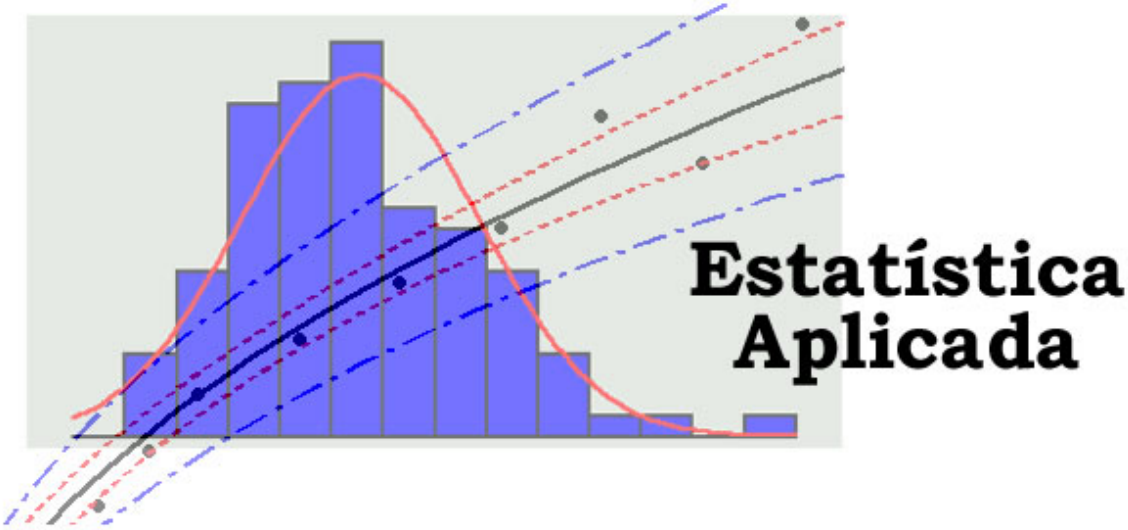
UNIFEI - Universidade Federal de Itajubá

Instituto de Engenharia de Produção & Gestão



Notas compiladas por

PEDRO PAULO BALESTRASSI

ANDERSON PAULO DE PAIVA

Itajubá/2007

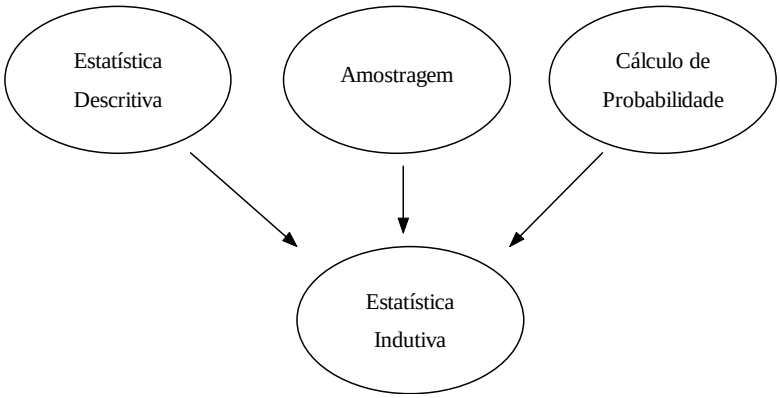
CAPÍTULO 1 - ESTATÍSTICA

1.1 - Do que trata a Estatística

A essência da ciência é a observação. A ciência que se preocupa com a organização, descrição, análise e interpretação dos dados experimentais é denominada de *Estatística*, um ramo da *Matemática Aplicada*. A palavra estatística provêm de "*Status*".

1.1.1 - Grandes áreas

O diagrama seguinte mostra o contexto em que se situa o estudo da Estatística, aqui subdividido em Estatística *Descritiva* e Estatística *Indutiva* (ou *Inferencial*).



A Estatística Descritiva está relacionada com a organização e descrição de dados associada a cálculos de médias, variâncias, estudo de gráficos, tabelas, etc.. É a parte mais conhecida .

A Estatística Indutiva é o objetivo básico da ciência. A ela está associada, *Estimação de Parâmetros*, *Testes de Hipóteses*, *Modelamento*, etc..

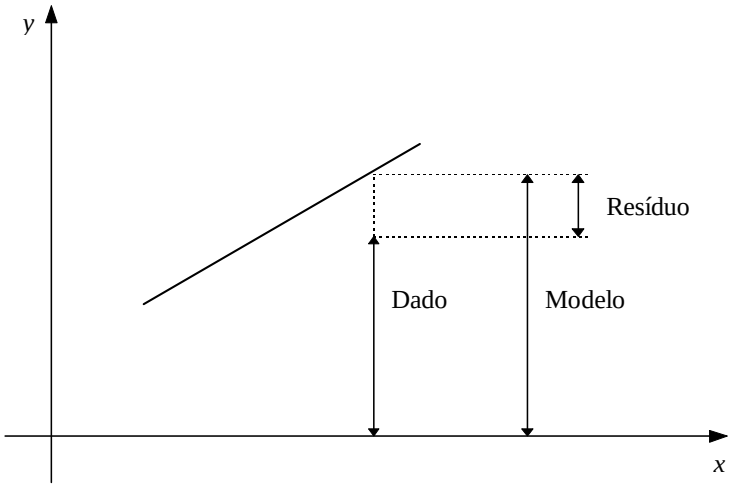
No *Cálculo de Probabilidades*, está a essência dos modelos *Não-Determinísticos* e a corroboração de que toda inferência estatística está sujeita a erros.

A *Amostragem* é o ponto de partida (na prática) para todo um Estudo Estatístico.

Aqui pode se ter origem um problema bastante comum em Engenharia: “Análise profunda sobre dados superficiais ! ”.

1.1.2 - Modelos

Um dos objetivos da análise de dados é buscar um modelo para as observações. Na figura abaixo, duas variáveis estão representadas.



De um modo esquemático, podemos então escrever que:

Dado = Modelo + Resíduo (D = M + R)

Tukey (1977) chama *M* de parte “suave” dos dados e *R* de parte “grosseira”.

A análise de *R* é tão importante quanto a de *M*.

Os modelos podem ser essencialmente *determinísticos* ou *não-determinísticos* (*probabilísticos* ou *estocásticos*). Nos determinísticos as condições sob as quais um experimento é executado determina o resultado do experimento.

Ex.: *I* pode ser determinado por *V/R* em um circuito elétrico resistivo elementar.

Nos modelos não determinísticos usa-se uma *Distribuição de Probabilidade*.

Ex.: Peças são fabricadas até que *x* peças perfeitas sejam produzidas; o número total de peças fabricadas é contado. Usa-se uma distribuição, no caso a *Geométrica*, para a tomada de decisões

1.1.3 - População e Amostra

O estudo de qualquer fenômeno, seja ele natural, social, econômico ou biológico, exige a coleta e a análise de dados estatísticos. A coleta de dados é, pois, a fase inicial de qualquer pesquisa.

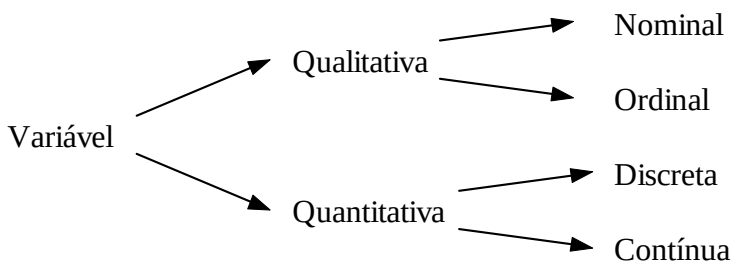
A *População* é a coleção de todas as observações potenciais sobre determinado fenômeno. O conjunto de dados efetivamente observados, ou extraídos, constitui uma *Amostra* da população. É sobre os dados da amostra que se desenvolvem os estudos, com o objetivo de se fazer inferências sobre a população.

1.2 - Análise exploratória dos Dados

Os erros e inconsistências ocorridos na coleta de dados devem ser corrigidos. As amostras de dados devem ser agrupadas de forma que seu manuseio, visualização e compreensão sejam simplificados.

1.2.1 - Tipos de variáveis

Os dados coletados em uma primeira fase podem ser definidos como variáveis qualitativas ou quantitativas, de acordo com a seguinte figura:



Ex.: Para uma população de peças produzidas em um determinado processo, poderíamos ter:

Variável	***Tipo***
Estado: Perfeita ou defeituosa	Qualitativa Nominal
Qualidade: 1ª, 2ª ou 3ª categoria	Qualitativa Ordinal
Nº de peças defeituosas	Quantitativa Discreta
Diâmetro das peças	Quantitativa Contínua

1.2.2 - Agrupamentos de Dados e Distribuição de Freqüências

Quando se vai fazer um levantamento de uma população, um dos passos é retirar uma amostra desta população e obter dados relativos à variável desejada nesta amostra.

Cabe à Estatística sintetizar estes dados na forma de tabelas e gráficos que contenham, além dos valores das variáveis, o número de elementos correspondentes a cada variável.

A este procedimento está associado o conceito de:

- **Dados brutos**: É o conjunto de dados numéricos obtidos e que ainda não foram organizados.
- **Rol**: É o arranjo dos dados brutos em ordem crescente (ou decrescente).
- **Amplitude** (*H*): É a diferença entre o maior e o menor dos valores observados.
- **Freqüência absoluta** (n_i): É o número de vezes que um elemento aparece na amostra:

$\sum_{1}^{K} n_i = n$ onde *n* é o número total de dados da amostra e *k* é o número de valores diferentes na amostra.

- **Freqüência Relativa** (f_i):

$$f_i = \frac{n_i}{n} \quad \text{e} \quad \sum_{i=1}^{K} f_i = 1$$

- **Freqüência Absoluta Acumulada** (N_i): É a soma da freqüência absoluta do valor da variável *i* com todas as freqüências absolutas anteriores.

- **Freqüência Relativa Acumulada** (F_i):

$$F_i = \frac{N_i}{n}$$

Ex.: População = Diâmetro de determinada peça (em mm).

Dados brutos: { 168, 164, 164, 163, 165, 168, 165, 164, 168, 168 }

Rol : { 163, 164, 164, 164, 165, 165, 168, 168, 168, 168 }

H = 168 - 163 = 5

X	n_i	f_i	N_i	F_i
163	1	0.1	1	0.1
164	3	0.3	4	0.4
165	2	0.2	6	0.6
168	4	0.4	10	1.0
Σ	10	1		

Obs.: Esse tipo de tabela é denominada de "Distribuição de Freqüências".

Questão: Como implementar um programa computacional que forneça o rol de 2000 valores de uma determinada variável, de maneira ótima?

1.2.3 - Classes

As classes são um artifício para condensar o número de elementos diferentes de uma amostra. Imagine construir uma tabela para 200 valores diferentes, nos moldes do problema anterior?

Os principais pré-requisitos para uma boa definição de classes em um conjunto de dados são:

a) As classes devem abranger todas as observações;

b) O extremo superior de uma classe é o extremo inferior da classe subseqüente (simbologia: ├──, intervalo fechado à esquerda e aberto à direita);

c) Cada valor observado deve enquadrar-se em apenas uma classe;

d) $k \leq 25$, de um modo geral, sendo k o número de classes;

e) As unidades das classes devem ser as mesmas dos dados.

Cálculo de k (opções não rígidas):

• Fórmula de Sturges:

$k = 1 + \log_2^N$

• $k \cong \sqrt{N}$

Obs.: N é o número de elementos diferentes da amostra e em muitas vezes pode ser considerado $N = n$

Geralmente, temos ainda:

• Intervalo da classe (h): $h \cong H / k$

• Ponto médio da classe (x_i): Ponto médio entre o limite inferior e o limite superior de cada classe .

Ex.: Construção da distribuição de freqüências contendo classes:

X	x_i	n_i	f_i	$f\%$	N_i	F_i	$F\%$
10 ├── 20	15	2	0.04	**4**	2	**0.04**	**4**
20 ├── **30**	**25**	**12**	**0.24**	24	14	**0.28**	**28**
30 ├── 10	**35**	18	**0.36**	**36**	**32**	**0.64**	**64**
40 ├── **50**	45	**13**	0.26	**26**	**45**	**0.9**	**90**
50 ├── 60	**55**	**5**	**0.1**	**10**	**50**	**5.0**	**100**
Σ		**50**	1				

Obs.: A partir dos dados que não estão em negrito pode-se montar toda a tabela. Confira!.

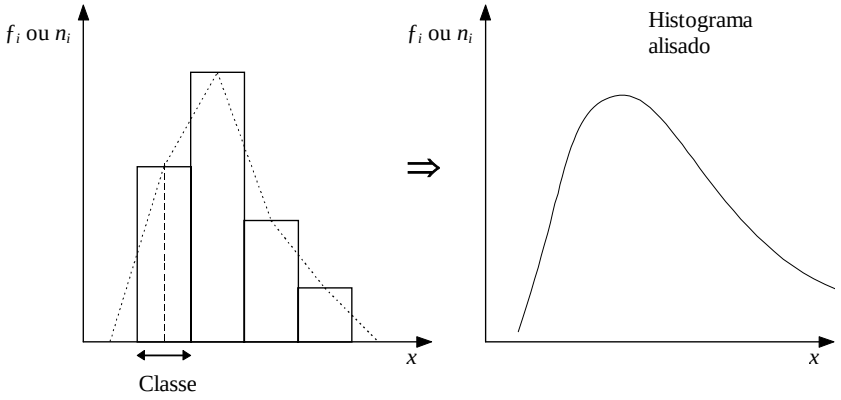
1.2.4 - Gráficos

Tradicionalmente, uma análise descritiva dos dados se limita a calcular algumas medidas de posição e variabilidade, como média e variância, por exemplo.

Contrária a esta tendência, uma corrente mais moderna, liderada por Tukey , utiliza principalmente técnicas visuais, representações pictóricas dos dados, em oposição aos dados numéricos.

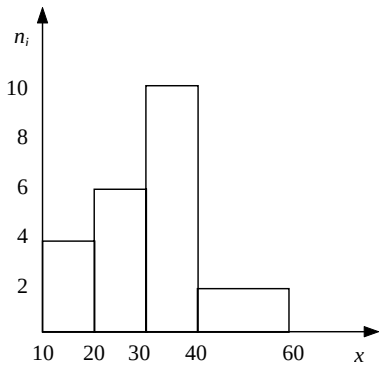
Os principais gráficos serão sucintamente descritos a seguir. É importante lembrar que os modernos programas computacionais de Edição de Texto, Planilha Eletrônica e Banco de Dados facilitam em muito a manipulação com gráficos. Alguns desses programas são: Word, Wordperfect, QuatroPro, Lotus, Dbase, Excell, ...

- Histograma e polígono de freqüência.



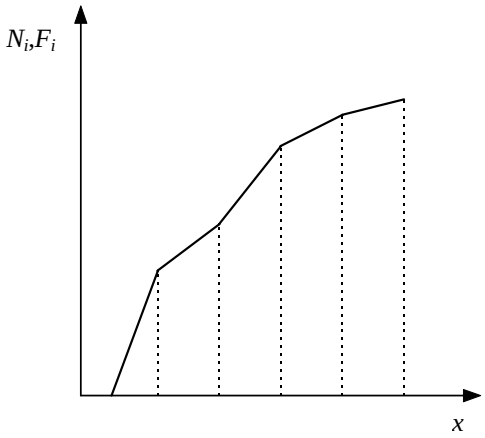
As áreas dos retângulos são proporcionais às freqüências e o polígono utiliza os pontos médios das classes. O histograma alisado é obtido quando o intervalo da classe é diminuído suficientemente para que uma curva contínua possa ser traçada. Tal curva é muito útil para ilustrar rapidamente qual o tipo de comportamento que se espera para a distribuição de uma dada variável. No capítulo referente a variáveis aleatórias contínuas, voltar-se-á a estudar esse histograma sob um ponto de vista mais matemático.

Ex.: Construção da tabela de distribuição de freqüências a partir do histograma de classes desiguais.



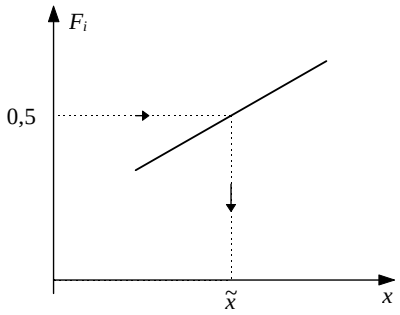
X	n_i	f_i
10 ├── 20	4	4/24
20 ├── 30	6	6/24
30 ├── 40	10	10/24
40 ├── 60	4	4/24 !!
Σ	24	1

- Ogiva (De Galton)

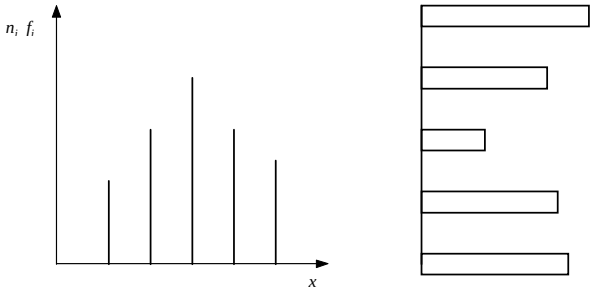


A ogiva utiliza os pontos extremos das classes e é usado em freqüências acumuladas. Tal gráfico pode fornecer informações adicionais por meio de simples operações gráficas.

Ex.: Para um valor de F_i correspondendo a 0.5 (50%) pode-se chegar à mediana ($\tilde{x}$) do conjunto de observações:

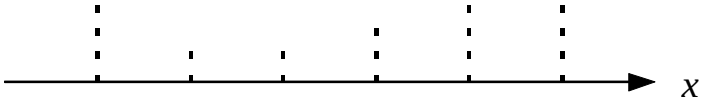


- Gráficos em barras (ou colunas)



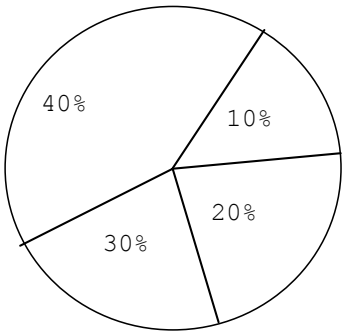
As distribuições não envolvem classes ou são qualitativas.

• Gráfico de pontos



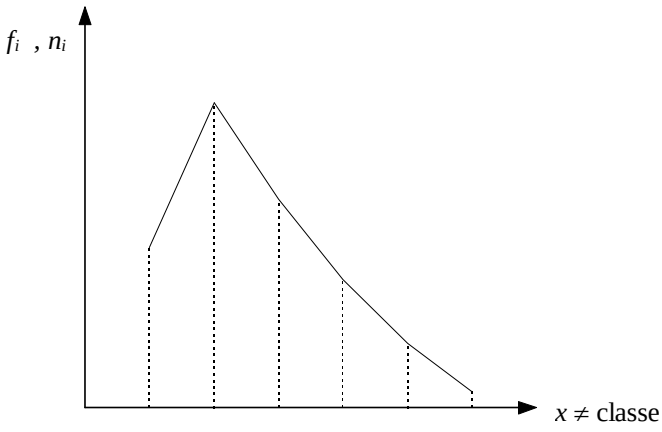
Para pequena quantidade de elementos.

• Gráfico em setores



As subdivisões são mensuráveis.

• Gráfico em linha



Um dos mais utilizados. Não há observações intermediárias.

<u>1.2.5 - Ramo-e-folhas</u>

Alternativo ao gráfico, o Ramo-e-Folhas é uma boa prática (sempre que possível), ao se iniciar uma análise de dados.Temos aqui:

	—	x					
Ramos	—	x	x				Folhas
	—	x	x	x	x	x	
	—	x	x	x			

Ex.: A ingestão diária média, per capita, em gramas, de proteínas para 33 países desenvolvidos é:

81	113	108	74	79	78	90	93	105	109	93
106	103	100	100	100	101	101	101	95	90	94
90	91	92	93	87	89	78	89	85	94	86

Construção do ramo-e-folhas:

11	3										
10	8	5	9	6	3	0	0	0	1	1	1
9	0	3	3	5	0	4	0	1	2	3	4
8	1	7	9	9	5	6					
7	4	9	8	8							

ou

11	3										
10^+	8	5	9	6							
10^-	3	0	0	0	1	1	1				
9^-	0	3	3	5	0	4	0	1	2	3	4
8	1	7	9	9	5	6					
7	4	9	8	8							

Exercícios I

01) Os pesos dos jogadores de um time de futebol variam de 75 a 95 quilos. Quais seriam os extremos se quiséssemos grupá-los em 10 classes?

02) em certa época, os salários mensais dos operários de uma indústria eletrônica variavam de 1.500 a 3.250 unidades monetárias. Quais seriam os extremos se quiséssemos grupá-los em seis classes?

03) Os números de lugares vagos em vôos entre duas cidades foram agrupados nas classes 0├──5 5├──10 10├──20 20├──25 25├──30 30 ou mais.
Com esta distribuição é possível determinar o número de vôos em que há:
(a) menos de 20 assentos vagos; (b) mais de 20; (c) ao menos 9; (d) no máximo 9; (e) exatamente 5; (f) entre 10 e 25?

04) Os pontos médios de uma distribuição de leituras de temperaturas são 16, 25, 34, 43, 52, 61. Determinar os limites de classe e o intervalo de classe.

05) Para agrupar os dados relativos ao número de dias chuvosos em julho durante os últimos 50 anos, um meteorologista utilizou as classes 0—5, 6—11, 12—16, 18—24, 24—31. Explique onde pode surgir dificuldades.

06) Construir o histograma para a distribuição de freqüências abaixo:

Nº Empregados	n_i
0 ├── 10	5
10 ├── 20	20
20 ├── 30	35
30 ├── 40	40
40 ├── 60	50
60 ├── 80	30
80 ├── 100	20
100 ├── 140	20
140 ├── 180	15
180 ├── 260	15

Representar o Histograma alisado.

07) Para distribuição de freqüência abaixo, construir a ogiva de Galton.

Salários	n_i
4,00 ├── 8,00	10
8,00 ├── 12,00	12
12,00 ├── 16,00	8
16,00 ├── 20,00	5
20,00 ├── 24,00	1

Aproximadamente, qual o salário *S* tal que 50% dos funcionários ganham menos que *S*.

08) Os dados abaixo referem-se à produção em toneladas, de dado produto, para 20 companhias químicas (numeradas de 1 a 20).

(1,50),	(2,280),	(3,50),	(4,170),	(5,180)
(6,500),	(7,250),	(8,200),	(9,1050),	(10,240)
(11,180),	(12,1000),	(13,1100),	(14,120),	(15,4200)
(16,5100),	(17,480),	(18,90),	(19,870),	(20,360)

Construir o ramo-e-folhas das produções.

1.3 - Medidas Estatísticas

A redução dos dados através de ramo-e-folhas e tabelas de freqüências fornece muito mais informações sobre o comportamento de uma variável do que a própria série original de dados. Contudo, muitas vezes queremos resumir ainda mais esses dados, apresentando um ou alguns valores que sejam representativos da série toda. Quando usamos um só valor, obtemos uma redução drástica dos dados.

As principais medidas estatísticas (ou simplesmente estatísticas) referem-se às medidas de posição (locação ou tendência central) ou às medidas de dispersão (ou variabilidade):

1.3.1 - Medidas de posição

Mostram o valor representativo em torno do qual os dados tendem a agrupar-se com maior ou menor freqüência.

- **Média Aritmética simples ($\bar{x}$) –(ou simplesmente média)**

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

Usada em dados não agrupados em classes .

- **Média Aritmética ponderada (também $\bar{x}$)**

$$\bar{x} = \frac{x_1 p_1 + x_2 p_2 + \cdots + x_n p_n}{p_1 + p_2 + \cdots + p_n} = \frac{\sum_{i=1}^{n} x_i p_i}{\sum_{i=1}^{n} p_i} \quad \text{onde,}$$

p_i = peso da amostra x_i

Agora, para dados agrupados em classes, temos:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i n_i}{\sum_{i=1}^{n} n_i} = \frac{1}{n} \sum_{i=1}^{n} x_i n_i = \sum_{i=1}^{n} x_i f_i$$

A média aritmética simples pode ser vista como a média ponderada com todos os pesos iguais. Para efeito de nomenclatura sempre trataremos a média aritmética simples ou ponderada simplesmente por média ($\bar{x}$).

• **Mediana ($\tilde{x}$)**

É o valor "do meio" de um conjunto de dados, quando os dados estão dispostos em ordem crescente ou decrescente.

Para dados não agrupados em classes:

Se n é ímpar $\rightarrow \tilde{x} = \left(\frac{n+1}{2}\right)^{o}$ termo

Se n é par $\rightarrow \tilde{x} = \dfrac{\left(\frac{n}{2}\right)^{o} \text{termo} + \left(\frac{n}{2}+1\right)^{o} \text{termo}}{2}$

Ex.: $\{35, 36, 37, 38, 40, 40, 41, 43, 46\} \Rightarrow \tilde{x} = 40$

$\{12, 14, 14, 15, 16, 16, 17, 20\} \Rightarrow \tilde{x} = \frac{15+16}{2} = 15{,}5$

• **Média × Mediana**

A média é muito sensível a valores extremos de um conjunto de observações, enquanto a mediana não sofre muito com a presença de alguns valores muito altos ou muito baixos. A mediana é mais "robusta" do que a média.

Devemos preferir a mediana como medida sintetizadora quando o histograma do conjunto de valores é assimétrico, isto é, quando há predominância de valores elevados em uma das caudas.

Ex.: { 200, 250, 250, 300, 450, 460, 510 }

$\bar{x} = 345{,}7$ $\tilde{x} = 300$

Tanto $\bar{x}$ como $\tilde{x}$ são boas medidas de posição.

Ex.: { 200, 250, 250, 300, 450, 460, 2300 }

$\bar{x} = 601$ $\tilde{x} = 300$

Devido ao valor 2300, $\tilde{x}$ é preferível a $\bar{x}$.

• **Percentis**

“ O percentil de ordem 100.p, $0 \leq p \leq 1$, de um conjunto de valores dispostos em ordem crescente é um valor tal que (100 p)% das observações estão nele ou abaixo dele e 100 (1 - p)% estão nele ou acima dele.”

Os percentis de ordem 25,50 e 75 são chamados de *quartis*. Representam-se por Q_1, Q_2 e Q_3. Naturalmente $Q_2 = \tilde{x}$.

Os *decis* são os percentis de ordem 10, 20, ... 90.

Ex.: Para valores de 51 a 100, ordenados crescentemente:

Q_1 deixa 25% dos dados (12,5 $\Rightarrow$ 13 valores) nele ou abaixo dele e 75% dos dados (37,5 $\Rightarrow$ 38 valores) nele ou acima dele. Assim: $Q_1 = 63$.

Similarmente, P_{80} deixa 80% dos dados (40 valores) nele ou abaixo dele e 20% dos dados (10 valores) nele ou acima dele. Assim: $P_{80} = \dfrac{90+91}{2} = 90{,}5$

Obs.: Os programas computacionais usam regras de interpolação para o cálculo de percentis e diferem ligeiramente dos valores acima (No Excel, por exemplo, $Q_1 = 63{,}25$)

Para dados agrupados em classes, os percentis podem ser obtidos por interpolação linear (regra de três simples). Para $P_{50} = Q_2 = \tilde{x}$, por exemplo, temos:

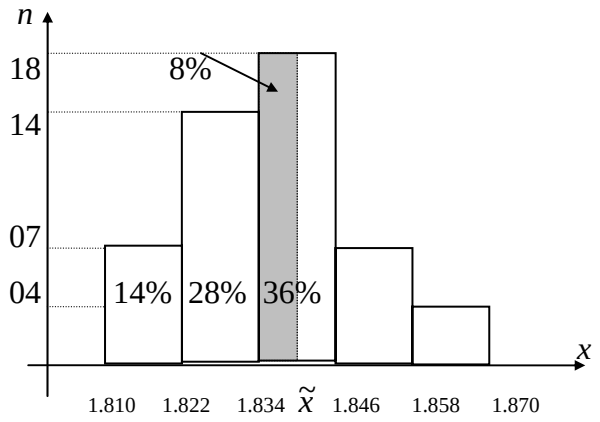
$$\tilde{x} = L_i + \frac{(n/2) - N_a}{n_i} \cdot h$$

O exemplo a seguir explica os termos desta “fórmula”. Faça a associação!

Ex.: Dada a distribuição de freqüência de uma variável X qualquer:

X	x_i	n_i	N_i
1,810 |— 1,822	1,816	7	7
1,822 |— 1,834	1,828	14	21
1,834 |— 1,846	1,840	18	39
1,846 |— 1,858	1,852	7	46
1,858 |— 1,870	1,864	4	50

Temos que, para ($\tilde{x}$), o 25º elemento (50% de 50), está na terceira classe:



$$\frac{1.846 - 1.834}{36\%} = \frac{\tilde{x} - 1.834}{8\%}$$

Logo $\tilde{x}$ =1.837

Um outro processo gráfico pode ser usado para o cálculo desses percentis. (Veja Ogiva de Galton). Tal processo exige rigor no traçado e deve-se preferir papel milimetrado.

Obs.*:* As calculadoras geralmente não fornecem mediana e percentis.

• **A Média Aparada (ou média interna)**

É obtida eliminando do conjunto as *m* maiores e as *m* menores observações. Geralmente, $2{,}5\% \leq m \leq 5\%$ dos dados. Esta eliminação corresponde, na realidade, à supressão dos valores extremos - muito altos ou muito baixos. Tal média representa um valor entre $\bar{x}$ e $\tilde{x}$.

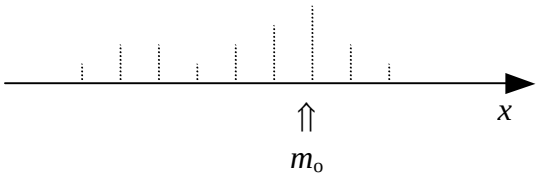
Ex.: { 200, 250, 250, 300, 450, 460, 2300 }

$$\bar{x}(m=1) = \frac{250 + 250 + 300 + 450 + 460}{5} = 342$$

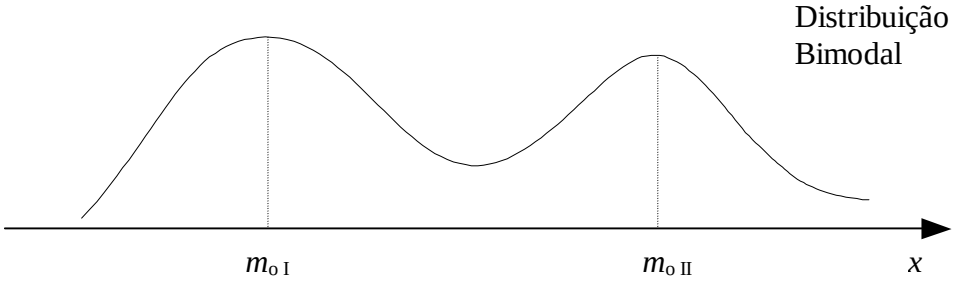
• **A moda e a classe modal (m_o)**

É o valor que representa a maior freqüência em um conjunto de observações individuais. Para dados agrupados temos a classe modal. Em alguns casos pode haver mais de uma moda. Assim temos uma distribuição bimodal, trimodal, etc...

Ex.:



X	x_i	n_i	
10 |— 20	15	2	
20 |— 30	25	4	
30 |— 40	35	10	⇒ Classe Modal
40 |— 50	45	6	
50 |— 60	55	2	



Obs.: A moda geralmente não é fornecida em calculadoras.

1.3.2 - Medidas de dispersão ou Variabilidade

Quase nunca uma única medida é suficiente para descrever de modo satisfatório um conjunto de dados. Torna-se então necessário estabelecer medidas que indiquem o grau de dispersão em relação ao valor central.

Nos seguintes conjuntos de dados, por exemplo:

A = { 3, 4, 5, 6, 7 }

B = { 1, 3, 5, 7, 9 }

C = { 5, 5, 5, 5 }

D = { 3, 5, 5, 7 }

E = { 3.5, 5, 6.5 }

F = { -1000, 15, 1000 }

Temos em todos eles a mesma média. A identificação de cada um desses

conjuntos de dados pela sua média nada informa sobre as diferentes variabilidades dos mesmas. Algumas medidas que sintetizam essa variabilidade são:

- **Amplitude (*H*):**

Como anteriormente definida, *H* tem o inconveniente de levar em conta somente os dois valores extremos, o maior e o menor deles.

- **Desvio Médio (*DM*(*x*)), Variância (S^2,*Var(X) ou* σ^2) e Desvio Padrão (*S*,*DP(X) ou* σ):**

Aqui o princípio básico é analisar os desvios das observações em relação à média das observações.

Em *A* = { 3, 4, 5, 6, 7 },por exemplo, os desvios x_i - $\bar{x}$ são: -2, -1, 0, 1, 2. É fácil ver que a soma dos desvios, é identicamente nula e que portanto não serve como medida de dispersão:

$$\sum_{i=1}^{n}(x_1-\bar{x})=\sum_{i=1}^{n}x_1-\sum_{i=1}^{n}\bar{x}=n\bar{x}-n\bar{x}\equiv 0$$

Duas opções para analisar os desvios das observações são:

a) considerar o total dos desvios em valor absoluto ou;

b) considerar o total dos quadrados dos desvios.

Assim, para o conjunto A, teríamos, respectivamente

$$\sum_{i=1}^{5}\left|x_i-\bar{x}\right|=2+1+0+1+2=6 \quad \text{e}$$

$$\sum_{i=1}^{5}\left(x_i-\bar{x}\right)^2=4+1+0+1+4=10$$

Associando estas medidas à média, temos:

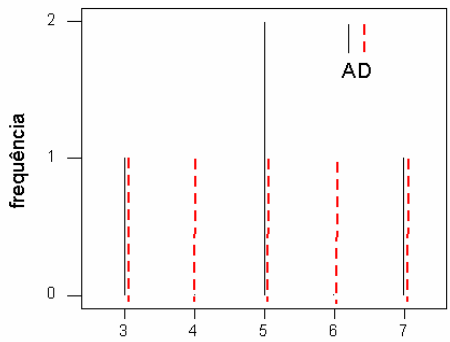
$$DM(x)=\frac{\sum_{i=1}^{n}\left|x_i-\bar{x}\right|}{n}$$ que é o desvio médio.

$$S^2 = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}{n}$$ que é a variância ($Var(x)$)

Sendo a variância uma medida que expressa um desvio quadrático médio, é conveniente usar uma medida que expresse a mesma unidade dos dados originais. Tal medida é o desvio padrão S (ou $DP(x)$), dada por:

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

O uso do *DM(x)* pode causar dificuldades quando comparamos conjuntos de dados com número diferentes de observações.



Ex.: Em $A = \{ 3, 4, 5, 6, 7 \}$ temos:

$DM(x) = 6/5 = 1.2$ e

$S^2 = 10/5 = 2$

Em $D = \{ 3, 5, 5, 7 \}$ temos:

$DM(x) = 1{,}0$ e

$S^2 = 2{,}0$

Assim, podemos dizer que, segundo o Desvio Médio, o Grupo *D* é mais homogêneo (tem menor dispersão) do que *A*, enquanto que ambos tem a mesma homogeneidade segundo a variância. O desvio médio possui pequena utilização em estatística e em geral vale 0.8 vezes o desvio padrão

O cálculo do desvio padrão exige o cálculo prévio da variância e uma fórmula alternativa para S^2 é dada por:

$$S^2 = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}{n} = \frac{\sum_{i=1}^{n} x_i^{\,2}}{n} - \bar{x}^2$$ (Confira!)

Relacionados à inferência estatística, alguns autores usam (n - 1) como divisor para a variância:

$$S^2 = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}{n-1}$$, e isto será visto adiante (tendenciosidade).

Obs.: Muitas calculadoras científicas possuem duas medidas para desvio padrão. Uma associada à divisão por n (simbolizada geralmente por σ ou σ_n) e outra associada à divisão por n - 1 (chamada também de não-polarizada, simbolizada geralmente por S ou σ_{n-1}). Verifique a simbologia usada pela sua calculadora, caso você possua uma!

Para dados agrupados em classes, a variância é dada por:

$$S^2 = \frac{\sum_{i=1}^{K} \left(x_i - \bar{x}\right)^2 \cdot n_i}{n} \quad \text{ou} \quad S^2 = \sum_{i=1}^{K} \left(x_i - \bar{x}\right)^2 \cdot f_i$$

$$S^2 = \sum_{i=1}^{K} x_i^{\,2} \cdot f_i - \bar{x}^2 \quad \text{ou} \quad S^2 = \sum_{i=1}^{K} x_i^{\,2} \cdot \frac{n_i}{n} - \bar{x}^2$$

É intuitivamente claro que a multiplicação de cada parcela $(x_i - \bar{x})^2$ por n_i significa a repetição dos desvios quadrados $(x_i - \bar{x})^2$, n_i vezes.

- **Momentos de uma distribuição de frequências**

Definimos o momento de ordem t de um conjunto de dados como

$$M_t = \frac{\sum_{i=1}^{n} x_i^{\,t}}{n}$$

Definimos o momento de ordem t centrado em relação a uma constante a como

$$M_t = \frac{\sum_{i=1}^{n} (x_i - a)^t}{n}$$

Especial interesse tem o caso do momento centrado em relação a $\bar{x}$, dado por

$$m_t = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^t}{n}$$

Conforme já vimos nos casos da média e da variância, as expressões precedentes podem ser reescritas levando-se em consideração as frequências dos diferentes valores existentes. Temos então respectivamente,

$$M_t = \frac{\sum_{i=1}^{n} x_i^{\,t} f_i}{n}$$

$$M_t^a = \frac{\sum_{i=1}^{n} (x_i - a)^t f_i}{n}$$

$$m_t = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^t f_i}{n}$$

É fácil ver que: $M_1 = \bar{x}$; $m_1 = 0$; $m_2 = s^2$

• **Escores padronizados (z_i)**

No contexto de um único conjunto de dados, o desvio padrão pode ser interpretado intuitivamente como unidade natural de dispersão de dados. Essa interpretação é utilizada na construção de "*escores padronizados*" de larga utilização:

$$z_i = \frac{x_i - \bar{x}}{s}$$

Repare que:

i) $x_i - \bar{x}$ considera o afastamento de x_i em relação à média.

ii) A divisão por s toma s como unidade ou padrão de medida.

Ex.: Dois grupos de pessoas acusam os seguintes dados:

Grupo	Peso médio	Desvio Padrão
A	66.5 kg	6.38 kg
B	72.9 kg	7.75 kg

Nesses grupos há duas pessoas que pesam respectivamente, 81.2 kg e 88.0 kg. Usando o conceito de escore padronizado podemos ver que:

em A : $z_A = \dfrac{81{,}2 - 66{,}5}{6{,}38} = 2{,}3$ e

em B : $z_B = \dfrac{88 - 72{,}9}{7{,}75} = 1{,}95$

Logo, a pessoa de A revela um maior excesso relativo de peso.

- **Coeficiente de variação (*cv*)**

$$cv = \frac{S}{\bar{x}}$$

cv exprime a variabilidade em termos relativos. É uma medida adimensional e sua grande utilidade é permitir a comparação das variabilidades em diferentes conjuntos de dados.

Ex.: Testes de resistência à tração aplicados a dois tipos diferentes de aço:

	$\bar{x}$ (kg/mm^2)	s (kg/mm^2)
Tipo I	27,45	2,0
Tipo II	147,00	17,25

$$cv_I = \frac{2}{27{,}45} = 7{,}29\% \qquad cv_{II} = \frac{17{,}25}{147} = 11{,}73\%$$

Assim, apesar do Tipo I ser menos resistente, é ele mais estável, mais consistente.

O uso do coeficiente de variação pode ser pensado considerando a questão: Um desvio padrão de 10 se a média é 10.000 é bem diferente se a média é 100!

- **Outros Tópicos:**
  - **Assimetria e Curtose**
  - **Box Plots (Método dos cinco números)**

Exercícios II

01) Mostre que:

(a) $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$

(b) $\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 = \sum_{i=1}^{n}x_1^2 - n\bar{x}^2 = \sum_{i=1}^{n}x_i^2 - \frac{\left(\sum x_i\right)^2}{n}$

(c) $\sum_{i=1}^{k}n_i\left(x_i - \bar{x}\right)^2 = \sum_{i=1}^{k}n_i x_i^2 - n\bar{x}^2$

(d) $\sum_{i=1}^{k}f_i\left(x_i - \bar{x}\right)^2 = \sum_{i=1}^{k}f_i x_i^2 - \bar{x}^2$

02) O conjunto abaixo representa as notas do exame final de uma determinada turma:

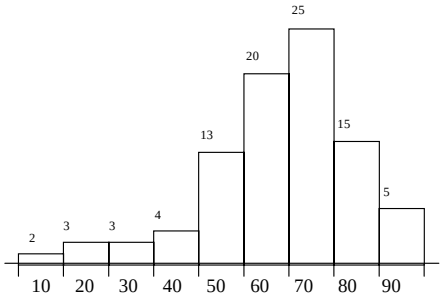
54	50	15	65	70	42	77	33	55	92
61	50	35	54	75	71	64	10	51	86
70	66	60	60	71	64	63	77	75	70
81	48	34	73	65	62	66	75	60	85
64	57	74	60	63	85	47	67	79	37
66	45	58	67	71	53	23	61	66	88
58	48	73	76	81	83	62	75	69	68
66	71	66	67	50	76	60	45	61	74

a) Construir uma distribuição de freqüência, adotando um intervalo de classe conveniente, o histograma e o polígono de freqüência. Construir o “ramo-e-folha”. (Sugerem-se classes de tamanho 10, a partir de 10.)

b) Calcular a média, o desvio padrão, a mediana, o 1º quartil e o 65º percentil.

c) Resolver (b) graficamente no que couber, com auxílio da ogiva de freqüência acumuladas.

03) Dado o histograma abaixo, calcular a média, a variância, a mediana e o 3º quartil. Sabe-se que o número total de observações é 90.



04) Dado o conjunto de observações

120 107 95 118 150 130 132 109 136

(a) Determinar os quartis.

(b) Calcular a média aparada com $m = 1$.

(c) A média aparada praticamente coincide com a média aritmética simples. Por quê?

05) Esboce o histograma de uma distribuição que tenha média e mediana iguais.

06) Construir dois conjuntos de dados de mesma amplitude mas com variabilidade diferentes.

07) Por engano, um professor omitiu uma nota no conjunto de notas de 10 alunos. Se as nove notas restantes são 48, 71, 79, 95, 45, 57, 75, 83, 97 e a média das 10 notas é 72, qual o valor da nota omitida.

08) Em certo ano, uma universidade pagou a cada um de seus 45 instrutores um salário médio mensal de R$ 1.500, a cada um de seus 67 assistentes R$ 2.000, a cada um de seus 58 adjuntos R$ 2.600 e a cada um de seus 32 titulares R$ 3.100. Qual o salário médio mensal dos 202 docentes?

09) São os seguintes os números de alunos e respectivos QI médios em três estabelecimentos de ensino:

Estabelecimento	**Nº de Alunos**	**QI Médio**
A	790	104
B	155	110
C	530	106

Qual é o QI médio global dos três estabelecimentos? Que tipo de média foi utilizada?

10) Alega-se que, para amostras de tamanho $n = 10$, a amplitude amostral deve ser aproximadamente três vezes o desvio padrão. Verifique a alegação com relação aos seguintes dados que representam a velocidade de 10 carros cronometrados em um posto de controle:

97 80 88 62 93 105 74 86 83 (Km/h).

11) Calcule $\bar{x}$, S^2 e S para os dados abaixo, onde os valores de x são pontos médios de intervalos. Determine o escore padronizado para $x = 80$:

x_i	15	25	35	45	55	65	75	85	95	105
n_i	1	5	12	18	21	19	10	7	6	1

12) Tempos, em minutos, de espera numa fila de ônibus, durante 13 dias, de um cidadão que se dirige diariamente ao seu emprego:

15 10 2 17 6 8 3 10 2 9 5 9 13

Calcular a média, a mediante e a moda.

13) Dada a seguinte distribuição de idades dos membros de uma sociedade,

Idade	x_i	n_i	**Idade**	x_i	n_i
15 ├── 20	17	16	35 ├── 40	37	17
20 ├── 25	22	35	40 ├── 45	42	8
25 ├── 30	27	44	45 ├── 50	47	2
30 ├── 35	32	27	50 ├── 55	52	1

Calcular:

a) $\bar{x}$ e s;

b) $\tilde{x}$, Q_1, Q_3;

c) o quartil médio dado por $\frac{1}{2}(Q_1 + Q_3)$;

d) o intervalo interquartil $Q_3 - Q_1$;

e) o intervalo semi-interquartil dado por $\frac{1}{2}(Q_3 - Q_1)$.

Nota: o quartil médio é às vezes usado como medida de tendência central em lugar da média. O intervalo interquartil e o intervalo semi-interquartil são usados como medidas de variabilidade.

14) Para comparar a precisão de dois micrômetros, um técnico estuda medidas tomadas com ambos os aparelhos. Com um mediu repetidamente o diâmetro de uma pequena esfera de rolamento; as mensurações acusam média de 5,32 mm e d.p. de 0,019 mm. Com outro mediu o comprimento natural de uma mola, tendo as mensurações acusado média

de 6,4 cm e d.p. de 0,03 cm. Qual dos dois aparelhos é relativamente mais preciso?

15) Uma pesquisa sobre consumo de gasolina deu os seguintes valores para a quilometragem percorrida por três marcas de carro (de mesma classe), em cinco testes com um tanque de 40 l:

Carro A	400	397	401	389	403
Carro B	403	401	390	378	395
Carro C	399	389	403	387	401

Qual a medida mais adequada para comparar o desempenho dos carros?

16) O que acontece com a mediana, a média e o desvio padrão de uma série de dados quando:

a) Cada observação é multiplicada por 2;

b) Soma-se 10 a cada observação;

c) Subtrai-se a média geral $\bar{x}$ de cada observação;

d) De cada observação subtrai-se $\bar{x}$ e divide-se pelo desvio padrão.

17) A tabela abaixo mostra o tempo gasto por empregados numa determinada operação em uma fábrica:

Tempo (min)	Freqüência
10 ├── 15	5
15 ├── 20	57
20 ├── 25	42
25 ├── 30	28
30 ├── 40	18

a) Esboce o histograma correspondente;

b) Calcule a média, o desvio padrão;

c) O presidente decide separar os funcionários mais rápidos (com tempo inferior a um desvio padrão abaixo da média) para receberem promoção. Qual a porcentagem desses funcionários?

(tempo de 150 funcionários)

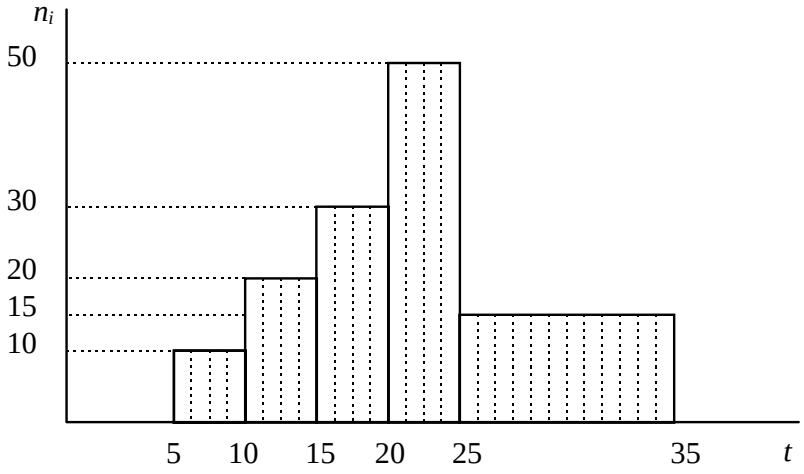
18) Em testes de resistência à tração aplicados a um tipo de aço, obteve-se um coeficiente de variação de 8%. O escore padronizado obtido para uma resistência de 28 kg/mm^2 foi de -0,4.

a) Qual a resistência média?

b) Qual o desvio padrão?

c) O que significa o resultado negativo do escore padronizado?

19) O histograma abaixo descreve o tempo (em min) gasto por funcionários de uma fábrica em uma determinada operação:



a) Construa a distribuição para o histograma;

b) Calcule a média e o desvio padrão;

c) O presidente decide demitir os funcionários mais lentos (com tempo superior a meio desvio padrão acima da média). Qual a porcentagem desses funcionários?

1.4 - Análise Bidimensional

Aqui, o interesse reside em analisar o comportamento conjunto de duas variáveis e de um modo geral, podemos representar tais variáveis utilizando uma tabela de dupla entrada, ou seja, uma *Distribuição Conjunta* das freqüências das variáveis.

Assim, para duas variáveis

$X = x_i$, com i = 1, 2, ...k e

$Y = y_j$ com j = 1, 2, ... l, temos:

X \ Y	1	2	$\cdots$	j	$\cdots$	L	Total
1	n_{11}	n_{12}	$\cdots$	n_{1j}	$\cdots$	n_{1L}	$n_{1\bullet}$
2	n_{21}	n_{22}	$\cdots$	n_{2j}	$\cdots$	n_{2L}	$n_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	n_{i1}	n_{i2}	$\cdots$	n_{ij}	$\cdots$	n_{iL}	$n_{i\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
k	n_{k1}	n_{k2}	$\cdots$	n_{kj}	$\cdots$	n_{kL}	$n_{k\bullet}$
Total	$n_{\bullet 1}$	$n_{\bullet 2}$	$\cdots$	$n_{\bullet j}$	$\cdots$	$n_{\bullet L}$	$n_{\bullet\bullet}$

Onde:

n_{ij} = número de elementos pertencentes ao i-ésimo nível da variável X e j-ésimo nível da variável Y.

$n_{i\bullet} = \sum_{j=1}^{L} n_{ij}$ = número de elementos do i-ésimo nível da variável X.

$n_{\bullet j} = \sum_{i=1}^{K} n_{ij}$ = número de elementos do j-ésimo nível da variável Y.

$n_{\bullet\bullet} = \sum_{i=1}^{K} \sum_{j=1}^{L} n_{ij} = \sum_{i=1}^{K} n_{i\bullet} = \sum_{j=1}^{L} n_{\bullet j}$ = nº total de elementos.

De modo análogo, podemos definir as freqüências relativas (proporções) do seguinte modo:

$$f_{ij} = \frac{n_{ij}}{n_{\bullet\bullet}}, \quad f_{i\bullet} = \frac{n_{i\bullet}}{n_{\bullet\bullet}} \quad \text{e} \quad f_{\bullet j} = \frac{n_{\bullet j}}{n_{\bullet\bullet}}.$$

Uma outra freqüência que pode ser construída é aquela para a qual se mantém fixa uma linha. Para a i-ésima linha fixa, temos então:

$f_{j/i} = \dfrac{n_{ij}}{n_{i\bullet}}$ (leia-se freqüência de *j*, dado *i*), que é a proporção dentre os indivíduos do i-ésimo nível *X* que possuem a j-ésima característica de *Y*.

Analogamente, temos:

$$f_{i/j} = \frac{n_{ij}}{n_{\bullet j}}$$

Ex.: Sejam as seguintes variáveis:

$X \equiv$ Curso escolhido

$Y \equiv$ Alunos segundo sexo.

Xi ***Yj***	**Masculino**	**Feminino**	**Total**
Economia	85	35	120
Administração	55	25	80
Total	140	60	200

(Tabela A)

Distribuição conjunta das freqüências das variáveis *X* e *Y*.

A linha e coluna de totais são chamados de **marginais**.

Uma tabela (dentre outras possíveis) poderia ser escrita a partir da Tabela A:

Xi ***Yj***	**Masculino**	**Feminino**	**Total**
Economia	61%	58%	60%
Administração	39%	42%	40%
Total	100%	100%	100%

(Tabela B)

Distribuição conjunta das proporções em relação aos totais de cada coluna.

Nesse caso, temos $\left(f_{i/j} = \dfrac{n_{ij}}{n_{\bullet j}} \right)$

1.4.1 - Independência de variáveis

Um dos principais objetivos de uma distribuição conjunta é descrever a

associabilidade existente entre as variáveis, isto é, queremos conhecer o grau de dependência entre elas, de modo que possamos prever melhor o resultado de uma delas, quando conhecemos a realização da outra.

Ex.: A partir de uma renda familiar podemos estimar a classe social de uma pessoa, pois sabemos da existência de dependência entre essas duas variáveis.

Usando uma distribuição conjunta, como a fornecida pela Tabela A, verificamos que fica relativamente difícil tirar alguma conclusão, devido à diferença entre os totais marginais. Por outro lado, usando a Tabela B, podemos observar que independentemente do sexo, 60% das pessoas preferem economia e 40%, administração (observe na coluna de total). Não havendo dependência entre as variáveis, esperaríamos estas mesmas proporções para cada sexo. De fato, vemos que as proporções do sexo masculino (61% e 39%) e do feminino (58% e 42%) são próximas das marginais (60% e 40%). Esses resultados parecem indicar não haver dependência entre as duas variáveis. *X* e *Y* são portanto variáveis independentes.

Observe agora a dependência quando os cursos são Engenharia e Ciências Sociais.

Y ***X***	**Masculino**	**Feminino**	**Total**
Engenharia	100 (83%)	20 (17%)	120 (100%)
C. Sociais	20 (25%)	60 (75%)	80 (100%)
Total	120 (60%)	80 (40%)	200 (100%)

(Tabela C)
Distribuição conjunta das proporções em relação aos totais de cada linha.

Nesse caso, temos $\left(f_{j/i} = \dfrac{n_{ij}}{n_{i\bullet}} \right)$

1.4.2 - Coeficiente de Contingência (*C*)

Existem algumas medidas que quantificam a dependência entre variáveis nominais. Uma delas é o chamado coeficiente de contingência, *C*, devido a Pearson. Tal coeficiente varia entre 0 e 1, e a proximidade de 0 indica total independência.

Ex.: O cálculo de *C* pode ser entendido por meio do seguinte exemplo, onde queremos verificar se a criação de determinado tipo de cooperativa está associada com algum fator regional. A tabela a seguir, sintetiza os dados coletados (Dados observados, o_{ij}).

Estado	Tipo de Cooperativa				
	Consumidor	**Produtor**	**Escola**	**Outros**	**Total**
São Paulo	214 (33%)	237 (37%)	78 (12%)	119 (18%)	648 (100%)
Paraná	51 (17%)	102 (34%)	126 (42%)	22 (7%)	301 (100%)
Rio G.Sul	111 (18%)	304 (51%)	139 (23%)	48 (8%)	602 (100%)
Total	376 (24%)	643 (42%)	343 (22%)	189 (12%)	1551 (100%)

(Tabela D) Distribuição con-junta das propor-ções em relação aos totais de cada linha.

$$\left(f_{j/i} = \frac{n_{ij}}{n_{i\bullet}} \right)$$

Tal tabela mostra a existência de uma certa dependência entre as variáveis. Caso houvesse independência, esperaríamos que em cada estado tivéssemos 24% de cooperativas de consumidores, 42% de produtores, 22% de escolas e 12% de outros.

Uma nova tabela, tabela E, poderia ser agora construída, onde teríamos todos os valores esperados (e_{ij}) caso as variáveis fossem independentes. É ela dada a seguir:

Estado	Tipo de Cooperativa				
	Consumidor	**Produtor**	**Escola**	**Outros**	**Total**
São Paulo	156 (24%)	272 (42%)	142 (22%)	78 (12%)	648 (100%)
Paraná	72 (24%)	127 (42%)	66 (22%)	36 (12%)	301 (100%)
Rio G.Sul	144 (24%)	254 (42%)	132 (22%)	72 (12%)	602 (100%)
Total	376 (24%)	643 (42%)	343 (22%)	189 (12%)	1551 (100%)

(Tabela E) Distribuição conjun-ta dos valores es-perados em relação aos totais das linhas.

$(n_{ij} = f_{i\bullet}\, n_{\bullet j})$

Comparando as Tabelas D e E, podemos verificar as discrepâncias existentes entre os valores observados (o_{ij}) e os esperados (e_{ij}). Na próxima tabela, (F), evidenciamos essa discrepância :

Estado	Tipo de Cooperativa			
	Consumidor	**Produtor**	**Escola**	**Outros**
São Paulo	58	-35	-64	41
Paraná	-21	-25	60	-14
Rio G. Sul	-33	50	7	-24

(Tabela F)

Desvios entre dados observados e esperados.

$(n_{ij} = o_{ij} - e_{ij})$

Observe que:

i) A soma em cada linha é nula (*soma dos resíduos*)

ii) Em Escola - São Paulo, temos: $o_{ij} = 78$

$e_{ij} = 142$

$o_{ij} - e_{ij} = -64$

Em Escola - Paraná, temos: $o_{ij} = 126$

$e_i = 66$

$o_{ij} - e_{ij} = 60$

É fácil ver que em relação aos valores esperados (e_{ij}), o desvio em Escola - Paraná (60) é maior do que o desvio em Escola - São Paulo (-64).Uma maneira de observar melhor esses desvios relativos a e_{ij} é considerar a seguinte medida:

$$\frac{\left(o_{ij} - e_{ij}\right)^2}{e_{ij}},$$

que para Escola - São Paulo fica:

$$\frac{(-64)^2}{142} = 28{,}84 \quad \text{e}$$

para Escola - Paraná fica:

$$\frac{(60)^2}{66} = 54{,}54$$

A tabela a seguir indica todos os valores desses desvios relativos a e_{ij}:

Estado	Tipo de Cooperativa			
	Consumidor	**Produtor**	**Escola**	**Outros**
São Paulo	21,56	4,50	28,84	21,55
Paraná	6,12	4,92	54,54	5,44
Rio G. Sul	7,56	9,84	0,37	8,00

(Tabela G)

Desvios entre dados observados e esperados.

$$\left(n_{ij} = \frac{\left(o_{ij} - e_{ij}\right)^2}{e_{ij}} \right)$$

Uma medida de afastamento global pode ser dada pela soma de todos os valores da tabela G. Essa medida possui distribuição de χ^2 (qui-quadrado) e, no nosso exemplo temos:

$$\chi^2 = \sum_i \sum_j \frac{\left(o_{ij} - e_{ij}\right)^2}{e_{ij}} = 21{,}56 + 6{,}12 + \cdots + 8{,}00 = 173{,}24$$

Quanto maior for o valor de χ^2, maior será o grau de associação existente entre as duas variáveis. Pearson propôs o chamado coeficiente de contingência *C* definido por

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

onde $0 \le C \le 1,$ sendo $C = 0$ quando $\chi^2 = 0$.

Uma melhor aproximação de *C* para avaliar a independência de variáveis é dado por:

$$C^* = \frac{C}{\sqrt{\frac{t-1}{t}}}$$

Onde *t* = mínimo entre o número de colunas e o número de linhas da tabela.

Assim, para o nosso exemplo, temos:

$$C = \sqrt{\frac{173{,}24}{173{,}24 + 1551}} = 0{,}32 \quad \text{e}$$

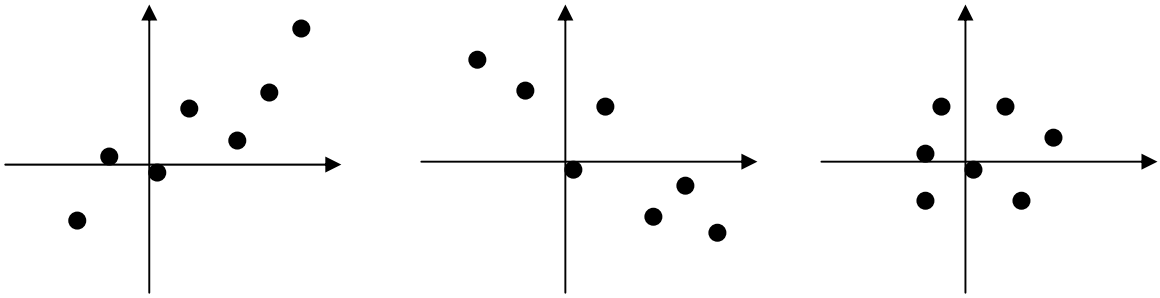
$$C^* = \frac{0{,}32}{\sqrt{\frac{2}{3}}} = 0{,}40$$

A quantificação dessa independência poderá ser vista futuramente em Testes de Hipóteses

1.4.3 - Coeficiente de correlação (r)

Um procedimento útil para se verificar a associação entre variáveis quantitativas é o ***gráfico de dispersão***, que nada mais é do que a representação dos pares de valores (nuvem de pontos) em um sistema cartesiano.

Admitamos um gráfico de dispersão como os dados a seguir, onde, através de uma transformação conveniente a origem foi colocada no centro da nuvem de pontos:



(A) (B) (C)

Em (A) os dados possuem uma certa associação linear direta (ou positiva). Observando-os, notamos que a grande maioria dos pontos está situada no primeiro e terceiro quadrantes, onde as coordenadas têm o mesmo sinal e portanto o produto será sempre positivo. Se somarmos o produto das coordenadas de todos os pontos, o resultado será um número positivo.

Em (B), usando o mesmo procedimento, o resultado será negativo. Em (C), o resultado será nulo.

Obs.: A soma dos produtos das coordenadas depende, e muito, do número de pontos e torna-se difícil comparar essa medida para dois conjuntos diferentes de pontos. Isto é atenuado usando-se "a média da soma dos produtos das coordenadas".

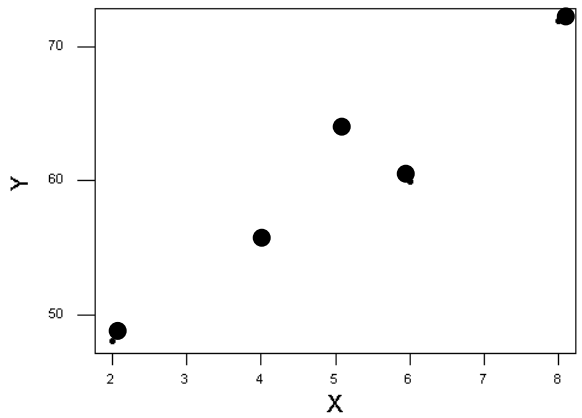
Ex.: Suponha que o nosso desejo seja o de quantificar a associabilidade entre duas variáveis relacionadas a cinco agentes de uma seguradora. Assim, temos:

$X \equiv$ Anos de experiência do agente.

$Y \equiv$ Número de clientes do agente.

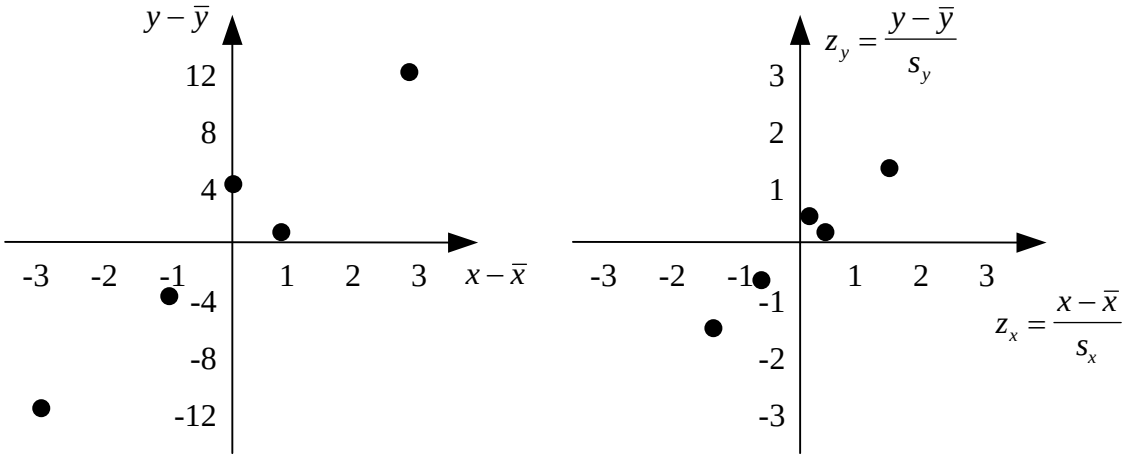
A tabela e o gráfico de dispersão seguintes mostram os dados obtidos:

Agente	*X*	*Y*
A	2	48
B	4	56
C	5	64
D	6	60
E	8	72
Total	25	300



O primeiro problema que devemos resolver é a mudança da origem do sistema para o centro da nuvem pontos. O ponto mais conveniente é aquele formado pelas duas médias $(\bar{x},\bar{y})$. A transformação dos eixos *x* em $x-\bar{x}$ e *y* em $y-\bar{y}$ resolve esse problema.

Observando esses valores centrados, verificamos que ainda existe um problema quanto à escala. A variabilidade em *Y* é muito maior do que em *X*, e o produto das coordenadas ficará muito mais afetado pelos resultados de *Y* do que de *X*. Para corrigir isso, podemos reduzir as duas variáveis a uma mesma escala. Isso é obtido dividindo-se os desvios pelos respectivos desvios padrões. Observe isso na figura a seguir:



Assim, podemos montar a seguinte tabela:

Agente	x	y	$x-\bar{x}$	$y-\bar{y}$	$\frac{x-\bar{x}}{s_x}=z_x$	$\frac{y-\bar{y}}{s_y}=z_y$	$z_x \cdot z_y$
A	2	48	-3	-12	-1.5	-1.5	2,25
B	4	56	-1	-4	-0.5	-0.5	0,25
C	5	64	0	4	0	0.5	0
D	6	60	1	0	0.5	0	0
E	8	72	3	12	1.5	1.5	2,25
Total	25	300	0	0	0	0	4,75

$\bar{x}=5$ $\quad S_x=2$ $\quad \bar{y}=60$ $\quad S_y=8$

Para completar a definição da medida descrita no início desse tópico, basta calcular a média dos produtos das coordenadas reduzidas, isto é a correlação linear entre *X* e *Y*, dada por:

$$r = \text{Correlação } (X,Y) = \frac{4{,}75}{5} = 0{,}95 = 95\%$$

Portanto, para este exemplo, o grau de associabilidade está quantificado em 95%.

Da discussão feita até aqui, podemos definir o coeficiente de correlação entre duas variáveis *X* e *Y*, dados *n* pares ordenados $(x_1,y_1),\ (x_2,y_2)\cdots(x_n,y_n)$ como sendo:

$$r = \text{Corr } (X,Y) = \frac{1}{n}\sum_{i=1}^{n} z_{x_i} z_{y_i} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right)$$

ou seja, a média dos produtos dos valores reduzidos da variável.

Costuma-se usar outras fórmulas para *r*:

$$r = \frac{1}{n}\frac{\sum (x_i-\bar{x})(y_i-\bar{y})}{s_x \cdot s_y} = \frac{\text{Covariância } (X,Y)}{s_x \cdot s_y}$$, ou ainda

$$r = \frac{\sum x_i y_i - n\bar{x}\,\bar{y}}{\sqrt{\left(\sum x_i^2 - n\bar{x}^2\right)\left(\sum y_i^2 - n\bar{y}^2\right)}}$$

Obs.*:* Pode-se demonstrar que $-1 \le r \le 1$.

Atenção*:* A correlação apresentada aqui é linear. Existem outros tipos de correlação!

Exercícios III

01) No estudo de uma certa comunidade verificou-se que:

I. A proporção de indivíduos solteiros é de 0,4;

II. A proporção de indivíduos que recebem até 10 salários mínimos é de 0,2;

III. A proporção de indivíduos que recebem até 20 salários mínimos é de 0,7;

IV. A proporção de indivíduos casados entre os que recebem mais de 20 salários mínimos é de 0,7;

V. A proporção de indivíduos que recebem até 10 salários mínimos entre os solteiros é de 0,3.

a) Construa a distribuição conjunta das variáveis estado civil e faixa salarial e as respectivas distribuições marginais;

b) Você diria que existe relação entre as duas variáveis consideradas?

02) Uma amostra de 200 habitantes de uma cidade foi colhida para analisar a atitude frente a um certo projeto governamental. O resultado foi o seguinte:

	Local de Residência			
Opinião	Urbano	Suburbano	Rural	Total
A favor	30	35	35	100
Contra	60	25	15	100
Total	90	60	50	200

a) Calcule as proporções em relação ao total das colunas

b) Você diria que a opinião independe do local de residência?

c) Encontre uma medida de dependência entre as variações

03) Numa amostra de 5 operários de uma dada empresa, foram observadas duas variáveis; sendo X os anos de experiência num dado cargo e Y o tempo, em minutos, gasto na execução de uma certa tarefa relacionada com esse cargo.

X	1	2	4	4	5
Y	7	8	3	2	2

$\sum x = 16$ $\quad$ $\sum x^2 = 62$

$\sum y = 22$ $\quad$ $\sum y^2 = 130$

$\sum x.y = 53$

Usando um critério estatístico, você diria que a variável *X* pode ser usada para explicar a variação de *Y*? Justifique.

04) Muitas vezes, a determinação da capacidade de produção instalada para certo tipo de indústria em certas regiões é um processo difícil e custoso. Como alternativa, pode-se estimar a capacidade de produção através da escolha de uma outra variável de medida mais fácil e que esteja linearmente relacionada com ela.

Suponha que foram observados os valores para as variáveis: capacidade de produção instalada, potência instalada e área construída. Com base num critério estatístico, qual das variáveis você escolheria para estimar a capacidade de produção instalada?

X capacidade de produção instalada (ton)	4	5	4	5	8	9	10	11	12	12
Y potência instalada (1.000 kW)	1	1	2	3	3	5	5	6	6	6
Z área construída (100 m)	7	7	10	10	11	9	12	10	11	14

$\sum x = 80$; $\sum y = 38$; $\sum z = 100$; $\sum x^2 = 736$; $\sum y^2 = 182$; $\sum z^2 = 1048$;
$\sum x.y = 361$; $\sum x.z = 848$; $\sum y.z = 411$.

Capítulo 2 - Probabilidades e Variáveis Aleatórias

2.1 - Introdução à Probabilidade

A *probabilidade* em termos práticos é uma medida que exprime a incerteza presente em situações onde os resultados são variáveis.

Um *experimento aleatório* (ou experimento não-determinístico) é o processo de coleta de dados relativos a um fenômeno que acusa variabilidade em seus resultados.

Um *Espaço Amostral* (E) é o conjunto de todos os resultados de um experimento aleatório.

Um *Evento* é um subconjunto de um espaço amostral. Os principais tipos de eventos são (considere $A \subset E$ e $B \subset E$):

- Evento *Simples*: O número de elementos do evento é igual a 1, $n(A) = 1$.
- Evento *Impossível*: $A = \Phi$, $n(A)=0$
- Evento *Certo*: $A = E$
- Evento *Complementar*: $\overline{A} = E - A = C_E^A$
- Evento *Interseção*: $A \cap B$
- Evento *União*: $A \cup B$
- Eventos *Mutuamente Excludentes* (ou *Exclusivos*): $A \cap B = \Phi$

Obs.*:* Devido ao cálculo de probabilidade envolver operações entre conjuntos, o anexo I, ao final desse capítulo, descreve sucintamente a *Teoria Axiomática da Álgebra de Boole* aplicada a conjuntos.

A probabilidade, em sua definição clássica pode ser assim definida:

Dado um espaço amostral E com $n(E)$ elementos e um evento A de E com $n(A)$ elementos, a probabilidade do evento A é o número $P(A)$ tal que: $P(A) = \dfrac{n(A)}{n(E)}$

Na prática: $P(A) = \dfrac{\text{nº casos favoráveis}}{\text{nº casos possíveis}}$

Obs.: A feqüência relativa $f_i = n_i / n$ comumente é associada à probabilidade. Essa associação nos leva a diversas distribuições de probabilidade.

Atualmente, principalmente devido a Kolmogorov, um grande matemático russo, o cálculo de probabilidades constitui uma Teoria Matemática com suas definições, axiomas, teoremas e demonstrações. As regras básicas que constituem essa teoria são vistas a seguir.

Consideremos um espaço amostral E com A e B eventos de E, com os seguintes axiomas:

A_1 : $0 \leq P(A) \leq 1$

A_2 : $P(E) = 1$

A_3: Para qualquer $A_1, A_2, \ldots A_i$, Exclusivos $\left(A_i \cap A_j = \Phi \ \ \forall \ i \neq j\right)$, temos que

$$P\left(\sum \bigcup_{i=1}^{k} A_i\right) = \sum_{i=1}^{k} P\left(A_i\right)$$

Alguns teoremas dessa teoria são:

T_1: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

T_2: $P\left(\overline{A}\right) = 1 - P(A)$

T_3: Se $A \subset B$ então $P(A) \leq P(B)$

T_4: $P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap C) - P(A \cap B)$
$- P(B \cap C) + P(A \cap B \cap C)$

Obs.: Demonstre os teoremas T_1 a T_4..

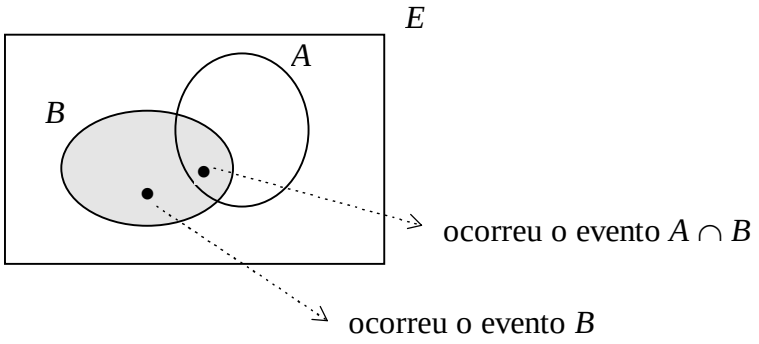
Obs.: Uma breve revisão de *Análise Combinatória* está descrita no anexo II. Este conhecimento é muito útil na determinação do número de elementos de um evento qualquer.

2.2 - Probabilidade Condicional e Independência de Eventos

Para dois eventos quaisquer A e B, sendo $P(B) > 0$ definimos a *probabilidade condicional de A, dado B, P(A | B)*, como sendo

$$\boxed{P(A \mid B) = \frac{P(A \cap B)}{P(B)}}$$

Obs.: Entenda tal definição a partir do seguinte gráfico:



$$P(A|B) = \frac{n(A \cap B)}{n(B)} = \frac{n(A \cap B)/n(E)}{n(B)/n(E)} = \frac{P(A \cap B)}{P(B)}$$

A partir da definição de probabilidade condicional obtemos a chamada regra do produto de probabilidades:

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

de larga utilização.

Dois eventos A e B são denominados *independentes* se a ocorrência de um deles não interfere na ocorrência do outro. Neste caso:

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

Obs.*:* É freqüente, aqui, adotarmos $P(A \cap B) = P(AB)$

Se E_1, E_2, ... E_n são eventos de um espaço amostral E, então tais eventos são *mutuamente independentes* se:

$$P(E_1 \cap E_2 \cap E_3 \ldots \cap E_n) = P(E_1) \,.\, P(E_2) \,.\, P(E_3) \ldots P(E_n)$$

Ex.: A fundação de um alto edifício pode ceder devido a falhas do tipo "Bearing"ou do tipo "Settlement". Seja B e S os respectivos modos de falha. Se $P(B)$ = 0.1%, $P(S)$ = 0.8% e $P(B|S)$ = 10% (Probabilidade de falha B devido a falha S).

A probabilidade de falha da fundação pode ser dada por:

$$
\begin{aligned}
P\,(\text{Falha}) &= P\,(B \cup S) = P\,(B) + P\,(S) - P\,(B \cap S) \\
&= P\,(B) + P\,(S) - P\,(B \mid S)\,.\,P(S) \\
&= 0.001 + 0.008 - (0.1)\,(0\ 008) \\
&= 0.082 = 8{,}2\%
\end{aligned}
$$

A probabilidade de que o edifício tenha falha S mas não tenha falha B é:

$$
\begin{aligned}
P\left(S \cap \overline{B}\right) &= P\left(\overline{B} \cap S\right) = P(S) \cdot P\left(\overline{B} | S\right) \\
&= P(S)\left[1 - P\left(B | S\right)\right] \\
&= 0.008\,[1 - 0.1] \\
&= 0.0072 = 0.72\%
\end{aligned}
$$

Em um grande número de aplicações, teremos que adotar a hipótese de independência de dois eventos A e B, e depois empregar essa suposição para calcular $P(A \cap B)$ como igual a $P(A)\,.\,P(B)$. Geralmente, condições físicas sob as quais o experimento seja realizado tornarão possível decidir se tal suposição será justificada ou aproximadamente justificada.

Ex.: Consideremos um grande lote de peças, digamos 10.000. Admitamos que 10% dessas peças sejam defeituosas e 90% perfeitas. Duas peças são extraídas. Qual a probabilidade de que ambas sejam perfeitas?

Definamos os eventos A e B, assim

A = { A primeira peça é perfeita}

B = { A segunda peça é perfeita}

Se admitirmos que a primeira peça seja reposta, antes que a segunda seja escolhida, então os eventos A e B podem ser considerados independentes e, portanto,

$P(A \cap B) = P(A)\,.\,P(B) = (0.9)\,(0.9) = 0.81.$

Na prática, contudo, a segunda peça é escolhida sem a reposição da primeira peça; neste caso,

$$P(A \cap B) \quad = \quad P(A) \,.\, P(B \mid A) = \ (0.9) \,.\, \frac{8999}{9999}$$

que é aproximadamente igual 0.81.

Assim, muito embora A e B não sejam independentes no segundo caso, a hipótese de independência (que simplifica consideravelmente os cálculos) acarreta apenas um erro desprezível. Se existissem somente poucas peças no lote, a hipótese de independência poderia acarretar um erro grande.

2.2.1 - Confiabilidade

A *teoria da confiabilidade* estuda sistemas e seus componentes, como, por exemplo, sistemas mecânicos ou eletrônicos (um automóvel ou um computador). O objetivo desta teoria é estudar relações entre o funcionamento dos componentes e do sistema. A figura a seguir ilustra um sistema composto de dois componentes ligados em série.

——①——②——

O sistema funciona se os componentes 1 e 2 funcionam simultaneamente. Se um dos componentes não funciona, o sistema também não funciona. Supondo que os componentes funcionem independentemente, e se p_i é a probabilidade do componente i (i=1,2) funcionar, então a probabilidade do sistema funcionar é $p_1 p_2$.

Chamando:

E : o sistema funciona

A_i : o componente i funciona, i = 1,2

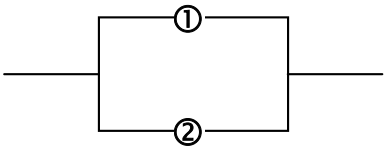
então,

$P(E) = P(A_1 \cap A_2) = P(A_1)\, P(A_2) = p_1\, p_2$

Cada p_i é chamada a *confiabilidade do componente* ***i*** e $P(E) = h\,(p_1, p_2) = p_1\, p_2$ é

chamada a *confiabilidade do sistema*.

Se os componentes 1 e 2 estiverem ligados em paralelo, como na figura a seguir, então o sistema funciona se pelo menos um dos dois componentes funciona. Ou seja,

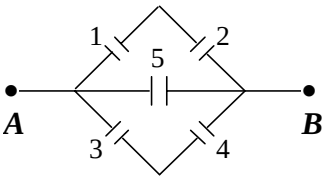


$P(E) = P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2)$

e a confiabilidade do sistema é

$h(p_1, p_2) = p_1 + p_2 - p_1 p_2$

Ex.: Suponhamos que para o circuito da figura a seguir, a probabilidade de que cada relé esteja fechado seja p, e que todos os relés funcionem independentemente. Qual será a probabilidade de que o circuito permita a passagem de corrente entre A e B?



Empregando a mesma definição dos circuitos anteriores:

$$\begin{aligned} P(E) &= P(A_1 A_2) + P(A_5) + P(A_3 A_4) - P(A_1 A_2 A_5) - P(A_1 A_2 A_3 A_4) - P(A_5 A_3 A_4) \\ &\quad + P(A_1 A_2 A_3 A_4 A_5) \\ &= p^2 + p + p^2 - p^3 - p^4 - p^3 + p^5 \\ &= p + 2p^2 - 2p^3 - p^4 + p^5 \end{aligned}$$

Uma bastante comum, mas errônea, resolução de um problema, pode ser vista a seguir.

Ex.: Admita-se que dentre seis parafusos,.dois sejam menores do que um comprimento especificado. Se dois dos parafusos forem escolhidos ao acaso, qual será a probabilidade p de que os dois parafusos mais curtos sejam retirados?

Adote que, A_i : o i-ésimo parafuso é curto (i = 1,2)

A solução correta é dada por:

$$p = P\,(A_1 \cap A_2) = \mathrm{P}(A_2 \mid A_1)\ .\ P(A_1) = \frac{1}{5} \cdot \frac{2}{6} = \frac{1}{15}$$

ou ainda, usando o conceito de probabilidade;

$$p = \frac{\text{n° casos favoraveis}}{\text{n° casos possiveis}} = \frac{1}{\binom{6}{2}} = \frac{1}{15}$$

A solução comum, mas incorreta, é obtida escrevendo-se

$$p = P\left(A_1 \cap A_2\right) = P\left(A_2\right) \cdot P\left(A_1\right) = \frac{1}{5} \cdot \frac{2}{6} = \frac{1}{15}.$$

Naturalmente, o importante é que, muito embora a resposta esteja numericamente correta, a identificação de 1/5 com $P(A_2)$ é incorreta; 1/5 representa $P(A_2 \mid A_1)$.

Para calcular $P(A_2)$ corretamente, escrevemos

$$\begin{aligned} P(\mathrm{A}_2) &= P(A_2 \cap A_1) + P(A_2 \cap \overline{A}_1) \\ &= P(A_2 | A_1) \cdot P\left(A_1\right) + P\left(A_2 | \overline{A}_1\right) \cdot P\left(\overline{A}_1\right) \\ &= \frac{1}{5} \cdot \frac{2}{6} + \frac{2}{5} \cdot \frac{4}{6} = \frac{1}{3} \end{aligned}$$

<u>2.2.2 - Teorema de Bayes</u>

Uma das relações mais importantes envolvendo probabilidades condicionais é dada pelo teorema de Bayes que expressa uma probabilidade condicional em termos de outras probabilidades condicionais.

Se E_1, E_2, ... E_n são eventos mutuamente excludentes ($E_i \cap E_j = \Phi$ para todo $i \neq j$) de um espaço amostral *Exaustivo E*, $\left(\bigcup_{i=1}^{n} E_i = E\right)$, então cada E_i é denominado de *partição* de *E*.

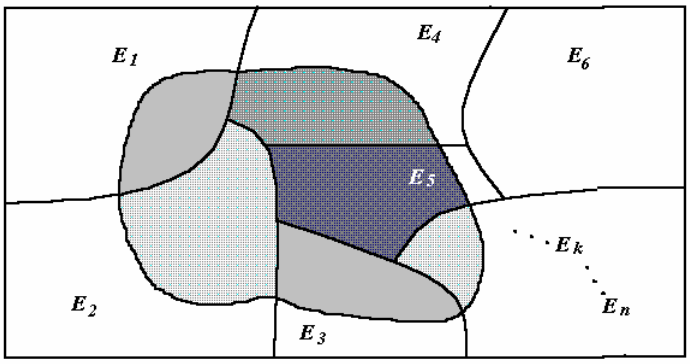
Se *A* é um evento arbitrário de *E* com $p(A) > 0$, então a probabilidade de ocorrência de uma das partições E_k, dado que ocorreu o evento *A*, pode ser expressa por:

$$P(E_k | A) = \frac{P\left(E_k \cap A\right)}{P(A)} = \frac{P\left(E_K \cap A\right)}{\sum_{i=1}^{n} P\left(A \cap E_i\right)}$$

ou, mais simplesmente (!?), por:

$$P(E_k \mid A) = \frac{P(E_k)\cdot P(A|\, E_K)}{\sum_{i=1}^{n} P(E_i)\cdot P(A|\, E_i)} \text{ , e } i = 1, 2, ...n$$

Sendo que $P(A|\, E_K) = \dfrac{P(A\cap E_K)\cdot}{P(E_k)}$ e $P(A|\, E_i) = \dfrac{\cdot\, P(A\cap E_i)}{P(E_i)}$



Ex.: Uma companhia produz circuitos integrados em três fábricas, I, II e III. A fábrica I produz 30% dos circuitos, enquanto II e III produzem 45% e 25% respectivamente. As probabilidades de que um circuito integrado produzido por estas fábricas não funcione são 0.01, 0.02 e 0.04, respectivamente. Escolhido um circuito da produção com defeito conjunta das três fábricas, qual a probabilidade de ele ter sido fabricado em I?

A tabela a seguir sintetiza os dados, *F* = I, II, III

Fábrica	**Produção**	***d* (defeituoso)**	***P*(*F*).*P*(*d*|*F*)**
I	30%	0,01	0,003
II	45%	0,02	0,009
III	25%	0,04	0,01
			0,022

Assim:

$$P(I|d) = \frac{P(I)\cdot P(d|I)}{\sum_{F=\mathrm{I}}^{\mathrm{III}} P(F)\cdot P(d|F)} = \frac{0.003}{0.022} = 13.63\%$$

Exercícios IV

1 - Dão-se as seguintes probabilidades para A e B: $P(A) = 1/2$, $P(B) = 1/4$, $P(A \mid B) = 1/3$.

a) Utilizando as leis de probabilidade, calcular $P(\overline{A})$, $P(AB)$, $P(A \cup B)$.

b) Usar diagrama para calcular $P(A\overline{B})$ e $P(\overline{A} \cup \overline{B})$.

2 - Se A, B e C são eventos arbitrários, exprima em notação de conjuntos os seguintes eventos:

a) ocorrem apenas 2 eventos;

b) ocorrem não mais de 2 eventos;

c) ocorrem A e B mas não C.

d) ocorre ao menos um

e) não ocorre nenhum;

f) ocorre apenas um.

3 - Numa amostra de 40 indivíduos, 10 acusam pressão alta. Estime a probabilidade de outro indivíduo escolhido ao acaso, no mesmo grupo do qual foi extraída a amostra, também ter pressão elevada.

4 - Jogam-se dois dados. Qual a probabilidade de o produto dos números das faces superiores estar entre 12 e 15 inclusive?

5 - Uma moeda é viciada de tal forma que a probabilidade de ocorrer "cara" é duas vezes a de ocorrer "coroa". Jogada três vezes a moeda, qual a probabilidade de aparecerem exatamente duas "coroas"?

6 - Admitindo que a probabilidade de uma criança ser menino (H) seja 0,51, determinar a probabilidade de uma família de seis filhos ter:

a) ao menos um H;

b) ao menos uma M.

7 - Escolhido ao acaso um ponto no disco unitário, determinar a probabilidade de o ponto estar no setor $[0, \pi/6]$.

8 - A experiência indica que 15% dos que reservam mesa em um restaurante nunca aparecem. Se o restaurante têm 60 mesas e aceita 62 reservas, qual a probabilidade de poder acomodar todos os que comparecerem?

9 - Um sistema tem dois componentes que operam independentemente. Suponhamos que as probabilidades de falha dos componentes 1 e 2 sejam 0,1 e 0,2, respectivamente. Determinar a probabilidade de o sistema funcionar nos dois casos seguintes:

a) os componentes são ligados em série (isto é, ambos devem funcionar para que o sistema funcione);

b) os componentes são ligados em paralelo (basta um funcionar para que o sistema funcione)

10 - Após seleção preliminar, uma lista de jurados consiste em 10 homens e sete mulheres. Escolhidos dentre esses os cinco membros de um júri, a escolha recai somente sobre homens. O processo sugere discriminação contra as mulheres? Responda calculando a probabilidade de não haver nenhuma mulher no júri.

11 - Qual o número mínimo de jogadas de uma moeda necessário para assegurar uma probabilidade superior a 0,74 de obter ao menos uma cara ?

12 - De acordo com certa tábua de mortalidade, a probabilidade de José estar vivo daqui a 20 anos é 0,6, e a mesma probabilidade para João é 0,9. Determinar:

a) P (ambos estarem vivos daqui a 20 anos)

b) P (nenhum estar vivo daqui a 20 anos)

c) P (um estar vivo e outro estar morto daqui a 20 anos)

13 - Uma tábua de mortalidade acusa as seguintes taxas de mortalidade q_x (isto é, probabilidade de um indivíduo de idade x morrer antes de atingir a idade $x + 1$):

x	30	31	32	33	34	35
q	0,00213	0,00219	0,00225	0,00232	0,00240	0,00251

a) Dado um indivíduo de 30 anos, qual a probabilidade de ele atingir a idade de 31 anos?

b) Para o mesmo indivíduo, qual a probabilidade de morrer antes de completar 35 anos?

14 - Determinar a probabilidade de n pessoas ($n \leq 365$) fazerem aniversário em datas diferentes.

15 - As probabilidades de um estudante passar em Álgebra (A), em Literatura (L) e ambas (AL) são 0,75, 0,84 e 0,63 respectivamente. Qual a probabilidade de passar em Álgebra, sabendo-se que passou em Literatura?

16 - A caixa I tem duas bolas brancas e duas pretas; a caixa II tem duas bolas brancas e uma preta; a caixa III tem uma bola branca e três pretas.

a) Tira-se uma bola de cada caixa. Determinar a probabilidade de serem todas brancas.

b) Escolhe-se uma caixa ao acaso e tira-se uma bola. Qual a probabilidade de ser branca ?

c) Em (b), calcular a probabilidade de ter sido escolhida a caixa I, sabendo-se que a bola extraída é branca.

17 - Suponha um teste para câncer em que 95% dos que têm o mal reagem positivamente, enquanto que 3% dos que não têm o mal reagem positivamente. Suponhamos ainda que 2% dos internos de um hospital tenham câncer. Qual a probabilidade de um doente escolhido ao acaso, e que reaja positivamente ao teste, ter de fato o mal?

18 - Duas sacolas têm, cada uma, *m* moedas de R$ 5,00 e *n* moedas R$ 2,00. Extrai-se uma moeda de cada sacola.

a) Calcular a probabilidade de ambas serem de R$ 5,00.

b) Mostrar que essa probabilidade é maior do que a de extrair duas moedas de R$ 5,00 de uma única sacola com todas as moedas.

19 - Quando surgiu a LOTO, estabeleceu-se que o valor a ser cobrado por cartão apostado seria proporcional ao número de quinas presentes no cartão. Um cartão com seis dezenas tem $\binom{6}{5} = 6$ quinas. Um com sete dezenas tem $\binom{7}{5} = 21$ quinas. Como o número de dezenas permitido varia de 5 a 10, temos os seguintes resultados:

Nº de dezenas	5	6	7	8	9	10
Nº de quinas	1	6	21	56	126	252

Assim, um cartão com cinco dezenas custava Cr$ 1,00, enquanto um com 10 dezenas custava Cr$ 252,00. Com estes preços, tornava-se muito caro apostar em várias dezenas. O prêmio acumulou várias semanas! A dificuldade com o sistema adotado é que ele considera apenas o prêmio da quina, quando, na realidade, existem prêmios também para as quadras e ternos. Considerando-se estes outros prêmios, as relações entre os preços dos cartões são outras. Seu cálculo, entretanto, é bem mais complicado.

Em um volante com 100 dezenas, 00, 01, 02 ... 98, 99, o apostador pode marcar de 5 a 10 dezenas, concorrendo ao sorteio de uma quina, cinco quadras e 10 ternos. No menor jogo, cinco dezenas, ele paga (em fevereiro de 1986) Cr$ 1,20 e, pelo maior jogo, Cr$ 43,00.

Qual a probabilidade de acertar a quina com um cartão de cinco dezenas? de 10 dezenas?

20 - Com relação ao problema anterior, se estamos interessados apenas no prêmio pago ao acertador da quina, o que é mais vantajoso: um cartão com 10 dezenas ou 35 cartões com cinco dezenas cada?

21 - Uma determinada peça é manufaturada por três fábricas, I, II e III. Sabe-se que I produz o dobro que II, II e III produziram o mesmo número de peças. 2% das peças produzidas por I e II são defeituosas; em III tem-se 4% de defeituosas. Todas as peças produzidas são colocadas em um depósito e depois uma peça é extraída ao acaso. Qual a probabilidade de que esta peça seja defeituosa?

22 - Um sistema automático de alarme contra incêndio utiliza três células sensíveis ao calor que agem independentemente uma da outra. Cada célula entra em funcionamento com probabilidade 0,8 quando a temperatura atingir 60º C. Se pelo menos uma das células entrar em funcionamento o alarme soa. Calcular a probabilidade do alarme soar quando a temperatura atingir 60º C.

23 - Uma montagem eletromecânica é formada por dois subsistemas *A* e *B*. De procedimentos de ensaios anteriores, as seguintes probabilidades se admitem conhecidas:

$P(A \text{ falhe}) = 0.20$

$P(A \text{ e } B \text{ falhem}) = 0.15$

$P(B \text{ falhe sozinho}) = 0.25$

Calcule a probabilidade: $P(A \text{ falhe} \mid B \text{ falhe})$.

24 - Um certo tipo de motor elétrico falha, com probabilidade de 10% devido aos fatores: emperramento de mancais, queima de enrolamentos, desgaste das escovas. Suponha que o emperramento seja duas vezes mais provável do que a queima, e a queima, quatro vezes mais provável do que o desgaste. Qual a probabilidade de que a falha se dê por desgaste das escovas?

(Obs.: a probabilidade de que haja falha devido aos três fatores pode ser desprezada e os fatores são independentes)

25 - Uma companhia produz circuitos integrados em três fábricas, I, II e III. A fábrica I produz 40% dos circuitos, a fábrica II 15% e a fábrica III 45%. As probabilidades de que um circuito integrado produzido por estas fábricas não funcionem são: 0.01, 0,04 e 0.03, respectivamente. Escolhido um circuito da produção conjunta das três fábricas, qual a

probabilidade de o mesmo não funcionar?

26 - Um sistema de proteção atua devido a falhas em duas subestações A e B, que agem conjuntamente. As probabilidades de falha nessas duas subestações são $P(A) = 0.002$ e $P(B) = 0.008$. sabe-se também que a probabilidade de falha em A devido a falha em B é 0.2. Determine:

a) A probabilidade do sistema de proteção atuar;

b) A probabilidade de que o sistema atue devido a falha em A mas não falhe em B.

27 - Uma companhia produz cilindros em três fábricas I, II e III. A fábrica I produz 30% dos cilindros, a fábrica II 45% e a fábrica III 25%. As probabilidades de que um cilindro produzido por estas fábricas seja defeituoso são 0.01, 0.02 e 0.04, respectivamente. Escolhido um cilindro da produção conjunta das três fábricas, qual a probabilidade de o mesmo ser da fábrica I?

28 - Numa indústria de enlatados, as linhas de produção I, II e III respondem por 50%, 30% e 20% da produção, respectivamente. As proporções de latas com defeito de produção nas linhas I, II e III são 0.4%, 0.6% e 1.2%. Qual a probabilidade de uma lata defeituosa (descoberta ao final da inspeção do produto acabado) provir da linha I.

29 - Sabe-se que na fabricação de um certo artigo, defeitos de um tipo ocorrem com probabilidade 0.1 e defeitos de outro tipo com probabilidade 0.05. Qual será a probabilidade de que:

a) Um artigo não tenha defeito?

b) Um artigo seja defeituoso?

c) Um artigo tenha apenas um tipo de defeito, sabido que é defeituoso?

2.3 - Variáveis Aleatórias

2.3.1 - Variáveis Aleatórias Unidimensionais

Variável Aleatória (*va*) é uma função que associa números reais aos pontos de um espaço amostral. Ou seja, os resultados do experimento aleatório são associados numericamente. Geralmente usamos letras maiúsculas (X, Y, Z...) para designar as variáveis aleatórias, e minúsculas (x, y, z...) para indicar particulares valores dessas variáveis. O comportamento de uma variável aleatória é descrito por *sua distribuição de probabilidade*.

Ex.: No teste de dois componentes eletrônicos, podemos ter o seguinte espaço amostral.

$E = \{ PP, PN, NP, NN \}$

sendo P resultado positivo e N negativo.

Seja X o número de resultados positivos do teste. X é uma variável aleatória. Temos então:

Pontos de *E*	***PP***	***PN***	***NP***	***NN***
X	2	1	1	0

A distribuição de probabilidades de X é dada pela seguinte tabela:

X	**2**	**1**	**0**
$P(X)$	1/4	1/2	1/4

sendo $P(X)$ a sua probabilidade

As variáveis aleatórias podem ser discretas, contínuas ou mistas. No caso discreto a distribuição de probabilidade pode ser caracterizada por uma *função de probabilidade*, que indica diretamente as probabilidades associadas a cada valor. No caso contínuo temos a *função densidade de probabilidade* (*fdp*).

Tais distribuições possuem as seguintes propriedades:

	Caso discreto	Caso contínuo
a)	$f(x_i) \geq 0$	$f(x) \geq 0$
b)	$\sum_{i=1}^{n} f(x_i) = 1$	$\int_{-\infty}^{\infty} f(x)\,dx = 1$
c)	$P(X = x_i) = f(x_i)$	$P(a < X \leq b) = \int_a^b f(x)\,dx \qquad (b > a)$

Ex.: No teste de partida de válvulas eletrônicas a probabilidade de que haja um resultado positivo é P (+) = 3/4 e um resultado negativo P (-) = (1/4). Os testes prosseguem até que aconteça o primeiro resultado positivo. Temos então o seguinte espaço amostral:

E = { +, - +, - - +, - - - +, ...}

Definindo a variável aleatória X como:

X : número de testes do experimento. X = 1, 2, 3, ... temos a seguinte distribuição de probabilidade.

X	1	2	3 ...
$P(X)$	$\frac{3}{4}$	$\frac{3}{16}$	$\frac{3}{48}$ ...

$P(X)$ foi calculado usando a seguinte função de probabilidade (aceitando que uma válvula não influencie na outra):

$$P(X = x_n) = f(x_n) = (1/4)^{n-1}\,(3/4) \quad \text{com } n = 1, 2, 3, \ldots$$

observe que ocorre um resultado positivo precedido de n - 1 negativos.

As propriedades **a)** e **c)** são imediatas, **b)** pode ser vista assim:

$$\sum_{i=1}^{n} f(x_i) = \frac{3}{4} + \frac{3}{16} + \frac{3}{48} + \ldots = \frac{3}{4}\left(1 + \frac{1}{4} + \frac{1}{16} + \ldots\right)\frac{3}{4}.\frac{1}{1 - \frac{1}{4}} = 1$$

Obs.*:* Estamos empregando o resultado de que a série geométrica $1 + r + r^2 + \ldots$ converge para $1/(1-r)$ sempre que $|r| < 1$.

Ex.: Seja X a duração da vida (em horas) de um certo tipo de lâmpada. Admitindo que X seja uma variável aleatória contínua, suponha que a *fdp* f de X seja dada por

$$f(x) = a/x^3, \quad 1.500 \leq x \leq 2.500$$
$$= 0 \text{ , para quaisquer outros valores.}$$

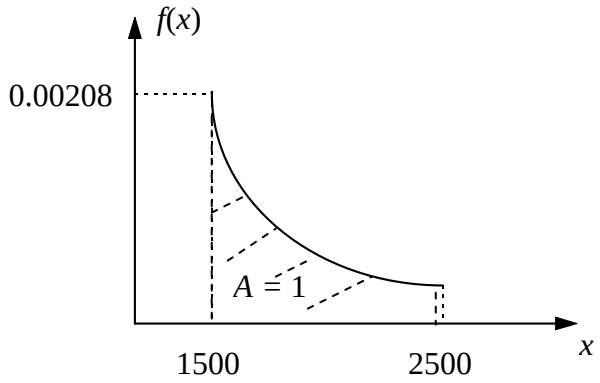
(Isto é, está se atribuindo probabilidade zero aos eventos { $x < 1500$} e {$x > 2500$}).

Para calcular a constante a, recorre-se à condição

$$\int_{-\infty}^{\infty} f(x)\, dx = 1, \text{ que neste caso se torna:}$$

$$\int_{1.500}^{2.500} \left(a/x^3\right) dx = 1.$$

Daí obtemos que a = 7.031.250 e o gráfico de f é dado a seguir:



Em capítulo posterior estudaremos melhor muitas variáveis aleatórias importantes, tanto discretas como contínuas.

2.3.2 - Função de Repartição ou de Distribuição Acumulada ($F(x)$)

Tal função é definida por:

$$F(x) = P\,(X \leq x),$$

servindo como alternativa para a caracterização da distribuição de probabilidade de qualquer tipo de variável aleatória.

No caso discreto temos que

$$F(x) = \sum_{x_i}^{x} P(x_i) \qquad \forall \quad x_i \le x$$

e no caso contínuo,

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

com $$f(x) = \frac{d}{dx} F(x).$$

São propriedades da função de repartição:

a) $0 \le F(x) \le 1$

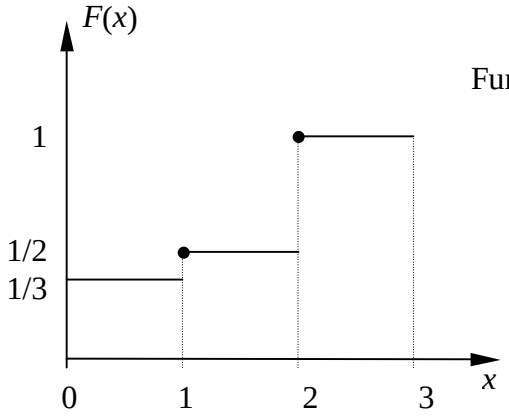
b) $F(-\infty) = 0$

c) $F(+\infty) = 1$

d) *F(x)* é sempre não-decrescente

e) $F(b) - F(a) = P(a < x \le b) \qquad (b > a)$

Ex.: Suponhamos que a variável aleatória *X* assuma os três valores 0,1 e 2, com probabilidade 1/3, 1/6 e 1/2, respectivamente. Então, temos o seguinte gráfico:



Função degrau ou escada

Observe que $P(X = x_i)$ é igual ao salto que a função *F(x)* dá em x_i; por exemplo, $P(X=1)= 1/6 = F(1) - F(1-)$. De um modo geral $P(X = x_i) = F(x_i) - F(x_i-)$

Obs.: $F(x_i -) = \lim_{y \to x-} F(y)$

Ex.: Suponhamos que *X* seja uma variável aleatória contínua com *fdp*

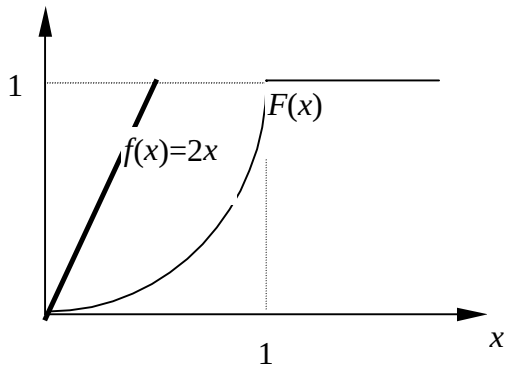
$f(x) = 2x,\ 0 < x < 1$

$= 0$, para quaisquer outros valores.

Portanto, *F(x)* é dada por:

$$F(x) = 0 \text{ se } x \le 0$$
$$= \int_0^x 2s\,ds = x^2 \qquad \text{se } 0 < x \le 1$$
$$= 1 \text{ se } x > 1$$

Assim, temos o seguinte gráfico:



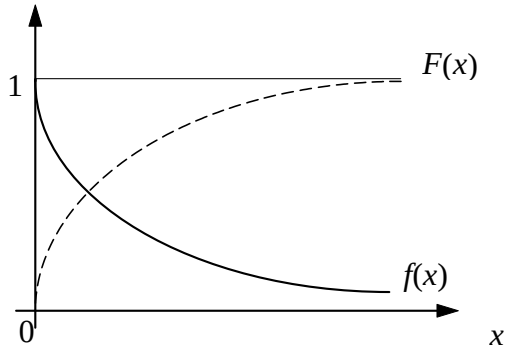
Ex.: Suponha que uma variável aleatória contínua tenha $F(x)$ dado por:

$$F(x) = 0 \quad , x \le 0$$
$$= 1 - e^{-x} , x > 0$$

Nesse caso,

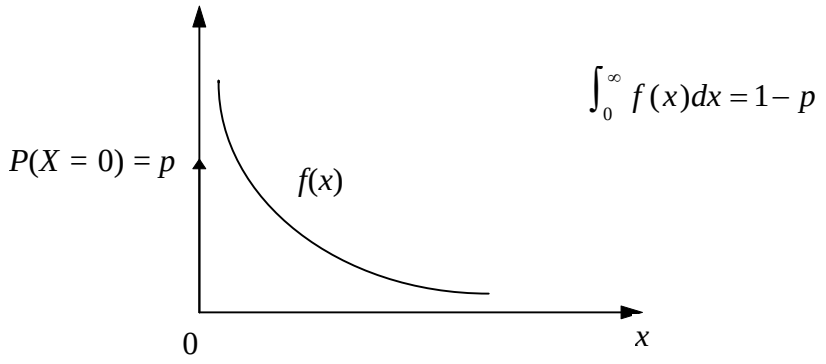
$$f(x) = \frac{d\ F(x)}{dx} = e^{-x} \qquad x \ge 0$$
$$= 0, \text{ para quaisquer outros valores.}$$

Assim temos:



Ex.: Suponha que estejamos ensaiando um equipamento e façamos igual a X o tempo de funcionamento, uma *v.a.*contínua com *fdp f (x)* para $x \ge 0$. No entanto, podem surgir situações

nas quais exista uma probabilidade não nula p de que o equipamento não funcione no momento $x = 0$. Nesse caso, podemos modificar o nosso modelo e atribuir $P(X = 0) = p$ e $P(X > 0) = 1 - p$. Deste modo, p descreve a distribuição de X no ponto 0 (*v.a.* discreta), enquanto $f(x)$ descreve a distribuição para valores de $x > 0$ (*v.a.* contínua). Tal distribuição é um exemplo de distribuição mista. Observe o gráfico:



2.3.3 - Variáveis aleatórias bidimensionais

Um par ordenado de valores aleatórios define uma *v.a.* bidimensional. Uma distribuição de probabilidade bidimensional pode ser discreta, sendo caracterizada por uma função probabilidade bidimensional tal que

$$\sum_i \sum_j P(x_i, y_j) = 1; \quad P(x_i, y_j) \geq 0 \qquad \text{para todo } (x, y)$$

ou contínua, sendo caracterizada por uma função densidade de probabilidade bidimensional tal que

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\, dx\, dy = 1 \qquad f(x, y) \geq 0$$

No caso discreto, definem-se as *distribuições marginais* de X e Y (os totais em cada linha e coluna - ver cap1) respectivamente, por:

$$P(X = x_i) = \sum_j P(x_i, y_j),$$

$$P(Y = y_j) = \sum_i P(x_i, y_j),$$

e no caso contínuo, por

$$g(x) = \int_{-\infty}^{\infty} f(x, y)\, dy$$

$$h(y) = \int_{-\infty}^{\infty} f(x, y)\, dx$$

Duas *v.a.* discretas X e Y são independentes se, para todos os pares (x_i, y_j)

$P(x_i, y_j) = P(X = x_i) \,.\, P(Y = y_j)$

Analogamente, no caso contínuo, X e Y são independentes se, para todos os pares (x, y),

$f(x, y) = g(x) \,.\, h(y)$

Ex.: Duas linhas de produção fabricam um certo tipo de peça. Suponha que a capacidade (em qualquer dia) seja de 5 peças na linha I e 3 peças na linha II. Admita que o número de peças realmente produzidas em qualquer linha seja uma variável aleatória, e que (x, y) represente a variável aleatória bidimensional que fornece o número de peças produzidas pela linha I e a linha II, respectivamente. A tabela a seguir fornece a distribuição de probabilidade conjunta de (x, y) . Cada casa representa

$p = (x_i\,, y_i) = P\,(X = x_i\,, Y = y_i)$

Assim, $p\,(2,3) = P\,(X = 2;\ Y = 3) = 0.04$, etc...

Se B for definido, por exemplo, como:

B : Mais peças são produzidas pela linha I do que pela linha II

então encontraremos que:

$P\,(B) = 0.01 + 0.03 + 0.05 + 0.07 + 0.09 + 0.04 + 0.05 + 0.06 + 0.08 + 0.05$
$+ 0.05 + 0.06 + 0.06 + 0.05 = 0.75$

Y \ X	0	1	2	3	4	5	Total
0	0	0.01	0.03	0.05	0.07	0.09	0.25
1	0.01	0.02	0.04	0.05	0.06	0.08	0.26
2	0.01	0.03	0.05	0.05	0.05	0.06	0.25
3	0.01	0.02	0.04	0.06	0.06	0.05	0.24
Total	0.03	0.08	0.16	0.21	0.24	0.28	1.00

Suponhamos que se deseje calcular a probabilidae condicionada $P(X = 2 \mid Y = 2)$.

Temos então:

$$P(X = 2 \mid Y = 2) = \frac{P(X = 2, Y = 2)}{P(Y = 2)} = \frac{0.05}{0.25} = 0.20$$

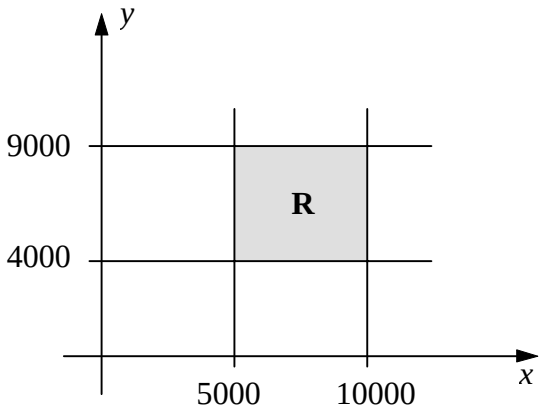
e generalizando:

$$P(x_i \mid y_j) = \frac{P(x_i, y_j)}{P(y_j)} = P(X = x_i \mid Y = y_j)\ , \quad P(y_j) > 0$$

Ex.: Suponha que um fabricante de lâmpadas esteja interessado no número de lâmpadas encomendadas a ele durante os meses de janeiro e fevereiro. Sejam X e Y, respectivamente, o número de lâmpadas encomendadas durante esses dois meses. Admitiremos que (X , Y) seja uma *v.a* contínua bidimensional, com a seguinte *fdp* conjunta

$f(x, y) = c \qquad 5000 \le x \le 10.000 \quad , 4.000 \le y \le 9.000$

$\qquad\quad = 0 \qquad$ para quaisquer outros valores



Assim, c é calculado como

$$\int_{4.000}^{9.000} \int_{5.000}^{10.000} c\,dx\,dy = c[5000]^2 = 1 \Rightarrow c = (5.000)^{-2}$$

Se, por exemplo,

$B \quad = \{ X \ge Y \}$, teremos

$$P(B) = 1 - c\int_{5.000}^{9.000}\int_{5.000}^{y} dx\,dy = 1 - c\int_{5.000}^{9.000}[y - 5.000]\,dy = \frac{17}{25}$$

Ex.: Seja (x, y) uma *v.a.* com *fdp* dada por:

$$f(x, y) = x^2 + \frac{xy}{3} \qquad 0 \le x \le 1 \qquad 0 \le y \le 2$$

$$= 0 \qquad \text{para outros valores}$$

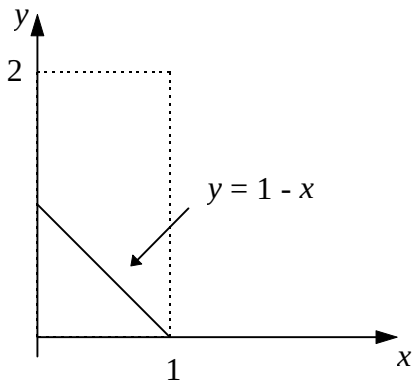
$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y)\, dx\, dy = 1 \qquad \text{(Verifique!)}$$

Se $B = \{X + Y \ge 1\}$ podemos calcular $P(B)$ como:

$P(B) = 1 - P(\overline{B})$ onde $\overline{B} = \{X + Y \le 1\}$.

Portanto,

$$P(B) = 1 - \int_0^1 \int_0^{1-x} \left(x^2 + \frac{xy}{3} \right) dy\, dx = \frac{65}{72}$$



As distribuições marginais $g(x)$ e $h(y)$ são dadas por:

$$g(x) = \int_0^2 \left(x^2 + \frac{xy}{3} \right) dy = 2x^2 + \frac{2}{3}x$$

$$h(y) = \int_0^1 \left(x^2 + \frac{xy}{3} \right) dx = \frac{y}{6} + \frac{1}{3}$$

e as probabilidades condicionadas $g(x \mid y)$ e $h(y \mid x)$ são dadas por:

$$g(x \mid y) = \frac{f(x, y)}{h(y)} = \frac{x^2 + xy/3}{1/3 + y/6} = \frac{6x^2 + 2xy}{2 + y} \qquad 0 \le x \le 1,\ 0 \le y \le 2$$

$$h(y \mid x) = \frac{f(x, y)}{g(x)} = \frac{x^2 + xy/3}{2x^2 + \frac{2}{3}x} = \frac{3x + y}{6x + 2} \qquad 0 \le x \le 1,\ 0 \le y \le 2$$

Pode-se verificar também que $g(x \mid y)$ e $h(y \mid x)$ são *fdp*. Observe:

$$\int_0^1 \frac{6x^2 + 2xy}{2+y}\ dx = \frac{2+y}{2+y} = 1$$

$$\int_0^2 \frac{3x+y}{6x+2}\ dy = 1$$

Ex.: Duas características do desempenho do motor de um foguete são o empuxo X e a taxa de mistura Y. Suponha que (X, Y) seja uma variável aleatória bidimensional com função:

$f(x, y) = 2\,(x + y - 2xy) \qquad 0 \le x \le 1, \qquad 0 \le y \le 1$

$\qquad = 0$ para quaisquer outros valores.

A função de probabilidade marginal de X é dada por:

$$g(x) = \int_0^1 2\left(x + y - 2xy\right)\ dy = 2\left(xy + \frac{y^2}{2} - xy^2\right)\Big|_0^1$$

$g(x) = 1 \qquad 0 \le x \le 1$

Isso significa que X é *distribuída uniformemente* sobre [0,1].

Para Y, a função marginal é dada por:

$$h(y) = \int_0^1 2\left(x + y - 2xy\right)\ dx = 2\left(x^2/2 + xy - x^2 y\right)\Big|_0^1$$

$= 1 \qquad 0 \le y \le 1$

Portanto, Y é também uniformemente distribuída sobre [0,1].

Obs.*:* Dizemos que a *v.a.* contínua bidimensional é uniformemente distribuída sobre uma região R do plano euclideano quando

$f(x, y) = c^{te} \qquad$ para $(x, y) \in R$

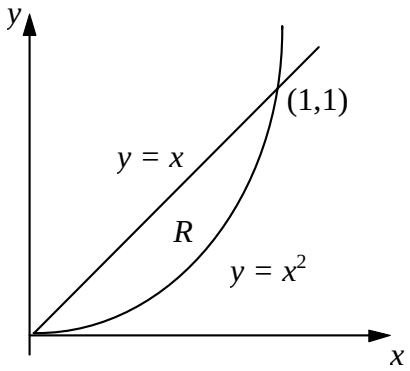
$\qquad = 0 \qquad$ para qualquer outra região

Em virtude da condição $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f\left(x, y\right) dxdy = 1$

a definição acima acarreta que a constante será igual a 1/área (R). Estamos supondo que R seja uma região com área finita, não nula.

Ex.: Suponhamos que a *v.a.* (x, y) seja uniformemente distribuída sobre a região R indicada na figura a seguir. Portanto,

$$f(x, y) = \frac{1}{\text{area}\,(R)} \quad , \quad (x, y) \in R$$



Assim

$$\text{área}\ (R) = \int_0^1 \left(x - x^2\right) dx = 1/6$$

e a *fdp* será dada por $f(x, y) \ = 6 \quad , (x, y) \in R$

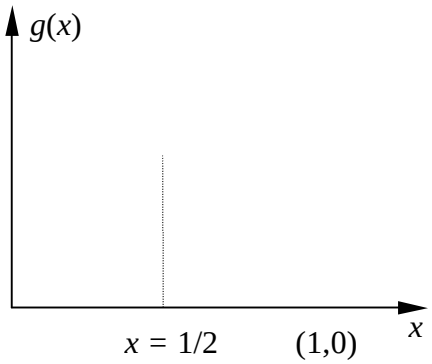
$\qquad\qquad\qquad\qquad\quad = 0 \quad , (x, y) \notin R$

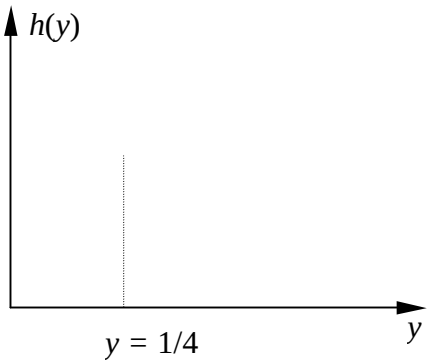
As *fdp* marginais de X e Y são dadas por:

$$g(x) = \int_{-\infty}^{\infty} f(x, y)\, dy \quad = \int_{x^2}^{x} 6\, dy = 6\left(x - x^2\right), \qquad 0 \le x \le 1,$$

e $$h(y) = \int_{-\infty}^{\infty} f(x, y)\, dy = \int_{y}^{\sqrt{y}} 6\, dx = 6\left(\sqrt{y} - y\right), \qquad 0 \le y \le 1$$

e seus respectivos gráficos são:





Ex.: Suponhamos que uma máquina seja utilizada para determinada tarefa durante a manhã e

para uma tarefa diferente durante a tarde. Representemos por X e Y, respectivamente o número de vezes que a máquina pára por defeito de manhã e à tarde. A tabela a seguir dá a distribuição de probabilidade conjunta de (x, y).

Y \ X	0	1	2	$p(y_j)$
0	0.1	0.2	0.2	0.5
1	0.04	0.08	0.08	0.2
2	0.06	0.12	0.12	0.3
$p(x_i)$	0.2	0.4	0.4	1.0

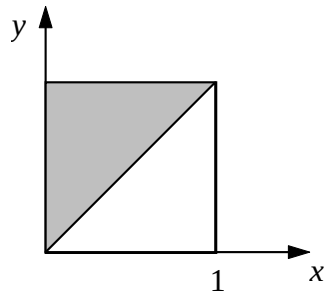
Um cálculo simples mostra que, para todas as casas temos $p(x_i, y_j) = p(x_i) \cdot p(y_j)$. Portanto, X e Y são *v.a.* independentes.

Ex.: Sejam X e Y a duração da vida de dois dispositivos eletrônicos. Suponha que a *fdp* conjunta seja dada por:

$$f(x, y) = e^{-(x+y)} \qquad x \geq 0 \quad y \geq 0$$

Desde que podemos fatorar $f(x, y) = e^{-x} \cdot e^{-y}$, a independência entre X e Y fica estabelecida.

Ex.: Seja agora $f(x, y) = 8\,x\,y$, $0 \leq x \leq y \leq 1$ uma *fdp*. O domínio é indicado pela região da figura seguinte.



Muito embora f já seja escrita na forma fatorada, X e Y não são independentes já que o campo de definição é dependente.

Exercícios V

1 - Uma moeda é viciada de tal forma que $P(K) = 2P(C)$. Jogando-a três vezes, determinar a distribuição de probabilidade do número de caras, X. Determinar a função cumulativa e calcular: a) $P(1 \le X \le 3)$ b) $P(X > 2)$

2 - Calcular $P(a < X < b)$, conhecida $F(x)$.

3 - No exercício anterior, determinar uma aproximação para: $p(|X - 8| \ge 6)$.

4 - Em cada caso, determinar se os valores dados podem ser considerados como integrantes de uma *v.a.* X que tome apenas os valores 1, 2, 3, 4:

a) $f(1) = 0{,}23$ $f(2) = 0{,}28$ $f(3) = 0{,}39$ $f(4) = 0{,}18$

b) $f(1) = 1/7$ $f(2) = 2/7$ $f(3) = 1/14$ $f(4) = 3/7$

c) $f(1) = 0{,}15$ $f(2) = 0{,}2$ $f(3) = 0{,}38$ $f(4) = 0{,}27$

d) $f(1) = 0{,}26$ $f(2) = 0{,}49$ $f(3) = 0{,}08$ $f(4) = 0{,}33$

5 - Determinar a distribuição de probabilidade e a distribuição acumulada do número de discos de *rock*, quando se escolhem aleatoriamente quatro discos de uma coleção com cinco discos de *rock*, três de músicas popular e dois clássicos.

6 - O diâmetro X de um cabo elétrico supõe-se ser uma *v.a.* contínua X, com *fdp* $f(x) = 6x(1-x),\ 0 \le x \le 1$.

a) Verifique se $f(x)$ é realmente uma *fdp* e esboce o seu gráfico;

b) Obtenha uma expressão para $F(x)$;

c) Calcule $P(X \le 1/2 \,|\, 1/3 < X < 2/3)$

7 - A porcentagem de álcool em certo composto pode ser considerada uma *v.a* X., onde X, $0 < X < 1$, tem a seguinte função de probabilidade:

$$f(x) = 20x^3(1-x) \quad 0 < X < 1$$

a) Estabeleça a expressão de $F(x)$;

b) Calcule $p(x \le 2/3)$;

c) Suponha que o preço de venda desse composto dependa do conteúdo de álcool. Especificamente, se $1/3 < x < 2/3$, o composto se vende por C_1 dólares/galão; caso contrário, ele se vende por C_2 dólares/galão. Se o custo for C_3 dólares/galão, calcule a distribuição de probabilidade do lucro líquido por galão.

2.4 - Esperança Matemática

As distribuições de probabilidade são modelos teóricos onde as probabilidades dos valores assumidos pela *v.a.* podem ser interpretadas como freqüências relativas (f_i) em um número sempre crescente de provas, ou seja, como limites de freqüências relativas.

Podemos, pois, definir, para as distribuições de probabilidade, as mesmas medidas de tendência central e de dispersão utilizadas nas distribuições de freqüências do Capítulo 1.

Guardando essa semelhança podemos definir a média de uma *v.a. X* como:

$$E(X) = \sum x_i \, f\left(x_i\right) \qquad \text{(caso discreto)}$$

$E(X)$ é a *Esperança Matemática*, também chamada de *Valor Esperado*, *Expectância* ou ainda *Média Populacional* (μ). Repare que $E(X)$ é uma média ponderada dos x_i, em que os pesos são as probabilidades associadas.

Se X é uma *v.a.* contínua com *fdp* f, $E(X)$ é dada por:

$$E(X) = \int_{-\infty}^{\infty} x \, f(x) \, dx \qquad \text{(caso contínuo)}$$

Sendo (X, Y) uma *v.a.* bidimensional, façamos $Z = H(X, Y)$.

- Se (X, Y) for discreta e se

$p(x_i, y_j) = p(X = x_i, Y = y_j)$, $\quad i, j = 1, 2, 3, \ldots$ teremos:

$$E(Z) = \sum_{j=1}^{\infty} \sum_{i=1}^{\infty} H\left(x_i, y_j\right) p\left(x_i, y_j\right)$$

- Se (x, y) for contínua com *fdp* f, teremos:

$$E(Z) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} H\left(x, y\right) f\left(x, y\right) dx \, dy$$

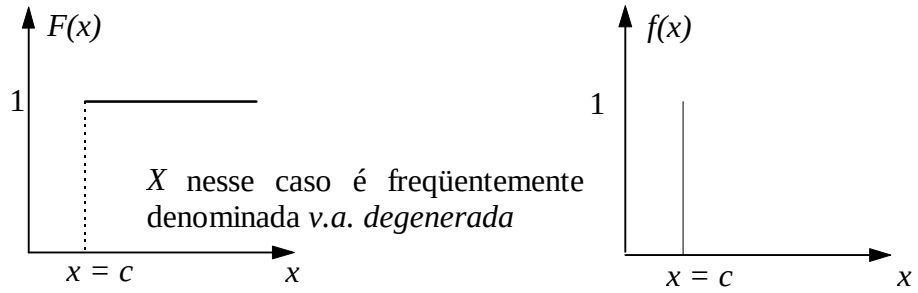
São propriedades de $E(X)$ (as verificações são aqui dadas apenas para o caso contínuo. Verifique tais propriedades também para o caso discreto):

i) Se X é uma *v.a.* com $X = c$, onde c é uma constante, então

$E(X) = E(c) = c.$

Pois: $E(X) = \int_{-\infty}^{\infty} c\ f(x)\, dx = c\int_{-\infty}^{\infty} f(x)\, dx = c$

Observe o significado de $X = c$



ii) $E(cX) = c\,E(X)$

Nesse caso:

$$E(cX) = \int_{-\infty}^{\infty} cx\ f(x)\, dx = c\int_{-\infty}^{\infty} x\ f(x)\, dx = cE(X)$$

Obs.: Se multiplicarmos todos os termos de uma *v.a.* X por c, o seu valor médio fica também multiplicado por c.

iii) Seja (X, Y) uma *v.a.* bidimensional. Seja $Z = H_1(x, y)$ e $W = H_2(x, y)$. Então $E(Z + W) = E(Z) + E(W)$

Realmente:

$$E(Z+W) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left[H_1(x,y) + H_2(x,y)\right] f(x,y)\, dxdy = E(Z) + E(W)$$

iv) Se X e Y são duas *v.a.* quaisquer. Então

$E(X + Y) = E(X) + E(Y).$

E isso decorre imediatamente de *iii*.

v) Sejam n *v.a.* $X_1, X_2, \ldots X_n$.. Então*:*

$$E(X_1 + X_2 + \cdots + X_n) = E(X_1) + E(X_2) + \cdots + E(X_n)$$

Isso decorre de *iv*, pela aplicação da indução matemática.

vi) Seja (*X*,*Y*) uma *v.a.* bidimensional. Se *X* e *Y* são *independentes*, então $E(XY) = E(X).E(Y)$

De fato:

$$E(XY) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\, f(x,y)\, dxdy = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\, g(x)h(y)\, dxdy$$
$$= \int_{-\infty}^{\infty} x\, g(x)dx \int_{-\infty}^{\infty} yh(y)\, dy = E(X).E(Y)$$

Ex.: Seja *X* uma *v.a.* que assume os seguintes valores $X = \{15,\ 10,\ 5\ \ -5\}$ e tenha a seguinte distribuição de probabilidade:

X	*f*(*x*)
15	0,56
10	0,23
5	0,02
-5	0,19

Assim, para o cálculo da Esperança Matemática, temos:

$$E(X) = \sum x_i\, f(x_i) = 15(0.56) + 10(0.23) + 5(0.02) + (-5)(0.19) = 9{,}85$$

Se fizermos $Z_1 = 2X$ teremos uma nova *v.a.* com os seguintes valores $Z_1 = \{30,\ 20,\ 10,\ -10\}$ e $E(Z_1) = E(2X) = 2E(X) = 19.7$ (O valor esperado E(*X*) foi duplicado)

Se fizermos $Z_2 = X + 2$ teremos uma nova *v.a.* com os seguintes valores, $Z_2 = \{17,\ 12,\ 7,\ -3\}$ e $E(Z_2) = E(X+2) = 2 + E(X)$ = 11,85.

Ex.: Suponhamos que a procura *D*, por semana, de um certo produto seja uma *v.a.* com uma certa distribuição de probabilidade, digamos $P(D = n) = p(n)$, *n* = 0, 1, 2, ... Admitamos que

o custo para o vendedor seja C_1 dólares por peça, enquanto ele vende cada peça por C_2 dólares. Qualquer peça que não tenha sido vendida até o fim de semana deve ser estocada ao custo de C_3 dólares por peça. Se o vendedor decidir produzir N peças no início da semana, qual será seu lucro esperado por semana? Para que valor de N, o lucro esperado se torna máximo?

Se L for o lucro semanal, temos

$$\begin{aligned} L &= NC_2 - NC_1 && D > N \\ &= DC_2 - C_1N - C_3(N-D) && D \le N \end{aligned}$$

Reescrevendo:

$$\begin{aligned} L &= N\left(C_2 - C_1\right) && D > N \\ &= (C_2 + C_3)D - N\left(C_1 + C_3\right) && D \le N \end{aligned}$$

O lucro esperado será dado por:

$$\begin{aligned} E(L) &= N(C_2 - C_1).P(D > N) + \left[(C_2 + C_3)D - N(C_1 + C_3)\right]P(D \le N) \\ &= N(C_2 - C_1).P(D > N) + (C_2 + C_3)\sum_{n=0}^{N} np(n) - N(C_1 + C_3)P(D \le N) \\ &= N(C_2 - C_1)\sum_{n=N+1}^{\infty} p(n) + (C_2 + C_3)\sum_{n=0}^{N} np(n) - N(C_1 + C_3)\sum_{n=0}^{N} p(n) \\ &= N(C_2 - C_1) + (C_2 + C_3)\left[\sum_{n=0}^{N} np(n) - N\sum_{n=0}^{N} p(n)\right] \end{aligned}$$

$$E(L) = N(C_2 - C_1) + (C_2 + C_3)\sum_{n=0}^{N} p(n)\,(n - N)$$

Suponhamos que se saiba a seguinte distribuição de probabilidade seja apropriada para D:

$P(D = n) = 1/5$; $\qquad n = 1, 2, 3, 4, 5.$

Então:

$$\begin{aligned} E(L) &= N(C_2 - C_1) + \frac{(C_2 + C_3)}{5}\left[N(N+1)/2 - N^2\right] && N \le 5 \\ &= N(C_2 - C_1) + (C_2 + C_3)\frac{1}{5}(15 - 5N) && N > 5 \end{aligned}$$

Suponhamos ainda que:

$C_1 = \$3,\ C_2 = \$9$ e $C_3 = \$1.$

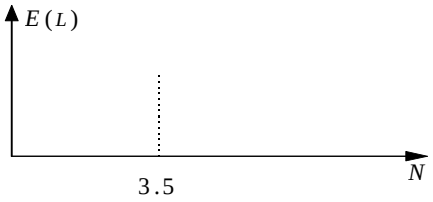
Logo,

$$E(L) = 6N + 2\left[N(N+1)/2 - N^2\right] \qquad N \le 5$$
$$\quad = 6N + 2(15 - 5N) \qquad N > 5$$

ou ainda

$$E(L) = 7N - N^2 \qquad N \le 5$$
$$\quad = 30 - 4N \qquad N > 5$$

Portanto, o máximo ocorre para N = 3,5. Para N = 3 ou 4, teremos $E(L) = 16$, que é o máximo atingível, posto que N é inteiro.



2.4.1 - A variância de uma *v.a.*

Seja X uma *v.a.*. Definimos a variância de X denotada por $Var(X)$, S^2 ou σ^2, da seguinte maneira:

$$Var(X) = E\left[(X - E(X))^2\right] \text{ ou } Var(X) = E\left[(X - \mu)^2\right]$$

A raiz quadrada positiva de $Var(X)$ é o desvio padrão de X, *DP(X), S ou* σ.

Uma outra expressão para a variância é:

$$Var(X) = E(X^2) - \left[E(X)\right]^2 = E(X^2) - \mu^2 \qquad \text{Comprove!}$$

Obs.: $Var(X) = E\left[(X - E(X))^2\right]$ é um caso especial da seguinte noção mais geral. O k-ésimo *momento* da *v.a. X*, em relação à sua esperança, é definida como sendo $\mu_k = E\left[(X - E(X))^k\right]$. Evidentemente, para k = 2 obtemos a variância (ver Cap.1).

Ex.: O serviço de meteorologia classifica o tipo de céu que é visível, em termos de "graus de

nebulosidade". Uma escala de 11 categorias é empregada: 0, 1, 2, ... 10, onde 0 representa o céu perfeitamente claro, enquanto os outros valores representam as diferentes condições intermediárias. Suponha que uma determinada estação meteorológica faça tal classificação em um determinado dia e hora. Seja *X* a *v.a.* que pode tomar um dos 11 valores acima. Admita-se que a distribuição de probabilidade de *X* seja:

$$p_0 = p_{10} = 0.015$$
$$p_1 = p_2 = p_8 = p_9 = 0.15$$
$$p_3 = p_4 = p_5 = p_6 = p_7 = 0.06$$

Portanto,

$$E(X) = 1.(0.15) + 2.(0.15) + \cdots + 10.(0.05) = 5.0$$

A fim de calcular *Var*(*X*), necessitamos calcular $E(X^2)$

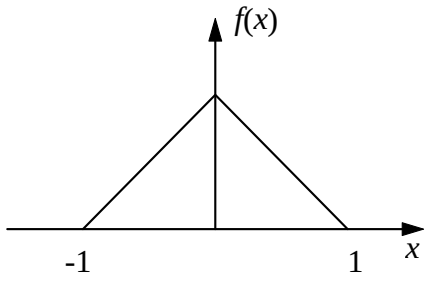
$$E(X^2) = 1(0.15) + 4(0.15) + \cdots + 100(0.05) = 35.6$$

e

$$Var(X) = E(X^2) - \mu^2 = 35.6 - 25 = 10.6$$
$$DP(X) = 3.25$$

Ex.: Suponhamos que *X* seja uma *v.a.* contínua com *fdp*:

$$f(x) = 1 + x \quad -1 \le x \le 0$$
$$\quad = 1 - x \quad 0 \le x \le 1$$



$$E(X) = \int_{-1}^{1} x f(x)dx = \int_{-1}^{0} x(1+x)dx + \int_{0}^{1} x(1-x)dx = 0$$
$$E(X^2) = \int_{-1}^{0} x^2(1+x)dx + \int_{0}^{1} x^2(1-x)dx = 1/6$$
$$Var(X) = 1/6$$

São propriedades da Variância:

i) $Var(X+c) = Var(X)$, $c = c^{\underline{te}}$

Pois: $Var(X+c) = E\left\{\left[(X+c)-E(X+c)\right]^2\right\} = E\left\{\left[(X+c)-E(X)-c\right]^2\right\}$

$= E\left[(X-E(X))^2\right] = Var(X)$

Obs.: Esta propriedade é intuitivamente evidente, porque somar uma constante a um resultado *X* não altera sua variabilidade, que é aquilo que a variância mede. Apenas "desloca" os valores de *X* para a direita ou à esquerda, dependendo do sinal de *c*.

ii) $Var(cX) = c^2\, Var(X)$

Pois: $Var(cX) = E\left[(cX)^2\right] - \left[E(cX)\right]^2 = c^2E(X^2) - c^2\left[E(X)\right]^2$

$= c^2\left\{E(X)^2 - \left[E(X)\right]^2\right\} = c^2Var(X)$

iii) Se *X*, *Y* forem independentes então:

$Var(X+Y) = Var(X) + Var(Y)$

Pois: $Var(X+Y) = E[(X+Y)^2] - E[(X+Y)]^2$

$= E[X^2 + 2XY + Y^2] - [E(X) + E(Y)]^2$

$= E[X^2 + 2XY + Y^2] - [E(X)]^2 - 2E(X)E(Y) - [E(Y)^2]$

$= E(X^2) - [E(X)]^2 + E(Y^2) - [E(Y)]^2$

$= Var(X) + Var(Y)$

Obs.: É importante lembrar que a variância não é, em geral, aditiva, como o é $E(X)$.

iv) Sejam $X_1, X_2, \cdots X_n$, *n v.a.* independentes.

$Var(X_1 + X_2 + \cdots + X_n) = Var(X_1) + Var(X_2) + \cdots + Var(X_n)$

v) Seja *X* uma *v.a.* com variância finita. Para $\alpha \in \mathrm{R}$, temos:

$Var(X) = E[(X-\alpha)^2] - [E(X)-\alpha]^2$

Obs.: Se interpretarmos *Var*(*X*) como o *momento de inércia* e *E*(*X*) como o *centro de uma unidade de massa*, então a propriedade acima é uma formulação do *teorema dos eixos*

paralelos, em Mecânica. O momento de inércia em relação a um ponto arbitrário é igual ao momento de inércia em relação ao centro de gravidade, mais o quadrado da distância desse ponto arbitrário ao centro de gravidade.

$E(X-\alpha)$ se torna mínimo se $\alpha = E(X)$. Isto decorre imediatamente da propriedade acima. Portanto, o momento de inércia (de uma unidade de massa distribuída sobre uma reta), em relação a um eixo que passe por um ponto arbitrário, torna-se mínimo se esse ponto escolhido for o centro de gravidade.

2.4.2 - Desigualdade de Tchebichev

Existe uma bem conhecida desigualdade, devida ao matemático russo Tchebichev, que nos fornece meios de compreender precisamente como a variância mede a variabilidade em relação ao valor esperado de uma *v.a.*. É ela:

$$P(|X-\mu| \geq \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2} \qquad \varepsilon \geq 0$$

ou, com $\varepsilon = k\sigma$ $\quad P(|X-\mu| \geq k\sigma) \leq \dfrac{1}{k^2}$

onde:

X é uma *v.a.* contínua ou discreta com média μ e variância σ^2, ambas finitas.

Ex.: Fazendo $k = 2$ na desigualdade de Tchebichev, vemos que

$$P(|X-\mu| \geq 2\sigma) \leq 0.25 \quad \text{ou}$$

$$P(|X-\mu| < 2\sigma) \geq 0.75$$

Em palavras, a probabilidade de X diferir de sua média em mais de dois desvios padrões é inferior ou, no máximo, igual a 0.25; equivalentemente, a probabilidade de X estar compreendido no intervalo de dois desvios padrões a contar de sua média, é superior ou, no mínimo, igual a 0.75. Tal fato é tanto mais notável quanto se considerarmos que nem sequer especificamos a distribuição de probabilidade de X.

Exercícios VI

1 - Se uma pessoa recebe R$ 5,00 se aparecerem somente caras ou somente coroas na jogada de três moedas, e paga R$ 2,00 em caso contrário, qual seu ganho esperado?

2 - Determinar o número esperado de químicos em um comitê de três elementos escolhidos entre quatro químicos e três biólogos.

3 - X é uma *v.a.* com a seguinte distribuição de probabilidade:

x	0	1	2	3	4
$f(x)$	1/4	1/8	1/4	1/8	1/4

Determinar o valor esperado de $Y = (X+1)^2$

4 - Mostre que a variável padronizada tem média 0 e variância 1.

5 - O tempo T, em minutos, necessário para um operário de uma indústria processar certa peça é uma *v.a.* com a seguinte distribuição de probabilidade:

T	2	3	4	5	6	7
P	0.1	0.1	0.3	0.2	0.2	0.1

Para cada peça processada, o operário ganha um fixo de U$ 2.0 mas se ele processa uma peça em menos de 6 minutos, ganha U$ 0.5 por minuto poupado (por exemplo, se ele processa a peça em quatro minutos, recebe a quantia adicional de U$ 1.0).

Qual a média e a variância da quantia ganha por peça?

6 - Suponha que um mecanismo eletrônico tenha um tempo de vida X (em unidades de 1000 horas) que tenha uma função de probabilidade dada por: $f(x) = e^{-x}$, $x > 0$.

Suponha que o custo de fabricação de um item seja U$ 200 e o preço de venda seja U$ 500. O fabricante garante total devolução se $x \leq 0{,}8$.

Qual o lucro esperado por item?

7 - O tempo de vida (em horas) de um mancal é uma variável aleatória, com função de probabilidade dada por:

$$f(t) = \begin{cases} \dfrac{1}{800} e^{-t/800} & , t \geq 0 \\ 0 & , t < 0 \end{cases}$$

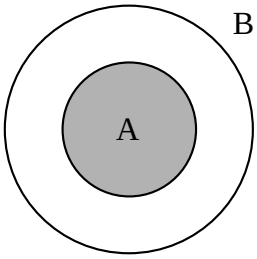
A vida média do mancal é de 800 horas. Qual a probabilidade de que o seu tempo de vida seja maior do que a média?

Anexo 1 - Noções de Conjuntos e Álgebra de Boole

A partir do estudo de elementos da *teoria de conjuntos* os seguintes conceitos, entre outros, podem ser facilmente entendidos:

Sendo A e B dois conjuntos quaisquer, temos:

i) $A \subset B$ (lê-se "A está incluído em B" ou "A é um subconjunto de B") se e somente se para todo $a \in A$ então $a \in B$.

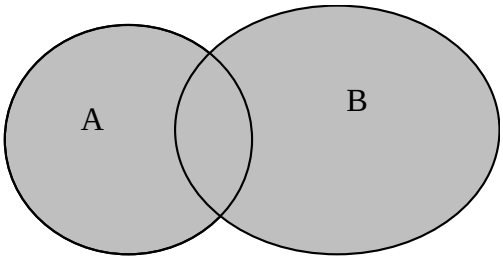


Portanto $A \subset A$ e, denotando por Φ o conjunto vazio, $\Phi \subset A$ qualquer que seja o conjunto A.

ii) $A = B$ se e somente se $A \subset B$ e $B \subset A$.

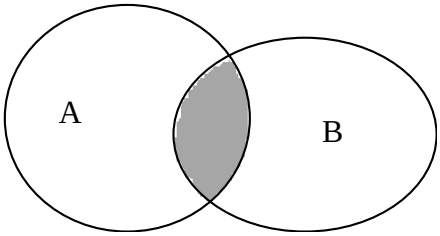
iii) $A \cup B$ (lê-se "A união B") é o conjunto formado pelos elementos que pertencem a A, a B ou a ambos, isto é:

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



iv) $A \cap B$ (lê-se "A interseção B") é o conjunto formado pelos elementos que pertencem a A e a B, isto é:

$A \cap B = \{ x : x \in A \text{ e } x \in B\}$



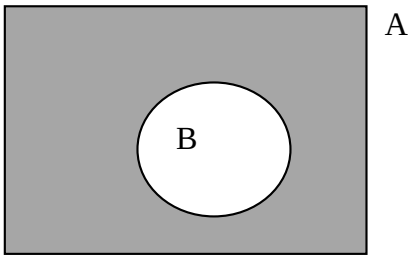
v) P (A) (lê-se "partes de A) é o conjunto formado por todos os subconjuntos de A.

Ex.: Se $A = \{ a, b, c\}$, então:

$P(A) = \{\Phi, \{a\}, \{b\}, \{c\}, \{a,b\}, \{a,c\}, \{b,c\}, A\}$

vi) $C_A B$ (lê-se "complemento de B com relação a A) é o conjunto formado por aqueles elementos de A que não pertencem a B, isto é:

$C_A B = \{ x \in A : x \notin B\}$



Quando A é o conjunto universo, $C_A B = B' = \overline{B}$.

<u>Álgebra Booleana</u>

Um modelo matemático da teoria axiomática da Álgebra de Boole pode ser encontrado dentro da teoria de conjuntos, usando os conceitos acima definidos. Assim, temos:

Sendo, $X, Y, Z, ...$ subconjuntos de um conjunto universal A, isto é,

$X, Y, Z... \in P(A)$

definidas as operações binárias:

$$a: \quad P(A) \times P(A) \rightarrow P(A)$$
$$(X, Y) \quad \rightarrow X + Y = X \cup Y$$
$$me: \quad P(A) \times P(A) \rightarrow P(A)$$
$$(X, Y) \quad \rightarrow X \,.\, Y = X \cap Y$$

e a função:

$$\sigma: \quad P(A) \rightarrow P(A)$$
$$X \quad \rightarrow \sigma(X) = C_A X = X' = \overline{X},$$

tendo ainda os seguintes axiomas:

1)	$X + Y = Y + X$	$X.Y = YX$
2)	$X\,.\,(Y + Z) = (XY) + (XZ)$	$X + (YZ) = (X + Y)(X + Z)$
3)	$X + 0 = X$	$X.1 = X$
4)	$X + \overline{X} = 1$	$X.\overline{X} = 0$
5)	$1 \neq 0$	

e associando *1* a *A* e *0* a Φ

temos uma importante ferramenta para operação de conjuntos.

Alguns teoremas desse modelo são:

a) $X + X = X$; $X.X = X$

b) Se $\begin{cases} X + Y = 1 \\ X.Y = 0 \end{cases}$ então $Y = \overline{X}$

c) $\overline{\overline{X}} = X$

d) $X\,.\,0 = 0$; $X + 1 = 1$

e) $X(X + Y) = X$; $X + (XY) = X$

f) $X + (Y + Z) = (X + Y) + Z$; $X(YZ) = (XY)Z$

g) $\overline{(X + Y)} = \overline{X}.\overline{Y}$; $\overline{X.Y} = \overline{X} + \overline{Y}$ (Leis de De Morgan)

h) $\overline{0} = 1$; $\overline{1} = 0$

Obs.: Demonstre os teoremas acima. Perceba que essa teoria também se aplica a um modelo tecnológico relacionado com chaveamento.

Ex.: Simplifique a operação de conjuntos abaixo:

$$\left[(A \cap B') \cup (A' \cap B)\right]' \cap \left[(A \cap B) \cup (A' \cap B')\right]'$$

Uma possível solução:

$\overline{\left(A.\overline{B}+\overline{A}.B\right)}.\overline{\left(A.B+\overline{A}.\overline{B}\right)}$	fazendo $\cap \equiv \cdot$, $\cup \equiv +$, $X' \equiv \overline{X}$
$\overline{\left(A.\overline{B}+\overline{A}.B\right)+\left(A.B+\overline{A}.\overline{B}\right)}$	usando (*g*)
$\overline{A\left(\overline{B}+B\right)+\overline{A}\left(B+\overline{B}\right)}$	evidenciando A e $\overline{A}$
$\overline{A.1+\overline{A}.1}=\overline{A+\overline{A}}=\overline{1}$	usando o axioma (*4*) e (*3*)
$\overline{1}=0\equiv\Phi$ (vazio)	usando (*h*)

Anexo 2 - Revisão de Análise Combinatória

O cálculo efetivo da probabilidade de um evento depende freqüentemente do uso dos resultados da Análise Combinatória. Esta seção é, pois, um sumário dos principais resultados dessa área de matemática elementar.

A lista de exercícios, referente a essa parte, ao final, sintetiza esse sumário.

- **Princípio Fundamental da Contagem**

Os problemas de Análise Combinatória são, basicamente, problemas de contagem. A abordagem destes problemas é baseada num fato, de fácil comprovação, denominado Princípio Fundamental da Contagem ou, simplesmente, Regra do Produto, que enunciaremos e exemplificaremos a seguir.

Um acontecimento é composto por k estágios sucessivos e independentes, com, respectivamente, $n_1, n_2, n_3 ..., n_k$ possibilidades cada. O número total de maneiras distintas de ocorrer este acontecimento é $n_1, n_2, n_3 ..., n_k$.

Ex.: Um engenheiro, ao se inscrever em um programa de mestrado deve escolher o curso e a faculdade que deseja cursar. Sabe-se que existem cinco cursos possíveis: Mecânica, Elétrica, Telecomunicações, Civil e Robótica. Cada curso pode ser feito em três faculdades possíveis: no Brasil, no Japão, nos EUA. Qual é o número total de opções que o engenheiro pode fazer?

$n = 5 . 3 = 15$ opções

- **Técnicas de Contagem**

Seja $A = \{ a; b; c; d; ... j \}$ um conjunto formado por 10 elementos distintos, e consideremos os "agrupamentos" *abc*, *abd* e *acb*.

Os agrupamentos *abc* e *abd* são considerados sempre distintos, pois diferem pela natureza de um elemento.

Os agrupamentos *abc* e *acb*, podem ser considerados distintos ou não.

Se por exemplo, os elementos do conjunto A forem pontos**,** $A = \{ A_1, A_2, ... A_{10} \}$, e ligando estes pontos desejarmos obter retas, então os agrupamentos $A_1 A_2$ e $A_2 A_1$ são iguais, pois representam a mesma reta. É isso um típico caso de combinação simples.

Se por outro lado, os elementos do conjunto A forem algarismos, $A = \{0, 1, ..., 9\}$, e com estes algarismos desejarmos obter números, então os agrupamentos 12 e 21 são distintos, pois representam números diferentes. É isso um típico caso de arranjo simples.

Apresentaremos, a seguir, o estudo sucinto dos três tipos de agrupamentos comumente utilizados em problemas de contagem, que são: Arranjos, Permutações (são um tipo particular de arranjo) e combinações.

- **Arranjos simples**

Seja A um conjunto com n elementos e k um número natural, com $k \leq n$. Chama-se arranjo simples k a k, dos n elementos de A, a cada "subconjunto" de k elementos de A.

Obs.: Note que, de acordo com a definição, os arranjos são agrupamentos que diferem entre sí, tanto pela natureza como também pela ordem de seus elementos.

De acordo com a definição dada, na formação de todos os arranjos simples k a k dos n elementos de A, temos:

– n possibilidades na escolha do 1º elemento;
– n - 1 possibilidades na escolha do 2º elemento, pois um elemento já foi utilizado;
– n - 2 possibilidades na escolha do 3º elemento, pois dois já foram utilizados;

⋮

– n - (k + 1**)** possibilidades na escolha do kº elemento.

De acordo com o Princípio Fundamental da contagem, o número total de arranjos simples k a k dos n elementos de A é $n(n - 1)(n - 2) ... (n - k + 1)$**.** Representando esse número por $A_{n,k}$, temos:

$A_{n,k} = n(n - 1)(n - 2) ... (n - k + 1)$, ou ainda

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

Ex.: $A_{10,\,4}$ = 10. 9. 8. 7. = 5040

- **Permutações simples**

Seja A um conjunto com n elementos. Os arranjos simples n a n de n elementos de A são chamados permutações simples de n elementos, representado por P_n. Assim:

$$P_n = A_{n,\,n} = n!$$

Ex.: $P_6 = 6! = 720$

- **Combinações simples**

Seja A um conjunto com n elementos e k um número natural, com $k \leq n$. Chama-se combinação simples k a k, dos n elementos de A, a cada subconjunto de k elementos de A.

$C_{n,\,k}$ é o símbolo utilizado para representar o número total de combinações de n, k a k. Permutando os k elementos de uma combinação k a k, obtemos P_k arranjos distintos. Permutando os k elementos das $C_{n,k}$ combinações, obtemos P_k**.** $C_{n,k}$ arranjos simples, que são exatamente todos os arranjos simples de **n** elementos k a k.

Assim sendo, $C_{n,k}$. $P_k = A_{n,k}$ e portanto:

$$C_{n,k} = \frac{A_{n,k}}{P_k} \quad \text{ou} \quad C_{n,k} = \frac{n!}{k!\,(n\ -\ k!)} \quad C_{n,k} = \binom{n}{k}$$

- **Permutações com repetição**

Sejam **a** elementos iguais a α, **b** elementos iguais a β**, g** elementos iguais a γ ..., **l** elementos iguais a λ, num total de **a+b +g+ ... + l = n** elementos

O número total de permutações distintas que podemos obter com estes **n** elementos é:

$$P^* = P_n^{a,b,g,\ldots l} = \frac{n!}{a!b!g!\ldots l!}$$

- **Permutações circulares**

$$P' = (n - 1)!$$

- **Arranjos com repetição**

$$A^*_{n,K} = n^K$$

- **Combinação com repetição**

$$C^*_{n,K} = C_{n+K-1,K} = \binom{n+k-1}{k}$$

Obs.: Os exercícios da série são um ótimo treinamento para se começar a "pensar" probabilisticamente. Faça todos!

Exercícios VII

01) Há 6 estradas entre *A* e *B* e 4 entre *B* e *C*. De quantas maneiras pode-se ir de *A* a *C*, passando por *B*.

02) Com relação ao problema anterior, de quantas maneiras pode-se ir e voltar de A a C, sem usar a mesma estrada mais de uma vez?

03) Dados os conjuntos $A = \{ 1, 2, 3, 4 \}$ e $B = \{ 1, 2, 3, 5, 6, 7 \}$, calcular o número de funções injetoras de *A* em *B*. Obs.: Uma função é injetora se para $x \neq y \rightarrow f(x) \neq f(y)$ *com* $f: A \rightarrow B$.

04) Sendo $A = \{ 1, 2, 3, 4 \}$ e $B = \{ 7, 8, 9, 10 \}$, calcular o número de funções bijetoras de *A* em *B*. Obs.: Uma função é bijetora se $x \neq y \rightarrow f(x) \neq f(y)$ e $im(A)=B$

05) Quantos números de algarismos distintos e compreendidos entre 100 e 1000, podem ser obtidos utilizando os algarismos 1, 2, 3, 5, 6?

06) Quatro livros diferentes de Matemática, seis de Física e dois de Química devem ser arrumados em uma prateleira.
Quantas arrumações diferentes podem ser feitas, se:
a) os livros de cada matéria devem ficar juntos?
b) apenas os de matemática devem ficar juntos?

07) Quantas comissões de 3 elementos podemos formar com um grupo de 8 pessoas?

08) Um fabricante de doces dispõe de embalagens com capacidade de 4 doces cada uma . Sabendo-se que ele fabrica 10 tipos diferentes de doces, pergunta-se: Quantos tipos de embalagens com 4 doces diferentes ele poderá oferecer?

09) Vamos supor que uma placa de motocicleta contenha duas letras distintas do alfabeto completo, seguidas por três dígitos. Quantas placas diferentes podem ser impressas?

10) Quantos são os anagramas da palavra AEROPORTO? *(Anagramas são palavras com ou sem sentido a partir de uma palavra primitiva)*

11) De quantas maneiras, uma oficina pode pintar 5 automóveis iguais, recebendo cada um, tinta de uma única cor, se a oficina dispõe apenas 3 cores e não quer misturá-las?

12) Resolver a equação: $x \cdot (x + 1) \cdot A_{16,\, x-1} = A_{16,\, x+1}$

13) Quantos números ímpares, compreendidos entre 2000 e 7000, podemos formar com os algarismos 2, 3, 4, 6, 8 e 9, de modo que não figurem algarismos repetidos?

14) Um trem de passageiros é constituído de uma locomotiva e 6 vagões distintos, sendo um deles restaurante. Sabendo que a locomotiva deve ir à frente e que o vagão restaurante não pode ser colocado imediatamente após a locomotiva, o número de modos diferentes de montar a composição é:

15) Com dez consoantes diferentes e com cinco vogais diferentes, quantas "palavras" (com sentido ou não) podem ser formadas, tendo cada palavra três consoantes diferentes e duas vogais diferentes?

16) Uma empresa tem 3 diretores e 5 gerentes. Quantas comissões de 5 pessoas podem ser formadas, contendo no mínimo 1 diretor?

17) São dados 12 pontos em um plano, dos quais 5 e somente 5 estão alinhados. Quantos triângulos podem ser formados com vértices em 3 dos 12 pontos?

18) Uma urna contém 10 bolas brancas e 6 pretas. De quantos modos é possível tirar 7 bolas, das quais pelo menos 4 bolas sejam pretas?

19) No sistema decimal, quantos números de cinco algarismos (sem repetição) podemos escrever, de modo que os algarismos 0 (zero), 2 (dois) e 4 (quatro) apareçam agrupados?

Obs.: Considerar somente números de cinco algarismos em que o primeiro algarismo é diferente de zero.

20) Calcular o número máximo de planos determinados por 8 pontos do espaço dos quais 4 são coplanares.

21) Existem n maneiras de distribuir 7 moedas de valores diferentes entre duas pessoas. Excluindo-se possibilidade de uma só receber todas as moedas, o valor de *n* será?

22) Com os algarismos 1, 2, 3, 4 e 5, e sem repetição, pode-se escrever *x* números maiores que 2.500. Calcular *x*.

23) Em um avião de oito lugares viajam oito pessoas, das quais quatro têm condições de operar como piloto ou co-piloto. De quantas maneiras diferentes estas oito pessoas podem se distribuir no avião?

24) Quantos são, ao todo, os anagramas da palavra MATEMÁTICA que começam com

vogal? (Não levar em conta o acento).

25) O número de soluções inteiras e não negativas da equação $x + y + z + w = 5$ é:

26) Um feirante possui, em sua banca, maçãs, peras e laranjas em grande quantidade. Desejando atender melhor a sua clientela, o feirante resolveu empacotar todas as sua frutas, de modo que cada pacote contivesse exatamente 5 frutas. Quantos tipos de pacotes poderá o feirante oferecer, no máximo, à sua clientela?

27) Numa loteria são sorteados 6 objetos. Sabe-se que a urna contém exatamente 20 bilhetes. Uma pessoa retira da urna 4 bilhetes. Qual o número de possibilidades que essa pessoa tem de retirar, pelo menos 2 bilhetes premiados entre os quatro retirados.

28) O número de comissões diferentes, de 2 pessoas, que podemos formar com os n diretores de uma firma, é k. Se, no entanto, ao formar estas comissões, tivermos de indicar uma das pessoas para presidente e a outra para suplente podemos forma $k + 3$ comissões diferentes. Então, n vale:

29) Dado um polígono convexo P de n lados, calcular o número de polígonos convexos, cujos vértices são vértices de P.

30) Se x e y são números naturais maiores que 1 e tais que $\begin{cases} A_{x+y,2} = 56 \\ C_{x-y,2} = 1 \end{cases}$,então $x \,.\, y$ é igual ao valor de :

31) Dentre os anagramas distintos que podemos formar com n letras, das quais somente duas são iguais, 120 apresentam estas duas letras iguais juntas. Então, n vale:

32) Uma prateleira tem 6 compartimentos separados. De quantas maneiras podemos colocar nos compartimentos 4 bolinhas indistinguíveis? (Este tipo de problema surge na Física em relação à estatística de Bose-Einstein).

33) Uma prateleira tem 6 compartimentos separados. De quantas maneiras podemos distribuir 11 bolinhas indistinguíveis pelos compartimentos, de modo que nenhum deles fique vazio? (Este tipo de problema surge na Física em relação à estatística de Fermi-Dirac).

CAPÍTULO 3 - DISTRIBUIÇÕES DE PROBABILIDADE

Nesse capítulo serão apresentadas várias distribuições de probabilidade, desenvolvendo sua forma analítica a partir de certos conceitos básicos sobre fenômenos reais. Tais distribuições tem extensivas aplicações em engenharia, ciência básica, etc. Serão elas subdivididas em discretas e contínuas.

3.1 - Distribuições Discretas de Probabilidade

Relembrando: Uma distribuição discreta de probabilidade é sintetizada por ter:

i) $f(x_i) \geq 0$

ii) $\sum_{i=1}^{n} f(x_i) = 1$

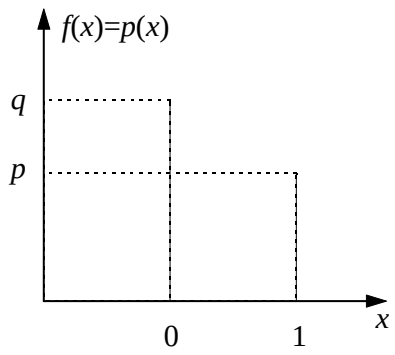
iii) $P(X = x_i) = f(x_i)$

3.1.2 - Provas de Bernoulli e Distribuição de Bernoulli

Existem muitos problemas nos quais os experimentos consistem de *n* provas ou subexperimentos. Aqui o interesse é apenas em uma prova individual onde haja duas possibilidades de resultado: sucesso (*S*) ou Falha (*F*).

Por conveniência, iremos definir uma *v.a.* $X = 1$ se o resultado de uma prova for *S* e $X = 0$ se o resultado for *F*. Assim temos o espaço amostral $E = \{S, F\}$ e:

Resultado da prova	S	F
X	1	0
P(x)	p	q



A distribuição acima, ou

$$\begin{cases} P(X=1)=p \\ P(X=0)=q \end{cases}$$

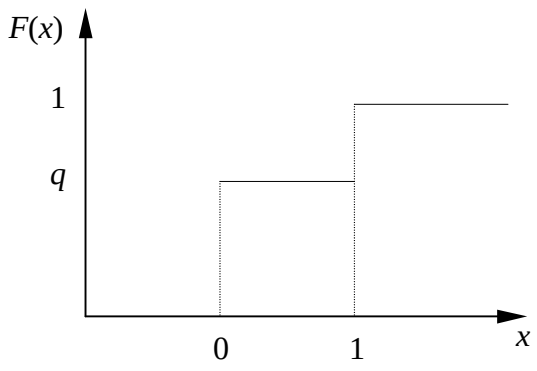
É denominada de *distribuição de Bernoulli*. A sua média e variância são calculadas como segue:

$E(X) = 0 \,.\, q + 1 \,.\, p = p$

$Var(X) = E(X^2)-[E(X)]^2 = [(0^2.q) + (1^2.p)] - p^2 = p\,(1-p) = pq$

A função acumulada $F(x)$ será então:

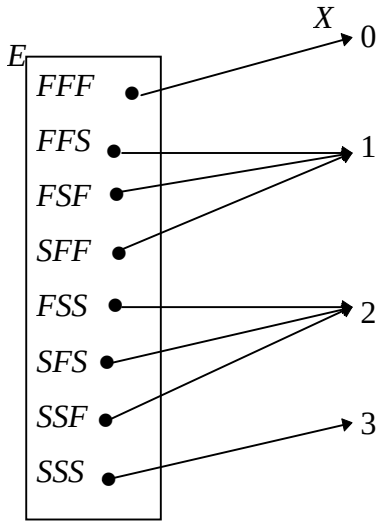
$$F(x)=\begin{cases} 0, \text{ se } x<0 \\ 1-p=q, \text{ se } 0\le x<1 \\ 1, \text{ se } x\ge 1 \end{cases}$$



Observe que nessas provas (provas de Bernoulli) as condições são idênticas nas n provas. Os resultados S e F são mutuamente excludentes com probabilidades constantes p e q (com $p + q = 1$). As provas são independentes umas das outras (o conhecimento do resultado de uma prova, S ou F, não influencia no resultado seguinte).

Ex.: Suponha um experimento que consiste em três provas de Bernoulli com probabilidade p de sucesso em cada prova. A *v.a.* X faz a seguinte associação ao espaço amostral E:

X: Número de sucessos



A distribuição de X pode ser determinada como segue:

x	$P(x)$
0	$P\{FFF\} = P(X = 0) = qqq = q^3$
1	$P\{FFS\} + P\{FSF\} + P\{SFF\} = P(X = 1) = 3pq^2$
2	$P\{FSS\} + P\{SFS\} + P\{SSF\} = P(X = 2) = 3p^2q$
3	$P\{SSS\} = P(X = 3) = p^3$

3.1.3 - A Distribuição Binomial

Uma forma de descrever $P(x)$ para o exemplo anterior pode ser dada por (confira!):

$$P(X = x) = \binom{3}{x} p^x q^{3-x} \qquad x = 0, 1, 2, 3$$

A *v.a. X* que representa o número de sucessos em *n* provas de Bernoulli tem uma distribuição denominada *Binomial* dada por:

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \qquad x = 0, 1, 2, \cdots n$$
$$= 0 \qquad \text{para outros valores}$$

Para que tal distribuição seja de probabilidade devemos ter que

$$\sum_{i=0}^{n} f(x_i) = 1$$

De fato:

$$\sum_{k=0}^{n} P(X = k) = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} = \left[p + (1-p)\right]^n = 1^n = 1$$

Os parâmetros dessa distribuição são *n*, $n \in N$, e *p*, $0 \le p \le 1$. Uma simples demonstração da distribuição Binomial é dada a seguir:

Seja:

P (*X* = *x*) = *P* {"*x* sucessos em *n* provas"}

A probabilidade para um resultado particular onde ocorre *S* nas *x* primeiras provas e *F* nas *n* - *x* provas restantes é

$$P\{\underbrace{SSSS\ldots}_{x}\underbrace{FF\ldots FF}_{n-x}\} = p^x q^{n-x} \quad , (q = 1-p)$$

Existem $\binom{n}{x} = \dfrac{n!}{x!(n-x)!}$ resultados tendo exatamente *x S* e (*n* - *x*) *F*; justificando a distribuição Binomial.

A média de tal distribuição é dada por:

$$E(X) = \sum_{x=0}^{n} x \cdot \frac{n!}{x!(n-x)!} p^x q^{n-x}$$
$$= np \sum_{x=1}^{n} \frac{(n-1)!}{(x-1)!(n-x)!} p^{x-1} q^{n-x}$$

Fazendo *y* = *x* - 1

$$E(X) = np \sum_{y=0}^{n-1} \frac{(n-1)!}{y!(n-1-y)!} p^y q^{n-1-y}$$

que resulta em $\qquad E(X) = np$

Usando uma aproximação similar, podemos calcular a variância como:

$$Var(X) = \sum_{x=0}^{n} \frac{x^2 n!}{x!(n-x)!} p^x q^{n-x} - (np)^2$$
$$= n(n-1)p^2 \sum_{y=0}^{n-2} \frac{(n-2)!}{y!(n-2-y)!} p^y q^{n-2-y} + np - (np)^2$$

que resulta em $Var\ (X) = npq$

Obs.:Uma bem mais simples aproximação poderia ter sido considerar X como a soma de n independentes variáveis aleatórias de Bernoulli, cada uma com média p e variância pq. Teríamos assim:

$$X = X_1 + X_2 + \cdots + X_n$$
$$E(X) = E(X_1) + E(X_2) + \cdots + E(X_n) = p + p + \cdots + p = np$$
$$Var(X) = Var(X_1) + Var(X_2) + \cdots + Var(X_n) = pq + pq + \cdots + pq = npq$$

Ex.: A probabilidade de um teste "Burn in / Burn out" queimar um componente eletrônico é 0,2. Colocando-se três componentes sob teste, qual a probabilidade de que pelo menos dois deles se "queime"?

Adotando (Q) para o componente que se queime e (N) para o caso dele não se queimar, temos o seguinte espaço amostral E:

$E = \{QQQ, QQN, QNQ, NQQ, NNQ, NQN, QNN, NNN\}$

Associando p à probabilidade do componente se queimar e empregando as suposições de independência de um resultado em relação ao outro, temos as seguintes probabilidades associadas ao espaço amostral E:

$\{ \quad (0.2)^3,$

$(0.8).(0.2)^2, (0.8)\ (0.2)^2, (0.8).(0.2)^2,$

$(0.2)(0.8)^2, (0.2)(0.8)^2, (0.2)\ (0.8)^2,$

$(0.8)^3 \quad \}$

Nosso interesse não está dirigido para os resultados individuais de E. Ao contrário, desejamos tão somente conhecer quantos componentes com resultado Q sejam encontrados (não interessando a ordem em que tenham ocorrido). Isto é, desejamos estudar a

v.a X a qual atribui a cada resultado de *E* o número de componentes queimados.

Consequentemente o conjunto dos valores possíveis de *X* é { 0, 1, 2, 3 }, com a seguinte distribuição de probabilidade:

x	$P(x)$
0	$P\{NNN\} = P(X = 0) = q^3 = (0.8)^3$
1	$P\{NNQ\} + P\{NQN\} + P\{QNN\} = P(X = 1) = 3pq^2 = 3(0.2)(0.8)^2$
2	$P\{QQN\} + P\{QNQ\} + P\{NQQ\} = P(X = 2) = 3p^2q = 3(0.2)^2(0.8)$
3	$P\{QQQ\} = P(X = 3) = p^3 = (0.2)^3$

Observe que a soma de *p(x)* é igual a 1 porque tal soma pode ser escrita como $(0.8 + 0.2)^3$.

Para o nosso problema: Calcular a probabilidade de que pelo menos dois deles se queime, temos:

$$P(X \geq 2) = P(X=2) + P(X= 3) = 3p^2q + p^3 = 0.104 = 10{,}4\%$$

Se *n* for pequeno, os termos individuais da distribuição binomial serão relativamente fáceis de calcular. Contudo, se *n* for relativamente grande, os cálculos se tornam bastante incômodos. Felizmente foram preparadas tabelas de probabilidades binomiais; existem várias dessas tabelas. (Veja Tabela 1, em anexo).

Ex.: Suponha que uma válvula eletrônica, instalada em determinado circuito, tenha probabilidade 0.2 de funcionar mais do que 500 horas. Se ensaiarmos 20 válvulas, qual a probabilidade de que delas, exatamente *k*, funcionem mais que 500 horas (*k* = 0, 1, 2, ... 20)?

Seja $X \equiv$ número de válvulas que funcionem mais de 500 horas.

$p = 0.2$

$X = 0, 1, 2, \ldots 20$

Admitindo a distribuição binomial

$$P(X = k) = \binom{20}{k} (0.2)^k (0.8)^{20-k}$$

Os valores podem ser obtidos por tabela para $n = 20$ e $p = 0.20$. Observe que a tabela fornece valores acumulados, assim, por exemplo $P(X = 4) = F(4) - F(3)$ ou seja P (X =

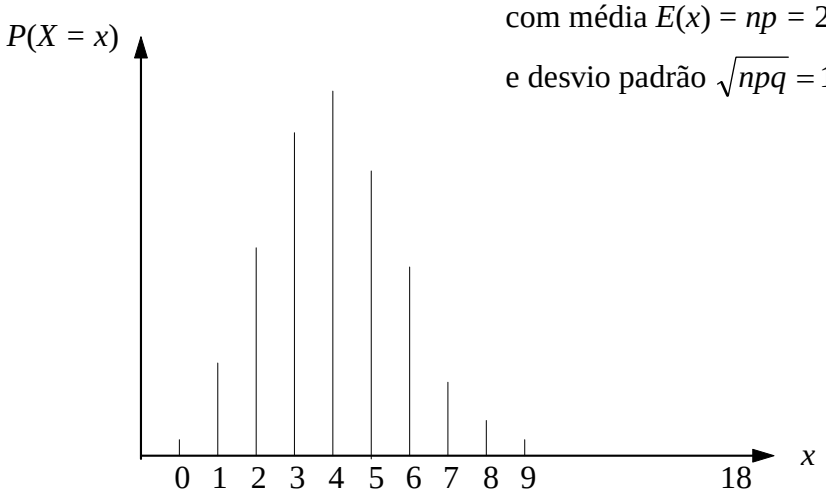
4) = 0.6296 - 0.4114 = 0.2182 (Confira!)

Obs.: A tabela encerra valores até n = 15

Se marcarmos os valores dessa distribuição obteremos o seguinte gráfico:

com média $E(x) = np = 20.(0.2) = 4$

e desvio padrão $\sqrt{npq} = 1.788$



Obs.:

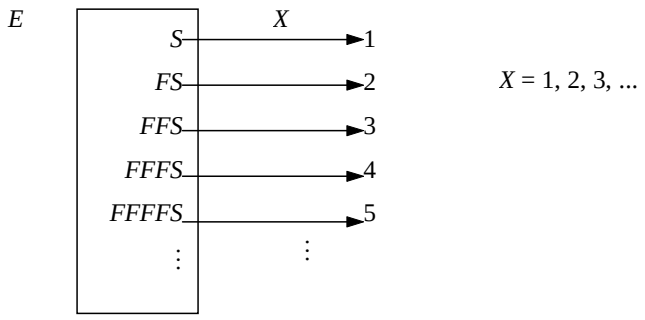
Relembrando: $(x+y)^n = a_0x^n + a_1x^{n-1}y + a_2x^{n-2}y^2 + \cdots + a_{n-1}x\ y^{n-1} + a_ny^n$

$$= \sum_{i=0}^{n} a_i x^{n-i} y^i = \sum_{i=0}^{n} \binom{n}{i} x^{n-i} y^i$$

onde cada a_i é obtido também pelo triângulo de Tartáglia.

3.1.4 - A Distribuição Geométrica

A distribuição geométrica está também associada à seqüência de uma prova de Bernoulli excetuando-se que o número de provas não é fixada, e, na verdade, a *v.a.* de interesse X é definida como o número de provas necessárias para obter o primeiro sucesso. A figura a seguir ilustra a associação entre o espaço amostral e a *v.a.* X:



A distribuição de probabilidade de X é dada por:

$$f(x) = P(X = x) = q^{x-1}p \qquad x = 1,\ 2,\ 3,\ \ldots$$
$$= 0 \qquad \text{outros valores}$$

É fácil verificar que:

$$\sum_{x=1}^{\infty} f(x) = 1 \text{ pois } \sum_{x=1}^{\infty} pq^{x-1} = p\sum_{k=0}^{\infty} q^k = p\left[\frac{1}{1-q}\right] = 1$$

e $f(x) \geq 0$ para todo x.

A Esperança e variância da distribuição geométrica podem ser encontradas como segue:

$$\mu = E(X) = \sum_{x=1}^{\infty} xpq^{x-1} = p\frac{d}{dq}\sum_{x=1}^{\infty} q^x$$
$$= p\frac{d}{dq}\left[\frac{q}{1-q}\right] = \frac{1}{p}$$

$$\sigma^2 = Var(X) = \sum_{x=1}^{\infty} x^2 pq^{x-1} - \left(\frac{1}{p}\right)^2 = p\sum_{x=1}^{\infty} x^2 q^{x-1} - \frac{1}{p}$$
$$= \frac{q}{p^2}$$

Ex.: Um certo experimento é repetido até que um determinado resultado seja obtido. As provas são independentes e o custo de executar um experimento é de $ 25.000. Entretanto, se o resultado a alcançar (Sucesso) não for atingido, um custo de $ 5.000 é necessário para o "setup" da próxima prova. Busca-se determinar o custo esperado do projeto.

Se X é o número de provas necessárias para obter sucesso no experimento, então a

função do custo é dada por:

X	$C(X)$
1	25.000
2	25.000 *2- 5.000
3	25.000* 3-5.000*2
...	...

$C(X)$ = \$ 25.000 X + \$ 5.000 (X - 1)

= (30.000) X - 5.000

Assim:

$E[C(X)] = E[30.000\,X] - E(5.000)$

$$= 30.000\,E(X) - 5.000 = 30.000 \cdot \frac{1}{p} - 5.000$$

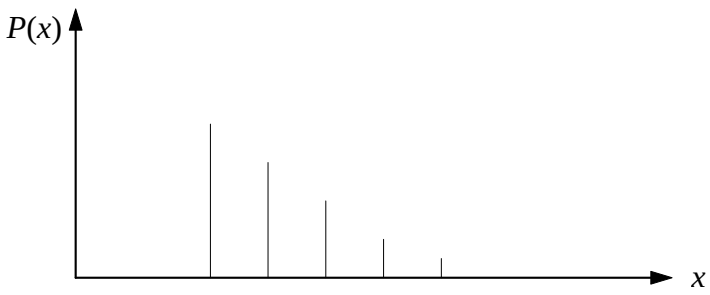
Se a probabilidade de sucesso em uma simples prova é de, digamos 0,25, então E[C(X)] = \$ 115.000.

Suponha agora que se tenha somente \$ 500.000 para investir no experimento. Qual a probabilidade do custo ultrapassar essa quantia?

Temos então que:

$$\begin{aligned}
P\left(C(X) > 500000\right) &= P\left(30000X - 5000 > 500000\right) \\
&= P\left(X > \frac{505000}{30000}\right) \\
&= P\left(X > 16.833\right) \\
&= 1 - P\left(X \le 16\right) \\
&= 1 - \sum_{x=1}^{16} (0.25)(0.75)^{x-1} \\
&= 1 - (0.25)\sum_{x=1}^{16} (0.75)^{x-1} \\
&\cong 0.01 = 1\%
\end{aligned}$$

A distribuição geométrica decresce, isto é, $p(x) < p(x - 1)$ para $x = 2, 3, ...$ Isto é mostrado no seguinte gráfico.



Obs.*:* Uma interessante e útil propriedade da distribuição geométrica é que ela *não tem memória*, isto é,

$$P\ (X > x + s \mid X > s) = P\ (X > x)$$

ou seja: Suponha-se que o evento S não tenha ocorrido durante as primeiras s repetições do experimento. Então a probabilidade de que ele não ocorra durante as próximas x repetições será a mesma probabilidade que não tivesse ocorrido durante as primeiras x repetições. Em outras palavras, a informação de nenhum S é "esquecida", no que interessa aos cálculos subseqüentes.

Exercícios VIII

1 - Um estudo mostra que 70% dos pacientes que vão a uma clínica médica devem esperar no mínimo 15 minutos para serem atendidos. Determinar as probabilidades de que, de oito pacientes, 0, 1, 2, 3, ..., 7, 8 tenham de sujeitar-se àquela espera. Fazer um histograma dessa distribuição de probabilidade.

2 - Utilizando a tábua de probabilidades binomiais, determinar as seguintes probabilidades:

a) 3 sucessos em 9 provas, com $p = 0,4$;

b) 6 falhas em 15 provas, com $p = 0,6$;

c) 5 ou mais sucessos em 8 provas, com $p = 0,5$;

d) entre 7 e 12 (ambos inclusive) sucessos em 15 provas, com $p = 0,7$.

3 - Se há 0,2 de probabilidade de uma pessoa dar crédito a um rumor sobre a vida privada de certo político, qual a probabilidade de que a quinta pessoa a ouvir tal rumor seja a primeira a dar-lhe crédito?

4 - Segundo a teoria de Mendel, no enxerto de duas espécies de plantas com flores amarelas e brancas, 30% das plantas resultantes têm flor amarela. Em sete pares de plantas enxertadas, qual a probabilidade:

a) de não resultar flor amarela;

b) de haver quatro ou mais plantas com flor amarela.

5 - Se a probabilidade de nascimento de menino é 0,52, qual a probabilidade de o terceiro filho de uma família ser o primeiro homem (H)?

6 - Em um grupo de 1500 asilados, 570 têm 75 anos ou mais. Em uma amostra aleatória de 11, qual a probabilidade:

a) de quatro terem 75 anos ou mais;

b) de três ou menos estarem naquela faixa etária.

7 - Se a probabilidade de um eleitor, escolhido aleatoriamente, votar em uma eleição é 0,75, qual a probabilidade de sete entre 10 eleitores comparecerem à eleição

8 - Em 1500 famílias de três filhos, quantas podemos esperar:

a) com ao menos um menino (H);

b) com exatamente duas meninas (M);

c) com nenhuma menina.

9 - Um depósito recebe três lotes de pilhas para lanternas. Sabe-se que as percentagens de pilhas defeituosas no 1º, no 2º e no 3º lote são, respectivamente, 0,015, 0,02 e 0,04. Escolhida ao acaso uma pilha de cada lote, qual a probabilidade de:

a) haver no máximo duas defeituosas;

b) não haver nenhuma defeituosa.

10 - (Bhattacharyva & Johnson) Um laboratório é contratado para fornecer, a um distribuidor, lotes de vacina para gado. Ocasionalmente, algumas vacinas se revelam ineficazes. O distribuidor quer proteger-se contra o risco de receber um número indesejável de vacinas ineficazes. Como não é possível testar todas as vacinas (já que o teste inutiliza a vacina), o distribuidor adota o seguinte processo de seleção: extrai de cada lote uma amostra aleatória de 10 ampolas, contendo um número X de vacinas estéreis. se $X = 0$, o lote é aceito. Se $X \geq 1$, o lote é rejeitado. Este processo é designado *plano de amostragem simples com tamanho de amostra n = 10 e número de aceitação c = 0*.

Admitamos que o tamanho do lote seja suficientemente grande para que a distribuição de X seja (aproximadamente) binomial, com $n = 10$ e p = fração desconhecida de vacinas estéreis em cada lote.

a) Se $p = 0{,}2$ qual a probabilidade de o lote ser aceito?

b) Calcule a probabilidade $P(A)$ de aceitar um lote, para $p = 0{,}05$, $p = 0{,}1$, $p = 0{,}2$, $p = 0{,}4$. Grafe $P(A)$ como função de p, unindo os pontos do gráfico por uma curva contínua (esta curva é chamada *curva característica de operação (CCO) para o plano de amostragem*).

3.2 - Outras Distribuições Discretas de Probabilidade

3.2.1 - A Distribuição de Pascal

A distribuição de *Pascal* também tem a sua base nas provas de Bernoulli. É uma extensão da distribuição geométrica. Nesse caso, a *v.a X* representa o número de repetições necessárias a fim de que um resultado bem sucedido possa ocorrer exatamente *r* vezes. É evidente que se *r* = 1, *X* terá a distribuição geométrica.

Se $X = k$ então o sucesso, de probabilidade *p*, ocorre na *k*-ésima repetição e nas (*k* - 1) repetições anteriores tal sucesso ocorreu (*r* - 1) vezes. A probabilidade desse evento é simplesmente

$$\binom{k-1}{r-1} p^{r-1} . q^{(k-1)-(r-1)} . p \ ,$$

desde que o que acontece nas primeiras (*k* - 1) repetições é independente daquilo que acontece na *k*-ésima repetição. Portanto,

$$f(x) = P(X = k) = \binom{k-1}{r-1} p^{r} \ . q^{k-r} \ , \qquad k = r, r+1, \ldots$$

Ex.: A probabilidade de um bem sucedido lançamento de foguete é igual a 0,8. Suponha que tentativas de lançamento sejam feitas até que tenham ocorrido 3 lançamentos bem sucedidos. Qual a probabilidade de que exatamente 6 tentativas sejam necessárias?

Temos aqui que:

p = 0,8 e o espaço amostral pode ser assim definido:

E = {*SSFFFS, SFSFFS*, --------*S*, ...}

com $\binom{6-1}{3-1} p^2 q^3$ representando a distribuição binomial antes do terceiro sucesso

Assim, temos:

$$P(X = 3) = \binom{5}{2} (0.8)^2 (0.2)^3 (0.8) = \binom{5}{2} (0.8)^3 (0.2)^3$$

Se estivermos interessados em calcular a probabilidade de que menos de 6 tentativas sejam necessárias, temos:

$$P(X<6)=P(X=5)+P(X=4)+P(X=3)$$
$$=\binom{4}{2}(0.8)^3(0.2)^2+\binom{3}{2}(0.8)^3(0.2)^1+\binom{2}{2}(0.8)^3(0.2)^0$$

Obs.: Quando $r = 1$, a distribuição de Pascal reduz-se à distribuição geométrica.

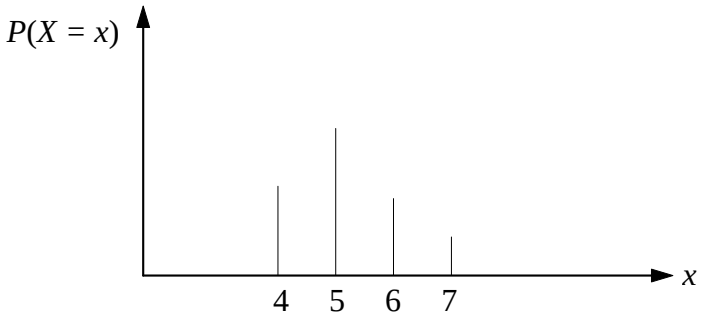
Pode-se provar que para a distribuição de Pascal, a média e a variância são dadas por:

$$E(X)=\frac{r}{p} \text{ e } Var(X)=\frac{rq}{p^2}$$

Ex.: A probabilidade de que um experimento seja bem sucedido é 0,8. Se o experimento for repetido até que quatro resultados bem sucedidos tenham ocorrido, qual será o número esperado de repetições necessárias?

Temos simplesmente que:

$$E(X)=\frac{4}{(0.8)} = 5$$



Obs.: A distribuição de Pascal é conhecida também como distribuição *Binomial Negativa*.

3.2.2 - A Distribuição Multinomial

Vamos examinar agora uma importante *v.a* discreta de maior número de dimensões, a qual pode ser considerada como uma generalização da binomial. Considere um experimento $\mathcal{E}$, seu espaço amostral E, e a partição de E em k eventos mutuamente

excludentes A_1, A_2,... A_k. (isto é, quando $\mathcal{E}$ for realizado, um, e somente um dos eventos A_i ocorrerá.). Considere agora n repetições de $\mathcal{E}$. Seja $p_i = P(A_i)$ e suponha que p_i permaneça constante durante todas as repetições. Evidentemente, teremos:

$$\sum_{i=1}^{k} p_i = 1$$

Definindo as *v.a.* X_1, X_2, ... X_k como:

$X_i \equiv$ número de vezes que A_i ocorre nas n repetições de $\mathcal{E}$, $i = 1, 2, ...k$

Os X_i não são *v.a.* independentes, porque $\sum_{i=1}^{k} X_i = n$. Em conseqüência, logo que os valores de quaisquer (k - 1) dessas *v.a.* sejam conhecidos, o valor da outra ficará determinado. Chegamos então ao seguinte resultado, a distribuição multinomial.

$$P\left(X_1 = n_1, X_2 = n_2, \cdots, X_k = n_k\right) = \frac{n!}{n_1!n_2!\cdots n_k!} p_1^{n_1} p_2^{n_2} \cdots p_k^{n_k}$$

onde $\sum_{i=1}^{k} n_i = n$

Obs.: Se k = 2, a expressão acima se reduz à distribuição binomial. Nesse caso os dois eventos serão denominados de "sucesso" (S) com probabilidade p e "falha"(F) com probabilidade q = 1 - p.

Obs.: A distribuição multinomial também é denominada de *polinomial*.

A média e variância de X_i, uma componente particular, são dadas por:

$E(X_i) = np_i$ e $Var(X_i) = np_i(1 - p_i)$ $\quad i = 1, 2, \cdots k$,

onde X_i é uma componente particular de X.

Ex.: Uma barra de comprimento especificado é fabricada. Admita-se que o comprimento real X (polegadas) seja uma *v.a.* uniformemente distribuída sobre [10,12]. Suponha que se deseje investigar a ocorrência de três eventos:

$A_1 = \{X < 10.5\}$, $A_2 = \{10.5 \le X \le 11.8\}$ e $A_3 = \{X > 11.8\}$ com probabilidades:

$p_1 = P(A_1) = 0.25$, $p_2 = P(A_2) = 0.65$ e $p_3 = P(A_3) = 0.1$

Se 10 dessas barras forem fabricadas, a probabilidade de termos 5 barras de comprimento dado por A_1, e 2 dado por A_3 é obtida por:

$$P(X_1 = 5,\ X_2 = 3,\ X_3 = 2) = \frac{10!}{5!\,3!\,2!}(0.25)^5(0.65)^3(0.1)^2$$

3.2.3 - A Distribuição Hipergeométrica

No estudo da distribuição binomial, vimos que ela se aplica quando a probabilidade de sucesso se mantém constante em todas as provas do experimento. A binomial corresponde, pois, ao esquema de extrações com reposição.

A distribuição hipergeométrica é adequada quando consideramos *extrações casuais* feitas, *sem reposição*, de uma população dividida segundo dois atributos. Aqui as probabilidades de sucesso e falha não se mantém constante.

Um exemplo antes da formalização:

Ex.:Em problemas de controle de qualidade, lotes com $N = 100$ elementos são examinados. O número de elementos com defeito é $r = 10$. Escolhendo-se $n = 5$ peças sem reposição, a probabilidade de não se obter peças defeituosas é, por raciocínio simples:

$$p = \frac{\binom{10}{0}\binom{90}{5}}{\binom{100}{5}} \cong 0.584$$

A probabilidade de se obter 2 defeituosas é:

$$p = \frac{\binom{10}{2}\binom{90}{3}}{\binom{100}{5}} \cong 0.007$$

Generalizando: Considere uma população de N elementos, r dos quais sejam sucessos e $N - r$ sejam falha. Escolhendo ao acaso n elementos da população ($n \leq N$), sem reposição, e adotando X como o número de sucessos obtidos, temos a seguinte distribuição de probabilidade hipergeométrica.

$$P(X=k)=\frac{\binom{r}{k}\binom{N-r}{n-k}}{\binom{N}{n}} \qquad k=0,\ 1,\ 2,\ \cdots$$

Ex.: Pequenos motores elétricos são expedidos em lotes de 50 unidades. Antes que uma remessa seja aprovada, um inspetor escolhe 5 desses motores e os inspeciona. Se nenhum dos motores inspecionados for defeituoso, o lote é aprovado. Se um ou mais forem verificados defeituosos, todos os motores da remessa são inspecionados. Suponha que existam, de fato, três motores defeituosos no lote. Qual a probabilidade de que a inspeção 100% seja necessária?

Se fizermos igual a X o número de motores defeituosos encontrados, a inspeção 100% será necessária se $X \geq 1$.

Logo:

$$P(X \geq 1)=1-P(X=0)=1-\frac{\binom{3}{0}\binom{47}{5}}{\binom{50}{5}} \cong 0.28$$

No cálculo da média e variância, adotamos $p = r / N$ e $q = 1 - p$ e obtemos:

$$E(X)=np \ \text{ e } \ Var(X)=npq\frac{N-n}{N-1}$$

Se $N >> n$, podemos aproximar

$$P(X=k)=\binom{n}{k}p^k(1-p)^{n-k} \qquad \text{(binomial)}$$

Com $E(X) = np$ e $Var(X) = npq$

Obs.: Em geral a aproximação da distribuição hipergeométrica pela binomial é bastante boa se $n / N \leq 0.1$.

3.2.4 - A Distribuição de Poisson

Seja X uma *v.a.* discreta, tomando os seguintes valores: 0, 1, 2, ... Se

$P(X = k) = \dfrac{e^{-\alpha}\alpha^k}{k!}$ $\quad X = 0,\ 1,\ 2, \cdots$ diremos que X tem distribuição de Poisson, com parâmetro $\alpha > 0$.

Observe que a expressão acima é uma distribuição de probabilidade pois:

$$\sum_{k=0}^{\infty} P(X = k) = \sum_{k=0}^{\infty} \frac{e^{-\alpha}\alpha^k}{k!} = e^{-\alpha}e^{\alpha} = 1$$

Uma interessante propriedade dessa distribuição é que a sua média e variância são iguais. Observe:

$$E(X) = \sum_{k=0}^{\infty} \frac{k\, e^{-\alpha}\alpha^k}{k!} = \sum_{k=1}^{\infty} \frac{e^{-\alpha}\alpha^k}{(k-1)!}$$

Fazendo $s = k - 1$ obtemos:

$$E(X) = \sum_{s=0}^{\infty} \frac{e^{-\alpha}\alpha^{s+1}}{s!} = \alpha \sum_{s=0}^{\infty} \frac{e^{-\alpha}\alpha^{s}}{s!} = \alpha$$

Semelhantemente, para a variância, obtemos:

$E(X^2) = \alpha^2 + \alpha$ (Verifique!) e $Var(X) = E(X^2) - [E(X)]^2 = \alpha$

ou seja:

$E(X) = Var(X) = \alpha$

Ex.: Uma complicada peça de maquinaria, quando funciona adequadamente, dá um lucro de C dólares por hora ($C > 2$) a uma firma. Contudo, esta máquina tende a falhar em momentos inesperados e imprevisíveis. Suponha que o número de falhas durante qualquer período de duração t horas, seja uma *v.a.* com distribuição de Poisson com parâmetro t. Se a máquina falhar X vezes durante as t horas o prejuízo sofrido (parada da máquina mais reparação) é igual a ($X^2 + X$) dólares.

Assim, o lucro total P, durante qualquer período de t horas, é igual a:

$P = Ct - (X^2 + X)$, onde X é a *v.a* que representa o número de falhas da máquina.

Se desejarmos escolher o valor de t de tal maneira que o lucro esperado se torne

máximo, teremos:

$E(P) = E(Ct) - E(X^2 + X) = Ct - E(X^2) - E(X)$

Sabendo que $E(X^2) = t + t^2$ e $E(X) = t$,

$E(P) = Ct - 2t - t^2$

Para achar o valor de t para o qual $E(P)$ se torna máximo, derivamos $E(P)$ e fazemos a expressão resultante igual a zero. Assim,

$C - 2 - 2t = 0$,

o que fornece $t = ½ \ (C - 2)$ horas.

Existem essencialmente, duas razões pelas quais a distribuição de Poisson representa papel tão importante em muitas aplicações: Como *aproximação* da Distribuição Binomial e no estudo de "eventos raros" através do *Processo de Poisson*. Vejamos o primeiro caso:

• **<u>Aproximação da Distribuição Binomial</u>**

Seja X uma *v.a.* distribuída binomialmente com parâmetro p (baseado em n repetições de um experimento). Isto é,

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Admita-se que quando $n \to \infty$, p $\to 0$ e $np \to \alpha$.

Nessas condições é possível demonstrar uma importante consideração:

$$\lim_{n\to\infty} P(X = k) = \frac{e^{-\alpha}\alpha^k}{k!}$$

Ou seja, poderemos obter uma aproximação das probabilidades binomiais com a aproximação de Poisson, toda vez que n seja grande e p seja pequeno.

Já verificamos que se X tiver uma distribuição binomial, $E(X) = np$, enquanto se X tiver uma distribuição de Poisson (com parâmetro α), $E(X) = \alpha$. Nessa aproximação adotamos então $\alpha = np$.

Ex.: A probabilidade de um indivíduo ter reação negativa a certa injeção é de 0,001. Determinar a probabilidade de que de 2.000 indivíduos injetados, exatamente 3 tenham

reação negativa.

Usando a distribuição binomial com n = 2.000 e p = 0.001 temos:

$$P(X=3)=\binom{2000}{3}(0.001)^3(0.999)^{1997}$$

O cálculo desses números dá origem a considerável dificuldade. Pela aproximação de Poisson temos:

$$\alpha = np = (2000)(0.001) = 2$$

$$P(X=3)=\frac{e^{-2}2^3}{3!}=0.1804$$

Para determinar a probabilidade de que mais de quatro tenham reação negativa temos:

$$P(X>4)=1-[P(X=4)+P(X=3)+P(X=2)+P(X=1)+P(X=0)]$$

$$=1-\left[\frac{e^{-2}2^4}{4!}+\frac{e^{-2}2^3}{3!}+\frac{e^{-2}2}{1!}+\frac{e^{-2}2^0}{0!}\right]$$

$$=1-e^{-2}\left[\frac{16}{24}+\frac{8}{6}+\frac{4}{2}+2+1\right]=0.055$$

Com o advento das modernas calculadoras, tais cálculos se fazem com extrema rapidez. Contudo, existem tabelas de valores da distribuição de Poisson que facilitam em muito essas operações. Em anexo, estão inseridas essas tabelas e, assim como para a distribuição binomial, são elas acumuladas, ou seja:

$$P(X\le k_0)=\sum_{k=0}^{k_0}\frac{e^{-\alpha}\alpha^k}{k!}$$ (Ver tabela 2)

Ex.: Consideremos um experimento binomial com n = 200, p = 0.04 em que se pede a probabilidade de, no máximo, 5 sucessos.

Temos pela binomial:

$$P(X\le 5)=\sum_{k=0}^{5}\binom{200}{k}(0.04)^k(0.96)^{5-k}$$

Tal cálculo ultrapassa em muito o âmbito das tábuas binomiais usuais, em vista do elevado valor de n. O cálculo direto é impraticável. Usando a aproximação de Poisson temos:

$\alpha = np$ = (200) (0.04) = 8

$P(X \leq 5) = 0.1912$, valor obtido da Tab. 2 na interseção da coluna 8 com o valor 5.

Se desejarmos por exemplo $P(X = 7)$ usando a tabela, temos:

$$P(X = 7) = F(7) - F(6) = 0.4530 - 0.3134 = 0.1396$$

Obs.: A tabela de Poisson encerra valores até $\alpha = 20$.

• O Processo de Poisson

A distribuição de Poisson exerce por si mesma um papel extremamente importante porque ela representa um modelo probabilístico adequado para um grande número de fenômenos observáveis.

Suponhamos que determinados eventos estejam ocorrendo em certos intervalos de tempo e que as hipóteses abaixo sejam válidas:

i) O número de ocorrências de um evento em um intervalo de tempo é independente do número de ocorrências do evento em qualquer outro intervalo disjunto;

ii) A probabilidade de duas ou mais ocorrências simultâneas (seqüenciais) é praticamente zero;

iii) O número médio de ocorrências por unidade de tempo α é constante ao longo do tempo.

Temos então aqui um processo de Poisson. Uma apresentação mais formal de tal processo pode ser vista na Referência 1.

Ex.: Em uma experiência de laboratório passam, em média, por um contador, quatro partículas radioativas por milissegundo. Qual a probabilidade de entrarem no contador seis partículas em determinado milissegundo?

Utilizando a distribuição de Poisson com $\alpha = 4$, temos então que:

$$P(X = 6) = \frac{e^{-4} 4^6}{6!} = 0.1042$$

(Verifique que as condições do processo foram satisfeitas.)

Ex.: Chegam, em média, 10 navios-tanque por dia a um movimentado porto, que tem capacidade para 15 desses navios. Qual a probabilidade de que, em determinado dia, um

ou mais navios tanque tenham de ficar ao largo, aguardando vaga?

Temos aqui que, para $\alpha = 10$:

$$P(X > 15) = 1 - P(X \leq 15) = 1 - 0.9513 = 0.0487$$

Ex.: Uma central telefônica recebe em média 300 chamadas por hora e pode processar no máximo 10 ligações por minuto. Estimar a probabilidade de a capacidade da mesa ser ultrapassada.

Temos agora:

$\alpha = 300/60 = 5$ chamadas/minuto em média

$$P(X > 10) = 1 - P(X \leq 10) = 1 - 0.986 = 0.014 = 1{,}4\%$$

É importante compreender que a distribuição de Poisson surgiu como uma conseqüência de algumas hipóteses que fizemos. Isto significa que sempre que essas hipóteses sejam válidas (ou ao menos aproximadamente válidas), a distribuição de Poisson pode ser empregada como um modelo adequado. Uma grande classe de fenômenos podem empregar esse modelo.

***Obs.*:** O processo de Poisson pode ser aplicado não somente ao número de ocorrências de um evento, durante um período de tempo fixado, mas também ao número de ocorrências de um evento dentro das fronteiras de uma área ou volume fixados.

Ex.: Suponha que um astrônomo investigue uma parte da Via-Láctea e admita que, na parte considerada, a densidade de estrelas λ seja constante. [Isto significa que em um volume V (unidades cúbicas), podem ser encontradas, em média, λV estrelas]. Seja X_V o número de estrelas encontradas em uma parte da Via-Láctea que tenha o volume V. Se as hipóteses do Processo de Poisson forem satisfeitas (com "volume" substituindo "tempo"), então

$$P(X_V = n) = \frac{e^{-\alpha}\alpha^n}{n!} \text{ com } \alpha = \lambda V$$

As hipóteses interpretadas neste contexto, afirmariam essencialmente que o número de estrelas que aparecem em partes não sobrepostas da Via-Láctea, representam *v.a.* independentes e que a probabilidade de mais do que uma estrela aparecer em uma parte bastante pequena do céu é zero.

Exercícios IX

1 - Nos sinais de um transmissor ocorrem distorções aleatórias a uma taxa média de 1 por minuto. Para mensagens de 3 minutos, a taxa média de ocorrência de distorção é de 3 por mensagem. Qual a probabilidade de o número de distorções em uma mensagem de 3 minutos ser:

a) 2;

b) 1;

c) 3 ou mais.

2 - Na fabricação de resistores de 50 ohms são considerados bons os que têm resistência entre 45 e 55 ohms. Sabe-se que a probabilidade de um deles ser defeituoso é 0,2%. Os resistores são vendidos em lotes de 100 com a garantia de serem todos bons. Qual a probabilidade de um lote violar esta garantia?

3 - Uma central telefônica recebe em média 300 chamadas por hora e pode processar no máximo 10 ligações por minuto. Utilizando a distribuição de Poisson, estimar a probabilidade de a capacidade da mesa ser ultrapassada.

4 - A probabilidade de uma pessoa sofrer intoxicação alimentar na lanchonete de um parque de diversões é 0,001. Com auxílio da aproximação de Poisson, determinar a probabilidade de que, em 1500 pessoas que passam o dia no parque, no máximo duas tenham intoxicação.

5 - Uma urna contém duas bolas brancas e seis pretas. extrai-se uma, anota-se a cor e devolve-se a bola à urna. Determinar a probabilidade de aparecer bola branca exatamente três vezes em oito extrações.

6 - Jogam-se dois dados muitas vezes. Determinar a probabilidade de a soma 7 aparecer pela primeira vez na sétima jogada.

7 - Jogam-se dois dados muitas vezes. determinar a probabilidade de a soma 7 aparecer pela quarta vez na sétima jogada.

8 - Conhecem-se a seguintes probabilidades:

0,6, de uma declaração de imposto de renda ser preenchida corretamente;

0,2, de uma declaração conter somente erros que favorecem o contribuinte;

0,1, de uma declaração conter somente erros que favorecem o fisco;

0,1, de uma declaração conter ambos os tipos de erro.

Calcular a probabilidade de que, em 10 declarações selecionadas aleatoriamente, seis estejam corretas, duas contenham erros a favor do contribuinte, uma contenha erros a favor do fisco e uma contenha ambos os tipos de erro.

9 - Um agricultor planta seis sementes escolhidas aleatoriamente de uma caixa com cinco sementes de tulipa e quatro de crisântemo

Qual a probabilidade de ele plantar duas sementes de crisântemo e quatro de tulipa?

3.3 - Distribuições Contínuas de Probabilidade

Uma distribuição contínua de probabilidade (ou função densidade de probabilidade - *fdp*) tem, como sabemos, as seguintes propriedades:

i) $f(x) \geq 0$

ii) $\int_{-\infty}^{\infty} f(x) = 1$

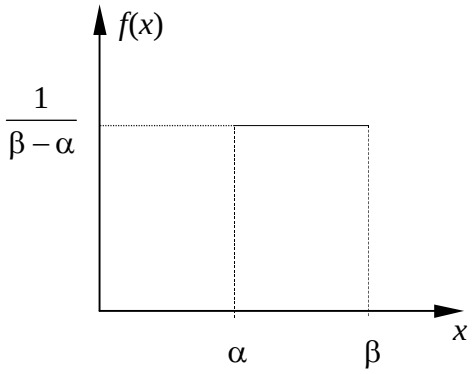
iii) $P(a \leq X \leq b) = \int_a^b f(x)dx \qquad (b > a)$

3.3.1 - Distribuição Uniforme

A *fdp* é aqui definida por:

$$f(x) = \frac{1}{\beta - \alpha}, \quad \alpha \leq x \leq \beta \quad \text{com } \alpha \text{ e } \beta \in \text{R}, \ \alpha < \beta$$
$$= 0 \quad , \quad \text{outros valores}$$

com o seguinte gráfico representativo:



Observe que as propriedades das *fdp* são satisfeitas com a definição acima. Isto é:

$$\int_{-\infty}^{\infty} f(x)dx = \int_{\alpha}^{\beta} \frac{dx}{\beta - \alpha} = 1$$

e $$P(a \leq X \leq b) = \int_a^b \frac{dx}{\beta - \alpha} = \frac{b - a}{\beta - \alpha}$$

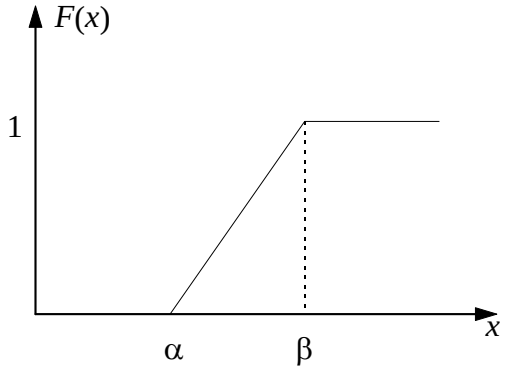
A média e variância da distribuição uniforme são:

$$E(X) = \int_{\alpha}^{\beta} \frac{xdx}{\beta - \alpha} = \frac{1}{2(\beta - \alpha)} x^2 \Big|_{\alpha}^{\beta} = \frac{\beta + \alpha}{2}$$

e $$Var(X) = \int_{\alpha}^{\beta} \frac{x^2 dx}{\beta - \alpha} - \left[\frac{\beta + \alpha}{2}\right]^2 = \frac{(\beta - \alpha)^2}{12}$$

A função de distribuição acumulada é dada por

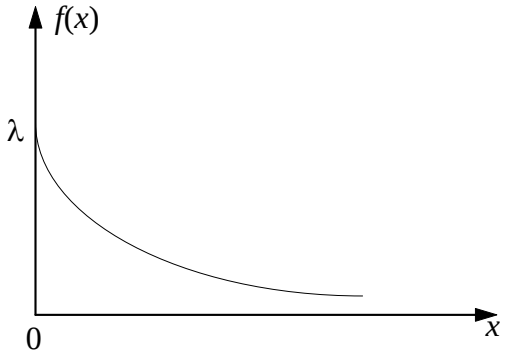
$$\begin{aligned} F(x) &= 0 && , \ x < \alpha \\ &= \int_{\alpha}^{x} \frac{dx}{\beta - \alpha} = \frac{x - \alpha}{\beta - \alpha} && , \ \alpha \le x \le \beta \\ &= 1 && , \ x \ge \beta \end{aligned}$$



3.3.2 - Distribuição Exponencial

Temos agora,

$$\begin{aligned} f(x) &= \lambda e^{-\lambda x} && , \ x \ge 0 && \text{com } \lambda \in \mathrm{R}, (\lambda > 0) \\ &= 0 && , \ \text{outros valores} \end{aligned}$$



Uma integração imediata mostra que

$$\int_0^\infty \lambda\, e^{-\lambda x} dx = -e^{-\lambda x}\Big|_0^\infty = 1$$

A média e variância da distribuição exponencial são dadas por:

$$E(X) = \int_0^\infty x\lambda\, e^{-\lambda x} dx = -x\, e^{-\lambda x}\Big|_0^\infty + \int_0^\infty e^{-\lambda x} dx = \frac{1}{\lambda}$$

(usou-se integração por partes)

e

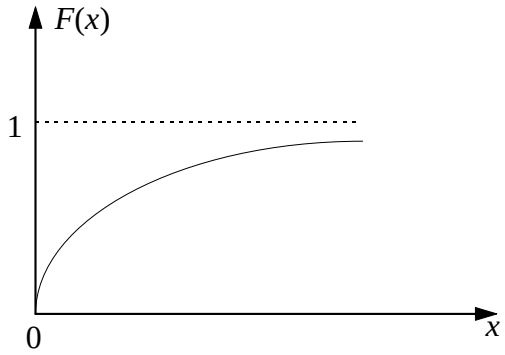
$$\begin{aligned} Var(X) &= \int_0^\infty x^2 \lambda\, e^{-\lambda x} dx = -\left(\frac{1}{\lambda}\right)^2 \\ &= \left[-x^2\, e^{-\lambda x}\Big|_0^\infty + 2\int_0^\infty x e^{-\lambda x} dx \right] - \left(\frac{1}{\lambda}\right)^2 \\ &= \frac{1}{\lambda^2} \end{aligned}$$

Observe a interessante propriedade da distribuição exponencial:

$$E(X) = DP(X) = 1/\lambda$$

A função acumulada $F(x)$ é dada por:

$$\begin{aligned} F(x) &= 0 && ,\ x < 0 \\ &= \int_0^x \lambda\, e^{-\lambda t} dt = 1 - e^{-\lambda x} && ,\ x \geq 0 \end{aligned}$$

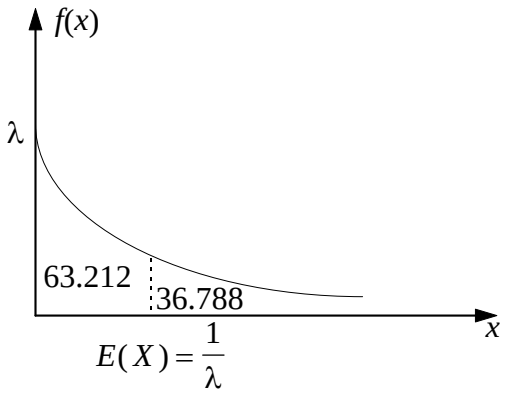


Ex.: Um componente eletrônico é conhecido por ter sua vida útil representada por uma fdp exponencial com tempo médio de falha $E(X)$ de 10^5 horas (logo $\lambda = 10^{-5}$). Suponha que desejamos determinar a fração de componentes que poderão falhar antes da vida média ou valor esperado.

Temos então,

$$P\left(T \le \frac{1}{\lambda}\right) = \int_0^{1/\lambda} \lambda\, e^{-\lambda x} dx = -e^{-\lambda x}\Big|_0^{1/\lambda} = 1 - e^{-1}$$
$$= 0.63212$$

Esse resultado indica que 63,212% dos componentes irão falhar antes de 10^5 horas.



Ex.: Suponha que um fusível tenha duração de vida *X*, a qual pode ser considerada uma *v.a.* contínua com distribuição exponencial. Existem dois processos pelos quais o fusível pode ser fabricado. O processo I apresenta uma duração de vida esperada de 100 horas (isto é, $\lambda = 100^{-1}$), enquanto o processo II apresenta uma duração de vida esperada de 150 horas ($\lambda = 150^{-1}$). Suponha que o processo II seja duas vezes mais custoso (por fusível) que o processo I, que custa *C* dólares por fusível. Admita-se, além disso, que se um fusível durar menos que 200 horas, uma multa de *K* dólares seja lançada? Obs.: ($C > K$)

Vamos calcular o custo esperado para cada processo. Para o processo I, teremos

C_1 = custo (por fusível) $= C$ se $X > 200$

$= C + K$ se $X \le 200$

logo,
$$E(C_1) = C\,P(X > 200) + (C + K)\,P(X \le 200)$$
$$= Ce^{-(1/100)200} + (C + K)\left(1 - e^{-(1/100)200}\right)$$
$$= Ce^{-2} + (C + K)(1 - e^{-2})$$
$$= K(1 - e^{-2}) + C$$

por um cálculo semelhante encontraremos

$$E(C_{II}) = K\left(1 - e^{-4/3}\right) + 2C$$

Portanto $E(C_{II}) - E(C_I) = C + K\left(e^{-2} - e^{-4/3}\right) = C - 0.13K$

Conseqüentemente, preferiremos o processo I, visto que $C > 0.13K$.

3.3.3 - A Distribuição Gama

A função usada na definição da distribuição gama é definida por:

$$\Gamma(n) = \int_0^\infty x^{n-1} e^{-x} dx \qquad n > 0 \ ,$$

denominada de função gama.

Uma importante propriedade de tal função é que:

$\Gamma(n) = (n-1)\Gamma(n-1)$, ou seja

$\Gamma(n) = (n-1)!$

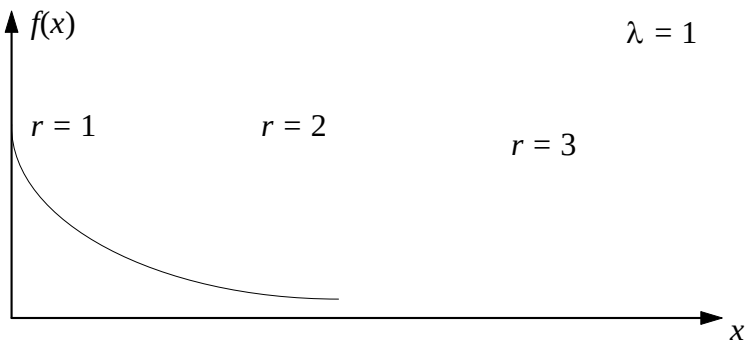
desde que $\Gamma(1) = \int_0^\infty e^{-x} dx = 1$ e $n > 0$.

***Obs.*:** Podemos verificar que $\Gamma(1/2) = \int_0^\infty x^{-1/2} e^{-x} dx = \sqrt{\pi}$

Com o uso da função gama, podemos introduzir a distribuição de probabilidade gama como:

$$f(x) = \frac{\lambda}{\Gamma(r)} (\lambda x)^{r-1} e^{-\lambda x} \qquad , \ x > 0$$
$$= 0 \qquad , \ \text{para outros valores}$$

Os parâmetros de tal fdp são $r > 0$ e $\lambda > 0$. O parâmetro r é chamado de parâmetro de forma" e λ "parâmetro de escala". A figura seguinte mostra uma distribuição gama para $\lambda = 1$ e $r = 1$, 2 e 3.



Pode ser observado que $f(x) \geq 0$ para todo x e

$$\int_{-\infty}^{\infty} f(x)dx = \int_0^x \frac{1}{\Gamma(r)} (\lambda x)^{r-1} e^{-\lambda x} dx$$
$$= \frac{1}{\Gamma(r)} \int_0^\infty y^{r-1} e^{-y} dy = \frac{1}{\Gamma(r)} \Gamma(r) = 1$$

***Obs.*:** A distribuição exponencial é um caso particular da distribuição gama, pois

para r = 1, $f(x) = \lambda e^{-\lambda x}$. Para r = v/2 e λ = 1/2 teremos uma importante distribuição chamada $\chi^2_{(v)}$ (Qui-quadrado) que será vista adiante, em estimação de parâmetros.

Pode ser mostrado que a média e variância da distribuição gama são dadas por:

$$E(X) = r/\lambda \quad \text{e} \quad Var(X) = r/\lambda^2$$

e

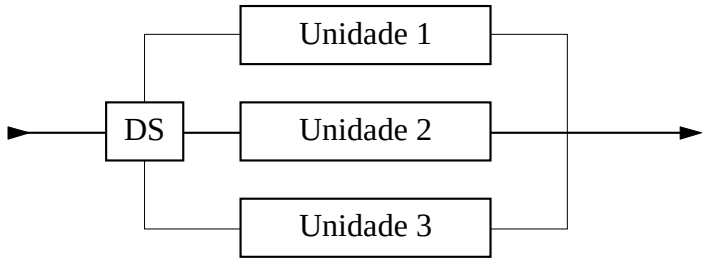
$$F(x) = 1 - \sum_{k=0}^{r-1} e^{-\lambda x} \frac{(\lambda x)^k}{k!} \quad x > 0,\ \text{r} > 0$$

Uma bem conhecida utilização da distribuição Gama decorre da seguinte proposição:

" Se uma variável aleatória X é a soma de r independentes distribuições exponenciais, com parâmetro λ, então X tem uma distribuição Gama com parâmetros λ e r. "

Ex.: Um sistema opera como mostrado na figura seguinte. Inicialmente, a unidade 1 está ativa enquanto as unidades 2 e 3 estão inativas ("stand by"). Quando a unidade 1 falha, um determinado equipamento de decisão (DS) ativa a unidade 2. Caso a unidade 2 falhe, a unidade 3 é, do mesmo modo, ativada. O equipamento de decisão é assumido perfeito e o tempo de vida do sistema pode ser representado como a soma dos tempos de vida das unidades ($X = X_1 + X_2 + X_3$). Se os tempos de vida das unidades são independentes um do outro e se as unidades tem cada uma um tempo de vida (X_j, j = 1, 2, 3), tendo densidade $g(x_j) = (1/100)e^{-x_j/100}$, então X terá uma distribuição gama com r = 3 e λ = 0.01. Isto é,

$$f(x) = \frac{0.01}{2!}(0.01x)^2 e^{-0.01x}$$



A probabilidade de que o sistema opere pelo menos x horas é dado por $R(x)$ e é chamado de "*função de confiabilidade*", dado por:

$R(x)$=1-$F(x)$

Aqui,

$$R(x) = 1 - F(x) = \sum_{k=0}^{2} \left[e^{-0.01x} (0.01x)^k \right] / k!$$
$$= e^{-0.01x} \left[1 + (0.01x) + \frac{(0.01x)^2}{2} \right]$$

***Obs.*:** A distribuição gama é muito útil na *teoria da confiabilidade*.

3.3.4 - A Distribuição de Weibull

A distribuição de Weibull tem sido largamente utilizada em muitos fenômenos randômicos. A principal utilidade de tal distribuição é que ela proporciona uma excelente aproximação para ser a fdp de muitas variáveis aleatórias.

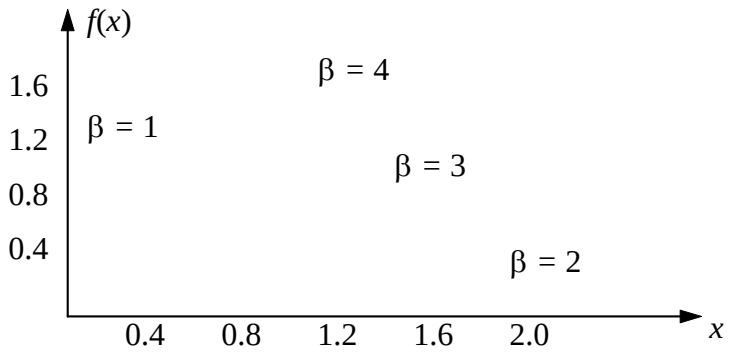
Em resumo, as propriedades da distribuição de Weibull são:

- $f(x) = \frac{\beta}{\delta} \left(\frac{x-\gamma}{\delta} \right)^{\beta-1} \exp\left[-\left(\frac{x-\gamma}{\delta} \right)^{\beta} \right], x \geq \gamma$

com $-\infty < \gamma < \infty$, $\delta > 0$ e $\beta > 0$

- $E(X) = \gamma + \delta\Gamma\left(1 + \frac{1}{\beta}\right)$
- $Var(X) = \delta^2 \left[\Gamma\left(1 + \frac{2}{\beta}\right) - \Gamma\left(1 + \frac{1}{\beta}\right)^2 \right]$
- $F(x) = 1 - \exp\left[-\left(\frac{x-\gamma}{\delta} \right)^{\beta} \right]$

A figura seguinte mostra algumas dessas distribuições para $r = 0$, $\delta = 1$ e $\beta = 1, 2, 3$ e 4. Note que quando $\beta = 1$ temos a distribuição exponencial com $\lambda = 1/\delta$



Exercícios X

1 - A dureza de uma peça de cerâmica é proporcional ao tempo de queima. Suponha que se conseguiu um processo de medição dessa dureza e que a medida respectiva seja uma *v.a.* distribuída uniformemente entre 0 e 10. Se a dureza de uma peça de cozinha deve estar no intervalo [5: 9], qual a probabilidade de uma peça escolhida ao acaso ser adequada ao uso na cozinha?

2- Um geólogo define o granito como uma rocha que contém quartzo, feldspato e pequena quantidade de outros minerais, desde que não contenha mais de 75% de quartzo. Se todas as percentagens são igualmente prováveis, qual a proporção das amostras coletadas pelo geólogo em sua vida profissional que conterão de 50 a 65% de quartzo?

3 - Uma companhia de seguros segurou um grupo de 10.000 indivíduos de 28 anos de idade. Seus seguros entraram em vigor em 31 de agosto e o ano fiscal termina em 31 de dezembro. Pela tábua de mortalidade adotada, a probabilidade de um indivíduo de 28 anos morrer dentro de um ano é 0,0015. Se o valor do seguro é de R$ 5.000,00 por morte, e se as mortes se distribuem uniformemente durante o ano, qual a reserva que a companhia deve constituir para fazer face aos sinistros que ocorrerem até o fim do ano fiscal?

4 - Um montanhista está perdido no flanco de uma montanha, fazendo sinal com sua lanterna para atrair a atenção de uma patrulha de salvamento. Se o tempo, em horas, que a lanterna é capaz de emitir um feixe de luz visível embaixo é uma *v.a.* exponencial com parâmetro $\lambda = \frac{1}{6}$:

a) qual a probabilidade de o sinal luminoso ser visto por ao menos 3 horas?

b) construa uma tábua mostrando a resposta da parte (a) quando $= \frac{1}{\lambda} = 1,\ 2,\ 3,\ 4,\ 5,\ 6.$

5 - O tempo de uma conversa telefônica tem distribuição exponencial com $\lambda = 1/5$. Se uma pessoa está com pressa de fazer uma ligação e outra pessoa chega antes dela à cabine, qual a probabilidade de a primeira pessoa ter de esperar menos de 3 minutos para poder fazer a ligação?

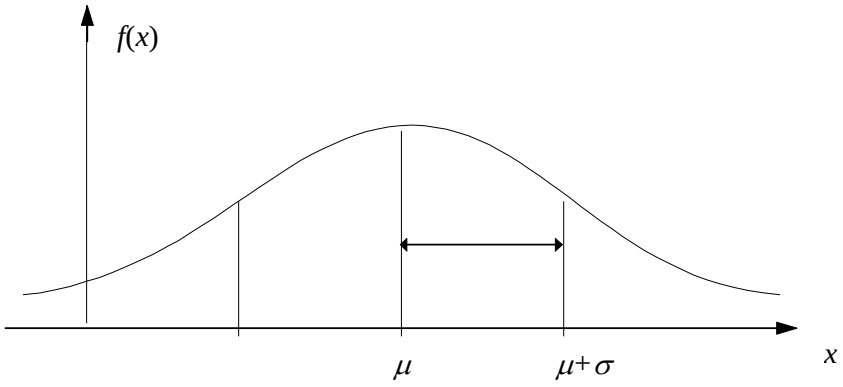
3.4 - A Distribuição Normal

Uma variável aleatória X possui uma distribuição Normal (ou Gaussiana) se sua *fdp* for da forma

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(1/2)\left[\frac{x-\mu}{\sigma}\right]^2} \qquad , \quad -\infty < x < \infty$$

Os parâmetros μ e σ devem satisfazer às condições $-\infty < \mu < \infty$, $\sigma > 0$. X terá a notação $X: N(\mu;\sigma)$ ou $X: N(\mu;\sigma^2)$.

A figura seguinte, a curva do sino ("the bell curve"), representa tal distribuição.



As principais propriedades de $f(x)$ são:

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

b) $f(x) \geq 0$

c) $\lim_{x\to\infty} f(x) = 0$ e $\lim_{x\to-\infty} f(x) = 0$

d) $f(\mu + x) = f(\mu - x)$

e) Máx $f(x)$ ocorre em $x = \mu$

f) Os pontos de inflexão são $x = \mu \pm \sigma$

g) $E(X) = \mu$

h) $Var(X) = \sigma^2$

***Obs.*:** As demonstrações dessas propriedades são dadas na Referência 1.

Tabela Φ(*z*)

A função de distribuição acumulada $F(x)$ é dada por

$$F(x) = P(X \le x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left[\frac{u-\mu}{\sigma}\right]^2} du$$

É impossível avaliar essa integral sem ajuda de métodos numéricos, e para cada avaliação há a necessidade do par de valores (μ;σ). Entretanto, uma simples transformação de variáveis,

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

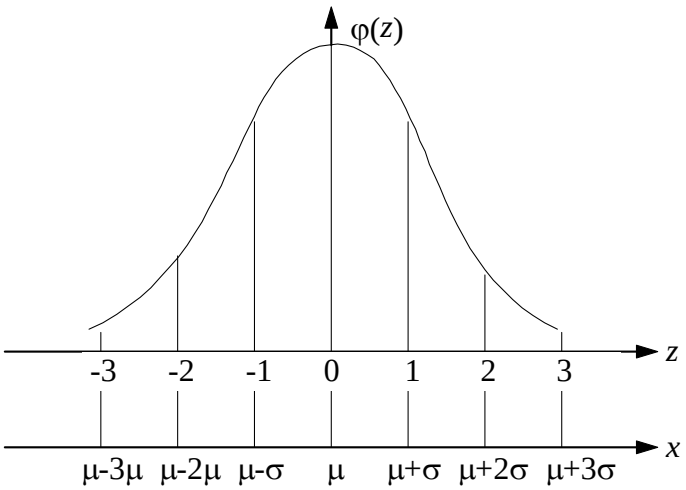
faz com que haja a necessidade apenas do parâmetro *z*, isto é,

$$F(x) = P(X \le x) = P\left(Z \le \frac{x-\mu}{\sigma}\right) = \int_{-\infty}^{\frac{x-\mu}{\sigma}} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz$$
$$= \int_{-\infty}^{\frac{x-\mu}{\sigma}} \varphi(z)\, dz = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

A *fdp*

$$\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \qquad , \ -\infty < z < \infty$$

tem média μ = 0 e variância $\sigma^2 = 1$, isto é *Z*: *N*(0; 1), e dizemos que *Z* tem uma "Distribuição Normal Padronizada". O gráfico de φ (*z*) e a transformação efetuada é mostrada na figura seguinte.



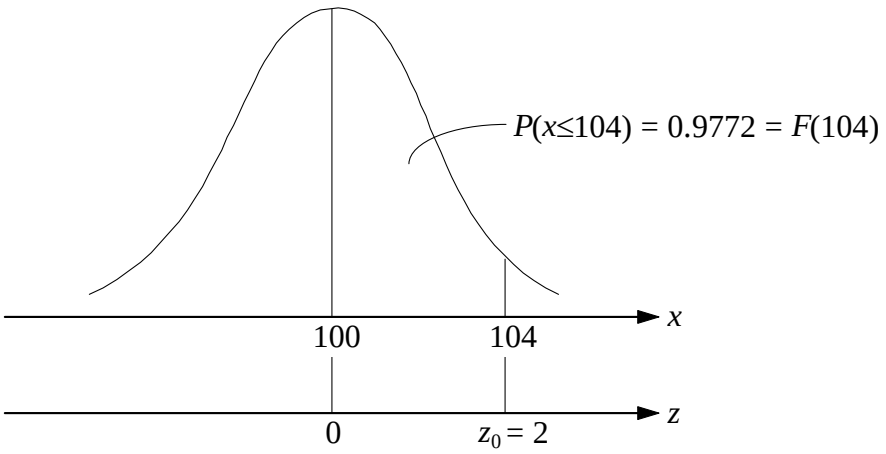
A função acumulada para *Z* é dada por:

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du$$

e tal função tem sido tabelada (Tabela 3).

O procedimento usado na solução de problemas envolvendo consulta a tabela normal reduzida é bastante simples.

Ex.: Suponha que X: $N(100; 2)$ e que desejamos avaliar $P(X \leq 104)$. Temos então



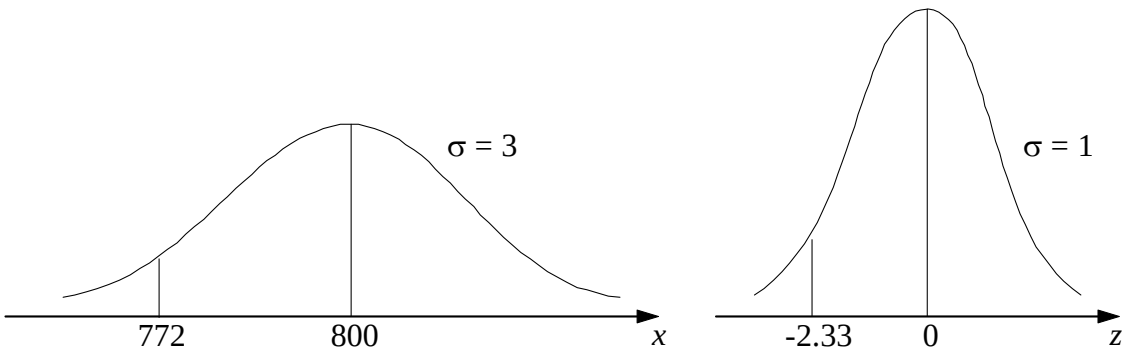
Pois $z_0 = \dfrac{104-100}{2} = 2$ e $\Phi(2) = 0.9772$ (por tabela)

Observe a relação entre o argumento de F e o argumento de Φ, o qual mede a distância de x à média μ em unidades de desvio padrão (σ). No caso considerado $F(104) = \Phi(2)$ indica que 104 está dois desvios padrões ($\sigma = 2$) acima da média. Na solução de problemas, algumas vezes necessitamos usar a propriedade de simetria de $\varphi(z)$.

Ex.: A tensão de ruptura (em newtons) de uma fibra sintética é representada por X e distribuída como $N(800; 12)$. O controle de qualidade na fabricação da fibra exige uma tensão de no mínimo 772 N. Uma amostra da fibra é randomicamente testada. A probabilidade de obtermos $P(X \geq 772)$ é obtido a partir de:

$$\begin{aligned} P(X < 772) &= P\left(\frac{x-\mu}{\sigma} < \frac{772-800}{12}\right) \\ &= P(Z < -2.33) \\ &= \Phi(-2.33) = 0.01 \end{aligned}$$

Assim, a probabilidade desejada, $P(X \geq 772)$ é igual a $1 - P(X < 772) = 0.99$ como mostra a figura seguinte.



Ex.: O tempo requerido para reparar uma máquina de empacotamento automático de alimentos em um complexo processo de produção é X minutos. Estudos mostram que $X \sim N(120, 4)$ é uma boa aproximação para X. Se o processo é interrompido por mais de 125 minutos todos os equipamentos devem ser limpos com a perda dos alimentos em processo. O custo total desta perda é aproximadamente $ 10.000. No sentido de determinar a probabilidade desta ocorrência podemos proceder como a seguir:

$$\begin{aligned} P(X > 125) &= P\left(Z > \frac{125-120}{4}\right) = P(Z > 1.25) \\ &= 1 - \Phi(1.25) = 1 - 0.8944 \\ &= 0.1056 \end{aligned}$$

Assim, dada uma falha na máquina, o custo esperado $E(C)$ será dado por:

$E(C) = 0.1056\ (10000 + C_{R1}) + 0.8944\ C_{R1}$

onde

C representa o custo total

C_{R1} representa o custo de reparo

Simplificando temos:

$E(C) = C_{R1} + 1056$

Suponha que um gerenciamento pode reduzir a média da distribuição do tempo de reparo para 115 minutos pela adição de mais pessoal de manutenção. O novo custo de reparo será $C_{R2} > C_{R1}$.

Temos agora que:

$$\begin{aligned} P(X > 125) &= P\left(Z > \frac{125-115}{4}\right) = P(Z > 2.5) \\ &= 1 - \Phi(2.5) = 1 - 0.9938 \\ &= 0.0062 \end{aligned}$$

fazendo com que o custo esperado seja:

$E(C) = C_{R2} + 62$

Poderíamos agora avaliar se esta segunda alternativa é melhor por meio da confirmação de que:

C_{R2} + 62 < C_{R1} + 1056 ou

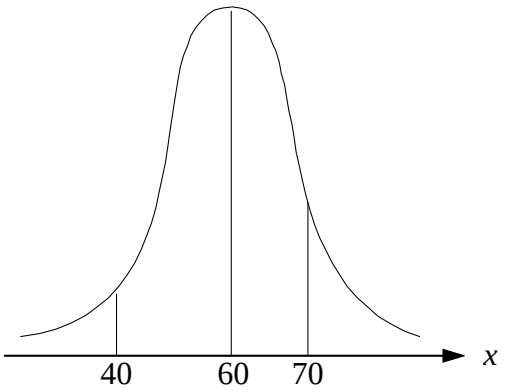
C_{R2} - C_{R1} < \$ 994

Ex.: Suponha que as freqüências indesejáveis para um determinado sinal elétrico tenham uma variação normal com média 60 Hz e desvio padrão 15 Hz.

A) Qual a probabilidade desse sinal elétrico possuir componentes entre 40 e 70 Hz devido a essas freqüências indesejáveis?

Temos que:

$$\begin{aligned} P(40 \le X \le 70) &= P(-1.33 < Z < 0.67) \\ &= \Phi(0.67) - \Phi(-1.33) \\ &= 0.748 - 0.09 \\ &= 0.65 = 65\% \end{aligned}$$



B) Qual a maior freqüência do sinal para que a probabilidade de contaminação por freqüências indesejáveis seja de 10%?

Nesse caso temos um outro caminho:

$$P(Z < z_0) = P(X < x_0) = 0.10 \text{ , e}$$

pela tabela encontramos que z_0 = - 1.27. Assim:

$$Z = \frac{X - \mu}{\sigma} \Rightarrow -1.27 = \frac{x_0 - 60}{15} \Rightarrow x_0 \cong 40.9 \text{ Hz}$$

***Obs.*:** Existem alguns intervalos simétricos da curva normal que são usados freqüentemente. São eles:

$P(\mu - 1.00\ \sigma \leq X \leq 1.00\ \sigma) = 0.6826$

$P(\mu - 1.645\ \sigma \leq X \leq \mu + 1.645\ \sigma) = 0.90$

$P(\mu - 1.96\ \sigma \leq X \leq \mu + 1.96\ \sigma) = 0.95$

$P(\mu - 2.57\ \sigma \leq X \leq \mu + 2.57\ \sigma) = 0.99$

$P(\mu - 3.00\ \sigma \leq X \leq \mu + 3.00\ \sigma) = 0.9978$

***Obs.*:** Combinação de Distribuições Normais:

"A combinação linear de variáveis aleatórias com Distribuições Normais independentes é também uma distribuição Normal."

Por exemplo: Se $X{:}N(\mu_x, \sigma_x)$ e $Y{:}N(\mu_y, \sigma_y)$ então

$$W = aX + bY + c$$

terá distribuição normal com:

$\mu_w = a\mu_x + b\mu_y + c$ e

$$\sigma_w^2 = a^2\sigma_x^2 + b^2\sigma_y^2$$

Exercícios XI

1 - Se X é uma *v.a.* $N(95; 7{,}5)$, determinar b tal que:

a) $P(X > b) = 0{,}975$;

b) $P(X > b) = 0{,}8708$

2 - A vida média de certo aparelho é de oito anos, com d.p. de 1,8 ano. O fabricante substitui os aparelhos que acusam defeito dentro do prazo de garantia. Se ele deseja substituir no máximo 5% dos aparelhos que apresentem defeito, qual deve ser o prazo de garantia?

3 - Um processo industrial produz peças com diâmetro médio de 2.00” e d.p. de 0,01”. As peças com diâmetro que se afaste da média por mais de 0,03” são consideradas defeituosas. admitida a normalidade:

a) qual a percentagem das peças defeituosas?

b) qual a probabilidade de encontrar duas peças defeituosas em seqüência?

c) qual probabilidade de encontrar duas peças perfeitas em seqüência?

4 - Em um exame de estatística, a nota média foi 70, com d.p. 4,5. Todos os alunos que obtiveram nota 75 a 89 receberam conceito *B*. Se as notas têm distribuição aproximadamente normal, e se 10 estudantes obtiveram conceito *B*, quantos se submeteram ao exame?

5 - As alturas de 500 estudantes de uma faculdade têm média 1,70 m e d.p. 0,75 m. Supõe-se normal a distribuição das alturas.

a) Quantos têm altura inferior a 1,75 m?

b) Entre 1,72 e 1,80 m?

c) Igual a 1,74 m?

6 - Se um conjunto de valores tem distribuição normal, qual a percentagem dos valores que distam da média:

a) Por mais de 1,3 σ;

b) Por menos de 0,52σ?

7 - Uma empresa usa anualmente milhares de lâmpadas elétricas, que permanecem acesa continuamente, dia e noite. A vida de uma lâmpada pode ser considerada uma *v.a.* normal, com média de 50 dias e d.p. de 15 dias. Em 1º de janeiro a companhia instalou 8.000 lâmpadas novas. Aproximadamente quantas deverão ser substituídas em 1º de

fevereiro?

8 - Seja X o número de sucessos em n provas de Bernoulli com probabilidade p de sucesso. Pede-se:

a) Determinar as probabilidades exatas de $X \leq 6$ se $n = 25$ e $p = 0{,}35$.

9 - Se a ocorrência de falhas é uma *v.a.* de Poisson com média 0,2 por m^2, qual a probabilidade de o número de falhas em 5.000 m^2 de tecido:

a) estar entre 975 e 1025;

b) ser superior a 980.

10 - Determinar a probabilidade de obter sete caras em 14 jogadas de uma moeda.

11 - Um laboratório alega que um produto seu cura 80% dos casos de certa doença. Para verificar a afirmação, uma equipe médica do governo testou o produto em 100 indivíduos portadores da doença e decidiu aprová-lo se houver no mínimo 75% de curas.

a) qual a probabilidade de a afirmação do laboratório ser rejeitada, quando a probabilidade de cura é de fato 0.8?

b) qual a probabilidade de a afirmação ser aceita, se a probabilidade de cura é de fato 0,7?

12 - Joga-se um par de dados 180 vezes. Qual a probabilidade de ocorrência do total 7:

a) ao menos 25 vezes?

b) entre 33 e 41 vezes?

c) exatamente 30 vezes?

13 - Joga-se o par de dados 200 vezes. Qual a probabilidade de um número de duplas (dois ‘1’, dois ‘2’, ...).

a) estar entre 40 e 50?

b) ser igual ou superior a 25?

c) ser inferior ou igual a 40?

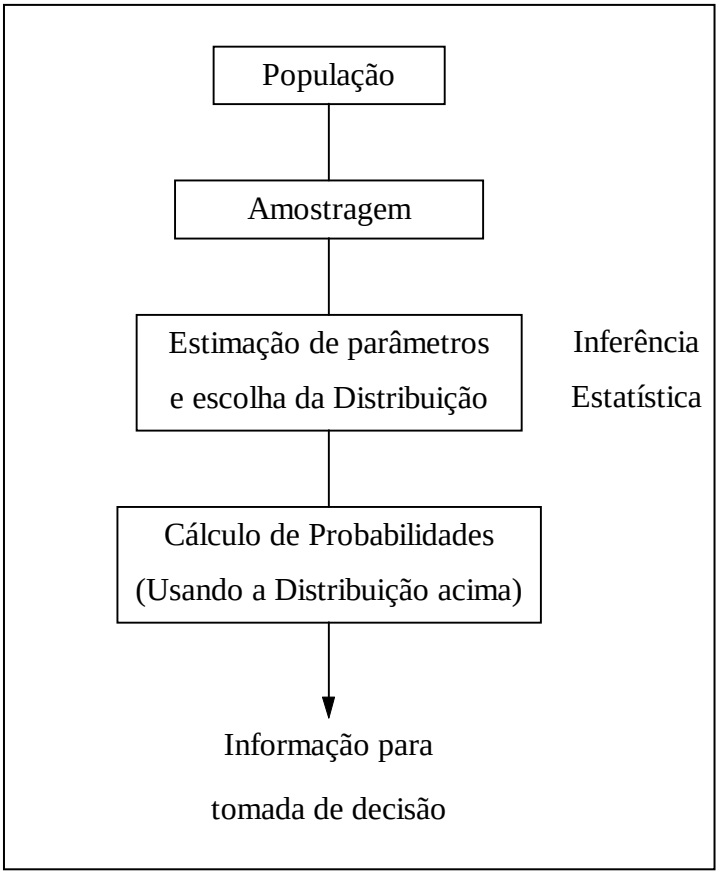
14 - Em uma turma de história, o exame final acusou média 72,5 e d.p. 11,2 (nota máxima 100). Se a distribuição das notas pode ser considerada aproximadamente normal, que percentagem delas excede 80?

15 - Em um posto de grande movimento, os automóveis entram à razão de três em cada dois minutos, em média. Determinar a probabilidade de mais de 100 automóveis procurarem o posto no período de uma hora.

CAPÍTULO 4 - ESTIMAÇÃO DE PARÂMETROS

4.1 - Introdução à Inferência Estatística

A figura abaixo descreve a regra de inferência estatística em um processo de tomada de decisão.



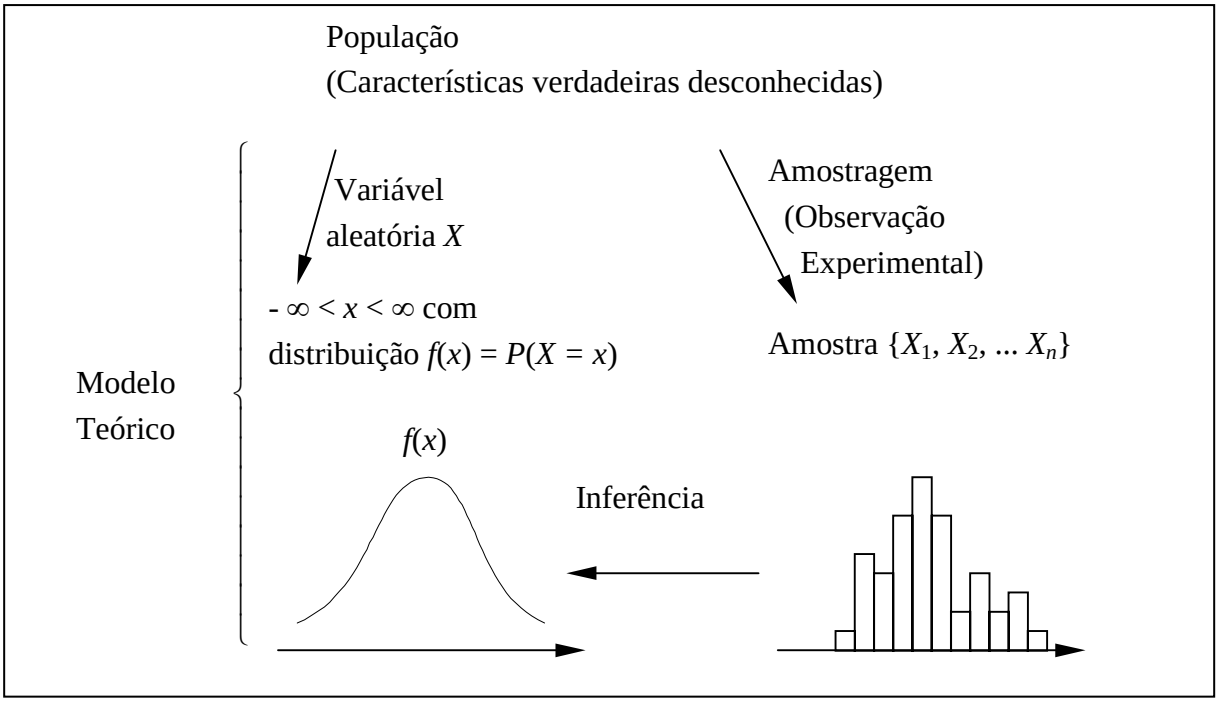
O Objetivo da estimação é determinar os parâmetros da distribuição a serem utilizados para uma determinada população. **Ex.**: Para a distribuição normal os parâmetros são μ e σ^2, para as distribuições de Poisson e Exponencial o parâmetro é α, etc.

Aqui os termos "população" e "distribuição" são equivalentes. As amostras, obtidas da população são aleatórias, onde todos os elementos da população tem a mesma probabilidade de serem escolhidos. Com isso tentamos garantir que toda a variabilidade presente na população estará refletida na amostra. Outra conseqüência desse caráter aleatório é que o conhecimento de um elemento qualquer nada nos informa quanto ao valor dos outros elementos. (O processo de amostragem será visto adiante - Cap.6).

A figura seguinte descreve o processo de inferência estatística aplicado a uma determinada população.

Qualquer valor calculado com base nos elementos de uma amostra é chamado uma "estatística". Por exemplo, a média amostral, ou seja, a média dos elementos da amostra ($\overline{X}$), é uma estatística, assim como o são a mediana amostral ($\tilde{X}$), a amplitude amostral (h), etc.

As estatísticas variam de uma para outra amostra, sendo pois, elas próprias variáveis aleatórias. Podemos, assim, falar de população de médias amostrais, de medianas amostrais ou, equivalentemente, de distribuição das médias amostrais, das medianas amostrais, etc.



Ex.: Média $\mu \cong \overline{X}$

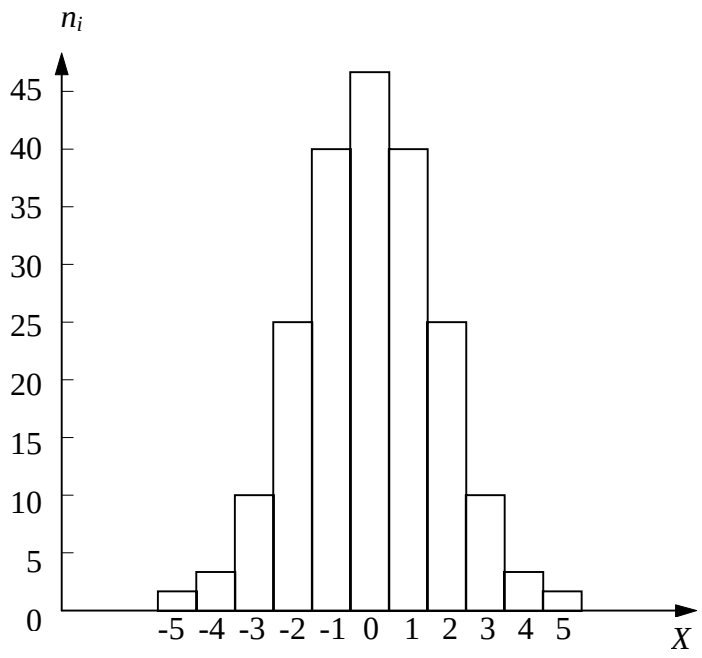
Variância $\sigma^2 \cong S^2$

Estimação

$$\overline{X} = \frac{1}{n}\sum X_i \quad \text{(média amostral)}$$

$$S^2 = \frac{1}{n-1}\sum (X_i - \overline{X})^2 \quad \text{(variância amostral)}$$

Ex.: Para uma população de 200 valores que variam entre -5 e 5 temos a seguinte distribuição freqüencial e o seu respectivo histograma.



x_i	n_i	N_i
-5	1	1
-4	3	4
-3	10	14
-2	23	37
-1	39	76
0	48	124
1	39	163
2	23	186
3	10	196
4	3	199
5	1	200

Extraindo uma amostra aleatória de tamanho 10 desta população obtemos as seguintes oito amostras que estão relacionadas juntamente com as respectivas médias e medianas amostrais.

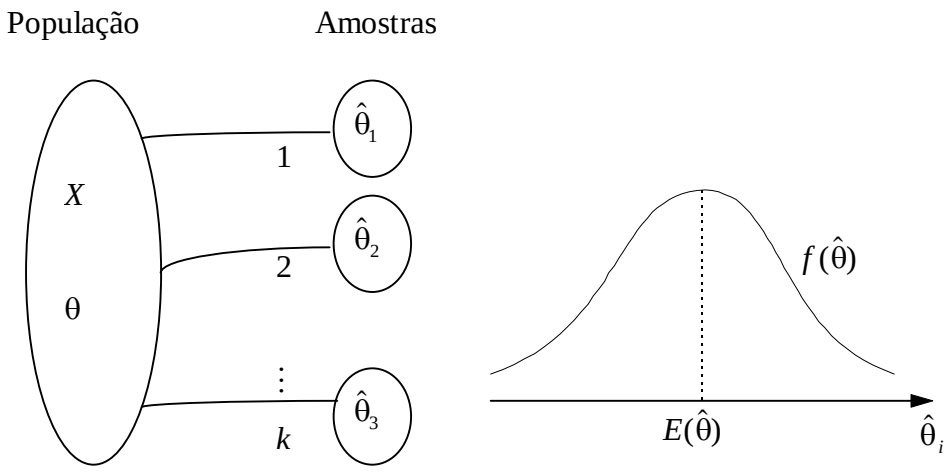
Amostra	Valores										$\overline{X}$	$\tilde{X}$
1	1	1	0	0	-3	1	-1	0	-2	2	-0.10	0
2	2	1	-1	0	2	0	2	0	1	1	0.80	1
3	-1	4	-2	-2	-1	1	2	1	-1	-1	0.00	-1
4	-1	0	2	2	-3	0	1	-5	1	-3	0.90	-0.5
5	-2	1	0	0	0	-2	1	-2	1	-2	-0.40	0
6	-1	-1	0	0	3	-2	-1	-4	2	-1	-040	-1
7	0	1	-1	-1	0	0	-5	1	0	1	-0.10	0.5
8	0	1	0	0	3	1	2	0	3	-2	080	0.5

Em tal extração pode ser utilizada uma tábua de números aleatórios (Ver Tabela 4) ou um processo computadorizado (tipo “random”) para gerar uma seqüência aleatória de números. Se de alguma forma, por exemplo, obtermos os números aleatórios 097, 019, 065, 060, 124, 104, 196, 058, 179, 155 podemos associá-los aos valores em N_i da tabela e extrair a seguinte amostra: 0, -2, -1, -1, 0, 0, 3, -1, 2, 1. (confira)

Para facilitar a linguagem usada em Inferência Estatística, iremos diferenciar as características da amostra e da população. Os símbolos mais comuns são dados a seguir:

	Estatística	População
Média	$\overline{X}$	μ
Variância	S^2	σ^2
Nº elementos	n	N
Proporção	$\hat{p}$	p
Geral	$\hat{\theta}$	θ

θ (theta) representa o parâmetro populacional de interesse. Para estimá-lo extraímos uma amostra de tamanho *n* da população (X_1, X_2, X_3 ... X_n) e procuramos construir uma função desses valores, ou seja, uma estatística. A estatística que estima θ é comumente $\hat{\theta}$. E como diferentes amostras originam valores distintos de $\hat{\theta}$, $\hat{\theta}$ é, ele próprio, uma variável aleatória. Nesse último caso temos uma *Distribuição Amostral* da estatística $\hat{\theta}$. O gráfico seguinte ilustra tal distribuição:



onde para cada amostra, calculamos o valor $\hat{\theta}_i$ da estatística $\hat{\theta}$. Os valores $\hat{\theta}_i$ formam uma nova população, a Distribuição Amostral da estatística $\hat{\theta}$.

Quando propomos $\hat{\theta}$ como uma estimativa de θ, não esperamos realmente que $\hat{\theta}$ venha a ser igual a θ. (Recordemos que $\hat{\theta}$ é uma *v.a.* e, por isso, pode tomar diferentes valores). Este dilema origina duas importantes questões:

1. Quais as características que desejamos que uma "boa" estimativa apresente?
2. Como decidiremos que uma estimativa é "melhor" do que a outra?

Tais questões não admitem uma análise tão simples. Em primeiro lugar, o próprio valor do parâmetro θ não é verdadeiramente correto (ao menos em uma situação concreta). Assim, como dizer que $\hat{\theta}$ o é? Segundo, se tivermos duas estimativas de θ, por exemplo $\hat{\theta}_0$ e $\hat{\theta}_1$, qual o critério de comparação para decidirmos entre $\hat{\theta}_0$ e $\hat{\theta}_1$? Observe que isto não é um problema como, por exemplo, no cálculo de raízes da equação $3x^5 - 4x^2 + 13x - 7 = 0$. Uma vez que se tenha achado uma solução, é muito simples verificar se ela é a solução correta: precisaríamos apenas substituí-la na equação dada. Se tivermos duas respostas aproximadas, digamos r_1 e r_2, será também assunto simples de decidir qual das aproximações é a melhor.

A estimação de um parâmetro populacional pode ser de dois tipos: *estimação pontual* e *estimação intervalar*. A estimação pontual procura fixar um valor numérico único que esteja satisfatoriamente próximo do verdadeiro valor do parâmetro. A estimativa intervalar procura determinar intervalos com limites aleatórios, que abranjam o valor do parâmetro populacional, com uma margem de segurança prefixada.

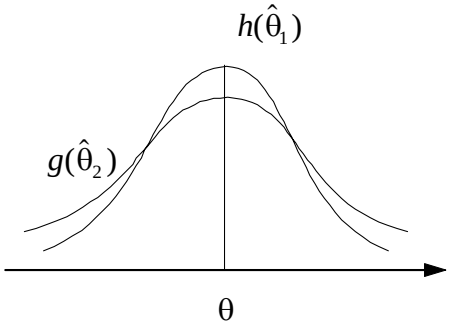
4.2 - Critérios para Estimativas

Há vários métodos de obtenção de estimadores, destacando-se entre eles o Método da Máxima Verossimilhança. Seu estudo, entretanto, ultrapassa o nível deste texto. Como existem muitos estimadores para um mesmo parâmetro de uma população, é necessário estudar algumas propriedades que os distingam uns dos outros. Duas delas são: não-tendenciosidade e variância mínima.

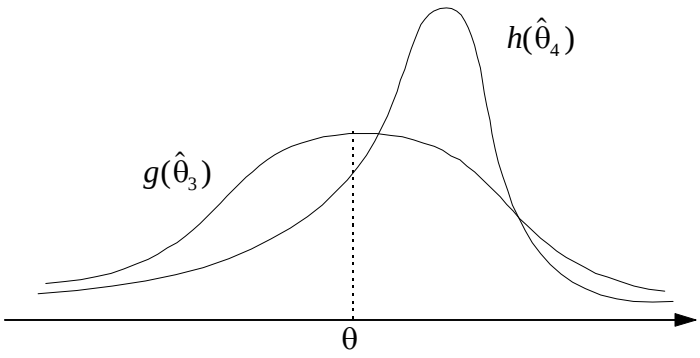
Seja X uma *v.a.* cuja distribuição dependa do parâmetro θ. Seja X_1, X_2, ... X_n uma amostra aleatória de X e seja $\hat{\theta}$ uma função da amostra. (Portanto, $\hat{\theta}$ é uma *v.a.* unidimensional cuja distribuição de probabilidade depende de θ). Nós diremos que $\hat{\theta}$ é uma estimativa não-tendenciosa de θ se $E(\hat{\theta}) = \theta$ para todo θ.

Seja $\hat{\theta}$ uma estimativa não-tendenciosa de θ. Diremos que $\hat{\theta}$ é uma estimativa de variância mínima de θ, se para todas as estimativas $\hat{\theta}^*$ tais que $E(\hat{\theta}^*) = \theta$, tivermos $Var(\hat{\theta}) \leq Var(\hat{\theta}^*)$ para todo θ.

Assim, se $\hat{\theta}_1$ e $\hat{\theta}_2$ forem duas estimativas de θ, cuja *fdp* seja representada na figura seguinte, presumivelmente preferiríamos $\hat{\theta}_1$ a $\hat{\theta}_2$.



No caso de estimativas $\hat{\theta}_3$ e $\hat{\theta}_4$, (ver figura abaixo) a decisão não é tão evidente, porque $\hat{\theta}_3$ é não-tendenciosa, enquanto $\hat{\theta}_4$ é tendenciosa. Todavia $Var(\hat{\theta}_3) > Var(\hat{\theta}_4)$. Isto significa que, enquanto em média $\hat{\theta}_3$ será próxima de θ, sua variância revela que desvios consideráveis em relação a θ não serão de surpreender. $\hat{\theta}_4$ por sua vez, tende a ser um tanto maior do que θ, em média, e no entanto poderá ser mais próxima de θ do que $\hat{\theta}_3$.



Ex.: Suponha que desejemos uma estimativa não-tendenciosa da variância σ^2 de uma *v.a.*, baseada em uma amostra X_1, X_2, ... X_n.

Muito embora, intuitivamente, pudéssemos considerar $\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$, verifica-se que esta estatística tem um valor esperado diferente de σ^2, ou seja

Para $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 \qquad E(\hat{\theta}) = \frac{(n-1)}{n}\sigma^2$

Por isso, uma estimativa não-tendenciosa de σ^2 será obtida tomando-se

$\hat{\theta} = S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ e $E(\hat{\theta}) = E(S^2) = \sigma^2$

De fato:

$$\begin{aligned}
\sum(X_i - \overline{X})^2 &= \sum(X_i - \mu + \mu - \overline{X})^2 = \sum[(X_i - \mu) + (\mu - \overline{X})]^2 \\
&= \sum\left[(X_i - \mu)^2 + 2(\mu - \overline{X})(X_i - \mu) + (\mu - \overline{X})^2\right] \\
&= \sum(X_i - \mu)^2 + 2(\mu - \overline{X})\sum(X_i - \mu) + n(\mu - \overline{X})^2 \\
&= \sum(X_i - \mu)^2 - 2n(\mu - \overline{X})^2 + n(\mu - \overline{X})^2 \\
&= \sum(X_i - \mu)^2 - n(\overline{X} - \mu)^2 = n\sigma^2 - \sigma^2 = \sigma^2(n-1)
\end{aligned}$$

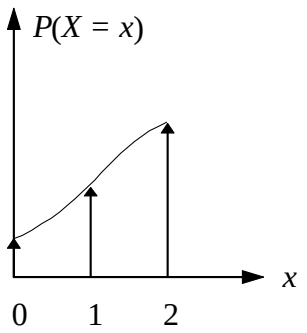
Portanto:

$$E\left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2\right) = \frac{1}{n-1}[\sigma^2(n-1)] = \sigma^2$$

4.3 - O Teorema Central do Limite

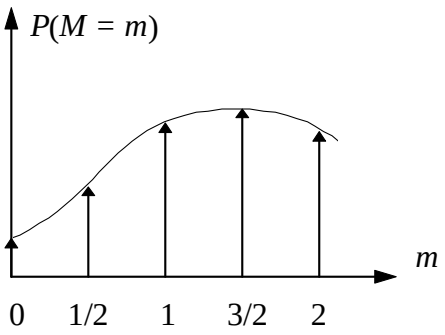
Ex.: Considere uma urna que contenha três espécies de objetos, identificados como 0, 1 e 2, em quantidades 20, 30 e 50 respectivamente. Um valor é extraído aleatoriamente e seu valor, digamos *X*, é registrado. *X* tem então a seguinte distribuição:

X	0	1	2
$P(X=x)$	0.2	0.3	0.5



Suponha agora que o objeto extraído em primeiro lugar seja reposto na urna e a seguir, um segundo objeto seja extraído e seu valor, digamos *Y*, seja registrado. Considere a *v.a.* $M = \dfrac{X+Y}{2}$ e sua distribuição

M	0	1/2	1	3/2	2
$P(M=m)$	0.04	0.12	0.29	0.30	0.25



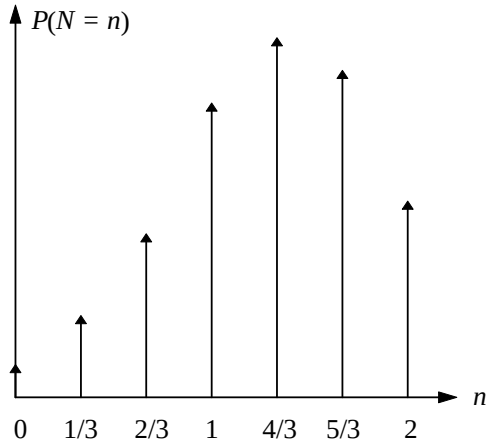
onde os valores de $P(M = m)$ foram calculados da seguinte forma:

$$P(M=0) = P(X=0, Y=0) = (0.2)^2 = 0.04$$
$$P(M=1/2) = P(X=0, Y=1) = P(X=1, Y=0) = (0.2)(0.3) + (0.3)(0.2) = 0.12$$

etc.

Finalmente, considere pelo mesmo processo uma terceira extração, digamos *Z*, e a seguinte *v.a.* $N = \dfrac{X+Y+Z}{3}$ com sua respectiva distribuição de probabilidade:

N	0	1/3	2/3	1	4/3	5/3	2
	0.008	0.036	0.114	0.207	0.285	0.225	0.125



As distribuições de probabilidades das *v.a. M* e *N* já estão mostrando sinais de "normalidade", isto é, a aparência da curva em sino da distribuição normal está começando a surgir. Partindo da distribuição de *X*, a qual é quase assimétrica verificamos que a média de apenas três observações tem uma distribuição que já está mostrando "indicações de normalidade".

Naturalmente, o exemplo acima não demonstra coisa alguma. No entanto, ele representa uma ilustração numérica do importante resultado dado a seguir:

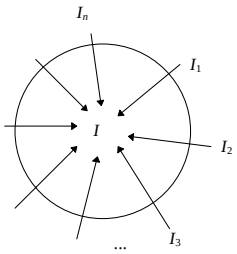
"Para uma população não normal com média μ e desvio padrão σ, a *distribuição da média amostral* $\overline{X}$ para amostras de tamanho *n* suficientemente grande é aproximadamente normal com média μ e desvio padrão $\sigma/\sqrt{n}$, isto é:

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N:(0,1) \text{ "}$$

Tal resultado, bastante surpreendente, constitui o chamado Teorema Central do Limite. Assim, mesmo para populações não normais, a distribuição da média $\overline{X}$ é aproximadamente normal. (Alguns textos referem-se erroneamente ao "Teorema do Limite

Central”; o que é central é o teorema e não o limite.) A demonstração de tal teorema foge em muito ao escopo desse texto.

Ex.: Suponha que temos algumas correntes elétricas independentes, por exemplo I_i, $i = 1, 2 \ldots n$, as quais são recebidas em um “somador”, como na figura seguinte. Seja I a soma das correntes recebidas. Isto é $I = \sum_{i=1}^{n} I_i$. Suponha que cada variável aleatória I_i seja *uniformemente distribuída* sobre o intervalo [0,10] mA. daí, $E(I_i) = 5$ mA e variância $Var(I_i) = 100/12$. (Confira!)



De acordo com o Teorema Central do Limite, se n for suficientemente grande a variável aleatória $Z = \dfrac{\bar{I} - \mu}{\sigma/\sqrt{n}}$ terá uma distribuição aproximadamente normal $N(0,1)$.

Assim:

$$E(\bar{I}) = E\left[\frac{1}{n}\sum I_i\right] = \frac{1}{n}E\left[I_1 + I_2 + \ldots + I_n\right] = \frac{1}{n}\left[E(I_1) + E(I_2) + \ldots + E(I_n)\right]$$
$$= \frac{5}{n}n = 5$$

$$Var(\bar{I}) = Var\left(\frac{\sum I_i}{n}\right) = \frac{1}{n^2}Var(I_i + I_2 + \ldots + I_n) = \frac{100}{12n}$$

Desta forma

$$\bar{I} : N\left(5; \frac{10}{\sqrt{12}} \Big/ \sqrt{n}\right)$$

e:
$$Z = \frac{\bar{I} - 5}{\frac{10}{\sqrt{12}} \cdot \frac{1}{\sqrt{n}}} \sim N(0,1)$$

Se desejarmos calcular $P(I>105)$ para $n = 20$, por exemplo, teremos:

$$Z = \frac{\frac{I}{n} - 5}{\frac{10}{\sqrt{12}}\frac{\sqrt{n}}{n}} = \frac{\frac{1}{n}[I - 5n]}{\frac{1}{n}\left(\frac{10\sqrt{n}}{\sqrt{12}}\right)} = \frac{I - 5n}{\frac{10}{\sqrt{12}}\sqrt{n}}$$

Assim, para $I = 105$ e $n = 20$

$$Z = \frac{105 - 100}{\frac{10}{\sqrt{12}}\sqrt{20}} \simeq 0.388$$

$$P(I > 105) = P(Z > 0.388) = 1 - \Phi(0.388) = 0.352 = 35{,}2\,\%$$

Uma outra forma de enunciar o teorema central do limite é:

"Seja X_1, X_2, ... X_n uma seqüência de variáveis aleatórias independentes, com $E(X_i) = \mu$ e $Var(X_i) = \sigma^2$, i = 1, 2, ... n. Façamos $X = X_1 + X_2 + ... + X_n$. Então:

$$Z = \frac{X - n\mu}{\sigma.\sqrt{n}}$$

tem distribuição aproximadamente normal $N(0,1)$"

Observe que a expressão acima simplifica os cálculos do exemplo anterior.

Um outro importante resultado relacionado ao teorema central do limite é:

"Quando a população é normal $N(\mu;\sigma)$, a *média amostral* $\overline{X}$ de amostras de tamanho n tem distribuição também normal com média μ e desvio padrão $\sigma/\sqrt{n}$, isto é:

$$Z = \frac{\overline{X} - u}{\sigma/\sqrt{n}} \sim N : (0,1)$$ "

Os resultados apresentados nesta seção serão amplamente utilizados nos próximos capítulos.

Exercícios XII

01 - Uma variável aleatória X tem distribuição normal, com média 100 e desvio padrão 10.

a) Qual a $P(90 < X < 110)$?

b) Se $\overline{X}$ é a média de uma amostra de 16 elementos retirados dessa população, calcule $P(90 < \overline{X} < 110)$.

c) Desenhe, num único gráfico, as distribuições de X e $\overline{X}$.

d) Que tamanho deveria ter a amostra para que $P(90 < \overline{X} < 110) = 95\%$?

02 - A máquina de empacotar um determinado produto o faz segundo uma distribuição normal, com média μ e desvio padrão 10 g.

a) Em quanto deve ser regulado o peso médio μ para que apenas 10% dos pacotes tenham menos do que 500 g.

b) Com a máquina assim regulada, qual a probabilidade de que o peso total de 4 pacotes escolhidos ao acaso seja inferior a 2 kg?

03 - No exemplo anterior, e após a máquina estar regulada, programou-se uma carta de controle de qualidade. De hora em hora, será retirada uma amostra de 4 pacotes, e estes serão pesados. Se a média da amostra for inferior a 495 g ou superior a 520 g, pára-se a produção para reajustar a máquina, isto é, reajustar o peso médio.

a) Qual a probabilidade de ser feita uma parada desnecessária?

b) Se o peso médio da máquina desregulou-se para 500 g, qual a probabilidade de continuar-se a produção fora dos padrões desejados?

04 - A capacidade máxima de um elevador é de 500 kg. Se a distribuição X dos pesos dos usuários é suposta $N(70, 10)$:

a) Qual a probabilidade de 7 passageiros ultrapassarem esse limite?

b) E seis passageiros?

05 - Uma variável X tem distribuição normal, com média 10 e desvio padrão 4. Aos participantes de um jogo, é permitido observar uma amostra de qualquer tamanho e calcular a média amostral. Ganha um prêmio aquele cuja média amostral for maior que 12.

a) Se um participante escolher uma amostra de tamanho 16, qual a probabilidade dele ganhar um prêmio?

b) Escolha um tamanho de amostra diferente de 16 para participar do jogo. Qual a

probabilidade de você ganhar um prêmio?

c) Baseado nos resultados acima, qual o melhor tamanho de amostra para participar do jogo?

06 - A distribuição dos comprimentos dos elos de corrente de bicicleta é normal, com média 2 cm e variância igual a 0,01 cm^2. Para que uma corrente se ajuste à bicicleta, deve ter comprimento total igual entre 58 e 61 cm.

a) Qual a probabilidade de uma corrente com 30 elos não se ajustar à bicicleta?

b) E uma corrente com 29 elos?

07 - Cada seção usada para construção de um oleoduto tem um comprimento médio de 5 m e desvio padrão de 20 cm. O comprimento total do oleoduto será de 8 km.

a) Se a firma construtora do oleoduto encomendar 1.600 seções, qual a probabilidade de terem que comprar mais do que uma seção adicional (isto é, das 1.600 seções somarem 7.995 m ou menos)?

b) Qual a probabilidade de uso exato de 1.599 seções, isto é, a soma das 1.599 seções estar entre 8.000 e 8.005 m?

08 - A distribuição dos salários (em salários mínimos) de operários do sexo masculino de uma grande fábrica é $N(5{,}4;\ 1.3)$, e a de operários do sexo feminino é $N(5{,}4;\ 1.5)$. Sorteiam-se duas amostras, uma com 16 homens e outra com 16 mulheres. S D é a diferença entre o salário médio dos homens e das mulheres:

a) Calcule $P\left(\left|D\right| > 0{,}5\right)$.

b) Qual o valor de d tal que $P\left(\left|D\right| > d\right) = 0{,}05$?

c) Que tamanho comum deveriam ter ambas as amostras para que $P\left(\left|D\right| > 0{,}4\right) = 0{,}05$?

4.4 - Algumas Estimativas Pontuais

4.4.1 - Média e variância de uma população normal $N(\mu; \sigma)$

Não é de se admirar que o melhor estimador da média populacional (μ) seja a média amostral ($\overline{X}$):

$$\overline{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n),$$

sendo X_i variáveis aleatórias independentes com $E(X_i) = \mu$ e $Var(X_i) = \sigma^2$, i = 1, 2, ... n.

Assim,

$$E(\overline{X}) = \frac{1}{n}E(X_1 + X_2 + \ldots + X_n) = \frac{1}{n}\left[E(X_1) + E(X_2) + \ldots + E(X_n)\right]$$
$$= \frac{1}{n}\left[\mu + \mu + \cdots + \mu\right] = \frac{n\mu}{n} = \mu$$

$$Var(\overline{X}) = \frac{1}{n^2}Var(X_1 + X_2 + \ldots + X_n) = \frac{1}{n^2}\left[Var(X_1) + Var(X_2) + \ldots + Var(X_n)\right]$$
$$= \frac{1}{n^2}\left[\sigma^2 + \sigma^2 + \cdots + \sigma^2\right] = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

$E(\overline{X}) = \mu$ revela que $\overline{X}$ é um estimador não tendencioso de μ. $Var(\overline{X}) = \sigma^2/n$ revela que quanto maior o valor de n, menor a sua variabilidade. Assim, neste caso, sendo $\theta = \mu$, adotaremos $\hat{\theta} = \overline{X}$.

Para a variância populacional (σ^2) o estimador S^2, com $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$, é não tendencioso, conforme já foi exemplificado. Adotaremos assim, para $\theta = \sigma^2$, $\hat{\theta} = S^2$.

Ex.: Se de uma população normal $N(\mu; \sigma)$ extraímos amostras cujos valores são:

1.1 0.9 0.3 -0.2 -3.1 1.5 -2.7 0.5 -1.5 2.1,

obtenha estimativas pontuais de μ, σ^2 e $P(X > 2.5)$.

Estimativa de μ:

$$\overline{X} = \frac{1}{10}[1.1 + 0.9 + 0.3 + \cdots + 2.1] = -0.11$$

Estimativa de σ^2:

$$S^2 = \frac{1}{9}[(1.1+0.11)^2 + (0.9+0.11)^2 + \cdots (2.1+0.11)^2] = 3.17$$

Estimativa de $P(X > 2.5)$:

Sendo $\sigma^2 \simeq 3.17 \quad \sigma = 1.78$

Assim para $X = 2.5 \quad z = \dfrac{X-\mu}{\sigma} = \dfrac{2.5-(-0.11)}{1.78} = 1.466$

e $P(X > 2.5) = P(Z > 1.466) = 1 - P(Z < 1.466) = 1 - 0.9286 = 7.14\%$

4.4.2 - Média e variância de uma proporção (*p*)

Consideremos agora o caso em que $\theta = p$, de uma população que apresenta certa característica. Extrai-se da população uma amostra de tamanho *n*. *X* será o número de elementos da amostra que apresentam a característica em estudo. É intuitivo que um estimador da proporção *p* seja a proporção amostral $\hat{p}$:

$$\hat{p} = \frac{X}{n}$$

As observações dos *n* elementos podem ser considerados como *n* provas de Bernoulli com probabilidade de sucesso *p*, ou seja, *X* tem distribuição binomial com média *np* e variância *npq*. Conseqüentemente, temos:

$E(\hat{p}) = E\left(\dfrac{X}{n}\right) = \dfrac{1}{n}(np) = p$. Assim $\hat{p}$ é não tendencioso.

$$Var(\hat{p}) = Var\left(\frac{X}{n}\right) = \frac{1}{n^2}npq = \frac{pq}{n}$$

O desvio padrão de $\hat{p}$ é também denominado de *Erro-Padrão* de $\hat{p}$, representado por $EP(\hat{p}) = \sqrt{\dfrac{pq}{n}}$, sendo $p \simeq \hat{p}$ e $q \simeq \hat{q} = 1 - p$.

Ex.: Para se avaliar a taxa de desemprego em determinado Estado, escolhe-se uma amostra aleatória de 1000 habitantes em idade de trabalho e contam-se os desempregados: 87. Estimar a proporção de desempregados em todo o Estado (população). Avaliar o erro padrão de estimativa.

Temos então:

$$\hat{p} = \frac{X}{n} = \frac{87}{1000} = 0.087 = 8.7\% \qquad \text{Logo: } p \simeq \hat{p} = 8.7\%$$

$$\hat{q} = 1 - \hat{p} = 0.913$$

$$EP(\hat{p}) = \sqrt{\frac{\hat{p}\hat{q}}{n}} = \sqrt{\frac{(0.087)(0.913)}{1000}} = 0.009 \qquad \text{Logo } \sigma \simeq 0.009$$

4.5 - Intervalos de Confiança

Até agora estivemos tratando apenas de obter uma estimativa pontual de um parâmetro desconhecido. Como mencionado anteriormente, há outra maneira de tratar o assunto que freqüentemente conduz a resultados muito significativos.

Seja então $X_1, X_2, \cdots, X_n$ uma amostra aleatória de uma população e θ o parâmetro de interesse. Sejam $\hat{\theta}_0$ e $\hat{\theta}_1$ estatísticas tais que:

$$P\left(\hat{\theta}_0 < \theta < \hat{\theta}_1\right) = 1 - \alpha$$

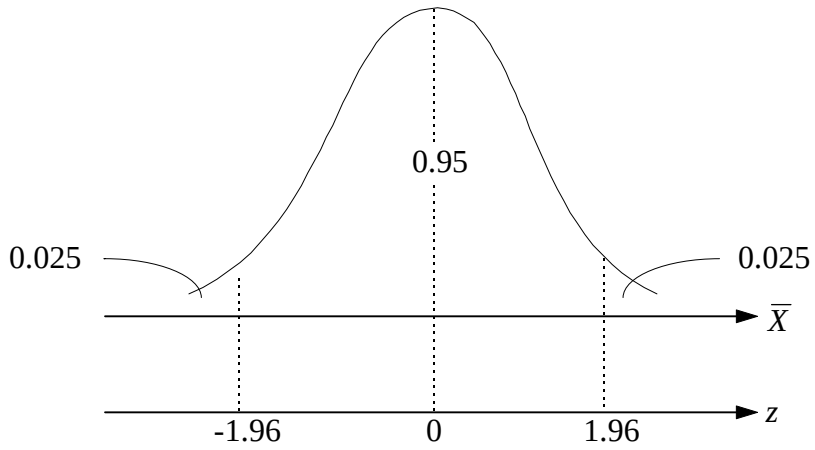
Então $\left[\hat{\theta}_0 ; \hat{\theta}_1\right]$ é chamado *intervalo de confiança de nível* 100(1 - α) % para o parâmetro θ. Usualmente toma-se 1 - α = 0.95 ou 0.99.

Muitas estatísticas consideram a construção de intervalos de confiança como o principal método de estudo de um parâmetro populacional através de uma amostra.

Ex.: Consideremos uma população normal com média μ e desvio padrão σ e uma amostra dessa população. Sabemos pelos resultados do Teorema Central do Limite que a média $\overline{X}$ desta amostra tem distribuição normal com média μ e desvio padrão $\sigma/\sqrt{n}$, ou seja:

$$\frac{\overline{X} - u}{\sigma/\sqrt{n}} \sim N : (0,1) \text{ ”}$$

Fixando α em 0.05, ou seja, $1 - \alpha = 0.95$, vemos pela tabela de distribuição normal padronizada *z*, que:



$$P(-1.96 < Z < 1.96) = 0.95$$

isto é:

$$P\left[-1.96 < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < 1.96\right] = 0.95$$

Reescrevendo as desigualdades entre parênteses, temos:

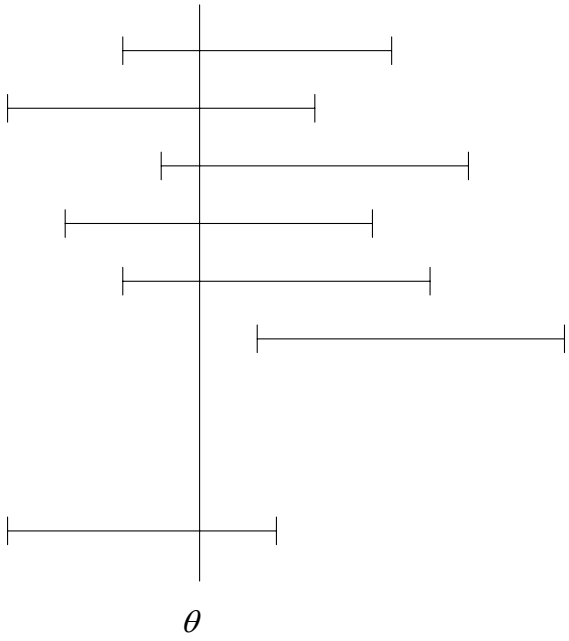
$$P\left[\overline{X} - 1.96(\sigma/\sqrt{n}) < \mu < \overline{X} + 1.96(\sigma/\sqrt{n})\right] = 0.95$$

Neste caso: $\theta = \mu$ é o parâmetro

$\left[\hat{\theta}_0; \hat{\theta}_1\right] = \left[\overline{X} - 1.96\left(\sigma/\sqrt{n}\right); \overline{X} + 1.96\left(\sigma/\sqrt{n}\right)\right]$ é o intervalo de confiança de 95% para μ, $IC(\mu{:}95\%)$

É importante observar que o *nível de confiança* (1 - α) se aplica ao "processo" de construção de intervalos, e não a um intervalo específico. Para explicitar o conceito de intervalo de confiança, suponha que retiremos um grande número de amostras de tamanho *n*, fixo, da população em estudo e, para cada amostra, construamos um intervalo. Os limites dos intervalos resultantes serão diferentes.

O verdadeiro valor do parâmetro estará contido, em média, em 100(1 - α)% desses intervalos, ou seja, 100(1 - α)% dos intervalos construídos abrangerão o verdadeiro valor do parâmetro, no caso μ, conforme ilustrado na figura seguinte, mas cada valor contém, ou não contém, o parâmetro.



Obs.: A expressão do tipo $P\left[\overline{X}-1.96\left(\sigma/\sqrt{n}\right)<\mu<\overline{X}+1.96\left(\sigma/\sqrt{n}\right)\right]=0.95$ deve ser interpretada muito cuidadosamente. Ela não significa que a probabilidade do parâmetro μ cair dentro de um intervalo especificado seja igual a 0.95. μ sendo o parâmetro, está ou não está dentro do intervalo acima. De preferência a expressão acima deve ser interpretada assim:

"0.95 é a probabilidade de que um intervalo aleatório $\left[\overline{X}-1.96\left(\sigma/\sqrt{n}\right);\overline{X}+1.96\left(\sigma/\sqrt{n}\right)\right]$ contenha μ."

Ex.: Para uma amostra de 50 observações de uma população normal com média desconhecida μ e desvio padrão $\sigma=6$, seja 20.5 a média amostral $\overline{X}$. Construir um intervalo de 95% de confiança para a média populacional.

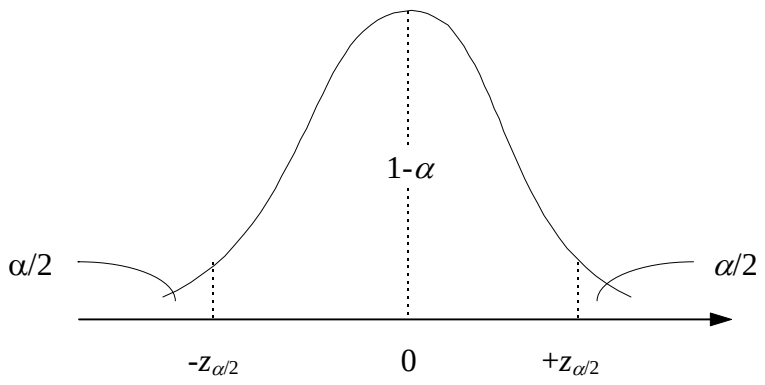
Temos, de imediato que:

$$P\left[\overline{X}-1.96\left(\sigma/\sqrt{n}\right)<\mu<\overline{X}+1.96\left(\sigma/\sqrt{n}\right)\right]=0.95$$

Assim, tal intervalo é [18.84; 22.16].

4.6 - Estimativa Intervalar da Média de uma população normal, σ conhecido

A figura seguinte representa a curva $N(0,1)$ e a notação que iremos utilizar, onde $P\left[|Z| < Z_{\alpha/2}\right] = 1-\alpha$



Na situação aqui apresentada, $N(\mu;\sigma)$ com σ conhecido, sabemos que:

$$z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0;1) \text{ e } P\left[-Z_{\alpha/2} < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < Z_{\alpha/2}\right] = 1-\alpha$$

Assim: $P\left[\overline{X} - Z_{\alpha/2}\left(\sigma/\sqrt{n}\right) < \mu < \overline{X} + Z_{\alpha/2}\left(\sigma/\sqrt{n}\right)\right] = 1-\alpha$

e o intervalo de confiança para μ será

$$\boxed{IC\left(\mu:(1-\alpha)100\right) = \left[\overline{X} - Z_{\alpha/2}\left(\sigma/\sqrt{n}\right); \overline{X} + Z_{\alpha/2}\left(\sigma/\sqrt{n}\right)\right]}$$

***Obs.*:** Para $\alpha = 5\%$ $\alpha/2 = 2.5\%$ $Z_{\alpha/2} = 1.96$,

$\alpha = 1\%$ $\alpha/2 = 0.5\%$ $Z_{\alpha/2} = 2.58$. (*Confira!*)

***Obs.*:** A situação aqui considerada é pouco concreta, pois dificilmente se conhece a variância de uma população quando sua média é desconhecida.

Exercícios XIII

01 - Um procedimento de controle de qualidade foi planejado para garantir um máximo de 10% de itens defeituosos na produção. A cada 15 minutos sorteia-se uma amostra de 20 peças, e, havendo mais de 15% de defeituosos, pára-se a produção para verificações. Qual a probabilidade de uma parada desnecessária?

02 - Se uma amostra com 36 observações é tomada de uma população, qual deve ser o tamanho de uma outra amostra para que o erro padrão desta amostra seja 2/3 do erro padrão da média da primeira amostra?

03 - Definimos a variável $e = \overline{X} - \mu$ como sendo o erro amostral da média. Suponha que a variância dos salários de uma certa região seja 400 unidades ao quadrado.

a) Determine $E(e)$ e $Var(e)$.

b) Que proporção das amostras de tamanho 25 terão erro amostral absoluto maior do que 2 unidades?

c) E que proporção das amostras de tamanho 100?

d) Neste último caso, qual o valor de d, tal que $P\left(\left|e\right| > d\right) = 1\%$?

e) Qual deve ser o tamanho da amostra para que 95% dos erros amostrais absolutos sejam inferiores a uma unidade?

04 - Um professor dá um teste rápido, constante de 18 questões do tipo certo-errado. Para testar a hipótese de o estudante estar adivinhando a resposta, ele adota a seguinte regra de decisão: "Se 12 ou mais estão corretas, ele não está adivinhando". Qual a probabilidade de rejeição da hipótese, quando verdadeira?

05 - Um distribuidor de sementes determina, através de testes, que 5% das sementes não germinam. Ele vende pacotes de 200 sementes com garantia de 90% de germinação. Qual a probabilidade de um pacote não satisfazer a garantia?

06 - A experiência indica que 60% dos passageiros de vôos interestaduais preferem refeições quentes, enquanto os restantes 40% preferem refeições frias. Para cada vôo, acham-se disponíveis 72 refeições quente e 48 frias. Se 100 passageiros tomam o avião, pergunta-se:

a) Qual o número máximo de passageiros que podem receber refeições quentes? Neste caso, quantos receberão refeições frias?

b) Qual o número máximo de passageiros que podem receber refeições frias? Neste caso,

qual o número mínimo de passageiros que podem pedir refeições quentes?

c) Quais os números mínimos e máximos de refeições quentes que podem ser pedidas de modo que todos recebem a refeição de sua preferência?

d) Qual a probabilidade de que cada passageiro receba a refeição de sua preferência?

07 - Obtenha a distribuição de $\hat{p}$ quando p=0,2 e n=5. Depois calcule $E(\hat{p})$ e $Var(\hat{p})$.

08 - Suponha um experimento consistindo de n provas de Bernoulli. Seja X o número de sucessos, e considere os estimadores:

a) $\hat{p}_1 = \dfrac{X}{n}$;

b) $\hat{p}_2 = \begin{cases} 1, \text{ se aprimeira prova resulta sucesso,} \\ 0, \text{ caso contrario.} \end{cases}$

Determine a esperança e a variância de cada estimador. Por que $\hat{p}_2$ não é um "bom" estimador?

09 - Calcule o intervalo de confiança para a média em cada um dos casos abaixo.

Média Amostral	*Tamanho da Amostra*	*Desvio Padrão da População*	*Coeficiente de Confiança*
170 cm	100	15 cm	95%
165 cm	184	30 cm	85%
180 cm	225	30 cm	70%

10 - De 50.000 válvulas fabricadas por uma companhia retira-se uma amostra de 400 válvulas, e obtêm-se a vida média de 800 horas e o desvio padrão de 100 horas.

a) Qual o intervalo de confiança de 99% para a vida média da população?

b) Com que confiança dir-se-ia que a vida média é 800 ± 0,98?

c) Que tamanho deve ter a amostra para que seja de 95% a confiança na estimativa 800 ± 7,84?

(Que suposições você fez para responder às questões acima?)

11- Qual deve ser o tamanho de uma amostra cujo desvio padrão é 10 para que a diferença da média amostral para a média da população, em valor absoluto, seja menor que 1, com coeficiente de confiança igual a:

a) 95% b) 99%

12 - Uma população tem desvio padrão igual a 10.

a) Que tamanho deve ter uma amostra para que, com probabilidade 8%, o erro em estimar a média seja superior a 1 unidade?

b) Supondo-se colhida a amostra no caso anterior, qual o intervalo de confiança, se $\bar{x} = 50$?

13 - Uma amostra aleatória de 625 donas-de-casa revela que 70% delas preferem a marca *X* de detergente. Construir um intervalo de confiança para *p* = proporção das donas-de-casa que preferem *X* com c.c. $\gamma = 90\%$

14 - Encontre os intervalos de confiança para *p* se $k/n = 0{,}3$, com c.c. $\gamma = 0{,}95$. Utilize os dois enfoques apontados na seção 9.5, com $n = 400$.

15 - Antes de uma eleição, um determinado partido está interessado em estimar a proporção *p* de eleitores favoráveis ao seu candidato. Uma amostra piloto de tamanho 100 revelou que 60% dos eleitores eram favoráveis ao candidato em questão.

a) Determine o tamanho da amostra necessário para que o erro cometido na estimação seja de, no máximo 0,01 com probabilidade de 80%.

b) Se na amostra final, com tamanho igual ao obtido em (a), observou-se que 55% dos eleitores eram favoráveis ao candidato em questão, construa um intervalo de confiança para a proporção *p*. (Utilize $\gamma = 0{,}95$).

16 - Suponha que estejamos interessados em estimar a porcentagem de consumidores de um certo produto. Se uma amostra de tamanho 300 forneceu 100 indivíduos que consomem o dado produto, determine:

a) O intervalo de confiança de *p*, com coeficiente de confiança de 95% (interprete o resultado).

b) O tamanho da amostra para que o erro da estimativa não exceda a 0,02 unidades com probabilidade de 95% (interprete o resultado)

17 - Da experiência passada, sabe-se que o desvio padrão da altura de crianças de 5ª série do 1º grau é 5 cm.

a) Colhendo uma amostra de 36 dessas crianças, observou-se a média de 150 cm. Qual o intervalo de confiança de 95% para a média populacional?

b) Que tamanho deve ter uma amostra para que o intervalo $150 \pm 0{,}98$ tenha 95% de confiança?

18 - 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 $\gamma = 0{,}90$.

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?

c) Suponha que no item (a) não fosse conhecido o desvio padrão. Como você procederia para determinar o intervalo de confiança, e que suposições você faria para isso?

19 - Estime o salário médio dos empregados de uma indústria têxtil, sabendo-se que uma amostra de 100 indivíduos apresentou os seguintes resultados:

(Use $\gamma = 0{,}95$)

U.m.	*Freqüência*
150,00 ├── 250,00	8
250,00 ├── 350,00	22
350,00 ├── 450,00	38
450,00 ├── 550,00	28
550,00 ├── 650,00	2
650,00 ├── 750,00	2

20 - Numa pesquisa de mercado para estudar a preferência da população de uma cidade em relação a um determinado produto, colheu-se uma amostra aleatória de 300 indivíduos, dos quais 180 preferiam esse produto.

a) Determine um intervalo de confiança para a proporção da população que prefere o produto em estudo (α=5%).

b) Determine a probabilidade de que a estimativa pontual dessa proporção não difira do verdadeiro valor em mais de 0,001.

c) É possível obter uma estimativa pontual dessa proporção que não difira do valor verdadeiro em mais de 0,0005 com probabilidade 0,95? Caso contrário, determine o que deve ser feito.

21 - 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, com c.c. $\gamma = 0{,}95$, para a porcentagem de eleitores favoráveis ao candidato A na época das eleições.

22 - Suponha que uma amostra de $n = 100$ de uma distribuição normal $N(\mu, \sigma)$ forneceu $\overline{X} = 510{,}6$. Supondo σ^2 conhecido e igual a 16, obtenha um intervalo de confiança para μ, com c.c. $\gamma = 0{,}95$.

4.7 - Estimativa Intervalar da Média de uma população normal, σ desconhecido (distribuição *t* de *Student*)

Sendo σ desconhecido poderemos estimá-lo usando a estimativa não tendenciosa

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \text{, ou seja } \theta = \sigma^2,\ \hat{\theta} = S^2,\ E(\hat{\theta}) = \theta .$$

Assim, teremos a seguinte variável aleatória:

$$t = \frac{(\overline{X} - \mu)}{S/\sqrt{n}}$$

É intuitivamente evidente que a distribuição de probabilidade da *v.a. t* deve ser muito mais complicada do que a de $Z = \left(\overline{X} - \mu\right)/\left(\sigma/\sqrt{n}\right)$, pois na definição de *t*, ambos, numerador e denominador, são *v.a.* enquanto *Z* é apenas uma função linear de $X_1, X_2, \cdots, X_n$.

A distribuição de *t* é dada por:

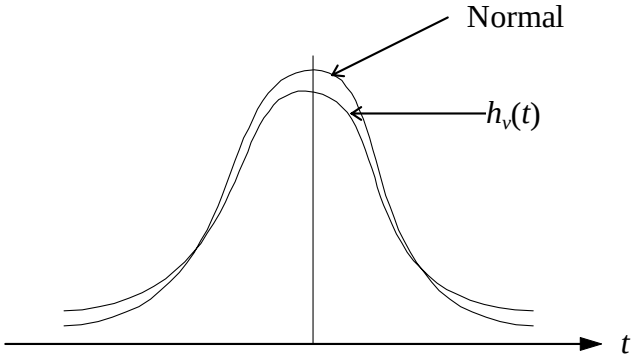
$$h_v(t) = \frac{\Gamma\left[(v+1)/2\right]}{\Gamma(v/2)\sqrt{\pi v}}\left(1 + \frac{t^2}{v}\right)^{-(v+1)/2}, \ -\infty < t < \infty$$

e é denominada de "*distribuição* ***t*** *de Student*", com *v* graus de liberdade.

O gráfico de $h_v(t)$ é simétrico, como está apresentado na figura seguinte. De fato, ele se assemelha ao gráfico da distribuição normal e pode-se verificar que:

$$\lim_{v \to \infty} h_v(t) = \left(1/\sqrt{2\pi}\right) e^{-t^2/2},$$

representando a curva normal.



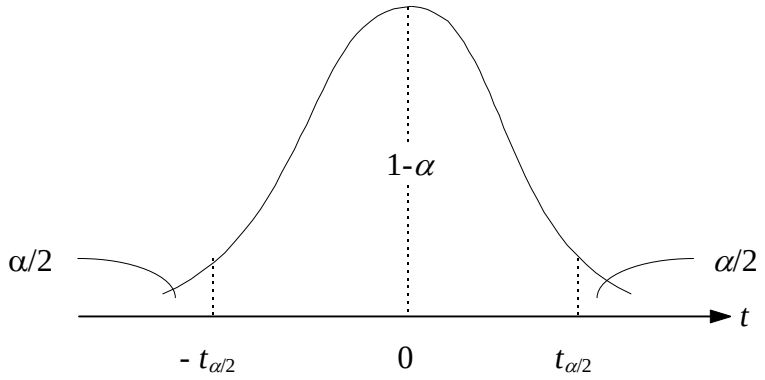
Para cada valor de *v*, temos uma distribuição diferente e, assim, em princípio, precisaríamos de um grande número de tabelas dessa distribuição. Uma situação intermediária

consiste em apresentar, apenas para algumas combinações de valores de v e $1-\alpha$, os valores b tais que $P[-b < t < b] = 1-\alpha$. b é denotado por $t_{\alpha/2}$.

Em virtude da importância desta distribuição, ela foi tabelada.(Ver Tabela 5)

Ex.: $t_{0.005;16} = 2.921$

A partir desse ponto podemos calcular o intervalo de confiança para a média μ, sendo σ desconhecido:



$$P\left[-t_{\alpha/2} < \frac{\overline{X}-\mu}{S/\sqrt{n}} < t_{\alpha/2}\right] = 1-\alpha$$

ou reescrevendo tal expressão:

$$P\left[\overline{X} - t_{\alpha/2}\left(S/\sqrt{n}\right) < \mu < \overline{X} + t_{\alpha/2}\left(S/\sqrt{n}\right)\right] = 1-\alpha, \quad \text{ou seja:}$$

$$\boxed{IC(\mu:(1-\alpha)100) = \left[\overline{X} - t_{\alpha/2}\left(S/\sqrt{n}\right); \overline{X} + t_{\alpha/2}\left(S/\sqrt{n}\right)\right]}$$

Obs.: o valor de v, graus de liberdade, depende fundamentalmente do nº de elementos da amostra e é dado por $v = n - 1$.

Ex.: Dez mensurações são feitas para a resistência elétrica de um certo tipo de fio, fornecendo os valores $X_1, X_2, \cdots, X_{10}$. Suponha que $\overline{X} = 10.48\,\Omega$ e

$$S = \sqrt{\frac{1}{9}\sum_{i=1}^{10}\left(X_i - \overline{X}\right)^2} = 1.36\,\Omega.$$

Vamos supor que X tenha distribuição $N(\mu;\sigma)$ e que desejamos obter um intervalo de confiança para μ, com coeficiente de confiança 0.90. Portanto $\alpha = 0.10$ e $\alpha/2 = 0.05$, $v = n - 1 = 9$.

Da tabela, encontramos que $t_{9;0.05} = 1.833$. Conseqüentemente, $IC(\mu;90\%)$ será:

$$\left[10.48 - (1.83)\left(1.36/\sqrt{10}\right);10.48 + (1.83)\left(1.36/\sqrt{10}\right)\right]$$

$$[9.65;11.31]$$

Obs.: Observe que aumentando o nível de confiança, o intervalo também aumenta. Confira para $IC(\mu;99\%)$.

Obs.: A distribuição aqui apresentada leva este nome em homenagem ao estatístico inglês W. S. Gosset, que publicou sua pesquisa sob o pseudônimo de "Student".

4.8 - Estimativa Intervalar da Variância de uma população normal (distribuição χ^2 - Qui-Quadrado)

Já vimos que uma estimativa pontual não-tendenciosa da variância populacional é a variância amostral S^2, com

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

Vamos estabelecer uma estimativa intervalar para σ^2 utilizando a estatística

$$Y = \frac{(n-1)S^2}{\sigma^2}$$

Essa estatística tem distribuição χ^2, chamada "qui-quadrado" (χ, letra do alfabeto grego que pode ser pronunciada aproximadamente como "qui" em português) se a população é normal. Ou seja:

$$Y: \chi^2_{(v)}$$

Conforme mencionado anteriormente (Capítulo 1) a distribuição χ^2 é um caso

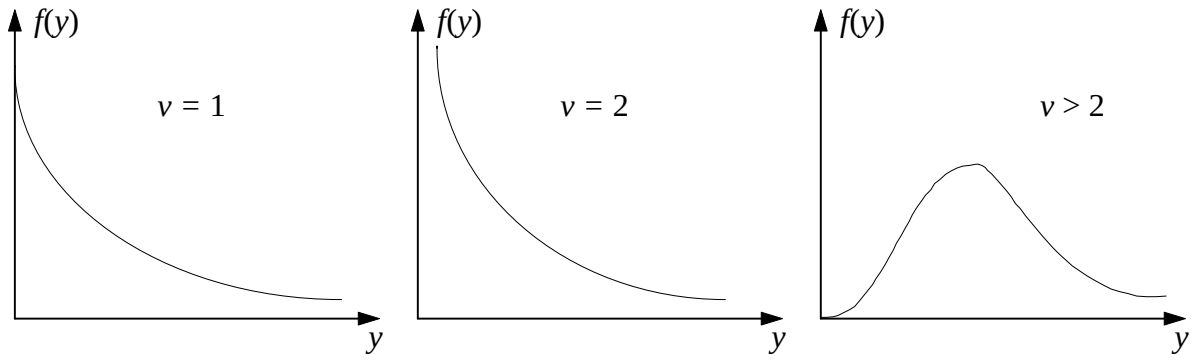
particular da distribuição Gama.

A distribuição χ^2 é dada por:

$$f(y) = \frac{1}{2^{v/2}\Gamma(v/2)} z^{v/2-1} e^{-y/2} \qquad y > 0$$

e se uma *v.a. Y*, que tenha *fdp* dada por tal equação, terá então uma distribuição qui-quadrado, com *v* graus de liberdade.

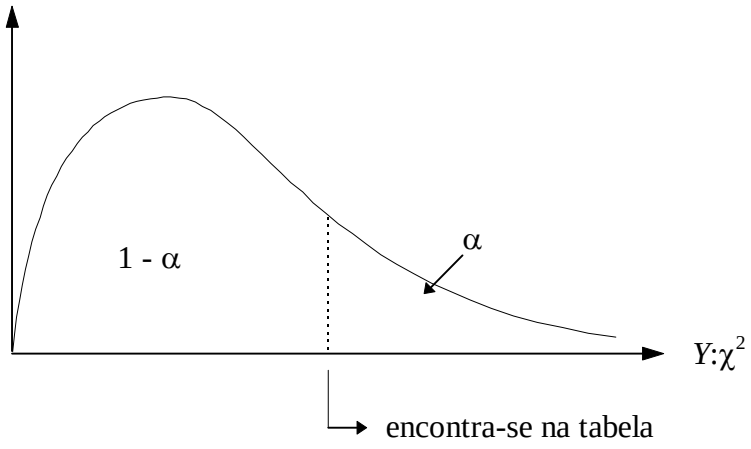
Na figura seguinte a *fdp* está apresentada para *v* = 1, 2 e *v* > 2.



Os valores da média e variância de tal distribuição são dados por:

$E(Y) = v$ $\qquad$ $Var(Y) = 2v$

Em virtude de sua importância, a distribuição χ^2 está tabelada para diversos valores do parâmetro *v* (Veja Tabela 6). Tal tabela fornece os valores da abscissa da distribuição para diversas áreas (probabilidades) da cauda à direita. Assim:

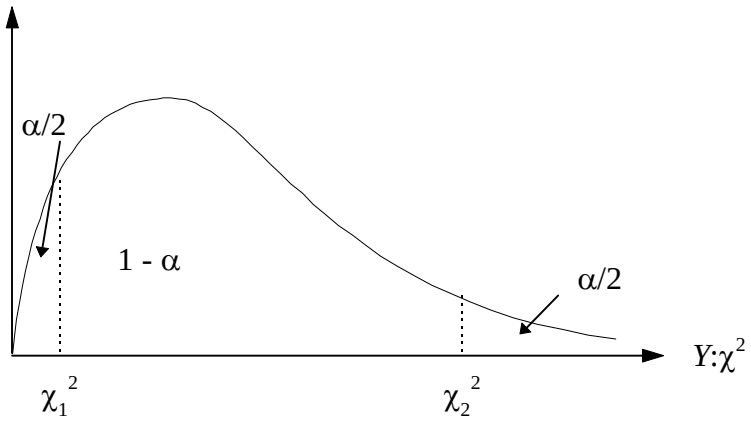


Obs.: Assim como na distribuição de *Student*, *v* = *n* - 1.

Ex.: $v = 9,\ \alpha = 5\% \Rightarrow \chi^2 = 16.9$

Podemos agora calcular o intervalo de confiança para a variância σ^2 por meio de:

$$P\left(\chi_1^2 < Y < \chi_2^2\right) = 1 - \alpha$$



Assim,

$$P\left(\chi_1^2 < \frac{(n-1)S^2}{\sigma^2} < \chi_2^2\right) = 1 - \alpha$$

que reescrita (confira!) fornece:

$$P\left(\frac{(n-1)S^2}{\chi_2^2} < \sigma^2 < \frac{(n-1)S^2}{\chi_1^2}\right) = 1 - \alpha$$

Um intervalo de $100(1-\alpha)\%$ de confiança para a variância de uma população normal é, pois,

$$\boxed{IC(\sigma^2 : (1-\alpha)100 = \left[\frac{(n-1)S^2}{\chi_2^2} ; \frac{(n-1)S^2}{\chi_1^2}\right]}$$

onde S^2 é a variância de uma amostra de tamanho n e χ_1^2 e χ_2^2 são valores obtidos da tabela de distribuição qui-quadrado com n - 1 graus de liberdade.

Ex.: Tem-se os seguintes pesos, em gramas, de 10 pacotes postais remetidos por

certa empresa:

46.4 46.1 45.8 47.0 46.1 45.9 45.8 46.9 45.2 46.0

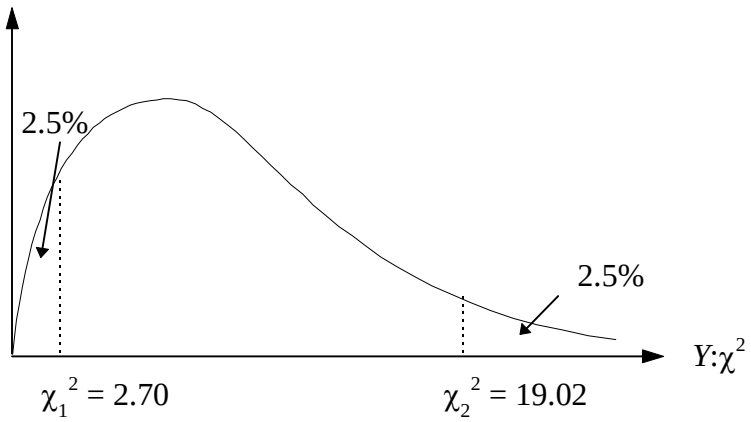
Admitindo normal a distribuição dos pesos, determinar um intervalo de 95% de confiança para a variância de todos os pacotes (população) expedidos pela empresa.

Temos nesse caso que:

$\bar{x} = 46.12$

$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = 0.286$

$\alpha = 5\%, \quad \frac{\alpha}{2} = 2.5\%, \quad v = 10 - 1 = 9$



e $\left[\frac{9.(0.286)}{19.02}; \frac{9.(0.286)}{2.70}\right] = [0.135; 0.953]$.

4.9 - Estimativa Intervalar de Duas variâncias de uma população normal (distribuição F - Fisher/Snedecor)

Uma terceira distribuição muito usada em inferência, envolve o quociente de duas variáveis do tipo χ^2. É ela a distribuição F, de Fisher-Snedecor.

Se duas. *v a* independentes *U* e *V*, cada uma distribuída segundo um χ^2, vom v_1 e

v_2 graus de liberdade, respectivamente, então a v.a. W, dada por:

$$W = \frac{U/v_1}{V/v_2}$$

tem distribuição F, com parâmetros v_1 e v_2 .

A *fdp* de F é dada por:

$$g(w) = \frac{\Gamma\left(\frac{v_1 + v_2}{2}\right)}{\Gamma\left(\frac{v_1}{2}\right)\Gamma\left(\frac{v_2}{2}\right)}\left(\frac{v_1}{v_2}\right)^{\frac{v_1}{2}} \frac{w^{\frac{v_1-2}{2}}}{\left(1+\frac{v_1 w}{v_2}\right)^{\frac{v_1+v_2}{2}}}, w > 0$$

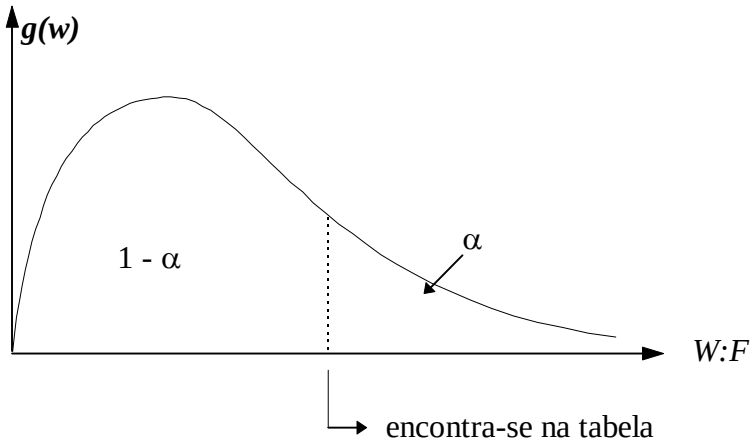
Os parâmetros v_1 e v_2 são chamados de graus de liberdade e se W possui distribuição F então dizemos que:

W:$F(v_1, v_2)$

cuja média e variância são dadas por:

$$\mathrm{E}(W) = \frac{v_2}{v_2 - v_1} \quad \text{e} \quad \mathrm{Var}(W) = \frac{2v_2^2(v_1 + v_2 - 2)}{v_1(v_2 - 2)^2(v_2 - 4)}$$

O gráfico típico da distribuição F está descrito na figura a seguir:



Na tabela 7 (Ver) são dados os pontos f_0 tais que:

$P(F(v_1, v_2) > f_0) = \alpha$ para α=5% e 1% e alguns valores de v_1 e v_2 .

Para encontrar os valores inferiores, usa-se a identidade:

$F(v_1, v_2) = 1/F(v_2, v_1)$

Ex.: Temos que o valor f_0 tal que $P(F(5,7) > f_0) = 5\%$ é 3.97

O valor f_0 tal que $P(F(5,7) < f_0) = 5\%$ pode ser obtido da seguinte forma:

$P(F(5,7) < f_0) = 5\%$

$P(1/F(7,5) < f_0) = 5\%$

$P(F(7,5) > 1/f_0) = 5\%$

Logo, $1/f_0 = 4.88$ (tabela) e $f_0 = 0.205$.

Exercícios XIV

01 - Suponha que $X : N(\mu, \sigma)$, μ e σ desconhecidos. Uma amostra de tamanho n = 600 forneceu $\overline{X} = 10{,}3$ e S^2 = 1,96. Supondo que a v.a. $Z = \dfrac{\overline{X} - \mu}{S/\sqrt{n}}$ seja aproximadamente normal, obtenha um *IC* para μ, com c.c. γ = 0,95. (Se n é pequeno, Z não é aproximadamente normal).

02 - Chamando de $t(v; \alpha)$ o valor y, tal que $P(t > y \mid v) = \alpha$, e usando os valores da Tabela 5, calcule:

a) $t(1; 5\%)$
b) $t(6; 10\%)$
c) $t(10; 95\%)$
d) $t(15; 2{,}5\%)$
e) $t(20; 80\%)$
f) $t(120; 0{,}1\%)$

03 - Da população $X : N(50{,}10)$ retirou-se uma amostra casual simples de tamanho n = 10, calculando-se o valor de $\overline{X}$, S e o respectivo valor de t.

a) Se $P(|\overline{X} - 50| < tS/\sqrt{10} = 10\%$, encontre o valor de t.
b) Se $\overline{X} = 48$ e $S^2 = 120$, qual a probabilidade de encontrar um valor de t menor que o produzido por essa amostra?
c) Se $S^2 = 120$, calcule a $P(|\overline{X} - 50| < 2)$

04 - Chamando de $\chi^2(v;\alpha)$ o valor de y, tal que $P(\chi^2 > y \mid v) = \alpha$, e usando os valores da Tabela 6, calcule:

a) $\chi^2(10; 50\%)$
b) $\chi^2(19; 1\%)$
c) $\chi^2(21; 10\%)$
d) $\chi^2(60; 0{,}1\%)$
e) $\chi^2(1; 2\%)$
f) $\chi^2(8; 30\%)$

05 - De uma população $X : N(50, 10)$ retira-se uma amostra de 10 elementos e calculam-se os valores de σ_*^2 e S^2. Encontre os valores pedidos abaixo, com a maior precisão possível.

a) Se $P(\sigma_*^2 > a) = 10\%$, encontre o valor de a.
b) Sabendo-se que $P(S^2 < a) = 5\%$ e $P(S^2 > b) = 5\%$, encontre a e b.
c) $P(S^2 < 163{,}16) = \alpha$, encontre α.
d) $P(S^2 > 100) = \alpha$, encontre α.
e) $P(S^2 < 80) = \alpha$, encontre α.
f) Se o valor observado de S^2 foi 180, qual a probabilidade de encontrar uma amostra que produza um S^2 maior do que o observado?

06 - Indicando por $F(v_1; v_2; \alpha)$ o número y, tal que $P(F > y \mid v_1, v_2) = \alpha$, calcule, usando a Tabela 7:

a) $F(2, 3, 5\%)$

b) $F(3, 2, 95\%)$

c) $F(1, \infty, 5\%)$

d) $F(120, 120, 5\%)$

e) $F(15, 15, 95\%)$

f) $F(28, 35, 5\%)$

07 - Da população $X : N(50,10)$ retirou-se uma amostra casual simples de $n = 10$ elementos. Da população $Y : N(60,10)$ retirou-se uma amostra casual simples de $m = 6$ indivíduos, independente da primeira. Obtemos as variâncias amostrais S_1^2 e S_2^2, respectivamente.

a) Encontre o valor de a tal que $P\left(\frac{S_1^2}{S_2^2} < a\right) = 95\%$.

b) Encontre o valor de b tal que $P\left(\frac{S_1^2}{S_2^2} > b\right) = 95\%$.

CAPÍTULO 5 -TESTES DE HIPÓTESES

5.1 - Introdução (Um Exemplo)

Comecemos com um exemplo.

Ex.: Determinado tipo de amplificador de isolação é fabricado em Taiwan e possui um determinado parâmetro, máxima temperatura de isolação, que em média vale 155 ^{o}C com desvio padrão 20°C. Os amplificadores com esta característica são vendidos no mercado americano. Caso haja uma falha durante o processo de fabricação com relação a tal parâmetro, os amplificadores são vendidos no mercado paraguaio e possuem uma máxima temperatura média de 145°C com desvio padrão 12°C.

Uma certa indústria utiliza os amplificadores com as características americanas em seus equipamentos produzidos e deseja obter um grande lote desses amplificadores. Entretanto, existe alguma razão para se temer que esse lote possa ser formado por amplificadores com as características paraguaias, o que seria indesejável.

Os amplificadores "americanos" e "paraguaios" são indistingüíveis quanto à aparência e portanto o comprador resolve testar 25 deles, tomados aleatoriamente do lote, para tomar a decisão de compra.

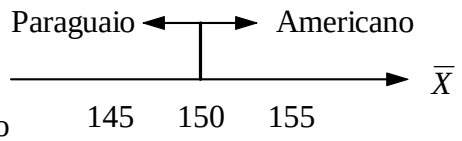
Algumas questões podem agora ser feitas:

A) Que regra de decisão devem usar os responsáveis da indústria para dizer se os amplificadores são compatíveis com o mercado americano ou paraguaio?

Uma resposta que ocorre imediatamente é a que considera como tipo de mercado aquele em que a média amostral ($\overline{X}$) mais se aproxima da média populacional (μ). Assim uma regra de decisão seria:

"Se $\overline{X} \geq 150$ (o ponto médio entre 145 e 155), diremos que os amplificadores são americanos, caso contrário, isto é $\overline{X} < 150$, são paraguaios."

Isto pode ser visto como



Se $\overline{X}$ = 148ºC, por exemplo, de acordo com essa regra de decisão diríamos que os amplificadores são do tipo paraguaio. Poderíamos, contudo estar enganados nesta decisão: Há a probabilidade de que 25 amplificadores do tipo americano apresentem média amostral $\overline{X} = 148^{o}\,\mathrm{C}$.

Então, para melhor entendermos a regra de decisão adotada, é interessante estudarmos os tipos de erros que podemos cometer e as respectivas probabilidades de cometermos estes erros. Tais erros são descritos assim:

Erro I - Dizer que os amplificadores são do tipo paraguaio quando na verdade são americanos. Isto ocorre quando uma amostra de 25 amplificadores do tipo americano apresenta média $\overline{X} \leq 150$ºC.

Erro II - Dizer que os amplificadores são do tipo americano quando na realidade são paraguaio. Isto ocorre quando uma amostra de 25 amplificadores paraguaios apresentam média $\overline{X} > 150$ºC.

Vamos usar a seguinte notação para facilitar a linguagem:

***H_o*:** Os amplificadores são do tipo americano. Isso equivale a dizer que a temperatura *X* de cada amplificador segue uma distribuição com média $\mu = 155^{o}\,\mathrm{C}$ e desvio σ = 20ºC.

Suponha *X: N*(155; 20)

***H_1*:** Os amplificadores são do tipo paraguaio. Analogamente, *X: N*(145; 12)

O Erro I pode ser entendido como o erro principal. Uma analogia pode ser aqui feita. Considere a importante afirmação: “Uma pessoa é considerada inocente até que se prove o contrário, pois é um erro muito mais grave condenar um inocente (Erro I) do que libertar um culpado (Erro II)”.

No nosso exemplo, definido o Erro I, podemos delimitar a região de valores, em que poderemos cometer tal erro. Esta região é definida como **região crítica *RC***, dada nesse caso por:

$RC = \{ y \in R / y \leq 150\}$

Com as anotações anteriores a probabilidade de cometer cada um dos dois erros pode ser assim descrita:

P (Erro I) = P ($\overline{X} \in RC \mid H_o$ é verdadeira) = $\alpha \equiv$ nível de significância e

P (Erro II) = P ($\overline{X} \notin RC \mid H_1$ é verdadeira) = β

α e β podem ser assim calculados:

i) Se H_o é verdadeira, pelos resultados do Teorema Central do Limite (Ver), a média amostral $\overline{X}$ terá distribuição aproximadamente normal, ou seja:

$\overline{X} : N\left(\mu ; \frac{\sigma}{\sqrt{n}}\right)$, isto é, $\overline{X} : N\left(155 ; \frac{20}{\sqrt{25}}\right)$ ou $\overline{X} : N(155 ; 4)$

Assim:

$$P(\text{Erro I}) = P\left(\overline{X} \leq 150 \mid \overline{X} : N(155;4)\right) = P\left(z \leq \frac{150-155}{4}\right)$$
$$= P(z \leq -1.25) = 0.105 = 10{,}5\,\% = \alpha$$

ii) De modo análogo, se H_1 é verdadeira, teremos:

$\overline{X} : N\left(145 ; \frac{12}{\sqrt{25}}\right)$ ou $\overline{X} : N(145 ; 2.4)$

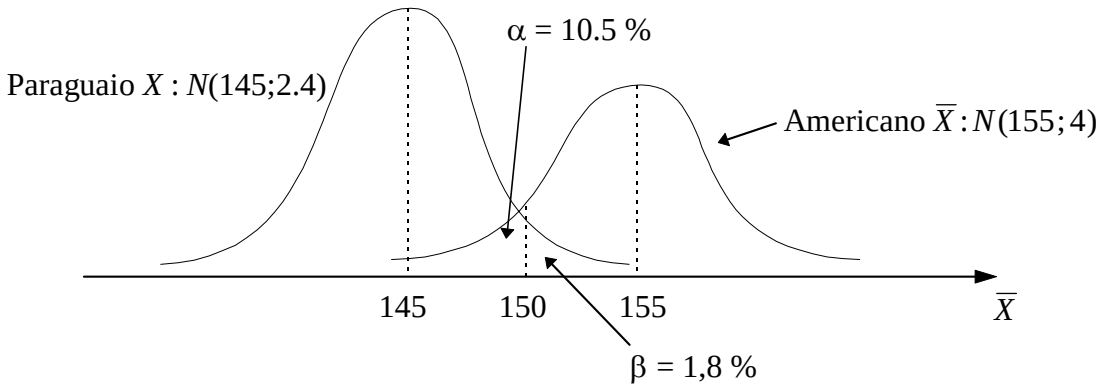
Assim:

$$P(\text{Erro II}) = P\left(\overline{X} > 150 \mid \overline{X} : N(145;2.4)\right) = P\left(z > \frac{150-145}{2.4}\right)$$
$$= P(z > 2.08) = 0.018 = 1{,}8\,\% = \beta$$

Observando estes dois resultados, notamos que, com a regra de decisão adotada, estaremos cometendo o Erro I com maior probabilidade do que o Erro II. De certo modo, esta regra de decisão privilegia a afirmação de que amplificadores são do tipo paraguaio. A tabela seguinte sintetiza tal regra de decisão:

RC ← 150 → $\overline{X}$		Tipo de Amplificador	
		Americano	Paraguaio
Decisão	Americano	Correta	Erro II β = 1,8 %
	Paraguaio	Erro I α = 10,5 %	Correta

E a figura seguinte ilustra todo esse processo:



Para cada regra de decisão adotada, isto é, se escolhermos um valor $\overline{x}_c$ em vez de 150, os valores de α e β mudarão. Observe que se $\overline{x}_c$ for menor que 150, α diminui e β aumenta.

Para $\overline{x}_c$ = 148.75 teremos que α = β = 5,94%.

Do exposto anteriormente, constatamos que, escolhido um valor $\overline{x}_c$ poderemos achar α e β. Mas também podemos proceder de modo inverso: fixar um dos erros, digamos α, e encontrar a regra de decisão para tal erro. Por exemplo, fazendo α = 5%, teremos:

$$P(\text{Erro I}) = P\left(\overline{X} \le \overline{x}_c \mid \overline{X} : N(155;4)\right) = 5\%$$
$$= P(Z \le -1.645)$$

usando a transformação: $Z = \dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}}$ temos:

$$-1.645 = \frac{\overline{x}_c - 155}{4} \Rightarrow \overline{x}_c = 148.42 \ ^{\circ}\text{C}$$

Então a regra de decisão será:

"Se $\overline{X}$ for inferior à 148.42, dizemos que o lote é do tipo paraguaio; caso contrário, dizemos que é americano, ao nível de 5%."

β pode ser assim calculado:

$$\beta = P(\text{Erro II}) = P\left(\overline{X} > 148.42 \mid \overline{X} : N(145;2.4)\right)$$
$$= P(Z > 1.425) = 7{,}93\%$$

Esse tipo de procedimento (fixar α) é o mais utilizado.

"**B**) Suponhamos agora que devido a grande procura desses amplificadores, diversas pequenas indústrias resolvem investir na sua fabricação, tentando imitar ao máximo os amplificadores do tipo americano. Sabe-se contudo que nenhuma dessas imitações consegue atingir para o parâmetro, máxima temperatura de isolação, a média de 155°C. Qual a regra de decisão que deve agora adotar a indústria compradora?"

A hipótese que nos interessa mais de perto nesse caso é:

H_o : Os amplificadores são do tipo americano ($X : N$ (155 ; 20)).

A hipótese alternativa é muito mais ampla e pode ser expressa como:

H_1 : Os amplificadores não são do tipo americano ($\mu < 155$).

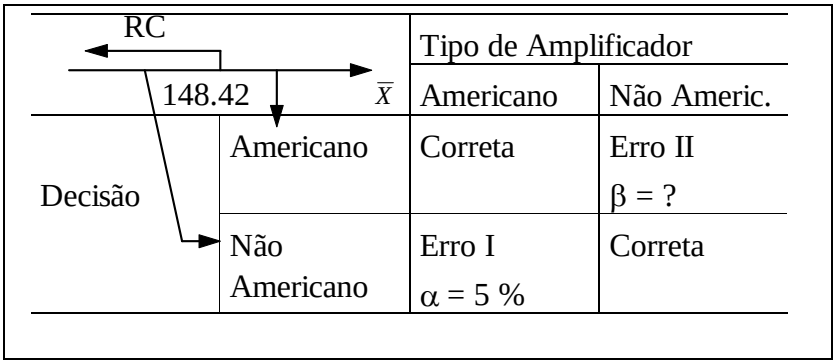
Isso significa que só iremos desconfiar de H_o se $\overline{X}$ for muito menor do que 155. Ou seja, a nossa regra de decisão deverá ser semelhante à vista anteriormente. Como os parâmetros, alternativos são muitos, a melhor solução para construir a regra de decisão é fixar α, a probabilidade de Erro I , ou seja,

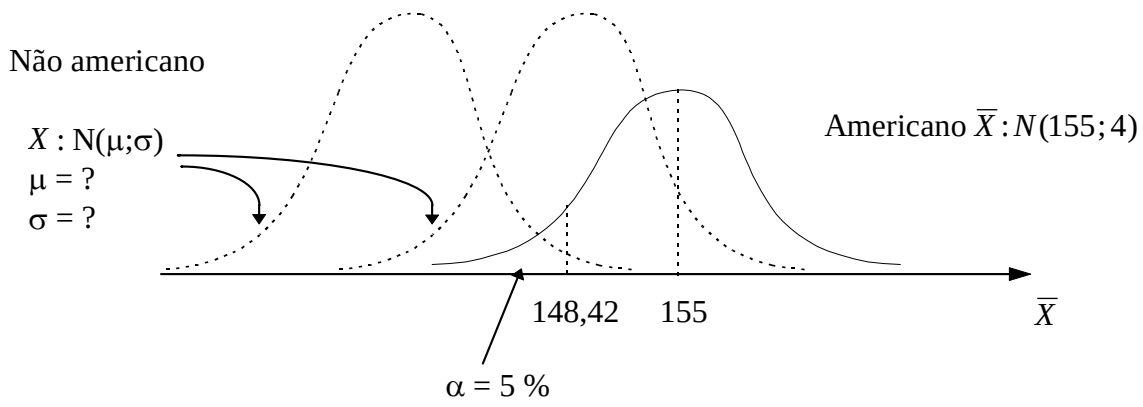
Erro I: Rejeitar H_o sendo H_o verdadeira.

Fixando α em 5%, por exemplo, teremos a seguinte regra de decisão:

"Se $\overline{X}$ for superior a 148.42 ^{o}C diremos que o lote é do tipo americano; caso contrário, diremos que não é, ao nível de 5%."

Com essa regra de decisão e com a hipótese alternativa mais ampla, não podemos encontrar β, pois não temos um único parâmetro μ como alternativa. Então, aqui não podemos controlar o Erro II. A tabela e a figura seguinte expressam a regra de decisão adotada:





"**C**) Admitamos agora que não exista razão alguma para acreditarmos que a máxima temperatura de isolação dos amplificadores, produzidos pelas indústrias concorrentes, seja menor do que os do tipo americano. Contudo, os compradores preferencialmente gostariam de ainda assim obter os amplificadores desse tipo. Qual regra deveria ser agora adotada?"

Essa nova situação irá nos levar a duvidar de que não são do tipo americano, se a média observada for muito maior ou muito menor do que 155°C. Isto equivale às seguintes hipóteses:

H_o: Os amplificadores são do tipo americano ($X : N$ (155; 20)).

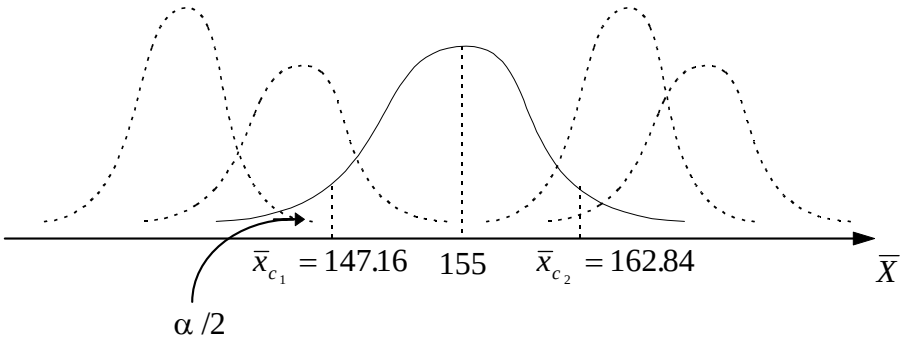
H_1: Os amplificadores não são do tipo americano ($\mu \neq 155$).

Assim, a regra de decisão deverá indicar dois pontos $\overline{x}_{c_1}$ e $\overline{x}_{c_2}$, tais que:

"Se $\overline{X}$ estiver entre $\overline{x}_{c_1}$ e $\overline{x}_{c_2}$, diremos que os amplificadores são do tipo americano; caso contrário, não são, ao nível de α%".

Fixando α, a probabilidade do Erro I, existirão muitos valores que satisfazem essa condição. A tabela e a figura seguintes esclarecem a regra de decisão para α = 5%.

RC 155 RC		Tipo de Amplificador	
147.16	162,84	Americano	Não Americ.
Decisão	Americano	Correta	Erro II β = ?
	Não Americano	Erro I α = 5 %	Correta



onde:

$$P(\text{Erro I}) = P(\bar{X} \le \bar{x}_{c_1} \text{ ou } \bar{X} > \bar{x}_{c_2} \mid \bar{X}: N(155\,;4)) = 5\%$$

$$= P(Z < -1.96 \text{ ou } Z > 1.96)$$

e

$$-1.96 = \frac{\bar{x}_{c_1} - 155}{4} \quad \therefore \quad \bar{x}_{c_1} = 147.16$$

$$-1.96 = \frac{\bar{x}_{c_2} - 155}{4} \quad \therefore \quad \bar{x}_{c_2} = 162.84.$$

A região de rejeição da hipótese H_o será:

$RC = \{ y \in \mathbf{R} \mid y < 147.16 \text{ ou } y > 162.84\}$

Exercícios XV

01- Para decidirmos se os habitantes de uma ilha são descendentes da civilização *A* ou *B*, iremos proceder do seguinte modo:

(i) selecionamos uma amostra de 100 moradores adultos da ilha, e determinamos a altura média deles;

(ii) se essa altura média for superior a 176, diremos que são descendentes de *B*: caso contrário, são descendentes de *A*.

Os parâmetros das alturas das duas civilizações são:

$A : \mu = 175$ e $\sigma = 10$

$B : \mu = 177$ e $\sigma = 10$

Definamos: Erro I - dizer que os habitantes da ilha são descendentes de *B* quando, na realidade, são de *A*.

Erro II - dizer que são de *A* quando, na realidade, são de *B*.

a) Qual a probabilidade do Erro I? E do Erro II?

b) Qual deve ser a regra de decisão se quisermos fixar a probabilidade do Erro I em 5%? Qual a probabilidade do Erro II, nesse caso?

c) Se $\sigma_A = 5$, como ficariam as respostas de (b)?

d) Quais as probabilidades do Erro II, nas condições da questão (b), se a média $\mu_B = 178$? E $\mu_B = 180$? E $\mu_B = 181$? Coloque num gráfico os pares (μ_B, P(Erro II/μ_B)).

02 - Fazendo o teste

$H_0 : \mu = 1.150$ ($\sigma = 150$), contra $H_1 : \mu = 1.200$ ($\sigma = 200$),

e $n = 100$ estabeleceu-se a seguinte região crítica: $RC = [1.170, +\infty[$.

a) Qual a probabilidade α de rejeitar H_0 quando verdadeira?

b) Qual a probabilidade β de aceitar H_0 quando H_1 é verdadeira?

c) Qual deve ser a região crítica para que $\alpha = \beta$?

03 - Nas situações abaixo, escolha como hipótese nula, H_0, aquela que para você leva a um erro de primeira espécie mais importante. Descreva quais os dois erros em cada caso.

a) O trabalho de um operador de radar é detectar aeronaves inimigas. Quando surge alguma coisa estranha na tela, ele deve decidir entre as hipóteses:

- está começando um ataque;

- tubo bem, apenas uma leve interferência.

b) Num júri, um indivíduo está sendo julgado por crime. As hipóteses sujeitas ao júri são:

- o acusado é inocente;
- o acusado é culpado.

c) Um pesquisador acredita que descobriu uma vacina contra resfriado. Ele irá conduzir uma pesquisa de laboratório para verificar a veracidade da afirmação. De acordo com o resultado, ele lançará ou não a vacina no mercado. As hipóteses que pode testar são:

- a vacina é boa;
- a vacina não é boa.

04 - Se, ao lançarmos três vezes uma moeda, aparecem 3 coroas, decidimos rejeitar a hipótese de que a moeda é "honesta". Quais as probabilidades de Erro I e Erro II?

05 - A variável X, custo de manutenção de um tear, pode ser considerada como tendo distribuição normal e média μ podem ser 200 ou 210. Para verificar qual dos dois valores é o mais provável, usar-se-á uma amostra de 25 teares. Defina:

a) uma hipótese a ser testada;

b) uma regra de decisão e encontre as probabilidades dos erros de tipo I e II.

5.2 - Procedimento Geral de Testes de Hipóteses

A construção de um teste de hipótese, para um parâmetro populacional, pode ser colocada do seguinte modo. Existe uma *v.a. X* em cada população. Tem-se uma hipótese sobre determinado parâmetro θ dessa população. Por exemplo, afirmamos que esse valor é um número θ_o. Colhe-se uma amostra aleatória de elementos dessa população, e através dela deseja-se comprovar ou refutar tal hipótese.

Como já vimos anteriormente, iniciamos explicitando claramente qual a hipótese que estamos colocando à prova, e a chamamos de Hipótese nula H_o. No nosso caso:

$H_o : \theta = \theta_o$

Em seguida convém explicitar também a hipótese alternativa H_1, que será considerada como aceitável, caso H_o seja rejeitada. A caracterização estatística de H_1 será dada por:

$H_1 : \theta \neq \theta_o$ (teste bilateral), ou

$H_1 : \theta > \theta_o$ (teste unilateral), ou

$H_1 : \theta < \theta_o$ (teste unilateral),

dependendo das informações que o problema traz.

Qualquer que seja a decisão tomada, já vimos que estamos sujeitos a cometer erros. Tais erros são:

- Erro do Tipo I $\equiv$ Rejeitar H_o sendo H_o verdadeira, com a seguinte probabilidade de ocorrência:

  P (Erro I) = P (rejeitar H_o | H_o é verdadeira) = α
- Erro do Tipo II $\equiv$ Não rejeitar H_o sendo H_o falsa, com a seguinte probabilidade.

  P (Erro II) = P (não rejeitar H_o | H_o é falsa) = β

O objetivo do teste de hipótese é dizer, através de uma estatística $\hat{\theta}$ obtida de uma amostra, se H_o é ou não aceitável. Operacionalmente isso é conseguido através da região crítica *RC*. Tal região é construída em função de α. A determinação de β já é mais difícil e pode ser melhor compreendida pela Curva Característica de Operação, que será vista adiante.

Uma seqüência lógica que pode ser usada sistematicamente para qualquer teste de hipótese é dada a seguir. Sua implementação computadorizada tem sido freqüentemente utilizada em forma de algoritmo:

Primeiro passo:	Definir H_o : $\theta = \theta_0$ e H_1
Segundo passo:	Escolher a Estatística de teste adequada, $\hat{\theta}$
Terceiro passo:	Escolher α e estabelecer *RC*
Quarto passo:	Calcular $\hat{\theta}$ com base em uma amostra de tamanho *n* extraída da população.
Quinto passo:	Rejeitar H_o caso $\hat{\theta} \in RC$. Não rejeitar H_o em caso contrário.

Ex.: Para o exemplo dos amplificadores "americanos" da seção anterior (caso B) teríamos: $X: N(155;20)$ Tipo "Americano"

$X: N(<155;20)$ Tipo "Não Americano"

Assim, o algoritmo acima pode ser aplicado:

1º passo: $H_0 : \mu = 155$ (Amplificadores são "Americanos")

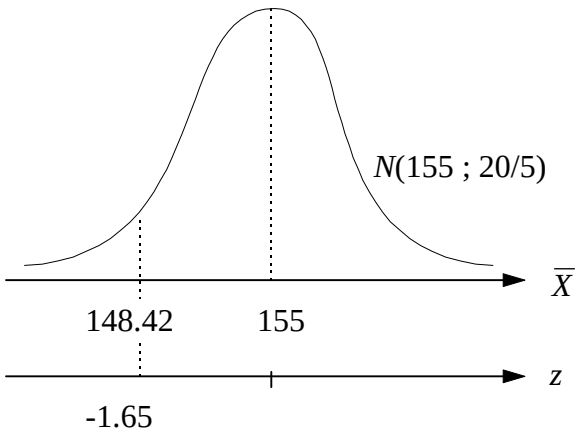
$H_1 : \mu < 155$ (Amplificadores são "Não Americanos")

2º passo: $X : N(155; 20)$

$\overline{X} : N\left(155; \frac{20}{\sqrt{25}}\right)$ (Média Amostral dada pelo Teor.Central do Limite)

$$z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}}$$

3º passo: $\alpha = 5\% \Rightarrow z = -1.65$ Teste unilateral



$RC = \{y \in \mathrm{R} \mid y < 148.42\}$

4º passo: Cálculo da média amostral com base nas 25 amostras

5º passo: Caso a média amostral pertença à região crítica, rejeitamos H_o e aceitamos H_1 (Dizemos que os amplificadores são do tipo "Não Americano"). Em caso contrário aceitamos H_o..

Algumas situações de testes são mais freqüentes do que outras. Assim, adiante, veremos exemplificados os seguintes testes:

1) Teste sobre a média de uma população normal com variância conhecida.
2) Teste sobre a proporção.
3) Teste sobre a média de uma população normal com variância desconhecida.
4) Teste sobre a variância de uma população normal.
5) Comparação das variâncias de duas populações normais.
6) Comparação de duas médias de populações normais, variâncias iguais, desconhecidas.
7) Comparação de duas médias de populações normais, variâncias desiguais, desconhecidas.
8) Observações emparelhadas.
9) Teste de independência.
10) Teste sobre coeficiente de correlação

5.3 - Curva Característica de Operação (CCO) e Poder de um Teste

Vimos que, na construção do teste de hipóteses, procuramos controlar o Erro I, fixando α, a sua probabilidade de ocorrência . Fixado esse número, a região crítica é construída de modo que

$P\,(\hat{\theta} \in RC \mid H_o$ é verdadeira) = α

Obs.: No exemplo dos amplificadores "Americanos", $\hat{\theta} = \overline{X}$

Já a probabilidade do Erro II, β, na maioria dos casos não é possível calcular, pois usualmente a hipótese alternativa não especifica uma única possibilidade, mas uma família de possibilidades alternativas. Temos para o Erro II que

$P\,(\hat{\theta} \notin RC \mid H_1$ é verdadeira) = β ou

$P\,(\hat{\theta} \in RC' \mid H_1$ é verdadeira) = β

sendo *RC'* o complementar de *RC* em relação a **R**.

Ex.: Sendo X o custo de manutenção de um tear uma variável aleatória $X : N(\mu\,; 20)$, para testar as hipóteses.

$H_o : \mu = 200$

$H_1 : \mu > 200$,

usando uma amostra de 25 teares e fixando α = 5%, temos, segundo os resultados do Teorema Central do Limite, que a média amostral $\overline{X}$ possui distribuição $\overline{X} : N\left(\mu, \sigma \mid \sqrt{n}\right)$,

ou seja:

$$\overline{X} : N\left(\mu; \frac{20}{\sqrt{25}}\right) \quad \text{ou} \quad \overline{X} : N\left(\mu\,; 4\right).$$

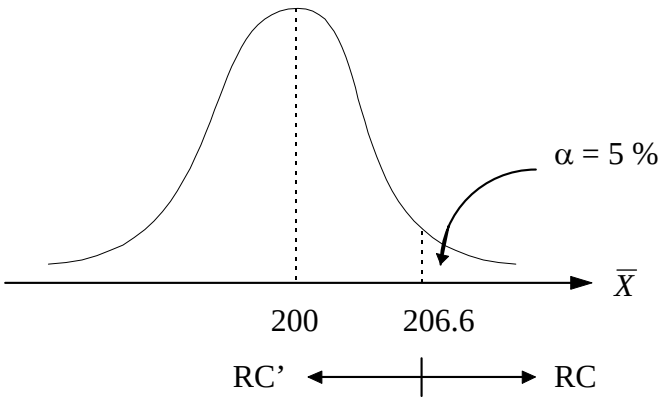
Aqui, a *RC* pode ser facilmente calculada a partir de α = 5% e da expressão:

$$z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \quad \text{com } 1.65 = \frac{\overline{x}_c - 200}{4}, \text{ resultando } \overline{x}_c = 206.6$$

Assim: $RC = \{\, y \in \mathbf{R} \mid y > 206.6 \,\}$ e a nossa regra de decisão será:

"Se $\overline{X} \in RC$, o custo de manutenção é superior a 200, caso contrário, não é."

A figura seguinte ilustra esse processo.



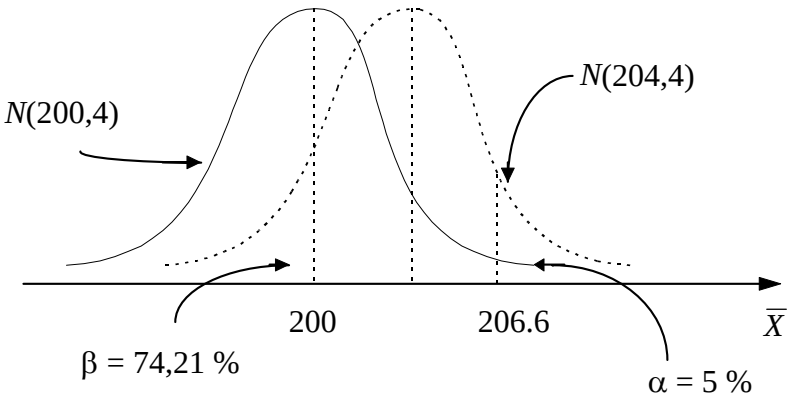
Vamos supor que, em realidade, a hipótese testada H_o seja falsa. A probabilidade do Erro II, não pode ser calculada, a menos que se especifique um valor alternativo para μ. Por exemplo, se $\mu = 204$, a média $\overline{X}$ de 25 teares terá distribuição.

$\overline{X} : N(204, 4)$

portanto,

$$\beta = P\left(\overline{X} \in RC' \mid \mu = 204\right) = 74{,}21\%$$

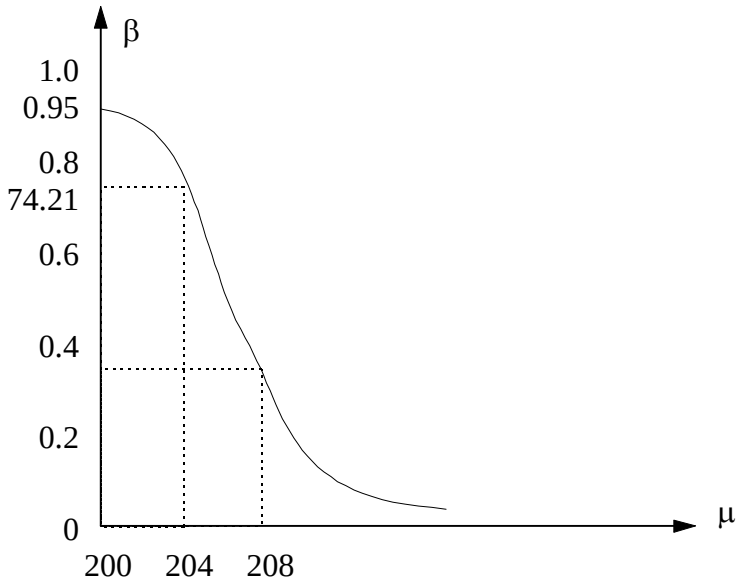
pois: $$z = \frac{\overline{x}_c - \mu}{\sigma/\sqrt{n}} = \frac{206.6 - 204}{4} = 0.65 \Rightarrow \beta = 74.21\%$$



Obs.: β é também a probabilidade de aceitar H_0 para os diversos valores de μ.

Atribuindo alguns valores alternativos a μ, obtemos a tabela e a figura seguinte:

μ	201	202	203	204	205	208
β	91.92	87.49	81.59	74.21	65.54	36.31



Tal curva é denominada Curva Característica de Operação (C.C.O.) deste teste . (H_1 : $\mu > 200$, $\alpha = 5\%$, σ conhecido).

Genericamente a Curva Característica de Operação costuma ser dada, para α fixo, em função da distância $\mu - \mu_o$ padronizada, isto é, medida em termos do desvio padrão da população, ou seja, designando por *d* essa distância padronizada, temos:

$$H_1: \mu > \mu_0 \quad \Rightarrow \quad d = \frac{\mu - \mu_0}{\sigma}$$

ou $$H_1: \mu < \mu_0 \quad \Rightarrow \quad d = \frac{\mu_0 - \mu}{\sigma}$$

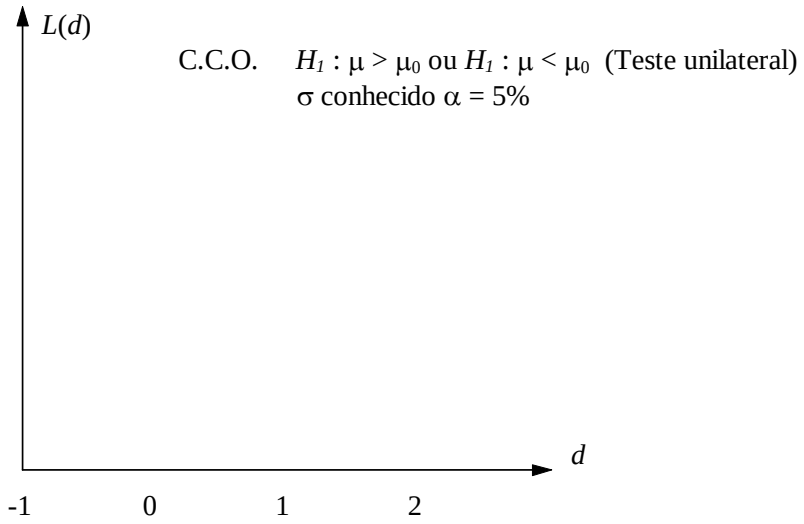
ou $$H_1: \mu \neq \mu_0 \quad \Rightarrow \quad d = \frac{|\mu - \mu_0|}{\sigma}$$

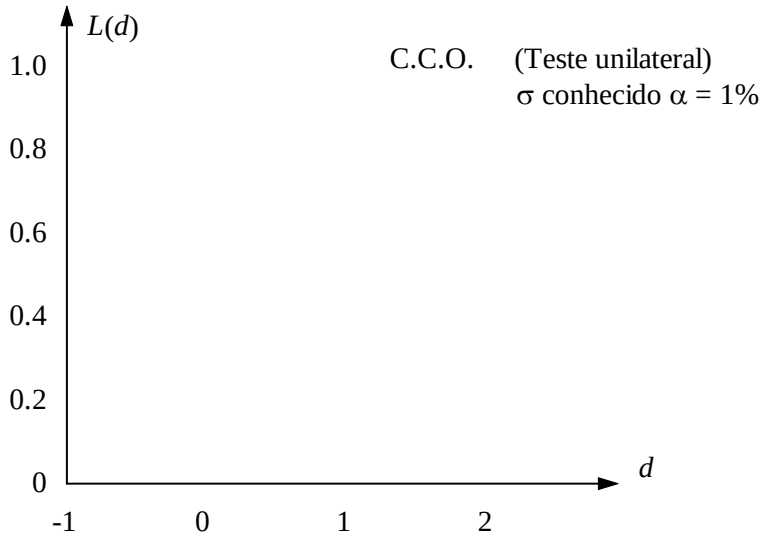
Em suma, o que se pretende quantificar é o afastamento entre o valor real do parâmetro (μ alternativo) e o valor testado (μ_o).

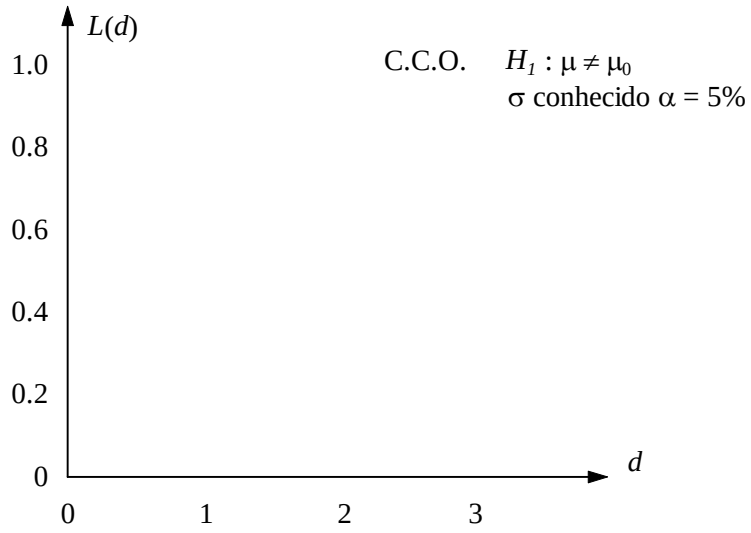
Com o intuito de tornar gerais as C.C.O., vamos também substituir β pela probabilidade de aceitar H_o, à qual designaremos por L (d) (Obs.: $L(d) = \beta$). Tal generalização tem a vantagem de tornar a curva utilizável, independentemente dos particulares valores de μ_o ou σ, devido *d* ser padronizado

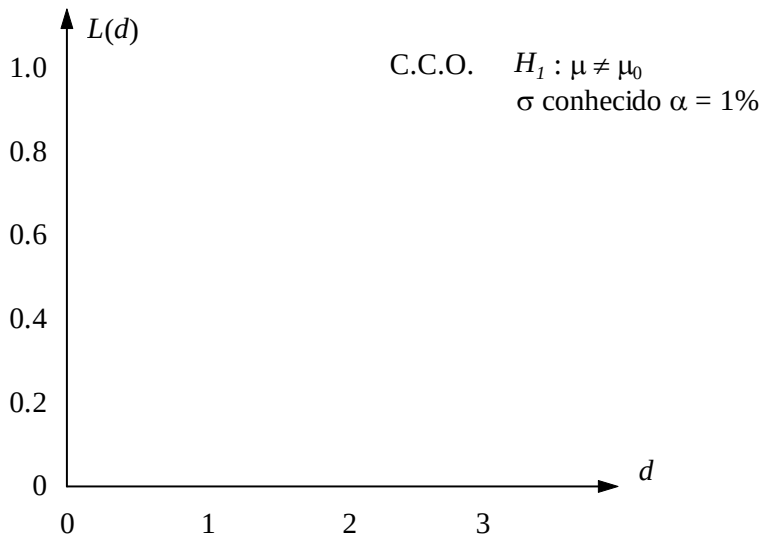
Entretanto, devemos considerar ainda que a variação de β (ou L (d)), depende, também, fundamentalmente, do tamanho da amostra *n*. Observando as curvas das figuras seguintes, notamos que o teste é mais sensível para os maiores valores de *n* (pequenas variações em *d* acarretam grandes variações em L (d)). Os testes mais sensíveis devem ser

preferidos. Este conceito está de acordo com a idéia de que um teste com amostras maiores deve levar a resultados melhores.









Uma análise dessas curvas mostra que, nas proximidades do valor testado ($d = 0$), a probabilidade de aceitação de H_o é sempre muito alta , bastante próxima de 1 - α. Logo se H_1 for verdadeira, a gravidade do Erro II não é relevante pois a diferença prática entre a realidade e a hipótese testada seria pequena. A gravidade do Erro II se acentua à medida que o verdadeiro valor do parâmetro se afasta do valor testado pressuposto em H_0. Nessas condições, aceitar H_0 pode ser altamente indesejável, mas para tais casos, β tende a diminuir.

Devemos, pois, estabelecer até que ponto uma divergência entre a realidade e H_0 pode ser tolerada. Seja μ' esse ponto e d' a distância padronizada correspondente. Fixamos, então, uma probabilidade β máxima de se cometer o Erro II se $d \geq d'$, caso em que esse erro causa preocupação. Com d' e β, caracterizamos um ponto no gráfico das CCOs. A curva que passa por esse ponto define o tamanho da amostra necessária à realização do teste com α, β e d' fixados.

As CCOs indicam o poder discriminatório dos testes. Assim, pela análise de suas figuras, vimos que, quanto maior a amostra, mais factível será distinguirmos, com probabilidade de erro fixada, uma pequena diferença entre o valor real do parâmetro e o valor testado. Graficamente, as curvas que correspondem aos testes mais poderosos são as que apresentam maior inclinação.

Ex.: Para um caso em que a carga média de ruptura especificada para certos parafusos seja de 50kg, sendo o desvio padrão dessas cargas igual a 4kg, suponhamos que um comprador especifique também que:

a) se o lote satisfaz à especificação, o comprador deseja limitar a 5% a

probabilidade de concluir que o lote é insatisfatório;

b) se o lote tiver uma resistência média ligeiramente menor que 50kg, tal fato não causa preocupação, porém deseja-se que, se a verdadeira resistência média for inferior a 48kg, tal fato seja identificado com pelo menos 90% de probabilidade.

Nessas condições, qual o tamanho da amostra mínima necessária e qual o limite da região crítica?

Solução: O teste a ser feito será:

$H_0 : \mu = 50\text{kg}$

$H_0 : \mu < 50\text{kg}$

sendo que as condições (a) e (b) indicam α=5% e β=10%, este associado a um *d*' tal que:

$$d' = \frac{\mu_0 - \mu'}{\sigma} = \frac{50 - 48}{4} = 0.5 \ .$$

Entrando na CCO com esses valores, vê-se que a amostra deverá ter cerca de 35 elementos.

Dimensionada a amostra, podemos, através do Teorema Central do Limite determinar o limite da região crítica:

$$\bar{x}_1 = \mu_0 - z_{5\%} \frac{\sigma}{\sqrt{n}} = 50 - 1.645 \frac{4}{\sqrt{35}} \cong 48.89\text{kg}$$

Expressões analíticas para *n*

Alternativamente às CCOs, expressões analíticas para a determinação do tamanho da amostra podem ser também usadas.

Ex.: Derivemos uma dessas expressões. Sejam as hipóteses

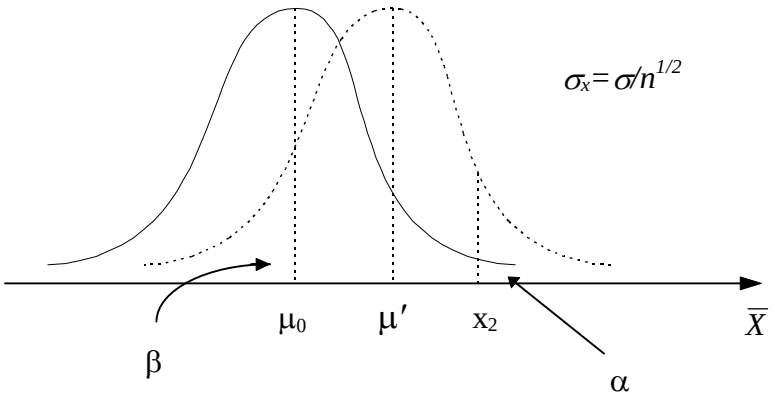
$H_0 : \mu = \mu_0$

$H_0 : \mu > \mu_0$

com σ conhecido e fixados α, β e μ. Conforme vimos, a fixação de $\mu' > \mu_0$ equivale a se admitir que, se o valor real do parâmetro μ for, em verdade, superior a μ_0 , porém não ultrapassando μ' , não nos importaremos em cometer o erro II, pois a aceitação de H_0 em tais casos não trará consequências consideráveis. Se porém, tivermos, em realidade, $\mu' < \mu$, desejaremos limitar a probabilidade de aceitar H_0 nessas condições a um valor máximo fixado β. Essa probabilidade será associada ao ponto μ' e, assim, garantimos que, se $\mu' > \mu$, $L(d) \leq \beta$.

Temos então que, se $\mu = \mu_0$, a probabilidade de se rejeitar H_0 deverá ser α, e, se

$\mu = \mu'$, a probabilidade de se aceitar H_0 deverá ser β. Admitindo-se que σ seja constante com μ, tal situação está mostrada na figura seguinte, onde $\bar{x}_2$ é o limite da região crítica e as curvas apresentadas representam as distribuições amostrais de $\bar{x}$ se $\mu = \mu_0$ e se $\mu = \mu'$.



A expressão seguinte, já vista, fornece

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

Por outro lado, também podemos escrever

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

Logo, igualando as duas expressões, podemos obter:

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

Essa expressão nos fornece o tamanho mínimo da amostra para satisfazer ás condições impostas. É evidente que a expressão de *n* pode ser utilizada indistintamente para testes unilaterais à direita ou à esquerda. Muitas outras dessas expressões podem ser obtidas de diferentes outros testes

Diversos autores preferem apresentar as C.C.O. sob a forma de Curvas de Poder do Teste, nas quais se plotam valores de 1 - *L* (*d*).em função de *d*.

Programas de computador como o ***Statgraf*** fornecem essas curvas de uma maneira bastante simples.

Exercícios XVI

01 - Suponha que estejamos testando $H_0 : p = 0{,}5$ contra $H_1 : p \neq 0{,}5$ e que, para uma amostra de tamanho $n = 10$, decidimos pela região crítica $RC = \{0, 1, 2, 8, 9, 10\}$.

a) Determine o nível de significância x.

b) Calcule a CCO do teste para $p = 0{,}2$, $p = 0{,}4$, $p = 0{,}6$ e $p = 0{,}8$. Faça um gráfico da CCO

c) Qual o poder do teste para $p = 0{,}5$?

02 - Sendo X o custo de manutenção de um tear, sabe-se que $X : N(\mu, 20)$. Para testar a hipótese $H_0 : \mu = 200$, contra a alternativa $H_1 : \mu > 200$, será usada uma amostra de 25 teares.

a) Fixando-se $\alpha = 5\%$, encontre a correspondente RC.

b) Atribuindo-se valores arbitrários para μ, esboce a CCO do teste.

c) Para que valores de μ, o poder será maior do que 50%?

5.4 - Testes paramétricos

5.4.1 - Teste sobre a Média de uma População Normal com Variância Conhecida

"**Ex.:** A associação dos proprietários de indústrias metalúrgicas está muito preocupada com o tempo perdido com acidentes de trabalho, cuja média, nos últimos anos, tem sido da ordem de 60 horas/homem por ano e desvio padrão de 20 horas/homem. Tentou-se um programa de prevenção de acidentes e, após o mesmo, tomou-se uma amostra de 9 indústrias e mediu-se o número de horas/homem perdidas por acidente, que foi de 50 horas. Você diria, ao nível de 5%, que há evidências de melhoria? Calcule um intervalo de confiança de 95% para a média."

Solução:

1º passo: $H_0 : \mu = 60$ (Não houve melhora)

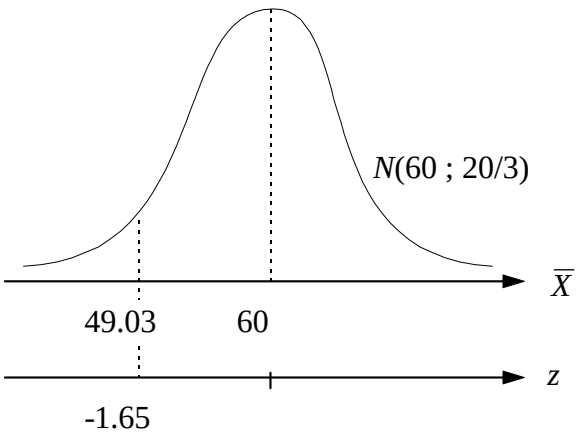
$H_1 : \mu < 60$ (Houve melhora)

2º passo: $X : N(60; 20)$

$$\overline{X} : N\left(60; \frac{20}{\sqrt{9}}\right)$$ (Dado pelo Teorema Central do Limite)

$$z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}}$$

3º passo: $\alpha = 5\% \Rightarrow z = -1.65$ Teste unilateral



$RC = \{y \in \mathrm{R} \mid y < 49.03\}$

4º passo: $n = 9$ $\overline{X} = 50$ (Das amostras)

5º passo: como $\overline{X} = 50 \notin RC$, não rejeitamos H_o.

"Não há evidência, ao nível de 5%, de que o programa tenha ocasionado melhoras".

Cálculo de *IC* (μ: 95%):

$$P\left(\overline{X} - z_{\alpha/2}\,\sigma/\sqrt{n} \le \mu \le \overline{X} + z_{\alpha/2}\,\sigma/\sqrt{n}\right) = 1-\alpha$$

$$IC(\mu{:}95\%) = \left(\overline{X} - z_{\alpha/2}\,\sigma/\sqrt{n} \le \mu \le \overline{X} + z_{\alpha/2}\,\sigma/\sqrt{n}\right) = \left[50 - 1.96\left(20/\sqrt{9}\right); 50 + 1.96\left(20/\sqrt{9}\right)\right]$$
$$= \left[36.94\,;63.07\right]$$

5.4.2 - Teste Sobre a Proporção

"**Ex.: a**) Um industrial deseja certificar-se de que a fração de mercado que prefere seu produto ao de seu concorrente é superior a 70%. Para tanto colheu uma amostra aleatória de 165 opiniões, das quais 122 lhe foram favoráveis. Pode o industrial ficar satisfeito com esse resultado, adotado o nível de significância de 5%?"

Solução:

1º passo: $H_0 : p = 70\% = 0.7$ (O industrial não está seguro)

$H_1 : p > 0.7$ (O industrial está seguro)

2º passo: O melhor estimador para p é $\hat{p}$ com:

$$\hat{p} = \frac{x}{n}$$

sendo: $n \equiv$ número de elementos da amostra

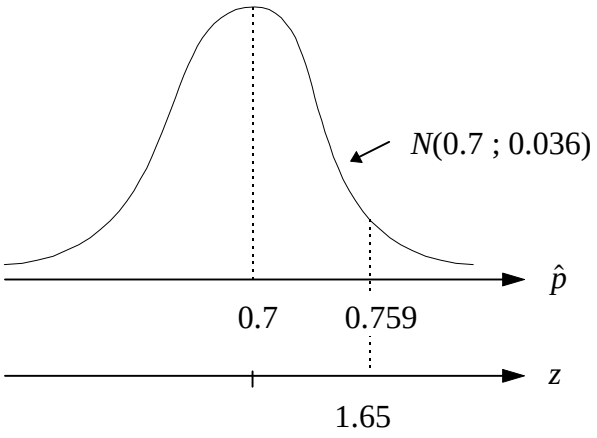
$x \equiv$ característica de interesse da amostra

$$\hat{p} : N\left(p; \sqrt{\frac{p.q}{n}}\right)$$

ou seja: $\hat{p} : N\left(0.7; \sqrt{\frac{(0.7)(0.3)}{165}}\right) = \hat{p} : N\left(0.7;\, 0.036\right)$

$$z = \frac{\hat{p} - \mu}{\sigma/\sqrt{n}} = \frac{\hat{p} - 0.7}{0.036}$$

3º passo: $\alpha = 5\% \quad \Rightarrow \quad z = 1.65$ Teste unilateral



$RC = \{y \in \mathrm{R} \mid y > 0.759\}$

4º passo: $n = 165 \qquad x = 122 \qquad \hat{p} = \dfrac{x}{n} = \dfrac{122}{165} = 0.739$

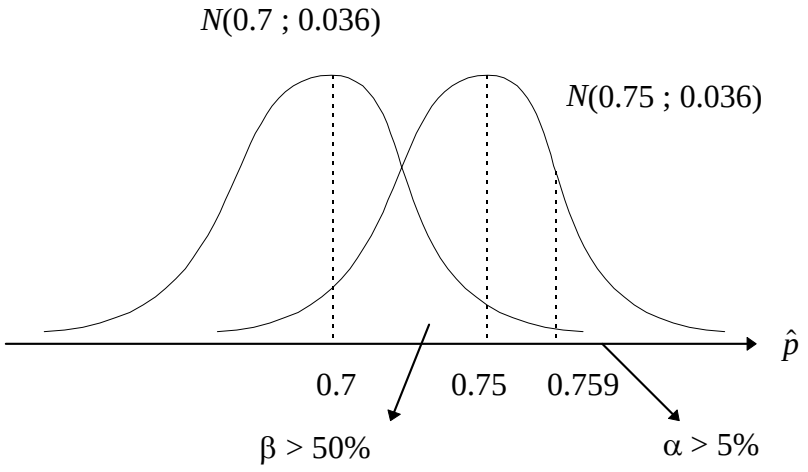
5º passo: como $\hat{p} = 0.739 \notin RC$, não rejeitamos H_0

"Não há evidência, ao nível de 5%, de que a proporção seja superior a 70%. O industrial não deve ficar satisfeito a esse nível."

"b) Por outro lado, o industrial considera um erro grave o chegar a se desiludir (no caso, admitir que não tem mais de 70% do mercado) quando, em verdade, ele tem mais de 75%. Ele gostaria que a probabilidade de cometer esse erro não superasse 10%. Pergunta-se se a amostra utilizada seria suficiente para atender a essa exigência e ao nível de significância adotado."

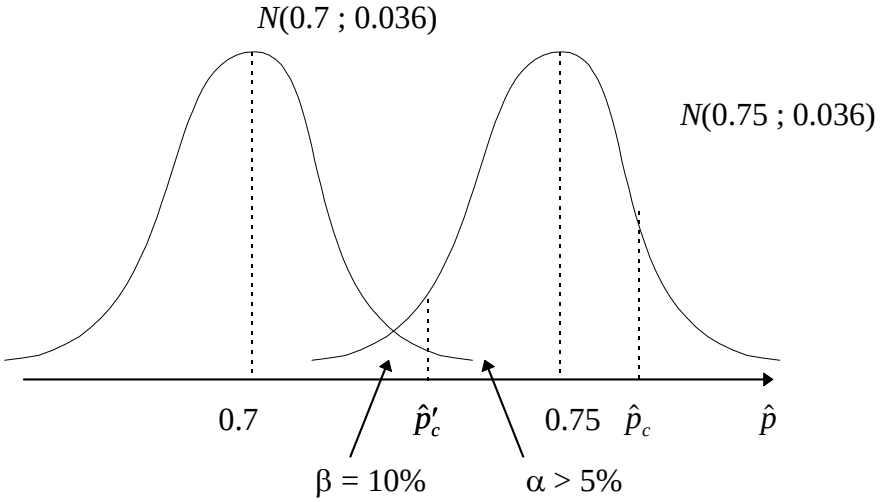
Solução:

Aqui, o problema consiste em avaliar o erro β para μ = 75%. Temos assim a seguinte situação:

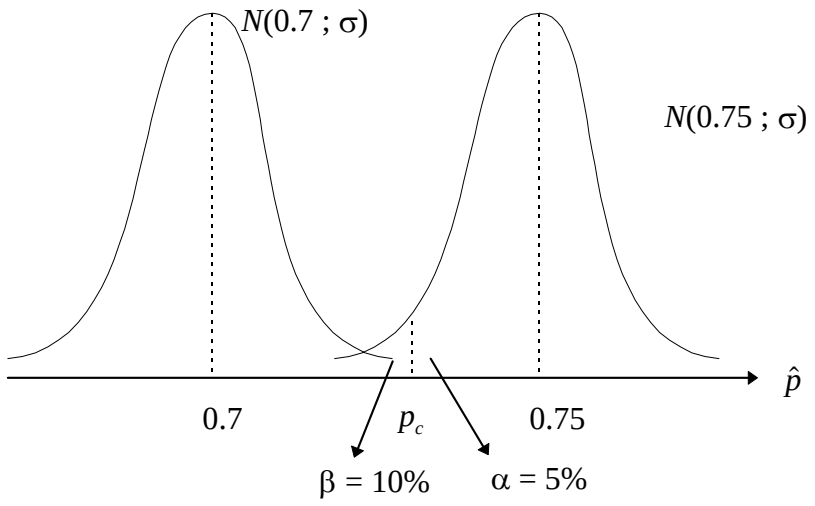


Observe que $\hat{p}_c = 0.759$ sendo maior que 0.75 implica em ocasionar um erro β superior a 50%.

Para diminuir β ao limite de 10% deveremos fazer com que $\hat{p}_c$ seja reduzido. Isso contudo provoca um aumento de α. Observe:



Para α = 5%, fixo, devemos atuar na variância, no intuito de conseguir β = 10%. Isso é alcançado aumentando-se o número de amostras *n* (e conseqüentemente teremos uma diminuição em $\sqrt{\frac{p.q}{n}}$). Veja isto na figura seguinte:



"Logo, a amostra utilizada não é suficiente para atender a exigência proposta pelo problema."

5.4.3 - Teste sobre a Média de uma População Normal com Variância Desconhecida

"**Ex.:** Sempre que o aumento da temperatura da água em uma câmara compressora superar 5°C, o processo de resfriamento deve ser recalibrado. Este processo é, entretanto caro e portanto deve ser feito apenas se necessário (adote um nível de 5%). Em oito experimentos independentes com a câmara, obtiveram-se os seguintes aumentos médios:

6.4 4.3 5.7 4.9 6.5 6.4 5.1 5.9

a) Estes dados sugerem a necessidade de recalibração?

b) Determine *IC* (μ : 95%)."

Solução:a)

1° passo: $H_o : \mu = 5°C$ (Não recalibração da câmara)

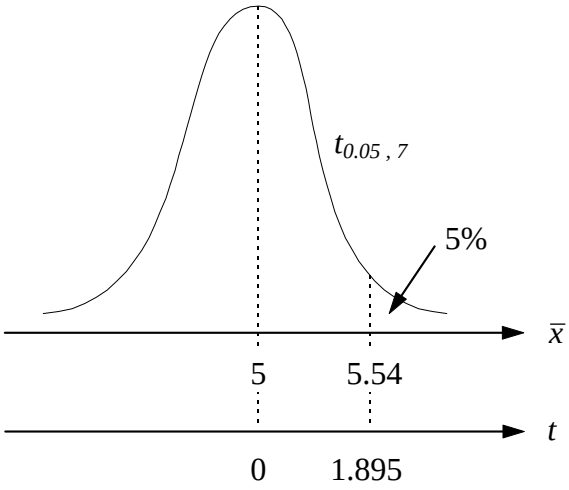
$H_1 : \mu > 5°C$ (Recalibração da câmara)

2° passo: Uso da estatística *t* de Student

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \quad \text{com} \quad \bar{x} = \frac{\sum x_i}{n}, \quad s = \sqrt{\frac{1}{n-1}\sum\left(x_i - \bar{x}\right)^2}$$

e n - 1 graus de liberdade

3º passo: $\alpha = 5\%$, $n = 8$, $v = n - 1 = 7 \Rightarrow t_{\alpha,v} = t_{0.05,7} = 1.895$ (tabela)



$\bar{x} = 5.65$

$s = 0.81$

$RC = \{y \in \mathrm{R} \mid y > 5.54\}$

4º passo: $\bar{x} = 5.65°\,\mathrm{C}$

5º passo: como $\bar{x} = 5.65 \quad \in \quad RC$, rejeitamos H_o

"Há evidências, ao nível de 5%, de que a câmara deva ser recalibrada."

b) $IC\,(\mu;\,95\%) \;=\; \left[\bar{x} - t_{\alpha/2}\;{}^{s}\!/\!\sqrt{n};\;\bar{x} + t_{\alpha/2}\;{}^{s}\!/\!\sqrt{n}\right]$

$$= \left[5.65 - (2.364).\frac{(0.81)}{\sqrt{8}};\, 5.65 + (2.364)\frac{(0.81)}{\sqrt{8}}\right]$$

$$= \left[4.97;\; 6.33\right]$$

5.4.4 - Teste sobre a Variância de uma População Normal

"**Ex.:** 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 10g e média de 500g. O peso de cada pacote X segue uma distribuição normal. Colhendo-se uma amostra de 16 pacotes obtivemos uma variância $s^2 = 169\ \mathrm{g}^2$.

Com esse resultado podemos dizer que a máquina está desregulada em relação à variância? (Considere $\alpha = 5\%$)."

Solução:

1º passo: $H_o : \sigma^2 = 100$ (A máquina está regulada)

$H_o : \sigma^2 > 100$ (A máquina está desregulada)

2º passo: Uso da estatística χ^2 (Qui Quadrado)

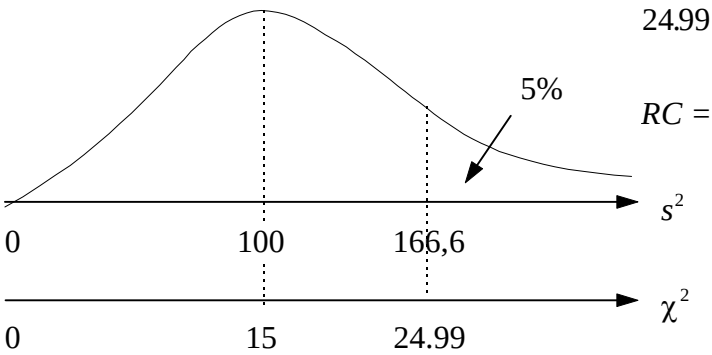
$$X^2 = \frac{(n-1)\,s^2}{\sigma^2} : \chi^2_{(v)} \quad \text{com } s^2 = \frac{1}{n-1}\sum\left(x_i - \bar{x}\right)^2 \quad \text{e}$$

$v = n - 1$ graus de liberdade.

3º passo: $\alpha = 5\%, n = 16, v = n - 1 = 15 \Rightarrow \chi^2_{0,05} = 24{,}99$ (tabela)

$$24.99 = \frac{(15)\,S_c^2}{100} \Rightarrow S_c = 166{,}6$$

$RC = \{y \in \mathrm{R} \mid y > 166{,}6\}$



4º passo: $s^2 = 169\ g^2$

5º passo: como $s^2 = 169 \in RC$, rejeitamos H_o

"Há evidência de que, ao nível de 5%, a máquina não esteja regulada."

Exercícios XVII

01 - Sabe-se que o consumo mensal *per capita* de um determinado produto tem distribuição normal, com desvio padrão de 2 kg. A diretoria de uma firma que fabrica esse produto resolveu que retiraria o produto da linha de produção se a média de consumo *per capita* fosse menor que 8 kg. Caso contrário, continuaria a fabricá-lo. Foi realizada uma pesquisa de mercado, tomando-se uma amostra de 25 indivíduos, e verificou-se que $\sum_{i=1}^{25} X_i = 180$ kg, onde X_i representa o consumo mensal do *i*-ésimo indivíduo da amostra.

a) Construa um teste de hipótese adequado, utilizando $\alpha = 0{,}05$, e com base na amostra colhida determine a decisão a ser tomada pela diretoria.

b) Qual a probabilidade β de se tomar uma decisão errada se, na realidade, a média populacional é $\mu = 7{,}8$ kg?

c) Se a diretoria tivesse fixado $\alpha = 0{,}01$, a decisão seria a mesma? (Justifique sua resposta).

d) Se o desvio padrão da população fosse 4 kg, qual seria a decisão? (Justifique sua resposta).

02 - A associação dos proprietários de indústrias metalúrgicas está muito preocupada com o tempo perdido com acidentes de trabalho, cuja média, nos últimos tempos, tem sido da ordem de 60 horas/homem por ano e desvio padrão de 20 horas/homem. Tentou-se um programa de prevenção de acidentes e, após o mesmo, tomou-se uma amostra de 9 indústrias e mediu-se o número de horas/homem perdidas por acidente, que foi de 50 horas. Você diria, ao nível de 5%, que há evidência de melhoria?

03 - O salário médio dos empregados das indústrias siderúrgicas é de 2,5 salários mínimos, com um desvio padrão de 0,5 salários mínimos. Se uma firma particular emprega 49 empregados com um salário médio de 2,3 salários mínimos, podemos afirmar que esta indústria paga salários inferiores, ao nível de 5%?

04 - Uma companhia de cigarros anuncia que o índice médio de nicotina dos cigarros que fabrica apresenta-se abaixo de 23 mg por cigarro. Um laboratório realiza 6 análises desse índice, obtendo: 27, 24, 21, 25, 26, 22. Sabe-se que o índice de nicotina se distribui normalmente, com variância igual a 4,86 mg^2. Pode-se aceitar, ao nível de 10%, a afirmação do fabricante?

05 - Uma pessoa gaba-se de adivinhar qual será o resultado do lance de uma moeda, mas é preciso que os presentes não o perturbem com pensamentos duvidosos. Para testar tal capacidade, lançou-se uma moeda perfeita 6 vezes, e o adivinhador acertou 5. Qual seria sua conclusão?

06 - O consumidor de um certo produto acusou o fabricante, dizendo que mais de 20% das unidades fabricadas apresentam defeito. Para confirmar sua acusação, ele usou uma amostra de tamanho 50, onde 27% das peças eram defeituosas. Mostre como o fabricante poderia refutar a acusação. Utilize um nível de significância de 10%.

07 - Um fabricante garante que 90% dos equipamentos que fornece a uma fábrica estão de acordo com as especificações exigidas. O exame de uma amostra de 200 peças desse equipamento revelou 25 defeituosas. Teste a afirmativa do fabricante, aos níveis de 5% e 1%.

08 - Os produtores de um programa de televisão pretendem modificá-lo se for assistido regularmente por menos de um quarto dos possuidores de televisão. Uma pesquisa encomendada a uma empresa especializada mostrou que, de 400 famílias entrevistadas, 80 assistem ao programa regularmente. Baseado nos dados, qual deve ser a decisão dos produtores?

09 - Suponha que queiramos testar $H_0 : \mu = 50$, contra $H_1 : \mu > 50$, onde μ é média de uma normal $N(\mu, 30)$. Extraída uma amostra de $n = 36$ elementos da população, obtemos $\bar{x} = 52$. Calcule o nível descritivo α do teste.

10 - Os novos operários de uma empresa são treinados a operarem uma máquina, cujo tempo X (em horas) de aprendizado é anotado. Observou-se que X segue de perto a distribuição $N(25,10)$. Uma nova técnica de ensino, que deve melhorar o tempo de aprendizado, foi testada em 16 novos empregados, os quais apresentaram 20,5 horas como tempo médio de aprendizado. Usando o nível descritivo, você diria que a nova técnica é melhor que a anterior?

11 - A precipitação pluviométrica anual numa certa região tem desvio padrão $\sigma = 3,1$ e média desconhecida. Para os últimos 9 anos, foram obtidos os seguintes resultados: 30,5; 34,1; 27,9; 35,0; 26,9; 30,2; 28,3; 31,7; 25,8.

a) Construa um teste de hipóteses para saber se a média da precipitação pluviométrica anual é maior que 30,0 unidades. Utilize um nível de significância de 5%.

b) Discuta o mesmo problema, considerando σ desconhecido.

c) Supondo que, na realidade, $\mu = 33{,}0$, qual a probabilidade de tirarmos uma conclusão errada?

12 - Supõe-se que determinado tipo de indústria deva ter, em média, 30 empregados. Para testar tal hipótese, colhe-se uma amostra de 50 indústrias, cujo resultado está abaixo. Caso rejeite a hipótese, dê um intervalo de confiança para a verdadeira média: (suponha que $s^2 = \sigma^2$)

Nº de Empregados	*Freqüência*
25 ├── 35	8
35 ├── 45	10
45 ├── 55	13
55 ├── 65	10
65 ├── 75	9

13 - Uma fábrica de automóveis anuncia que seus carros consomem, em média, 11 litros por 100 km, com desvio padrão de 0,8 litros. Uma revista resolve testar esta afirmação e analisa 35 automóveis dessa marca, obtendo 11,3 litros por 100 km como consumo médio (considerar distribuição normal). o que a revista pode concluir sobre o anúncio da fábrica, ao nível de 10%?

14 - Um dos maiores problemas de uma grande rede de vendas a varejo é verificar a adequação do estoque declarado com o real existente. Decidiu-se fazer a verificação através de procedimentos amostrais. Indicando por X o total em unidades monetárias de cada produto em estoque, verificou-se que $X : N(\mu, 20)$. Serão sorteados 4 produtos. O total X de cada um será verificado e calcular-se-á a média $\overline{X}$, que será a estatística de decisão. Numa determinada filial, o valor declarado de μ é 50. Havendo falta, este parâmetro deve ser 45; no caso de excesso, 58.

a) Defina H_0 e H_1.

b) Descreva os erros do tipo I e II.

c) Fixando $\alpha = 10\%$, qual a regra de decisão para julgar se o estoque está correto ou não?

d) Calcule o erro β.

e) Qual o significado de α e β neste problema?

15 - Duas máquinas, *A* e *B*, são usadas para empacotar pó de café. A experiência passada garante que o desvio padrão para ambas é de 10 g. Porém suspeita-se de que elas têm médias diferentes. Para verificar, sortearam-se duas amostras: uma com 25 pacotes da máquina *A* e outra com 16 pacotes da máquina *B*. As médias foram, respectivamente, $\bar{x}_A = 502{,}74$ e $\bar{x}_B = 496{,}60$ g. Com estes números, e ao nível de 5%, qual seria a conclusão do teste $H_0 : \mu_A = \mu_B$?

16 - Na região sul da cidade, 60 entre 400 pessoas preferem a bebida Meca-Mela entre as demais similares. Na região norte, a proporção é de 40 entre 225 entrevistados. Baseado no resultado dessa amostra, você diria que a proporção de todos os moradores nas duas regiões é a mesma?

17 - Uma pesquisa mercadológica sobre fidedignidade a um produto, foi realizada em dois anos consecutivos, com duas amostras independentes de 400 donas de casa em cada uma delas. A preferência pela marca em questão foi de 33% e 29%, respectivamente. Os resultados trazem alguma evidência de mudança de preferência?

18 - Seja X uma v.a. com distribuição binomial, com n = 15. Considere testar $H_0 : p \geq 0{,}5$ contra $H_1 : p < 0{,}5$, com $RC = \{0, 1, 2\}$.

a) Calcule a probabilidade do erro I.

b) Calcule a probabilidade do erro II quando $p = 0{,}3$.

c) Esboce o gráfico da CCO do teste.

19 - O tempo médio, por operário, para executar uma tarefa, tem sido 100 minutos, com um desvio padrão de 15 minutos. Introduziu-se uma modificação para diminuir esse tempo, e, após certo período, sorteou-se uma amostra de 16 operários, medindo-se o tempo de execução de cada um. O tempo médio da amostra foi 85 minutos, e o desvio padrão foi 12 minutos. Estes resultados trazem evidências estatísticas da melhora desejada? Em caso afirmativo, estime o novo tempo médio de execução. (Apresente as suposições teóricas usadas para resolver o problema)

20 - Estamos desconfiados de que a média das receitas municipais *per capita* das cidades pequenas (0 - 20.000 habitantes) é maior do que as receitas do estado, que é de 1.229 unidades.

Para comprovar ou não esta hipótese, sorteamos dez cidades pequenas, e obtivemos os seguintes resultados: 1.230; 582; 576; 2.093; 2.621; 1.045; 1.439; 717; 1.838; 1359.

Obs.: Para facilitar os cálculos, informamos que a soma das observações é 13.500, e a soma dos quadrados das observações é 22.335.650 (13.500^2 = 182.250.000).

a) Mostre que o teste de hipótese usado levará a aceitação de que a média das cidades pequenas é igual à do estado.

b) Você não acha estranha essa conclusão quando observa que a média da amostra obtida é bem maior do que a média do estado? Como você explicaria isso?

21 - Deseja-se estimar qual a porcentagem média da receita familiar gasta com alimentação pelos moradores de uma grande vila industrial. Para isso, selecionou-se uma amostra de 16 famílias, que apresentou os seguintes resultados:

41	44	35	42	34	22	42	42
38	62	29	63	38	45	48	40

a) Dê um *IC* de 95% para a porcentagem média de todas as famílias de moradores da vila.

b) Que suposição você faz para responder a pergunta anterior?

22 - Observou-se a produção mensal de uma indústria durante vários anos, verificando-se que ela obedecia a uma distribuição normal, com variância 300. Foi adotada uma nova técnica de produção e, durante 24 meses, observou-se a produção mensal. Após esse período, constatou-se que $\overline{X} = 10.000$ e s^2 = 400. Há razões para se acreditar que a variância mudou, ao nível de 20%?

23 - Numa linha de produção, é muito importante que o tempo gasto numa determinada operação não varie muito de empregado para empregado. Que parâmetro estatístico poderia ser usado para avaliar esse fato? Por quê? Se 11 empregados apresentam os tempos abaixo para realizar essa operação, qual seria a estimativa para o parâmetro acima?

125	135	115	120	150	130
125	145	125	140	130	

24 - Usando os dados do problema 23, verifique se também houve mudança de variância do processo antigo para o novo.

5.5 - Testes Paramétricos (Duas Amostras)

5.5.1 - Comparação das Variâncias de duas Populações Normais

"**Ex.:** A tabela seguinte sintetiza medidas de corrente elétrica (em mA) obtidas por dois engenheiros na análise de placas de circuito integrado.

I:	35	35	37	33	31	33
II:	32	34	34	31	32	

Há evidência para se afirmar que as medidas possuem a mesma homogeneidade?"

Solução:

1º passo: $H_0 = \sigma_I^2 = \sigma_{II}^2 = \sigma^2$ (Há uma mesma homogeneidade)

$H_1 = \sigma_I^2 \neq \sigma_{II}^2$ (Não há uma mesma homogeneidade)

2º passo: Uso da estatística F (Fisher/Snedecor- Ver)

3º passo: α = 10%,

$n_{\mathrm{I}} = 6$ ($v_{\mathrm{I}} = 5$),

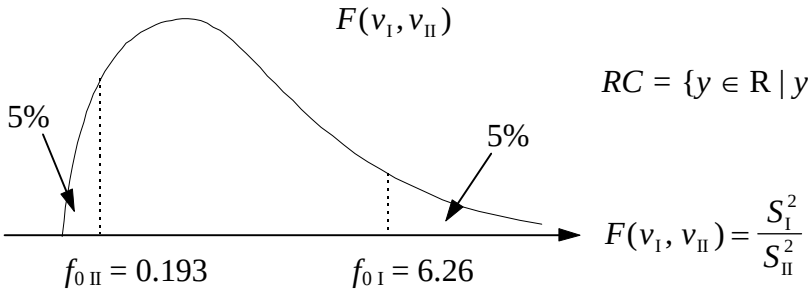
$n_{\mathrm{II}} = 5$ ($v_{\mathrm{II}} = 4$),

teste bilateral.

$$P\left[F(5,4) > f_{0_{\mathrm{I}}}\right] = 5\% \Rightarrow f_{0_{\mathrm{I}}} = 6.26$$

$$P\left[F(5,4) < f_{0_{\mathrm{II}}}\right] = 5\% \text{ , } P\left[\frac{1}{F(4,5)} < f_{0_{\mathrm{II}}}\right] = 5\% \text{ , } P\left[F(4,5) > \frac{1}{f_{0_{\mathrm{II}}}}\right] = 5\%$$

$$\frac{1}{f_{0_{\mathrm{II}}}} = 5.19 \Rightarrow f_{0_{\mathrm{II}}} = 0.193$$



$RC = \{y \in \mathrm{R} \mid y < 0.193 \text{ ou } y > 6.26\}$

4º passo: $S_{\mathrm{I}}^2 = 4.4$, $S_{\mathrm{II}}^2 = 1.8$ $\Rightarrow$ $F = \frac{S_{\mathrm{I}}^2}{S_{\mathrm{II}}^2} = 2.44$

5º passo: Como $F = 2.44 \notin RC$, não rejeitamos H_0

"Não há evidência ao nível de 10% de que as variâncias sejam diferentes, logo, há uma mesma homogeneidade."

5.5.2 - Comparação de Duas Médias de Populações Normais

Obs.: Seja $X:N(\mu_x,\sigma_x)$ e $Y:N(\mu_y,\sigma_y)$

Temos assim os seguintes estimadores:

$$\overline{X} = \frac{1}{n}\sum X_i \text{ e } S_x^2 = \frac{1}{n-1}\sum (X_i - \overline{X})^2,$$

$$\overline{Y} = \frac{1}{m}\sum X_i \text{ e } S_y^2 = \frac{1}{m-1}\sum (Y_i - \overline{Y})^2.$$

Definindo $\overline{D} = \overline{X} - \overline{Y}$, verifica-se que $\overline{D}:N\left(\mu_{\overline{D}},\sigma_{\overline{D}}\right)$

com $\mu_{\overline{D}} = \mu_x - \mu_y$ e $\sigma_{\overline{D}}^2 = \dfrac{\sigma_x^2}{n} + \dfrac{\sigma_y^2}{m}$.

Conseqüentemente:

$$Z = \frac{\overline{D} - \mu_{\overline{D}}}{\sigma_{\overline{D}}} = \frac{(\overline{X} - \overline{Y}) - (\mu_x - \mu_y)}{\sqrt{\dfrac{\sigma_x^2}{n} + \dfrac{\sigma_y^2}{m}}} \cong N(0;1)$$

Caso 1) Variâncias iguais Desconhecidas

"**Ex.:** No exemplo anterior, nas medidas de corrente elétrica dos dois engenheiros, tínhamos:

I:	35	35	37	33	31	33
II:	32	34	34	31	32	

e o teste revelou que as medidas possuem a mesma homogeneidade, ou seja, $\sigma_I^2 = \sigma_{II}^2 = \sigma^2$. Uma outra questão: Há evidência para se afirmar, ao nível de 5%, que a média é maior no conjunto de medidas do engenheiro I?"

Solução:

1º passo: $H_0:\mu_I = \mu_{II}$ $\quad H_1:\mu_I > \mu_{II}$ ou $H_0:\mu_D = 0$ $\quad H_1:\mu_D > 0$ com $\mu_{\overline{D}} = \mu_I - \mu_{II}$

2º passo: Uso da estatística

$$Z = \frac{\overline{D} - \mu_{\overline{D}}}{\sigma_{\overline{D}}} = \frac{(\overline{X}_{\mathrm{I}} - \overline{X}_{\mathrm{II}}) - (\mu_{\mathrm{I}} - \mu_{\mathrm{II}})}{\sqrt{\frac{\sigma_{\mathrm{I}}^2}{n} + \frac{\sigma_{\mathrm{II}}^2}{m}}}$$

Fazendo $\sigma_{\mathrm{I}}^2 = \sigma_2^2 = \sigma^2 = S_p^2 = \dfrac{(n-1)S_{\mathrm{I}}^2 + (m-1)S_{\mathrm{II}}^2}{n+m-2}$

temos, então, usando a estimativa de Student, que:

$$t = \frac{\left(\overline{X}_{\mathrm{I}} - \overline{X}_{\mathrm{II}}\right) - \left(\mu_{\mathrm{I}} - \mu_{\mathrm{II}}\right)}{S_p\sqrt{\frac{1}{n} + \frac{1}{m}}} \cong t_{\alpha, n+m-2}$$

3º passo: $\alpha = 5\%$, $n = 6$, $m = 5 \Rightarrow t_{5\%,\,9} = 1.833$

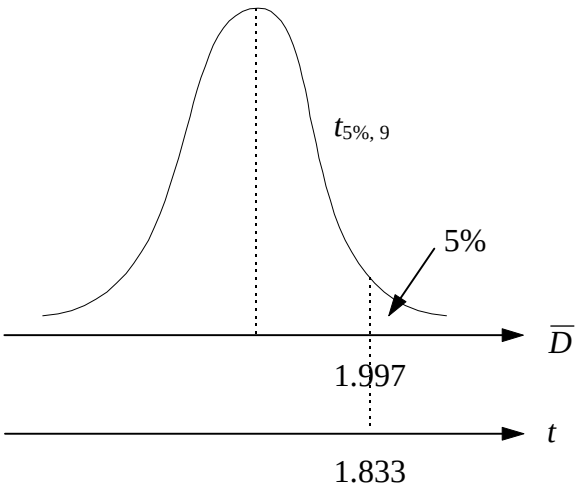
$\overline{X}_{\mathrm{I}} = 34 \qquad S_{\mathrm{I}}^2 = 4.4$

$\overline{X}_{\mathrm{II}} = 32.6 \qquad S_{\mathrm{II}}^2 = 1.8$

$$S_p^2 = \frac{5(4.4) + 4(1.8)}{9} = 3.24$$

$$S_p = 1.8$$

$$1.833 = \frac{\overline{D}_C - 0}{1.8\sqrt{\frac{1}{6} + \frac{1}{5}}} \Rightarrow \overline{D}_C = 1.997$$



$RC = \{y \in \mathrm{R} \mid y > 1.997\}$

4º passo: Das amostras,

$$\overline{D} = \overline{X}_{\mathrm{I}} - \overline{X}_{\mathrm{II}} = 34 - 32.6 = 1.4$$

5º passo: Como $\overline{D} = 1.4 \notin RC$, não rejeitamos H_o

"Não há evidência, ao nível de 5% de que as médias sejam diferentes."

Caso 2) Variâncias Desiguais Desconhecidas

"**Ex.:** Em uma experiência industrial, foi executado um trabalho por 10 operários, de acordo com o método I, e por 20 operários, de acordo com o método II. Os resultados levaram aos seguintes dados sobre a duração média e variabilidade do tempo, necessários à execução do trabalho:

Método I: $\bar{x}_I = 53$ min $\quad S_I = 6$ min
Método II: $\bar{x}_{II} = 57$ min $\quad S_{II} = 15$ min

Os dados permitem afirmar, ao nível de 5%, que o método I fornece um tempo médio menor que o método II?"

Solução:

De acordo com o teste de duas Variâncias (Fisher/Snedecor)

$H_0: \sigma_I^2 = \sigma_{II}^2 = \sigma^2$

$H_1: \sigma_I^2 \neq \sigma_{II}^2$

concluimos que H_1 é verdadeira .

A partir desse resultado podemos efetuar o seguinte teste das médias:

1º passo: $H_0: \mu_I = \mu_{II}$ $\qquad$ $H_0: \mu_D = 0$

$H_1: \mu_I < \mu_{II}$ $\qquad$ ou $\qquad$ $H_1: \mu_D < 0$ $\qquad$ $\mu_D = \mu_I - \mu_{II}$

2º passo: Uso da estatística

$$Z = \frac{\overline{D} - \mu_D}{\sigma_{\overline{D}}} = \frac{(\overline{X}_{\mathrm{I}} - \overline{Y}_{\mathrm{II}}) - (\mu_{\mathrm{I}} - \mu_{\mathrm{II}})}{\sqrt{\frac{\sigma_{\mathrm{I}}^2}{n} + \frac{\sigma_{\mathrm{II}}^2}{m}}}$$

Fazendo $\sigma_{\mathrm{I}}^2 \simeq S_{\mathrm{I}}^2$

$\sigma_{\mathrm{II}}^2 \simeq S_{\mathrm{II}}^2$

temos $t = \dfrac{\left(\overline{X}_{\mathrm{I}} - \overline{X}_{\mathrm{II}}\right) - \left(\mu_{\mathrm{I}} - \mu_{\mathrm{II}}\right)}{\sqrt{\frac{S_{\mathrm{I}}^2}{n} + \frac{S_{\mathrm{II}}^2}{m}}} \simeq t_{\alpha;v}$

com $v = \dfrac{\left(\frac{S_{\mathrm{I}}^2}{n} + \frac{S_{\mathrm{II}}^2}{m}\right)^2}{\frac{\left(\frac{S_{\mathrm{I}}^2}{n}\right)^2}{n-1} + \frac{\left(\frac{S_{\mathrm{II}}^2}{m}\right)^2}{m-1}} - 2$

3º passo: $\alpha = 5\%$, $v = 25{,}8 \simeq 26$ graus de liberdade.

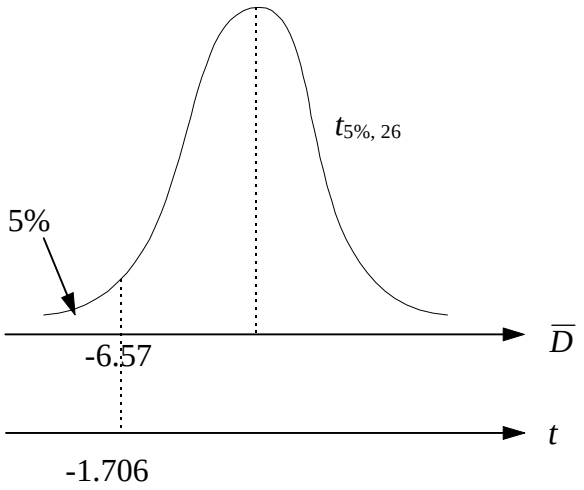
Logo, $t_{5\%\,,\,26} = -1.706$

$S_{\mathrm{I}}^2 = 36$

$S_{\mathrm{II}}^2 = 225$

$$-1.706 = \frac{\overline{D}_C - 0}{\sqrt{\frac{36}{10} + \frac{225}{20}}}$$

$\overline{D}_C = -6.57$



$RC = \{y \in \mathrm{R} \mid y < -6.57\}$

4º passo: $\overline{D} = \overline{X}_{\mathrm{I}} - \overline{X}_{\mathrm{II}} = -4$

5º passo: Como $\overline{D} \notin RC$, não rejeitamos H_o

"Não há evidência, ao nível de 5%, de que a média do método I seja menor do que a média do método II."

<u>Caso 3) Observações Emparelhadas</u>

"**Ex.:** Uma empresa deseja estudar o efeito de uma pausa de 10 minutos para um cafezinho sobre a produtividade de seus trabalhadores. Para isso, sorteou 6 operários e contou o número de peças produzidas durante uma semana sem intervalo e uma semana com intervalo.

Operário	1	2	3	4	5	6
Sem Intervalo	23	35	29	33	43	32
Com Intervalo	28	38	29	37	42	30

Os resultados sugerem melhora de produtividade em um dos casos? (α= 5%)"

<u>Solução</u>:

1º passo: $H_0: \mu_s = \mu_c$ $\qquad$ $H_0: \mu_D = 0$

$H_1: \mu_s < \mu_c$ $\qquad$ ou $\qquad$ $H_1: \mu_D < 0$ $\qquad$ $\mu_D = \mu_s - \mu_c$

2º passo: Uso da estatística $t = \dfrac{\overline{D} - \mu_D}{S_D / \sqrt{n}}$ pois

$D = X - Y$ é uma *v.a.* com σ_D desconhecido.

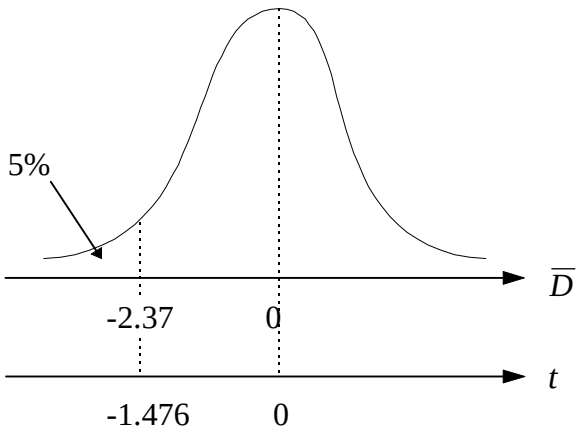
Obs.: $\begin{cases} X_1, X_2, \cdots X_n \\ Y_1, Y_2, \cdots Y_n \end{cases} \Rightarrow X_1 - Y_1, X_2 - Y_2, \cdots X_n - Y_n = D_1, D_2, \cdots D_n$

3º passo: $\alpha = 5\%$ $n = 6$ $v = n - 1 = 5 \Rightarrow t_{5\%,5} = -2.015$

$\overline{D} = \overline{X}_S - \overline{X}_C = 32.5 - 34 = -1.5$

$S_D = 2.88$

$-2.015 = \dfrac{\overline{D}_C - 0}{2.88/\sqrt{6}}, \overline{D}_C = -2.37$



$RC = \{y \in \mathrm{R} \mid y < -2.37\}$

4º passo: Das amostras, $\overline{D} = -1.5$

5º passo: Como $\overline{D} \notin RC$, não rejeitamos H_o

"Não há melhora de produtividade, ao nível de 5%, com o intervalo do cafezinho."

Exercícios XVIII

01 - Uma das maneiras de medir o grau de satisfação dos empregados de uma mesma categoria quanto à política salarial é através do desvio padrão de seus salários. A fábrica *A* diz ser mais coerente na política salarial do que a fábrica *B*. Para verificar essa afirmação, sorteou-se uma amostra de 10 funcionários não especializados de *A*, e 15 de *B*, obtendo-se os desvios padrões $s_A = 1{,}0\ SM$ e $s_B = 1{,}6\ SM$. Qual seria a sua conclusão?

02 - Deseja-se comparar a qualidade de um produto produzido por duas fábricas. esta qualidade será definida pela uniformidade com que é produzido o produto por cada fábrica. Tomaram-se duas amostras, uma de cada fábrica, medindo-se o comprimento dos produtos (o resumo dos resultados está no quadro abaixo). A qualidade das duas fábricas é a mesma? Caso sua resposta seja negativa, dê um intervalo de confiança para indicar a intensidade dessa desigualdade.

Estatísticas	*Fábrica A*	*Fábrica B*
Amostra	21	17
Média	21,15	21,12
Variância	0,0412	0,1734

03 - Num estudo comparativo do tempo médio de adaptação, numa amostra aleatória, de 50 homens e 50 mulheres de um grande complexo industrial, produziu os seguintes resultados:

Estatísticas	*Homens*	*Mulheres*
Médias	3,2 anos	3,7 anos
Desvios padrões	0,8 anos	0,9 anos

Que conclusões você poderia tirar para a população de homens e mulheres desta indústria? (Indique as suposições feitas para resolver o problema)

04 - Diversas políticas em relação às filiais de uma rede de supermercados estão associadas ao gasto médio dos clientes em cada compra. Deseja-se comparar esse parâmetro para duas novas filiais, através de duas amostras de 50 clientes cada. As médias obtidas foram 62 e 71, respectivamente. Sabe-se que o desvio padrão, em ambos os casos, deve ser da ordem de 20 unidades. É possível afirmar que o gasto médio das duas filiais é o mesmo? Caso não seja, dê um intervalo de confiança para a diferença.

05 - 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 resultados 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

06 - No problema 02, teste a hipótese de que as médias dos comprimentos do produto produzido pelas duas fábricas são iguais.

07 - Para investigar 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

08 - Para verificar a importância de um determinado cartaz nas compras de certos produto, procedeu-se do seguinte modo:

a) formaram-se 7 pares de lojas;

b) os pares foram formados de modo que tivessem as mesmas características quanto à localização, ao tamanho e ao volume de vendas;

c) num dos elementos do par, colocou-se o cartaz; no outro não;

d) as vendas semanais foram registradas, e os resultados estão abaixo

Qual seria sua conclusão sobre a eficiência do cartaz?

	par	1	2	3	4	5	6	7
vendas	*sem cartaz*	13	18	14	16	19	12	22
	com cartaz	16	24	18	14	26	17	29

5.6 -Testes Não Paramétricos

5.6.1 - Teste de Independência

"**Ex.:** A tabela seguinte representa a distribuição conjunta das proporções em relação aos totais de cada linha, associando tipo de cooperativa ao fator regional (veja Capítulo 1, Coeficiente de Contingência).

Estado	Tipo de Cooperativa				Total
	Consumidor	Produtor	Escola	Outros	
São Paulo	214 (33%)	237 (37%)	78 (12%)	119 (18%)	648 (100%)
Paraná	51 (17%)	102 (34%)	126 (42%)	22 (7%)	301 (100%)
Rio G. do Sul	111 (18%)	304 (51%)	139 (23%)	48 (8%)	602 (100%)
Total	376 (24%)	643 (42%)	343 (22%)	189 (12%)	1551 (100%)

Os dados evidenciam dependência entre os fatores: tipo de cooperativa e região de localização? (use α = 5%)"

Solução:

1º passo: $H_0: p_{ij} = (p_{i\bullet})(p_{\bullet j})$ (valor observado = valor esperado $\Rightarrow$ independência)

$H_1: p_{ij} \neq (p_{i\bullet})(p_{\bullet j})$ (valor observado≠valor esperado $\Rightarrow$ dependência)

Obs.: Recorde-se que:

P_{ij} é a probabilidade de um indivíduo pertencer à casa (i, j),

$P_{i\bullet}$ e $P_{\bullet j}$ são as probabilidades marginais. Assim, temos:

P_{ij} = 33%,17%,...,8%

$P_{i\bullet}$ = 24%, 42%, 22% e 12%

$P_{\bullet j}$ = 100%

2º passo: Uso da estatística $\chi^2_{\alpha,v}$ com $X^2 = \sum_{i=1}^{m}\sum_{j=1}^{n}\frac{\left(o_{ij} - e_{ij}\right)^2}{e_{ij}} : \chi^2_{\alpha,v}$ sendo:

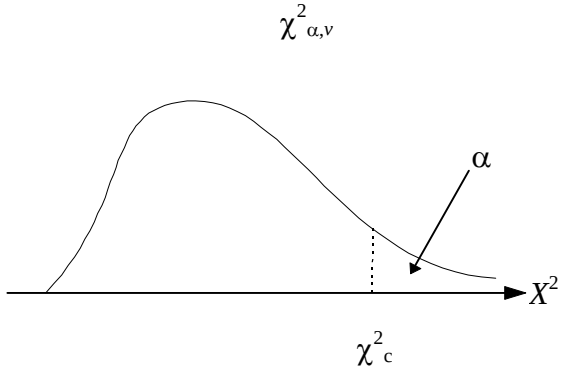
m = número de linhas

n = número de colunas

$v = (m - 1)(n - 1)$

$o_{ij} \equiv$ valor observado

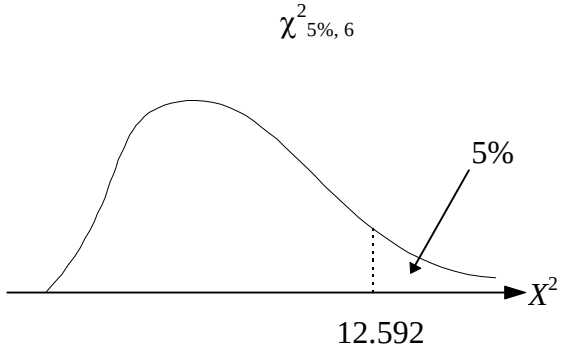
$e_{ij} \equiv$ valor esperado



Obs.: O teste aqui será sempre unilateral, a despeito do sinal "≠" em H_0 (Na verdade está se testando se $X^2 = 0$ ou $X^2 > 0$)

3º passo: $\alpha = 5\%$ $m = 3$, $n = 4$, $v = 6 \Rightarrow \chi^2_{5,6} = 12.592$

Logo, $RC = \{y \in \mathrm{R} \mid y > 12.592\}$



4º passo: Das amostras,

$X^2 = 21.56 + 6.12 + \cdots + 8.00 = 173.24$ (Cap.I-Análise Bidimensional)

5º passo: Como $X^2 \in RC$, rejeitamos H_o

"Os dados evidenciam uma forte dependência entre os fatores, ao nível de 5%."

Obs.: X^2 é uma distribuição aproximada e os valores de e_{ij} não devem ser inferiores a 5 na aplicação desse teste.

5.6.2 - Teste sobre Coeficiente de Correlação

Caso 1) ρ=0

"**Ex.:** Verificar se podemos, ao nível de 5% de significância, concluir pela existência de correlação positiva entre a tensão na saída de um determinado gerador (em V) e a rotação de um mecanismo acoplado ao seu eixo (em Hz).

Tensão:	174	161	170	180	182	164	156	168	176	175
Rotação:	73	66	64	94	79	72	62	64	90	81

Solução:

1º passo: $H_0: \rho(X,Y) = 0$ (Não existe correlação positiva))

$H_1: \rho(X,Y) > 0$ (Existe correlação positiva))

onde: $X \equiv$ Tensão

$Y \equiv$ Rotação

2º passo: Este teste pode ser feito através da quantidade

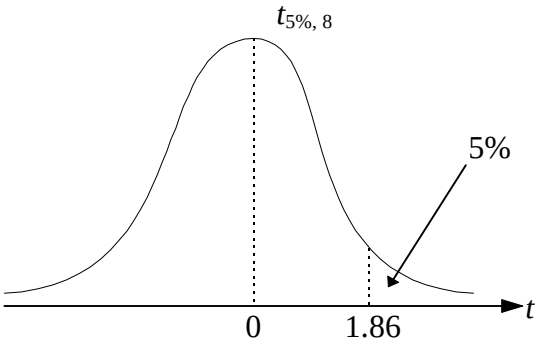
$$T = r\sqrt{\frac{n-2}{1-r^2}} : t_{\alpha,v}$$

onde: r = corr(X,Y) e

$n \equiv$ número de amostras

associada à distribuição de Student com $v = n - 2$ graus de liberdade.

3º passo: $\alpha = 5\%$, $n = 10$, $v = 8$, $t_{5\%,8} = 1.86$



$RC = \{y \in R \mid y > 1.86\}$

4º passo: Das amostras

$$r = \frac{1}{n}\sum\left(\frac{X_i - \overline{X}}{S_x} \cdot \frac{Y_i - \overline{Y}}{S_y}\right) \simeq 0.772$$

$$T = 0.772\sqrt{\frac{8}{1-(0.772)^2}} = 3.435$$

5º passo: Como $T \in RC$, rejeitamos H_o

"Há evidência, ao nível de 5%, de correlação positiva entre a tensão do gerador e a rotação do mecanismo."

Caso 2) $\rho = \rho_0$

"**Ex.:** Durante muito tempo, o coeficiente de correlação entre a nota final num curso de treinamento de operários e sua produtividade, após 6 meses do curso, foi de 0,50. Foram introduzidas modificações no curso, com o intuito de aumentar a correlação. Se o coeficiente de correlação de uma amostra de 28 operários submetidos ao novo curso foi de 0,65, você diria que os objetivos da modificação foram atingidos?"

Solução:

1º passo: $H_0: \rho(X, Y) = 0.50$ (Objetivos não atingidos)

$H_1: \rho(X, Y) > 0.50$ (Objetivos atingidos)

onde: $X \equiv$ Resultado do teste

$Y \equiv$ Produtividade

2º passo: Fisher, estudando a distribuição amostral de *r*, descobriu que a estatística

$$Z = \frac{1}{2}\ln\frac{1+r}{1-r}$$

tem distribuição muito próxima da Normal $N(\mu_z, \sigma_z)$ com

$$\mu_Z = \frac{1}{2}\ln\frac{1+\rho_0}{1-\rho_0}, \quad \sigma_Z^2 = \frac{1}{n-3} \quad \text{e} \quad \rho_0 \equiv \text{ parâmetro populacional.}$$

Obs.: $\rho \neq \pm 1$ e $\rho \neq 0$

3º passo: $\alpha = 5\%$

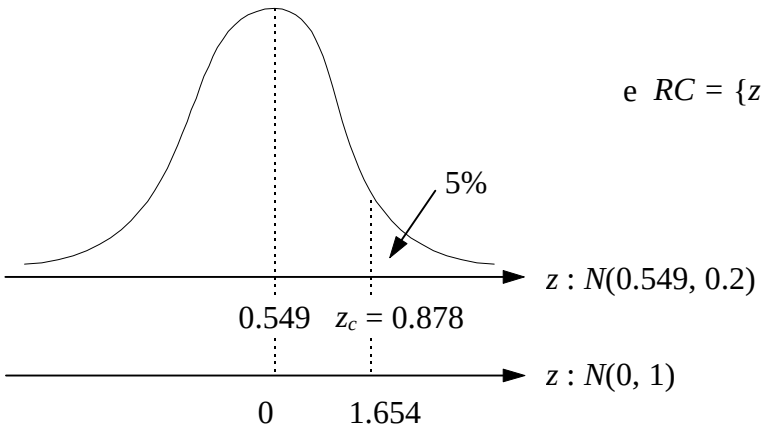
$$\mu_Z = \frac{1}{2}\ln\frac{1+0.5}{1-0.5} = 0.549$$

$$\sigma_Z^2 = \frac{1}{28-3} = 0.04 \quad \Rightarrow \quad \sigma_Z = 0.2$$

logo: $Z: N(0.549; 0.2)$

Relacionando à normal reduzida $N(0,1)$ com $\alpha = 5\%$, temos que

$$Z = \frac{Z_C - \mu}{\sigma}, \quad 1.654 = \frac{Z_C - 0.549}{0.2}, \quad Z_C = 0.878$$



e $RC = \{z \in \mathrm{R} \mid z > 0.878\}$

4º passo: Das amostras, com $r = 0.65$

$$Z_0 = \frac{1}{2}\ln\frac{1+0.65}{1-0.65} = 0.774$$

5º passo: Como $z_0 \notin RC$, não rejeitamos H_o

"Não há evidência, ao nível de 5%, de que tenha aumentado o coeficiente de correlação."

Exercícios XIX

01 - Investigando a “fidelidade” de consumidores de um produto, obteve-se uma amostra de 200 homens e 200 mulheres. Foram classificados como tendo alto grau de fidelidade 100 homens e 120 mulheres. Os dados fornecem evidência de possíveis diferenças de grau de fidelidade entre sexos?

02 - Uma pesquisa sobre a qualidade de certo produto foi realizada enviando-se questionários a donas-de-casa através do correio. Aventando-se a possibilidade de que os respondentes voluntários tenham um particular vício de respostas, fizeram-se mais duas tentativas com os não-respondentes. Os resultados estão indicados abaixo. Você acha que existe relação entre a resposta e o número de tentativas?

Opinião sobre o produto	*Nº de donas-de-casa*		
	1ª tentativa	*2ª tentativa*	*3ª tentativa*
Excelente	62	36	12
Satisfatório	84	42	14
Insatisfatório	24	22	24

03 - Estamos estudando se há ou não correlação entre as notas de diversas disciplinas de um curso de mestrado. Analisando uma amostra de 12 alunos, encontrou-se uma correlação de 0,60 entre as disciplinas de Estatística e Metodologia da Pesquisa. Teste a hipótese de não haver correlação entre as disciplinas. Caso a rejeite, dê um intervalo de confiança para o coeficiente de correlação populacional.

04 - Existe relação entre o volume de uma carga e o tempo gasto para acondicioná-la? Para investigar este fato, sortearam-se 9 pedidos de mercadorias, medindo-se as duas variáveis de interesse. Com os dados obtidos abaixo, quais seriam suas conclusões?

Tempo	84	108	110	133	144	152	180	196	231
Volume	48	72	63	82	88	109	112	123	140

CAPÍTULO 6 - AMOSTRAGEM

O texto que se segue é, a menos de pequenas adaptações, tradução de um texto clássico de William G.Cochran, estatístico com contribuições fundamentais à Teoria da Amostragem.

6.1- Os Principais Estágios de uma Pesquisa por Amostragem

Como preliminar para uma discussão do papel que a teoria desempenha em uma pesquisa por amostragem, é interessante descrever sucintamente os estágios envolvidos no planejamento e execução de uma pesquisa. As pesquisas variam grandemente em sua complexidade. Extrair uma amostra de um conjunto de 5.000 fichas numeradas e ordenadas em um arquivo é uma tarefa simples. Muito diferente é a extração de uma amostra dos habitantes da região amazônica, onde o transporte principal é fluvial e através de florestas e os habitantes olham com desconfiança qualquer estranho que comece a fazer perguntas. Problemas sérios em uma pesquisa podem ser triviais, ou mesmo não existir, em outras.

Os principais estágios de uma pesquisa acham-se grupados - de forma um tanto arbitrária - em 11 tópicos.

1º. Objetivos da pesquisa

Os objetivos da pesquisa devem ser claramente definidos. Sem isto, é fácil, numa pesquisa complexa, esquecer os objetivos, perdidos nos detalhes do planejamento, e tomar decisões em desacordo com os objetivos.

2º. População a ser submetida a amostragem

A palavra população denota o conjunto do qual se extrai a amostra. A definição da população pode não apresentar problemas, como no caso de um lote de lâmpadas, para determinar a vida média de uma lâmpada. Por outro lado, ao extrairmos uma amostra de um conjunto de fazendas, devemos definir claramente o que entendemos por fazenda, e atentar para os casos extremos. Essas regras de definição devem ser usadas na prática: o pesquisador deve ser capaz de decidir, no campo, sem muita hesitação, se determinado caso pertence ou não à população.

A população a ser submetida a amostragem deve coincidir com a população a cujo respeito desejamos informação (população alvo). Às vezes, por motivos práticos ou de conveniência, a população sob amostragem é mais restrita do que a população alvo. Em tais circunstâncias, cabe observar que as conclusões extraídas da amostra se aplicam à população sob amostragem. Para julgar se tais conclusões podem ser estendidas à população alvo, são necessárias outras fontes de informação. Serão úteis quaisquer informações que se possam obter sobre as diferenças entre a população submetida a amostragem e a população alvo.

3º. Dados a serem coletados

Convém verificar que todos os dados coletados sejam relevantes para a finalidade da pesquisa e que nenhum dado essencial seja omitido. Há uma tendência freqüente, sobretudo com populações humanas, de formular um número demasiadamente grande de questões, algumas das quais não chegam sequer a ser analisadas subseqüentemente. Um questionário muito extenso reduz a qualidade das respostas, tanto a questões importantes como a questões secundárias.

4º. Grau de precisão desejado

Os resultados de pesquisas por amostragem estão sempre sujeitos a certo grau de inconfiança, não só porque se estudou apenas uma parte da população, mas também em decorrência de erros de mensuração. Esta inconfiança pode ser reduzida aumentando-se o tamanho da amostra e utilizando-se instrumentos mais precisos de mensuração. Mas isto em geral exige tempo e custa dinheiro. Conseqüentemente, é importante especificar o grau de precisão exigido nos resultados. Isto é de responsabilidade da pessoa que vai usar os dados, e pode apresentar dificuldades, pois muitos administradores não estão acostumados a pensar em termos do valor do erro que pode ser tolerado em estimativas, consistente com a tomada de uma boa decisão. O estatístico pode prestar bom auxílio nesta fase.

5º. Métodos de mensuração

Há mais de uma escolha do aparato de mensuração e da forma de abordar a população. Dados sobre o estado de saúde da pessoa podem ser obtidos da declaração pessoal ou de relatório do exame médico. A pesquisa pode utilizar um questionário auto administrado, um entrevistador que lê um conjunto de questões sem qualquer discrição, ou um processo de entrevista que deixa margem ampla quanto à forma e ordenação das questões. A abordagem pode ser postal, telefônica, pessoal ou uma combinação das três.

Parte importante do trabalho preliminar é a construção de formulários onde se registrarão as perguntas e as respostas. No caso de questionários simples, as respostas podem às vezes ser pré-codificadas - isto é, registradas de maneira que possam ser transferidas sem dificuldade para um arquivo de computador. De fato, para construir arquivos de dados adequados, é necessário visualizar a estrutura das tabelas finais de resumo, que serão usadas para tirar conclusões.

6º. Sistemas de referência

Antes de selecionar a amostra, a população deve ser dividida em partes, chamadas unidades de amostragem, ou simplesmente unidades. Estas unidades devem abranger toda a população, sem qualquer superposição. Isto é: cada elemento da população deve pertencer a uma, e somente uma, unidade. Às vezes, a unidade apropriada é óbvia, como no caso de uma população de lâmpadas. Outras vezes, pode haver escolha da unidade. Ao extrair uma amostra da população de uma cidade, a unidade pode ser um indivíduo, os membros de uma família, ou o conjunto de todas as pessoas que moram em um quarteirão. Na amostragem de uma safra agrícola, a unidade pode ser um campo, uma fazenda ou uma área de terra de formato e dimensões à nossa escolha.A construção desta lista de unidades de amostragem, chamada Sistema de Referência, é, em geral, um dos principais problemas práticos. Com base na experiência, os pesquisadores adotam uma atitude critica em relação a listas que tenham sido compostas rotineiramente para algum propósito. A despeito de afirmações em contrário, tais listas se revelam freqüentemente incompletas, ou parcialmente ilegíveis, ou então apresentam duplicações. Uma boa estrutura pode ser difícil de obter quando a população é especializada.

7º. Escolha da amostra

Há uma variedade de planos para a escolha da amostra. Para cada plano considerado, podem-se fazer estimativas do tamanho da amostra com base no conhecimento do grau de precisão desejado. Devem-se comparar os custos relativos e o tempo envolvidos em cada plano antes de se tomar uma decisão.

8º. O pré-teste

É útil testar o questionário e o método de trabalho no campo de atividade, em uma escala reduzida. Isto quase sempre resulta em melhoria do questionário e pode revelar outros problemas que seriam sérios em uma escala maior, como, por exemplo, o fato do custo poder ser muito maior do que o esperado.

9º. Organização do trabalho de campo

Em pesquisas extensas defrontamo-nos com muitos problemas de caráter administrativo. O pessoal deve receber treinamento nos objetivos da pesquisa e nos métodos de mensuração a serem empregados, e seu trabalho deve ser supervisionado adequadamente. É de grande valia um processo para verificação preliminar da qualidade dos resultados. Os planos devem prever a não-resposta, isto é, a falha do entrevistador em obter informações de determinadas unidades da amostra.

10º. Resumo e análise dos dados

O primeiro passo é verificar o questionário completo, visando a corrigir erros ou, ao menos, eliminar dados obviamente errôneos. Tornam-se necessárias decisões sobre o processo de cálculo nos casos de omissão de respostas a certas questões, por parte dos entrevistados, ou de eliminação no processo de edição do questionário. Após isto, fazem-se os cálculos que conduzem às estimativas. Pode haver diferentes métodos de estimação para os mesmos dados.

11º. Informação ganha para futuras pesquisas

Quanto mais informações tivermos inicialmente sobre uma população, mais fácil se torna planejar uma amostra que dê estimativas precisas. Qualquer amostragem completada é potencialmente um guia para futuras amostragens melhoradas nos dados que ela fornece sobre as médias, os desvios padrão e a natureza da variabilidade das mensurações principais e o custo da obtenção dos dados. A prática da amostragem avança mais rapidamente quando se fazem provisões para reunir e registrar informações deste tipo.

Há outro aspecto importante em que qualquer amostra completada facilita a obtenção de futuras amostras. As coisas nunca se comportam tal como foram planejadas em uma pesquisa complexa. O pesquisador arguto aprende a reconhecer erros de execução e fazer com que eles não ocorram em pesquisas futuras.

6.2 - Questionários

1º. Tipos de questões

Dois tipos de questões são usualmente empregados na redação de questionários: questões de múltipla escolha e questões de resposta aberta.

As alternativas em uma questão de múltipla escolha devem ser claras, mutuamente excludentes e quando pedirem opiniões, fornecer opções dos dois lados do assunto. Idealmente, as opções devem cobrir todas as respostas prováveis. Se, entretanto, muitas alternativas são apresentadas, elas podem não ser suficientemente claras e confundir o respondente no momento de sua decisão. A grande desvantagem de questões de múltipla escolha é que tendem a sugerir uma resposta, já que limita as respostas possíveis, impedindo o respondente de dizer exatamente o que pensa.

Este tipo de limitação não ocorre nas questões de resposta aberta, onde o entrevistado usa suas próprias palavras para responder à pergunta. Uma pergunta deste tipo produz uma grande gama de respostas que devem ser classificadas em grupo homogêneos antes que se possa fazer uma análise estatística. Esta classificação é uma tarefa difícil quando o número de respostas a serem analisadas é muito grande. Por isso, questões de resposta aberta são mais freqüentemente empregadas em estudos pilotos ou nos estágios exploratórios, quando se procura determinar que tipos de resposta aparecerão. Essas informações são então usadas na construção do questionário a ser utilizado na obtenção dos dados de um grupo maior. Às vezes é inevitável misturar os dois tipos de pergunta, quando, por exemplo, colocamos a opção “Outros” e pedimos especificação. Se os dados forem analisados por computador, deve-se pensar na etapa da codificação ao redigir as perguntas.

2º. Ordem das questões

Um questionário consiste em uma bateria de questões arranjadas em certa ordem. As primeiras questões são para estabelecer contato com o respondente e devem ser bem simples. Quando vários tópicos estão envolvidos, deve-se completar um tópico antes de passar a outro. A ordem das questões freqüentemente afeta as respostas dadas pelo respondente, já que as perguntas chamam a atenção do entrevistado para um conjunto de pensamentos e sentimentos, em cujo contexto as outras perguntas serão respondidas. Em pesquisa de mercado, por exemplo, questões que mencionam um produto específico tendem a

viciar as perguntas que se seguem; conseqüentemente, estas questões identificando produtos ou firmas devem ser colocadas no final, sempre que possível.

3º. Tipo de abordagem

Muitas pessoas tendem a racionalizar ou exagerar suas respostas quando são questionadas diretamente sobre seus motivos, realizações ou outros assuntos que envolvam seu prestígio ou auto-estima. Para se evitar a introdução de tendenciosidade nessas respostas, usa-se freqüentemente uma abordagem indireta na elaboração de questões que envolvem prestígio. Por exemplo, ao invés de perguntar: "Você terminou o curso secundário?", pode-se perguntar: "Em que ano você estava quando deixou de estudar?" Na segunda pergunta tenta-se evitar constrangimento aos respondentes que não terminaram o curso secundário.

4º. Clareza nas perguntas

Uma pergunta deve ter aproximadamente o mesmo sentido para todos os entrevistados; caso contrário, os dados obtidos não terão grande utilidade. Termos com sentido dúbio devem ser evitados. As perguntas devem ser simples. Nem todos os entrevistados entenderão questões com enunciado complexo, originando, assim, resultados ruins.

5º. Não sugerir respostas

Na formulação das perguntas deve-se evitar um tipo de redação como esta: "Você concorda em que esta bebida, sendo a melhor, deve custar mais caro?" Esta pergunta sugere tão obviamente uma resposta que é praticamente inútil. Algumas vezes, entretanto, é difícil perceber que a redação de uma pergunta possa sugerir determinada resposta.

6º. A necessidade do pré-teste

Assim que um questionário tenha sido redigido, deve ser testado em um estudo piloto. Esta fase é fundamental para detectar dificuldades não observadas, como o "lay out" do questionário, ordem e redação das perguntas, necessidade de instruções mais claras para os entrevistadores etc. Naturalmente, a correção dessas imprecisões melhorará a qualidade do levantamento.

7º. A prática de pesquisas por amostragem

O leitor deve convencer-se de que é fundamental conhecer as características específicas da área onde pretende participar de pesquisas por amostragem. O significado

especial de algumas palavras, os melhores locais e horários para se fazer coleta de dados, o tipo de entrevistador são, entre outros, fatores importantes para o bom andamento do levantamento. Só lendo literatura na área especifica é que se pode, entretanto, conhecer estes detalhes. Por exemplo, para coletar amostras de um minério com vistas ao controle de algum componente indesejável, existem formas já bastante testadas internacionalmente, consagradas mesmo em normas de operação (cf. ABNT: Associação Brasileira de Normas Técnicas). Não se pode iniciar um trabalho nesta área sem um estudo detalhado deste material. O mesmo se aplica a pesquisas de origem-destino, técnica muito usada no planejamento de tráfego urbano. A forma de planejar uma pesquisa de opinião pública é uma das tecnologias mais valiosas de uma firma que vende pesquisas de mercado e de opinião. Cada firma reflete nelas sua experiência.

Para grandes levantamentos, o planejamento do plano amostral e dos instrumentos de coleta exige o concurso de um profissional.

6.3 - Amostragem Aleatória Simples

Neste caso a amostra é escolhida elemento a elemento. A população é numerada de 1 a *N*. Escolhem-se, em seguida, na tábela de números aleatórios (TNA -Tabela 4), *n* números compreendidos entre 1 e *N*. Esse processo é equivalente a um sorteio no qual se colocam todos os números misturados dentro de uma urna. Os elementos correspondentes aos números escolhidos formarão a amostra.

Ex.:A tabela a seguir refere-se aos diâmetros (mm) de 30 eixos produzidos por uma indústria automobilística (dados hipotéticos).

Eixo	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15
Diâm.	26	32	26	19	20	22	30	31	17	20	16	17	28	15	26
Eixo	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
Diâm.	19	14	16	16	26	27	31	13	26	18	29	18	16	21	24

Extrair, sem reposição, uma amostra aleatória de tamanho $n = 5$.

Solução:

Partindo da coluna 3 na TNA, procuramos os cinco primeiros números não superiores a 30, lendo os dois últimos algarismos à direita. Obtemos a amostra:

Leitura na TNA	26	15	03	07	06
Diâmetro	29	26	26	30	22

Poderíamos também tomar os cinco primeiros números não superiores a 30 formados pelos dois primeiros algarismos à esquerda.

Obteríamos a amostra:

Leitura na TNA	27	25	08	03	19
Diâmetro	18	18	31	26	16

6.4 - Amostragem Estratificada

Quando os elementos da população estão divididos em grupos não superpostos, é mais fácil e mais eficiente escolher, independentemente, uma amostra aleatória simples dentro de cada um destes grupos, os quais são chamados estratos.

Esta forma de amostragem é uma das mais utilizadas, já que a maioria das populações têm estratos bem definidos: os homens e as mulheres; os alunos das escolas X, Y, Z; os operários pertencentes às classes salariais 1, 2, 3, 4 etc.

O mais comum é utilizar-se a amostragem estratificada proporcional, que consiste em selecionar os elementos da amostra entre os vários estratos, em número proporcional ao tamanho de cada um dos estratos. Em outras palavras, sejam:

N o número de elementos da população

L o número de estratos

N_i o número de elementos do estrato i

n o tamanho da amostra a ser selecionada.

Note-se que:

$$N = N_1 + N_2 + ... + N_L$$

Calcula-se a fração de amostragem dada por

$$f = \frac{n}{N}$$

número de elementos a serem sorteados em cada estrato será:

$$N_1f, N_2f, ..., N_kf$$

Ex.:Na execução de uma rede elétrica, uma firma especializada utiliza eletrodutos de dois tipos: E e F. Em uma análise do custo do material foram consideradas 30 faturas,

representadas abaixo pelo preço de 10m de eletroduto:

Eletroduto (estrato) *E*

Fatura	01	02	03	04	05	06
Preço (R$)	710	710	715	715	755	760

Eletroduto (estrato) *F*

Fatura	01	02	03	04	05	06	07	08	09	10
Preço (R$)	750	750	750	750	755	760	760	765	765	765
Fatura (cont.)	11	12	13	14	15	16	17	18	19	20
Preço (R$)	765	765	770	770	770	785	785	790	790	795
Fatura (cont.)	21	22	23	24						
Preço (R$)	795	800	810	820						

Extrair, sem reposição, uma amostra estratificada proporcional de tamanho $n = 8$.

Solução:

$f = 8/30 = 0{,}27$

De cada estrato serão sorteadas respectivamente n_1 e n_2 unidades:

$n_1 = (0{,}27)(6) = 1{,}62 \cong 2$

$n_2 = (0{,}27)(24) = 6{,}48 \cong 6$

Lendo os dois últimos algarismos a partir do início da quarta coluna da TNA, obtemos o resultado:

Extrato	*E*		*F*					
Leitura na TNA	03	01	20	03	18	17	24	12
Fatura (R$)	715	710	795	750	790	785	820	765

Entre as vantagens da amostragem estratificada destacam-se:

a) Os dados são geralmente mais homogêneos dentro de cada estrato do que na

população como um todo;

b) O custo da coleta e análise dos dados é freqüentemente menor nesse tipo de amostragem do que na aleatória simples, devido a conveniências administrativas;

c) Podem-se obter estimativas separadas dos parâmetros populacionais para cada estrato sem selecionar outra amostra e, portanto, sem custo adicional.

6.5 - Amostragem por Conglomerado

Uma amostra por conglomerado é uma amostra aleatória simples na qual cada unidade de amostragem é um grupo, ou conglomerado, de elementos.

O primeiro passo para se usar este processo é especificar conglomerados apropriados. Os elementos em um conglomerado devem ter características similares. Como regra geral, o número de elementos em um conglomerado deve ser pequeno em relação ao tamanho da população, e o número de conglomerados, razoavelmente grande.

Tanto no caso da amostragem estratificada como no da amostragem por conglomerado, a população deve estar dividida em grupos. Na amostragem estratificada, entretanto, seleciona-se uma amostra aleatória simples dentro de cada grupo (estrato), enquanto que na amostragem por conglomerado selecionam-se amostras aleatórias simples de grupos, e todos os itens dentro dos grupos (conglomerados) selecionados farão parte da amostra.

A amostragem por conglomerado é recomendada quando:

a) Ou não se tem um sistema de referência listando todos os elementos da população, ou a obtenção dessa listagem é dispendiosa;

b) O custo da obtenção de informações cresce com o aumento da distância entre os elementos.

Ex.:Supondo que se deseje estimar o rendimento médio familiar em uma cidade grande, como deve ser escolhida a amostra?

Solução:

A amostragem aleatória simples é inviável, pois pressupõe uma listagem de todas as famílias da cidade, o que é praticamente impossível de se obter.

A alternativa da amostragem estratificada é também inviável, já que aqui também é necessária uma listagem dos elementos por estrato.

A melhor escolha é amostragem por conglomerado. O sistema de referência pode ser constituído por todos os quarteirões da cidade. cada quarteirão é um conglomerado. Extrai-se uma amostra aleatória simples de quarteirões e neles pesquisa-se a renda familiar em todas as casas.

6.6 - Amostragem Sistemática

Uma amostra sistemática de tamanho n é constituída dos elementos de ordem K, $K + r$, $K + 2r$, ..., onde K é um inteiro escolhido aleatoriamente entre 1 e n e r é o inteiro mais próximo da fração N/n. Por exemplo, se a população tem 100 elementos e vamos escolher uma amostra de tamanho 6, K é um inteiro escolhido aleatoriamente entre 1 e 6 e $r = 100/6 = 16{,}6 = 17$. Se $K = 3$, a amostra será composta pelos seguintes elementos: 3 20 37 54 71 88.

Se o tamanho da população é desconhecido, não podemos determinar exatamente o valor de r. Escolheremos intuitivamente um valor razoável para r.

Às vezes a amostragem sistemática é preferida à amostragem aleatória simples, porque é mais fácil de executar, estando portanto sujeita a erros, e proporciona mais informações com menor custo.

Ex.: Escolha a técnica adequada para extrair uma amostra de 50 compradores de uma loja.

Solução:

A amostragem aleatória simples não pode ser empregada neste caso, pois o entrevistador não pode determinar quais compradores serão incluídos na amostra, uma vez que não se conhece o tamanho N da população, até que todos os compradores tenham ido à loja. Assim, ele pode usar a amostragem sistemática (digamos 1 em cada 20 compradores) até obter a amostra do tamanho desejado.

6.7 - Cálculo do Tamanho da Amostra

Este tópico apresenta alguma dificuldade técnica e só um estatístico profissional, conhecedor dos detalhes da situação, poderá calcular o tamanho da amostra necessária em uma pesquisa específica. Limitar-nos-emos aqui a apresentar fórmulas para cálculo do tamanho da amostra em duas situações simples, admitindo que será usada a amostragem aleatória simples.

Proporção

Se a pergunta é do tipo SIM-NÃO e queremos, com nível (1 - α) de confiança, que a proporção estimada esteja, no máximo, a uma distância *d* da proporção verdadeira, ou seja, se queremos que

$$P\left[\,|p-\hat{p}| \le d\right] = 1-\alpha$$

o valor de *n* é dado por

$$\frac{z^z PQ}{d^2}$$

onde $z = Z_{1-\frac{\alpha}{2}}$, *P* é uma estimativa preliminar de *p*, *Q = 1 - P* e *d* é o maior desvio aceitável.

Ex:Deve-se realizar uma pesquisa de opinião pública para determinar a proporção de pessoas que sofreram lesões em conseqüência de algum tipo de acidente de trânsito. Quantas pessoas deverão ser ouvidas para que sejam satisfeitas as seguintes condições: *d* = 0,02, *p* = 10%, α = 5%?

Solução:

$$P = 0{,}1 \qquad Q = 1-P = 0{,}9 \qquad d = 0{,}02$$

$$\alpha = 0{,}05 \qquad z = 1{,}96$$

$$n = \frac{z^2 PQ}{d^2} = \frac{(1{,}96)^2(0{,}1)(0{,}9)}{(0{,}02)^2} = 865 \text{ pessoas}$$

Média

Há dois casos a considerar:

1. Controlar o erro relativo da estimativa da média de uma característica da população, com nível 1 - α de confiança: ou seja, queremos

$$P\left[\left|\frac{\bar{y}-\mu}{\bar{y}}\right| \le r\right] = 1-\alpha$$

onde $\bar{y}$ = média amostral, μ = média populacional. Neste caso o valor de *n* é dado por

$$n = \left(\frac{zs}{r\bar{y}}\right)^2$$

onde $z = z_{1-\frac{\alpha}{2}}$, s=desvio padrão preliminar, r=erro relativo, $\bar{y}$=média preliminar

Ex.:Quantos carros devem ser observados para se determinar ao nível α = 5% e com *r* = 10%, a velocidade média em uma avenida, com $\bar{y}$=40km/h e *s* = 10 Km/h?

Solução:

$\bar{y} = 40 \quad s = 10 \quad r = 0{,}1 \quad \alpha = 0{,}05 \quad z = 1{,}96$

$$n = \left(\frac{zs}{r\bar{y}}\right)^2 = \left[\frac{(1{,}96)\,(10)}{(0{,}1)\,(40)}\right]^2 = 24 \text{ carros}$$

Note-se a necessidade de estimativas preliminares dos parâmetros, o que constitui mais uma razão para se fazer um estudo piloto.

2. Se quisermos controlar o erro absoluto da média, ou seja, se quisermos que

$$P\left[\left|\bar{X}-\mu\right| \le d\right] = 1-\alpha$$

O valor de *n* é dado por

$$n = \left(\frac{z\sigma}{d}\right)^2$$

onde $z = z_{1-\frac{\alpha}{2}}$, σ = desvio padrão populacional, *d* = distância da média estimada à média verdadeira.

Ex.:Com os dados do exemplo anterior, determinar *n* para *d* = 5, σ = 12.

Solução:

$$n = \left(\frac{z\sigma}{d}\right)^2 = \left[\frac{(1{,}96)\,(12)}{5}\right]^2 = 23 \text{ carros}$$

Populações finitas

As expressões anteriores são válidas para populações infinitas ou, na prática, muito grandes. Para populações finitas o tamanho da amostra é dado por

$$n = \frac{n_0}{1 + \frac{n_o}{N}}$$

onde n_0 é o valor obtido pelas expressões anteriores e N o tamanho da população. Freqüentemente, em populações com pequeno número de indivíduos, todos terão de ser incluídos na “amostra”.

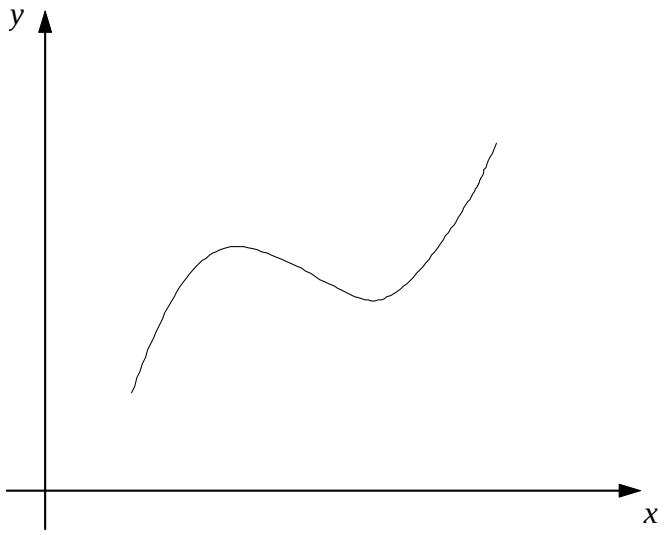
Capítulo 7 - Regressão

Muitas vezes a posição dos pontos experimentais no *diagrama de dispersão* sugere a existência de uma relação funcional entre as duas variáveis. Surge então o problema de se determinar uma função que exprima esse relacionamento.

Esse é o problema da regressão, conforme a denominação introduzida por Fisher e universalmente adotada.

Assim, se os pontos experimentais se apresentarem como na figura seguinte admitiremos existir um relacionamento funcional entre os valores y e x, responsável pelo aspecto do diagrama, e que explica grande parte da variação de y com x , ou vice-versa. Este relacionamento funcional corresponderia à linha existente na figura, que seria a "linha de regressão". Uma parcela da variação, entretanto, permanece em geral sem ser explicada, e será

atribuída ao acaso. Em outras palavras, admitimos existir uma função que justifica, em média, a variação de uma das variáveis com a outra. Na prática, os pontos experimentais terão uma variação em torno da linha representativa dessa função, devido à existência de uma variação aleatória adicional, que chamaremos de *variação residual.*



Linha de regressão

Essa função de regressão, portanto, nos dá o valor médio de uma das variáveis em função da outra; por exemplo, $\mu(Y/x)$. Posto dessa forma, o problema que vamos examinar será, dados os pontos experimentais, o de realizar uma indução quanto à expressão

matemática da função de regressão.

Evidentemente, tudo se simplificará se a forma da linha de regressão for suposta conhecida. O problema, então, se reduzirá apenas à estimação de seus parâmetros. Esse caso ocorrerá se existirem razões teóricas que permitam saber de antemão qual o modelo que rege o comportamento de uma variável em função da outra. A lei de Hook, por exemplo, afirma que, dentro de certos limites, as deformações de corpos metálicos variam linearmente com as tensões aplicadas. Na análise de um experimento desse tipo, portanto, o modelo linear para a linha de regressão poderia ser adotado de início. Pode também ocorrer o caso em que a forma da linha fica evidente da própria análise do diagrama de dispersão.

Caso a forma da linha de regressão não seja conhecida de antemão, ela deverá ser inferida juntamente com seus parâmetros. Teremos, então, além do problema de estimação dos parâmetros do modelo da linha de regressão, a dificuldade adicional de especificar a forma do modelo.

No estudo que segue, vamos admitir inicialmente que a forma da linha de regressão seja uma reta. Teremos então o problema da *regressão linear simples*, que veremos a seguir (O termo “simples” destina-se a frisar que temos apenas duas variáveis.). Posteriormente, estudaremos a regressão polinomial, em que a forma da função é suposta um polinômio de grau superior a 1. Nos dois casos a idéia e os princípios fundamentais serão os mesmos que discutiremos em seguida.

Vamos também admitir que a variável X seja suposta sem erro, ou seja, não-aleatória, enquanto que a variável Y apresenta uma parcela de variação residual, a qual é responsável pela dispersão dos pontos experimentais em torno da linha de regressão. Essa suposição permite utilizar um modelo que simplifica a solução do problema, e é justificável porque muitos casos práticos se aproximam dele. Na verdade, encontraremos, na prática, muitos casos em que a variável X pode ser medida com precisão muito maior do que Y, o que coloca o problema praticamente nas condições supostas.

A situação descrita corresponde, muitas vezes, a experimentos em que os valores de X são pré-determinados ou pré-escolhidos pelo experimentador, já que a variável X é suposta não-aleatória. No entanto, os valores de Y, sendo aleatórios, não podem ser exatamente previstos, e serão determinados experimentalmente. Podemos, por exemplo, medir as temperaturas de um forno em aquecimento de 5 em 5 min., a partir de um instante 0. Ora, a menos de pequenas imprecisões, totalmente desprezíveis, os tempos (valores de X) estão bem determinados, ao passo que as temperaturas deverão ser verificadas no decurso do

experimento. Vemos que, neste exemplo, os valores de X independem dos de Y, pois foram simplesmente arbitrados, enquanto que os valores de Y dependerão dos de X desde que exista regressão. Por essa razão, a variável X é dita *variável independente*, enquanto Y é dita *variável dependente*.

No caso da regressão múltipla, em que mais de duas variáveis são envolvidas, obteremos uma equação para prever valores de uma variável dependente em função de duas ou mais variáveis independentes.

O modelo acima descrito, portanto, considera que os valores da variável aleatória Y dependerão do(s) valor(es) assumido(s) pela(s) variável (is) independente(s) e também do acaso, isto é, estarão sujeitos a uma variação aleatória que se sobrepõe à variação explicada pela função de regressão. Isso pode ser expresso sob a forma.

$$y = \varphi(x) + \psi,$$

onde φ denota a função de regressão e ψ a componente aleatória da variação de Y. No caso de regressão múltipla, x deverá ser interpretado como um vetor de valores das variáveis independentes.

Ora, é perfeitamente coerente com a idéia contida no modelo admitir-se que a variável aleatória ψ tenha média 0, a fim de que toda a variação explicada de y fique concentrada em $\varphi(x)$. Isso significa que a função de regressão fornece a média de y para cada x considerado, conforme já mencionado.

Uma outra suposição básica que adotaremos no nosso estudo é a de que a variação residual da variável Y seja constante com X. Dito em outras palavras, isso significa admitir que a variação de Y em torno da linha teórica de regressão pode ser descrita por um desvio-padrão residual que independe do ponto considerado.

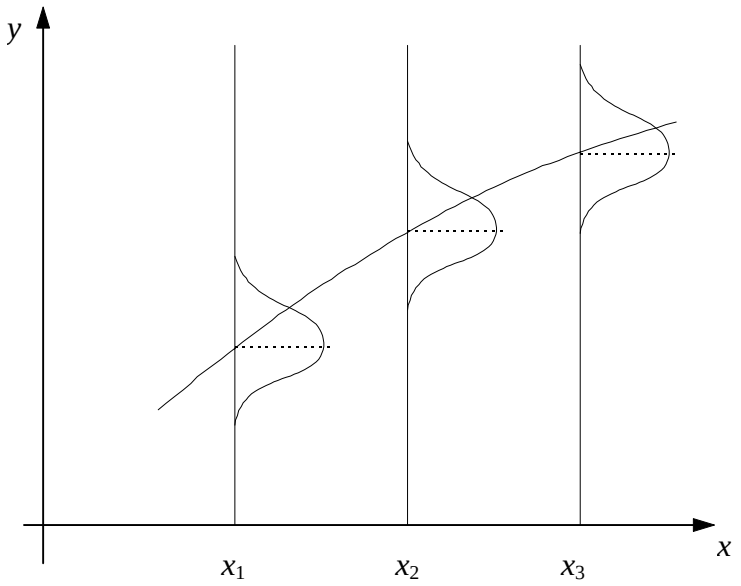
Por fim, admitiremos que a variação de Y em torno da linha teórica de regressão se dê segundo uma distribuição normal. Como a linha teórica de regressão dá os valores médios de Y em função de X, essa suposição implica considerar que a variação residual de Y seja normalmente distribuída com média zero e desvio-padrão constante. Uma tentativa de representar graficamente essa situação é feita na figura seguinte. Frise-se que a linha ali representada é a linha teórica de regressão, da qual obteremos uma estimativa a partir dos pontos experimentais.

7.1 - Regressão Linear Simples

Suponhamos ser a linha teórica de regressão uma reta e que queiramos estabelecer a regressão de Y em função de X. Logo, a função que desejamos obter é da forma

$$y = \alpha + \beta x$$

Estimaremos os parâmetros α e β da reta teórica através dos pontos experimentais



Variação residual em torno da linha teórica de regressão.

fornecidos pela amostra, obtendo uma reta estimativa na forma

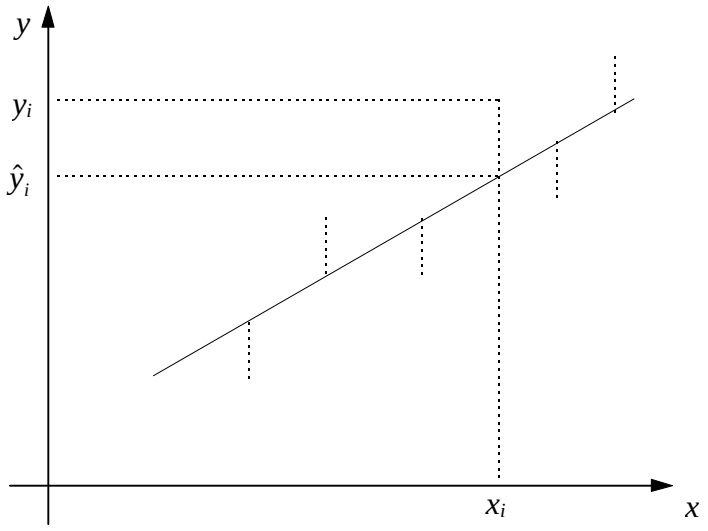
$$\hat{y} = a + bx,$$

onde a é a estimativa do parâmetro α, e b, também chamado *coeficiente de regressão linear*, é a estimativa do parâmetro β. O símbolo $\hat{y}$ é utilizado para uma conveniente distinção dos valores dados pela reta estimativa, das ordenadas dos pontos experimentalmente obtidos.

Existem diversos métodos para a obtenção da reta desejada. O mais simples de todos, que podemos chamar de "método do ajuste visual", consiste simplesmente em traçar diretamente a reta, com auxílio de uma régua, no diagrama de dispersão, procurando fazer, da melhor forma possível, com que essa reta passe por entre os pontos. Esse procedimento, entretanto, somente será razoável se a correlação linear for muito forte, caso contrário levará a

resultados subjetivos. Acima de tudo, ademais, merece a crítica de ser um procedimento nem um pouco científico.

Por outro lado, a aplicação do *princípio de máxima verossimilhança* leva, nas condições admitidas, ao chamado procedimento de *mínimos quadrados*, segundo o qual a reta a ser adotada deverá ser aquela que torna mínima a soma dos quadrados das distâncias da reta aos pontos experimentais, medidas no sentido da variação aleatória. Ou seja, devemos procurar a reta para a qual se consiga minimizar $\Sigma_{i=1}^{n} d_i^2$, sendo as distâncias d_i as indicadas na figura seguinte. A idéia central desse procedimento é simplesmente a de minimizar a variação residual em torno da reta estimativa.



Distâncias cuja soma dos quadrados deve ser minimizada.

Tendo em vista a expressão $\hat{y} = a + bx$, devemos, portanto, impor a condição

$$\min \Sigma_{i=1}^{n} d_i^2 = \min \Sigma_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2 = \min \Sigma_{i=1}^{n} \left(y_i - a - bx_i\right)^2 .$$

Os valores a e b que minimizam essa expressão serão aqueles que anulam as derivadas parciais dessa expressão. Ou seja, devemos ter

$$\frac{\partial}{\partial a} \sum_{i=1}^{n} d_i^2 = 0 \text{ e } \frac{\partial}{\partial b} \sum_{i=1}^{n} d_i^2 = 0.$$

Considerando a última forma da expressão, chega-se facilmente às expressões

$$-2\sum_{i=1}^{n} (y_i - a - bx_i) = 0,$$
$$-2\sum_{i=1}^{n} x_i (y_i - a - bx_i) = 0,$$

as quais imediatamente fornecem o seguinte sistema de duas equações a duas incógnitas:

$$\begin{cases} \sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i, \\ \sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 \end{cases}$$

Os pontos experimentais fornecem os elementos para a montagem desse sistema, cuja solução forneceria os coeficientes *a* e *b*. Entretanto é mais fácil considerar de uma vez a solução analítica do sistema, a qual fornece

$$\begin{cases} b = \dfrac{\sum_{i=1}^{n} (x_i - \bar{x}) y_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \dfrac{S_{xy}}{S_{xx}}, \\ a = \bar{y} - b\bar{x}, \end{cases}$$

expressões que dão diretamente os coeficientes *a* e *b* que desejávamos obter.

Algumas vezes torna-se interessante fazer codificações lineares nos valores das variáveis para simplificar os cálculos. Entretanto isso não acarreta maiores problemas, bastando, ao final, compensar as codificações feitas. Por exemplo, se fizermos as transformações

$$w_i = \frac{x_i - x_0}{h} \text{ e } z_i = y_i - y_0,$$

obteremos, aplicando o procedimento visto, a reta na forma

$$\hat{z} = a + bw, \ \therefore \ \hat{y} - y_0 = a + b \cdot \frac{x - x_0}{h}.$$

Essa expressão pode ser escrita

$$\hat{y} = \left(a + y_0 - \frac{bx_0}{h} \right) + \frac{b}{h} x,$$

que é a equação da reta desejada na forma final.

Ex.: 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 diagrama de dispersão. Calcular o coeficiente de correlação

linear.

Solução. Na tabela seguinte temos os valores necessários à determinação dos coeficientes da reta. Temos, aplicando as expressões anteriores,

$$S_{xy} = 50{,}5 - \frac{36 \cdot 9{,}2}{8} = 50{,}5 - 41{,}4 = 9{,}1,$$
$$S_{xx} = 204 - \frac{(36)^2}{8} = 204 - 162 = 42.$$

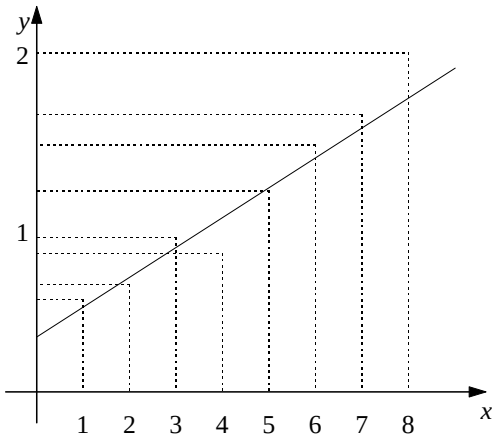
Tabela - Valores para o cálculo da reta e do coeficiente de correlação linear

x_i	y_i	$x_i y_i$	x_i^2	y_i^2
1	0,5	0,5	1	0,25
2	0,6	1,2	4	0,36
3	0,9	2,7	9	0,81
4	0,8	3,2	16	0,64
5	1,2	6,0	25	1,44
6	1,5	9,0	36	2,25
7	1,7	11,9	49	2,89
8	2,0	16,0	64	4,00
36	9,2	50,5	204	12,64

$$b = \frac{S_{xy}}{S_{xx}} = \frac{9{,}1}{42} \cong 0{,}217,$$
$$a = \bar{y} - b\bar{x} \cong \frac{9{,}2}{8} - 0{,}217 \cdot \frac{36}{8} = 1{,}150 - 0{,}976 = 0{,}174.$$

Logo, a equação da reta de mínimos quadrados, dada na figura seguinte é

$$\hat{y} = 0{,}174 + 0{,}217x$$



Reta de mínimos quadrados

Para o cálculo do coeficiente de correlação, é necessário usar os valores da coluna y_i^2 da tabela, calculando-se

$$S_{yy} = 12{,}64 - \frac{(9{,}2)^2}{8} = 12{,}64 - 10{,}58 = 2{,}06, \therefore r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{9{,}1}{\sqrt{42 \cdot 2{,}06}} \cong 0{,}98$$

Esse alto valor do coeficiente de correlação linear de Pearson justifica o traçado da reta de regressão.

7.1.1 - Reta passando pela origem

Em certos casos, sabemos que a reta teórica de regressão deve passar pela origem, ou seja, que sua equação é da forma

$$y = \beta x.$$

Nesse caso, temos apenas a estimar o coeficiente angular β da reta. A reta estimativa será

$$\hat{y} = bx.$$

e a aplicação do princípio dos mínimos quadrados leva a impor a condição

$$\min \sum_{i=1}^{n} d_i^2 = \min \sum_{i=1}^{n} (y_i - bx_i)^2 .$$

Anulando a derivada em relação a *b*, temos, de forma análoga à vista para o caso geral,

$$\frac{d}{db}\sum_{i=1}^{n}(y_i - bx_i)^2 = 0, \therefore -2\sum_{i=1}^{n} x_i(y_i - bx_i) = 0,$$

$$\therefore \sum_{i=1}^{n} x_i y_i = b\sum_{i=1}^{n} x_i^2, \therefore b = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}.$$

Assim, no exemplo anteriormente visto, se considerássemos que a reta de regressão devesse ultrapassar pela origem, o coeficiente *b* seria

$$b = \frac{50{,}5}{204} \cong 0{,}248,$$

e a equação da reta de mínimos quadrados seria

$$\hat{y} = 0{,}248x.$$

7.1.2 - Funções linearizáveis

O modelo de regressão linear simples que acabamos de estudar é útil em muitas situações reais. Um caso que merece menção e que recai naquele modelo é o das funções linearizáveis.

Certas funções, mediante transformações convenientes, linearizam-se, o que torna simples a solução do problema de regressão. Assim, por exemplo, se admitirmos que a função de regressão seja uma função exponencial do tipo

$$y = \alpha \cdot \beta^x ,$$

a aplicação de logaritmos promove a linearização da função na forma

$$\log y = \log \alpha + x \log \beta.$$

Um artifício simples para se saber se a transformação logarítmica promove uma boa linearização consiste em usar papéis monologarítmicos ou dilogarítmicos. No presente caso, a adequabilidade da transformação vista seria evidenciada se, plotando-se os valores de x segundo escala linear e os de y segundo a escala logarítmica de um papel monologarítmico, os pontos observados se aproximassem de uma reta.

Chamando $z = \log y$, $A = \log \alpha$ e $B = \log \beta$, passamos a ter o problema de estimar os parâmetros da reta

$$z = A + Bx,$$

o qual sabemos resolver. Para tanto, basta trabalhar com os valores x_i versus $z_i = \log y_i$, obtendo as estimativas de A e B, cujos antilogaritmos serão as estimativas de α e β. Analogamente, se tivermos $y = \alpha x^\beta$, o problema se resolverá trabalhando-se com $\log y_i$ versus $\log x_i$.

Outros casos de fácil linearização podem também ser encontrados, como, por exemplo, se $y = a + bx^2$, $y = (a + bx)^{-1}$, etc. Cuidado especial, entretanto, deve ser tomado quanto o processo de linearização envolve transformações na variável dependente Y.

7.2 - Regressão Polinomial

O mesmo princípio dos mínimos quadrados visto para a regressão linear poderá

ser aplicado se admitirmos que a função de regressão é um polinômio de grau $k > 1$. A diferença está em que temos $k + 1$ coeficientes a estimar.

Tomando as derivadas parciais em relação às $k + 1$ estimativas, chegamos a um sistema de $k + 1$ equações com $k + 1$ incógnitas o qual, resolvido, fornece a solução do problema. Uma dificuldade tradicional nesse tipo de problema era a solução do sistema no caso de valores relativamente grande de k. Com o surgimento dos computadores, essa dificuldade praticamente deixou de existir. A mesma observação é válida no caso de regressão linear múltipla. No caso de admitirmos que a função de regressão seja uma parábola na forma

$$y = \alpha + \beta x + yx^2,$$

devemos obter a parábola-estimativa

$$\hat{y}_p = a + bx + cx^2.$$

O sistema de equações ao qual chegamos é:

$$\begin{cases} \sum y_i = na + b\sum x_i + c\sum x_i^2, \\ \sum x_i y_i = a\sum x_i + b\sum x_i^2 + c\sum x_i^3, \\ \sum x_i^2 y_i = a\sum x_i^2 + b\sum x_i^{32} + c\sum x_i^4. \end{cases}$$

Se os valores de x_i forem igualmente espaçados, a solução desse sistema fica bastante simplificada se trabalharmos com $x_i - \bar{x}$ ao invés dos x_i, pois os somatórios de potenciais ímpares se anulam. O sistema ficará então

$$\begin{cases} \sum y_i = na + c\sum (x_i - \bar{x})^2, \\ \sum (x_i - \bar{x}) y_i = b\sum (x_i - \bar{x})^2, \\ \sum (x_i - \bar{x})^2 y_i = a\sum (x_i - \bar{x})^2 + c\sum (x_i - \bar{x})^4. \end{cases}$$

Vemos que o coeficiente b do sistema modificado seria obtido diretamente, sendo o mesmo da reta, e restaria apenas resolver um sistema de duas equações a duas incógnitas para se determinar a e c. A parábola seria, evidentemente, obtida na forma

$$\hat{y}_p = a + b(x - \bar{x}) + c(x - \bar{x})^2.$$

Ex. Ajustar a parábola de mínimos quadrados aos dados do exemplo anterior.

Solução. Já temos

$$
\begin{aligned}
n &= 8 \\
S_{xy} &= \sum (x_i - \bar{x}) y_i = 9{,}1, \\
S_{xx} &= \sum (x_i - \bar{x})^2 = 42, \\
\sum y_i &= 9{,}2, \\
b &= \frac{S_{xy}}{S_{xx}} = \frac{9{,}1}{42} \cong 0{,}2167 \cong 0{,}217, \\
\bar{x} &= 4{,}5.
\end{aligned}
$$

Falta obtermos $\sum \left(x_i - \bar{x}\right)^2 y_i$ e $\sum \left(x_i - \bar{x}\right)^4$, o que será feito com o auxílio da tabela seguinte. Temos, então, o sistema

$$
\begin{cases}
9{,}2 = 8a + 42c, \\
50{,}9 = 42a + 388{,}5c,
\end{cases}
$$

o qual, resolvido, fornece

$a = 1{,}0688$ e $c = 0{,}0155$.

Logo, a parábola de mínimos quadrados será

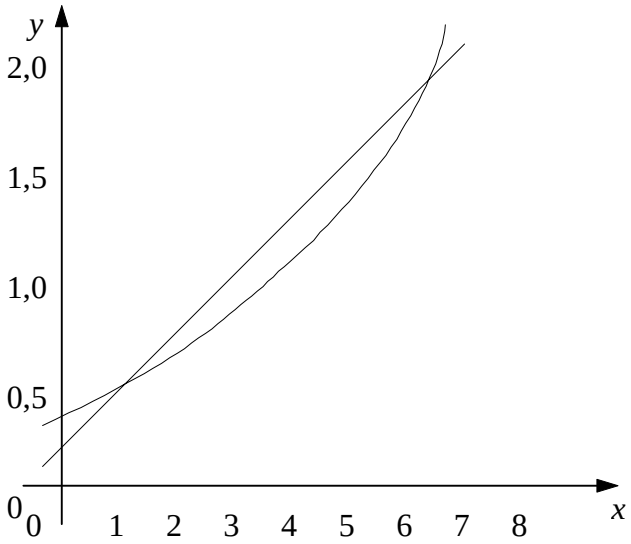
$$\hat{y}_p = 1{,}0688 + 0{,}2167(x - 4{,}5) + 0{,}0155(x - 4{,}5)^2,$$

x_i	y_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^2 y_i$	$(x_i - \bar{x})^4$
1	0,5	-3,5	12,25	6,125	150,0625
2	0,6	-2,5	6,25	3,750	39,0625
3	0,9	-1,5	2,25	2,025	5,0625
4	0,8	-0,5	0,25	0,200	0,0625
5	1,2	0,5	0,25	0,300	0,0625
6	1,5	1,5	2,25	3,375	5,0625
7	1,7	2,5	6,25	10,625	39,0625
8	2,0	3,5	12,25	24,500	150,0625
	9,2		42,00	50,900	388,5000

que pode ser colocada na forma

$$\hat{y}_p = 0{,}408 + 0{,}0772x + 0{,}0155x^2.$$

Essa parábola, juntamente com a reta já obtida, encontra-se traçada na figura seguinte.

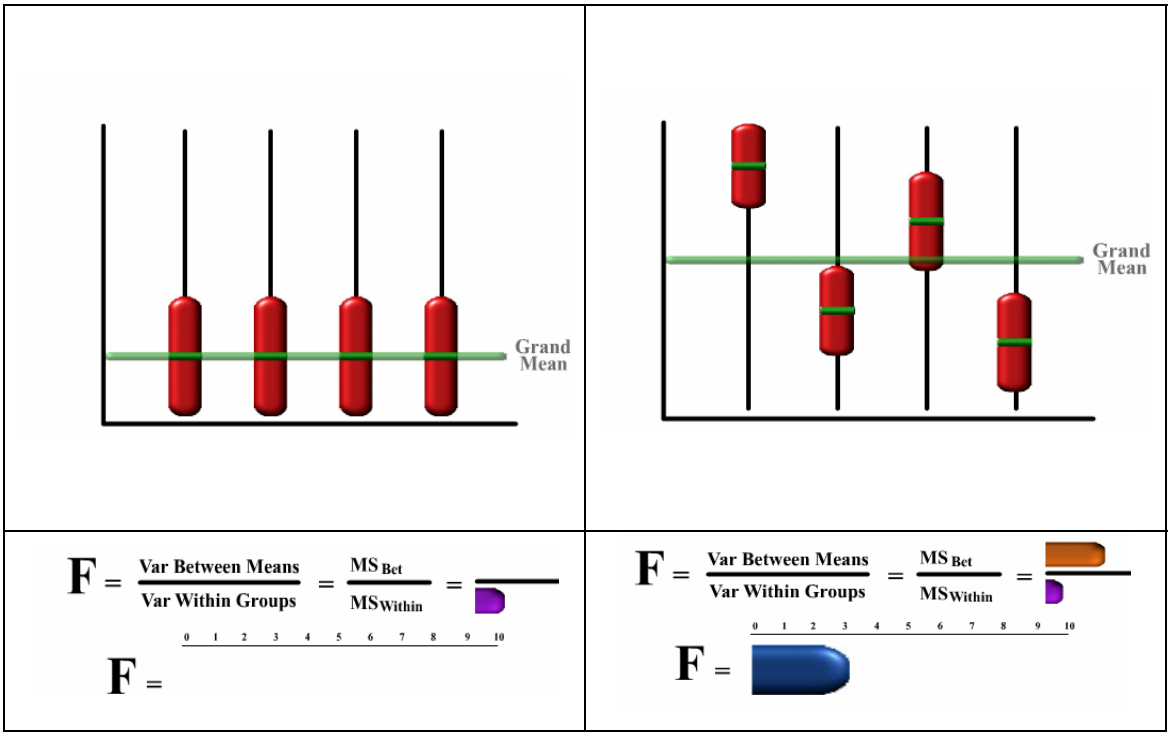

y
2,0
1,5
1,0
0,5
0
0
1
2
3
4
5
6
7
8
x

Capítulo 8 - Anova

A Análise de Variância (ANOVA) é uma técnica utilizada para se testar a igualdade entre médias de vários grupos. Supondo que amostras aleatórias e independentes de *k* populações estejam disponíveis, com *n* observações cada e, denotando-se as *k* médias por $\mu_1, \mu_2, \ldots, \mu_k$, uma análise de variância para um fator (*One-way*) pode ser conduzida para se testar a hipótese nula descrita da seguinte forma:

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$$
$$H_1 : \textbf{\textit{pelo menos uma média } } \mu_k \textbf{\textit{ é diferente das demais}}$$

Ao contrário do que sugere o nome, a Anova é um teste para médias e não para variâncias. Criada por Sir R. A. Fisher por volta de 1920, este procedimento matemático que desencadeou o desenvolvimento da Metodologia de Projeto de Experimentos e Superfície de Resposta, é baseado no relacionamento entre as duas formas básicas de variação: a variação dentro da amostra e a variação entre amostras.



A figura anterior mostra o princípio da análise de variância: uma comparação entre os dois tipos de variação existentes nas amostras provenientes de dados agrupados em níveis: a variação dentro (*Within*) e entre (*Between*).

A variação dentro da amostra (*Within*) corresponde à variância (ou o desvio padrão) que ocorre com as réplicas de um grupo. Do ponto de vista do controle estatístico de processo, a variância dentro da amostra representa as variações de causa comum que ocorrem com os processos produtivos, serviços etc. Como cada grupo (ou nível) terá uma média x_i para as n observações (repetições ou replicações), é possível se calcular uma média para as i médias de cada grupo. Dessa forma, considerando-se a média das médias como o valor de tendência central dos grupos, utilizando-se as diferenças entre cada média e a média geral, encontra-se a variância entre os grupos. Em termos de controle de qualidade, a "variância entre" (*Between*) é a expressão matemática para a variação de causas especiais. Dividindo-se a variância entre pela variância dentro obtém-se a estatística de teste F. A distribuição F (***Figura 8.2***) criada para expressar todos os possíveis valores desta razão entre variâncias, inicia-se em zero e estende-se no sentido positivo do eixo x. Como por princípio, a variância dentro deva ser constante para todos os grupos, a razão F aumentará muito quando as médias dos grupos forem muito diferentes. Quanto mais diferentes, maior o valor de F. Se este valor ultrapassar o valor crítico para F, então haverá evidência estatística suficiente para se rejeitar a hipótese nula de igualdade entre as múltiplas médias.

Este teste é muito utilizado para se verificar a influência de certos fatores sobre uma resposta de interesse. Em termos práticos, pode-se testar como os diversos fatores presentes nos processos industriais, comerciais e outros, exercem influência sobre uma característica de qualidade dos produtos e serviços. Quando dois níveis de um fator geram respostas médias estatisticamente iguais, isto significa que o fator não influencia a resposta. Quando, ao contrário, a resposta média em um nível específico é estatisticamente diferente daquela obtida nos outros níveis, o fator é significativo. Para um experimento de 2 fatores A e B, em dois níveis codificados (+1, -1), pode-se escrever a seguinte equação como modelo estatístico do experimento:

$$Y_{ijk} = \mu + A_i + B_j + (AB)_{ij} + \varepsilon_{ijk} \qquad (8.1)$$

Onde: i = número de níveis do fator A ($i = 1, 2, 3,..., a$); j = número de níveis do Fator B ($j = 1, 2, 3,..., b$); k = número de réplicas de cada combinação de fatores (k = 1, 2, 3,..., n).

O termo Y_{ijk} é a (ijk)-ésima observação obtida no experimento, μ é um parâmetro comum a todos os tratamentos, denominado de média geral, A_i é efeito do i-ésimo tratamento do Fator A, B_j o efeito do j-ésimo tratamento do Fator B, $(AB)_{ij}$, o efeito da ij-ésima interação AB entre os fatores e ε_{ijk}, um componente do erro aleatório. Considere o seguinte experimento formado por dois fatores de interesse (Temperatura e Pressão), em dois níveis cada um (40°C e 60°C, 25 e 35 PSI), com 4 réplicas:

TEMPERATURA	**PRESSÃO**						**Σ**
	25 PSI			**35 PSI**			
40°C	130	155	**539**	34	40	**229**	**768**
	74	180		80	75		
60 °C	150	188	**623**	136	122	**479**	**1102**
	159	126		106	115		
Σ	**1162**			**708**			**1870**

Tabela 8.1 – Dados para uma ANOVA com 2 fatores, 2 níveis e 4 réplicas.

O parâmetro de teste utilizado pela ANOVA para verificar a igualdade entre as médias baseia-se na relação existente entre a variação dentro de um tratamento (*Within*) e a variação entre tratamentos (*Between*) (***Figura 8.1***).

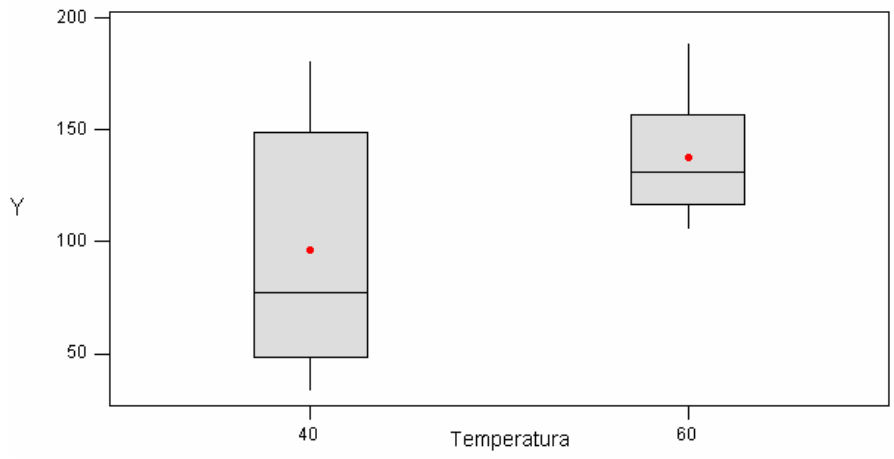


Figura 8.1 – Boxplot do Fator Temperatura.

A variação dentro origina-se das réplicas. Calculando-se a média das réplicas, obtém-se a média dentro do tratamento. Quanto mais as réplicas diferirem desta média, maior será a variação dentro deste tratamento.

Na ***tabela 8.1*** nota-se que as réplicas são muito diferentes, proporcionando uma variabilidade muito grande; por isso o *boxplot* é largo (o intervalo interquartílico é consideravelmente grande). A partir desse gráfico, percebe-se que, para a Temperatura, a variação no nível 40° é maior do que no nível 60°.

Cada um dos *boxplots* representa a variabilidade *dentro* de cada um dos dois níveis do fator, enquanto que a diferença entre as médias denota a variação *entre* os fatores. A *variação dentro* é também chamada de Erro.

Dividindo-se a *Variação Entre* pela *Variação Dentro* obtém-se a *Estatística de Teste F*, que, quando comparada com um valor de *F crítico*, definido segundo o nível de significância e o número de graus de liberdade (***Figura 8.2***), possibilita a aceitação ou rejeição da hipótese nula, acerca da igualdade entre as médias dos níveis dos fatores.

Sempre que se toma uma decisão ou se faz uma inferência sobre uma população a partir de dados amostrais corre-se o risco de se cometer algum tipo de erro. Hipótese é tão somente uma afirmação que se deseja testar, tendo como fonte de provas os dados provenientes da análise da amostra.

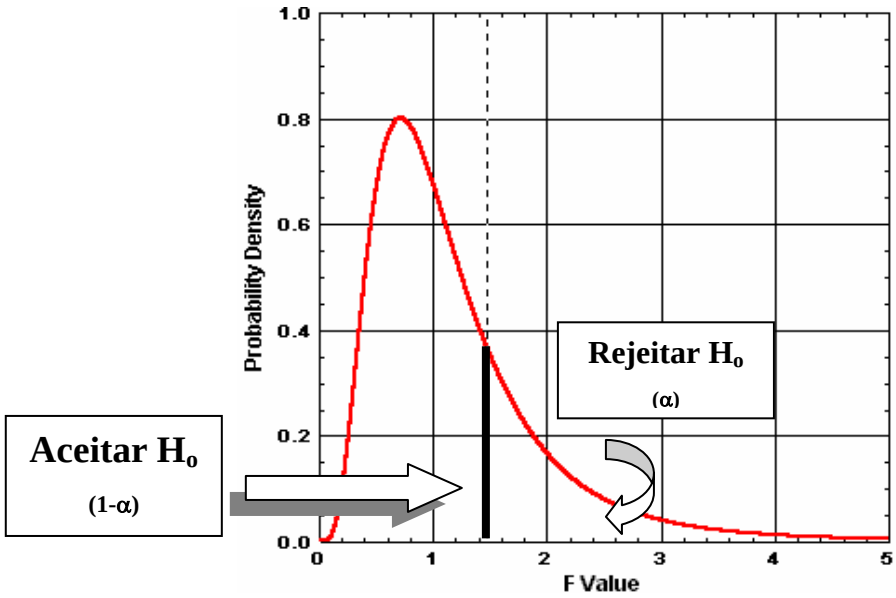


Figura 8.2 – Regiões de Rejeição e Aceitação de Ho para uma distribuição F.

A ***figura 8.3***, indica que os erros do Tipo I e II não podem ser cometidos simultaneamente.

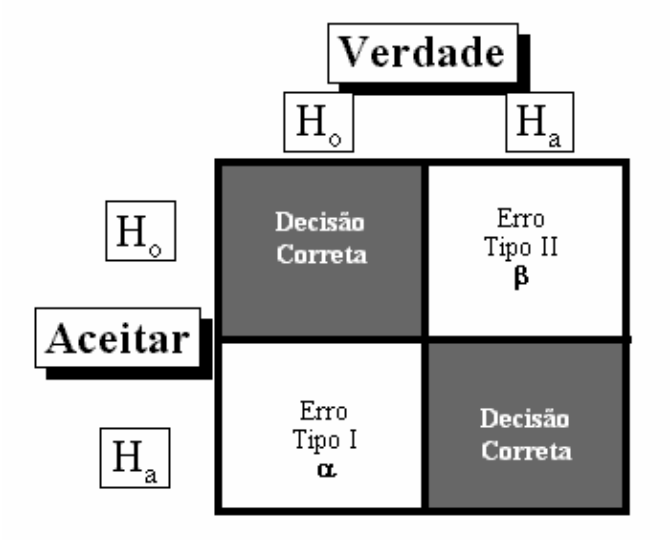


Figura 8.3 – Avaliação do Erro de Decisão.

Existem dois tipos de erro (mutuamente exclusivos) que podem ser cometidos em um teste de hipóteses ***(figura 8.3)***. O primeiro é o *Erro Tipo I*, que consiste em se rejeitar a hipótese nula quando ela é verdadeira. A probabilidade de se cometer este tipo de erro é α, denominado *nível de significância* do teste.

O segundo tipo é o *Erro Tipo II*, que consiste em se aceitar a hipótese nula quando ela for falsa. A probabilidade de se cometer este tipo de erro é β. A probabilidade de se aceitar a hipótese nula quando ela é verdadeira é $1-\alpha$, e é denominado *nível de confiança* do teste. A probabilidade de se rejeitar a hipótese nula quando ela for falsa é $1-\beta$, e é chamada de *Poder do Teste*.

A amostra fornece a *Estatística de Teste*, parâmetro que quando comparado ao *Valor Crítico* determinado por α, conduz à aceitação ou rejeição da hipótese nula. Ao conjunto de todos os valores da estatística de teste que levam à rejeição da hipótese nula dá-se o nome de Região Crítica.

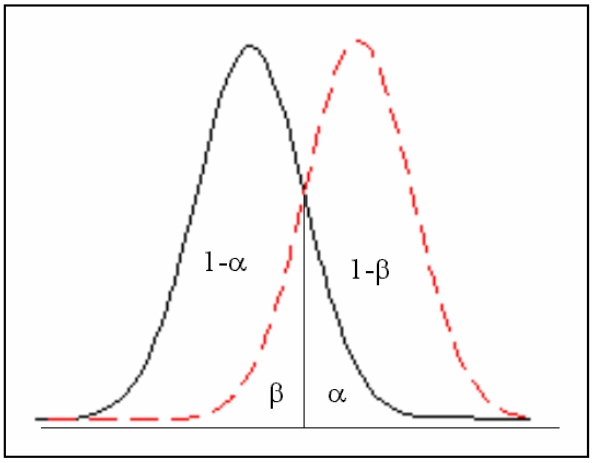


Figura 8.4 – Representação das Probabilidades α e β.

Cada estatística de teste está associada a uma distribuição de probabilidade (*t*-student, F-Fisher, Qui-Quadrado, Normal Padronizada, etc.). No caso da ANOVA, a distribuição utilizada para o teste de hipótese é a Distribuição F de Fisher-Snedecor.

8.1 – Anova para Dois Fatores

Para se avaliar a significância do efeito dos níveis de um tratamento A, bem como de um tratamento B, faz-se necessário testar a hipótese da igualdade entre as respostas médias obtidas com os níveis dos dois fatores. É importante se determinar também, se existe interação entre os dois tratamentos. As hipóteses necessárias neste caso estão representadas na ***tabela 3.2***.

HIPÓTESE	TRATAMENTOS	INTERAÇÕES
Hipótese Nula	H0: A1 = A2 = 0	H_0: $AB_{ij} = 0$
Hipótese Alternativa	H1: A1 ≠ A2	H1: $AB_{ij} \neq 0$

Tabela 3.2 – Hipóteses presentes em uma ANOVA.

Para os dois Tratamentos, a aceitação de H0 significa que os efeitos na resposta obtidos com os dois níveis dos fatores A e B são iguais; a Hipótese Alternativa, que os tratamentos são diferentes. Na Interação, a aceitação de H_0 indica que a Interação não é significativa.

	FATOR B				
FATOR A	**NÍVEL**	**1**	**2**	**...**	***b***
	1	$Y_{111}, y_{112},...,y_{11n}$	$Y_{121}, y_{122},..,y_{12n}$		$Y_1b_1, y_1b_2,...,y_1\beta n$
	2	$Y_{211}, y_{212},...,y_{21n}$	$Y_{221}, y_{222},..,y_{22n}$		$Y_2b_1, y_2b_2,...,y_2bn$
	...	...	...	...	...
	A	$Ya_{11}, ya_{12},...,y\alpha_{1n}$	$Ya_{21}, ya_{22},..,y\alpha_{2n}$		$Yab_1, yab_2,...,yabn$

Quadro 8.1 – Arranjo Geral para Fatoriais em dois níveis.

Para expressar genericamente um arranjo geral de dois tratamentos (Fatores), definir-se-á uma simbologia a ser adotada. Antes, porém, deve ser ressaltado que cada Fator A pode possuir ***a*** níveis (..., -2, -1, 0, 1, 2...), que cada Fator B, ***b*** níveis (..., -2, -1, 0, 1, 2...), e que ainda, um determinado tratamento pode possuir ***n*** réplicas (observações). Um arranjo geral poderia ser construído como sugere o ***quadro 8.2***, com as seguintes formulações para seus elementos:

Soma de cada linha (Y_{SL})	$= \sum_{j=1}^{b} \sum_{k=1}^{n} y_{ijk}$	Média de cada linha (Y_{ML})	$= \frac{1}{bn} \sum_{j=1}^{b} \sum_{k=1}^{n} y_{ijk}$
Soma de cada coluna (Y_{SC})	$= \sum_{i=1}^{a} \sum_{k=1}^{n} y_{ijk}$	Média de cada coluna (Y_{MC})	$= \frac{1}{an} \sum_{i=1}^{a} \sum_{k=1}^{n} y_{ijk}$
Soma dentro de cada célula (Y_{SCEL})	$= \sum_{k=1}^{n} y_{ijk}$	Média dentro de cada célula (Y_{MCEL})	$= \frac{1}{n} \sum_{k=1}^{n} y_{ijk}$
Soma Geral (Y_{ST})	$= \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n} y_{ijk}$	Média Geral (Y_{MT})	$= \frac{1}{abn} \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n} y_{ijk}$

Quadro 8.2 – Fórmulas para o Cálculo das Somas de Quadrados de Anova para 2 fatores.

Com estes somatórios, pode-se escrever a *Soma de Quadrados Total* para o modelo:

$$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk}-y_{MT})^2=\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}\left[(y_{ML}-y_{MT})+(y_{MC}-y_{MT})+(y_{MCEL}-y_{ML}-y_{MC}-y_{MT})+(y_{ijk}-y_{MT})\right]^2=$$

$$=bn\sum_{i=1}^{a}(y_{ML}-y_{MT})^2+an\sum_{j=1}^{b}(y_{MC}-y_{MT})^2+n\sum_{i=1}^{a}\sum_{j=1}^{b}(y_{MCEL}-y_{ML}-y_{MC}+y_{MT})+\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk}-y_{MC})^2$$

Dessa maneira, a Soma de Quadrados Total do arranjo pode ser dividida em somas de quadrados devido aos tratamentos de linha (Fator A) SS_A, aos tratamentos de coluna (Fator B) SS_B; a soma de quadrados devido à interação entre A e B, SS_{AB} e a soma de quadrados devido ao erro experimental SS_E.

Desse modo, pode-se escrever que: $SS_T = SS_A + SS_B + SS_{AB} + SS_E$

O ***quadro 8.3*** apresenta os cálculos necessários para se proceder a Análise de Variância para dois fatores. Ressalte-se que o número de graus de liberdade em uma soma de quadrados é igual ao número de elementos independentes naquela soma.

As equações das somas de quadrados podem ainda ser escritas de maneira simplificada, como:

$$SS_T=\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}y_{ijk}^2-\frac{y_{st}^{\ 2}}{abn},\quad SS_A=\frac{1}{bn}\sum_{i=1}^{a}y_{sl}^2-\frac{y_{st}^{\ 2}}{abn},\quad SS_B=\frac{1}{an}\sum_{j=1}^{b}y_{sc}^2-\frac{y_{st}^{\ 2}}{abn}$$

Para se obter uma simplificação da Soma de Quadrados referentes à interação é aconselhável calcular-se primeiramente a soma de quadrados entre os totais das células ab, o que se pode chamar de Soma de Quadrados Parcial, tal que:

$$SS_p=\frac{1}{n}\sum_{i=1}^{a}\sum_{j=1}^{b}y_{scel}^2-\frac{y_{st}^{\ 2}}{abn}$$

Em seguida, subtrai-se *SSA* e *SSB* de *SSP*, de modo que:

$$SS_{AB} = SS_P - SS_A - SS_B$$

Assim, pode-se encontrar a parcela de variação devido ao erro, tal que:

$$SS_E = SS_T - SS_{AB} - SS_A - SS_B$$

A Média Quadrática é o quociente entre a soma de quadrados e o número de graus de liberdade associados, respectivamente, a cada fonte de variação.

Fonte de Variação	Soma de Quadrados	Graus de liberdade	Média Quadrática	Estatística de Teste (F)
Tratamento A	SS_A	*a-1*	$MS_A = \frac{\boldsymbol{SS_A}}{\boldsymbol{a-1}}$	$F_0 = \frac{MS_A}{MS_E}$
Tratamento B	SS_B	*b-1*	$MS_B = \frac{\boldsymbol{SS_A}}{\boldsymbol{b-1}}$	$F_0 = \frac{MS_B}{MS_E}$
Interação AB	SS_{AB}	*(a-1) (b-1)*	$MS_{AB} = \frac{\boldsymbol{SS_A}}{\boldsymbol{(a-1)(b-1)}}$	$F_0 = \frac{MS_{AB}}{MS_E}$
Erro	SS_E	*ab (n-1)*	$MS_A = \frac{\boldsymbol{SS_E}}{\boldsymbol{ab(n-1)}}$	
Total	SS_T	*abn-1*		

Quadro 8.3 – Tabela de ANOVA.

F_o representa a estatística de teste para a análise de dois grupos de dados. Utilizando-se os dados do exemplo, obtêm-se os seguintes resultados:

$$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} y_{ijk}^2 - \frac{y_{st}^{\ 2}}{abn} = (130)^2 + (155)^2 + (74)^2 + ... + (115)^2 - \left(\frac{1870}{2x2x4}\right)^2 = 31868$$

$$SS_A = \frac{1}{bn}\sum_{i=1}^{a} y_{sl}^2 - \frac{y_{st}^{\ 2}}{abn} = \frac{1}{2x4}\left[(768)^2 + (1102)^2\right] - \left(\frac{1870}{2x2x4}\right)^2 = 6972,25$$

$$SS_B = \frac{1}{an}\sum_{j=1}^{b} y_{sc}^2 - \frac{y_{st}^{\ 2}}{abn} = \frac{1}{2x4}\left[(708)^2 + (1162)^2\right] - \left(\frac{1870}{2x2x4}\right)^2 = 12885,25$$

$$SS_A + SS_B = 6972,25 + 12885,25 = 19854,50$$

$$SS_{AB} = \frac{1}{n}\sum_{i=1}^{a}\sum_{j=1}^{b} y_{scel}^2 - \frac{y_{st}^{\ 2}}{abn} - SS_A - SS_B$$

$$SS_{AB} = \frac{1}{4}\left[(539)^2 + (623)^2 + (229)^2 + (479)^2\right] - 19854,25 - 218.556,25 = 1722,25$$

$SS_E = SS_T - SS_{AB} - SS_A - SS_B = 31868 - 19854{,}50 - 1722{,}25 = 10291{,}00$

Assim, pode-se compor a seguinte tabela de Análise de Variância:

Fonte	**SS**	**GL**	**Média Quadrática (MS)**	**F_o**
Tratamentos	19854,50	2	$MS_{AB} = \frac{19854{,}5}{2} = 9928{,}75$	$F_0 = \frac{9928{,}75}{857{,}58} = 11{,}58$
Interação AB	1722,25	1	$MS_{AB} = 1722{,}25$	$F_0 = \frac{1722{,}25}{857{,}58} = 2{,}01$
Erro	10291,00	12	$MS_E = \frac{10291{,}00}{12} = 857{,}58$	
Total	31867,50	15		

Tabela 8.2 – Tabela de ANOVA para os dados da tabela 8.1.

Da tabela para a Distribuição F da seção de anexos, com o nível de significância $\alpha = 0{,}05$, com os graus de liberdade do numerador e do denominador das duas fontes de Variação (Tratamentos e Interação), tem-se que:

a) $F_{(2,\ 12,\ alfa=0{,}05)} = 3{,}89 << 11{,}58$, portanto, rejeita-se a hipótese nula (de que os Efeitos dos dois tratamentos são iguais). Logo, aceita-se a hipótese alternativa H1 de que os efeitos são distintos. O *P-Value* também fornece essa conclusão. O *P-Value* é a probabilidade de se obter um valor da estatística amostral de teste no mínimo tão extremo como o que resulta dos dados amostrais, na suposição de a hipótese nula ser verdadeira. No presente exemplo, *P-value* é igual a 0,00158, valor bem menor que o nível de significância do teste, que é alfa = 0,05.

b) $F_{(1,\ 12,\ alfa=0{,}05)} = 4{,}75 > 2{,}01$, portanto, deve-se aceitar Ho, ou seja, a interação entre os tratamentos A e B não é estatisticamente significativa. O *P-Value* fornece a mesma conclusão (P-Value=0,182 >> alfa =0,05).

A Análise de Variância com mais de dois fatores é um tanto complexa, no entanto, não muito diferente do procedimento demonstrado neste item.

8.2 – Análise de Resíduos

A análise de variância considera que as observações analisadas sigam distribuição, com variância constante. Isto, em geral, é verificado através da análise dos resíduos (ou erros) calculados entre as observações e a média do grupo ao qual elas pertencem. Um resíduo é a diferença entre uma observação y_{ij} e o seu valor estimado (ou ajustado) a partir do modelo estatístico estudado, $\hat{y}_{ij}$. Para que o modelo de análise seja válido, os resíduos devem ser aleatórios, normais, independentes (não correlacionados) e identicamente distribuídos.

A ***tabela 8.3*** apresenta os resíduos para os dados da ***tabela 8.1***.

RESÍDUOS	**TEMP**	**PRESSÃO**					
		25 PSI			**35 PSI**		
	40°C	-4,75	20,25	$\hat{y}_{ijk}$ **= 134,75**	-23,25	-17,25	$\hat{y}_{ijk}$ **= 57,25**
		-60,75	45,25		22,75	17,75	
	60°C	-5,75	32,25	$\hat{y}_{ijk}$ **= 155,75**	16,25	-13,75	$\hat{y}_{ijk}$ **= 119,75**
		3,25	-29,75		2,25	-4,75	

Tabela 8.3 – Resíduos para cada tratamento.

Se um modelo está correto, os resíduos devem estar desestruturados; em particular, eles não devem estar correlacionados com qualquer variável incluída na resposta predita.

Uma anomalia que ocasionalmente ocorre neste tipo de gráfico é o da *Variação não constante*. Às vezes, a variância das observações aumenta com o incremento da magnitude das observações. Isto viola o princípio da homogeneidade da variância. Os gráficos que representam a heterocedasticidade apresentarão resíduos dispostos em forma *funil* ou de *borboleta*, o que caracteriza um padrão de tendência.

Variância não constante também ocorre quando os dados não são normalmente distribuídos, seguindo, portanto, distribuições assimétricas. Nesse tipo de distribuição a

variância tende a ser função da média. Nota-se, contudo, que o gráfico do exemplo (***Figura 8.5***) não apresenta este tipo de anormalidade.

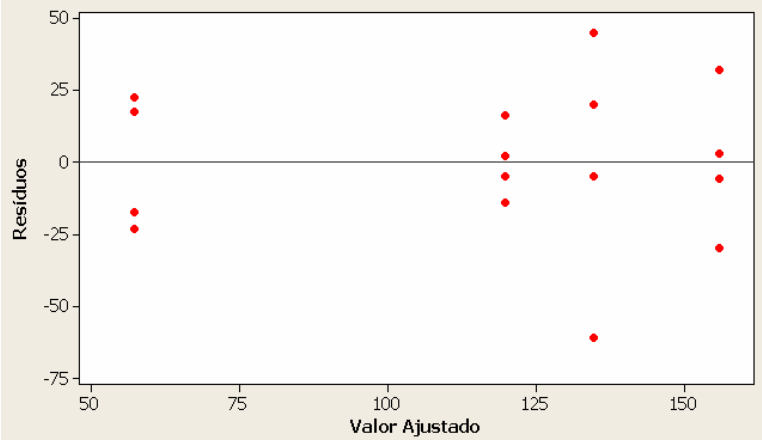


Figura 8.5 – Resíduos versus Média.

Os resíduos devem apresentar comportamentos aleatórios quando dispostos em uma ordem temporal. Se qualquer padrão não-aleatório for identificado nestes tipos de gráficos, uma transformação de dados (Log, Raiz Quadrada, Box-Cox) na resposta Y deve ser realizada. A tendência de se ter resíduos positivos ou negativos indicam uma correlação positiva, o que implicaria na violação da hipótese de independência dos resíduos. Uma aleatorização apropriada do experimento é um bom procedimento para a garantia dessa independência.

Algumas vezes, a habilidade do experimentador em executar o experimento pode mudar com o progresso da experimentação, ou, conquanto o processo esteja sendo estudado constantemente, pode se tornar mais inconsistente. Isto sempre resultará em uma mudança no erro da variância ao longo do tempo. O gráfico que surge dessa situação revela uma relação resíduo versus ordem com mais variação do que as outras.

A utilização dos *Resíduos Padronizados* gera mais informação sobre a qualidade da resposta do que os resíduos ordinários. Os *Resíduos Padronizados (d)* podem ser obtidos dividindo-se os resíduos das observações em relação à média das réplicas (*FIT*) pelo desvio padrão das réplicas (*SE Fit*). Deste modo, tais resíduos possuirão média igual a zero e desvio padrão unitário. Conseqüentemente, eles serão úteis na identificação de ***outliers***. A maior parte dos resíduos padronizados deve encontrar-se no intervalo $-3 \leq d \leq 3$, e qualquer observação com resíduo padronizado fora deste intervalo pode representar uma resposta observada potencialmente não usual ou

incorreta. Estes *outliers* devem ser examinados com cuidado: tanto pode indicar erros de registro, quanto algo mais preocupante.

Em relação ao exemplo em questão, não há evidências de padrões não-aleatórios presentes (***Figura 8.6***).

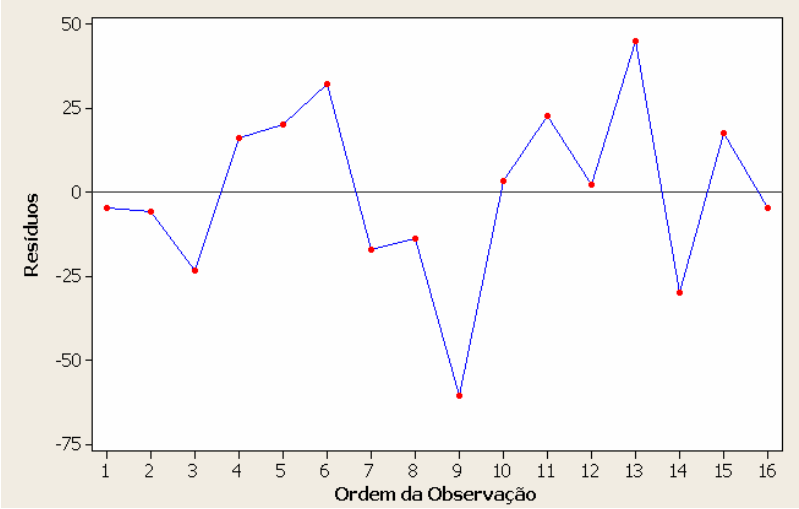


Figura 8.6 – Gráfico dos Resíduos Padronizados versus tempo (ordem da observação).

O gráfico da figura 8.7 representa o ajuste dos resíduos a uma Distribuição Normal e serve como indicador de sua normalidade.

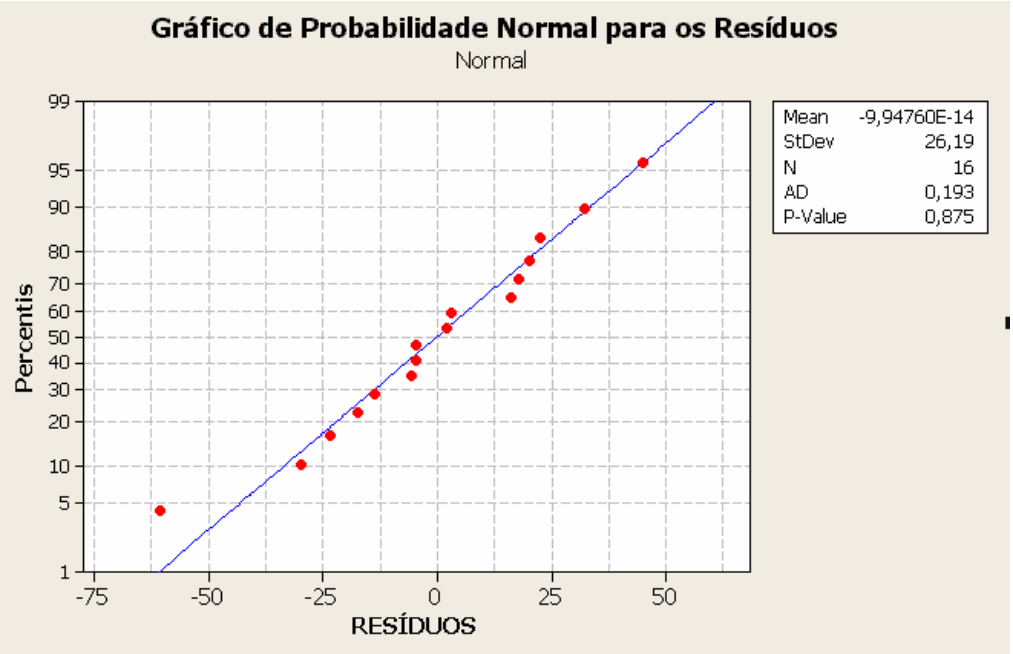


Figura 8.7 – Gráfico de Probabilidade Normal para os Resíduos.

Para o exemplo dado, a direção dos pontos parece constante, assemelhando-se a uma reta. Quanto maior for a concentração de pontos sob a reta na sua região central, com uma diminuição para as extremidades, mais ajustados estarão os resíduos.

Esse gráfico é construído dispondo-se os resíduos em ordem crescente sob o eixo das abscissas, enquanto que a variável reduzida Z é escrita sobre o eixo das ordenadas.

Uma outra maneira de verificar a normalidade dos resíduos é proceder ao teste de normalidade, disponível em muitos pacotes estatísticos computacionais. Um valor de *P-Value* maior do que o nível de significância de 0,05 (5%) indica que os resíduos são normais. No presente exemplo, *P-Value* é igual a 0,875, que é bem maior que o nível de significância. Portanto, não há problemas com os resíduos.

Respostas de Alguns Exercícios Selecionados

Exercícios I

(03) a)sim b)sim c)sim d)sim e)não f)sim

(07) s=10

Exercícios II

(02) $b)\bar{x} = 63.25, \quad s = 16.11, \quad \tilde{x} = 65.63, \quad Q_1 = 55, \quad P_{65} = 70.42$

(04) Q_1=107 Q_3=132 Q_2=120

(06) Uma solução: A={10,11,12,14,25}, B={30,33,36,39,42,45}

(07) 70

(08) R$2235,00

(09) 106

(10) A amplitude ultrapassa o valor alegado

(12)$\bar{x} = 8.38$ $\tilde{x} = 9$ não há moda

(13) a)$\bar{x}$=28.03 s=7.28 b) Q_1=37.5 Q_3=112.3 Q_2=75.5

c)26.9 d)11.4

(14) Aparentemente, o primeiro é mais preciso; mas a comparação pode estar prejudicada pelo fato de terem sido medidos objetos diferentes

(17) b) $\bar{x}$=22.7 s=6.02 c) 16.7%

(18) a) $\bar{x}$=28.92 b)s=2.31

Exercícios III

(01) a)

	<10	10 a 20	>20	Total
Solteiro	0.12	0.19	0.09	0.40
Casado	0.08	0.31	0.21	0.60
Total	0.20	0.50	0.30	1.00

b) não,

(02) a

33	58	70	50
67	42	30	50
100	100	100	100

b) não c)

Exercícios IV

(01) a) b)

(02) b) c) e)

(03) 0.25

(04) 1/6

(05) 2/9

(06) a) 0.986 b) 0.982

(07) 1/12

(08)

(09) a) 0.72 b) 0.98

(10) sim

(11) 3

(12) a) 0.54 b) 0.04 c) 0.42

(13) a) 0.9978 b) 0.0112

(15) 0.63

(16) a) 1/12 b) 13/36 c) 6/13

(17) 0.396

(18) a)

(19)

(20) Um cartão com 10 dezenas

(21) 2.5%

(23) 97.5%

(24) 0.8%

(25) 2.35%

(26) a) 0.84% b) 0.04%

(27) 13.63%

(28) 32.25%

(29) a) 99.5% b) 14.5% c) 96.5%

Exercícios V

(01) b) 26/27 c) 8/27

(02) F(b-1) - F(a) para o caso discreto

(03)

(04) a) não b) não c) sim d) não

(05)

(06) b) $F(x)=3x^2 -2x^3$ c)

(07) a) $f(x)=x^4(5-4x)$ b) c) $0.63C_2 + 0.42C_1 -C_3$

Exercícios VI

(01) -R$0.25

(02) 1.7

(05) E(G) = 2.75 e VAR(G) = 0.412

(06) U$ 25

(07) 36.78%

Exercícios VIII

(01) P(X=6)= 0.296

(02) a) 0.251 b) 0.126 c) 0.363 d) 0.858

(03) 0.0819

(04) a) 0.0858 b) 0.126

(05) 0.1198

(06) a) 0.38 b) 0.3455

(07) 0.25

(08) a) 1312 b) 562 c) 188

(09) a) 1 b) 0.9267

(10) a) 0.107

Exercícios IX

(01) a) 22.4% b) 14.9% c) 57.7%

(02) 18.13%

(03) 1.4%

(04) 80.9%

(05) 20.76%

(06) 5.6%

(07) 0.89%

(08) 4.7%

(09) 5/14

Exercícios X

(01) 40%

(02) 20%

(03) R$25.000,00

(04) a) 60.65%

(05) 45.11%

Exercícios XI

(01) a) 109.70 b) 86.56

(02) 5 anos

(03) a) 0.26% b) 7.10^{-6} c) 0.9948

(04) 63

(05) a) 263 b) 20 c) 0

(06) a) 19.36% b) 39.7%

(07) 734

(09) a) 0.582 b) 0.7422

(10) 0.21

(11) a) 12.92% b) 16.35%

(12) a) 86.43% b) 29.78% c) 7.96%

(13) a) 0.12 b) 0.9535 c) 0.8790

(14) 25.14%

(15) 0.1335

Exercícios XII

(01) a) 68% b) 100% d) 4.

(02) a) 512,80 b) 0,52%

(03) a) 5,82% b) 53,28%

(04) a) 35,20% b) 0,05%

(05) a) 2,27%

(06) a) 4,75% b) 50%

(07) a) 26,44% b) 16,01%

(08) a) 31,25% b) 0,97 c) 95

Exercícios XIII

(01) 23%

(02) 81

(03) a) 0 e 400/*n* b) 61,70% c) 31,74%

d) 5,16% e) 1537

(05) 0,06%

(08) $E\left(\hat{p}_1\right) = p$ $Var\left(\hat{p}_1\right) = \dfrac{p(1-p)}{n}$ $E\left(\hat{p}_2\right) = p$ $Var\left(\hat{p}_2\right) = p(1-p)$

(09) 170 ± 2,94 165 ± 3,18 180 ± 2,08

(10) a)]787,1; 812,9[b) 16% c) 625

(11) a) 384 b) 666

(12) a) 307 b)]49; 51[

(13)]0,677; 0,732[

(14)]0,25; 0,35[,]0,255; 0,345[

(15) a) 3932 b)]0,5345; 0,5655[

(16) a)]0,274; 0,386[b) 2401

(17) a)]148,37; 151,63[

(18) a)]5,13; 7,31[b) 107.584

(19) 400 ± 20,46

(20) a)]0,543; 0,657[c) 3.841.600

(21) 0,520 ± 0,049

(22) 510,6 ± 0,784

Exercícios XIV

(02) a) 6,314 b) 1,440 c) -1,812 d) 2,131 e) -0,860 f) 3,160

(04) a) 9,342 b) 36,191 c) 29,615 e) 5,412 f) 9,524

(06) a) 9,55 b) 0,105 c) 3,84 d) 1,35 e) 0,42 f) 1,8275

Exercícios XV

(02) a) 9,18% b) 6,68% c) *RC* = [1171; ∞[

(04) $\alpha = 1/8$

Exercícios XVII

(02) Não, pois z_0 = -1,5 e RC =] - ∞; - 1,645]

(03) Sim, pois z_0 = -2,8 e RC =] - ∞; - 1,645]

(06) Rejeita-se a hipótese p = 0,20, pois Z_0 = 1,36 e RC =]1,28; ∞[

(09) 27,43%

(15) H_0 não é rejeitado, pois z_0 = 1,94 e RC' = [- 1,96; 1,96]

(18) a) 0,3% b) 87,2%

(22) a) 41,56 ± 5,51

Exercícios XVIII

(01) São iguais, pois F_0 = 2,56

Referência Bibliográfica

1) Meyer, P. L., *Probabilidade: Aplicações à Estatística.*Livros Técnicos e Científicos Editora S.A., 1978

2) Costa Neto, P. L., *Estatística.*Edgard Blücher Ltda, 1992

3) Soares, J.F.; Farias, A.A.; Cesar, C.C.; *Introdução à Estatística.* Guanabara Koogan, 1991

4) Soong, T.T., *Modelos Probabilísticos em Engenharia e Ciências.* .Liv. Téc.e Científicos Editora S.A., 1986

5) Moretin , Bussab, *Estatística Básica.*Editora Atlas, 1992

6) Hoel, P.G. , *Introduction to Mathematical Statistics.* John Wiley, New York, 1962

7) Gibbons, J.D., *Nonparametric Statistical Inference.* McGraw-Hill, New York, 1971

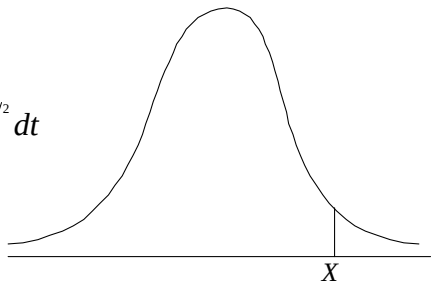
8) Walker, H. M., *Statistical Inference*. Henry Holt, New York, 1953

9) Bowker, A. H., Lieberman, G. J., *Engineering Statistics.* Prentice-Hall, New Jersey, 1972

10) Cochran, W. G., *Técnicas de Amostragem.* Fundo de Cultura, Rio de Janeiro, 1965

Tabela 3 - Distribuição normal

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^{2/2}} dt$$



x	**Φ(*x*)**	***x***	**Φ(*x*)**	***x***	**Φ(*x*)**	***x***	**Φ(*x*)**	***x***	**Φ(*x*)**
-3,10	0,000968	-2,54	0,005543	-1,98	0,023852	-1,42	0,077804	-0,86	0,194894
-3,09	0,001001	-2,53	0,005703	-1,97	0,024419	-1,41	0,079270	-0,85	0,197663
-3,08	0,001035	-2,52	0,005868	-1,96	0,024998	-1,40	0,080757	-0,84	0,200454
-3,07	0,001070	-2,51	0,006037	-1,95	0,025588	-1,39	0,082264	-0,83	0,203269
-3,06	0,001107	-2,50	0,006210	-1,94	0,026190	-1,38	0,083793	-0,82	0,206108
-3,05	0,001144	-2,49	0,006387	-1,93	0,026803	-1,37	0,085344	-0,81	0,208970
-3,04	0,001183	-2,48	0,006569	-1,92	0,027429	-1,36	0,086915	-0,80	0,211855
-3,03	0,001223	-2,47	0,006756	-1,91	0,028067	-1,35	0,088508	-0,79	0,214764
-3,02	0,001264	-2,46	0,006947	-1,90	0,028717	-1,34	0,090123	-0,78	0,217695
-3,01	0,001306	-2,45	0,007143	-1,89	0,029379	-1,33	0,091759	-0,77	0,220650
-3,00	0,001350	-2,44	0,007344	-1,88	0,030054	-1,32	0,093418	-0,76	0,223627
-2,99	0,001395	-2,43	0,007549	-1,87	0,030742	-1,31	0,095098	-0,75	0,226627
-2,98	0,001441	-2,42	0,007760	-1,86	0,031443	-1,30	0,096801	-0,74	0,229650
-2,97	0,001489	-2,41	0,007976	-1,85	0,032157	-1,29	0,098525	-0,73	0,232695
-2,96	0,001538	-2,40	0,008198	-1,84	0,032884	-1,28	0,100273	-0,72	0,235762
-2,95	0,001589	-2,39	0,008424	-1,83	0,033625	-1,27	0,102042	-0,71	0,238852
-2,94	0,001641	-2,38	0,008656	-1,82	0,034379	-1,26	0,103835	-0,70	0,241964
-2,93	0,001695	-2,37	0,008894	-1,81	0,035148	-1,25	0,105650	-0,69	0,245097
-2,92	0,001750	-2,36	0,009137	-1,80	0,035930	-1,24	0,107488	-0,68	0,248252
-2,91	0,001807	-2,35	0,009387	-1,79	0,036727	-1,23	0,109349	-0,67	0,251429
-2,90	0,001866	-2,34	0,009642	-1,78	0,037538	-1,22	0,111233	-0,66	0,254627
-2,89	0,001926	-2,33	0,009903	-1,77	0,038364	-1,21	0,113140	-0,65	0,257846
-2,88	0,001988	-2,32	0,010170	-1,76	0,039204	-1,20	0,115070	-0,64	0,261086
-2,87	0,002052	-2,31	0,010444	-1,75	0,040059	-1,19	0,117023	-0,63	0,264347
-2,86	0,002118	-2,30	0,010724	-1,74	0,040929	-1,18	0,119000	-0,62	0,267629
-2,85	0,002186	-2,29	0,011011	-1,73	0,041815	-1,17	0,121001	-0,61	0,270931
-2,84	0,002256	-2,28	0,011304	-1,72	0,042716	-1,16	0,123024	-0,60	0,274253
-2,83	0,002327	-2,27	0,011604	-1,71	0,043633	-1,15	0,125072	-0,59	0,277595
-2,82	0,002401	-2,26	0,011911	-1,70	0,044565	-1,14	0,127143	-0,58	0,280957
-2,81	0,002477	-2,25	0,012224	-1,69	0,045514	-1,13	0,129238	-0,57	0,284339
-2,80	0,002555	-2,24	0,012545	-1,68	0,046479	-1,12	0,131357	-0,56	0,287740
-2,79	0,002635	-2,23	0,012874	-1,67	0,047460	-1,11	0,133500	-0,55	0,291160
-2,78	0,002718	-2,22	0,013209	-1,66	0,048457	-1,10	0,135666	-0,54	0,294599
-2,77	0,002803	-2,21	0,013553	-1,65	0,049471	-1,09	0,137857	-0,53	0,298056
-2,76	0,002890	-2,20	0,013903	-1,64	0,050503	-1,08	0,140071	-0,52	0,301532
-2,75	0,002980	-2,19	0,014262	-1,63	0,051551	-1,07	0,142310	-0,51	0,305026
-2,74	0,003072	-2,18	0,014629	-1,62	0,052616	-1,06	0,144572	-0,50	0,308538
-2,73	0,003167	-2,17	0,015003	-1,61	0,053699	-1,05	0,146859	-0,49	0,312067
-2,72	0,003264	-2,16	0,015386	-1,60	0,054799	-1,04	0,149170	-0,48	0,315614
-2,71	0,003364	-2,15	0,015778	-1,59	0,055917	-1,03	0,151505	-0,47	0,319178
-2,70	0,003467	-2,14	0,016177	-1,58	0,057053	-1,02	0,153864	-0,46	0,322758
-2,69	0,003573	-2,13	0,016586	-1,57	0,058208	-1,01	0,156248	-0,45	0,326335
-2,68	0,003681	-2,12	0,017003	-1,56	0,059380	-1,00	0,158655	-0,44	0,329969
-2,67	0,003793	-2,11	0,017429	-1,55	0,060571	-0,99	0,161087	-0,43	0,333598
-2,66	0,003907	-2,10	0,017864	-1,54	0,061780	-0,98	0,163543	-0,42	0,337243
-2,65	0,004025	-2,09	0,018309	-1,53	0,063008	-0,97	0,166023	-0,41	0,340903
-2,64	0,004145	-2,08	0,018763	-1,52	0,064256	-0,96	0,168528	-0,40	0,344578
-2,63	0,004269	-2,07	0,019226	-1,51	0,065522	-0,95	0,171056	-0,39	0,348268
-2,62	0,004396	-2,06	0,019699	-1,50	0,066807	-0,94	0,173609	-0,38	0,351973
-2,61	0,004527	-2,05	0,020182	-1,49	0,068112	-0,93	0,176186	-0,37	0,355691
-2,60	0,004661	-2,04	0,020675	-1,48	0,069437	-0,92	0,178786	-0,36	0,359424
-2,59	0,004799	-2,03	0,021178	-1,47	0,070781	-0,91	0,181411	-0,35	0,363169
-2,58	0,004940	-2,02	0,021692	-1,46	0,072145	-0,90	0,184060	-0,34	0,366928
-2,57	0,005085	-2,01	0,022216	-1,45	0,073529	-0,89	0,186733	-0,33	0,370700
-2,56	0,005234	-2,00	0,022750	-1,44	0,074934	-0,88	0,189430	-0,32	0,374484
-2,55	0,005386	-1,99	0,023295	-1,43	0,076359	-0,87	0,192150	-0,31	0,378281

Tabela 3 - Distribuição normal (continuação)

x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$
-0,30	0,382089	0,38	0,648027	1,06	0,855428	1,74	0,959071	2,42	0,992240
-0,29	0,385908	0,39	0,651732	1,07	0,857690	1,75	0,959941	2,43	0,992451
-0,28	0,389739	0,40	0,655422	1,08	0,859929	1,76	0,960796	2,44	0,992656
-0,27	0,393580	0,41	0,659097	1,09	0,862143	1,77	0,961636	2,45	0,993857
-0,26	0,397432	0,42	0,662757	1,10	0,868334	1,78	0,962462	2,46	0,993053
-0,25	0,401294	0,43	0,666402	1,11	0,866500	1,79	0,963273	2,47	0,993244
-0,24	0,405165	0,44	0,670031	1,12	0,868643	1,80	0,964070	2,48	0,993431
-0,23	0,409046	0,45	0,673645	1,13	0,870762	1,81	0,964852	2,49	0,993613
-0,22	0,412936	0,46	0,677242	1,14	0,872857	1,82	0,965621	2,50	0,993790
-0,21	0,416834	0,47	0,680822	1,15	0,874928	1,83	0,966375	2,51	0,994963
-0,20	0,420740	0,48	0,684386	1,16	0,876976	1,84	0,967116	2,52	0,994132
-0,19	0,424655	0,49	0,687933	1,17	0,878999	1,85	0,967843	2,53	0,994297
-0,18	0,428576	0,50	0,691462	1,18	0,881000	1,86	0,968557	2,54	0,994457
-0,17	0,432505	0,51	0,694974	1,19	0,882977	1,87	0,969258	2,55	0,994614
-0,16	0,436441	0,52	0,698468	1,20	0,884930	1,88	0,969946	2,56	0,994766
-0,15	0,440382	0,53	0,701944	1,21	0,886860	1,89	0,970621	2,57	0,995915
-0,14	0,444330	0,54	0,705401	1,22	0,888767	1,90	0,971283	2,58	0,995060
-0,13	0,448283	0,55	0,708840	1,23	0,890651	1,91	0,971933	2,59	0,995201
-0,12	0,452242	0,56	0,712260	1,24	0,892512	1,92	0,972571	2,60	0,995339
-0,11	0,456205	0,57	0,715661	1,25	0,894350	1,93	0,973197	2,61	0,995473
-0,10	0,460172	0,58	0,719043	1,26	0,896165	1,94	0,973810	2,62	0,995604
-0,09	0,464144	0,59	0,722405	1,27	0,897958	1,95	0,974412	2,63	0,995731
-0,08	0,468119	0,60	0,725747	1,28	0,899727	1,96	0,975002	2,64	0,995855
-0,07	0,472097	0,61	0,729069	1,29	0,901475	1,97	0,975581	2,65	0,996975
-0,06	0,476078	0,62	0,732371	1,30	0,903199	1,98	0,976148	2,66	0,996093
-0,05	0,480061	0,63	0,735653	1,31	0,904902	1,99	0,976705	2,67	0,996207
-0,04	0,484047	0,64	0,738914	1,32	0,906582	2,00	0,977250	2,68	0,996319
-0,03	0,488034	0,65	0,742154	1,33	0,908241	2,01	0,977784	2,69	0,996427
-0,02	0,492022	0,66	0,745373	1,34	0,909877	2,02	0,978308	2,70	0,996533
-0,01	0,496011	0,67	0,748571	1,35	0,911492	2,03	0,978822	2,71	0,996636
0,00	0,500000	0,68	0,751748	1,36	0,913085	2,04	0,979325	2,72	0,996736
0,01	0,503989	0,69	0,754903	1,37	0,914657	2,05	0,979818	2,73	0,996833
0,02	0,507978	0,70	0,758036	1,38	0,916207	2,06	0,980301	2,74	0,996928
0,03	0,511967	0,71	0,761148	1,39	0,917736	2,07	0,980774	2,75	0,997020
0,04	0,515953	0,72	0,764238	1,40	0,919243	2,08	0,981237	2,76	0,997110
0,05	0,519939	0,73	0,767305	1,41	0,920730	2,09	0,981691	2,77	0,997197
0,06	0,523922	0,74	0,770350	1,42	0,922196	2,10	0,982136	2,78	0,997282
0,07	0,527903	0,75	0,773373	1,43	0,923641	2,11	0,982571	2,79	0,997365
0,08	0,531881	0,76	0,776373	1,44	0,925066	2,12	0,982997	2,80	0,997445
0,09	0,535856	0,77	0,779350	1,45	0,926471	2,13	0,983414	2,81	0,997523
0,10	0,539828	0,78	0,782305	1,46	0,927855	2,14	0,983823	2,82	0,997599
0,11	0,543795	0,79	0,785236	1,47	0,929219	2,15	0,984222	2,83	0,997673
0,12	0,547758	0,80	0,788145	1,48	0,930563	2,16	0,984614	2,84	0,997744
0,13	0,551717	0,81	0,791030	1,49	0,931888	2,17	0,984997	2,85	0,997814
0,14	0,555670	0,82	0,793892	1,50	0,933193	2,18	0,985371	2,86	0,997882
0,15	0,559618	0,83	0,796731	1,51	0,934478	2,19	0,985738	2,87	0,997948
0,16	0,563559	0,84	0,799546	1,52	0,935744	2,20	0,986097	2,88	0,998012
0,17	0,567495	0,85	0,802337	1,53	0,936992	2,21	0,986447	2,89	0,998074
0,18	0,571424	0,86	0,805106	1,54	0,938220	2,22	0,986791	2,90	0,998134
0,19	0,575345	0,87	0,807850	1,55	0,939429	2,23	0,987126	2,91	0,998193
0,20	0,579260	0,88	0,810570	1,56	0,940620	2,24	0,987455	2,92	0,998250
0,21	0,583166	0,89	0,813267	1,57	0,941792	2,25	0,987776	2,93	0,998305
0,22	0,587064	0,90	0,815940	1,58	0,942947	2,26	0,988089	2,94	0,998359
0,23	0,590954	0,91	0,818589	1,59	0,944083	2,27	0,988396	2,95	0,998411
0,24	0,594835	0,92	0,821214	1,60	0,945201	2,28	0,988696	2,96	0,998462
0,25	0,598706	0,93	0,823814	1,61	0,946301	2,29	0,988989	2,97	0,998511
0,26	0,602568	0,94	0,826391	1,62	0,947384	2,30	0,989276	2,98	0,998559
0,27	0,606420	0,95	0,828944	1,63	0,948449	2,31	0,989556	2,99	0,998605
0,28	0,610261	0,96	0,831472	1,64	0,949497	2,32	0,989830	3,00	0,998650
0,29	0,614092	0,97	0,833977	1,65	0,950529	2,33	0,990097	3,01	0,998694
0,30	0,617911	0,98	0,836457	1,66	0,951543	2,34	0,990358	3,02	0,998736
0,31	0,621719	0,99	0,838913	1,67	0,952540	2,35	0,990613	3,03	0,998777
0,32	0,625516	1,00	0,841345	1,68	0,953521	2,36	0,990863	3,04	0,998817
0,33	0,629300	1,01	0,843752	1,69	0,954486	2,37	0,991106	3,05	0,998856
0,34	0,633072	1,02	0,846136	1,70	0,955435	2,38	0,991344	3,06	0,998893
0,35	0,636831	1,03	0,848495	1,71	0,956367	2,39	0,991576	3,07	0,998930
0,36	0,640576	1,04	0,850830	1,72	0,957284	2,40	0,991802	3,08	0,998965
0,37	0,644309	1,05	0,853141	1,73	0,958185	2,41	0,992024	3,09	0,998999

Tabela 2 - Probabilidades de Poisson acumuladas

$$\sum_{x=0}^{r} \frac{e^{-\lambda}\lambda^{x}}{x!}$$

	λ													
r	**0,1**	**0,2**	**0,3**	**0,4**	**0,5**	**0,6**	**0,7**	**0,8**	**0,9**	**1,0**	**1,1**	**1,2**	**1,3**	**1,4**
0	0,9048	0,8187	0,7408	0,6703	0,6065	0,5488	0,4965	0,4493	0,4065	0,3679	0,3329	0,3012	0,2725	0,2466
1	0,9953	0,9824	0,9630	0,9384	0,9098	0,8781	0,8442	0,8087	0,7724	0,7358	0,6990	0,6626	0,6268	0,5918
2	0,9998	0,9988	0,9964	0,9920	0,9856	0,9768	0,9658	0,9525	0,9371	0,9197	0,9004	0,8795	0,8571	0,8335
3	1,0000	0,9999	0,9997	0,9992	0,9982	0,9966	0,9942	0,9909	0,9865	0,9810	0,9743	0,9662	0,9569	0,9463
4		1,0000	0,9999	0,9999	0,9998	0,9996	0,9992	0,9985	0,9976	0,9963	0,9946	0,9923	0,9893	0,9857
5			1,0000	1,0000	0,9999	0,9999	0,9999	0,9998	0,9996	0,9994	0,9990	0,9985	0,9978	0,9968
6					1,0000	1,0000	0,9999	0,9999	0,9999	0,9999	0,9999	0,9997	0,9996	0,9994
7							1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9999
8													1,0000	1,0000

	λ													
r	**1,5**	**1,6**	**1,7**	**1,8**	**1,9**	**2,0**	**2,1**	**2,2**	**2,3**	**2,4**	**2,5**	**2,6**	**2,7**	**2,8**
0	0,2231	0,2019	0,1827	0,1653	0,1496	0,1353	0,1225	0,1108	0,1003	0,0907	0,0821	0,0743	0,0672	0,0608
1	0,5578	0,5249	0,4932	0,4628	0,4337	0,4060	0,3796	0,3546	0,3309	0,3084	0,2873	0,2674	0,2487	0,2311
2	0,8088	0,7834	0,7572	0,7306	0,7037	0,6767	0,6496	0,6227	0,5960	0,5697	0,5438	0,5184	0,4936	0,4695
3	0,9344	0,9212	0,9068	0,8913	0,8747	0,8571	0,8386	0,8194	0,7993	0,7787	0,7576	0,7360	0,7141	0,6919
4	0,9814	0,9763	0,9704	0,9636	0,9559	0,9473	0,9379	0,9275	0,9162	0,9041	0,8912	0,8774	0,8629	0,8477
5	0,9955	0,9940	0,9920	0,9896	0,9868	0,9834	0,9796	0,9751	0,9700	0,9643	0,9580	0,9510	0,9433	0,9349
6	0,9991	0,9987	0,9981	0,9974	0,9966	0,9955	0,9941	0,9925	0,9906	0,9884	0,9858	0,9828	0,9794	0,9756
7	0,9998	0,9997	0,9996	0,9994	0,9992	0,9989	0,9985	0,9980	0,9974	0,9967	0,9958	0,9947	0,9934	0,9919
8	1,0000	1,0000	0,9999	0,9999	0,9998	0,9998	0,9997	0,9995	0,9994	0,9991	0,9989	0,9985	0,9981	0,9976
9			1,0000	1,0000	1,0000	1,0000	0,9999	0,9999	0,9999	0,9998	0,9997	0,9996	0,9995	0,9993
10							1,0000	1,0000	1,0000	1,0000	0,9999	0,9999	0,9999	0,9998
11											1,0000	1,0000	1,0000	1,0000

	λ													
r	**2,9**	**3,0**	**3,1**	**3,2**	**3,3**	**3,4**	**3,5**	**3,6**	**3,7**	**3,8**	**3,9**	**4,0**	**4,1**	**4,2**
0	0,0550	0,0498	0,0450	0,0408	0,0369	0,0334	0,0302	0,0273	0,0247	0,0224	0,0202	0,0183	0,0166	0,0150
1	0,2146	0,1991	0,1847	0,1712	0,1586	0,1468	0,1359	0,1257	0,1162	0,1074	0,0992	0,0916	0,0845	0,0780
2	0,4460	0,4232	0,4012	0,3799	0,3594	0,3397	0,3208	0,3027	0,2854	0,2689	0,2531	0,2381	0,2238	0,2102
3	0,6696	0,6472	0,6248	0,6052	0,5803	0,5584	0,5366	0,5152	0,4942	0,4735	0,4532	0,4335	0,4142	0,3954
4	0,8318	0,8153	0,7982	0,7806	0,7626	0,7442	0,7254	0,7064	0,6872	0,6678	0,6484	0,6288	0,6093	0,5898
5	0,9258	0,9161	0,9057	0,8946	0,8829	0,8705	0,8576	0,8441	0,8301	0,8156	0,8006	0,7851	0,7693	0,7531
6	0,9713	0,9665	0,9612	0,9554	0,9490	0,9421	0,9347	0,9267	0,9182	0,9091	0,8995	0,8893	0,8786	0,8675
7	0,9901	0,9881	0,9858	0,9832	0,9802	0,9769	0,9733	0,9692	0,9648	0,9599	0,9546	0,9489	0,9427	0,9361
8	0,9969	0,9962	0,9953	0,9943	0,9931	0,9917	0,9901	0,9883	0,9863	0,9840	0,9815	0,9786	0,9755	0,9721
9	0,9991	0,9989	0,9986	0,9982	0,9978	0,9973	0,9967	0,9960	0,9952	0,9942	0,9931	0,9919	0,9905	0,9889
10	0,9998	0,9997	0,9996	0,9995	0,9994	0,9992	0,9990	0,9987	0,9984	0,9981	0,9977	0,9972	0,9966	0,9959
11	0,9999	0,9999	0,9999	0,9999	0,9998	0,9998	0,9997	0,9996	0,9995	0,9994	0,9993	0,9991	0,9989	0,9986
12	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9999	0,9999	0,9999	0,9998	0,9998	0,9997	0,9997	0,9996
13						1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9999	0,9999	0,9999
14											1,0000	1,0000	1,0000	1,0000

Tabela 2 - Probabilidades de Poisson acumuladas (continuação)

	λ													
r	**4,3**	**4,4**	**4,5**	**4,6**	**4,7**	**4,8**	**4,9**	**5,0**	**5,1**	**5,2**	**5,3**	**5,4**	**5,5**	**5,6**
0	0,0136	0,0123	0,0111	0,0101	0,0091	0,0082	0,0074	0,0067	0,0061	0,0055	0,0050	0,0045	0,0041	0,0037
1	0,0719	0,0663	0,0611	0,0563	0,0518	0,0477	0,439	0,0404	0,0372	0,0342	0,0314	0,0289	0,0266	0,0244
2	0,1974	0,1851	0,1736	0,1626	0,1523	0,1452	0,1333	0,1247	0,1165	0,1088	0,1016	0,0948	0,0884	0,0824
3	0,3772	0,3594	0,3423	0,3257	0,3097	0,2942	0,2793	0,2650	0,2513	0,2381	0,2254	0,2133	0,2017	0,1906
4	0,5704	0,5512	0,5321	0,5132	0,4946	0,4763	0,4582	0,4405	0,4231	0,4061	0,3895	0,3733	0,3575	0,3422
5	0,7367	0,7199	0,7029	0,6858	0,6684	0,6510	0,6335	0,6160	0,5984	0,5809	0,5635	0,5461	0,5289	0,5119
6	0,8558	0,8436	0,8311	0,8180	0,8046	0,7908	0,7767	0,7622	0,7474	0,7324	0,7171	0,7017	0,6860	0,6703
7	0,9290	0,9214	0,9134	0,9049	0,8960	0,8867	0,8769	0,8666	0,8560	0,8449	0,8335	0,8217	0,8095	0,7970
8	0,9683	0,9642	0,9597	0,9549	0,9497	0,9442	0,9382	0,9319	0,9252	0,9181	0,9106	0,9027	0,8944	0,8857
9	0,9871	0,9851	0,9829	0,9805	0,9778	0,9749	0,9717	0,9682	0,9644	0,9603	0,9559	0,9512	0,9462	0,9409
10	0,9952	0,9943	0,9933	0,9922	0,9910	0,9896	0,9880	0,9863	0,9844	0,9823	0,9800	0,9775	0,9747	0,9718
11	0,9983	0,9980	0,9976	0,9971	0,9966	0,9960	0,9953	0,9945	0,9937	0,9927	0,9916	0,9904	0,9890	0,9875
12	0,9995	0,9993	0,9992	0,9990	0,9988	0,9986	0,9983	0,9980	0,9976	0,9972	0,9967	0,9962	0,9955	0,9949
13	0,9998	0,9998	0,9997	0,9997	0,9996	0,9995	0,9994	0,9993	0,9992	0,9990	0,9988	0,9986	0,9983	0,9980
14	1,0000	0,9999	0,9999	0,9999	0,9999	0,9999	0,9998	0,9998	0,9997	0,9997	0,9996	0,9995	0,9994	0,9993
15		1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9999	0,9999	0,9999	0,9999	0,9998	0,9998	0,9998
16							1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9999	0,9999
17												1,0000	1,0000	1,0000

	λ													
r	**5,7**	**5,8**	**5,9**	**6,0**	**6,1**	**6,2**	**6,3**	**6,4**	**6,5**	**6,6**	**6,7**	**6,8**	**6,9**	**7,0**
0	0,0033	0,0030	0,0027	0,0025	0,0022	0,0020	0,0018	0,0017	0,0015	0,0014	0,0012	0,0011	0,0010	0,0009
1	0,0224	0,0206	0,0189	0,0174	0,0159	0,0146	0,0134	0,0123	0,0113	0,0103	0,0095	0,0087	0,0080	0,0073
2	0,0768	0,0715	0,0666	0,0620	0,0577	0,0536	0,0498	0,0463	0,0430	0,0400	0,0371	0,0344	0,0320	0,0296
3	0,1800	0,1700	0,1604	0,1512	0,1425	0,1342	0,1264	0,1189	0,1118	0,1052	0,0988	0,0928	0,0871	0,0818
4	0,3272	0,3127	0,2987	0,2851	0,2719	0,2592	0,2469	0,2351	0,2237	0,2127	0,2022	0,1920	0,1823	0,1730
5	0,4950	0,4783	0,4619	0,4457	0,4298	0,4141	0,3988	0,3837	0,3690	0,3457	0,3406	0,3270	0,3137	0,3007
6	0,6544	0,6384	0,6224	0,6063	0,5902	0,5742	0,5582	0,5423	0,5265	0,5108	0,4953	0,4799	0,4647	0,4497
7	0,7841	0,7710	0,7576	0,7440	0,7301	0,7160	0,7017	0,6873	0,6728	0,6581	0,6433	0,6285	0,6136	0,5987
8	0,8766	0,8672	0,8574	0,8472	0,8367	0,8259	0,8148	0,8033	0,7916	0,7796	0,7673	0,7548	0,7420	0,7291
9	0,9352	0,9292	0,9228	0,9161	0,9090	0,9016	0,8939	0,8858	0,8774	0,8686	0,8596	0,8502	0,8405	0,8305
10	0,9686	0,9651	0,9614	0,9574	0,9531	0,9486	0,9437	0,9386	0,9332	0,9274	0,9214	0,9151	0,9084	0,9015
11	0,9859	0,9841	0,9821	0,9799	0,9776	0,9750	0,9723	0,9693	0,9661	0,9627	0,9591	0,9552	0,9510	0,9467
12	0,9941	0,9932	0,9922	0,9912	0,9900	0,9887	0,9873	0,9857	0,9840	0,9821	0,9801	0,9779	0,9755	0,9730
13	0,9977	0,9973	0,9969	0,9964	0,9958	0,9952	0,9945	0,9937	0,9929	0,9920	0,9909	0,9898	0,9885	0,9872
14	0,9991	0,9990	0,9988	0,9986	0,9984	0,9981	0,9978	0,9974	0,9970	0,9966	0,9961	0,9956	0,9950	0,9943
15	0,9997	0,9996	0,9996	0,9995	0,9994	0,9993	0,9992	0,9990	0,9988	0,9986	0,9984	0,9982	0,9979	0,9976
16	0,9999	0,9999	0,9999	0,9998	0,9998	0,9997	0,9997	0,9996	0,9996	0,9995	0,9994	0,9993	0,9992	0,9990
17	1,0000	1,0000	1,0000	0,9999	0,9999	0,9999	0,9999	0,9999	0,9998	0,9998	0,9998	0,9997	0,9997	0,9996
18				1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
19									1,0000	1,0000	1,0000	1,0000	1,0000	1,0000

Tabela 2 - Probabilidades de Poisson acumuladas (continuação)

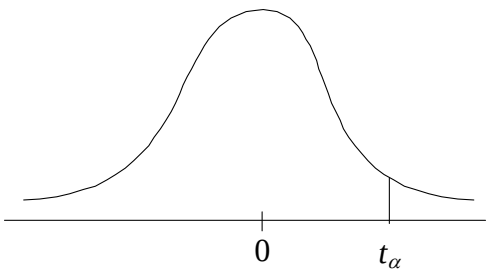
	λ													
r	**7,1**	**7,2**	**7,3**	**7,4**	**7,5**	**7,6**	**7,7**	**7,8**	**7,9**	**8,0**	**8,1**	**8,2**	**8,3**	**8,4**
0	0,0008	0,0007	0,0007	0,0006	0,0006	0,0005	0,0005	0,0004	0,0004	0,0003	0,0003	0,0003	0,0002	0,0002
1	0,0067	0,0061	0,0056	0,0051	0,0047	0,0043	0,0039	0,0036	0,0033	0,0030	0,0028	0,0025	0,0023	0,0021
2	0,0275	0,0255	0,0236	0,0219	0,0203	0,0188	0,0174	0,0161	0,0149	0,0138	0,0127	0,0118	0,0109	0,0100
3	0,0767	0,0719	0,0674	0,0632	0,0591	0,0554	0,0518	0,0485	0,0453	0,0424	0,0396	0,0370	0,0346	0,0323
4	0,1641	0,1555	0,1473	0,1395	0,1321	0,1249	0,1181	0,1117	0,1055	0,0996	0,0940	0,0887	0,0837	0,0789
5	0,2881	0,2759	0,2640	0,2526	0,2414	0,2307	0,2203	0,2103	0,2006	0,1912	0,1822	0,1736	0,1653	0,1573
6	0,4349	0,4204	0,4060	0,3920	0,3782	0,3646	0,3514	0,3384	0,3257	0,3134	0,3013	0,2896	0,2781	0,2670
7	0,5838	0,5689	0,5541	0,5393	0,5246	0,5100	0,4956	0,4812	0,4670	0,4530	0,4391	0,4254	0,4119	0,3987
8	0,7160	0,7027	0,6892	0,6757	0,6620	0,6482	0,6343	0,6204	0,6065	0,5925	0,5786	0,5647	0,5507	0,5369
9	0,8202	0,8096	0,7988	0,7877	0,7764	0,7649	0,7531	0,7411	0,7290	0,7166	0,7041	0,6915	0,6788	0,6659
10	0,8942	0,8867	0,8788	0,8707	0,8622	0,8535	0,8445	0,8352	0,8257	0,8159	0,8058	0,7955	0,7850	0,7743
11	0,9420	0,9371	0,9319	0,9265	0,9208	0,9148	0,9085	0,9020	0,8952	0,8881	0,8807	0,8731	0,8652	0,8571
12	0,9703	0,9673	0,9642	0,9609	0,9573	0,9536	0,9496	0,9454	0,9409	0,9362	0,9313	0,9261	0,9207	0,9150
13	0,9857	0,9841	0,9824	0,9805	0,9784	0,9762	0,9739	0,9714	0,9687	0,9658	0,9628	0,9595	0,9561	0,9524
14	0,9935	0,9927	0,9918	0,9908	0,9897	0,9886	0,9873	0,9859	0,9844	0,9827	0,9810	0,9791	0,9771	0,9749
15	0,9972	0,9969	0,9964	0,9959	0,9954	0,9948	0,9941	0,9934	0,9926	0,9918	0,9908	0,9898	0,9887	0,9875
16	0,9989	0,9987	0,9985	0,9983	0,9980	0,9978	0,9974	0,9971	0,9967	0,9963	0,9958	0,9953	0,9947	0,9941
17	0,9996	0,9995	0,9994	0,9993	0,9992	0,9991	0,9989	0,9988	0,9986	0,9984	0,9982	0,9979	0,9977	0,9973
18	0,9998	0,9998	0,9998	0,9997	0,9997	0,9996	0,9996	0,9995	0,9994	0,9993	0,9992	0,9991	0,9990	0,9989
19	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9998	0,9998	0,9998	0,9997	0,9997	0,9997	0,9996	0,9995
20	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9998	0,9998
21							1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9999
22													1,0000	1,0000

	λ													
r	**8,5**	**8,6**	**8,7**	**8,8**	**8,9**	**9,0**	**9,5**	**10,0**	**10,5**	**11,0**	**11,5**	**12,0**	**12,5**	**13,0**
0	0,0002	0,0002	0,0002	0,0002	0,0001	0,0001	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
1	0,0019	0,0018	0,0016	0,0015	0,0014	0,0012	0,0008	0,0005	0,0003	0,0002	0,0001	0,0000	0,0000	0,0000
2	0,0093	0,0086	0,0079	0,0073	0,0068	0,0062	0,0042	0,0027	0,0018	0,0012	0,0008	0,0004	0,0003	0,0002
3	0,0301	0,0281	0,0262	0,0244	0,0228	0,0212	0,0149	0,0103	0,0071	0,0049	0,0034	0,0022	0,0015	0,0010
4	0,0744	0,0701	0,0660	0,0621	0,0584	0,0550	0,0403	0,0292	0,0211	0,0151	0,0107	0,0075	0,0053	0,0037
5	0,1496	0,1422	0,1352	0,1284	0,1219	0,1157	0,0885	0,0670	0,0504	0,0375	0,0277	0,0203	0,0148	0,0107
6	0,2562	0,2457	0,2355	0,2256	0,2160	0,2068	0,1649	0,1301	0,1016	0,0786	0,0603	0,0457	0,0345	0,0259
7	0,3856	0,3728	0,3602	0,3478	0,3357	0,3239	0,2687	0,2202	0,1785	0,1432	0,1137	0,0894	0,0698	0,0540
8	0,5231	0,5094	0,4958	0,4823	0,4689	0,4557	0,3918	0,3328	0,2794	0,2320	0,1906	0,1549	0,1249	0,0997
9	0,6530	0,6490	0,6269	0,6137	0,6006	0,5874	0,5218	0,4579	0,3971	0,3405	0,2888	0,2423	0,2014	0,1658
10	0,7634	0,7522	0,7409	0,7294	0,7178	0,7060	0,6453	0,5830	0,5207	0,4599	0,4017	0,3471	0,2970	0,2517
11	0,8487	0,8400	0,8311	0,8220	0,8126	0,8030	0,7520	0,6967	0,6387	0,5793	0,5198	0,4615	0,4057	0,3531
12	0,9091	0,9029	0,8965	0,8898	0,8829	0,8758	0,8364	0,7915	0,7419	0,6887	0,6329	0,5759	0,5189	0,4631
13	0,9486	0,9445	0,9403	0,9358	0,9311	0,9261	0,8981	0,8644	0,8253	0,7813	0,7330	0,6815	0,6278	0,5730
14	0,9726	0,9701	0,9675	0,9647	0,9617	0,9585	0,9400	0,9165	0,8879	0,8540	0,8152	0,7719	0,7250	0,6751
15	0,9862	0,9848	0,9832	0,9816	0,9798	0,9780	0,9665	0,9512	0,9316	0,9074	0,8783	0,8443	0,8060	0,7636
16	0,9934	0,9926	0,9918	0,9909	0,9899	0,9889	0,9823	0,9729	0,9604	0,9441	0,9236	0,8986	0,8693	0,8355
17	0,9970	0,9966	0,9962	0,9957	0,9952	0,9947	0,9911	0,9857	0,9781	0,9678	0,9542	0,9370	0,9158	0,8904
18	0,9987	0,9985	0,9983	0,9981	0,9978	0,9976	0,9957	0,9928	0,9885	0,9823	0,9738	0,9625	0,9481	0,9301
19	0,9995	0,9994	0,9993	0,9992	0,9991	0,9989	0,9980	0,9965	0,9942	0,9907	0,9857	0,9786	0,9694	0,9573
20	0,9998	0,9998	0,9997	0,9997	0,9996	0,9996	0,9991	0,9984	0,9972	0,9953	0,9925	0,9883	0,9826	0,9750
21	0,9999	0,9999	0,9999	0,9999	0,9998	0,9998	0,9996	0,9993	0,9987	0,9977	0,9962	0,9939	0,9905	0,9859
22	1,0000	1,0000	1,0000	1,0000	0,9999	0,9999	0,9999	0,9997	0,9994	0,9989	0,9982	0,9969	0,9950	0,9923
23					1,0000	1,0000	1,0000	0,9999	0,9997	0,9995	0,9991	0,9984	0,9975	0,9960
24								1,0000	0,9999	0,9998	0,9996	0,9992	0,9988	0,9980
25									1,0000	0,9999	0,9998	0,9996	0,9994	0,9990
26										1,0000	0,9999	0,9998	0,9997	0,9995
27											1,0000	0,9999	0,9998	0,9998
28												0,9999	0,9999	0,9999
29													1,0000	1,0000

Tabela 2 - Probabilidades de Poisson acumuladas (continuação)

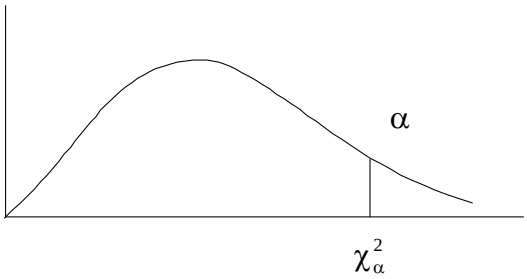
	λ													
r	**13,5**	**14,0**	**14,5**	**15,0**	**15,5**	**16,0**	**16,5**	**17,0**	**17,5**	**18,0**	**18,5**	**19,0**	**19,5**	**20,0**
0	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
1	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
2	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
3	0,0007	0,0004	0,0003	0,0002	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
4	0,0026	0,0017	0,0012	0,0008	0,0006	0,0003	0,0002	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
5	0,0077	0,0054	0,0039	0,0028	0,0019	0,0013	0,0009	0,0006	0,0003	0,0002	0,0002	0,0001	0,0000	0,0000
6	0,0192	0,0141	0,0104	0,0076	0,0055	0,0039	0,0028	0,0020	0,0013	0,0010	0,0007	0,0005	0,0003	0,0002
7	0,0415	0,0315	0,0239	0,0180	0,0134	0,0099	0,0073	0,0054	0,0038	0,0028	0,0020	0,0015	0,0010	0,0007
8	0,0790	0,0620	0,0483	0,0374	0,0288	0,0219	0,0166	0,0126	0,0093	0,0070	0,0052	0,0038	0,0027	0,0020
9	0,1352	0,1093	0,0877	0,0698	0,0552	0,0432	0,0337	0,0261	0,0200	0,0153	0,0116	0,0088	0,0066	0,0049
10	0,2112	0,1756	0,1448	0,1184	0,0961	0,0773	0,0618	0,0491	0,0386	0,0303	0,0236	0,0183	0,0140	0,0107
11	0,3044	0,2599	0,2201	0,1847	0,1538	0,1269	0,1040	0,0846	0,0683	0,0548	0,0437	0,0346	0,0272	0,0213
12	0,4093	0,3584	0,3110	0,2676	0,2282	0,1930	0,1620	0,1350	0,1115	0,0916	0,0747	0,0605	0,0486	0,0389
13	0,5182	0,4644	0,4125	0,3632	0,3171	0,2744	0,2357	0,2008	0,1697	0,1425	0,1188	0,0984	0,0808	0,0661
14	0,6233	0,5703	0,5175	0,4656	0,4154	0,3674	0,3225	0,2808	02425	0,2080	0,1771	0,1497	0,1256	0,1048
15	0,7178	0,6693	0,6191	0,5681	0,5170	0,4667	0,4180	0,3714	0,3274	0,2866	0,2490	0,2148	0,1839	0,1564
16	0,7975	0,7558	0,7111	0,6641	0,6154	0,5659	0,5164	0,4677	0,4203	0,3750	0,3321	0,2920	0,2549	0,2210
17	0,8609	0,8271	0,7897	0,7488	0,7052	0,6593	0,6120	0,5640	0,5159	0,4686	0,4225	0,3783	0,3363	0,2970
18	0,9084	0,8825	0,8529	0,8194	0,7824	0,7423	0,6996	0,6549	0,6088	0,5622	0,5155	0,4694	0,4245	0,3814
19	0,9421	0,9234	0,9012	0,8752	0,8455	0,8122	0,7757	0,7363	0,6944	0,6508	0,6060	0,5606	0,5150	0,4702
20	0,9649	0,9520	0,9362	0,9170	0,8943	0,8681	0,8384	0,8054	0,7693	0,7306	0,6897	0,6471	0,6033	0,5590
21	0,9795	0,9711	0,9603	0,9469	0,9304	0,9107	0,8877	0,8614	0,8317	0,7990	0,7635	0,7255	0,6853	0,6436
22	0,9885	0,9832	0,9762	0,9672	0,9558	0,9417	0,9247	0,9047	0,8814	0,8550	0,8255	0,7931	0,7579	0,7205
23	0,9938	0,9906	0,9863	0,9805	0,9729	0,9632	0,9513	0,9367	0,9192	0,8988	0,8754	0,8490	0,8195	0,7874
24	0,9968	0,9949	0,9923	0,9888	0,9840	0,9776	0,9695	0,9593	0,9467	0,9317	0,9139	0,8932	0,8696	0,8432
25	0,9984	0,9973	0,9959	0,9938	0,9908	0,9868	0,9815	0,9747	0,9660	0,9553	0,9423	0,9268	0,9086	0,8877
26	0,9992	0,9986	0,9978	0,9966	0,9949	0,9924	0,9892	0,9848	0,9790	0,9717	0,9626	0,9514	0,9379	0,9220
27	0,9996	0,9993	0,9989	0,9982	0,9973	0,9958	0,9938	0,9911	0,9874	0,9826	0,9764	0,9687	0,9590	0,9474
28	0,9998	0,9996	0,9994	0,9991	0,9986	0,9977	0,9966	0,9950	0,9926	0,9896	0,9856	0,9804	0,9738	0,9656
29	0,9999	0,9998	0,9997	0,9995	0,9993	0,9988	0,9982	0,9972	0,9958	0,9940	0,9915	0,9881	0,9837	0,9781
30	1,0000	0,9998	0,9998	0,9998	0,9996	0,9993	0,9990	0,9985	0,9977	0,9966	0,9951	0,9930	0,9901	0,9865
31		0,9999	0,9999	0,9999	0,9998	0,9996	0,9995	0,9992	0,9987	0,9981	0,9972	0,9960	0,9941	0,9918
32		1,0000	1,0000	1,0000	0,9999	0,9998	0,9997	0,9996	0,9993	0,9990	0,9985	0,9977	0,9966	0,9952
33					1,0000	0,9998	0,9998	0,9998	0,9996	0,9994	0,9992	0,9988	0,9981	0,9972
34						1,0000	0,9999	0,9999	0,9997	0,9997	0,9995	0,9993	0,9989	0,9984
35							1,0000	1,0000	0,9998	0,9999	0,9997	0,9996	0,9994	0,9991
36									0,9998	0,9999	0,9998	0,9998	0,9996	0,9995
37									1,0000	1,0000	0,9999	0,9999	0,9998	0,9997
38											1,0000	1,0000	0,9998	0,9998
39													1,0000	0,9999
40														1,0000

Tabela 5 - Distribuição *t*



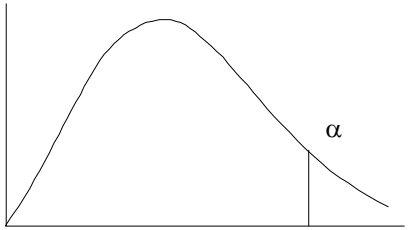
v	α					
g. l.	**0,25**	**0,10**	**0,05**	**0,025**	**0,01**	**0,005**
1	1,000	3,078	6,314	12,706	31,821	63,657
2	0,816	1,886	2,920	4,303	6,965	9,925
3	0,765	1,638	2,353	3,182	4,541	5,841
4	0,741	1,533	2,132	2,776	3,747	4,604
5	0,727	1,476	2,015	2,571	3,365	4,032
6	0,718	1,440	1,943	2,447	3,143	3,707
7	0,711	1,415	1,895	2,365	2,998	3,499
8	0,706	1,397	1,860	2,306	2,896	3,355
9	0,703	1,383	1,833	2,262	2,821	3,250
10	0,700	1,372	1,812	2,228	2,764	3,169
11	0,697	1,363	1,796	2,201	2,718	3,106
12	0,695	1,356	1,782	2,179	2,681	3,055
13	0,694	1,350	1,771	2,160	2,650	3,012
14	0,692	1,345	1,761	2,145	2,624	2,977
15	0,691	1,341	1,753	2,131	2,602	2,947
16	0,690	1,337	1,746	2,120	2,583	2,921
17	0,689	1,333	1,740	2,110	2,567	2,898
18	0,688	1,330	1,734	2,101	2,552	2,878
19	0,688	1,328	1,729	2,093	2,539	2,861
20	0,687	1,325	1,725	2,086	2,528	2,845
21	0,686	1,323	1,721	2,080	2,518	2,831
22	0,686	1,321	1,717	2,074	2,508	2,819
23	0,685	1,319	1,714	2,069	2,500	2,807
24	0,685	1,318	1,711	2,064	2,492	2,797
25	0,684	1,316	1,708	2,060	2,485	2,787
26	0,684	1,315	1,706	2,056	2,479	2,779
27	0,684	1,314	1,703	2,052	2,473	2,771
28	0,683	1,313	1,701	2,048	2,467	2,763
29	0,683	1,311	1,699	2,045	2,462	2,756
30	0,683	1,310	1,697	2,042	2,457	2,750
40	0,681	1,303	1,684	2,021	2,423	2,704
60	0,679	1,296	1,671	2,000	2,390	2,660
120	0,677	1,289	1,658	1,980	2,358	2,617
∞	0,674	1,282	1,645	1,960	2,326	2,576

Tabela 6 - Distribuição χ^2



	α							
g. l.	**0,995**	**0,990**	**0,975**	**0,950**	**0,050**	**0,025**	**0,010**	**0,005**
1	0,00004	0,00016	0,00098	0,00393	3,84146	5,02389	6,63490	7,87944
2	0,01003	0,02010	0,05064	0,10259	5,99147	7,37776	9,21034	10,5966
3	0,07172	0,11483	0,21580	0,35185	7,81473	9,34840	11,3449	12,8381
4	0,20699	0,29711	0,48442	0,71072	9,48773	11,1433	13,2767	14,8602
5	0,41174	0,55430	0,83121	1,14548	11,0705	12,8325	15,0863	16,7496
6	0,67573	0,87209	1,23735	1,63539	12,5916	14,4494	16,8119	18,5476
7	0,98927	1,23904	1,68987	2,16735	14,0671	16,0128	18,4753	20,2777
8	1,34419	1,64648	2,17973	2,73264	15,5073	17,5346	20,0902	21,9550
9	1,73493	2,08791	2,70039	3,32511	16,9190	19,0228	21,6660	23,5893
10	2,15585	2,55821	2,42697	3,94030	18,3070	20,4831	23,2093	25,1882
11	2,60321	3,05347	3,81575	4,57481	19,6751	21,9200	24,7250	26,7569
12	3,07382	3,57056	4,40379	5,22603	21,0261	23,3367	26,2170	28,2995
13	3,56503	4,10691	5,00874	5,89186	22,3621	24,7356	27,6883	29,8194
14	4,07468	4,66043	5,62872	6,57063	23,6848	26,1190	29,1413	31,3193
15	4,60094	5,22935	6,26214	7,26094	24,9958	27,4884	30,5779	32,8013
16	5,14224	5,81221	6,90766	7,96164	26,2962	28,8454	31,9999	34,2672
17	5,69724	6,40776	7,56418	8,67176	27,5871	30,1910	33,4087	35,7185
18	6,26481	7,01491	8,23075	9,39046	28,8693	31,5264	34,8053	37,1564
19	6,84398	7,63273	8,90655	10,1170	30,1435	32,8523	36,1908	38,5822
20	7,43386	8,26040	9,59083	10,8508	31,4104	34,1696	37,5662	39,9968
21	8,03386	8,89720	10,2829	11,5913	32,6705	35,4789	38,9321	41,4010
22	8,64272	9,54249	10,9823	12,3380	33,9244	36,7807	40,2894	42,7956
23	9,26042	10,1957	11,6885	13,0905	35,1725	38,0757	41,6384	44,1813
24	9,88623	10,8564	12,4011	13,8484	36,4151	39,3641	42,9798	45,5585
25	10,5197	11,5240	13,1197	14,6114	37,6525	40,6465	44,3141	46,9278
26	11,1603	12,1981	13,8439	15,3791	38,8852	41,9232	45,6417	48,2899
27	11,8076	12,8786	14,5733	16,1513	40,1133	43,1944	46,9630	49,6449
28	12,4613	13,5648	15,3079	16,9279	41,3372	44,4607	48,2782	50,9933
29	13,1211	14,2565	16,0471	17,7083	42,5569	45,7222	49,5879	52,3356
30	13,7867	14,9535	16,7908	18,4926	43,7729	46,9792	50,8922	53,6720
40	20,7065	22,1643	24,4331	26,5093	55,7585	59,3417	63,6907	66,7659
50	27,9907	29,7067	32,3574	34,7642	67,5048	71,4202	76,1539	74,4900
60	35,5346	37,4848	40,4817	43,1879	79,0819	83,2976	88,3794	91,9517
70	43,2752	45,4418	48,7576	51,7393	90,5312	95,0231	100,425	104,215
80	51,1720	53,5400	57,1532	60,3915	101,879	106,629	112,329	116,321
90	59,1963	61,7541	65,6466	69,1260	113,145	118,136	124,116	128,299
100	67,3276	70,0648	74,2219	77,9295	124,342	129,561	135,807	140,169

Tabela 7 - Distribuição ***F*** - 1%

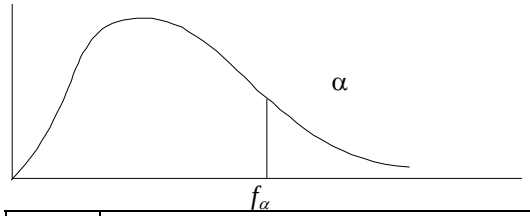


m **= g. lib. numerador**

n **= g. lib. denominador**

	m																		
n	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**12**	**15**	**20**	**24**	**30**	**40**	**60**	**120**	**∞**
1	4052,00	4999,50	5403,00	5625,00	5764,00	5859,00	5928,00	5981,00	6022,00	6056,00	6106,00	6157,00	6209,00	6235,00	6261,00	6287,00	6313,00	6339,00	6366,00
2	98,50	99,00	99,17	99,25	99,30	99,30	99,36	99,37	99,39	99,40	99,42	99,43	99,45	99,46	99,47	99,47	99,48	99,49	99,50
3	34,12	30,82	29,46	28,71	28,24	27,91	27,67	27,49	27,35	27,33	27,05	26,87	26,69	26,60	26,50	26,41	26,32	26,22	26,13
4	21,20	18,00	16,69	15,98	15,52	15,21	14,08	14,80	14,66	14,55	14,37	14,20	14,02	13,93	13,84	13,75	13,65	13,56	13,46
5	16,26	13,27	12,06	11,39	10,97	10,67	10,46	10,29	10,16	10,05	9,89	9,72	9,55	9,47	9,38	9,29	9,20	9,11	9,02
6	13,75	10,92	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87	7,72	7,56	7,40	7,31	7,23	7,14	7,06	6,97	6,88
7	12,25	9,55	8,45	7,85	7,46	7,19	6,99	6,84	6,72	6,62	6,47	6,31	6,16	6,07	5,99	5,91	5,82	5,74	5,65
8	11,26	8,65	7,59	7,01	6,63	6,37	6,18	6,03	5,91	5,81	5,67	5,52	5,36	5,28	5,20	5,12	5,03	4,95	4,86
9	10,56	8,02	6,99	6,42	6,06	5,80	5,61	5,47	5,35	5,26	5,11	4,96	4,81	4,73	4,65	4,57	4,48	4,40	4,31
10	10,04	7,56	6,55	5,99	5,64	5,39	5,20	5,06	4,94	4,85	4,71	4,56	4,41	4,33	4,25	4,17	4,08	4,00	3,91
11	9,65	7,21	6,22	5,67	5,32	5,07	4,89	4,74	4,63	4,54	4,40	4,25	4,10	4,02	3,94	3,86	3,78	3,69	3,60
12	9,33	6,93	5,95	5,41	5,06	4,82	4,64	4,50	4,39	4,30	4,16	4,01	3,86	3,78	3,70	3,62	3,54	3,45	3,36
13	9,07	6,70	5,74	5,21	4,86	4,62	4,44	4,30	4,19	4,10	3,96	3,82	3,66	3,59	3,51	3,43	3,34	3,25	3,17
14	8,86	6,51	5,56	5,04	4,69	4,46	4,28	4,14	4,03	3,94	3,80	3,66	3,51	3,43	3,35	3,27	3,18	3,09	3,00
15	8,68	6,36	5,42	4,89	4,56	4,32	4,14	4,00	3,89	3,80	3,67	3,52	3,37	3,29	3,21	3,13	3,05	2,96	2,87
16	8,53	6,23	5,29	4,77	4,44	4,20	4,03	3,89	3,78	3,69	3,55	3,41	3,26	3,18	3,10	3,02	2,93	2,84	2,75
17	8,40	6,11	5,18	4,67	4,34	4,10	3,93	3,79	3,68	3,59	3,46	3,31	3,16	3,08	3,00	2,92	2,83	2,75	2,65
18	8,29	6,01	5,09	4,58	4,25	4,01	3,84	3,71	3,60	3,51	3,37	3,23	3,08	3,00	2,92	2,84	2,75	2,66	2,57
19	8,18	5,93	5,01	4,50	4,17	3,94	3,77	3,63	3,52	3,43	3,30	3,15	3,00	2,92	2,84	2,76	2,67	2,58	2,49
20	8,10	5,85	4,94	4,43	4,10	3,87	3,70	3,56	3,46	3,37	3,23	3,09	2,94	2,86	2,78	2,69	2,61	2,52	2,42
21	8,02	5,78	4,87	4,37	4,04	3,81	3,64	3,51	3,40	3,31	3,17	3,03	2,88	2,801	2,72	2,64	2,55	2,46	2,36
22	7,95	5,72	4,82	4,31	3,99	3,76	3,59	3,45	3,35	3,26	3,12	2,98	2,83	2,75	2,67	2,58	2,50	2,40	2,31
23	7,88	5,66	4,76	4,26	3,94	3,71	3,54	3,41	3,30	3,21	3,07	2,93	2,78	2,70	2,62	2,54	2,45	2,35	2,26
24	7,82	5,61	4,72	4,22	3,90	3,67	3,50	3,36	3,26	3,17	3,03	2,89	2,74	2,66	2,58	2,49	2,40	2,31	2,21
25	7,77	5,57	4,68	4,18	3,85	3,63	3,46	3,32	3,22	3,13	2,99	2,85	2,70	2,62	2,54	2,45	2,36	2,27	2,17
26	7,72	5,53	4,64	4,14	3,82	3,59	3,42	3,29	3,18	3,09	2,96	2,81	2,66	2,58	2,50	2,42	2,33	2,23	2,13
27	7,68	5,49	4,60	4,11	3,78	3,56	3,39	3,26	3,15	3,06	2,93	2,78	2,63	2,55	2,47	2,38	2,29	2,20	2,10
28	7,64	5,45	4,57	4,07	3,75	3,53	3,36	3,23	3,12	3,03	2,90	2,75	2,60	2,52	2,44	2,35	2,26	2,17	2,06
29	7,60	5,42	4,54	4,04	3,73	3,50	3,33	3,20	3,09	3,00	2,87	2,73	2,57	2,49	2,41	2,33	2,23	2,14	2,03
30	7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17	3,07	2,98	2,84	2,70	2,55	2,47	2,39	2,30	2,21	2,11	2,01
40	7,31	5,18	4,31	3,83	3,51	3,29	3,12	2,99	2,89	2,80	2,66	2,52	2,37	2,29	2,20	2,11	2,02	1,92	1,80
60	7,08	4,98	4,13	3,65	3,34	3,12	2,95	2,82	2,72	2,63	2,50	2,35	2,20	2,12	2,03	1,94	1,84	1,73	1,60
120	6,85	4,79	3,95	3,48	3,17	2,96	2,79	2,66	2,56	2,47	2,34	2,19	2,03	1,95	1,86	1,76	1,66	1,53	1,38
∞	6,63	4,61	3,78	3,32	3,02	2,80	2,64	2,51	2,41	2,32	2,18	2,04	1,88	1,79	1,70	1,59	1,47	1,32	1,00

Tabela 7 - Distribuição ***F*** - 5%



m **= g. lib. numerador**

n **= g. lib. denominador**

	m																		
n	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**12**	**15**	**20**	**24**	**30**	**40**	**60**	**120**	**∞**
1	161,40	199,50	215,70	224,60	230,20	234,00	236,80	238,90	240,50	241,90	243,90	245,90	248,00	249,10	250,10	251,10	252,20	253,30	254,30
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40	19,41	19,43	19,45	19,45	19,46	19,47	19,48	19,49	19,50
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,74	8,70	8,66	8,64	8,62	8,59	8,57	8,55	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,91	5,86	5,80	5,77	5,75	5,72	5,69	5,66	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,68	4,62	4,56	4,53	4,50	4,46	4,43	4,40	4,36
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,00	3,94	3,87	3,84	3,81	3,77	3,74	3,70	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,57	3,51	3,44	3,41	3,38	3,34	3,30	3,27	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,28	3,22	3,15	3,12	3,08	3,04	3,01	2,97	2,93
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,07	3,01	2,94	2,90	2,86	2,83	2,79	2,75	2,71
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,91	2,85	2,77	2,74	2,70	2,66	2,62	2,58	2,54
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85	2,79	2,72	2,65	2,61	2,57	2,53	2,49	2,45	2,40
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75	2,69	2,62	2,54	2,51	2,47	2,43	2,38	2,34	2,30
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67	2,60	2,53	2,46	2,42	2,38	2,34	2,30	2,25	2,21
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60	2,53	2,46	2,39	2,35	2,31	2,27	2,22	2,18	2,13
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54	2,48	2,40	2,33	2,29	2,25	2,20	2,16	2,11	2,07
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,42	2,35	2,28	2,24	2,19	2,15	2,11	2,06	2,01
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45	2,38	2,31	2,23	2,19	2,15	2,10	2,06	2,01	1,96
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41	2,34	2,27	2,19	2,15	2,11	2,06	2,02	1,97	1,92
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38	2,31	2,23	2,16	2,11	2,07	2,03	1,98	1,93	1,88
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,28	2,20	2,12	2,08	2,04	1,99	1,95	1,90	1,84
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32	2,25	2,18	2,10	2,05	2,01	1,96	1,92	1,87	1,81
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30	2,23	2,15	2,07	2,03	1,98	1,94	1,89	1,84	1,78
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,27	2,20	2,13	2,05	2,01	1,96	1,91	1,86	1,81	1,76
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25	2,18	2,11	2,03	1,98	1,94	1,89	1,84	1,79	1,73
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24	2,16	2,09	2,01	1,96	1,92	1,87	1,82	1,77	1,71
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22	2,15	2,07	1,99	1,95	1,90	1,85	1,80	1,75	1,69
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31	2,25	2,20	2,13	2,06	1,97	1,93	1,88	1,84	1,79	1,73	1,67
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19	2,12	2,04	1,96	1,91	1,87	1,82	1,77	1,71	1,65
29	4,18	3,33	2,93	2,70	2,55	2,43	2,35	2,28	2,22	2,18	1,10	2,03	1,94	1,90	1,85	1,81	1,75	1,70	1,64
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16	2,09	2,01	1,93	1,89	1,84	1,79	1,74	1,68	1,62
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08	2,00	1,92	1,84	1,79	1,74	1,69	1,64	1,58	1,51
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99	1,92	1,84	1,75	1,70	1,65	1,59	1,53	1,47	1,39
120	3,92	3,07	2,68	2,45	2,29	2,17	2,09	2,02	1,96	1,91	1,83	1,75	1,66	1,61	1,55	1,50	1,43	1,35	1,25
∞	3,84	3,00	2,60	2,37	2,21	2,10	2,01	1,94	1,88	1,83	1,75	1,67	1,57	1,52	1,46	1,39	1,32	1,22	1,00