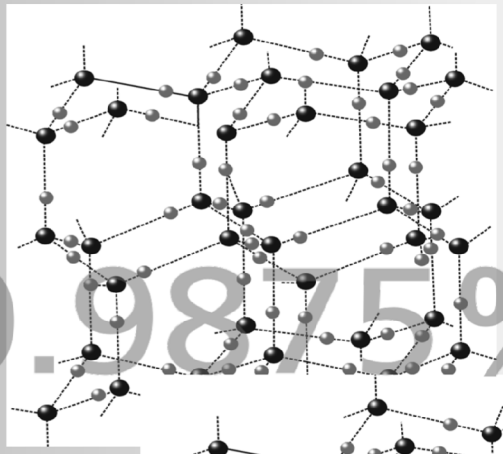
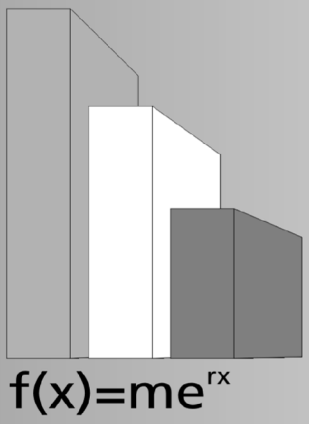
Módulos 2 e 3

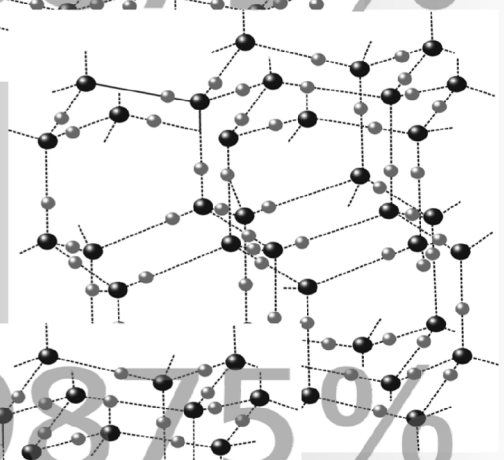
Jean Louis Valentin
Luiz Manoel Figueiredo
Mario Olivero
Mariza Ortegoza

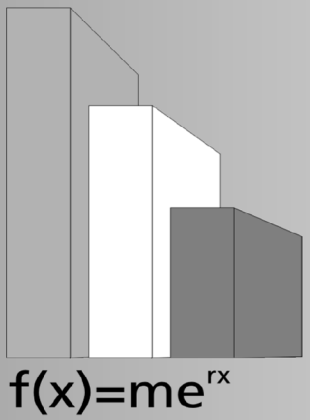
Volume 2
2ª edição

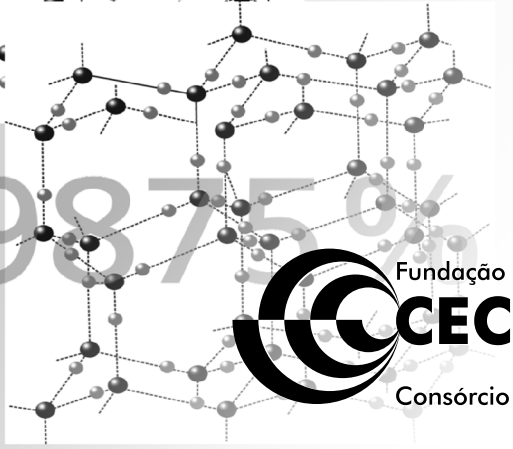
Elementos de Matemática e Estatística











Fundação CECIERJ
Consórcio cederj

Fundação CECIERJ

Consórcio cederj

Centro de Educação Superior a Distância do Estado do Rio de Janeiro

Elementos de Matemática e Estatística

Volume 2 - Módulos 2 e 3

2ª edição

Jean Louis Valentin
Luiz Manoel Figueiredo
Mario Olivero
Mariza Ortegoza

GOVERNO DO Rio de Janeiro

SECRETARIA DE CIÊNCIA E TECNOLOGIA

Ministério da Educação

BRASIL UM PAÍS DE TODOS GOVERNO FEDERAL

Apoio:

FAPERJ
Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro

Fundação Cecierj / Consórcio Cederj

Rua Visconde de Niterói, 1364 – Mangueira – Rio de Janeiro, RJ – CEP 20943-001
Tel.: (21) 2334-1569 Fax: (21) 2568-0725

Presidente
Masako Oya Masuda

Vice-presidente
Mirian Crapez

Coordenação do Curso de Biologia
UENF - Milton Kanashiro
UFRJ - Ricardo Iglesias Rios
UERJ - Cibele Schwanke

Material Didático

ELABORAÇÃO DE CONTEÚDO
Jean Louis Valentin
Luiz Manoel Figueiredo
Mario Olivero
Mariza Ortegoza

COORDENAÇÃO DE DESENVOLVIMENTO INSTRUCIONAL
Cristine Costa Barreto

DESENVOLVIMENTO INSTRUCIONAL E REVISÃO
Anna Maria Osborne
Anna Carolina da Matta Machado
Maria Helena Hatschbach

COORDENAÇÃO DE LINGUAGEM
Maria Angélica Alves

Departamento de Produção

EDITORA
Tereza Queiroz

COORDENAÇÃO EDITORIAL
Jane Castellani

REVISÃO TIPOGRÁFICA
Jane Castellani
Sandra Valéria Oliveira

COORDENAÇÃO DE PRODUÇÃO
Jorge Moura

PROGRAMAÇÃO VISUAL
Sanny Reis
Yozo Kono
Ronaldo d´Aguiar Silva

ILUSTRAÇÃO
Jefferson Caçador
Morvan Neto

CAPA
Eduardo Bordoni

PRODUÇÃO GRÁFICA
Patricia Seabra

V156e

Valentin, Jean Louis.
Elementos de matemática e estatística. v. 2 / Jean Louis Valentin. 2ª ed. – Rio de Janeiro: Fundação CECIERJ, 2009.
120p.; 19 x 26,5 cm.

ISBN: 85-7648-033-6

1. Variável aleatória. 2. Distribuição binomial. 3. Distribuição de frequência. 4. Análise de variância. 5. Estimativa. I. Figueiredo, Luiz Manoel. II. Olivero, Mario. III. Ortegoza, Mariza. IV. Título.

CDD: 519.5

2009/2

Referências Bibliográficas e catalogação na fonte, de acordo com as normas da ABNT.

Governo do Estado do Rio de Janeiro

Governador
Sérgio Cabral Filho

Secretário de Estado de Ciência e Tecnologia
Alexandre Cardoso

Universidades Consorciadas

UENF - UNIVERSIDADE ESTADUAL DO NORTE FLUMINENSE DARCY RIBEIRO
Reitor: Almy Junior Cordeiro de Carvalho

UERJ - UNIVERSIDADE DO ESTADO DO RIO DE JANEIRO
Reitor: Ricardo Vieiralves

UFF - UNIVERSIDADE FEDERAL FLUMINENSE
Reitor: Roberto de Souza Salles

UFRJ - UNIVERSIDADE FEDERAL DO RIO DE JANEIRO
Reitor: Aloísio Teixeira

UFRRJ - UNIVERSIDADE FEDERAL RURAL DO RIO DE JANEIRO
Reitor: Ricardo Motta Miranda

UNIRIO - UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO
Reitora: Malvina Tania Tuttman

Elementos de Matemática e Estatística

Volume 2

SUMÁRIO

Módulo 2

Módulo 3

Variável aleatória e valor esperado

AULA 14

objetivos

Nesta aula você deverá ser capaz de:

- Compreender um conceito muito importante nos estudos estatísticos: a variável aleatória.
- Saber como calcular seu valor esperado.

VARIÁVEL ALEATÓRIA

Os resultados de um experimento aleatório podem ser numéricos ou não. Experimentos como:

• anotar os tempos em uma maratona;

• medir a taxa de precipitação pluviométrica durante um período;

• lançar uma moeda três vezes e anotar a quantidade de coroas

que ocorrem, têm seus espaços amostrais constituídos de números. Muitos experimentos, porém, possuem resultados qualitativos (e não quantitativos). Por exemplo:

– entrevistar um eleitor, antes de uma eleição, para conhecer sua preferência;

– inspecionar uma lâmpada para verificar se é ou não defeituosa;

– lançar uma moeda e observar se dá cara ou coroa.

Podemos, então, classificar os resultados de um experimento como quantitativos ou qualitativos. Os estatísticos trabalham com os dois tipos, embora os quantitativos sejam mais comuns.

Em certos casos, é possível converter dados qualitativos em quantitativos, associando um valor numérico a cada resultado. Vamos ver alguns exemplos.

Exemplo 1

1. Experimento: "lançamento de duas moedas e observação do par obtido".

– Espaço amostral associado: $\Omega = \{(K,K), (K,C), (C,K), (C,C)\}$

– Um resultado numérico que podemos definir: contar o número de caras, isto é, fazer a seguinte associação:

resultado	valor numérico associado
(C, C)	0
(K, C)	1
(C, K)	1
(K, K)	2

2. Experimento: "retirada de uma lâmpada de um lote e observação se é (sim) ou não (não) defeituosa".

- espaço amostral associado: $\Omega = \{\text{sim}, \text{não}\}$

- um resultado numérico que podemos definir: contar o número de lâmpadas defeituosas, isto é:

resultado	valor numérico associado
sim	1
não	0

3. Experimento: "lançamento de um dado e observação da face de cima".

- espaço amostral associado: $\Omega = \{1, 2, 3, 4, 5, 6\}$

Note que, neste caso, os resultados do experimento já são numéricos. Mesmo assim, podemos associar-lhes outros números. Por exemplo, contar a ocorrência de números ímpares, isto é:

resultado	valor numérico associado
1	1
2	0
3	1
4	0
5	1
6	0

Pelos exemplos acima, você pode notar que, a cada resultado, corresponde um e apenas um valor numérico. Esse procedimento pode ser visto, matematicamente, como a criação de uma função. Tal função é chamada *variável aleatória*.

Temos, então, a seguinte definição:

> Variável aleatória é uma função numérica definida em um espaço amostral.

!

Você deve achar estranho chamar uma função de *variável* aleatória. E é mesmo! Mas essa terminologia já é consagrada na área e por isso vamos adotá-la. Não se esqueça, porém: apesar do nome, trata-se de uma função.

De modo geral, dado um experimento de espaço amostral Ω, uma variável aleatória X é uma função

$$X : \Omega \to \mathbb{R}$$

que associa cada evento elementar a um número real. Em nosso curso, trabalharemos apenas com as chamadas variáveis aleatórias discretas, que são aquelas que assumem valores num subconjunto enumerável de $\mathbb{R}$. Mais particularmente, as variáveis que estudaremos assumirão apenas uma quantidade finita de valores.

!

Um conjunto é *enumerável* quando é finito ou quando existe uma bijeção (relação 1 para 1) entre ele e um subconjunto do conjunto dos números naturais. Um conjunto enumerável pode ter seus elementos listados em seqüência: $\{x_1,..., x_n\}$ se finito ou $\{x_1,..., x_n...\}$ se infinito.

DISTRIBUIÇÃO DE PROBABILIDADE

Dado um certo experimento aleatório, podemos interpretar os valores assumidos por uma variável aleatória como eventos numéricos associados àquele experimento. Vamos deixar isso mais claro, retomando o exemplo do lançamento das duas moedas. Estamos interessados em contar o número de caras. Definimos, então, a variável aleatória

$$\begin{aligned} \Omega &\to X \\ (C, C) &\to 0 \\ (C, K) &\to 1 \\ (K, C) &\to 1 \\ (K, K) &\to 2 \end{aligned}$$

Para cada valor de X, identificamos os resultados do experimento que lhe são associados:

evento numérico		**eventos associados**
$X = 0$	$\to$	$\{(C, C)\}$
$X = 1$	$\to$	$\{(C, K), (K, C)\}$
$X = 2$	$\to$	$\{(K, K)\}$

Sendo Ω = {(*K*,*K*), (*K*,*C*), (*C*,*K*), (*C*,*C*)} eqüiprovável, cada resultado tem probabilidade 1/4. Podemos determinar a probabilidade de ocorrência de cada evento numérico, a partir das probabilidades dos eventos do experimento:

$$\begin{aligned} P(X=0) &= P\{(C,C)\} = 1/4 \\ P(X=1) &= P\{(C,K),(K,C)\} = 1/2 \\ P(X=2) &= P\{(K,K)\} = 1/4 \end{aligned}$$

e construir a tabela:

X	P
0	1/4
1	2/4
2	1/4

Note que essa tabela, na qual anotamos X e suas respectivas probabilidades, caracteriza uma função que a cada valor de X associa um número real do intervalo [0,1]. Esta função é denominada **distribuição de probabilidade** da variável aleatória X.

Observação: é importante destacar que foram definidas duas funções:

1. *variável aleatória*, que associa a cada resultado de um experimento um número real; e

2. *distribuição de probabilidade* de uma variável aleatória, que associa a cada valor assumido pela variável um número real restrito ao intervalo [0,1].

Resumindo: Dado um experimento aleatório de espaço amostral Ω, uma variável aleatória X é uma função

$$X : \Omega \to \{x_1, \ldots, x_n\}$$

A escolha dos números $P(x_i) = P(X = x_i),\ i = 1, \ldots, n$, é determinada a partir das probabilidades associadas aos eventos no espaço amostral Ω, no qual X está definida.

Exemplo 2

Retomemos os experimentos do exemplo 1, e vamos supor que os espaços amostrais sejam todos eqüiprováveis:

1.

X	evento	probabilidade
0	{(C,C)}	P(X = 0) = 1/4
1	{(K,C)},{(C,K)}	P(X = 1) = 2/4
2	{(K,K)}	P(X = 2) = 1/4

Distribuição de probabilidade da variável aleatória *X*:

X	P
0	1/4
1	2/4
2	1/4

2.

X	evento	probabilidade
0	{não}	P(X = 0) = 1/2
1	{sim}	P(X = 1) = 1/2

Distribuição de probabilidade da variável aleatória *X*:

X	P
0	1/2
1	1/2

3.

X	evento	probabilidade
0	{2, 4, 6}	P(X = 0) = 3/6
1	{1, 3, 5}	P(X = 1) = 3/6

Distribuição de probabilidade da variável aleatória *X*:

X	P
0	3/6
1	3/6

VALOR ESPERADO DE UMA VARIÁVEL ALEATÓRIA

Vamos definir uma grandeza que irá refletir nossa expectativa em relação ao valor de uma variável aleatória.

Suponha que lancemos um dado equilibrado 300 vezes e anotemos o resultado da face de cima. Queremos determinar a média dos valores observados.

Como os resultados possíveis são eqüiprováveis, é de se esperar que cada um ocorra uma quantidade de vezes próxima de 50 (já que são 300 lançamentos e 6 resultados possíveis). A média dos valores deve ser, então, um valor próximo de:

$$\text{média} = \frac{1 \times 50 + 2 \times 50 + 3 \times 50 + 4 \times 50 + 5 \times 50 + 6 \times 50}{300} = 3,5 .$$

Note que

$$\text{média} = (1 \times \tfrac{1}{6}) + (2 \times \tfrac{1}{6}) + (\times \tfrac{1}{6}) + (5 \times \tfrac{1}{6}) + (6 \times \tfrac{1}{6})$$

$$\text{média} = \sum_{k=1}^{6} k.P(K) .$$

A somatória dos produtos de cada resultado (numérico) do experimento pela sua probabilidade de ocorrência fornece um valor médio da variável aleatória. Esse valor é chamado **valor esperado ou esperança matemática** ou ainda **média** da variável aleatória.

Seja X uma variável aleatória que assume os valores $x_1, \ldots, x_n$ com probabilidades $p_i = P(X = x_i),\ i = 1, \ldots, n$. O **valor esperado** da variável aleatória X, representado por $E(X)$, é dado por:

$$E(X) = x_1.p_1 + \ldots + x_n.p_n\ .$$

!

$E(X)$ é a notação mais usual da Esperança de X. Alguns autores também usam a letra grega μ (lê-se "mi") para indicar o valor esperado.

Exemplo 3

Consideremos as famílias constituídas de três filhos. Representando por h a criança de sexo masculino e por m a de sexo feminino, o espaço amostral associado a essa observação pode ser representado por:

$$\Omega = \{mmm, mmh, mhm, hmm, mhh, hmh, hhm, hhh\}$$

O número de meninas é uma variável aleatória X que assume os valores 0, 1, 2 e 3. A tabela abaixo mostra a distribuição de probabilidade dessa variável:

X	evento associado	probabilidade
0	{*hhh*}	1/8
1	{*mhh, hmh, hhm*}	3/8
2	{*mmh, mhm, hmm*}	3/8
3	{*mmm*}	1/8

O número esperado de meninas é a soma dos produtos de cada valor de X pela sua probabilidade de ocorrência. Neste caso, temos que o valor esperado para essa variável é

$$0.(1/8) + 1.(3/8) + 2.(3/8) + 3.(1/8) = 12/8 = 3/2 = 1,5$$

Observe que o valor esperado para o número de meninas é impossível de ocorrer na realidade: nenhuma família de três filhos tem 1,5 menina!

Isso muitas vezes ocorre com o valor esperado de uma variável aleatória. Como dissemos anteriormente, ele indica uma média dos valores observados, se o experimento for realizado um grande número de vezes.

Os próximos exemplos ilustram aplicações do valor esperado na análise de alguns jogos.

Exemplo 4

Um jogador paga 1 real para jogar um dado. Se a face observada é 6, ele recebe 10 reais (lucra 9, visto que já pagou 1). Se sai qualquer outro número, ele nada recebe. Podemos definir a variável aleatória (X) que fornece o ganho do jogador em cada partida:

face observada	ganho (X)
1, 2, 3, 4, 5	-1
6	9

Supondo que ele vá jogar um grande número de vezes, vamos calcular o valor esperado de seu ganho, $E(X)$. Para isso, vamos completar a tabela anterior, acrescentando as colunas com as probabilidades de cada evento numérico e com o produto de cada valor assumido pela variável e sua probabilidade:

face observada	ganho	probabilidade	produto
	(X)	(P)	(XP)
1, 2, 3, 4, 5	-1	5/6	-5/6
6	9	1/6	9/6

Então o valor esperado é $E(X) = -5/6 + 9/6 = 4/6 \approx 0,67$ reais.

Vamos interpretar esse resultado: o jogador não vai receber 67 centavos em nenhuma jogada (pois, como vimos, ele ganha 9 ou perde 1) mas, se ele jogar muitas vezes, é de se esperar que ganhe, em média, 67 centavos de real por partida. Por exemplo, se ele jogar 100 vezes, ganhará algumas vezes, perderá outras, mas deverá ganhar, ao final, cerca de 100 x 0,67 = 67 reais.

!

Será que entre os jogos que envolvem sorte (jóquei, loterias, bingo etc.) há algum que seja desequilibrado a favor do jogador? Se você conhecer as regras de pontuação e pagamento de algum deles, pode determinar o valor esperado de ganho.

O exemplo 4 trata de um jogo em que o valor esperado do ganho é positivo. Jogos desse tipo são chamados *desequilibrados a favor do jogador*. Quando o valor esperado de ganho é nulo, o jogo é dito *equilibrado* e, quando é negativo, dizemos que o jogo é *desequilibrado* contra o jogador.

Exemplo 5

Numa loteria, o jogador paga 1 real para marcar cinco algarismos quaisquer numa cartela. O jogo paga 1.000 reais para o jogador que acerta os cinco algarismos. Vamos encontrar o valor esperado de ganho para o jogador nesse jogo.

Os algarismos podem ser repetidos e a escolha de cada um independe da escolha de outro qualquer. Logo, a probabilidade de escolher os cinco corretos é $\frac{1}{10}.\frac{1}{10}.\frac{1}{10}.\frac{1}{10}.\frac{1}{10} = \frac{1}{100.000}$ e a probabilidade de escolher pelo menos um algarismo errado é $1 - \frac{1}{100.000} = \frac{99.999}{100.000}$. Se ele acerta, lucra 1000-1 = 999 reais e se erra, perde 1 real. Então, o valor esperado de ganho é:

$$(999)\left(\frac{1}{100.000}\right) + (-1)\left(\frac{99.999}{100.000}\right) = -0,99 \text{ reais}.$$

Logo, a expectativa é de que o jogador perca cerca de 99 centavos de real por jogada. O jogo é desequilibrado contra o jogador.

RESUMO

Vimos que, dado um experimento aleatório de espaço amostral Ω,

- uma **variável aleatória** é uma função numérica definida em Ω.
- a **distribuição de probabilidade** de uma variável aleatória é uma função que associa, a cada valor assumido pela variável, um número real do intervalo [0,1].
- o **valor esperado** de uma variável aleatória fornece um valor médio dessa variável.
- se X é uma variável aleatória que assume os valores $x_1, \ldots, x_n$ com probabilidades $p_i = P(X = x_i),\ i = 1, \ldots, n$, o **valor esperado** da variável aleatória X é dado por $E(X) = x_1.p_1 + \ldots + x_n.p_n$.

EXERCÍCIOS

1. Uma urna contém 3 bolas brancas e 2 bolas azuis. Duas bolas são retiradas dessa urna, uma após a outra, sem reposição, e suas cores são anotadas. Seja a variável aleatória que associa a cada resultado desse experimento o número de bolas brancas observadas. Determine a distribuição de probabilidade dessa variável.

2. Para lançar um dado, um jogador paga 1 real. O jogo paga:

(a) 3 reais para o resultado 6.

(b) 2 reais para o resultado 5.

(c) 1 real para o resultado 4.

(d) O jogo não paga qualquer dos resultados 1, 2, 3.

Determine o valor esperado de ganho nesse jogo.

3. Uma loja de departamentos vende aparelhos de ar-condicionado. A tabela a seguir lista dados compilados sobre as vendas em um dia:

unidades de aparelhos	0	1	2	3	4
probabilidade de venda	0,10	0,35	0,30	0,20	0,05

Determine o valor esperado de vendas diárias.

AUTO-AVALIAÇÃO

É importante que você compreenda claramente cada uma das funções definidas nesta aula: a variável aleatória e a distribuição de probabilidade dessa variável. A partir desses conceitos foi possível definir o valor esperado da variável. Esse valor fornece uma média dos valores que a variável pode assumir. Se você sentir dificuldade para resolver os exercícios propostos, releia as definições e os exemplos resolvidos, com atenção. Se as dúvidas persistirem, solicite a ajuda do tutor da disciplina.

Distribuição binomial

AULA 15

objetivo

Nesta aula você deverá ser capaz de:

- Estudar a distribuição de probabilidades de experimentos com apenas dois tipos de resultados, realizados uma certa quantidade de vezes.

INTRODUÇÃO

Vamos considerar experimentos aleatórios que apresentam dois resultados possíveis aos quais denominaremos sucesso e fracasso.
Por exemplo:

1. Experimento: lançar uma moeda e observar se dá cara ou não
sucesso: cara
fracasso: coroa
2. Experimento: lançar um dado e observar se dá 5 ou 6 pontos, ou não
sucesso: sair 5 ou 6
fracasso: sair 1 ou 2 ou 3 ou 4
3. Experimento: retirar uma bola de uma urna que contém 10 bolas, sendo 7 bolas brancas e 3 bolas não-brancas, e observar se é branca ou não.
sucesso: branca
fracasso: não-branca

Representemos por p a probabilidade de ocorrer sucesso e q=1-p a probabilidade de fracasso. Nos exemplos anteriores, supondo a moeda e o dado equilibrados, temos:

1. $p = q = \frac{1}{2}$
2. $p = \frac{2}{6}$ e $q = \frac{4}{6}$
3. $p = \frac{7}{10}$ e $q = \frac{3}{10}$

Suponhamos que o experimento considerado seja repetido n vezes, e que o resultado de cada tentativa seja independente dos resultados das demais tentativas.

Vamos definir a variável aleatória

X= número de sucessos nas n tentativas

A variável aleatória X tem uma distribuição de probabilidade:

X	P
0	p_0
1	p_1
2	p_2
.	.
.	.
.	.
n	p_n

O problema que queremos resolver é: como calcular p_k, onde $p_k = P(X = k) = P$ (obter exatamente *k* sucessos nas *n* tentativas)?

Em outras palavras, como calcular a probabilidade de ocorrerem exatamente *k* sucessos nas *n* realizações do experimento?

• Uma possibilidade de ocorrerem *k* sucessos nas *n* tentativas é:

$$\underbrace{SSS\ldots S}_{k \text{ vezes}}\ \underbrace{FFF\ldots F}_{(n-k) \text{ vezes}}$$

onde *S* indica Sucesso; *F* indica Fracasso

• Devido à independência dos resultados de cada tentativa, a probabilidade de ocorrer o caso descrito acima é

$$\underbrace{(ppp\ldots p)}_{k \text{ vezes}}.\ \underbrace{(qqq\ldots q)}_{(n-k) \text{ vezes}}$$

ou seja, é $p^k.q^{n-k}$.

• Os *k* sucessos e os *n*–*k* fracassos podem ocorrer em qualquer ordem e ocupar quaisquer posições na seqüência. O total de possibilidades é o total de permutações de *n* elementos, com *k* e *n*–*k* elementos repetidos. Vimos no Módulo 2 que esse total é dado por

$$\frac{n!}{k!(n-k)!} = \left(\begin{array}{c} n \\ k \end{array}\right) = C_{n,k}$$

• Conclusão: Como são $\left(\begin{array}{c} n \\ k \end{array}\right)$ tentativas, temos:

$$p(k) = P(X = k) = \left(\begin{array}{c} n \\ k \end{array}\right) p^k q^{n-k}$$

A expressão $\left(\begin{array}{c} n \\ k \end{array}\right) p^k q^{n-k}$ fornece o termo geral do desenvolvimento do binômio $(p+q)^n$. Por isso, essa distribuição de probabilidade é denominada **distribuição binomial**.

Vamos retomar os experimentos do exemplo anterior:

Exemplo 6

Qual a probabilidade de, em cinco lançamentos de uma moeda equilibrada, serem observadas exatamente três caras?

Solução:

Sendo *X* o número de caras nos 5 lançamentos.

• Em cada lançamento:

sucesso: cara $\rightarrow p$ = 1/2

fracasso: coroa $\rightarrow q$ = 1/2

• Número de lançamentos (tentativas): *n*=5

• Número desejado de sucessos: *k*=3

• Probabilidade pedida:

$$P(X=k) = \left(\begin{array}{c} n \\ k \end{array} \right) p^k q^{n-k}$$

$$P(X=k) = \left(\begin{array}{c} 5 \\ 3 \end{array} \right) (\tfrac{1}{2})^3 (\tfrac{1}{2})^{5-3}$$

$$P(X=3) = \tfrac{5!}{3!2!} \cdot \tfrac{1}{2^3} \cdot \tfrac{1}{2^2} = \tfrac{10}{32}$$

Exemplo 7

Qual a probabilidade de, em dez lançamentos de um dado honesto, serem obtidos 5 ou 6 pontos em exatamente quatro das dez tentativas?

Solução:

Definimos a variável aleatória *X* = número de vezes em que são observadas as faces 5 ou 6, nos dez lançamentos. Queremos calcular *P*(*X*=4). Temos:

$$n = 10; \quad k = 4; \quad p = \tfrac{2}{6}; \quad q = \tfrac{4}{6}$$

Então

$$P(X=4) = \left(\begin{array}{c} 10 \\ 4 \end{array} \right) \left(\begin{array}{c} 2 \\ 6 \end{array} \right)^5 \left(\begin{array}{c} 4 \\ 6 \end{array} \right)^{10-4} = \frac{10!}{4!6!} \cdot \left(\begin{array}{c} 2 \\ 6 \end{array} \right)^5 \left(\begin{array}{c} 4 \\ 6 \end{array} \right)^6 = \frac{40}{243}$$

Exemplo 8

Uma urna contém dez bolas das quais sete, e apenas sete, são brancas. Cinco bolas são retiradas, uma a uma, com reposição. Qual a probabilidade de serem retiradas exatamente três bolas brancas?

Solução:

Definimos a variável aleatória *X*= número de bolas brancas nas 5 retiradas. Queremos calcular *P*(*X*=3). Temos:

$$n = 5; \quad k = 3; \quad p = \tfrac{7}{10}; \quad q = \tfrac{3}{10}$$

Então

$$P(X=3) = \left(\begin{array}{c} 5 \\ 3 \end{array} \right) \left(\begin{array}{c} 7 \\ 10 \end{array} \right)^3 \left(\begin{array}{c} 3 \\ 10 \end{array} \right)^{5-3} = \frac{5!}{3!2!} \cdot \left(\begin{array}{c} 7 \\ 10 \end{array} \right)^3 \left(\begin{array}{c} 3 \\ 10 \end{array} \right)^2 = \frac{3087}{10.000} = 0,3087$$

Exemplo 9

No lançamento de quatro moedas, dê a distribuição de probabilidade da variável aleatória "número de caras".

Solução:

• Em cada moeda:

sucesso: cara $p = \frac{1}{2}$

fracasso: coroa $q = \frac{1}{2}$

• Variável aleatória: X = número de caras nas 4 moedas

• Distribuição de probabilidade da variável aleatória X:

X	(interpretação)	P
0	(nenhuma cara e 4 coroas)	$C_{4,0}.(\frac{1}{2})^0.(\frac{1}{2})^4 = 1/16$
1	(1 cara e 3 coroas)	$C_{4,1}.(\frac{1}{2})^1.(\frac{1}{2})^3 = 6/16$
2	(2 caras e 2 coroas)	$C_{4,2}.(\frac{1}{2})^2.(\frac{1}{2})^2 = 6/16$
3	(3 caras e 1 coroa)	$C_{4,3}.(\frac{1}{2})^3.(\frac{1}{2})^1 = 4/16$
4	(4 caras e nenhuma coroa)	$C_{4,4}.(\frac{1}{2})^4.(\frac{1}{2})^0 = 1/16$

Exemplo 10

Numa prova de 10 questões objetivas, a probabilidade de que um aluno acerte uma pergunta qualquer, no "chute", é $\frac{1}{5}$. Para ser aprovado, ele tem que acertar pelo menos 6 questões. Qual a probabilidade deste aluno ser aprovado, apenas chutando as respostas?

Solução:

Interpretemos o problema:

• o experimento "responder a uma questão" será repetido 10 vezes;

• em cada tentativa:

- sucesso: acertar no chute; p = 1/5

- fracasso: errar ao chutar; q = 4/5

• variável aleatória X = número de acertos nas 10 tentativas

Queremos calcular a probabilidade de ocorrer X = 6 ou X = 7 ou X = 8 ou X = 9 ou X = 10. Como esses eventos são mutuamente exclusivos, a probabilidade da união desses eventos é a soma das probabilidades de cada um.

Então, a probabilidade pedida é:

$$P(X=6)+P(X=7)+P(X=8)+P(X=9)+P(X=10) = C_{10,6}.(1/5)^6.(4/5)^4 + C_{10,7}.(1/5)^7.(4/5)^3 + C_{10,8}.(1/5)^8.(4/5)^2 + C_{10,9}.(1/5)^9.(4/5)^1 + C_{10,10}.(1/5)^{10}.(4/5)^0 = 0,0064 = 0,64\%$$

Diante de uma probabilidade tão pequena, este exemplo pode ser usado para incentivar um aluno a estudar, não é?

RESUMO

Na Aula 14 vimos a distribuição de probabilidade uniforme, definida num espaço amostral eqüiprovável. Nesta aula, vimos uma outra distribuição de probabilidade: a distribuição binomial. Vamos listar suas características:

• Repete-se um experimento n vezes.

• Só há dois tipos de resultados possíveis em cada tentativa: designamos um deles sucesso e o outro fracasso.

• A probabilidade de resultar um sucesso em uma tentativa é p; logo, a de ocorrer um fracasso é $1-p=q$.

• As realizações são todas independentes.

• A variável aleatória X número de sucessos nas n tentativas tem distribuição binomial.

• A probabilidade de ocorrerem exatamente k sucessos nas n tentativas, $p(X=k)$, é dada por $\left(\begin{array}{c} n \\ k \end{array} \right) p^k q^{n-k}$.

EXERCÍCIOS

1. Uma pesquisa indicou que, numa cidade, 75% dos automóveis têm seguro. Se 6 automóveis sofrerem um acidente, qual a probabilidade de exatamente 2 deles terem seguro?

2. Uma moeda equilibrada é lançada 10 vezes. Qual a probabilidade de serem observadas exatamente 4 coroas?

3. Uma urna contém 4 bolas azuis e 6 bolas vermelhas. São retiradas 5 bolas, uma a uma, com reposição, e sua cor é anotada. Qual a probabilidade de, em todas as retiradas, a bola ser azul?

4. A probabilidade de um homem de 50 anos viver mais 20 anos é 0,6. Considerando um grupo de 8 homens de 50 anos, qual a probabilidade de que pelo menos 4 cheguem aos 70 anos?

5. 60% das pessoas de uma população têm olhos castanhos. Cinco pessoas são escolhidas ao acaso (pode-se supor sorteio com reposição). Qual o número esperado de pessoas com olhos castanhos nas 5 selecionadas?

(Sugestão: Faça X = número de pessoas com olhos castanhos nas 5 escolhidas. Forme a distribuição de probabilidade de X, depois calcule E(X). Comprove que E(X) = 60% de 5.)

AUTO-AVALIAÇÃO

Você deve identificar claramente as condições nas quais podemos definir uma distribuição binomial. Para resolver os exercícios, podemos usar as propriedades válidas para as probabilidades, vistas na Aula 10. Se você sentir dificuldades, solicite ajuda do tutor da disciplina.

FIM DESTA AULA...

Vimos os conceitos e resultados básicos da teoria das probabilidades:

• A identificação de um fenômeno aleatório.

• As definições de experimento, espaço amostral e evento.

• A definição de probabilidade e o estudo de suas propriedades (probabilidade do evento complementar, regra da adição, probabilidade condicional, regra da multiplicação, regra da probabilidade total).

• Os conceitos de eventos mutuamente exclusivos e de eventos independentes.

• O teorema de Bayes, que permite calcular a probabilidade das causas de um determinado efeito.

• A definição da função numérica variável aleatória e o cálculo de seu valor esperado.

• A distribuição binomial.

Para saber mais deste assunto ou entender melhor algum conceito visto, indicamos uma bibliografia básica:

MENDENHAL, W. *Probabilidade e Estatística*. Rio de Janeiro: Ed. Campus, vol.1, 1985.

MEYER, P.L. *Probabilidade – Aplicações à Estatística*. Rio de Janeiro: Ao Livro Técnico, 1974.

MORGADO, A.C.O. e outros. *Análise Combinatória e Probabilidade*. Rio de Janeiro: SBM - Coleção do Professor de Matemática, 1981.

TOLEDO, G.L. e OVALLE, I.I. *Estatística Básica*. São Paulo: Ed. Atlas S.A., 1978.

O que é Estatística? A distribuição de freqüências

AULA 16

objetivos

Ao final desta aula, você deverá ser capaz de:

- Definir estatística, população, amostra, variabilidade de uma medida, variável aleatória.
- Representar graficamente uma distribuiçãode freqüências.
- Calcular os parâmetros de posição e de dispersão de uma distribuição de freqüências.

Pré-requisitos

Nesta aula, bem como em todas as outras, você deverá dispor, além de papel e borracha, de uma pequena calculadora, das mais simples, para poder realizar os exercícios propostos.

INTRODUÇÃO

Você vai saber o que é a *Estatística* e como podemos representar os dados sob forma de gráficos de distribuição. Você aprenderá também a calcular alguns parâmetros básicos de uma distribuição de dados.

O QUE É ESTATÍSTICA?

Você deve ter percebido que o tema da aula é Bioestatística, porém vamos sempre usar a palavra Estatística durante as aulas. Qual é a diferença? Nenhuma! Na realidade, não existe uma Bioestatística, mas simplesmente uma Estatística aplicada à Biologia.

Você, certamente, ouviu falar de *Estatística* e até comentado sobre ela com amigos. Por exemplo, teve a oportunidade de ouvir no noticiário que:

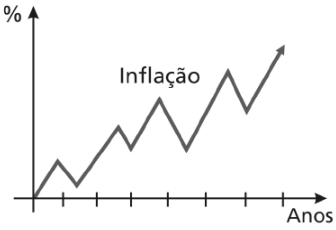
– tal candidato à eleição alcançou 40 pontos percentuais, com margem de erro de 2 pontos para cima e 2 pontos para baixo;

– seu time tem 30% de chance de ganhar o campeonato;

– a inflação subiu 0,5% esse mês.

Na área da Biologia também surgem perguntas, tais como:

– qual a eficiência de um novo remédio para a cura de uma determinada doença, ou de um novo inseticida para o combate a uma determinada praga?



– se parar de jogar esgoto nas águas da baía, haverá um aumento da pesca?

– fumar faz mal à saúde?

Essas perguntas importantes só podem ser respondidas fazendo pesquisas e coletas de dados sobre o assunto. Em seguida, esses dados devem ser analisados e interpretados para tirar as devidas conclusões. É aí que entra a Estatística.

Então, o que você entende por Estatística? Nesta aula, vamos definir alguns conceitos para compreender melhor o que é Estatística e qual a sua utilidade.

A Estatística é tão antiga como o primeiro homem, pois a necessidade de enumerar as coisas surgiu com ele. Embora a palavra Estatística ainda não existisse, há indícios de que 3.000 anos a.C. já se faziam censos na Babilônia, China e Egito. Em 1085, Guilherme, o Conquistador, ordenou um levantamento estatístico da Inglaterra visando

obter informações sobre terras, proprietários, empregados, animais etc., como base para o cálculo de impostos. Nasceu assim, por volta da metade do século XVIII, a palavra "Estatística" do latim *Status* (Estado) sobre a qual acumularam-se descrições e dados relativos ao Estado. A Estatística, nas mãos dos estadistas, constitui-se então verdadeira ferramenta administrativa.

Nascida como simples compilação de números, a Estatística tem evoluído até nossos dias como um poderoso instrumento destinado a pesquisar as relações de causa e efeito dos fenômenos, possibilitando suas previsões dentro de uma razoável margem de erro.

A Estatística encontra aplicações em quase todos os campos da atividade humana, desde a política, a sociologia, até a indústria, o comércio, a meteorologia, a geografia e a biologia.

A Estatística vem servindo, de maneira eficiente, à Biologia. Imensos são os serviços prestados por ela. Toda vez que são introduzidos novos métodos terapêuticos ou de diagnósticos, é preciso estabelecer se o novo método é, realmente, superior ao antigo. É comum ouvir dizer também que determinada ação ou atividade do homem provocou uma alteração "significativa" da qualidade das águas de um lago, por exemplo (veremos mais tarde o que representa a palavra "significativa").

Desse modo, podemos definir também a Estatística como uma ciência do processamento dos dados numéricos fornecidos pela observação ou pela experiência. Ela permite que você tire conclusões e verifique a validade dessas conclusões. Em sentido mais restrito, o termo estatístico é também usado para designar os próprios dados ou números deles derivados como, por exemplo, médias. Podemos falar assim de estatística de pesca, de vendas, de acidentes etc.

A ESTATÍSTICA E A PESQUISA CIENTÍFICA

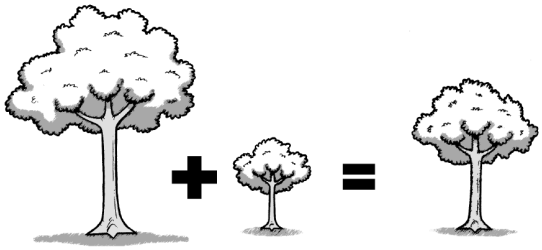
Os métodos estatísticos fornecem ferramentas úteis à investigação científica. Ela pode prestar auxílios fundamentais, nos casos de querer, por exemplo:

a) caracterizar quantitativamente um dado fenômeno. Ex:. encontrar a média de altura das árvores de uma floresta;

b) fazer previsões. Ex.: prever os danos causados na biota de uma baía pelo derramamento de óleo;

c) fazer comparações. Ex.: testar qual de duas substâncias é mais eficiente no tratamento de uma dada doença;

d) construir um modelo que descreva como um fenômeno é afetado por vários fatores. Ex.: estimar a taxa de crescimento de um peixe cultivado em aquário, de acordo com a temperatura da água e o tipo e quantidade de alimento.



POPULAÇÃO E AMOSTRAS

Como você pode constatar, a Estatística tem mil e uma aplicações nas diversas áreas da ciência, inclusive nas áreas da Biologia e da Ecologia. Mas o seu uso exige dados que são obtidos fazendo diversas medições e coletas de amostras dentro de uma determinada população. Esses dois termos, *amostra* e *população,* têm um significado bem preciso em estatística. Vamos ver qual.

O termo *população* é empregado para designar um conjunto de objetos que possuem propriedades comuns, passíveis de caracterização, que os diferenciem de qualquer outro não pertencendo ao dado conjunto. Do ponto de vista estatístico, a população constitui o universo sobre o qual pretendemos direcionar as pesquisas. Assim, o conjunto de todas as estaturas dos alunos de uma universidade constitui uma população de estaturas, o conjunto de todos os pesos de uma população de peixes constitui uma população de pesos.

Como geralmente não é possível trabalhar com todos os elementos da população (todos os peixes do mar, por exemplo), restringimos nosso estudo a uma porção deste universo, denominada *amostra*. A amostra é, portanto, a fração populacional a partir da qual vamos tirar conclusões sobre a população.

ESTATÍSTICA DESCRITIVA E INDUTIVA

A parte da estatística que procura somente descrever e analisar um certo grupo de elementos é chamada *estatística descritiva*. Por outro lado, se uma amostra é representativa de uma população, ou seja, se ela contém, em proporção, tudo o que a população possui qualitativa e quantitativamente, então podemos tirar conclusões importantes sobre a população a partir dessa amostra. Nesse caso, dizemos que vamos inferir os resultados obtidos com a amostra para a população, e realizamos uma estatística chamada *estatística indutiva* ou *inferencial*. Essa inferência (transferência de conclusões da amostra para a população) nunca dará um resultado exato, mas somente uma estimativa da população, estimativa que será sempre acompanhada de um certo erro, cujo tamanho pode ser avaliado pela *teoria das probabilidades*. Não se assuste nem se apresse, esse conceito de estimativa e suas aplicações serão vistos mais tarde.

!

Você deve ter visto a *teoria das probabilidades* em aula de Matemática. Mas não se preocupe, faremos uma pequena revisão sobre o assunto, na próxima aula.

VARIABILIDADE DE UMA MEDIDA

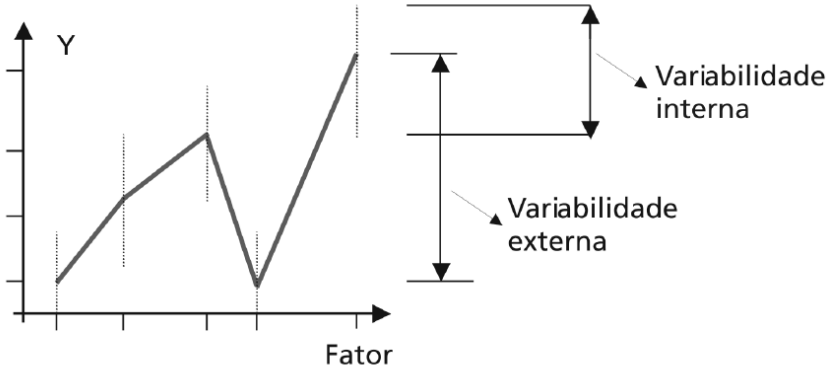
Uma *variável* é um caractere suscetível de variar. Vamos começar pelo exemplo para que você possa entender. Os organismos de uma comunidade podem ser estudados sob diversos ângulos. Por exemplo, podem ser classificados quanto ao sexo, ao tamanho dos indivíduos, ou seu peso etc. Sexo, tamanho, peso, são variáveis, isto é, são propriedades às quais podemos associar conceitos ou números e assim expressar informações sob forma de medidas.

Cada variável tem um domínio, que é o conjunto de valores entre os valores mínimo e máximo que ela pode alcançar. Uma variável pode conter diversos tipos de dados:

a) dados qualitativos ou categóricos. Eles apenas discriminam objetos em categorias ou atributos (ex.: sexo, cor dos olhos);

b) dados quantitativos que podem ser contínuos, assumindo qualquer valor entre duas observações (ex.: peso de um indivíduo), ou discretos, assumindo somente valores inteiros (ex.: número de crianças numa família).

Agora que conhecemos a variável, vamos descobrir quais as causas da variabilidade de uma medida. Existem dois tipos (fontes) de variabilidade, como mostrado na figura a seguir:



a) a *variabilidade externa,* ligada aos fatores e à variação das condições de experiência ou do meio natural, como por exemplo o efeito da temperatura sobre o crescimento de um organismo, da intensidade luminosa sobre a fotossíntese etc. O estudo do efeito desses fatores constitui geralmente o objetivo da pesquisa.

b) a *variabilidade interna* (intrínseca, própria) aparece sempre, mesmo quando todas as condições são mantidas constantes. Essa variabilidade interna pode mascarar a variabilidade externa que queremos estudar. O pesquisador deve tentar diminuir ao máximo essa fonte de variabilidade, aperfeiçoando os equipamentos ou o método de amostragem, por exemplo.

Em Biologia, ao inverso de outras áreas como a Física, quaisquer que sejam as precauções haverá sempre uma certa variabilidade interna, pois não existem dois seres vivos rigorosamente idênticos. Logo, é inútil tentar suprimir a variabilidade interna de um fenômeno biológico. É uma parte intrínseca do fenômeno. Não é um "erro"! Numa série de medidas, essas duas variabilidades são misturadas. Os métodos estatísticos (a "Estatística") têm como objetivo separar essas duas fontes de variabilidade e permitir, assim, o estudo do fenômeno.

AMOSTRAGEM ALEATÓRIA

Uma amostragem é considerada **ALEATÓRIA** quando todos os elementos de uma população têm igual chance de serem selecionados. É isso que acontece quando se faz o sorteio da loteria: cada número contido na bola tem a mesma chance de ser sorteada. A aleatorização é exigida em estatística inferencial, para poder extrapolar os resultados de uma amostra para a população.

ALEATÓRIO

Um evento que depende do acaso. Sua realização é regida pela teoria das probabilidades.

VARIÁVEL ALEATÓRIA

É chamada de "aleatória" a variável cujos valores variam mesmo após ter fixado todos os fatores possíveis de influenciar. Essa variável não pode ser prevista exatamente. Somente podemos ter uma idéia das suas características (média, dispersão).

DISTRIBUIÇÃO DE FREQÜÊNCIAS

Você concorda que as tabelas com grande número de dados são cansativas e não dão ao leitor uma visão rápida e global do conjunto de dados? Para facilitar essa visão podemos organizar os dados em uma tabela de *distribuição de freqüências*. Ela pode ser feita com dados qualitativos, quantitativos discretos ou quantitativos contínuos. Vamos ver um exemplo para cada tipo de dados.

Exemplo 1 – Dados qualitativos

Numa amostragem de 80 indivíduos, realizada numa população de camarões em cativeiro, obtivemos 35 machos e 45 fêmeas. Com esses dados foi possível elaborar a seguinte tabela:

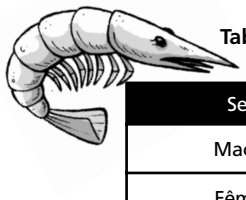


Tabela 16.1

Sexo	Freqüência	Freqüência relativa
Machos	35	0,44
Fêmeas	45	0,56
TOTAL	80	1,00

Na coluna "Freqüência" encontra-se o número de indivíduos de cada categoria (macho ou fêmea). É uma freqüência absoluta, que chamaremos simplesmente de "freqüência". O total da coluna corresponde ao tamanho da amostra (80 indivíduos). Na coluna "Freqüência relativa" o número de indivíduos é dividido pelo total de indivíduos observados (80). Neste caso o total da coluna é 1,00. Em outras palavras, podemos dizer que nessa amostra há 44% de machos (multiplicando 0,44 por 100). O cálculo da freqüência relativa é necessário quando se pretende comparar a razão sexual de duas ou mais amostras com número diferente de indivíduos.

Exemplo 2 – Dados quantitativos discretos

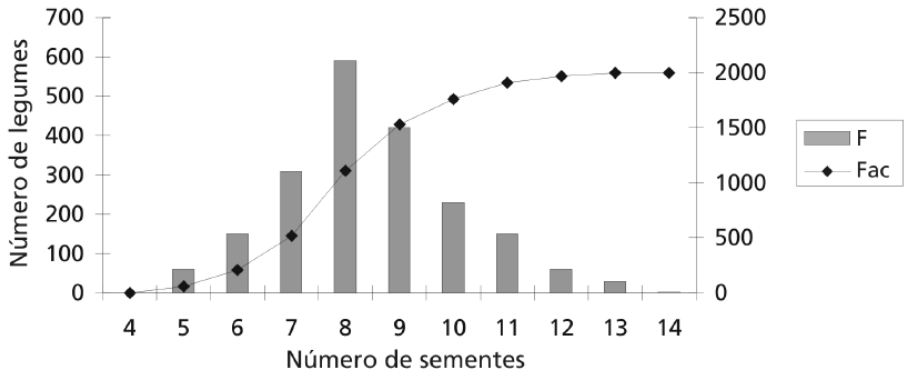
Numa amostra de 2000 legumes contamos o número de sementes (X) em cada legume. Na segunda coluna da tabela colocamos as freqüências (F), isto é, o número de legumes contendo aproximadamente 4, 5, 6 até 14 sementes. Na terceira coluna calculamos a freqüência relativa (Fr), dividindo cada freqüência pelo total de legumes (2000). Na quarta coluna são as freqüências acumuladas (Fac). Elas são obtidas somando as freqüências sucessivas de cada valor de semente.

A tabela pode ser interpretada da seguinte maneira: na amostra de 2000 legumes, 310, por exemplo, contêm 7 sementes (ou seja, 15,5% dos legumes), e 520 legumes contêm ATÉ 7 sementes (quer dizer, 4, 5, 6 ou 7 sementes).

Tabela 16.2

X	F	Fr	Fac
4	0	0,000	0
5	60	0,030	60
6	150	0,075	210
7	310	0,155	520
8	590	0,295	1110
9	420	0,210	1530
10	230	0,115	1760
11	150	0,075	1910
12	60	0,030	1970
13	28	0,014	1998
14	2	0,001	2000
N	2000	1,000	

Os dados dessa tabela de freqüências podem ser representados graficamente sob forma de um diagrama de barra ou histograma, como representado na figura abaixo.



Em vez de utilizar barras, podemos reunir o topo de cada barra por uma linha (como feito para as freqüências acumuladas), e obter um polígono de freqüências.

Exemplo 3 – Dados quantitativos contínuos

Para montar a tabela de distribuição de freqüências de uma variável contínua é preciso dividir a variável em classes e calcular a freqüência dos dados em cada classe. Vejamos como isso acontece.

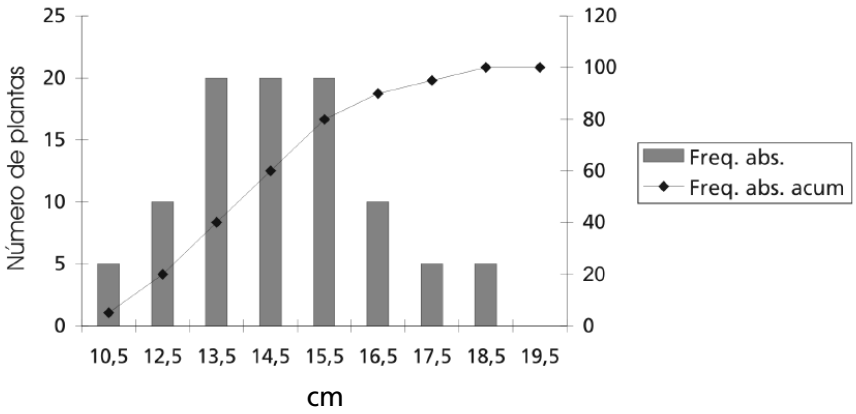
Medimos o comprimento de 100 plantas (cm) e elaboramos a tabela a seguir:

Tabela 16.3

Classes de tamanho	Ponto médio X	Freqoluta	Freq. abs. acumulada	Freq. relativa	Freq. relat. acumulada
10,0-10,9	**10.5**	5	5	0,05	0,05
11.0-11.9	**11.5**	5	10	0,05	0,10
12.0-12.9	**12.5**	10	20	0,10	0,20
13.0-13.9	**13.5**	20	40	0,20	0,40
14.0-14.9	**14.5**	20	60	0,20	0,60
15.0-15.9	**15.5**	20	80	0,20	0,80
16.0-16.9	**16.5**	10	90	0,10	0,90
17.0-17.9	**17.5**	5	95	0,05	0,95
18.0-18.9	**18.5**	5	100	0,05	1,00
19.0-19.9	**19.5**	0	100	0,00	1,00
Total		100		1,00	

Observamos que na primeira coluna são colocados os intervalos de classe. Eles devem ser escolhidos de maneira adequada, de acordo com os valores mínimo e máximo dos dados, a precisão das medidas e o número total de dados. Não deve haver ambigüidade nos limites inferiores e superiores de cada classe. Na segunda coluna é calculado o ponto médio das classes, isto é, a média entre o limite inferior e superior de cada classe.

Você pode então traçar a figura a seguir:



No caso hipotético onde o número de classes tende para o infinito e o intervalo de classe tende para o zero, o histograma de freqüências se transforma numa curva contínua, representada por Y = f(x). A área delimitada por essa curva corresponde à probabilidade de ocorrência de determinados valores dentro de um intervalo. Assim, a probabilidade de um valor exato x é nula. A probabilidade de que x esteja dentro de um intervalo muito pequeno é mínima. Ela é chamada *probabilidade elementar* no ponto x.

OS PARÂMETROS DE UMA DISTRIBUIÇÃO

Vimos que um conjunto de dados (chamado também de série estatística) pode ser representado por tabelas e gráficos. Porém, para obter uma descrição sumária mais objetiva que gráficos e tabelas, devemos encontrar descritores mensuráveis desses dados. Esses descritores são chamados de *Estatísticas* ou *Parâmetros* da distribuição dos dados. São medidas que servem para caracterizar a distribuição dos dados. Há dois grandes tipos de parâmetros: os de *posição e* os de *dispersão*.

PARÂMETROS DE POSIÇÃO

Os parâmetros de posição são também chamados de *medidas de tendência central,* pois são valores próximos ao centro de um conjunto de dados. Veremos três medidas de tendência central (ou parâmetros de posição): *a média aritmética*, a *mediana* e a *moda*.

Média aritmética

A média aritmética, ou *média*, de um conjunto de N números $X_1, X_2, \ldots, X_N$, é representada por m ou $\overline{X}$ (leia-se X-barra) e é definida por

$$m = \frac{\sum_{i=1}^{N} X_i}{N}$$

No símbolo $\sum_{i=1}^{N} X_i$ lê-se: somatório de todos os Xi (xis-i), quando i varia de 1 a N. Por exemplo, a média aritmética de 2, 5, 8, 13 é: m = (2+5+8+13) / 4 = 7. Cada valor corresponde a um Xi. Assim: (X_1 = 2), (X_2 = 5), (X_3 = 8) e (X_4 = 13)

Mediana

A mediana de um conjunto de dados organizados em ordem de grandeza, é o valor que divide o conjunto em 2 partes iguais, isto é, que contem o mesmo número de dados. Em outras palavras, corresponde ao valor central quando a série de dados é par, ou à média dos valores centrais quando ela é ímpar.

Exemplo 1

A mediana de 3, 4, 5, 6, 8, 8, 9 é 6

Exemplo 2

A mediana de 5, 5, 7, 9, 11, 12, 15, 18 é (9+11)/2 = 10

Moda

A moda de um conjunto de dados é o valor que ocorre com a maior freqüência, ou seja, é o valor mais comum. A moda pode não

existir e, quando existe, pode não ser única.

Exemplo 1

O conjunto 2, 2, 5, 7, **9, 9, 9,** 10, 10, 1, 12, 18 tem uma moda **9**. É denominado unimodal.

Exemplo 2

O conjunto 3, 5, 8, 10, 12, 15, 16 não tem moda.

Exemplo 3

O conjunto 2, 3, **4, 4, 4**, 5, 5, **7, 7, 7,** 9 tem duas modas **4** e **7**. Ele é chamado *bimodal*.

No caso de dados agrupados em classes, a moda é a classe de maior freqüência. Da mesma maneira, pode haver nenhuma, uma ou mais de uma moda.

PARÂMETROS DE DISPERSÃO

Se a natureza fosse estável, se as mesmas causas produzissem sempre os mesmos efeitos, é bem possível que o homem nunca tivesse desenvolvido a noção de variação, mas a realidade é outra: o mundo está em permanente oscilação. Ao conjunto das medidas, isto é, das Estatísticas, que medem as oscilações de uma variável deu-se o nome de *medidas de variabilidade*. Existem várias medidas de variabilidade, ou parâmetros de dispersão de uma distribuição de dados. Veremos agora a variância e o desvio padrão. Fique atento!

Variância

Imaginemos dois conjuntos de 8 amostras de peixes coletados em diversas pontos de dois lagos A e B. Medimos o número de indivíduos em cada amostra:

Lago A : 8, 9, 10, 8, 6,11, 7, 13

Lago B : 7, 3, 10, 6, 5, 13, 18, 10

O total de peixes é o mesmo nos dois conjuntos de amostras (= 72). Podemos fazer a seguinte pergunta: Qual dos dois lagos tem uma distribuição de peixes mais regular (homogênea)? Ou seja, em que lago a variação entre as amostras é menor? Observe que o cálculo da média de peixes entre amostras

não ajuda, pois é a mesma (m= 72/8 = 9 peixes). Porém verificamos que:

– no lago A, os dados variam entre 6 e 13 peixes, isto é, a amplitude total é 13 - 6 = 7 peixes;

– no lago B, a amplitude é 18 – 3 = 15 peixes.

Queremos quantificar essa variação e, para isso, precisamos de um ponto de referência que pode ser a média aritmética. Para cada conjunto de amostras vamos fazer o seguinte:

a) subtrair de cada valor x a média aritmética m do conjunto (x-m);

b) elevar cada diferença ao quadrado $(x-m)^2$;

c) somar os quadrados: $\Sigma(x-m)^2$;

d) dividir a soma dos quadrados pelo número n de amostras: $\Sigma(x-m)^2 /n$.

O resultado assim obtido é chamado de variância (v), cuja fórmula é:

$$v = \frac{\Sigma(x-m)^2}{n}, \text{ ou seja, } v = \frac{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}{n}$$

A segunda expressão de v é equivalente à primeira, mas de cálculo mais fácil, pois não necessita da média.

A fórmula da variância é constituída de dois termos:

– o numerador, $\Sigma(x-m)^2$, chamado de *dispersão* ou de *Soma dos Quadrados dos Desvios* (SQD).

– o denominador, n, é o número de dados. Para pequeno número de dados (<30), utilizar $n - 1$ (chamado *grau de liberdade*, geralmente indicado por *gl*, ou pela letra grega γ (lê-se *n*u).

Você deve estar imaginando o que é *grau de liberdade*? Em Estatística, uma exigência básica é a independência entre os dados de uma série, isto é, o conhecimento de um valor da série não permite descobrir o valor exato de um outro. Quando conhecemos a soma dos dados, é possível recalcular facilmente qualquer um dado que falta, ou seja, a introdução da soma resulta na perda de 1 grau de liberdade.

Quando n é grande, utilizar n ou $n - 1$ não acarreta grande diferença nos resultados, porém não é o caso com amostras pequenas. Por exemplo, se você precisar calcular a variância de uma série de 100 dados, tanto faz dividir por 100 ou por 99. O erro cometido será pequeno.

Mas se a série for de somente10 dados, então, nesse caso, dividir por n=10 ou por n-1=9 faz uma diferença importante e você deve dividir pelo grau de liberdade n-1.

Você pode então se perguntar: a partir de que n devo utilizar $\boldsymbol{n}$**-1**? Teoricamente, a partir de n = 30, mas, você quer um conselho? Utilize sempre $\boldsymbol{n}$**-1**. Assim você não correrá o risco de errar no seu cálculo de variância.

!

Por que n = 30? É que 30 é o número mínimo de dados para que a série estatística seja distribuída de acordo com a *Lei Normal*. Não fique espantado! Veremos mais adiante o que é uma *Lei Normal*.

Vamos então aprender a calcular a variância do número de peixes encontrados nos dois lagos A e B. Vamos utilizar a segunda fórmula da variância. Aquela que não necessita da média. Para isso, organizamos a tabela seguinte:

Tabela 16.4

	LAGO A		LAGO B	
Amostras	X	X^2	X	X^2
(1)	8	64	7	49
(2)	9	81	3	9
(3)	10	100	10	100
(4)	8	64	6	36
(5)	6	36	5	25
(6)	11	121	13	169
(7)	7	49	18	324
(8)	13	169	10	100
TOTAL	72	684	72	812

Ao calcular, concluímos que a variância no Lago A é igual a:

$$v_A = (684 - 72^2 / 8) / (8\text{-}1) = 5{,}14$$

e a variância no Lago B é igual a:

$$v_B = (812 - 72^2 / 8) / (8\text{-}1) = 23{,}43.$$

Desvio padrão

E o desvio padrão? O que representa e como calculá-lo?

No exemplo anterior, cada valor encontrado da variância corresponde à unidade "peixes ao quadrado", pois os dados foram elevados ao quadrado. Para restabelecer a unidade "peixes", extraímos a raiz quadrada positiva da variância. O resultado é chamado *desvio padrão, s*.

Logo,

o desvio padrão do lago A é $s_A = \sqrt{5{,}14} = 2{,}27$ peixes;

e o desvio padrão do lago B é $s_B = \sqrt{23{,}43} = 4{,}84$ peixes;

Você deve ter percebido que a distribuição dos peixes no lago A parece ser mais homogênea do que no lago B. Vamos confirmar isso com o cálculo do parâmetro seguinte que é o *coeficiente de variação*.

O COEFICIENTE DE VARIAÇÃO

Para poder comparar a variação (dispersão ou desvio padrão) de dois conjuntos de dados com unidades diferentes ou médias diferentes, é preciso calcular o *coeficiente de variação, Cv,* que é o desvio padrão ponderado, isto é, dividido pela média. Ele é expresso geralmente em percentagem. Observe:

$$Cv = \frac{s}{m}.100$$

Confira o exemplo!

O coeficiente de variação dos peixes do lago A é 2,27/9 x 100 = 25,2 % e o do lago B é: 4,84/9 x 100 = 53,8 %.

!

IMPORTANTE!

Por convenção, a média, a variância e o desvio padrão de uma série de dados amostras são representados, respectivamente, pelas letras *m* ou *x*-barra,*v* ou s^2 e *s*. Quando se trata da população, esses parâmetros devem ser representados pelas letras gregas μ^2, σ^2.

Tabela 16.5

Parâmetro	População	Amostra
Média	μ	*m* ou $\overline{x}$
Variância	σ^2	v ou s^2
Desvio padrão	σ	s

RESUMO

Você começou a perceber o que é Estatística e suas aplicações no cotidiano e também na Biologia. Aprendeu a organizar os dados em distribuição de freqüência e a calcular os parâmetros dessa distribuição. Observou que eles se dividem em parâmetros de posição (média, mediana, moda) e de dispersão (variância, desvio padrão). Aprendeu também que a amostra permite estimar os parâmetros de uma população.

EXERCÍCIOS

1. Na tabela abaixo figuram as medidas de altura (em mm) de 50 plantas:

70	68	64	70	67	68	79	76	72	82
71	78	64	78	74	81	73	79	79	70
77	91	79	63	64	72	71	85	66	67
76	75	70	82	65	84	69	76	74	72
78	82	77	75	76	78	79	64	82	64

a) Considere os dados e construa uma tabela em que aparecem as classes de altura com 5 mm de intervalo, o ponto médio das classes e as freqüências absolutas, relativas e acumuladas.

b) Trace o histograma de freqüências absolutas e de freqüências acumuladas.

2. Dada a série estatística: 1, 6, 6, 3, 7, 4, 10, calcule a média, o desvio padrão e o coeficiente de variação.

3. Em 10 amostragens numa praia foram coletadas 3 espécies de conchas com as seguintes abundâncias:

Amostras	Espécie 1	Espécie 2	Espécie 3
1	0	3	1
2	0	4	0
3	2	3	14
4	0	1	1
5	1	5	4
6	0	3	0
7	0	6	8
8	1	1	0
9	1	3	4
10	3	9	6

Calcule a média, a variância, o desvio padrão e o coeficiente de variação de cada espécie.

4. Num estudo de crescimento de uma população de moluscos (mexilhões), um pesquisador mediu o comprimento da concha em milímetros, os dias 5/1/92 e 5/6/92. Os resultados foram:

Dia 5/1/92

15-13-12-09-07-02-08-15-13-12-09-06-02-08-14-13-12-08-06-02-08-14-13-11-08-05-02-08-13-13-11-07-02-02-13-13-11-06-02-02-13-12-11-09-02-02-13-12-11-09-02-02-13-12-10-07-02-02-13-12-10-07-02-02

Dia 5/6/92

08-11-09-04-03-02-15-11-09-04-03-12-14-11-09-04-03-11-13-11-08-04-03-14-11-07-04-03-11-10-06-04-03-11-10-06-04-03-11-10-04-03-02-11-10-04-03-02-11-10-04-03-02

Para cada série de medidas, trace o histograma de freqüências (com intervalo de classe de 1 mm). Para verificar se houve um crescimento dos mexilhões entre as duas amostragens, esse pesquisador pensou calcular a média de cada série de amostras e compará-las. O procedimento seria correto? Justifique sua resposta. Caso seja negativo, sugere o que ele deve fazer.

Um modelo teórico de distribuição de freqüência: a distribuição normal

AULA 17

objetivos

Ao final desta aula, você deverá ser capaz de:

- Adquirir noções de probabilidade e saber a relação entre probabilidade e freqüência.
- Conhecer as características da distribuição normal, a equação da curva normal e o cálculo das probabilidades.

INTRODUÇÃO

Nesta aula você vai relembrar noções de probabilidades que você deve ter já visto nas aulas de matemática. Essas noções são importantes em estatística. Elas permitem entender como transformar freqüências em probabilidades, e com isso conhecer e aplicar um dos modelos teóricos de distribuição de freqüência mais utilizados em estatística que é a ***distribuição normal.*** Você vai ver o quanto essa distribuição teórica pode ajudá-lo na análise estatística dos seus dados de Biologia ou Ecologia.

NOÇÕES DE PROBABILIDADE

Na aula passada, trabalhamos com distribuições limitadas, isto é, com um certo número de dados (observações) oriundos de amostras e distribuídos entre um valor mínimo e um valor máximo. Para cada classe (evento) calculamos a freqüência de ocorrência, obtendo assim uma distribuição de freqüências observadas. Mas será que essa distribuição de freqüência é a imagem da realidade? Aprofundando nosso estudo e aumentando o número de observações, constatamos que as freqüências relativas são modificadas. Para ter a certeza de que um evento (classe) ocorrerá x vezes num total de n observações, seria necessário fazer uma infinidade de medidas, o que é impossível na prática. Para tanto, os matemáticos estabeleceram leis baseadas nas distribuições teóricas. Na natureza, a ocorrência ou não-ocorrência de um evento segue certas leis que permitem calcular as freqüências teóricas e traçar um diagrama de distribuição teórica. São as leis da probabilidade.

Na aula de hoje veremos que a probabilidade de ocorrência de um evento pode ser estimada ao fazermos um grande número de observações e calculando as freqüências de ocorrência desse evento.

!

A probabilidade pertence ao domínio do REALIZÁVEL, e a freqüência ao domínio do REALIZADO.

Ao comparar as distribuições observadas com as distribuições teóricas estabelecidas a partir das leis probabilísticas, poderemos saber se tal fenômeno observado se realiza segundo uma ou outra lei.

Em caso afirmativo, poderemos transformar nossas observações em previsões, ou seja, transformar as freqüências observadas em probabilidades; faremos, assim, uma inferência, com o conhecimento da margem do erro. Antes disto, porém, é necessário definir a probabilidade e descrever os diversos modelos teóricos de distribuição de freqüência.

Para conhecer as noções de probabilidade você precisará definir ou recordar alguns conceitos.

Acaso: conjunto de causas que agem em sentidos opostos e regem um evento aleatório.

Evento aleatório: evento imprevisível, embora se conheça todas as condições de realizar-se. É apenas o acaso que decide. Vejamos alguns exemplos:

a) Vamos jogar a cara ou coroa. Lançamos a moeda (um lançamento de moeda é chamada de "prova") e apostamos, por exemplo, em obter coroa (coroa corresponde ao "evento esperado"). Nós temos uma certa esperança de que o evento coroa se realize. Essa esperança é chamada de probabilidade p, que é calculada pela fórmula $p = \frac{x}{n}$, onde x = número de eventos favoráveis e n =número de eventos possíveis. Logo, a probabilidade de tirar "coroa" é igual a 1/2 = 0,5, ou seja, há 50% de chance de obter coroa.

b) No jogo de dados são 6 eventos possíveis. A probabilidade de tirar um determinado número é igual a 1/6.

c) No jogo de 32 cartas são 32 eventos possíveis. A probabilidade de tirar uma determinada carta é igual a 1/32.

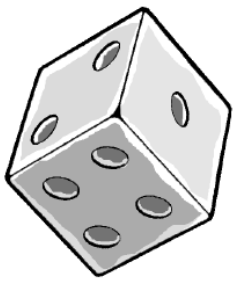
Os exemplos anteriores servem para mostrar que a probabilidade varia entre 0 (evento impossível) e 1 (evento certo). A probabilidade de um evento complementar (não-ocorrência do evento) é $q = 1 - \frac{p.}{m}$

Vamos conhecer um pouco mais sobre probabilidades.

Soma de probabilidades: a probabilidade de ocorrer um OU outro evento é igual à soma das probabilidades de cada evento. Por exemplo:

a) A probabilidade de tirar uma dama num jogo de 32 cartas = 4/32.

b) A probabilidade de tirar um rei de copas, ou um dez de espadas, ou um nove = 1/32+1/32+4/32 = 6/32.



Probabilidades compostas: a probabilidade de ocorrer um evento E outro evento em seguida, é igual ao produto das probabilidades de cada evento. Por exemplo: num jogo de dados, a probabilidade de obter 3 na primeira vez e um número ímpar na segunda é igual a 1/6 x 3/6 = 3/36 = 1/12.

Relação entre Probabilidade e Frequência: vamos supor que lançamos uma moeda 10 vezes. Pode acontecer de obtermos, por exemplo, 6 vezes coroa e 4 vezes cara. Nesse caso, a freqüência relativa observada de coroa é 6/10 = 0,60. De outro modo, lançando 100 vezes a mesma moeda podemos obter, por exemplo, 48 vezes coroa. A freqüência será então 0,48. Você deve estar concluindo que, aumentando o número de lances, a freqüência observada se aproxima de 0,50, o que corresponde à probabilidade do evento coroa. Assim, quando n tende para o infinito, a freqüência relativa f tende para a probabilidade p. É a lei dos grandes números: $n \to \infty \Rightarrow f \to p$.

!

A freqüência relativa é uma medida experimental da probabilidade e, inversamente, a probabilidade é uma freqüência prevista.

Pode-se aplicar então às probabilidades tudo que se aplica às freqüências relativas.

Distribuição aleatória: uma distribuição aleatória representa as probabilidades de uma variável aleatória. Ela constitui o limite para o qual tende a distribuição experimental quando se repete um grande número de vezes as observações.

Observe!

$$n \to \infty \Rightarrow f \to p \Rightarrow distr.freq. \to distr.aleat.$$

Agora, veremos um modelo teórico de distribuição de freqüência: a distribuição normal.

DISTRIBUIÇÃO NORMAL (DN)

A distribuição normal é, pelos recursos que ela oferece, uma das mais importantes distribuições de probabilidades conhecidas. Vamos conhecer suas características gerais:

a) A distribuição normal é uma distribuição contínua: x pode assumir qualquer valor entre + α e - α.

b) Ela é simétrica em relação à média aritmética, que corresponde também à mediana e à moda da distribuição.

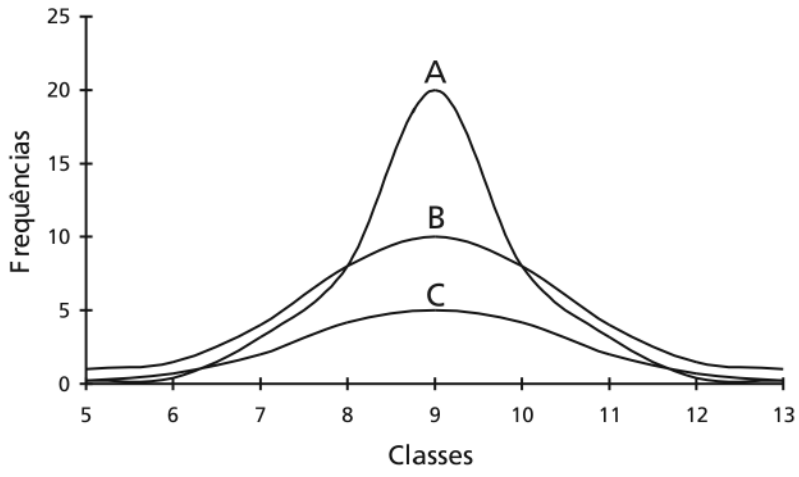
c) Ela é definida por dois parâmetros populacionais: a média (μ) e o desvio padrão (σ).

d) Ela é representada, graficamente, por uma curva contínua, em forma de "sino". A cada valor do desvio padrão σ corresponde uma forma de curva, mais fechada para pequeno σ, mais aberta para elevado σ. Veja na figura a seguir as 3 curvas que representam 3 séries de dados (A, B e C), com mesma média (na classe 9), porém com 3 desvios padrões diferentes. Olhe para a curva A, ela é mais fechada e corresponde a dados com menor desvio padrão, isto é, os dados são muito mais próximos à média.

A Distribuição Normal é também chamada de Distribuição de GAUSS ou de LAPLACE-GAUSS, em homenagem aos seus inventores, o astrônomo e matemático francês Pierre Simon Laplace (1749-1827) e o alemão Karl Friedrich Gauss (1777-1855).

!

Você se lembra do que é o desvio padrão? Senão, dê uma olhada na aula passada. Esse conceito é básico e muito importante para toda a disciplina.



Você deve lembrar, das suas aulas de matemática, que cada curva (como aparece na figura acima) corresponde a uma equação matemática. Pois existe também um equação que define a curva de distribuição normal. Vamos ver como ela é definida.

EQUAÇÃO DA CURVA NORMAL

A distribuição normal é definida pela equação:

$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

, onde μ e σ são a média e o desvio padrão da população.

Ao conhecer os parâmetros μ e σ, podemos calcular para cada valor de x a probabilidade que nos interessa. Porém, é preciso refazer o cálculo para cada μ e σ diferentes. Para evitar essa dificuldade, aplicamos uma transformação que torna a equação independente de μ e σ, multiplicando os dois lados da equação por σ e fazendo

$Z = \frac{x-m}{\sigma}$ e $Y = \sigma.y$. Nesse caso, temos uma nova equação da curva normal, chamada "equação da curva normal reduzida":

$$Y = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}Z^2}$$

Essa curva é única. Ela tem por média $\mu = 0$ e desvio padrão $\sigma = 1$. A transformação de x em Z faz com que o eixo das abscissas seja independente das unidades e represente somente unidades de desvio em relação à média.

Agora que conhecemos a Lei Normal e sua equação, vamos ver como calcular as probabilidades.

CÁLCULO DAS PROBABILIDADES

Primeiro, você deve saber de uma coisa importante: as probabilidades são representadas pela área sob a curva. Assim, a probabilidade p_x de um valor x ocorrer entre dois valores x1 e x2 é igual à área sob a curva entre x1 e x2. Do mesmo modo, a área total $(+\infty < x < -\infty)$ corresponde à probabilidade total ($p_X = 1$).

Você deve estar pensando: como vou medir a área sob a curva? Você poderia usar a equação da curva que dá as probabilidades Y para cada valor de x, mas não se assuste. Felizmente, os estatísticos fizeram todos os cálculos para nós e colocaram os resultados em tabela que você vai poder consultar no final do livro. Vamos falar um pouco mais sobre essas probabilidades e o uso da Tabela 1.

No final do volume há um Anexo com as Tabelas de 1 a 5.

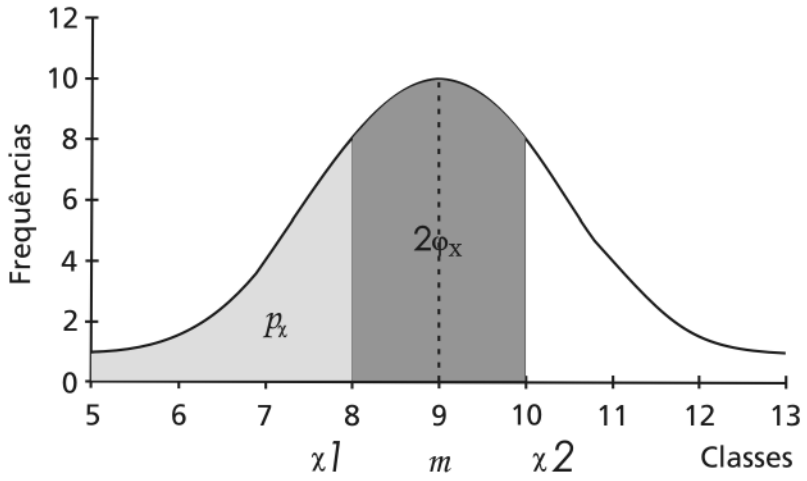
Você sabe como utilizar a tabela? São duas maneiras de "ler" uma probabilidade nessa tabela. Vejamos:

a) a primeira é "ler" a probabilidade, chamada $p_{X,}$ para um valor x estar no intervalo entre $-\infty$ e x1 (ou seja, até o limite x1).

b) a segunda é "ler" a probabilidade, chamada $2\varphi_X$, de um valor x estar incluído num intervalo formado por dois valores x1 e x2 simétricos em relação à média.

Fora desses intervalos, as probabilidades são 1 – px ou 1 – $2\varphi_X$.

Você vai entender melhor olhando a figura abaixo:



Acompanhe também os exemplos a seguir. Você vai entender melhor como se calcula a probabilidade de um determinado valor de x, tirado ao acaso na população, estar incluído entre determinados limites da distribuição.

Exemplo

Sabendo que a distribuição de tamanho da microalga Skeletonema costatum segue a lei normal de média 10 μm e desvio padrão 4 μm, calcule, para uma série de 100 células, o número de células tendo um tamanho:

(a) menor que 10 μm;

(b) maior que 14 μm;

(c) compreendido entre 8 μm e 13 μm.

Respostas:

Para entrar na tabela que dá as probabilidades ligadas a cada limite de tamanho (x) é preciso em primeiro lugar transformar x em $Z = \frac{x - \mu}{\sigma}$, ou seja:

para x1=10 $\Rightarrow Z = \frac{10-10}{4} = 0 \Rightarrow$ a tabela nos dá: $p_{x1} = 0{,}50$

para x2=14 $\Rightarrow Z = \frac{14-10}{4} = +1 \Rightarrow p_{x2} = 0{,}84$

para x3=8 $\Rightarrow Z = \frac{8-10}{4} = -0.50 \Rightarrow p_{x3} = 0{,}31$ e

para x4=13 $\Rightarrow Z = \frac{13-10}{4} = 0.75 \Rightarrow p_{x4} = 0{,}77.$

Logo, concluímos que:

a) Há 50% das células de tamanho entre -∞ e 10 μm. Resultado óbvio, pois 10 μm corresponde à média.

b) Há 84% das células de tamanho entre -∞ e 14 μm, e conseqüentemente, 100 – 84 = 16% acima de 14 μm.

c) Há 31% das células de tamanho até 8 μm e 77% de tamanho até 13μm, logo 77 – 31 = 46% de células de tamanho entre 8 e 13 μm.

Você está vendo que não é tão difícil! O importante é lembrar que, antes de consultar a tabela normal, você deve transformar os valores de x em Z, pois são os valores de Z que estão na tabela e não os valores de x.

Aproveitando esse comentário, vamos procurar na tabela alguns valores peculiares de Z e verificar a qual probabilidade eles correspondem.

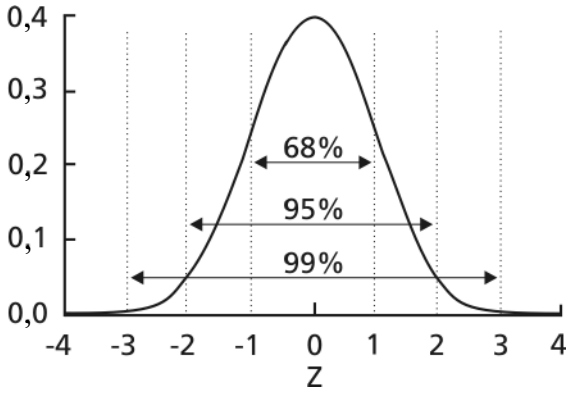
PROBABILIDADES PARA ALGUNS VALORES PECULIARES DE Z

a) Se Z = 1 a tabela dá 2φ=0,68 , isto é, há 68% de probabilidade de que x se encontre entre Z=+1 e Z = –1 (pontos simétricos em relação à média 0). Como = 1, temos x = μ + σ. ⇒ A probabilidade para que x se encontre entre (μ + σ) e (μ -σ) é de 0,68, e 0,32 fora deste intervalo.

b) Se Z = 2 a tabela dá 2φ = 0,95. Seguindo o mesmo raciocínio anterior temos x = μ + 2σ.⇒ A probabilidade para que x se encontre entre (μ + 2σ) e (μ -2σ) é de 0,95, e 0,05 fora deste intervalo.

c) Se Z = 2,6 a tabela dá 2φ = 0,99 ⇒ A probabilidade para que x se encontre entre (μ + 2.6σ) e (μ -2.6σ) é de 0,99, e de 0,01 fora deste intervalo.

Veja como esses intervalos são representados na figura abaixo:



Assim, retirando um dado ao acaso de uma série estatística de dados aleatórios, de média m e desvio padrão s, há 99% de chance de que ele seja compreendido entre (m + 2,6s) e (m – 2,6s), 95% entre (m + 2s) e (m – 2s).e 68% entre (m + s) e (m – s).

!

Os valores 2 e 2,6 foram arredondados. Os valores exatos seriam 1,96 e 2,58. Daí em diante usaremos sempre 2 e 2,6 para facilitar os cálculos.

Você deve começar a perceber o quanto a Lei Normal é interessante para fazer previsões e vai ser convencido disso com os exercícios que irá resolver a seguir.

RESUMO

Você aprendeu a calcular a probabilidade de um evento ocorrer. Essa probabilidade, que varia entre 0 (evento impossível de ser realizado) e 1 (evento sempre realizado), é regida por leis. Uma delas é a Lei Normal. Seu uso é freqüente em Biologia. Você observou que a Lei Normal se aplica a dados contínuos com distribuição simétrica em relação à média. Ela permite calcular a probabilidade de um determinado valor ocorrer dentro de um determinado intervalo, a partir da média e do desvio padrão.

EXERCÍCIOS

1. A média de uma população é de 21,65 cm e desvio padrão 3,21. Qual a probabilidade de que um indivíduo retirado ao acaso tenha um valor maior do que 28,55 cm ou menor do que 14,75 cm?

2. A média de chuva em certa cidade é de 18,75 milímetros com desvio padrão de 6,25 milímetros. Considerando normal a distribuição desses dados, calcule a probabilidade de nos próximos anos as chuvas estarem entre 15,00 e 25,00 milímetros.

3. Após uma pesca de camarões medimos as fêmeas ovadas. A média é 12 cm e o desvio padrão 1,5 cm. Suponhamos que os tamanhos estejam seguindo a Lei Normal. Pergunta-se:

a) Qual a proporção de indivíduos de tamanho superior a 14 cm, a 8,5 cm e compreendido entre 14 e 8,5 cm ?

b) Qual o tamanho ultrapassado por 10% dos camarões?

c) Entre quais tamanhos, simétricos em relação à média, estão compreendidos 95% dos camarões?

Estimativas e testes de hipóteses

AULA 18

objetivos

Ao final desta aula, você deverá ser capaz de:

- Estimar a média de uma população a partir de uma amostra.
- Realizar testes de hipóteses e de significância.
- Comparar duas médias e duas variâncias.

Pré-requisitos

Para compreender bem esta aula, você deve recordar as noções de variância, desvio padrão e grau de liberdade que vimos na Aula 16, bem como do cálculo das probabilidades de uma distribuição normal visto na Aula 17.

INTRODUÇÃO

Na Aula 6, você viu as origens e as primeiras noções da Estatística. Hoje, verá um dos principais objetivos da Estatística que é estimar os parâmetros de uma população a partir de uma amostragem. Com esse conhecimento, você vai poder comparar populações a partir de amostras e testar suas hipóteses.

ESTIMATIVAS DE UMA POPULAÇÃO

Certamente você sabe o que é estimar. Quando fala "eu acho que...", você está estimando, isto é, você tem alguma idéia sobre o assunto, mas nenhuma certeza. Na Estatística o conceito é o mesmo, porém não basta "achar", é preciso também saber se a estimativa é válida ou não. Em outros termos, qual a minha chance de errar quando eu afirmo alguma coisa?

A validade da estimativa depende da amostra ser uma boa imagem da população (representativa). Isso acontece quando todos os indivíduos têm a mesma chance de serem coletados, o que é possível quando a amostra não é viciada e as coletas são feitas ao acaso.

Observe que a relação população–amostra tem duplo sentido:

População ⟶ Amostra: seja uma população de razão sexual (RS) conhecida = 0,5 (sabemos que há 50% de fêmeas e 50% de machos). Pegamos uma amostra de 100 indivíduos. Podemos prever, com antecipação, que a RS desta amostra será de 0,5 ou próxima deste valor.

Amostra ⟶ População: seja agora uma população de RS desconhecida. Pegamos uma amostra de 100 indivíduos. Nela encontramos 50 machos e 50 fêmeas. Calculamos que a RS é igual a 0,5. Porém, não podemos afirmar que a RS da população é realmente de 0,5, podemos somente dizer que a RS é provavelmente de 0,5. Nesse caso nós estimamos que a RS é de 0,5.

Para tentar avaliar essa probabilidade, definimos um intervalo dentro do qual supomos que se encontra a RS desconhecida. A probabilidade de acertar depende da amplitude do intervalo: escolhendo um intervalo de 0 a 1, temos, obviamente, certeza de não errar. A probabilidade é máxima (p=1), o coeficiente de êxito (segurança)

é 100% e o de risco 0%. O intervalo escolhido é chamado *intervalo de confiança*. Diminuindo o intervalo de confiança, aumentamos os riscos de errar na estimativa.

Para você entender melhor, vamos raciocinar ao contrário. Se eu aceitar cometer um erro de 5% (isto é, uma probabilidade de errar de p = 0,05), qual deve ser o intervalo de confiança? Ou seja, entre que limites mínimo e máximo deve se encontrar a razão sexual da população? É fácil entender que, se eu não quisesse errar nada (0% de erro), posso afirmar que a RS da população está compreendida entre 0 e 1. O que é óbvio e não me ajuda em nada. Então tenho que aceitar um erro maior para poder ficar mais preciso na definição do intervalo. Entendeu?

Nas áreas da Biologia e Ecologia, consideramos aceitável errar até 5% , ou seja, ter uma segurança de 95%. Utilizaremos esse limite máximo de erro em todos nossos exemplos e exercícios.

!

Estimar = definir um intervalo de confiança associado a um coeficiente de segurança, ou de risco.

Então, vamos ver, agora, como calcular a *estimativa da média* de uma população desconhecida.

ESTIMATIVA DE UMA MÉDIA

Você entendeu que precisamos definir um intervalo dentro do qual a média real da população tem uma certa probabilidade de se encontrar. Qual é esse intervalo? Vamos logo utilizar um exemplo prático...

Seja uma amostra de 100 plantas. Medimos o tamanho de cada uma e calculamos a média m=10 cm e o desvio padrão s=3 cm. Qual é a estimativa da média da população?

Vamos raciocinar! Se eu repetisse várias vezes a amostragem e recalculasse a média a cada vez, teria uma série de médias um pouquinho diferentes, algumas abaixo de 10 cm, outras acima de 10 cm. Então, precisamos calcular os valores mínimo e máximo, simétricos em relação à média da amostra (10 cm), que definem um intervalo dentro do qual eu gostaria que ocorressem, por exemplo, 95% das médias (eu cometeria um erro de 5% afirmando que a média da população se encontra nesse intervalo).

Como fazer isso?

Vamos relembrar a aula passada onde vimos que, retirando um dado ao acaso de uma série estatística de *n* dados aleatórios, de média *m* e desvio padrão *s*, há 95% de chance de que ele seja compreendido entre (*m* + 2*s*) e (*m* - 2*s*). É isso mesmo que precisamos! Vamos então somar e subtrair duas vezes o desvio padrão da média. Porém, atenção, estamos tratando aqui da variação da média e não de todos os dados, logo, devemos utilizar o desvio padrão da média, que se chama erro padrão da média, Sm. Ele não é conhecido, mas os estatísticos demonstram que ele pode ser calculado pela fórmula: $S_m = \frac{s}{\sqrt{n}}$, onde *s* é o desvio padrão da amostra e *n* o número de dados.

!

ATENÇÃO: não confunda *erro padrão* com desvio padrão. Você já sabe que *desvio padrão* é o quanto os dados estão afastados da média, enquanto que o *erro padrão* é o quanto a média pode variar, se a gente repetir várias vezes a amostragem.

Voltemos a nosso exemplo e calculemos o erro padrão da média. Ele vale: $S_m = \frac{3}{\sqrt{100}} = 0{,}3$

Em seguida, basta somar e subtrair da média 2 vezes o valor 0,3 e conseguimos um intervalo de

$m \pm 2.S_m = 10 \pm 2 \cdot 0{,}3 =$ [9,4 – 10,6].

Em conclusão, o tamanho médio da população de plantas, de onde foi retirada a amostra, tem 95% de chance de se encontrar entre 9,4 cm e 10,6 cm. Com outras palavras, estimo que a média da população está entre 9,4 cm e 10,6 cm. Afirmando isso, estou correndo um risco de errar em 5% (probabilidade de erro de $p = 0{,}05$).

Se você achar que 5% de erro é muito, calcule o intervalo para somente 1% de erro. Neste caso, deve multiplicar S_m por 2,6, em vez de 2,0. Faça você mesmo o novo cálculo para 1% e veja o que acontece com o intervalo.

Lembre que os valores 2,0 e 2,6 que utilizamos, até agora, para o cálculo do intervalo de confiança, nas probabilidades de p = 0,05 e de p = 0,01, respectivamente, foram obtidos a partir da Lei Normal. Isso é válido quando o número de dados é relativamente grande (> 30). Mas quando n é pequeno (<30), o que acontece freqüentemente nas pesquisas, a distribuição dos dados não é normal. A curva é mais achatada (chamada *hipernormal*), e existe uma curva diferente para cada valor de n, até n=30. Em conseqüência, seria necessário recalcular as probabilidades para cada curva e cada valor de n. Não se preocupe, o matemático **STUDENT** fez esses cálculos para você. Ele calculou, para cada valor de n, um valor limite chamado pela letra t, que deve ser usado no lugar de 2,0 ou 2,6. Esses valores de t estão na Tabela 2, no Anexo.

A distribuição t de Student foi criada por um pesquisador irlandês de nome William Sealy Gosset (1876-1936). Ele publicou seus trabalhos sob o pseudônimo de Student.

Como usar a tabela? Você deve escolher uma probabilidade de erro (coluna 0,05 ou 0,01, por exemplo) e a linha correspondendo ao grau de liberdade que vale n – 1. O valor encontrado na intersecção dessa coluna e dessa linha é o valor de t que você vai utilizar para o cálculo do intervalo de confiança da média. Veremos mais tarde outras aplicações dessa tabela.

!

Você se lembra do que é o ***grau de liberdade***? Não? então recorde esse conceito na Aula 6.

Vamos exemplificar com o mesmo exercício, mas supondo que a amostragem foi somente de 9 plantas em vez de 100, a média e o desvio padrão permanecendo os mesmos. Nesse caso, qual será a estimativa da média da população?

O intervalo de confiança *I*c da média é agora igual a $\pm t.S_m$, onde t = 2,31 para n – 1 = 8 graus de liberdade (veja na tabela de Student na interseção da coluna 0,05 e da linha 8), e o erro padrão da média, $S_m = \frac{3}{\sqrt{9}} = 1,0$.

Observe que, para n = 9 plantas, o intervalo dentro do qual a média da população tem 95% de chance de se encontrar é agora: $10 \pm 2{,}31.1{,}0 \Rightarrow [7{,}69\text{-}12{,}31]$.

Verifique que, para uma mesma probabilidade de erro de $p = 0{,}05$, o intervalo de confiança da média aumentou, comparativamente ao exercício anterior com 100 plantas, isto é, a estimativa da média ficou menos precisa, o que é coerente, pois trabalhamos com menos dados. Você percebe a importância de uma boa amostragem em Estatística?

Qual seria o *n* ideal para uma boa estimativa? Como você viu, a distribuição é normal a partir de *n*=3, quando o valor de *t* estabiliza em volta de 2,0 (para *p*=0,05). Mas, não é sempre possível conseguir um *n* tão grande. Então qual seria o *n* mínimo possível? Observe a tabela de Student: para n–1 = 1 (isto é, 2 dados), o valor de *t* é elevadíssimo (=12,7), mas cai bastante (= 4,3) com *n*–1=2 (isto é, 3 dados). Em conclusão, aconselha-se ter pelo menos *n*-3.

TESTES DE HIPÓTESES E SIGNIFICÂNCIA

Na prática, somos freqüentemente levados a tomar decisões acerca de populações, baseadas nas informações das amostras. São decisões chamadas de *decisões estatísticas*. Por exemplo, podemos querer decidir se um novo remédio é eficaz na cura de uma doença, se um método de coleta é melhor que outro etc. Ao se tentar chegar às decisões, é conveniente a formulação de hipóteses ou de conjecturas acerca das populações interessadas. Essas suposições, que podem ser verdadeiras ou não, são denominadas *hipóteses estatísticas*.

Em alguns casos, formula-se uma hipótese estatística com o único propósito de rejeitá-la. Por exemplo, desejando decidir se uma área da floresta é mais rica em espécies de plantas do que outra, formula-se a hipótese de que *não há diferença* entre elas (isto é, que a diferença observada é somente devida à flutuação das amostras provenientes da mesma população). Essa hipótese é denominada *hipótese nula* e é representada por H_0. Qualquer hipótese diferente de H_0 é chamada de *hipótese alternativa*.

Estamos inclinados a rejeitar uma hipótese nula (H_o = não há diferença entre amostras), quando a diferença observada é grande demais para ser unicamente devido ao acaso.

Concluímos, então, que a diferença é *significativa*. Essa decisão é tomada na base do cálculo da probabilidade de errar ao rejeitar H_o e afirmar que duas amostras são significativamente diferentes. Os processos que permitem tomar essa decisão são chamados de *testes de hipóteses* ou de *significância*.

Em razão do caráter aleatório da amostragem, uma tomada de decisão é sempre acompanhada de um probabilidade de erro. Existem dois tipos de erro:

– Erro do Tipo I = rejeitar uma hipótese quando ela é verdadeira.

– Erro do Tipo II = aceitar (= não rejeitar) uma hipótese quando ela é falsa.

Ao testar uma hipótese, a probabilidade máxima com a qual estaremos dispostos a correr o risco de um erro tipo I é denominada *nível de significância*. Essa probabilidade é geralmente representada por α. Na prática, é usual escolher um nível de significância de 0,05 ou 0,01, embora possam ser usados outros valores. Assim, ao dizer que uma população A é significativamente diferente de uma população B, ao nível de significância de $\alpha = 0,05$, temos 5% de chance de errar, ou seja, temos uma confiança de 95% de que essa decisão esteja certa.

Na sua tomada de decisão, o pesquisador procura sempre diminuir ao máximo o erro. Para isso ele deve aumentar o número n de amostras, o que não é sempre possível.

COMPARAÇÃO DE DUAS MÉDIAS

Agora, você acumulou conhecimento suficiente para começar a aplicar a estatística como ferramenta de decisão. Pode, por exemplo, comparar as médias $m1$ e $m2$ de duas populações, obtidas a partir de 2 amostras de efetivo $n1$ e $n2$ e variâncias $v1$ e $v2$. Queremos saber se as duas médias são significativamente diferentes.

Calculamos a diferença d entre as médias, $d = m1-m2$.

O raciocínio é o seguinte: se as duas amostras tivessem sido retiradas da mesma população, haveria pouca diferença entre as médias,

e *d* seria pequeno. Então, será que o valor d obtido entre as duas amostras é grande demais para concluirmos que se trata da mesma população ou de duas populações iguais?

Vamos aplicar um teste ao valor de *d*, chamado teste *t* de Student.

Vamos imaginar que você repita a amostragem, nas duas populações, um grande número de vezes e calcule as diferenças. Você terá uma série de valores de d distribuídos normalmente de média d_m e desvio padrão S_d, chamado erro padrão das diferenças. Considerando a hipótese nula ($d = 0$), isto é, se as duas populações fossem iguais, haveria 95% dos valores de d dentro do intervalo:

$d_m \pm 2.S_d$

Você já viu isso quando estimamos a média. Agora é a mesma coisa, pois trata-se da média das diferenças.

Logo, há uma probabilidade $2\varphi = 0{,}95$ de que um valor de *d*, tirado ao acaso, seja situado no intervalo $\pm 2.S_d$ e uma probabilidade $1–2\varphi = 0{,}05$, que seja situado fora desse intervalo. Podemos dizer, então, que 95% dos valores de *d* devem ser inferiores ou iguais ao limite $2\ S_d$ do intervalo, ou seja, $\frac{d}{S_d} \leq 2$.

!

Você deve saber que 2φ é *a* probabilidade de ocorrer um valor dentro de um intervalo simétrico em relação à média. Retorne à Aula 17 sobre distribuição normal para relembrar.

Observe que, se *d* for muito grande, poderemos ter $\frac{d}{S_d}$ maior do que 2, logo, fora do intervalo dos 95%, e, nesse caso, deveremos rejeitar a hipótese nula de igualdade das médias. Poderemos, então, afirmar que as duas médias são, significativamente, diferentes, com probabilidade de errar de $p = 0{,}05$.

Ao acompanhar os exemplos anteriores você, certamente, percebeu que precisamos do *erro padrão das diferenças*, S_m. Ele é desconhecido, mas os estatísticos demonstram que ele pode ser calculado pela fórmula: $S_d = \sqrt{\left(\frac{v_1}{n_1} + \frac{v_2}{n_2}\right)}$, onde $v1$ e $v2$ são as variâncias e $n1$ e $n2$ o número de dados das amostras.

Você vai se familiarizar com esse importante teste observando o exemplo a seguir:

Sejam 2 amostras, A e B, com os seguintes parâmetros na Tabela 18.1.

Tabela 18.1

	amostra A	amostra B
número de dados (n)	200	100
média (m)	2,5	2,8
variância (v)	0,8	0,6

Será que a diferença observada entre as 2 amostras permite dizer que elas foram coletadas em 2 populações diferentes? Ou seja, que as duas médias são significativamente diferentes, ao nível de probabilidade de errar de p = 0,05?

Vamos efetuar os cálculos:

d = 2,8-2,5 = 0,3,

$S_d = \sqrt{\left(\frac{0,8}{200} + \frac{0,6}{100}\right)} = 0,1$, logo temos $\frac{d}{S_d} = \frac{0,3}{0,1} = 3$, que é

superior a t = 2.

Podemos rejeitar a hipótese nula e afirmar, com probabilidade de erro p<0,05, que as duas médias são significativamente diferentes. Como $\frac{d}{S_d}$ é superior a 2,6, podemos também concluir que a probabilidade de erro é p<0,01. Dizemos que a diferença entre as médias é *altamente significativa*.

Nós trabalhamos, neste exemplo, com número elevado de dados em cada amostra. O que aconteceria se tivéssemos poucos dados (<30)? O princípio do teste é o mesmo, o que muda é a fórmula do erro padrão da diferença S_m e o valor limite do t. Vamos exemplificar...

Seja o mesmo exercício que o anterior, mas com n1=10 e n2 =12.

Neste caso, a fórmula do erro padrão é:

$$S_d = \sqrt{\left[\frac{n_1 v_1 + n_2 v_2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right]} = 0{,}37, \text{ logo,}$$

$$\frac{|d|}{S_d} = \frac{0{,}3}{0{,}37} \approx 1.0$$

Comparamos esse valor com o valor limite *t* da tabela de Student, para $n1 + n2 - 2 = 20$ graus de liberdade. A tabela de t indica 2,09 ($p = 0{,}05$). O valor calculado é bem inferior a 2,09. Não é possível rejeitar a hipótese nula. Se eu afirmasse que as médias são diferentes, cometeria um erro maior do que 5%.

!

Observe que quando as duas amostras têm o mesmo número de dados ($n1 = n2$), a fórmula do erro padrão se simplifica em $S_d = \sqrt{\frac{v_1 + v_2}{n - 1}}$, e entramos na tabela de *t* com *n*-1 graus de liberdade.

Importante observação

No caso de não rejeitar a hipótese nula (para uma determinada probabilidade de erro), o pesquisador não poderá tirar uma conclusão. Ele não poderá afirmar que as 2 populações são iguais sem correr um elevado erro tipo II, pois, como vimos nos exemplos anteriores, a diferença pode ser ou não significativa, dependendo do número *n* de dados. Ele deverá somente concluir que: "*nas condições estabelecidas de amostragem, não foi possível evidenciar uma diferença significativa entre as populações*".

!

A Estatística testa unicamente diferenças, nunca igualdades.

!

Para relembrar: cometer um erro tipo II consiste em aceitar a hipótese nula (igualdade entre médias) quando deveria ter sido rejeitada. Ao contrário, o erro tipo I consiste em rejeitar uma hipótese nula quando deveria ter sido aceita.

Até o momento, você aprendeu a comparar duas populações utilizando as médias como critério de comparação. A média é um parâmetro da população, mas você já viu que existem outros, como a variância. Por que não usar a variância para comparar duas populações? Isso é possível, já que, se duas populações são iguais, não apenas as médias devem ser iguais, mas também as variâncias. Vamos então comparar as variâncias...

COMPARAÇÃO DE 2 VARIÂNCIAS

O princípio da comparação de variâncias é o mesmo que para as médias. Formulamos a hipótese de que as variâncias são iguais e vamos ver se, com o teste, podemos rejeitar essa hipótese, dentro de uma certa probabilidade de erro, por exemplo 5% ($p < 0,05$).

O teste, porém, é diferente. Em vez de testar a diferença entre variâncias, vamos testar a razão F entre a maior e a menor variância.

SNEDECOR & FISHER

Snedecor foi estatístico americano (1882-1974) e Sir Ronald Fisher, estatístico inglês (1890-1962). O nome do teste (teste F) foi dado em homenagem a este último.

Sejam 2 amostras com $n1$ e $n2$ indivíduos e variância v1 e v2, respectivamente. Calculamos $F = \frac{v_1}{v_2}$.

A distribuição de F não segue a lei normal. Existe uma infinidade de valores de F, dependendo do grau de liberdade de cada amostra $\gamma_1 = n_1 - 1$ e $\gamma_2 = n_2 - 1$.

Todas as distribuições de F foram calculadas por **SNEDECOR & FISHER**. Para cada par de graus de liberdade $\gamma 1$,$\gamma 2$, eles calcularam um valor limite de F abaixo do qual encontram-se 95% (ou 99%) dos valores de F da população. Veja a tabela de F em anexo (Tabela 3).

Vamos praticar o teste F com o exercício a seguir:

Sejam 2 lotes de mudas de uma espécie de planta, com adubação diferente. Medimos os indivíduos de cada lote para verificar se houve uma diferença de crescimento significativa, isto é, se houve influência do adubo:

Tabela 18.2

	lote 1	lote 2
efetivo (n)	31	17
variância (v)	10	3

Calculamos:

$$F = \frac{v_1}{v_2} = \frac{10}{3} = 3,33$$

e vamos comparar com o valor da tabela de F.

Como usar a tabela de F? Escolha a probabilidade (0,05 ou 0,01), entre na coluna com grau de liberdade da maior variância (30) e na linha com o grau de liberdade da menor variância (16).

A tabela de F dá, para $\gamma_1 = n_1 - 1 = 30$ e $\gamma_2 = n_2 - 1 = 16$, um valor limite de F = 2,2 (para p=0,05) e F= 3,1 (para p=0,01). Logo, F é altamente significativo, pois é maior que 3,1. Podemos rejeitar a hipótese nula e afirmar, com uma probabilidade de 0,01 de errar, que a adubação tem influência sobre o crescimento.

RESUMO

Você aprendeu a *estimar* a média de uma população a partir de uma amostra. Percebeu que estimar consiste em calcular um *intervalo de confiança*, dentro do qual a média verdadeira da população tem uma certa probabilidade de ocorrer.

Em seguida, aprendeu a comparar duas populações. Para isso, viu que podemos comparar duas médias com o *teste t de Student*, ou duas variâncias com o *teste F de Snedecor*. Esses testes são chamados *testes de hipóteses*. Eles consistem em verificar se podemos rejeitar a hipótese de igualdade entre populações, chamada *hipótese nula*, dentro de uma certa probabilidade de erro.

EXERCÍCIOS

1. Calcule o intervalo de confiança das médias do exercício 4 da Aula 16, ao nível de 0,05 de probabilidade.

2. Considere um levantamento de larvas de camarão na área de Cabo Frio (RJ) feito em 40 estações repartidas em 2 áreas: uma área 1 (estações de 1 a 20) e uma área 2 (estações de 21 a 40). Queremos saber se existe uma diferença de abundância de larvas entre essas duas áreas:

Área 1		Área 2	
Estação	Nº larvas	Estação	Nº larvas
1	2	21	74
2	17	22	132
3	25	23	18
4	188	24	22
5	89	25	31
6	123	26	24
7	31	27	49
8	132	28	48
9	94	29	80
10	26	30	39
11	28	31	39
12	22	32	11
13	63	33	2
14	83	34	12
15	70	35	3
16	26	36	13
17	31	37	0
18	35	38	0
19	11	39	2
20	65	40	0

3. Peixes são submetidos a dois regimes alimentares com taxa de proteínas diferentes. Medimos o aumento de peso dos indivíduos de cada lote:

Regime	Nº de peixes	Aumento de peso
Rico em proteínas	12	134,146,104,119,124,161,113,129,107,83,97,123
Pobre em proteínas	7	70,118,101,85,107,132,94

Há uma diferença de aumento de peso entre os lotes?

A análise de variância

AULA 19

objetivo

Ao final desta aula, você deverá ser capaz de:

- Conhecer os princípios, os cálculos e as condições de aplicação de uma análise de variância.

Pré-requisitos

Você deve ter entendido perfeitamente como realizar o teste F de comparação de variâncias visto na Aula 18.

INTRODUÇÃO

Na aula passada, você aprendeu a comparar duas médias e duas variâncias para saber se duas populações podiam ser consideradas estatisticamente diferentes. Entretanto, freqüentemente, é necessário comparar mais de duas populações. Como fazer isso? Você poderia sugerir aplicar um teste *t* ou um teste F às populações tomadas duas a duas. Seria possível, mas não é a melhor opção e o trabalho vai ser imenso! Por essa razão, os estatísticos criaram um jeito que possibilita comparar todas as médias de uma só vez. Trata-se da técnica conhecida por *análise de variância*.

Nesta aula, você vai aprender a aplicar a análise de variância. Esse método permite verificar a influência de um determinado fator sobre o crescimento de um organismo, como por exemplo, diversos tipos de alimentos, diferentes temperaturas, salinidades etc. A análise de variância é um método muito eficiente, sobretudo em ciências experimentais, mas que exige uma perfeita adequação do desenho experimental para poder tirar proveito dessa eficiência.

O PRINCÍPIO DA ANÁLISE DE VARIÂNCIA

Você deve se lembrar da primeira aula, quando falamos da variabilidade de uma medida. Lembre que essa variabilidade tem duas origens, uma é a variabilidade devido a um determinado fator, outra é a variabilidade devido ao acaso. A análise de variância vai permitir separar essas duas fontes de variação e verificar se o fator estudado tem um efeito significativo. Em outras palavras, a análise visa a comparar a variância fatorial $\mathbf{v}_f$ com a variância residual $\mathbf{v}_r$, para testar a significância do efeito do fator por meio do teste $F=\frac{V_f}{V_r}$ (você já viu esse teste F na aula passada, lembra?).

O pesquisador vai planejar sua experiência e organizar seus dados de acordo com o fator que ele quer estudar. Para cada nível do fator (tratamento), ele vai realizar medidas sucessivas independentes (grupos de réplicas). Como já falamos, a variabilidade geral dessas medidas tem duas origens: a variabilidade devida ao fator (intergrupos) e a variabilidade própria, residual, devida ao acaso (intragrupos).

Vamos ver como calcular cada uma dessas variâncias, para depois compará-las com o teste F.

CÁLCULO DAS VARIÂNCIAS

Você já sabe calcular uma variância. Vamos relembrar a fórmula:

$$v = \frac{\sum x^2 - \frac{(\sum x)^2}{n-1}}{n-1}$$

O numerador é chamado de "Dispersão *D*", e o denominador de "Grau de liberdade γ".

Logo, vamos calcular a variância fatorial $v_f = \frac{D_f}{\gamma_f}$ e a variância residual $v_r = \frac{D_r}{\gamma_r}$.

Observe, na tabela a seguir, todas as fórmulas que você vai precisar para fazer os cálculos.

Tabela 19.1

	Dispersão	Grau de liberdade	Variância
Geral	$D_g = \sum_g x^2 - \frac{T_g^2}{N}$	$\gamma_g = N - 1$	$V_g = \frac{D_g}{\gamma_g}$
Fatorial	$D_f = \sum \frac{T^2}{n} - \frac{T_g^2}{N}$	$\gamma_f = k - 1$	$V_f = \frac{D_f}{\gamma_f}$
Residual	$D_r = \sum_g x^2 - \sum \frac{T^2}{n}$	$\gamma_r = N - k$	$V_r = \frac{D_r}{\gamma_r}$

Vejamos o que representa cada letra nas fórmulas:

- *x* é o valor de cada dado;
- T_g é o total geral dos dados;
- T é o total dos dados em cada tratamento;
- N é o número total de dados;
- n é o número de dados em cada tratamento;
- k é o número de tratamentos;
- γ é o grau de liberdade.

Verifique que $D_g = D_f + D_r$ e $\gamma_g = \gamma_f + \gamma_r$

Nada melhor do que um pequeno exercício para você entender o desenvolvimento dos cálculos.

EXERCÍCIO

Sejam medidas de intensidade fotossintética (em μl de oxigênio por mg de peso seco e por hora) de uma espécie de alga marinha incubada em 3 intensidades de luz diferentes L1, L2 e L3. Foram efetuadas 8 medidas para cada experimento. Queremos verificar se existe uma influência significativa do fator Luz sobre a fotossíntese dessa planta:

Você deve organizar os seus dados conforme a tabela a seguir:

Tabela 19.2

	L1		L2		L3		
Réplicas	**X**	**X^2**	**X**	**X^2**	**X**	**X^2**	
(1)	6.14		4.47		9.63		
(2)	3.86		9.90		6.38		
(3)	10.04		5.75		13.40		
(4)	7.49		11.80		14.50		
(5)	6.80		4.95		14.50		
(6)	10.00		6.49		10.20		
(7)	11.60		5.44		17.70		
(8)	5.80		9.90		12.30		
T	62.09		58.70		98.61		$Tg^2/N=2005,68$
n	8		8		8		N = 24
$\sum x^2$		531.30		484.51		1302.12	$\sum_g x^2=2317,93$
$\frac{T^2}{n}$	481,71		430,71		1215,49		$\sum \frac{T^2}{n}=2127,91$
Var.	7.05		7.69		12.38		
Média	7.76		7.34		12.33		

Observe que, na última coluna, você tem todas as informações necessárias para calcular as variâncias.

Com isso, chega aos resultados finais que são expressos na **Tabela 19.3.**

Tabela 19.3

Fonte de variação	Dispersão D	Grau de Lib. γ	Variância v	F	$F_{0,05}$	$F_{0,01}$
Geral	312,25	23	13,58			
Fatorial	122,23	2	61.12	6,76	3,47	5,78
Residual	190,02	21	9,04			

Observe:

Dispersão geral = Dispersão fatorial + Dispersão residual; de fato temos: 312,25=122,23 + 190,02.

Grau de liberdade geral = Grau de liberdade fatorial + Grau de liberdade residual; de fato temos: 23 = 2+21.

O valor F foi obtido fazendo $F = \frac{v_f}{v_r} = \frac{61,12}{9,04} = 6,76$

Esse valor é comparado ao valor da tabela de F para p=0,05 e p=0,01.

Como utilizar a tabela? Você já viu isso na aula anterior.

Agora você escolherá a coluna para o grau de liberdade da variância fatorial e a linha para o grau de liberdade da variância residual. Ou seja, encontrará o valor de F que está na intersecção da coluna 2 e da linha 21. Você concluirá que o valor é 3,47 na tabela com probabilidade 0,05 e 5,78 na tabela com probabilidade 0,01.

Observe que 6,76 > 5,78. Podemos rejeitar a hipótese nula e afirmar ao nível de probabilidade de 0,01 que existe uma influência altamente significativa do fator "luz" sobre a fotossíntese da alga.

A análise que você acaba de fazer mostrou uma influência global do fator luz. Mas você pode se perguntar: qual das três intensidades de luz tem influência mais significativa?

Para isso você pode aplicar o *Teste da MDS (Menor Diferença Significativa)*.

Esse teste consiste em:

a) escolher um nível de significância, por exemplo 0,05;

b) calcular a Menor Diferença Significativa MDS = $t.S_d$, onde t = valor da tabela de Student para o grau de liberdade do resíduo, e S_d = erro padrão da diferença. Ele é calculado pela fórmula: $S_d = \sqrt{\frac{kv_r}{n}}$ quando o número de dados n é igual em todos os tratamentos, sendo k o número de tratamentos, e v_r a variância residual, ou $S_d = \sqrt{v_r\left(\frac{1}{n_1} + \frac{1}{n_2} + ... + \frac{1}{n_k}\right)}$, quando os n são diferentes.

No exercício, temos 3 tratamentos (as 3 intensidades de luz) com 8 dados em cada, logo $S_d = \sqrt{\frac{3x9,04}{8}} = 1,84$.

A Menor Diferença Significativa vale, então, MDS = 2,09 x 1,84 = 3,85.

c) comparar as diferenças das médias com a MDS. Se a diferença for maior que a MDS, afirmamos, então, que essa diferença é significativa na probabilidade de 0,05.

Veja o que acontece com as médias dos 3 tratamentos L1, L2 e L3 do exercício:

L2 – L1 = 0,42 ⇒ diferença não significativa a p<0,05, pois 0,42 < 3,85;
L3 – L1 = 4,99 ⇒ diferença significativa a p<0,05, pois 4,99>3,85;
L3 – L2 = 4,57 ⇒ diferença significativa a p<0,05, pois 4,57 >3,85.

Você pode, então, concluir que é somente a intensidade L3 de luz que proporciona um aumento significativo da fotossíntese das algas.

CONDIÇÕES DE APLICAÇÃO DA ANÁLISE DA VARIÂNCIA

Você está vendo que a análise de variância é uma ferramenta bastante importante para tirar conclusões sobre resultados de experimentos. Mas, atenção, ela deve ser aplicada adequadamente e tem uma exigência que deve ser respeitada: as variâncias de cada grupo devem ser, teoricamente, idênticas (homogêneas).

Existem testes para verificar a homogeneidade das variâncias. Um deles é o teste de Hartley, bastante simples. Ele consiste em:

a) calcular a razão F entre a variância máxima e a variância mínima;
b) comparar F ao valor limite da tabela de F para 2 e n-1 graus de liberdade;
c) se o F calculado for inferior ao F da tabela, podemos aceitar a hipótese de homogeneidade das variâncias.

Observe que, no exercício, temos F = 12,38 / 7,05 = 1,76 bem inferior ao valor da tabela de F que é F = 19,36 a p = 0,05 (para 2 e 7 graus de liberdade). Logo, podemos aceitar a hipótese de homogeneidade das variâncias. A análise de variância pode ser realizada sem problema.

E se o teste mostrar que as variâncias não são homogêneas, o que fazer? Procure verificar por que isso aconteceu. Às vezes, um dado pode estar bem diferente dos outros por motivo conhecido (um erro de experiência, por exemplo). Nesse caso é permitido eliminar esse dado.

!

Não seja esperto demais eliminando todos os dados que divergem muito da média. Isso se chama "trapacear" e não condiz com a ética de uma pesquisa honesta.

É possível também tentar homogeneizar as variâncias, isto é, reduzir as diferenças entre dados, aplicando uma transformação aos dados, como por exemplo, uma transformação logarítmica. Em último caso, a sugestão é utilizar um *teste não paramétrico* que independe de qualquer exigência.

!

A estatística que estamos aprendendo nesta disciplina é chamada de *paramétrica*, pois, como você está percebendo, utilizamos nos cálculos os parâmetros das distribuições (média, variância). Existe uma estatística chamada *não paramétrica* que não utiliza esses parâmetros, e conseqüentemente, não exige a normalidade dos dados ou a homogeneidade das variâncias. Ela é de fácil aplicação, mas com muitos menos recursos que a estatística paramétrica. Você terá oportunidade de conhecer essa estatística em outros cursos mais avançados.

A análise de variância que você acaba de conhecer tem por objetivo estudar o efeito de um único fator, por exemplo a luz, como vimos no exercício (análise monofatorial). Esse mesmo tipo de análise pode ser aplicado também para testar o efeito de 2 fatores (análise bifatorial). Os cálculos são complexos demais para serem realizados sem ajuda do computador, mas os princípios são os mesmos. Vamos explicar um pouco.

Suponha que você quisesse testar a ação de um fator A com "**a**" níveis (tratamentos) diferentes e de um fator B com "**b**" tratamentos diferentes. Você teria então **k = a.b** tratamentos (grupos de dados), e um total de **N = n.a.b** dados (sendo **n** o número de repetições igual em cada grupo).

No caso de 2 fatores A e B, a dispersão fatorial se subdivide em:

a) dispersão D_A, devida ao fator A, com graus de liberdade γ_A ;

b) dispersão D_B, devida ao fator B, com graus de liberdade γ_A;

c) dispersão D_{AB}, devida à interação AB, com graus de liberdade γ_{AB}.

Em razão da maior complexidade, as fórmulas das dispersões e grau de liberdade não serão apresentadas aqui. Porém, os demais cálculos e a interpretação são idênticos à análise monofatorial. As variâncias devidas ao fator A, ao fator B e à Interação AB são calculadas dividindo as dispersões pelo respectivo grau de liberdade, e testadas calculando os valores de **F** para cada uma: $F = \frac{V_A}{V_r}$, $F = \frac{V_B}{V_r}$ e $F = \frac{V_{AB}}{V_r}$. Quando F é superior ou igual ao F da tabela, a ação do fator é significativa.

Uma diferença em relação à análise monofatorial é a inclusão do efeito da "interação". Uma "interação AB" significa que a influência do fator A depende do nível do fator B, ou vice-versa.

Baseado no mesmo raciocínio podemos efetuar uma análise com mais de 2 fatores (análise multifatorial). O procedimento é o mesmo, apenas aumentando o número de interações. Por exemplo, com três fatores A, B e C, teremos o efeito das interações AB, AC, BC e ABC.

Por fim, o que é essencial para realizar uma análise bi ou multifatorial, é um bom planejamento do experimento. Todas as combinações de tratamento devem ser previstas. Por exemplo, se você testar 3 níveis de luz L1, L2 e L3 e 2 níveis de temperatura, T1 e T2, sobre o crescimento de uma planta, você deverá planejar 6 tratamentos, com um certo número de repetições em cada um. Veja como deverá ser montada a tabela de dados com 4 repetições (x).

Tabela 19.4

T1			T2		
L1	L2	L3	L1	L2	L3
x	x	x	x	x	x
x	x	x	x	x	x
x	x	x	x	x	x
x	x	x	x	x	x

A tabela final dos resultados de uma análise bifatorial será mostrada a seguir:

Tabela 19.5

Fonte de variação	Dispersão D	Grau de Lib. γ	Variância v	F	$F_{0,05}$	$F_{0,01}$
Fator T	D_T	γ_T	v_T	F_T		
Fator L	D_L	γ_L	v_L	F_L		
Interação TL	D_{TL}	γ_{TL}	v_{TL}	F_{TL}		
Residual	D_r	γ_r	v_r			

RESUMO

Nós vimos, nesta aula, como realizar uma análise de variância. É uma técnica muito usada em experiência para verificar, através de vários tratamentos e repetições, a influência de um ou mais fatores sobre uma determinada variável. Ela permite comparar, simultaneamente, várias médias através de testes F.

EXERCÍCIOS

1. Uma espécie de camarão apresenta duas variedades com o mesmo valor econômico. A fim de rentabilizar ao máximo a produção, estuda-se a idade média da desova de cada variedade. Constata-se que, para cada variedade, a idade média de 7 lotes de fêmeas é a seguinte:

Variedade I	72	71	91	72	73	72	75
Variedade II	88	75	77	76	74	67	72

Compare as duas variedades com um teste t e com uma análise de variância.

2. O crescimento do mexilhão serve de teste numa experiência sobre a toxidade de detergentes. Compõe-se 5 lotes de 10 indivíduos do mesmo tamanho:

Lote 1 = sem detergente (testemunho);
Lote 2 = com detergente DT1;
Lote 3 = com detergente DT2;
Lote 4 = com detergente DT3.
Lote 5 = com mistura de detergentes.

Mede-se o crescimento dos mexilhões após um mês de experiência:

Lote 1	Lote 2	Lote 3	Lote 4	Lote 5
73	57	58	62	58
67	58	61	66	59
70	60	56	65	58
72	59	58	63	61
65	62	57	64	57
71	60	56	62	56
67	60	61	65	58
67	57	60	65	57
70	59	57	62	57
68	61	58	67	59

Quais são as conclusões?

O teste do Qui-quadrado

AULA 20

objetivos

Nesta aula você deverá ser capaz de:

- Compreender um conceito muito importante nos estudos estatísticos: a variável aleatória.
- Saber como calcular seu valor esperado.

INTRODUÇÃO

Você aprendeu, até agora, a realizar testes de hipóteses, comparando duas médias pelo teste ***t,*** e mais de duas médias pela análise de variância. Podemos também comparar, não apenas os parâmetros de duas distribuições, mas as duas distribuições por completo, isto é, o conjunto dos dados de duas distribuições. Existe um teste para isso que se chama ***teste do QUI-QUADRADO***. Vamos ver como ele funciona e quais suas aplicações.

Como você já sabe, de acordo com as regras da probabilidade, os resultados obtidos por meio de amostras nem sempre concordam exatamente com os teóricos esperados. Por exemplo, embora a teoria permita esperar 50 caras e 50 coroas num lançamento de 100 moedas, raramente obtemos um resultado exato. Deseja-se, então, saber se as freqüências observadas diferem de modo significativo das esperadas. O teste do χ^2 permitirá responder a essa pergunta.

QUI-QUADRADO

Indicado por χ^2, é uma estatística concebida por Karl Pearson em 1899.

DEFINIÇÃO

O teste do χ^2 serve para comparar uma distribuição de freqüência observada com um distribuição de freqüências teóricas. Esse tipo de teste é chamado de *prova de aderência* ou *prova de independência*.

Seja uma distribuição de **k** freqüências observadas, f1, f2, f3...fk, e uma distribuição de **k** freqüências teóricas, φ1, φ2, φ3 ... φk. Queremos verificar se a distribuição de f segue a lei teórica φ. Em outros termos, queremos verificar se as divergências entre f e φ podem ser atribuídas unicamente à flutuação da amostragem (hipótese nula).

A fórmula para o cálculo do χ^2 é bastante simples:

$$x^2 = \sum_1^k \frac{(f-\varphi)^2}{\varphi}$$

Vejamos alguns exemplos de aplicação dessa fórmula.

EXEMPLOS DE APLICAÇÃO DO QUI-QUADRADO

Exemplo 1: ajustamento à Lei Normal

PROVA DE ADERÊNCIA

É, como diz o nome, para ver se a distribuição observada adere ou se ajusta à distribuição teórica.

Este primeiro exemplo é uma **PROVA DE ADERÊNCIA,** para ver se seus dados estão distribuídos de acordo com a Lei Normal.

Você sabe o quanto é importante que dados observados, oriundos de amostragens, estejam distribuídos de acordo com a lei normal (curva em sino) para poder fazer estimativas e aplicar testes de hipóteses (teste t, por exemplo). Desse modo, é essencial que você verifique inicialmente se seus dados seguem a Lei Normal. Para isso, você deve calcular as freqüências teóricas, e comparar suas freqüências observadas com essas freqüências teóricas através do cálculo χ^2, como você vai ver no exercício a seguir.

Imagine que você tenha medido o tamanho de uma amostra de 100 indivíduos. Os tamanhos se distribuem em 11 classes de 1 em 1 cm até 11 cm, da maneira seguinte (veja coluna f da **Tabela 20.1**)

Tabela 20.1

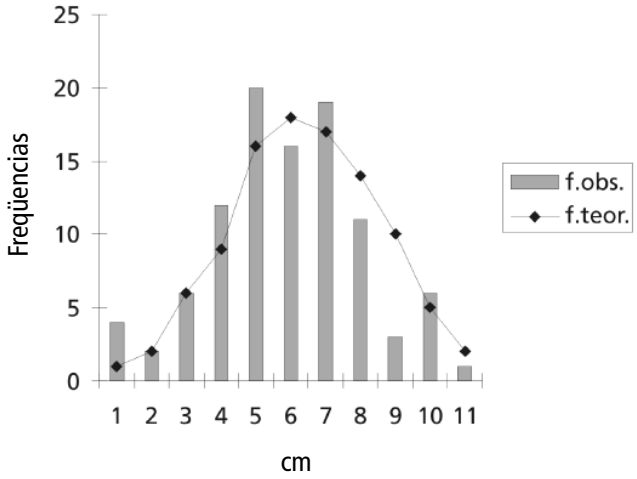
Classes (X) cm	f	φ
1	4	1
2	2	2
3	6	6
4	12	9
5	20	16
6	16	18
7	19	17
8	11	14
9	3	10
10	6	5
11	1	2
Total	100	100

Na terceira coluna da tabela foram colocadas as freqüências teóricas, já calculadas para você, a partir da **Tabela 20.1**.

A sua pergunta é: posso dizer que meus dados de tamanho estão distribuídos de acordo com a Lei Normal?

Vamos dar uma olhada, primeiro, no gráfico das distribuições de freqüências observadas e de freqüências teóricas:

Figura 10.1



Observe que a curva teórica é bastante parecida com as freqüências observadas (em barra), mas que existem algumas discrepâncias. Será que essa diferença entre freqüências é grande demais para ser somente devida ao acaso? Ou seja, devo rejeitar a hipótese nula (igualdade entre distribuições)?

Você vai, então, aplicar a prova de aderência calculando o. χ^2.

Mas antes, ATENÇÃO! O cálculo do . χ^2 tem uma exigência que você deve respeitar: as freqüências teóricas não devem ser inferiores a 5.

!

Os estatísticos mostram que se j for inferior a 5, o valor do Qui-quadrado é superestimado, o que quase sempre leva você a rejeitar erradamente a hipótese nula.

Observe que as classes de 1 cm, 2 cm e 11 cm estão com j <5. O que você deve fazer? Simplesmente, reunir as classes 1, 2 e 3 cm em uma única classe, bem como as classes 10 e 11 cm. Assim você terá a nova tabela a seguir:

Tabela 20.2

Classes (X) cm	f.	φ
1-3	12	9
4	12	9
5	20	16
6	16	18
7	19	17
8	11	14
9	3	10
10-11	7	7
Total	100	100

Agora sim, você pode fazer o cálculo do Qui-quadrado. Veja a seguir:

$$x^2 = \frac{(12-9)^2}{9} + \frac{(12-9)^2}{9} + \frac{(20-16)^2}{16} + \frac{(7-7)^2}{7} = 9,0$$

E agora, o que você pode dizer desse valor 9,0? Você deve testar sua significância, comparando-o com o valor limite que está na Tabela 4 em anexo. Vamos ver como se faz isso.

LIMITES DE SIGNIFICÂNCIA

Os valores do χ^2 variam entre 0 e $+\infty$. Existe uma distribuição de χ^2 diferente para cada valor de *k*.

Para cada valor de k, Pearson calculou a probabilidade de ter χ^2 superior ou inferior a um determinado limite. Para entrar na Tabela 4, você deve escolher um nível de probabilidade (coluna 0,05 por exemplo), e o grau de liberdade (linha). Qual é esse grau de liberdade?

!

Lembre que ***k*** representa o número de freqüências, isto é, o número de linhas da sua tabela, ok!

!

Você está constatando que sempre aparece esse famoso *grau de liberdade*. Se você esqueceu o seu significado, volte à Aula 11.

GRAU DE LIBERDADE

Existe uma curva de distribuição do χ^2 para cada valor de k, mas certos valores não são independentes. O grau de liberdade χ^2 é igual a k menos o número de relações entre os dados. O número de relações varia de acordo com o problema e deve ser calculado a cada caso. Em alguns casos, já sabemos calcular.

MENDEL

Você já teve aulas de Genética? Então deve saber quem era Gregor J, Mendel, o monge e naturalista austríaco (1822-1884) que estabeleceu as leis básicas da hereditariedade das características biológicas.

Caso 1

Para comparar as freqüências observadas com a Lei de **MENDEL**, só precisamos conhecer o número de indivíduos utilizados, ou seja, a soma das freqüências. Logo, só temos uma relação, e assim χ^2 =k-1.

Caso 2

Para comparar com a Lei Normal, precisamos da soma, da média e do desvio padrão, para calcular as freqüências teóricas. São 3 relações, logo χ^2 = k-3.

Vamos, então, entrar na tabela do χ^2 escolhendo a coluna 0,05 (para 5% de probabilidade de erro) e a linha χ^2 = k-3 = 8-3 = 5 (como se trata de comparação com a Lei Normal, vimos que tem 3 graus de liberdade a menos). O valor limite é 11,07.

Como 9,0 < 11,07, não podemos rejeitar a hipótese nula (f = φ). Não há diferença significativa entre a distribuição de tamanho dos indivíduos e a curva da Lei Normal. Podemos concluir que a distribuição de tamanho dos indivíduos coletados segue a Lei Normal.

Exemplo 2: ajustamento sobre duas classes (k=2).

Seja uma amostra com N=100 plantas em que se encontram 20 plantas masculinas e 80 femininas. Sabendo, pela literatura, que a proporção de plantas masculinas desta população é normalmente de 30%, podemos dizer que nossa amostra é conforme à população da literatura?

Vamos calcular o χ^2 para comparar essas duas distribuições de freqüências:

	f.	φ.	χ^2
Masculinas	20	30	$(20-30)^2/30=3,33$
Femininas	80	70	100/70 =1,43
			χ^2 **= 4,76**

No caso de 2 freqüências apenas, k=2, você entra na tabela com χ^2 = k-1 = 1 grau de liberdade (linha 1). Na probabilidade de p=0,05, a tabela do χ^2 dá 3,84 como valor limite. Como 4,76>3,84, podemos rejeitar a hipótese nula: a amostra não é conforme à teoria. A percentagem de plantas masculinas é significativamente menor.

Exemplo 3: teste de independência entre duas distribuições de freqüências observadas.

Seja uma amostra A com 240 plantas e uma amostra B com 130 plantas. Contando as plantas masculinas e femininas, obtivemos os seguintes resultados.

	Masculinas	Femininas	Total
Amostra A	91	149	240
Amostra B	59	71	130
Total	150	220	370

Nós formulamos a hipótese nula **Ho**: não há diferença entre as duas amostras em termos de proporção sexual. Podemos rejeitar esta hipótese? Há uma diferença significativa entre as duas amostras, em termos de proporção sexual?

Você percebe que, nesse exemplo, não temos freqüências teóricas. Como podemos então aplicar um Qui-quadrado a esses dados?

É simples! Vamos deduzir as freqüências teóricas das freqüências observadas, fazendo o seguinte raciocínio. Preste atenção! Num total de 370 plantas, temos 150 masculinas. Respeitando essa proporção, deveríamos ter, teoricamente, na amostra A: (150 x 240)/370 = 97,3 masculinas e 240 – 97,3 = 142,7 femininas. Da mesma maneira, para a amostra B, deveríamos ter, teoricamente, (150 x 130)/370 = 52,7 masculinas e 130 – 52,7 = 77,3 femininas. Pronto, conseguimos nossas freqüências teóricas (aquelas esperadas em caso de independência entre sexo e amostra) que colocamos na **Tabela 20.3.**

Tabela 20.3: Tabela das freqüências teóricas.

	Masc.	Fem.	Total
Amostra A	97,3	142,7	240
Amostra B	52,7	77,3	130
Total	150	220	370

e calculamos o χ^2 =

$$x^2 = \frac{(91-97,3)^2}{97,3} + \frac{(149-142,7)^2}{142,7} + \frac{(59-52,7)^2}{52,7} + \frac{(71-77,3)^2}{77,3} = 1,95$$

No caso de uma tabela com a colunas e b linhas, o grau de liberdade é:

$\chi^2 = (a\text{-}1)(b\text{-}1) = 1$. Pela tabela do Qui-quadrado, para 0,05 e $\chi^2 = 1$, temos $\chi^2 = 3,84$. Como conseqüências, não podemos rejeitar a hipótese nula. Não há diferença significativa entre as duas amostras, em termos de proporção sexual.

O mesmo procedimento pode ser aplicado para mais de 2 amostras e mais de 2 caracteres, como no exemplo a seguir:

Exemplo 4: teste de independência entre 2 caracteres.

Sejam 3 espécies parasitas A1, A2, A3 sobre 3 espécies hospedeiras B1, B2, B3. Para cada hospedeiro conta-se o número de parasitas de cada espécie. Os resultados estão na tabela seguinte:

	A1	A2	A3	Total
B1	130 (120)	160 (150)	10 (30)	300
B2	180 (160)	180 (200)	40 (40)	400
B3	90 (120)	160 (150)	50 (30)	300
Total	400	500	100	1000

Queremos verificar se existe uma "associação" ou uma "independência" entre esses dois caracteres, ou seja, se determinadas espécies parasitam preferencialmente determinados hospedeiros.

! Diz-se que 2 caracteres são "associados" quando a ocorrência de um está ligada à ocorrência do outro. Caso contrário, eles são "independentes".

Seguindo o mesmo procedimento que no exemplo anterior para calcular as freqüências teóricas (entre parênteses na tabela), calculamos χ^2= 40,2. Esse valor é significativo?

Par χ^2= (a-1)(b-1)= (3-1)(3-1)=4 graus de liberdade, a tabela dá um valor limite de χ^2 = 13,28 a 0.01. Logo, podemos rejeitar a hipótese de independência. Certas espécies parasitam preferencialmente certas hospedeiras.

EXIGÊNCIAS DO TESTE

Você deve ter reparado que o uso do teste do Qui-quadrado tem algumas exigências que devem ser respeitadas:

a) as freqüências teóricas não podem ser muito pequenas. Se $\varphi < 5$, você deve agrupar classes, como fizemos no exemplo 1;

b) devem ser utilizadas as freqüências brutas, e não as relativas;

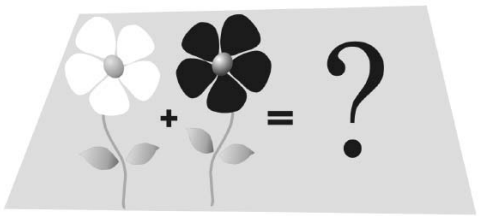
c) no caso de ter somente duas freqüências (k=2, ou tabela de 2 colunas e 2 linhas), se as freqüências forem baixas (f<5 ou total < 50), o χ^2 é geralmente superestimado. Para contornar esse obstáculo, é preciso aplicar uma correção, chamada *correção de Yates*, que consiste em subtrair 0,5 da diferença entre freqüências. Veja a fórmula doχ^2 com a correção de Yates:

$$x^2 = \sum_1^k \frac{\left[|f - \varphi| - 0,5\right]^2}{\varphi}$$

RESUMO

Você aprendeu a aplicar o teste do Qui-quadrado para comparar uma distribuição observada e uma distribuição teórica (teste de aderência), ou duas distribuições observadas (teste de independência). Verificou que a hipótese nula de igualdade entre distribuições é testada com os valores limites da tabela e o grau de liberdade varia caso a caso. Observamos que uma correção é necessária quando as freqüências teóricas são pequenas (<5).

EXERCÍCIOS



1. Numa experiência de cruzamento de flores de cor rosa, obtém-se 22 flores de cor branca, 20 de cor vermelha e 58 de cor rosa. De acordo com as leis de Mendel, as proporções deveriam ser de, respectivamente, 25, 25 e 50%. Podemos dizer que os resultados do cruzamento estão de acordo com a lei, aos níveis de significância de 0,01 e 0,05 ?

2. Estuda-se a resposta de um crustáceo a uma excitação luminosa. Foram testados 20 indivíduos. Foi constatado que 75% deles têm uma atividade à luz. Aplicar o teste do Qui-quadrado para verificar se a luz tem um efeito significativo sobre a atividade deste crustáceo. Consideraremos que, em caso de independência ao fator luz, a resposta seria de 50%.

3. Para conhecer a importância relativa de diversas doenças em relação ao meio social foi feito uma pesquisa que deu os seguintes resultados:

Doenças	Meios sociais		
	A	B	C
Tuberculose	70	79	38
Alcoolismo	24	14	15
Obesidade	27	50	40
Doença cardíaca	14	13	18
Doença venérea	19	35	26
Câncer	24	31	23

Há independência ou associação entre o meio social e os tipos de doenças?

Regressão e correlação

AULA 21

objetivos

Ao final desta aula, você deverá ser capaz de:

- Analisar e representar graficamente uma série estatística dupla.
- Calcular a reta de regressão e o coeficiente de correlação linear.

INTRODUÇÃO

Nas aulas passadas, você aprendeu a aplicar os métodos estatísticos a uma série de dados independentes. Você viu como representar uma distribuição de freqüências, estimar e comparar uma ou mais médias e comparar duas distribuições. Nesta aula, você vai ver uma abordagem um pouco diferente da Estatística. Vamos considerar, não mais uma única série de dados, mas sim duas séries de dados medidos simultaneamente em cada amostra, ou seja, vamos analisar uma série estatística dupla.

Nas áreas da Biologia e da Ecologia, somos freqüentemente levados a estudar o comportamento conjunto de duas séries de dados, ou seja da possível relação entre duas variáveis. Por exemplo, o aumento do peso de um animal em função da idade ou do tamanho, a relação entre o metabolismo de um organismo e a temperatura, da fotossíntese em função da luz etc. Para esse tipo de estudo é preciso obter uma série estatística dupla, chamada também de ***variável bidimensional***, em que cada amostra é descrita por duas variáveis entre as quais queremos verificar a existência de uma relação e quantificar a intensidade dessa relação.

REPRESENTAÇÃO GRÁFICA DE UMA SÉRIE ESTATÍSTICA DUPLA

A melhor maneira de visualizar uma possível relação entre duas variáveis Y e X é, num primeiro passo, elaborar um gráfico chamado *diagrama de dispersão*, onde Y é plotado em função de X num sistema de eixos perpendiculares. Cada par de dados YX é representado por um ponto. Constitui-se, assim, uma "nuvem" de pontos cuja forma é reveladora da relação entre Y e X. Você vai entender melhor observando os gráficos dos exemplos seguintes.

Exemplo 1: Observe a **Figura 21.1**. A forma da nuvem de pontos parece revelar uma certa relação entre Y e X. Há uma tendência de aumento de Y quando X aumenta. Podemos prever o que se chama de *correlação linear positiva*.

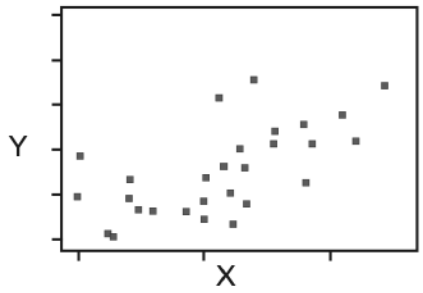


Figura 21.1

Na **Figura 21.2** os pontos são alinhados. Há uma perfeita correlação linear positiva.

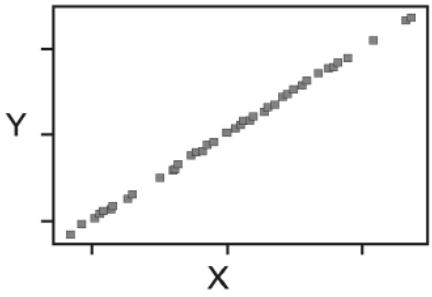


Figura 21.2

Exemplo 2: Neste exemplo, Y diminui quando X aumenta. Diz-se que há uma *correlação linear negativa*.

Exemplo 3: Os pontos estão distribuídos dentro de um círculo. Não existe relação entre Y e X, nenhuma tendência de aumento ou diminuição de Y quando X varia. A correlação entre Y e X é nula:

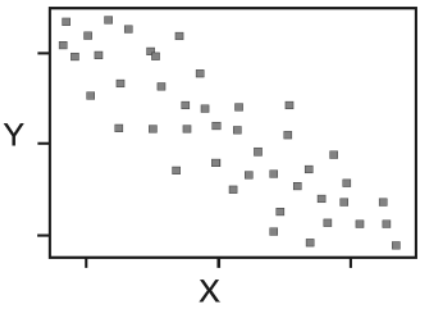


Figura 21.3

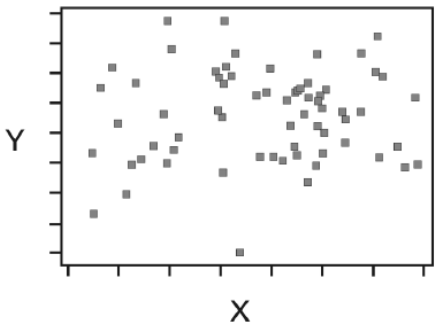


Figura 21.4

Nestas figuras, onde cada ponto representa um par de dados Y-X, você já pode perceber se existe ou não uma certa relação de dependência entre as duas variáveis. Vamos estudar melhor estas relações, e ver alguns exemplos de aplicações.

NOÇÃO DE RETA DE REGRESSÃO

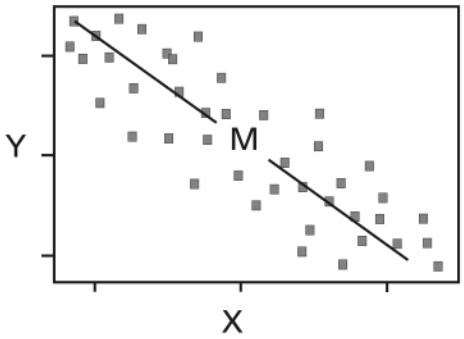


Figura 21.5

Veja a **Figura 21.5.** Ela é igual à figura do exemplo 2, acrescentada de uma linha traçada entre todos os pontos e passando pelo ponto médio M de Y e X. Essa linha imaginária é chamada de *reta de regressão*. Ela representa a relação matemática entre Y e X. Sua equação é $Y = aX + b$.

!

Lembra o que representam os coeficientes ***a*** e ***b*** da equação da reta, que você deve ter visto em aula de Matemática? *a* é o coeficiente angular, também chamado de coeficiente de regressão em estatística; ele representa a inclinação da reta. *b* é o coeficiente linear; é o valor onde a reta corta o eixo Y. Ele indica a posição da reta.

É chamado de resíduo, a soma dos quadrados das distâncias d entre os pontos e a reta, $\sum d^2$. A reta de regressão é traçada de tal maneira que essa soma seja mínima.

Essa equação permite estimar Y a partir de um dado valor de X. Em razão da dispersão dos pontos em volta da reta, o valor de Y não será exatamente conhecido, mas sim, estimado com uma certa probabilidade de erro. Esse erro depende de quanto os pontos estão afastados da reta. Y é exatamente conhecido unicamente quando os pontos são perfeitamente alinhados, isto é, quando o *resíduo* é igual a zero.

Para traçar essa reta devemos então calcular os parâmetros *a* e *b* da equação cujas fórmulas são:

$a = \mathrm{Cov}_{YX} / v_X$

onde, $Cov_{yx} = \dfrac{\sum (y-\bar{y})(x-\bar{x})}{n} = \dfrac{\sum yx - \dfrac{\sum y \sum x}{n}}{n}$ é

chamado de covariância e v_X é a variância de x

$$b_{yx} = \bar{y} - a\bar{x}.$$

Veja, no exemplo a seguir, como você deve desenvolver os cálculos de uma reta de regressão.

Exemplo

Sejam duas variáveis aleatórias Y e X medidas simultaneamente em *n*=10 amostras. Vamos, em primeiro lugar, fazer o gráfico de dispersão de Y (ordenadas) em função de X (abscissas) e verificar visualmente se aparece alguma relação linear entre Y e X. Em caso afirmativo, vamos calcular a equação da reta de regressão e traçar essa reta.

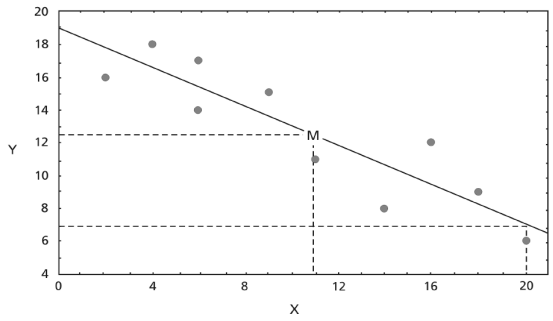


Figura 21.6

Observe que a distribuição dos pontos revela uma certa relação linear negativa entre Y e X. Vamos então traçar a reta de regressão e, para facilitar, elaborar a tabela seguinte que vai permitir fazer os cálculos preliminares que servirão para calcular a equação da reta. Vejamos:

!

Não se justifica traçar a reta se não há nenhuma relação linear aparente entre as duas variáveis. Você vai aprender mais adiante como verificar se essa relação existe com o cálculo do coeficiente de correlação linear *r* de Pearson. Mas não se apresse, vamos chegar lá!

	Y	X	Y^2	X^2	YX
	16	2	256	4	32
	18	4	324	16	72
	17	6	289	36	102
	14	6	196	36	84
	15	9	225	81	135
	11	11	121	121	121
	8	14	64	196	112
	12	16	144	256	192
	9	18	81	324	162
	6	20	36	400	120
Total	126	106	1736	1470	1132
Média	12,6	10,6			
Variância (v)	14,84	34,64			
D. padrão (s)	3,85	5,89			

Covariância:

$$Cov_{yx} = \frac{\sum (y-\bar{y})(x-\bar{x})}{n} = \frac{\sum yx - \frac{\sum y \sum x}{n}}{n} = -20{,}34$$

Cálculo da reta de regressão:

$$y = a_{yx}X + b_{yx} \Rightarrow a_{yx} = \frac{-20,34}{34,64} = -0,59$$

e $\quad b_{yx} = 12,6 - (-0,59 \,.\, 10,6) = 18,9$

Logo, a equação da reta de regressão é: **Y = -0,59 X + 18,9.**

E agora, como você vai traçar a reta? Você sabe que ela passa pelo ponto médio M cujas coordenadas são Y = 12,6 e X = 10,6. Então basta você calcular um segundo ponto para traçá-la. Para isso, você deve atribuir um valor qualquer a X e calcular Y pela equação. Por exemplo, se X=20, temos pela equação Y=7,1.

Chegou o momento de conferir se podemos afirmar, dentro de uma probabilidade de errar de, por exemplo, p =0,05, que essa relação linear entre Y e X existe realmente.

Como vamos fazer isso? Calculando o coeficiente de correlação linear *r* de Pearson.

KARL PEARSON (1857-1936)

Desenvolveu esse coeficiente, mas foi Bravais que lançou o conceito de correlação. De tal forma que esse coeficiente é também chamado de Bravais-Pearson.

O COEFICIENTE DE CORRELAÇÃO LINEAR

Definição e cálculo

Chamado *r* de Pearson, o coeficiente de correlação linear é a imagem da dispersão dos pontos em relação à reta de regressão. Ele mede a intensidade da relação linear entre duas variáveis e varia entre -1 e +1.

Quando *r* = +1, a dispersão é nula, todos os pontos estão perfeitamente alinhados, e Y é diretamente proporcional a X.

Quando *r* = -1, Y é inversamente proporcional a X.

Quando *r* = 0, a dispersão é máxima e não existe correlação linear entre Y e X. As duas variáveis são consideradas independentes.

Observe a fórmula de cálculo de *r* e vamos aplicá-la ao exemplo anterior:

$$r = \frac{Cov_{yx}}{s_y . s_x}$$

No exemplo anterior, temos: $r = \frac{-20,36}{3,85 * 5,89} = -0,898$.

O coeficiente é negativo e parece bastante elevado, próximo de 1, indicando uma forte correlação negativa. Porém, considerando a hipótese nula r = 0, podemos afirmar, com uma probabilidade de erro inferior a 0,05, que -0,898 é estatisticamente diferente de zero e, por conseqüência, que existe uma correlação linear negativa significativa entre Y e X?

• Este teste de significância é realizado utilizando a tabela dos valores limites que *r* deve ultrapassar para serem significativamente diferente de zero. Esses limites dependem da probabilidade escolhida (0,05 ou 0,01) e do grau de liberdade n-2.

• No nosso exemplo, *r* = -0,898 é altamente significativo, pois superior ao valor limite da tabela a 0,01 que é de 0,765 para y = n-2 = 8 graus de liberdade.

Em conclusão, podemos afirmar que existe uma correlação linear negativa altamente significativa entre as variáveis Y e X, na probabilidade p < 0,01.

ALGUMAS RECOMENDAÇÕES

O coeficiente de correlação linear, *r* de Pearson, é um dos mais utilizados pelo biólogos e ecólogos. Seu uso adequado exige, entretanto, certos cuidados:

• Ele expressa exclusivamente a intensidade da relação **linear** entre duas variáveis. Ele varia, em valor absoluto, entre 0 (nenhuma relação linear) e 1 (relação linear perfeita, seja direta para *r* = +1, seja inversa para *r* = -1).

• A ausência de significância (baixo valor de *r*) revela somente a ausência de relação linear entre duas variáveis, podendo ser não-linear. É sempre aconselhado traçar o diagrama de dispersão dos pontos e, assim, visualizar a existência de uma possível relação não-linear. Por exemplo, na **Figura 21.7**, os pontos não são distribuídos ao longo de uma linha, mas de uma curva que parece uma parábola. Nesse caso, o cálculo de *r* não pode ser feito com os dados brutos, mas somente após aplicar uma transformação que lineariza a relação. A transformação sugerida aqui seria da forma $y=ax^2+bx+c$, que é a equação da parábola. Outros tipos de transformações são possíveis. Por exemplo, quando os pontos seguem uma exponencial. A transformação indicada seria log(x).

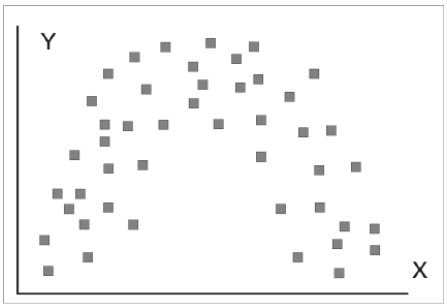


Figura 21.7

• O teste de significância de *r* não pode ser aplicado se as distribuições das variáveis não forem normais, isto é, seguindo a lei de Gauss. Em certos casos, é possível tentar uma transformação que normalize os dados, por exemplo log(x); nesse caso concluímos que Y é uma função logarítmica de x, e a equação da reta de regressão será: Y = a. Log(x) + b.

• A existência de uma correlação significativa não representa obrigatoriamente uma relação de causa/efeito, pois é possível encontrar correlação entre quaisquer variáveis, sem significado biológico ou ecológico.

RESUMO

Você aprendeu a verificar se existe uma relação linear entre duas variáveis, através do cálculo do coeficiente de correlação linear de Pearson *r* , que varia entre +1 e -1. A hipótese nula ($r = 0$) é testada com os valores limites da tabela de *r.* Se *r* for significativo podemos traçar a reta de regressão que representa a relação matemática entre as duas variáveis, e permite estimar uma em função da outra. Esse método tem muitas aplicações em Biologia e Ecologia, mas deve ser usado com cuidado, respeitando a exigência de normalidade dos dados.

EXERCÍCIOS

1. Abaixo estão relacionados os comprimentos X e as larguras Y de 10 folhas tiradas ao acaso de determinada árvore. Há correlação entre as duas características?

Folha Nº	Comprimento (X)	Largura (Y)
1	12	10
2	12	14
3	11	9
4	16	13
5	13	10
6	12	12
7	10	8
8	9	7
9	17	13
10	15	14

2. A taxa de conversão alimentar (Y) varia com o peso (X) de um animal. Essas duas variáveis foram medidas numa amostra de 6 animais (tabela abaixo). Verificar se existe uma correlação linear significativa entre essas duas variáveis e traçar a reta de regressão de Y em função de X.

Peso (X)	Taxa de conversão (Y)
35	3,8
40	3,4
45	3,2
50	2,8
55	2,6
60	2,3

3. Um estudo da diluição de compostos nitrogenados, rejeitados no mar por uma usina, foi realizado medindo de hora em hora a salinidade (S) e os teores em nitrogênio (N):

Horas	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
S	11	12	13	10	10	9	13	9	14	10	12	10	15	14	12	15
N	20	17	14	21	19	24	15	?	17	22	16	20	15	16	19	15

a) existe uma relação linear significativa entre o nitrogênio e a salinidade?

b) na oitava hora faltou o dado de nitrogênio. Estimar este valor a partir da salinidade.

Anexo

TABELA 1

Tabela de Gauss: probabilidades p_x de um valor x estar incluído no intervalo entre $-\infty$ e Z.

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

Uso da tabela:

a) entrar com valor de Z utilizando a linha (para a dezena) e a coluna (para o centésimo). Por exemplo: a probabilidade para Z=1,96 será lida na interseção da linha 1,9 e da coluna 0,06.

b) para valores negativos de Z, fazer 1 – px. Por exemplo, se Z = –1,96 temos px = 1 – 0,9750 = 0,0250.

c) para encontrar os valores de 2 x (probabilidade entre dois valores simétricos em relação à média), fazer 2(px - 0,5) se Z > 0 e 2(0,5 - px) se Z < 0.

TABELA 2

Valores limites do *t* de Student para níveis de significância de 0,05 e 0,01.

gl	.05	.01
1	12.706	63.657
2	4.303	9.925
3	3.182	5.841
4	2.776	4.604
5	2.571	4.032
6	2.447	3.707
7	2.365	3.499
8	2.306	3.355
9	2.262	3.250
10	2.228	3.169
11	2.201	3.106
12	2.179	3.055
13	2.160	3.012
14	2.145	2.977
15	2.131	2.947
16	2.120	2.921
17	2.110	2.898
18	2.101	2.878
19	2.093	2.861
20	2.086	2.845
21	2.080	2.831
22	2.074	2.819
23	2.069	2.807
24	2.064	2.797
25	2.060	2.787
26	2.056	2.779
27	2.052	2.771
28	2.048	2.763
29	2.045	2.756
30	2.042	2.750
40	2.021	2.704
60	2.000	2.660
120	1.980	2.617
∝	1.960	2.576

TABELA 3

Valores críticos de F ao nível de significância de 0,05 para ν_1 (numerador) e ν_2 (denominador) graus de liberdade.

ν_1	1	2	3	4	5	6	7	8	9	10
ν_2										
1	161.40	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.39	19.40
3	10.130	9.552	9.277	9.117	9.013	8.941	8.887	8.845	8.812	8.786
4	7.709	6.944	6.591	6.388	6.256	6.163	6.094	6.041	5.999	5.964
5	6.608	5.786	5.409	5.192	5.050	4.950	4.876	4.818	4.772	4.735
6	5.987	5.143	4.757	4.534	4.387	4.284	4.207	4.147	4.099	4.060
7	5.591	4.737	4.347	4.120	3.972	3.866	3.787	3.726	3.677	3.637
8	5.318	4.459	4.066	3.838	3.687	3.581	3.500	3.438	3.388	3.347
9	5.117	4.256	3.863	3.633	3.482	3.374	3.293	3.230	3.179	3.137
10	4.965	4.103	3.708	3.478	3.326	3.217	3.135	3.072	3.020	2.978
11	4.844	3.982	3.587	3.357	3.204	3.095	3.012	2.948	2.896	2.854
12	4.747	3.885	3.490	3.259	3.106	2.996	2.913	2.849	2.796	2.753
13	4.667	3.806	3.411	3.179	3.025	2.915	2.832	2.767	2.714	2.671
14	4.600	3.739	3.344	3.112	2.958	2.848	2.764	2.699	2.646	2.602
15	4.543	3.682	3.287	3.056	2.901	2.790	2.707	2.641	2.588	2.544
16	4.494	3.634	3.239	3.007	2.852	2.741	2.657	2.591	2.538	2.494
17	4.451	3.592	3.197	2.965	2.810	2.699	2.614	2.548	2.494	2.450
18	4.414	3.555	3.160	2.928	2.773	2.661	2.577	2.510	2.456	2.412
19	4.381	3.522	3.127	2.895	2.740	2.628	2.544	2.477	2.423	2.378
20	4.351	3.493	3.098	2.866	2.711	2.599	2.514	2.447	2.393	2.348
21	4.325	3.467	3.072	2.840	2.685	2.573	2.488	2.420	2.366	2.321
22	4.301	3.443	3.049	2.817	2.661	2.549	2.464	2.397	2.342	2.297
23	4.279	3.422	3.028	2.796	2.640	2.528	2.442	2.375	2.320	2.275
24	4.260	3.403	3.009	2.776	2.621	2.508	2.423	2.355	2.300	2.255
25	4.242	3.385	2.991	2.759	2.603	2.490	2.405	2.337	2.282	2.236
26	4.225	3.369	2.975	2.743	2.587	2.474	2.388	2.321	2.265	2.220
27	4.210	3.354	2.960	2.728	2.572	2.459	2.373	2.305	2.250	2.204
28	4.196	3.340	2.947	2.714	2.558	2.445	2.359	2.291	2.236	2.190
29	4.183	3.328	2.934	2.701	2.545	2.432	2.346	2.278	2.223	2.177
30	4.171	3.316	2.922	2.690	2.534	2.421	2.334	2.266	2.211	2.165
50	4.034	3.183	2.790	2.557	2.400	2.286	2.199	2.130	2.073	2.026
100	3.936	3.087	2.696	2.463	2.305	2.191	2.103	2.032	1.975	1.927

Tabela 3 (continuação)

ν_1	11	12	13	14	15	16	17	18	19	20
ν_2										
1	243.0	243.9	244.7	245.4	245.9	246.5	246.9	247.3	247.7	248.01
2	19.41	19.41	19.42	19.42	19.43	19.43	19.44	19.44	19.44	19.45
3	8.763	8.745	8.729	8.715	8.703	8.692	8.683	8.675	8.667	8.660
4	5.936	5.912	5.891	5.873	5.858	5.844	5.832	5.821	5.811	5.803
5	4.704	4.678	4.655	4.636	4.619	4.604	4.590	4.579	4.568	4.558
6	4.027	4.000	3.976	3.956	3.938	3.922	3.908	3.896	3.884	3.874
7	3.603	3.575	3.550	3.529	3.511	3.494	3.480	3.467	3.455	3.445
8	3.313	3.284	3.259	3.237	3.218	3.202	3.187	3.173	3.161	3.150
9	3.102	3.073	3.048	3.025	3.006	2.989	2.974	2.960	2.948	2.936
10	2.943	2.913	2.887	2.865	2.845	2.828	2.812	2.798	2.785	2.774
11	2.818	2.788	2.761	2.739	2.719	2.701	2.685	2.671	2.658	2.646
12	2.717	2.687	2.660	2.637	2.617	2.599	2.583	2.568	2.555	2.544
13	2.635	2.604	2.577	2.554	2.533	2.515	2.499	2.484	2.471	2.459
14	2.565	2.534	2.507	2.484	2.463	2.445	2.428	2.413	2.400	2.388
15	2.507	2.475	2.448	2.424	2.403	2.385	2.368	2.353	2.340	2.328
16	2.456	2.425	2.397	2.373	2.352	2.333	2.317	2.302	2.288	2.276
17	2.413	2.381	2.353	2.329	2.308	2.289	2.272	2.257	2.243	2.230
18	2.374	2.342	2.314	2.290	2.269	2.250	2.233	2.217	2.203	2.191
19	2.340	2.308	2.280	2.256	2.234	2.215	2.198	2.182	2.168	2.155
20	2.310	2.278	2.250	2.225	2.203	2.184	2.167	2.151	2.137	2.124
21	2.283	2.250	2.222	2.197	2.176	2.156	2.139	2.123	2.109	2.096
22	2.259	2.226	2.198	2.173	2.151	2.131	2.114	2.098	2.084	2.071
23	2.236	2.204	2.175	2.150	2.128	2.109	2.091	2.075	2.061	2.048
24	2.216	2.183	2.155	2.130	2.108	2.088	2.070	2.054	2.040	2.027
25	2.198	2.165	2.136	2.111	2.089	2.069	2.051	2.035	2.021	2.007
26	2.181	2.148	2.119	2.094	2.072	2.052	2.034	2.018	2.003	1.990
27	2.166	2.132	2.103	2.078	2.056	2.036	2.018	2.002	1.987	1.974
28	2.151	2.118	2.089	2.064	2.041	2.021	2.003	1.987	1.972	1.959
29	2.138	2.104	2.075	2.050	2.027	2.007	1.989	1.973	1.958	1.945
30	2.126	2.092	2.063	2.037	2.015	1.995	1.976	1.960	1.945	1.932
50	1.986	1.952	1.921	1.895	1.871	1.850	1.831	1.814	1.798	1.784
100	1.886	1.850	1.819	1.792	1.768	1.746	1.726	1.708	1.691	1.676

Tabela 3 (continuação). Nível de significância: 0,01

ν_1	1	2	3	4	5	6	7	8	9	10
ν_2										
1	4052.	4999.	5403.	5624.	5763.	5859.	5928.	5981.	6022.	6055.
2	98.50	99.00	99.16	99.24	99.30	99.33	99.35	99.37	99.38	99.39
3	34.11	30.81	29.45	28.71	28.23	27.91	27.67	27.48	27.34	27.22
4	21.19	18.00	16.69	15.97	15.52	15.20	14.97	14.79	14.659	14.54
5	16.25	13.27	12.06	11.39	10.96	10.67	10.45	10.28	10.15	10.05
6	13.74	10.92	9.780	9.148	8.746	8.466	8.260	8.102	7.976	7.874
7	12.24	9.547	8.451	7.847	7.460	7.191	6.993	6.840	6.719	6.620
8	11.25	8.649	7.591	7.006	6.632	6.371	6.178	6.029	5.911	5.814
9	10.56	8.022	6.992	6.422	6.057	5.802	5.613	5.467	5.351	5.257
10	10.04	7.559	6.552	5.994	5.636	5.386	5.200	5.057	4.942	4.849
11	9.646	7.206	6.217	5.668	5.316	5.069	4.886	4.744	4.632	4.539
12	9.330	6.927	5.953	5.412	5.064	4.821	4.640	4.499	4.388	4.296
13	9.074	6.701	5.739	5.205	4.862	4.620	4.441	4.302	4.191	4.100
14	8.862	6.515	5.564	5.035	4.695	4.456	4.278	4.140	4.030	3.939
15	8.683	6.359	5.417	4.893	4.556	4.318	4.142	4.004	3.895	3.805
16	8.531	6.226	5.292	4.773	4.437	4.202	4.026	3.890	3.780	3.691
17	8.400	6.112	5.185	4.669	4.336	4.102	3.927	3.791	3.682	3.593
18	8.285	6.013	5.092	4.579	4.248	4.015	3.841	3.705	3.597	3.508
19	8.185	5.926	5.010	4.500	4.171	3.939	3.765	3.631	3.523	3.434
20	8.096	5.849	4.938	4.431	4.103	3.871	3.699	3.564	3.457	3.368
21	8.017	5.780	4.874	4.369	4.042	3.812	3.640	3.506	3.398	3.310
22	7.945	5.719	4.817	4.313	3.988	3.758	3.587	3.453	3.346	3.258
23	7.881	5.664	4.765	4.264	3.939	3.710	3.539	3.406	3.299	3.211
24	7.823	5.614	4.718	4.218	3.895	3.667	3.496	3.363	3.256	3.168
25	7.770	5.568	4.675	4.177	3.855	3.627	3.457	3.324	3.217	3.129
26	7.721	5.526	4.637	4.140	3.818	3.591	3.421	3.288	3.182	3.094
27	7.677	5.488	4.601	4.106	3.785	3.558	3.388	3.256	3.149	3.062
28	7.636	5.453	4.568	4.074	3.754	3.528	3.358	3.226	3.120	3.032
29	7.598	5.420	4.538	4.045	3.725	3.499	3.330	3.198	3.092	3.005
30	7.562	5.390	4.510	4.018	3.699	3.473	3.305	3.173	3.067	2.979
50	7.171	5.057	4.199	3.720	3.408	3.186	3.020	2.890	2.785	2.698
100	6.895	4.824	3.984	3.513	3.206	2.988	2.823	2.694	2.590	2.503

Tabela 3 (continuação). Nível de significância: 0,01

ν_1	11	12	13	14	15	16	17	18	19	20
ν_2										
1.	6083	. 6106.	6125.	6142.	6157.	6170.	6181.	6191.	6200.	6208.
2.	99.40	99.41	99.42	99.42	99.43	99.43	99.44	99.44	99.44	99.44
3.	27.13	27.05	26.98	26.92	26.87	26.82	26.78	26.75	26.71	26.69
4.	14.45	14.37	14.30	14.24	14.19	14.15	14.11	14.08	14.04	14.02
5.	9.963	9.888	9.825	9.770	9.722	9.680	9.643	9.610	9.580	9.553
6.	7.790	7.718	7.657	7.605	7.559	7.519	7.483	7.451	7.422	7.396
7.	6.538	6.469	6.410	6.359	6.314	6.275	6.240	6.209	6.181	6.155
8.	5.734	5.667	5.609	5.559	5.515	5.477	5.442	5.412	5.384	5.359
9.	5.178	5.111	5.055	5.005	4.962	4.924	4.890	4.860	4.833	4.808
10.	4.772	4.706	4.650	4.601	4.558	4.520	4.487	4.457	4.430	4.405
11.	4.462	4.397	4.342	4.293	4.251	4.213	4.180	4.150	4.123	4.099
12.	4.220	4.155	4.100	4.052	4.010	3.972	3.939	3.909	3.883	3.858
13.	4.025	3.960	3.905	3.857	3.815	3.778	3.745	3.716	3.689	3.665
14.	3.864	3.800	3.745	3.698	3.656	3.619	3.586	3.556	3.529	3.505
15.	3.730	3.666	3.612	3.564	3.522	3.485	3.452	3.423	3.396	3.372
16.	3.616	3.553	3.498	3.451	3.409	3.372	3.339	3.310	3.283	3.259
17.	3.519	3.455	3.401	3.353	3.312	3.275	3.242	3.212	3.186	3.162
18.	3.434	3.371	3.316	3.269	3.227	3.190	3.158	3.128	3.101	3.077
19.	3.360	3.297	3.242	3.195	3.153	3.116	3.084	3.054	3.027	3.003
20.	3.294	3.231	3.177	3.130	3.088	3.051	3.018	2.989	2.962	2.938
21.	3.236	3.173	3.119	3.072	3.030	2.993	2.960	2.931	2.904	2.880
22.	3.184	3.121	3.067	3.019	2.978	2.941	2.908	2.879	2.852	2.827
23.	3.137	3.074	3.020	2.973	2.931	2.894	2.861	2.832	2.805	2.781
24.	3.094	3.032	2.977	2.930	2.889	2.852	2.819	2.789	2.762	2.738
25.	3.056	2.993	2.939	2.892	2.850	2.813	2.780	2.751	2.724	2.699
26.	3.021	2.958	2.904	2.857	2.815	2.778	2.745	2.715	2.688	2.664
27.	2.988	2.926	2.871	2.824	2.783	2.746	2.713	2.683	2.656	2.632
28.	2.959	2.896	2.842	2.795	2.753	2.716	2.683	2.653	2.626	2.602
29.	2.931	2.868	2.814	2.767	2.726	2.689	2.656	2.626	2.599	2.574
30.	2.906	2.843	2.789	2.742	2.700	2.663	2.630	2.600	2.573	2.549
50.	2.625	2.562	2.508	2.461	2.419	2.382	2.348	2.318	2.290	2.265
100.	2.430	2.368	2.313	2.265	2.223	2.185	2.151	2.120	2.092	2.067

Tabela 4. Valores limites do Qui-quadrado em função do grau de liberdade (GL) e do nível de significância.

G.L.	.99	.98	.95	.90	.80	.70	.50	.30	.20	.10	**.05**	.02	**.01**	.001
1	.00016	.00063	.0039	.016	.064	.15	.46	1.07	1.64	2.71	**3.84**	5.41	**6.64**	10.83
2	.02	.04	.10	.21	.45	.71	1.39	2.41	3.22	4.60	**5.99**	7.82	**9.21**	13.82
3	.12	.18	.35	.58	1.00	1.42	2.37	3.66	4.64	6.25	**7.82**	9.84	**11.34**	16.27
4	.30	.43	.71	1.06	1.65	2.20	3.36	4.88	5.99	7.78	**9.49**	11.67	**13.28**	18.46
5	.55	.75	1.14	1.61	2.34	3.00	4.35	6.06	7.29	9.24	**11.07**	13.39	**15.09**	20.52
6	.87	1.13	1.64	2.20	3.07	3.83	5.35	7.23	8.56	10.64	**12.59**	15.03	**16.81**	22.46
7	1.24	1.56	2.17	2.83	3.82	4.67	6.35	8.38	9.80	12.02	**14.07**	16.62	**18.48**	24.32
8	1.65	2.03	2.73	3.49	4.59	5.53	7.34	9.52	11.03	13.36	**15.51**	18.17	**20.09**	26.12
9	2.09	2.53	3.32	4.17	5.38	6.39	8.34	10.66	12.24	14.68	**16.92**	19.68	**21.67**	27.88
10	2.56	3.06	3.94	4.86	6.18	7.27	9.34	11.78	13.44	15.99	**18.31**	21.16	**23.21**	29.59
11	3.05	3.61	4.58	5.58	6.99	8.15	10.34	12.90	14.63	17.28	**19.68**	22.62	**24.72**	31.26
12	3.57	4.18	5.23	6.30	7.81	9.03	11.34	14.01	15.81	18.55	**21.03**	24.05	**26.22**	32.91
13	4.11	4.76	5.89	7.04	8.63	9.93	12.34	15.12	16.98	19.81	**22.36**	25.47	**27.69**	34.53
14	4.66	5.37	6.57	7.79	9.47	10.82	13.34	16.22	18.15	21.06	**23.68**	26.87	**29.14**	36.12
15	5.23	5.98	7.26	8.55	10.31	11.72	14.34	17.32	19.31	22.31	**25.00**	28.26	**30.58**	37.70
16	5.81	6.61	7.96	9.31	11.15	12.62	15.34	18.42	20.46	23.54	**26.30**	29.63	**32.00**	39.29
17	6.41	7.26	8.67	10.08	12.00	13.53	16.34	19.51	21.62	24.77	**27.59**	31.00	**33.41**	40.75
18	7.02	7.91	9.39	10.86	12.86	14.44	17.34	20.60	22.76	25.99	**28.87**	32.35	**34.80**	42.31
19	7.63	8.57	10.12	11.65	13.72	15.35	18.34	21.69	23.90	27.20	**30.14**	33.69	**36.19**	43.82
20	8.26	9.24	10.85	12.44	14.58	16.27	19.34	22.78	25.04	28.41	**31.41**	35.02	**37.57**	45.32
21	8.90	9.92	11.59	13.24	15.44	17.18	20.34	23.86	26.17	29.62	**32.67**	36.34	**38.93**	46.80
22	9.54	10.60	12.34	14.04	16.31	18.10	21.24	24.94	27.30	30.81	**33.92**	37.66	**40.29**	48.27
23	10.20	11.29	13.09	14.85	17.19	19.02	22.34	26.02	28.43	32.01	**35.17**	38.97	**41.64**	49.73
24	10.86	11.99	13.85	15.66	18.06	19.94	23.34	27.10	29.55	33.20	**36.42**	40.27	**42.98**	51.18
25	11.52	12.70	14.61	16.47	18.94	20.87	24.34	28.17	30.68	34.38	**37.65**	41.57	**44.31**	52.62
26	12.20	13.41	15.38	17.29	19.82	21.79	25.34	29.25	31.80	35.56	**38.88**	42.86	**45.64**	54.05
27	12.88	14.12	16.15	18.11	20.70	22.72	26.34	30.32	32.91	36.74	**40.11**	44.14	**46.96**	55.48
28	13.56	14.85	16.93	18.94	21.59	23.65	27.34	31.39	34.03	37.92	**41.34**	45.42	**48.28**	56.89
29	14.26	15.57	17.71	19.77	22.48	24.58	28.34	32.46	35.14	39.09	**42.56**	46.69	**49.59**	58.30
30	14.95	16.31	18.49	20.60	23.36	25.51	29.34	33.53	36.25	40.26	**43.77**	47.96	**50.89**	59.70

TABELA 5

Valores críticos do *r* de Pearson aos níveis de significância de 0,05 e 0,01 para n – 2 graus de liberdade.

GL	0,05	0,01
1	0.997	0,999
2	0,950	0,990
3	0,878	0,959
4	0,811	0,917
5	0,755	0,875
6	0,707	0,834
7	0,666	0,798
8	0.632	0.765
9	0.602	0.735
10	0.576	0.708
11	0.553	0.684
12	0.532	0.661
13	0.514	0.641
14	0.497	0.623
15	0.482	0.606
16	0.468	0.590
17	0.456	0.575
18	0.444	0.561
19	0.433	0.549
20	0.423	0.537
25	0.381	0.487
30	0.349	0.449
35	0.325	0.418
40	0.304	0.393
45	0.288	0.372
50	0.273	0.354
60	0.250	0.325
70	0.232	0.302
80	0.217	0.283
90	0.205	0.267

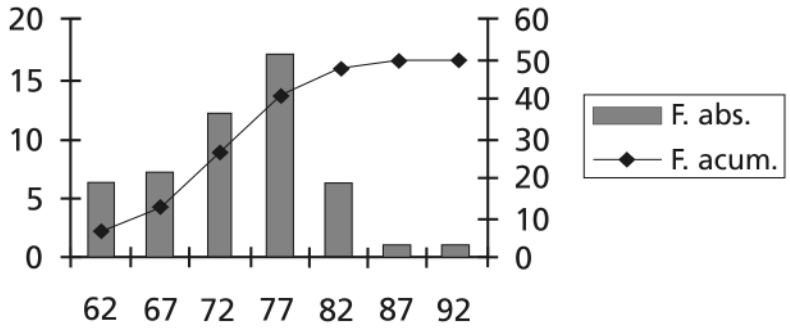
Elementos de Matemática e Estatística

Gabarito

1,a) tabela

Classes (mm)	Ponto médio X	Freq. absolutas	Freq. relativas	Freq. abs. acumuladas
60-64	62	6	0.12	6
65-69	67	7	0.14	13
70-74	72	12	0.24	25
75-79	77	17	0.34	42
80-84	82	6	0.12	48
85-89	87	1	0.02	50
90-94	92	1	0.02	50
	TOTAL	50	1.00	

b) histograma



2. Média: 5,29

   Desvio padrão: 2,93

   Coeficiente de variação: 0,55

   Obs.: O desvio padrão foi calculado com n-1 graus de liberdade em razão do pequeno número de dados.

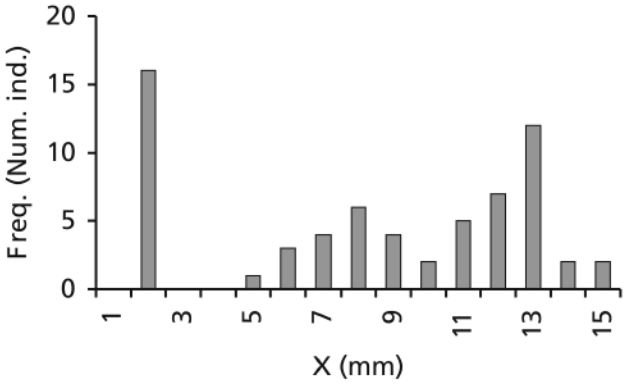
3.

Espécie	Média	Desvio padrão	Coef. variação
1	0.8	1.03	1.29
2	3.8	2.39	0.63
3	3.8	4.54	1.19

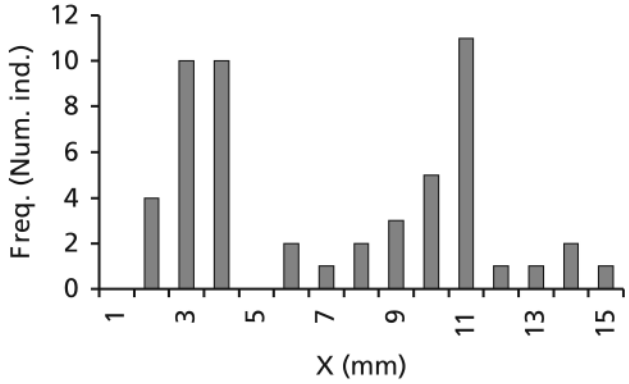
4.

Classes (mm)	Ponto médio X	Freq. absolutas	
		Dia 05/01	Dia 05/06
0.5-1.4	1	0	0
1.5-2.4	2	16	4
2.5-3.4	3	0	10
3.5-4.4	4	0	10
4.5-5.4	5	1	0
5.5-6.4	6	3	2
6.5-7.4	7	4	1
7.5-8.4	8	6	2
8.5-9.4	9	4	3
9.5-10.4	10	2	5
10.5-11.4	11	5	11
11.5-12.4	12	7	1
12.5-13.4	13	12	1
13.5-14.4	14	2	2
14.5-15.4	15	2	1

Histograma do dia 05/01



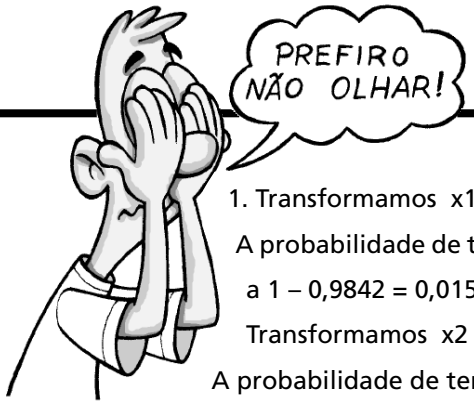
Histograma do dia 05/06



A distribuição de freqüência é **multimodal**. O cálculo da média geral não faz sentido, pois a série de medidas revela a presença de várias populações que correspondem a gerações sucessivas. O pesquisador deverá calcular o crescimento de cada uma separadamente. Para isso ele poderá:

• considerar o deslocamento da moda. Por exemplo, a primeira tem moda 2 mm em 05/01 e moda 3,5 mm em 05/06, 5 meses depois, dando um crescimento de cerca de 1 mm em 5 meses (0,20 mm/mês). A segunda geração tem moda 8 mm em 05/01 e 11 mm em 05/06. Uma terceira aparece com moda 13 mm em 05/01 e 14 mm em 05/06, ou

• reunir as classes de comprimento de cada geração e calcular a média de cada uma.



1. Transformamos x1 =28,55 em z1 = (28,55 – 21,65)/3,21 = 2,15.
A probabilidade de ter x até x1 é igual a 0,9842, e acima de x1 é igual a 1 – 0,9842 = 0,0158
Transformamos x2 = 14,75 em z2 = (14,75 – 21,65)/3,21 = -2,15
A probabilidade de ter x até x2 é também igual a 1 – 0,9842 = 0,0158
Logo, a probabilidade de tirar um individuo com valor abaixo de 14,75 ou acima de 28,55 é: 0,0158 + 0,0158 = 0,0316 (isto é, há 3,16 % de chance)

2. Transformar x1 = 15 em z1 = (15 – 18,75)/6,25 = – 0,60
Transformar x2 = 25 em z2 = (25 – 18,75)/6,25 = 1,00
px1 = 1 – 0,7257 = 0,2743
px2 = 0,8643
Probabilidade entre px1 e px2 é p = 0,8643 – 0,2743 = 0,59
Há 59% de chance de que as chuvas dos próximos anos estejam entre 15 e 25mm.

3. a) Para x1 = 14, temos z1 = (14 – 12)/1,5 = 1,333 e p_{x1} = 0,9082 e 1-p_{x1} = 0,0918.
Cerca de 9,2 % dos indivíduos têm tamanho maior que 14cm;
para x2 = 8,5, temos z2 = -2,33 e p_{x2} = 1 – 0,9901 = 0,0099 (probabilidade até 8,5cm). Logo, acima de 8,5cm a probabilidade é 1 – 0,0099 = 0,9901.
0,9082 – 0,0099 = 0,8983. Cerca de 89,8% dos indivíduos estão entre 8,5 e 14cm de tamanho.

b) O problema é inverso ao anterior. Conhecemos a probabilidade 1- p_x = 1 – 0,10 = 0,90 correspondendo a *p*x. Procuramos esse valor na tabela Normal. Encontramos o valor mais próximo (0,8997) na intersecção da linha 1,2 e da coluna 0,08, que corresponde a Z = 1,28. Sabemos que Z = (x – m)/s, ou seja, x = Z.s + m = 1,28.1,5 + 12 = 13,9cm. Logo, concluímos que 10% dos indivíduos têm tamanho superior a 13,9cm.

c) Utilizamos a propriedade da distribuição normal, segundo a qual adicionando e subtraindo 1,96.*s* à média de uma amostra, obtemos um intervalo dentro do qual encontram-se 95% dos dados. Logo 95% dos indivíduos têm tamanho entre 12 – 1,96.1,5 = 9,1cm e 12 + 1,96.1,5 = 14,9cm.

1.

$$0{,}8 \pm 2{,}23.\frac{1{,}03}{\sqrt{10}} \Rightarrow Ic = \langle 0{,}07 \ - \ 1{,}53 \rangle$$

$$3{,}8 \pm 2{,}23.\frac{2{,}39}{\sqrt{10}} \Rightarrow Ic = \langle 2{,}09 \ - \ 5{,}51 \rangle$$

$$3{,}8 \pm 2{,}23.\frac{4{,}54}{\sqrt{10}} \Rightarrow Ic = \langle 0{,}60 \ - \ 7{,}00 \rangle$$

2.

	ÁREA 1		ÁREA 2
Nº de estações	20		20
Média	58,05		29,95
Variância	2318		1152,8
Diferença das médias		28,10	
Erro padrão		13,51	
Teste t		2,08	
t a 0,05		2,09	

O teste *t* é praticamente igual ao valor de *t* da tabela de Student. Podemos dizer que as duas médias são, significativamente, diferentes ao nível de probabilidade de 0,05

3.

	Regime rico		**Regime pobre**
Nº de peixes	12		7
Média	120		101
Variância	457,45		425,33
Diferença das médias		19	
Erro padrão		10,61	
Teste t		1,79	
t a 0,05		2,11	

O teste *t* é inferior ao *t* de Student. Rejeitando a hipótese nula cometeríamos um erro do tipo I maior do que 5%. Nas condições experimentais utilizadas não podemos dizer que o enriquecimento do alimento em proteínas proporcionou um aumento de peso.

1. Teste *t*:

   d = 0,429

   S_d = 3,90

   t = 0,11 (diferença não significativa a p = 0,05)

Análise de variância – Resultados

Fonte de variação	Dispersão	G. de L.	Variância	F calculado	F a 0,05
FATOR	0,64	1	0,64	0,014	4,75
RESÍDUO	548,58	12	45,71		

Não podemos rejeitar a hipótese nula. Não foi possível verificar, nas condições da experiência, a influência do fator "variedade" sobre a idade da desova.

2.

Fonte de variação	Dispersão	G. de L.	Variância	F calculado	F a 0,01
FATOR	907,5	4	226,87	62,8	3,78
RESÍDUO	162,6	45	3,61		

Existe uma influência altamente significativa ($p < 0,01$) do fator "Detergente" sobre o crescimento do mexilhão. Pelo cálculo da Menor Diferença Significativa, $MDS = 2.\sqrt{3,61\frac{5}{10}} = 2,69$ verificamos que todos os detergentes têm efeito significativo, porém, o de laboratório (DT3) tem ação significativamente menor que os comerciais, entre os quais não há diferença.

1. $\chi 2 = 2,64$, inferior ao valor crítico da tabela de $\chi 2$ que é de 5,99 a p=0,05 para 3-1=2 graus de liberdade. Logo, rejeitando a hipótese nula, faríamos um erro superior a 5%. Consideramos os resultados coerentes com a lei de Mendel.

2.

	LUZ	ESCURO
Freq. observada	15	5
Freq. teórica	10	10

Como k = 2 e N = 20 (<50), aplica-se a correção de Yates

$\chi 2 = 4,05$

Para (2-1)(2-1) = 1 grau de liberdade, temos $\chi 2 = 3,84$ (p=0,05) e $\chi 2 = 6,63$ (p = 0,01). Rejeita-se a hipótese nula ao nível de 0,05, mas não ao nível de 0,01. Em outros termos, aceitando cometer um erro de 5%, podemos dizer que a luz tem influência sobre o comportamento dos crustáceos. Porém, para um erro de somente 1%, não é possível concluir sobre esta influência.

3. $\chi 2 = 22,17$

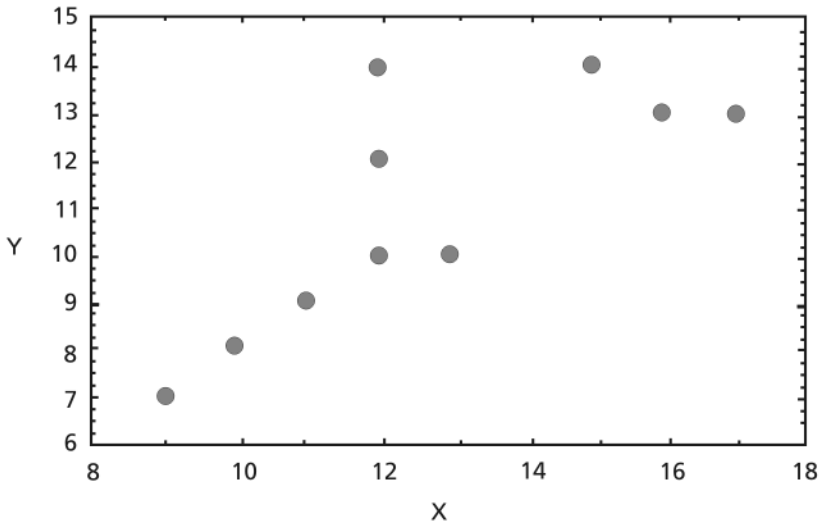
   Para (3-1)(6-1) = 10 graus de liberdade, temos $\chi 2 = 18,31$ a 0,05 e $\chi 2 = 23,20$ a 0,01

   Significativo a p=0,05, mas não a p=0,01

   Para uma probabilidade de erro de 0,05, podemos dizer que existe uma certa associação entre as doenças e o meio social. Porém, para um erro de somente 1% não é possível verificar essa associação.

Aula 21

1. Verificamos se pode existir uma relação entre Y e X pelo diagrama de dispersão.



Parece haver uma correlação linear. Devemos calcular o coeficiente de correlação *r* de Pearson entre Y e X, e testar sua significância, isto é, se podemos rejeitar a hipótese nula $r = 0$.

Folha Nº.	Comp. (X)	X^2	Largura (Y)	Y^2	XY
1	12	144	10	100	120
2	12	144	14	196	168
3	11	121	9	81	99
4	16	256	13	169	208
5	13	169	10	100	130
6	12	144	12	144	144
7	10	100	8	64	80
8	9	81	7	49	63
9	17	289	13	169	221
10	15	225	14	196	210
TOTAL	127	1673	110	1268	1443

Var (x) = $(1673 - 127^2/10)/9 = 6,68$ s (x) = 2,58

Var (y) = $(1268 - 110^2/10)/9 = 6,44$ s (y) = 2,54

Cov (xy) = 1443 – 127 x 110/10)/9 = 5,11

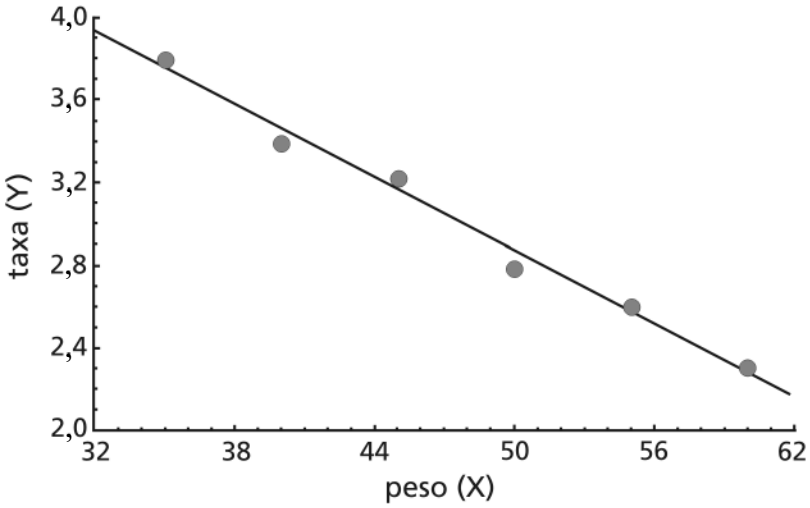
r = 5,11 / (2,58 x 2,54) = 0,780

Para n-2 = 8 graus de liberdade, $r_{95} = 0,632$ e $r_{99} = 0,735$

O coeficiente de correlação linear é significativamente diferente de zero ($p<0,01$). Podemos dizer que existe uma correlação linear entre o comprimento e a largura das folhas.

2. $r = -0,996$

   Para 4 graus de liberdade, $r_{99} = 0,917$. Logo, existe uma correlação linear negativa altamente significativa ($p<0,01$) entre o peso e a taxa de conversão alimentar



A equação da reta de regressão é: Y = -0,0589 X + 5,812

3. r = - 0,868 (significativo a p < 0,01).
   Equação da reta N = - 1,31 S + 33,7.
   Para S = 9, o valor de N é estimado em 21,9 pela equação acima.
   Atenção: nos cálculos de r e da reta, não devem ser considerados os dados da oitava hora em razão da falta de dado de nitrogênio. Logo, fazer n = 15.

ISBN 85-7648-033-6

UENF
Universidade Estadual
do Norte Fluminense

SECRETARIA DE
CIÊNCIA E TECNOLOGIA

Ministério
da Educação