rightarrow \chi^2_{20} = 25,7$ ns .

Un versus Arma(1,1) $\rightarrow |725,9 - 767,0| \rightarrow \chi^2_{25} = 41,1$ (P <0,05).

Un versus CS $\rightarrow |725,9 - \mathbf{769,5}| \rightarrow \chi^2_{26} = \mathbf{43,6}$ (P <0,05).

Un versus Ar1 $\rightarrow |725,9 - 773,7| \rightarrow \chi^2_{26} = \mathbf{47,8}$ (P <0,05).

Un versus Vc $\rightarrow |725,9 - 801,9| \rightarrow \chi^2_{27} = 76,0$ (P <0,05).

Resposta do exercício 6

No *split-plot*, são utilizados dois níveis de tratamentos ou fatores experimentais – tratamentos A que são distribuídos aleatoriamente às parcelas e tratamentos B que são distribuídos nas subparcelas. Geralmente a característica é avaliada apenas uma vez na unidade experimental. Os erros são independentes, identicamente distribuídos com média zero e com distribuição pelo menos aproximada da normal.

Quando as avaliações são realizadas periodicamente no indivíduo ao longo do tempo ou no espaço, como nesse estudo, tem-se as análises de medidas repetidas. As várias medidas são avaliadas na mesma unidade experimental, são correlacionadas e produzem uma estrutura de covariância. O procedimento Mixed disponibiliza cerca de 40 tipos de dessas estruturas de covariâncias, e para cada análise tem de selecionar a estrutura mais adequada.

Capítulo 9

Resposta do exercício 1

Quando se consegue um ambiente homogêneo onde pode instalar todos os tratamentos, o delineamento inteiramente (IC) é um dos mais utilizados e eficientes. Esse delineamento tem várias vantagens: não há limite para o número de tratamentos que se quer avaliar, tem maior flexibilidade, o número de repetições pode variar entre tratamentos, a análise estatística é simples, o número de graus de liberdade associado ao erro geralmente é grande, perda de parcela não é problema, o modelo

requer poucas suposições e o custo na condução do experimento é um dos menores. Dentre as desvantagens do IC tem-se a sua ineficiência **quando não se consegue um ambiente homogêneo, pois, nessa situação não ocorre homogeneidade entre as parcelas e não possibilitam a obtenção de resultados abrangentes**. O modelo matemático é do tipo $y_{ik} = \mu + t_i + r_j + \varepsilon_{ik}$, em que y_{ik} é o valor observado no tratamento i (t_i) da repetição j (r_j); μ é o efeito médio global; e ε_{ik} é o erro aleatório**. Supõe-se que eles ajustam a uma distribuição normal, são independentes e identicamente distribuídos com variância σ^2.**

Resposta do exercício 2

Para esse delineamento, o modelo matemático é:

$$y_{ijk} = \mu + G_i + P_j + GP_{ij} + \varepsilon_{ijk}$$

$$(i = 1, 2, 3; j = 1, 2, 3; k = 1, ..., 8)$$

em que:

y_{ijk} = peso corporal da novilha do grupo genético i, que recebeu o nível de proteína j e pertencente à repetição k.

μ = efeito médio global.

G_i = efeito do grupo genético i.

P_j = efeito do nível de proteína j.

GP_{ij} = efeito da interação entre grupo genético e nível de proteína.

ε_{ijk} = erro aleatório associado ao valor y_{ijk}; $\varepsilon_{ijk} \sim NIID\ (0, \sigma^2)$.

Anova

Fator de variação	GL
Grupo genetico – g	2
Proteina bruta – p	2
Interação g x p	4
Erro	63
Total	71

Rotina SAS

Admitindo, como dados fictícios, que o peso da novilha (y) para rep = 1, G = 1 e P = 1, seja 220 kg e para a última repetição: rep = 8, G = 3, P = 3, seja 330 kg, a análise de variância fica:

```
data;
input rep g p y;
cards;
1 1 1 220
...
8 3 3 330
;
proc glm; class g p;
model y = g p g*p/ss3;
lsmeans g p g*p / tukey; run;
```

Resposta do exercício 3

Como na figura os cinco gráficos foram construídos com a mesma escala para o eixo y, é possível interpretar Fibra em detergente neutro,%, (FDN) em um esquema fatorial 2x5 x 5 (duas fontes de N: ureia e nitrato de amônio; cinco doses de nitrogênio (N): [(0 kg ha^{-1}, 25 kg ha^{-1}, 50 kg ha^{-1}, 100 kg ha^{-1}e 200 kg ha^{-1}) e cinco cortes (1 a 5)].

Existem variações para ureia e nitrato de amônio que dependem da dose e do corte. Observa superioridade do nitrato de amônia em relação à ureia no corte 4. Nos cortes de 1 a 4 observa redução da FDN, tanto para ureia quanto para nitrato de amônio, à medida que as doses de N aumentam. Já no corte 5, observa leve aumento para as duas variáveis à medida que as doses aumentam. Porém, nos cortes 2, 3 e 5, há superioridade da ureia.

Resposta do exercício 4

A alternativa correta é:

(X) Todas as respostas acima são verdadeiras.

Resposta do exercício 5

Consultando o valor de F tabelado (Tabela 4.1, Anexo 1) para 4 GL no numerador e 25 GL no denominador, para α = 0,05, tem-se: $F_{4,25}$ = 2, 78. Como o F calculado ($F_{4,25}$ = 31,23), é muito superior, uma maneira correta de encontrar o verdadeiro valor de α e, consequentemente, do grau de certeza é por meio da rotina SAS abaixo. Observa-se que o valor de α (α = 2.1553E-9) foi praticamente zero e o grau de confiança ou de certeza (1- α = 1). Conclui-se que o teste F foi altamente significativo, e que as cinco variedades de batata diferem entre si em relação à produção, com um grau de confiança superior a 99% de probabilidade.

```
data;
input x n1 n2;
area = probf(x, n1, n2);
alfa = 1- area;
datalines;
31.23 4  25
;
proc print; var alfa area; run;
output
alfa            area
2.1553E-9   1.00000
```

Capítulo 10

Resposta do exercício 1

No delineamento tradicional em blocos casualizados (BC), que é um dos mais utilizados em agricultura, uma exigência é que o ambiente seja o mais homogêneo possível dentro dos blocos, **porém, pode haver grande variação entre os blocos**. **Os tratamentos são sorteados dentro de cada bloco para garantir independência entre os erros**. No DBC, **todos os tratamentos, que podem ser apenas um fator ou organizados na forma fatorial são distribuídos dentro de cada bloco**. Uma vantagem do delineamento em DBC, **quando comparado com o delineamento inteiramente casualizado é que, em algumas situações é mais fácil encontrar um local homogêneo para instalar um bloco do que para instalar um experimento**. No delineamento em DBC com parcela dividida (*split-plot*), **que tem aplicação em várias áreas**, a parcela principal é

dividida em subunidades chamada de subparcelas. No *split-plot*, na parcela é distribuído o tratamento principal A, enquanto a subparcela recebe o tratamento secundário B. Na análise de variância deste delineamento, **para testar o tratamento principal A, utiliza o quadrado médio residual que é formado pela interação tratamentos × blocos (erro a). Para testar o tratamento secundário B, que é distribuído nas subparcelas, utiliza o erro aleatório entre subparcelas (erro b).**

Resposta do exercício 2

Na Tabela 10 tem-se os resultados da análise de variância de três efeitos principais quanto à produtividade de matéria seca (PMS), e dentro de cada efeito principal a diferença foi altamente significativa (P < 0,0001). Nas duas fontes de N, o nitrato de amônio foi significativamente superior em relação à ureia nos dois anos agrícolas (1998−1999, 1999−2000). Quanto às cinco doses de N, observa uma resposta crescente e praticamente linear na PMS com o aumento das doses, em ambos os anos agrícolas. Quanto aos cortes, observa-se oscilações na PMS. Para uma interpretação mais precisa desses resultados, a primeira etapa é realizar uma análise exploratória com a elaboração de gráficos em linhas x-y, para cada corte e cada ano agrícola (1998−1999) e (1999−2000), com PMS no eixo *y*, doses no eixo *x*, e duas linhas (1 − ureia; e 2 − nitrato de amônio), resultando em dez gráficos. Com a elaboração desses gráficos teria uma visão detalhada dos resultados do experimento e, posteriormente, para os dados utilizados na elaboração de cada gráfico, calcular regressão linear, quadrática, etc., com PMS como variável dependente (y) e doses de N (x) como variável explanatória.

Resposta do exercício 3

( ) Houve perda de uma parcela.

(X) Houve perda de uma subparcela.

( ) Houve perda de duas subparcelas.

( ) Não houve perda de parcela e nem de subparcela.

Houve perda de uma subparcela, pois com o experimento balanceado, tem-se grau de liberdade do erro b = 32 e total = 59.

Resposta do exercício 4

No BIB, se existem k tratamentos a serem instalados e em cada bloco tem ***a*** tratamentos (a < k), então o número de blocos necessários para instalar todos os k tratamentos é calculado pela fórmula:

$$\binom{k}{a} = \frac{k!}{a!(k-a)!}$$

Como temos k = 6 e a = 5, então:

$$\binom{6}{5} = \frac{6!}{5!(6-5)!} = \frac{6!}{5!1!} = 6 \text{ blocos}$$

Resposta do exercício 5

O croqui de campo para blocos incompletos balanceados com seis tratamentos, seis blocos e cinco tratamentos em cada bloco, é apresentado abaixo. Cada par de tratamentos ocorre em cinco blocos.

Bloco	Tratamento				
1	a_1	a_2	a_3	a_4	a_5
2	a_1	a_3	a_5	a_6	a_2
3	a_1	a_4	a_5	a_6	a_3
4	a_1	a_5	a_6	a_2	a_4
5	a_1	a_6	a_2	a_3	a_4
6	a_2	a_3	a_4	a_5	a_6

Para elaborar o croqui de campo, associamos algumas variáveis:

k = número de tratamentos (k = 6).

b = número de blocos (b = 6).

a = número de tratamentos por bloco (a = 5).

r = número de repetições de um tratamento ou o número de vezes que um tratamento ocorre no experimento (r = 5).

n = número de observações (n = ba = kr = 30).

Finalmente, para elaborar o croqui de campo, calculamos λ, que é o número de vezes que cada par de tratamento ocorre no mesmo bloco:

$$\lambda = r(a - 1)/(k - 1) = 5(5 - 1)/(6 - 1) = 20/5 = 4$$

Capítulo 11

Resposta do exercício 1

O delineamento em quadrado latino (QL) é bastante útil **em várias situações experimentais e, principalmente, na agricultura.** Em geral, o fator distribuído na linha é o sujeito que receberá o tratamento, enquanto as colunas representam o tempo, ocasião, período em que os tratamentos são avaliados. Nesse desenho, as unidades experimentais ou parcelas onde os tratamentos são aplicados possuem dois controles locais – linhas e colunas. **Dentro de linhas e dentro de colunas, as condições precisam ser as mais homogêneas possíveis**. Dentre as desvantagens dos delineamentos em QL, tem-se o fato de que o tipo de sorteio de um experimento é único e também o fato de que não se pode calcular efeito de interação entre linhas e colunas e tratamentos. O desenho em quadrado latino possui número igual de linhas, de colunas e de tratamentos, com os *t* tratamentos distribuídos em *t* linhas e *t* colunas, surgindo daí a denominação QL t x t, por exemplo, QL do tipo 3 x 3, 4 x 4, 5 x 5, 6 x 6, 7 x 7, 8 x 8, 9 x 9. **Quanto ao tamanho do QL, caso as necessidades do pesquisador seja um QL 3 x 3 ou QL 4 x 4, algumas modificações experimentais precisam ser feitas, tais como, aumentar o número de indivíduos (linhas), considerar três QL 3 x 3 ou dois QL 4 x 4, além de outras modificações no desenho de campo.**

Resposta do exercício 2

Nesse caso, o experimento seria inviabilizado. Somente pode tolerar a morte de um animal se ela ocorrer no último período. Nessa situação, a análise de variância pode ser efetuada sem problemas pelo procedimento GLM do SAS, utilizando a SQ do tipo III para testar todas as hipóteses.

Resposta do exercício 3

A análise de variância do QL 4 x 4 realizado em três épocas será a seguinte:

Fator de variação	Graus de liberdade
Animais	3
Períodos (QL)	9
Tratamentos	3
Erro	32
Total	47

Resposta do exercício 4

(X) QL 6 x 6 com perda de uma parcela.

No mesmo experimento caso não ocorresse perda de parcela, o grau de liberdade (GL) do erro seria 20 e GL total seria 35.

Resposta do exercício 5

```
data;
input x n1 n2;
area = probf(x, n1, n2);
alfa = 1 - probf(x, n1, n2);
datalines;
1.54  4  12
0.25 4  12
9.81 4 12
;
proc print; var area alfa; run;
output
    area       alfa
  0.74731    0.25269
  0.09586    0.90414
  0.99908    0.00092
```

A Tabela 13 (Capítulo 11) com os valores corretos para Pr > F será:

Fator de variação	Grau de liberdade	Soma de quadrados	Quadrado médio	F	Pr > F
Linhas	4	167,84	41,96	1,54	0,25269
Colunas	4	27,04	6,76	0,25	0,90414
Tratamentos	4	1.072,64	268,16	9,81	0,00092
Erro	12	327,92	27,33		
Total	24	1.595,44			

Capítulo 12

Resposta do exercício 1

Regressão linear, covariância e correlação são técnicas estatísticas utilizadas no estudo entre duas variáveis *x* e *y*, que necessariamente devem ser contínuas. A regressão linear tem por objetivo estimar y_i em função de x_i em uma relação de n pares de variáveis aleatórias (y_i, x_i, i = 1, 2, ... , n), porém o modelo estimado é mais importante quando **existe uma relação linear entre *x* e *y***. Na covariância é determinado o parentesco entre estas duas variáveis **aleatória**s, isto é, como elas se covariam. É **uma medida da variabilidade conjunta dessas duas variáveis**. A regressão linear estuda a relação entre *x* e *y* e deve existir uma relação de causa e efeito entre elas; **já a correlação estuda a relação linear entre *x* e *y*. Exige apenas que essas duas variáveis estejam associadas, porém, sem uma relação de causa e efeito entre elas**. **Em qualquer tipo de estudo, os valores obtidos e significâncias estatísticas dessas três técnicas são dependentes do tamanho amostral.** Quando se deseja saber apenas o parentesco da relação linear entre duas variáveis aleatórias *x* e *y* obtidas de uma amostra de tamanho *n*, principalmente quando elas são medidas em um mesmo indivíduo, o coeficiente de correlação *r* (-1 < r < 1) é suficiente.

Resposta do exercício 2

Com base na Figura 1, verificam-se doses crescentes do teor de fósforo (x) (0,0 g kg^{-1}; 2,0 g kg^{-1}; 4,0 g kg^{-1}; 6,0 g kg^{-1}; 8,0 g kg^{-1}), que proporcionaram um crescimento linear na produção de matéria seca (y), g $vaso^{-1}$, que foi estimada pela regressão linear $\hat{y}$ = -0,414 + 0,600x, e a correlação linear entre as duas variáveis aleatórias x e y foi de

r = 0,89**. Embora na Figura 1 mostram-se alguns pontos discrepantes, acima e abaixo da reta, a notação "**" indica que a significância foi ao nível de 1% de probabilidade.

Resposta do exercício 3

```
data;
input x media;
y = 56 + 0.1*x;
res = y - media;
cards;
0    50.0
25   60.0
50   70.0
75   65.0
100  60.0
;
proc print; var x y media res;run;
```

Os cinco valores estimados (ŷ) e os cinco resíduos estão abaixo.

x	ŷ	media	resíduo
0	56.0	50	6,0
25	58,5	60	-1,5
50	61,0	70	-9,0
75	63,5	65	-1,5
100	66,0	60	6,0

Resposta do exercício 4

Para verificar se a regressão linear foi suficiente para explicar se doses crescentes do teor de fósforo (x) (0,0 g kg^{-1}; 2,0 g kg^{-1}; 4,0 g kg^{-1}; 6,0 g kg^{-1}; 8,0 g kg^{-1}) explicam o crescimento linear na produção de matéria seca (y), g $vaso^{-1}$, inicialmente calculamos os dois valores de F:

Devido à regressão linear: $F_{1,10} = 26,78$.

Desvio da regressão linear: $F_{3,10} = 22,50$.

Com esses dois valores calculamos a probabilidade de F pela rotina SAS a seguir:

```
data;
input x n1 n2;
area = probf(x, n1, n2);
alfa = 1 - probf(x, n1, n2);
datalines;
26.78 1 10
22.50 3 10
;
proc print; var area alfa; run;
output
area        alf
0.99958   .00041611
0.99991   .00009146
```

Conclusão: observa-se que a regressão linear foi altamente significativa (p < ,00041). Também se observa que o desvio da regressão linear foi altamente significativo (p <,00009), indicando que o cálculo da regressão linear deve ser revisto e substituído para uma regressão curvilinear.

Resposta do exercício 5

Abaixo é apresentada a Tabela 10 (Capítulo 12) com a interpretação correta para os coeficientes de correlação. Onde houve mudanças está em negrito.

Coeficiente de correlação	Interpretação
r = -1	**Correlação negativa perfeita**
-0,2 ≤ r < 0,1	**Correlação praticamente nula**
0,2 ≤ r ≤ 0,4	**Correlação positiva fraca** ou razoavelmente positiva
0,7 ≤ r ≤ 0,9	Correlação positiva forte ou altamente positiva
-0,3 ≤ r ≤ -0,1	**Correlação negativa fraca ou razoavelmente negativa**
r = 0	Correlação nula
r = 1	Correlação positiva perfeita

Resposta do exercício 6

Quanto à correlação linear entre as duas variáveis aleatórias − TMA dia^{-1} (y) e idade em dias (x), para ambos os modelos (Brody e Von Bertalanffy), ela seria negativa, com correlação negativa quase perfeita para o modelo Brody e uma correlação negativa

mais fraca para o modelo Von Bertalanffy. Quanto ao ajuste de uma regressão linear de y = f(x), certamente ela teria um nível de significância de 5% de probabilidade para o modelo Brody e seria não significativa a este nível para o modelo Von Bertalanffy. Para este modelo, uma regressão curvilinear seria mais adequada.

Capítulo 13

Resposta do exercício 1

Dados categorizados representam categorias ou características, que são susceptíveis de medida, porém, não de classificação. A análise de dados categorizados organizados em tabelas de contingência é comum nas mais diversas áreas do conhecimento. Um exemplo são os questionários que são utilizados para as mais diversas finalidades, e as perguntas fechadas geralmente possibilitam respostas do tipo sim ou não; múltipla escolha, com respostas do tipo escalas de avaliação, perguntas com escores e ou notas representam dados categorizados. Os dados categorizados são organizados em tabelas de contingências e analisados pelo teste de qui-quadrado de Pearson (χ^2) e testes correlatos. O teste de χ^2 deve ser usado quando, ambos, o número de observações em cada célula ($\mathbf{n}_{ij}$) e a menor frequência esperada ($\mathbf{e}_{ij}$) forem maiores ou iguais **a 5 (nij > 5, eij > 5).** Quando *n* < 40 ou uma das células nij for menor que **5,** utiliza-se a correção de Yates, a qual ajusta a fórmula do teste de χ^2 por subtrair 0,5 da diferença entre cada valor observado e seu valor esperado. Essa correção **reduz o valor de χ^2 obtido e aumenta o valor de *p***, evitando a superestimação da significância estatística para **frequências pequenas**. Existem vários testes, destacando-se a estatística de Cochran-Mantel-Haenszel que testa a hipótese **Ho: as respostas em cada estrato são independentes de tratamentos. versus Ha: as respostas em cada estrato são dependentes de tratamentos**. Já o teste de McNemar é apropriado para tabelas de contingência 2 x 2 **quando uma amostra de n indivíduos é submetida a uma situação antes e após**, em que cada paciente serve como seu próprio controle; quando se deseja testar a diferença entre proporções pareadas.

Resposta do exercício 2

A primeira observação a ser feita é se a hipótese nula para homogeneidade marginal se verifica, isto é:

$$n_{1.} = n_{.1} \text{ e } n_{2.} = n_{.2}$$

Caso ocorre a hipótese nula de homogeneidade marginal, significa que não houve efeito no tratamento. Verifica-se que isso não ocorre, pois 11 ≠ 18 e 64 ≠ 57.

O teste de qui-quadrado com 1 grau de liberdade (GL) é calculado por:

$$\chi^2_1 = (n_{12} - n_{21})^2 / (n_{12} + n_{21}) = (2 - 9)^2 / (2 + 9) = 4{,}45$$

$$\chi^2_1 = 4{,}45$$

Como $(n_{12} + n_{21}) < 25$, recomenda-se a correção de continuidade no cálculo do qui-quadrado do teste de McNemar.

$$\chi^2_1 = (|n_{12} - n_{21}| - 1)^2 / (n_{12} + n_{21}) = 5{,}81$$

Consultando a Tabela 2 (Anexo), o valor de qui-quadrado com 1 GL para α = 0,05 é 3,841 e para α = 0,01 é 6,635. Assim, rejeita-se ao nível de 5% de probabilidade a hipótese nula de que o tratamento não foi eficiente.

Resposta do exercício 3

Como tem sete colunas calcula-se o teste de χ^2 com seis graus de liberdade:

$$\chi^2_6 = \sum_{i=1}^{c} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$$

em que:

$$n_{11} = 25,\ n_{12} = 28;\ n_{13} = 30;\ n_{14} = 29;\ n_{15} = 31;\ n_{16} = 27;\ n_{17} = 32$$

$$e_{11} = e_{12} = e_{13} = e_{14} = e_{15} = e_{16} = e_{17} = 30$$

A estatística de χ^2 com seis graus de liberdade é calculada por:

$$\chi^2_6 = \sum_{i=1}^{7} \frac{(35-30)^2}{35} + \frac{(28-30)^2}{28} + \frac{(30-30)^2}{30} + \frac{(29-30)^2}{29} + \frac{(31-30)^2}{31} + \frac{(27-30)^2}{27} + \frac{(32-30)^2}{32} = 1{,}3822$$

Consultando a Tabela 2 (Anexo), o valor de χ^2 com seis graus de liberdade para α = 0,05 é 12,592, indicando que o teste de qui-quadrado calculado ($\chi^2_6 = 1{,}3822$) é não

significativo e a hipótese de nulidade não é rejeitada. Portanto, pode-se afirmar que o nascimento de bezerros é constante e com uma frequência de 30 diariamente.

Resposta do exercício 4

Considerando as duas proporções de abscessos hepáticos ausente:

$\hat{p}_1 = x_1/n_1 = 6/390 = 0,0154$ e $\hat{p}_2 = x_2/n_2 = 11/268 = 0,0410$

O interesse é testar a hipótese:

- H_0: $p_1 - p_2 = 0$ versus Ha_1: $p_1 - p_2 > 0$
- H_0: $p_1 - p_2 = 0$ versus Ha_2: $p_1 - p_2 < 0$

Para isso calcula-se o teste z:

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{p_0(1-p_0)\,(1/n_1 + 1/n_2)}}$$

em que:

$x_1 = 6$; $x_2 = 11$; $n_1 = 390$; $n_2 = 268$

$\hat{p}_1 = x_1/n_1 = 6/390 = 0,0154$

$\hat{p}_2 = x_2/n_2 = 11/268 = 0,0410$

$$p_0 = \frac{x_1 + x_2}{n_1 + n_2} = \frac{6 + 11}{390 + 268} = 0,0258$$

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{p_0(1-p_0)\,(1/n_1 + 1/n_2)}} = \frac{(0,0154 - 0,0410) - 0}{\sqrt{0,0258(1 - 0,0258)\,(1/268 + 1/390)}} = -2,0385$$

Como $|z| > 1,96$, o valor z = -2,0385 é significativo ao nível de 5% de probabilidade. Portanto, rejeita-se a hipótese de nulidade: H_0: $p_1 - p_2 = 0$ e aceita-se a hipótese alternativa: Ha_2: $p_1 - p_2 < 0$. Conclui-se que ausência de abscessos hepáticos de caprinos jovens (0−12 meses) é menor do que de caprinos adultos (> 12 meses).

Resposta do exercício 5

O interesse é verificar se a proporção de animais recuperados no grupo vacinado ($\hat{p}_{11}$ = 5/9 = 0,55) é superior à proporção de animais recuperados no grupo-controle ($\hat{p}_{21}$ = 5/25 = 0,20).

No teste exato de Fisher, o valor de probabilidade p é calculado por:

$$P = n_{1.}!\, n_{2.}!\, n_{.1}!\, n_{.2}!/(\, n_{11}!\, n_{12}!\, n_{21}!n_{22}!\, n!)$$

A probabilidade p é calculada por meio da rotina SAS.

```
data;
input n11 n12 n21 n22;
n = n11+ n12+ n21+ n22;
p = fact(n11+n12)*fact(n21+n22)*fact(n11+n21)*fact(n12+n22)/
    (fact(n11)*fact(n12)*fact(n21)*fact(n22)* fact(n));
cards;
5 4 5 20
;
proc print; var p;run;
output
 p

0.0510
```

Para p = 0,0510, o teste exato de Fisher é significativo ao nível de 5% de probabilidade. Rejeita-se, portanto, a hipótese de nulidade e conclui-se que a amostra forneceu evidência de que o grupo de animais tratado difere do grupo de animais-controle. Realmente a proporção de animais recuperados no grupo vacinado ($\hat{p}$11 = 5/9 = 0,55) é bastante superior à proporção de animais recuperados no grupo-controle ($\hat{p}_{21}$ = 5/25 = 0,20).

A obra *Estatística Experimental na Agropecuária* aborda tópicos fundamentais da estatística experimental. Embora as discussões e as aplicações sejam mais direcionadas a estudantes e profissionais ligados às ciências agrárias, o conteúdo do livro é destinado a todos que necessitam fazer uso e aplicação da estatística experimental.

Em todo livro, são discutidas várias aplicações da estatística, na maioria das vezes com dados reais, por meio de exercícios resolvidos e propostos, usando recursos computacionais na análise de dados e solução de problemas com o auxílio do software Statistical Analysis System (SAS), utilizado por grandes corporações em análises de delineamentos e procedimentos.

MINISTÉRIO DA
AGRICULTURA, PECUÁRIA
E ABASTECIMENTO

PÁTRIA AMADA
BRASIL
GOVERNO FEDERAL

ISBN 978-65-89957-26-3
9 786589 957263
CGPE 017808