ESCOLA POLITÉCNICA DA UNIVERSIDADE DE SÃO PAULO
DEPARTAMENTO DE ENGENHARIA DE PRODUÇÃO

ESTATÍSTICA I

Prof. Alberto W. Ramos

SÃO PAULO, 2016

ESTATÍSTICA

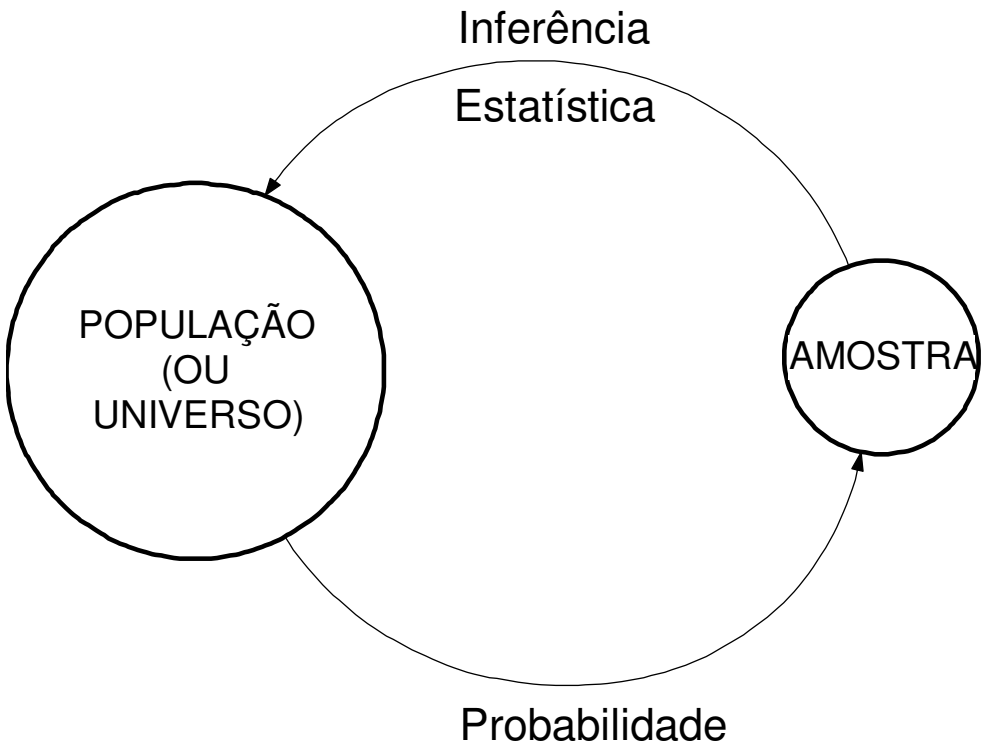
É a ciência que trata da coleta, organização, descrição, análise e interpretação de dados experimentais

Para que precisamos de informações ?

© ALBERTO W. RAMOS 2016

TERMOS BÁSICOS

Termo	***Significado***
População (ou universo)	É a coleção de todas as unidades sobre as quais desejamos informações
Amostra	É parte da coleção total de unidades
Censo	É a pesquisa que envolve 100% da população
Variável	Aquela característica na qual estamos interessados

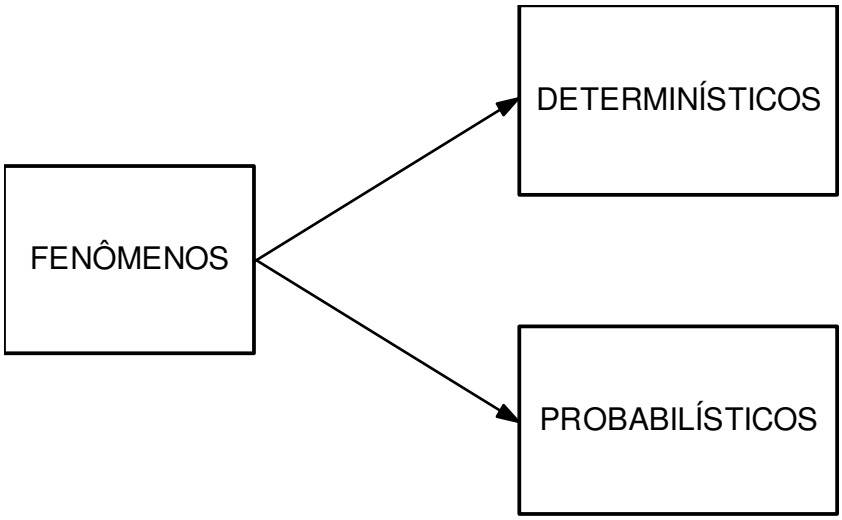


© ALBERTO W. RAMOS 2016

Cálculo de Probabilidades

© ALBERTO W. RAMOS 2016

PROBABILIDADE



Definições:

a) Espaço Amostral (S): conjunto de todos os resultados possíveis de um fenômeno probabilístico.

   Ex.: lançamento de dado → S = {1,2,3,4,5,6}

b) Evento (A,B,C,...): qualquer subconjunto de S.

   Ex.: P = ponto par = {2,4,6}
   I = ponto ímpar = {1,3,5}
   T = ponto maior que três = {4,5,6}

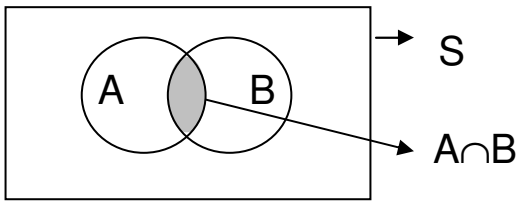
   Obs.: S = evento certo
   Ø = evento impossível

© ALBERTO W. RAMOS 2016

OPERAÇÕES COM EVENTOS

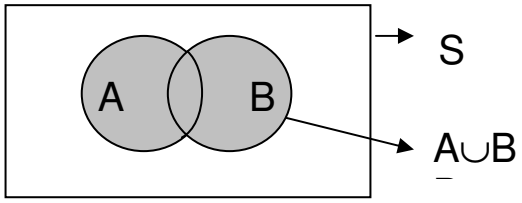
a) Evento intersecção: $A \cap B$

Ex.: $P \cap T = \{4,6\}$ (ambos ocorrem)



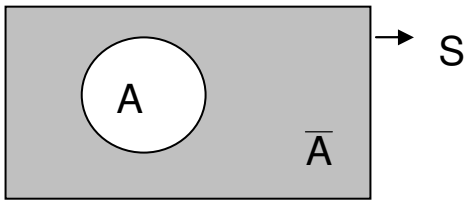
b) Evento união: $A \cup B$

Ex.: $P \cup I = \{1,2,3,4,5,6\} = S$ (pelo menos um ocorre)



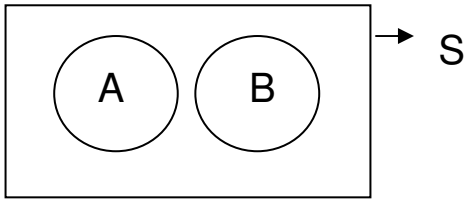
c) Evento complementar: $\overline{A}$

Ex.: $\overline{P} = \{1,3,5\} = I$ (P não ocorre)



c) Eventos mutuamente exclusivos: $A \cap B = \varnothing$

Ex.: $P \cap I = \varnothing$ (P e I não ocorrem ao mesmo tempo)



© ALBERTO W. RAMOS 2016

DEFINIÇÃO DE PROBABILIDADE

É um número real, associado a um evento, que mede sua chance de ocorrência:

$$P(A) = \frac{m}{n}$$

onde:

- m é o número de resultados favoráveis a A
- n é o número de resultados possíveis, desde que *igualmente prováveis*

Observações:

a) $0 \leq P(E) \leq 1$

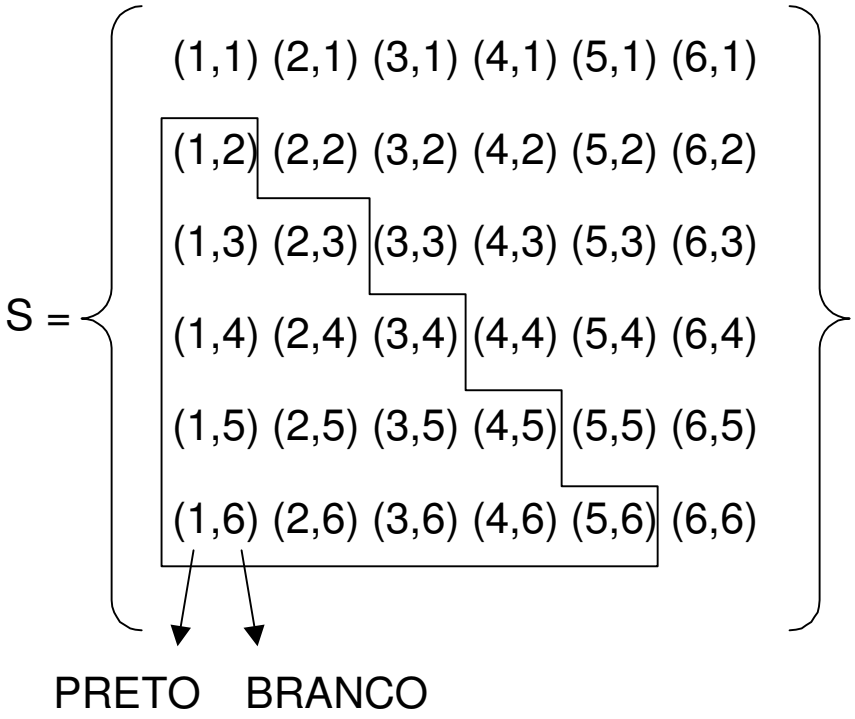
b) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

c) $P(\overline{A}) = 1 - P(A)$

© ALBERTO W. RAMOS 2016

EXEMPLO

Seja o lançamento de dois dados: um preto e outro banco. Qual a probabilidade de se obter ponto no dado preto menor que o branco?



$$P(A) = \frac{15}{36}$$

© ALBERTO W. RAMOS 2016

PROBABILIDADE CONDICIONADA

Notação: $P(A/B) \rightarrow$ probabilidade do evento A, sabendo-se que o evento B ocorreu

Ex.: $A \rightarrow$ chuva

$B \rightarrow$ previsão de chuva

$P(A/B) \rightarrow$ probabilidade de chuva dado que houve previsão de chuva

Definição:

$$P(A/B) = \frac{P(A \cap B)}{P(B)}, \; P(B) \neq 0$$

ou

$$P(B/A) = \frac{P(A \cap B)}{P(A)}, \; P(A) \neq 0$$

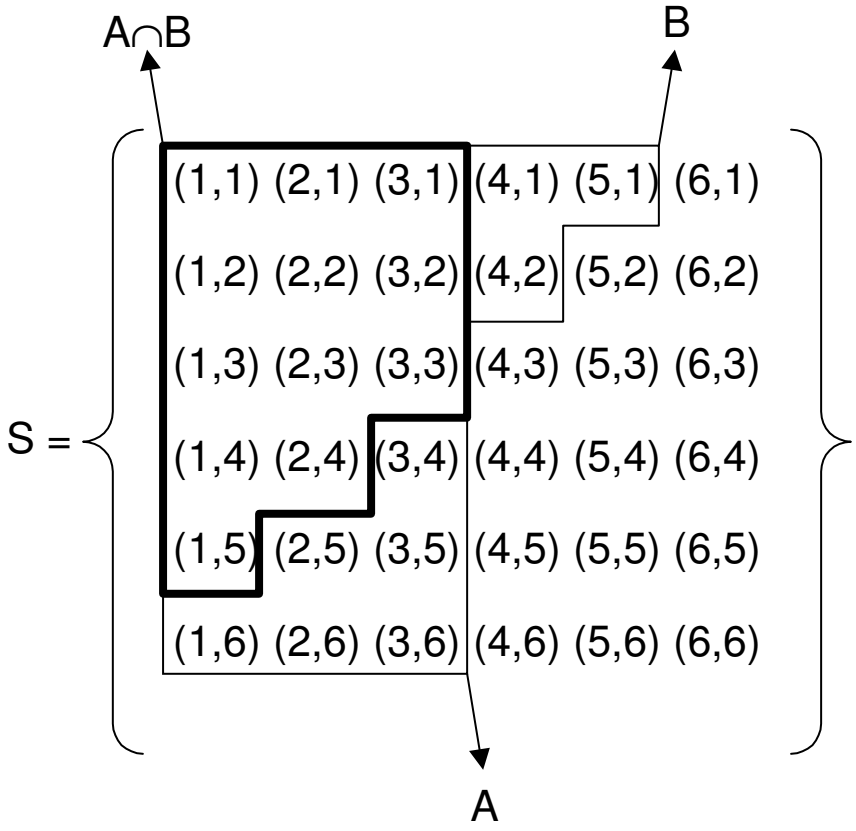
logo:

$$P(A \cap B) = P(A) \; . \; P(B/A) = P(B) \; . \; P(A/B)$$

© ALBERTO W. RAMOS 2016

EXEMPLO

Seja o lançamento de dois dados, com A: dar ponto 1, 2 ou 3 no primeiro dado e B: dar soma ≤ 6. Calcular P(A/B) e P(B/A).



$$P(A) = \frac{18}{36} = \frac{1}{2} \qquad P(B) = \frac{15}{36} = \frac{5}{12} \qquad P(A \cap B) = \frac{12}{36} = \frac{1}{3}$$

$$P(A/B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{1}{3}}{\frac{5}{12}} = \frac{12}{15} = \frac{4}{5}$$

$$P(B/A) = \frac{P(A \cap B)}{P(A)} = \frac{\frac{1}{3}}{\frac{1}{2}} = \frac{2}{3}$$

© ALBERTO W. RAMOS 2016

EVENTOS INDEPENDENTES

Se $P(A/B) = P(A/\overline{B}) = P(A) \Rightarrow$ o evento A é estatisticamente independente de $B \Rightarrow P(B/A) = P(B/\overline{A}) = P(B)$

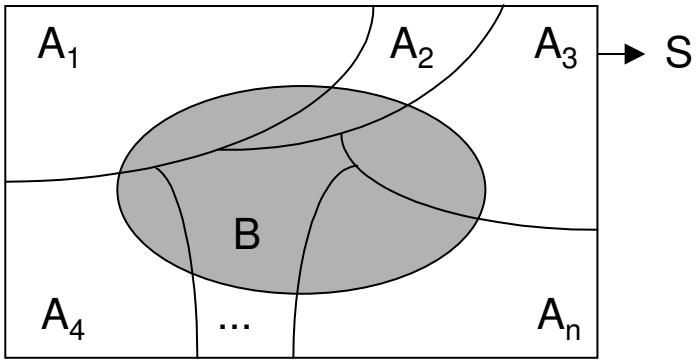
Neste Caso:

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

© ALBERTO W. RAMOS 2016

TEOREMA DA PROBABILIDADE TOTAL

Sejam A_1, A_2, ..., A_n eventos mutuamente exclusivos e exaustivos (partição) e seja B um evento qualquer de S.



$$B = \bigcup_{i=1}^{n} A_i \cap B \Rightarrow P(B) = \sum_{i=1}^{n} P(A_i \cap B)$$

$$\therefore \boxed{P(B) = \sum_{i=1}^{n} P(A_i).P(B/A_i)} \quad \textit{(TPT)}$$

© ALBERTO W. RAMOS 2016

TEOREMA DE BAYES

Nas mesmas condições do Teorema da Probabilidade Total.

$$P(A_j / B) = \frac{P(A_j \cap B)}{P(B)}$$

$$\boxed{P(A_j / B) = \frac{P(A_j).P(B / A_j)}{\sum_{i=1}^{n} P(A_i).P(B / A_i)}} \quad \textit{(TB)}$$

© ALBERTO W. RAMOS 2016

Variáveis Aleatórias

© ALBERTO W. RAMOS 2016

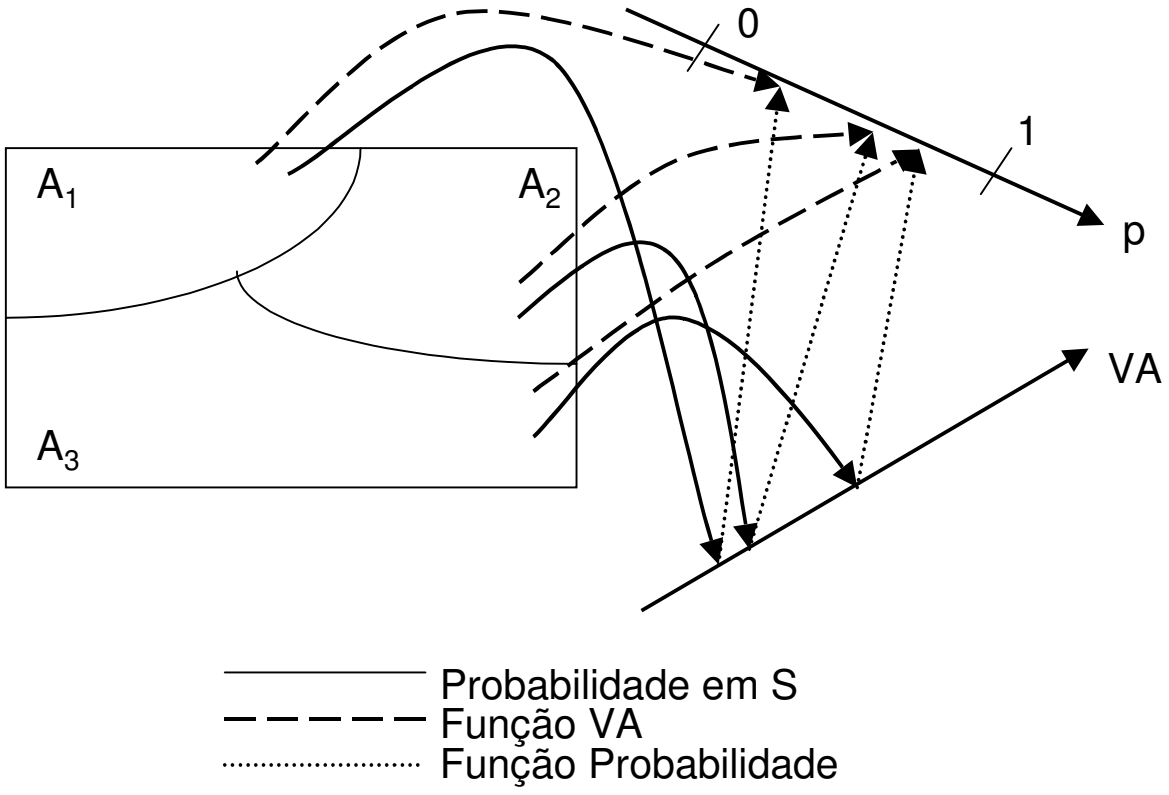
VARIÁVEIS ALEATÓRIAS

Uma variável aleatória (VA) é a representação dos eventos de uma partição de **S** através de números reais.

Exemplos:

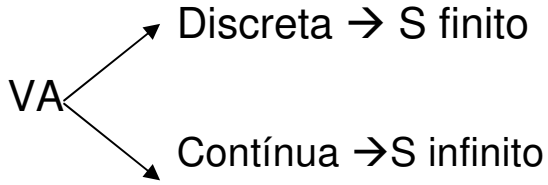
a) número de caras obtidas no lançamento de três moedas.

b) soma de pontos obtida no lançamento de dois dados.



© ALBERTO W. RAMOS 2016

TIPOS DE VARIÁVEIS ALEATÓRIAS (VA)



VA Discretas:

A distribuição de probabilidade é representada pela função probabilidade, tal que:

a) $P(X=x_i) \geq 0, \forall x_i$

b) $\sum_i P(X = x_i) = 1$

c) $\sum_{x_i > a}^{b} P(X = x_i) = P(a < X \leq b)$

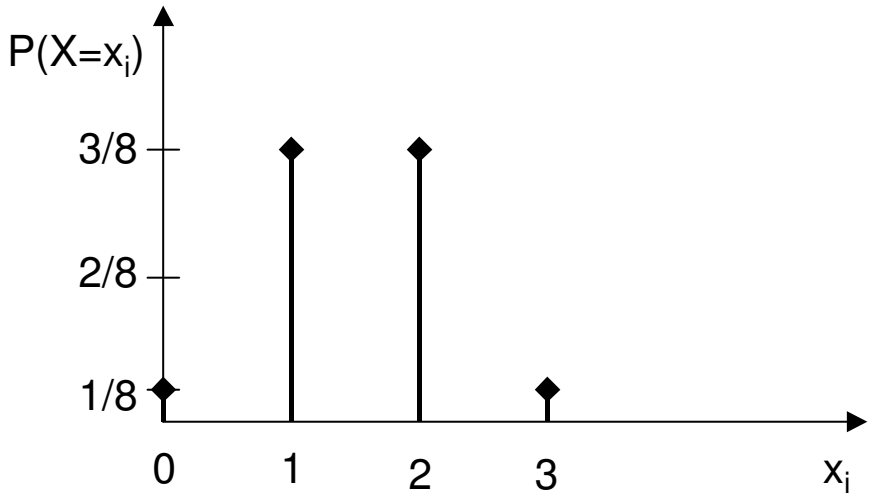
© ALBERTO W. RAMOS 2016

EXEMPLO

Seja X o número de caras (K) obtidas no lançamento de três moedas.

$$S = \left\{ \begin{array}{cccc} CCC & KCC & KKC & KKK \\ & CKC & KCK & \\ & CCK & CKK & \end{array} \right\}$$

x_i	0	1	2	3
$P(X=x_i)$	1/8	3/8	3/8	1/8



© ALBERTO W. RAMOS 2016

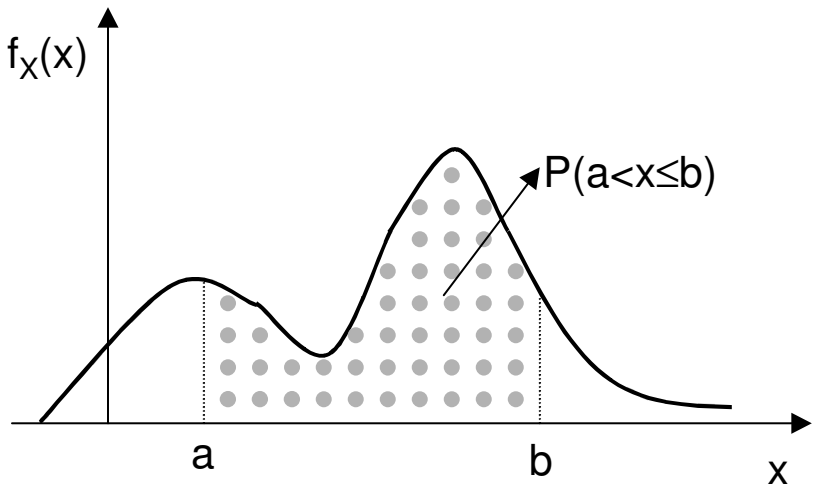
VA Contínuas:

S é infinito e a probabilidade de cada resultado individual é zero (mas não teoricamente impossível). A distribuição de probabilidade é representada pela função densidade de probabilidade $f_X(x)$.

a) $f_X(x) \geq 0$

b) $\int_{-\infty}^{+\infty} f_X(x)dx = 1$

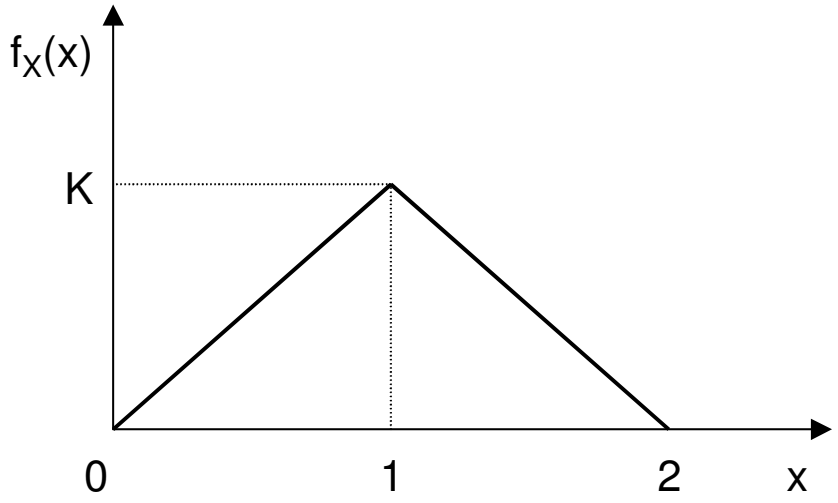
a) $\int_{a}^{b} f_X(x)dx = P(a < x \leq b), b > a$



© ALBERTO W. RAMOS 2016

EXEMPLO

Seja uma função densidade de probabilidade definida como:



a) determinar o valor de K.

b) equacionar esta fdp.

© ALBERTO W. RAMOS 2016

FUNÇÃO DE REPARTIÇÃO OU DE DISTRIBUIÇÃO ACUMULADA

É definida por:

$$F_X(x) = P(X \leq x) \qquad -\infty < x < +\infty$$

Para VAB discretas tem-se:

$$F_X(a) = \sum_{x_i \leq a} P(X = x_i)$$

Para VAB contínuas tem-se:

$$F_X(a) = \int_{-\infty}^{a} f_X(x)dx$$

Propriedades:

a) $F_X(-\infty) = 0$

b) $F_X(+\infty) = 1$

c) $P(a < X \leq b) = F_X(b) - F_Y(a)$

© ALBERTO W. RAMOS 2016

EXEMPLO

Obter as funções de distribuição acumulada dos dois exercícios anteriores.

© ALBERTO W. RAMOS 2016

PARÂMETROS DE POSIÇÃO

Indicam onde se localiza o centro da distribuição.

1) Média ou Valor Esperado: $\mu(X)$

- VA Discreta: $\mu(X) = \sum x_i.P(X = x_i)$
- VA Contínua: $\mu(X) = \int x.f_X(x)dx$

Propriedades:

a) $\mu(K) = K$, K = constante

b) $\mu(K.X) = K.\ \mu(X)$

c) $\mu(X+Y) = \mu(X) + \mu(Y)$

d) $\mu(X-Y) = \mu(X) - \mu(Y)$

e) $\mu(X \pm K) = \mu(X) \pm K$

f) Se X e Y são independentes $\Rightarrow \mu(X.Y) = \mu(X)\ .\ \mu(Y)$

2) Mediana: MD

É o ponto tal que: $P(X<MD) = P(X>MD) = ½$.

3) Moda: M_O

É o ponto de máxima probabilidade ou densidade de probabilidade.

© ALBERTO W. RAMOS 2016

PARÂMETROS DE DISPERSÃO

Indicam a variabilidade da distribuição de probabilidade.

1) Variância: $\sigma^2(X)$, V(X)

$$\sigma^2(X) = \mu[(X-\mu)^2] = \mu(X^2) - [\mu(X)]^2$$

- VA Discreta:

$$\sigma^2(X) = \sum_i (x_i - \mu)^2 . P(X = x_i) = \sum_i x_i^2 . P(X = x_i) - \left[\sum_i x_i . P(X = x_i) \right]^2$$

- VA Contínua:

$$\sigma^2(X) = \int_{-\infty}^{+\infty} (x-\mu)^2 . f_X(x) dx = \int_{-\infty}^{+\infty} x^2 . f_X(x) dx - \left[\int_{-\infty}^{+\infty} x . f_X(x) dx \right]^2$$

Propriedades:

a) $\sigma^2(K) = 0$, K = constante

b) $\sigma^2(K.X) = K^2 . \sigma^2(X)$

c) Se X e Y são independentes:

$$\sigma^2(X+Y) = \sigma^2(X) + \sigma^2(Y)$$

$$\sigma^2(X-Y) = \sigma^2(X) + \sigma^2(Y)$$

d) $\sigma^2(X \pm K) = \sigma^2(X)$

© ALBERTO W. RAMOS 2016

2) Desvio-Padrão: $\sigma(X)$

$$\sigma(X) = \sqrt{\sigma^2(X)}$$

3) Coeficiente de Variação: CV

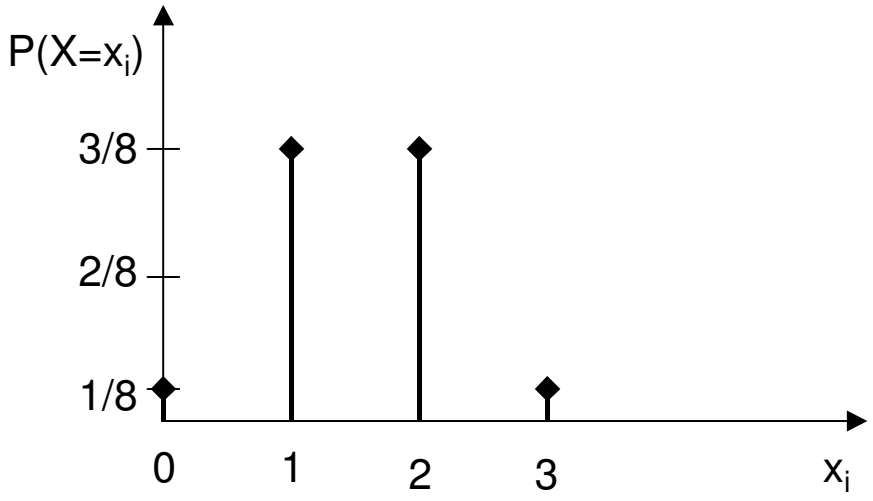
$$CV = \frac{\sigma(X)}{\mu(X)}$$

© ALBERTO W. RAMOS 2016

EXEMPLOS

Seja X o número de caras (K) obtidas no lançamento de três moedas.

x_i	0	1	2	3
$P(X=x_i)$	1/8	3/8	3/8	1/8



$$\mu(X) = \sum x_i.P(X = x_i) = 0\ x\frac{1}{8} + 1\ x\frac{3}{8} + 2\ x\frac{3}{8} + 3\ x\frac{1}{8} = \frac{12}{8} = 1{,}5$$

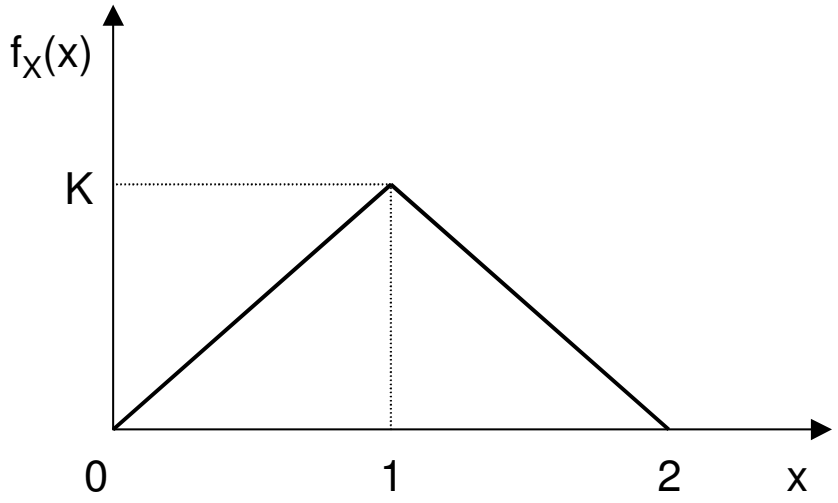
$$\sum_i x_i^2.P(X = x_i) = 0^2\ x\frac{1}{8} + 1^2\ x\frac{3}{8} + 2^2\ x\frac{3}{8} + 3^2\ x\frac{1}{8} = \frac{24}{8} = 3$$

$$\sigma^2(X) = \sum_i x_i^2.P(X = x_i) - \left[\sum_i x_i.P(X = x_i)\right]^2 = 3 - (1{,}5)^2 = 0{,}75$$

© ALBERTO W. RAMOS 2016

EXEMPLO

Seja uma função densidade de probabilidade definida como:



Determinar a média $\mu(X)$ e a variância $\sigma^2(X)$.

© ALBERTO W. RAMOS 2016

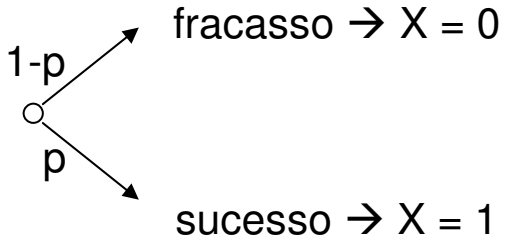
Principais Distribuições de Probabilidade

© ALBERTO W. RAMOS 2016

DISTRIBUIÇÕES DISCRETAS

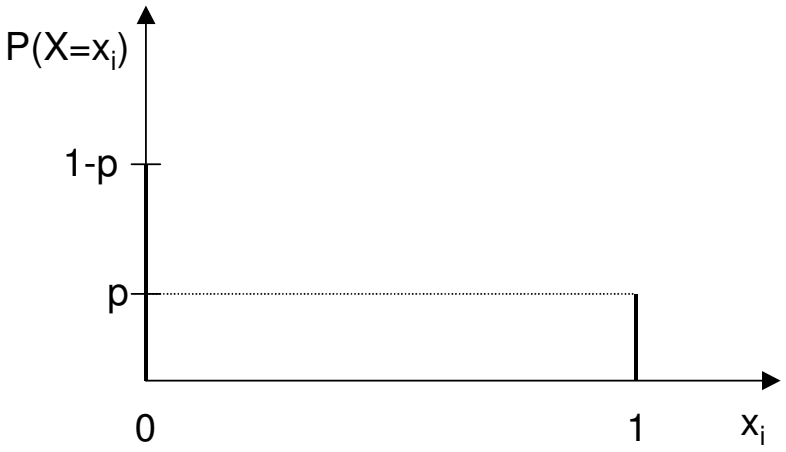
- ***Distribuição de Bernoulli***

Seja uma prova que só possa ter dois resultados:



com X = número de sucessos (0 ou 1)

x_i	0	1
$P(X = x_i)$	1-p	p



- $\mu(X) = 0 \times (1-p) + 1 \times p = p$
- $\mu(X^2) = 0^2 \times (1-p) + 1^2 \times p = p$
- $\sigma^2(X) = \mu(X^2) - [\mu(X)]^2 = p - p^2 = p(1-p)$

© ALBERTO W. RAMOS 2016

EXEMPLO

Um dado é lançado cinco vezes. Qual é a probabilidade de sair ponto “seis” duas vezes ?

Eventos:

S (sucesso) → sair ponto seis
F (fracasso) → não sair ponto seis

Existem 10 maneiras (possibilidades) disto ocorrer:

SSFFF
SFSFF
SFFSF
SFFFS
FSFFS
FFSFS
FFFSS
FSSFF
FSFSF
FFSSF

Contudo existe uma maneira mais fácil, sem ter necessidade de listar todas as possibilidades:

$$\binom{n}{x} = C_{n,x} = \frac{n!}{x!(n-x)!}$$

n = número de tentativas (5 lançamentos)
x = número de sucessos (2 pontos “seis”)

© ALBERTO W. RAMOS 2016

- ***Distribuição Binomial***

$$P(X = x) = C_{n,x}.p^{x}.(1-p)^{n-x}$$

onde:
n = número de tentativas
x = número de sucessos
p = probabilidade de sucesso

A média e a variância desta distribuição são:

- $\mu(X) = n.p$
- $\sigma^2(X) = n.p.(1-p)$

Um lote de 1000 peças foi recebido na empresa e sabe-se que este tem 200 itens defeituosos. Se for retirada (com reposição) uma amostra de 10 peças, qual a chance desta conter 1 defeituosa ?

© ALBERTO W. RAMOS 2016

- ***Distribuição de Poisson***

$$P(X = x) = \frac{e^{-\lambda t}.(\lambda t)^x}{x!}$$

onde:
e = número de Euler (2,72)
λ = freqüência média de sucessos
t = intervalo de observação

Nesta distribuição a média e a variância são:

- $\mu(X) = \lambda t$
- $\sigma^2(X) = \lambda t$

Em um restaurante, no horário do almoço, em média chegam 10 clientes por minuto. Qual é a probabilidade de em dez minutos chegarem exatamente 122 clientes ?

© ALBERTO W. RAMOS 2016

APROXIMAÇÕES DE DISTRIBUIÇÕES

Se $\mathbf{p} \leq 0{,}10 \Rightarrow$ pode-se empregar a Distribuição de Poisson no lugar da Binomial, fazendo-se $\mathbf{n.p = \lambda.t}$.

Ex.: $n = 20$ e $p = 0{,}02$

$P_{BINOMIAL}\ (X=1) = 0{,}273$ e $P_{POISSON}\ (X=1) = 0{,}268$

© ALBERTO W. RAMOS 2016

DISTRIBUIÇÕES CONTÍNUAS

- ***Distribuição Exponencial***

Seja **T** o intervalo decorrido entre dois sucessos consecutivos de um fenômeno de Poisson, com parâmetro **λ**:

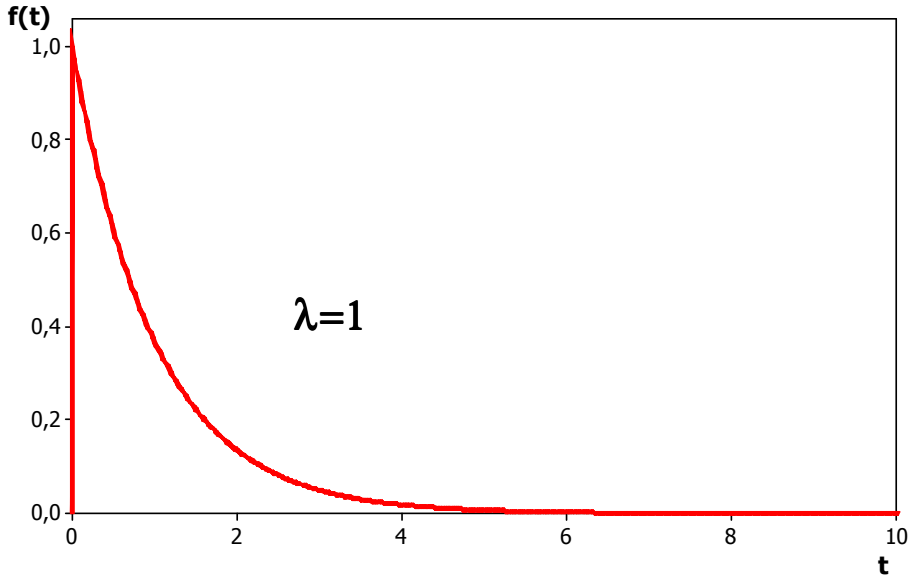
$$P(X=0)=e^{-\lambda t}$$

P(X=0) é a probabilidade de nenhum sucesso no intervalo de observação **t**. Significa também a probabilidade do primeiro sucesso levar mais do que **t** para ocorrer.

$$P(X=0)=P(T>t)=e^{-\lambda t}$$

$$\Rightarrow P(T\le t)=F_T(t)=1-e^{-\lambda t}$$

$$\Rightarrow f_T(t)=\frac{d}{dt}F_T(t)=\lambda e^{-\lambda t},\ t\ge 0$$



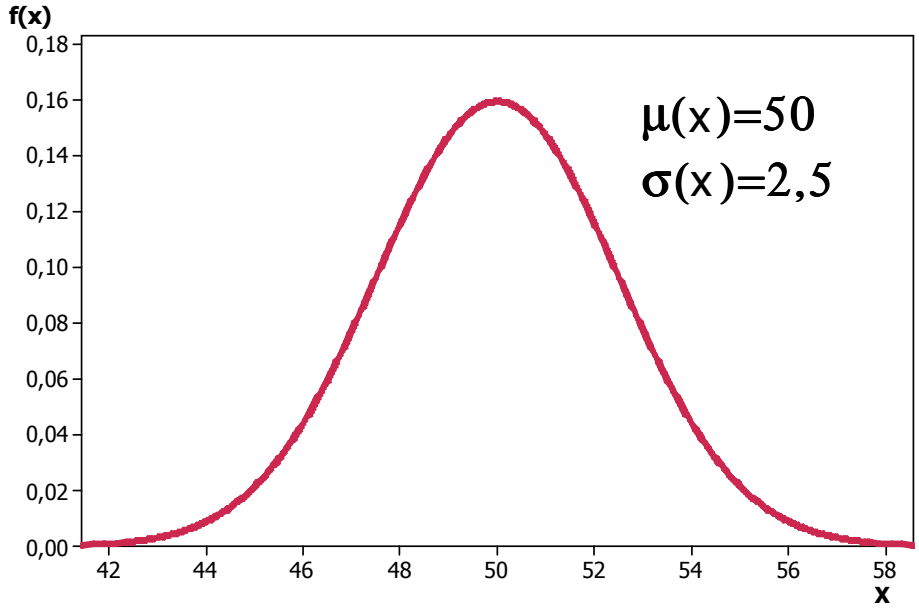
$$\mu(T)=\frac{1}{\lambda}\qquad e\qquad \sigma^2(T)=\frac{1}{\lambda^2}$$

© ALBERTO W. RAMOS 2016

- ***Distribuição Normal (ou de Gauss)***

Seja **X** uma variável aleatória contínua com a seguinte distribuição:

$$f_X(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma} \right)^2 \right], \; -\infty < x < +\infty$$



Esta distribuição tem média e variância:

- $\mu(X) = \mu$
- $\sigma^2(X) = \sigma^2$

Obs.: a) exp z = e^z

b) É comum escrever-se: X ~ N(μ;σ^2)

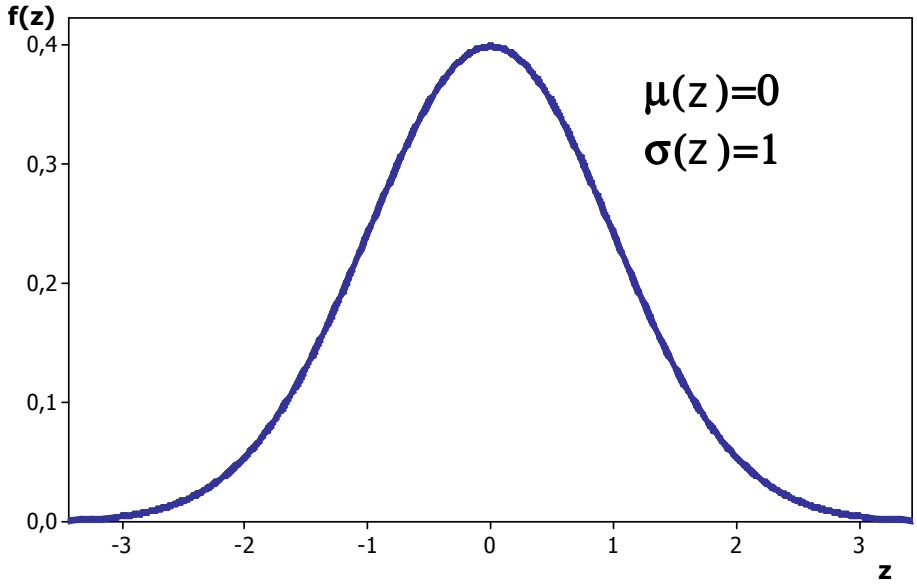
© ALBERTO W. RAMOS 2016

DISTRIBUIÇÃO NORMAL REDUZIDA (OU PADRONIZADA)

Seja **X** uma variável aleatória tal que $X \sim N(\mu;\sigma^2)$ e seja **Z** definida como:

$$z = \frac{x - \mu}{\sigma}$$

Então: $Z \sim N(0;1)$, com:



- $\mu(Z) = 0$
- $\sigma^2(Z) = 1$

Obs.: a) $P(\mu \le x \le x_0) \equiv P(0 \le z \le z_0)$

b) $P(X \le x) = P\left(Z \le \frac{x - \mu}{\sigma}\right) \equiv \Phi\left(\frac{x - \mu}{\sigma}\right)$

© ALBERTO W. RAMOS 2016

TABELA DA DISTRIBUIÇÃO NORMAL

valores de $P(0 < Z < z_0)$

z_0	0	1	2	3	4	5	6	7	8	9
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,7	0,2580	0,2611	0,2642	0,2673	0,2703	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1,0	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1,1	0,3643	0,3665	0,3685	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319
1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441
1,6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545
1,7	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633
1,8	0,4641	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699	0,4706
1,9	0,4713	0,4719	0,4726	0,4732	0,4738	0,4744	0,4750	0,4756	0,4761	0,4767
2,0	0,4772	0,4778	0,4783	0,4788	0,4793	0,4798	0,4803	0,4808	0,4812	0,4817
2,1	0,4821	0,4826	0,4830	0,4834	0,4838	0,4842	0,4846	0,4850	0,4854	0,4857
2,2	0,4861	0,4864	0,4868	0,4871	0,4875	0,4878	0,4881	0,4884	0,4887	0,4890
2,3	0,4893	0,4896	0,4898	0,4901	0,4904	0,4906	0,4909	0,4911	0,4913	0,4916
2,4	0,4918	0,4920	0,4922	0,4925	0,4927	0,4929	0,4931	0,4932	0,4934	0,4936
2,5	0,4938	0,4940	0,4941	0,4943	0,4945	0,4946	0,4948	0,4949	0,4951	0,4952
2,6	0,4953	0,4955	0,4956	0,4957	0,4959	0,4960	0,4961	0,4962	0,4963	0,4964
2,7	0,4965	0,4966	0,4967	0,4968	0,4969	0,4970	0,4971	0,4972	0,4973	0,4974
2,8	0,4974	0,4975	0,4976	0,4977	0,4977	0,4978	0,4979	0,4979	0,4980	0,4981
2,9	0,4981	0,4982	0,4982	0,4983	0,4984	0,4984	0,4985	0,4985	0,4986	0,4986
3,0	0,4987	0,4987	0,4987	0,4988	0,4988	0,4989	0,4989	0,4989	0,4990	0,4990
3,1	0,4990	0,4991	0,4991	0,4991	0,4992	0,4992	0,4992	0,4992	0,4993	0,4993
3,2	0,4993	0,4993	0,4994	0,4994	0,4994	0,4994	0,4994	0,4995	0,4995	0,4995
3,3	0,4995	0,4995	0,4995	0,4996	0,4996	0,4996	0,4996	0,4996	0,4996	0,4997
3,4	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4998
3,5	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998
3,6	0,4998	0,4998	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999
3,7	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999
3,8	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999
3,9	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000

*FONTE***:** COSTA NETO, P.L.O. ***Estatística***. São Paulo, Edgard Blucher, 1978.

© ALBERTO W. RAMOS 2016

TEOREMA DO LIMITE CENTRAL

Uma variável aleatória resultante da soma de n variáveis aleatórias independentes, no limite, quando $n \rightarrow \infty$, tem distribuição Normal.

TEOREMA DAS COMBINAÇÕES LINEARES

Uma variável aleatória obtida pela combinação linear de variáveis aleatórias Normais independentes tem também distribuição Normal.

© ALBERTO W. RAMOS 2016

APROXIMAÇÕES PELA NORMAL

- ***Binomial pela Normal***

  Se $np \geq 5$ e $n.(1-p) \geq 5 \Rightarrow$ vale a aproximação da distribuição Binomial pela Normal.

- ***Poisson pela Normal***

  Se $\lambda t \geq 5 \Rightarrow$ vale a aproximação da distribuição de Poisson pela Normal.

- ***Correção da Continuidade***

  Como se aproxima uma distribuição discreta (Binomial ou Poisson) por uma contínua (Normal), é necessária esta correção:

  discreta: $P(X=x_i) \rightarrow$ Normal: $P(x_i - \frac{1}{2} \leq X \leq x_i + \frac{1}{2})$

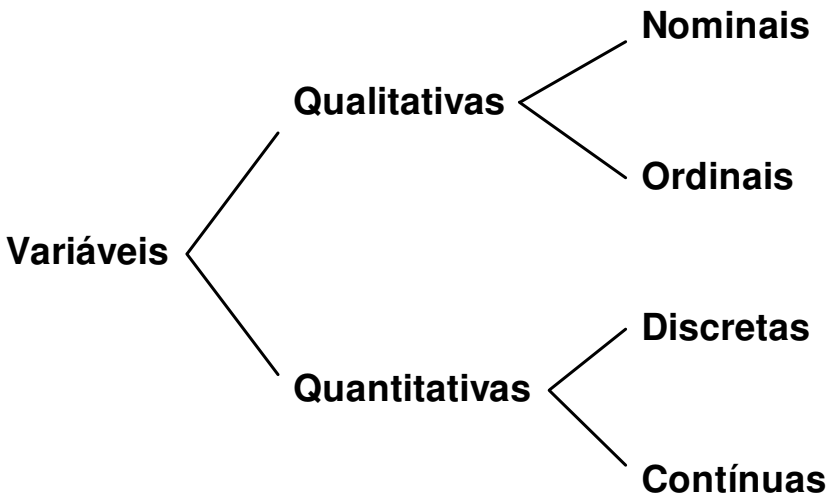
  discreta: $P(x_1 \leq X \leq x_2) \rightarrow$ Normal: $P(x_1 - \frac{1}{2} \leq X \leq x_2 + \frac{1}{2})$

© ALBERTO W. RAMOS 2016

Estatística Descritiva

© ALBERTO W. RAMOS 2016

TIPOS DE VARIÁVEIS



Exemplo (Variáveis em uma ficha cadastral PF)

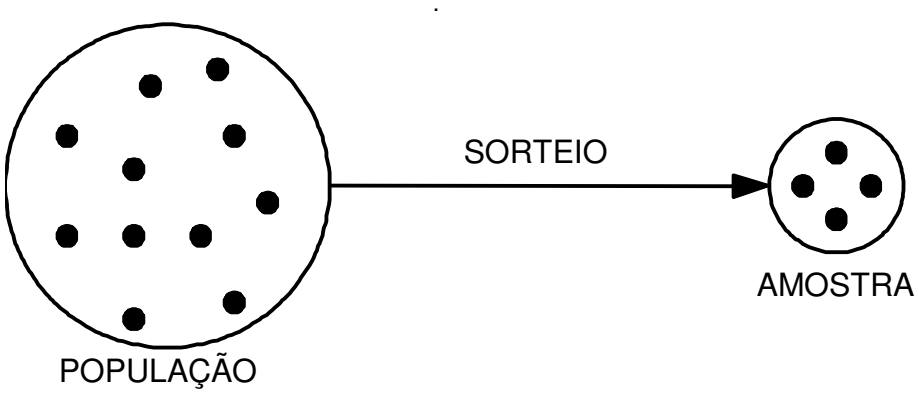
	Variável	Tipo
1	Número de dependentes	
2	Idade	
3	Local de nascimento	
4	Nível educacional	
5		Qualitativa, nominal
6		Qualitativa, ordinal
7		Quantitativa, discreta
8		Quantitativa, contínua

© ALBERTO W. RAMOS 2016

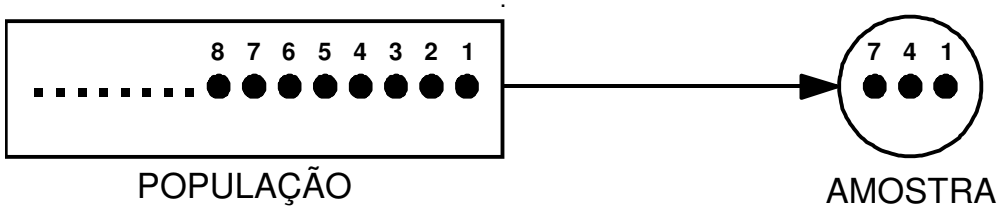
AMOSTRAGEM

Tão importante quanto determinar quantos itens devem ser examinados na amostragem (tamanho da amostra), é determinar como coletar estes itens.

1. Amostragem simples (aleatória ou casual): todos itens da população têm igual chance de pertencer à amostra (sorteio)

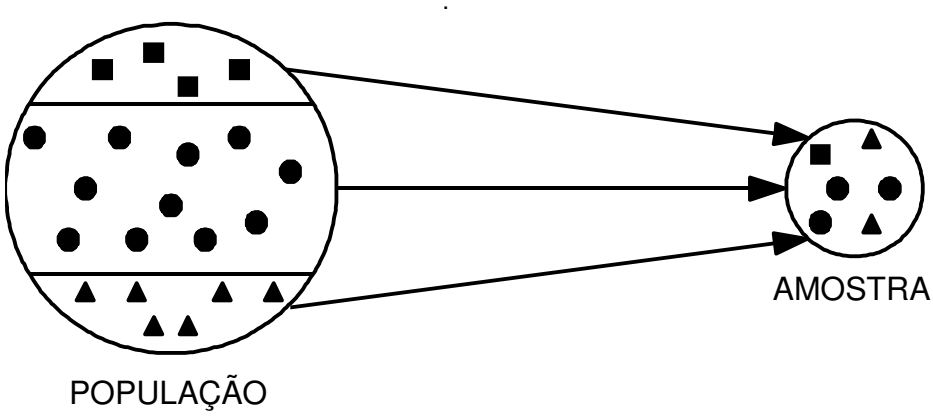


2. Amostragem sistemática: os itens encontram-se ordenados e a retirada de elementos da amostra é feita periodicamente

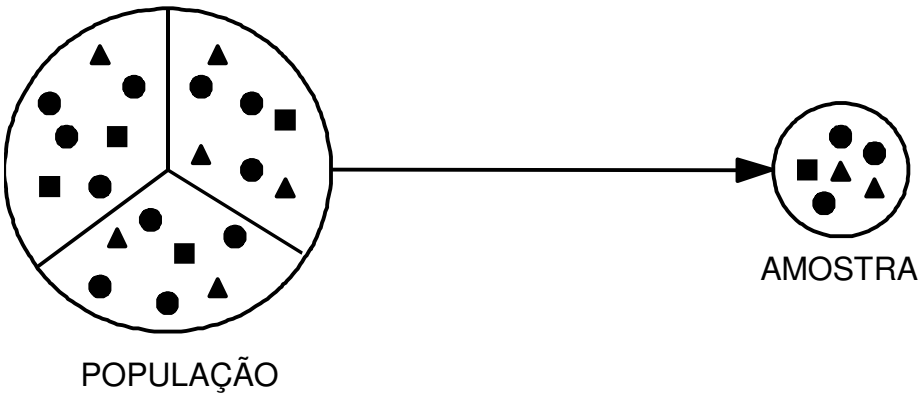


© ALBERTO W. RAMOS 2016

3. Amostragem estratificada: a população encontra-se dividida em vários estratos e as amostras são coletadas aleatoriamente de cada estrato.



4. Amostragem por agrupamentos: a população encontra-se fisicamente dividida em pequenos grupos, que são sorteados para formar a amostra.



© ALBERTO W. RAMOS 2016

Distribuições de Frequências

© ALBERTO W. RAMOS 2016

DISTRIBUIÇÕES DE FREQUÊNCIAS

Quando há uma grande quantidade de dados disponíveis ($n \geq 30$) é mais adequado trabalhar-se com estes agrupados do que com os valores individualmente

Distribuições de frequências permitem observar :

- Quantas vezes ocorre um certo resultado
- Simetria ou assimetria dos dados
- Onde se concentram mais os valores
- Qual é a variabilidade (dispersão) dos dados
- Existência de valores discrepantes (dados suspeitos)
- Estratificação (diferentes subgrupos de dados)

© ALBERTO W. RAMOS 2016

EXEMPLO

Trinta empresas foram selecionadas aleatoriamente e classificadas de acordo com seu tamanho:

Tamanho	Categoria
1 a 100 funcionários	P
101 a 500 funcionários	M
mais de 500 funcionários	G

Os resultados foram os seguintes:

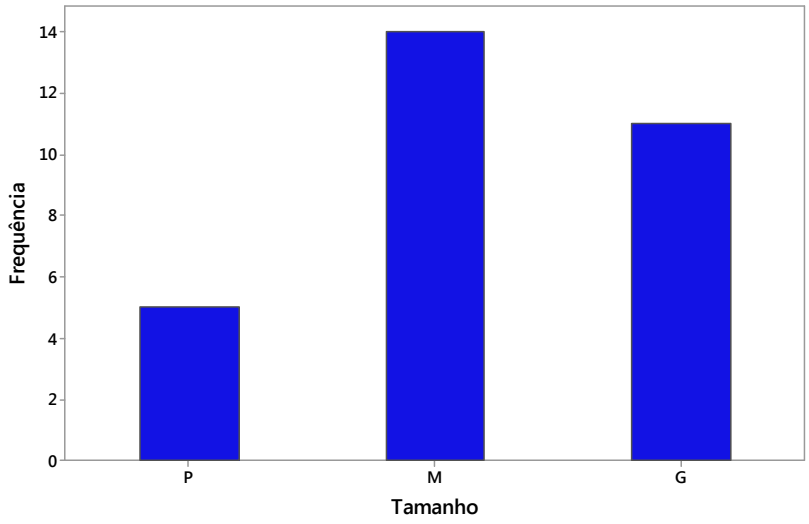
M	P	M	M	M	P	G	G	M	G
M	G	M	M	M	M	M	P	G	M
G	P	G	G	G	G	M	M	P	G

Distribuição de frequências:

Porte	f	p’
P	5	0,16
M	14	0,47
G	11	0,37
Total	30	1,00

f : frequência absoluta

p’ : frequência relativa



© ALBERTO W. RAMOS 2016

EXEMPLO

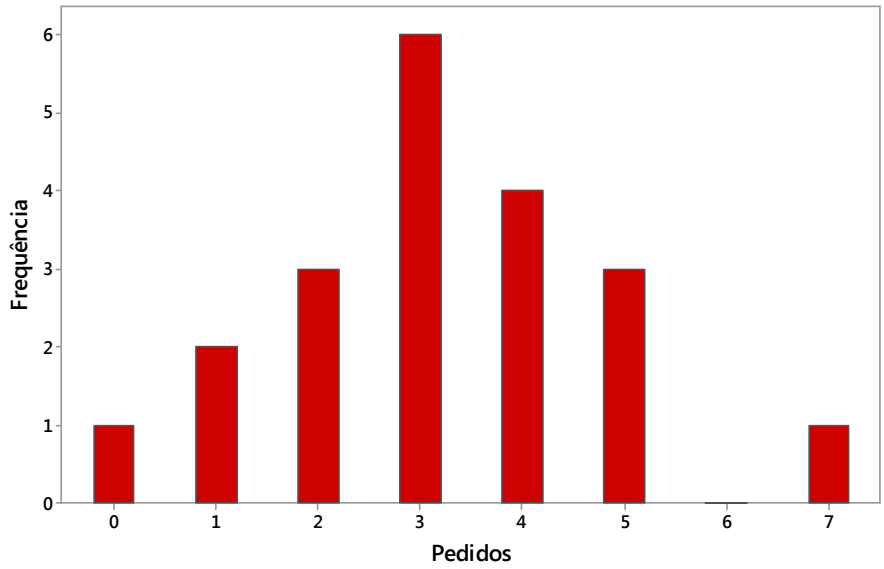
Número de pedidos de concessão de empréstimo recebidos por uma agência nas últimas 20 semanas.

(1) 2	(2) 3	(3) 5	(4) 4
(5) 7	(6) 4	(7) 2	(8) 5
(9) 1	(10) 3	(11) 3	(12) 5
(13) 3	(14) 4	(15) 0	(16) 3
(17) 4	(18) 1	(19) 2	(20) 3

Min = 0 Máx = 7 (8 valores distintos)

Distribuição de frequências :

pedidos	0	1	2	3	4	5	6	7
frequência	1	2	3	6	4	3	0	1



© ALBERTO W. RAMOS 2016

EXEMPLO

Porcentagem do faturamento recolhido sob a forma de tributos federais por 50 empresas

Souza Cruz	36,9
Autolatina	28,4
General Motors	25,7
Brahma	65,4
Philip Morris	46,7
Shell	3,4
Gessy Lever	15,8
IBM	20,6
Fiat Automóveis	14,2
Nestlé	9,0
Goodyear	15,1
Esso	1,8
Mercedes-Benz	1,9
Firestone	32,7
Pirelli	17,6
Texaco	3,8
Atlantic	4,5
Skol	37,5
Consul	14,3
Santa Marina	20,3
CBA	7,7
Antarctica Paulista	46,5
Brastemp	12,1
Suzano	12,7
Philips	4,8

Petróleo Ipiranga	3,1
Johnson & Johnson	10,1
Avon	17,8
Antarctica – Rio	28,9
Alcan	7,8
Bosch	18,9
Klabin	11,0
Glasurit	11,1
Kaiser – SP	56,1
Krupp	27,0
Carrefour	2,4
Usiminas	5,2
3M	26,0
Hoechst	8,0
Poliolefinas	22,9
Cebrasp	30,0
Arno	13,9
MBR	8,0
Estrela	3,4
Solvay	13,3
Kodak	10,2
Metal Leve	14,2
Champion	9,1
Rhodia	4,8
Antarctica – Nordeste	29,4

© ALBERTO W. RAMOS 2016

Como existem 50 (n=50) diferentes valores, vamos agrupar os dados em classes, seguindo o seguinte procedimento:

a) Obter uma amostra de 50 a 100 dados ($50 < n < 100$)

b) Determinar o maior e o menor valor (x_{max} e x_{min})

c) Calcular a amplitude total dos dados $R_T = x_{max} - x_{min}$

d) Determinar o número de classes $k = \sqrt{n}$

e) Calcular a amplitude das classes $h = R/k$

f) Determinar os limites das classes

g) Construir uma tabela de freqüências

h) Traçar o diagrama

Resolução:

- $x_{min} = 1,8$ e $x_{max} = 65,4$
- $R_T = 65,4 – 1,8 = 63,6$
- $k = \sqrt{n} = \sqrt{50} \cong 7$
- $h = 63,6/7 \cong 10$
- Limites das classes:
  Limite inferior da 1ª classe = 0
  Limite inferior da 2ª classe = 0 + 10 = 10
  Limite inferior da 3ª classe = 10+ 10 = 20 ....

© ALBERTO W. RAMOS 2016

TABELA DE FREQUÊNCIAS

Classes	***c***	***f***	***p'***	***F***	p_A
0≤x<10	4,95	17	0,34	17	0,34
10≤x<20	14,95	16	0,32	33	0,66
20≤x<30	24,95	9	0,18	42	0,84
30≤x<40	34,92	4	0,08	46	0,92
40≤x<50	44,95	2	0,04	48	0,96
50≤x<60	54,95	1	0,02	49	0,98
60≤x<70	64,95	1	0,02	50	1,00
Total		50	1,00		

Notação:

c : ponto médio (centro) da classe

f : Frequência absoluta

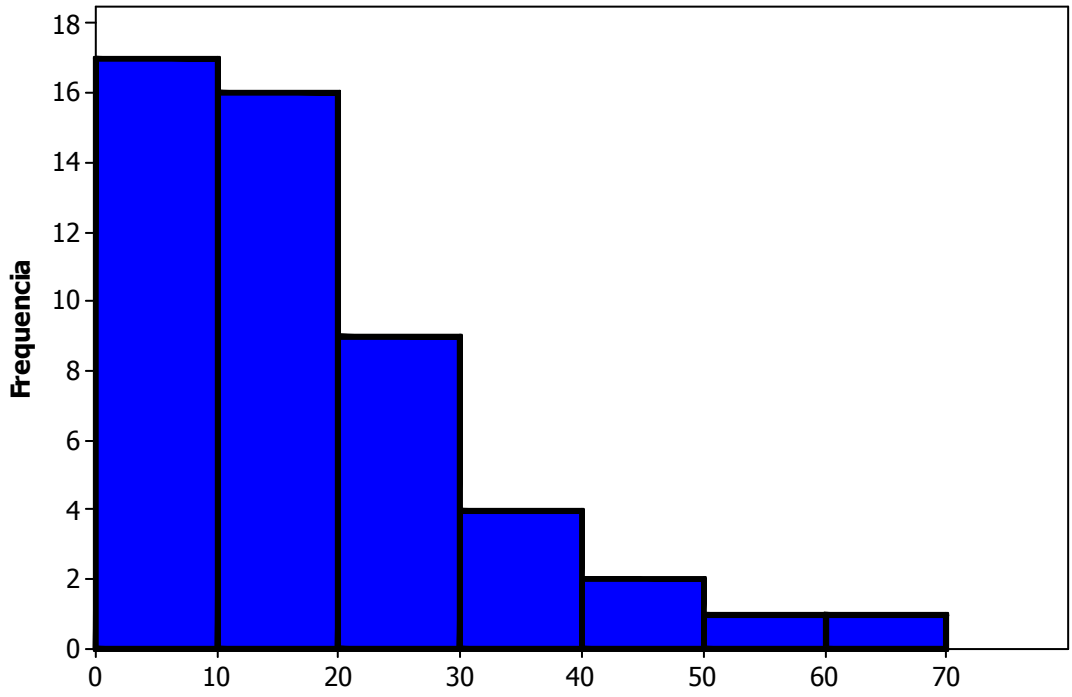
p': Proporção (frequência relativa)

F : Frequência absoluta acumulada

p_A : Frequência relativa acumulada

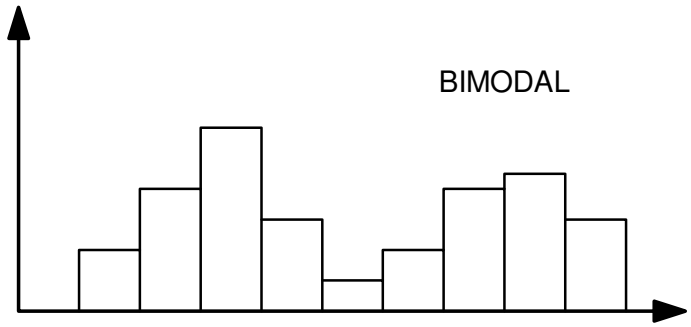
© ALBERTO W. RAMOS 2016

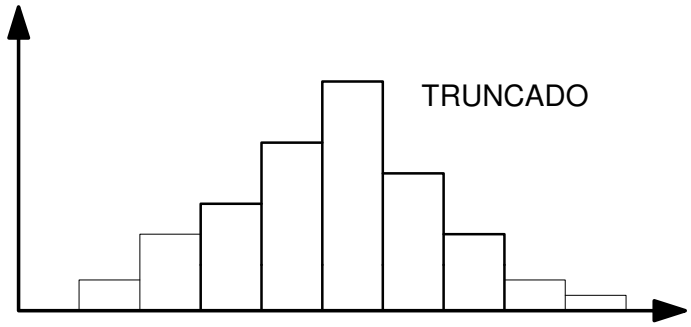
HISTOGRAMA

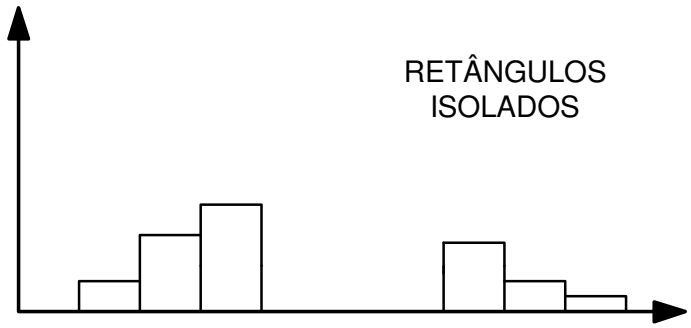


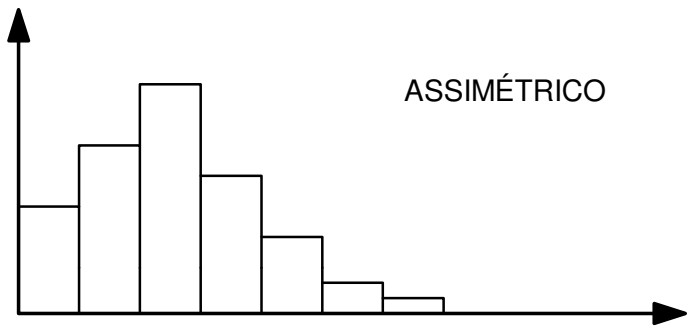
© ALBERTO W. RAMOS 2016

ANÁLISE DE HISTOGRAMAS



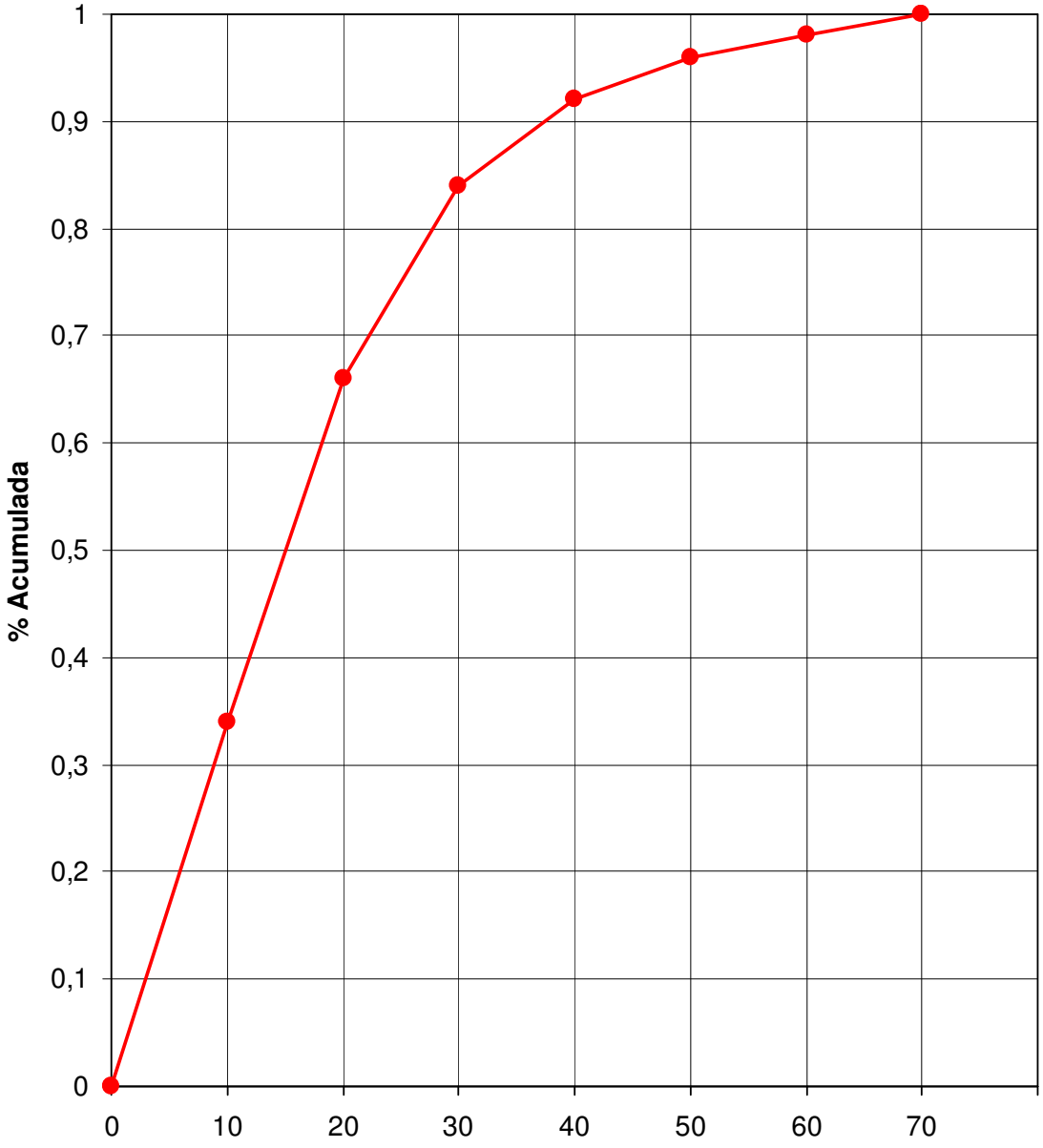






© ALBERTO W. RAMOS 2016

POLÍGONO DE FREQUÊNCIAS ACUMULADAS



© ALBERTO W. RAMOS 2016

Medidas Descritivas

© ALBERTO W. RAMOS 2016

MEDIDAS DESCRITIVAS

Quando queremos caracterizar uma amostra, podemos caracterizá-la mediante o cálculo de certas quantidades chamadas de medidas descritivas

1. Medidas de Posição

- Mediana (md ou Q_2) e Quartis (Q_1 e Q_3)
- Média ($\bar{x}$)
- Moda (m_o)

2. Medidas de Dispersão

- Amplitude (R)
- Variância (s^2)
- Desvio-padrão (s)

© ALBERTO W. RAMOS 2016

MEDIANA

A mediana é a quantidade que divide o conjunto de dados ordenados da amostra em duas metades com igual número de elementos. Quartil, por sua vez, divide em quatro metades com iguais quantidades de elementos.

Exemplo

Participação de mercado das 12 maiores seguradoras em % do valor total dos prêmios emitidos. (Fonte: FENASEG in Exame, Fev / 93)

1,9 2,0 2,1 2,5 3,0 3,1 3,3 3,7 6,1 7,7 17,1 18,7

Mediana $md = 3,2$

Primeiro Quartil $Q_1 = 2,3$

Terceiro Quartil $Q_3 = 6,9$

© ALBERTO W. RAMOS 2016

EXEMPLO

Participação de mercado das 11 principais modalidades de seguros em % do valor total dos prêmios emitidos (outras modalidades correspondem à 6,9%)

RAMO	%	
Automóvel	33,6	
Saúde	14,0	
Incêndio	12,9	⇐ Terceiro Quartil
Vida	12,2	
Riscos Diversos	5,5	
Habitação	5,3	⇐ Mediana
Transporte	3,1	
Acidentes Pessoais	2,9	
Obrigatório Veículos	1,7	⇐ Primeiro Quartil
Riscos de Engenharia	1,0	
Responsabilidade Civil *	0,9	

Fonte (Fenaseg, in Exame, Fev / 93)

© ALBERTO W. RAMOS 2016

MÉDIA ARITMÉTICA

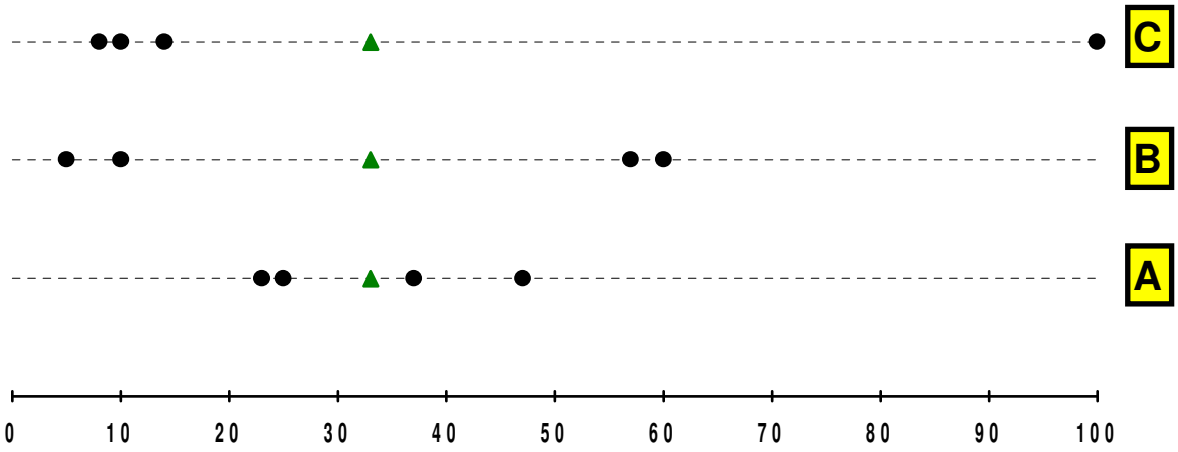
$$\bar{x} = \frac{\Sigma x}{n}$$

Exemplos

A) 23 25 37 47 $\Rightarrow \bar{x}$ (A) = 33

B) 5 10 57 60 $\Rightarrow \bar{x}$ (B) = 33

C) 8 10 14 100 $\Rightarrow \bar{x}$ (C) = 33



Quais as diferenças entre os conjuntos de dados ?

© ALBERTO W. RAMOS 2016

- Não se deve tomar decisões baseadas apenas na média

- A média nem sempre é o valor "mais comum" ou "mais típico" ou "ponto de simetria".

- A média é influenciada por dados suspeitos (*outliers*)

  Salários de uma empresa (US$)

  800 800 1000 1400 96000 $\Rightarrow$ $\bar{x}$ = 20000

© ALBERTO W. RAMOS 2016

VARIÂNCIA E DESVIO-PADRÃO

Variância:

$$s^2 = \frac{\Sigma(x - \bar{x})^2}{n - 1}$$

Desvio-Padrão:

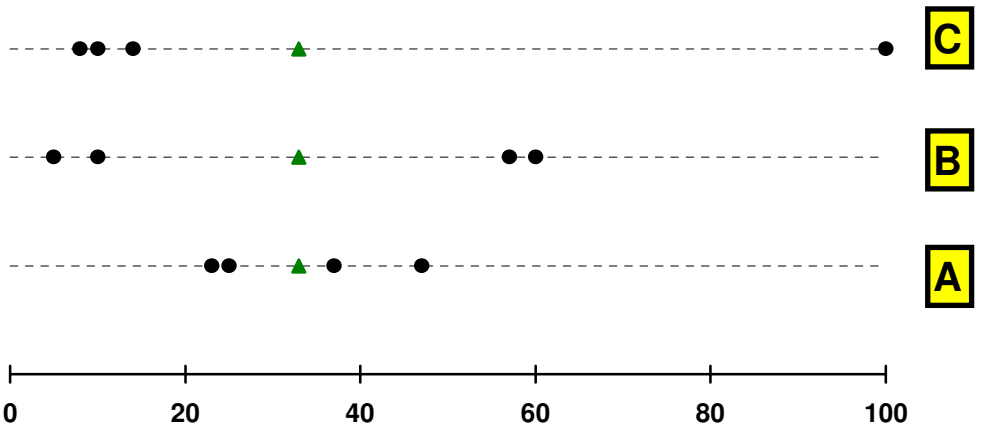
$$s = \sqrt{s^2}$$

Exemplo

A) 23 25 37 47 $\Rightarrow s^2 = 125{,}33$ $s = 11{,}2$

B) 5 10 57 60 $\Rightarrow s^2 = 872{,}67$ $s = 29{,}5$

C) 8 10 14 100 $\Rightarrow$ [ ]



© ALBERTO W. RAMOS 2016

DADOS AGRUPADOS

Só devem ser utilizados quando não se dispuser dos dados originais, uma vez que são valores aproximados. Em geral, a aproximação é boa.

Classes	***c***	***f***	***p'***	***F***	p_A
0≤x<10	4,95	17	0,34	17	0,34
10≤x<20	14,95	16	0,32	33	0,66
20≤x<30	24,95	9	0,18	42	0,84
30≤x<40	34,92	4	0,08	46	0,92
40≤x<50	44,95	2	0,04	48	0,96
50≤x<60	54,95	1	0,02	49	0,98
60≤x<70	64,95	1	0,02	50	1,00
Total		50	1,00		

$$\bar{x} = \frac{\Sigma f.c}{n} \Rightarrow \bar{x} = \text{———} =$$

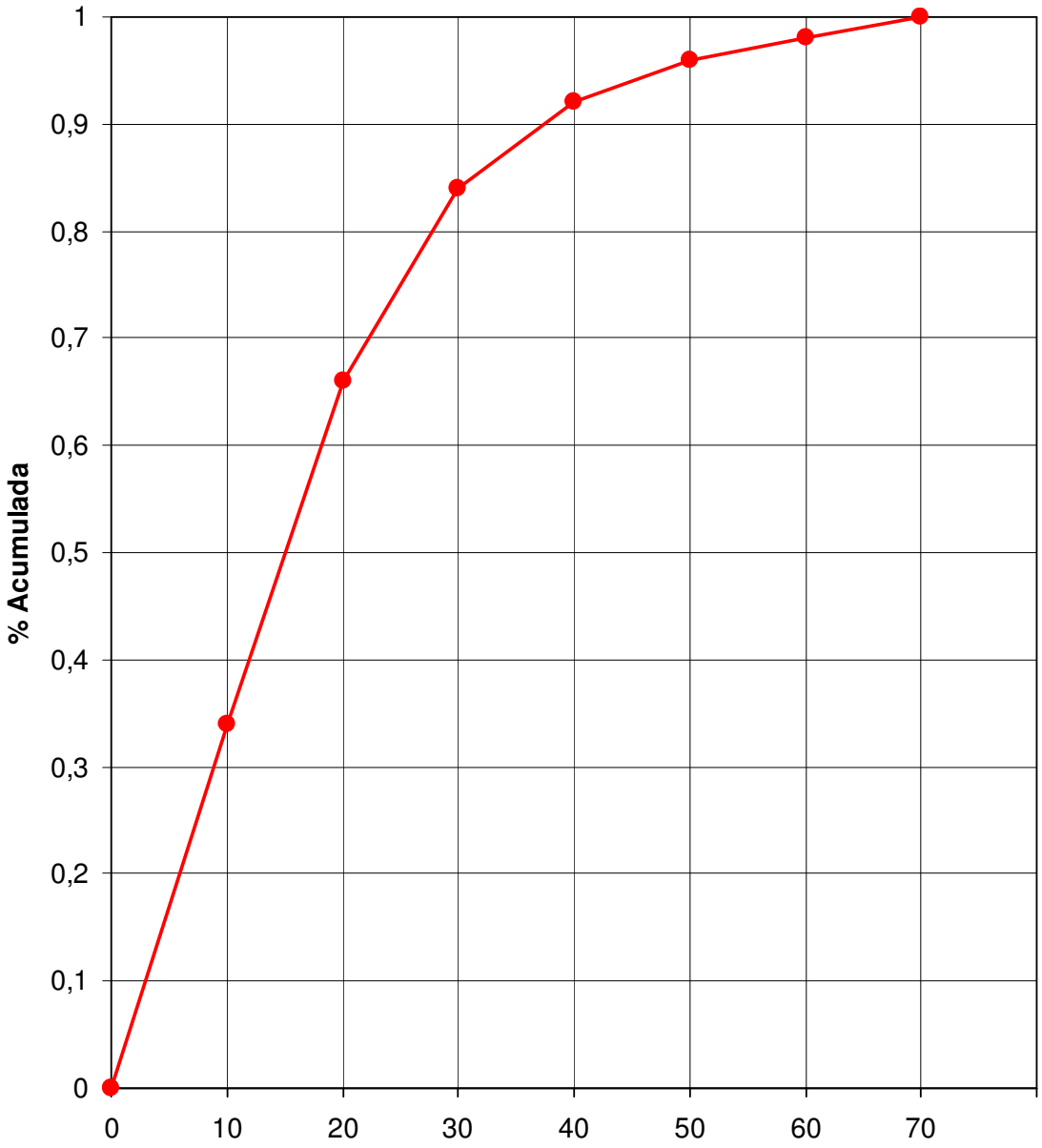
$$s^2 = \frac{\sum (c - \bar{x})^2 f}{n-1} \Rightarrow s^2 = \frac{\Sigma f.c^2 - \dfrac{(\Sigma f.c)^2}{n}}{n-1}$$

$$s^2 = \text{———————————} =$$

© ALBERTO W. RAMOS 2016

MEDIANA E QUARTIS

Pode-se calculá-los utilizando o polígono de frequências acumuladas e observando o valor no eixo das abscissas.



© ALBERTO W. RAMOS 2016

Distribuições Amostrais

© ALBERTO W. RAMOS 2016

DISTRIBUIÇÃO DA MÉDIA ($\bar{x}$)

Como:

$$\bar{x} = \frac{\sum x_i}{n} = \frac{1}{n}(x_1 + x_2 + x_3 + ... + x_n)$$

então:

$$\mu(\bar{x}) = \frac{1}{n}[\mu(x_1) + \mu(x_2) + ... + \mu(x_n)]$$

$$= \frac{1}{n}[\mu + \mu + ... + \mu] = \frac{n.\mu}{n} = \mu$$

por outro lado:

$$\sigma^2(\bar{x}) = \left(\frac{1}{n}\right)^2 [\sigma^2(x_1) + \sigma^2(x_2) + ... + \sigma^2(x_n)]$$

$$= \frac{1}{n^2}[\sigma^2 + \sigma^2 + ... + \sigma^2] = \frac{n.\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

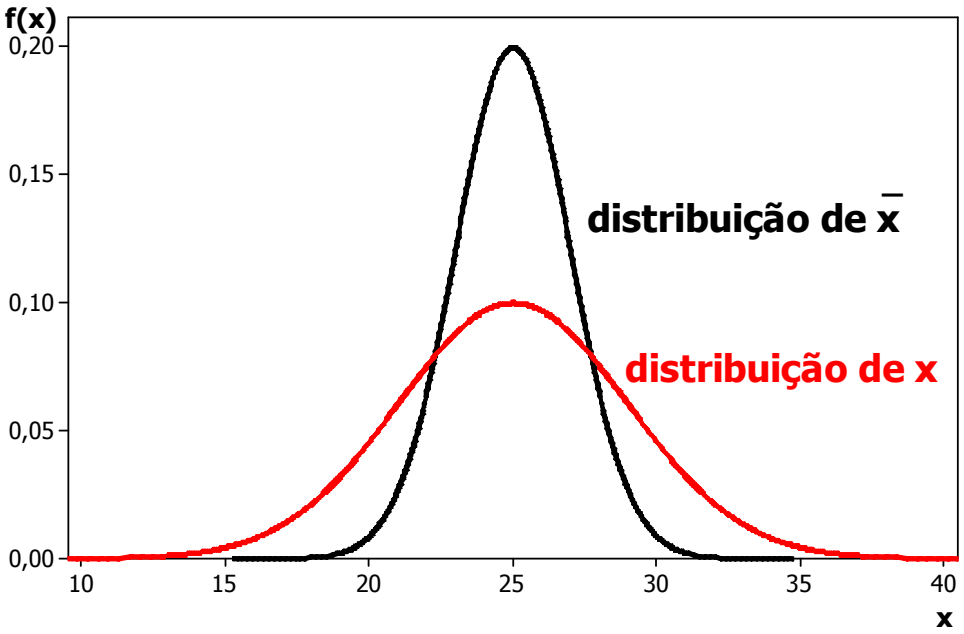
$$\sigma(\bar{x}) = \frac{\sigma}{\sqrt{n}}$$

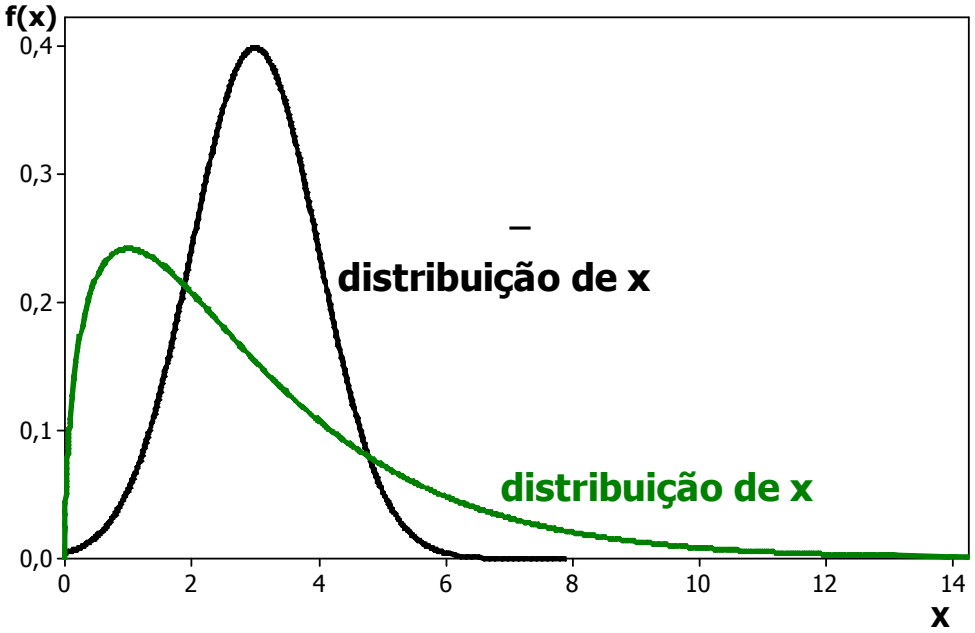
Obs.: se a amostragem é sem reposição e a população é finita, então:

$$\sigma^2(\bar{X}) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$$

© ALBERTO W. RAMOS 2016

A distribuição de $\bar{x}$ é normal, devido ao Teorema Limite Central, desde que n seja suficientemente grande.





© ALBERTO W. RAMOS 2016

DISTRIBUIÇÃO χ^2 (QUI-QUADRADO)

$$\chi^2_\nu = \sum_{i=1}^{\nu}\left(\frac{x_i - \mu}{\sigma}\right)^2 = \sum_{i=1}^{\nu} z_i^2$$

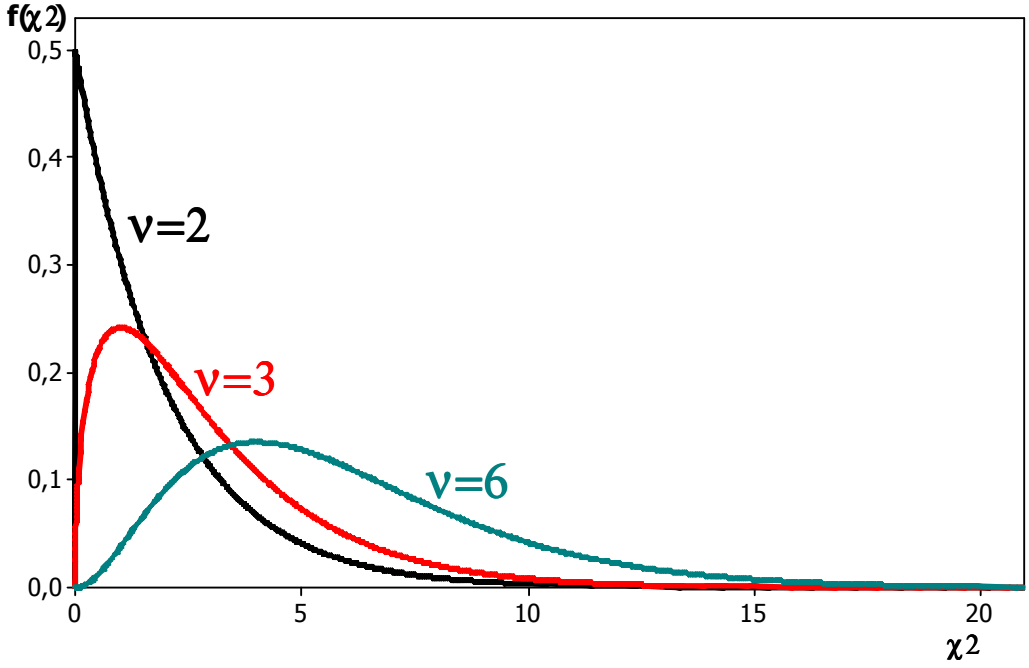
$$\mu(\chi^2_\nu) = \mu(\sum z_i^2) = \nu.\mu(z_i^2) = \nu$$

$$\sigma^2(\chi^2_\nu) = 2.\nu$$

Propriedades:

a) Quando $\nu \to \infty$, $\chi^2_\nu \to$ Normal (Teorema do Limite Central)

b) Se as variáveis são independentes $\Rightarrow \chi^2_{\nu_1} + \chi^2_{\nu_2} = \chi^2_{\nu_1+\nu_2}$



© ALBERTO W. RAMOS 2016

TABELA DA DISTRIBUIÇÃO χ^2

	α					
ν	**0,99**	**0,95**	**0,90**	**0,10**	**0,05**	**0,01**
1	0,0002	0,0039	0,0158	2,706	3,841	6,635
2	0,0201	0,103	0,211	4,605	5,991	9,210
3	0,115	0,352	0,584	6,251	7,815	11,345
4	0,297	0,711	1,064	7,779	9,488	13,277
5	0,554	1,145	1,610	9,236	11,070	15,086
6	0,872	1,635	2,204	10,645	12,592	16,812
7	1,239	2,167	2,833	12,017	14,067	18,475
8	1,646	2,733	3,490	13,362	15,507	20,090
9	2,088	3,325	4,168	14,684	16,919	21,666
10	2,558	3,940	4,865	15,987	18,307	23,209
11	3,053	4,575	5,578	17,275	19,675	24,725
12	3,571	5,226	6,304	18,549	21,026	26,217
13	4,107	5,892	7,042	19,812	22,362	27,688
14	4,660	6,571	7,790	21,064	23,685	29,141
15	5,229	7,261	8,547	22,307	24,996	30,578
16	5,812	7,962	9,312	23,542	26,296	32,000
17	6,408	8,672	10,085	24,769	27,587	33,409
18	7,015	9,390	10,865	25,989	28,869	34,805
19	7,633	10,117	11,651	27,204	30,144	36,191
20	8,260	10,851	12,443	28,412	31,410	37,566
21	8,897	11,591	13,240	29,615	32,671	38,932
22	9,542	12,338	14,041	30,813	33,924	40,289
23	10,196	13,091	14,848	32,007	35,172	41,638
24	10,856	13,848	15,659	33,196	36,415	42,980
25	11,524	14,611	16,473	34,382	37,652	44,314
26	12,198	15,379	17,292	35,563	38,885	45,642
27	12,879	16,151	18,114	36,741	40,113	46,963
28	13,565	16,928	18,939	37,916	41,337	48,278
29	14,256	17,708	19,768	39,087	42,557	49,588
30	14,953	18,493	20,599	40,256	43,773	50,892
40	22,164	26,509	29,051	51,805	55,758	63,691
50	29,707	34,764	37,689	63,167	67,505	76,154
60	37,485	43,188	46,459	74,397	79,082	88,379

*FONTE***:** COSTA NETO, P.L.O. ***Estatística***. São Paulo, Edgard Blucher, 1978.

© ALBERTO W. RAMOS 2016

DISTRIBUIÇÃO DA VARIÂNCIA (s^2)

Relembrando:

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

Se $X \sim N(\mu;\sigma) \Rightarrow s^2 \sim \chi^2_{n-1}$. Note que μ foi substituído por $\bar{x}$.

$$\chi^2_{n-1} = \sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\sigma}\right)^2 = \frac{\sum (x_i - \bar{x})^2}{\sigma^2} = \frac{\sum (x_i - \bar{x})^2}{n-1} \times \frac{n-1}{\sigma^2} = \frac{(n-1)s^2}{\sigma^2}$$

Logo:

$$s^2 = \frac{\sigma^2}{n-1}\chi^2_{n-1}$$

$$\mu(s^2) = \frac{\sigma^2}{n-1}\mu(\chi^2_{n-1}) = \frac{\sigma^2}{n-1}(n-1) = \sigma^2$$

$$\sigma^2(s^2) = \frac{\sigma^4}{(n-1)^2}\sigma^2(\chi^2_{n-1}) = \frac{\sigma^4}{(n-1)^2} \times 2(n-1) = \frac{2.\sigma^4}{n-1}$$

© ALBERTO W. RAMOS 2016

DISTRIBUIÇÃO DA FREQUÊNCIA (f)

Se a população é infinita ou amostragem com reposição $\Rightarrow$ p (probabilidade de sucesso) é constante.

Distribuição de f: Binomial.

$$\mu(f) = n.p$$

$$\sigma^2(f) = n.p.(1-p)$$

DISTRIBUIÇÃO DA PROPORÇÃO (p')

$$p' = \frac{f}{n}$$

Distribuição de p': Binomial.

$$\mu(p') = \mu\left(\frac{f}{n}\right) = \frac{\mu(f)}{n} = \frac{n.p}{n} = p$$

$$\sigma^2(p') = \sigma^2\left(\frac{f}{n}\right) = \frac{\sigma^2(f)}{n^2} = \frac{n.p.(1-p)}{n^2} = \frac{p.(1-p)}{n}$$

© ALBERTO W. RAMOS 2016

DISTRIBUIÇÃO t-STUDENT

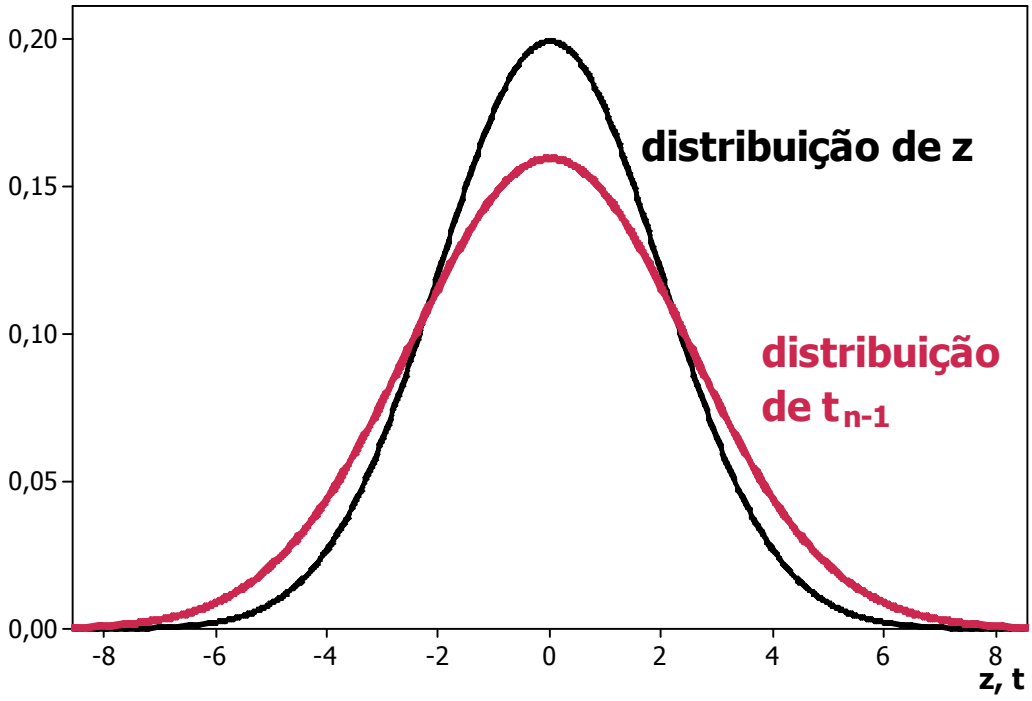
Se X tem distribuição normal com média μ e desvio-padrão σ $\Rightarrow$ $\bar{x}$ também terá distribuição com média μ, mas com desvio-padrão $\sigma/\sqrt{n}$

Pode-se definir:

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Se no lugar de σ for usado **s**, então:

$$t_{n-1} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$



© ALBERTO W. RAMOS 2016

TABELA DA DISTRIBUIÇÃO t-STUDENT

	α			
ν	**0,100**	**0,050**	**0,025**	**0,010**
1	3,078	6,314	12,706	31,821
2	1,886	2,920	4,303	6,965
3	1,638	2,353	3,182	4,541
4	1,533	2,132	2,776	3,747
5	1,476	2,015	2,571	3,365
6	1,440	1,943	2,447	3,143
7	1,415	1,895	2,365	2,998
8	1,397	1,860	2,306	2,896
9	1,383	1,833	2,262	2,821
10	1,372	1,812	2,228	2,764
11	1,363	1,796	2,201	2,718
12	1,356	1,782	2,179	2,681
13	1,350	1,771	2,160	2,650
14	1,345	1,761	2,145	2,624
15	1,341	1,753	2,131	2,602
16	1,337	1,746	2,120	2,583
17	1,333	1,740	2,110	2,567
18	1,330	1,734	2,101	2,552
19	1,328	1,729	2,093	2,539
20	1,325	1,725	2,086	2,528
21	1,323	1,721	2,080	2,518
22	1,321	1,717	2,074	2,508
23	1,319	1,714	2,069	2,500
24	1,318	1,711	2,064	2,492
25	1,316	1,708	2,060	2,485
30	1,310	1,697	2,042	2,457
50	1,299	1,676	2,009	2,403
80	1,292	1,664	1,990	2,374
120	1,289	1,657	1,980	2,351
∞	1,282	1,645	1,960	2,326

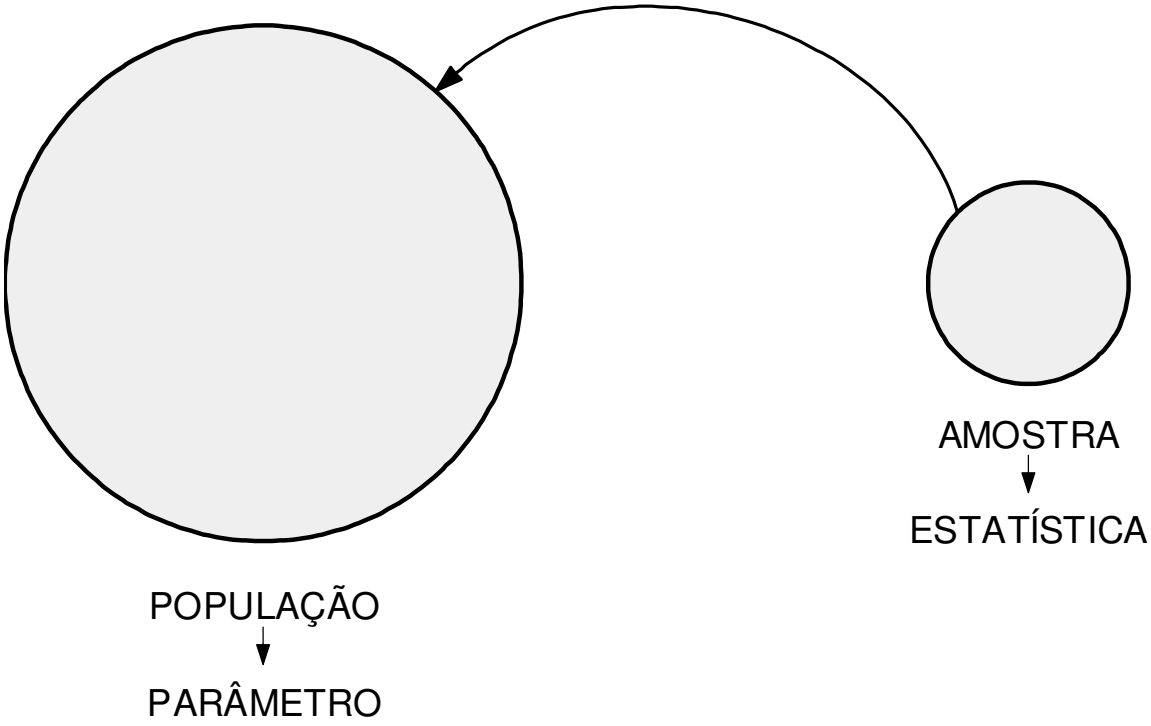
*FONTE***:** COSTA NETO, P.L.O. ***Estatística***. São Paulo, Edgard Blucher, 1978.

© ALBERTO W. RAMOS 2016

Estimação de Parâmetros

© ALBERTO W. RAMOS 2016

Quando um parâmetro de uma população é desconhecido, vamos estimá-lo a partir das estatísticas fornecidas pelas amostras.



© ALBERTO W. RAMOS 2016

ESTIMADOR E ESTIMATIVA

Estimador: Quantidade calculada em função dos elementos da amostra, que será usada na estimação do parâmetro.

Estimativa: Um certo valor de um estimador.

EXEMPLO

113 - 124 - 115 - 107 - 120 - 115 – 110

estimador (de μ)	estimativa
$\bar{x}$	114,9
$\tilde{x}$	115
m_0	115

© ALBERTO W. RAMOS 2016

ESTIMAÇÃO POR PONTO

A estimação por ponto consiste em fornecer um único valor, que é a melhor estimativa para o parâmetro da população.

a) Estimação com base em uma amostra

Parâmetro Estimado	Melhor Estimador	Observações
μ	$\overline{x} = \frac{\sum x_i}{n}$	
σ^2	$s^2 = \frac{\sum (x_i - \mu)^2}{n}$ $s^2 = \frac{\sum (x_i - \overline{x})^2}{n-1}$	μ conhecido μ desconhecido
σ	$s = \sqrt{s^2}$ $s = \frac{1}{c_4}\sqrt{s^2}$	$n \geq 30$ $n < 30$
p	$p' = \frac{f}{n}$	

© ALBERTO W. RAMOS 2016

b) Estimação com base em várias (k) amostras

Amostra	Valores	$\bar{x}$	s^2
1	$x_{11}\ x_{12}\ x_{13}\ \ldots\ x_{1n}$	$\bar{x}_1$	s_1^2
2	$x_{21}\ x_{22}\ x_{23}\ \ldots\ x_{2n}$	$\bar{x}_2$	s_2^2
3	$x_{31}\ x_{32}\ x_{33}\ \ldots\ x_{3n}$	$\bar{x}_3$	s_3^2
.	.	.	.
.	.	.	.
.	.	.	.
k	$x_{k1}\ x_{k2}\ x_{k3}\ \ldots\ x_{kn}$	$\bar{x}_k$	s_k^2

$$\bar{\bar{x}} = \frac{n_1.\bar{x}_1 + n_2.\bar{x}_2 + n_3.\bar{x}_3 + \ldots + n_k.\bar{x}_k}{n_1 + n_2 + n_3 + \ldots + n_k}$$

Se $n_1 = n_2 = n_3 = \ldots = n_k \Rightarrow \bar{\bar{x}} = \frac{\sum \bar{x}_i}{k}$

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \ldots + (n_k - 1)s_k^2}{n_1 + n_2 + \ldots + n_k - k}$$

Se $n_1 = n_2 = n_3 = \ldots = n_k \Rightarrow s_p^2 = \frac{\sum s_i^2}{k}$

$$p'_p = \frac{n_1.p'_1 + n_2.p'_2 + \ldots + n_k.p'_k}{n_1 + n_2 + \ldots + n_k}$$

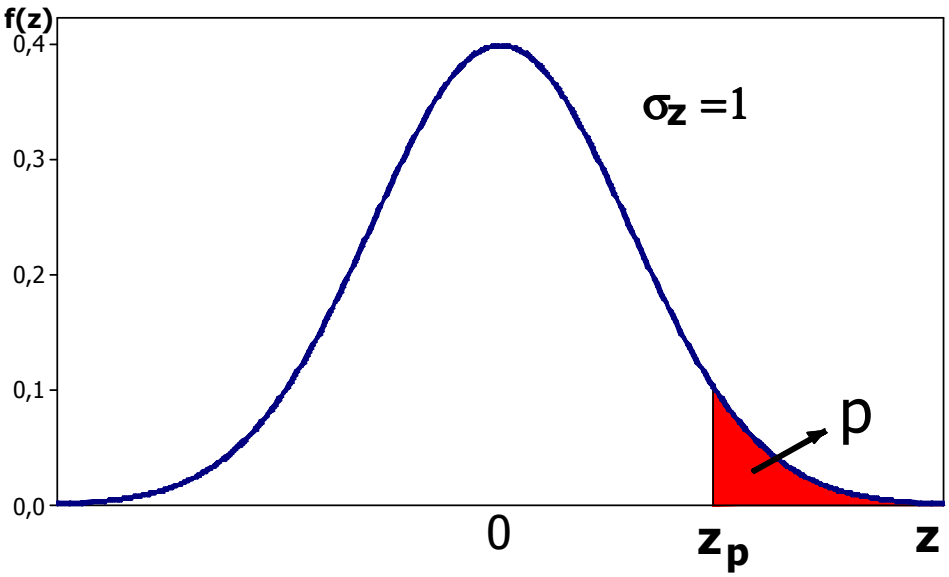
Se $n_1 = n_2 = n_3 = \ldots = n_k \Rightarrow p'_p = \frac{\sum p'_i}{k}$

© ALBERTO W. RAMOS 2016

ESTIMAÇÃO POR INTERVALO

Todas as estimativas por ponto contêm um erro, pois são diferentes do valor do parâmetro, embora próximas. Para avaliar a magnitude do erro de estimação, constrói-se um "Intervalo de Confiança (IC)" em torno da estimativa, com probabilidade conhecida.

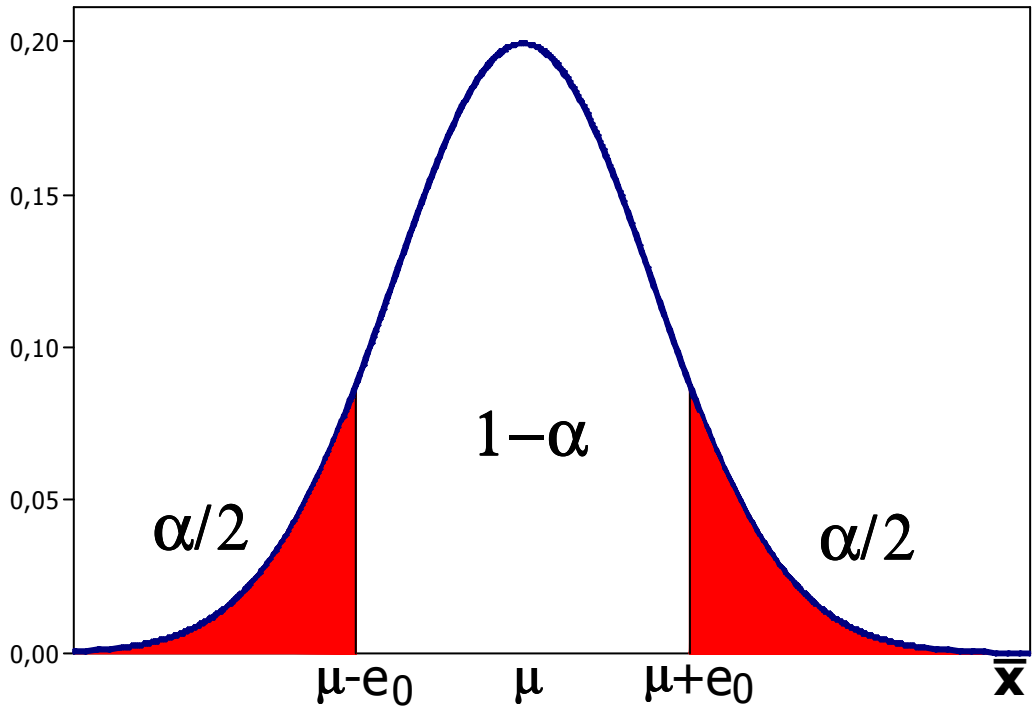
Notação:



$\mu \Rightarrow$ média da população
$\bar{x} \Rightarrow$ média da amostra
$\sigma \Rightarrow$ desvio-padrão da população
$s \Rightarrow$ desvio-padrão da amostra
$n \Rightarrow$ tamanho da amostra
$e_0 \Rightarrow$ semi-amplitude do IC $\Rightarrow$ IC = $2.e_0$

© ALBERTO W. RAMOS 2016

a) IC para μ com σ conhecido:



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

$$\mu - e_0 \leq \bar{x} \qquad \bar{x} \leq \mu + e_0$$

$$\mu \leq \bar{x} + e_0 \qquad \bar{x} - e_0 \leq \mu$$

$$\bar{x} - e_0 \leq \mu \leq \bar{x} + e_0$$

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

© ALBERTO W. RAMOS 2016

$$\frac{(\mu + e_0) - \mu}{\frac{\sigma}{\sqrt{n}}} = z_{\alpha/2}$$

$$\therefore e_0 = z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}$$

IC para μ:

$$\bar{x} \pm z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}$$

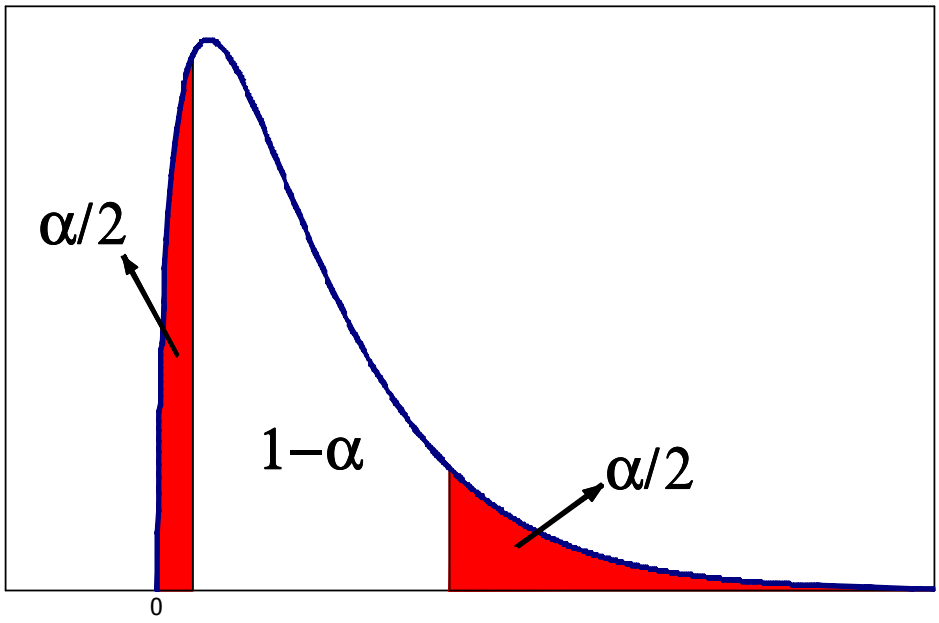
b) IC para μ com σ desconhecido:

Se σ é desconhecido, substitui-se por s na fórmula anterior e, neste caso, passa-se a ter um t-Student com n-1 graus de liberdade.

$$\bar{x} \pm t_{n-1;\alpha/2} \times \frac{s}{\sqrt{n}}$$

© ALBERTO W. RAMOS 2016

c) IC para σ^2:



$$\chi^2_{n-1} = \frac{(n-1)}{\sigma^2} s^2$$

$$P(\chi^2_{n-1;1-\alpha/2} \leq \chi^2_{n-1} \leq \chi^2_{n-1;\alpha/2}) = 1-\alpha$$

$$\chi^2_{n-1;1-\alpha/2} \leq \frac{(n-1)s^2}{\sigma^2} \leq \chi^2_{n-1;\alpha/2}$$

$$\frac{(n-1)s^2}{\chi^2_{n-1;\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{n-1;1-\alpha/2}}$$

$$P\left(\frac{(n-1)s^2}{\chi^2_{n-1;\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{n-1;1-\alpha/2}} \right) = 1-\alpha$$

© ALBERTO W. RAMOS 2016

d) IC para σ

$$P\left(\sqrt{\frac{(n-1)s^2}{\chi^2_{n-1;\alpha/2}}} \le \sigma \le \sqrt{\frac{(n-1)s^2}{\chi^2_{n-1;1-\alpha/2}}}\right) = 1-\alpha$$

e) IC para p

p' tem Distribuição Binomial $\rightarrow$ $\mu(p') = p$

$$\sigma^2(p') = \frac{p(1-p)}{n}$$

Se n.p ≥ 5 e n.(1-p) ≥ 5 $\Rightarrow$ vale aproximação pela Normal.

$$e_0 = z_{\alpha/2} \cdot \sqrt{\frac{p(1-p)}{n}}$$

Como não conhecemos p, usamos p':

$$p' \pm z_{\alpha/2} \cdot \sqrt{\frac{p'.(1-p')}{n}}$$

© ALBERTO W. RAMOS 2016

TAMANHO DE AMOSTRAS (PARA ESTIMAÇÃO)

a) Média:

. Se σ conhecido:

$$e_0 = z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}$$

$$n = \left(\frac{z_{\alpha/2}}{e_0} \sigma \right)^2$$

. Se σ desconhecido:

$$n = \left(\frac{t_{n'-1;\alpha/2}}{e_0} s \right)^2$$

n' = tamanho da amostra-piloto

b) Proporção Populacional (probabilidade):

$$n = \left(\frac{z_{\alpha/2}}{e_0} \right)^2 p'(1-p')$$

Se não há estimativa para p, adotar p' = 1/2.

© ALBERTO W. RAMOS 2016

Testes de Hipóteses

© ALBERTO W. RAMOS 2016

Com base nos resaultados da amostra, quer se testar uma certa hipótese (considerada como válida, até prova em contrário), a respeito de um parâmetro da população.

Notação:

H_0 = hipótese nula, a ser testada
H_1 = hipótese alternativa

Exemplo:

H_0 = o réu é inocente
H_1 = o réu é culpado

Vai se obter uma amostra e:

- aceito H_0 (fraco)
- rejeito H_0 e **afirmo** H_1 (forte)

© ALBERTO W. RAMOS 2016

TIPOS DE ERROS

Dois tipos de erros podem ser cometidos em testes de hipóteses:

a) Erro tipo I: rejeitar H_0 quando H_0 é verdadeira.
Ex.: juiz condenar um réu inocente.

b) Erro tipo II: aceitar H_0 quando H_0 é falsa.
Ex.: juiz absolve um réu culpado.

Cada tipo de erro tem uma certa probabilidade de ser cometido (α e β, respectivamente).

		REALIDADE	
		H_0 verdadeira	H_0 falsa
DECISÃO	aceitar H_0	**decisão correta** **$1 - \alpha$**	**erro tipo II** **β**
	rejeitar H_0	**erro tipo I** **α**	**decisão correta** **$1 - \beta$**

© ALBERTO W. RAMOS 2016

EXEMPLO

Um certo fabricante de pneus afirma que estes têm duração média de 45.000 Km e desvio-padrão de 3.000 Km. Uma empresa adquiriu um lote deste produto, retirou e testou uma amostra de 16 pneus que forneceu $\bar{x}$ = 44.175 Km. Qual a decisão: aceitar ou rejeitar o lote?

DICA: ***quando montar as hipóteses, colocar sempre em H_1 aquilo que se deseja afirmar ou provar.***

© ALBERTO W. RAMOS 2016

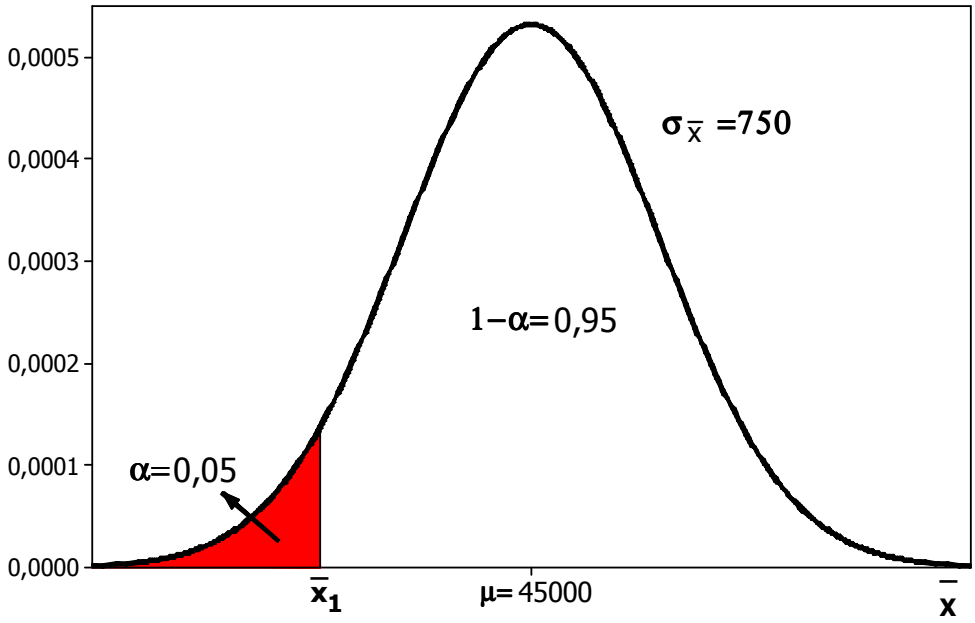
Hipóteses:

$$H_0: \mu \geq 45.000$$
$$H_1: \mu < 45.000$$

Se H_0 é verdadeira $\Rightarrow \bar{x} \sim N(\mu; \sigma/\sqrt{n})$

$$\mu = 45.000$$

$$\frac{\sigma}{\sqrt{n}} = \frac{3.000}{\sqrt{16}} = 750$$



$$z_1 = -1{,}645 \Rightarrow -1{,}645 = \frac{\bar{x}_1 - 45.000}{750} \Rightarrow \bar{x}_1 \cong 43.766$$

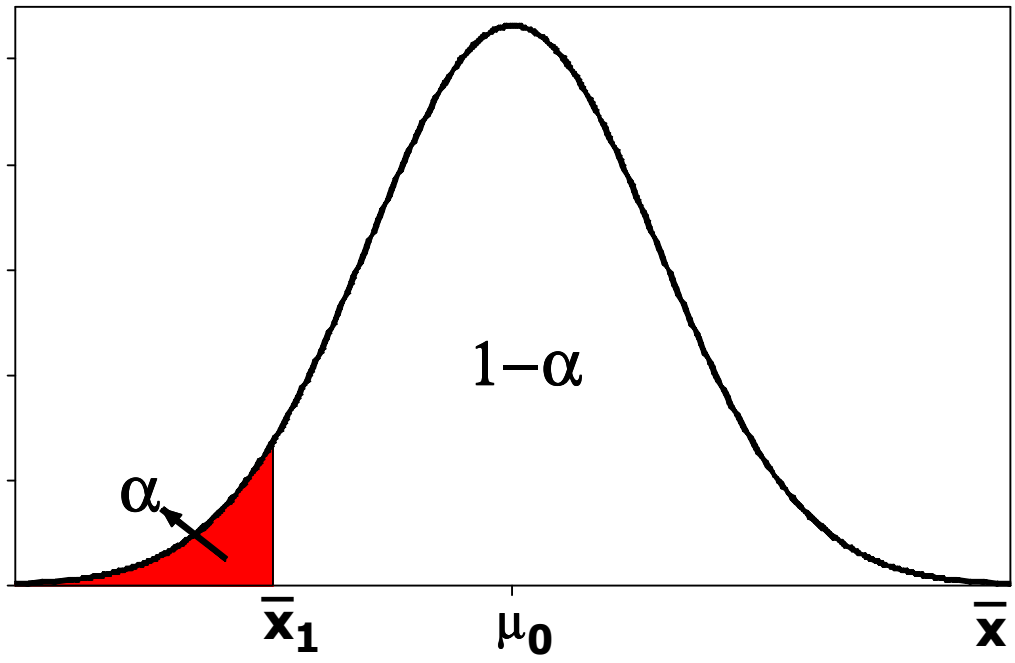
Como $\bar{x} = 44.175$, ainda cai na região de probabilidade 1 - α = 95% $\Rightarrow$ aceito H_0

© ALBERTO W. RAMOS 2016

TESTES PARA A MÉDIA

A) σ conhecido:

1º Caso: $H_0: \mu = \mu_0$
$H_1: \mu < \mu_0$



$$\bar{x}_1 = \mu_0 - z_\alpha \cdot \frac{\sigma}{\sqrt{n}}$$

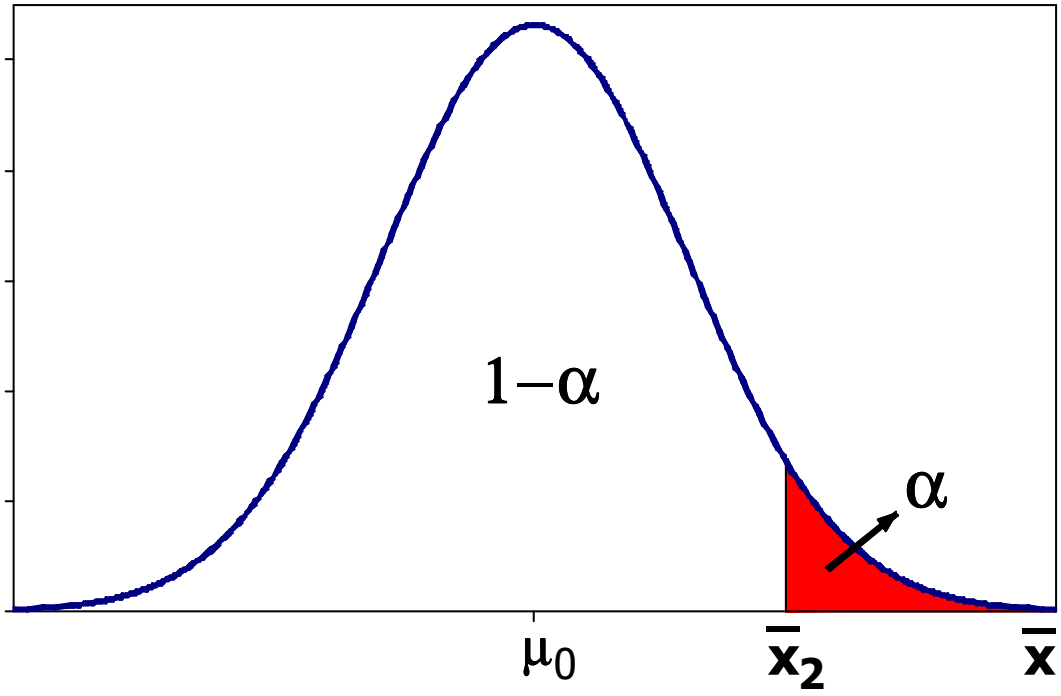
Se $\bar{x}_{CALC} < \bar{x}_1 \Rightarrow$ rejeito H_0

ou se $z_{CALC} = \dfrac{\bar{x}_{CALC} - \mu_0}{\dfrac{\sigma}{\sqrt{n}}} < -z_\alpha \Rightarrow$ rejeito H_0

© ALBERTO W. RAMOS 2016

2º Caso: $H_0: \mu \leq \mu_0$

$H_1: \mu > \mu_0$



$$\bar{x}_2 = \mu_0 + z_\alpha \cdot \frac{\sigma}{\sqrt{n}}$$

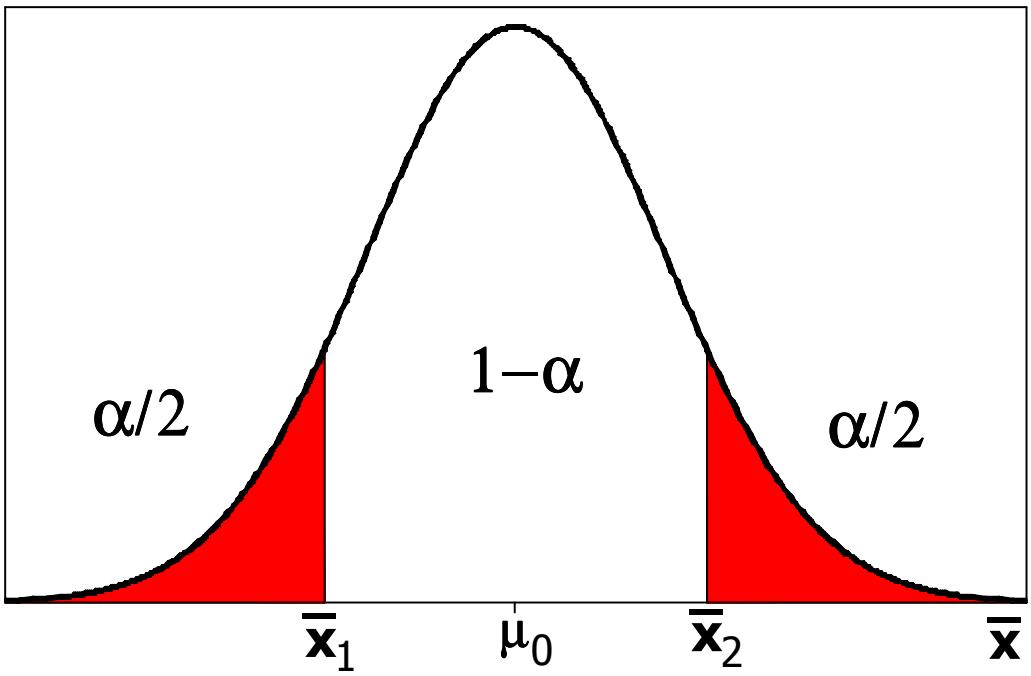
Se $\bar{x}_{CALC} > \bar{x}_2 \Rightarrow$ rejeito H_0

ou se $z_{CALC} > z_\alpha \Rightarrow$ rejeito H_0

© ALBERTO W. RAMOS 2016

3º Caso: $H_0: \mu = \mu_0$

$H_1: \mu \neq \mu_0$



$$\overline{x}_1 = \mu_0 - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

$$\overline{x}_2 = \mu_0 + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

Se $\overline{x}_{CALC} < \overline{x}_1$ ou $\overline{x}_{CALC} > \overline{x}_2 \Rightarrow$ rejeito H_0

ou se $z_{CALC} < -z_{\alpha/2}$ ou $z_{CALC} > z_{\alpha/2} \Rightarrow$ rejeito H_0

© ALBERTO W. RAMOS 2016

B) σ desconhecido:

1º Caso: $H_0: \mu = \mu_0$

$H_1: \mu < \mu_0$

$$\bar{x}_1 = \mu_0 - t_{n-1;\alpha} \cdot \frac{s}{\sqrt{n}} \quad \text{ou} \quad t_{CALC} = \frac{\bar{x}_{CALC} - \mu_0}{\frac{s}{\sqrt{n}}}$$

$$\text{Se } \bar{x}_{CALC} < \bar{x}_1 \text{ ou } t_{CALC} < -t_{n-1;\alpha} \Rightarrow \text{rejeito } H_0$$

2º Caso: $H_0: \mu \leq \mu_0$

$H_1: \mu > \mu_0$

$$\bar{x}_2 = \mu_0 + t_{n-1;\alpha} \cdot \frac{s}{\sqrt{n}}$$

$$\text{Se } \bar{x}_{CALC} > \bar{x}_2 \text{ ou } t_{CALC} > t_{n-1;\alpha} \Rightarrow \text{rejeito } H_0$$

© ALBERTO W. RAMOS 2016

3º Caso: $H_0: \mu = \mu_0$

$H_1: \mu \neq \mu_0$

$$\bar{x}_1 = \mu_0 - t_{n-1;\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

$$\bar{x}_2 = \mu_0 + t_{n-1;\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

Se $\bar{x}_{CALC} < \bar{x}_1$ ou $\bar{x}_{CALC} > \bar{x}_2 \Rightarrow$ rejeito H_0

Ou se $t_{CALC} < -t_{n-1;\alpha/2}$ ou $t_{CALC} > t_{n-1;\alpha/2} \Rightarrow$ rejeito H_0

© ALBERTO W. RAMOS 2016

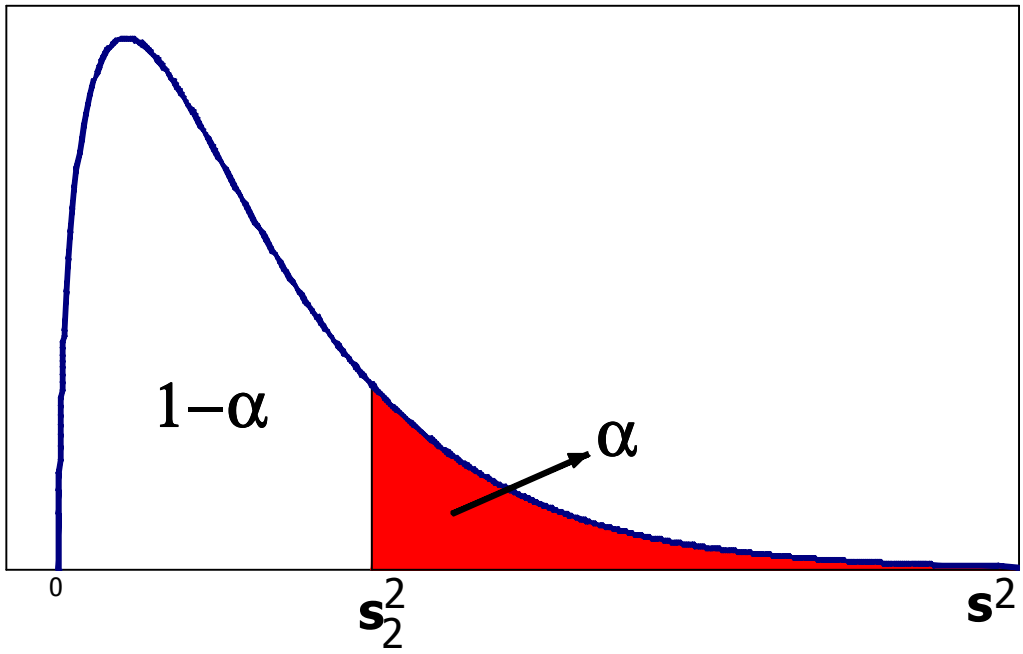
TESTES PARA VARIÂNCIA

1º Caso: H_0: $\sigma^2 = \sigma_0^2$

H_1: $\sigma^2 > \sigma_0^2$

Se H_0 for verdadeira ($\sigma^2 = \sigma_0^2$), resulta:

$$\frac{(n-1)s^2}{\sigma_0^2} = \chi^2_{n-1}$$



Se $s^2_{CALC} > s^2_2 \Rightarrow$ rejeito H_0

Se $\frac{(n-1)s^2_{CALC}}{\sigma_0^2} > \chi^2_{n-1;\alpha} \Rightarrow$ rejeito H_0

Se $\chi^2_{CALC} > \chi^2_{n-1;\alpha} \Rightarrow$ rejeito H_0

© ALBERTO W. RAMOS 2016

2º Caso: $H_0: \sigma^2 = \sigma_0^2$

$H_1: \sigma^2 < \sigma_0^2$

Se $\chi^2_{CALC} < \chi^2_{n-1;1-\alpha} \Rightarrow$ rejeito H_0

3º Caso: $H_0: \sigma^2 = \sigma_0^2$

$H_1: \sigma^2 \neq \sigma_0^2$

Se $\chi^2_{CALC} < \chi^2_{n-1;1-\alpha/2} \Rightarrow$ rejeito H_0

ou

Se $\chi^2_{CALC} > \chi^2_{n-1;\alpha/2} \Rightarrow$ rejeito H_0

© ALBERTO W. RAMOS 2016

TESTES PARA PROPORÇÃO

Sabemos que p' tem distribuição binomial. Porém, se $n.p_0 \geq 5$ e $n.(1-p_0) \geq 5 \Rightarrow$ a distribuição de p' poderá ser aproximada por uma distribuição normal*

1º Caso: $H_0: p = p_0$

$H_1: p < p_0$

$$p_1 = p_0 - z_\alpha \cdot \sqrt{\frac{p_0(1-p_0)}{n}}$$

$$\text{Se } p' < p_1 \text{ ou } z_{CALC} = \frac{p' - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} < -z_\alpha \Rightarrow \text{rejeito } H_0$$

2º Caso: $H_0: p = p_0$

$H_1: p > p_0$

$$p_2 = p_0 + z_\alpha \cdot \sqrt{\frac{p_0(1-p_0)}{n}}$$

$$\text{Se } p' > p_2 \text{ ou } z_{CALC} = \frac{p' - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} > z_\alpha \Rightarrow \text{rejeito } H_0$$

Nota: se esta aproximação não for válida, deve-se empregar o teste exato de Fischer, cuja teoria foge ao escopo deste curso.

© ALBERTO W. RAMOS 2016

3º Caso: $H_0: p = p_0$

$H_1: p \neq p_0$

$$p_1 = p_0 - z_{\alpha/2} \cdot \sqrt{\frac{p_0(1-p_0)}{n}}$$

$$p_2 = p_0 + z_{\alpha/2} \cdot \sqrt{\frac{p_0(1-p_0)}{n}}$$

$$\text{Se } p' < p_1 \text{ ou se } p' > p_2 \Rightarrow \text{rejeito } H_0$$

ou ainda, se $z_{CALC} < -z_{\alpha/2}$ ou se $z_{CALC} > z_{\alpha/2} \Rightarrow$ rejeito H_0

© ALBERTO W. RAMOS 2016

TAMANHO DE AMOSTRA PARA TESTES DE HIPÓTESES

Sejam as hipóteses:

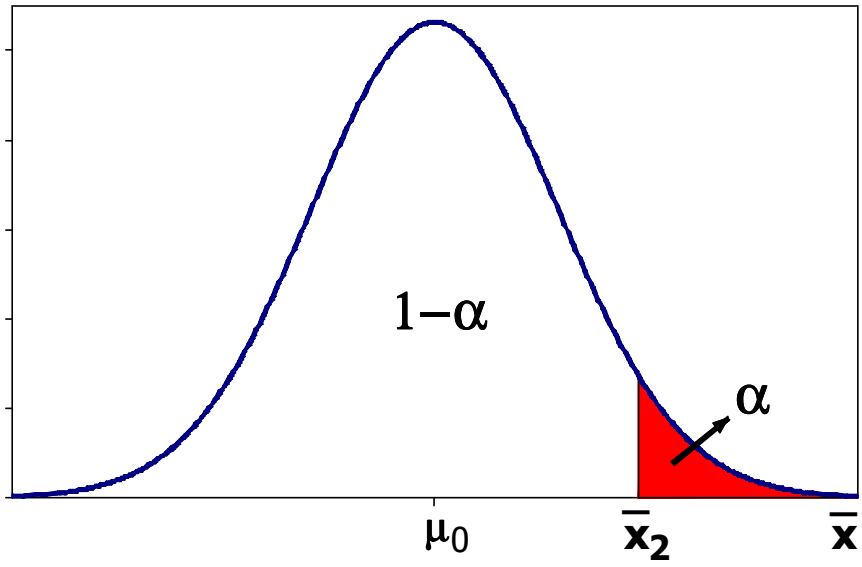
$$H_0: \mu \leq \mu_0$$

$$H_1: \mu > \mu_0$$

e vamos assumir que:

- σ é conhecido e;
- α e β estâo fixados (determinados).

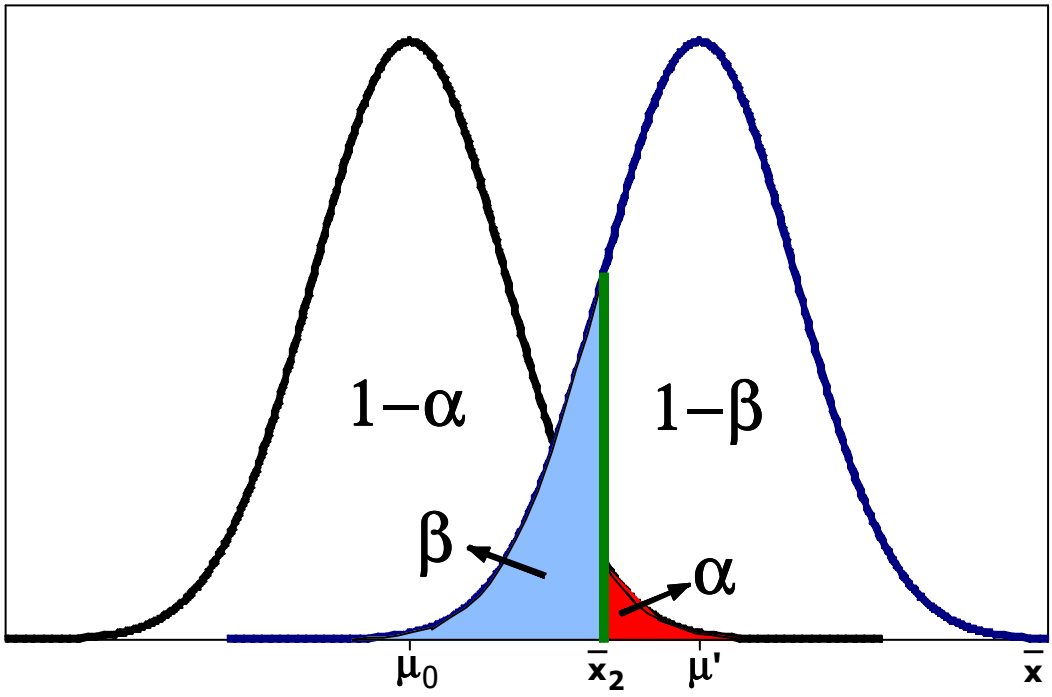
Se H_0 é verdadeira, ou seja, se $\mu = \mu_0$



$$\bar{x}_2 = \mu_0 + z_\alpha \cdot \frac{\sigma}{\sqrt{n}}$$

© ALBERTO W. RAMOS 2016

Mas, se em realidade $\mu = \mu' > \mu_0$, então



E, consequentemente

$$\bar{x}_2 = \mu' - z_\beta \cdot \frac{\sigma}{\sqrt{n}}$$

igualando-se ambas expressões, resulta em

$$n = \left(\frac{z_\alpha + z_\beta}{\mu' - \mu_0} \sigma \right)^2$$

ou, se σ é desconhecido

$$n = \left(\frac{t_{n'-1;\alpha} + t_{n'-1;\beta}}{\mu' - \mu_0} s \right)^2$$

© ALBERTO W. RAMOS 2016

Se, alternativamente, as hipóteses testadas fossem:

$$H_0: \mu = \mu_0$$

$$H_1: \mu \neq \mu_0$$

Então:

$$n = \left(\frac{z_{\alpha/2} + z_{\beta}}{\mu' - \mu_0} \sigma \right)^2$$

ou, se s fôr desconhecido

$$n = \left(\frac{t_{n'-1;\alpha/2} + t_{n'-1;\beta}}{\mu' - \mu_0} s \right)^2$$

Analogamente, para o caso de testes unilaterais da proporção, considerada válida a aproximação da distribuição da binomial pela normal:

$$n = \left(\frac{z_{\alpha} \sqrt{p_0(1-p_0)} + z_{\beta} \sqrt{p'(1-p')}}{p' - p_0} \right)^2$$

Como ficaría a expressão acima para teste bilateral da proporção?

© ALBERTO W. RAMOS 2016

Comparações Múltiplas

© ALBERTO W. RAMOS 2016

COMPARAÇÃO DE VÁRIAS VARIÂNCIAS

Sejam várias amostras, de mesmo tamanho (**n**), retiradas de **k** populações Normais.

Se quisermos testar as hipóteses:

H_0: $\sigma_1^2 = \sigma_2^2 = ... = \sigma_k^2$

H_1: pelo menos um σ_i^2 diferente dos demais

obtemos:

$$g_{CALC} = \frac{\max s_i^2}{\sum s_i^2} \quad (i = 1, 2,..., k)$$

e também:

$$g_{CRIT} \rightarrow \text{função de n, k e } \alpha$$

Se $g_{CALC} > g_{CRIT} \Rightarrow$ rejeito H_0 e afirmo H_1

© ALBERTO W. RAMOS 2016

TABELA g

(α = 5%)

	n = 2	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
k = 2	0,9985	0,9750	0,9392	0,9057	0,8772	0,8534	0,8332	0,8159	0,8010
3	0,9669	0,8709	0,7977	0,7457	0,7071	0,6771	0,6530	0,6333	0,6167
4	0,9065	0,7679	0,6841	0,6287	0,5895	0,5598	0,5365	0,5157	0,5017
5	0,8412	0,6838	0,5931	0,5441	0,5065	0,4783	0,4564	0,4387	0,4241
6	0,7808	0,6161	0,5321	0,4803	0,4447	0,4184	0,3980	0,3817	0,3682
7	0,7271	0,5612	0,4800	0,4307	0,3974	0,3726	0,3535	0,3384	0,3259
8	0,6798	0,5157	0,4377	0,3910	0,3595	0,3362	0,3185	0,3043	0,2926
9	0,6385	0,4775	0,4027	0,3584	0,3286	0,3067	0,2901	0,2768	0,2659
10	0,6020	0,4450	0,3733	0,3311	0,3029	0,2823	0,2666	0,2541	0,2439

Observações:

k = quantidade de amostras

n = tamanho da amostra

© ALBERTO W. RAMOS 2016

EXEMPLO

Cinco amostras com quatro elementos cada forneceram s_i^2: 3,7 - 2,5 - 5,1 - 6,0 - 3,2. Ao nível de significância de 5%, existe evidência que alguma σ_i^2 seja diferente das demais?

$$n = 4 \quad e \quad k = 5$$

$$\max s_i^2 = 6,0$$

$$\sum s_i^2 = 20,5$$

Com isso, temos:

$$g_{CALC} = \frac{6,0}{20,5} = 0,2927$$

$$g_{CRIT} = g_{5;4;5\%} = 0,5931$$

$\Rightarrow$ aceito que as variâncias são iguais, ou seja:

$$\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = \sigma_5^2$$

© ALBERTO W. RAMOS 2016

COMPARAÇÃO DE VÁRIAS MÉDIAS

Vamos imaginar que temos três grupos de pessoas e que se quer verificar se o seu QI (quociente de inteligência) médio é igual. Para tanto, sorteiou-se oito indivíduos de cada grupo e a estes foi aplicado um certo teste.

Os resultados obtidos serão apontados numa tabela da seguinte forma:

Grupo	Notas	$\bar{x}$	s^2
1	x_{11} x_{12} x_{13} .... x_{18}	$\bar{x}_1$	s_1^2
2	x_{21} x_{22} x_{23} .... x_{28}	$\bar{x}_2$	s_2^2
3	x_{31} x_{32} x_{33} .... x_{38}	$\bar{x}_3$	s_3^2

Notação empregada:

n - tamanho da amostra (8, no caso)
k - quantidade de médias comparadas (3, no caso)
$\bar{x}_i$ - média da amostra do grupo i
$\bar{\bar{x}}$ - média geral (média das médias)
s_i^2 - variância da amostra do grupo i
s_d^2 - variância dentro da amostra (ou residual)
s_e^2 - variância entre amostras
s_t^2 - variância total

© ALBERTO W. RAMOS 2016

Como não se conhece a variância da população, chamada de σ^2, pode-se estimá-la mediante três métodos diferentes:

<u>Método 1</u>: através dos s^2 obtidos em cada grupo

$$\bar{s}^2 = s_d^2 = \frac{\sum s_i^2}{k} = \frac{\sum\sum (x_{ij} - \bar{x}_i)^2}{k.(n-1)}$$

<u>Método 2</u>: através das médias dos grupos

$$s_e^2 = n.\frac{\sum (\bar{x}_i - \bar{\bar{x}})^2}{k-1}$$

<u>Método 3</u>: através de todos os dados individuais

$$s_t^2 = \frac{\sum\sum (x_{ij} - \bar{\bar{x}})^2}{n.k-1}$$

Como toda esta notação é muito complicada, vamos mostrar os conceitos mediante aplicação ao exemplo do QI.

© ALBERTO W. RAMOS 2016

Imagine que os resultados obtidos tenham sido os seguintes:

Grupo	Notas	$\bar{x}$	s^2
1	4 5 5 4 8 4 3 7	5,0	2,9
2	2 4 3 7 5 4 2 5	4,0	2,9
3	3 6 6 4 5 4 6 6	5,0	1,4

Método 1: através dos s^2 obtidos em cada grupo

$$s_d^2 = \frac{2,9 + 2,9 + 1,4}{3} = 2,4$$

Método 2: através das médias dos grupos

$$\bar{\bar{x}} = \frac{5,0 + 4,0 + 5,0}{3} = 4,7$$

$$s_e^2 = 8.\frac{[(5,0 - 4,7)^2 + (4,0 - 4,7)^2 + (5,0 - 4,7)^2]}{(3 - 1)} = 2,7$$

Método 3: através de todos os dados individuais

$$s_t^2 = \frac{[(4 - 4,7)^2 + (5 - 4,7)^2 + (5 - 4,7)^2 + ... + (6 - 4,7)^2]}{8.3 - 1} = 2,4$$

Pode-se perceber que:

- as médias $\bar{x}$ são próximas;
- os valores de s_d^2, s_e^2 e s_t^2 também são próximos.

© ALBERTO W. RAMOS 2016

Imagine, agora, que os resultados obtidos fossem:

Grupo	Notas	$\bar{x}$	s^2
1	4 5 5 4 8 4 3 7	5,0	2,9
2	0 2 1 5 3 2 0 3	2,0	2,9
3	7 10 10 8 9 8 10 10	9,0	1,4

Método 1: através dos s^2 obtidos em cada grupo

$$s_d^2 = \frac{2,9 + 2,9 + 1,4}{3} = 2,4$$

Método 2: através das médias dos grupos

$$\bar{\bar{x}} = \frac{5,0 + 2,0 + 9,0}{3} = 5,3$$

$$s_e^2 = 8.\frac{[(5,0 - 5,3)^2 + (2,0 - 5,3)^2 + (9,0 - 5,3)^2]}{(3 - 1)} = 98,7$$

Método 3: através de todos os dados individuais

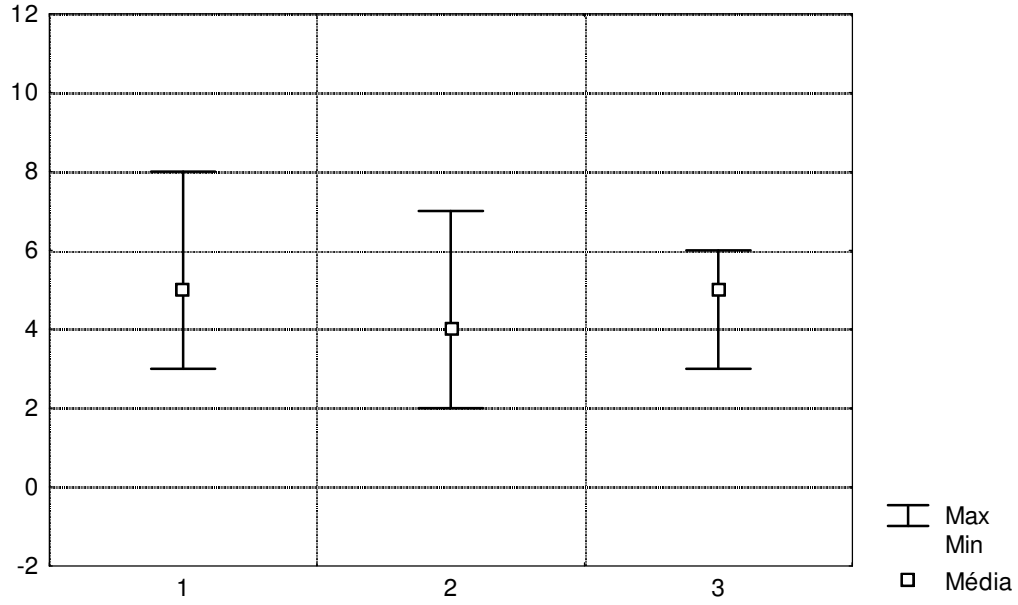
$$s_t^2 = \frac{[(4 - 5,3)^2 + (5 - 5,3)^2 + (5 - 5,3)^2 + ... + (6 - 5,3)^2]}{8.3 - 1} = 10,8$$

Pode-se perceber, neste novo conjunto de resultados, que:

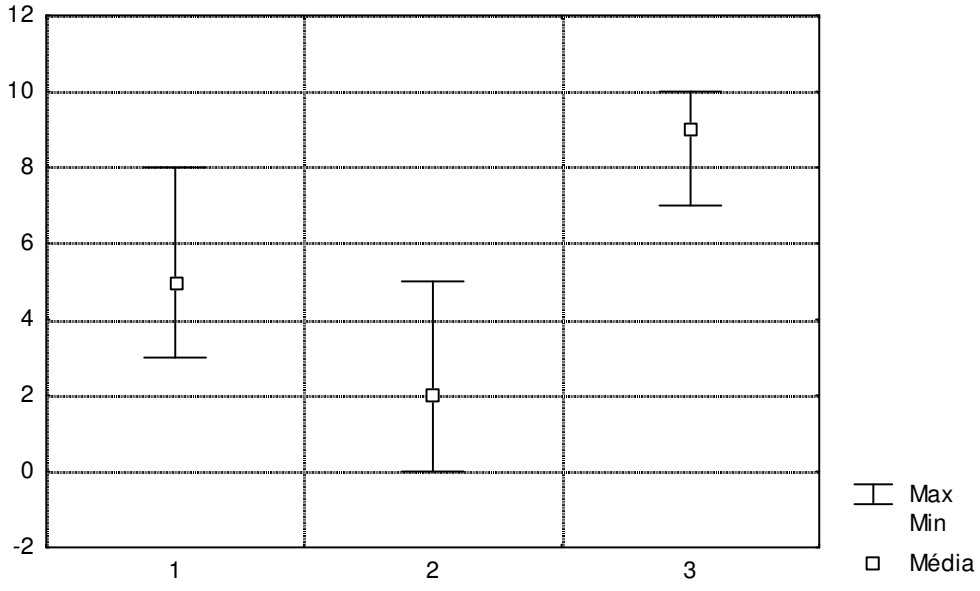
- as médias $\bar{x}$ não mais são próximas;
- o valor de s_d^2 não se alterou;
- os valores de s_e^2 e s_t^2 aumentaram muito.

© ALBERTO W. RAMOS 2016

Os gráficos abaixo ajudam na interpretação dos resultados. No primeiro conjunto de dados as médias estavam próximas:



Já no outro conjunto, as médias apresentavam-se mais afastadas umas em relação às outras:



© ALBERTO W. RAMOS 2016

COMENTÁRIOS

1. Se as médias das populações são iguais, os valores de $\bar{x}_i$ serão próximos e tanto faz estimar-se σ^2 através de s_d^2, s_e^2 ou s_t^2, pois todos eles fornecerão valores próximos.

2. Mas, quando as médias das populações são diferen-tes, os valores de $\bar{x}_i$ divergirão entre si. Embora s_d^2 continue sendo um bom estimador de σ^2, s_e^2 e s_t^2 não mais o serão, pois são afetados pela diferença entre as médias.

3. Assim, pode-se comparar as médias das diversas popula-ções (k) através da comparação de variâncias: s_e^2 e s_d^2, respectivamente. Este teste é chamado de teste F, onde:

$$F_{calc} = \frac{s_e^2}{s_d^2}$$

4. s_e^2 tem (k-1) graus de liberdade, s_d^2 tem [k.(n-1)] graus de liberdade. Portanto, F_{calc} terá (k-1) no seu numerador; [k.(n-1)] graus de liberdade no seu denominador.

5. Quanto maior o valor de F_{calc} maior é a probabilidade de que as médias sejam diferentes entre si. Para chegar a uma conclusão, F_{calc} é comparado contra um F_{crit}, obtido a partir de uma tabela.

6. Se $F_{calc} < F_{crit}$, então admite-se que as médias são iguais.

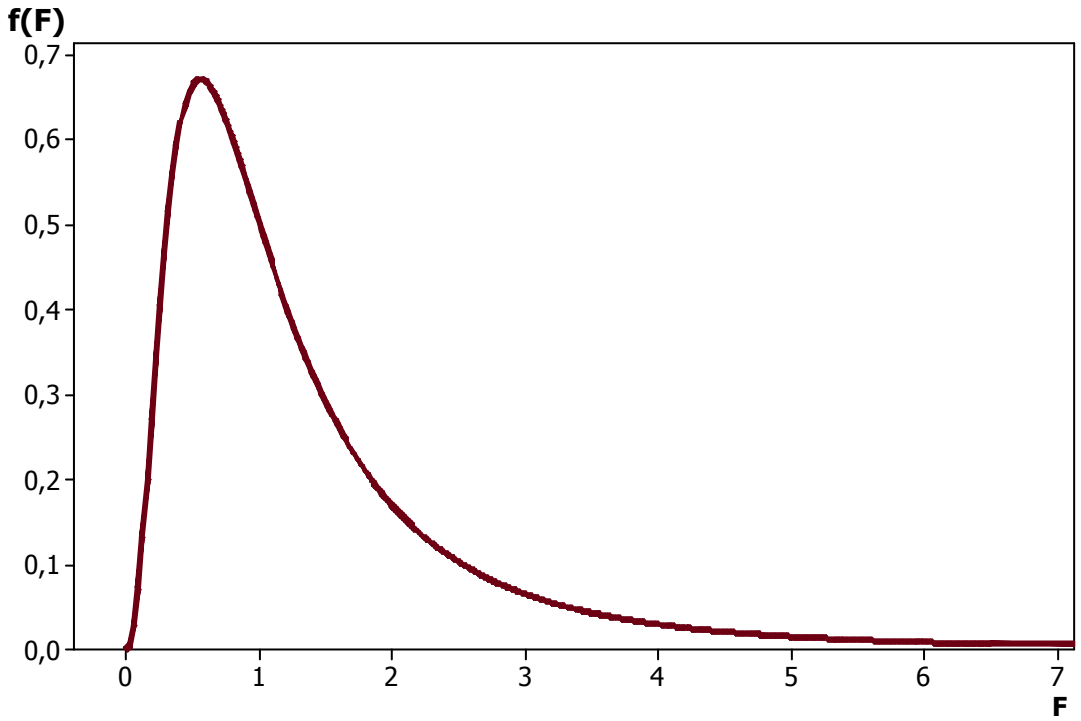
7. A análise de variância assume a hipótese de que as populações possuem a mesma variância (σ^2). Se isto não ocorrer, os resultados não serão válidos.

© ALBERTO W. RAMOS 2016

DISTRIBUIÇÃO F-SNEDECOR

Sejam duas amostras independentes, retiradas de populações Normais, com mesma variância (σ^2), que forneceram estimativas s_1^2 e s_2^2, respectivamente. Ao quociente de s_1^2 por s_2^2, chamamos de:

$$F_{n_1-1;n_2-1} = \frac{s_1^2}{s_2^2}$$



© ALBERTO W. RAMOS 2016

TABELA F-SNEDECOR

(α = 5%)

	ν_1									
ν_2	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	161,4	199,5	215,7	224,6	230,2	234,0	236,8	238,9	240,5	241,9
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32

*FONTE***:** COSTA NETO, P.L.O. ***Estatística***. São Paulo, Edgard Blucher, 1978.

© ALBERTO W. RAMOS 2016

EXEMPLO

O segundo conjunto de resultados do teste de QI, forneceu:

$$s_e^2 = 98,7 \qquad e \qquad s_d^2 = 2,4$$

logo

$$F_{calc} = \frac{98,7}{2,4} = 41,1$$

F_{calc} tem (3 -1) = 2 GL no numerador e [3 x (8-1)] = 21 GL no denominador.

F_{crit} (para um α=5%) será $F_{2,\ 21,\ 5\%}$ = 3,47 => pelo menos uma turma é diferente das demais.

© ALBERTO W. RAMOS 2016

TABELA DA ANÁLISE DE VARIÂNCIA

É comum apresentar-se os resultados da análise de variância na forma de uma tabela, similar à de baixo:

Fonte	***SQ***	***GL***	***QM***	F_{CALC}
Entre	$n.\sum(\bar{x}_i - \bar{\bar{x}})^2$	(k-1)	s_e^2	s_e^2 / s_d^2
Dentro	$\sum\sum(x_{ij} - \bar{x}_i)^2$	k.(n-1)	s_d^2	
Total	$\sum\sum(x_{ij} - \bar{\bar{x}})^2$	k.n-1	s_t^2	

onde:

SQ - é a soma de quadrados
GL - são os graus de liberdade das estimativas
QM - é o quadrado médio = SQ/GL

No caso de nosso exemplo do teste de QI, com o segundo conjunto de dados, tem-se:

Fonte	***SQ***	***GL***	***QM***	F_{CALC}
Entre	197,3	2	98,7	41,4
Dentro	50,0	21	2,4	
Total	247,3	23	10,8	

© ALBERTO W. RAMOS 2016

AMOSTRAS DE TAMANHOS DIFERENTES

Há situações onde, eventualmente, pode-se estar trabalhando com amostras de tamanho diferente. Neste caso, a tabela da Análise de Variância é modificada da seguinte forma:

Fonte	***SQ***	***GL***	***QM***	F_{CALC}
Entre	SQE=SQT-SQD	(k-1)	$s_E^2 = \frac{SQE}{k-1}$	s_E^2/s_R^2
Residual	$SQR = \sum (n_i - 1)s_i^2$	$\sum n_i - k$	$s_R^2 = \frac{SQR}{\sum n_i - k}$	
Total	$SQT = \sum\sum (x_{ij} - \bar{\bar{x}})^2$	$\sum n_i - 1$	s_T^2	

Obs: neste caso

$$s_R^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + ... + (n_k - 1)s_k^2}{n_1 + n_2 + ... + n_k - k}$$

© ALBERTO W. RAMOS 2016

ALGUNS CUIDADOS

A análise de variância tem algumas hipóteses básicas que são assumidas para sua validade:

- o modelo válido é do tipo $x_{ij} = \mu + \alpha_i + \varepsilon_{ij}$, onde μ é a média geral, α_i é o efeito do nível i do fator e ε_{ij} é o erro;
- as populações são homocedásticas, ou seja, possuem a mesma variância em comum σ^2;
- as populações podem ser adequadamente representadas por uma distribuição de probabilidade normal;
- conseqüentemente, $\varepsilon_{ij} \sim N(0; \sigma^2)$.

1. A condição de homocedasticidade é fundamental para que os resultados sejam válidos e pode ser verificada mediante uma análise de resíduos ou, então, pelo teste de Cochran ou de Bartlett.

2. A condição de normalidade dos dados não é essencial, pois a análise de variância fornece bons resultados mesmo quando a população não é normal. Ela pode ser verificada através do papel de probabilidade normal.

© ALBERTO W. RAMOS 2016

Correlação e Regressão

© ALBERTO W. RAMOS 2016

CORRELAÇÃO

Quando duas (ou mais) variáveis apresentam tendência de variação conjunta, diz-se que estas se correlacionam

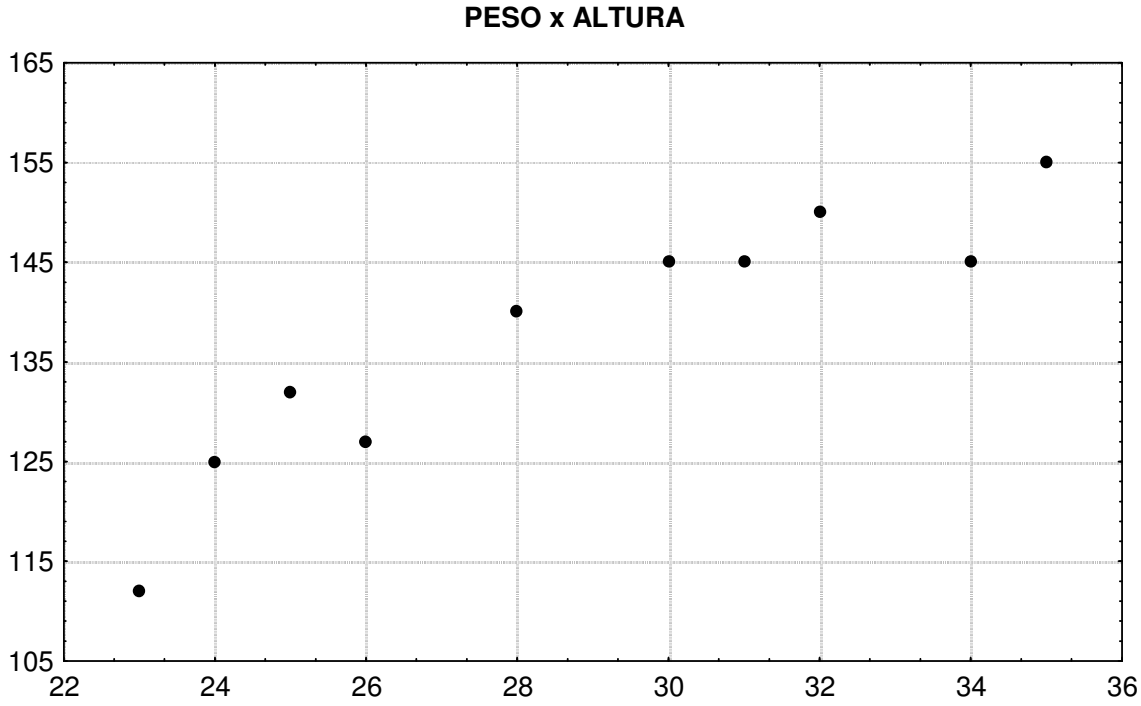
EXEMPLO

Criança	***Peso***	***Altura***
1	30	145
2	32	150
3	24	125
4	28	140
5	26	127
6	34	145
7	25	132
8	23	112
9	35	155
10	31	145

Existe correlação entre peso e altura ?

© ALBERTO W. RAMOS 2016

O diagrama de dispersão é uma maneira rápida e eficiente de verificar a existência de correlação

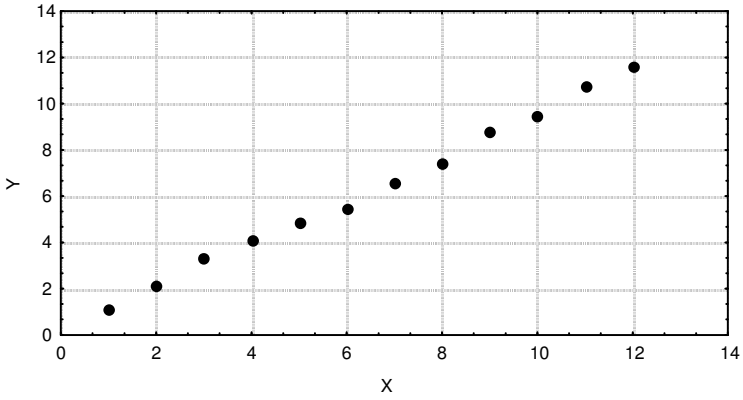


Qual é a conclusão ?

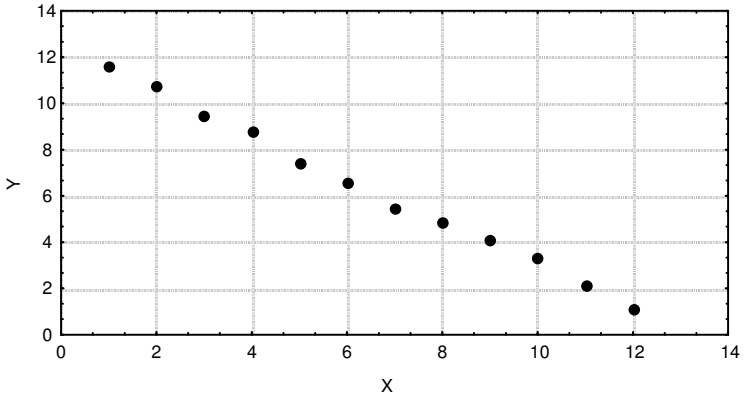
© ALBERTO W. RAMOS 2016

TIPOS DE CORRELAÇÃO

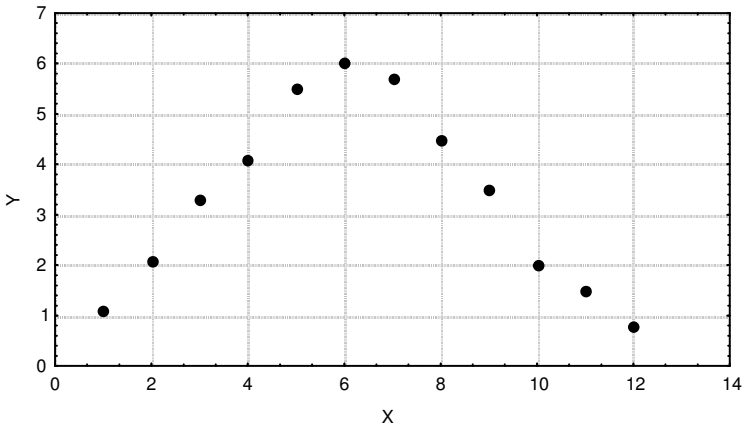
CORRELAÇÃO LINEAR POSITIVA
(QUANDO X AUMENTA => Y TAMBÉM AUMENTA)



CORRELAÇÃO LINEAR NEGATIVA
(QUANDO X AUMENTA => Y DIMINUI)



CORRELAÇÃO NÃO LINEAR



© ALBERTO W. RAMOS 2016

COEFICIENTE DE CORRELAÇÃO LINEAR

Para medir o grau de correlação entre duas variáveis, pode-se utilizar o coeficiente de correlação linear:

$$r = \frac{\Sigma(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\Sigma(x_i - \overline{x})^2 \Sigma(y_i - \overline{y})^2}} = \frac{S_{XY}}{\sqrt{S_{XX}.S_{YY}}}$$

onde:

$$S_{XY} = \sum x_i.y_i - \frac{(\sum x_i.\sum y_i)}{n}$$

$$S_{XX} = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$$

$$S_{YY} = \sum y_i^2 - \frac{(\sum y_i)^2}{n}$$

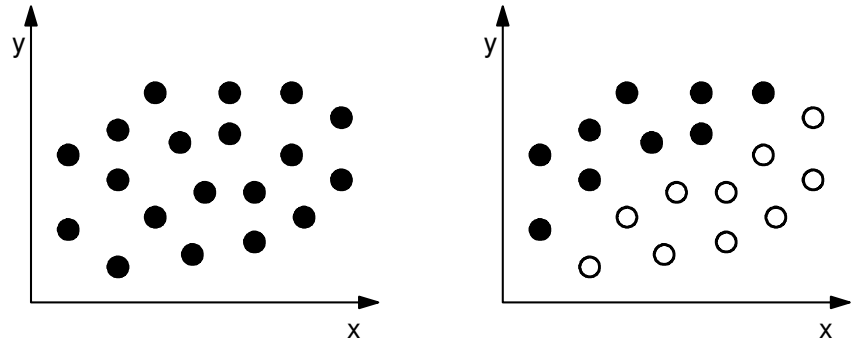
A sua interpretação é a seguinte:

- o valor de r varia de -1 a +1 (inclusive)
- r > 0 indica correlação linear positiva
- r < 0 indica correlação linear negativa
- r = -1 indica correlação linear negativa perfeita
- r = +1 indica correlação linear positiva perfeita

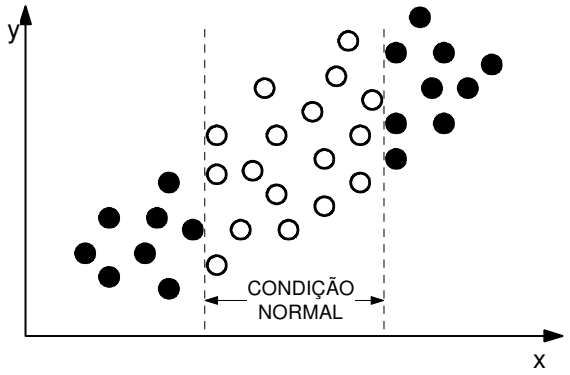
© ALBERTO W. RAMOS 2016

CUIDADOS NA CORRELAÇÃO

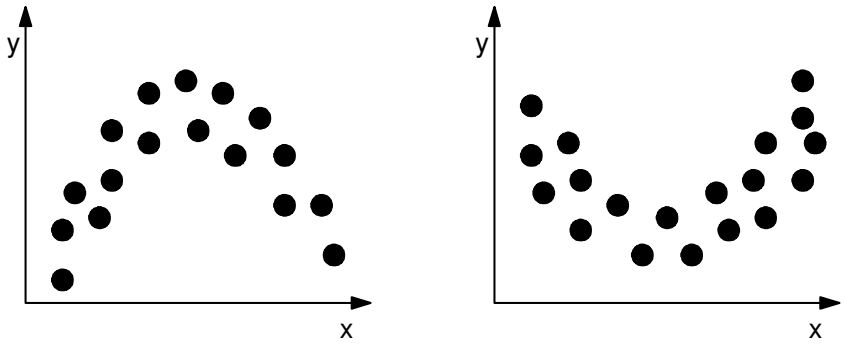
a) Estratificação de Dados



b) Amplitude do Estudo de Correlação



c) Picos e Vales



© ALBERTO W. RAMOS 2016

TESTE DA CORRELAÇÃO

Para determinar se a correlação é estatisticamente significativa ou não, faz-se um teste de hipóteses quanto ao coeficiente de correlação populacional (ρ). Assim:

$H_0: \rho = 0$ (não há correlação)

$H_1: \rho \neq 0$ (há correlação)

Estas hipóteses serão testadas mediante o cálculo de um t de Student, definido como:

$$t_{calc} = t_{n-2} = r\sqrt{\frac{n-2}{1-r^2}}$$

que será comparado contra $t_{crit} = t_{n-2;\alpha/2}$.

Se $|t_{calc}| > t_{crit}$ ➔ H_0 será rejeitada e pode-se afirmar que a correlação existe.

© ALBERTO W. RAMOS 2016

REGRESSÃO

O objetivo fundamental da regressão é descobrir a equação que relaciona duas (ou mais) variáveis, ou seja:

$$y = f(x_1, x_2, \ldots, x_k) + \varepsilon$$

onde:

$x_1, x_2, \ldots, x_k$ são chamadas de fatores;

$f(x_1, x_2, \ldots, x_k)$ indica uma função de várias variáveis;

ε é chamado de erro.

EXEMPLOS

Resposta	***Variáveis***
Pressão	Volume e Temperatura
Rendimento	Temperatura, Tempo e Quantidade de Catalisador
Viscosidade	Temperatura, Pressão, Velocidade e Vazão

© ALBERTO W. RAMOS 2016

REGRESSÃO LINEAR SIMPLES

Admite que uma equação do primeiro grau representa satisfatoriamente o modelo:

$$y = \beta_0 + \beta_1.x$$

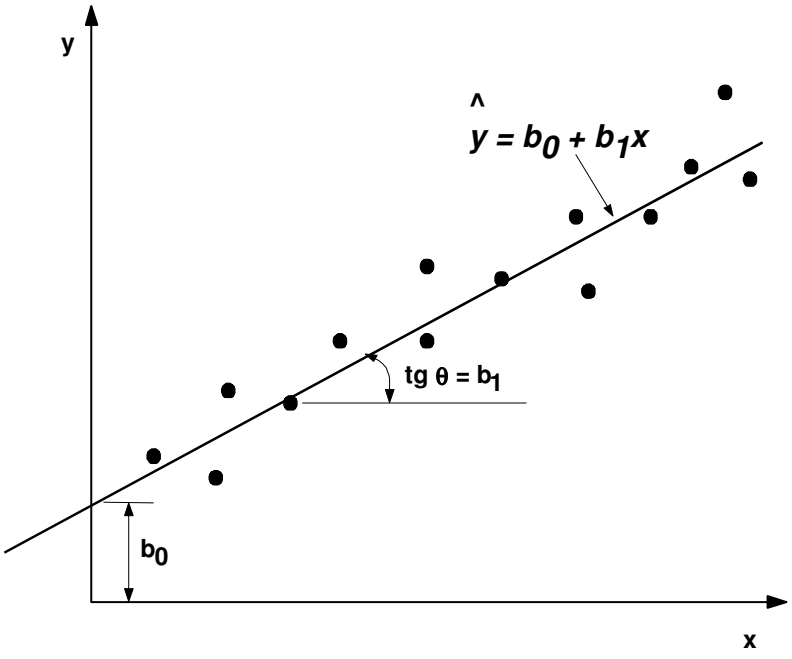
como as constantes β_0 e β_1 são desconhecidas, então a equação da reta será estimada através de:

$$\hat{y} = b_0 + b_1.x$$

onde:

b_0 - é o intercepto da reta

b_1 - é a coeficiente angular da reta



© ALBERTO W. RAMOS 2016

EXEMPLO

Foi feito um levantamento de diversos modelos de automóveis quanto a potência do motor (Hp) e o consumo médio (km/l).

Carro	***Potência***	***Consumo***
1	130	10,1
2	81	10,5
3	93	11,3
4	113	10,5
5	90	11,6
6	63	12,4
7	55	15,0
8	102	11,3
9	92	12,4
10	81	12,0
11	103	10,9
12	90	11,6
13	74	12,4
14	73	13,1
15	102	10,9
16	78	12,0
17	100	10,5
18	100	10,5

© ALBERTO W. RAMOS 2016

DETERMINAÇÃO DA EQUAÇÃO DA RETA

A equação da reta é determinada a partir dos dados da tabela anterior, através do método dos mínimos quadrados, utilizando-se as seguintes fórmulas:

$$b_1 = \frac{S_{XY}}{S_{XX}}$$

onde:

$$S_{XY} = \sum x_i y_i - \frac{\left(\sum x_i . \sum y_i\right)}{n}$$

e

$$S_{XX} = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}$$

O outro termos é obtido através de:

$$b_0 = \bar{y} - \bar{b_1}.x$$

© ALBERTO W. RAMOS 2016

Para facilitar os cálculos, usa-se uma tabela auxiliar:

Carro	Potência (x_i)	Consumo (y_i)	x_i^2	y_i^2	$x_i y_i$
1	130	10,1	16900	102,01	1313,0
2	81	10,5	6561	110,25	850,5
3	93	11,3	8649	106,09	957,9
4	113	10,5	12769	110,25	1186,5
5	90	11,6	8100	134,56	1044,0
6	63	12,4	3969	153,76	781,2
7	55	15,0	3025	225,00	825,0
8	102	11,3	10404	127,69	1152,6
9	92	12,4	8464	153,76	1140,8
10	81	12,0	6561	144,00	972,0
11	103	10,9	10609	118,81	1122,7
12	90	11,6	8100	134,56	1044,0
13	74	12,4	5476	153,76	917,6
14	73	13,1			
15	102	10,9			
16	78	12,0			
17	100	10,5			
18	100	10,5			
TOTAL	1620	208,0	151404	2429,42	18411,9

S_{XY} =

S_{XX} =

b_1 =

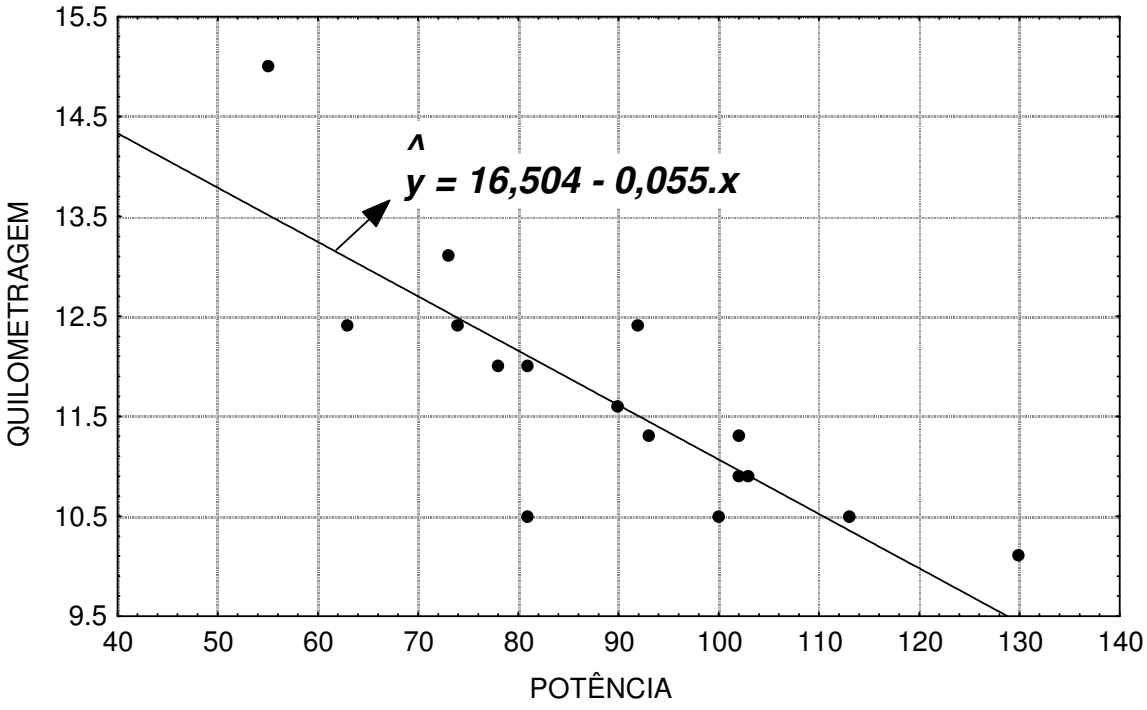
b_0 =

© ALBERTO W. RAMOS 2016

Assim sendo, resulta que a reta (de regressão) é:

$$\hat{y} = 16{,}504 - 0{,}055.x$$

Abaixo é mostrado um gráfico (diagrama de dispersão) com os pontos e a reta marcados.



Qual a melhor estimativa do consumo médio do carro quando seu motor tem 90 Hp de potência?

© ALBERTO W. RAMOS 2016

COMENTÁRIOS:

- o método dos mínimos quadrados busca traçar a melhor reta através dos pontos, ou seja, aquela que torna mínima a distância destes à reta;

- sempre é possível obter a equação de uma reta que passa por um conjunto de pontos, mas isto não significa que o modelo seja necessariamente adequado;

- para se verificar a adequação do modelo, emprega-se a análise de variância (ANOVA).

- é recomendável também fazer uma análise de resíduos para completar a análise de adequação do modelo.

© ALBERTO W. RAMOS 2016

ANÁLISE DE VARIÂNCIA APLICADA À REGRESSÃO

Para verificar se a regressão linear é estatisticamente significativa, deve-se testar o seguinte conjunto de hipóteses:

$H_0: \beta_1 = 0$ (não há regressão)

$H_1: \beta_1 \neq 0$ (há regressão)

Este teste pode ser feito mediante a aplicação do método da análise de variância. Pode-se identificar dois tipos de variância diferentes: a total e a residual.

A variância total é estimada através de:

$$s_T^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1} = \frac{S_{YY}}{n-1}$$

A variância residual (ou em torno da reta de regressão) é estimada através de:

$$s_R^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{n-2} = \frac{S_{YY} - b_1^2 S_{XX}}{n-2}$$

© ALBERTO W. RAMOS 2016

À diferença entre estas duas variâncias chama-se de variância devido ao modelo de regressão, que é estimada através de:

$$s_M^2 = \frac{b_1{}^2 S_{XX}}{1}$$

Se a regressão for significativa, então a variância residual (ou devida ao erro) deve ser pequena quando comparada com a variância devida a regressão. Consequentemente, o quociente das duas variâncias (regressão/erro) pode ser testado mediante um F-Snedecor. Em termos de tabela, este teste fica:

Fonte	*GL*	*SQ*	*QM*	*Fcalc*
Regressão	1	$b_1{}^2 S_{XX}$	$b_1{}^2 S_{XX}$	$\frac{b_1{}^2 S_{XX}}{s_R^2}$
Erro	n-2	$S_{YY} - b_1{}^2 S_{XX}$	$s_R^2 = \frac{S_{YY} - b_1{}^2 . S_{XX}}{n-2}$	
Total	n-1	S_{YY}		

F_{calc} será comparado contra $F_{crit} = F_{1;\, n-2;\, \alpha}$ e se $|F_{calc}| > F_{crit}$ ➔ rejeita-se H_0

© ALBERTO W. RAMOS 2016

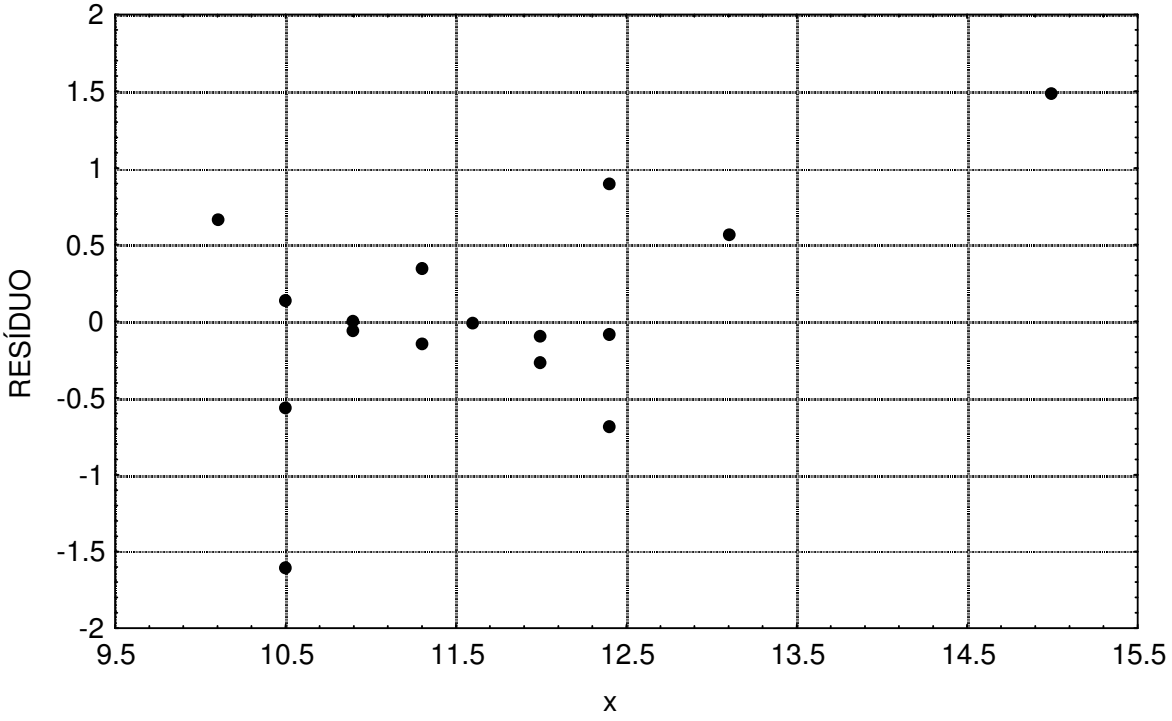
ANÁLISE DE RESÍDUOS

O resíduo (ou erro) é definido como sendo a diferença entre o valor observado (y) e o valor previsto pela equação obtida (y-chapéu). Assim, no exemplo:

Amostra	x	y	$\hat{y}$	$e = y - \hat{y}$
1	130	10,1	9,4	0,7
2	81	10,5	12,0	-1,5
3	93	11,3	11,4	-1,1
4	113	10,5	10,3	0,2
5	90	11,6	11,6	0
6	63	12,4	13,0	-0,6
7	55	15,0	13,5	1,5
8	102	11,3	10,9	0,4
9	92	12,4	11,4	1,0
10	81	12,0	12,0	0
11	103	10,9	10,8	0,1
12	90	11,6	11,6	0
13	74	12,4	12,4	0
14	73	13,1	12,5	0,6
15	102	10,9	10,9	0
16	78	12,0	12,2	-0,2
17	100	10,5	11,0	-0,5
18	100	10,5	11,0	-0,5

© ALBERTO W. RAMOS 2016

Se o modelo (linear) ajustado aos dados for adequado, então os resíduos devem se apresentar distribuídos aleatoriamente em torno do valor zero, quando marcados num gráfico cartesiano como o abaixo.



Padrões estranhos observados na forma em que os resíduos se distribuem neste gráfico podem indicar problemas.

© ALBERTO W. RAMOS 2016

Papel de Probabilidade Normal

© ALBERTO W. RAMOS 2016

PAPEL DE PROBABILIDADE NORMAL (PPN)

O PPN tem por objetivo verificar graficamente se dados experimentais aderem à distribuição normal.

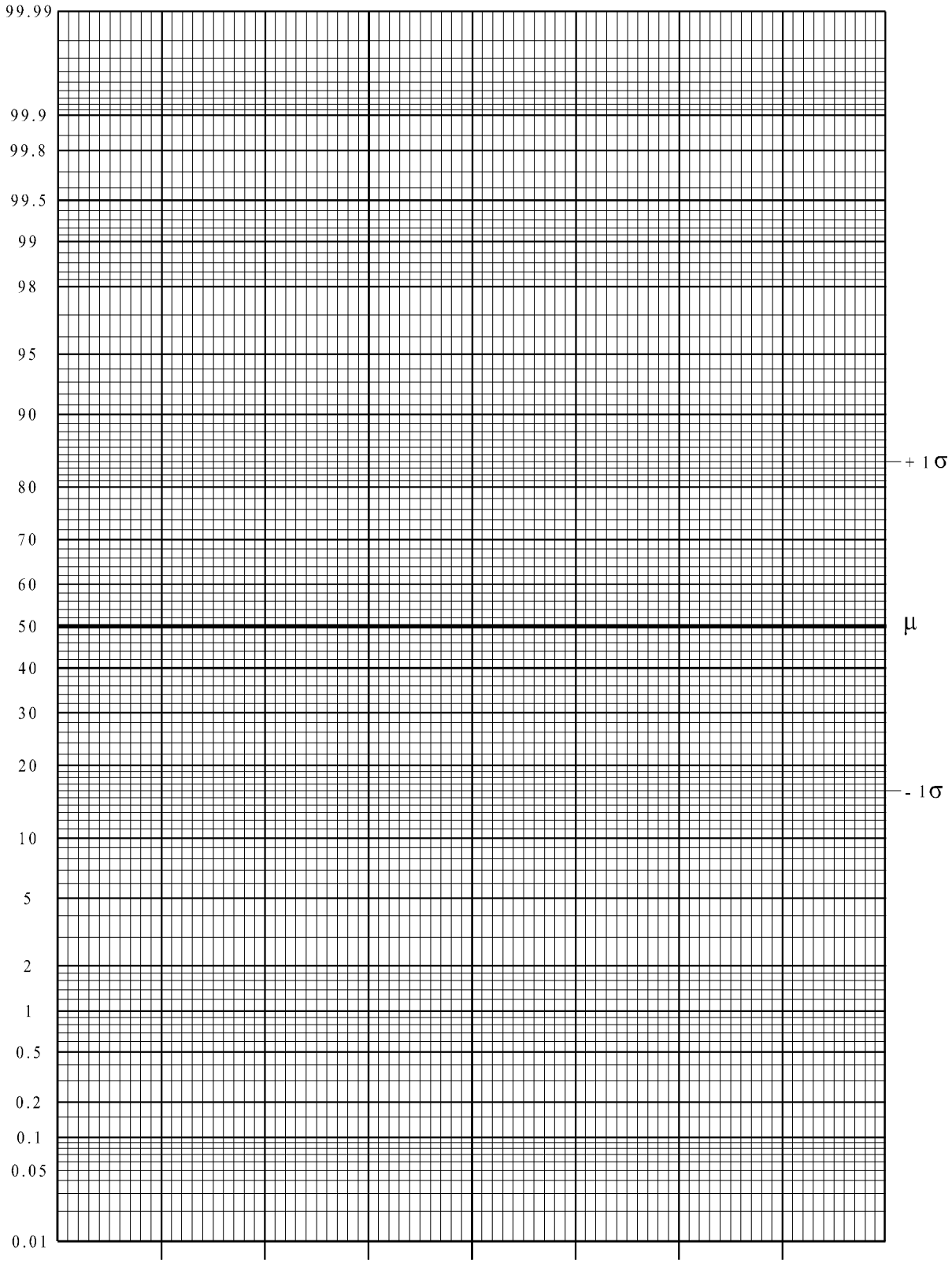
1º Caso: muitos dados (n> 30)

EXEMPLO

Duração	*Quantidade*	*% acumulada*
$50 \leq x < 55$	5	5,0
$55 \leq x < 60$	23	28,0
$60 \leq x < 65$	36	64,0
$65 \leq x < 70$	27	91,0
$70 \leq x < 75$	8	99,0
$75 \leq x < 80$	1	100,0

© ALBERTO W. RAMOS 2016

Papel de Probabilidade Normal



© ALBERTO W. RAMOS 2016

2º Caso: poucos dados (n<30)

EXEMPLO

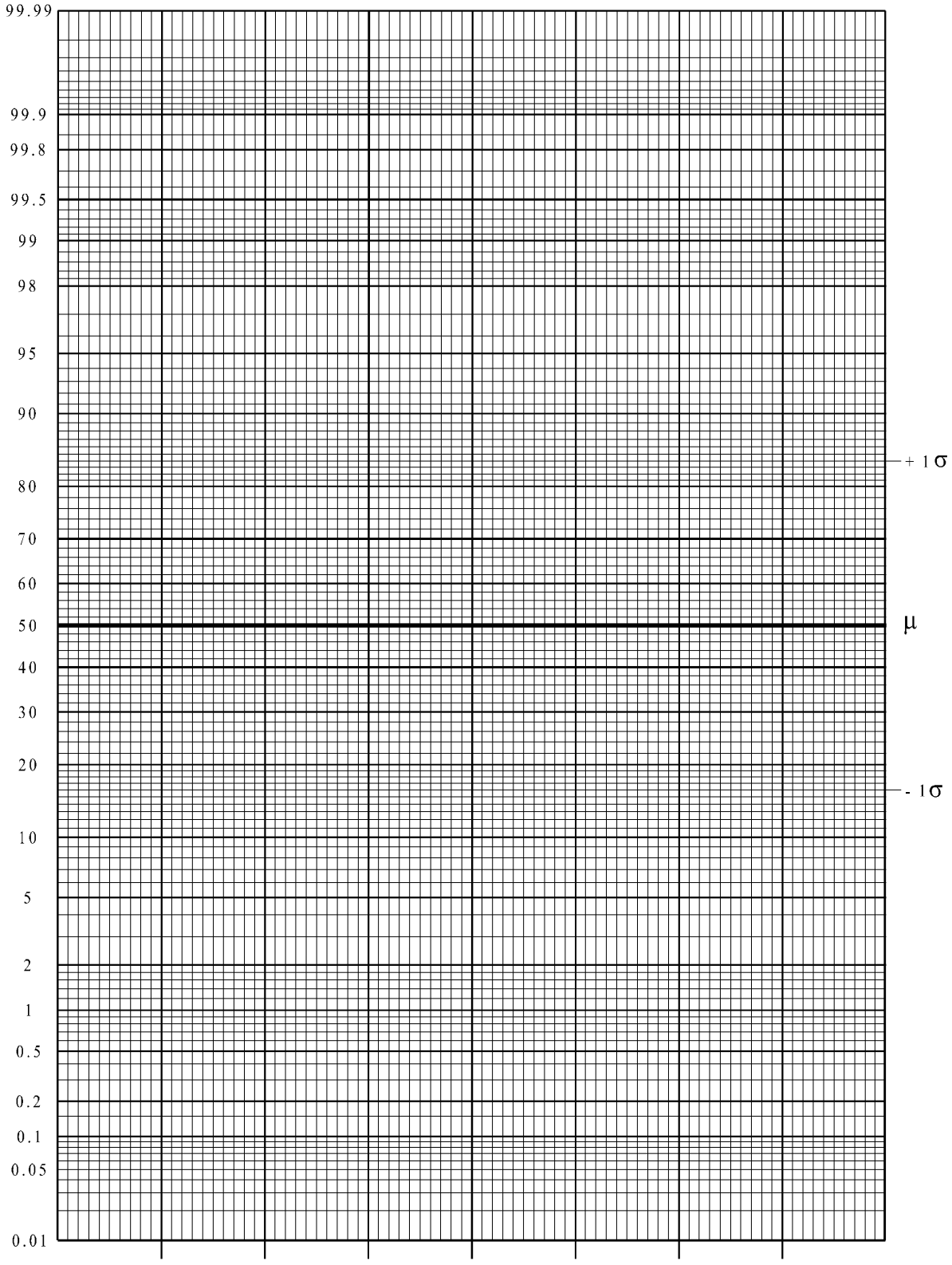
Valores
-8,75
23,75
-1,75
-6,25
0,75
5,25
-1,25

$$P = \frac{(i - 0{,}5)}{7} x100\%$$

i (posto)	Valor	P (percentil)
1	-8,75	7,1
2	-6,25	21,4
3	-1,75	35,7
4	-1,25	50,0
5	0,75	64,3
6	5,25	78,6
7	23,75	92,9

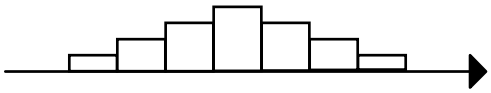
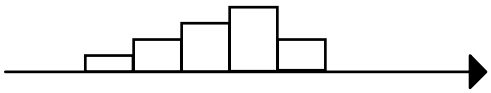
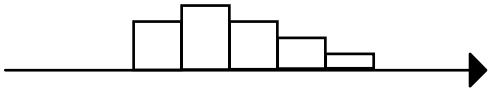
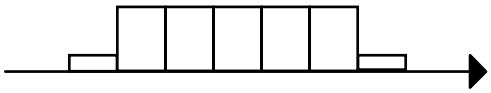
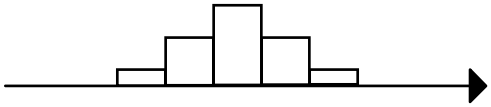
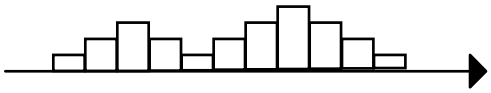
© ALBERTO W. RAMOS 2016

Papel de Probabilidade Normal



© ALBERTO W. RAMOS 2016

INTERPRETAÇÃO

HISTOGRAMA	*PPN*
NORMAL 	
ASSIMÉTRICO A ESQUERDA 	
ASSIMÉTRICO A DIREITA 	
ACHATADO 	
ALONGADO 	
BIMODAL 	

© ALBERTO W. RAMOS 2016

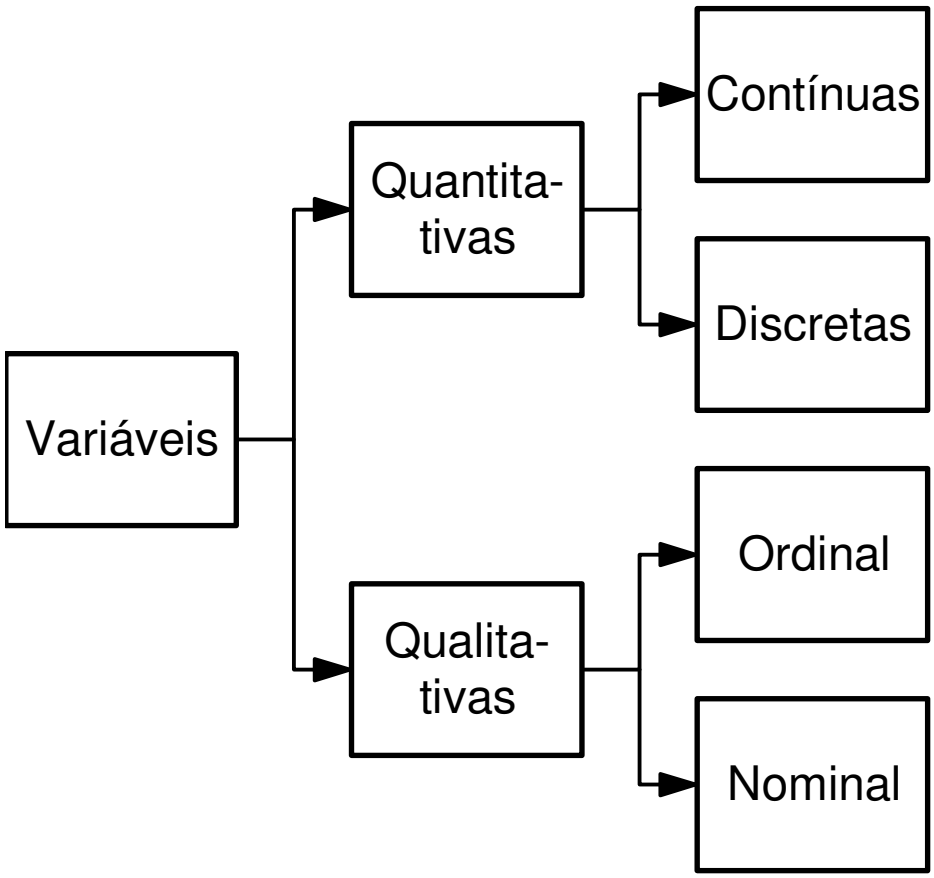
Teste Qui-Quadrado

© ALBERTO W. RAMOS 2016

TESTE QUI-QUADRADO

O teste qui-quadrado serve para avaliar se duas variáveis qualitativas (também chamadas de categóricas) são ou não independentes entre si.

.



Logo, o conjunto de hipóteses que está sendo testado é:

H_0: as variáveis são independentes
H_1: as variáveis são dependentes

© ALBERTO W. RAMOS 2016

EXEMPLO

Uma amostra de 300 estudantes de uma universidade foi obtida, e estes foram classificados quanto a :

Área de concentração : Exatas ; Humanas
Jornal preferido : A, B, C, Outros

Obtendo-se os seguintes resultados.

O_{ij}	Jornal A	Jornal B	Jornal C	Outros	Total
Exatas	60	20	90	20	190
Humanas	30	40	30	10	110
Total	90	60	120	30	300

Existem evidências de que Área de Concentração e Jornal Preferido estejam relacionados (dependência) ?

© ALBERTO W. RAMOS 2016

As hipóteses testadas, neste caso, são:

H_o : Área e Jornal Preferido são independentes
H_1 : Área e Jornal Preferido não são independentes

Na amostra havia 190 alunos de Exatas, num total de 300, ou seja, 190/300 = 0,633 ou 63,3% e, conseqüentemente, havia 110/300 = 0,366 ou 36,6% de alunos de Humanas.

Na coluna do Jornal A obteve-se um total de 90 alunos. Logo, se não houver dependência entre área e jornal, espera-se que:

Proporção de Exatas ➔ $\frac{190}{300} x90 = 57$

Proporção de Humanas ➔ $\frac{110}{300} x90 = 33$

Analogamente, para as demais colunas, obtem-se os valores entre parênteses da tabela abaixo.

O_{ij} (E_{ij})	Jornal A	Jornal B	Jornal C	Outros	Total
Exatas	60(57)	20(38)	90(76)	20(19)	190
Humanas	30(33)	40(22)	30(44)	10(11)	110
Total	90	60	120	30	300

Ou, genericamente:

$$E_{ij} = \frac{L_i \times C_j}{n}$$

L_i = Total da linha i $\qquad$ C_j = Total da coluna j

© ALBERTO W. RAMOS 2016

Assim, pode-se obter as diferenças entre o observado (O_{ij}) e o esperado (E_{ij}) na tabela, conforme abaixo:

O_{ij} - E_{ij}	Jornal A	Jornal B	Jornal C	Outros	Total
Exatas	+3	-18	+14	+1	0
Humanas	-3	+18	-14	-1	0
Total	0	0	0	0	0

Define-se como qui-quadrado, à estatística:

$$\chi^2_{calc} = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

χ^2	Jornal A	Jornal B	Jornal C	Outros	Total
Exatas	0,158	8,526	2,579	0,053	11,316
Humanas	0,273	14,727	4,455	0,091	19,545
Total	0,431	23,254	7,033	0,144	30,861

que, para a decisão, será comparado contra

$$\chi^2_{crítico} = \chi^2_{(L-1)(C-1);\alpha} = \chi^2_{3;5\%} = 7{,}815$$

como $\chi^2_{calc} > \chi^2_{crítico} \Rightarrow$ Rejeito H_o

© ALBERTO W. RAMOS 2016

BIBLIOGRAFIA

BUSSAB, W. O.; MORETTIN, P. A. ***Estatística básica***. 5 ed. São Paulo, Saraiva, 2004.

COSTA NETO, P. L. O. ***Estatística***. 2 ed. São Paulo, Edgard Blucher, 2002.

DOWNING, D.; CLARK, J. ***Estatística aplicada***. São Paulo, Saraiva, 1999.

FONSECA, J. S.; MARTINS, G. A.; TOLEDO, G. L. ***Estatística aplicada.*** São Paulo, Atlas, 1982.

FREUND, J. E.; SIMON, G. A. ***Estatística aplicada.*** Porto Alegre, Bookman, 2000.

LAREDO, A. Notas de aula do curso de estatística I. São Paulo, FGV, 1998.

© ALBERTO W. RAMOS 2016