Apostila

Bioestatística

Maria Ivanilde Araújo
Antônio Alcirley da Silva Balieiro

Bioestatística

Conteúdo

1. Os dados e a Estatística

Definiremos de maneira simples e concisa alguns elementos que usaremos no decorrer do curso.

Dados: é um (ou mais) conjunto de valores, numéricos ou não.

Estatística: é um conjunto de técnicas desenvolvidas com a finalidade de auxiliar a responder, de forma objetiva e segura, as situações que envolvem uma grande quantidade de informações. Pode ser usada para analisar situações complexas ou não. Permite, de forma sistemática, organizar, descrever, analisar e interpretar dados oriundos de estudo ou experimentos realizados em qualquer área do conhecimento.

Podemos dividir a estatística em três partes:

a) Estatística Descritiva
b) Probabilidade
c) Inferência Estatística.

1.1. A Estatística na Prática

Porque a estatística é importante?

Os métodos estatísticos são usados hoje em quase todos os campos de investigação científica, já que eles nos capacitam a responder a um vasto número de questões, tais como as listadas abaixo:

a) Como os cientistas avaliam a validade de novas teorias?
b) Como os pesquisadores médicos testam a eficiência de novas drogas?
c) Como os demógrafos prevêem o tamanho da população do mundo em qualquer tempo futuro?
d) Como pode um economista verificar se a mudança atual no Índice de Preços ao Consumidor é a continuação de uma tendência secular ou simplesmente um desvio aleatório?
e) Como é possível para alguém predizer o resultado de uma eleição entrevistando apenas algumas centenas de eleitores?
f) Como os pesquisadores na educação testam a eficiência de um novo método de ensino?

Estes são poucos exemplos nos quais a aplicação da estatística é necessária. Por isso, a estatística tornou-se uma ferramenta cotidiana para todos os tipos de profissionais que entram em contato com dados quantitativos ou tiram conclusões a partir destes.

A Estatística, além de servir como apoio científico à quase todas as áreas do conhecimento (Engenharia, Economia, Agronomia, Medicina, Física, Ciências Humanas em geral, etc.), proporciona mecanismos para diagnosticar e aperfeiçoar a gestão e operação de diversos sistemas complexos, desde os sistemas humanos aos sistemas físicos, possibilitando criar modelos que descrevam o comportamento de algumas variáveis em função de outro conjunto de variáveis. Por exemplo, através de Métodos de Regressão podemos relacionar ou criar uma relação entre, a variabilidade de diversas variáveis estocásticas como o valor nominal dos imóveis de uma cidade em função de características previamente especificadas desses imóveis.

1.2. Os Dados

1.2.1. Coleta de Dados

Após a definição do problema a ser estudado e o estabelecimento do planejamento da pesquisa (forma pela qual os dados serão coletados; cronograma das atividades, custos envolvidos; exame das informações disponíveis; delineamento da amostra etc.), o passo seguinte é a coleta de dados, que consiste na busca ou compilação dos dados das variáveis, componentes do fenômeno a ser estudado.

A coleta de dados pode ser direta ou indireta.

Coleta direta: Quando os dados são obtidos na fonte originária. Os valores assim compilados são chamados de dados primários, como, por exemplo, nascimentos, casamentos e óbitos, todos registrados no Cartório de Registro Civil; opiniões obtidas em pesquisas de opinião pública, ou ainda, quando os dados são coletados pelo próprio pesquisador.

A coleta direta pode ser classificada relativamente ao fator tempo em:

- **Contínua** – Quando feita continuamente, como por exemplo, nascimentos e óbitos, freqüência dos alunos às aulas;

- **Periódica** – Quando é feita em intervalos constantes de tempo, como os censos (de 10 em 10 anos);
- **Ocasional** – Quando é feita sem época preestabelecida.

Coleta indireta: Quando os dados obtidos provêm da coleta direta. Os valores assim compilados são denominados de dados secundários, como, por exemplo, o cálculo do tempo de vida média, obtido pela pesquisa, nas tabelas demográficas publicadas pela

Fundação Instituto Brasileiro de Geografia e Estatística – IBGE constitui-se em uma coleta indireta.

Apresentação dos Dados

Após a crítica, os dados devem ser apresentados sob forma adequada (tabelas ou gráficos), para o melhor entendimento do fenômeno que está sendo estudado.

Análise dos Resultados

Realizadas as fases anteriores, faz-se uma análise dos resultados obtidos, através dos métodos da Estatística Indutiva ou Inferência, e tiram-se as conclusões e previsões.

1.2.2. Tipos de Variáveis

Cada uma das características observadas ou mensuradas em um fenômeno é denominada de *variável*. Para o fenômeno "sexo" são dois os resultados possíveis: sexo masculino e sexo feminino;

Para a variável "número de filhos" há um número de resultados possíveis expressos através dos números naturais: 0, 1, 2, 3, ..., n;

Para a variável "estatura" temos uma situação diferente, pois os resultados podem tomar um número infinito de valores numéricos dentro de um determinado intervalo.

As variáveis podem ser:

Variáveis Quantitativas - Referem-se às quantidades e podem ser medidas em uma escala numérica. Exemplos: idade das pessoas, preço dos produtos, peso dos recém nascidos. Elas subdividem-se em dois grupos:

- **Variáveis Quantitativas Discretas:** são aquelas que assumem apenas determinados valores tais como 0, 1, 2, 3, 4, 5, 6 dando saltos de descontinuidade entre seus valores. Normalmente refere-se a contagens. Por exemplo: número de vendas diárias em uma empresa, número de pessoas por família, quantidade de doentes por hospital.
- **Variáveis Quantitativas Contínuas**: são aquelas cujos valores assumem uma faixa contínua e não apresentam saltos de descontinuidade. Exemplos dessas variáveis são: Os pesos de pessoas, a renda familiar, o consumo mensal de energia elétrica, o preço de um produto agrícola.

Variáveis Qualitativas - Refere-se a dados não numéricos. Exemplos dessas variáveis são: O sexo das pessoas, a cor, o grau de instrução. Elas subdividem-se também em dois grupos:

- **Variáveis Qualitativas Ordinais**: São aquelas que definem um ordenamento ou uma hierarquia. Exemplos são: O grau de instrução, a classificação de um estudante no curso de Estatística, as posições das 100 empresas mais lucrativas, etc.
- **Variáveis Qualitativas Nominais:** Estas por sua vez, não definem qualquer ordenamento ou hierarquia. São exemplos destas: A cor, o sexo, o local de nascimento, etc.

População: É o conjunto de elementos a serem observados. Exemplo: Todas as imobiliárias em uma dada cidade; todos os imóveis à venda em certo período em uma dada região, as empresas de engenharia de Manaus, todas as peças não-conformes em certo período na produção de um produto em uma determinada indústria, etc.

Amostra: É uma pequena parte selecionada de uma população que se pretende estudar. Fazemos uma amostragem quando:

- O número de elementos da população é muito grande;
- Quando queremos economizar tempo e dinheiro;
- Não é possível acessar todos os elementos da população.

2. Estatística Descritiva

É a parte mais conhecida. Quem vê os noticiários na televisão ou nos jornais, sabe quão freqüente é o uso de médias, índices e gráficos nas notícias.

É a parte da Estatística que coleta, descreve, organiza e apresenta os dados. É nesta etapa que são tiradas conclusões.

Exemplos:

a) O INPC, Índice Nacional de Preços ao Consumidor, é um índice de maior importância em nossa sociedade. Sua constituição envolve a sintetização, em um único número, dos aumentos dos produtos da cesta básica. No fundo é um sucessivo cálculo de médias, da mesma forma o INCC, Índice Nacional de Construção Civil.

b) Anuário Estatístico Brasileiro. O Instituto Brasileiro de Geografia e Estatística - IBGE publica a cada ano este anuário apresentando, em várias tabelas, os mais diversos dados sobre o Brasil: Educação, transporte, economia, cultura, etc. Embora simples e fáceis de serem entendidas, as tabelas são o produto de um processo extremamente demorado de coleta e apuração e dados.

2.1. Métodos Tabulares e Métodos Gráficos

As técnicas aqui estudadas permitem detectar e corrigir erros e inconsistências ocorridos durante um processo de coleta de dados, determinar as principais características destes mesmos dados e propiciar familiaridade com eles.

Tabela é um quadro que resume um conjunto de observações.

Ela é composta de:

- **Título**: Conjunto de informações, as mais completas possíveis, respondendo às perguntas: O que? (referente ao fato), Quando? (correspondente à época), Onde? (relativo ao lugar);
- **Corpo**: Conjunto de linhas e colunas que contém informações sobre a variável em estudo;

- **Cabeçalho**: Parte superior da tabela que especifica o conteúdo das colunas;
- **Rodapé**: Reservado para as observações pertinentes, bem como a identificação da fonte dos dados.

Exemplos:

Estimativas para o ano de 2006 de número de casos novos por câncer, em homens e mulheres, segundo localização primária, no Brasil.

Localização Primária Neoplasia Maligna	Estimativa dos Casos Novos		
	Masculino	Feminino	Total
Mama Feminina	-	48.930	48.930
Traquéia, Brônquio e Pulmão	17.850	9.320	27.170
Estômago	14.970	8.230	23.200
Colo de Útero	-	19.260	19.260
Próstata	47.280	-	47.280
Cólon e Reto	11.390	13.970	25.360
Esôfago	7.970	2.610	10.580
Leucemias	5.330	4.220	9.550
Cavidade Oral	10.060	3.410	13.470
Pele Melanoma	2.710	3.050	5.760
Outras Localizações	61.530	63.320	124.850
Subtotal	**179.090**	**176.320**	**355.410**
Pele Não Melanoma	55.480	61.160	116.640
Todas as Neoplasias	**234.570**	**237.480**	**472.050**

Fonte: Estimativa 2006 – Incidência de Câncer no Brasil - INCA

Prevalência dos que experimentaram cigarro, pelo menos uma ou duas tragadas, segundo sexo.

Experimentou Cigarro	Masculino		Feminino		Total	
	Número	Percentual	Número	Percentual	Número	Percentual
Sim	50	39,37%	153	32,69%	203	34,12%
Não	77	60,63%	315	67,31%	392	65.88%

Fonte: Questionário sobre Prevalência de Tabagismo e seus Determinantes, em profissionais da área de saúde da cidade de Manaus.

A representação gráfica dos dados tem por finalidade dar uma idéia, a mais imediata possível, dos resultados obtidos, Nos permitindo chegar a conclusões sobre a evolução do fenômeno ou sobre como se relacionam os valores da série. Não há apenas uma maneira de representar graficamente uma série estatística. A escolha do gráfico mais apropriado ficará a critério do analista.

Contudo, os elementos: Simplicidade, clareza e veracidade devem ser consideradas quanto à elaboração de um gráfico.

Sintetizando Dados Quantitativos /Qualitativos

Tabelas: O objetivo é apresentar os dados agrupados de forma que seu manuseio, visualização e compreensão sejam simplificados. Esta apresentação pode ser feita de forma tabular ou gráfica.

As tabelas, dependendo do tipo de dados, podem ser:

a) Simples
b) Dupla entrada
c) Distribuição de freqüência

Distribuição de Freqüências: Um estudo completo das distribuições de freqüências é imprescindível porque este é o tipo de tabela mais importante para a Estatística Descritiva. A seguir são descritos os procedimentos usuais na construção dessas tabelas. Primeiramente vamos ver alguns conceitos fundamentais:

a) Dados brutos: É o conjunto dos dados numéricos obtidos após a crítica dos valores coletados. Os seguintes valores poderiam ser os dados brutos: 24, 23, 22, 28, 35, 21, 23, 33.

b) Rol: É o arranjo dos dados brutos em ordem de freqüência crescente ou decrescente. Os dados brutos anteriores ficariam assim: 21, 22, 23, 23, 24, 28, 33, 35.

c) Amplitude Total ou "Range" (*R*). É a diferença entre o maior e o menor valor observado. No exemplo, R = 35 - 21 = 14.

d) Classe: É cada um dos grupos de valores em que se subdivide a amplitude total do conjunto de valores observados da variável.

e) **Limite de Classe:** São os valores extremos do intervalo de classe.

Exemplo: No intervalo de classe 75|-----85, o limite inferior (LI) é representado pelo valor 75, inclusive, e o valor 85 representa o limite superior (LS), exclusive, do intervalo de classe.

f) **Ponto Médio do Intervalo de Classe (x_i):** É o valor que representa a classe para o cálculo de certas medidas. Na distribuição de freqüência com dados agrupados em intervalos de classe considera-se que os dados distribuem-se de maneira uniforme no intervalo. Sua fórmula é bem simples, vejamos:

$$x_i = \frac{\mathrm{LS} - \mathrm{LI}}{2}$$

Tipos de Freqüências

Freqüência Simples Absoluta (F_i)

É o número de vezes que o elemento aparece na amostra, ou o número de elementos pertencentes a uma classe.

Freqüência Absoluta Acumulada (F_{ac})

É a soma da freqüência absoluta da classe com a freqüência absoluta das classes anteriores.

Freqüência Simples Relativa (f_r)

A freqüência relativa é o valor da freqüência absoluta dividido pelo número total de observações: $f_r = \frac{F_i}{n}$

Freqüência Relativa Acumulada (f_{ra})

A freqüência acumulada relativa é o valor da freqüência acumulada dividido pelo número total de observações: $f_{ra} = \frac{F_{ac}}{n}$

Distribuição de Freqüências

Utilizamos esse tipo de distribuição quando estamos interessados em agrupar o conjunto de dados.

Exemplo: Considere o seguinte conjunto de dados: 21, 21, 21, 22, 22, 23, 23, 24, 25, 25, 25, 25, 26, 26, 26, 28, 30. Construa uma distribuição com todas as freqüências.

X	F_i	F_{ac}	f_r	f_{ra}
21	3	3	3/17	3/17
22	2	5	2/17	5/17
23	2	7	2/17	7/17
24	1	8	1/17	8/17
25	4	12	4/17	12/17
26	3	15	3/17	15/17
28	1	16	1/17	16/17
30	1	17	1/17	17/17
$\sum$	17		1	

Intervalos de Classes

Conjunto de observações apresentadas na forma contínua, sem superposição de intervalos, de tal modo que cada valor do conjunto de observação possa ser alocado em um, e apenas um, dos intervalos.

O número k de intervalos para cada conjunto de observações com n valores pode ser calculado por diversas formas. O método de Sturges é um dos métodos e estabelece que:

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

Onde:

k = número de classes

n = número total de observações

Ex.: para um conjunto com 50 observações obtemos $\log_{10} 50 \cong 1{,}699$;

$k = 1 + 3{,}322 \times 1{,}699 \cong 6{,}6 \cong 7$ intervalos

O tamanho (w) de cada intervalo é obtido pela divisão do valor da diferença entre o maior e o menor valor, R, pelo número de intervalos k: $w = \frac{R}{k}$

Exemplo 1		
Se utilizar a fórmula de Sturges $R = 53 - 19 = 34\, e\, n = 50$ Então: $k = 1 + 3{,}322 \text{ x } 1{,}699 \cong 7$ intervalos $w = \frac{34}{7} \cong 5$ idades, Então o tamanho do intervalo é de 5 idades.	Intervalos de Classe	Freqüência
	17 |------- 23	04
	23 |------- 29	14
	29 |------- 35	12
	35 |------- 41	08
	41 |------- 47	04
	47 |------- 53	07
	53 |------- 59	01

Exemplo 2		
ou, se construirmos os intervalos empiricamente.	Intervalo de classe	Freqüência
	10 |------- 20	02
	20 |------- 30	20
	30 |------- 40	12
	40 |------- 50	12
	50 |------- 60	04

Gráficos

Os gráficos são representações pictóricas dos dados, muito valiosas na visualização dos resultados. É importante saber representar os dados na forma gráfica corretamente, pois se forem representados de forma errada acarretam ao analista uma idéia falsa dos dados chegando até mesmo a confundi-lo. Os principais tipos de gráficos usados na representação estatística são:

a) Histograma e Polígono de freqüência (para dados agrupados em distribuições de freqüências)

Histograma

Histograma é uma representação gráfica de uma tabela de distribuição de freqüências. Desenhamos um par de eixos cartesianos e no eixo horizontal (abscissa) colocamos os valores da variável em estudo e no eixo vertical (ordenadas) colocamos

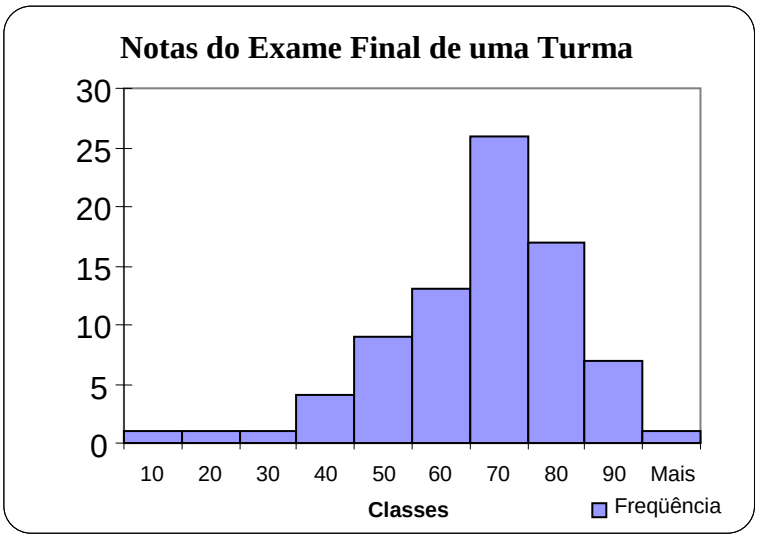
os valores das freqüências. O histograma tanto pode ser representado para as freqüências absolutas como para as freqüências relativas.

Exemplo: O conjunto abaixo representa as notas do exame final de uma 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

Notas	*Freqüência*
0 ----| 10	1
10 ----| 20	1
20 ----| 30	1
30 ----| 40	4
40 ----| 50	9
50 ----| 60	13
60 ----| 70	26
70 ----| 80	17
80 ----| 90	7
Mais	1

Fonte: dados hipotéticos

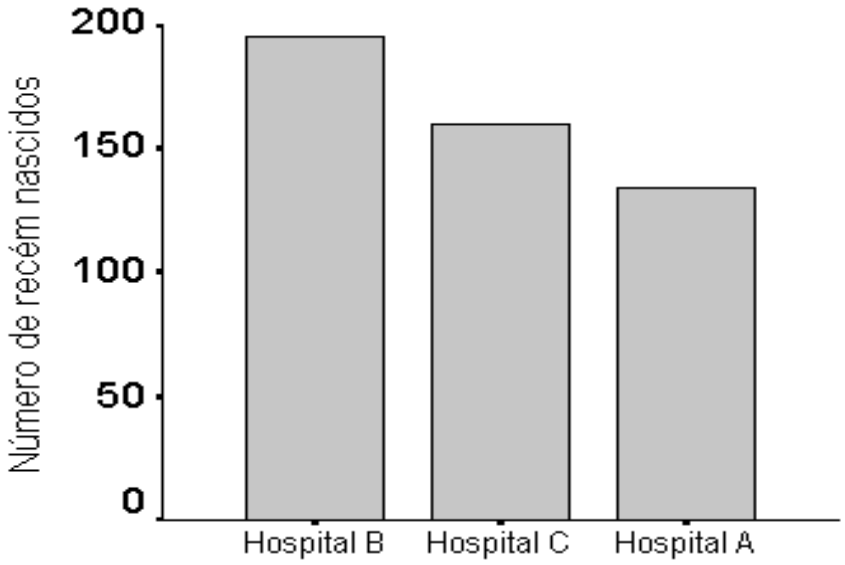


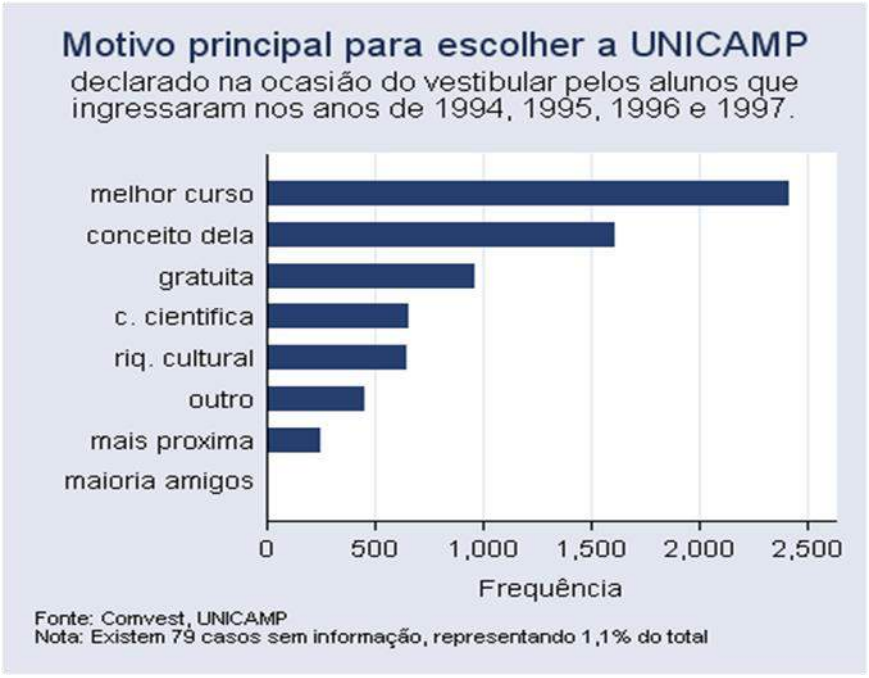
Fonte: dados hipotéticos

b) Gráfico em barras ou colunas (verticais e horizontais);

Esses tipos de gráficos têm como finalidade a comparação de grandezas e prestam-se em especial à representação de dados relacionados a séries de tempo, como por exemplo:

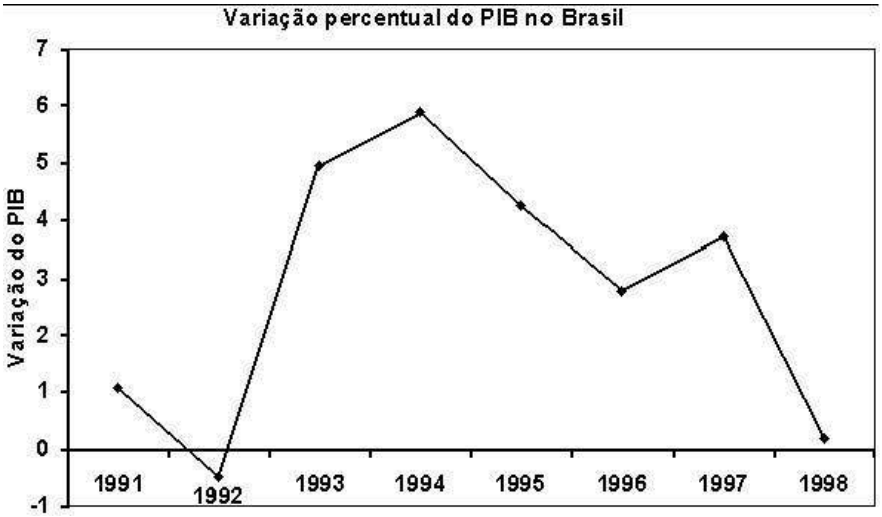
Gráfico de barras representando a percentagem de recém nascidos por hospital



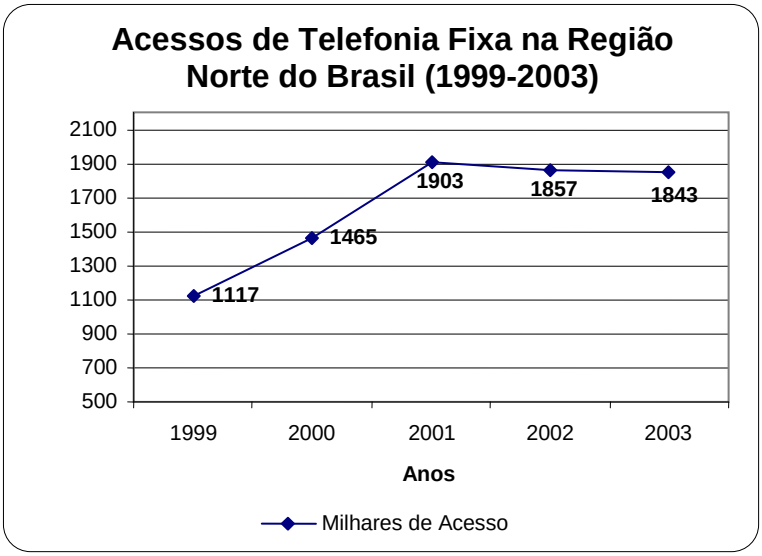


c) Gráfico em linhas ou lineares;

São freqüentemente usados para representar séries de tempo ou quando um dos fatores seja o tempo, pois uma vez tratando-se de um grande período de tempo a representação em colunas acaba sendo inviável devido à alta concentração de dados. Sua construção é de forma muito simples bastando marcar os pontos correspondentes aos valores observados em cada período e ligá-los por meio de um traço.



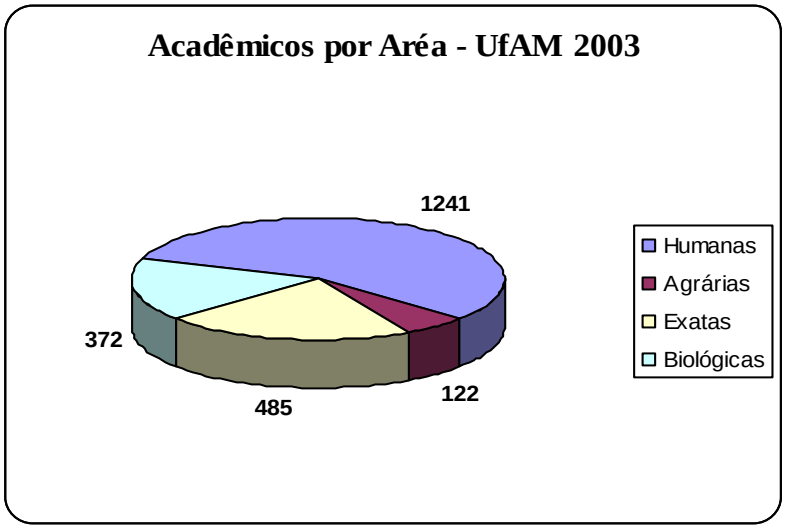
Fonte: Fictício



Fonte: Dados fornecidos pela Telemar

d) Gráfico em setores

Construção de um gráfico de setores, dos acadêmicos por área – UFAM 2003.



Fonte: questionário aplicado aos alunos da UFAM, com ingresso no ano de 2003. *134 alunos não responderam esta questão

2.2. Métodos Numéricos

2.2.1. Medidas de Posição, Locação ou Tendência Central

Mostram o valor representativo em torno do qual os dados tendem a agrupar-se com maior ou menor freqüência. São utilizadas para sintetizar em um único número o conjunto de dados observados. Esta sintetização é necessária, por exemplo, na construção do INPC (Índice Nacional de Preços ao Consumidor). Embora, em um dado mês, cada artigo registre um aumento específico, é necessário sintetizar esses aumentos em um único número para ser usado nos vários setores da economia.

a) Média Aritmética Simples

Podemos dizer que esta é a mais importante medida de locação e que é mais comumente usada para descrever um conjunto de observações. A média aritmética

simples de um conjunto de *n* observações é o quociente entre a soma dos dados e a quantidade dessas observações. É denotada por $\overline{X}$.

$$\text{Média} = \frac{\text{Soma dos dados (valores observados)}}{\text{Número total de observações}}$$

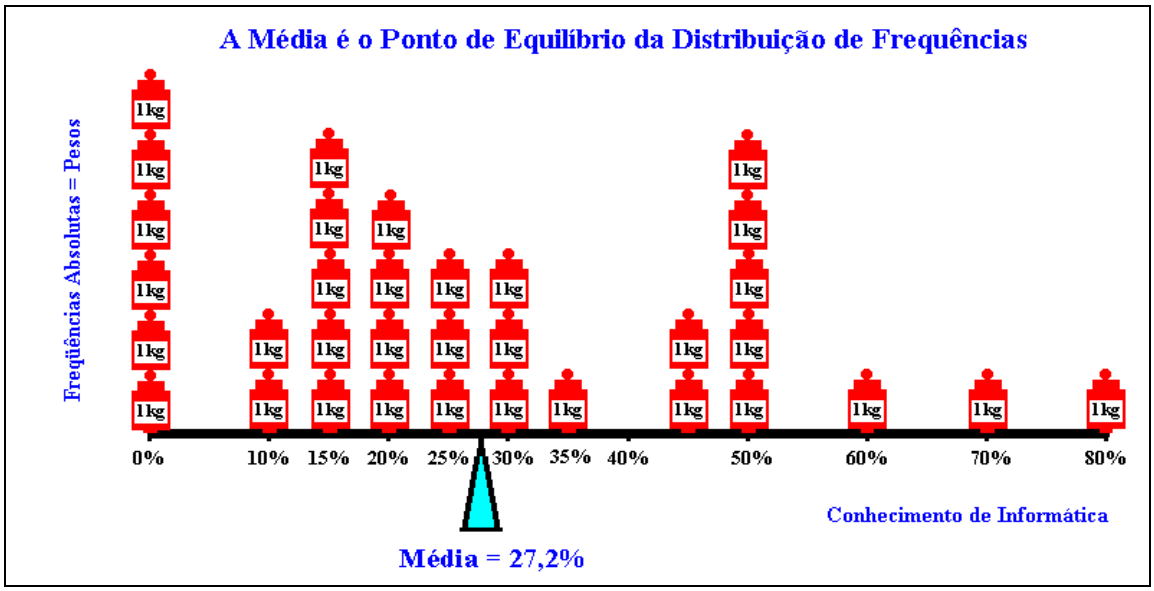
Em linguagem matemática, a **média amostral** se expressa de forma seguinte:

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

Exemplo: Calcule a média da variável X: 3, 5, 8, 12, 7, 12, 15, 18, 20, 20.

$$\overline{X} = \frac{3+5+8+12+7+12+15+18+20+20}{10} = \frac{120}{10} = 12$$

Diante da pergunta "**Como interpretar a média?**", as respostas mais comuns são: "*Representa a posição da maioria*" ou "*É o valor que está no meio da amostra*". Ambas estão **erradas!** Quem representa a posição da(s) maioria(s) locais é a **moda**, e quem está no meio do rol é a **mediana**. A média é outra coisa! O gráfico seguinte nos ajudará a responder a questão.



b) Média Aritmética Ponderada

Em algumas situações os números que queremos sintetizar têm graus de importância diferentes, usa-se então a média aritmética ponderada.

A média aritmética ponderada de um conjunto de *n* observações é o quociente da divisão pela soma dos pesos da soma das observações multiplicadas por seu respectivo peso.

Com intervalos de Classe

Neste caso, convencionamos que todos os valores incluídos em um determinado intervalo de classe coincidem com o seu ponto médio, e determinamos a média aritmética ponderada, por meio da fórmula:

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i F_i}{\sum_{i=1}^{n} F_i}$$

Onde x_i é ponto médio da classe.

Exemplo: Consideremos a distribuição relativa a 34 famílias de 4 filhos, seja X o número de filhos do sexo masculino:

Nº de meninos	F_i	$x_i F_i$
0	2	0
1	6	6
2	10	20
3	12	36
4	4	16
	Σ = 34	Σ = 78

Fonte: dados hipotéticos.

Temos, então:

$\sum_{i=1}^{n} x_i F_i = 78$ e $\sum_{i=1}^{n} F_i = 34$

Logo:

$$\overline{X} = \frac{78}{34} = 2{,}29 \rightarrow \overline{X} = 2{,}3$$

isto é: $\overline{X} = 2{,}3$ meninos

Nota: sendo X uma variável discreta, como interpretar o resultado obtido, 2 meninos e 3 décimos de menino?

O valor médio 2,3 meninos sugere, neste caso que o maior número de famílias tem 2 meninos e 2 meninas, sendo, porém, a tendência geral de uma leve superioridade numérica em relação ao número de meninos.

Exemplo: Suponha que tenhamos feito uma coleta de dados relativos às estaturas de 40 alunos, que compõem uma amostra dos alunos de um colégio A, resultando a seguinte tabela de valores.

i	Estaturas (cm)	F_i	x_i	x_iF_i
01	150 |---- 154	04	152	608
02	154 |---- 158	09	156	1404
03	158 |---- 162	11	160	1760
04	162 |---- 166	08	164	1312
05	166 |---- 170	05	168	840
06	170 |---- 174	03	172	516
		Σ = 40		Σ = 6440

Fonte: dados hipotéticos

Temos, neste caso:

$\sum_{i=1}^{n} x_i F_i = 6440$ e $\sum_{i=1}^{n} F_i = 40$

Logo:

$$\overline{X} = \frac{6440}{40} = 161 \rightarrow \overline{X} = 161\text{cm}$$

c) Mediana

A mediana de um conjunto de *n* observações é o valor "do meio" do conjunto, quando os dados estão ordenados. Se *n* é ímpar esse valor é único; se *n* é par, a mediana é a média aritmética simples dos dois valores centrais.

Exemplo: Determinar a mediana do conjunto X: 2, 20, 12, 23, 20, 8, 12.

Ordenando os termos: 2, 8, 12, **12**, 20, 20, 23.

A mediana será o número 12, pois ele divide o conjunto em duas partes iguais. Portanto, Md = 12.

Exemplo: Determinar a mediana da série X: 7, 21, 13, 15, 10, 8, 9, 13.

Ordenando os termos: 7, 8,9, **10, 13**, 13, 15, 21.

A mediana será:

$$Md = \frac{10 + 13}{2} = 11{,}5$$

Com intervalos de Classe

Neste caso, o problema consiste em determinar o ponto do intervalo em que está compreendida a mediana.

Para tanto, temos inicialmente que determinar a classe na qual se encontra a mediana – classe mediana: É o valor que divide as observações em duas partes, onde 50% dos dados ficam acima dele e o restante abaixo. Tal classe será, evidentemente, aquela corresponde à freqüência acumulada imediatamente superior a $\sum_{i=1}^{n} F_i/2$

Na prática seguimos os seguintes passos:

1) Determinamos as freqüências acumuladas.
2) Calculamos $\sum_{i=1}^{n} F_i/2$
3) Marcamos a classe correspondente à freqüência acumulada imediatamente superior a $\sum_{i=1}^{n} F_i/2$ – classe mediana – e, em seguida, empregamos a fórmula:

$$Md = LI + h\,\frac{\left[\frac{\sum_{i=1}^{n} F_i}{2} - f_{ant}\right]}{f_{Md}}$$

Onde:

LI é o limite inferior da classe mediana

f_{ant} é a freqüência acumulada da classe anterior à classe mediana

f_{Md} é a freqüência simples da classe mediana

h é a amplitude da classe mediana

Exemplo: Tomemos a distribuição relativa à tabela do nº de meninos, completando-a com a coluna correspondente à freqüência acumulada:

Nº de meninos	F_i	F_{ac}
0	2	2
1	6	8
2	10	18
3	12	30
4	4	34
	Σ = 34	

Fonte: dados hipotéticos

Sendo:

$\sum_{i=1}^{n} F_i/2 = 34/2 = 17$

A menor freqüência acumulada que supera este valor é 18, que corresponde ao valor 2 da variável, sendo este o valor mediano.

Logo:

Md = 2 meninos

Exemplo: Tomemos a distribuição relativa à tabela da estatura dos alunos, completando-a com a coluna correspondente à freqüência acumulada:

i	Estaturas (cm)	F_i	F_{ac}
1	150 |---- 154	4	4
2	154 |---- 158	9	13
3	**158 |---- 162**	**11**	**24** → **Classe Mediana**
4	162 |---- 166	8	32
5	166 |---- 170	5	37
6	170 |---- 174	3	40
		Σ = 40	

Fonte: dados hipotéticos

Temos:

$\sum_{i=1}^{n} F_i/2 = 40/2 = 20$

Como há 24 valores incluídos nas três primeiras classes da distribuição e como pretendemos determinar o valor que ocupa o 20º lugar, a partir do inicio da série, vemos que este deve estar localizado na terceira classe ($i = 3$), supondo que as freqüências dessas classes estejam uniformemente distribuídas.

Como há 11 elementos nessa classe e sendo o intervalo de classe igual a 4, devemos tomar do limite inferior, a distância:

$$\frac{20 - 13}{11} x4 = \frac{7}{11} x4$$

e a mediana será dada por:

$$Md = 158 + \frac{7}{11} x4 = 158 + \frac{128}{11} = 158 + 2{,}54 = 160{,}54$$

Logo: Md = 160,5 cm

d) Moda (Mo)

É o valor de maior freqüência em um conjunto de dados. Ela é denotada por Mo.

Exemplo: Determinar a moda dos conjuntos de dados:

X: 2, 8, 3, 5, 4, 5, 3, 5, 5, 1.

O elemento de maior freqüência é 5. Portanto, Mo = 5. É uma seqüência unimodal, pois só temos uma moda. X: 6, 10, 5, 6, 10, 2.

Este conjunto de dados apresenta o elemento 6 e 10 como elementos de maior freqüência. Portanto, Mo = 6 e Mo = 10. Por isso é chamada de bimodal.

Quando não houver elementos que se destaquem pela maior freqüência, dizemos que a série é amodal.

Exemplo: X: 3, 3, 3, 4, 4, 4.

Não há moda, pois os elementos têm a mesma freqüência.

Com intervalos de Classe

A classe que apresenta a maior freqüência é denominada classe modal. Pela definição, podemos afirmar que a moda, neste caso, é o valor dominante que está compreendido entre os limites da classe modal.

Par determinação da moda, Czuber criou a seguinte expressão denominada fórmula de Czuber e, na qual:

$$Mo = LI + h\frac{D_1}{D_1 + D_2}$$

LI é o limite inferior da classe modal

h é a amplitude da classe modal

$\boldsymbol{D_1 = f_{Mo} - f_{ant}}$

$\boldsymbol{D_2 = f_{Mo} - f_{post}}$

Onde:

$\boldsymbol{f_{Mo}}$ é a freqüência simples da classe modal

$\boldsymbol{f_{ant}}$ é a freqüência simples da classe anterior à classe modal

$\boldsymbol{f_{post}}$ é a freqüência simples da classe posterior à classe modal

Exemplo: Tomemos a distribuição relativa à tabela da estatura dos alunos:

i	Estaturas (cm)	F_i
1	150 |---- 154	4
2	154 |---- 158	9
3	**158 |---- 162**	**11**
4	162 |---- 166	8
5	166 |---- 170	5
6	170 |---- 174	3
		Σ = 40

→ Classe Modal

Fonte: dados hipotéticos

Temos:

A classe modal é $i = 3$

$D_1 = f_{Mo} - f_{ant} \Rightarrow D_1 = 11 - 9 \Rightarrow D_1 = 2$

$D_2 = f_{Mo} - f_{post} \Rightarrow D_2 = 11 - 8 \Rightarrow D_2 = 3$

E como:

$$Mo = LI + h\frac{D_1}{D_1 + D_2}$$

Temos:

$$Mo = 158 + \frac{2}{2+3}x4 = 158 + \frac{2x4}{5} = 158 + \frac{8}{5} = 158 + 1{,}6 = 159{,}6$$

Logo: $Mo = 159{,}6\ cm$

e) Os Quartis

Denominamos quartis os valores de uma série que a dividem em quatro partes iguais.

Há, portanto, três quartis:

- O primeiro quartil (Q_1) que é o valor que está situado de tal modo na série que uma quarta parte (25%) dos dados é menor e as três quartas partes restantes (75%) maiores do que ele;
- O segundo quartil (Q_2) que é, evidentemente, coincidente com a mediana (Q_2 = Md);
- O terceiro quartil (Q_3), que é o valor situado de tal sorte que as três quartas partes (75%) dos termos são menores e uma quarta parte (25%), maior que ele.

Quando os dados são agrupados para determinar os quartis, usamos a mesma técnica do cálculo da mediana, bastando substituir, na fórmula da mediana,

$\frac{\sum_{i=1}^{n} F_i}{2}$ por $k\frac{\sum_{i=1}^{n} F_i}{4}$

Sendo k o número de ordem do quartil.

Assim, temos:

$$Q_k = LI + h\,\frac{\left[k\,\frac{\sum_{i=1}^{n} F_i}{4} - f_{ant}\right]}{f_Q}$$

Exemplo: Tomemos a distribuição relativa à tabela da estatura dos alunos:

i	Estaturas (cm)	F_i	F_{ac}	
1	150 |---- 154	4	4	
2	**154 |---- 158**	**9**	**13**	→ Q_1
3	158 |---- 162	11	24	
4	**162 |---- 166**	**8**	**32**	→ Q_2
5	166 |---- 170	5	37	
6	170 |---- 174	3	40	
		Σ = 40		

Fonte: dados hipotéticos

Primeiro quartil

Temos:

$$\frac{\sum_{i=1}^{n} F_i}{4} = \frac{40}{4} = 10$$

$$Q_1 = 154 + 4\frac{(10-4)}{9} = 154 + \frac{24}{9}$$

$$= 154 + 2.66 = 156{,}66$$

$$Q_1 = 156{,}66$$

Terceiro quartil

Temos:

$$3\frac{\sum_{i=1}^{n} F_i}{4} = 3\frac{40}{4} = 30$$

$$Q_3 = 162 + 4\frac{(30-24)}{8} = 162 + \frac{24}{8}$$

$$= 162 + 3 = 165$$

$$Q_3 = 165$$

f) Os Percentis

Denominamos percentis aos noventa e nove valores que separam uma série em 100 partes iguais.

O cálculo de um percentil segue a mesma técnica do cálculo da mediana, porém, a fórmula:

$$\frac{\sum_{i=1}^{n} F_i}{2} \text{ será substituída por } k\frac{\sum_{i=1}^{n} F_i}{100}$$

sendo k o número de ordem do percentil.

Assim, para o k-ésimo percentil, temos:

$$P_k = LI + h\frac{k\frac{\sum_{i=1}^{n} F_i}{100} - f_{ant}}{f_p}$$

Exemplo: Considerando a distribuição relativa à tabela da estatura dos alunos, temos para oitavo percentil:

$$k = 8 \rightarrow 8\ x\ \frac{\sum_{i=1}^{n} F_i}{100} = 8\ x\ \frac{40}{100} = 3{,}2$$

Logo:

$$P_8 = 150 + \frac{(3{,}2 - 0)x4}{4} = 150 + \frac{12{,}58}{4} = 150 + 3{,}2 = 153{,}2$$

Então, $P_8 = 153{,}2\text{cm}$

Utilização das Medidas de Tendência Central

Na maioria das situações não necessitamos calcular as três medidas de tendência central, normalmente precisamos de apenas uma das medidas para caracterizar o centro da série. A medida ideal em cada caso é aquela que melhor representa a maioria dos dados da série. Quando houver forte concentração de dados na área central da série, devemos optar pela média. Quando houver forte concentração de dados no início e no final da série, devemos optar pela mediana. A Moda deve ser a opção como medida de tendência central apenas em séries que apresentam um

elemento típico, isto é, um valor cuja freqüência é muito superior à freqüência dos outros elementos da série.

2.2.2. Medidas de Variabilidade ou Dispersão

Raramente uma única medida é suficiente para descrever de modo satisfatório um conjunto de dados. Tomemos como exemplo o caso da média aritmética, que é uma medida de locação largamente empregada, e consideremos dois conjuntos de observações:

A: 25 28 31 34 37 B: 17 23 30 39 46

Ambos têm a mesma média, 31. No entanto, percebe-se intuitivamente que o conjunto B acusa dispersão muito maior do que o conjunto A. Torna-se então necessário estabelecer medidas que indiquem o grau de dispersão, ou variabilidade, em relação ao valor central.

As medidas de dispersão são medidas que mostram o grau de concentração os dados em torno da média. As principais medidas de dispersão são: variância, desvio padrão e coeficiente de variação.

a) Variância

É a soma dos quadrados dos desvios em relação à média. Com ela estabeleceremos uma medida de variabilidade para um conjunto de dados. É denotada por S^2 no caso amostral ou σ^2 no caso populacional.

Para Dados Brutos:

$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1} = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}\right]$$

Para Dados Agrupados em Intervalos de Classe:

$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2 F_i}{n-1} = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 f_i - \frac{(\sum_{i=1}^{n} x_i F_i)^2}{n}\right]$$

Variância Populacional:

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{N}$$

b) Desvio-padrão

É a raiz quadrada positiva da variância, representado por *S* ou *DP* no caso amostral ou σ no caso da população.

$$\sigma = \sqrt{\sigma^2}, para\ população.$$

$$S = \sqrt{S^2}, para\ amostra.$$

Exemplo: Calcule a variância e o desvio padrão da série abaixo, representativa de uma população.

x_i	F_i	x_iF_i	$(x_i - \overline{x})^2$	$(x_i - \overline{x})^2F_i$
2	3	6	2,72	8,17
3	5	15	0,42	2,11
4	8	32	0,12	0,98
5	4	20	1,82	7,29
$\sum$	20	73	-	18,55

Primeiro, calculamos a média:

$$\overline{X} = \frac{\sum_{i=1}^{n} x_iF_i}{\sum_{i=1}^{n} F_i} = \frac{73}{20} = 3{,}65$$

Como estamos trabalhando com uma população a variância é dada por:

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{N} = \frac{18{,}55}{20} = 0{,}9275$$

O desvio padrão será:

$$\sigma = \sqrt{0{,}9275} = 0{,}963$$

c) Coeficiente de Variação de Pearson

Por vezes é conveniente exprimir a variabilidade em termos relativos, isto porque, por exemplo, um desvio padrão de 10 pode ser insignificante se a observação típica é 10.000, mas altamente significativo para uma observação típica de 100.

Toma-se então uma medida relativa da variabilidade, comparando o desvio padrão com a média. Esta medida é o Coeficiente de Variação.

$$CV = \frac{Desvio\ Padrão\ Amostral}{Média\ Amostral}\ x\ 100\% \rightarrow CV = \frac{S}{\overline{X}}\ x\ 100\%$$

Já vimos que o desvio padrão tem a mesma unidade de medida que os dados, de modo que o coeficiente de variação é adimensional.

A grande utilidade do coeficiente de variação é permitir a comparação da variabilidade de diferentes conjuntos de dados.

Se: $CV \leq 15\%$ $\Rightarrow$ Baixa dispersão – Homogênea, estável, regular.

$15\% < CV < 30\%$ $\Rightarrow$ Média dispersão.

$CV > 30\%$ $\Rightarrow$ Alta dispersão – Heterogênea.

Exemplo: Dois grupos de 50 alunos de Estatística foram submetidos a uma avaliação de probabilidade e o resultado foram os seguintes:

Grupo	Média das notas	Desvio-padrão	CV
A	6	2	$CV = \frac{S}{\overline{X}} = \frac{2}{6} = 0{,}33$
B	6,2	1,5	$CV = \frac{S}{\overline{X}} = \frac{1{,}5}{6} = 0{,}25$

Como podemos observar o grupo B apresentou um nível de dispersão menor do que o grupo A, para confirmar nossa análise utilizaremos o Coeficiente de Variação de Pearson conforme a tabela acima.

Exercícios propostos

1) As notas de 50 alunos em um teste foram:

   75 89 66 52 90 68 83 94 77 60 38 47 87 65 97 58 82 49 65 70 73 81 85 77 83 56 63 79 82 84 69 70 63 62 75 29 88 74 37 81 76 74 63 69 73 91 87 76 71 71.

   Calcule as medidas de locação a as medidas de dispersão.

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

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 par comparar o desempenho dos carros?

3) Em turma de 9 alunos, as notas em matemática e história foram:

Nº do aluno	1	2	3	4	5	6	7	8	9
Matemática	6	4	5	7	8	3	5	5	7
História	7	8	9	10	6	7	8	9	5

Em qual disciplina os alunos são mais consistentes?

4) Em cinco testes, um estudante obteve média 63,2 com desvio padrão 3,1. Outro estudante teve média 78,5 com desvio padrão 5,5. Qual dos dois é mais consistente?

5) No exercício 3, temos as notas de 9 alunos em matemática e história. Sabe-se que os critérios adotados em cada exame não são comparáveis, por isso decidiu-se usar o *desempenho relativo* em cada exame. Essa medida será obtida do seguinte modo:

   - Para cada exame serão calculados a média $\overline{X}$ e o desvio padrão *dp*(*X*);
   - A nota *X* de cada aluno será padronizada do seguinte modo:

$$Z = \frac{X - \overline{X}}{dp(X)}$$

   a) Calcule as notas padronizadas dos alunos em matemática;
   b) Com os resultados obtidos em (a), calcule a média $\overline{Z}$ e o *dp*(*Z*);
   c) Se alguma das notas padronizadas estiver acima de 2dp(Z) ou abaixo de -2dp(Z), esse aluno deve ser considerado um caso atípico. Existe algum nessa situação?
   d) O aluno nº 6 saiu-se relativamente melhor em história ou matemática?

3. Análise Combinatória

O cálculo efetivo da probabilidade de um evento depende freqüentemente do uso dos resultados da análise combinatória.

3.1. Princípio Fundamental da Contagem

Se determinada operação pode ser realizada de n_1 maneiras distintas, e se, para cada uma dessas maneiras, uma segunda operação pode ser realizada de n_2 maneiras distintas, então as duas operações podem ser realizadas, conjuntamente, de (n_1 x n_2) maneiras. Cada um dos modos de realização da primeira pode associar-se a cada um dos modos de realização da segunda.

Exemplo: Quantos números pares de 3 algarismos podemos formar com os algarismos 1, 4, 5, 8 e 9, se cada algarismo só puder figurar uma vez?

Solução: Como o número deve ser par, só há duas escolhas para o algarismo das unidades: 4 e 8. A cada uma dessas escolhas correspondem quatro escolhas do algarismo das centenas e três do algarismo das dezenas. Podemos, então, formar 2 x 4 x 3 = 24 números, que são:

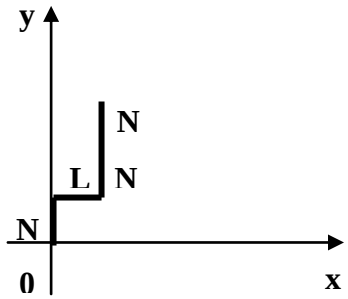
154 184 194 514 814 914 584 594
854 894 984 148 158 198 418 518
918 458 498 548 598 948 958 954

Neste exemplo estamos aplicando o princípio fundamental da contagem estabelecido acima para uma operação constante de duas etapas, a uma operação com três etapas. As etapas são as escolhas dos algarismos que formarão os números.

Exemplo: Um homem encontra-se na origem de um sistema cartesiano ortogonal de eixos *0x* e *0y*. Ele pode dar um passo de cada vez, para norte (N) ou para leste (L). Quantas trajetórias ele pode percorrer, se der exatamente 4 passos?

Solução: Notemos que cada trajetória consiste em uma quádrupla ordenada (a_1, a_2, a_3, a_4) em que $a_1 \in \{N, L\}$, $a_2 \in \{N, L\}$, $a_3 \in \{N, L\}$ e $a_4 \in \{N, L\}$.

Por exemplo, (N, L, N, N) corresponde graficamente a:



Logo o princípio fundamental da contagem, o número de trajetórias (quádruplas ordenadas) é 2 x 2 x 2 x 2 = 16.

3.2. Arranjos e Permutações

Dado o conjunto de *n* objetos, o número de disposições desses elementos tomados *k* de cada vez, constitui o que chamamos arranjos de *n* elementos *k* a *k* ($1 \leq k \leq n$). Símbolo: A_n^k.

Os arranjos distinguem-se entre si, não só pela natureza dos elementos que os compõem, mas também pela ordem em que figuram.

Exemplo: Os arranjos das quatro letras *a*, *b*, *c*, *d* tomadas 3 a 3 são:

Abc acb bca Bac cab cba
abd adb bda Bcd dab dba
acd adc cad Cda dac bdc
bdd bdc cbd CDB dcb

Pode-se mostrar que o número de arranjos de *n* elementos tomados *k* de cada vez é

$$A_n^k = \frac{k!}{(n-k)!}$$

O símbolo *n*! chamado fatorial de *n*, é o produto dos *n* primeiros números inteiros consecutivos, de 1 até *n*. Assim, 3! = 1 x 2 x 3 = 6; 7! = 1 x 2 x 3 x 4 x 5 x 6 x 7 = 5040; etc. Por convenção, 0! = 1.

Em particular $\begin{cases} A_n^1 = n \ \forall \ n \in \mathrm{N}^* \\ \dfrac{n!}{(n-1)!} = n \ \forall \ n \in \mathrm{N}^*, \end{cases}$

e a fórmula $A_n^k = \dfrac{n!}{(n-k)!}$ é valida $\forall \ n \in \mathrm{N}^*, \ \forall \ k \in \mathrm{N}^* \ com \ k \leq n.$

Exemplo: Quantos números ímpares de três algarismos podemos formar com algarismos significativos, sem os repetir?

Solução: O número há de terminar em 1 ou 3 ou 5 ou 7 ou 9. Em quaisquer hipóteses, restam oito algarismos que podem ser escolhidos 2 a 2 de A_8^2 maneiras. Podemos formar:

$$5A_8^2 = 5x\frac{8!}{6!} = 280 \ números$$

Note que:

$$A_8^2 = \frac{8!}{6!} = \frac{8x7x6x5x4x3x2x1}{6x5x4x3x2x1} = 8x7 = 56$$

Se os arranjos abrangem a totalidade dos elementos, temos o que se chama permutação de *n* elementos. Símbolo: P_n. Obviamente, $P_n = n!$

Em particular $\begin{cases} P_1 = 1 \\ 1! = 1 \end{cases}$

e a fórmula $P_n = n!$ é válida $\forall \ n \in \mathrm{N}^*$

Exemplo: Quantos números de cinco algarismos podemos formar com os algarismos 2, 3, 7, 8 e 9, sem os repetir?

Solução: Considerando todas as ordenações possíveis daqueles cinco algarismos, Temos:

$$5! = 1 \ x \ 2 \ x \ 3 \ x \ 4 \ x \ 5 = 120$$

3.3. Combinações

Há casos em que só interessam os elementos que compõem o grupamento, não importando a ordem em que ali figuram. Temos então o que se chama combinações de *n* elementos *k* a *k*. Símbolo:

$$C_n^k = \binom{n}{k} = \frac{n!}{k!(n-k)!} \qquad \forall\, n, k \in \mathrm{N}^*\ k < n$$

Casos particulares:

1º *caso*: $n,\ k \in \mathrm{N}^*\ e\ k = n$

$$\begin{cases} C_n^n = 1 & \text{(O único subconjunto com 0 elemento é o vazio)} \\ \dfrac{n!}{n!(n-n)!} = 1 \end{cases}$$

2º *caso*: $n \in \mathrm{N}^*\ e\ k = 0$

$$\begin{cases} C_n^0 = 1 & \text{(O único subconjunto do conjunto vazio é o próprio conjunto)} \\ \dfrac{n!}{0!\,(n-0)!} = 1 \end{cases}$$

3º *caso*: $n = 0\ e\ k = 0$

$$\begin{cases} C_0^0 = 1 \\ \dfrac{0!}{0!\,(0-0)!} = 1 \end{cases}$$

Dados os casos particulares, conclui-se que a fórmula

$C_n^k = \binom{n}{k} = \dfrac{n!}{k!(n-k)!}$ é válida $\forall\ n,\ k \in \mathrm{N}\ k \leq n$.

Os grupamentos básicos do exemplo das letras**,** *abc*, *abd*, *acd*, *bcd*, são precisamente às combinações dos quatro elementos *a*, *b*, *c*, *d* tomados 3 a 3. Ali, qualquer permutação que façamos com os elementos de determinado agrupamento origina a mesma combinação. São idênticas as combinações *abc* e *acb*, *bcd* e *cbd*, etc.

Vê-se que, para determinar o número de combinações de *n* elementos *k* a *k*, basta considerar o número de arranjos desses elementos de cada grupamento:

$$C_n^k = \frac{1}{k!} x \frac{n!}{(n-k)!} = \frac{n!}{k!\,(n-k)!}$$

Exemplo: De um grupo de sete indivíduos, quantas comissões de três elementos podemos formar?

Solução: Evidentemente, só interessam os indivíduos em si, e não a ordem em que os consideramos. Temos, pois, um problema de combinações de 7 elementos 3 a 3 cujo número é:

$$\binom{7}{3} = \frac{7!}{3!(7-3)!} = 35$$

Exemplo: De um grupo de sete indivíduos, quantas comissões podemos formar, compostas de um presidente, um tesoureiro e um secretário?

Solução: Aqui já não interessam apenas os indivíduos em si, mas também os cargos que vão ocupar (ou seja, a ordem em que os consideramos). Uma comissão em que João é presidente, José é tesoureiro e Alberto secretário é diferente da comissão composta dos mesmos indivíduos, mas em que José é presidente, Alberto tesoureiro e João secretário. Estamos em face de um problema de arranjos de 7 elementos 3 a 3. Seu número é:

$$A_7^3 = \frac{7!}{(7-3)!} = 210$$

É o número de combinações de 7 elementos 3 a 3 do exemplo anterior multiplicado pelo número de permutações dos 3 elementos de cada agrupamento:

$$35x3! = 210$$

Exemplo: De quantas formas podemos escolher 4 cartas de um baralho de 52 cartas, sem levar em conta a ordem delas, de modo que em cada escolha haja pelo menos um rei?

Solução: Como não levamos em conta a ordem das cartas, cada escolha é uma combinação. O número total de combinações é $\binom{52}{4}$. O número de combinações em que não comparece o rei é $\binom{48}{4}$. Logo a diferença $\binom{52}{4}$ - $\binom{48}{4}$ é o número de combinações em que comparece ao menos um rei.

4. Probabilidade

4.1. Experimento Aleatório

Em quase tudo, em maior ou menor grau, vislumbramos o **acaso**. Assim, da afirmação "é provável que meu time ganhe a partida hoje" pode resultar:

a. Que, apesar do favoritismo, ele perca;

b. Que, como pensamos, ele ganhe;

c. Que empate.

Como vimos, o resultado final depende do **acaso**. Fenômenos como esse são chamados de **fenômenos aleatórios** ou **experimentos aleatórios**.

Experimentos ou fenômenos aleatórios são aqueles que, mesmo repetidos várias vezes sob condições semelhantes, apresentam resultados imprevisíveis.

4.2. Espaço Amostral

A cada experimento correspondem, em geral, vários resultados possíveis. Assim, ao lançarmos uma moeda, há dois resultados possíveis: ocorrer cara ou ocorrer coroa. Já ao lançarmos um dado há seis resultados possíveis: 1, 2, 3, 4, 5 ou 6.

Ao conjunto desses resultados possíveis damos o nome de **espaço amostral** ou **conjunto universo**, representados por Ω.

Os dois experimentos citados anteriormente têm os seguintes espaços amostrais:

- lançamento de uma moeda: Ω = {Cara, Coroa};

- lançamento de um dado: Ω = {1, 2, 3, 4, 5, 6}.

Do mesmo modo, como em **dois** lançamentos sucessivos de uma moeda podemos obter **cara** nos dois lançamentos, ou **cara** no primeiro e **coroa** no segundo, ou **coroa** no primeiro e **cara** no segundo, ou **coroa** nos dois lançamentos, o espaço amostral é:

Ω = {(Cara, Cara), (Cara, Coroa), (Coroa, Cara), (Coroa, Coroa)}.

Cada um dos elementos de Ω que corresponde a um resultado recebe o nome de ponto amostral. Assim:

$$\{(\text{Cara, Cara})\} \subset \Omega \Rightarrow (\text{Cara, Cara}) \text{ é um ponto amostral de } \Omega.$$

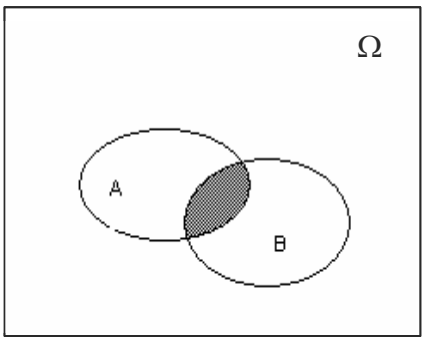
4.3. Eventos

Chamamos de eventos a qualquer subconjunto do espaço amostral Ω de um experimento aleatório.

Operações com Eventos

Interseção

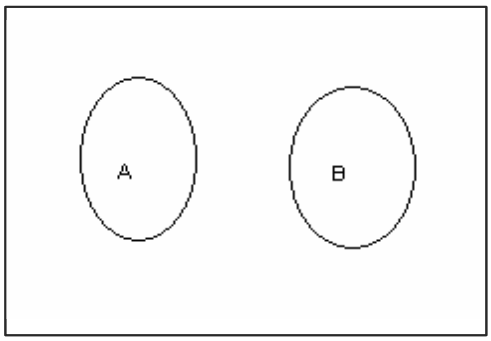
O evento interseção de dois eventos A e B equivale à ocorrência de ambos. Ela contém todos os pontos do espaço amostral comum a A e a B. Denota-se por $A \cap B$ (ou às vezes, por AB). A interseção é ilustrada pela área hachurada do diagrama abaixo.



Exemplo: Seja A o conjunto de alunos de uma instituição que freqüentam o curso secundário, e B o conjunto dos que freqüentam um curso facultativo de interpretação musical. A interseção $A \cap B$ é o conjunto dos alunos que fazem o curso secundário e freqüentam o curso facultativo.

Exclusão

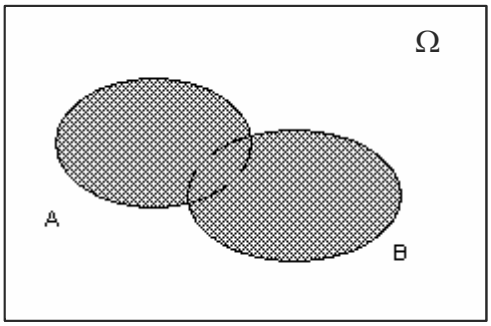
Dois eventos A e B dizem-se mutuamente exclusivos ou mutuamente excludentes quando a ocorrência de um deles impossibilita a ocorrência do outro. Os eventos não têm nenhum elemento em comum. Exprime-se isto escrevendo $A \cap B = \varnothing$. O diagrama a seguir ilustra esta situação.



Exemplo: Na jogada de um dado, seja A o evento "aparecer número par" e B o evento "aparecer número ímpar". A e B são mutuamente excludentes; $A \cap B = \varnothing$; nenhum número pode ser par e ímpar ao mesmo tempo.

União

O evento *união* de A e B equivale à ocorrência de A, *ou* de B, *ou* de ambos. Contém os elementos do espaço amostral que estão em pelo menos um dos dois conjuntos. Denota-se por $A \cup B$. A área hachurada do diagrama ilustra a situação.

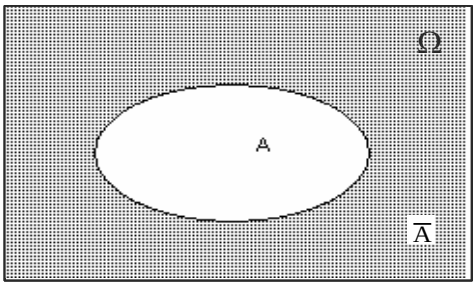


Nota-se que à interseção está associada à conjunção *e*, enquanto que à união está associada à conjunção *ou*.

Exemplo: Se A é o conjunto dos alunos de um estabelecimento que freqüentam o curso de ciências contábeis e B é o conjunto de aluno do mesmo estabelecimento que fazem administração de empresas, então $A \cup B$ é o conjunto dos alunos que fazem pelo menos um daqueles dois cursos.

Negação (Complementar)

A negação do evento A, denotada por $\overline{A}$ é chamada de evento complementar de A. É ilustrada na parte hachurada.

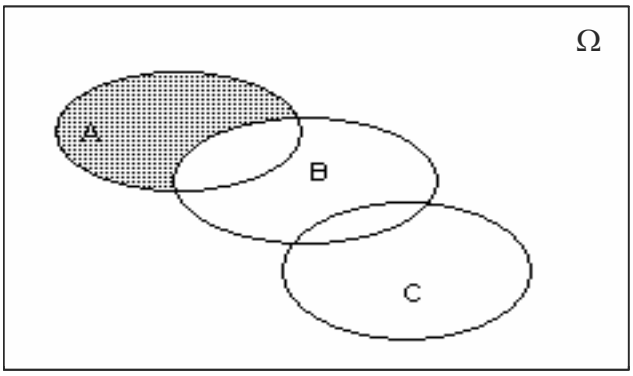


Exemplo: Se, na jogada de um dado, o evento E_1 consiste no aparecimento das faces 1, ou 2, ou 5, ou 6. Então: $E_1 = \{1, 2, 5, 6\}$ e $\overline{E_1} = \{3, 4\}$

Exemplo: Sejam A, B e C eventos arbitrários. Exprimir, em notação de conjuntos, os eventos: (a) apenas A ocorre, (b) todos os três ocorrem, (c) ao menos dois ocorrem.

Solução:

(a) Se só A ocorre, então B não ocorre, C não ocorre. O evento é representado por $A \cap \overline{B} \cap \overline{C}$. É ocorrência simultânea, ou interseção, de A, $\overline{B}$, $\overline{C}$. Ilustração a seguir.



(b) $A \cap B \cap C$.

(c) $(A \cap B \cap \overline{C}) \cup (A \cap \overline{B} \cap \overline{C}) \cup (\overline{A} \cap B \cap C) \cup (A \cap B \cap C)$. Isto é, ocorrem A, B, $\overline{C}$, ou A $\overline{B}$ C, ou $\overline{A}$, B, C ou A, B, C. É uma união de interseções.

Probabilidade

Não é possível fazer inferências estatísticas sem utilizar alguns resultados da teoria das probabilidades. Esta teoria, embora intimamente associada à estatística,

tem suas características próprias. Ela procura quantificar as incertezas existentes em determinada situação, ora usando um número, ora uma função matemática.

Definimos probabilidade clássica como:

$$Probabilidade = \frac{nº\ de\ casos\ favoráveis}{nº\ de\ resultados\ possíveis}$$

Suponha o lançamento de um dado. Qual a probabilidade da face superior ser 6? O nº de resultados favoráveis é 1, uma vez que existe somente um 6. O nº total de resultados possíveis são 6 (1, 2, 3, 4, 5, 6). Então a probabilidade é 1/6.

Outra definição de probabilidade é da freqüência relativa de ocorrência de um evento em um grande nº de repetições. Utilizando o caso do dado, calculamos a probabilidade de aparecer 6 lançando o dado um grande número de vezes e então observando a proporção de vezes que o número 6 apareceu, esta proporção nos dará a probabilidade do nº da face superior ser 6.

Obviamente, ninguém calcula probabilidade desta maneira. Todavia existe um caso conhecido de Kerrick, que foi internado na Dinamarca durante a 2º Guerra, que realizou vários destes experimentos. Por exemplo, ele lançou uma moeda 10.000 vezes: Inicialmente a freqüência relativa de caras flutuou muito, mas finalmente convergiu para valores próximos de 0,5 com um valor de 0,507 no final do experimento.

A probabilidade freqüentista, dada por Poisson, diz que se n é o n.º de ensaios e n(E) o n.º de ocorrências do evento E, então a probabilidade de E, denotada por P(E) é:

$$P(E) = \lim_{n\to\infty} \frac{n(E)}{n}$$

Experimento	Tipo	Espaç$o amostral	Variáveis aleatórias
Lançar uma moeda três vezes	Discreto	$X = \{CCC, CCX, CXC, XXC, XXX\}$	Número de caras Número de lançamentos até aparecer cara
Escolher aleatoriamente uma amostra de três alunos da disciplina de estatística	Discreto	$X_1 = \{ABC, ACD, ABE, \dots\}$ $X_2 = \{AAA, AAB, AAC, \dots\}$	Número de alunos de sexo masculino

Lançar dois dados	Discreto	$X = \{(1,1),(1,2),\dots,(6,6)\}$	Soma dos valores das faces Diferença entre os valores das faces
Escolher aleatoriamente eleitores e perguntar em quem irão votar para presidente	Discreto	$X = \{Lula, FHC, CG, E, H\}$	Número de eleitores quem votarão no candidato X
Escolher aleatoriamente uma mulher e anotar o número de filhos vivos	Discreto Contínuo	$X = \{0,1,2,\dots\}$ $X = \{X; 12 \leq X \leq 50\}$	Número de filhos vivos por mulher Idade em que engravidou a primeira vez
Observar o número de casos de meningite por mês	Discreto	$X = \{0,1,2,\dots\}$	Número de casos de meningite por mês Número de casos por sexo, por faixa etária, ...
Aplicar a escala de atitudes frente à Matemática e observar a pontuação	Contínuo	$X = \{X; 20 \leq X \leq 80\}$	Valor na escala de atitudes
Observar a reprovação em Matemática dos alunos de 5ª séries por turma	Contínuo	$X = \{X; 0\% \leq X \leq 100\%\}$	Porcentagem de alunos reprovados por turma
Observar o tempo de vida (até queimar) de uma lâmpada	Contínuo	$X = \{X; X \geq 0\}$	Tempo de vida da lâmpada (em horas)
Observar a quantidade de chuva mensal	Contínuo	$X = \{X; 0 \leq X \leq M\}$ M suficientemente grande, porém limitado.	Quantidade de chuva mensal (em mm)

2.1.1. Regras de Probabilidade

Independente do ponto de vista de probabilidade (clássico ou freqüentista) as regras para o cálculo de probabilidade são as mesmas. Antes das regras precisamos de algumas definições. Eventos A_1, A_2, A_3,... são ditos **mutuamente exclusivos** se, quando um ocorre os outros não ocorrem. Eles são ditos exaustivos se exaurem todas as

possibilidades. No caso do lançamento de um dado, os eventos A_1, A_2, A_3,..., A_6 de que o dado mostre 1, 2, 3, 4, 5 e 6 são mutuamente exclusivos e exaustivos.

Podemos escrever $P(A \cup B)$ como a probabilidade de que os eventos A ou B ou ambos ocorram, a isto denominamos união de eventos, neste caso união de A e B.

Escrevemos $P(A \cap B)$ como a probabilidade da ocorrência conjunta de A e B, e denominamos de interseção dos eventos A e B.

Propriedades da Probabilidade

Sendo A um evento qualquer.

- $P(A) = 1 - P(\overline{A})$
- $0 \leq P(A) \leq 1$

Sendo A e B dois eventos quaisquer, temos:

- $P(B) = P(B \cap A) + P(B \cap \overline{A})$
- Se $A \subset B$ então $P(A) \leq P(B)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Exemplo: Sejam os eventos

A: O dado mostra 1, 3 ou 5

B: O dado mostra 3

Então

$A \cup B$: O dado mostra 1, 3 ou 5

$A \cap B$: O dado mostra 3

A regra de adição de probabilidade afirma que

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

Se A e B são mutuamente exclusivos não podem ocorrer conjuntamente, assim $P(A \cap B) = 0$. Então para eventos mutuamente exclusivos

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

Se, em adição, A e B são exaustivos, $P(A) + P(B) = 1$.

Nós denotamos por $\overline{A}$ o complementar de A. $\overline{A}$ representa a não ocorrência de A. Porque A ocorre ou não (isto é, $\overline{A}$ ocorre), A e $\overline{A}$ são mutuamente exclusivos e exaustivos.

Então $P(A) + P(B) = \ ou\ P(\overline{A}) = 1 - P(A)$

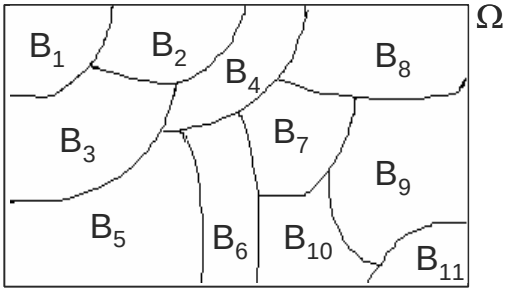
Teorema da Probabilidade Total

Inicialmente, consideremos *n* eventos B_1, B_2, ..., B_n. Diremos que eles formam uma partição do espaço amostral Ω, quando:

1) $P(B_k) > 0\ \forall\ k;$
2) $B_i \cap B_j = \emptyset\ para\ i \neq j;$
3) $\bigcup_{i=1}^{n} B_i = \Omega$

Isto é, os eventos B_1, B_2, ..., B_n são dois a dois mutuamente exclusivos e exaustivos (sua união é Ω).

Ilustração para *n* = 11:

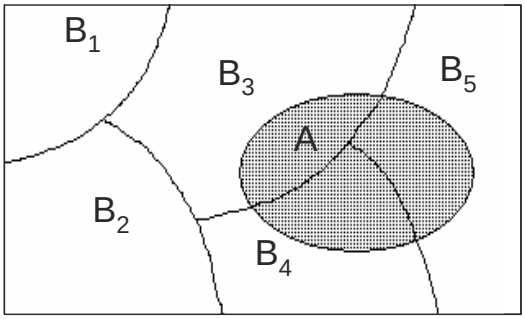


Seja Ω um espaço amostral, A um evento qualquer de Ω e B_1, B_2, ..., B_n uma partição de Ω.

É válida a seguinte relação:

$A = (B_1 \cap A) \cup (B_2 \cap A) \cup (B_3 \cap A) \cup ... \cup (B_n \cap A)$.

A figura ilustra o fato para *n* = 5



Nesse caso:

$A = (B_1 \cap A) \cup (B_2 \cap A) \cup (B_3 \cap A) \cup ... \cup (B_5 \cap A)$.

Notemos que $(B_1 \cap A) \cup (B_2 \cap A) \cup ... \cup (B_n \cap A)$ são dois a dois mutuamente exclusivos, portanto:

$P(A) = P(B_1 \cap A) + P(B_2 \cap A) + ... + P(B_n \cap A)$.

Exemplo: Na tabela abaixo temos dados referentes a alunos matriculados em quatro cursos de uma universidade em dado ano.

Tabela: Distribuição de alunos segundo sexo e escolha de curso.

Sexo / Curso	Homens (H)	Mulheres (F)	Total
Matemática Pura (M)	70	40	110
Matemática Aplicada (A)	15	15	30
Estatística (E)	10	20	30
Computação (C)	20	10	30
Total	115	85	200

Vamos indicar por M o evento que ocorre quando, escolhendo-se ao acaso um aluno do conjunto desses quatro cursos, ele for estudante de Matemática Pura. A, E, C, H e F têm significados análogos. Dessa maneira, vemos que $P(E) = 30/200$, ao passo que $P(H) = 115/200$

Dados os eventos A e H, podemos considerar dois novos eventos:

$A \cup H$, chamado a *reunião* de A e H, quando pelo menos um dos eventos ocorre;

$A \cap H$, chamado a *intersecção* de A e H, quando A e H ocorrem simultaneamente.

É fácil ver que $P(A \cap H) = 15/200$, pois o aluno escolhido terá de estar, ao mesmo tempo, matriculado no curso de matemática Aplicada e ser homem.

Vemos que $P(A) = 30/200$ e $P(H) = 115/200$; suponha que nosso cálculo para $P(A \cup H)$ fosse:

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

$$P(A \cup H) = \frac{30}{200} + \frac{115}{200} = \frac{145}{200}$$

Se assim o fizéssemos, estariamos contando duas vezes os alunos que são homens e estão matriculados no curso de Matemática Aplicada, como destacado na Tabela. Portanto a resposta correta é

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

$$P(A \cup H) = \frac{30}{200} + \frac{115}{200} - \frac{15}{200} = \frac{130}{200}$$

No entanto, considerando-se os eventos A e C, vemos que $P(A) = 30/200$, $P(C) = 30/200$ *e* $P(A \cup C) = 60/200 = P(A) + P(C)$. Neste caso, os eventos A e C são disjuntos ou *mutuamente exclusivos*, pois se A ocorre, então C não ocorre e vice-versa.

Exemplo: Uma urna contém 100 bolinhas numeradas, de 1 a 100. Uma bolinha é escolhida e observa-se seu número. Admitindo probabilidades iguais a $\frac{1}{100}$ para todos os eventos elementares, qual a probabilidade de:

a) Observarmos um múltiplo de 6 e de 8 simultaneamente?
b) Observarmos um múltiplo de 6 ou de 8?
c) Observarmos um número não múltiplo de 5?

Solução: temos Ω = {1, 2, 3, ..., 99, 100}

a) Um múltiplo de 6 e 8 simultaneamente terá que ser múltiplo de 24; portanto, o evento que nos interessa é: A = {24, 48, 72, 96}.

$$P(A) = \frac{1}{100} + \frac{1}{100} + \frac{1}{100} + \frac{1}{100} = \frac{4}{100} = \frac{1}{25}$$

b) Sejam os eventos:

B: o número é múltiplo de 6.

C: o número é múltiplo de 8.

O evento que nos interessa é $B \cup C$, então:

B = {6, 12, 18, 24, 30, 36, 42, 48, 54, 60, 66, 72, 78, 84, 90, 96 }

e $P(B) = \frac{16}{100} = \frac{4}{25}$

C = {8, 16, 24, 32, 40, 48, 56, 64, 72, 80, 88, 96}

e $P(C) = \frac{12}{100} = \frac{3}{25}$

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

Ora, B $\cap$ C nada mais é do que o evento A (do item a).

Logo, P(B ∩ C) = $\frac{1}{25}$.

Segue-se então que: P(B ∪ C) = $\frac{4}{25}+\frac{3}{25}-\frac{1}{25}=\frac{6}{25}$.

c) Seja D o evento, o número é múltiplo de 5.

Temos:

D = {5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100}

P(D) = $\frac{20}{100}=\frac{1}{5}$.

O evento que nos interessa é $\overline{D}$. Logo,

$$P(\overline{D}) = 1 - P(D) = 1 - \frac{1}{5} = \frac{4}{5}$$

2.1.2. Probabilidade Condicional e Regra Da Multiplicação

Às vezes, nós restringimos nossa obtenção ao subconjunto de todos os eventos possíveis. Por exemplo, suponha que ao lançarmos um dado, os casos 1, 2 e 3 não sejam levados em consideração; considere o evento B o dado mostrar 4, 5 ou 6. Considere o evento A de que o dado mostre 6. A probabilidade de A é agora 1/3 porque o número total de resultados é 3 e não 6. A probabilidade condicional é definida como segue: A probabilidade de um evento A dado que outro evento B ocorreu, é denotada por $P(A|B)$ e é definido por

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

No caso acima, $P(A \cap B) = 1/6,\ P(B) = 3/6, e\ então\ P(A|B) = 1/3$

Definiremos agora eventos independentes. A e B são ditos independentes se, a probabilidade de ocorrência de um não depende se o outro ocorreu ou não.

Então se A e B são independentes, as probabilidades condicional e não-condicional são as mesmas, isto é, $P(A|B) = P(A)$ e $P(B|A) = P(B)$. Porque $P(A|B) = P(A \cap B)/P(B)$ e $P(A|B) = P(A)$ temos a **regra da multiplicação**, que diz que: $P(A \cap B) = P(A)P(B)$ se A e B são independentes.

Como exemplo, suponha o lançamento do dado duas vezes:

A = evento de que o 1º lançamento mostre um 6

B = evento de que o 2º lançamento mostre um 6

Claramente, A e B são eventos independentes. Então a probabilidade (de conseguirmos duplo 6) é $P(A \cap B) = P(A)P(B) = 1/36$

Independência de eventos

Dado dois eventos A e B de um espaço amostral Ω, diremos que A *independe de* B se:

P(A | B) = P(A)

Isto é, A *independe de* B se a ocorrência de B não afeta a probabilidade de A. Observemos que, se A *independe de* B (P(A) > 0), então B *independe de* A, pois:

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

Em suma, se A independe de B, então B independe de A e, além disso:

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

Isso sugere a definição:

Dois eventos A e B são chamados independentes se

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

Exemplo: Existem 100 baterias em uma caixa, 60 delas são baterias de 12 volts e 40 delas com menos de 12 volts. 20 das baterias de 12 volts têm conexões duplas; 5 das baterias com menos de 12 volts têm conexões duplas.

Baterias na caixa	12 volts	<12volts	Total
Conectores Duplos	20	5	25
Conectores Simples	40	35	75
TOTAL	60	40	100

- ✓ Qual é a probabilidade que uma bateria selecionada ao acaso da caixa seja de 12 volts? p = 60/100 = 0,6
- ✓ Qual é a probabilidade que uma bateria selecionada de 12 volts também tenha conector duplo? p = 20/60 = 1/3
- ✓ Qual é a probabilidade que uma bateria selecionada seja de 12 volts e tenha conector duplo? p = 20/100 = 1/5

Exemplo: Uma urna I contém 3 bolas vermelhas e 4 brancas, a urna II tem 6 vermelhas e 2 brancas. Uma urna é escolhida ao acaso e nela é escolhida uma bola, também ao acaso.

a) Qual probabilidade de observarmos urna I e bola vermelha?

b) Qual probabilidade de observarmos bola vermelha?

c) Se a bola observada foi vermelha, qual probabilidade que tenha vindo da urna I?

Solução:

Sejam U_I o evento sair urna I, U_{II} o evento sair urna II e V o evento sair bola vermelha;

Então temos:

a) $P(U_I \cap V) = P(U_I)P(V|U_I) = \frac{1}{2}x\frac{3}{7} = \frac{3}{14}$

b) $P(V) = P(U_I \cap V) + P(U_{II} \cap V) = \frac{1}{2}x\frac{3}{7} + \frac{1}{2}x\frac{6}{8} = \frac{33}{56}$

c) Estamos interessados em $P(U_I|V)$. Por definição

$$P(U_I|V) = \frac{P(U_I \cap V)}{P(V)}$$

Usando os resultados dos itens a e b,

$$P(U_I|V) = \frac{3/14}{33/56} = \frac{4}{11}$$

2.1.3. Teorema de Bayes

O Teorema de Bayes é baseado na probabilidade condicional.

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \; e \; P(B|A) = \frac{P(A \cap B)}{P(A)}$$

Escrevemos a 2º equação como $P(A \cap B) = P(B|A)P(A)$

Então,

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

é conhecido como **Teorema de Bayes.**

Seja H_1 e H_2, duas hipóteses e D os dados observados. Substituindo-se H_1 por A e D por B, teremos

$$P(H_1|D) = \frac{P(D|H_1)P(H_1)}{P(D)}$$

Analogamente substitui-se H_2 por A e D por B, teremos

$$P(H_2|D) = \frac{P(D|H_2)P(H_2)}{P(D)}$$

Daí,

$$\frac{P(H_1|D)}{P(H_2|D)} = \frac{P(D|H_1)P(H_1)}{P(D|H_2)P(H_2)}$$

A parte à direita na equação é chamada de "razão de chances a posteriori". O primeiro termo do lado direito é chamado "razão de verossimilhança" e o segundo termo do lado direito chamado "razão de chances à priori"

Faz-se uso destas equações nos problemas de escolha entre dois modelos.

Exemplo: Tem-se 2 urnas: a primeira tem 1 bola vermelha e 4 brancas, e a segunda tem 2 bolas vermelhas e 2 brancas. Uma urna é escolhida ao acaso e uma bola é retirada. A bola é branca. Qual é a probabilidade de que esta bola tenha vindo da primeira urna?

Defina:

H_1: A primeira urna foi escolhida

H_2: A segunda urna foi escolhida

D: Dado, isto é, a bola é branca.

Temos $P(H_1) = P(H_2) = 1/2$. Também $P(D|H_1) = 4/5$ e $P(D|H_2) = 1/2$.

então $P(H_1|D) = 8/5$ ou $P(H_1|D) = 8/13$ e $P(H_2|D) = 5/13$. A resposta é então $P(H_2|D) = 8/13$.

Como $P(H_1|D) = \frac{P(D|H_1)P(H_1)}{P(D)} = \frac{P(D|H_1)P(H_1)}{P(D|H_1)P(H_1) + P(D|H_2)P(H_2)}$

$$P(H_1|D) = \frac{(4/5).(1/2)}{[(4/5).(1/2)] + [(1/2).(1/2)]} = \frac{4/10}{[4/10] + [1/4]} = \frac{4/10}{13/20} = \frac{8}{13}$$

Exemplo: Para selecionar seus funcionários, uma empresa oferece aos candidatos um curso de treinamento durante uma semana. No final do curso, eles são submetidos a uma prova e 25% são classificados como bons (B), 50% como médios (M) e os

restantes 25% como fracos (F). Para facilitar a seleção, a empresa pretende substituir o treinamento por um teste contendo questões referentes a conhecimentos gerais e específicos. Para isso, gostaria de conhecer qual a probabilidade de um indivíduo aprovado no teste ser considerado fraco, caso fizesse o curso. Assim, neste ano, antes do início do curso, os candidatos foram submetidos ao teste e receberam o conceito aprovado (A) ou reprovado (R). No final do curso, obtiveram-se as seguintes probabilidades condicionais:

P(A|B) = 0,80; P(A|M) = 0,50; P(A|F) = 0,20

Queremos encontrar P(F|A) e, pelo Teorema de Bayes, essa probabilidade é dada por

$$P(F \mid A) = \frac{P(A|F)P(F)}{P(A|B)P(B)+P(A|M)P(M)+P(A|F)P(F)}$$
$$= \frac{(0,20)(0,25)}{(0,80)(0,25)+(0,50)(0,50)+(0,20)(0,25)} = 0,10$$

Então, apenas 10% dos aprovados é que seriam classificados como fracos durante o curso. De modo análogo podemos encontrar P(B|A) = 0,40 e P(M|A) = 0,50, que poderiam fornecer subsídios para ajudar na decisão de substituir o treinamento pelo teste.

Exemplo: A urna I contém 3 fichas vermelhas e 2 azuis, e a urna II contém 2 vermelhas e 8 azuis. Joga-se uma moeda "honesta". Se a moeda der cara, extrai-se uma ficha da urna I; se der coroa, extrai-se uma ficha da urna II. Uma ficha vermelha é extraída. Qual a probabilidade de ter saído cara no lançamento?

Solução:

Sejam V o evento sair bola vermelha, A o evento sair bola azul, C o evento sair cara e K o evento sair coroa.

Nas urnas I e II temos:

I: {3V, 2A} e II: {2V, 8A}

Queremos: P(C | V)

Onde, P(C) = $\frac{1}{2}$, P(K) = $\frac{1}{2}$, P(V | C) = $\frac{3}{5}$ e P(V | K) = $\frac{2}{10}$

Como:

$P(V) = P(C \cap V) + P(K \cap V)$

Temos:

$P(V) = P(C) \times P(V \mid C) + P(K) \times P(V \mid K)$

$$P(V) = \frac{1}{2} \times \frac{3}{5} + \frac{1}{2} \times \frac{2}{10} = \frac{4}{10}$$

Calculamos agora $P(C \mid V)$:

$$P(C \mid V) = \frac{P(C \cap V)}{P(V)} = \frac{3/10}{4/10} = \frac{3}{4}$$

Exercícios propostos

1) Em um grupo de 500 estudantes, 80 estudam Matemática, 150 estudam Estatística e 10 estudam Matemática e Estatística. Se o aluno é escolhido ao acaso, qual a probabilidade de que:

a) Ele estude Estatística e Matemática?

b) Ele estude somente Matemática?

c) Ele estude somente Estatística?

d) Ele não estude Matemática e nem Estatística?

e) Ele estude Matemática ou Estatística?

2) Nove livros são colocados ao acaso numa estante. Qual a probabilidade de que 3 livros determinados fiquem juntos?

3) De um grupo de 10 pessoas, entre elas Regina, cinco são escolhidas ao acaso e sem reposição. Qual a probabilidade de que Regina compareça entre as cinco?

4) Um grupo é constituído de 10 pessoas, entre elas Jonas e César. O grupo é disposto ao acaso em uma fila. Qual a probabilidade de que haja exatamente 4 pessoas entre Jonas e César?

5) Um homem encontra-se na origem de um sistema cartesiano ortogonal. Ele só pode andar uma unidade de cada vez, para cima ou para a direita. Se ele andar 10 unidades, qual a probabilidade de chegar no ponto P(7, 3)?

6) De um total de 100 alunos que se destinam aos cursos de Matemática, Física e Química sabe-se que:

   - 30 destinam-se à Matemática e, destes, 20 são do sexo masculino.
   - O total de alunos do sexo masculino é 50, dos quais 10 destinam-se à Química.

- Existem 10 moças que se destinam ao curso de Química.

Nessas condições, sorteando um aluno ao acaso do grupo total e sabendo que é do sexo feminino, qual a probabilidade de que ele se destine ao curso de Matemática?

7) O mês de outubro tem 31 dias. Numa certa localidade, chove 5 dias no mês de outubro. Qual a probabilidade de não chover nos dias 1º e 2 de outubro?
8) Em uma universidade, o número de homens é igual ao de mulheres. 5% dos homens são calouros e 0,25% das mulheres são calouros. Uma pessoa é selecionada ao acaso e verifica-se que é calouro. Qual a probabilidade de que ela seja mulher?
9) As probabilidades de que duas pessoas A e B resolvam um problema são: P(A) =1/3 e P(B) = 3/5. Qual a probabilidade de que:
   a) Ambos resolvam o problema?
   b) Ao menos um resolva o problema?
   c) Nenhum resolva o problema?
   d) "A" resolva o problema, mas "B" não?
   e) "B" resolva o problema, mas "A" não?
10) A probabilidade de um homem sobreviver mais 10 anos, a partir de uma certa data, é 0,4, e de que sua esposa sobreviva mais 10 anos a partir da mesma data é 0,5. Qual a probabilidade de:

a) Ambos sobreviverem mais 10 anos a partir daquela data?
b) Ao menos um deles sobreviver mais 10 anos a partir daquela data?

11) Um prédio de três andares, com dois apartamentos por andar, tem apenas três apartamentos ocupados. Qual a probabilidade de que um dos três andares tenha exatamente um apartamento ocupado?
12) Entre 20 universitários, apenas um cursou o ensino médio em escola pública. 10 universitários entre os 20 são escolhidos ao acaso. Qual a probabilidade do universitário que cursou o ensino médio em escola pública estar entre os 10?
13) Em uma mesa existem 100 provas, sendo 80 corrigidas. Se 5 provas forem escolhidas ao acaso, sem reposição, qual a probabilidade de 4 terem sido corrigidas?
14) Um colégio tem 1000 alunos. Destes: 200 estudam Matemática, 180 estudam Física, 200 estudam Química, 20 estudam Matemática, Física e Química,

50 estudam Física e Química, 70 estudam somente Química e 50 estudam Matemática e Física.

Um aluno do colégio é escolhido ao acaso. Qual a probabilidade de:

a) Ele estudar só Matemática?

b) Ele estudar só Física?

c) Ele estudar Matemática e Química?

15) Cinco algarismos são escolhidos ao acaso, com reposição, entre os algarismos 0, 1, 2, 3, 4, 5, 6, 7, 8, 9. Qual a probabilidade de os cinco algarismos serem diferentes?

16) Há 60 candidatos a um emprego. Alguns têm curso superior (*S*), outros não; alguns têm no mínimo três anos de experiência (*T*), outros não. A distribuição é:

	S	$\overline{S}$	Total
T	12	6	18
$\overline{T}$	24	18	42
Total	36	24	60

Se a ordem de entrevista é aleatória, *S* é o evento o primeiro entrevistado tem curso superior e *T* é o evento o primeiro entrevistado tem experiência mínima de três anos, calcule as probabilidades:

(a) P(*S*); (b) P(*T* | *S*); (c) P(S $\cap$T); (d) P($\overline{S}$ $\cap$T)

17) Sabe-se que um soro da verdade, quando ministrado a um suspeito, é 90% eficaz quando a pessoa é culpada e 99% eficaz quando é inocente. Em outras palavras, 10% dos culpados são julgados inocentes, e 1% dos inocentes é julgado culpado. Se o suspeito foi retirado de um grupo em que 95% jamais cometeram qualquer crime, e o soro indica culpado qual a probabilidade de o suspeito ser inocente?

4.4. Distribuição de Probabilidade

Variáveis Aleatórias e Distribuições de Probabilidade

A variável X é dita **variável aleatória** se para todo n.º real a existe uma probabilidade $P(X \leq a)$ que X assuma os valores menores ou iguais que a, ou seja, é a

variável que associa um número real ao resultado de um experimento aleatório. **Exemplo:** Valor de mercado de um lote com área de 250m^2 em Manaus.

Denotaremos variáveis aleatórias por letras maiúsculas X, Y, Z e assim por diante. Usaremos letras minúsculas x, y, z para denotar valores particulares de uma variável aleatória. Então $P(X = x)$ é a probabilidade de que a variável aleatória X assuma o valor, x. $P(x_1 \leq X \leq x_2)$ é a probabilidade de que a variável aleatória X assuma valores entre x_1 e x_2, inclusive ambos.

Se a variável aleatória X pode assumir somente um particular conjunto de valores (finito ou infinito enumerável), diz-se que é uma **variável aleatória discreta**.

Uma variável aleatória é dita **contínua** se pode assumir qualquer valor em certo intervalo.

Um exemplo de uma variável aleatória discreta é o número de consumidores chegando numa loja durante certo período (digamos, à primeira hora comercial). Um exemplo de variável aleatória contínua é a renda das famílias brasileiras. Na prática o uso de variáveis aleatórias contínuas é mais popular porque a teoria matemática é mais simples. Por exemplo, quando dizemos que renda é uma variável aleatória contínua (de fato, estritamente falando, ela é discreta) o fazemos porque tratá-la desta forma é uma conveniente aproximação.

A fórmula dando as probabilidades para diferentes valores da variável aleatória X é chamada **distribuição de probabilidade** no caso discreto e **função densidade de probabilidade (f. d. p.)** no caso de variável aleatória contínuas, usualmente denotadas por $f(x)$.

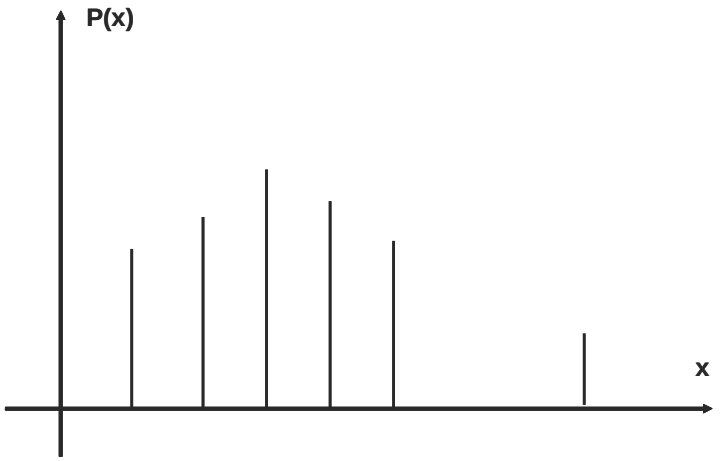


Figura: Distribuição de probabilidade de uma variável aleatória discreta.

Em geral, para variáveis aleatórias contínuas, a ocorrência de qualquer valor exato de X pode ser visto como tendo probabilidade zero. Probabilidades são discutidas em termos de intervalos. Estas probabilidades são obtidas por integração de $f(x)$ sobre o intervalo desejado. Por exemplo, se queremos $P(a \leq X \leq b)$ isto é dado por: $P(a \leq X \leq b) = \int_a^b f(x)\,dx$

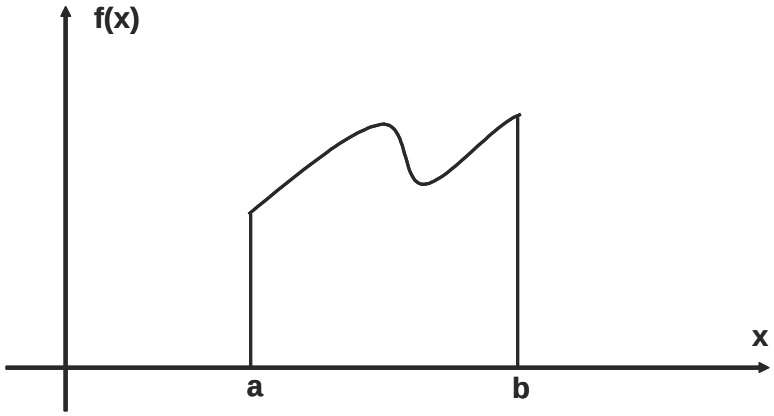


Figura: Distribuição de probabilidade de variável aleatória contínua.

A probabilidade de que a v. a. X assuma valores menores ou iguais a c é freqüentemente escrita como$F(c) = P(X \leq c)$. A função $F(x)$ representa, para diferentes valores de x a probabilidade acumulada e é chamada de **função distribuição acumulada.** Assim,

No caso discreto:

$$F(c) = P(X \leq c) = \sum_{t \leq c} f(t)$$

No caso contínuo:

$$P(X \leq c) = \int_{-\infty}^{c} f(x)\,dx$$

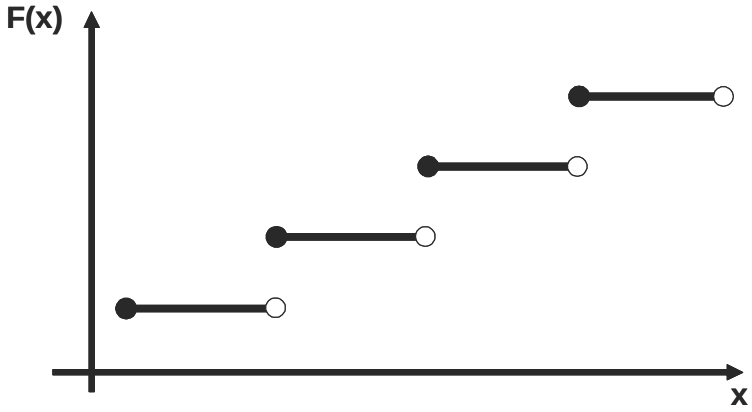


Figura: Distribuição acumulada de uma variável aleatória.

Esperança de uma Variável Aleatória

Sejam X e Y variáveis aleatórias. Uma função matemática que calcula a média de X ou de Y chama-se esperança ou valor esperado de X ou de Y é denotada por $E(X)$ou $E(Y)$, e satisfaz as seguintes propriedades:

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

(2) $E(aX) = aE(X), a \in \mathbb{R}$

Variância de uma Variável Aleatória

Seja X uma variável aleatória. Defini-se a variância de X (denotada por$Var(X)$), como: $Var(X) = E[X - E(X)]^2 = E(X^2) - [E(X)]^2$

A variância satisfaz as seguintes propriedades:

(1) $Var(aX) = a^2 Var(X), a \in \mathbb{R}$

(2) Se $X_1 e\ X_2$ são variáveis aleatórias independentes então,

$Var(X_1 + X_2) = Var(X_1) + Var(X_2)$

Esperança e Variância de uma variável aleatória discreta

Esperança de X no caso discreto; Seja $P(X_i) = P(X = x_i), onde\ i = 1,2,3 \ldots, n, \ldots$

$$E(X) = \sum_{i=1}^{\infty} x_i P(X_i)$$

Variância de X,

$$Var(X) = \sum_{i=1}^{\infty} [x_i - E(X)]^2 P(X_i) \; ou \; Var(X) = E(X^2) - [E(X)]^2$$

Onde:

$$E(X^2) = \sum_{i=1}^{\infty} x_i^2 P(X_i)$$

Esperança e Variância de uma variável aleatória contínua

Esperança de X no caso contínuo,

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

Variância de X,

$$Var(X) = \int_{-\infty}^{\infty} [x - E(X)]^2 f(x)\, dx \; \; ou \; Var(X) = E(X^2) - [E(X)]^2$$

Onde:

$$E(X^2) = \int_{-\infty}^{\infty} x^2 f(x)\, dx$$

4.4.1 Distribuições Discretas de Probabilidade

Algumas variáveis aleatórias adaptam-se muito bem a uma série de problemas práticos. Um estudo dessas variáveis é de grande importância para a construção de modelos probabilísticos para situações reais e a conseqüente estimação de seus parâmetros. Para algumas destas distribuições, existem tabelas que facilitam o cálculo das probabilidades em função dos seus parâmetros. Nesta seção estudaremos os dois modelos discretos mais importantes: a distribuição *binomial* e a distribuição de *Poisson*.

4.4.1.1 Distribuição Binomial

Uma das mais comuns em estatística. Deriva de um processo conhecido como teste de Bernoulli em que cada tentativa tem duas possibilidades excludentes de ocorrência chamada de sucesso e falha (ex. moeda).

Um experimento aleatório é chamado binomial se em *n* repetições:

1) Os ensaios são independentes;

2) Cada resultado do ensaio pode assumir somente uma de duas possibilidades:

sucesso ou fracasso;

3) A probabilidade de sucesso em cada ensaio, denotado por p, permanece constante.

O Processo de Bernoulli:

Definição:

Uma seqüência de testes de Bernoulli forma um Processo de Bernoulli, sob as seguintes condições:

a) Cada tentativa resulta em um de dois resultados mutuamente excludentes. Um dos resultados possíveis é chamado (arbitrariamente) de sucesso e o outro de falha;

b) A probabilidade de sucessos, denotada p, permanece constante em todas as tentativas. A probabilidade da falha, $1 - p$, é denotada por q;

c) As tentativas são independentes; isto é, o resultado de uma tentativa particular não é afetado pelos resultados das outras tentativas.

Assim, a probabilidade de obtermos exatamente X sucessos em n tentativas é a distribuição binomial:

$$P(X = x) = \binom{n}{x} p^x q^{n-x}, x = 0, 1, \dots, n$$

Com esperança e variância dada por:

$$E(X) = np \text{ e } Var(X) = npq$$

Exemplo: De um lote de produtos manufaturados, extraímos 100 itens ao acaso; se 10% dos itens do lote são defeituosos, calcule a probabilidade de 12 itens serem defeituosos.

Solução:

$$X \sim binomial(100; 0{,}1)$$

$$P(X = 12) = \binom{100}{12} (0{,}1)^{12} (0{,}9)^{88} = 0{,}0988$$

Exemplo: Na linha de produção de uma fábrica, em condições normais de funcionamento cada uma das peças pode ser considerada como produzida independentemente das demais. Se retirarmos uma amostra de *n* peças da linha de produção e se chamarmos de *p* a fração de peças defeituosas que são produzidas, então X, o número de peças defeituosas na amostra, é uma variável aleatória com distribuição binomial com parâmetros *n* e *p*. Para *n* = 5 e *p* = 0,1 calculemos as probabilidades que haja:

a) Uma peça defeituosa:

$$P(X = 1)\binom{5}{1}(0{,}1)(0{,}9)^4 = 0{,}328$$

b) Duas peças defeituosas:

$$P(X = 2)\binom{5}{2}(0{,}1)^2(0{,}9)^3 = 0{,}365$$

c) Pelo menos uma peça defeituosa:

$$P(X \geq 1) = 1 - P(X = 0) = 1 - (0{,}9)^5 = 0{,}4095$$

Exemplo: Uma moeda é lançada 20 vezes. Qual a probabilidade de saírem 8 caras?

Solução:

X: Número de sucessos (caras)

$$X = 0{,}1{,}2{,}\ldots{,}20 \rightarrow p = P(c) = \frac{1}{2} \rightarrow X \sim binomial\left(20, \frac{1}{2}\right)$$

$$P(X = 8) = \binom{20}{8}\left(\frac{1}{2}\right)^8\left(\frac{1}{2}\right)^{12} = 0{,}12013$$

Exemplo: Numa criação de coelhos, 40% são machos. Qual a probabilidade de que nasçam pelo menos 2 coelhos machos num dia em que nasceram 20 coelhos?

Solução:

X: Número de coelhos machos

$$X = 0{,}1{,}2{,}\ldots{,}20 \rightarrow P(X) = p = 0{,}4 \rightarrow X \sim binomial(20; 0{,}4)$$

$$P(X \geq 2) = 1 - P(X < 2) = 1 - \{P(X = 0) + P(X = 1)\} =$$

$$= 1 - \left\{\binom{20}{0}(0{,}4)^2(0{,}6)^{20} + \binom{20}{1}(0{,}4)^1(0{,}6)^{19}\right\} =$$

$$= 1 - \{0{,}00003 + 0{,}00049\} = 0{,}99948$$

Exemplo: Uma urna tem 20 bolas pretas e 30 brancas. Retiram-se 25 bolas com reposição. Qual a probabilidade de que:

a) 2 sejam pretas?

b) Pelo menos 3 sejam pretas?

Solução:

$X: Número\ de\ bolas\ pretas \rightarrow P = 20/50 = 0{,}4$

Logo: $X \sim binomial(25; 0{,}4)$

$$P(X = 2) = \binom{25}{2}(0{,}4)^2(0{,}6)^{23} = 0{,}00038$$

$$P(X \geq 3) = 1 - P(X < 3) = 1 - \{P(X = 0) + P(X = 1) + P(X = 2)\}$$
$$= 1 - \left\{\binom{25}{0}(0{,}4)^0(0{,}6)^{25} + \binom{25}{1}(0{,}4)^1(0{,}6)^{24} + \binom{25}{2}(0{,}4)^2(0{,}6)^{23}\right\} =$$
$$= 1 - \{0 + 0{,}00005 + 0{,}00038\} = 1 - 0{,}00043 = 0{,}99957$$

Exemplo: A taxa de imunização de uma vacina é 80%. Se um grupo de 20 pessoas foi vacinado, desejamos saber o comportamento probabilístico do número de pessoas imunizadas desse grupo.

Seja X a variável de interesse. Para cada pessoa do grupo, a probabilidade de estar imunizada é 0,8 e admitimos, ainda, independência entre os resultados das várias pessoas vacinadas.

Dessa forma, teremos $X \sim Binomial(20; 0{,}8)$, em que sucesso corresponde à imunização. Por exemplo, a probabilidade de 15 estarem imunizados é dada por:

$$P(X = 15) = \binom{20}{15} 0{,}8^{15} 0{,}2^{20-15} = 0{,}175$$

Outras probabilidades podem ser calculadas de modo análogo.

4.4.1.2 Distribuição de Poisson

Dizemos que a variável aleatória X tem distribuição de Poisson com parâmetro θ se:

$$P(X = k) = e^{-\theta}\frac{\theta^k}{k!}, para\ \theta > 0\ e\ k = 0, 1, 2, \ldots$$

Propriedades:

1) $E(X) = Var(X) = \theta$

2) Se $X_1 \sim Poisson(\theta_1)$ e $X_2 \sim Poisson(\theta_2)$ e são variáveis aleatórias independentes, então $X_1 + X_2 \sim Poisson(\theta_1 + \theta_2)$. Podemos ver a distribuição de *Poisson* como aproximação da binomial quando *n* é grande e *p* é pequeno (evento raro), ou quando contamos ocorrências de um certo evento em um intervalo de tempo: nº de acidentes em frente à entrada da UFAM por semana, nº de ligações recebidas por uma central telefônica das 12:00 às 13:00. Uma região do espaço (área, volume): nº de glóbulos vermelhos em uma amostra de sangue, nº de plantas de uma certa espécie em uma região, de modo que o número médio de ocorrências seja pequeno (usualmente < 15).

Exemplo: 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 central ser ultrapassada.

Solução:

Seja X o nº de ligações por minuto.

$\theta = 300/60 = 5$ chamadas por minuto.

$$P(X > 10) = 1 - P(X \leq 10) = 1 - \left(\frac{e^{-5}5^{10}}{10!}\right) = 1 - 0{,}986 = 0{,}014 = 1{,}4\%$$

Exemplo: 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 que ficar ao largo, aguardando a vaga?

Solução:

Terão de esperar ao largo os navios-tanque que excederem o máximo de 15:

Seja X o nº de navios-tanque que chegam ao porto por dia.

$$P(X > 15) = 1 - P(X \leq 15) = 1 - \sum_{x=0}^{15} P(X \leq x) = 1 - 0{,}9512 = 0{,}0488$$

Exercícios propostos

1) Um estudante tem probabilidade p = 0,8 de acertar cada problema que tenta resolver. Numa prova de 8 problemas, qual a probabilidade de que ele acerte exatamente 6.
2) Um exame consta de 20 questões tipo *certo* e *errado*. Se o aluno "chutar" todas as respostas, qual a probabilidade de ele acertar exatamente 10 *questões*?
3) Um time de futebol tem probabilidade p = 3/5 de vencer todas as vezes que joga. Se disputar 5 partidas, qual a probabilidade de que vença ao menos uma?
4) Um teste tipo certo e errado consta de 6 questões. Se o aluno "chutar" as respostas ao acaso, qual a probabilidade de que ele acerte mais do que 2 testes?
5) Um casal planeja ter 5 filhos. Admitindo que sejam igualmente prováveis os resultados: filho do sexo masculino e filho do sexo feminino, qual a probabilidade de o casal ter:

   (a) 5 filhos do sexo masculino?

   (b) Exatamente 3 filhos do sexo masculino?

   (c) No máximo um filho do sexo masculino?

   (d) O 5º filho do sexo masculino, dado que os outros 4 são do sexo feminino?
6) Uma central telefônica recebe em média 300 chamadas por hora e pode processar no máximo 10 ligações por minuto. Utilizando a distribuição de Poisson, estime a probabilidade de a capacidade da central ser ultrapassada?
7) Se $X \sim Binomial(n, p)$ prove que $E(X) = np$ e $Var(X) = np(1 - p)$.
8) O número de acidentes pequenos durante uma partida de futebol é uma v. a. de Poisson com média $\theta = 4{,}5$. Qual a probabilidade de ocorrerem, no máximo, dois acidentes durante uma partida?
9) Em uma excursão ao Pantanal de Mato Grosso, certa ave é avistada em número que é uma v. a. de Poisson com média $\theta = 0{,}8$. Determinar a probabilidade de que, em uma excursão: (a) não se aviste nenhuma dessas aves; (b) seja avistada apenas uma; (c) sejam avistada duas; (d) sejam avistadas mais de três.

10) O número de reclamações que uma lavanderia recebe por dia é uma v. a. de Poisson com média $\theta = 3{,}5$. Qual a probabilidade de a lavanderia receber apenas uma reclamação em determinado dia?

4.4.2 Distribuições Contínuas de Probabilidade

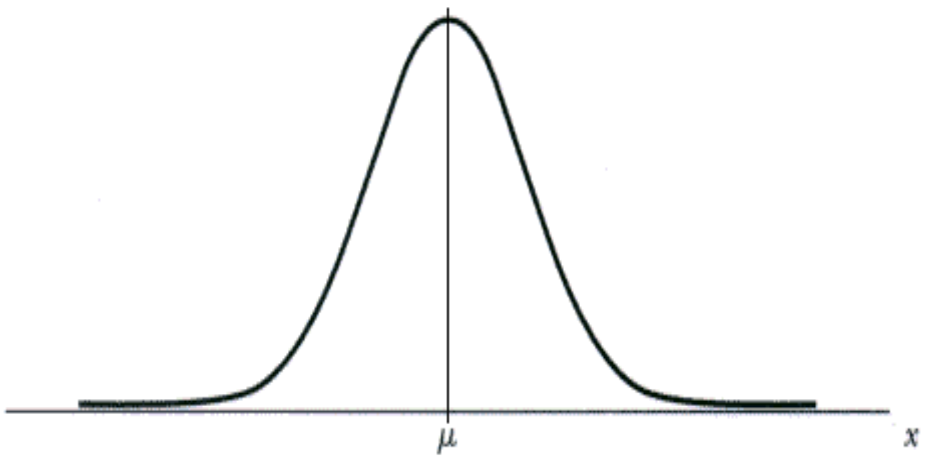
De modo geral, podemos dizer que as variáveis aleatórias cujos valores resultam de algum processo de mensuração são variáveis aleatórias contínuas.

4.4.2.1 Distribuição Normal

A distribuição normal é uma distribuição em forma de sino que é usado muito extensivamente em aplicações estatísticas em campos bem variados. Sua densidade de probabilidade (f.d.p.) é dada por:

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

Sua média é μ e sua variância é σ^2. Quando X tem uma distribuição normal com média μ e variância σ^2, escrevemos, de forma compacta, $X \sim N(\mu, \sigma^2)$



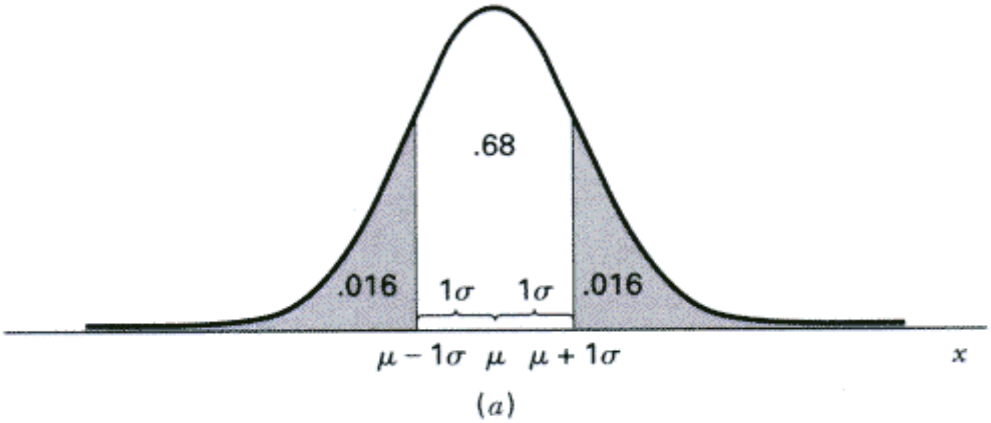
Uma importante propriedade da distribuição normal é que qualquer função linear de variável aleatória normalmente distribuída é também normalmente distribuída.

Características:

1. Simétrica em relação à média
2. A média, moda e mediana são iguais

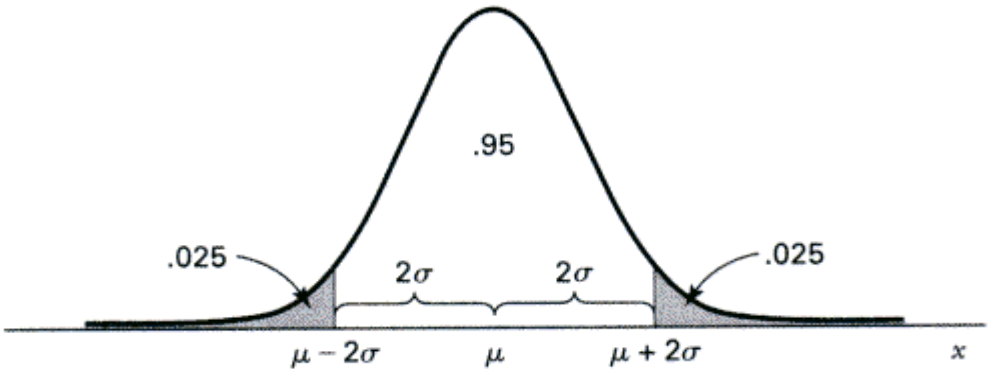
3. A área total sob a curva é igual a 1, 50% à esquerda e 50% à direita da média

a) A área entre $\mu - 1\sigma$ e $\mu + 1\sigma$ é aproximadamente 68%



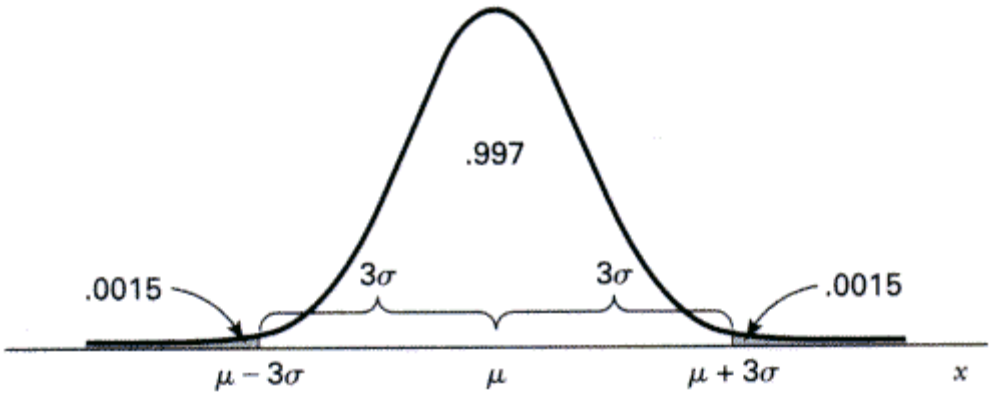
(a)

b) A área entre $\mu - 2\sigma$ e $\mu + 2\sigma$ é aproximadamente 95%



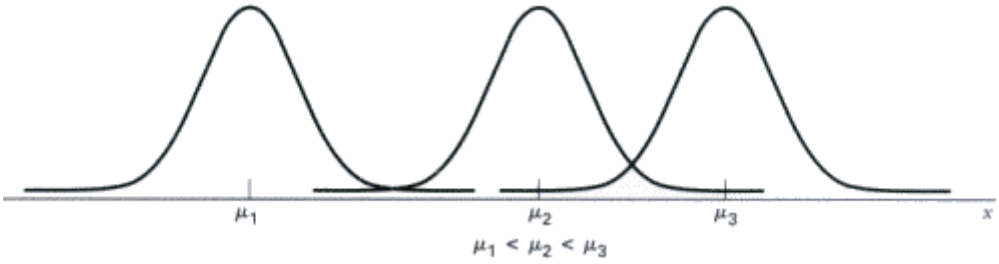
(b)

c) A área entre $\mu - 3\sigma$ e $\mu + 3\sigma$ é aproximadamente 99,7%

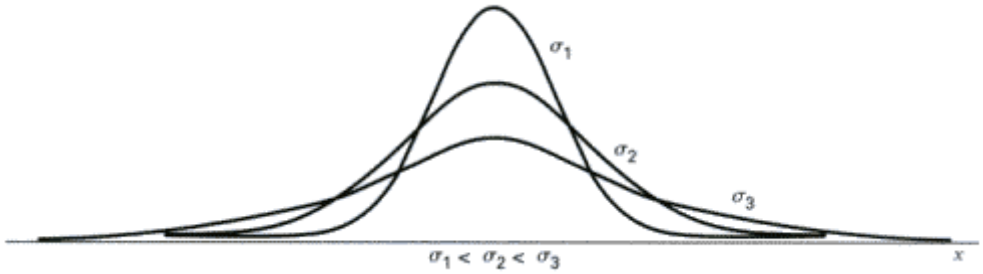


(c)

A distribuição normal é completamente determinada pelos parâmetros μ e σ



μ parâmetro de locação.



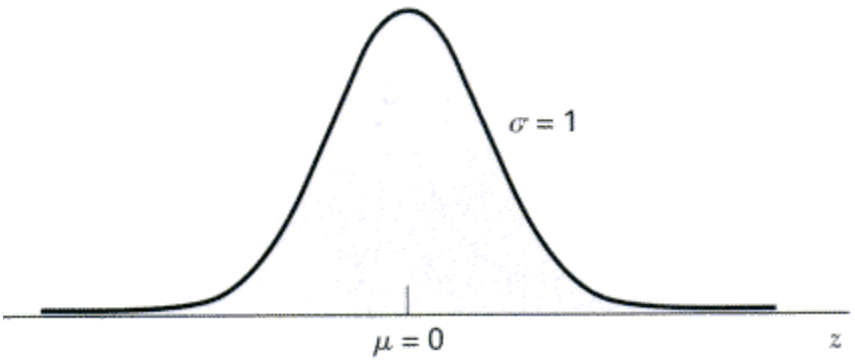
σ parâmetro de forma.

4.4.2.2 Distribuição Normal Padrão:

Caracterizada pela média igual a zero e desvio padrão igual a 1.

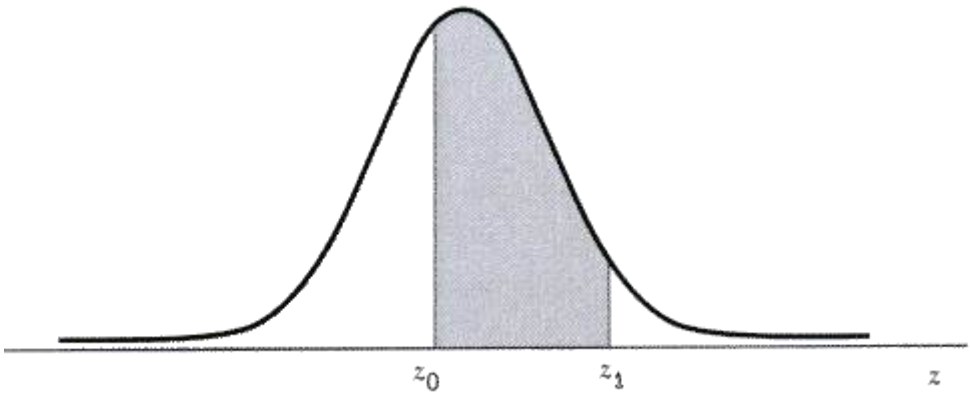
Se X tem distribuição normal com média μ e variância $\boldsymbol{\sigma^2}$, então $\boldsymbol{Z = \frac{(x-\mu)}{\sigma}}$

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}, -\infty < z < \infty$$

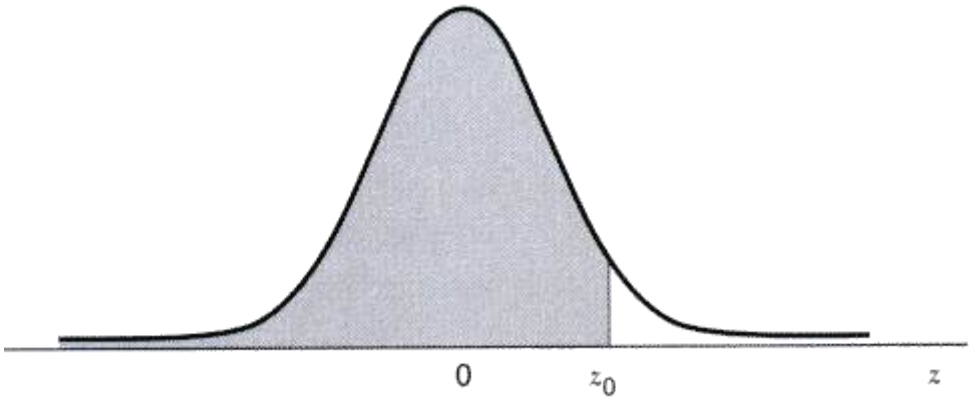


A área entre Z_0 e Z_1 é calculada por

$$\int_{z_0}^{z_1} \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} dz$$



As áreas estão tabeladas: veja a tabela z



Exemplo: Suponha que as notas (em pontos) do exame de seleção para uma universidade segue uma distribuição normal com média 500 e desvio padrão 100. Determine a probabilidade de um estudante ter nota:

a) Acima de 650?

b) Menos do que 250?

c) Entre 325 e 675?

X= nº de pontos (nota) do candidato

μ =500 $X \sim N(500; 100^2$

σ = 100

a) P(X > 650)

= P(X > 650) = $P\left(\frac{X-500}{100} > \frac{650-500}{100}\right)$

= P(Z > 1,5)

= 0,5 – 0,4332

= 0,0668

b) P(X < 250)

= P(X < 250) = $P\left(\frac{X-500}{100} > \frac{250-500}{100}\right)$

= P(Z > -2,5)

= 0,5 – 0,4938

= 0,0062

c) P(325 < X < 675)

= P(325 < X < 675) = $P\left(\frac{325-500}{100} < \frac{X-500}{100} < \frac{675-500}{100}\right)$

= P(-1,75 < Z < 1,75)

= 0,4599 + 0,4599

= 0,9198

Exemplo: Os depósitos efetuados no Banco da Ribeira durante o mês de janeiro são distribuídos normalmente, com média de $10.000,00 e desvio padrão de $1.500,00. Um depósito é selecionado ao acaso dentre todos os referentes ao mês em questão. Encontre a probabilidade de que o depósito seja:

a)$10.000,00 ou menos;

b)Pelo menos $10.000,00;

c)Um valor entre $12.000,00 e $15.000,00;

d)Maior do que $20.000,00.

Temos que μ = 10.000 e σ = 1.500. Seja a v.a. *X* = depósito.

$$a) P(X \le 10.000) = P\left(Z \le \frac{10.000-10.000}{1.500}\right) = P(Z \le 0) = 0,5$$

$$b) P(X \ge 10.000) = P(Z \ge 0) = 0,5$$

$$c) P(12.000 < X < 15.000) = P\left(\frac{12.000-10.000}{1.500} < Z < \frac{15.000-10.000}{1.500}\right)$$

$$= P(4/3 < Z < 10/3)$$

$$= P(1,33 < Z < 3,33) = 0,09133$$

$$d) P(X > 20.000) = P\left(Z > \frac{20.000-10.000}{1.500}\right) = P(Z > 6,67) \cong 0.$$

Exemplo: Seja $X \sim N(100,25)$. Calcular:

a) P(100 ≤ X ≤ 106)

b) P(89 ≤ X ≤ 107)

c) P(112 ≤ X ≤ 116)

d) P(X ≥ 108)

Solução:

$\therefore \mu = 100$ e $\sigma = 5 \rightarrow Z = \frac{X - 100}{5}$

a) $P(100 \le X \le 106) = P\left(\frac{100-100}{5} \le Z \le \frac{106-100}{5}\right) = P(0 \le Z \le 1,2) = 0,384930$

b) $P(89 \le X \le 107) = P\left(\frac{89-100}{5} \le Z \le \frac{107-100}{5}\right) = P(-2,2 \le Z \le 1,4) =$

$= P(-2,2 \le Z \le 0) + P(0 \le Z \le 1,4) =$

$= 0,486097 + 0,419243 = 0,90534$

c) $P(112 \le X \le 116) = P\left(\frac{112-100}{5} \le Z \le \frac{116-100}{5}\right) = P(2,4 \le Z \le 3,2) =$

$= P(0 \le Z \le 3,2) - P(0 \le Z \le 2,4) =$

$= 0,499313 - 0,491803 = 0,007510$

d) $P(X \ge 108) = P\left(Z \ge \frac{108-100}{5}\right) = P(Z \ge 1,6) = 0,5 - P(0 \le Z \le 1,6) =$

$= 0,5 - 0,445201 = 0,054799$

Exemplo: Sendo $X \sim N(50,16)$, determinar X_α tal que:

a) $P(X \ge x_\alpha) = 0,05$

b) $P(X \ge x_\alpha) = 0,99$

Solução:

$\mu = 50$, $\sigma = 4$

a) Procurando no corpo da tabela 0,45 (0,5 – 0,05), encontramos: $Z_\alpha = 1,64$

$\therefore$ como $Z_\alpha = \frac{X_\alpha - \mu}{\sigma} \rightarrow 1,64 = \frac{X_\alpha - 50}{4}$ $\therefore X_\alpha = 56,56$ $\therefore P(X \ge 56,56) = 0,05$

b) Procurando no corpo da tabela 0,49 (0,5 – 0,01) encontramos: $Z_\alpha = 2,32$

$2,32 = \frac{X_\alpha - 50}{4}$ $\therefore X_\alpha = 59,28$ $\therefore P(X \le 59,28) = 0,99$

Exemplo: Uma fábrica de carros sabe que os motores de sua fabricação têm duração normal com média igual a 150.000 km e desvio-padrão de 5.000 km. Qual a probabilidade de que um carro, escolhido ao acaso, dos fabricados por essa firma, tenha um motor que dure:

a) Menos de 170.000 km?

b) Entre 140.000 km e 165.000 km?

c) Se a fábrica substitui o motor que apresenta duração inferior à garantia, qual deve ser esta garantia para que a porcentagem de motores substituídos seja inferior a 0,2%?

Solução:

X: duração do motor em km onde μ = 150.000 km , σ = 5.000 km

a) $P(X < 170.000) = P\left(Z \leq \frac{170.000 - 150.000}{5000}\right) = P(Z \leq 4) = 0,5 + P(0 \leq Z \leq 4) =$

$= 0,5 + 0,499968 = 0,999968$

b) $P(140.000 < X < 165.000) = P\left(\frac{140.000 - 150.000}{5000} \leq Z \leq \frac{165.000 - 150.000}{5000}\right)$

$= P(-2 \leq Z \leq 3) = P(-2 \leq Z \leq 0) + P(0 \leq Z \leq 3) =$

$= 0,477250 + 0,498650 = 0,97590$

c) Procurando no corpo da tabela 0,498 (0,5 – 0,002), encontramos: Z_α= -2,87

$\therefore -2,32 = \frac{X_\alpha - 150.000}{5000} \ \therefore X_\alpha = 135.650$

A garantia deve ser de 135.650 km.

Exercícios propostos

1) Determinar a média e o desvio padrão de um exame em que as notas 75 e 88 correspondem aos valores padronizados -0,4 e 1,3 respectivamente.
2) Dada uma v. a. $N(18; 2,5)$, determinar: (a) $P(X \leq 15)$; (b) k tal que $P(X < k) = 0,2578$.
3) O total de pontos obtidos no vestibular de uma faculdade é uma v. a. N(550; 110). Determinar a probabilidade de um estudante obter: (a) mais de 600 pontos; (b) menos de 350 pontos; (c) entre 300 e 600 pontos.

4) Os QI de 500 candidatos a certa faculdade têm distribuição aproximadamente $N(110; 11)$. Se a faculdade exige QI mínimo de 95 para admissão, qual a percentagem provável de reprovação.
5) Um teste de aptidão para exercício de certa profissão exige uma seqüência de operações a serem executadas rapidamente uma após a outra. Para passar no teste, o candidato deve completá-lo em 80 minutos no máximo. Admita que o tempo para completar o teste seja uma v. a. $N(90; 20)$(minutos). (a) Que percentagem dos candidatos tem chance de ser aprovado? (b) Os 5% melhores receberão um certificado especial. Qual o tempo máximo para fazer jus a tal certificado?
6) Para $Z = N(0, 1)$, determinar $P(|Z| \geq 1)$.
7) Determinar z_0 tal que a área entre a média (0) e z_0 seja 0,40.
8) Para uma v. a. $N(2; 3)$, determinar: (a) um valor tal que a probabilidade do intervalo da média a esse valor seja 0,4115; (b) outro valor tal que a probabilidade do intervalo $x = 3,5$ a esse valor seja 0,2307.
9) Se um conjunto de valores tem distribuição normal, qual a percentagem dos valores que distam da média: (a) por mais de $1,3\sigma$; (b) por menos de $0,52\sigma$?
10) Em um exame de matemática, a nota média foi 70, com desvio padrão de 4,5. Todos os alunos que obtiveram nota de 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?

Distribuições Relacionadas

Em adição a distribuição normal, existem outras distribuições de probabilidade que serão usadas freqüentemente. São as distribuições t, χ^2 e F tabuladas em vários textos. Estas distribuições são derivadas da distribuição normal e são definidas como descritas abaixo.

4.4.2.3 Distribuição χ^2

Uma v. a. contínua Y, com valores positivos, tem uma distribuição qui-quadrado com n graus de liberdade (denotada por $\chi^2_{(n)}$), se sua densidade for dada por

$$f(y,n) = \begin{cases} \dfrac{1}{\Gamma(n/2)2^{n/2}} y^{\frac{n}{2}-1} e^{-\frac{y}{2}}, y > 0 \\ 0, \qquad\qquad y < 0 \end{cases}$$

$$E(Y) = n, Var(Y) = 2n$$

Se $X_1, \dots, X_n$ são variáveis independentes normais com médias zero e variância 1, isto é, X_i independente $\sim N(0,1), i = 1, \dots, n$ então $Z = \sum_{i=1}^{n} X_i^2$ tem distribuição χ^2 com n graus de liberdade, e escreveremos $Z \sim \chi^2$. A distribuição χ_n^2 é a distribuição da soma de n variáveis normais independentes padronizadas.

Propriedade: Se $Z_1 \sim \chi_n^2$ *e* $Z_2 \sim \chi_m^2$ e Z_1 *e* Z_2 independentes, então $Z_1 + Z_2 \sim \chi_{n+m}^2$

4.4.2.4 Distribuição t

Se $X \sim N(0,1)$ *e* $Y \sim \chi_n^2$ e X e Y são independentes, então

$$Z = \frac{X}{\sqrt{\frac{Y}{n}}}$$

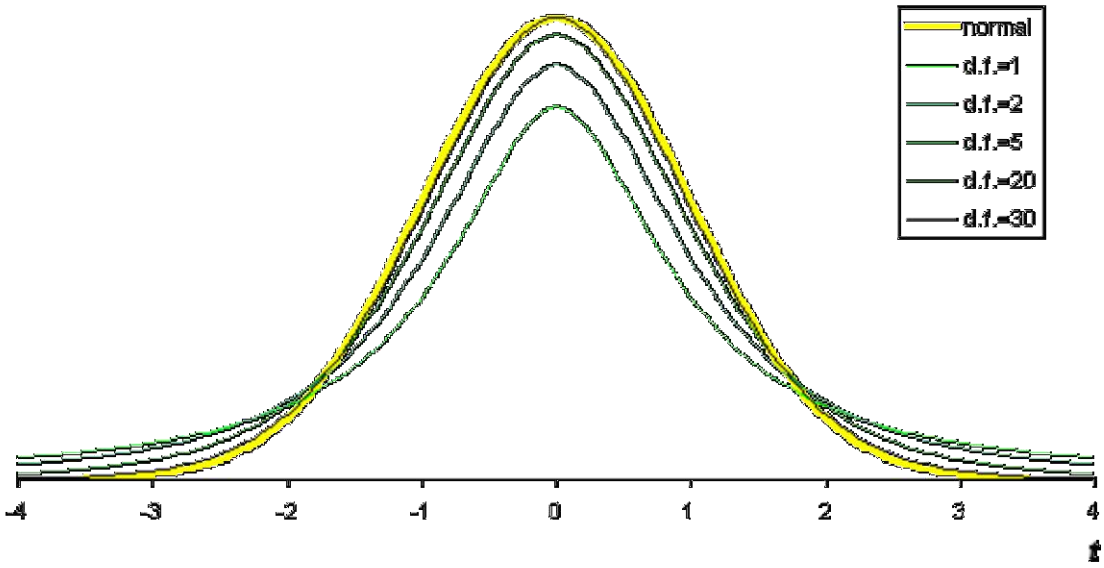
tem densidade dada por:

$$f(z,n) = \frac{\Gamma((n+1)/2)}{\Gamma(n/2)\sqrt{\pi n}} (1 + z^2/n)^{-(n+1)/2}, -\infty < z < \infty.$$

tal v. a. tem distribuição *t* com n graus de liberdade.

$$E(Z) = 0, Var(Z) = \frac{n}{n-2}$$

Escrevemos $Z \sim t_n$. A distribuição t é a distribuição de uma variável normal padrão dividida pela raiz quadrada de uma variável χ^2 independente dividida pelos graus de liberdade. A distribuição t é uma distribuição simétrica como a normal, porém um pouco mais achatada e com caudas mais longas que a normal. Quando os graus de liberdade tendem a infinito, a distribuição *t* aproxima-se da normal.



4.4.2.5 Distribuição F

Se $Y_1 \sim \chi^2_{n_1}$ e $Y_2 \sim \chi^2_{n_2}$, onde Y_1 e Y_2 são independentes, então $W = \frac{Y_1/n_1}{Y_2/n_2}$ tem densidade dada por:

$$g(w; n_1, n_2) = \frac{\Gamma((n_1 + n_2)/2)}{\Gamma(n_1/2)\Gamma(n_2/2)} \left(\frac{n_1}{n_2}\right)^{n_1/2} \cdot \frac{w^{(n_1-2)/2}}{(1 + n_1 f/n_2)^{(n_1+n_2)/2}}, w > 0.$$

tal v. a. tem distribuição F com graus de liberdade n_1 e n_2. Escrevemos $W \sim F_{n_1,n_2}$

$$E(W) = \frac{n_2}{n_2 - 2}, Var(W) = \frac{2n_2^2(n_1 + n_2 - 2)}{n_1(n_2 - 2)^2(n_2 - 4)}$$

O 1º índice n_1 refere-se aos graus de liberdade do numerador e o 2º, n_2, refere-se aos graus de liberdade do denominador. A distribuição F é, então, a razão entre duas variáveis χ^2 independentes divididas pelos seus respectivos graus de liberdade.

5. Inferência Estatística

Consiste em um conjunto de procedimentos por meio dos quais as informações obtidas com base em dados amostrais são utilizadas para o estabelecimento de conclusões e a tomada de decisões sobre a população da qual a amostra foi extraída.

Os problemas básicos da inferência estatística são: o chamado *teste de hipótese* e a *estimação.* O problema de estimação apresenta-se em todas as situações, seja no

cotidiano ou em qualquer ciência. A estimativa pode ser de uma média de uma medida de variabilidade ou de uma proporção.

Distribuição Amostral $\Rightarrow$ É a distribuição que descreve o padrão de variação dos valores de uma estatística, para diferentes amostras extraídas da população de interesse, é denominada distribuição amostral.

Amostra Aleatória $\Rightarrow$ As observações X_1, X_2, ..., X_n constituem uma amostra aleatória de tamanho *n* da população, se cada observação resulta de seleções independentes dos elementos da população e se cada X_i tem a mesma distribuição da população da qual foi extraída.

A distribuição da média amostral $\overline{X}$, de uma amostra aleatória de tamanho *n* extraída de uma população que tem média μ e desvio padrão σ, tem as seguintes características:

Média = $E(\overline{X}) = \mu_{\bar{x}} = \mu$

Variância = $Var(\overline{X}) = \sigma^2_{\bar{x}} = \dfrac{\sigma^2}{n}$

Desvio Padrão = $Dp(\overline{X}) = \sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$

Teorema Central do Limite

A distribuição da média amostral $\overline{X}$, de uma amostra aleatória de tamanho *n* extraída de uma população NÃO NORMAL, com média μ e desvio padrão σ, é APROXIMADAMENTE NORMAL com média μ e desvio padrão $\sigma/\sqrt{n}$.

Este resultado significa que:

$Z = \dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}}$ é aproximadamente *N*(0, 1).

Estimação $\Rightarrow$ Estudo de métodos para obter medidas representativas da população calculadas a partir da amostra.

Parâmetro $\Rightarrow$ É uma medida numérica que descreve alguma característica de uma população.

Estatística $\Rightarrow$ É uma função das observações amostrais, que não depende de parâmetros desconhecidos. Seja (X_1, X_2, ..., X_n) uma amostra aleatória e (x_1, x_2, ..., x_n) os valores tomados pela amostra; então $y = H(x_1 , x_2, ..., x_n)$ é uma estatística.

Principais estatísticas:

- Média Amostral

- Proporção Amostral

- Variância Amostral

Estimador $\Rightarrow$ É a função da amostra que corresponde a um parâmetro populacional.

Estimativa $\Rightarrow$ É o valor do estimador, calculado a partir de uma amostra.

Média, *moda* e *mediana* são estimadores do valor central.

População:	Média =		μ
	Variância =		σ^2
	Proporção =		π
Amostra:	Média =	$\overline{X}$	*estimador de* μ
	Variância =	S^2	*estimador de* σ^2
	Proporção =	p	*estimador de* π

Tipos de Estimações de Parâmetros

i) Estimação Pontual

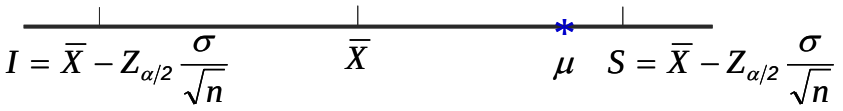
ii) Estimação Intervalar

Estimação Pontual $\Rightarrow$ É usada quando, a partir da amostra, procura-se obter um único valor de certo parâmetro populacional, ou seja, obter estimativas a partir dos valores amostrais.

Estimatição Intervalar $\Rightarrow$ É o intervalo definido pela estimativa pontual mais/menos o erro máximo da estimativa.

Erro Máximo da Estimativa $\Rightarrow$ Representa a diferença (erro) máxima que será permitida entre a estimativa pontual ($\overline{X}$) e o valor verdadeiro do parâmetro que está sendo estudado (μ).

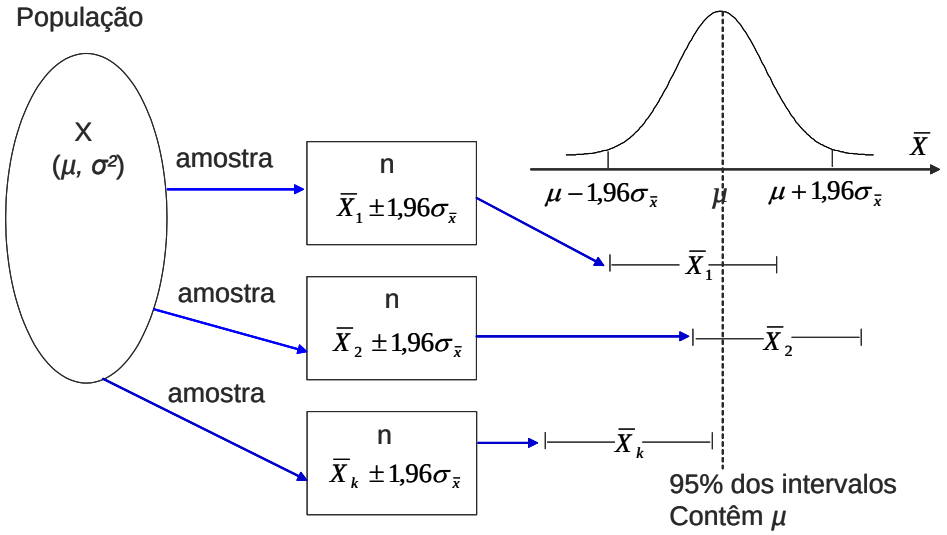
$$Erro = \left| \overline{X} - \mu \right|$$



Erro Cometido na Estimação de μ por $\overline{X}$, segundo Montgomery, D.C. & Runger, G.C. (1994).

5.1 Intervalo de Confiança

Dado a limitação da estimação pontual, que reside no desconhecimento da magnitude do erro que se está cometendo. Surge à idéia da construção de um intervalo que contenha, com um nível de confiança conhecido, o valor verdadeiro do parâmetro, baseado na distribuição amostral do estimador pontual.



Intervalo de Confiança para Média (μ) Para Variância populacional (σ^2) conhecida

$$\overline{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Podemos afirmar com 100(1 – α)% de confiança que o intervalo de $\overline{X} \pm e$ para contém a média populacional que estamos procurando estimar.

O grau de confiança mais utilizado é 95% e o valor correspondente $Z_{\alpha/2}$ é 1,96.

O intervalo de confiança é definido pelo grau de confiança e pela variabilidade.

Passos para construção do intervalo de confiança para Média (μ) onde a Variância populacional (σ^2) é conhecida

- Colete uma amostra de tamanho *n* da população de interesse;
- Calcule o valor de $\overline{X}$;
- Escolha o valor do coeficiente de confiança $1 - \alpha$;
- Obter o valor de $Z_{\alpha/2}$ da tabela da distribuição normal;
- Calcule os limites do intervalo de confiança.

Exemplo: Para uma amostra de 50 observações de uma população normal com média desconhecida e desvio padrão σ = 6, seja 20,5 a média amostral $\overline{X}$. Construir um intervalo de 95% de confiança para a média populacional.

Solução:

$\overline{X}$ = 20,5; n = 50; σ = 6

$$\left[20{,}5 - 1{,}96\frac{6}{\sqrt{50}};\ \ 20{,}5 + 1{,}96\frac{6}{\sqrt{50}}\right] = [18{,}84;\ 22{,}16]$$

O resultado obtido [18,84; 22,16] é um intervalo de confiança de 95% de confiança para a média populacional μ, calculado com base na amostra observada.

Não se deve escrever $P(18{,}84 < \mu < 22{,}16) = 0{,}95$ porque a expressão entre parênteses não contém nenhuma variável aleatória, já que μ é valor fixo, e, embora desconhecido, está, ou não, dentro do intervalo.

Distribuição Amostral da Estatística *t*

Se $X_1, X_2, ..., X_n$ é uma amostra aleatória de uma população normal com média μ e desvio padrão σ, então a distribuição de

$$t = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

é denominada distribuição *t* de Student com $n - 1$ graus de liberdade.

Intervalo de Confiança para Média (μ) para Variância populacional (σ^2) desconhecida

$$\bar{X} - t_{n-1;\alpha/2}\frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{n-1;\alpha/2}\frac{S}{\sqrt{n}}$$

- Neste caso precisa-se calcular a estimativa *S* (desvio padrão amostral) a partir dos dados;
- O coeficiente *t* segue a distribuição "*t*" de Student, no caso com (*n* – 1) graus de liberdade.

Passos para construção do intervalo de confiança para Média (μ) onde a Variância populacional (σ^2) é desconhecida

- Colete uma amostra de tamanho $n \leq 30$ da população de interesse;
- Calcule os valores de $\bar{X}$ e *S*;
- Escolha o valor do coeficiente de confiança $1 - \alpha$;
- Determine os valores de $t_{\alpha/2}$; $n - 1$ apartir da tabela da distribuição *t* de Student;
- Calcule os limites do intervalo de confiança.

Exemplo: Uma amostra de tamanho 9, extraída de uma população normal acusa $\bar{X}$ = 1,0 e *S* = 0,264. Construir intervalos de 98% e 95% de confiança para média populacional.

Solução:

Para $1 - \alpha = 98\% \Rightarrow \alpha = 0{,}02$; $\alpha/2 = 0{,}01$ graus de liberdade = 9 – 1 = 8.

Intervalo: $\left[\bar{X} \pm t_{0,01;8}\left(S/\sqrt{n}\right)\right] \Rightarrow [1 \pm 2{,}896(0{,}264/3)] \Rightarrow [0{,}745;\ 1{,}255]$

Para $1 - \alpha = 95\% \Rightarrow \alpha = 0{,}05$; $\alpha/2 = 0{,}025$ graus de liberdade = 9 – 1 = 8.

Intervalo: $\left[\bar{X} \pm t_{0,025;8}\left(S/\sqrt{n}\right)\right] \Rightarrow [1 \pm 2{,}306(0{,}264/3)] \Rightarrow [0{,}797;\ 1{,}203]$.

Note que, aumentando o nível de confiança, o tamanho do intervalo também aumenta.

Intervalo de Confiança para Média de uma População Não-Normal – Grandes Amostras

Enquanto, nos casos anteriores, se conhecia a distribuição da estatística com base na qual se obteve o intervalo, aqui, não se passa o mesmo. Usaremos o Teorema

Central do Limite para afirmar que, se *n* é suficientemente grande, $\frac{\overline{X}-\mu}{S/\sqrt{n}}$ tem distribuição aproximadamente normal *N*(0, 1). Portanto,

$$\overline{X} - Z_{\alpha/2}\frac{S}{\sqrt{n}} \leq \mu \leq \overline{X} + Z_{\alpha/2}\frac{S}{\sqrt{n}}$$

é um intervalo de confiança para média com nível de aproximado de 100(1 – α)%

Os passos para construção deste intervalo é análogo ao intervalo para média populacional com σ^2 conhecida, com exceção do fato do cálculo para determinar S^2.

Exemplo: A resistência à tração de 20 corpos de prova é (valores já ordenados):

131 132 134 135 135 138 138 139 139 140

142 143 144 144 145 146 147 148 149 150

Estabelecer uma estimativa intervalar de 95% para média populacional.

Solução:

n = 20, $\overline{X}$ = 140,95 e *S* = 5,73

$\left[140{,}95 \pm 1{,}96(5{,}73/\sqrt{20})\right]$ = [138,44; 143,46].

Assim um intervalo de confiança a 95% para μ é dado por (138,44; 143,46)

Tamanho da Amostra para Estimar a Média

A partir da fórmula do erro máximo de estimação, mediante aplicação de um cálculo algébrico podemos reformular a fórmula isolando a variável "n".

Para que seja possível ter 100(1 – α)% de confiança que o erro de estimação $\left|\overline{X} - \mu\right|$ é inferior a um valor predeterminado ***e***, o tamanho da amostra necessário é:

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

Etapas para determinação do tamanho da amostra necessária para estimação da média populacional μ

- Especificar o coeficiente de confiança 1 – α;
- Obter uma estimativa preliminar do desvio padrão σ;

- Especificar o erro máximo ***e*** permitido na estimação;
- Obter o valor de $Z_{\alpha/2}$ da tabela da distribuição normal;

Exemplo: Suponha que desejamos estimar uma média populacional com erro amostral 0,5 e probabilidade de confiança 0,95. Como σ^2 (variância populacional) não é conhecida, foi retirada uma amostra de 10 observações da população para nos dar uma idéia sobre valor de σ^2, obteve-se $S^2 = 16$. Determine o tamanho da amostra necessário para atender estas especificações.

$$n = \frac{(1{,}96)^2 16}{(0{,}5)^2} = 245{,}86 \Rightarrow n \cong 246$$

Estimativa de Proporções

Suponha que há interesse na proporção de elementos da população que possuem alguma característica de interesse (*p*).

Se o tamanho da amostra (*n*) for suficientemente grande, é possível fazer mensurações para: Intervalo de Confiança e Teste de Hipótese.

Erro Máximo da Estimativa para *p*

$$e = \mathbf{Z}_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

Intervalo de Confiança para *p*

$$p - \mathbf{Z}_{\alpha/2}\sqrt{\frac{p(1-p)}{p}} \leq P \leq p + \mathbf{Z}_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

Etapas para a Construção de um Intervalo de 100(1 – α)% de Confiança para *p*

- Colete uma amostra aleatória de tamanho *n* da população de interesse;
- Determine o valor de *y*;

Onde *y* = nº de elementos na amostra que possuem a característica de interesse;

- Calcule $\hat{p} = \overline{p} = \frac{y}{n}$;
- Escolha o valor do coeficiente de confiança $1 - \alpha$;

- Determine os valores de $Z_{\alpha/2}$ apartir da tabela da distribuição normal;
- Calcule os limites do intervalo de confiança.

Exemplo: Entrevistam em uma cidade 1.500 pessoas em idade de trabalho, e constata-se que 145 estão desempregadas.

1) Estimar a taxa de desempregado com base nos dados,
2) Estabelecer um intervalo de 95% de confiança para a taxa populacional.

Solução:

1) $\hat{p} = \frac{145}{1500} = 0{,}097 = 9{,}7\%$

2) $\alpha = 0{,}05$; $\alpha/2 = 0{,}025$; $Z_{0{,}025} = 1{,}96$.

$\left[0{,}097 \pm 1{,}96\sqrt{0{,}097(1-0{,}097)/1500}\right]$ = [0,097 ± 0,0149] = [0,082; 0,112] = [8,2%; 11,2%].

Assim um intervalo de confiança a 95% para *p* (taxa populacional) é dado por (0,082; 0,112*)*.

Tamanho da Amostra para Estimar *p*

Para que seja possível ter 100(1 – α)% de confiança que o erro de estimação $|\hat{p} - p|$ é inferior a um valor predeterminado ***e***, o tamanho da amostra necessário é:

$$n = \frac{Z_{\alpha/2}^2 p(1-p)}{e^2}$$

Etapas para determinação do tamanho da amostra necessária para estimação da proporção *p*

- Especificar o coeficiente de confiança 1 – *α*;
- Obter uma estimativa preliminar da proporção *p*;
- Especificar o erro máximo ***e*** permitido na estimação;
- Obter o valor de $Z_{\alpha/2}$ da tabela normal

Estimativa preliminar da proporção *p*

- Dados históricos sobre a população de interesse;

- Resultados obtidos em estudos similares ao que está sendo realizado;
- Extração de uma amostra-piloto;
- Utilizar p = 0,5 (atitude conservadora), valor que corresponde a um máximo para n.

Exemplo: Suponhamos que, em uma amostra de 500 alunos que possuem computador em casa em uma cidade, haja 340 com Internet. Se quisermos estimar o número de alunos que possuem computador com Internet, qual o tamanho da amostra necessário para que tenhamos 95% de confiança em que erro o de nossa estimativa não seja superior a 0,02 (a) utilizando a informação dos 500 alunos (b) sem utilizar esta informação?

Solução:

(a) Tratemos os 500 alunos como uma amostra preliminar que fornece a estimativa $\hat{p}$ = 0,68. Então,

$$n = \frac{(1{,}96)^2 (0{,}68)(0{,}32)}{(0{,}02)^2} = 2.090$$

(b) Como não temos uma estimativa par p, substituímos $\hat{p}$ (1 – $\hat{p}$) por 1/4. Então, de acordo com a segunda expressão de n acima.

$$n = \frac{(1{,}96)^2}{4(0{,}02)^2} = 2.401$$

O tamanho da amostra é grande, pois o erro máximo admitido é pequeno (0,02).

Distribuição Amostral da Estatística χ^2

Se $X_1, X_2, ..., X_n$ é uma amostra aleatória de uma população normal com média μ e desvio padrão σ, então a distribuição de

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

segue uma distribuição qui-quadrado (χ^2) com $n - 1$ graus de liberdade

Intervalo de Confiança para Variância e o Desvio Padrão de uma População Normal Baseados na Distribuição Qui-Quadrado.

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

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

Etapas para a construção de Intervalo de 100(1 – α)% de Confiança para σ^2 e σ – População Normal.

- Colete uma amostra de tamanho *n* da população de interesse;
- Calcule o valor de S^2;
- Escolha o valor do coeficiente de confiança $1-\alpha$;
- Determine os valores de $\chi^2_{1-\alpha/2;n-1}$ e $\chi^2_{\alpha/2;n-1}$ apartir da tabela da distribuição qui-quadrado;
- Calcule os limites do intervalo de confiança para σ^2, para σ se extrai a raiz quadrada dos limites de confiança da equação (1).

Exemplo: Uma amostra de tamanho $n = 15$ de uma população normal tem média $\overline{X}$ = 26 e desvio padrão *S* = 3,32. Construa um intervalo de 95% de confiança para σ e para σ^2.

Solução:

n = 15; $\alpha/2$ = 0,025; graus de liberdade = 15 – 1 = 14; $1-\alpha/2$ = 0,975

Intervalo para σ^2: $\left[\frac{(14)(11{,}022)}{26{,}1} \leq \sigma^2 \leq \frac{(14)(11{,}022)}{5{,}629}\right] \Rightarrow$ [5,908; 27,414].

Intervalo para σ: [2,431; 5,236].

Exercícios propostos

1) Dado o seguinte conjunto de medidas,

0,0105 0,0193 0,0152 0,0229 0,0244

0,0190 0,0208 0,0279 0,0253 0,0276

Determinar: (a) um intervalo de 99% de confiança para a média populacional; (b) um intervalo de 95% de confiança para a variância populacional.

2) Admitindo as amostras extraídas de populações com variâncias conhecidas, determinar intervalos de confiança para as médias populacionais nos casos a seguir:

   (a) $n = 9$, $\overline{X} = 20$, $\sigma^2 = 9$ nível de confiança 95%;

   (b) $n = 25$, $\overline{X} = 120$, $\sigma^2 = 400$ nível de confiança 90%;

3) Supondo desconhecidas as variâncias e normais às populações, determinar intervalos de 95% de confiança para as médias populacionais nos casos a seguir:

   (a) $n = 9$, $\Sigma X_i = 36$, $\Sigma(X_i - \overline{X})^2 = 288$;

   (b) $n = 9$, $\Sigma X_i = 450$, $\Sigma(X_i - \overline{X})^2 = 32$.

4) Uma amostra de tamanho $n = 25$ de uma população normal com variância conhecida, determinar: (a) o nível de confiança $1 - \alpha$ se $n = 16$, $\sigma = 8$ e a amplitude do intervalo de confiança é 3,29; (b) o tamanho da amostra se $\sigma^2 = 100$ e o intervalo de 95% de confiança para a média é [17,2; 22,8]

5) Se uma população normal com desvio padrão conhecido $\sigma = 0{,}075$; qual deve ser o tamanho da amostra para que a amplitude total do intervalo de 95% de confiança para a média populacional seja no máximo igual a 0,04?

6) Em uma amostra aleatória de 200 eleitores, 114 são a favor de determinado projeto de lei. Determine um intervalo de 96% de confiança para a fração da população do município favorável ao projeto.

7) Uma amostra de 80 motoristas de determinado estado indica que o automóvel roda, em média, 22.000 km por ano, com desvio padrão de 3.800 km. Construa um intervalo de 98% de confiança para a rodagem anual média dos carros.

8) Encontre o coeficiente de confiança para p, se $n = 100$, $\hat{p} = 0{,}6$ e a amplitude do intervalo deve ser igual a 0,090.

9) Suponha que $X \sim N(\mu, \sigma^2)$, μ e σ^2 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 intervalo de 95% de confiança para μ.

10) Um psicólogo deseja fazer um teste em 25 pessoas de um grande grupo. Sabe-se que os resultados desse teste têm distribuição normal com desvio padrão $\sigma = 16$. A média dos resultados das 25 pessoas submetidas ao teste será usada como

estimativa global de toda população, sujeita à condição de que o “erro” máximo da estimativa seja 5. (a) Determinar a probabilidade de a diferença entre a estimativa e o verdadeiro valor da média ser no máximo 5. (b) Determinar o tamanho da amostra necessário para que a probabilidade em questão seja de 0,98.

5.2 Testes de Hipóteses

Conceitos:

Idéia básica: procurar condições que garantam que os resultados de experimentos possam ser generalizados além da situação experimental

Hipótese estatística: Consideração feita acerca de um parâmetro (ou característica) na população estudada.

O teste de hipóteses consiste na comparação de duas hipóteses, chamadas **Hipótese nula** e **Hipótese alternativa**.

Hipótese nula (H_0):

- Hipótese sobre a qual o teste é montado.
- Na maior parte dos casos é a hipótese de que "não há diferença".
- Em geral não é a hipótese que se deseja comprovar.

Hipótese alternativa (H_A):

Conclusões possíveis de um Teste de hipóteses:

- **Rejeita a Hipótese nula** (em favor da Hipótese alternativa considerada).

OU

- **Não rejeita a Hipótese nula** (em relação à Hipótese alternativa).

Planejamento amostral:

A comparação de duas hipóteses é feita baseada em evidências experimentais (amostras), sujeitas a erros amostrais e/ou erros não-amostrais.

Erros amostrais

Erros previstos no planejamento amostral, oriundos de flutuações amostrais - podem ser controlados e medidos.

Erros não-amostrais

Quaisquer erros que não estejam previstos no planejamento amostral - não

podem ser controlados nem medidos.

Erros na conclusão do teste de hipóteses:

Por causa das flutuações amostrais, ao comparar duas hipóteses e tomar uma decisão, pode-se tomar a decisão errada.

Dois tipos de erro:

- ***Erro Tipo I (α)***: Rejeitar a hipótese nula quando na realidade ela é verdadeira.
- ***Erro Tipo II (β)***: Não rejeitar a hipótese nula quando na realidade ela é falsa.

Como é possível rejeitar uma hipótese que é verdadeira?

O teste de hipóteses se baseia numa situação experimental (amostra), sujeita a flutuações na amostra. Devido a essas flutuações, pode-se ter uma amostra que não represente bem a população, levando a uma conclusão que não corresponde à realidade.

		Realidade (na população)	
		H_0 é verdadeira	***H_0 é falsa***
Conclusão do Teste (ação baseada na amostra)	***Aceitar H_0***	**correto**	***β*** **Erro Tipo II**
	Rejeitar H_0	***α*** **Erro Tipo I**	**Correto**

Como esses erros são evitados?

Montando testes que tornem esses erros os menores possíveis.

Não é possível minimizar ambos os erros ao mesmo tempo.

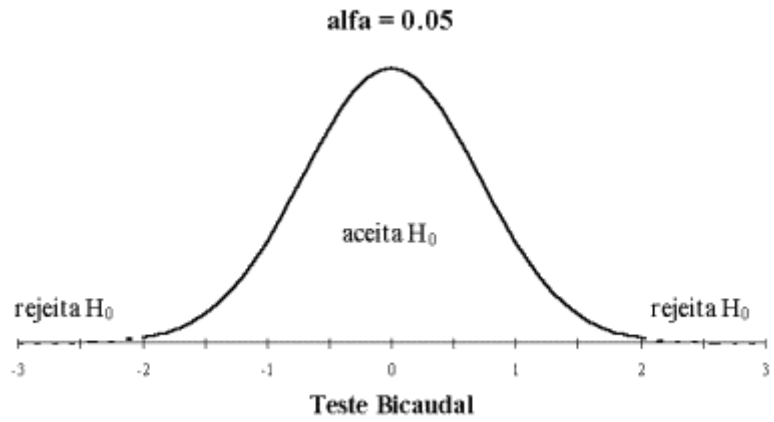
Os testes de hipóteses são montados de forma que, fixado o Erro Tipo I que se está disposto a cometer, o Erro Tipo II seja o menor.

Tipos de testes de hipóteses

1. Para detectar se existe alguma diferença (não importa em que direção)

$$H_0: \ \mu_1 = \mu_2 \quad \text{versus} \quad H_A: \ \mu_1 \neq \mu_2$$

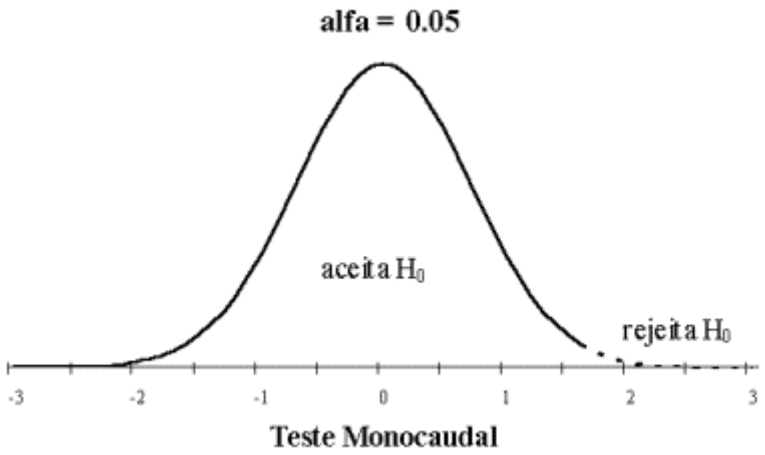
Teste de Hipóteses bicaudal



2. Para detectar se existe alguma diferença **E** em que direção ela está

$$H_0: \mu_1 \geq \mu_2 \quad \text{versus} \quad H_A: \mu_1 < \mu_2$$

Teste de Hipóteses monocaudal



Passos para a Construção de um Teste de Hipóteses

Passo 1. Fixe qual a hipótese H_0 a ser testada e qual a hipótese alternativa H_A;

Passo 2. Use a teoria estatística e as informações disponíveis para decidir qual estatística (estimador) será usada para testar a hipótese H_0. Obter as propriedades dessa estatística (distribuição, média, desvio padrão).

Passo 3. Fixe a probabilidade α de cometer o erro tipo I e use este valor para construir a região crítica (regra de decisão). Lembre-se esta região é construída para a estatística definida no passo 2, usando os valores do parâmetro hipotetizado por H_0;

Passo 4. Use as observações da amostra para calcular o valor da estatística de teste;

Passo 5. Se o valor da estatística calculado com os dados da amostra não pertencer à região crítica, não rejeite H_0 ; caso contrário, rejeite H_0.

O valor *p*

Definição: é o nome que se dá à probabilidade de se observar um resultado tão ou mais extremo que o da amostra, supondo que a hipótese nula seja verdadeira:

$$\boldsymbol{p} = P(Z \geq z \mid H_0)$$

Em outras palavras, ***p*** é a probabilidade de, supondo que a hipótese nula seja verdadeira, observar-se o valor registrado em um experimento por acaso.

Note que o valor ***p*** é calculado com base na amostra, enquanto que α é o maior valor ***p*** que leva à rejeição da hipótese nula.

Teste de Hipótese para Média em Populações Normais

Para as Hipóteses (σ^2 é conhecido):

H_0: $\mu = \mu_0$ $\begin{cases} vs\ (a)\ H_A : \mu > \mu_0\ \ (\mu_0\ \ é\ \ um\ \ valor\ \ especificado) \\ vs\ (b)\ H_A : \mu < \mu_0 \\ vs\ (c)\ H_A : \mu \neq \mu_0 \end{cases}$

Estatística de Teste: $Z = \dfrac{\overline{X} - \mu}{\sigma / \sqrt{n}}$

Regra de Decisão:

a) Rejeitamos H_0 se $\overline{X} \geq \mu_0 + z(\alpha)\dfrac{\sigma}{\sqrt{n}}$

b) Rejeitamos H_0 se $\overline{X} \leq \mu_0 + z(\alpha)\dfrac{\sigma}{\sqrt{n}}$

c) Rejeitamos H_0 se $\overline{X} \leq \mu_0 + z(\alpha)\dfrac{\sigma}{\sqrt{n}}$ ou $\overline{X} \geq \mu_0 + z(\alpha)\dfrac{\sigma}{\sqrt{n}}$

Exemplo 1: Considere que desejamos testar as hipóteses abaixo com respeito a uma média populacional μ.

H_0: $\mu = 40$

H_A: $\mu \neq 40$

Supondo que a população seja normalmente distribuída com variância $\sigma^2 = 9$ e que para uma amostra de tamanho $n = 25$ obteve-se uma média igual a 37,5. Realize um teste, ao nível de 10% de significância para essas hipóteses.

$\overline{X} = 37{,}5 \quad \overline{X} \sim N(\mu, \sigma^2/n)$

Rejeito H_0 se $\overline{X} \leq 40 - 1{,}64\frac{3}{\sqrt{25}} = 39{,}016$

ou se $\overline{X} \geq 40 + 1{,}64\frac{3}{\sqrt{25}} = 40{,}984$

Teste bicaudal;

Decisão: Rejeito H_0

Usando o valor p = 0,00001 rejeitamos H_0;

Conclusão: Ao nível de significância de 0,10 rejeitamos a hipótese de que a média μ é igual a 40, ou seja, aceitamos a hipótese alternativa.

Exemplo 2: O vice-presidente de produção de uma fábrica abriu uma filial em Pittsburgh, 15 meses atrás ele mediu a taxa de produção nos 3 primeiros meses, e encontrou uma média mensal de produção de 2000 unidades com desvio padrão de 21 unidade.

Ele quer saber se o nível de produção em Pittsburgh é diferente de 2000 no último ano ao nível de significância de 0,01.

Hipóteses: H_0: $\mu = 2000$

H_A: $\mu \neq 2000$ (teste bilateral)

α = a probabilidade de cometer o erro tipo 1, i.é. a probabilidade de rejeitar a hipótese nula quando verdadeira.

Os dados de produção do último ano foram: 2005, 2212, 2015, 2065, 2013, 2056, 2012, 2104, 2145, 2002, 2105, 2106.

A produção média dos últimos 12 meses é 2070, o desvio padrão da produção é 72

$$z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} = \frac{2070 - 2000}{72\sqrt{12}} = 3.369$$

Neste caso rejeita-se hipótese nula. Usando o valor p = 0,00037 rejeita-se a hipótese nula. Assume-se que a alternativa é correta e a produção não é 2000.

Exemplo 3: Uma companhia de cigarros anuncia que o índice médio de nicotina dos cigarros que fabrica apresenta-se abaixo de 23mg 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?

Solução:

Passo 1: H_0: $\mu \leq 23$ versus H_1: $\mu > 23$.

Passo 2: $\overline{X} \sim N(\mu;0{,}9^2)$.

Passo 3: $\alpha = 0{,}10$ $\quad \frac{\overline{X}_C - 23}{0{,}9} = 1{,}282 \Leftrightarrow \overline{X}_C = 24{,}15$. $RC =]24{,}15;+\infty[$

Passo 4: $\overline{X} = 24{,}17$.

Passo 5: Como o valor observado pertence à RC, rejeita-se H_0, ou seja, há evidências de que a informação do fabricante é falsa, ao nível significância de 10%.

Teste de Hipótese para Média em Populações Normais

Para as Hipóteses (σ^2 é desconhecido):

$$H_0: \mu = \mu_0 \begin{cases} vs\ (a)\ H_A: \mu > \mu_0\ \ (\mu_0\ \acute{e}\ um\ valor\ especificado) \\ vs\ (b)\ H_A: \mu < \mu_0 \\ vs\ (c)\ H_A: \mu \neq \mu_0 \end{cases}$$

Estatística de Teste: $\sigma\ desconhecido \rightarrow S \rightarrow T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$

Regra de Decisão:

a) Rejeitamos H_0 se $\overline{X} \geq \mu_0 + t_{(n-1;\alpha)} \frac{S}{\sqrt{n}}$

b) Rejeitamos H_0 se $\overline{X} \leq \mu_0 - t_{(n-1;\alpha)} \frac{S}{\sqrt{n}}$

c) Rejeitamos H_0 se $\overline{X} \leq \mu_0 - t_{(n-1;\alpha)} \frac{S}{\sqrt{n}}$ ou $\overline{X} \geq \mu_0 + t_{(n-1;\alpha)} \frac{S}{\sqrt{n}}$

Exemplo: Um fabricante de baterias declarou que a capacidade média de um tipo de bateria que sua empresa produz é de pelo menos 140 amperes por hora. Um órgão independente de proteção ao consumidor deseja testar a credibilidade da declaração do fabricante e mede a capacidade de uma amostra aleatória de 20 baterias, de um lote produzido recentemente. Os resultados em amperes/hora são os seguintes:

137,4 140,0 138,8 139,1 144,4 139,2 141,8 137,3 133,5 138,2

141,1 139,7 136,7 136,3 135,6 138,0 140,9 140,6 136,7 134,1

H_0: $\mu = 140$ n = 20 $\alpha = 0,05$

H_A: $\mu < 140$ $S^2 = 7,0706$ $S = 2,66$

$$\text{Rejeito } H_0 \text{ se } \overline{X} \leq 140 - 1,729\frac{2,66}{\sqrt{20}} = 138,97$$

Decisão: rejeita H_0

Conclusão: há evidências que levam a crer que a declaração do fabricante está superestimada e o órgão de proteção ao consumidor deveria dar início a uma ação corretiva contra a firma.

Teste de Hipótese com grandes amostras para uma Proporção populacional

Para as Hipóteses

$$H_0: p = p_0 \begin{cases} vs\ (a)\ H_A : p > p_0\ \ (p_0\ \ é\ um\ valor\ especificado) \\ vs\ (b)\ H_A : p < p_0 \\ vs\ (c)\ H_A : p \neq p_0 \end{cases}$$

Se n é grande $\hat{p} \approx Normal\left(p, \frac{p(1-p)}{n}\right)$

Estatística de Teste: $Z_0 = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$

Regra de Decisão:

a) Rejeitamos H_0 se $Z_0 \geq z_{(\alpha)}$

b) Rejeitamos H_0 se $Z_0 \leq -z_{(\alpha)}$

c) Rejeitamos H_0 se $Z_0 \geq z_{(\alpha/2)}$ ou $Z_0 \leq -z_{(\alpha/2)}$

Exemplo: 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, nos níveis de 5% e 1%.

p = proporção de equipamentos que estão de acordo com as especificações exigidas.

H_0: $p = 0,9$ $\qquad\qquad$ $\hat{p} = 1 - \left(\frac{25}{200}\right) = 0,88$

H_A: $p < 0,9$

$\alpha = 0,05$ e $\alpha = 0,01$ $\qquad$ $p_0 = 0,90$ $\qquad$ n = 200

Como a amostra é igual (n = 200), vale a aproximação normal

$$Z_0 = \frac{0,88 - 0,9}{\sqrt{\frac{0,9(1-0,9)}{200}}} = -0,94$$

Rej. H_0 (90% dos equip. de acordo c/ especf.) se $Z_0 \leq -z_{(\alpha)}$

Como –0,94 > -1,96 e –0,94 > -2,57

Decisão: Não rejeitamos H_0

Conclusão: Aos níveis de significância de 5% e 1% não há evidências de que os equipamentos estejam fora das especificações exigidas.

Teste de Hipótese para uma variância populacional (com população normal)

Para as Hipóteses

H_0: $\sigma^2 = \sigma_0^2$ $\begin{cases} vs\ (a)\ H_A : \sigma^2 > \sigma_0^2 \ \ (\sigma_0^2\ é\ um\ valor\ especificado) \\ vs\ (b)\ H_A : \sigma^2 < \sigma_0^2 \\ vs\ (c)\ H_A : \sigma^2 \neq \sigma_0^2 \end{cases}$

Estatística de Teste: $\chi_0^2 = \frac{(n-1)S^2}{\sigma_0^2}$

Regra de Decisão:

a) Rejeitamos H_0 se $\chi_0^2 \geq \chi^2_{(1-\alpha)(n-1)}$

b) Rejeitamos H_0 se $\chi_0^2 \leq \chi^2_{\alpha(n-1)}$

c) Rejeitamos H_0 se $\chi_0^2 \geq \chi^2_{(1-\alpha/2)(n-1)}$ ou $\chi_0^2 \leq \chi^2_{\alpha/2(n-1)}$

Exemplo: Uma indústria deseja controlar flutuações da espessura dos plásticos, há informações que se o verdadeiro desvio padrão da espessura excede 1,5mm, então há motivos para preocupação com a qualidade do produto. Foi retirada uma amostra de 10 folhas de plásticos em uma particular produção, mediram-se as espessuras dessas folhas onde S^2 = 5,155. Os dados dão suspeitos de que a variabilidade da espessura excede o nível estabelecido nessa produção?

σ = 1,5mm H_0: σ^2 = 2,25

n = 10 H_A: $\sigma^2 > 2,25$

σ^2 = 2,25mm2 α = 0,05

Admitimos que as espessuras das folhas dos plásticos seguem uma distribuição normal.

$$\chi_0^2 = \frac{(10-1)5,155}{2,25} = 20,62$$

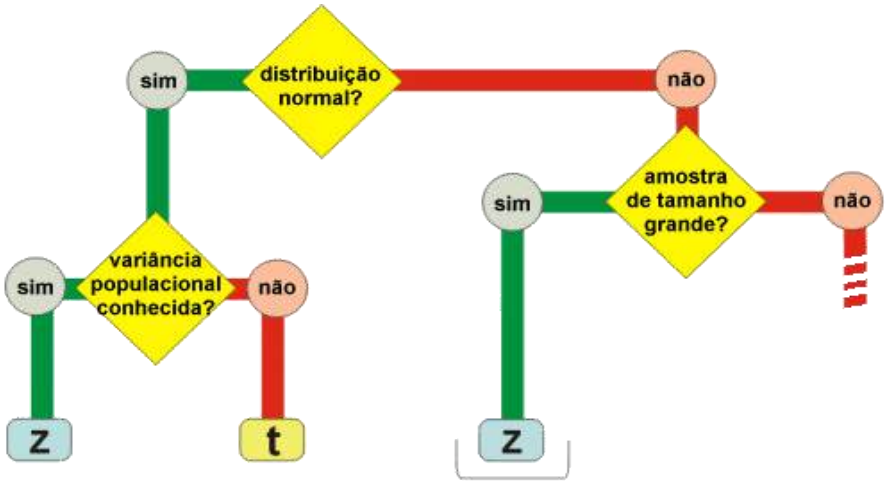
$$\chi^2_{(9;0,05)} = 16,919$$

Rej. H_0 (o desv. Pad. não excede 1,5mm)

Decisão: rejeita H_0

Conclusão: ao nível de significância de 5% há evidências que a variabilidade está excedendo ao valor especificado.

Resumo para o Teste de Hipóteses



Teorema Central do Limite

Exercícios propostos

1) Uma amostra aleatória de tamanho n = 18 de uma população normal tem média $\overline{X}$ = 31,5 e desvio padrão S = 4,2. No nível de significância de 5% esses dados sugerem que a média populacional seja superior a 30?
2) Sabe-se, por experiência passada, que uma população se distribui normalmente, com variância igual a 100, e quer-se saber se, através de uma amostra de n = 25, pode-se admitir que a verdadeira média populacional seja maior que 50. Qual a região de rejeição do teste? Adotar α = 0,05.
3) Identifique a hipótese nula e alternativa em cada caso explique os erros tipo I e tipo II.
   (a) Um agrônomo acredita que uma nova variedade de sementes produz plantas mais resistentes a uma determinada doença do que a variedade existente. Ele planeja expor os dois tipos. Os dados coletados serão usados para verificar a veracidade de sua conjectura.
   (b) A Secretaria de Trabalho de um estado deseja determinar se a taxa de desempregado no estado varia significativamente da previsão de 6% feita dois meses antes.
4) 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 33% e 29%, respectivamente. Os resultados trazem alguma evidência de mudança de preferência?
5) O número médio diário de clientes de um posto de gasolina tem sido 250, com um desvio padrão de 80 clientes. Durante uma campanha de 25 dias, em que os clientes recebiam um brinde, o número médio de clientes foi de 280, com um desvio padrão de 50. Você diria que a campanha modificou a distribuição do número de clientes do posto? Descreva as suposições feitas para a resolução do problema.
6) Um jornal alega que 25% dos seus leitores pertencem à classe A. Que regra de decisão você adotaria para testar essa hipótese, contra a alternativa de que a percentagem verdadeira não é 25% no nível de 5% de significância? Se em uma amostra de 740 leitores encontramos 156 de classe A, qual a sua decisão a respeito da veracidade da alegação veiculada pelo jornal?

5.3 Comparações de Parâmetros: O caso de duas Populações

As comparações de dois grupos geralmente podem ser traduzidas, na linguagem estatística, em comparações de duas médias, duas variâncias ou duas proporções, conforme será apresentado nas próximas seções deste capítulo.

Comparação de duas Médias – Amostras Independentes

Duas amostras coletadas de duas populações são independentes se a extração da amostra de uma das populações não afeta a extração da amostra da outra população.

Estrutura dos Dados – Amostras Aleatórias Independentes

(a) $x_{11}, x_{12}, \ldots, x_{1n_1}$ é uma amostra aleatória de tamanho n_1 da população 1, cuja média é representada por μ_1 e cuja variância é representada por σ_1^2.

(b) $x_{21}, x_{22}, \ldots, x_{2n_2}$ é uma amostra aleatória de tamanho n_2 da população 2, cuja média é representada por μ_2 e cuja variância é representada por σ_2^2.

(c) $x_{11}, x_{12}, \ldots, x_{1n_1}$ são independentes de $x_{21}, x_{22}, \ldots, x_{2n_2}$. Ou seja, as respostas obtidas para uma população não têm qualquer relação com as respostas coletadas para a outra população.

Testes para $\mu_1 - \mu_2$ – Grandes Amostras

Quando $n_1 > 30$ e $n_2 > 30$, o teste de $H_0 : \mu_1 = \mu_2$ é baseado na estatística de teste:

$$Z_0 = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

Que tem, aproximadamente, distribuição normal padronizada. A região crítica, para um teste com nível de significância α, depende da hipótese alternativa:

Hipótese Alternativa Região Crítica

$$H_1 : \mu_1 \neq \mu_2 \quad \Rightarrow \quad Z_0 > Z_{\alpha/2} \text{ ou } Z_0 < -Z_{\alpha/2}$$

$$H_1 : \mu_1 > \mu_2 \quad \Rightarrow \quad Z_0 > Z_{\alpha}$$

$$H_1 : \mu_1 < \mu_2 \quad \Rightarrow \quad Z_0 < -Z_{\alpha}$$

Etapas para Realização do Teste de Hipótese

1) Identifique o parâmetro de interesse;
2) Estabeleça a hipótese nula H_0 e a alternativa H_1;
3) Escolha o nível de significância;
4) Determine a estatística de teste apropriada;
5) Determine a região crítica do teste;
6) Faça os cálculos necessários a partir dos dados amostrais e determine o valor da estatística de teste;
7) Decida se a hipótese nula H_0 deve ser ou não rejeitada;

   Uma abordagem alternativa para tomada da decisão quanto à rejeição, ou não, de H_0 consiste em utilizar o p-valor da estatística de teste observada (utilizando "softwares" conforme capítulo 9 e 10);
8) Apresente a decisão no contexto do problema que está sendo analisado.

Intervalo de Confiança para a Diferença entre Duas Médias

Um intervalo de 100(1 – α)% de confiança para $\mu_1 - \mu_2$, no caso de grandes amostras, é dado por:

$$\left(\overline{X}_1 - \overline{X}_2\right) \pm Z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

É possível mostrar que, no teste de

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

ou, o que é equivalente,

$$H_0 : \mu_1 - \mu_2 = 0$$

$$H_1 : \mu_1 - \mu_2 \neq 0$$

a hipótese nula H_0 será rejeitada ao nível de significância α se o número zero não pertencer ao intervalo de 100(1 – α)% de confiança para $\mu_1 - \mu_2$.

Testes para $\mu_1 - \mu_2$ – Pequenas Amostras, Variâncias Desconhecidas mas Supostamente Iguais

Suponha que as populações de interesse tenham distribuições normais com médias μ_1 e μ_2 e variâncias desconhecidas mas supostas iguais $\sigma_1^2 = \sigma_2^2 = \sigma^2$. Quando são extraídas amostras aleatórias independentes $n_1 < 30$ e $n_2 < 30$ destas populações, o teste de $H_0 : \mu_1 = \mu_2$ é baseado na estatística de teste

$$t_0 = \frac{\overline{X}_1 - \overline{X}_2}{S_c\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \text{ onde } S_c = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}} \qquad (1)$$

que tem distribuição *t* de Student com $n_1 + n_2 - 2$ graus de liberdade. A região crítica, para um teste com nível de significância α, depende da hipótese alternativa:

Hipótese Alternativa		Região Crítica
$H_1 : \mu_1 \neq \mu_2$	$\Rightarrow$	$t_0 > t_{\alpha/2;\, n_1+n_2-2}$ ou $t_0 < -t_{\alpha/2;\, n_1+n_2-2}$
$H_1 : \mu_1 > \mu_2$	$\Rightarrow$	$t_0 > t_{\alpha;\, n_1+n_2-2}$
$H_1 : \mu_1 < \mu_2$	$\Rightarrow$	$t_0 < -t_{\alpha;\, n_1+n_2-2}$

Para a realização do teste de hipótese siga os passos explicados anteriormente.

Intervalo de Confiança para a Diferença entre Duas Médias – Pequenas Amostras, Variâncias Desconhecidas, mas Suposta Iguais

Um intervalo de 100(1 – α)% de confiança para $\mu_1 - \mu_2$, no caso de pequenas amostras e variância populacionais desconhecidas mas supostas iguais, é dado por:

$$\left(\overline{X}_1 - \overline{X}_2\right) \pm t_{\alpha/2;\, n_1+n_2-2} S_c \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Exemplo: Em um ensaio clínico comparou-se dois anorexígenos e as perdas de peso foram registradas na tabela abaixo:

Paciente	Droga 1	Paciente	Droga 2
1	0,9	7	3,8

2	1,3	8	4,9
3	1,5	9	5,9
4	2,4	10	6,6
5	2,9	11	6,7
6	3,0	12	7,1
		13	7,0

Existe diferença de pesos pela administração das drogas?

Solução:

Hipóteses a serem testadas: $H_0 : \mu_1 = \mu_2$ vs. $H_1 : \mu_1 \neq \mu_2$

Consideramos que σ_1^2 e σ_2^2 são similares então:

$S_1 = 0{,}784$ $S_2 = 1{,}520$

$$S_c = \sqrt{\frac{5(0{,}784) + 6(1{,}520)}{5+6}} = \sqrt{1{,}185} = 1{,}089$$

Estatística de teste: $t_0 = \dfrac{2-6}{1{,}089\sqrt{\frac{1}{6}+\frac{1}{7}}} = \dfrac{-4}{0{,}606} = -6{,}60$

graus de liberdade = 6 + 7 – 2 = 11; $t_{0{,}025;\,11} = 2{,}2010$

Região crítica: $|\, t_0 \,| > 2{,}2010$. P-valor $\Rightarrow P(t_{0{,}025;\,11} \geq |\, -6{,}60 \,|\,) \cong 0{,}0000$

O valor calculado de T está na região crítica de rejeição. Portanto, os dados nos dão evidências para rejeição de H_0, e concluímos que há diferença significativa entre as médias de pesos dos dois grupos (droga 1 e droga2).

Testes para $\mu_1 - \mu_2$ – Pequenas Amostras Variâncias Desconhecidas e Desiguais

Em algumas situações, não podemos considerar que as variâncias desconhecidas σ_1^2 e σ_2^2 das duas populações normais de interesse sejam iguais. Quando são extraídas amostras aleatórias independentes $n_1 < 30$ e $n_2 < 30$ destas populações, o teste de $H_0 : \mu_1 = \mu_2$ é baseado na estatística de teste

$$t_0^* = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

que tem, aproximadamente, distribuição *t* de Student com v graus de liberdade, onde

$$v = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\left(S_1^2/n_1\right)^2/(n_1+1) + \left(S_2^2/n_2\right)^2/(n_2+1)} - 2$$

. A região crítica, para um teste com nível de significância α, depende da hipótese alternativa:

Hipótese Alternativa		Região Crítica
$H_1 : \mu_1 \neq \mu_2$	$\Rightarrow$	$t_0^* > t_{\alpha/2;\, v}$ ou $t_0^* < -t_{\alpha/2;\, v}$
$H_1 : \mu_1 > \mu_2$	$\Rightarrow$	$t_0^* > t_{\alpha;\, v}$
$H_1 : \mu_1 < \mu_2$	$\Rightarrow$	$t_0^* < -t_{\alpha;\, v}$

Para a realização do teste de hipótese siga os passos dados anteriormente.

Intervalo de Confiança para a Diferença entre Duas Médias – Pequenas Amostras Variâncias Desconhecidas e Desiguais

Um intervalo de 100(1 – α)% de confiança para $\mu_1 - \mu_2$, no caso de pequenas amostras e variância populacionais desconhecidas e desiguais, é dado por:

$$\left(\overline{X}_1 - \overline{X}_2\right) \pm t^*_{\alpha/2;\, v} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

Exemplo: Comparação das alturas de crianças de 4 anos de duas populações diferentes:

	População 1	População 2
Média altura: $\overline{X}$	42	45
Variância: S^2	4	25
Amostra: *n*	10	10

Considerando as variâncias amostrais diferentes

Hipóteses a serem testadas: H_0: $\mu_1 = \mu_2$ vs. H_1: $\mu_1 \neq \mu_2$; $\alpha = 5\%$

Calcula-se o valor de t: $t_0^* = \frac{42-45}{\sqrt{\frac{4}{10}+\frac{25}{10}}} = -\frac{3}{1{,}7} = -1{,}76$

Graus de liberdade = 12; $t_{0{,}025;\,12} = 2{,}1788$

Região crítica: $|\ t_0^*\ | > 2{,}1788$. P-valor $\Rightarrow P(t_{0{,}025;\,12} \geq |\ -1{,}76\ |\ ) \cong 0{,}0519$

O valor calculado de T não está na região crítica de rejeição. Portanto, os dados nos dão evidências para não rejeição de H_0, e concluímos que as crianças não têm altura média diferente.

Comparação de duas Médias – Amostras Emparelhadas

As amostras extraídas de duas populações são emparelhadas quando seus elementos são coletados em pares, de modo que os dois elementos de um mesmo par são muito similares ou homogêneos em relação às características que possam estar relacionadas à variável de interesse no estudo.

Teste t Emparelhado

Suponha que $x_{11}, x_{12}, ..., x_{1n_1}$ e $x_{21}, x_{22}, ..., x_{2n_2}$ são duas amostras aleatórias, só que agora as observações estão emparelhadas, isto é, a amostra é formada pelos pares

$$(x_{11}, x_{21}); (x_{12}, x_{22}); ...;(x_{1n_1}, x_{2n_2})$$

onde os pares são independentes entre si.

Seja D_i as diferenças entre cada par de observações como

$$D_i = x_{1i} - x_{2i}$$

passamos a ter a amostra aleatória

$$D_1, D_2, ..., D_n$$

É possível mostrar que podemos testar hipóteses sobre a diferença entre as duas médias populacionais μ_1 e μ_2 por meio da realização de teste sobre a média da distribuição das diferenças (μ_D), sendo válida a relação

$$\mu_D = \mu_1 - \mu_2$$

Logo, testar as hipóteses:

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

é equivalente a testar as hipóteses:

$$H_0 : \mu_D = 0$$

$$H_1 : \mu_D \neq 0$$

O teste de $H_0 : \mu_D = 0$ é baseado na estatística de teste

$$t_0 = \frac{\overline{D}}{S_D/\sqrt{n}}, \text{ onde}$$

- $\overline{D} = \frac{1}{n}\sum_{i=1}^{n} D_i$ é média da amostra das diferenças;
- $S_D = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(D_i - \overline{D})^2}$ é o desvio padrão da amostra das diferenças.

t_0 tem distribuição *t* de Student com n – 1 graus de liberdade. A região crítica, para um teste com nível de significância α, depende da hipótese alternativa:

Hipótese Alternativa		Região Crítica
$H_1 : \mu_1 \neq \mu_2$	$\Rightarrow$	$t_0 > t_{\alpha/2;\, n-1}$ ou $t_0 < -t_{\alpha/2;\, n-1}$
$H_1 : \mu_1 > \mu_2$	$\Rightarrow$	$t_0 > t_{\alpha;\, n-1}$
$H_1 : \mu_1 < \mu_2$	$\Rightarrow$	$t_0 < -t_{\alpha;\, n-1}$

Para a realização do teste de hipótese siga os passos dados anteriormente.

Intervalo de Confiança para a Diferença entre Duas Médias – Amostras Emparelhadas

Um intervalo de 100(1 – α)% de confiança para μ_D, é dado por:

$$\overline{D} \pm t_{\alpha/2;\, n-1}\frac{S_D}{\sqrt{n}}$$

Exemplo: Alega-se que certa dieta reduz de 4,0 kg, em média, o peso de uma pessoa em duas semanas. Sete pessoas submetem-se à dieta. Os pesos respectivos, antes e depois da dieta, são:

Antes	62,0	64,0	61,0	67,0	66,0	59,5	65,0
Depois	63,0	56,0	57,5	61,8	63,0	54,5	64,0
Diferença	-1,0	8,0	3,5	5,2	3,0	5,0	1,0

Testar em um nível de 5% as hipóteses: $H_0 : \mu_D = 4$ vs. $H_1 : \mu_D > 4$.

Solução:

Teste unilateral:

$\overline{D} = 3{,}53 \quad S_D = 2{,}95$

Estatística de teste:

$$T = \frac{3{,}53 - 4{,}0}{2{,}95/\sqrt{7}} = \frac{-0{,}47}{1{,}11} = -0{,}42$$

graus de liberdade = 7 – 1 = 6; $t_{0{,}05;\,6} = 1{,}943$

Região crítica: t > 1,943. P-valor $\Rightarrow$ $P(t_{0{,}05;\,6} \geq -0{,}42) = 0{,}3435$.

T calculado está na região de não rejeição. Portanto, os dados nos dão evidências para não rejeitarmos H_0. Concluímos que a redução média é igual a 4 kg.

A Distribuição F – Distribuição Amostral do Quociente de Duas Variâncias

Suponha que duas amostras independentes, de tamanho n_1 e n_2, são retiradas de duas populações com distribuição normal com mesma variância σ^2. Representaremos as variâncias por S_1^2 e S_2^2, respectivamente. Nestas condições, o quociente

$$F_0 = \frac{S_1^2}{S_2^2}$$

tem distribuição F com ($n_1 - 1$, $n_2 - 1$) graus de liberdade.

Como no caso da distribuição t, a qualificação com ($n_1 - 1$, $n_2 - 1$) graus de liberdade é necessária porque para cada valor diferente dos tamanhos das amostras n_1

e n_2 (ou, o que equivalente, para cada valor diferente de ($n_1 - 1$, $n_2 - 1$)) existe uma distribuição F específica.

A distribuição F é contínua e assimétrica, com F assumindo apenas valores maiores ou iguais a zero. Assim como as distribuições, normal padronizada e t, a distribuição F também é tabelada.

Testes para Comparação de duas Variâncias – Populações Normais

O procedimento estatístico para comparar as variâncias de duas populações é baseado no seguinte conjunto de suposições:

(a) $x_{11}, x_{12}, \ldots, x_{1n_1}$ é uma amostra aleatória de uma população normal $N(\mu_1, \sigma_1^2)$

(b) $x_{21}, x_{22}, \ldots, x_{2n_2}$ é uma amostra aleatória de uma população normal $N(\mu_2, \sigma_2^2)$

(c) As duas amostras são independentes.

O teste de $H_0: \sigma_1^2/\sigma_2^2 = 1$ é baseado na estatística de teste

$$F_0 = \frac{S_1^2}{S_2^2}$$

que tem distribuição F com ($n_1 - 1$, $n_2 - 1$) graus de liberdade. A região crítica, para um teste com nível de significância α, depende da hipótese alternativa:

Hipótese Alternativa		Região Crítica
$H_1: \sigma_1^2 \neq \sigma_2^2$	$\Rightarrow$	$F_0 > F_{\alpha/2;\, n_1-1,\, n_2-1}$ ou $F_0 < F_{1-\alpha/2;\, n_1-1,\, n_2-1}$
$H_1: \sigma_1^2 > \sigma_2^2$	$\Rightarrow$	$F_0 > F_{\alpha;\, n_1-1,\, n_2-1}$
$H_1: \sigma_1^2 < \sigma_2^2$	$\Rightarrow$	$F_0 < F_{1-\alpha;\, n_1-1,\, n_2-1}$

Intervalo de Confiança para a Razão das Variâncias de duas Populações Normais

Um intervalo de 100(1 – α)% de confiança para σ_1^2/σ_2^2 é dado por:

$$\frac{S_1^2}{S_2^2} F_{1-\alpha/2;\ n_1-1,\ n_2-1} \leq \frac{\sigma_1^2}{\sigma_2^2} \leq \frac{S_1^2}{S_2^2} F_{\alpha/2;\ n_1-1,\ n_2-1}$$

Exemplo: Diferença entre dois hipnóticos (horas de sono frente a duas terapias):

Paciente	terapia tradicional	Paciente	terapia nova
1	8,2	9	8,1
2	9,4	10	8,8
3	8,9	11	8
4	8,5	12	9,5
5	9,8	13	9,5
6	8,9	14	9,6
7	10,4	15	9,5
8	10,7	16	7,8
$\overline{X}$	9,35		8,85
S	0,89		0,78
N	8		8

Hipóteses a serem testadas: $H_0 : \sigma_1^2 = \sigma_2^2$ vs $H_1 : \sigma_1^2 \neq \sigma_2^2$; α = 5%

Estatística de teste: $F = \frac{0{,}89^2}{0{,}78^2} = 1{,}30$

Graus de liberdade = (8 – 1), (8 – 1) = 7 numerador, 7 denominador ; $F_{0{,}975;\ 7,\ 7}$ = 3,7880

P-valor $\Rightarrow$ P($F_{0{,}975;\ 7,\ 7} \geq 1{,}30$) = 0,3609.

Os dados nos dão evidências para não rejeitarmos H_0. Concluímos que as variâncias são homogêneas.

Testes para Comparação de duas Variâncias – Populações Não Necessariamente Normais – Grandes Amostras

Se $n_1 \geq 40$ e $n_2 \geq 40$, o teste de $H_0 : \sigma_1^2 = \sigma_2^2$ é baseado na estatística de teste

$$Z_0 = \frac{S_1 - S_2}{S_c \sqrt{\frac{1}{2n_1} + \frac{1}{2n_2}}} \qquad (2)$$

que tem, aproximadamente, distribuição normal padronizada. A região crítica, para um teste com nível de significância α, depende da hipótese alternativa:

Hipótese Alternativa		Região Crítica
$H_1 : \sigma_1^2 \neq \sigma_2^2$	$\Rightarrow$	$Z_0 > Z_{\alpha/2}$ ou $Z_0 < -Z_{\alpha/2}$
$H_1 : \sigma_1^2 > \sigma_2^2$	$\Rightarrow$	$Z_0 > Z_{\alpha}$
$H_1 : \sigma_1^2 < \sigma_2^2$	$\Rightarrow$	$Z_0 < -Z_{\alpha}$

Na equação (2) S_c é a estimativa combinada do desvio padrão σ, que é apresentada na equação (1).

Testes para Comparação de duas Proporções p_1 e p_2 no Caso de Grandes Amostras

Em muitas situações pode ser de interesse avaliar a veracidade de alguma hipótese sobre as proporções de elementos de duas populações que possuem alguma característica de interesse (p_1 e p_2).

Se n_1p_1, $n_1(1 – p_1)$, n_2p_2 e $n_2(1 – p_2)$ forem maiores ou iguais a 5, o teste de $H_0 : p_1 = p_2$ é baseado na estatística de teste

$$Z_0 = \frac{\bar{p}_1 - \bar{p}_2}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

que tem, aproximadamente, distribuição normal padronizada. A região crítica, para um teste com nível de significância α, depende da hipótese alternativa:

Hipótese Alternativa		Região Crítica
$H_1 : p_1 \neq p_2$	$\Rightarrow$	$Z_0 > Z_{\alpha/2}$ ou $Z_0 < -Z_{\alpha/2}$
$H_1 : p_1 > p_2$	$\Rightarrow$	$Z_0 > Z_{\alpha}$
$H_1 : p_1 < p_2$	$\Rightarrow$	$Z_0 < -Z_{\alpha}$

Intervalo de Confiança para a Diferença entre duas Proporções

Um intervalo de 100(1 – α)% de confiança para $p_1 – p_2$ é dado por:

$$(\bar{p}_1 - \bar{p}_2) \pm Z_{\alpha/2}\sqrt{\frac{\bar{p}_1(1-\bar{p}_1)}{n_1} + \frac{\bar{p}_2(1-\bar{p}_{21})}{n_{21}}}$$

Exemplo: Em um estudo sobre a incidência de abortos naturais entre médicas anestesistas e de outras especialidades, obtiveram-se os seguintes resultados:

Tabela: Incidência de abortos naturais entre médicas anestesistas e de outras especialidades.

	Anestesistas	Outras Especialidades	Total
Gestações normais	23	52	75
Abortos naturais	14	6	20
Total	37	58	95

Os estimadores de p_1 e p_2, proporções de abortos naturais, são dados por:

$$\hat{p}_1 = \frac{14}{37} = 0{,}378 \quad \hat{p}_2 = \frac{6}{58} = 0{,}103.$$

Para o desvio padrão da diferença, temos:

$$\hat{p} = \frac{14+6}{37+58} = \frac{20}{95} = 0{,}211$$

$$\hat{DP}(\hat{p}_1 - \hat{p}_2) = \sqrt{(0{,}211)(0{,}789)\left(\frac{1}{37} + \frac{1}{58}\right)} = 0{,}086$$

A estatística Z é então dada por: $Z = \frac{0{,}378 - 0{,}103}{0{,}086} = 3{,}198$

Consultando a tabela normal, vemos que a hipótese nula pode ser rejeitada em um nível de significância aproximado $\alpha = 0{,}001$.

6. Tabelas de Contingência

6.1 Teste Qui-Quadrado

No período de 4 semanas iniciando no dia **1º de maio**, o Pronto Socorro Médico do Hospital das Clínicas recebeu 225 pacientes em crise hipertensiva distribuídos segundo os dias de atendimento conforme a tabela abaixo:

Dias da semana	Dom	Seg	Ter	Qua	Qui	Sex	Sab
Pac. Atendidos	38	25	30	28	33	31	40

Pergunta-se:

A distribuição de pacientes ao longo da semana é homogênea?

O coeficiente qui-quadrado, χ^2, mede a "distância" entre os valores observados em uma amostra e os valores hipotéticos esperados (a diferença é relativa ao valor esperado).

Em geral, se O_1, O_2, ..., O_n são valores observados na amostra, de cada um dos **n** atributos, e E_1, E_2, ..., E_n são os respectivos valores esperados, definimos χ^2 como

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Para grandes amostras, χ^2 aproxima-se da distribuição qui-quadrado com (n – 1) graus de liberdade. Rejeitamos a hipótese nula para valores elevados de χ^2.

Voltando ao exemplo, queremos saber se:

H_0: A freqüência observada obedece à distribuição $p_1 = p_2 = ... = p_n$;

H_1: Os pacientes compareceram preferencialmente em determinados dias da semana.

Categorias	p_i	o_i	e_i	$o_i - e_i$	$(o_i - e_i)^2/e_i$
Dom	1/7	38	32,1	5,86	1,067
Seg	1/7	25	32,1	-7,14	1,587
Ter	1/7	30	32,1	-2,14	0,142
Qua	1/7	28	32,1	-4,14	0,534
Qui	1/7	33	32,1	0,86	0,022
Sex	1/7	31	32,1	-1,14	0,041
Sab	1/7	40	32,1	7,86	1,921
Total	1	225	225	0	5,316

Portanto,

$$\chi^2 \cong 5{,}316$$

O número de graus de liberdade é dado por:

$$\text{g.l.} = n - 1 = 7 - 1 = 6$$

ou seja, o número de categorias menos 1.

Da tabela qui-quadrado, para g.l.= 6 e nível de significância 5%, temos:

$$\chi^2 \text{ (g.l.} = 6, \alpha = 0{,}05) = 12{,}6$$

O p-valor é dado por $P(\chi^2 \geq 5{,}316) \cong 0{,}6$ ou 60%

Desta forma, não podemos rejeitar a hipótese nula de que a distribuição de pacientes é homogênea ao longo da semana.

Teste χ^2 para grandes amostras

Uma importante aplicação do teste χ^2 ocorre quando se quer estudar as relações entre duas ou mais variáveis de classificação. A representação das freqüências observadas, nesse caso, pode ser feita por meio de uma tabela de contingência.

Hipóteses a serem testadas:

Ho: As variáveis são independentes (não estão associadas)

H_1: As variáveis não são independentes (estão associadas)

Estatística de teste:

$$\chi^2_{cal} = \sum_{i=1}^{L}\sum_{j=1}^{C} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}, \text{ onde } E_{ij} = \frac{(\text{soma da linha i})(\text{soma da coluna j})}{\text{total de observações}}$$

O número de graus de liberdade é dado por:

(L - 1) (C - 1), onde L é o número de linhas e C o número de colunas da tabela de contingência.

Regra de descisão:

Fixando um nível α rejeita-se H_0

se $\chi^2_{cal} > \chi^2_{\alpha;\,(L-1)(C-1)}$;

O p-valor é dado a partir da $P(\chi^2 \geq \chi^2_{cal})$.

Exemplo: Efeito de uma droga comparada com placebo (tabela de contingência 2 x 2). Uma droga nova foi testada em 56 pacientes, outros 51 receberam placebo. 48 dos que receberam a droga melhoraram e 38 dos que receberam placebo melhoraram.

	Efeito		
Estímulo	positivo	negativo	Total marginal
Droga	48	8	56
Placebo	38	13	51
Total marginal	86	21	107

Pergunta: A resposta foi contingente ao estímulo?

Note que esta tabela, tomando-se a percentagem de efeito obtido com droga e placebo (i.e., as proporções tomadas em relação aos totais marginais das linhas) resulta em:

	Efeito	
Estímulo	positivo	negativo
Droga	0,875	0,143
Placebo	0,745	0,255

Esta tabela, obtida a partir dos valores observados acima, tomando-se a percentagem em relação aos **valores esperados**:

	Efeito		
Estímulo	positivo	negativo	Total marginal
Droga	**45,01**	**10,99**	56
Placebo	**40,99**	**10,01**	51
Total marginal	86	21	107

resulta em:

	Efeito	
Estímulo	Positivo	negativo
Droga	0,804	0,196
Placebo	0,804	0,196

H_0: Não há diferença entre os grupos que receberam a droga e placebo em relação à melhora dos pacientes.

Solução:

Neste exemplo

$$E_{11} = \frac{86x56}{107} = 45{,}01 \quad E_{21} = \frac{86x51}{107} = 40{,}99 \quad E_{12} = \frac{21x56}{107} = 10{,}99 \quad E_{22} = \frac{21x51}{107} = 10{,}01$$

Estímulo	positiva	negativa	Total marginal
Droga	48 45,01	8 10,99	56
Placebo	38 40,99	13 10,01	51
Total marginal	86	21	107

Determinam-se as freqüências esperadas, (E_{ij}),

Sendo que $\chi^2_{cal} = \sum_{i=1}^{L}\sum_{j=1}^{C} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$, temos

$$\chi^2_{cal} = \frac{(48-45{,}01)^2}{45{,}01} + \frac{(38-40{,}99)^2}{40{,}99} + \frac{(8-10{,}99)^2}{10{,}99} + \frac{(13-10{,}01)^2}{10{,}01} \cong 2{,}123$$

Pela tabela de distribuição de qui-quadrado, obtemos:

$$\chi^2 \text{ (g.l. = 1, } \alpha = 0{,}05) = 3{,}84$$

O p-valor é $P(\chi^2 \geq 2{,}123) \cong 0{,}15$ ou 15%.

Como o valor obtido é menor que o valor crítico não podemos rejeitar a hipótese nula para erro α = 5%. Logo, não podemos dizer que os efeitos da droga e do placebo sejam significativamente diferentes.

Restrições ao uso do Teste Qui-quadrado

Ao utilizar-se o teste qui-quadrado supõe-se que o tamanho das amostras seja "grande".

"Em situações práticas, o valor de χ^2 calculado é aproximado porque se utiliza amostras de tamanho finito e o valor da freqüência observada pode assumir somente números inteiros. Por exemplo, você jamais vai encontrar 2,87 casos de pacientes doentes."

Portanto é necessário observar as seguintes restrições:

(1) Regra geral, o teste χ^2 pode ser usado se o número de observações em cada casela da tabela for maior ou igual a 5 e a menor freqüência esperada for maior ou igual a 5 ($E_{ij} \geq 5$, $O_{ij} \geq 5$).

(2) Tabela 2x2 (g.l. *= 1*): **Fórmula corrigida de Yates**

Neste caso tem-se somente 4 caselas, e alguns autores recomendam que se utilize a fórmula corrigida de Yates, χ^2, mesmo para os casos em que se tem N grande.

O assunto é controverso, recomendando-se utilizar a fórmula corrigida se N < 40 ou uma das caselas tiver $E_{ij} < 5$, e optar pela fórmula tradicional se N ≥ 40 e nenhuma casela tiver freqüência esperada menor que 5.

$$\chi^2_{cal} = \sum_{i=1}^{L}\sum_{j=1}^{C} \frac{(|O_{ij} - E_{ij}| - 0{,}5)^2}{E_{ij}}$$

(3) Para N pequeno

Preconiza-se a utilização do **teste exato de Fisher** nas seguintes situações:

(i) se **N** < 20, ou

(ii) se 20 < **N** < 40 **e** a menor freqüência esperada for menor que 5.

(4) Tabelas com mais de 2 colunas ou 2 linhas (*g.l.* > 1):

Nesta situação pode-se utilizar o teste χ^2 se o número de caselas com freqüência esperada inferior a 5 é menor que 20% do total de caselas e nenhuma freqüência esperada é zero.

Por exemplo, imagine que a tabela de contingência de pesquisa de pressão arterial em que a freqüência observada no grupo das mulheres tenha sido ligeiramente diferente:

	Mulheres	Homens	Total marginal
Hipertensos	45	47	92
Normotensos	252	148	400
Hipotensos	3	5	8
Total marginal	300	200	500

Determinando-se as freqüências esperadas teremos:

	Mulheres	Homens	Total marginal
Hipertensos	45	47	92
	55,2	*36,8*	
Normotensos	252	148	400
	240	*160*	
Hipotensos	3	5	8
	4,8	*3,2*	
Total marginal	300	200	500

Note que a tabela apresenta duas caselas (33,3% das caselas) com freqüências esperadas menores que 5.

Há alternativas para se contornar a situação e aplicar o teste χ^2:

(a) planejar de antemão a utilização de espaço amostral maior, o que requer algum conhecimento prévio da população estudada;

(b) combinar duas ou mais categorias para aumentar as freqüências esperadas, devendo observar se as combinações ainda permitem testar a hipótese de interesse.

Em nosso exemplo, podemos combinar as categorias de Normotensos e Hipotensos criando, assim, a categoria combinada de indivíduos não-Hipertensos.

	Mulheres	Homens	Total marginal
Hipertensos	45	47	92
Não-Hipertensos	255	153	408
Total marginal	300	200	500

Podemos testar se há diferença de distribuição de hipertensos e não-hipertensos entre mulheres e homens:

H_o: Não há diferença entre os grupos de mulheres e homens no que diz respeito à proporção de hipertensos e não-hipertensos.

Recalculando as freqüências esperadas:

	Mulheres	Homens	Total marginal
Hipertensos	45	47	92
	55,2	*36,8*	
Não-Hipertensos	255	153	408
	244,8	*163,2*	
Total marginal	300	200	500

pela fórmula tradicional,

$$\chi^2 = \frac{(O_{11}\cdot O_{22} - O_{12}\cdot O_{21})^2 . N}{O_{*1}\cdot O_{*2}\cdot O_{1*} O_{2*}}$$
$$= \frac{(45\cdot 153 - 47\cdot 255)^2 . 500}{300\cdot 200\cdot 92\cdot 408}$$
$$= 5{,}774$$

e pela fórmula corrigida de Yates,

$$\chi^2_{cc} = \frac{\left(\left|O_{11}\cdot O_{22} - O_{12}\cdot O_{21}\right| - \frac{1}{2}N\right)^2 . N}{O_{*1}\cdot O_{*2}\cdot O_{1*}\cdot O_{2*}}$$
$$= \frac{(|45\cdot 153 - 47\cdot 255| - 500/2)^2 \cdot 500}{300\cdot 200\cdot 92\cdot 408}$$
$$\cong 5{,}222$$

Da tabela de qui-quadrado, $\chi^2_c = (\text{g.l.} = 1,\ \alpha = 0{,}05) = 3{,}84$ e portanto rejeitamos a hipótese nula para ambos os valores, χ^2 e χ^2_{cc}.

Note que o valor obtido com a fórmula corrigida de Yates é sempre menor que o valor da estatística χ^2 tradicional.

Desta forma, ele tende a rejeitar a hipótese nula com maior freqüência.

6.2 Coeficiente de Contingência

Medida do grau de relacionamento, associação ou dependência das classificações em uma tabela de contingência.

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

Quanto maior o valor de C, maior o grau de associação, O máximo valor de C dependerá do número de linhas e colunas da tabela e pode variar de 0 (independência) a 1 (dependência total).

Exemplo: A tabela abaixo mostra os resultados de um ensaio com 154 pacientes que apresentavam dor abdominal. Administrou-se Brometo de Pinavérico (dois comprimidos/dia), (grupo tratamento). Ao grupo Controle foi administrado um placebo.

Tabela: Ingestão de brometo de pinavérico e alívio da dor abdominal.

	Permanência da dor abdominal		
Grupo	Sim	Não	**Total**
Tratamento	6	57	**63**
Controle	30	61	**91**
Total	**36**	**118**	**154**

Hipóteses a serem testadas:

H_0: $p_C = p_T$ versus H_A: $p_C \neq p_T$

A hipótese nula se refere a que os resultados dos grupos T e C sejam iguais. Dessa forma, do ponto de vista clínico, interessa a rejeição de H_0, que indicaria a eficiência terapêutica da droga.

$$E_{11} = \frac{63x36}{154} \cong 14{,}73;\ E_{12} = \frac{63x118}{154} \cong 48{,}27$$

$$E_{21} = \frac{91x36}{154} \cong 21{,}27;\ E_{22} = \frac{91x118}{154} \cong 69{,}73$$

	Permanência da dor abdominal		
Grupo	Sim	Não	**Total**
Tratamento	14,3	48,27	**63**

Controle	21,27	69,73	**91**
Total	**36**	**118**	**154**

$$\chi^2_{cal} = \sum_{i=1}^{L}\sum_{j=1}^{C}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} = \frac{(6-14{,}73)^2}{14{,}73} + \frac{(57-48{,}27)^2}{48{,}27} + \frac{(30-21{,}27)^2}{21{,}27} + \frac{(61-69{,}73)^2}{69{,}73}$$

$$= 11{,}4290$$

$$p-valor = P(\chi^2 \geq 11{,}4290) < 0{,}1\%$$

$$C = \sqrt{\frac{(11{,}4290)^2}{(11{,}4290)^2 + 154}} = 0{,}6774$$

A amostra fornece evidências para rejeição de H_0, ou seja, fica comprovado o efeito terapêutico do Brometo de Pinavérico no alívio das dores abdominais.

6.3 Risco Relativo

Tabela: Classificação Cruzada: Uso de aspirina X infarto do miocárdio

	Infarto		
Grupo	Sim	Não	Total
Placebo	189	10845	11034
Aspirina	104	10933	11037

Seja,

Π_1→ Probabilidade de um indivíduo do grupo placebo ter infarto;

Π_2→ Probabilidade de um indivíduo do grupo aspirina ter infarto.

Estimação pontual de Π_1 e Π_2

Estimador Pontual

Interesse: comparar Π_1 e Π_2

Risco Relativo: $R = \frac{\Pi_1}{\Pi_2}$

$\Pi_1 = \Pi_2 \Leftrightarrow \Pi_1 / \Pi_2 = 1$

Estimador do Risco Relativo

$$\hat{R} = \frac{\hat{\Pi}_1}{\hat{\Pi}_2} \Rightarrow \hat{\Pi}_1 = \hat{R}.\hat{\Pi}_2$$

O risco relativo para os dados da tabela uso da Aspirina e Infarto no miocárdio é

$$\hat{R} = \frac{0{,}0171}{0{,}0094} = 1{,}82$$
$$\hat{\Pi}_1 = 1{,}82.\hat{\Pi}_2$$

A probabilidade de infarto estimada para o grupo placebo é, 1,82 vezes a probabilidade de infarto estimada para o grupo Aspirina.

Estimação por Intervalo

Intervalo de Confiança para logR com coeficiente aproximado 1 – α

$$\log\hat{R} \pm z_{\alpha/2}\sqrt{\frac{1-\hat{\Pi}_1}{n_1\hat{\Pi}_1} + \frac{1-\hat{\Pi}_2}{n_2\hat{\Pi}_2}}$$

P(I ≤ logR ≤ S) → 1 – α

Onde I (inferior) e S (superior); P(I ≤ logR ≤ S) = P(e^I ≤ R ≤ e^S) (e^I; e^S) é um intervalo de confiança para R com coeficiente aproximado 1 – α

O i.c. com 95% de confiança para logR é

(0,3607; 0,8369)

O i.c. com 95% de confiança para R é

R→($e^{0,3607}$; $e^{0,8369}$) = (1,4343; 2,3092)

6.4 Teste Exato de Fisher

Exemplo: 2 drogas A e B: testando dois grupos de 10 indivíduos

A → 30% de melhora, B → 90% de melhora

	Droga		
	A	B	Total
Resp	3	9	12
Não resp	7	1	8
Total	10	10	20

Hipótese:

$$H_0: p_A = p_B$$

A probabilidade de obtermos uma dada configuração de tabela 2 x 2 no teste de Fisher é dada por:

$$p_1 = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{N!\,a!\,b!\,c!\,d!} = \frac{12!8!10!10!}{20!3!7!9!1!} = 0{,}009526$$

onde:

a	*B*	*a+b*
c	*D*	*c+d*
a+c	*b+d*	*N*

Para estimarmos **p** temos que encontrar a probabilidade de obtermos um resultado mais extremo $\Rightarrow$ reduzir de 1 a casela com menos dados:

$$\begin{matrix} 3 & 9 \\ & \\ 7 & 1 \end{matrix} \quad \rightarrow \quad \begin{matrix} 2 & 10 \\ & \\ 8 & 0 \end{matrix}$$

$$p_2 = \frac{12!8!10!10!}{20!2!10!8!0!} = 0{,}000357$$

$\mathbf{p}_1 + \mathbf{p}_2 = 0{,}0099$ é a probabilidade de obtermos o valor observado ou o valor mais extremo por acaso. Monocaudal $\rightarrow$ **p** = 0,0099; bicaudal $\rightarrow$ **p** = 2 x (0,0099) = 0,0198$\Rightarrow$ $\mathbf{H_0}$ é rejeitada!

7. Bioestatística não paramétrica

7.1 Teste do Sinal

O Teste do Sinal é o mais antigo de todos os testes não-paramétricos. Trata-se de um teste binomial com p = 1/2.

Os dados consistem de n pares de observações (X_i, Y_i), onde X_i representa uma situação "pré" e Y_i uma situação "pós", ou então, (X_i, Y_i) são pareados de acordo com suas afinidades e os objetivos da pesquisa.

Dentro de cada par (X_i, Y_i) a comparação é feita e o par é classificado como "+" (mais) se $X_i < Y_i$ e como "–" (menos) se $X_i > Y_i$ ou como "0" (empate) se $X_i = Y_i$.

Estatística de teste:

T = é número total de pares com "+" (são desconsiderados os casos de empate, de modo que: n = número de pares restantes).

Hipótese a ser testada:

$$H_0 : P(+) = P(-)$$

A região crítica, para um teste com nível de significância α, depende da hipótese alternativa:

Hipótese Alternativa Região Crítica

$H_1 : P(+) \neq P(-) \quad \Rightarrow \quad T \leq t \text{ ou } T \geq n - t$

$H_1 : P(+) > P(-) \quad \Rightarrow \quad T \geq n - t$

$H_1 : P(+) < P(-) \quad \Rightarrow \quad T \leq t$

Os valores de T é um indicativo para a hipótese alternativa mais provável. Para $n \leq 20$ utiliza-se a *tabela do teste do sinal* (t é encontrado nesta tabela). Caso contrário $n > 20$ usar aproximação normal, onde

$t = \frac{1}{2}\left(n + Z_{\alpha/2}\sqrt{n}\right)$ caso bilateral; $t = \frac{1}{2}\left(n + Z_{\alpha}\sqrt{n}\right)$ caso unilateral.

Exemplo: Vinte e quatro pacientes foram submetidos a uma dieta para emagrecimento, obtendo os seguintes resultados:

Paciente	Antes(X_i)	Depois(Y_i)	Paciente	Antes(X_i)	Depois(Y_i)
1	83,5	80,0	13	70,4	72,0
2	95,4	95,0	14	75,6	71,8
3	80,0	81,5	15	85,2	80,0
4	90,7	90,0	16	84,0	84,3
5	87,6	83,0	17	96,0	91,4
6	91,3	85,6	18	81,0	76,0
7	103,8	90,4	19	77,3	80,0
8	88,2	86,0	20	108,5	96,0
9	75,4	77,2	21	97,5	95,0
10	86,2	82,5	22	89,0	82,3
11	93,5	90,0	23	98,0	88,0
12	110,0	104,0	24	95,0	92,0

A dieta foi eficiente?

Solução:

"+" se $X_i < Y_i$; "–" se $X_i > Y_i$; "0" se $X_i = Y_i$;

Paciente	Antes(X)	Depois(Y)	Sinal	Paciente	Antes(X)	Depois(Y)	Sinal
1	83,5	80,0	–	13	70,4	72,0	+
2	95,4	95,0	–	14	75,6	71,8	–
3	80,0	81,5	+	15	85,2	80,0	–
4	90,7	90,0	–	16	84,0	84,3	+
5	87,6	83,0	–	17	96,0	91,4	–
6	91,3	85,6	–	18	81,0	76,0	–
7	103,8	90,4	–	19	77,3	80,0	+
8	88,2	86,0	–	20	108,5	96,0	–
9	75,4	77,2	+	21	97,5	95,0	–
10	86,2	82,5	–	22	89,0	82,3	–
11	93,5	90,0	–	23	98,0	88,0	–
12	110,0	104,0	–	24	95,0	92,0	–

Hipóteses a serem testadas: H_0: P(+) ≥ P(–) vs H_1: P(+) < P(–)

n = 24; p = 1/2; α = 5%; número total de "–" = 19

Estatística de teste: T = número total de "+" = 5

Como n > 20 usa-se aproximação normal $Z_\alpha = 1{,}64 \Rightarrow t = \frac{1}{2}\left(24 + 1{,}64\sqrt{24}\right) = 16{,}01$

Regra de decisão: Rejeita H_0 se $T \leq t$; Como $5 \leq 16{,}01 \Rightarrow$ Portanto os dados nos dão evidência para rejeição de H_0, ou seja, no nível de significância de 5% há diferença entre os pesos antes e depois com reduzidos "+". Logo a dieta é eficiente.

7.2 Teste U de Mann-Whitney ⇔ *t* "não-pareado":

Exemplo: Idade de primigestas de duas localidades:

Loc A	Loc B
16	15
26	25
23	17
19	40
45	22
30	27
23	21
23	20
29	18
24	16
35	14
32	31

Combinam-se todos os dados de ambas as localidades em ordem crescente.

Dado	Posto	Loc	Dado	Posto	Loc
14	1	B	23	12	A
15	2	B	24	14	A
16	3,5	A	25	15	B
16	3,5	B	26	16	A
17	5	B	27	17	B
18	6	B	29	18	A
19	7	A	30	19	A
20	8	B	31	20	B
21	9	B	32	21	A
22	10	B	35	22	A
23	12	A	40	23	B
23	12	A	45	24	A

Hipóteses a serem testadas:

H_0: $\mu_A = \mu_B$

H_1: $\mu_A \neq \mu_B$

Estatística de teste:

$$U_A = N_A N_B + \frac{N_B(N_B+1)}{2} - R_B$$

$$U_B = N_A N_B + \frac{N_A(N_A+1)}{2} - R_A$$

onde

N_A e N_B são os tamanhos da amostras A e B

R_A e R_B são as somas dos postos de A e B

R_A = 3,5+7+12+12+12+14+16+18+19+21+22+24 = 180,5

R_B = 1+2+3,5+5+6+8+9+10+15+17+20+23 = 119,5

$$U_A = 12x12 + \frac{12(12+1)}{2} - 119{,}5 = 102{,}5$$

$$U_B = 12x12 + \frac{12(12+1)}{2} - 180{,}5 = 41{,}5$$

Estatística U: o menor entre U_A e U_B = 41,5

$$\mu_U = \frac{N_A N_B}{2} \qquad \sigma_U = \sqrt{\frac{N_A N_B (N_A + N_B + 1)}{12}}$$

$$\mu_U = 72 \qquad \sigma_U = 17{,}32$$

$$U \rightarrow Z$$

$$Z_U = \frac{41{,}5 - 72}{17{,}32} = -1{,}76$$

H_0 não é rejeitada (considerando α = 5% e o teste bicaudal com Z_c = 1,96).

Mann-Whitney para Pequenas Amostras

A estatística Z_u tem distribuição aproximadamente normal apenas quando $N_A, N_B \geq$ 10. Caso isto não se verifique o procedimento para o teste U é diferente:

- Tome as somas de postos;
- Consulte a tabela de Mann-Whitney para pequenas amostras:

- Localize nas colunas desta tabela o valor de α e o tamanho da amostra de tamanho inferior a 10;
- Localize nas linhas o tamanho da outra amostra;
- Rejeite H_0 se os valores das somas de postos estiverem fora do intervalo fornecido na tabela.

Calculando *p* a partir da soma de postos

Quando ambas as amostras têm n_1, $n_2 > 10$ há uma forma alternativa de calcular o valor de p:

Escolha uma das duas somas de postos (a escolha é arbitrária);

- Caso não existam empates de postos, calcule:

$$T = \frac{\left[\left|R_1 - \frac{n_1(n_1+n_2+1)}{2}\right| - \frac{1}{2}\right]}{\sqrt{\left(\frac{n_1 n_2}{12}\right)(n_1+n_2+1)}}$$

onde

R_1 é a soma de postos

n_1,n_2 corresponde ao tamanho das amostras.

- Caso existam empates de postos, calcule:

$$T = \frac{\left[\left|R_1 - \frac{n_1(n_1+n_2+1)}{2}\right| - \frac{1}{2}\right]}{\sqrt{\left(\frac{n_1 n_2}{12}\right)\left[n_1+n_2+1-\frac{\sum_{i=1}^{g} t_i(t_i^2-1)}{(n_1+n_2)(n_1+n_2-1)}\right]}}$$

onde

g é o número de grupos empatados;

$\boldsymbol{t_i}$ é número de observações com o mesmo valor no ***i***-ésimo grupo.

- A hipótese nula:

$$H_0: \mu_1 = \mu_2$$

é rejeitada se:

$$T > Z_{1-\alpha/2}$$

- e **p** é dado por:

$$p = 2 \times [1 - \Phi(T)]$$

onde Φ(T) é a área sob a curva normal acumulada definida por T (veja a tabela Z).

7.3 Teste de Wilcoxon ⇔ *t* pareado:

Exemplo: Células tumorais, 2 lotes de cada órgão (% de resposta a um tipo de tratamento):

Tipo de célula	Amostra A	Amostra B
Pulmão	90	85
Fígado	94	90
Laringe	78	80
Intestino	79	70
Cérebro	84	84
Pele	75	65
Linfonodo	74	63
Pâncreas	99	89
Ossos	72	73

Calcula-se a diferença entre cada par de amostras para estabelecer a ordem dos postos:

Amostra A	Amostra B	≠ s	Postos (R)
90	85	5	4
94	90	4	3
78	80	-2	2
79	70	9	5
84	84	0	?
75	65	10	6,5
74	63	11	8
99	89	10	6,5
72	73	-1	1

$$H_0 : \mu_D = 0$$

Σ**R+** soma associada com postos positivos das ≠s

Σ**R-** soma associada com postos negativos das ≠s

Σ**R+** = 4 + 3 + 5 + 6,5 + 8 + 6,5 = 33

Σ**R-** = 2 + 1 = 3

Estatística V:

O menor entre ΣR**+** e ΣR-

então V = 3

calacula-se: $\mu_V = \frac{N(N+1)}{4}$ e $\sigma_V = \sqrt{\frac{N(N+1)(2N+1)}{24}}$

onde N = n° de pares testados (posto zero não é testado!).

$$\mu_V = \frac{8(8+1)}{4} = 18 \quad \sigma_V = \sqrt{\frac{8(8+1)(2.8+1)}{24}} = 7{,}14$$
$$V \to Z$$
$$Z_V = \frac{V - \mu_V}{\sigma_V} = \frac{3-18}{7{,}14} = -2{,}10$$

H_0 seria rejeitada (considerando α = 5% e o teste bicaudal com z_c = 1,96)

Nota: A estatística de teste Z_v tem distribuição aproximadamente normal apenas quando há mais que 15 pares de valores com diferença não nula e a variável é contínua (este exemplo tem apenas 8 pares): o procedimento para o teste pode ser um pouco diferente, veja abaixo.

Wilcoxon para Pequenas Amostras

A estatística de teste Z_v tem distribuição aproximamente normal apenas quando há mais que 15 pares de valores com diferença não nula e a variável é contínua. Caso isto não se verifique, recomenda-se que a decisão sobre

$$H_0 : \mu_D = 0$$

seja tomada utilizando-se uma tabela para Wilcoxon para pequenas amostras:

H_0 é rejeitada se ΣR- ou ΣR**+** está(ão) fora do intervalo dado pelos valores críticos inferior e superior.

7.4 Teste de Kruskal-Wallis

Suponha que temos 3 amostras provenientes de 3 populações (desconhecidas). Queremos testar se as populações são as mesmas.

Pressuposições

- As observações são todas independentes;
- Dentro de uma dada amostra, todas as observações são provenientes da mesma população;
- As k populações são aproximadamente da mesma forma e contínuas.

Hipóteses a serem testadas:

H_0: as amostras são provenientes da mesma população;

H_1: pelo menos duas amostras diferem entre si;

Estatística de teste:

- Caso não existam empates de postos, calcule:

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1)$$

onde

R_i é a soma de postos na amostra i

n_i corresponde ao tamanho da amostra i

$N = \sum n_i$

- Caso existam empates de postos, calcule:

$$H = \frac{\frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1)}{1 - \frac{\sum_{i=1}^{g} T_i}{N^3 - N}}$$

onde

g é o número de grupos empatados;

$T_i = t_i^3 - t_i$;

$\mathbf{t_i}$ é número de observações com o mesmo valor no i-ésimo grupo.

Rejeitamos H_0 ao nível α de significância se $H \geq h$, onde $P(H \geq h) = \alpha$

Se H_0 for verdadeira, e n_i for grande ($n_i \geq 6$), $\forall$ i = 1, ..., k; k > 3, então H tem distribuição aproximadamente χ^2 com k – 1 graus de liberdade.

Exemplo: Vinte pacientes com mesmo tipo de câncer e no mesmo estágio da doença são submetidos a 4 tipo de tratamentos. Eles são separados aleatoriamente em 4

grupos de 5, a cada grupo é aplicado um dos tratamentos e a eficiência é medida pelo tempo de sobrevida (em anos).

Tratamento	Sobrevida (em anos)				
A	14,2	10,6	9,4	5,6	2,4
B	12,8	12,3	6,4	6,1	1,6
C	11,5	10,1	5,1	5,0	4,8
D	14,9	13,7	8,5	7,7	5,9

Existe diferença entre os tratamentos?

Solução:

Hipóteses a serem testadas:

H_0: $T_A = T_B = T_C = T_D$;

H_1: pelo menos dois tratamentos diferem entre si;

Combinam-se todos os dados dos 4 tratamentos e soma os seus postos:

Tratamento	Sobrevida (em anos)					Soma dos Postos
A	14,2	10,6	9,4	5,6	2,4	
Posto	**19**	**14**	**12**	**6**	**2**	**$R_A = 53$**
B	12,8	12,3	6,4	6,1	1,6	
Posto	**17**	**16**	**9**	**8**	**1**	**$R_B = 51$**
C	11,5	10,1	5,1	5,0	4,8	
Posto	**15**	**13**	**5**	**4**	**3**	**$R_C = 40$**
D	14,9	13,7	8,5	7,7	5,9	
Posto	**20**	**18**	**11**	**10**	**7**	**$R_D = 66$**

Temos k = 4 e $n_1 = n_2 = n_3 = n_4 = 5$;

Estatística de teste:

$$H = \frac{12}{20(20+1)}\left[\frac{53^2}{5}+\frac{51^2}{5}+\frac{40^2}{5}+\frac{66^2}{5}\right]-3(20+1)$$

$$H = \frac{12}{420}\left[2273{,}2\right]-63 \Rightarrow H = 1{,}95$$

Como k > 3 então usamos a aproximação pela distribuição Qui-quadrado.

Graus de liberdade = 4 – 1 = 3

P-valor $\Rightarrow$ P($\chi^2_3 \geq 1{,}95$) = 0,5828

Os dados evidenciam a não rejeição de H_0, ou seja, não há diferenças entre os tratamentos.

Teste Qui-quadrado

No período de 4 semanas iniciando no dia **1º de maio**, o Pronto Socorro Médico do Hospital das Clínicas recebeu 225 pacientes em crise hipertensiva distribuídos segundo os dias de atendimento conforme a tabela abaixo:

Dias da semana	Dom	Seg	Ter	Qua	Qui	Sex	Sab
Pac. Atendidos	38	25	30	28	33	31	40

Pergunta-se:

A distribuição de pacientes ao longo da semana é homogênea?

O coeficiente qui-quadrado, χ^2, mede a "distância" entre os valores observados em uma amostra e os valores hipotéticos esperados (a diferença é relativa ao valor esperado).

Em geral, se O_1, O_2, ..., O_n são valores observados na amostra, de cada um dos **n** atributos, e E_1, E_2, ..., E_n são os respectivos valores esperados, definimos χ^2 como

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Para grandes amostras, χ^2 aproxima-se da distribuição qui-quadrado com (n – 1) graus de liberdade. Rejeitamos a hipótese nula para valores elevados de χ^2.

Voltando ao exemplo, queremos saber se:

H_0: A freqüência observada obedece à distribuição $p_1 = p_2 = ... = p_n$;

H_1: Os pacientes compareceram preferencialmente em determinados dias da semana.

Categorias	p_i	o_i	e_i	$o_i - e_i$	$(o_i - e_i)^2/e_i$
Dom	1/7	38	32,1	5,86	1,067
Seg	1/7	25	32,1	-7,14	1,587
Ter	1/7	30	32,1	-2,14	0,142
Qua	1/7	28	32,1	-4,14	0,534
Qui	1/7	33	32,1	0,86	0,022
Sex	1/7	31	32,1	-1,14	0,041
Sab	1/7	40	32,1	7,86	1,921
Total	1	225	225	0	5,316

Portanto,

$$\chi^2 \cong 5{,}316$$

O número de graus de liberdade é dado por:

$$\text{g.l.} = n - 1 = 7 - 1 = 6$$

ou seja, o número de categorias menos 1.

Da tabela qui-quadrado, para g.l.= 6 e nível de significância 5%, temos:

$$\chi^2 \text{ (g.l.} = 6, \alpha = 0{,}05) = 12{,}6$$

O p-valor é dado por $P(\chi^2 \geq 5{,}316) \cong 0{,}6$ ou 60%

Desta forma, não podemos rejeitar a hipótese nula de que a distribuição de pacientes é homogênea ao longo da semana.

Teste χ^2 para grandes amostras

Uma importante aplicação do teste χ^2 ocorre quando se quer estudar as relações entre duas ou mais variáveis de classificação. A representação das freqüências observadas, nesse caso, pode ser feita por meio de uma tabela de contingência.

Hipóteses a serem testadas:

Ho: As variáveis são independentes (não estão associadas)

H_1: As variáveis não são independentes (estão associadas)

Estatística de teste:

$$\chi^2_{cal} = \sum_{i=1}^{L}\sum_{j=1}^{C} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}, \text{ onde } E_{ij} = \frac{(\text{soma da linha i})(\text{soma da coluna j})}{\text{total de observações}}$$

O número de graus de liberdade é dado por:

(L - 1) (C - 1), onde L é o número de linhas e C o número de colunas da tabela de contingência.

Regra de descisão:

Fixando um nível α rejeita-se H_0

se $\chi^2_{cal} > \chi^2_{\alpha;\,(L-1)(C-1)}$;

O p-valor é dado a partir da $P(\chi^2 \geq \chi^2_{cal})$.

Exemplo: Efeito de uma droga comparada com placebo (tabela de contingência 2 x 2). Uma droga nova foi testada em 56 pacientes, outros 51 receberam placebo. 48 dos que receberam a droga melhoraram e 38 dos que receberam placebo melhoraram.

Estímulo	Efeito		Total marginal
	positivo	negativo	
Droga	48	8	56
Placebo	38	13	51
Total marginal	86	21	107

Pergunta: A resposta foi contingente ao estímulo?

Note que esta tabela, tomando-se a percentagem de efeito obtido com droga e placebo (i.e., as proporções tomadas em relação aos totais marginais das linhas) resulta em:

Estímulo	Efeito	
	positivo	negativo
Droga	0,875	0,143
Placebo	0,745	0,255

Esta tabela, obtida a partir dos valores observados acima, tomando-se a percentagem em relação aos **valores esperados**:

Estímulo	Efeito		Total marginal
	positivo	negativo	
Droga	**45,01**	**10,99**	56
Placebo	**40,99**	**10,01**	51
Total marginal	86	21	107

resulta em:

Estímulo	Efeito	
	positivo	negativo
Droga	0,804	0,196
Placebo	0,804	0,196

H_0: Não há diferença entre os grupos que receberam a droga e placebo em relação à melhora dos pacientes.

Solução:

Neste exemplo

$$E_{11} = \frac{86x56}{107} = 45{,}01 \quad E_{21} = \frac{86x51}{107} = 40{,}99 \quad E_{12} = \frac{21x56}{107} = 10{,}99 \quad E_{22} = \frac{21x51}{107} = 10{,}01$$

Estímulo	positiva	negativa	Total marginal
Droga	48 45,01	8 10,99	56
Placebo	38 40,99	13 10,01	51
Total marginal	86	21	107

Determinam-se as freqüências esperadas, (E_{ij}),

Sendo que $\chi^2_{cal} = \sum_{i=1}^{L}\sum_{j=1}^{C} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$, temos

$$\chi^2_{cal} = \frac{(48-45{,}01)^2}{45{,}01} + \frac{(38-40{,}99)^2}{40{,}99} + \frac{(8-10{,}99)^2}{10{,}99} + \frac{(13-10{,}01)^2}{10{,}01} \cong 2{,}123$$

Pela tabela de distribuição de qui-quadrado, obtemos:

$$\chi^2 \text{ (g.l. = 1, } \alpha = 0{,}05) = 3{,}84$$

O p-valor é $P(\chi^2 \geq 2{,}123) \cong 0{,}15$ ou 15%.

Como o valor obtido é menor que o valor crítico não podemos rejeitar a hipótese nula para erro $\alpha = 5\%$. Logo, não podemos dizer que os efeitos da droga e do placebo sejam significativamente diferentes.

Restrições ao uso do Teste Qui-quadrado

Ao utilizar-se o teste qui-quadrado supõe-se que o tamanho das amostras seja "grande".

"Em situações práticas, o valor de χ^2 calculado é aproximado porque se utiliza amostras de tamanho finito e o valor da freqüência observada pode assumir somente números inteiros. Por exemplo, você jamais vai encontrar 2,87 casos de pacientes doentes."

Portanto é necessário observar as seguintes restrições:

(1) Regra geral, o teste χ^2 pode ser usado se o número de observações em cada casela da tabela for maior ou igual a 5 e a menor freqüência esperada for maior ou igual a 5 ($E_{ij} \geq 5$, $O_{ij} \geq 5$).

(2) Tabela 2x2 (g.l. *= 1*): **Fórmula corrigida de Yates**

Neste caso tem-se somente 4 caselas, e alguns autores recomendam que se utilize a fórmula corrigida de Yates, χ^2, mesmo para os casos em que se tem N grande.

O assunto é controverso, recomendando-se utilizar a fórmula corrigida se N < 40 ou uma das caselas tiver $E_{ij} < 5$, e optar pela fórmula tradicional se N ≥ 40 e nenhuma casela tiver freqüência esperada menor que 5.

$$\chi^2_{cal} = \sum_{i=1}^{L}\sum_{j=1}^{C} \frac{\left(\left|O_{ij} - E_{ij}\right| - 0{,}5\right)^2}{E_{ij}}$$

(3) Para N pequeno

Preconiza-se a utilização do **teste exato de Fisher** nas seguintes situações:

(i) se **N** < 20, ou

(ii) se 20 < **N** < 40 **e** a menor freqüência esperada for menor que 5.

(4) Tabelas com mais de 2 colunas ou 2 linhas (*g.l.* > 1):

Nesta situação pode-se utilizar o teste χ^2 se o número de caselas com freqüência esperada inferior a 5 é menor que 20% do total de caselas e nenhuma freqüência esperada é zero.

Por exemplo, imagine que a tabela de contingência de pesquisa de pressão arterial em que a freqüência observada no grupo das mulheres tenha sido ligeiramente diferente:

	Mulheres	Homens	Total marginal
Hipertensos	45	47	92
Normotensos	252	148	400
Hipotensos	3	5	8
Total marginal	300	200	500

Determinando-se as freqüências esperadas teremos:

	Mulheres	Homens	Total marginal
Hipertensos	45	47	92
	55,2	*36,8*	
Normotensos	252	148	400
	240	*160*	
Hipotensos	3	5	8
	4,8	*3,2*	
Total marginal	300	200	500

Note que a tabela apresenta duas caselas (33,3% das caselas) com freqüências esperadas menores que 5.

Há alternativas para se contornar a situação e aplicar o teste χ^2:

(a) planejar de antemão a utilização de espaço amostral maior, o que requer algum conhecimento prévio da população estudada;

(b) combinar duas ou mais categorias para aumentar as freqüências esperadas, devendo observar se as combinações ainda permitem testar a hipótese de interesse.

Em nosso exemplo, podemos combinar as categorias de Normotensos e Hipotensos criando, assim, a categoria combinada de indivíduos não-Hipertensos.

	Mulheres	Homens	Total marginal
Hipertensos	45	47	92
Não-Hipertensos	255	153	408
Total marginal	300	200	500

Podemos testar se há diferença de distribuição de hipertensos e não-hipertensos entre mulheres e homens:

Ho: Não há diferença entre os grupos de mulheres e homens no que diz respeito à proporção de hipertensos e não-hipertensos.

Recalculando as freqüências esperadas:

	Mulheres	Homens	Total marginal
Hipertensos	45	47	92
	55,2	*36,8*	
Não-Hipertensos	255	153	408
	244,8	*163,2*	
Total marginal	300	200	500

pela fórmula tradicional,

$$\chi^2 = \frac{(O_{11}O_{22} - O_{12}O_{21})^2 N}{O_{*1}O_{*2}O_{1*}O_{2*}} = \frac{(45 \cdot 153 - 47 \cdot 255)^2 500}{300 \cdot 200 \cdot 92 \cdot 408} = 5{,}774$$

e pela fórmula corrigida de Yates,

$$\chi^2_{cc} = \frac{\left(|O_{11}O_{22} - O_{12}O_{21}| - \frac{1}{2}N\right)^2 \cdot N}{O_{*1}O_{*2}O_{1*}O_{2*}} = \frac{(|45 \cdot 153 - 47 \cdot 255| - 500/2)^2 \cdot 500}{300 \cdot 200 \cdot 92 \cdot 408}$$

= 5,22

Da tabela de qui-quadrado, $\chi^2_c = (\text{g.l.} = 1,\ \alpha = 0{,}05) = 3{,}84$ e portanto rejeitamos a hipótese nula para ambos os valores, χ^2 e χ^2_{cc}.

Note que o valor obtido com a fórmula corrigida de Yates é sempre menor que o valor da estatística χ^2 tradicional.

Desta forma, ele tende a rejeitar a hipótese nula com maior freqüência.

Teste Exato de Fisher

Exemplo: 2 drogas A e B: testando dois grupos de 10 indivíduos

$A \rightarrow$ 30% de melhora, $B \rightarrow$ 90% de melhora

	Droga		
	A	B	Total
Resp	3	9	12
Não resp	7	1	8
Total	10	10	20

Hipótese:

$$H_0: p_A = p_B$$

A probabilidade de obtermos uma dada configuração de tabela 2 x 2 no teste de Fisher é dada por:

$$p_1 = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{N!\,a!\,b!\,c!\,d!} = \frac{12!8!10!10!}{20!3!7!9!1!} = 0{,}009526$$

onde:

a	*b*	*a+b*
c	*d*	*c+d*
a+c	*b+d*	*N*

Para estimarmos **p** temos que encontrar a probabilidade de obtermos um resultado mais extremo $\Rightarrow$ reduzir de 1 a casela com menos dados:

$$\begin{matrix} 3 & 9 \\ 7 & 1 \end{matrix} \rightarrow \begin{matrix} 2 & 10 \\ 8 & 0 \end{matrix}$$

$$p_2 = \frac{12!8!10!10!}{20!2!10!8!0!} = 0{,}000357$$

$\mathbf{p}_1 + \mathbf{p}_2 = 0{,}0099$ é a probabilidade de obtermos o valor observado ou o valor mais extremo por acaso.

monocaudal $\rightarrow$ **p** = 0,0099

bicaudal $\rightarrow$ **p** = 2 x (0,0099) = 0,0198

$\Rightarrow$ $\mathbf{H_0}$ é rejeitada!

8. Teorema de Bayes em Bioestatística

Eventos Independentes

Diz-se que dois eventos são independentes quando a ocorrência de um deles não influencia a probabilidade do outro ocorrer (não confunda com eventos disjuntos).

$$P(E_1 \text{ e } E_2) = P(E_1).\ P(E_2)$$

sendo **p** a probabilidade, e E_1 e E_2 dois eventos independentes, lê-se: a probabilidade de ocorrência conjunta dos dois eventos é o produto das probabilidades de ocorrência individuais.

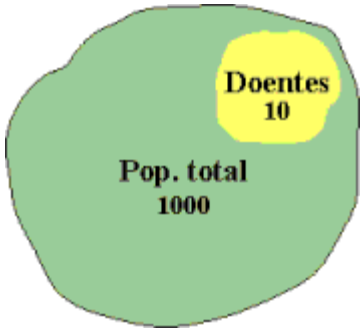
Eventos Não Independentes

$$P(E_1 \text{ e } E_2) = P(E_2).\ P(E_1 | E_2)$$

sendo **p** a probabilidade, e E_1 e E_2 dois eventos não independentes, lê-se: a probabilidade de ocorrência conjunta dos dois eventos é a probabilidade de ocorrência de um evento E_2 multiplicada pela probabilidade do outro evento E_1 dado que o E_2 ocorreu.

Prevalência de um evento ≡ *Probabilidade de Ocorrência*

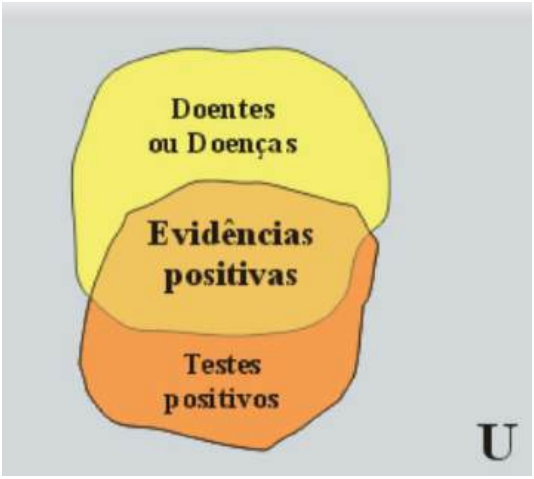
Exemplo: Prevalência de casos de tuberculose em uma comunidade



$$\text{Prev(Tb)} = \frac{\text{número de casos}}{\text{total da população}}$$

$$\text{Prev(Tb)} = \frac{10}{1000} = 0{,}01 = 1\%$$

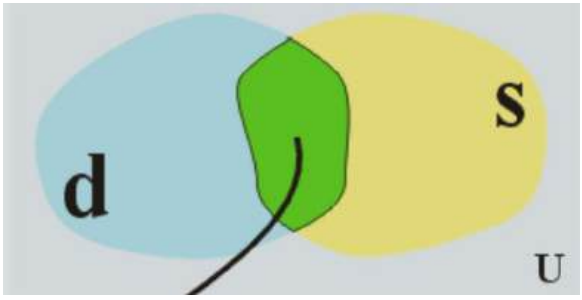
Evidência em Medicina ≡ *Informação para conduta diagnóstica, prognostica ou terapêutica.*



Diagnóstico ≡ *Processo Classificatório.*

Conjunto de Doenças ↔ Conjunto de Evidências { Sintomas / Sinais / Exames Auxiliares }

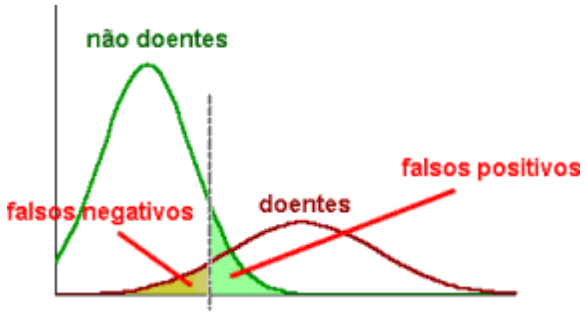
d s



$$P(d|s) = \frac{|(d \cap s)|}{|s|} \Rightarrow P(d|s) = \frac{P(d).P(s|d)}{P(s)}$$

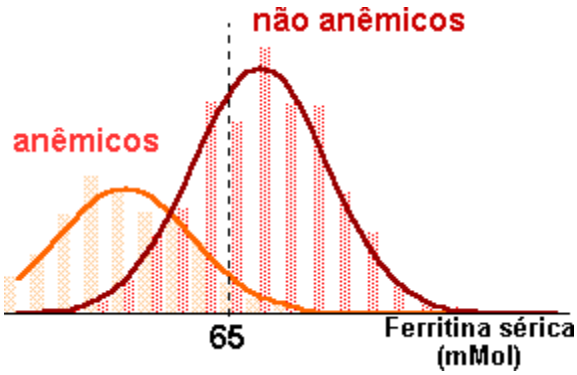
	com doença (d)	sem doença ($\bar{d}$)	
Teste + (s)	A	B falsos positivos	(A+B) **total de Positivos**
Teste - ($\bar{s}$)	C falsos negativos	D	(C+D) **total de**

			Negativos
	(A+C) **total dos *verdadeiramente* doentes**	(B+D) **total dos *verdadeiramente* não doentes**	A+B+C+D **total**
	("*Gold standard*")		



Exemplo: Anemia ferropriva

	com anemia ferropriva (d)	sem anemia ferropriva ($\bar{d}$)	Total
Ferritina sérica + (s)	731	270	1001
Ferritina sérica - ($\bar{s}$)	78	1500	1578
Total	809	1770	2579



Ferritina sérica ≥ 65 ⇒ **–**

Ferritina sérica ≤ 65 ⇒ **+**

"*Cut off*" (valor de corte) = 65 mMol

Sensibilidade

Capacidade de identificar os positivos (doentes) entre os verdadeiramente doentes

$$\text{Sensibilidade} = \frac{A}{A+C} = P(s|d)$$

No exemplo da anemia ferropriva:

Sensibilidade estimada do teste da Ferritina Sérica $= \frac{731}{809} = 0{,}904 \text{ ou } 90{,}4\%$

78 indivíduos (9,6%) de falsos negativos

Especificidade

Capacidade de identificar os negativos (não-doentes) entre os verdadeiramente negativos

$$\text{especificidade} = \frac{D}{B+D} = P(\bar{s}|\bar{d})$$

No exemplo da anemia ferropriva:

Especificidade do teste da Ferritina Sérica $= \frac{1500}{1770} = 0{,}847 \text{ ou } 84{,}7\%$

270 (15,3%) de falsos positivos

Valor Preditivo Positivo – *VPP*

Probabilidade de o indivíduo ser portador da doença ***dado*** que o teste é positivo.

$$VPP = P(d|s) = \frac{(\text{sensibilidade do teste}) x (\text{prevalência da doença})}{(\text{prevalência de positivos ao teste})}$$

No exemplo da anemia ferropriva:

$$\text{VPP do teste Ferritina Sérica} = \frac{0{,}904 x \frac{809}{2579}}{\frac{1001}{2579}} = 0{,}732 \text{ ou } 73{,}2\%$$

Um indivíduo desta população positiva ao teste da Ferritina Sérica tem 73,2% de probabilidade de ser portador da anemia ferropriva.

$$P(d|s) = \frac{P(s|d)P(d)}{P(d)P(s|d) + P(\bar{d})P(s|\bar{d})} = \frac{A}{A+B}$$

Como calcular a prevalência de positivos P(s)

A prevalência de positivos é o denominador do **VPP**.

$$VPP \equiv (d|s) = \frac{P(d)P(s|d)}{P(s)} = \frac{P(s|d)P(d)}{P(d)P(s|d) + P(\bar{d})P(s|\bar{d})}$$

Para entender como chegar ao denominador destacado em cinza, abaixo:

$$p(d|s)=\frac{p(s|d)p(d)}{p(d)p(s|d)+p(\bar{d})p(s|\bar{d})}=\frac{A}{A+B}$$

é necessário utilizar os conceitos de intersecção de conjuntos e de probabilidade condicional.

Considere que entre os indivíduos positivos ao teste, há indivíduos doentes e não doentes:

$$[\mathrm{s}]=[\mathrm{s}\cap\mathrm{d}]\cup[\mathrm{s}\cap\bar{\mathrm{d}}]\Rightarrow \mathrm{P}(\mathrm{s})=\mathrm{P}(\mathrm{s}\cap\mathrm{d})+\mathrm{P}(\mathrm{s}\cap\bar{\mathrm{d}})$$

onde a probabilidade da união é igual à somas das probabilidades dos eventos $[\mathrm{s}\cap\mathrm{d}]$ e $[\mathrm{s}\cap\bar{\mathrm{d}}]$ porque estes são disjuntos.

Usamos probabilidade condicional para escrever:

$$\mathrm{P}(\mathrm{s}\cap\mathrm{d})=\mathrm{P}(\mathrm{s}|\mathrm{d})\mathrm{P}(\mathrm{d})$$

$$\mathrm{P}(\mathrm{s}\cap\bar{\mathrm{d}})=\mathrm{P}(\mathrm{s}|\bar{\mathrm{d}})\mathrm{P}(\bar{\mathrm{d}})$$

o que nos dá o denominador. Note que

P(d) ...prevalência da doença

$\mathrm{P}(\bar{\mathrm{d}})$ ...prevalência de sadios (complemento do anterior)

P(s | d) ...sensibilidade

$\mathrm{P}(\bar{\mathrm{s}}\mid\bar{\mathrm{d}})$ ...especificidade, então:

$\mathrm{P}(\mathrm{s}\mid\bar{\mathrm{d}})=1-\mathrm{P}(\bar{\mathrm{s}}\mid\bar{\mathrm{d}})$ é seu complemento.

Similarmente, para **VPN**

$$p(\bar{d}|\bar{s})=\frac{p(\bar{s}|\bar{d})\times p(\bar{d})}{p(\bar{d})p(\bar{s}|\bar{d})+p(d)p(\bar{s}|d)}$$

o denominador é a prevalência de negativos ao teste, e uma demonstração semelhante é possível.

Valor Preditivo Negativo – *VPN*

Probabilidade de o indivíduo não ser portador da doença dado que o teste é negativo.

$$\mathrm{VPN}=\mathrm{P}(\bar{\mathrm{d}}|\mathrm{s})=\frac{(\mathrm{especificidade\ do\ teste})\mathrm{x}(\mathrm{prevalência\ de\ não\ doentes})}{(\mathrm{prevalência\ de\ negativos\ ao\ teste})}$$

No exemplo da anemia ferropriva:

$$\text{VPN do teste Ferritina Sérica } = \frac{\frac{1770}{2579} \text{x} 0{,}547}{\frac{1578}{2579}} = 0{,}949 \text{ ou } 94{,}9\%$$

Um indivíduo desta população negativo ao teste da Ferritina Sérica tem 94,9% de probabilidade de não ser portador da anemia ferropriva.

$$P(\bar{d}|\bar{s}) = \frac{P(\bar{s}|\bar{d}).P(\bar{d})}{P(\bar{s}|\bar{d}).P(\bar{d}) + P(\bar{s}|d).P(d)} = \frac{D}{C+D} \qquad \text{ou}$$

$$VPN = \frac{(1-P(d)) \text{ x especificidade}}{(1-P(d)) \text{ x especificidade} + (1-\text{sensibilidade}) \text{ x } P(d)}$$

a demonstração relacionada a este denominador é similar à feita para o VPP acima.

Razão de Verossimilhança Positiva

Probabilidade de que dado resultado de teste fosse esperado em um paciente portador da doença, comparado com a probabilidade de que o mesmo resultado fosse esperado em um paciente sem a doença.

$$RVP = \frac{\text{sensibilidade}}{1-\text{especificidade}}$$

No exemplo da anemia ferropriva:

$$RVP = \frac{0{,}9}{1-0{,}85} = 6{,}0$$

Quanto melhor o teste, maior a **RVP.**

Razão de Verossimilhança Negativa

Probabilidade de que dado resultado de teste fosse esperado em um paciente não portador da doença, comparado com a probabilidade de que o mesmo resultado fosse esperado em um paciente com a doença.

$$RVN = \frac{1-\text{sensibilidade}}{\text{especificidade}}$$

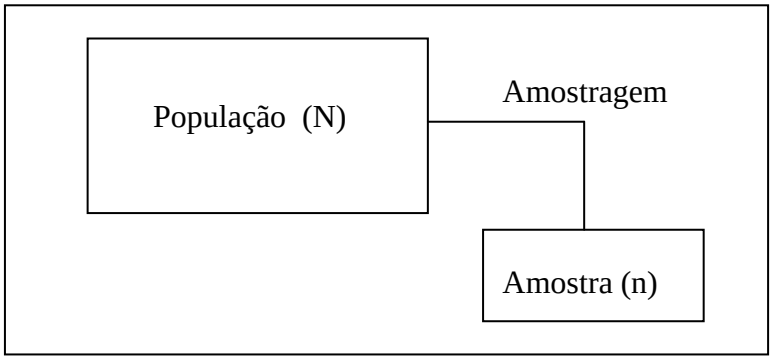
No exemplo da anemia ferropriva:

$$RVN = \frac{1-0,9}{0,85} = 0,1$$

Quanto melhor o teste, menor a **RVN.**

9. Amostragem

É o processo de retirada de informações dos "n" elementos amostrais, na qual deve seguir um método adequado (tipos de amostragem).



Plano de Amostragem

1º) Definir os Objetivos da Pesquisa

2º) População a ser Amostrada

- Parâmetros a ser Estimados (Objetivos)

3º) Definição da Unidade Amostral

- Seleção dos Elementos que farão parte da amostra

4º) Forma de seleção dos elementos da população

- Tipo de Amostragem

a) Estratificada

b) Aleatória Simples

c) Sistemática

d) por Conglomerados

5º) Tamanho da Amostra

Exemplo: Moradores de uma Cidade (população alvo)

- Objetivo: *Tipo de Residência*

  Própria.....um piso

  Alugada.....dois pisos

Emprestada.....três ou mais pisos

- Unidade Amostral: *Domicílios* (*residências*)
- Elementos da População: *Família por domicílio*
- Tipo de Amostragem:

  aleatória simples

  estratificada

  sistemática

Amostragem "COM" e "SEM" reposição

Seja "N" o número de elementos de uma população, e seja "n" o número de elementos de uma amostra, então:

Se o processo de retirada dos elementos for COM reposição (pop. infinita (f $\leq$5%), onde f é fator de correção), o número de amostras possíveis será:

nº de amostras = N^n

Se o processo de retirada de elementos for SEM reposição (pop. finita (f > 5%) onde f é fator de correção), o

número de amostras possíveis será:

nº de amostras = $C_{N,n} = \frac{N!}{n!(N-n)!}$

Ex.: Supondo N = 8 e n = 4

com reposição: no de amostras = $N^n = 8^4 = 4096$

sem reposição: no de amostras = $C_{N,n} = \frac{N!}{n!(N-n)!} = C_{8,4} = \frac{8!}{4!(8-4)!} = 70$

9.1 Tipos de Amostragem

Amostragem Simples ou Ocasional

É o processo mais elementar e freqüentemente utilizado. Todos os elementos da população têm igual probabilidade de serem escolhidos. Para uma população finita o processo deve ser sem reposição. Todos os elementos da população devem ser

numerados. Para realizar o sorteio dos elementos da população devemos usar a **Tabela de Números Aleatórios**.

Amostragem Sistemática

Trata-se de uma variação da Amostragem Aleatória Ocasional, conveniente quando a população está naturalmente ordenada, como fichas em um fichário, lista telefônica, etc.

Exemplo: N = 5000, n = 50, r = N / n = 10 (P. A. de razão 5)

Sorteia-se usando a **Tabela de Números Aleatórios** um número entre 1 e 10, (x = 3), o número sorteado refere-se ao 1º elemento da amostra, logo os elementos da amostra serão:

3 13 23 33 43 .....

Para determinar qualquer elemento da amostra podemos usar a fórmula do termo geral de uma P.A.

$$\mathbf{a_n = a_1 + (n - 1).\ r}$$

Amostragem Estratificada

É um processo de amostragem usado quando nos depararmos com populações heterogêneas, na qual pode-se distinguir subpopulações mais ou menos homogêneas, denominados estratos.

Após a determinação dos estratos, seleciona-se uma amostra aleatória de cada uma subpopulação (estrato).

As diversas subamostras retiradas das subpopulações devem ser proporcionais aos respectivos números de elementos dos estratos, e guardarem a proporcionalidade em relação à variabilidade de cada estrato, obtendo-se uma estratificação ótima.

Tipos de variáveis que podem ser usadas em estratificação: idade, classes sociais, sexo, profissão, salário, procedência, etc.

Tamanho da Amostra

Os pesquisadores de todo o mundo, na realização de pesquisas científicas, qualquer setor da atividade humana, utilizam as técnicas de amostragem no planejamento de seus trabalhos, não só pela impraticabilidade de poderem observar, numericamente, em sua totalidade determinada população em estudo, como devido ao aspecto econômico dessas investigações, conduzida com um menor custo operacional, dentro de um menor tempo, além de possibilitar maior precisão nos respectivos resultados, ao contrário, do que ocorre com os trabalhos realizados pelo processo censitário (COCHRAN, 1965; CRUZ, 1978). A técnica da amostragem, a despeito de sua larga utilização, ainda necessita de alguma didática mais adequada aos pesquisadores iniciantes.

Na teoria da amostragem, são consideradas duas dimensões:

1ª) Dimensionamento da Amostra;

2ª) Composição da Amostra.

9.2 Procedimentos para determinar o tamanho da amostra

1º) Analisar o questionário, ou roteiro da entrevista e escolher uma variável que julgue mais importante para o estudo. Se possível mais do que uma;

2º) Verificar o nível de mensuração da variável: nominal, ordinal ou intervalar;

3º) Considerar o tamanho da população: infinita ou finita

4º) Se a variável escolhida for:

Intervalar e a população considerada infinita, você poderá determinar o tamanho da amostra pela fórmula:

$$n = \left(\frac{Z.\sigma}{d} \right)^2$$

onde:

Z = abscissa da curva normal padrão, fixado um nível de confiança (1- □□

Z = 1,65 □(1 - □) = 90%

Z = 1,96 □(1 - □) = 95%

Z = 2,0 □(1 - □) = 95.5%

Z = 2,57 □(1 - □) = 99%

Geralmente usa-se Z = 2

□□= desvio padrão da população, expresso na unidade variável, onde poderá ser

determinado por:

- □Especificações Técnicas
- Resgatar o valor de estudos semelhantes
- Fazer conjeturas sobre possíveis valores

d = erro amostral, expresso na unidade da variável. O erro amostral é a máxima diferença que o investigador admite suportar entre □□e $\overline{X}$, isto é: |□□- $\overline{X}$ | < d .

Intervalar e a população considerada finita, você poderá determinar o tamanho da amostra pela fórmula:

$$n = \frac{Z^2.\sigma^2.N}{d^2.(N-1)+Z^2.\sigma^2}$$

onde:

Z = abscissa da normal padrão

σ^2 = variância populacional

N = tamanho da população

d = erro amostral

Nominal ou ordinal, e a população considerada infinita, você poderá determinar o tamanho da amostra pela fórmula:

$$n = \frac{Z^2.\hat{p}.\hat{q}}{d^2}$$

onde:

Z = abscissa da normal padrão

$\hat{p}$ = estimativa da verdadeira proporção de um dos níveis da variável escolhida. Por exemplo, se a variável escolhida for parte da empresa, $\hat{p}$ poderá ser a estimativa

da verdadeira proporção de grandes empresas do setor que está sendo estudado. $\hat{p}$ será expresso em decimais ($\hat{p}$ = 30% □ $\hat{p}$ = 0.30).

$\hat{q}$ = 1 - $\hat{p}$

d = erro amostral, expresso em decimais. O erro amostral neste caso será a máxima diferença que o investigador admite suportar entre □□□e $\hat{p}$, isto é: |□□□- $\hat{p}$ | < d, em que □□□é a verdadeira proporção (*p* é a freqüência relativa do evento a ser calculado a partir da amostra).

Nominal ou ordinal, e a população considerada finita, você poderá determinar o tamanho da amostra pela fórmula:

$$n = \frac{Z^2.\hat{p}.\hat{q}.N}{d^2(N-1)+Z^2.\hat{p}.\hat{q}}$$

onde:

Z = abscissa da normal padrão

N = tamanho da população

$\hat{p}$ = estimativa da proporção

$\hat{q}$ = 1 - $\hat{p}$

d = erro amostral

Estas fórmulas são básicas para qualquer tipo de composição da amostra; todavia, existem fórmulas específicas segundo o critério de composição da amostra.

Se o investigador escolher mais de uma variável, poderá acontecer de ter que aplicar mais de uma fórmula, assim deverá optar pelo maior valor de "n".

Obs.: Quando não tivermos condições de prever o possível valor para $\hat{p}$, admita $\hat{p}$ = 0.50, pois, dessa forma, você terá o maior tamanho da amostra, admitindo-se constantes os demais elementos.

10. Regressão

10.1. Correlação

Coeficiente de Correlação

O coeficiente de correlação (*r*) representa a relação entre duas ou mais variáveis. O valor de *r* está sempre entre –1 e +1. Quando *r* = 0 não há correlação entre as variáveis. Temos *correlação positiva* quando $r > 0$ e *correlação negativa* quando $r < 0$. Em geral, um coeficiente maior que *0,3* é suficiente para indicar correlação.

Medida de Correlação de Pearson (r)

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{n\sum x^2 - \left(\sum x\right)^2}\sqrt{n\sum y^2 - \left(\sum y\right)^2}}$$

Se existe relação direta, *r* é positivo. Se a relação é inversa, *r* é negativo. Existem várias fórmulas para cálculo do coeficiente de correlação. O modo mais simples de obtê-lo é pela raiz quadrada do coeficiente de determinação (R^2). Costuma-se classificar o coeficiente, conforme seu valor, como apresentado na tabela a seguir.

Níveis de correlação

Valor	*Correlação*
R = 0	nula
0 < |R| < 0.30	fraca
0.30 < |R| < 0.60	média
0.60 < |R| < 0.90	forte
0.90 < |R| < 1	fortíssima
|R| = 1	perfeita

10.2. Análise de Regressão

Regressão é um modelo estatístico para descrever relações entre variáveis. Os objetivos principais são identificar a relação e fazer inferências sobre os parâmetros do modelo. A análise de regressão é utilizada principalmente com o objetivo de previsão. Nosso propósito na análise de regressão é o desenvolvimento de um modelo

estatístico que possa ser utilizado para prever os valores de uma variável dependente, com base nos valores de pelo menos uma variável independente.

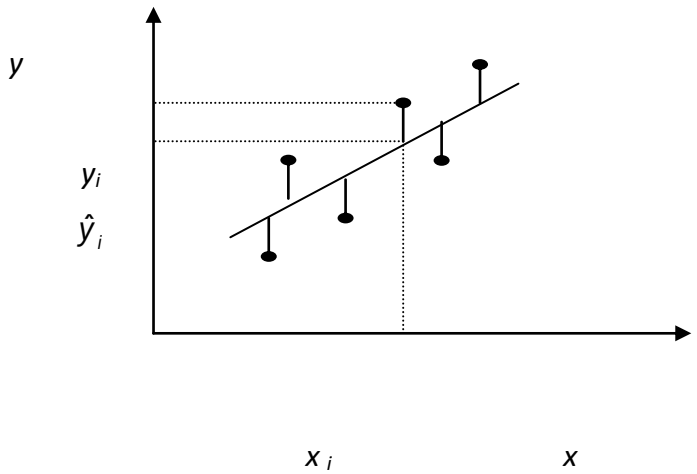
10.2.1. Regressão Linear Simples

Dado um conjunto de valores observados de X e Y, construir um modelo de regressão linear de Y sobre X consiste em obter, a partir desses valores, uma reta que melhor represente a relação verdadeira entre essas variáveis. Os parâmetros do modelo são obtidos apartir de uma amostra utilizando o método dos mínimos quadrados que analisa as *n* diferenças entre cada valor Y e o valor na reta, correspondente ao respectivo valor X e apresenta a menor soma de quadrados de tal diferença.

A análise de regressão linear significa encontrar a reta que melhor se ajuste aos dados. O melhor ajuste significa a tentativa de encontrar a reta para qual as diferenças entre os valores reais (Y_i) e os valores previstos da regressão ajustada $\left(\hat{Y}_i\right)$ sejam as menores possíveis.

A reta ajustada é representada por $\hat{Y} = \beta_0 + \beta_1 X$, onde:

- β_0 é o valor onde a reta intercepta o eixo *Y*
- β_1 é a inclinação



Distâncias cuja soma dos quadrados deve ser minimizada

O modelo estatístico:

$$Y = \beta_0 + \beta_1 X + \varepsilon,$$

onde β_0 e β_1 são os parâmetros do modelo e $\varepsilon \sim iid\ N(0, \sigma^2)$;

ε é o erro aleatório dado pela diferença entre o valor observado *y* e o valor obtido pela reta $\beta_0 + \beta_1 x$ (leva em conta a falha do modelo em ajustar exatamente os dados);

A equação $Y = \beta_0 + \beta_1 X + \varepsilon$ é denominada modelo de regressão linear simples. Usualmente *X* é denominado variável independente, explicativa regressora ou preditora e *Y* é denominada variável dependente ou variável resposta.

Teste de Hipóteses sobre β_1

Um caso muito importante das hipóteses é:

Para as hipóteses: $H_0: \beta_1 = 0$ versus $H_A: \beta_1 \neq 0$;

Estatística de Teste: $t_0 = \dfrac{\hat{\beta}_1}{\sqrt{\sum (y_i - \hat{y}_i)^2 / n-2}} \sqrt{\sum (x_i - \bar{x})^2}$, que tem distribuição *t* de Student, com n-2 graus de liberdade, sob H_0;

Para um teste com nível de significância α, a região de rejeição de H_0 é da forma:

$|t_0| > t_{\alpha/2,\, n-2}$.

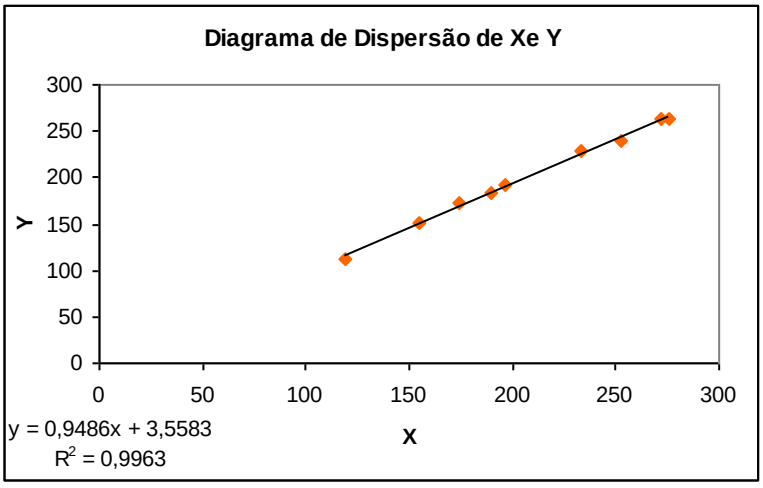
Neste teste, não rejeitar H_0 é equivalente a concluir que não existe relação linear entre *X* e *Y*. Note que este resultado pode implicar que *X* é pouco importante para explicar a variação em *Y* e que a relação verdadeira entre *X* e *Y* não é linear. Caso contrário, se H_0 é rejeitada, isto implica que *X* é importante para explicar a variabilidade em *Y* e pode indicar que o modelo linear é adequado ou que, apesar de haver um efeito linear de *X*, um modelo melhor poderá ser obtido se forem incluídos termos de ordem mais elevada em *X*.

Suposições Associadas ao Modelo de Regressão Linear Simples:

1. Os erros têm média zero e a mesma variância desconhecida σ^2 (Homocedasticidade).
2. Os erros são não correlacionados, ou seja, o valor de um erro não depende de qualquer outro erro.
3. A variável explicativa *X* é controlada pelo experimentador e é medida com erro desprezível (erro não significante do ponto de vista prático), ou seja, não é uma variável aleatória.
4. Os erros têm distribuição normal.

Exemplo: A mensuração exata (*Y*) de uma substância poluente é realizada por meio de uma análise química muito cara. Um novo método mais barato resulta na medida *X*, que supostamente pode ser usada para prever o valor de *Y*. Nove amostras foram obtidas e avaliadas pelos dois métodos, obtendo as medidas abaixo.

X	119	155	174	190	196	233	272	253	276
Y	112	152	172	183	192	228	263	239	263



O modelo ajustado $\hat{y} = 3{,}5583 + 0{,}9486x$

Para testar a hipótese $H_0 : \beta_1 = 0$, usamos o **valor p** como regra de decisão.

	Coeficientes	*Erro padrão*	*Stat t*	*valor-P*
Interseção	3,5583	4,6989	0,7573	0,4736
Variável X	0,9486	0,0220	43,1630	0,0000

Logo, há evidências para rejeitar H_0, ou seja, a média da variável Y é influenciada pela variável X.

Exemplo: O custo de um lote de uma certa peça depende do número de peças produzidas, ou seja, do tamanho do lote. Em uma amostra de dez lotes, observaram-se os seguintes resultados:

Tamanho	5	10	15	20	25	30	35	40	45	50
Custo	65	120	210	260	380	450	510	555	615	660

	Coeficientes	*Erro Padrão*	*Stat t*	*valor-P*
Interseção	3,6667	16,5735	0,2212	0,8305
Tamanho	13,7758	0,5342	25,7871	0,0000

O modelo ajustado:

$\hat{y}$ (custo)= 13,776(tamanho) + 3,6667

Teste de Significância de β_1

Há evidências para rejeição de H_0.. Assim os dados fornecem evidências que o tamanho do lote influência em média no custo.

Estimação da média da variável resposta:

O custo esperado correspondente a um lote de tamanho =12 é estimado por:

$\hat{y}$ (Custo) = 13,776(12) + 3,6667

$\hat{y}$ (Custo) = 168,98

Estimação por intervalo:

Seja *X* e *Y* duas v. a. quantitativas a esperança condicional de *Y*, dado *X* = *x*, por exemplo, é denotado por *E* (*Y* | *x*), é uma função de *x*, ou seja,

E (*Y* | *x*) = $\mu(x)$ = $\beta_0 + \beta_1 x$,

$$\hat{\mu}(x) \pm t_{n-2;\alpha}\left[QM\,\mathrm{Res}\left(\frac{1}{n}+\frac{(x-\bar{x})^2}{S_x^2}\right)\right]^{1/2}$$

onde $\hat{\mu}(x)$ é o estimador para $\mu(x)$ e *QMRes* = $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \Big/ n-2$

Assim um intervalo de confiança a 95% para μ (12) é dado por (142,95; 195,01*)*

Previsão de uma observação futura da variável resposta:

Suponha que desejamos prever o custo para o lote de tamanho = 55:

A estimativa do custo médio é:

$\hat{y}$ (custo) = 13,776(55) + 3,6667

$\hat{y}$ (custo) = 731,33

E o intervalo de previsão a 95% é dado por:

$$\hat{\mu}(x^*) \pm t_{n-2;\alpha/2}\left\{QM\,\mathrm{Re}s\left[1+\frac{1}{n}+\frac{(x^*-\bar{x})^2}{S_x^2}\right]\right\}^{1/2}$$

onde x^*: é um dado valor (nível) *x*; de $\hat{\mu}(x^*)$ é o estimador para $\mu(x^*)$ e *QMRes* = $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \Big/ n-2$

= (693,58; 829,09)

10.2.2. Regressão Linear Múltipla

Em geral, a variável resposta pode estar relacionada a *p* variáveis explicativas, isto é, *Y* apresenta-se como função de várias variáveis regressoras. Assim não teremos mais uma regressora, mas um vetor de variáveis regressoras, ou seja, $X = (X_1, X_2, ..., X_p)$. Da mesma natureza como foi tratado no modelo simples, o processo de estimação dos parâmetros será análogo, porém generalizando de forma matricial. Em suma a regressão linear simples é um caso particular da regressão linear múltipla.

O modelo ajustado é representado por $\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p$, Onde $\beta_0, \beta_1, \beta_2, ..., \beta_p$ são os parâmetros do modelo.

Na definição do nome, o adjetivo **linear** é usado para indicar que o modelo é **linear nos parâmetros** $\beta_0, \beta_1, \beta_2, ..., \beta_p$, e não porque *Y* é função linear dos *x*'s.

Modelo estatístico:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon,$$

onde β_0, β_1, β_2, ..., β_p. Neste modelo estes parâmetros representam a alteração esperada na variável resposta *Y* quando a variável x_j sofre um acréscimo unitário, enquanto todas as demais variáveis explicativas x_i *(i≠j)* são mantidas constantes denominados assim como coeficientes parciais de regressão e $\varepsilon \sim iid\ N(0, \sigma^2)$;

Teste da Significância da Regressão

Com objetivo de determinar se existe um relacionamento linear entre a variável resposta *Y* e o conjunto de variáveis regressoras $x_1, x_2, \ldots, x_p$ pode ser utilizado o teste de significância da regressão:

Para as hipóteses: $H_0: \beta_1 = \beta_2 = \ldots = \beta_p = 0$

$H_A: \beta_j \neq 0$ para pelo menos um j;

Estatística de Teste: $F_0 = \dfrac{QM\ \mathrm{Re}\, g}{QM\ \mathrm{Re}\, s}$, que tem distribuição F, com *(p, n-p-1)* graus de liberdade, sob H_0.

Para um teste com nível de significância α, a região de rejeição de H_0 é da forma:

$F_0 > F_\alpha$, (*p, n-p-1*).

A rejeição de H_0 significa que pelo menos uma das variáveis regressoras $x_1, x_2, \ldots, x_p$ contribui de modo significativo para o modelo.

Análise de Variância para a Significância da Regressão Linear Múltipla:

Causa de Variação	Graus de Liberdade	Soma de Quadrados	Quadrados Médios	F_o
Regressão	p	SQReg	QMReg	QMReg / QMRes
Resíduo	n-p-1	SQRes	QMR	
Total	n-1	SQT		

Onde: SQReg = Soma de Quadrados de Regressão, SQR= Soma de Quadrados de Resíduos, SQT= Soma de Quadrados Total, QMReg= Quadrado Médio de Regressão, QMRes= Quadrado Médio dos Resíduos.

Onde:

$$SQ\ \mathrm{Re}\, s = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2,\ SQT = \sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2,\ SQ\ \mathrm{Re}\, g = SQT - SQ\ \mathrm{Re}\, s,$$

$$QM\ \mathrm{Re}\, g = SQ\ \mathrm{Re}\, g / p,\ QM\ \mathrm{Re}\, s = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2 \Big/ n - p - 1$$

Exemplo: Uma indústria produtora de alumínio esta interessada em conhecer o tipo de relacionamento existente entre as variáveis razão $Al_2O_3/NaOH$ (Nesta razão, o símbolo Al_2O_3 está representado a massa de óxido de alumínio proveniente de da bauxita que entra no processo de produção, e o símbolo *NaOH* de refere a massa de hidróxido de sódio, um dos reagentes do processo, que é empregada na fabricação da alumina) (x_1), temperatura de reação (x_2) e o teor de Na_2O (y) ocluído na alumina. Com este objetivo, a equipe da indústria construiu uma tabela a partir de registros históricos de dados existentes. Suspeita-se da existência de relacionamento linear entre o teor de Na_2O e as variáveis razão $Al_2O_3/NaOH$ e temperatura de reação.

Y	x_1	x_2	y	x_1	x_2
0,43	0,647	77,1	0,39	0,633	76,5
0,39	0,638	78,3	0,41	0,645	78,6
0,44	0,651	76,0	0,43	0,642	74,7
0,42	0,648	77,9	0,40	0,638	75,5
0,43	0,640	74,1	0,39	0,635	78,2
0,42	0,643	74,6	0,40	0,639	75,9
0,41	0,643	76,0	0,40	0,639	76,6
0,46	0,651	73,3	0,42	0,645	78,0
0,42	0,650	78,6	0,44	0,650	77,2
0,40	0,639	78,7	0,40	0,642	78,0
0,39	0,636	77,8	0,43	0,648	76,1
0,41	0,641	75,8	0,42	0,642	74,6
0,43	0,649	77,3	0,39	0,633	77,5

Para a hipótese: H_0: $\beta_0 = \beta_1 = \beta_2 = 0$ (as variáveis x_1 e x_2 não são significativas para explicar *Y*).

Resumos dos resultados:

Causas de variação	*Gl*	*SQ*	*MQ*	*F*	*F de significação*
Regressão	2	0,0081	0,0040	118,3168	0,0000
Resíduo	23	0,0008	0,0000		
Total	25	0,0088			

	Coeficientes	*Erro padrão*	*Stat t*	*valor-P*
Interseção	-0,9878	0,1543	-6,4024	0,0000
Variável *X 1*	2,7904	0,2129	13,1037	0,0000
Variável *X 2*	-0,0051	0,0008	-6,6721	0,0000

Portanto, o modelo de regressão ajustado foi expresso por:

$$\hat{y} = -0{,}9878 + 2{,}7904x_1 - 0{,}0051x_2$$

Esta equação indica que, em média, para cada aumento de uma unidade na razão *Al_2O_3/NaOH*, com a temperatura mantida constante, o teor de *Na_2O* ocluído na alumina aumenta de 2,790 %. De modo similar, para cada aumento de 1 ^{0}C na temperatura de reação, com a razão *Al_2O_3/NaOH* mantida constante, o teor de *Na_2O* ocluído na alumina sofre uma diminuição, em média, de 0,0051%.

Como F_0 = 118,32 > $F_{0,05}(2,23)$ = 3,42, há evidencias para rejeição de H_0 com isso concluímos que o teor de *Na_2O* ocluído na alumina está relacionado à razão *Al_2O_3/NaOH* e/ou temperatura de reação.

10.3. Coeficiente de Determinação (R^2)

O coeficiente de determinação mede a proporção da variação, que é explicada pela variável independente no modelo de regressão.

Quando R^2 = 0, a variação explicada de *Y* é zero, ou seja, a reta ajustada é paralela ao eixo da variável *X*. Se R^2 = 1, a reta ajustada explicará toda a variação de *Y*. Assim sendo, quanto mais próximo da unidade estiver o valor de R^2, melhor será "a qualidade" do ajuste e quanto mais próximo de zero pior será "a qualidade" do ajuste.

Se o poder explicativo for, por exemplo, 98%, isto significa que 98% das variações de *Y* são explicadas por *X* através da função escolhida para relacionar as duas variáveis e 2% são atribuídas a causas aleatórias.

11. Análise de Sobrevivência

Análise Estatística de Sobrevivência é um método estatístico usado para analisar dados provenientes de situações médicas envolvendo dados censurados. Os mesmos métodos têm aplicações na confiabilidade industrial, ciências sociais e negócios, e neste caso leva o nome de Teoria da Confiabilidade.

A variável aqui será definida como o tempo de falha e consideram-se os tempos entre falhas que em sobrevivência podem ser morrer, recair, recuperar e etc, e na confiabilidade pode ser falha de itens eletrônicos, um mau funcionamento especificado de um produto, entre outros.

11.1 Organização da Pesquisa Médica

É comum ouvirmos através dos meios de comunicação relatos sobre a saúde e os avanços da medicina, vamos estudar meios para compreensão e avaliação crítica desses relatos.

Começaremos expondo as principais formas básicas de pesquisa utilizadas: estudos descritivos, estudos caso-controle, estudos tipo coorte e ensaio clínico aleatorizado.

Estudo Descritivo

Neste estudo o principal objetivo é puramente a descrição de um fato médico, onde sua principal característica é a ausência de um grupo para comparação.

Estudo Caso-Controle

É uma forma de estudo que visa verificar se os indivíduos, selecionados porque têm uma doença – Os casos – diferem significativamente, em relação à exposição a um dado fator de risco, de um grupo de indivíduos comparáveis, mas que não possuem a doença – Os controles.

Estudos Tipo Coorte

Aqui o objetivo é verificar se indivíduos, selecionados porque foram expostos ao fator de risco, desenvolvem a doença em questão em maior ou menor proporção do que um grupo de indivíduos, comparáveis, mas não expostos ao fator de risco.

Ensaios Clínicos Aleatorizados

É um experimento médico realizado com o objetivo de verificar, entre dois ou mais tratamentos, qual produz mais efeito.

11.2 Função de Sobrevivência

Seja *T* uma variável aleatória representando o tempo de falha do paciente. Então, a função de sobrevivência é a probabilidade de uma observação não falhar até um certo tempo *t*, ou seja, a probabilidade de uma observação sobreviver a um tempo maior ou igual a *t*.

$$F(t) = P(T \leq t), para\ t \geq 0 \ \rightarrow Função\ Distribuição\ de\ T$$

$$S(t) = 1 - F(t) = P(T > t), para\ t \geq 0 \ \rightarrow Função\ de\ Sobrevivência\ de\ T$$

11.3 Função de Taxa de Falha ou de Risco

Podemos expessar o intervalo de tempo $[t_1, t_2)$ em termos da função de sobrevivência como:

$$S(t_1) - S(t_2)$$

De maneira geral definindo o intervalo como $[t, t + \Delta t)$ a taxa de falha fica definida como sendo:

$$h(t) = \frac{S(t) - S(t + \Delta t)}{\Delta t \cdot S(t)}$$

11.4 Método Não-parámetrico para Análise de Sobrevivência

Estimador Kaplan-Meier

Também chamado de estimador limite-produto foi proposto por Kaplan e Meier (1958) o qual permite a presença de informações censuradas em que os tempos de sobrevivência são ordenados, isto é, $t_1 \leq t_2 \leq \cdots \leq t_n$.

Sua função é estimada como:

$$\widehat{S(t)} = \prod_{i|t_i<t}^{k} \left(\frac{n_i - d_i}{n_i}\right) = \prod_{i|t_i<t}^{k} \left(1 - \frac{d_i}{n_i}\right)$$

Onde:

d_i: *Números de falha no tempo* t_i

n_i: *Números de observações sob risco (não falhou e não foi censurado) até o tempo* t_i *(exclusive).*

t_i: *É o maior tempo de sobrevivência menor ou igual a t.*

Exemplo: Um estudo clínico foi realizado para investigar o efeito da terapia com esteróide no tratamento de hepatite viral aguda (gregory et al., 1995). Vinte e nove pacientes com esta doença foram aleatorizados para receber um placebo ou o tratamento com esteróide. Cada paciente foi acompanhado por 16 semanas ou até a morte (evento de interesse) ou até a perda do acompanhamento. Os tempos de sobrevivência observados, em semanas, para os dois grupos foram (+ incida censura):

Grupo controle (15 pacientes): +, 2+, 3, 3, 3+, 5+, 16+, 16+, 16+, 16+, 16+, 16+, 16+, 16+.

Grupo Esteróide (14 pacientes): 1, 1, 1, 1+, 4+, 5, 7, 8, 10, 10+, 12+, 16+, 16+, 16+.

Kaplan-Meier (Product-limit) analysis (esteróide)

Note: Censored cases are marked with +Cumulatv Standard

Time	di	ni	Survival	Error
1	3	14	0,785714	0,109664
5	1	9	0,698413	0,127581
7	1	8	0,611111	0,138315
8	1	7	0,523810	0,143486
10	1	6	0,436508	0,143696

12. Tabelas

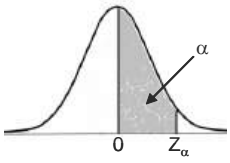


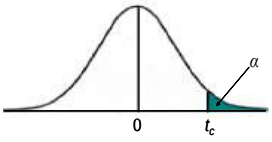
Tabela da distribuição Normal (0, 1),

$P(0 \leq Z \leq Z_\alpha) = \alpha$

z_α	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,00000	0,00399	0,00798	0,01197	0,01595	0,01994	0,02392	0,02790	0,03188	0,03586
0,1	0,03983	0,04380	0,04776	0,05172	0,05567	0,05962	0,06356	0,06749	0,07142	0,07535
0,2	0,07926	0,08317	0,08706	0,09095	0,09483	0,09871	0,10257	0,10642	0,11026	0,11409
0,3	0,11791	0,12172	0,12552	0,12930	0,13307	0,13683	0,14058	0,14431	0,14803	0,15173
0,4	0,15542	0,15910	0,16276	0,16640	0,17003	0,17364	0,17724	0,18082	0,18439	0,18793
0,5	0,19146	0,19497	0,19847	0,20194	0,20540	0,20884	0,21226	0,21566	0,21904	0,22240
0,6	0,22575	0,22907	0,23237	0,23565	0,23891	0,24215	0,24537	0,24857	0,25175	0,25490
0,7	0,25804	0,26115	0,26424	0,26730	0,27035	0,27337	0,27637	0,27935	0,28230	0,28524
0,8	0,28814	0,29103	0,29389	0,29673	0,29955	0,30234	0,30511	0,30785	0,31057	0,31327
0,9	0,31594	0,31859	0,32121	0,32381	0,32639	0,32894	0,33147	0,33398	0,33646	0,33891
1,0	0,34134	0,34375	0,34614	0,34849	0,35083	0,35314	0,35543	0,35769	0,35993	0,36214
1,1	0,36433	0,36650	0,36864	0,37076	0,37286	0,37493	0,37698	0,37900	0,38100	0,38298
1,2	0,38493	0,38686	0,38877	0,39065	0,39251	0,39435	0,39617	0,39796	0,39973	0,40147
1,3	0,40320	0,40490	0,40658	0,40824	0,40988	0,41149	0,41308	0,41466	0,41621	0,41774
1,4	0,41924	0,42073	0,42220	0,42364	0,42507	0,42647	0,42785	0,42922	0,43056	0,43189
1,5	0,43319	0,43448	0,43574	0,43699	0,43822	0,43943	0,44062	0,44179	0,44295	0,44408
1,6	0,44520	0,44630	0,44738	0,44845	0,44950	0,45053	0,45154	0,45254	0,45352	0,45449
1,7	0,45543	0,45637	0,45728	0,45818	0,45907	0,45994	0,46080	0,46164	0,46246	0,46327

1,8	0,46407	0,46485	0,46562	0,46638	0,46712	0,46784	0,46856	0,46926	0,46995	0,47062
1,9	0,47128	0,47193	0,47257	0,47320	0,47381	0,47441	0,47500	0,47558	0,47615	0,47670
2,0	0,47725	0,47778	0,47831	0,47882	0,47932	0,47982	0,48030	0,48077	0,48124	0,48169
2,1	0,48214	0,48257	0,48300	0,48341	0,48382	0,48422	0,48461	0,48500	0,48537	0,48574
2,2	0,48610	0,48645	0,48679	0,48713	0,48745	0,48778	0,48809	0,48840	0,48870	0,48899
2,3	0,48928	0,48956	0,48983	0,49010	0,49036	0,49061	0,49086	0,49111	0,49134	0,49158
2,4	0,49180	0,49202	0,49224	0,49245	0,49266	0,49286	0,49305	0,49324	0,49343	0,49361
2,5	0,49379	0,49396	0,49413	0,49430	0,49446	0,49461	0,49477	0,49492	0,49506	0,49520
2,6	0,49534	0,49547	0,49560	0,49573	0,49585	0,49598	0,49609	0,49621	0,49632	0,49643
2,7	0,49653	0,49664	0,49674	0,49683	0,49693	0,49702	0,49711	0,49720	0,49728	0,49736
2,8	0,49744	0,49752	0,49760	0,49767	0,49774	0,49781	0,49788	0,49795	0,49801	0,49807
2,9	0,49813	0,49819	0,49825	0,49831	0,49836	0,49841	0,49846	0,49851	0,49856	0,49861
3,0	0,49865	0,49869	0,49874	0,49878	0,49882	0,49886	0,49889	0,49893	0,49896	0,49900
3,1	0,49903	0,49906	0,49910	0,49913	0,49916	0,49918	0,49921	0,49924	0,49926	0,49929
3,2	0,49931	0,49934	0,49936	0,49938	0,49940	0,49942	0,49944	0,49946	0,49948	0,49950
3,3	0,49952	0,49953	0,49955	0,49957	0,49958	0,49960	0,49961	0,49962	0,49964	0,49965
3,4	0,49966	0,49968	0,49969	0,49970	0,49971	0,49972	0,49973	0,49974	0,49975	0,49976
3,5	0,49977	0,49978	0,49978	0,49979	0,49980	0,49981	0,49981	0,49982	0,49983	0,49983
3,6	0,49984	0,49985	0,49985	0,49986	0,49986	0,49987	0,49987	0,49988	0,49988	0,49989
3,7	0,49989	0,49990	0,49990	0,49990	0,49991	0,49991	0,49992	0,49992	0,49992	0,49992
3,8	0,49993	0,49993	0,49993	0,49994	0,49994	0,49994	0,49994	0,49995	0,49995	0,49995
3,9	>0,49995	etc ...								
z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09

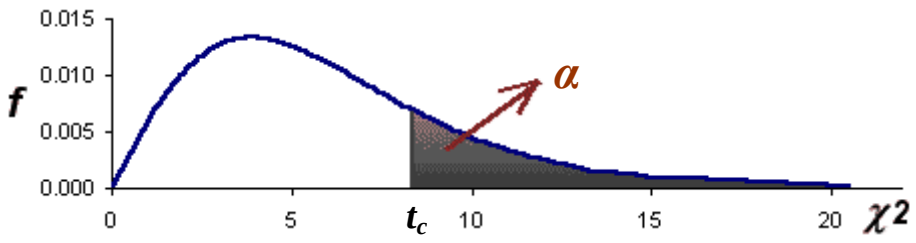
Tabela da distribuição *t* de Student, corpo da tabela dá valores t_c tais que P(t > t_c) = α , para graus de liberdade > 120, usar a aproximação normal.



α	0,2	0,15	0,1	0,05	0,025	0,02	0,015	0,01	0,005	α
g. l.	$t_{0.800}$	$t_{0.850}$	$t_{0.900}$	$t_{0.950}$	$t_{0.975}$	$t_{0.980}$	$t_{0.985}$	$t_{0.990}$	$t_{0.995}$	g. l.
1	1,3764	1,9626	3,0777	6,3138	12,7062	15,8945	21,2051	31,8210	63,6559	1
2	1,0607	1,3862	1,8856	2,9200	4,3027	4,8487	5,6428	6,9646	9,9250	2
3	0,9785	1,2498	1,6378	2,3534	3,1825	3,4819	3,8961	4,5407	5,8409	3
4	0,9410	1,1896	1,5332	2,1319	2,7765	2,9985	3,2976	3,7469	4,6041	4
5	0,9195	1,1558	1,4759	2,0151	2,5706	2,7565	3,0029	3,3649	4,0321	5
6	0,9057	1,1342	1,4398	1,9432	2,4469	2,6122	2,8289	3,1427	3,7074	6
7	0,8960	1,1192	1,4149	1,8946	2,3646	2,5168	2,7146	2,9980	3,4995	7
8	0,8889	1,1082	1,3968	1,8596	2,3060	2,4490	2,6338	2,8965	3,3554	8
9	0,8834	1,0997	1,3830	1,8331	2,2622	2,3984	2,5738	2,8214	3,2498	9
10	0,8791	1,0931	1,3722	1,8125	**2,2281**	2,3593	2,5275	2,7638	3,1693	10
11	0,8755	1,0877	1,3634	1,7959	2,2010	2,3281	2,4907	2,7181	3,1058	11
12	0,8726	1,0832	1,3562	1,7823	2,1788	2,3027	2,4607	2,6810	3,0545	12
13	0,8702	1,0795	1,3502	1,7709	2,1604	2,2816	2,4359	2,6503	3,0123	13
14	0,8681	1,0763	1,3450	1,7613	2,1448	2,2638	2,4149	2,6245	2,9769	14
15	0,8662	1,0735	1,3406	1,7531	2,1315	2,2485	2,3970	2,6025	2,9467	15
16	0,8647	1,0711	1,3368	1,7459	2,1199	2,2354	2,3816	2,5835	2,9208	16
17	0,8633	1,0690	1,3334	1,7396	2,1098	2,2238	2,3681	2,5669	2,8982	17
18	0,8621	1,0672	1,3304	1,7341	2,1009	2,2137	2,3562	2,5524	2,8784	18
19	0,8610	1,0655	1,3277	1,7291	2,0930	2,2047	2,3457	2,5395	2,8609	19
20	0,8600	1,0640	1,3253	1,7247	2,0860	2,1967	2,3363	2,5280	2,8453	20

21	0,8591	1,0627	1,3232	1,7207	2,0796	2,1894	2,3278	2,5177	2,8314	**21**
22	0,8583	1,0615	1,3212	1,7171	2,0739	2,1829	2,3202	2,5083	2,8188	**22**
23	0,8575	1,0603	1,3195	1,7139	2,0687	2,1770	2,3132	2,4999	2,8073	**23**
24	0,8569	1,0593	1,3178	1,7109	2,0639	2,1716	2,3069	2,4922	2,7970	**24**
25	0,8562	1,0584	1,3164	1,7081	2,0595	2,1666	2,3011	2,4851	2,7874	**25**
26	0,8557	1,0575	1,3150	1,7056	2,0555	2,1620	2,2958	2,4786	2,7787	**26**
27	0,8551	1,0567	1,3137	1,7033	2,0518	2,1578	2,2909	2,4727	2,7707	**27**
28	0,8547	1,0560	1,3125	1,7011	2,0484	2,1539	2,2864	2,4671	2,7633	**28**
29	0,8542	1,0553	1,3114	1,6991	2,0452	2,1503	2,2822	2,4620	2,7564	**29**
30	0,8538	1,0547	1,3104	1,6973	2,0423	2,1470	2,2783	2,4573	2,7500	**30**
35	0,8520	1,0520	1,3062	1,6896	2,0301	2,1332	2,2622	2,4377	2,7238	**35**
40	0,8507	1,0501	1,3031	1,6839	2,0211	2,1229	2,2503	2,4233	2,7045	**40**
45	0,8497	1,0485	1,3007	1,6794	2,0141	2,1150	2,2411	2,4121	2,6896	**45**
50	0,8489	1,0473	1,2987	1,6759	2,0086	2,1087	2,2338	2,4033	2,6778	**50**
60	0,8477	1,0455	1,2958	1,6707	2,0003	2,0994	2,2229	2,3901	2,6603	**60**
70	0,8468	1,0442	1,2938	1,6669	1,9944	2,0927	2,2152	2,3808	2,6479	**70**
80	0,8461	1,0432	1,2922	1,6641	1,9901	2,0878	2,2095	2,3739	2,6387	**80**
90	0,8456	1,0424	1,2910	1,6620	1,9867	2,0839	2,2050	2,3685	2,6316	**90**
100	0,8452	1,0418	1,2901	1,6602	1,9840	2,0809	2,2015	2,3642	2,6259	**100**
110	0,8449	1,0413	1,2893	1,6588	1,9818	2,0784	2,1986	2,3607	2,6213	**110**
120	0,8446	1,0409	1,2887	1,6577	1,9799	2,0763	2,1962	2,3578	2,6174	**120**
∞	0,8420	1,0370	1,2824	1,6464	1,9623	2,0564	2,1732	2,3301	2,5808	∞
α	**0,2** $t_{0.800}$	**0,15** $t_{0.850}$	**0,1** $t_{0.900}$	**0,05** $t_{0.950}$	**0,025** $t_{0.975}$	**0,02** $t_{0.980}$	**0,015** $t_{0.985}$	**0,01** $t_{0.990}$	**0,005** $t_{0.995}$	α

Tabela da distribuição *Qui-Quadrada*, corpo da tabela dá valores t_c tais que P($\chi^2 > t_c$) = α , para graus de liberdade > 120, usar a aproximação normal.



A	0,995	0,975	0,9	0,5	0,1	0,05	0,025	0,01	0,005	0,001
g.l. 1	0,0000	0,0010	0,0160	0,4550	2,7060	3,8410	5,0240	6,6350	7,8790	10,8270
2	0,0100	0,0510	0,2110	1,3860	4,6050	5,9910	7,3780	9,2100	10,5970	13,8150
3	0,0720	0,2160	0,5840	2,3660	6,2510	7,8150	9,3480	11,3450	12,8380	16,2660
4	0,2070	0,4840	1,0640	3,3570	7,7790	9,4880	11,1430	13,2770	14,8600	18,4660
5	0,4120	0,8310	1,6100	4,3510	9,2360	11,0700	12,8320	15,0860	16,7500	20,5150
6	0,6760	1,2370	2,2040	5,3480	10,6450	12,5920	14,4490	16,8120	18,5480	22,4570
7	0,9890	1,6900	2,8330	6,3460	12,0170	14,0670	16,0130	18,4750	20,2780	24,3210
8	1,3440	2,1800	3,4900	7,3440	13,3620	15,5070	17,5350	20,0900	21,9550	26,1240
9	1,7350	2,7000	4,1680	8,3430	14,6840	16,9190	19,0230	21,6660	23,5890	27,8770
10	2,1560	3,2470	4,8650	9,3420	15,9870	18,3070	20,4830	23,2090	25,1880	29,5880
11	2,6030	3,8160	5,5780	10,3410	17,2750	19,6750	21,9200	24,7250	26,7570	31,2640
12	3,0740	4,4040	6,3040	11,3400	18,5490	21,0260	23,3370	26,2170	28,3000	32,9090
13	3,5650	5,0090	7,0410	12,3400	19,8120	22,3620	24,7360	27,6880	29,8190	34,5270
14	4,0750	5,6290	7,7900	13,3390	21,0640	23,6850	26,1190	29,1410	31,3190	36,1240
15	4,6010	6,2620	8,5470	14,3390	22,3070	24,9960	27,4880	30,5780	32,8010	37,6980
16	5,1420	6,9080	9,3120	15,3380	23,5420	26,2960	28,8450	32,0000	34,2670	39,2520

17	5,6970	7,5640	10,0850	16,3380	24,7690	27,5870	30,1910	33,4090	35,7180	40,7910
18	6,2650	8,2310	10,8650	17,3380	25,9890	28,8690	31,5260	34,8050	37,1560	42,3120
19	6,8440	8,9070	11,6510	18,3380	27,2040	30,1440	32,8520	36,1910	38,5820	43,8190
20	7,4340	9,5910	12,4430	19,3370	28,4120	31,4100	34,1700	37,5660	39,9970	45,3140
21	8,0340	10,2830	13,2400	20,3370	29,6150	32,6710	35,4790	38,9320	41,4010	46,7960
22	8,6430	10,9820	14,0410	21,3370	30,8130	33,9240	36,7810	40,2890	42,7960	48,2680
23	9,2600	11,6890	14,8480	22,3370	32,0070	35,1720	38,0760	41,6380	44,1810	49,7280
24	9,8860	12,4010	15,6590	23,3370	33,1960	36,4150	39,3640	42,9800	45,5580	51,1790
25	10,5200	13,1200	16,4730	24,3370	34,3820	37,6520	40,6460	44,3140	46,9280	52,6190
26	11,1600	13,8440	17,2920	25,3360	35,5630	38,8850	41,9230	45,6420	48,2900	54,0510
27	11,8080	14,5730	18,1140	26,3360	36,7410	40,1130	43,1950	46,9630	49,6450	55,4750
28	12,4610	15,3080	18,9390	27,3360	37,9160	41,3370	44,4610	48,2780	50,9940	56,8920
29	13,1210	16,0470	19,7680	28,3360	39,0870	42,5570	45,7220	49,5880	52,3350	58,3010
30	13,7870	16,7910	20,5990	29,3360	40,2560	43,7730	46,9790	50,8920	53,6720	59,7020
31	14,4580	17,5390	21,4340	30,3360	41,4220	44,9850	48,2320	52,1910	55,0020	61,0980
32	15,1340	18,2910	22,2710	31,3360	42,5850	46,1940	49,4800	53,4860	56,3280	62,4870
33	15,8150	19,0470	23,1100	32,3360	43,7450	47,4000	50,7250	54,7750	57,6480	63,8690
34	16,5010	19,8060	23,9520	33,3360	44,9030	48,6020	51,9660	56,0610	58,9640	65,2470
35	17,1920	20,5690	24,7970	34,3360	46,0590	49,8020	53,2030	57,3420	60,2750	66,6190
36	17,8870	21,3360	25,6430	35,3360	47,2120	50,9980	54,4370	58,6190	61,5810	67,9850
37	18,5860	22,1060	26,4920	36,3360	48,3630	52,1920	55,6680	59,8930	62,8830	69,3480
38	19,2890	22,8780	27,3430	37,3350	49,5130	53,3840	56,8950	61,1620	64,1810	70,7040
A	**0,995**	**0,975**	**0,9**	**0,5**	**0,1**	**0,05**	**0,025**	**0,01**	**0,005**	**0,001**
g.l. 39	19,9960	23,6540	28,1960	38,3350	50,6600	54,5720	58,1200	62,4280	65,4750	72,0550
40	20,7070	24,4330	29,0510	39,3350	51,8050	55,7580	59,3420	63,6910	66,7660	73,4030
41	21,4210	25,2150	29,9070	40,3350	52,9490	56,9420	60,5610	64,9500	68,0530	74,7440
42	22,1380	25,9990	30,7650	41,3350	54,0900	58,1240	61,7770	66,2060	69,3360	76,0840
43	22,8600	26,7850	31,6250	42,3350	55,2300	59,3040	62,9900	67,4590	70,6160	77,4180
44	23,5840	27,5750	32,4870	43,3350	56,3690	60,4810	64,2010	68,7100	71,8920	78,7490
45	24,3110	28,3660	33,3500	44,3350	57,5050	61,6560	65,4100	69,9570	73,1660	80,0780
46	25,0410	29,1600	34,2150	45,3350	58,6410	62,8300	66,6160	71,2010	74,4370	81,4000
47	25,7750	29,9560	35,0810	46,3350	59,7740	64,0010	67,8210	72,4430	75,7040	82,7200
48	26,5110	30,7540	35,9490	47,3350	60,9070	65,1710	69,0230	73,6830	76,9690	84,0370
49	27,2490	31,5550	36,8180	48,3350	62,0380	66,3390	70,2220	74,9190	78,2310	85,3500
50	27,9910	32,3570	37,6890	49,3350	63,1670	67,5050	71,4200	76,1540	79,4900	86,6600
51	28,7350	33,1620	38,5600	50,3350	64,2950	68,6690	72,6160	77,3860	80,7460	87,9670
52	29,4810	33,9680	39,4330	51,3350	65,4220	69,8320	73,8100	78,6160	82,0010	89,2720
53	30,2300	34,7760	40,3080	52,3350	66,5480	70,9930	75,0020	79,8430	83,2530	90,5730
54	30,9810	35,5860	41,1830	53,3350	67,6730	72,1530	76,1920	81,0690	84,5020	91,8710
55	31,7350	36,3980	42,0600	54,3350	68,7960	73,3110	77,3800	82,2920	85,7490	93,1670
56	32,4910	37,2120	42,9370	55,3350	69,9190	74,4680	78,5670	83,5140	86,9940	94,4620
57	33,2480	38,0270	43,8160	56,3350	71,0400	75,6240	79,7520	84,7330	88,2370	95,7500
58	34,0080	38,8440	44,6960	57,3350	72,1600	76,7780	80,9360	85,9500	89,4770	97,0380
59	34,7700	39,6620	45,5770	58,3350	73,2790	77,9300	82,1170	87,1660	90,7150	98,3240
60	35,5340	40,4820	46,4590	59,3350	74,3970	79,0820	83,2980	88,3790	91,9520	99,6080
61	36,3000	41,3030	47,3420	60,3350	75,5140	80,2320	84,4760	89,5910	93,1860	100,8870
62	37,0680	42,1260	48,2260	61,3350	76,6300	81,3810	85,6540	90,8020	94,4190	102,1650
63	37,8380	42,9500	49,1110	62,3350	77,7450	82,5290	86,8300	92,0100	95,6490	103,4420
64	38,6100	43,7760	49,9960	63,3350	78,8600	83,6750	88,0040	93,2170	96,8780	104,7170
65	39,3830	44,6030	50,8830	64,3350	79,9730	84,8210	89,1770	94,4220	98,1050	105,9880
66	40,1580	45,4310	51,7700	65,3350	81,0850	85,9650	90,3490	95,6260	99,3300	107,2570
67	40,9350	46,2610	52,6590	66,3350	82,1970	87,1080	91,5190	96,8280	100,5540	108,5250

68	41,7140	47,0920	53,5480	67,3350	83,3080	88,2500	92,6880	98,0280	101,7760	109,7930
69	42,4930	47,9240	54,4380	68,3340	84,4180	89,3910	93,8560	99,2270	102,9960	111,0550
70	43,2750	48,7580	55,3290	69,3340	85,5270	90,5310	95,0230	100,4250	104,2150	112,3170
71	44,0580	49,5920	56,2210	70,3340	86,6350	91,6700	96,1890	101,6210	105,4320	113,5770
72	44,8430	50,4280	57,1130	71,3340	87,7430	92,8080	97,3530	102,8160	106,6470	114,8340
73	45,6290	51,2650	58,0060	72,3340	88,8500	93,9450	98,5160	104,0100	107,8620	116,0920
74	46,4170	52,1030	58,9000	73,3340	89,9560	95,0810	99,6780	105,2020	109,0740	117,3470
75	47,2060	52,9420	59,7950	74,3340	91,0610	96,2170	100,8390	106,3930	110,2850	118,5990
76	47,9960	53,7820	60,6900	75,3340	92,1660	97,3510	101,9990	107,5820	111,4950	119,8500
77	48,7880	54,6230	61,5860	76,3340	93,2700	98,4840	103,1580	108,7710	112,7040	121,1010
78	49,5810	55,4660	62,4830	77,3340	94,3740	99,6170	104,3160	109,9580	113,9110	122,3470
79	50,3760	56,3090	63,3800	78,3340	95,4760	100,7490	105,4730	111,1440	115,1160	123,5950
80	51,1720	57,1530	64,2780	79,3340	96,5780	101,8790	106,6290	112,3290	116,3210	124,8390
81	51,9690	57,9980	65,1760	80,3340	97,6800	103,0100	107,7830	113,5120	117,5240	126,0840
82	52,7670	58,8450	66,0760	81,3340	98,7800	104,1390	108,9370	114,6950	118,7260	127,3240
83	53,5670	59,6920	66,9760	82,3340	99,8800	105,2670	110,0900	115,8760	119,9270	128,5650
84	54,3680	60,5400	67,8760	83,3340	100,9800	106,3950	111,2420	117,0570	121,1260	129,8020
85	55,1700	61,3890	68,7770	84,3340	102,0790	107,5220	112,3930	118,2360	122,3240	131,0430
86	55,9730	62,2390	69,6790	85,3340	103,1770	108,6480	113,5440	119,4140	123,5220	132,2760
87	56,7770	63,0890	70,5810	86,3340	104,2750	109,7730	114,6930	120,5910	124,7180	133,5110
88	57,5820	63,9410	71,4840	87,3340	105,3720	110,8980	115,8410	121,7670	125,9120	134,7460

A	0,995	0,975	0,9	0,5	0,1	0,05	0,025	0,01	0,005	0,001
g.l. 89	58,3890	64,7930	72,3870	88,3340	106,4690	112,0220	116,9890	122,9420	127,1060	135,9770
90	59,1960	65,6470	73,2910	89,3340	107,5650	113,1450	118,1360	124,1160	128,2990	137,2080
91	60,0050	66,5010	74,1960	90,3340	108,6610	114,2680	119,2820	125,2890	129,4900	138,4370
92	60,8150	67,3560	75,1000	91,3340	109,7560	115,3900	120,4270	126,4620	130,6810	139,6670
93	61,6250	68,2110	76,0060	92,3340	110,8500	116,5110	121,5710	127,6330	131,8710	140,8940
94	62,4370	69,0680	76,9120	93,3340	111,9440	117,6320	122,7150	128,8030	133,0590	142,1180
95	63,2500	69,9250	77,8180	94,3340	113,0380	118,7520	123,8580	129,9730	134,2470	143,3430
96	64,0630	70,7830	78,7250	95,3340	114,1310	119,8710	125,0000	131,1410	135,4330	144,5660
97	64,8780	71,6420	79,6330	96,3340	115,2230	120,9900	126,1410	132,3090	136,6190	145,7890
98	65,6930	72,5010	80,5410	97,3340	116,3150	122,1080	127,2820	133,4760	137,8030	147,0090
99	66,5100	73,3610	81,4490	98,3340	117,4070	123,2250	128,4220	134,6410	138,9870	148,2300
100	67,3280	74,2220	82,3580	99,3340	118,4980	124,3420	129,5610	135,8070	140,1700	149,4490
101	68,1460	75,0840	83,2670	100,3340	119,5890	125,4580	130,7000	136,9710	141,3510	150,6660
102	68,9650	75,9460	84,1770	101,3340	120,6790	126,5740	131,8380	138,1340	142,5320	151,8840
103	69,7850	76,8090	85,0870	102,3340	121,7690	127,6890	132,9750	139,2970	143,7120	153,1000
104	70,6060	77,6720	85,9980	103,3340	122,8580	128,8040	134,1110	140,4590	144,8910	154,3140
105	71,4280	78,5360	86,9090	104,3340	123,9470	129,9180	135,2470	141,6200	146,0690	155,5270
106	72,2510	79,4010	87,8210	105,3340	125,0350	131,0310	136,3820	142,7800	147,2470	156,7400
107	73,0740	80,2670	88,7330	106,3340	126,1230	132,1440	137,5170	143,9400	148,4240	157,9500
108	73,8990	81,1330	89,6450	107,3340	127,2110	133,2570	138,6510	145,0990	149,5990	159,1640
109	74,7240	82,0000	90,5580	108,3340	128,2980	134,3690	139,7840	146,2570	150,7740	160,3710
110	75,5500	82,8670	91,4710	109,3340	129,3850	135,4800	140,9160	147,4140	151,9480	161,5820
111	76,3770	83,7350	92,3850	110,3340	130,4720	136,5910	142,0490	148,5710	153,1210	162,7870
112	77,2040	84,6040	93,2990	111,3340	131,5580	137,7010	143,1800	149,7270	154,2950	163,9950
113	78,0330	85,4730	94,2130	112,3340	132,6430	138,8110	144,3110	150,8820	155,4670	165,2020
114	78,8620	86,3420	95,1280	113,3340	133,7290	139,9210	145,4410	152,0370	156,6370	166,4060
115	79,6910	87,2130	96,0430	114,3340	134,8130	141,0300	146,5710	153,1900	157,8080	167,6090
116	80,5220	88,0840	96,9580	115,3340	135,8980	142,1380	147,7000	154,3440	158,9770	168,8130
117	81,3530	88,9550	97,8740	116,3340	136,9820	143,2460	148,8290	155,4970	160,1460	170,0140
118	82,1850	89,8270	98,7900	117,3340	138,0660	144,3540	149,9570	156,6480	161,3140	171,2160

119	83,0180	90,7000	99,7070	118,3340	139,1490	145,4610	151,0840	157,7990	162,4810	172,4170
120	83,8520	91,5730	100,6240	119,3340	140,2330	146,5670	152,2110	158,9500	163,6480	173,6180

12. Bibliografia

BUSSAB W. O., MORETTIN P. A. *Estatística Básica*. 5ª Edição. São Paulo: Editora Saraiva, 2002.

CHARNET R, FREIRE C. A. D., CHARNET E. M. R., BONVINO H. *Análise de Modelos de Regressão Linear (com Aplicações)*. São Paulo: Editora da Unicamp, 1999.

FARIAS A. A., CÉSAR C. C., SOARES J. F. *Introdução à Estatística*. 2ª Edição. Rio de Janeiro: Editora LTC, 2003.

HAZZAN S. *Fundamentos de Matemática Elementar: Combinatória, Probabilidade*, volume 5, 6ª edição. São Paulo: Editora Atual Editora Ltda 1996.

MEYER, P. L., *Probabilidade: Aplicações à Estatística*. 2ª Edição. Rio de Janeiro: LTC - Livros Técnicos e Científicos Editora, 1983.

MORETTIN L. G. *Estatística básica*, volume I, 7ª edição. São Paulo: Editora Markon Books do Brasil Editora Ltda 1999.

SOARES J. F., FARIAS A. A., CESAR C. C. *Introdução à Estatística*, Rio de Janeiro: Editora Guanabara Koogan S.A., 1991.

WERKEMA M. C. C., AGUIAR S. *Série Ferramentas da Qualidade: Análise de Regressão: Como Entender o Relacionamento entre as Variáveis de um Processo.*Volume 7, Minas Gerais: Editora Líttera Maciel Ltda, 1996.