ESTATÍSTICA

Romeu Magnani

Marisa Veiga Capela

INSTITUTO DE QUÍMICA

UNESP 1999 **ARARAQUARA**

I. ESTATÍSTICA DESCRITIVA

1. INTRODUÇÃO

A *Estatística Descritiva* trata da maneira de apresentar um conjunto de dados em tabelas ou gráficos e do modo de resumir as informações contidas nesses dados, através de certas medidas como média, variância, desvio padrão, coeficiente de variação, etc.

2. TIPOS DE VARIÁVEIS

Algumas variáveis são *qualitativas* e outras *quantitativas*. Uma variável qualitativa pode ser apenas um nome (variável qualitativa *nominal*) ou estabelecer uma ordem (variável qualitativa *ordinal*). As variáveis quantitativas, mais importantes neste curso, são classificadas em *discreta* (se referem em geral a contagens) ou *contínua* (podem assumir qualquer valor de um intervalo de números reais).

Exemplo 1: Na tabela abaixo são apresentados 60 valores de cada uma de 6 variáveis, que representam informações sobre alunos do sexo masculino cursando graduação em Química, em determinado ano (classifique essas variáveis conforme o tipo)

No. do aluno	No. de irmãos	Altura	Peso	Idade	Origem*	Grau de instrução do pai
1	2	1,71	70,9	18	AR	2o. grau
2	3	1,72	76,2	20	AR	2o. grau
3	2	1,69	72,6	18	OL	Superior
4	1	1,62	60,0	22	CP	2o. grau
5	3	1,77	71,3	19	CP	2o. grau
6	0	1,55	53,6	19	OL	2o. grau
7	0	1,66	65,8	20	AR	2o. grau
8	5	1,63	65,0	19	OL	2o. grau
9	3	1,73	87,8	19	OL	Superior
10	5	1,70	73,8	22	AR	Superior
11	4	1,82	81,3	20	OL	2o. grau
12	3	1,73	72,2	19	OL	Superior
13	2	1,80	74,7	24	AR	2o. grau
14	3	1,77	73,4	19	OL	2o. grau
15	2	1,73	69,1	21	OL	2o. grau
16	3	1,71	98,1	21	AR	2o. grau
17	2	1,74	71,2	18	OL	Superior
18	2	1,71	67,3	19	OE	2o. grau
19	3	1,74	69,0	21	AR	Superior
20	3	1,71	79,7	18	OL	2o. grau
21	2	1,88	85,7	18	OL	2o. grau
22	3	1,76	83,4	19	CP	Superior
23	2	1,62	64,0	20	OL	Superior
24	1	1,67	72,1	23	AR	Superior
25	3	1,64	63,5	19	CP	Superior
26	2	1,77	69,2	19	OE	1o. grau
27	2	1,73	76,8	23	OL	Superior
28	1	1,80	91,2	20	OL	2o. grau
29	2	1,73	64,8	21	OE	Nenhum
30	2	1,66	68,2	19	OL	Superior
31	2	1,79	82,5	20	OL	Superior
32	3	1,80	105,7	20	AR	1o. grau

No. do aluno	No. de irmãos	Altura	Peso	Idade	Origem*	Grau de instrução do pai
33	3	1,63	61,8	21	OL	2o. grau
34	2	1,77	79,4	20	OL	2o. grau
35	1	1,86	87,2	19	AR	Superior
36	0	1,66	59,9	25	OL	2o. grau
37	1	1,82	82,2	20	OL	2o. grau
38	6	1,85	79,2	21	AR	2o. grau
39	2	1,69	69,4	22	CP	Superior
40	3	1,58	62,0	22	OL	1o. grau
41	3	1,77	80,6	18	CP	Superior
42	0	1,76	70,4	19	OL	Superior
43	4	1,67	65,9	18	OL	Superior
44	4	1,75	74,9	21	CP	1o. grau
45	1	1,80	83,4	18	OL	2o. grau
46	2	1,71	77,4	18	OL	Superior
47	3	1,78	78,6	19	OL	Superior
48	2	1,70	78,6	24	CP	2o. grau
49	1	1,75	81,9	22	CP	2o. grau
50	3	1,75	74,0	21	AR	2o. grau
51	1	1,81	77,2	23	AR	Superior
52	4	1,71	70,0	22	CP	2o. grau
53	2	1,74	79,0	18	AR	Superior
54	1	1,78	83,4	21	OL	2o. grau
55	5	1,89	92,2	21	CP	Superior
56	2	1,82	94,6	20	AR	2o. grau
57	0	1,76	67,1	20	OL	2o. grau
58	4	1,76	72,0	19	CP	Superior
59	2	1,64	65,2	20	OL	2o. grau
60	0	1,65	71,7	18	OL	1o. grau

*AR: Araraquara e região (até 50km) CP: Capital
OL: Outros Locais do Estado OE: Outros Estados

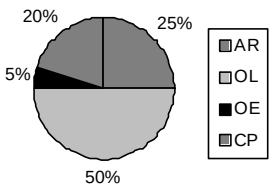
3. DISTRIBUIÇÃO DE FREQÜÊNCIAS

Muitas vezes, obtém-se informações relevantes sobre uma variável através de sua distribuição de freqüências. Esta é uma tabela contendo valores distintos da variável e as freqüências correspondentes. A freqüência pode ser *absoluta* (nº de vezes que o valor aparece no conjunto de dados) ou *relativa* (nº de vezes que o valor aparece dividido pelo total de valores) ou *percentua*l (a freqüência relativa multiplicada por 100). Pode ser útil também o gráfico da distribuição. Os gráficos recomendados dependem do tipo de variável.

No caso das variáveis quantitativas, em especial a variável contínua, são observadas as freqüências em intervalos de valores, em vez de freqüências individuais. Para variável quantitativa é de grande importância a *distribuição de freqüências acumuladas*. Uma f*reqüência acumulada* é a soma das freqüências até determinado valor (ou intervalo de valores)

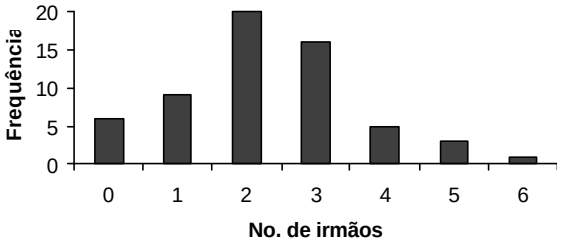
Exemplo 2: Distribuições de freqüências da variável *origem* do exemplo 1 e gráfico em *pizza*.

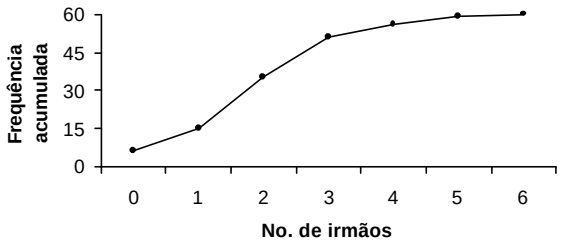
Origem	Freqüência	Freq. Relativa	Freq. Percentual
AR	15	0,25	25%
OL	30	0,50	50%
OE	3	0,05	5%
CP	12	0,20	20%
Total	60	1,00	100%



Exemplo 3: Distribuições de freqüências da variável discreta *número de irmãos* da tabela do exemplo 1, gráfico de freqüências e gráfico de freqüências acumuladas.

Nº de irmãos	Freqüência	Freqüência acumulada	Freqüência relativa	Freq. relativa acumulada
0	6	6	0,100	0,100
1	9	15	0,150	0,250
2	20	35	0,333	0,583
3	16	51	0,267	0,850
4	5	56	0,083	0,933
5	3	59	0,050	0,983
6	1	60	0,017	1,000
Total	60		1,000	





Observação: Os gráficos de freqüência absoluta, freqüência relativa e freqüência percentual têm o mesmo aspecto. Isso ocorre porque essas freqüências são proporcionais.

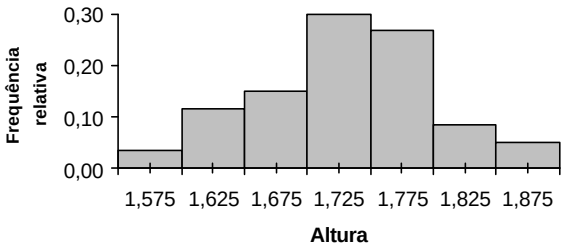
Uma distribuição de freqüências de variável contínua é diferente. A faixa que engloba todos os valores da variável é dividida em diversos intervalos, de preferência de mesma amplitude. A freqüência se refere ao número de valores da variável em cada intervalo. Um critério empregado aqui é o de considerar os intervalos fechados à direita, isto é, incluem o valor da extrema direita e não incluem o valor à esquerda. Ás vezes é conveniente substituir o intervalo pelo seu ponto médio.

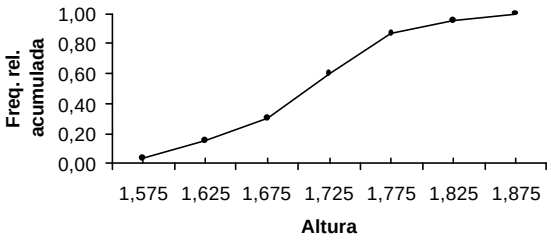
Exemplo 4: As alturas da tabela do exemplo 1, colocadas em ordem crescente, são:

1,55; 1,58; 1,62; 1,62; 1,63; 1,63; 1,64; 1,64; 1,65; 1,66; 1,66; 1,66; 1,67; 1,67; 1,69; 1,69; 1,70; 1,70; 1,71; 1,71; 1,71; 1,71; 1,71; 1,71; 1,72; 1,73; 1,73; 1,73; 1,73; 1,73; 1,74; 1,74; 1,74; 1,75; 1,75; 1,75; 1,76; 1,76; 1,76; 1,76; 1,77; 1,77; 1,77; 1,77; 1,77; 1,78; 1,78; 1,79; 1,80; 1,80; 1,80; 1,80; 1,81; 1,82; 1,82; 1,82; 1,85; 1,86; 1,88; 1,89;

Variação total: 1,89-1,55=0,34 metros. Uma sugestão é usar $\sqrt{60} \cong 7$ ou 8 intervalos. Tomando como variação total 0,35m e adotando 7 intervalos, cada um terá amplitude 0,35/7=0,05 m. A distribuição de freqüências absolutas (simples e acumulada) e a distribuição de freqüências relativas (simples e acumulada) são dadas abaixo, assim como os gráficos das distribuições de freqüências relativas.

Intervalos de alturas	Ponto médio	Freq.	Freq. acum.	Freq. relativa	Freq. relativa acumulada	Densidade de freq. rel.
1,55 —| 1,60	1,575	2	2	0,033	0,033	0,667
1,60 —| 1,65	1,625	7	9	0,117	0,150	2,333
1,65 —| 1,70	1,675	9	18	0,150	0,300	3,000
1,70 —| 1,75	1,725	18	36	0,300	0,600	6,000
1,75 —| 1,80	1,775	16	52	0,267	0,867	5,333
1,80 —| 1,85	1,825	5	57	0,083	0,950	1,667
1,85 —| 1,90	1,875	3	60	0,050	1,000	1,000
Total		60		1,000		

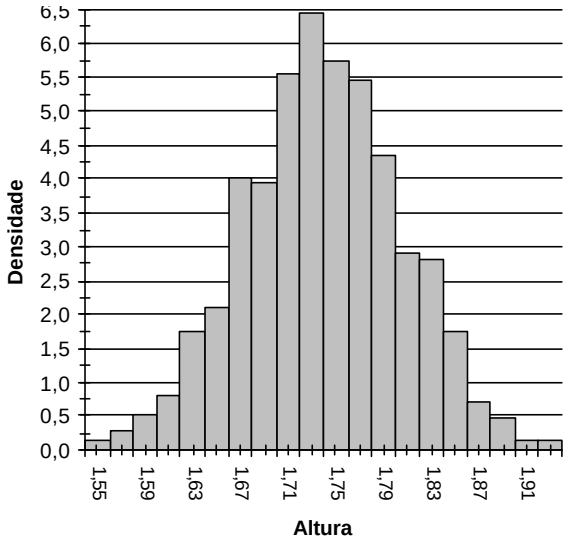




O gráfico em colunas retangulares acima é chamado *Histograma*, enquanto que o gráfico de freqüências acumuladas recebe o nome de *Ogiva de Galton*. No gráfico de freqüências simples, as alturas dos retângulos são proporcionais as alturas dos retângulos do gráfico de freqüências relativas. Portanto, eles têm o mesmo aspecto. Para as freqüências acumuladas também ocorre uma proporcionalidade das alturas.

Na tabela de distribuições de freqüências da variável *altura* foi incluída uma coluna de *densidade de freqüência relativa*. Esta é obtida pela divisão da freqüência relativa pela amplitude do intervalo de alturas correspondente. Desse modo, no *histograma* da *densidade* de freqüência, a área de cada retângulo é igual a freqüência relativa correspondente e a área total é igual a soma das freqüências relativas que é 1. Em termos percentuais, a área de cada retângulo é a porcentagem de alturas no intervalo base do retângulo.

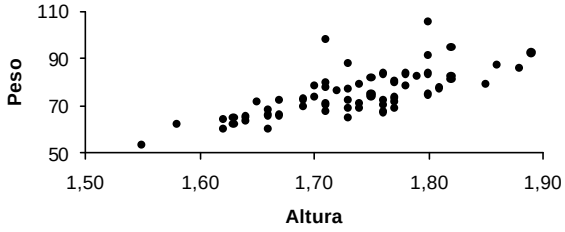
Atenção: A compreensão do conceito de densidade de freqüência relativa é fundamental para o entendimento de tópicos mais avançados de Estatística. Na figura tem-se o histograma da densidade de freqüências relativas das alturas de um grande número de alunos de graduação do sexo masculino. A base de cada retângulo (intervalo de alturas) é igual a 0,02 m e os números indicados representam uma parte dos pontos médios dos intervalos. No eixo vertical estão representadas as densidades de freqüências relativas, cuja unidade é 1/m. Então, a área do retângulo de ponto médio 1,71 é aproximadamente igual a 0,02 x 5,5= 0,11. Em outras palavras, 11% dos alunos têm alturas no intervalo de 1,70 a 1,72 m. No intervalo de 1,72 a 1,78 m estão aproximadamente 35,5% das alturas. Um problema interessante é determinar a altura, tal que, o conjunto de todas as alturas menores do que ela representa 2% do total. A resposta é a altura de aproximadamente 1,60 m.



4. RELAÇÃO ENTRE DUAS VARIÁVEIS

Até aqui as variáveis foram analisadas individualmente. Muitas vezes interessa verificar se há alguma associação entre duas ou mais variáveis. Com apenas duas variáveis pode ser usado o *gráfico de dispersão*.

Exemplo 5: Na figura abaixo está representado o *gráfico de dispersão* das variáveis *altura* e *peso* da *tabela do exemplo 1*. Parece haver uma dependência entre as variáveis, pois conforme a altura aumenta, o peso também aumenta.



4. USANDO O EXCEL

Funções

CONT.SE(matriz*; valor)	Conta o nº de vezes que determinado valor (nº ou não) aparece em uma matriz de dados.
FREQÜÊNCIA(matriz; valores de referência)	Quando o valor de referência é uma célula, dá a Freqüência acumulada. Para a freqüência absoluta é preciso marcar primeiro o intervalo de saída, inserir a função FREQUÊNCIA e pressionar ao mesmo tempo CONTROL+SHIFT+ENTER
MÁXIMO(matriz) MÍNIMO(matriz) CONT.VALORES(matriz)	valor máximo de uma matriz de dados valor mínimo de uma matriz de dados Total de valores numéricos de uma matriz de dados

*conjunto de células de uma planilha dispostos só em linha, só em coluna ou tanto em linha como em coluna.

Ferramentas de análise

HISTOGRAMA	Forma a distribuição de freqüência e constrói o Histograma.

PROBLEMAS:

1) Abra uma *pasta* no Excel e coloque a tabela do exemplo 1 em uma planilha. Em seguida, use as *funções* indicadas acima para resolver os exemplos de 2 a 5.

2) Resolva novamente o exemplo 4 usando a *ferramenta* HISTOGRAMA.

3) Estude as distribuições de freqüências das outras variáveis da tabela do exemplo 1: *peso*, *idade* e *grau de instrução do pai* (neste caso, use o gráfico de colunas agrupadas).

PROBLEMA PROPOSTO

PP1) Considere os dados da tabela abaixo, referentes a 50 estudantes do sexo feminino matriculadas no curso de Química do IQAr em 1998. Construa para cada variável as distribuições de freqüências e os respectivos gráficos. Faça o gráfico de dispersão para o par de variáveis altura e peso. Que conclusões podem ser obtidas se os resultados para as variáveis da tabela do exemplo 1 forem comparados com os obtidos aqui?

Nº	Peso (kg)	Altura (m)	idade (anos)
1	55,6	1,64	20
2	62,0	1,70	22
3	61,0	1,68	23
4	70,0	1,69	21
5	67,0	1,65	23
6	49,0	1,60	22
7	70,0	1,68	23

Nº	Peso (kg)	Altura (m)	idade (anos)
26	53,0	1,65	22
27	63,0	1,72	21
28	70,0	1,78	22
29	48,0	1,59	20
30	51,0	1,59	21
31	85,0	1,73	19
32	57,0	1,65	21

8	63,0	1,64	21
9	60,0	1,71	22
10	52,0	1,65	21
11	58,0	1,70	20
12	50,0	1,62	27
13	55,0	1,65	21
14	57,0	1,67	18
15	50,0	1,56	21
16	70,0	1,59	23
17	48,0	1,60	19
18	70,0	1,70	19
19	54,0	1,61	25
20	48,5	1,55	20
21	52,0	1,70	22
22	42,0	1,58	19
23	67,0	1,62	19
24	58,0	1,68	18
25	57,0	1,66	18

33	65,0	1,60	21
34	48,0	1,65	21
35	60,0	1,68	32
36	64,0	1,58	20
37	49,0	1,60	19
38	65,0	1,70	22
39	57,0	1,67	19
40	55,0	1,55	21
41	54,0	1,65	22
42	57,0	1,80	19
43	45,0	1,60	20
44	62,0	1,70	24
45	89,0	1,65	31
46	50,0	1,70	21
47	51,0	1,60	18
48	48,0	1,62	21
49	53,0	1,64	21
50	73,0	1,74	22

5. MEDIDAS DE POSIÇÃO

As medidas de posição mais conhecidas são: *média*, *mediana* e *moda*. São valores em torno dos quais os dados se distribuem, por isso são conhecidas como medidas de tendência central.

Se uma variável x possui os n valores: x_1, x_2, ..., x_n, a *média* aritmética, que representaremos aqui por *m*, ou *m*(x) quando houver necessidade de identificar a variável x, é

$$m(x) = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

A *mediana*, med, é o valor que ocupa a posição central da série de dados, quando estes são colocados em ordem crescente ou decrescente, e a *moda*, mo, é o valor com maior freqüência. Pode haver mais de uma moda.

Exemplo 6: Se uma variável têm valores iguais a: 10, 15, 18, 22, 22, 30, a média *m*, a mediana med e a moda são, respectivamente, iguais a

$$m = \frac{10 + 15 + 18 + 22 + 22 + 30}{6} = 19{,}5$$

$$med = \frac{18 + 22}{2} = 20 \text{ (pois existem dois valores centrais)}$$

moda = 22

Exemplo 7: Considerando as *alturas* dos alunos na tabela do exemplo 1, tem-se, em metros,

$$m = \frac{1}{60}(1{,}71 + 1{,}72 + 1{,}69 + 1{,}62 + \ldots + 1{,}64 + 1{,}65) = \frac{103{,}95}{60} = 1{,}733$$

med = 1,735

moda = 1,71

Essas medidas de posição podem ser determinadas pela distribuição de freqüências do exemplo 4 tomando o ponto médio dos intervalos. Tem-se:

$$m = \frac{1}{60}(2 \cdot 1{,}575 + 7 \cdot 1{,}625 + 9 \cdot 1{,}675 + 18 \cdot 1{,}725 + 16 \cdot 1{,}775 + 5 \cdot 1{,}825 + 3 \cdot 1{,}875)$$

$$= \frac{103{,}80}{60} = 1{,}730$$

med = 1,725

moda = 1,725

6. MEDIDAS DE DISPERSÃO

As medidas dispersão são valores que mostram o quanto os dados estão dispersos em relação ao centro da distribuição de freqüência (em geral, a média). As principais medidas de dispersão são: *variância* e *desvio padrão*, mas existem outras, tais como: *amplitude total*, *desvio médio* e *coeficiente de variação*.

Se uma variável x possui os n valores: x_1, x_2, ..., x_n, a *variância*, indicada por *Var* ou *Var*(x), é definida por

$$Var(x) = \frac{1}{n}[(x_1 - m)^2 + (x_2 - m)^2 + \cdots + (x_n - m)^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - m)^2$$

Entendendo (x_i - *m*) como o desvio de x_i em relação à média *m*, então a *variância* é a média

desses desvios ao quadrado. O *desvio padrão*, *dp*(x), é a raiz quadrada da variância, isto é,

$$\boxed{dp(x) = \sqrt{Var(x)}}$$

Quanto as outras medidas de dispersão, a *amplitude total* é a diferença entre o maior e o menor valor da série de dados, o *desvio médio* é a média dos desvios tomados sempre como positivos e o *coeficiente de variação*, CV, é o quociente entre o desvio padrão e a média, multiplicado por 100.

$$\boxed{CV = \frac{dp(x)}{\bar{x}} 100\%}$$

Exemplo 8: Considerando os dados do exemplo 6, tem-se

$$\begin{aligned} Var &= \frac{1}{6}[(10-19{,}5)^2 + (15-19{,}5)^2 + (18-19{,}5)^2 + (22-19{,}5)^2 \\ &\qquad + (22-19{,}5)^2 + (30-19{,}5)^2] \\ &= \frac{1}{6}[(-9{,}5)^2 + (-4{,}5)^2 + (-1{,}5)^2 + (-2{,}5)^2 + (-2{,}5)^2 + (10{,}5)^2] \\ &= \frac{235{,}5}{6} = 39{,}25 \end{aligned}$$

Observe que os desvios são iguais a -9,5; -4,5; -1,5; 2,5; 2,5; 10,5 e a soma desses desvios é igual a zero (isso acontece sempre). O valor 235,5 é a *Soma de Quadrados dos Desvios*.

O desvio padrão é igual a $dp = \sqrt{39{,}25} = 6{,}2650$

amplitude total = 30 - 10 = 20

desvio médio = $\text{desvio médio} = \frac{9{,}5 + 4{,}5 + 1{,}5 + 2{,}5 + 2{,}5 + 10{,}5}{6} = 5{,}1667$

coeficiente de variação = CV = $\frac{6{,}2650}{19{,}5} 100 = 32{,}13\%$

Exemplo 9: Para a distribuição de freqüências da variável x = altura do exemplo 4, tem-se:

$$Var = \frac{1}{60}[2 \cdot (1{,}575 - 1{,}730)^2 + 7 \cdot (1{,}625 - 1{,}730)^2 + \cdots + 3.(1{,}875 - 1{,}730)^2]$$

$$Var = \frac{0{,}2935}{60} = 0{,}0049 \text{ m}^2$$

$$\text{Desvio padrão} = \sqrt{0{,}0049} = 0{,}070 \text{ m}$$

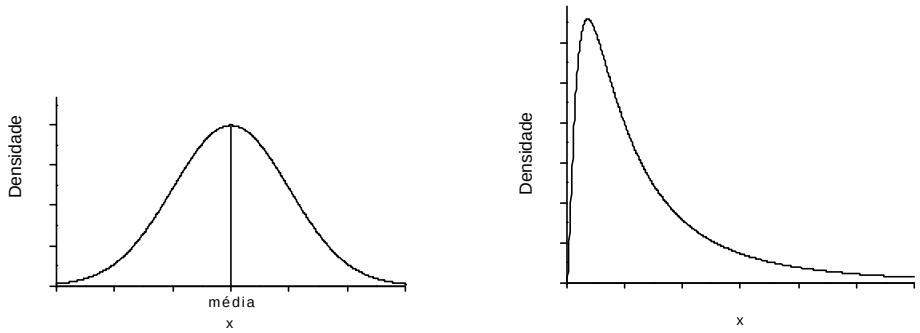
$$CV = \frac{0{,}070}{1{,}730} 100 = 4{,}04\%$$

Amplitude Total= 1,875 – 1,575 = 0,030 m

7. POPULAÇÃO E AMOSTRA

Os métodos estatísticos são próprios para o estudo de *populações*. *População* é um conjunto de dados que descreve algum fenômeno de interesse, ou seja, dados que têm, em comum, determinada característica. *Amostra* é um subconjunto de dados selecionados de uma população. Pretende-se, a partir da amostra, estudar a população. Portanto, uma amostra deve ter as mesmas características que a população de onde foi retirada. Existem procedimentos adequados de *amostragem.*

Considerando uma *população* formada por um conjunto muito grande de valores, é fácil imaginar que o gráfico da densidade de freqüência (ver exemplo 4) poderia ser representado por uma linha contínua como nas figuras abaixo. Em cada uma delas a área abaixo da curva é igual a 1. O gráfico a esquerda é simétrico em torno do eixo que contém a média e representa uma densidade de freqüência teórica, chamada *distribuição normal*, que será estudada adiante.



As medidas de posição e de dispersão, definidas nos itens 5 e 6, são válidas tanto para população como para amostra, mas, para a amostra, a variância e o desvio padrão tem como denominador (n–1) em lugar de n.

Exemplo 10: No exemplo 8, o correto seria $Var = \frac{235{,}5}{5} = 47{,}1000$ e $dp = 6{,}8629$.

Entretanto, no exemplo 9 faz pouca diferença dividir por 60 ou 60 -1=59.

8. MEDIDAS DE ASSIMETRIA E CURTOSE

O c*oeficiente de assimetria* e o *coeficiente de curtose* são medidas relacionadas com a forma da distribuição de freqüência ou da densidade de freqüência. A *assimetria* é uma medida da falta de simetria da distribuição. A *curtose* indica o grau de achatamento de uma densidade de freqüência em relação à distribuição normal citada no item anterior. Nos gráficos acima, o primeiro tem coeficiente de assimetria e coeficiente de curtose iguais a zero (pois trata-se de uma distribuição normal). No outro gráfico, tanto o coeficiente de assimetria como o de curtose são grandes.

Para um conjunto de valores x_i, com i=1,2,...,n, o *coeficiente de assimetria* é definido por

$$\frac{n}{(n-1)(n-2)}\sum\left(\frac{x_i-\bar{x}}{s}\right)^2$$

onde s = dp(x) é o desvio padrão do conjunto x_i considerado como amostra.

O *coeficiente de curtose* é dado por

$$\left[\frac{n(n+1)}{(n-1)(n-2)(n-3)}\sum\left(\frac{x_i-\bar{x}}{s}\right)^4\right]-\frac{3(n-1)^2}{(n-2)(n-3)}$$

9. USANDO O EXCEL

Funções:

MÉDIA(matriz)	Média de um conjunto de dados
MED(matriz)	Mediana
MODO(matriz)	Moda
DESVQ(matriz)	Soma de quadrados dos desvios em relação à média
DESVPAD(matriz)	Desvio padrão amostral
VAR(matriz)	Variância de uma amostra
CURT(matriz)	Coeficiente de curtose
DISTORÇÃO(matriz)	Coeficiente de assimetria
Observação: as funções a seguir se referem a população e usam *n* em vez de *n-1* no denominador.	
VARP(matriz)	Variância de uma população
DESVPADP(matriz)	Desvio padrão populacional

Ferramentas de análise

ESTATÍSTICA DESCRITIVA	Fornece informações sobre a tendência central e dispersão dos dados

PROBLEMAS: Todas as questões a seguir se referem aos dados da tabela do exemplo 1 (considerados como amostra).

4) Determine as medidas de tendência central e de dispersão para a variável *n^{0} de irmãos.* Use as funções apropriadas.

5) Repita o problema anterior para a variável *peso*.

6) Use a *ferramenta ESTATÍSTICA DESCRITIVA* para resolver os problemas 4) e 5)

PROBLEMAS ADICIONAIS:

7) Acione a *ajuda* do Excel para conhecer as funções ALEATÓRIO e ALEATÓRIOENTRE. Use essas funções para sortear 10 alunos da tabela do exemplo 1. Determine a média, variância e desvio padrão das *idades* dos alunos sorteados. Obtenha ajuda sobre a função PROCV e verifique como usá-la para copiar as idades dos alunos sorteados.

PROBLEMAS PROPOSTOS

PP2) Complete o problema proposto 1 com as medidas expostas aqui. Como ficam as conclusões anteriores?

PP3) Procure na literatura um conjunto de dados (mais de 30) de uma variável e faça um estudo usando os procedimentos da Estatística Descritiva. Escreva um pequeno relatório contendo:

- **a)** Do que se trata o conjunto de dados
- **b)** de onde foi tirado
- **c)** Coloque os resultados em tabelas e gráficos de acordo com as normas da ABNT (consulte a Biblioteca)
- **d)** tire conclusões.

COMPLEMENTOS

10. TEOREMA DE CHEBYSHEV (aplicação do desvio padrão)

Dado um número k, maior do que 1, então pelo menos $(1-1/k^2)$ dos valores de uma amostra ou população pertencerão ao intervalo de k desvios padrão antes e k desvios padrão além da média. Este intervalo tem extremos $(m - k \cdot dp)$ e $(m + k \cdot dp)$.

Exemplo 11: Para as alturas da tabela do exemplo 1, obteve-se no exemplos 7 e 9, a média 1,73 e o desvio padrão 0,070, respectivamente. Seja o intervalo 1,73 ± k. 0,070

Pelo teorema de Chebyshev tem-se:
Se k=2, pelo menos 1-1/4 = 3/4 (75%) dos valores estão no intervalo 1,73±2(0,070) (isto é, entre 1,59 m e 1,87 m). Na realidade, este intervalo contém 93,3% das alturas, como pode ser verificado pela tabela do exemplo 1.

Se k=3, pelo menos 1-1/9 = 8/9 (88,9%) das alturas estão no intervalo 1,73±3(0,070) (isto é, entre 1,52 e 1,94). Na realidade este intervalo contém 100% das alturas.

11. MEDIDAS DE ORDENAMENTO

A *mediana* é uma medida de ordem tal que metade das observações são menores que ela. Existem outras *medidas de ordenamento* que podem ser úteis. Para cada uma dessas medidas, uma proporção *p* das observações é menor do que ela. Por exemplo, os *quartis* dividem uma série de dados em quatro partes. Para cada p, entre 0 e 1, é determinado um *percentil*.

Exemplo 11: Seja a série de valores: 45; 33; 40; 36; 31; 49; 37; 30; 48; 38; 43

Série ordenada	30	31	33	36	37	38	40	43	45	48	49
ordem	1	2	3	4	5	6	7	8	9	10	11
ordem porcentual	0,0	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1,0

Tomando, por exemplo, o nº 43, 70% dos valores da série são menores que ele e 30% maiores. O *percentil* de p=0,70 (ou 70%) é 43.

Os quartis são :
1º quartil (ou percentil de 0,25) = 34,5 (25% dos valores são menores do que 34,5)
2º quartil (ou mediana) = 38 (50% dos valores são menores do que 38)
3º quartil (ou percentil de 0,75) = 44 (75% dos valores são menores do que 44)

Funções

ORDEM(nº ; matriz; ordem*)	Posição de um nº em uma matriz de dados
ORDEM.PORCENTUAL(matriz; nº; decimais**)	Posição percentual de um nº
PERCENTIL(matriz; *p*)	o percentil em matriz de dados correspondente a *p* (0<p<1)
QUARTIL(matriz; quartil)	Quartil de uma matriz de dados: 0= 100%, 1=75%; 2=50%; 3=25%; 4=0%.

*vazio ou zero = ordem decrescente, outro nº = ordem crescente
** nº de casas decimais. Vazio = 3 casas decimais

Ferramenta de análise

ORDEM E PERCENTIL	Tabela que contém a ordem percentual e ordinal de cada valor de um intervalo de dados

Exemplo 12: Aplicando a ferramenta ORDEM E PERCENTIL ao conjunto de dados do exemplo 11, sem classificá-los, obtém-se

*Ponto**	*Dados*	*Ordem*	*Porcentagem*
6	**49**	1	100
9	**48**	2	90
1	**45**	3	80
11	**43**	4	70
3	**40**	5	60
10	**38**	6	50
7	**37**	7	40
4	**36**	8	30
2	**33**	9	20
5	**31**	10	10
8	**30**	11	0

* *Ponto* indica a posição de cada elemento da série inicial.

PROBLEMAS:

8) Forme uma série de valores com alguns números repetidos e verifique como ficam as ordens.

9) Determine os quartis para as alturas da tabela do exemplo 1. Interprete o resultado.

II. DISTRIBUIÇÃO DE PROBABILIDADE

1. PROBABILIDADE

Chama-se *experimento* aleatório o experimento cujo resultado não pode ser previsto. Em outras palavras, um experimento é *aleatório* se, quando executado diversas vezes, produz resultados diferentes. Entretanto, pode-se descrever todos os resultados possíveis de um experimento aleatório. A noção de *probabilidade* está ligada diretamente a esse tipo de experimento.

Exemplo 1: Seja o lançamento de uma moeda três vezes. Representando por 0 o aparecimento de coroa e por 1 o aparecimento de cara, os resultados possíveis deste experimento são:

(0; 0; 0), (0; 0; 1), (0; 1; 0), (0; 1; 1), (1; 0; 0), (1; 0; 1), (1; 1; 0) e (1; 1; 1)

O conjunto de todos esses resultados forma o *espaço amostral* e cada um dos 8 resultados é um *ponto amostral*. Qualquer conjunto de pontos amostrais é um *evento*.

Se o espaço amostral é finito, a probabilidade de ocorrer qualquer ponto amostral é um número entre 0 e 1, de modo que a soma das probabilidades de todos os pontos amostrais que compõem o espaço amostral seja igual a 1. Um *evento* é qualquer conjunto de pontos amostrais. A probabilidade de ocorrer um evento é a soma das probabilidades de seus pontos amostrais.

O evento sem pontos amostrais tem probabilidade zero e o evento com todos os pontos amostrais (o próprio espaço amostral) tem probabilidade 1.

Exemplo 2: Quando uma moeda é lançada parece razoável atribuir probabilidade igual a 0,5, tanto de sair cara como de sair coroa. Assim, na execução do experimento: lançar uma moeda três vezes, cada ponto amostral também deve ter a mesma probabilidade de ocorrência. Para ilustrar, tem-se:

a) O ponto amostral: coroa no 1^{0} lançamento, cara no 2^{0} e cara no 3^{0}, isto é, o ponto (0; 1; 1), tem probabilidade igual a 1/8 = 0,125 (ou 12,5%) de ocorrer.

b) O evento: exatamente duas caras, isto é, um ponto do conjunto (0; 1; 1), (1; 0; 1), (1; 1; 0), tem probabilidade igual a 3/8 =0,375 (37,5%) de ocorrer.

c) O evento menos de duas caras, isto é, um ponto do conjunto (0; 0; 0), (0; 0; 1), (0; 1; 0), (1; 0; 0) tem probabilidade igual a 4/8=0,5 (50%)

Exemplo 3: Lançando-se uma moeda um número grande de vezes, deverá aparecer cara em metade dos lançamentos e coroa no restante. A *freqüência relativa* de caras se aproxima de 0,5 conforme é aumentado o número de lançamentos da moeda (Ver Problema 1). Portanto, a freqüência relativa de um ponto amostral pode ser tomada, aproximadamente, como sua a probabilidade.

Se dois eventos, de um mesmo espaço amostral, não têm pontos em comum, a probabilidade de ocorrer um ou o outro é a soma de suas probabilidades. Se a probabilidade do primeiro não depende da probabilidade do segundo e vice-versa, a probabilidade desses dois eventos ocorrerem simultaneamente é o produto de suas probabilidades individuais.

Exemplo 4: No lançamento de um dado, a probabilidade de sair 2 ou 5 é 1/6+1/6=1/3=0,3333. No lançamento de dois dados, a probabilidade de sair 2 e 5 é 1/6.1/6=1/36=0,0278.

2. VARIÁVEL ALEATÓRIA DISCRETA

Variável aleatória discreta é uma variável cujos valores x_1; x_2; x_3; ...; x_n ocorrem respectivamente com probabilidades $p(x_1)$; $p(x_2)$; $p(x_3)$; ...; $p(x_n)$ de modo que a soma dessas probabilidades seja igual a 1. Uma variável aleatória discreta segue uma *distribuição de probabilidades*, dada por uma fórmula, tabela ou gráfico, que corresponde a uma distribuição de freqüências relativas teórica.

Exemplo 5: No experimento do exemplo 1, a variável x = *n^{0} de caras* no lançamento da moeda três vezes é uma variável aleatória discreta. Pode assumir os valores 0; 1; 2 ou 3, com probabilidade respectivamente iguais a p(0)=1/8; p(1)=3/8; p(2)=3/8 e p(3)=1/8.

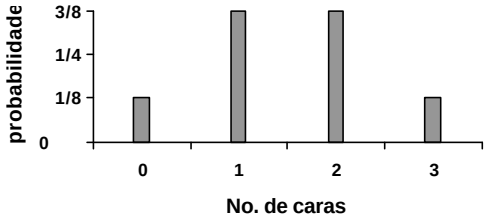
Essa distribuição pode ser dada por

Fórmula: (x=0,1,2,3) $\quad p(x) = \dfrac{3!}{8(3-x)!\,x!}$

Tabela:

x	0	1	2	3
p(x)	1/8	3/8	3/8	1/8

Gráfico -->



Uma distribuição de probabilidade tem média e desvio padrão representados pelas letras gregas μ e σ, respectivamente. A variância é representada por σ^2. A média e a variância da distribuição de probabilidade de uma variável x podem ser indicadas também por E(x) e V(x), respectivamente.

Definem-se

$$\mu = E(x) = \sum_i x_i \cdot p(x_i) \qquad \sigma^2 = V(x) = \sum_i (x_i - \mu)^2 p(x_i)$$

Observa-se que, se as probabilidades $p(x_i)$ forem todas iguais, essas fórmulas são semelhantes as de distribuição de freqüências. Na verdade, como visto no exemplo 3, uma distribuição de probabilidades pode ser construída aproximadamente por uma distribuição de freqüência.

Exemplo 6: Para a variável do exemplo 5, a média, a variância e o desvio padrão são:

$$\mu = 0.\frac{1}{8} + 1.\frac{3}{8} + 2.\frac{3}{8} + 3.\frac{1}{8} = \frac{3}{2} = 1{,}5$$

$$\sigma^2 = (0-\frac{3}{2})^2\frac{1}{8} + (1-\frac{3}{2})^2\frac{3}{8} + (2-\frac{3}{2})^2\frac{3}{8} + (3-\frac{3}{2})^2\frac{1}{8} = \frac{3}{4} = 0{,}75$$

$$\sigma = \sqrt{0{,}75} = 0{,}8660$$

3. DISTRIBUIÇÃO DE BERNOULLI

Uma variável aleatória discreta tem distribuição de Bernoulli quando ela representa um experimento cujo resultado pode ser um sucesso (se ocorrer o evento de interesse) ou um insucesso (o evento de interesse não ocorre). A probabilidade de sucesso é p e a probabilidade de insucesso é q=p-1.

Exemplo 7: No lançamento de uma moeda pode ocorrer cara (sucesso) ou coroa (insucesso). Portanto, o experimento de lançar uma moeda segue uma distribuição de Bernoulli.

4. DISTRIBUIÇÃO BINOMIAL

Uma variável aleatória tem distribuição binomial quando representa a execução de n vezes um experimento de Bernoulli, sendo cada execução independente da outra. Portanto, uma variável aleatória com distribuição Binomial descreve um experimento onde interessa o número de sucessos em n tentativas (ou provas) independentes, tendo cada prova apenas dois resultados possíveis; sucesso ou insucesso. Em cada tentativa a probabilidade de sucesso é p e de insucesso é q=1-p.

Se x é uma variável com distribuição Binomial, a probabilidade de *x* assumir um valor *k* é dada por

$$p(x = k) = C_{n,k} p^k q^{n-k}$$

A média da distribuição Binomial é $\mu = np$ e o desvio padrão é $\sigma = \sqrt{npq}$

Exemplo 8. Seja x = nº de caras no lançamento de uma moeda 3 vezes do exemplo 5. Os valores de x são: 0, 1, 2 e 3. Em cada lançamento a probabilidade de sucesso (cara) é p=0,5 e de insucesso (coroa) é q=0,5. Cada lançamento (tentativa) é independente do outro.

Então, a probabilidade de x assumir um valor k (k=0,1,2,3) quando uma moeda é lançada 3 vezes é:

$$p(k) = C_{3,k} \left(\frac{1}{2}\right)^k \left(\frac{1}{2}\right)^{3-k} = \frac{1}{8} C_{3,k} = \frac{3!}{8(3-k)!\,k!}$$

que é a mesma fórmula usada no exemplo 5 e, portanto, os resultados são os mesmos.

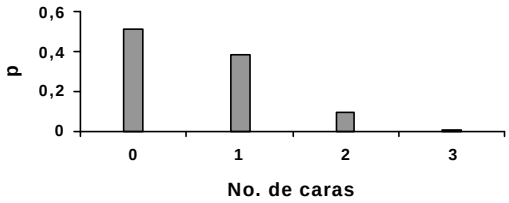
Quando a distribuição é binomial tem-se uma fórmula simples para o cálculo da média e do desvio padrão. A média é $\mu = 3 \cdot (0{,}5) = 1{,}5$ caras por execução do experimento completo (lançamento da moeda 3 vezes) e o desvio padrão

$$\sigma = \sqrt{3 \cdot (0{,}5) \cdot (0{,}5)} = 0{,}8660$$

Esses resultados já foram obtidos no exemplo 5.

Exemplo 9: Supondo que a moeda seja defeituosa, de tal forma que a probabilidade de sair cara em cada lançamento é 0,2, a distribuição de probabilidade da variável x= *nº de caras* é

x	Probabilidade
0	0,512
1	0,384
2	0,096
3	0,008



5. DISTRIBUIÇÃO DE POISSON

A distribuição de Poisson é uma caso particular da distribuição binomial, quando é difícil ou sem sentido calcular o número de insucessos ou o número total de tentativas (p é pequeno e n muito grande). A média é, $\lambda = np$ que também é igual a variância. A probabilidade da variável x com distribuição de Poisson assumir o valor k é

$$p(x = k) = e^{-\lambda} \frac{\lambda^k}{k!}$$

onde *e* é o número irracional 2,71828...

Exemplo 10: Seja um telefone que recebe em média duas chamadas por hora. Então:

a) a probabilidade deste telefone não receber nenhuma chamada em uma hora é

$$p(x=0) = e^{-2}\frac{2^0}{0!} = e^{-2} = 0{,}1353 \qquad (\lambda=2)$$

b) a probabilidade de receber no máximo 2 chamadas em 30 minutos é

$$p(x \le 2) = p(x=0) + p(x=1) + p(x=2) \qquad (\lambda = 1)$$

$$= e^{-1}\frac{1^0}{0!} + e^{-1}\frac{1^1}{1!} + e^{-1}\frac{1^2}{2!} = 0{,}9197$$

6. USANDO O EXCEL

Funções

DISTRBINOM(x; n; p; acumulada) POISSON(x, média; acumulada)	Ambas fornecem a *probabilidade* exata p(=x) se acumulada = FALSO e a *probabilidade acumulada* p($\le$ x) se acumulada=VERDADEIRO

PROBLEMAS:

1) Utilizando as funções ALEATÓRIO ou ALEATÓRIOENTRE simule o lançamento de uma moeda 50, 100, 200, 500 e 1000 vezes. Determine a freqüência relativa de caras. Compare as freqüências relativa de caras obtidas com os valores teóricos (probabilidades).

2) Considere o experimento de lançar uma moeda 3 vezes e observar o número de caras. Repita este experimento 1000 vezes. Construa a distribuição de freqüência do nº de caras, calcule a média e desvio padrão. Compare os resultados com os valores teóricos.

3) Considere o lançamento de uma moeda perfeita 30 vezes. Construa a distribuição de probabilidade e o gráfico da variável *nº de caras* nos 30 lançamentos. Determine a média, variância e desvio padrão. Que porcentagem dos valores estão no intervalo de 2 desvios padrão em torno da média. Compare com o valor dado pelo teorema de Chebyshev.

4) Um casal pretende ter 5 filhos e acredita que a probabilidade de ter um filho homem é 0,55. Nessas condições, qual a probabilidade dos 3 filhos do casal serem:
   a) 3 homens e 2 mulheres?
   b) pelo menos uma mulher
   c) mais de dois homens?

5) a) Considere ainda a probalidade de um filho homem igual a 0,55. Escolhendo-se ao acaso 200 casais em uma cidade com 5 filhos, quantos deverão ter exatamente 3 filhos homens?
   b) Qual a média de filhos homens de casais desta cidade?

6) Um recipiente contém 5000 bactérias. A probabilidade de que uma bactéria escape do recipiente é 0,0008. Qual a probabilidade de que mais de 6 bactérias escapem?

7) Estude no Excel as funções DIN.BIN.NEG e DIST.HIPERGEOM. Dê exemplos.

7. VARIÁVEL ALEATÓRIA CONTÍNUA

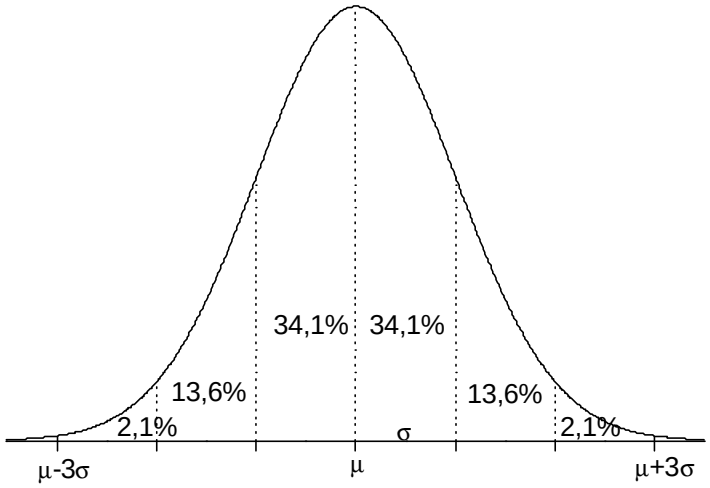
Variável aleatória contínua é uma variável cujos intervalos de valores ocorrem com uma certa probabilidade. Uma variável aleatória contínua possui uma distribuição de probabilidade que é dada por uma *função densidade de probabilidade* f(x) ou seu gráfico.

8. DISTRIBUIÇÃO NORMAL (ou de GAUSS)

Uma variável aleatória x tem distribuição normal se a sua função densidade de probabilidade é

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

onde μ é a média e σ o desvio padrão.



O gráfico de uma distribuição normal tem a forma de sino e a área total abaixo da curva é igual a 1. Qualquer fração da área total representa a probabilidade da variável x assumir um valor entre os extremos que definem esta área. Na figura, a probabilidade de um valor de x estar entre um desvio padrão antes da média e um desvio padrão depois é 0,341+0,341=0,682. Em outras palavras, 68,2% dos valores de x estão entre μ-σ e μ+σ.

Exemplo 8: Quanto por cento dos valores de x estão entre dois desvios padrão antes da média e dois desvios padrão depois? E entre três desvios padrão?

Observando-se o gráfico anterior pode-se responder facilmente às questões propostas: Estão entre 2 desvios padrão em torno da média 2(34,1+13,6)=95,4% dos valores. Entre 3 desvios padrão em torno da média tem-se 2(34,1+13,6+2,1)=99,6%

Exemplo 9: Considerando que a distribuição normal é simétrica em torno da média, praticamente 100% dos valores se localizam entre 3 desvios padrão antes da média e três desvios padrão depois da média e quanto maior o desvio padrão mais espalhados estão os valores em torno da média, esboce em um mesmo sistema de coordenadas os gráficos de três distribuições normais, todas de média 10, e desvios padrão 0,5; 1,0 e 1,5.

Exemplo 10: Suponha que uma população de estudantes tenha altura média 1,62 m e desvio padrão 0,08 m. Interprete a variação das alturas desta população.

Uma variável z de distribuição normal de média 0 e desvio padrão 1 é chamada *distribuição normal padrão*. Toda variável x com distribuição normal de média μ e variância σ^2 pode ser transformada para uma variável normal padrão z, definida por $z = \frac{x - \mu}{\sigma}$

Existem tabelas que fornecem áreas da distribuição normal padrão correspondentes a diversos valores de z. Uma delas, dada no apêndice, dá áreas da normal padrão acumulada.

Exemplo 11 No exemplo 10, a) qual a probabilidade de uma pessoa escolhida ao acaso da população ter altura menor que 1,74 m? b) Quanto por cento das pessoas da população têm altura menor do que 1,74 m? c) Quanto por cento têm alturas entre 1,58 e 1,66 m? Em que intervalo simétrico em torno da média estão 86% das alturas?

9. USANDO O EXCEL

Funções

DIST.NORM(x; μ; σ; acumulada)	*Probabilidade acumulada* F(<x) se acumulada =VERDADEIRO e Função densidade f(x) se acumulada=FALSO
INVNORM(p; μ; σ)	Inversa da normal: dá x tal que a área até ele é p
DIST.NORMP(z)	Normal padrão acumulada: da área até z
INVNORMP(p)	Inversa da normal padrão: dá z para área p

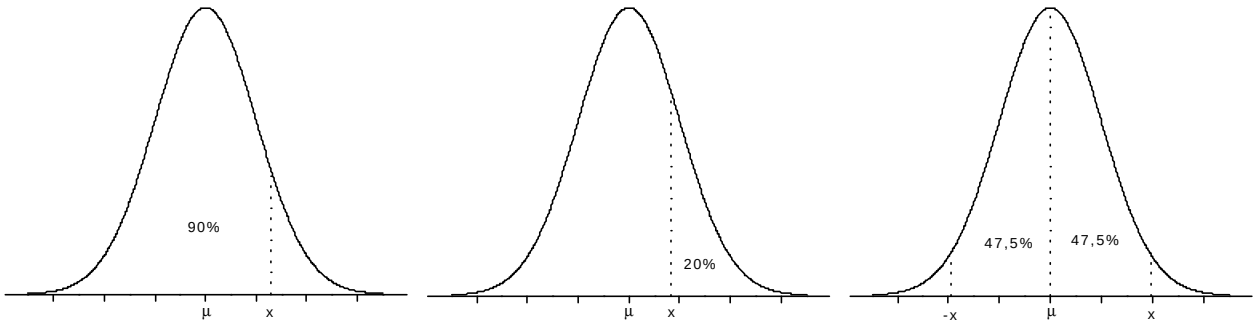
PROBLEMAS:

8) Se z é uma variável com distribuição normal padrão, calcule a probabilidade de z assumir um valor
a) menor do que 1,26 b) maior do que 1,26 c) maior do que -2 d) entre -0,80 e 1,78 e) entre -1,96 e 1,96

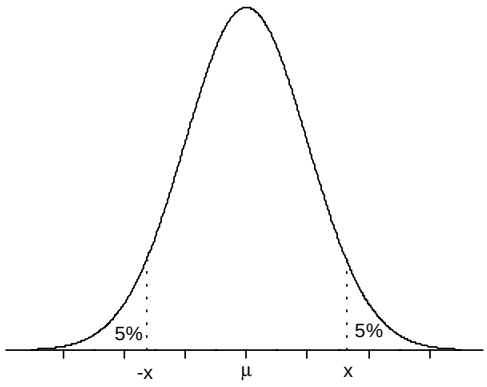
9) Se x tem distribuição normal de média μ=10 e σ=2, calcule a probabilidade de x assumir um valor
a) menor do que 12,5 b) maior do que 6,5 c) entre 6,5 e 12,5

10) Resolva o problema 8 usando a distribuição normal padrão

11) Os gráficos da figura 1 são de uma variável x com distribuição normal de média 320 e desvio padrão 25. Calcule os valores de x.



12) A figura abaixo representa uma distribuição normal padrão. Calcule o valor de x



13) Uma variável x tem distribuição normal de média 0,6 e desvio padrão 0,04. Em que intervalo simétrico em torno da média se encontram 95% dos valores de x? e 99%?

14) Simule valores das distribuições contínuas constantes da *ferramenta de análise* GERAÇÃO DE NÚMEROS ALEATÓRIOS.

PROBLEMA PROPOSTO

PP4) Suponha que a taxa de glicose no sangue das pessoas normais tenha distribuição normal de média 90 mg/dl e desvio padrão 9 mg/dl.

a) Quando uma pessoa poderia ser considerada com glicemia fora dos padrões normais?
b) Em geral, são aceitos como referência para uma pessoa sã os limites 70 e 110 mg/dl. Que área da distribuição normal acima é abrangida por esses limites?
c) Ainda considerando essa distribuição normal, 90% das pessoas deveriam ter a taxa de glicose em que intervalo simétrico em torno da média?
d) Simule 1000 valores desta distribuição, construa uma distribuição de freqüência e, a partir desta, responda as questões a) b) e c).

III. DISTRIBUIÇÃO AMOSTRAL

1. AMOSTRAGEM ALEATÓRIA

Dada uma população, à qual está associada uma variável de interesse, pretende-se retirar uma amostra de n elementos e, a partir desta amostra, estimar valores populacionais desconhecidos, tais como a média, proporção, desvio padrão, etc. Um modo simples de amostragem é a retirada da amostra de tal forma que, durante o processo de seleção, cada elemento da população tenha igual probabilidade de ser escolhido.

Seja uma população de média μ e variância σ^2. Para uma amostra com valores x_1, x_2, ..., x_n, a média e a variância serão indicadas respectivamente por $\bar{x}$ e s^2, de modo a distinguir dos valores populacionais μ e σ^2. A média e a variância da amostra são definidas por:

$$\boxed{\bar{x} = \frac{1}{n}\sum x_i} \quad \text{e} \quad \boxed{s^2 = \frac{1}{n-1}\sum (x_i - \bar{x})^2}.$$

Esses valores baseados na amostra são chamados de *estatísticas*.

Antes de considerar uma amostra individual, tomar-se-á para estudo todas as diferentes amostras de tamanho n que podem ser obtidas da população. Neste curso, quando a população for finita, a amostragem será com reposição. Para populações infinitas, ou muito grandes, não importa se a amostragem é com ou sem reposição.

2. DISTRIBUIÇÃO AMOSTRAL DA MÉDIA

A média amostral é uma variável aleatória e possui uma distribuição de probabilidades chamada *distribuição amostral da média*. O mesmo acontece para variância, desvio padrão, etc

Exemplo 1: Uma caixa possui a mesma quantidade de bolas com o números 10, 20, 30, 40 e 50. Seja a variável *x = nº da bola* e todos os modos possíveis de serem retiradas duas bolas desta caixa (isto é, amostras de tamanho n=2), com reposição da primeira.

Amostras n=2	Média amostral
(10 ; 10)	10
(10 ; 20)	15
(10 ; 30)	20
(10 ; 40)	25
(10 ; 50)	30
(20 ; 10)	15
(20 ; 20)	20
(20 ; 30)	25
(20 ; 40)	30
(20 ; 50)	35
(30 ; 10)	20
(30 ; 20)	25
(30 ; 30)	30
(30 ; 40)	35
(30 ; 50)	40
(40 ; 10)	25
(40 ; 20)	30

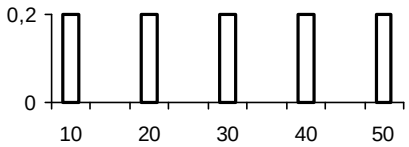
População (variável x): (10; 20; 30; 40; 50)

Distribuição de probabilidades

x	10	20	30	40	50
prob	0,2	0,2	0,2	0,2	0,2

média $\mu = 30$

variância $\sigma^2 = 200$



Distribuição amostral de médias (n=2)

$\bar{x}$ = média amostral

$\bar{x}$	10	15	20	25	30	35	40	45	50
prob	0,04	0,08	0,12	0,16	0,20	0,16	0,12	0,08	0,04

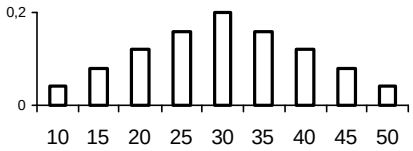
(40 ; 30)	35
(40 ; 40)	40
(40 ; 50)	45
(50 ; 10)	30
(50 ; 20)	35
(50 ; 30)	40
(50 ; 40)	45
(50 ; 50)	50
Média	30
Variância	100

média = $\mu(\bar{x}) = \mu = 30$

variância = $\sigma^2(\bar{x}) = \frac{\sigma^2}{n} = \frac{200}{2} = 100$

desvio padrão = $\sigma(\bar{x}) = \frac{\sigma}{\sqrt{n}} = 10$

gráfico da distribuição de médias



Exemplo 2: Na população do exemplo 1, qual a probabilidade de uma amostra de tamanho 2 ter média menor ou igual a 40? E entre 25 e 40, inclusivos? (R: 0,88 e 0,44)

TEOREMA DO LIMITE CENTRAL

Para amostras aleatórias relativas a uma variável x associada a uma população com média μ e variância σ^2, a distribuição amostral da média $\bar{x}$ de amostras de tamanho n tem média μ e variância σ^2/n. Se x é normal, então $\bar{x}$ também é normal. Mesmo que x não seja normal, $\bar{x}$ se aproxima da normal a partir de determinados tamanhos da amostra (n>30).

O desvio padrão $\frac{\sigma}{\sqrt{n}}$ é chamado *erro padrão* da média.

Exemplo 3: Na população do exemplo 1, qual a probabilidade de uma amostra de tamanho 64 ter média menor ou igual a 40? E entre 25 e 40? (Resp.: 0,9772 e 0,8186)

3. DISTRIBUIÇÃO AMOSTRAL DA PROPORÇÃO (ou freqüência relativa)

Exemplo 4: Uma caixa contém 1/3 de bolas amarelas e 2/3 de bolas brancas (população). Duas bolas são retiradas, uma a uma com reposição da primeira (amostras de tamanho 2), e é observada a proporção (ou frequência relativa) de bolas brancas.

Amostras n=2	Proporção amostral
(A ; A)	0
(A ; B_1)	0,5
(A ; B_2)	0,5
(B_1 ; A)	0,5
(B_1 ; B_1)	1
(B_1 ; B_2)	1
(B_2 ; A)	0,5
(B_2 ; B_1)	1
(B_2 ; B_2)	1
média	2/3
variância	1/9

população: variável x tal que: x=1 a bola é branca
x=0 a bola não é branca

p= proporção de bolas brancas = 2/3

x	0	1
prob	1-p	p

média = $\mu(x) = p = 2/3 = 0{,}6667$

variância = $\sigma^2 = p(1-p) = 2/9 = 0{,}2222$

Distribuição amostral de proporções (n=2)

$\hat{p}$ = proporção de bolas brancas na amostra (n=2)

$\hat{p}$	0	0,5	1
prob	1/9	4/9	4/9

média = $\mu(\hat{p}) = p = 2/3$

variância = $\sigma^2(\hat{p})$
= p(1-p)/n =1/9
= 0,1111

PROPRIEDADE

Se n>30 a distribuição amostral de $\hat{p}$ se aproxima de uma distribuição normal de média $\mu = p$ e variância $\sigma^2 = p(1-p)/n$.

Exemplo 5: No exemplo anterior, retirando-se 200 bolas da caixa, com reposição de cada bola, qual a probabilidade da proporção de bolas brancas ser menor do que 60%? (R: 0,0228)

4. USANDO O EXCEL

PROBLEMAS:

1) Uma caixa contém bolas numeradas 6 e 9, na mesma proporção. Forme a distribuição amostral de médias de amostras aleatórias de tamanho 3. Calcule a média e a variância da distribuição.
2) Qual a probabilidade da média de uma amostra de tamanho 100 retirada da população do problema anterior estar entre 6,5 e 7,8?
3) (Amostragem normal) Com a ferramenta GERAÇÃO DE NÚMERO ALEATÓRIO obter 1000 alturas de uma distribuição normal de média 1,62 m e desvio padrão 0,08 m. Forme a distribuição de freqüência, calcule a média e o desvio padrão.
4) Considere as alturas do problema 3 como sendo uma população. Com a ferramenta AMOSTRAGEM, sorteie amostras de tamanhos 5, 10, 30 e 120. Calcule a média e desvio padrão de cada amostra.
5) Considerando o problema 1, forme a *distribuição amostral de variâncias*. Calcule a média dessa distribuição amostral. Observe que a média das variâncias amostrais é igual a variância populacional. Isso justifica a divisão por (n-1) em lugar de (n) no cálculo da variância da amostra.
6) Estude no Excel, com a *Ferramenta de Análise* AMOSTRAGEM, como funciona o método de amostragem *periódico*.

IV. ESTIMAÇÃO DE PARÂMETROS

1. INTERVALO DE CONFIANÇA PARA A MÉDIA POPULACIONAL m

1º caso: A variância populacional s^2 é conhecida

Seja x uma variável aleatória de média μ (desconhecida) e desvio padrão σ (conhecido). Do capítulo anterior tem-se que a distribuição amostral de médias $\bar{x}$ de amostras de tamanho n, quando x é normal ou n é suficientemente grande, também é normal de média μ e desvio padrão $\frac{\sigma}{\sqrt{n}}$.

Na figura 1 é apresentado um intervalo simétrico em torno da média μ, de extremos $\mu - e_0$ e $\mu + e_0$, de tal modo que a probabilidade de $\bar{x}$ estar neste intervalo é $1-\alpha$, isto é,

$$P(\mu - e_0 \le \bar{x} \le \mu + e_0) = 1 - \alpha$$

Pela distribuição normal padrão calcula-se e_0

$$\frac{(\mu + e_0) - \mu}{\sigma/\sqrt{n}} = z_0, \text{ portanto } e_0 = z_0 \frac{\sigma}{\sqrt{n}}.$$

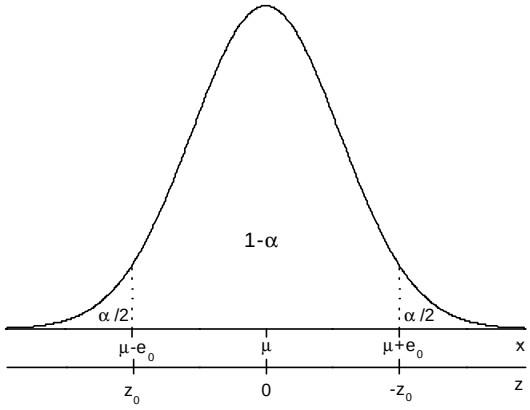


Figura 1. Intervalo de probabilidade (1-α) para a média

Assim $\boxed{P(\bar{x} - z_0 \frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + z_0 \frac{\sigma}{\sqrt{n}}) = 1 - \alpha}$ e fica definido um intervalo de extremos

$$\boxed{\bar{x} \pm z_0 \frac{\sigma}{\sqrt{n}}}$$

que poderá conter ou não a média populacional μ. Como esta é um parâmetro e não uma variável aleatória, não tem sentido dizer que "a probabilidade μ cair no intervalo é 1-α", por isso diz-se que os extremos acima definem um *intervalo de confiança* para a média μ. A interpretação será reforçada no exemplo a seguir.

Exemplo 1: Sabe-se que uma variável x =altura de alunos tem desvio padrão $\sigma = 0{,}09$m. Se em uma amostra de 36 alunos foi encontrada a média $\bar{x}$=1,70 m, qual o intervalo de 95% de confiança para a média μ de x? E o intervalo de 90%? (com uma amostra grande como esta não é necessário conhecer o desvio padrão populacional, pode ser usado o desvio padrão amostral *s*)

Se 1-α=0,95 → α=0,05, então z_0=1,96 (ver tabela no apêndice) e um intervalo de 95% de confiança para μ tem extremos $1{,}70 \pm 1{,}96 \frac{0{,}09}{\sqrt{36}} = \boxed{1{,}70 \pm 0{,}029}$, ou seja

1,670<m<1,729

Isso significa que 95% dos intervalos construídos com amostras de tamanho n=36, retiradas ao acaso desta população conterão a média μ.
Se α=0,10 obtém-se um intervalo de 90% de confiança **1,675<m<1,725**

2^0 caso: A variância populacional s^2 é desconhecida

Neste caso, não se conhece a variância populacional σ^2. Se a amostra é suficientemente grande, toma-se o desvio padrão da amostra como um valor aproximado do desvio padrão populacional. Então, emprega-se a metodologia anterior com *s* em lugar de σ.

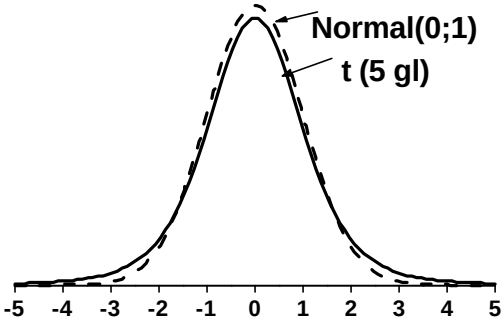
Entretanto, se a amostra é pequena, desde que a distribuição da população seja normal, usa-se a distribuição *t de Student*. O intervalo terá extremos definidos por

$$\bar{x} \pm t_0 \frac{s}{\sqrt{n}}$$

onde t_0 é obtido da distribuição de t com n-1 graus de liberdade (ver Tabela 2 anexa).

Observação: Enquanto z_0 depende apenas de $\bar{x}$, t_0 depende de $\bar{x}$ e *s*. A distribuição de t é simétrica em torno da média t=0 e tem a forma de sino. Ela se aproxima da normal conforme n cresce.

Exemplo 2: A cronometragem de certa operação forneceu os seguintes valores para n=6 determinações: 4; 5; 5; 6; 8 e 8 (em minutos). Supondo a cronometragem uma variável com distribuição aproximadamente normal, calcule intervalos de 95% e 99% de confiança para a média populacional μ..



(R: média $\bar{x} = 6$, variância s^2 = 2,8 e erro padrão $\frac{s}{\sqrt{n}} = \sqrt{\frac{2,8}{6}} = 0,6831$, com 5 G.L.

Se α=0,05 → t_0=2,4469 e **4,3<m<7,7**
α=0,01 → t_0=4,0321 e **3,2<m<8,8**)

2. INTERVALO DE CONFIANÇA PARA A PROPORÇÃO

Para estimar a proporção ρ de elementos da população com uma certa característica usa-se a proporção $\hat{p}$ com que essa característica foi observada em uma amostra. Desde que a amostra seja grande, pode-se tomar a distribuição normal como aproximação para a binomial.

Um intervalo de confiança aproximado para p, ao nível de confiança 1-α, é dado por

$$\hat{p} \pm z_0 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Exemplo 3: Retirando-se uma amostra de 100 itens da produção de uma máquina, verificou-se que 10 eram defeituosas. Encontre um intervalo de 95% de confiança para a proporção p de peças defeituosas dessa máquina.
(R: entre 4% e 16%)

3. INTERVALO DE CONFIANÇA PARA A VARIÂNCIA

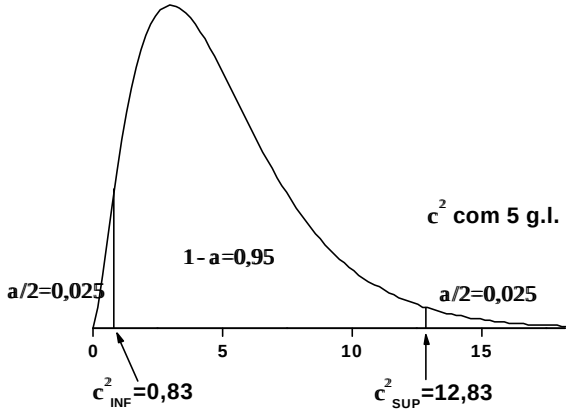
Seja uma população normal de média μ e variância σ^2. Considerando-se as amostras de tamanho n, com variância s^2, desta população, prova-se que a estatística $\chi_0^2 = \frac{(n-1)s^2}{\sigma^2}$ tem distribuição de qui-quadrado (χ^2) com n-1 graus de liberdade

Um intervalo de confiança para σ^2, com base em uma amostra de tamanho n e variância s^2, ao nível confiança 1-α, é dado por

$$\boxed{\frac{(n-1)s^2}{\chi^2_{SUP}} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{INF}}}$$

onde χ^2_{INF} e χ^2_{SUP} definem na limites da distribuição de qui-quadrado correspondentes à probabilidade 1-α.

Exemplo 3: Determine um intervalo de 95% de confiança para variância populacional da variável cronometragem do exemplo 2.



(R: s^2=2,8, n=6 e $\boxed{\frac{5(2,8)}{12,83} \le \sigma^2 \le \frac{5(2,8)}{0,83}}$ ou **1,091<s²<16,867** Tomando a raiz quadrado dos elementos dessa desigualdade determina-se um intervalo de confiança aproximado para o desvio padrão: **1,044<s²<4,107**)

4. TAMANHO DAS AMOSTRAS

Pode-se estabelecer o tamanho *n* de uma amostra para obter um intervalo de confiança com uma semi-amplitude e_0 pré-fixada. Por exemplo, no caso da média

$$e_0 = z_0 \frac{\sigma}{\sqrt{n}} \Rightarrow \boxed{n = \left(\frac{z_0\sigma}{e_0}\right)^2}$$

Em geral, σ é desconhecido e utiliza-se o desvio padrão de uma amostra piloto suficientemente grande.

Exemplo 4: Em relação à variável altura do exemplo 1, qual o tamanho de uma amostra para se obter um intervalo de 95% de confiança com e_0 (semi-amplitude) aproximadamente igual a 2 cm?
(R: $n \cong 78$)

5. INTERVALO DE CONFIANÇA PARA A DIFERENÇA ENTRE DUAS MÉDIAS de populações normais.

Sejam duas populações:

População 1: variável x_1 com distribuição normal de média μ_1 e variância σ_1^2.

População 2: variável x_2 com distribuição normal de média μ_2 e variância σ_2^2

São retiradas aleatoriamente duas amostras de tamanhos n_1 e n_2, uma de cada população, cuja médias são $\bar{x}_1$ e $\bar{x}_2$ e cujas variâncias são s_1^2 e s_2^2, respectivamente. Pretende-se estabelecer um intervalo de confiança para a diferença entre as médias populacionais, desconhecidas, $\mu_1 - \mu_2$. Conforme o nível de confiança 1-α adotado, são usados valores z_0 da distribuição normal, quando as variâncias populacionais são conhecidas, e valores t_0 da distribuição de t, quando se usa as variâncias das amostras

1^0) As variâncias populacionais são conhecidas

Suposição: as amostras são obtidas independentemente

$$(\bar{x}_1 - \bar{x}_2) \pm z_0 \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

2^0) As variâncias populacionais são desconhecidas

Suposições: as variâncias populacionais podem ser consideradas iguais, isto é, $\sigma_1^2=\sigma_2^2=\sigma^2$ e as amostras são obtidas independentemente

$$(\bar{x}_1 - \bar{x}_2) \pm t_0 . s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \quad \text{onde} \quad s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

OBS: Quando não é possível assumir que $\sigma_1^2=\sigma_2^2=\sigma^2$, é calculado um intervalo de confiança aproximado ao nível de 1-α de confiança:

$$(\bar{x}_1 - \bar{x}_2) \pm t_0 \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$ onde t_0 tem $\dfrac{(s_1^2/n_1 + s_2^2/n_2)^2}{\dfrac{(s_1^2/n_1)^2}{n_1 + 1} + \dfrac{(s_2^2/n_2)^2}{n_2 + 1}}$ graus de liberdade

7. USANDO O EXCEL

Funções	DIST.NORM(x; μ; p; acumulada)	*Probabilidade acumulada* se acumulada =VERDADEIRO e Função densidade se acumulada=FALSO
	INVNORM(α; μ; p)	Inversa da normal
	DIST.NORMP(z)	Normal padrão acumulada
	INVNORMP(p)	Inversa da normal padrão
	DIST.QUI(x; graus de liberdade)	Qui-quadrado
	INV.QUI(p; graus de liberdade)	Inversa da Qui-quadrado

PROBLEMAS:

1) Usando o Excel resolva os exemplos de 1 a 4.
2) Usando a ferramenta de análise GERAÇÃO DE NÚMERO ALEATÓRIO obtenha 1000

valores de uma variável normal de média 6 e desvio padrão 1,5. Faça de conta que os valores simulados são da variável: cronometragem de certa operação (exemplo 2). Tirando uma amostra de tamanho 6 desta população (ver problema 4, página 19) determine intervalos de 90, 95 e 99% para a média

3) Em uma pesquisa de opinião sobre a transformação de um jardim em estacionamento, foram consultados aleatoriamente 250 habitantes de uma cidade e 80 se motraram favoráveis. Encontre os limites de confiança de 90% e 95% para a proporção da população favorável a construção do estacionamento

PROBLEMAS ADICIONAIS DE LIVROS TEXTO

FONSECA, J.S.; MARTINS, G.A. *Curso de Estatística.* 3 ed. São Paulo: Ed. Atlas, 1981.

4) Foram retiradas 25 peças da produção diária de uma máquina, encontrando-se para uma certa medida uma média 5,2 mm. Sabendo-se que as medidas têm distribuição normal com desvio padrão 1,2 mm, construir intervalos de confiança para a média aos níveis de 90%, 95% e 99%. (R: $4,81 \leq \mu \leq 5,59$; $4,73 \leq \mu \leq 5,67$; $4,58 \leq \mu \leq 5,82$)

5) Em uma fábrica, colhida uma amostra de certa peça, obtiveram-se as seguintes medidas para os diâmetros:
10; 11; 11; 11; 12; 12; 12; 12; 13; 13; 13; 13; 13; 13; 13; 13; 13; 13; 13; 14; 14; 14; 14; 14; 15; 15; 15; 16; 16.
a) Estimar a média e variância
b) Construir um intervalo de confiança para a média ao nível de 5% de significância
(R: a) $\bar{x} = 13,13$; $s^2 = 2,05$ b) $12,60 \leq \mu \leq 13,66$)

6) Uma amostra de 300 habitantes de uma cidade mostrou que 180 desejavam a água fluorada. Encontrar os limites de confiança de 90% e 96% para a proporção da população favorável a fluoração. (R: $0,55 \leq p \leq 0,65$; $0,54 \leq p \leq 0,66$)

7) Uma amostra de tamanho 36 foi extraída de uma população normal de média μ_1 e variância σ^2 = 9, dando média $\bar{x}_1 = 70$. Uma outra amostra de tamanho 25 foi extraída de outra população normal de variância 16, dando $\bar{x}_2 = 60$. Determinar o intervalo para $\mu_1 - \mu_2$ ao nível de 96%. (R: $8,07 \leq \mu_1 - \mu_2 \leq 11,93$)

8) Supondo populações normais, construir o intervalo de confiança para a variância ao nível de 90% para as amostras:
a) 44,9; 44,1; 43,0; 42,9; 43,2; 44,5
b) 2; 2; 2; 3; 3; 5; 5; 5; 5; 6; 6; 7; 7; 8.
(R: a) $0,32 \leq \sigma^2 \leq 3,13$ b) $2,25 \leq \sigma^2 \leq 8,13$)

BUSSAB, O.B., MORETTIN, P.A. *Estatística básica*. São Paulo: Ed. Atual. 1987.

9) Um pesquisador está estudando a resistência de um determinado material sob determinadas condições. Ele sabe que essa variável é normalmente distribuída com desvio padrão de 2 unidades.
a) Utilizando os valores 4,9; 7,0; 8,1; 4,5; 5,6; 6,8; 7,2; 5,7; 6,2 unidades, obtidos de uma amostra de tamanho 9, determine o intervalo de confiança para a resistência média com um coeficiente de confiança 0,90. (R: 5,13<média< 7,32)
b) Qual o tamanho da amostra necessário para que o erro cometido, ao estimarmos a resistência média, não seja superior a 0,01 unidades com probabilidade 0,90? (R: n=108222)
c) Suponha que no item (a) não fosse conhecido o desvio padrão. Como você procederia para determinar o intervalo de confiança? (R: 5,50<média< 6,94)

10) Estão sendo estudados dois processos A e B para conservar alimentos, cuja principal

variável de interesse é o tempo de duração dos mesmos. Nos dois processos o tempo segue uma distribuição normal de variância é 100 e médias, respectivamente, μ_A e μ_B. Sorteiam-se duas amostras independentes: a amostra de A, com 16 latas, apresentou tempo médio de duração igual a 50, e a de B, com 25 latas, duração média igual a 60.

a) Construa um intervalo de confiança para μ_A e μ_B separadamente (R: 50±4,9 e 60±3,9)

b) Para verificar se os dois processos podem ter o mesmo desempenho, decidiu-se construir um intervalo de confiança para a diferença μ_A - μ_B. Caso o zero pertença ao intervalo, pode-se concluir que existe evidência de igualdade dos processos. Qual seria a sua resposta? (R: 10±6,3, não inclui o zero)

11) Antes de uma eleição em que existiam 2 candidatos A e B, foi feita uma pesquisa com 400 eleitores escolhidos ao acaso e verificou-se que 208 deles pretendiam votar no candidato A. Construa um intervalo de confiança, ao nível de 95%, para a porcentagem de eleitores favoráveis ao candidato A na época das eleições. (R: 0,520±0,049)

COSTA NETO, P.L.O. *Estatística.* São Paulo: Ed. Edfgard Blucher, 1977.

12) Uma amostra extraída de uma população normal forneceu os seguintes valores: 3,0; 3,2; 3,4; 2,8; 3,1; 2,9; 3,0; 3,2. Construa intervalos de 95% de confiança para a

a) variância da população (R: 2,92<média<3,23)

b) média da população (R: 0,0159<variância<0,1509)

13) Dadas duas amostras aleatórias de tamanhos 10 e 12, extraídas de duas populações normais independentes, as quais forneceram, respectivamente, $\bar{x}_1 = 20$, $\bar{x}_2 = 24$, $s_1 = 5{,}0$ e $s_2 = 3{,}6$; estabeleça um intervalo de 95% de confiança para a diferença entre as médias populacionais. (R: 4±3,9)

V. TESTE DE HIPÓTESES

1. INTRODUÇÃO

Problema ilustrativo: um fabricante de fruta em conserva afirma que os pesos das latas com o seu produto têm média 600 g e desvio padrão 30 g. Suspeita-se, entretanto, que o peso médio é menor do que o anunciado. Pretende-se decidir se a suspeita sobre a média tem procedência ou não, usando-se uma amostra aleatória, por exemplo, de 36 latas (por enquanto, o desvio padrão será considerado correto).

Existem duas hipóteses quanto a média μ da população de pesos: uma, chamada **hipótese nula**, H_0, de que μ = 600 g (ou $\mu - 600 = 0$) e outra, mais ampla, chamada **hipótese alternativa**, H_1, de que μ < 600 g.

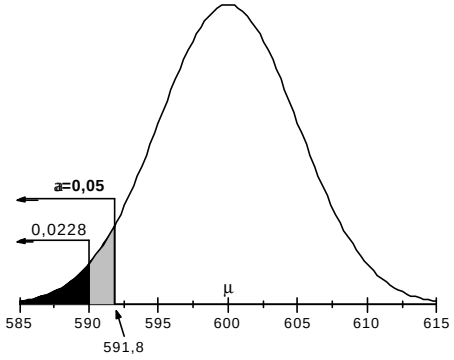
Com base na média de uma amostra de aleatória de n = 36 pesos de latas com fruta em conserva, será enunciado um critério para decidir se H_0 pode ser contrariada ou não. Portanto, feita uma determinada hipótese sobre um parâmetro de uma população, pretende-se saber se os resultados de uma amostra de tamanho n contrariam ou não tal afirmação.

Seja a variável x=peso, com média μ=600g e desvio padrão σ=30g. A variável aleatória $\bar{x}$, média de amostras de n=36 pesos, terá distribuição aproximadamente normal de média 600g e desvio padrão $\frac{30}{\sqrt{36}} = 5$ g.

Se a hipótese nula for verdadeira, o gráfico da figura representa a distribuição amostral de médias de 36 pesos. Por exemplo, a probabilidade da média de uma amostra ser menor do que 590 g é:

$$P(\bar{x} < 590) = P(z < \frac{590 - 600}{5}) = P(z < -2) = 0{,}0228$$

isto é, se o fabricante estiver certo, 2,28% das amostras de 36 latas possuem peso médio menor que 590 g.

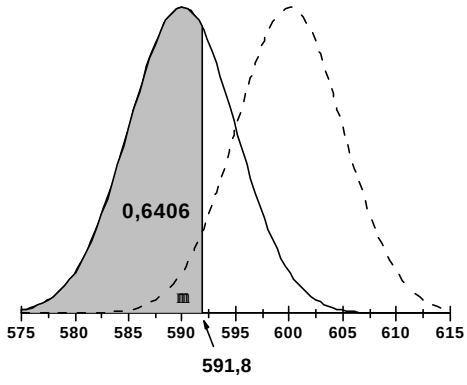


Pode-se fixar uma probabilidade α e determinar um valor $\bar{x}_c$ de modo (100.α)% das médias amostrais sejam menores do que ele, ou seja, tal que $P(\bar{x} < \bar{x}_c) = \alpha$. Escolhendo α = 0,05 tem-se:

$$P(\bar{x} < \bar{x}_c) = P(z < \frac{\bar{x}_c - 600}{5}) = 0{,}05$$

Como $P(z < -1{,}64) = 0{,}05$, então,

$$\frac{\bar{x}_c - 600}{5} = -1{,}64 \Rightarrow \boxed{\bar{x}_c = 591{,}8 \text{ g}}$$



Portanto, a probabilidade de uma média amostral de 36 pesos ser menor que 591,8g é 0,05. Desde que a hipótese nula seja verdadeira, apenas 5% das médias amostrais serão menores do que 591,8g.

Se a informação do fabricante é incorreta, então a média real é menor do que 600g e a probabilidade de uma média de 36 pesos ser menor do que 591,8g é superior a 5%. Por exemplo, supondo que a média correta seja 590g, a probabilidade de obter uma amostra de média menor do que 591,8 é 64,06% (ver figura)

Conclusão: Se a média $\bar{x}_0$ de uma amostra de 36 pesos for menor que $\bar{x}_c = 591{,}8$ g, tem-se uma das duas alternativas abaixo:

a) O fabricante está certo, a média da população de pesos é μ=600 g e foi obtida uma amostra com tão pouca chance de ocorrer por puro acaso.

b) O fabricante não diz a verdade, pois obteve-se tal média amostral porque a probabilidade de sua acorrência não era tão pequena, ou seja, a média da população é menor do que 600 g

(μ<600 g).

Com qual alternativa ficar?

Critério: Observe que foi fixado um valor razoavelmente pequeno para α, no caso α=0,05, determinou-se $\bar{x}_c = 591{,}8$, tal que a probabilidade de qualquer média de amostra de tamanho n=36 ser menor que $\bar{x}_c$ é 0,05 (5%), quando a média da população é μ=600g e o desvio padrão σ=30g. Retirando-se uma amostra, cuja média é $\bar{x}_0$, pode-se estabelecer o seguinte:

Se $\bar{x}_0 > \bar{x}_c$ aceita-se H_0
Se $\bar{x}_0 \leq \bar{x}_c$ rejeita-se H_0, aceitando H_1

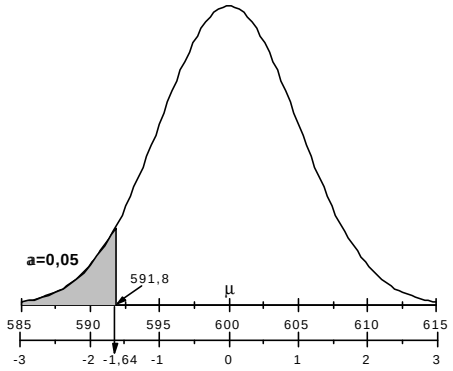
Pelo que foi discutido, rejeitando H_0 pode-se estar cometendo um erro, chamado **erro do tipo I** (rejeitar H_0 quando ela deveria ser aceita). A probabilidade de cometer um erro do tipo I é igual a α. Em geral, α = 0,05 ou α = 0,01 e é chamado **nível de significância** do teste.

Aceitando-se H_0, também pode-se estar cometendo um erro, chamado **erro do tipo II** (aceitar H_0 quando ela deveria ser rejeitada). Para calcular a probabilidade de cometer um erro do tipo II é preciso conhecer a média populacional, o que raramente ocorre na prática.

Portanto, em um teste de hipótese a maior preocupação é com o erro do tipo I, cuja probabilidade α é conhecida. Tem-se uma decisão estatísticamente forte quando se rejeita H_0.

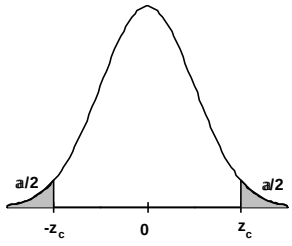
Observações:

1º) Em vez de verificar se $\bar{x}_0 < \bar{x}_c$ pode-se verificar se $z_0 = \dfrac{\bar{x}_0 - 600}{5} \leq -1{,}64$, isto é, se $\dfrac{\bar{x}_0 - m_0}{\dfrac{s}{\sqrt{n}}} \leq z_c$, onde $\bar{x}_0$ é a média da amostra, μ_0 é o valor hipotético da média e z_c é o valor da normal padrão para o nível de significância α. A correspondência entre $\bar{x}_c = 591{,}8$ e $z_c = -1{,}64$ pode ser observada na figura.



2º) A região em que se rejeita H_0, quando a média da amostra pertencer a ela, é chamada **região crítica**.

3º) No exemplo ilustrativo acima foi utilizado um teste uni-caudal. Em geral interessa um teste bi-caudal, isto é, testar H_0: $\mu = \mu_0$ contra H_1: $\mu \neq \mu_0$. Neste caso a região crítica é como da figura, isto é, rejeita-se H_0 se $z_0 \leq -z_c$ ou $z_0 \geq z_c$ onde $z_0 = \dfrac{\bar{x}_0 - \mu_0}{\dfrac{\sigma}{\sqrt{n}}}$



4º) Tomando-se o intervalo $-z_c \leq z_0 \leq z_c$, tem-se $-z_c \leq \dfrac{\bar{x}_0 - \mu_0}{\dfrac{\sigma}{\sqrt{n}}} \leq z_c$ ou fazendo $\mu = \mu_0$

$\bar{x} - z_c \dfrac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_c \dfrac{\sigma}{\sqrt{n}}$ que é o intervalo de 1-α de confiança para a média μ.

5º) A distribuição de $\bar{x}$ deve ser normal, ou próxima dela. As hipóteses e o nível de significância do teste devem ser escolhidos antes das observações serem obtidas. As hipóteses sugeridas pelas observações não têm valor científico.

2. TESTE DE UMA MÉDIA

Os passos que compõem o procedimento de um teste de média estão resumidos abaixo. Aqui, foi incluído o caso de não se conhecer o desvio padrão, o que é mais comum na prática.

(I) Enunciar as hipóteses H_0: $\mu = \mu_0$ contra H_1: $\mu \neq \mu_0$ (ou $\mu < \mu_0$, ou ainda, $\mu > \mu_0$)

(II) Fixar o nível de significância α

(III) Determinar a região crítica (região de rejeição de H_0). Se σ for conhecido, usar a variável normal padrão z e se σ for desconhecido usar a variável t de Student com n-1 graus de liberdade.

(IV) Calcular a estatística do teste (t de Student ou normal padrão)

$$\boxed{t_0 = \frac{\bar{x}_0 - \mu_0}{\frac{s}{\sqrt{n}}}}$$

onde μ_0 é o valor hipotético da média μ, enquanto, $\bar{x}_0$, *s* e *n* são, respectivamente, a média, o desvio padrão e o tamanho da amostra.

(V) Se t_0 pertencer à região crítica, rejeitar H_0, caso contrário, aceitar H_0.

Exemplo 1: Em indivíduos sadios, o consumo renal de oxigênio distribui-se normalmente em torno de 12 cm^3/min. Deseja-se investigar, com base em 9 indivíduos portadores de certa moléstia, se esta tem influência sobre o consumo renal de oxigênio. O consumo médio para os 9 pacientes foi $\bar{x}$ =12,84 cm^3/min e o desvio padrão s = 0,9 cm^3/min. Qual a conclusão ao nível de 5% de significância? E ao nível de 1%?
(R: t_0=2,8. Rejeita-se H_0 ao nível de 5% mas não a 1%. O valor de t_c, com 8 g.l., é obtido da tabela anexa: t=2,31 para α=0,05 e t=3,36 para α=0,01)

OBSERVAÇÃO: Devido a facilidade do uso de computadores, vem sendo adotado outro procedimento para a construção da região crítica. Consiste em determinar o **p-valor**. No exemplo anterior, obteve-se a estatística t_0 igual a 2,8. Então, o p-valor corresponde a uma região crítica limitada por -2,8 e 2,8 (se o teste é unicaudal usa-se apenas um desses valores como limite). Neste exemplo, o p-valor é 0,0232 e, portanto, rejeita-se a hipótese nula ao nível de 0,05 de significância, mas não ao nível de 0,01 (faça uma figura para interpretar este resultado)

3. TESTE DE UMA VARIÂNCIA POPULACIONAL

Hipóteses: $\boxed{H_0 : \sigma^2 = \sigma_0^2}$ contra $\boxed{H_1 : \sigma^2 \overset{<}{\underset{>}{\neq}} \sigma_0^2}$

Estatística do teste (qui-quadrado): $\boxed{\chi_0^2 = \frac{(n-1)s_1^2}{\sigma_0^2}}$ com n-1 graus de liberdade, onde n é o tamanho da amostra.

4. TESTE DA DIFERENÇA DE VARIÂNCIAS

Hipóteses: $\boxed{H_0 : \sigma_1^2 - \sigma_2^2 = 0}$ contra $\boxed{H_1 : \sigma_1^2 - \sigma_2^2 \overset{<}{\underset{>}{\neq}} 0}$

Estatística do teste (F de Snedcor): $\boxed{F_0 = \frac{s_1^2}{s_2^2}}$ com $n_1 - 1$ graus de liberdade para o

numerador e $n_2 - 1$ graus de liberdade para o denominador.

Observação: chamamos de s_1^2 a maior das duas variâncias amostrais

5. TESTES DA DIFERENÇA DE MÉDIAS

Pretende-se determinar se existe diferença entre as médias μ_1 e μ_2 (desconhecidas) de duas populações de variâncias σ_1^2 e σ_2^2 (conhecidas ou não).

hipóteses: $\boxed{H_0 : \mu_1 - \mu_2 = 0}$ contra $\boxed{H_1 : \mu_1 - \mu_2 \overset{<}{\underset{>}{\neq}} 0}$

São obtidas duas amostras aleatórias, uma de cada população, de médias $\bar{x}_1$ e $\bar{x}_2$, variâncias s_1^2 e s_2^2 (isto é, desvios padrão s_1 e s_2) e tamanhos n_1 e n_2. Condições: as populações têm distribuição normal ou as amostras são grandes (maiores que 30)

1) Duas amostras independentes e as variâncias populacionais são conhecidas

Estatística do teste (normal padrão): $\boxed{z_0 = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}}$

2) Duas amostras independentes presumindo variâncias populacionais equivalentes

Estatística do teste (t de Student):

$\boxed{t_0 = \dfrac{\bar{x}_1 - \bar{x}_2}{s\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}}$ onde $\boxed{s^2 = \dfrac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$ com $\boxed{gl = n_1 + n_2 - 2}$

3) Duas amostras independentes presumindo variâncias populacionais diferentes

Estatística do teste (t de Student): $\boxed{t_0 = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}}$ com $\boxed{\dfrac{(s_1^2/n_1 + s_2^2/n_2)^2}{\dfrac{(s_1^2/n_1)^2}{n_1 + 1} + \dfrac{(s_2^2/n_2)^2}{n_2 + 1}} - 2}$ graus de liberdade

4) Duas amostras (dependentes) cujos valores podem ser colocados em par.

Obtém-se as diferenças dos n pares de valores $d_i = x_i - y_i$. Calcula-se a média $\bar{d}$ e o desvio padrão s_d.

Estatística do teste (t de Student): $\boxed{t_0 = \dfrac{\bar{d}}{s_d/\sqrt{n}}}$ com n-1 graus de liberdade

6. USANDO O EXCEL

Funções

TESTEZ(**matriz**; μ_0; **sigma**)	*COMPARA UMA MÉDIA COM UM VALOR* μ_0 **matriz** é o intervalo de dados; μ_0 é o valor do teste; **sigma** é o desvio padrão da população (se omitido, o testez usa o desvio padrão da amostra)
TESTET(**matriz1; matriz2; caudas; tipo**)	*COMPARA DUAS MÉDIAS* (usa desvio padrão da amostra) **matriz1** e **matriz2** são os dois conjuntos de dados; se **caudas**=1 retorna o t uno-caudal e se caudas=2 retorna o t bi-caudal; **tipo** se refere ao teste de diferença de médias a ser executado: tipo=1, par, tipo=2, variâncias iguais e tipo=3, variâncias desiguais
TESTEF(**matriz1; matriz2**)	*COMPARA DUAS VARIÂNCIAS* **matriz1** e **matriz2** são os dois conjuntos de dados.

Ferramentas de análise

TesteZ: duas amostras para médias
TesteT: duas amostras presumindo variâncias equivalentes
TesteT: duas amostras presumindo variâncias diferentes
TesteT: duas amostras em par para médias
TesteF: duas amostras para variâncias

PROBLEMAS: (use sempre que possível as fórmulas do capítulo e depois a funções do Excel. Determine os valores críticos das distribuições teóricas de probabilidade tanto pelo Excel como pelas tabelas do apêndice. Depois de resolver a lista toda empregue as ferramentas de análise adequadas)

1) Em relação ao problema apresentado na introdução, suponha que tenha sido obtida uma amostra de 36 latas com os seguintes pesos: **613,6; 581,4; 640,9; 621,8; 635,6; 580,7; 625,2; 541,0; 607,6; 557,6; 593,1; 616,1; 618,5; 591,5; 601,9; 552,9; 583,6; 595,0; 561,7; 602,0; 626,0; 597,8; 597,3; 601,9; 564,6; 561,4; 649,0; 586,6; 572,0; 573,5; 605,7; 607,7; 609,4; 593,7; 599,9; 569,9**. Usando as fórmulas dadas na introdução, pede-se:
   a) ao nível de 1% de significância teste a hipótese de que a média é 600g contra a alternativa de que é menor do que 600g (suponha o desvio padrão populacional igual a 30g).
   b) e ao nível de 5%?
   c) Tomando como base esta amostra, qual o nível de significância abaixo do qual o fabricante teria razão de afirmar que a média é 600g, isto é, abaixo do qual a hipótese nula é aceita?

2) Resolva o problema anterior usando a *função TESTEZ*, Se o desvio padrão populacional, σ= 30g não fosse conhecido, ainda assim poderia ser usada a *função TESTEZ*?

3) Resolva o exemplo 1 tendo sido obtida a seguinte amostra do consumo renal de oxigênio: **12,3; 13,1; 11,9; 11,2; 11,6; 11,9; 11,6; 11,0; 10,5**. Observação: a amostra é proveniente de uma distribuição normal.

4) Determine os intervalos de confiança para a média populacional do consumo renal de oxigênio com os dados do exemplo 1 e com os dados do problema 3. Compare os intervalo de confiança para a média com os intervalos de confiança.

5) Pretende-se testar hipóteses, ao nível de 5% de significância, sobre a variância populacional referente ao problema 3.
   a) Use um teste uni-caudal para verificar se a variância é menor do que 0,6.
   b) Use um teste bi-caudal para verificar se a variância é diferente de 0,6. Observação: na prática, apenas uma dessas hipóteses é testada

6) Uma máquina enche automaticamente latas pequenas com fermento. Em certo dia retira-se 12 latas da produção obtendo-se os seguintes pesos das latas (em gramas): **59,4; 57,4; 60,5; 62,6; 62,3; 63,5; 55,6; 59,5; 62,3; 57,8; 58,6; 56,6**. No dia seguinte retira-se uma amostra de 15 latas obtendo-se os pesos**: 60,5; 58,0; 61,5; 62,9; 56,7; 61,2; 62,3; 60,9; 61,3; 62,1; 63,1; 62,0; 63,7; 60,7; 59,2**.

a) Teste se a variância do primeiro conjunto de dados é maior do que a do segundo, ao nível de 5% de significância? (como a máquina é a mesma, em princípio, a variabilidade deve ser a mesma)
b) Qual o p-valor e o que significa?

7) No problema anterior sabe-se que a variabilidade dos pesos é, em qualquer dia, σ=4 g.

a) Há evidência, ao nível de 5% de significância, de que as médias dos pesos das latas mudaram de um dia para o outro (para mais ou para menos, não importa)?
b) Com essas amostras, até que nível significância a hipótese nula pode ser rejeitada? Na prática, este problema tem sentido porque a máquina pode sofrer uma desregulagem quanto ao peso de enchimento das latas.

8) Responda as questões a) e b) do problema 7 considerando σ desconhecido.

9) Duas máquinas de marcas diferentes estão sendo testadas quanto ao enchimento de latas de fermento. A primeira delas encheu 10 latas dando os pesos: **54,9; 59,0; 57,9; 53,6; 57,3; 56,6; 56,3; 60,4; 57,5; 55,3**. A segunda, mais moderna, encheu também 10 latas com os seguintes pesos: **59,0; 58,9; 58,6; 59,4; 60,6; 60,4; 59,9; 59,1; 58,8; 60,6**.

a) Supõe-se que a precisão da máquina mais moderna é maior do que a outra. Isso é verdade a que nível de significância?
b) Pode-se afirmar que os pesos médios de enchimento da duas máquinas são significativamente diferentes, ao nível de 5%?
c) Qual o maior nível de significância para o qual pode-se afirmar que as médias são diferentes.

10) Sete pessoas obesas foram submetidas a uma determinada dieta de emagrecimento durante um mês. Os pesos, em quilogramas, no início e no fim do tratamento são dados na tabela abaixo.

Indivíduo	1	2	3	4	5	6	7
Peso inicial	178	155	116	188	135	127	162
Peso final	130	141	136	155	128	96	154

a) Ao nível de 5% de significância , pode-se concluir que a dieta é eficiente no emagrecimento de pessoa obesas?
b) E ao nível de 1%?
c) Qual o p-valor?

PROBLEMAS ADICIONAIS DE LIVROS TEXTO

11) Um fabricante de cigarros afirma que seu produto não contém mais que 25 miligramas de nicotina. Uma amostra de 16 cigarros dessa marca revelou uma média de 26,4 e desvio padrão de 2,0 mg de nicotina. Estes dados indicam, com evidência suficiente, que o fabricante está mentindo? Considere α=0,05

12)* Simule uma amostra de 16 valores de nicotina em cigarros, supondo que o teor de nicotina siga uma distribuição normal de média 25 e desvio padrão 2,0 mg. Com base nesta amostra, resolva o problema anterior.

13) Os resíduos industriais jogados nos rios, muitas vezes, absorvem oxigênio, reduzindo

* Problema baseado no problema 1 (não consta do livro)

assim o conteúdo de oxigênio necessário à respiração dos peixes e outras formas de vida aquática. Uma lei estadual exige um mínimo de 5 partes por milhão de oxigênio dissolvido, a fim de que o conteúdo de oxigênio seja suficiente para manter a vida aquática. Seis amostras de água retiradas de um rio de uma localidade específica, durante a maré baixa, revelaram 4,9; 5,1; 4,9; 5,0; 5,0 e 4,7 partes por milhão de oxigênio dissolvido. Estes dados têm evidencia suficiente para assegurar que o conteúdo de oxigênio dissolvido é menor que 5 partes por milhão? Use o nível de significância 0,05.

14) Retorne ao problema anterior. Um fiscal de controle de poluição suspeitou de que esse rio estava recebendo águas semitratadas do esgoto de uma cidade situada à sua margem. Para verificar suas suspeitas, recolheu 5 amostras de água desse rio, em uma localidade situada ao norte e 5 amostras de locais ao sul dessa cidade. Obteve os seguintes dados em partes por milhão (ppm):

Locais ao Norte	4,8	5,2	5,0	4,9	5,1
Locais ao Sul	5,0	4,7	4,9	4,8	4,9

Esses dados indicam evidência suficiente de que o conteúdo médio de oxigênio dissolvido nas águas do trecho do rio que passa nos locais situados ao norte da cidade que está sendo considerada é menor que o conteúdo médio de oxigênio das águas de locais ao sul da cidade? Teste considerando α=0,05

15) Uma das maneiras de manter sob controle a qualidade de um produto é controlar a sua variância. Uma máquina de encher pacotes de café está regulada para enchê-los com um desvio padrão de 10 g e média 500 g. O peso de cada pacote segue uma distribuição normal. Colheu-se uma amostra de 16 pacotes e observou-se uma variância s^2 = 169 g^2. Com esse resultado, você diria que a máquina está desregulada em relação à variância? (nível de 5%)

16) Uma fábrica de embalagens para produtos químicos está estudando dois processos para combater a corrosão de suas latas especiais. Para verificar o efeito dos tratamentos, foram usadas amostras cujos valores estão no quadro abaixo. Qual seria a conclusão sobre os dois tratamentos?

Método	Amostra	Média	Desvio padrão
A	15	48	10
B	12	52	15

17) Para verificar a influência da opção profissional sobre o salário inicial de recém-formados, investigaram-se dois grupos de profissionais: um de liberais em geral e outro de formados em Administração de Empresas. Com os resultados abaixo, expressos em salários mínimos, quais seriam suas conclusões?

Liberais	6,6	10,3	10,8	12,9	9,2	12,3	7,0		
Administradores	8,1	9,8	8,7	10,0	10,2	10,8	8,2	8,7	10,1

18) Um médico deseja saber se uma certa droga reduz a pressão arterial média. Para isso, mediu a pressão arterial de cinco voluntários, antes e depois da ingestão da droga, obtendo os dados do quadro abaixo. Você acha que existe evidência estatística de que a droga realmente reduz a pressão arterial média? Que suposições você fez para resolver o problema?

Voluntário	A	B	C	D	E
Antes	68	80	90	72	80
Depois	60	71	88	74	76

PROBLEMA PROPOSTO

PP5) Encontre na literatura especializada problemas aos quais podem ser aplicados métodos deste capítulo.

VI. COMPARAÇÃO DE VÁRIAS MÉDIAS

1. ANÁLISE DE VARIÂNCIA (ANOVA)

1.1. Classificação simples ou experimento de um fator

Problema ilustrativo: *Uma industria pode optar entre três máquinas distintas, A, B e C para realizar a mesma tarefa e pretende escolher uma delas com base no menor tempo de execução da tarefa.* Supõe-se, neste problema, que o tempo de execução depende de *um único fator*, o tipo de máquina. Este *fator* possui 3 *níveis*: máquina A, máquina B e máquina C.

Para a tomada de decisão, convocaram-se 12 operários, os quais foram divididos aleatoriamente em três grupos de 4 operários, sendo cada grupo designado para executar a tarefa em uma máquina. O tempo, em minutos, gasto pelos operários na execução da tarefa estão na tabela abaixo.

	Máquina			média
	A	B	C	geral
	6,1	5,5	10,0	
	7,0	5,1	9,2	
	8,1	7,8	7,8	
	5,6	6,4	10,2	
média	6,7	6,2	9,3	7,4

Considerando os resultados das máquinas A, B e C como amostras de populações distintas de médias desconhecidas, respectivamente iguais a μ_A, μ_B e μ_C, pretende-se testar a hipótese nula de que essas médias são iguais, contra a hipótese alternativa de que pelo menos duas médias são diferentes entre si. Em símbolos, a hipótese nula é indicada por H_0: $\mu_A=\mu_B=\mu_C=\mu$

De modo geral, o *fator* em estudo é chamado *tratamento*, com k níveis e n repetições em cada nível, dispostos como na tabela abaixo.

repetição	tratamento				média
	1	2	...	k	geral
1	x_{11}	x_{21}		x_{k1}	
2	x_{12}	x_{22}		x_{k2}	
...	...	...		...	
n	x_{1n}	x_{2n}		x_{kn}	
média	$\bar{x}_1$	$\bar{x}_2$		$\bar{x}_k$	$\bar{x}$

Cada x_{ij} representa o valor da repetição j do tratamento i, sendo i=1,2,...,k e j=1,2,...,n. A hipótese nula a ser testada é H_0: $\mu_1=\mu_2= \ldots =\mu_k = \mu$

Para o exposto a seguir deve-se ter: *as repetições nos níveis dos tratamentos são amostras de populações com distribuições normais de variâncias todas iguais a* σ^2.

A base da Análise de Variância está no seguinte: *se a hipótese nula H_0 é verdadeira*, existem três modos de estimar a variância σ^2, comum às k=3 populações.

1º modo) As k=3 amostras podem ser consideradas como provenientes de uma única população de média μ e variância σ^2. Assim, os $kn = 3 \cdot 4 = 12$ valores de tempos de execução da tarefa podem ser reunidos para formar uma só amostra. Com base nesta amostra uma estimativa da variância σ^2, indicada por s^2_{total}, é

$$s^2_{total} = \frac{1}{kn-1}\sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij}-\bar{x})^2 = \frac{1}{11}[(6{,}1-7{,}4)^2 + (7{,}0-7{,}4)^2 + ... + (10{,}2-7{,}4)^2]$$

$$= \frac{33{,}64}{11} = 3{,}0582$$

2º modo) A variância é estimada pelas médias $\bar{x}_1 = 6{,}7$; $\bar{x}_2 = 6{,}2$; $\bar{x}_3 = 9{,}3$ das k=3 amostras, as quais podem ser consideradas como provenientes da mesma população de variância σ^2. Como visto anteriormente, a variância das médias será $\sigma^2_{\bar{x}} = \frac{\sigma^2}{n}$ ou $\sigma^2 = n \cdot \sigma^2_{\bar{x}}$. Uma estimativa de σ^2, indicada por s^2_{entre}, é obtida multiplicando-se n=4 por uma estimativa da variância das k=3 médias amostrais. Obtém-se:

$$s^2_{entre} = \frac{n}{k-1}\sum_{i=1}^{k}[(\bar{x}_i - \bar{x})^2] = \frac{4}{2}[(6{,}7-7{,}4)^2 + (6{,}2-7{,}4)^2 + (9{,}3-7{,}4)^2]$$

$$= \frac{22{,}16}{2} = 11{,}08$$

3º modo) Uma estimativa da variância σ^2 é dada pela média das k=3 variâncias das n=4 amostras. Esta estimativa, indicada por s^2_{dentro}, é

$$s^2_{dentro} = \frac{1}{k}\cdot\frac{1}{n-1}\cdot\sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij}-\bar{x}_i)^2 =$$

$$= \frac{1}{9}[(6{,}1-6{,}7)^2 + ... + (5{,}6-6{,}7)^2 + (5{,}5-6{,}2)^2 + ... + (10{,}0-9{,}3)^2 + ... + (10{,}2-9{,}3)^2]$$

$$= \frac{11{,}48}{9} = 1{,}2756$$

Como o método só é valido quando as variâncias das k populações são iguais a σ^2, esta última estimativa independe de H_0 ser verdadeira. Quando H_0 for <u>falsa</u>, s^2_{entre} tende a estimar um valor maior que σ^2, ou seja, pelo menos uma média populacional deve ser diferente das demais. Portanto, a hipótese original pode ser substituída pela hipótese de que s^2_{entre} e s^2_{dentro} estimem a mesma variância σ^2. Pode-se provar que, se H_0 for <u>verdadeira</u>, as estimativas s^2_{entre} e s^2_{dentro} são independentes e, assim,. é apropriado o teste F para verificar se elas diferem significativamente de 1. Tem-se um F_0 amostral dado por:

$$F_0 = \frac{s^2_{entre}}{s^2_{dentro}} = \frac{11{,}08}{1{,}2756} = 8{,}69$$

Ao nível de 5% de significância, o valor crítico é $F_c = 4{,}26$ (ver tabela anexa) e, então $F_0 > F_c$. Isso quer dizer que s^2_{entre} é significativamente maior do que s^2_{dentro} e, portanto, pelo menos duas médias diferem significativamente entre si, ou seja, rejeita-se H_0. Mais adiante será discutido quais médias são diferentes.

OBSERVAÇÕES IMPORTANTES:

a) Foi realizada uma comparação de variâncias, mas as conclusões de interesse são sobre as médias
b) As três somas que aparecem nas expressões das estimativas das variâncias são chamadas de **Somas de Quadrados** (SQ): SQTotal= 33,64; SQEntre= 22,16 e SQDentro= 11,48. Os denominadores são os **graus de liberdade** (gl) dessas somas, respectivamente, 11, 2 e 9. As estimativas das variâncias, também chamadas de **Médias Quadráticas** (MQ), representam o quociente entre as somas de quadrados e os respectivos graus de

liberdade.

c) Ë válida a seguinte relação: $\boxed{SQDentro = SQTotal - SQEntre}$. Uma relação deste tipo ocorre também entre os graus de liberdade associados a essas somas. Portanto, basta calcular SQTotal e SQEntre e obter SQDentro por subtração. De modo análogo é obtido o $n^{\underline{0}}$ de graus de liberdade associado à SQDentro.

d) Outros nomes são atribuídos às somas de quadrados ou às médias quadráticas
   SQEntre = SQEntre Grupos = SQTratamento = SQMáquinas
   SQDentro = SQDentro de Grupos = SQResíduo = SQErro

e) Deve-se entender que os resultados de um experimento variam por diversos motivos, Na análise de um fator, a variação total é identificada por duas fontes (ou causas): uma devido aos tratamentos (máquinas) e outra, o resíduo (ou erro) que reuni todas as fontes restantes da variação.

f) Os valores necessários à análise costumam ser indicados em uma tabela de Análise de Variância.

Fonte de Variação	SQ	gl	MQ	F_0	F crítico
Máquina	22,16	2	11,08	8,69	4,26
Resíduo	11,48	9	1,2756		
Total	33,64	11			

g) Não é utilizada a Média Quadrática Total porque ela não é independente das demais.

1.2. Classificação dupla ou experimento de dois fatores

Problema ilustrativo: O experimento de um fator do item anterior, onde uma indústria está testando a eficiência de três máquinas, pode ser planejado de forma a isolar, além da variação devida às máquinas, a variação causada pela menor ou maior habilidade individual dos operários. Supõe-se, neste problema, que o tempo de execução depende de *dois fatores*, o tipo de máquina e o operário.

Para a tomada de decisão, são selecionados, por exemplo, 4 operários para atuarem em todas as máquinas. O tempo, em minutos, gasto por cada operário na execução da tarefa nas k=3 máquinas estão na tabela abaixo. Foram usados os mesmos dados do exemplo anterior para efeito de comparação, mas deve-se entender que o planejamento é diferente e os resultados seriam outros.

	Máquina			
Operário	A	B	C	média
1	6,1	5,5	10,0	7,2
2	7,0	5,1	9,2	7,1
3	8,1	7,8	7,8	7,9
4	5,6	6,4	10,2	7,4
média	6,7	6,2	9,3	7,4

Neste caso, existem duas hipóteses nulas a serem testadas, que são: igualdade dos tempos médios de máquina H_{01}: $\mu_A=\mu_B=\mu_C$ e igualdade de tempos médios de operários H_{02}: $\mu_1=\mu_2=\mu_3=\mu_4$.

De modo geral, se um fator possui k níveis e o outro n níveis, os resultados podem ser apresentados como na tabela abaixo.

Na tabela, cada x_{ij} é o resultado de um *tratamento*, o qual corresponde ao nível i do fator 1 (de média $\bar{x}_{i\bullet}$) combinado com o nível j do fator 2 (de média $\bar{x}_{\bullet j}$), sendo i=1,2,...,k e j=1,2,...,n. As hipóteses nulas a serem testadas são H_{01}: $m_{1\bullet} = m_{2\bullet} = ... = m_{k\bullet}$, referente ao fator 1 e H_{02}: $m_{\bullet 1} = m_{\bullet 2} = ... = m_{\bullet n}$, referente ao fator 2.

	Fator 1				
Fator 2	1	2	...	k	média
1	x_{11}	x_{21}	...	x_{k1}	$\bar{x}_{\bullet 1}$
2	x_{12}	x_{22}	...	x_{k2}	$\bar{x}_{\bullet 2}$
...	...	...	...		...
n	x_{1n}	x_{2n}	...	x_{kn}	$\bar{x}_{\bullet n}$
média	$\bar{x}_{1\bullet}$	$\bar{x}_{2\bullet}$	...	$\bar{x}_{k\bullet}$	$\bar{x}$

Sob a hipótese de que as observações são provenientes de uma distribuição normal de variância σ^2 e se as hipóteses nulas forem verdadeiras, esta variância pode ser estimada de quatro formas. Aparecem, agora, duas Somas de Quadrados Entre: a SQEntre Linhas = SQMáquina e a SQEntre Colunas = SQOperário. Assim, as estimativas da variância comum σ^2 são dadas por (as duas primeiras foram calculadas anteriormente)

$$s^2_{total} = \frac{1}{kn-1}\sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij}-\bar{x})^2 = MQTotal = \frac{SQTotal}{kn-1} = \frac{33{,}64}{11} = 3{,}0582$$

$$s^2_{coluna} = \frac{n}{k-1}\sum_{i=1}^{k}[(\bar{x}_{i\bullet}-\bar{x})^2] = \frac{SQColunas}{k-1} = \frac{22{,}16}{2} = 11{,}08$$

$$s^2_{linha} = \frac{k}{n-1}\sum_{j=1}^{n}[(\bar{x}_{\bullet j}-\bar{x})^2] = \frac{SQLinhas}{n-1} =$$

$$= \frac{1}{3}[3(7{,}2-7{,}4)^2 + 3(7{,}1-7{,}4)^2 + 3(7{,}9-7{,}4)^2 + 3(7{,}4-7{,}4)^2] =$$

$$\frac{1{,}14}{3} = 0{,}38$$

$$s^2_{dentro} = s^2_R = MQ\,Re\,síduo = \frac{SQ\,Re\,síduo}{(k-1)(n-1)} = \frac{10{,}34}{6} = 1{,}7233$$

onde $SQ\,Re\,síduo = SQTotal - SQColunas - SQLinhas = 33{,}64 - 22{,}16 - 1{,}14 = 10{,}34$ e o número de graus de liberdade correspondente é igual a $(kn-1)-(k-1)-(n-1) = (k-1)(n-1)$.

A hipótese H_{01}: $\mu_A=\mu_B=\mu_C$ é testada por

$$F_{01} = \frac{MQColuna}{MQ\,Re\,síduo} = \frac{11{,}08}{1{,}7233} = 6{,}43$$

Ao nível de 5% de significância; 2 e 6 graus de liberdade para o numerador e denominador, respectivamente, o F crítico vale $F_{c1} = 5{,}14$. Como $F_{01}>F_{c1}$, rejeita-se H_{01}. Portanto, ao nível de 5% de significância, pode-se concluir que pelo menos um efeito médio de máquina é diferente dos outros.

A hipótese H_{02}: $\mu_1=\mu_2=\mu_3=\mu_4$, por sua vez, é testada pela comparação das Médias Quadráticas Entre Linhas e do Resíduo, ou seja,

$$F_{02} = \frac{MQLinha}{MQ\,Re\,síduo} = \frac{0{,}38}{1{,}7233} = 0{,}22$$

Ao nível de 5% de significância; 3 e 6 graus de liberdade, tem-se F_{c2}=4,76 e não se rejeita a hipótese nula. Portanto, aceita-se que não há diferença significativa nos tempos médios dos operários. Se esta hipótese não for de interesse, não precisa ser testada.

O quadro da análise de variância fica:

Fonte de Variação	SQ	gl	MQ	F_0	F crítico
Máquina	22,16	2	11,08	6,43*	5,14
Operário	1,14	3	0,38	0,22	4,76
Resíduo	10,34	6	1,7233		
Total	33,64	11			

* significativo ao nível de 5%

1.3. Classificação dupla ou experimento de dois fatores, com repetição

Problema ilustrativo: O experimento, apresentado como ilustração deste capítulo, pode ser planejado de modo a medir a interação maquina x operário. Isto é, verificar se os tempos de execução da tarefa sofrem influência da maior ou menor dificuldade que um determinado operário enfrenta ao lidar com alguma máquina.

Por exemplo, observa-se que o operário 4 levou 5,6 min para executar a tarefa na máquina A e um tempo maior, 6,4 min, para executar a mesma tarefa na máquina B. Os outros três operários, ao contrário, levaram mais tempo na máquina A e menos na máquina B. Pode estar havendo uma *interação* dos operários com o tipo de máquina. Para medir esse efeito é necessário que os operários repitam as operações nas máquinas.

Supondo que tenham sido obtidos os resultados da tabela

	Máquina						
Operário	A		B		C		média
1	6,9		6,5		10,9		
	5,3	6,1	4,5	5,5	9,1	10,0	7,2
2	6,0		6,0		10,2		
	8,0	7,0	4,2	5,1	8,2	9,2	7,1
3	9,2		6,8		8,6		
	7,0	8,1	8,8	7,8	7,0	7,8	7,9
4	6,5		5,6		11,1		
	4,7	5,6	7,2	6,4	9,3	10,2	7,4
Média		6,7		6,2		9,3	7,4

$$s^2_{total} = \frac{1}{24-1}[(6,9-7,4)^2 + (5,3-7,4)^2 + (6,0-7,4)^2 + ... + (9,3-7,4)^2] = \frac{88,02}{23} = 3,6675$$

$$s^2_{coluna} = \frac{1}{3-1}8[(6,7-7,4)^2 + (6,2-7,4)^2 + ... + (9,3-7,4)^2] = \frac{44,32}{2} = 22,1600$$

$$s^2_{linha} = \frac{1}{4-1}6[(7,2-7,4)^2 + (7,1-7,4)^2 + ... + (7,4-7,4)^2] = \frac{2,28}{3} = 0,7600$$

$$s^2_{interação} = \frac{1}{(3-1)(4-1)}2[(6,1-6,7-7,2+7,4)^2 + (7,0-6,7-7,1+7,4)^2 + ... + (10,2-9,3-7,4+7,4)^2]$$

$$= \frac{20,68}{6} = 3,4467$$

A Soma de Quadrados do Resíduo e os correspondentes graus de liberdade são obtidos por subtração da Soma de Quadrados Total. Assim, o quadro da análise da variância fica

Fonte de Variação	SQ	gl	MQ	F_0	$F_{crítico}$ ($F_{5\%}$)
Máquina (coluna)	44,32	2	22,1600	12,82*	3,89
Operário (linha)	2,28	3	0,7600	0,45	3,49
Interação	20,68	6	3,4467	1,99	3,00
Resíduo	20,74	12	1,7283		
Total	88,02	23			

* significativo ao nível de 5%

Portanto, há apenas efeito de máquina e a conclusão deve ser a mesma obtida anteriormente. Não há efeito significativo de interação e então os resultados do operário 4 em relação às máquinas A e B, discutido acima, foram diferentes por puro acaso.

Quando a interação é significativa, o comportamento de um fator depende dos níveis do outro e a análise deve ser mudada.

2. COMPARAÇÕES MÚLTIPLAS

Quando a ANOVA identifica diferenças entre médias, pode-se determinar quais são diferentes pelo método de Scheffé. Duas médias $\bar{x}_p$ e $\bar{x}_q$, de duas linhas (ou colunas) p e q são consideradas distintas se sua diferença, em valor absoluto, for maior do que uma diferença mínima significativa (DMS), isto é, se

$$\left|\bar{x}_p - \bar{x}_q\right| > \text{DMS}$$

Para o cálculo da diferença mínima significativa tem-se:

a) Experimento de um fator, com k tratamentos e n_p e n_q repetições para as médias $\bar{x}_p$ e $\bar{x}_q$, respectivamente.

$$\text{DMS} = \sqrt{(\frac{1}{n_p} + \frac{1}{n_q})(k-1).(\text{MQ Re síduo})\, F_{(k-1);(n-k)}}$$

onde o índice de F indica os graus de liberdade, isto é, F é calculado com (k-1) e (n-k) g.l. Se o $n^{\underline{0}}$ de repetições é o mesmo (n), então

$$\text{DMS} = \sqrt{\frac{2}{n}(k-1).(\text{MQ Re síduo})\, F_{(k-1);(n-k)}}$$

b) Experimento de dois fatores (sem repetição). Sejam n_A e n_B os $n^{\underline{0s}}$ de níveis dos fatores A e B. Para comparar as médias do fator A duas a duas, tem-se

$$\text{DMS} = \sqrt{\frac{2}{n_B}(n_A - 1).(\text{MQ Re síduo})\, F_{(nA-1);(nA-1)(nB-1)}}$$

Para o fator B, muda-se n_A por n_B e vice-versa.

c) Experimento de dois fatores (com repetição). n_A e n_B têm o mesmo significado anterior e r é o $n^{\underline{0}}$ de repetições. Para médias do fator A, tem-se

$$\text{DMS} = \sqrt{\frac{2}{n_B r}(n_A - 1).(\text{MQ Re síduo})\, F_{(nA-1);nAnB(r-1)}}$$

Para o fator B, muda-se A por B e vice-versa.

No experimento de um fator que compara as máquinas, tem-se

$\text{DMS} = \sqrt{\frac{2}{4}(3-1)(4{,}26)(1{,}2756)} = \sqrt{5{,}4341} = 2{,}33$. Duas médias são significativamente distintas se a diferença entre elas (em valor absoluto) for maior do que 2,33. Então

$\left|\bar{x}_A - \bar{x}_B\right| = \left|6{,}7 - 6{,}2\right| = 0{,}5$

$\left|\bar{x}_A - \bar{x}_C\right| = \left|6{,}7 - 9{,}3\right| = 2{,}6$ Significativa ao nível de 5%

$\left|\bar{x}_B - \bar{x}_C\right| = \left|6{,}2 - 9{,}3\right| = 3{,}1$ Significativa ao nível de 5%

Conclusão: A média da máquina C é significativamente distinta das demais. A máquina C é a menos eficiente, porque os operários levam, em média, mais tempo para executarem a tarefa com ela.

4. USANDO O EXCEL

Ferramentas

Anova: fator único Anova: fator duplo sem repetição Anova: fator duplo com repetição	Obs: O número de repetições dos tratamentos não precisam ser iguais Neste caso os rótulos de linha e coluna são obrigatórios

PROBLEMAS:

1) Um experimento foi desenvolvido para testar o efeito de dois fatores sobre um produto agrícola: fertilizante (F) e irrigação (A), cada um em dois níveis (ausente e presente). As produções resultantes (em uma certa unidade) são apresentadas na tabela, onde o índice 0 indica a ausência do fertilizante ou irrigação e o índice 1 indica a presença.

	Irrigação	
Fertilizante	A_0	A_1
F_0	9	14
	15	18
	12	16
F_1	10	27
	8	22
	12	23

a) Considere os resultados dos quatro tratamentos A_0F_0, A_1F_0, A_0F_1 e A_1F_1 como de um delineamento de um fator e faça a análise de variância usando as fórmulas apropriadas.
b) Considere agora o delineamento como de dois fatores com repetição. Faça a análise de variância, também usando as fórmulas apropriadas deste capítulo (o procedimento utilizado no item *a* só é correto se a interação não for significativa)

2) Na tabela é apresentado o consumo de gasolina (km/L) de duas marcas de automóveis, que em um mesmo trajeto, perfazendo a mesma quilometragem, trafegaram somente na rodovia, somente na cidade, na rodovia e cidade. Empregando as fórmulas apropriadas faça análise de variância.

	automóvel A	automóvel B
Rodovia	14,0	13,8
Cidade	8,7	9,7
Rodovia/cidade	11,2	11,0

3) Resolva os problemas 1 e 2 usando as ferramentas de análise do Excel.

PROBLEMAS ADICIONAIS DE LIVROS TEXTO

MENDENHALL, W. *Probabilidade e Estatística.* Vol 2. Rio de Janeiro: Ed. Campus, 1985.

4) Realizou-se uma experiência a fim de examinar o efeito da idade sobre o número de batidas do coração, quando uma pessoa é submetida a certo tipo de exercício. Dez homens foram aleatoriamente escolhidos nas faixas etárias de 10-19, 20-39, 40-50 e 60-69 anos. Cada um andou sobre uma pista fixa (comandada pelo movimento dos pés) durante 12 minutos, numa intensidade pré-determinada. O aumento das batidas do coração de cada pessoa (as diferenças entre os totais antes e depois do exercício) foi anotado para cada homem, obtendo-se os resultados da tabela (em batidas por/minuto)

Esses dados apresentam evidência suficiente para indicar uma diferença entre o aumento médio de batidas para os quatro grupos?

Faixa etária			
10-19	20-39	40-59	60-69
29	24	37	28
33	27	25	29
26	33	22	34
27	31	33	36
39	21	28	21
35	28	26	20
33	24	30	25
29	34	34	24
36	21	27	33
22	32	33	32

VIEIRA S. Bioestatística, 1987

5) Na tabela são apresentadas as taxas de glicose, em miligramas por 100 ml de sangue, segundo o grupo, em ratos machos da raça Wistar, com 60 dias de idade. Testar a hipótese de que as médias relativas aos três grupos são iguais.

Grupo		
Parotidec-tomizado	Pseudo parotidectomizado	Normal
96,0	90,0	86,0
95,0	93,0	85,0
100,0	89,0	105,0
108,0	88,0	105,0
120,0	87,0	90,0
110,5	92,5	100,0
97,0	87,5	95,0
92,5	85,0	95,0

6) A tabela apresenta valores de pressão arterial de 6 cães decorridos 20, 40 e 60 minutos após a administração de 10 mg de prilocaína por quilo de peso vivo. Testar a hipótese de que a pressão arterial não se altera, quer decorridos 20, 40 ou 60 minutos após a administração de prilocaína.

	Tempo decorrido		
Cão	20	40	60
1	62	62	62
2	110	110	110
3	140	155	150
4	85	90	100
5	140	125	130
6	95	90	70

7) Realizou-se um experimento para investigar o efeito tóxico de 3 produtos químicos, A, B e C, sobre a pele de ratos. Uma polegada quadrada da pele de cada rato foi tratada com os três produtos, medindo-se a irritação resultante por escores de 0 a 10. Foram marcadas 3

áreas de uma polegada quadrada em cada um de 8 ratos, aplicando-se um produto a uma área de cada rato. Por conseguinte, a experiência foi feita em blocos, visando-se eliminar a variação da sensibilidade da pele de rato para rato. Os dados obtidos foram:

			Rato				
1	2	3	4	5	6	7	8
B 5	A 9	A 6	C 6	B 8	C 5	C 5	B 7
A 6	C 4	B 9	B 8	C 8	A 5	B 7	A 6
C 3	B 9	C 3	A 5	A 7	B 7	A 6	C 7

Esses dados têm evidencia suficiente que garanta haver diferença entre o efeito tóxico desses produtos? (nível de 5%. E a 1%)?

BEIGUELMAN, B. *Curso prático de Bioestatística*. Ribeirão Preto: Revista Brasileira de Genética, 1991.

8) Numa pesquisa para investigar os efeitos dos fatores alcoolismo e esforço físico sobre a produção de um determinado metabólito, tomaram-se duas amostras, uma de 20 alcoólatras e outra de 20 abstêmios, todos adultos e do sexo masculino. Em cada uma delas fez-se o sorteio de 10 indivíduos mantidos em repouso e de 10 indivíduos mantidos em pé durante quatro horas. Os resultados da pesquisa estão apresentados na tabela

	Repouso	Atividade
Alcoólatras	4,41	5,51
	3,43	0,64
	3,74	2,87
	0,67	0,51
	3,37	2,59
	2,94	0,32
	0,53	0,71
	3,4	0,68
	0,71	3,91
	4,71	2,87
Abstêmicos	6,75	6,92
	3,98	2,73
	6,2	6,01
	2,81	2,01
	5,32	6,04
	5,01	2,9
	2,67	1,94
	4,01	2,01
	2,8	5,42
	6,84	4,33

VIEIRA, S.; HOFFMANN, R. Estatística Experimental. São Paulo: Atlas, 1989.

9) Um professor conduziu um experimento para comparar a eficiência de quatro fontes de informação: jornais, televisão, revistas e rádio. Participaram desse experimento 24 alunos. Como os alunos eram de idades diferentes, o professor separou os alunos em dois blocos, de acordo com a faixa de idade. Depois sorteou, dentro dos blocos, uma fonte de informação para cada aluno. Os alunos então se submeteram ao experimento, isto é, tomaram conhecimento sobre determinado assunto apenas pela fonte de informação que lhes havia sido sorteada. Depois, fizeram um teste de conhecimento (em uma escala de 0 a 100) e as notas estão na tabela

	Jornal	TV	Rev.	Rádio
Faixa etária I	65	56	58	38
	69	49	65	30
	73	54	57	34
Faixa etária II	72	73	76	71
	79	77	69	65
	80	69	71	62

PROBLEMA PROPOSTO

PP6) Encontre na literatura especializada problemas aos quais podem ser empregados métodos deste capítulo.

VII. REGRESSÃO E CORRELAÇÃO

1. REGRESSÃO LINEAR SIMPLES

1.1. A reta de regressão

Problema ilustrativo 1: Um motorista submeteu-se a um teste onde deveria percorrer um trajeto a uma velocidade constante, durante determinado tempo. Ele não conseguiu manter exatamente uma velocidade constante, algumas vezes precisou aumentar a velocidade e outras diminuir. As distâncias percorridas de acordo com o tempo, em minutos, estão na tabela abaixo.

x= tempo (min)	0	1	2	3	4	5
y= Distância percorrida (km)	0	1,3	3,8	4,3	6,7	7,3

Sabe-se, da Física, que há uma relação linear entre a distância y* percorrida por um carro em velocidade exatamente constante e o tempo de deslocamento, chamado movimento uniforme, dada por

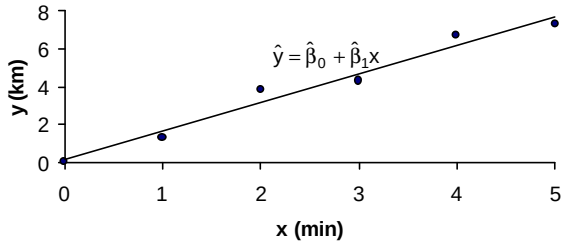
$$y^* = \beta_0 + \beta_1 x$$

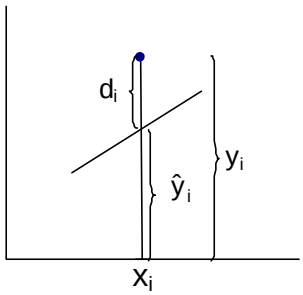
onde β_0 é o coeficiente linear da reta, representando a distância que o carro já havia percorrido quando x=0 , e β_1 é o coeficiente angular da reta, representando a velocidade constante com que o carro está se deslocando. Esta relação fornece um modelo matemático para descrever a distancia percorrida por um carro em movimento uniforme.

Neste exemplo, a velocidade não é constante e pretende-se estudar o movimento a partir dos dados experimentais. Considerando-se que o modelo acima é válido para cada par de valores conhecidos $(x_i; y_i)$, exceto por um erro experimental u_i , tem-se:

$$y_i = \beta_0 + \beta_1 x_i + u_i \qquad (i=1,2,...,n)$$

O erros u_i dependem dos valores dos parâmetros β_0 e β_1, que não são conhecidos exatamente. Então, com base nas observações experimentais é preciso um modo de determinar valores aproximados, chamados ***estimativas***, de β_0 e β_1, indicadas respectivamente por $\hat{\beta}_0$ e $\hat{\beta}_1$. O método mais empregado é o ***método dos mínimos quadrados***, descrito a seguir.





As estimativas $\hat{\beta}_0$ e $\hat{\beta}_1$ são os coeficientes de uma reta que se ajusta aos pontos experimentais, conforme a figura, tal que

$$\boxed{\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i}$$

A diferença entre cada y_i experimental e cada $\hat{y}_i$ da reta é chamado desvio ou resíduo. Então, cada desvio ou resíduo d_i é dado por: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

$$d_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

De todas as retas que podem ser traçadas entre os pontos experimentais, a reta que usa as

estimativas de mínimos quadrados $\hat{\beta}_0$ e $\hat{\beta}_1$ é a que dá a menor soma de quadrados dos resíduos.

Pode-se calcular $\hat{\beta}_0$ e $\hat{\beta}_1$ pelos métodos do Cálculo Diferencial e Integral, determinando o mínimo da função Soma de Quadrados (SQ) seguinte:

$$SQ = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Obtém-se

$$\boxed{\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}} \quad \text{e} \quad \boxed{\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}}$$

onde $\bar{y}$ e $\bar{x}$ são as médias dos n valores y_i e x_i, respectivamente.

Exemplo 1: Na tabela estão expostos os cálculos necessários à determinação da reta de mínimos quadrados para os pontos do problema ilustrativo inicial (tempo em minutos e deslocamento em km).

	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
	0	0,0	-2,5	-3,9	6,25	9,75
	1	1,3	-1,5	-2,6	2,25	3,90
	2	3,8	-0,5	-0,1	0,25	0,05
	3	4,3	0,5	0,4	0,25	0,20
	4	6,7	1,5	2,8	2,25	4,20
	5	7,3	2,5	3,4	6,25	8,50
Soma	15	23,4			17,5	26,6
Média	2,5	3,9				

Portanto,

$$\hat{\beta}_1 = \frac{26,6}{17,5} = 1,52 \text{ km/min} = 91,2 \text{ km/h} \quad \text{e} \quad \hat{\beta}_0 = 3,9 - (1,52)(2,5) = 0,1 \text{ km}$$

A equação da reta de regressão que melhor descreve a distância percorrida em função do tempo, tomando a velocidade como constante, é

$$\boxed{\hat{y} = 0,1 + 1,52x}$$

A partir dessa reta pode-se predizer a distância percorrida em qualquer tempo x. Assim, depois x= 2,5 min toma-se como distância percorrida $\hat{y} = 0,1 + 1,52(2,5) = 3,9$ km. Ou para x= 4 min tem-se: $\hat{y}$ = 4,66 km. O cálculo da distância percorrida após 5 min depende do modelo continuar válido.

1.2. Suposições sobre o termo de erro

Para introduzir as técnicas estatísticas, deve-se considerar que o experimento realizado é apenas uma amostra de uma população de resultados. Essa amostra poderia ser uma entre os possíveis resultados se o mesmo motorista repetisse o teste, ou poderia ser uma amostra tomada com um motorista dentre um conjunto grande de motoristas, dependendo do objetivo do experimento.

Assim, adota-se o modelo

$$y_i = \beta_0 + \beta_1 x_i + u_i \qquad (i=1,2,...,n)$$

onde x_i representa valores estabelecidos a *priori*, isto é, são valores fixos, e os y_i são valores

de uma variável aleatória. Nessas condições, supõe-se que o erro é uma variável aleatória de média zero e variância constante σ^2. Uma estimativa dessa variância é dada pelo quociente entre a soma de quadrados dos desvios (ou resíduos) por (n-2) graus de liberdade (2 é o número de parâmetros)

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}$$

Esta variância residual é a variância em torno da reta de regressão.

Exemplo 2: Na tabela abaixo são apresentados os valores ***previstos*** para y pela reta de regressão nos tempos de 0 a 5 minutos, os ***resíduos*** (ou desvios da regressão) e os ***resíduos padrão***.

Tempo	Distância real	Distância prevista	Resíduo	Resíduo Padrão
0	0,0	0,10	-0,10	-0,1900
1	1,3	1,62	-0,32	-0,6080
2	3,8	3,14	0,66	1,2540
3	4,3	4,66	-0,36	-0,6840
4	6,7	6,18	0,52	0,9880
5	7,3	7,70	-0,40	-0,7600

A *Soma de Quadrados dos Resíduos* **(SQRes)**,

$$SQRes = (-0,10)^2 + (-0,32)^2 + ... + (-0,40)^2 = 1,1080$$

dividida por n-2= 4 dá a *Média Quadrática dos Resíduos* (**MQRes**), que é uma estimativa da variância do erro experimental

$$s^2 = MQRes = \frac{SQ\,Res}{n-2} = \frac{1,1080}{4} = 0,2770$$

A raiz quadrada da MQRes. é chamada de ***Erro padrão***.

$$s = \sqrt{MQRes} = \sqrt{0,2770} = 0,5263$$

Os ***Resíduos padrão***, também apresentados na tabela, são obtidos pelo quociente dos resíduos pelo erro padrão. É uma forma de obter resíduos sem uma unidade de medida.

1.3. Intervalos de confiança para os parâmetros

Como visto acina, o resultado do teste realizado pelo motorista é apenas uma amostra de uma infinidade de resultados possíveis. Portanto, existe uma reta ideal com os parâmetros β_0 e β_1, que seriam obtidos se a velocidade fosse constante. Como esses parâmetros são desconhecidos, procura-se determinar intervalos nos quais deposita-se uma confiança de 1-α de contê-los. Ou seja, o processo é tal que em (1-α)100% dos testes que forem realizados obtêm-se intervalos que contêm esses valores ideais.

Os intervalos de confiança para β_0 e β_1 são da forma:

Estimativa do parâmetro $\pm$ t_c * erro padrão do parâmetro

onde t_c é o valor da distribuição t de Student com n-2 graus de liberdade e os erros padrão serão definidos abaixo.

Prova-se que, se o erro experimental tem distribuição normal de média zero e variância

σ^2, estimada pela variância residual s^2, as estimativas dos parâmetros também têm distribuição normal.

Um intervalo de (1-α) de confiança para o coeficiente linear β_0 é

$$\boxed{\hat{\beta}_0 \pm t_c\, s(\hat{\beta}_0)} \text{ onde } \boxed{s(\hat{\beta}_0) = \sqrt{s^2\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}\right]}}$$

e para β_1 é

$$\boxed{\hat{\beta}_1 \pm t_c s(\hat{\beta}_1)} \text{ onde } \boxed{s(\hat{\beta}_1) = \sqrt{\frac{s^2}{\sum (x_i - \bar{x})^2}}}$$

Exemplo 3: Considerando o problema ilustrativo, tem-se

$$s(\hat{\beta}_0) = \sqrt{0{,}2770\left(\frac{1}{6} + \frac{(2{,}5)^2}{17{,}5}\right)} = \sqrt{0{,}1451} = 0{,}3809$$

Ao nível de 5% de significância, com n-2=4 g.l., t_c = 2,7765 e um intervalo de 95% de confiança para o coeficiente linear β_0 é dado por:

$0{,}1 \pm 2{,}7765(0{,}3809) = 0{,}1 \pm 1{,}06$, ou seja, -0,96<$\beta_0$<1,16 (unidade km)

Para β_1 tem-se

$$s(\hat{b}_1) = \sqrt{\frac{0{,}277}{17{,}5}} = \sqrt{0{,}0158} = 0{,}1258$$

e um intervalo de 95% de confiança para o coeficiente angular é dado por

$1{,}52 \pm 2{,}7765(0{,}1258) = 1{,}52 \pm 0{,}35$, ou seja, 1,17<$\beta_0$<1,87 (em km/min)

Em km/h tem-se 70,2<β_0<112,2

1.4. Testes de hipóteses sobre os parâmetros

Pode-se testar hipóteses sobre β_0 e β_1 usando a distribuição t de Student, com n-2 graus de liberdade, ao nível de significância α.

Para testar $H_0 : \beta_0 = \beta_0^*$, a estatística é $\boxed{t_0 = \frac{\hat{\beta}_0 - \beta_0^*}{s(\hat{\beta}_0)}}$

e

para testar $H_0 : \beta_1 = \beta_1^*$ a estatística é $\boxed{t_0 = \frac{\hat{\beta}_1 - \beta_1^*}{s(\hat{\beta}_1)}}$

Exemplo 4: Considerando o problema ilustrativo, pretende-se testar $H_0 : \beta_0 = 0$ (se a reta passa pela origem) e $H_0 : \beta_1 = 0$ (se há regressão)

Nos dois casos, ao nível de 5% de signicância, com n-2=4 g.l., t_c = 2,7765

Para o coef. linear, $t_0 = \frac{0{,}1}{0{,}3809} = 0{,}2625$ e, portanto, aceita-se H_0

Para o coef. angular, $t_0 = \frac{1{,}52}{0{,}1258} = 12{,}081$ e rejeita-se H_0, comprovando que há regressão de y sobre x.

1.5. Intervalo de confiança para $\beta_0 + \beta_1 x_0$ e intervalo de previsão

A um valor x_0 de x corresponde na reta de regressão o valor

$$\hat{y}_0 = \beta_0 + \beta_1 x_0$$

sendo $\hat{y}_0$ uma estimativa de $y_0^* = \beta_0 + \beta_1 x_0$ da reta verdadeira. Um intervalo de confiança para y_0^* é dado por

$$\boxed{\hat{y}_0 \pm t_c s(\hat{y}_0)} \text{ com } \boxed{s(\hat{y}_0) = \sqrt{s^2[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}]}}$$

onde t_c é o valor da distribuição t de Student com n-2 g.l., ao nível de significânca α.

Um intervalo de previsão é um intervalo que, com uma confiança (1-α), contem um próximo valor experimental y_0 correspondente a x_0. É dado por

$$\boxed{\hat{y}_0 \pm t_c \sqrt{s^2[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}]}}$$

Exemplo 5: Considerando o problema ilustrativo, pretende-se calcular um intervalo de confiança para o valor na reta verdadeira e um intervalo de previsão para um valor experimental correspondentes a $x_0 = 2$, ao nível de 95%.

Para x_0=2, $\hat{y}_0 = 0{,}1 + 1{,}52(2) = 3{,}14$ e considerando os valores calculados anteriormente t_c=2,7765; $\bar{x} = 2{,}5$; $s^2 = 0{,}2770$ e $\sum (x_i - \bar{x})^2 = 17{,}5$, tem-se

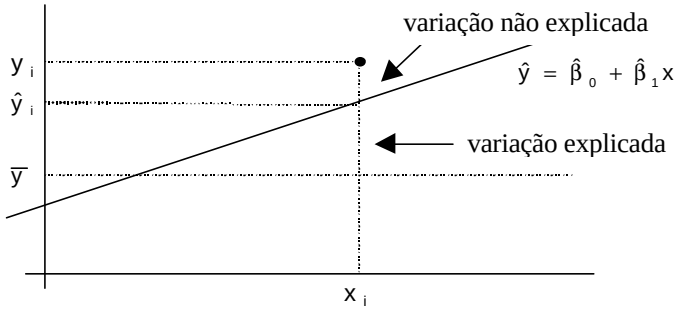
Intervalo de 95% de confiança para o valor na reta (em km)

$$3{,}14 \pm 2{,}7765 \sqrt{0{,}2770[\frac{1}{6} + \frac{(2 - 2{,}5)^2}{17{,}5}]} = 3{,}14 \pm 0{,}18$$

intervalo de previsão (em km)

$$3{,}14 \pm 2{,}7765 \sqrt{0{,}2770[1 + \frac{1}{6} + \frac{(2 - 2{,}5)^2}{17{,}5}]} = 3{,}14 \pm 1{,}59$$

1.6. Análise de variância aplicada à regressão



Ajustada a reta de regressão, definem-se:

Variação total de y, independente de x: $SQTotal = \sum_{i=1}^{n} (y_i - \bar{y})^2$

Variação explicada pela regressão $SQRegr = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$

Variação residual (variação não explicada pela regressão)

$$SQRes = \sum_{i=1}^{n} (y_i - \hat{y})^2$$

Pode-se provar que SQTotal = SQRegr + SQRes, ou seja, a variação total pode ser dividida em duas parcelas, uma correspondente à variação explicada pela reta de mínimos quadrados e outra residual, devida à variação do acaso.

De acordo com esta expressão, não havendo regressão, a variação total é praticamente igual a variação residual e, então, a variância do erro experimental pode ser estimada tanto pela variação total

$$s^2_{Total} = \frac{\sum (y_i - \bar{y}_i)^2}{n-1} = \frac{SQTotal}{n-1}$$

como pela variação residual, vista anteriormente e que independe de haver regressão,

$$s^2 = s^2_{Res} = \frac{\sum (y_i - \hat{y}_i)^2}{n-2} = \frac{SQ\,Res}{n-2} = MQ\,Res$$

Se a hipótese nula é falsa,

$$s^2_{Regr} = \frac{\sum (\hat{y}_i - \bar{y}_i)^2}{2-1} = MQ\,Regr$$

tende a crescer e prova-se que o quociente de s^2_{regr} por s^2_{Res} tem distribuição F. Então a estatística do teste é:

$$\boxed{F_0 = \frac{s^2_{Regr}}{s^2_{Res}} = \frac{MQ\,Regr}{MQ\,Res}}$$

Este teste é equivalente ao teste de hipótese para o coeficiente angular dado em 1.4.

Exemplo 6: Considerando o problema ilustrativo, pretende-se efetuar a ANOVA. Aproveitando resultados das páginas 48 e 49, tem-se

$SQTotal = (-3{,}9)^2 + (-2{,}6)^2 + ... + (3{,}4)^2 = 41{,}5400$

SQRes = 1,1080 e, portanto,

SQRegr = 41,5400 – 1,1080 = 40,4320

O quadro da análise de variância fica

Fonte de Variação	SQ	gl	MQ	F_0	$F_{5\%}$
Regressão	40,4320	1	40,4320	145,96	7,71
Resíduo	1,1080	4	0,2770		
Total	41,5400	5			

Conclui-se que, ao nível de 5% de significância, existe regressão de y sobre x.

2. CORRELAÇÃO LINEAR E COEFICIENTE DE DETERMINAÇÃO

Dadas duas variáveis x e y, das quais se conhecem n valores, tem-se:

Variância de x: $s^2_x = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ ==> desvio padrão de x é s_x

Variância de y: $s^2_y = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$ ==> desvio padrão de y é s_y

Covariância de x e y: $s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$

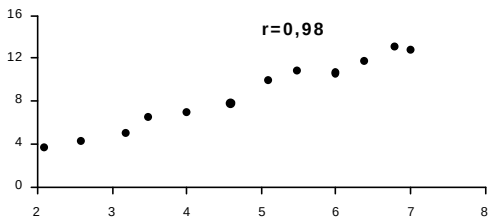
Uma medida do grau de associação linear entre as duas variáveis, que independe das

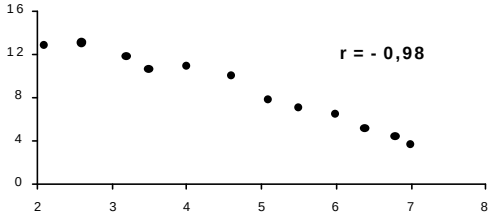
unidades de medidas de x e y, é o **coeficiente de correlação**, r, dado por

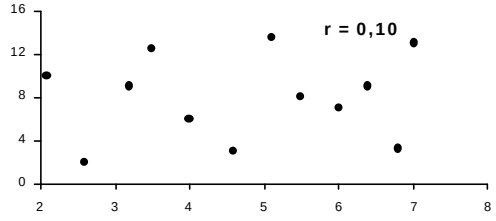
$$r = \frac{s_{xy}}{s_x \cdot s_y}$$

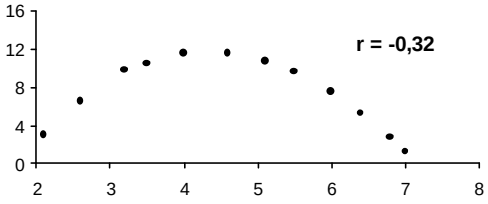
O coeficiente de correlação r varia de –1 a 1 e quanto mais próximo de –1 ou 1, maior será a associação linear entre x e y

Nas figuras abaixo são apresentados alguns conjuntos de pontos experimentais e o coeficiente de correlação linear









Define-se o ***coeficiente de determinação*** r^2 por

$$r^2 = \frac{\text{variação explicada}}{\text{variação total}}$$

$$r^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{SQ\,Regressão}{SQTotal}$$

O coeficiente de determinação pode ser interpretado como a proporção da variação total na variável y que é explicada pela reta de regressão. Ele é o quadrado do coeficiente de correlação r. O coeficiente de correlação é indicado para medir o grau de associação linear entre duas variáveis, enquanto o coeficiente de determinação é mais apropriado para definir quanto a reta de regressão explica o ajuste da reta.

Exemplo 7: Considerando o problema ilustrativo, aproveitando os cálculos anteriores, tem-se

$$s_x^2 = \frac{17,5}{5} = 3,5 \;;\; s_{xy} = \frac{26,6}{5} = 5,32 \text{ e } s_y^2 = \frac{41,54}{5} = 8,308$$

Portanto, o coeficiente de correlação entre x e y é

$$r = \frac{5,32}{\sqrt{3,5}\sqrt{8,308}} = 0,9866$$

o coeficiente de determinação é

$r^2 = (0,9866)^2 = 0,9734$ **,** ou usando os resultados do quadro da análise de variância

$$r^2 = \frac{40{,}4320}{41{,}5400} = 0{,}9754$$

Isso significa que 97,54% da variação total é explicada pela regressão.

3. REGRESSÃO MÚLTIPLA

O modelo de regressão múltipla envolve mais do que uma variável independente x. É da forma

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \text{erro}$$

onde os parâmetros são estimados pelo método dos mínimos quadrados, isto é, as estimativas minimizam a soma de quadrados dos resíduos

$$SQRes = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i} - ... - \hat{\beta}_{ki})^2$$

Problema ilustrativo 2: (apresentado no Excel) Suponha que um empresário esteja pensando em comprar um grupo de prédios de salas comerciais em um bairro comercial. O empresário pode usar a análise de regressão linear múltipla para fazer uma estimativa do valor de um prédio em uma determinada área, de acordo com as variáveis a seguir

Variável	refere-se a
y	valor estimado do prédio
x_2	área útil em metros quadrados
x_3	número de salas
x_4	número de entradas
x_5	idade do prédio em anos

Este exemplo considera que existe uma relação de linha reta entre cada uma das variáveis independentes (x_1, x_2, x_3 e x_4) e a variável dependente (y), o valor dos prédios comerciais no bairro. O empresário escolhe aleatoriamente uma amostra de 11 prédios a partir de um conjunto de 1500 prédios possíveis e obtém os seguintes dados ("Meia entrada" significa que o prédio só dispõe de uma entrada para entregas):

Área	salas	entradas	idade	valor(R$ 1000)
2310	2	2	20	142
2333	2	2	12	144
2356	3	1,5	33	151
2379	3	2	43	150
2402	2	3	53	139
2525	4	2	23	169
2448	2	1,5	99	126
2471	2	2	34	142
2494	3	3	23	163
2517	4	4	55	169
2540	2	3	22	149

Exemplo 8: Considerando o problema ilustrativo 2, obtém-se pelo Excel

$$y = 27{,}64*x_1 + 12.530*x_2 + 2.553*x_3 - 234{,}24*x_4 + 52.318$$

Agora, o empresário poderá fazer uma estimativa do valor de um prédio na mesma área com 272 metros quadrados, três salas e duas entradas, e que tem 25 anos de idade, usando a seguinte equação:

y = 27,64 *272 + 12.530*3 + 2.553*2 - 234,24*25 + 52.318 = $158.261

Exemplo 9: Fazendo a análise de variância obtém-se os resultados apresentados no quadro abaixo. O nº de graus de liberdade para a regressão é igual a p-1, onde p é o nº de parâmetros e para a regressão, n-p. Neste exemplo n = 11 e p = 5.

Fonte de Variação	SQ	gl	MQ	F_0	$F_{5\%}$
Regressão	1741,863	4	435,4658	294,76	4,53
Resíduo	8,8640	6	1,4773		
Total	1750,727	10			

A regressão é altamente significativa.

4. CORRELAÇÃO LINEAR MÚLTIPLA

Para calcular **o coeficiente de correlação múltipla** de y sobre x_1, x_2, ..., x_k usa-se o coeficiente de determinação:

$$r\ (\text{múltiplo}) = \sqrt{\frac{\text{variação explicada}}{\text{variação total}}} = \sqrt{r^2}$$

O coeficiente de determinação recebe um ajuste quando se emprega a regressão múltipla. O **coeficiente de determinação ajustado** é dado por

$$r^2(\text{ajust}) = r^2 - \frac{k}{n-k-1}(1-r^2)$$

onde n é o número de observações e k o número de variáveis independentes.

Exemplo 10: No problema ilustrativo 2, o coeficiente de correlação linear múltipla de y em relação a x_1, x_2, x_3 e x_4 é

$$r^2 = \frac{1741,863}{1750,727} = 0,9949 \text{ e}$$

$$r^2(\text{ajustado}) = 0,9949 - \frac{4}{6}(1 - 0,9949) = 0,9916$$

Portanto, a equação obtida explica 99,16% da variação de y.

5. USANDO O EXCEL

Funções

INTERCEPÇÃO(valores y; valores x)	estimativa $\hat{\beta}_0$ do coef. linear β_0 ($\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$)
INCLINAÇÃO(valores y; valores x)	estimativa $\hat{\beta}_1$ do coef. angular β_1
PREVISÃO(x; valores y; valores x)	valor de y correspondente a x
CORREL(valores y; valores x)	coeficiente de correlação
RQUAD(valores y; valores x)	coeficiente de determinação r^2
PROJ.LIN(valores y; valores x; constante; estatística)	constante = verdadeiro ou omitido ➔ $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ constante = falso ➔ $\hat{y} = \hat{\beta}_1 x$ estatística = falso ou omitido retorna apenas os coeficientes da reta estatística = verdadeiro retorna dados adicionais (ver ajuda do Excel)

Ferramentas de análise

REGRESSÃO
CORRELAÇÃO

PROBLEMAS:

1) Considere o problema ilustrativo 1 onde foi dada a tabela da distância percorrida pelo motorista, após cada minuto, em função do tempo:

x= tempo (min)	0	1	2	3	4	5
y= Distância percorrida (km)	0	1,3	3,8	4,3	6,7	7,3

a) Use as funções do Excel, I**NCLINAÇÃO** e **INTERCEPÇÃO**, para calcular o coeficiente linear e o coeficiente angular da reta de regressão.
b) Use a função **PREVISÃO** para calcular valores de distância percorrida quando x=1,3; x=4,7; x=6, de acordo com a reta de regressão.
c) Forme no Excel uma tabela de valores previstos, resíduos e resíduos padrão. Calcule a soma de quadrados dos resíduos.
d) Calcule intervalos de 90% de confiança para β_0 e β_1. Interprete.
e) Teste a hipótese de que $\beta_1 = 1,5$. Interprete

2) Use a Ferramenta de análise **Regressão** do Excel para estudar o problema da introdução. Na caixa de diálogo Regressão considere:
   Intervalo y de entrada: Indique coluna dos valores de y
   Intervalo x de entrada: Indique coluna de valores de x
   Rótulos: optativo
   Nível de confiança: 95%
   Constante é zero: NÃO ATIVE (no próximo problema será ativado)
   Intervalo de saída: Escolha uma célula
   Resíduos: ATIVE
   Resíduos padronizados: ATIVE
   Plotar resíduos: ATIVE
   Plotar ajuste de linha: ATIVE
   Plotagem de probabilidade normal: NÃO ATIVE

3) Repetir o problema anterior considerando a constante igual a zero. Faça uma interpretação cuidadosa deste problema.

4) Considere o problema ilustrativo 2 do item 3 (regressão linear múltipla). Use a ***ferramenta*** **Regressão** para resolvê-lo. Interprete cada resultado.

5) Ajuste aos dados abaixo uma reta e, depois, uma parábola (considere um modelo de regressão múltipla com $x_1 = x$ e $x_2 = x^2$). Use o coeficiente de determinação para decidir pelo melhor ajuste.

x	1,2	1,2	2,4	2,4	3,6	3,6	4,8	4,8	6,0	6,0
y	5,2	6,0	2,0	3,2	2,5	3,1	5,2	5,6	12,1	10,8

PROBLEMAS ADICIONAIS DE LIVROS TEXTO

COSTA NETO, P.L.O. Estatística. São Paulo: Ed. Edgard Blucher Ltda, 1978

6) O faturamento de uma loja durante seus primeiros oito meses de atividades é dado a seguir, em milhares de reais.

Meses	Faturamento
Março	20
Abril	22
Maio	22
Junho	25
Julho	10
Agosto	40
Setembro	45
Outubro	60

a) Ajuste uma reta de regressão e tire conclusões do ponto de vista estatístico.
b) Elimine o dado referente ao mês de julho, considerando que foi anormalmente baixo devido a uma brusca, porém passageira, recessão de mercado e, com base nos demais pontos, equacione a reta de regressão que melhor se adapte aos dados.

7) Ajustar uma parábola de mínimos quadrados aos dados do problema anterior

8) Oito alunos sorteados entre os da segunda série de um curso de Engenharia obtiveram as seguintes notas nos exames de Cálculo e Física:

Aluno	1	2	3	4	5	6	7	8
Cálculo	4,5	6,0	3,0	2,5	5,0	5,5	1,5	7,0
Fisica	3,5	4,5	3,0	2,0	5,5	5,0	1,5	6,0

Com base nesses dados, pode-se ter praticamente 99% de certeza de que os alunos mais bem preparados em Cálculo também o sejam em Física?

OBS: A estatística do teste é $t_0 = r\sqrt{\frac{n-2}{1-r^2}}$ com n-2 g.l. Este teste de correlação é equivalente ao teste do coeficiente angular da regressão igual a zero.

9) Obter a equação da reta de mínimos quadrados para os seguintes pontos experimentais:

x	1	2	3	4	5	6	7	8
y	0,5	0,6	0,9	0,8	1,2	1,5	1,7	2,0

Traçar a reta no digrama de dispersão. Calcular o coeficiente de correlação linear.

10) Uma reação química foi realizada sob seis pares de diferentes condições de pressão e temperatura. Em cada caso, foi medido o tempo necessário para que a reação se completasse. Os resultados obtidos são os que seguem:

Condição	Temperatura (°C)	Pressão (atm)	Tempo (s)
1	20	1,5	9,4
2	30	1,5	8,2
3	30	1,2	9,7
4	40	1,0	9,5
5	60	1,0	6,9
6	80	0,8	6,5

Obter a equação da função de regressão linear do tempo (y) em relação à temperatura (x_1) e à pressão (x_2).

LAPPONI, J.C. *Estatística Usando o Excel 5 e 7*. São Paulo: Lapponi Ed., 1997

11) Os dados abaixo se referem aos 10 maiores e melhores grupos de supermercados de acordo com o Censo 1990/1991 – Estrutura do Varejo Brasileiro-Nielsen

	Vendas $bilhões	Nº de caixas	Área 1000m^2	Nº de lojas	Funcionários 1000
Carrefour	164,1	1669	207,6	22	11,2
Cia Bras. de Distribuição	154,5	4670	458,1	416	26,9
Paes Mendonça	116,2	2968	314,3	132	20,6
Casas Sendas	63,9	1327	149,6	53	13,4
Bompreço	61,3	1648	155,9	103	11,3
Casas da Banha	43,8	1910	192,0	175	14,2
Eldorado	35,9	451	100,5	7	9,5
Cia Real de Distribuição	25,8	1183	93,0	62	7,5
Comercial Gentil Moreira	24,6	492	48,3	36	4,4
Rede Barateiro	21,8	505	51,3	25	5,1

Use a ferramenta Correlação. Interprete a maior correlação e também a menor.

PROBLEMA PROPOSTO

PP7) Encontre na literatura especializada problemas aos quais podem ser empregados métodos deste capítulo.

VIII. MODELOS LINEARIZÁVEIS

1. MODELO EXPONENCIAL

Nos modelos de regressão do capítulo anterior os parâmetros aparecem linearmente em suas expressões. Em alguns modelos onde isso não ocorre, um transformação pode tornar o modelo linear. Algum cuidado deve ser tomado com o termo do erro nessas transformações, como será visto a seguir.

Problema ilustrativo 1: Seja o modelo de regressão não-linear, com variável independente z, variável dependente x, parâmetros θ_0 e θ_1 e erro multiplicativo w_i

$$\boxed{z_i = \theta_0(\theta_1)^{x_i} \cdot w_i} \qquad \text{onde i=1,2,...n}$$

Aplicando logaritmo em ambos os membros da igualdade, obtém-se

$$\log(z_i) = \log(\theta_0) + \log(\theta_1) \cdot x_i + \log(w_i)$$

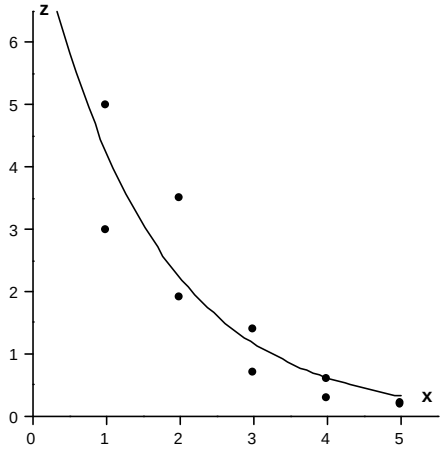
que é uma reta em um sistema de coordenadas log(z) contra x, ou seja, o modelo é da forma

$$\boxed{y_i = \beta_0 + \beta_1 x_i + \text{erro}}$$

onde $y_i = \log(z_i)$; $\beta_0 = \log(\theta_0)$; $\beta_1 = \log(\theta_1)$ e erro = $u_i = \log(w_i)$

Os valores numéricos para ilustrar este problema foram simulados. Primeiro supôs-se que $\theta_0 = 8$ e $\theta_1 = 0{,}5$. Em seguida fixou-se 10 valores de x: 0; 0; 1; 1; 2; 2; 3; 3; 4 4; 5 e 5 obtendo-se os valores $z_i^* = 8(0{,}5)^{x_i}$ (valores da variável independente sem erro). Em seguida foram criados os erros $u_i = \log(w_i)$ com distribuição normal de média zero e desvio padrão 0,1. Finalmente, obteve-se z_i multiplicando z_i^* pelo erro u_i.

x_i	z_i^*	$u_i=\log(w_i)$	$w_i=10^{u_i}$	z_i
0	8	-0,09	0,82	6,6
0	8	0,08	1,21	9,7
1	4	0,10	1,26	5,0
1	4	-0,13	0,75	3,0
2	2	0,24	1,74	3,5
2	2	-0,02	0,96	1,9
3	1	0,15	1,41	1,4
3	1	-0,18	0,66	0,7
4	0,5	0,11	1,28	0,6
4	0,5	-0,17	0,68	0,3
5	0,25	-0,19	0,65	0,2
5	0,25	-0,07	0,86	0,2



Supõe-se então que os pontos experimentais são os abaixo (ver figura acima)

x_i	0	0	1	1	2	2	3	3	4	4	5	5
z_i	6,6	9,7	5,0	3,0	3,5	1,9	1,0	0,7	0,6	0,3	0,2	0,2

Pretende-se ajustar o modelo linearizável, $z_i = \theta_0(\theta_1)^{x_i} \cdot w_i$ (observe pela figura o que significa erro multiplicativo). Aplicando logaritmos o modelo fica: $y_i = \beta_0 + \beta_1 x_i + u_i$ com os parâmetros já definidos acima.

A metodologia de regressão linear pode ser aplicada,obtendo-se para o modelo transformado:

Coeficientes	*Erro padrão*	*Intervalo de 95% de confiança*	
		Limite Inferior	Limite Superior
$\hat{\beta}_0$	0,9412	0,7872	1,0951
$\hat{\beta}_1$	-0,3261	-0,3769	-0,2753

com $s^2 = 0{,}01822$. Para os parâmetros originais basta considerar que $\beta_0 = \log(\theta_0)$ e, portanto, $\hat{\theta}_0 = 10^{\hat{\beta}_0}$ e analogamente para o outro parâmetro. Os resultados estão no quadro abaixo.

Coeficientes	*Estimativa*	*Intervalo de 95% de confiança*	
		Limite Inferior	Limite Superior
$\hat{\theta}_0$	8,734	6,126	12,448
$\hat{\theta}_1$	0,472	0,420	0,531

OBSERVAÇÃO: Se o erro fosse aditivo, não teria sentido aplicar logaritmo. O modelo seria considerado *intrinsecamente* não-linear e seria adotada uma metodologia própria desses modelos.

2. USANDO O EXCEL

O Excel ajusta, no módulo gráfico, linhas de tendência a um conjunto de dados, com as seguintes funções:

Linear simples	$y = b_0 + b_1 x$
Polinomial	$y = b_0 + b_1 x + b_2 x^2 + ... + b_k x^k, \text{para } k \geq 2$
Logarítmica	$y = \theta_0 + \theta_1 \text{Ln}(x)$
Potência	$y = \theta_0 x^{\theta_1}$
Exponencial	$y = \theta_0 e^{\theta_1 x}$ onde e=2,7182...

PROBLEMAS:

1) Estude o ajuste do modelo $y = a + \frac{b}{x} + \text{erro}$ aos dados abaixo

x	0,2	0,3	0,4	0,5	0,6	0,8	1
y	6,2	4,1	3,3	3,0	2,3	2,0	1,7

Determine intervalos de confiança para os parâmetros a e b.

2) Seja a função $y = e^{(a - b/x)}$.
   a) Que transformação deve ser feita para que as fórmulas de regressão linear simples possam ser usadas para ajustar essa função.
   b) Simule uma amostra aleatória de uma distribuição normal e estude o ajuste desse modelo com erro multiplicativo.

3) Simule um experimento análogo ao do problema introdutório com a função potência. Considere três repetições para cada valor da variável independente.

PROBLEMAS ADICIONAIS DE LIVROS TEXTO

COSTA NETO, P.L.O. Estatística. São Paulo: Ed. Edgard Blucher Ltda, 1978

4) Uma Teoria física faz prever que y dependerá de x segundo a expressão $y + C = \frac{x^2}{2p}$, onde C e p são duas constantes numéricas. Sabendo-se que x é medido sem erro e que a precisão da medida de y no intervalo experimental aqui considerado é constante, estime os melhores valores de C e p a partir dos seguintes dados:

x	1	2	3	4	5	6	7
y	0,2	0,6	0,8	1,4	2,6	3,2	5,0

5) Um certo fenômeno físico segue a lei $x(y + \gamma) = C$ (x e y variáveis; C e γ constantes). Sabendo-se que a determinação experimental de x é muito mais precisa do que a de y, estime o melhor valor para a constante C a partir dos pares de valores experimentais dados a seguir. Com base nesses dados, ao nível 5% de significância, existe evidência de que a constante γ seja realmente diferente de zero?

x	1	2	5	10	20	50
y	27,0	12,0	10,0	6,0	6,3	4,8

6) Estabeleça a equação da regressão para os dados que seguem, sabendo que a equação teórica é da forma $z = ay^{bx+c}$

x	1	1	2	3
y	2	3	2	1
z	4,0	7,5	16,0	1,8

PROBLEMA PROPOSTO

P8) Faça um estudo estatístico para os problemas de 4 a 6, determinando intervalos de confiança para os parâmetros, verificando se a regressão é significativa pelo teste t e pela análise de variância, calculando o coeficiente de determinação e construindo gráficos de resíduos. Para um valor arbitrário da variável independente (dentro do intervalo experimental) estime a resposta experimental correspondente e determine um intervalo de confiança.

APÊNDICE

TABELAS

As tabelas abaixo fornecem valores das distribuições normal padrão (z_0), t de Student (t_0), qui-quadrado (χ^2_0) e F (F_0), correspondentes a uma probabilidade p (área abaixo da curva).

Tabela 1: Distribuição normal acumulada

p	**0,9**	**0,95**	**0,975**	**0,9**	**0,995**
1-p	0,1	0,05	0,025	0,1	0,005
z_0	1,28	1,64	1,96	1,28	2,58

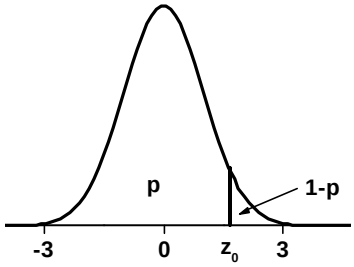


Tabela 2: Distribuição t de Student

	p				
g.l.	**0,1**	**0,05**	**0,025**	**0,01**	**0,005**
1	6,314	12,71	25,45	63,66	127,3
2	2,920	4,303	6,205	9,925	14,09
3	2,353	3,182	4,177	5,841	7,453
4	2,132	2,776	3,495	4,604	5,598
5	2,015	2,571	3,163	4,032	4,773
6	1,943	2,447	2,969	3,707	4,317
7	1,895	2,365	2,841	3,499	4,029
8	1,860	2,306	2,752	3,355	3,833
9	1,833	2,262	2,685	3,250	3,690
10	1,812	2,228	2,634	3,169	3,581
11	1,796	2,201	2,593	3,106	3,497
12	1,782	2,179	2,560	3,055	3,428
13	1,771	2,160	2,533	3,012	3,372
14	1,761	2,145	2,510	2,977	3,326
15	1,753	2,131	2,490	2,947	3,286
16	1,746	2,120	2,473	2,921	3,252
17	1,740	2,110	2,458	2,898	3,222
18	1,734	2,101	2,445	2,878	3,197
19	1,729	2,093	2,433	2,861	3,174
20	1,725	2,086	2,423	2,845	3,153
21	1,721	2,080	2,414	2,831	3,135
22	1,717	2,074	2,405	2,819	3,119
23	1,714	2,069	2,398	2,807	3,104
24	1,711	2,064	2,391	2,797	3,091
25	1,708	2,060	2,385	2,787	3,078
26	1,706	2,056	2,379	2,779	3,067
27	1,703	2,052	2,373	2,771	3,057
28	1,701	2,048	2,368	2,763	3,047
29	1,699	2,045	2,364	2,756	3,038
30	1,697	2,042	2,360	2,750	3,030
40	1,684	2,021	2,329	2,704	2,971
60	1,671	2,000	2,299	2,660	2,915
120	1,658	1,980	2,270	2,617	2,860
∞	1,645	1,960	2,242	2,576	2,808

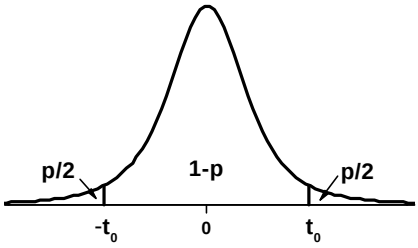
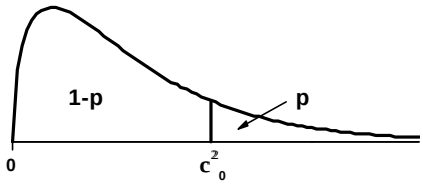
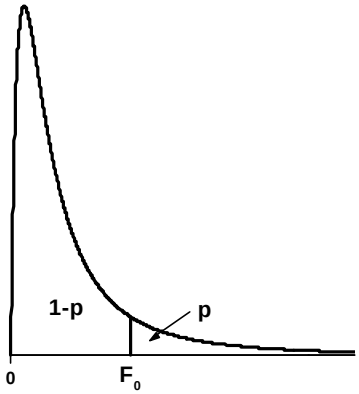


Tabela 3: Distribuição Qui-quadrado



	p									
g.l.	**0,995**	**0,99**	**0,975**	**0,95**	**0,9**	**0,1**	**0,05**	**0,025**	**0,01**	**0,005**
1	0,000	0,000	0,001	0,004	0,016	2,71	3,84	5,02	6,63	7,88
2	0,010	0,020	0,051	0,103	0,211	4,61	5,99	7,38	9,21	10,60
3	0,072	0,115	0,216	0,352	0,584	6,25	7,81	9,35	11,34	12,84
4	0,207	0,297	0,484	0,711	1,064	7,78	9,49	11,14	13,28	14,86
5	0,41	0,55	0,83	1,15	1,61	9,24	11,07	12,83	15,09	16,75
6	0,68	0,87	1,24	1,64	2,20	10,64	12,59	14,45	16,81	18,55
7	0,99	1,24	1,69	2,17	2,83	12,02	14,07	16,01	18,48	20,28
8	1,34	1,65	2,18	2,73	3,49	13,36	15,51	17,53	20,09	21,95
9	1,73	2,09	2,70	3,33	4,17	14,68	16,92	19,02	21,67	23,59
10	2,16	2,56	3,25	3,94	4,87	15,99	18,31	20,48	23,21	25,19
11	2,60	3,05	3,82	4,57	5,58	17,28	19,68	21,92	24,73	26,76
12	3,07	3,57	4,40	5,23	6,30	18,55	21,03	23,34	26,22	28,30
13	3,57	4,11	5,01	5,89	7,04	19,81	22,36	24,74	27,69	29,82
14	4,07	4,66	5,63	6,57	7,79	21,06	23,68	26,12	29,14	31,32
15	4,60	5,23	6,26	7,26	8,55	22,31	25,00	27,49	30,58	32,80
16	5,14	5,81	6,91	7,96	9,31	23,54	26,30	28,85	32,00	34,27
17	5,70	6,41	7,56	8,67	10,09	24,77	27,59	30,19	33,41	35,72
18	6,26	7,01	8,23	9,39	10,86	25,99	28,87	31,53	34,81	37,16
19	6,84	7,63	8,91	10,12	11,65	27,20	30,14	32,85	36,19	38,58
20	7,43	8,26	9,59	10,85	12,44	28,41	31,41	34,17	37,57	40,00
21	8,03	8,90	10,28	11,59	13,24	29,62	32,67	35,48	38,93	41,40
22	8,64	9,54	10,98	12,34	14,04	30,81	33,92	36,78	40,29	42,80
23	9,26	10,20	11,69	13,09	14,85	32,01	35,17	38,08	41,64	44,18
24	9,89	10,86	12,40	13,85	15,66	33,20	36,42	39,36	42,98	45,56
25	10,52	11,52	13,12	14,61	16,47	34,38	37,65	40,65	44,31	46,93
26	11,16	12,20	13,84	15,38	17,29	35,56	38,89	41,92	45,64	48,29
27	11,81	12,88	14,57	16,15	18,11	36,74	40,11	43,19	46,96	49,65
28	12,46	13,56	15,31	16,93	18,94	37,92	41,34	44,46	48,28	50,99
29	13,12	14,26	16,05	17,71	19,77	39,09	42,56	45,72	49,59	52,34
30	13,79	14,95	16,79	18,49	20,60	40,26	43,77	46,98	50,89	53,67

Tabela 4: Distribuição F



p=0,05	**g.l. num..**							
g.l. den.	1	2	3	4	5	6	7	8
1	161	199	216	225	230	234	237	239
2	18,5	19,0	19,2	19,2	19,3	19,3	19,4	19,4
3	10,1	9,55	9,28	9,12	9,01	8,94	8,89	8,85
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10
120	3,92	3,07	2,68	2,45	2,29	2,18	2,09	2,02
∞	3,84	3,00	2,61	2,37	2,21	2,10	2,01	1,94

p=0,01	**g.l. num..**							
g.l. den.	1	2	3	4	5	6	7	8
1	4052	4999	5404	5624	5764	5859	5928	5981
2	98,5	99,0	99,2	99,3	99,3	99,3	99,4	99,4
3	34,1	30,8	29,5	28,7	28,2	27,9	27,7	27,5
4	21,2	18,0	16,7	16,0	15,5	15,2	15,0	14,8
5	16,3	13,3	12,1	11,4	11,0	10,7	10,5	10,3
6	13,8	10,9	9,78	9,15	8,75	8,47	8,26	8,10
7	12,3	9,55	8,45	7,85	7,46	7,19	6,99	6,84
8	11,3	8,65	7,59	7,01	6,63	6,37	6,18	6,03
9	10,6	8,02	6,99	6,42	6,06	5,80	5,61	5,47
10	10,0	7,56	6,55	5,99	5,64	5,39	5,20	5,06
12	9,33	6,93	5,95	5,41	5,06	4,82	4,64	4,50
15	8,68	6,36	5,42	4,89	4,56	4,32	4,14	4,00
20	8,10	5,85	4,94	4,43	4,10	3,87	3,70	3,56
30	7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17
60	7,08	4,98	4,13	3,65	3,34	3,12	2,95	2,82
120	6,85	4,79	3,95	3,48	3,17	2,96	2,79	2,66
∞	6,64	4,61	3,78	3,32	3,02	2,80	2,64	2,51

PROBLEMA ESPECIAL 1

Entendendo o significado do nível de confiança de um intervalo de confiança

Em uma planilha do Excel

a) crie com a ferramenta de análise GERAÇÃO DE NÚMERO ALEATÓRIO 1000 valores de uma população normal de média 1,62 e desvio padrão 0,08 (Problema 3 – Cap. III, pag. 19). Em seguida enumere-os de 1 a 1000 (coluna A o nº e coluna B o valor)

b) Considere os valores criados em a) como sendo a própria população. Ache a média e desvio padrão e considere-os como sendo μ e σ, respectivamente.

c) Sorteie uma amostra de tamanho n=10 dos nºs da coluna A e coloque-os na coluna C. Use a função PROCV para encontrar os valores correspondentes na coluna B e coloque-os na coluna C (veja o problema 8, cap I, pag. 8). Cada vez que a tecla F9 é pressionada obtém-se uma nova amostra da população.

d) Determine os limites de um intervalo de 95% de confiança para a média μ da população.
Faça como na planilha apresentada abaixo, onde esses limites estão nas células C20 e E20.

e) Cada vez que é apertada a tecla F9, a célula C22 (construída com a função E) irá apresentar a mensagem VERDADEIRO se o intervalo contiver o valor μ e a mensagem FALSO se μ estiver fora do intervalo. Aperte F9 um número grande de vezes e conte quantas vezes aparece a mensagem FALSO. Ela deverá aparecer em torno de 5% das vezes . A célula D18, pode ser alterada para outros níveis de confiança.

	A	B	C	D	E
1	Nº	Altura Pop.	Nº sorteado	amostra	
2	1	1,60	874	**1,69**	
3	2	1,52	839	**1,55**	
4	3	1,64	83	**1,72**	
5	4	1,72	736	**1,60**	
6	5	1,72	743	**1,65**	
7	6	1,76	368	**1,72**	
8	7	1,45	512	**1,64**	
9	8	1,60	74	**1,57**	
10	9	1,71	354	**1,67**	
11	10	1,53	617	**1,64**	
12	11	1,56	n=	10	
13	12	1,48	média=	1,65	
14	13	1,47	Dev. pad.=	0,0588	
15	14	1,54	Erro padrão=	0,0186	
16	15	1,56	**Média Pop.=**	**1,62**	
17	16	1,45	***Intervalo de conf.***		t =
18	17	1,57	nível=	**0,95**	2,2622
19	18	1,59	*L.Inf*	**média**	**L.Sup**
20	19	1,63	**1,604**	**1,646**	**1,688**
21	20	1,59	**resultado:**		
22	21	1,59	**VERDADEIRO**		
23	22	1,59			
24	23	1,73			
25	24	1,61			

FUNÇÕES E FÓRMULAS	
A2	=1
A3	=A2+1
C2	=ALEATÓRIOENTRE(1;1000)
D2	=PROCV(C2;$A$2:$B$1000)
D13	=MÉDIA(D2:D11)
D14	=DESVPAD(D2:D11)
D15	=D14/RAIZ(D12)
D16	=MÉDIA(B2:B1001)
E18	=INVT(1-D18;D12-1)
C20	=D20-E18*D15
D20	=D13
E20	=D20+E18*D15
C22	=E(D16>C20;E20)

Continua até 1000

C22=VERDADEIRO -->intervalo inclui μ e C22=FALSO --> intervalo não inclui μ