

Probabilidade e Estatística Aplicado à Engenharia



Tales Jesus Fernandes

PREFÁCIO

Esta apostila foi gerada com base nas minhas notas de aula da disciplina de Estatística Aplicada à Engenharia (68 horas-aula) lecionada para os cursos de graduação na Universidade Federal de Lavras.

Os exemplos e exercícios foram retirados de listas disponíveis na internet bem como de livros específicos sobre o tema, os quais aparecem listados nas referências ao final.

Espero que este material ajude no acompanhamento de disciplinas de estatística com carga horária e conteúdos semelhantes. Cabe ressaltar que ele não cobre todo o conteúdo da parte básica de estatística. A intenção é oferecer um material resumido com enfoque nas principais técnicas estatísticas na área básica ou introdutória.

Quaisquer dúvidas e ou sugestões, por favor me escreva.

BONS ESTUDOS!!!

“Não conseguimos encontrar respostas para todos os nossos problemas. As que encontramos apenas nos levam a formular novas questões. De certa maneira, sentimo-nos tão confusos quanto antes, mas agora acreditamos que estamos confusos em um nível mais alto e sobre coisas mais importantes.”

Aviso colocado na porta do Departamento de Matemática da Universidade de Tromso, na Noruega.

SUMÁRIO

1	Introdução	6
1.1	Definições Iniciais	6
2	Estatística Descritiva	8
2.1	Técnicas de amostragem	8
2.2	Tipos de Variável	10
2.3	Tabela de distribuição de frequências - TDF	10
2.4	Representação Gráfica	12
2.5	LISTA DE EXERCÍCIOS 1: Coleta, organização e apresentação de dados	14
2.6	Medidas de Posição	19
2.7	Medidas de dispersão	21
2.8	Assimetria	23
2.9	LISTA DE EXERCÍCIOS 2: Medidas de Posição e de Dispersão	25
3	Probabilidade	29
3.1	Noções de Probabilidade	29
3.2	Teorema de Bayes	32
3.3	Diagrama de Árvore	35
3.4	LISTA DE EXERCÍCIOS 3: Noções de probabilidade	36
4	Variáveis Aleatórias Discretas	40
4.1	Distribuição de Probabilidades	40
4.2	Função de Distribuição Acumulada	41
4.3	Distribuição Binomial	43
4.4	Distribuição de Poisson	44
4.5	Relação entre as distribuições Binomial e Poisson.	45
4.6	LISTA DE EXERCÍCIOS 4: Distribuições de Probabilidade de Variáveis Aleatórias Discretas	47
5	Variáveis Aleatórias Contínuas	50
5.1	Função de Distribuição Acumulada	51
5.2	A distribuição normal	52
5.3	A distribuição normal padrão	54
5.4	LISTA DE EXERCÍCIOS 5: Variáveis aleatórias contínuas e a distribuição normal	57
5.5	Aproximação de distribuições de probabilidade discretas à normal	61
5.5.1	Binomial → Normal	61
5.5.2	Poisson → Normal	63
5.6	A distribuição Gama	65
5.7	A distribuição exponencial	66
5.8	A distribuição Weibull	68
5.9	LISTA DE EXERCÍCIOS 6: Distribuições de Probabilidade Contínuas: Exponencial, Gama e Weibull.	69
5.10	Amostragem da distribuição normal: distribuições amostrais	71
5.10.1	A distribuição <i>t</i> de Student	72
5.10.2	A distribuição χ^2	73
5.10.3	A distribuição F de Snedecor	74
5.10.4	Distribuição amostral de \bar{X}	75

5.10.5 Distribuição amostral de \hat{p}	77
5.10.6 Distribuição amostral de S^2	78
5.10.7 Gráficos Q-Q plots	78
5.11 LISTA DE EXERCÍCIOS 7: Distribuições amostrais e Teorema Central do Limite	80
6 Inferência Estatística	82
6.1 Estimação Pontual	82
6.2 Estimação Intervalar	82
6.2.1 IC para a média (μ)	82
6.2.2 IC para a proporção (p)	84
6.2.3 IC para a variância (σ^2)	86
6.3 Margem de Erro e dimensionamento de amostras	86
6.3.1 Exercícios	88
6.4 LISTA DE EXERCÍCIOS 8: Intervalos de Confiança, margem de erro e dimensionamento de amostras	90
6.5 Testes de Hipóteses	93
6.5.1 Teste de Hipóteses para a média com σ conhecido	95
6.5.2 Teste de Hipóteses para a média com σ desconhecido	95
6.5.3 Teste de Hipóteses para uma proporção, p	96
6.5.4 Teste de Hipóteses para duas proporções, p_1 e p_2	97
6.5.5 Teste de Hipóteses para a variância, σ^2 de uma população normal .	98
6.5.6 Teste de Hipóteses para a razão de variâncias, σ_1^2/σ_2^2 ;	99
6.5.7 Teste de Hipóteses para duas médias μ_1 e μ_2	100
6.6 Relação entre testes de hipóteses bilaterais e intervalos de confiança	104
6.7 Nível descritivo de um teste de hipóteses (p-valor)	106
6.8 Teste de Normalidade	108
6.8.1 O teste de Shapiro-Wilk	109
6.9 LISTA DE EXERCÍCIOS 9: Testes de hipóteses	110
7 Correlação e Regressão Linear Simples	115
7.1 Correlação	115
7.2 Regressão	116
7.2.1 Regressão Linear Simples	116
7.3 LISTA DE EXERCÍCIOS 10: Correlação e Regressão Linear Simples	119
8 Referências Bibliográficas	121

1 Introdução

Em algum momento, seja qual for a área de conhecimento, o profissional vai se deparar com situações na qual precisará tomar decisões e tirar conclusões com base nas informações presentes em um conjunto de dados.

Estatística é a ciência que nos ajuda a extrair estas informações dos dados, consequentemente, todas as áreas de estudo necessitam de uma análise estatística.

1.1 Definições Iniciais

Para compreender os conceitos envolvidos na disciplina de Estatística, existem três definições que precisam estar muito claras ao estudante. São elas:

- **População:** é o conjunto total de valores sobre o qual deseja-se conhecer alguma característica.
- **Amostra:** é um subconjunto da população.
- **Variável:** é uma característica da população, de interesse do pesquisador.

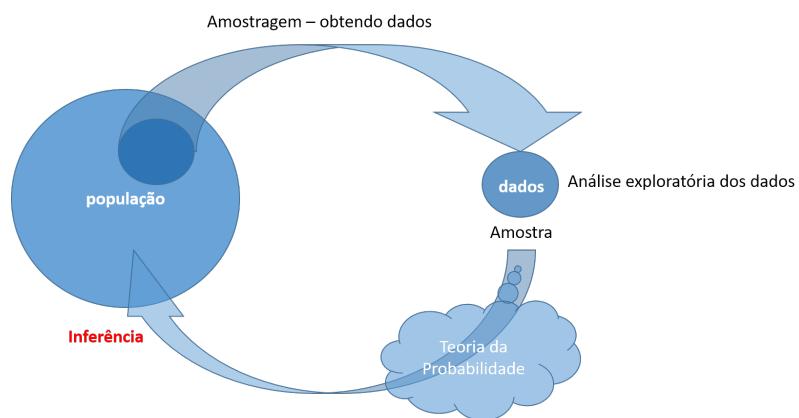


Figura 1: Esquema estatístico

O campo da estatística lida com a coleta, organização, apresentação, análise e uso dos dados para tomar decisões, resolver problemas e planejar produtos e processos. Em termos simples, estatística é a ciência dos dados. Uma definição mais formal para esta área de estudos é apresentada abaixo.

Definição: Estatística é a ciência e arte de extrair informações a partir de dados, com o objetivo de resolver problemas reais.

Especificamente na engenharia, as técnicas estatísticas podem ser uma ajuda poderosa no planejamento de novos produtos e sistemas, melhorando os projetos existentes e planejando, desenvolvendo e melhorando os processos de produção.

Os engenheiros são profissionais que resolvem problemas de interesse da sociedade, utilizando princípios científicos. Tais soluções são obtidas por meio do refinamento de produtos ou processos existentes ou pelo projeto de um novo produto. As etapas do método de engenharia, ou método científico são:

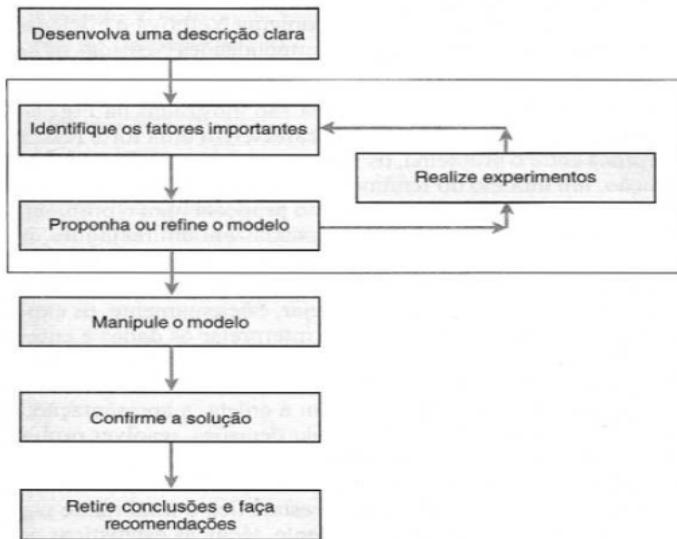


Figura 2: Ilustração do método científico para resolver problemas.

A estatística pode ser definida em duas partes:

- **Estatística Descritiva:** coleta e organização dos dados, resumo das informações, cálculo de probabilidades;
- **Inferência:** tirar conclusões, tomar decisões para a população com base na amostra.

2 Estatística Descritiva

2.1 Técnicas de amostragem

Em estatística, trabalhamos sempre com amostras, por não ser possível, ou ser exaustivo a análise de todos os elementos da população. A partir de agora, sempre que comentarmos sobre o tamanho de amostras será usado a letra “n”.

A finalidade de uma amostra é a de descrever, indiretamente, a população. Portanto, é necessário que as amostras coletadas guardem características as mais próximas possíveis da população. Em outras palavras devemos selecionar amostras que sejam representativas da população. A amostragem aleatória ou probabilística é a maneira mais correta de garantir esta representatividade.

Se a obtenção da amostra foi feita por algum tipo de sorteio, ela é chamada de amostra aleatória.

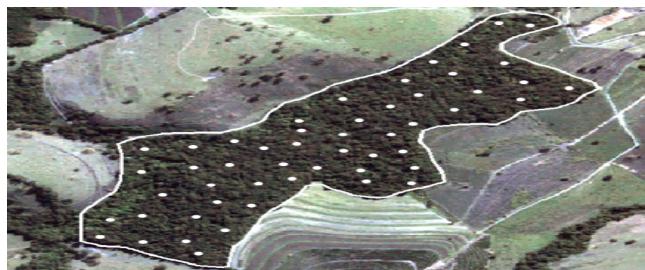
Existem diferentes maneiras de coletar a amostra, sendo que as principais são:

Amostragem Aleatória Simples - AAS

Deve ser realizada em populações estritamente homogêneas.

Consiste em sortear “n” elementos aleatórios de uma população de tamanho “N” de modo que todos tenham a mesma chance de ocorrer ($1/N$).

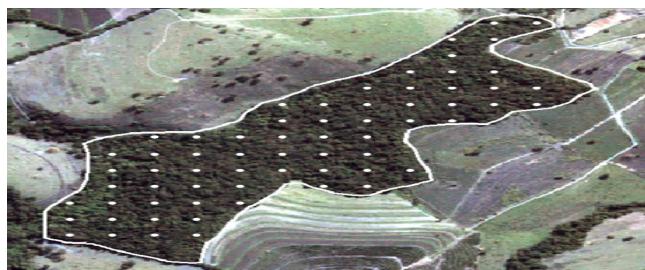
Este sorteio pode ser feito pelo computador, calculadora, papelzinho, etc...



Amostragem Aleatória Sistemática - AS

Utilizada em populações homogêneas que possuam indivíduos dispostos em uma ordem.

A única exigência é que os elementos estejam organizados em uma série. Sorteia-se apenas o 1º elemento e os demais são tomados a uma distância k do inicial. O valor de k pode ser definido como: $k = \frac{N}{n}$.



Amostragem Aleatória Estratificada - AAE

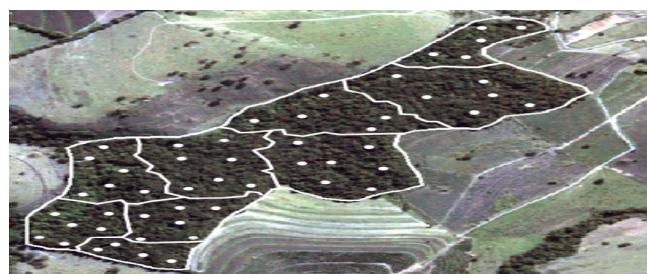
Utilizada quando a população é heterogênea, mas pode ser dividida em estratos homogêneos. Geralmente estes estratos são subdivisões bem particulares da população, grupos com características distintas.

Realiza-se então uma AAS dentro de cada estrato. Uma vez determinado o tamanho da amostra n , é necessário definir o tamanho da amostra em cada estrato (n_i), que deve ser proporcional ao tamanho do estrato (N_i). Determina-se este tamanho por:

$$n_i = \frac{N_i}{N},$$

desde que:

$$\sum n_i = n.$$



Amostragem Aleatória por Conglomerado - AAC

Deve ser utilizado em populações heterogêneas e consiste na divisão da população em grupos com as mesmas características da população (ao contrário dos estratos), ou seja, é feita divisão em subpopulações.

Não é necessário realizar amostragem em todos os conglomerados, uma vez que cada um deles devem reproduzir bem as características da população. A divisão em conglomerados não deve comprometer a representatividade, em razão da não-observação dos outros conglomerados.

A partir daí sorteia-se uma amostra (AAS, AS, AAE) dentro de alguns conglomerados. Seu principal objetivo é a economia de tempo e recursos.



2.2 Tipos de Variável

Uma vez definido como serão coletadas as amostras, o pesquisador deve se atentar ao tipo de variável que deseja estudar, pois naturalmente, cada uma tem suas peculiaridades e maneiras diferentes de serem abordadas.

As variáveis podem ser classificadas em:

- Variáveis Qualitativas
 - Nominais: não possuem uma ordem
 - Ordinais: passíveis de ordenação
- Variáveis Quantitativas
 - Discretas: números inteiros, provem de contagem.
 - Contínuas: números reais, provém de medição.

Exemplo: cor dos olhos, cor do cabelo, estado civil, escolaridade, quantidade de peças com defeito, número de interrupções, altura de alunos, peso de determinado componente em uma mistura, etc...

Para cada tipo de variável existem técnicas mais adequadas para resumir as informações.

2.3 Tabela de distribuição de frequências - TDF

É uma tabela que suas linhas trazem os dados organizados em classes ou categorias. Nas colunas são indicadas as frequências em cada classe (linha).

Frequência: medida que quantifica, contando, a ocorrência de variável em determinada classe. A frequência pode ser dividida em:

- absoluta - F_i
- relativa - f_{r_i}
- percentual - f_{p_i}

em que:

$$f_{r_i} = \frac{F_i}{n} \quad \text{e} \quad f_{p_i} = 100 * f_{r_i}$$

Dependendo do tipo da variável e dos objetivos, a TDF também pode receber colunas com as frequências acumuladas, que é a soma das frequências ocorridas até aquela classe.

TDF para Variáveis Qualitativas

No caso de variáveis Qualitativas a elaboração de uma tabela de frequências é simples, pois as classes são os próprios níveis na qual a variável esta dividia.

Exemplo: Considere os dados de um estudo na indústria alimentícia, no qual foram avaliados a preferência por iogurte de 145 clientes de grande supermercado. A variável em estudo é sabor de iogurte preferido com os seguintes níveis: SaladaF, Morango, Pêssego, Ameixa e Côco.

Tabela 1: Tabela de frequências para o tipo de iogurte preferido.

	F	fr	fp
saladaF	50	0,3448	34,48
morango	70	0,4828	48,28
pêssego	10	0,0690	6,90
ameixa	5	0,0345	3,45
côco	10	0,0690	6,90
Total	145	1	100

TDF para Variáveis Quantitativas com muitos níveis

Quando a variável é do tipo quantitativa contínua ou quantitativa discreta com muitos valores diferentes a divisão das classes não fica tão simples. Existem alguns algoritmos de modo que o principal deles apresenta os seguintes passos:

1º Organizar os dados em ordem crescente (rol).

2º Determinar a amplitude dos dados:

$$A \Rightarrow obs- < obs$$

3º Determinar o número de classes:

$$k = \sqrt{n} \text{ se } n \leq 100$$

$$k = 5 \log n \text{ se } n > 100$$

4º Determinar a amplitude de classes:

$$c = \frac{A}{k-1}$$

5º Determinar o limite inferior - LI da primeira classe:

$$LI_1 = < obs - \frac{c}{2}$$

6º Determinar o limite superior - LS da primeira classe:

$$LS_1 = LI_1 + c$$

7º Proceder de forma iterativa até a última classe: $LI_2 = LS_1$

$$LS_2 = LI_2 + c \quad \text{e assim por diante...}$$

2.4 Representação Gráfica

Uma boa maneira de realçar as características importantes em um conjunto de dados é através de um gráfico. Em qualquer boa análise estatística sempre aparecem gráficos.

Gráfico de Setores

É indicado para variáveis qualitativas. Deve ser feito com a frequência percentual (f_{p_i}).

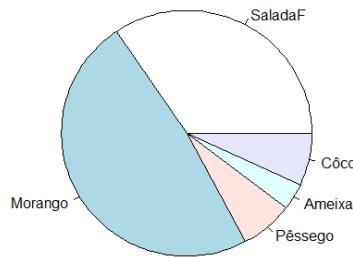


Figura 3: Gráfico de setores para a variável sabor de iogurte preferido.

Gráfico de Barras

É indicado para variáveis Qualitativas e Quantitativas Discretas. Pode ser feito tanto com a frequência absoluta (F_i), quanto com a relativa (f_{r_i}).

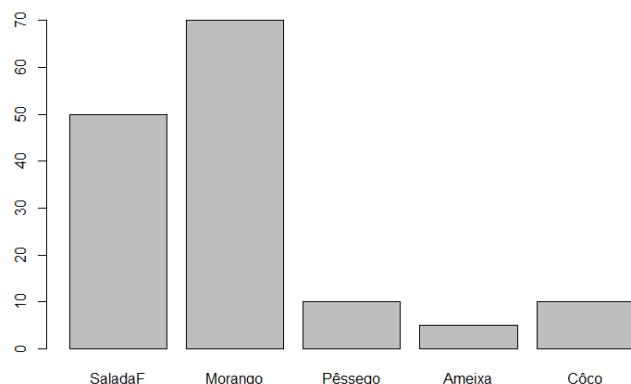


Figura 4: Gráfico de barras para a variável sabor de iogurte preferido.

Histograma

É indicado para variáveis Quantitativas Contínuas. Pode ser feito tanto com a frequência absoluta (F_i), quanto com a relativa (f_{r_i}).

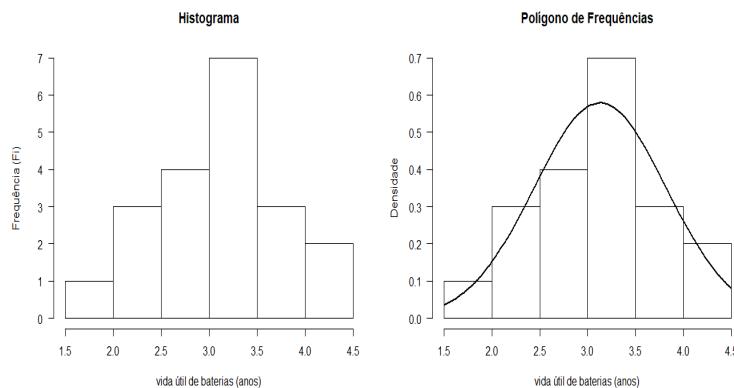


Figura 5: Histograma para a vida útil de baterias de automóveis.

Série Temporal

É indicado para variáveis Quantitativas Contínuas quando é observada ao longo do tempo.

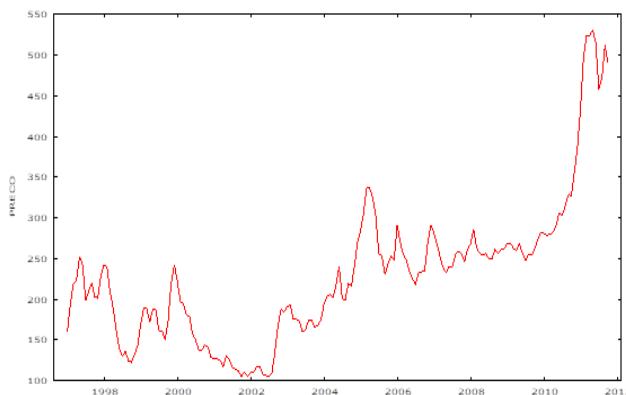


Figura 6: Gráfico do preço da saca de café entre 1997 e 2012.

Aplicativo em Shiny

Foi desenvolvido um aplicativo em “Shiny” para fazer gráficos de setores e gráfico de barras para variáveis qualitativas, bem como a tabela de frequências e o histograma para variáveis quantitativas contínuas. Podem ser acessados nos seguintes endereços:

<https://talesufla.shinyapps.io/GraficoBarraseSetores/>

<https://talesufla.shinyapps.io/TDFhist/>

Este é um servidor Shiny, que é um pacote do software estatístico R para interfaces gráficas interativas, sendo que o aplicativo fica disponível online. Basta salvar uma planilha do Excel com a extensão .csv ou no bloco de notas (.txt) e fazer o upload no aplicativo.

2.5 LISTA DE EXERCÍCIOS 1: Coleta, organização e apresentação de dados

1) A *Guerra das Colas* é o termo popular para a intensa competição entre Coca-Cola e Pepsi mostrada em suas campanhas de marketing. As campanhas têm estrelas do cinema e televisão, vídeos de rock, apoios de atletas e afirmações das preferências dos consumidores com base em testes de sabor. Como uma parte de uma campanha de marketing da Pepsi, suponha que 1000 consumidores de refrigerante sabor cola submetam-se a um teste cego de sabor (isto é, as marcas estão encobertas). Cada consumidor é questionado quanto à sua preferência em relação à marca A ou B.

- a. Descreva a população.
- b. Descreva a variável de interesse.
- c. Descreva a amostra.

2) Uma certa cadeia de fast-food possui 6289 unidades equipadas com serviço de drive-thru. Para atrair mais clientes para este serviço, a empresa está considerando oferecer um desconto de 50% àqueles que aguardarem mais de um certo tempo em minutos para receber seus pedidos. A empresa decidiu, então, estimar o tempo médio de espera em um drive-thru em Dallas, Texas, para estimar o limite de tempo a ser estipulado na campanha.

- a. Descreva a variável de interesse e classifique-a.
- b. Descreva a população.
- c. Com qual técnica de amostragem poderia ser coletada a amostra neste caso?

3) O Windows é um software produzido pela Microsoft Co. Na elaboração do Windows XP, a Microsoft telefonou para milhares de usuários da versão anterior e perguntou a eles como o produto poderia ser melhorado. Considere que as seguintes perguntas foram feitas aos clientes:

- a. Você sempre usa o Windows em sua casa?
- b. Qual é a sua idade?
- c. Os tutoriais e instruções que acompanham o Windows são úteis?
- d. Ao imprimir com o Windows, você sempre usa uma impressora a laser ou outro tipo de impressora?
- e. Se a velocidade do Windows pudesse ser alterada, qual das seguintes mudanças você preferiria: mais lento, inalterado, ou mais rápido?
- f. Quantas pessoas em sua casa usaram o Windows pelo menos uma vez?

Cada uma dessas perguntas define uma variável de interesse para a empresa. Classifique e diga qual seria o tipo de gráfico mais adequado para cada uma delas.

4) Todas as pontes de estradas nos Estados Unidos são inspecionadas periodicamente pela Federal Highway Administration (FHWA) para detectar deficiências estruturais. Os dados das inspeções da FHWA são compilados para o National Bridge Inventory (NBI). Algumas das quase 100 variáveis mantidas pelo NBI estão listadas abaixo. Classifique-as.

- a. Extensão máxima do vão (em pés).
- b. Número de veículos que a atravessam.
- c. Ponte com pedágio (sim ou não).
- d. Média diária de tráfego.
- e. Condição da pista (boa, regular ou sofrível).
- f. Extensão do retorno ou desvio (em milhas).
- g. Tipo de rota (federal, interestadual, estadual, regional ou municipal).

5) Considere os dados a seguir sobre os tipos de queixas de saúde (J = inflamação de articulações, F = fadiga, B = dor nas costas, M = fadiga muscular, C = tosse, N = irritação nasal/coriza, O = outros) feitas por agricultores. Construa uma tabela de frequências com frequência absoluta e frequência relativa sobre o tipo de queixas de saúde e desenhe um gráfico de setores.

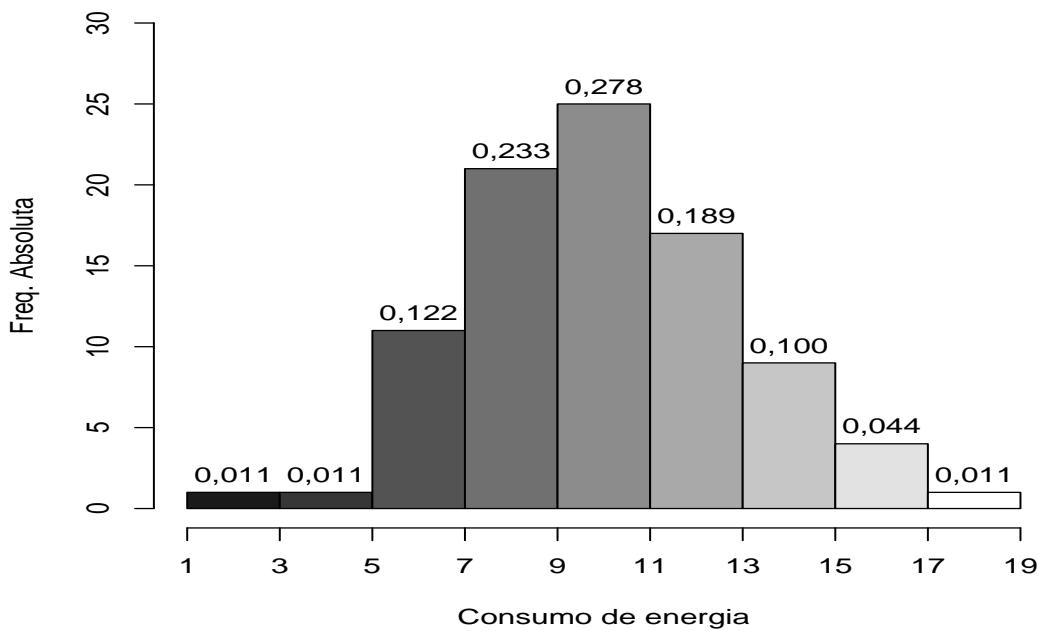
O	O	N	J	C	F	B	B	F	O	J	O	O	M
O	F	F	O	O	N	O	N	J	F	J	B	O	C
J	O	J	J	F	N	O	B	M	O	J	M	O	B
O	F	J	O	O	B	N	C	O	O	O	M	B	F
J	O	F	N										

6) O Conselho Norte-americano de Investigação de Riscos e Segurança Química é responsável por determinar as causas principais de acidentes industriais. Desde sua criação, em 1998, identificou 83 incidentes que foram causados por falhas em sistemas de gestão (*Process Safety Progress*, dez. 2004). A tabela a seguir dá um detalhamento das causas principais desses 83 incidentes.

Categoria de causa de sistema de gestão	Número de incidentes
Engenharia & Design	27
Procedimentos & Práticas	24
Gestão & Supervisão	22
Treinamento & Comunicação	10
Total	83

Construa um gráfico de barras com a frequência absoluta e um gráfico de setores com a frequência relativa para esses dados.

7) As empresas de energia necessitam de informações sobre o consumo de seus clientes para obterem previsões precisas da demanda. Investigadores da Wisconsin Power and Light determinaram que o consumo de energia (BTUs) dura um determinado período para uma amostra de 90 residências aquecidas a gás. O valor de consumo é apresentado no histograma a seguir. Suponha que os investigadores dessa companhia desejam visitar 8% das residências que apresentaram o consumo mais alto. Então, a partir de qual valor de consumo a residência deverá ser visitada, ou seja, qual é o quantil cuja área acima é 8%?



8) A corrosão das barras de aço da armação é um problema sério em estruturas de concreto localizadas em ambientes afetados por condições climáticas extremas. Por esse motivo, os pesquisadores têm investigado a utilização de barras de reforço feitas de material composto. Um estudo foi executado para desenvolver diretrizes sobre a aderência de barras plásticas reforçadas com fibra de vidro ao concreto (“Design Recommendations for Bond of GFRP Rebars to Concrete,” J. of Structural Engr., 1996, p. 247-254). Considere a Tabela de Distribuição de Frequências para a resistência da aderência medida. Preencha a tabela com os dados faltantes e faça um esboço do histograma de densidades.

Classe	freq.	Ponto médio ¹	Amplitude de classe (c)	f_r	Densidade ²
[2, 4)	9				
[4, 6)	15				
[6, 8)	5				
[8, 12)	9				
[12, 20)	8				
[20, 30)	2				

¹ Ponto médio da classe i : $(L_i + S_i)/2$, para $i = 1, 2, \dots, k$.

² Densidade da classe i : (f_{r_i}/c_i) , para $i = 1, 2, \dots, k$.

OBS: a densidade de classe é utilizada quando, por algum motivo, a TDF não apresenta classes de mesma amplitude, tornando-se uma medida mais coerente/realística nestes casos.

9) O número de partículas de contaminação de uma pastilha de silício antes de certo processo de limpeza foi determinado para cada pastilha em uma amostra de tamanho 100, resultando nas frequências a seguir:

Nº partículas	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Frequência	1	2	3	12	11	15	18	10	12	4	5	3	1	2	1

- a. Que proporção das pastilhas da amostra tinha ao menos uma partícula? Ao menos cinco partículas?
- b. Que proporção das pastilhas da amostra tinha entre cinco e 10 (inclusive) partículas? Estritamente entre cinco e 10 partículas?
- c. Desenhe um gráfico de barras usando a frequência relativa no eixo vertical.

10) O índice de céu claro foi determinado para o céu de Bagdá, compreendendo cada um dos 365 dias de um dado ano (“Contribution to the Study of the Solar Radiation Climate of the Baghdad Environment”, Solar Energy, 1990, p. 7-12). A tabela a seguir fornece os resultados.

Classe	Frequência
0,15I–0,25	8
0,25I–0,35	14
0,35I–0,45	28
0,45I–0,50	24
0,50I–0,55	39
0,55I–0,60	51
0,60I–0,65	106
0,65I–0,70	84
0,70I–0,75	11

- a. Determine as frequências relativa, percentual, acumulada percentual e desenhe o histograma.
- b. Dias nublados são aqueles com o índice de céu limpo inferior a 0,35. Em que porcentagem dos dias o céu esteve nublado?
- c. Dias de céu claro são aqueles para os quais o índice é no mínimo 0,65. Em que porcentagem dos dias o céu esteve limpo?

11) A transformação de valores de dados por meio de uma função matemática, como \sqrt{x} ou $1/x$, normalmente resulta em um conjunto de números com “melhores” propriedades estatísticas do que os dados originais. Como exemplo, o artigo “Time Lapse Cinematographic Analysis of BerylliumLung Fibroblast Interactions” (Environ. Research, 1983, p. 34-43) relatou os resultados de experimentos projetados para estudar o comportamento de algumas células que foram expostas ao berílio. Uma característica importante de tal célula individual é seu tempo de interdivisão (IDT). Os IDTs foram determinados para um grande número de células em condições de exposição (tratamento) e não-exposição (controle). Os autores do artigo usaram uma transformação logarítmica, isto é, valor transformado = log (valor original). Considere os seguintes dados representativos de IDT:

IDT	$\log_{10}(\text{IDT})$	IDT	$\log_{10}(\text{IDT})$	IDT	$\log_{10}(\text{IDT})$
28,1	1,45	60,1	1,78	21,0	1,32
31,2	1,49	23,7	1,37	22,3	1,35
13,7	1,14	18,6	1,27	15,5	1,19
46,0	1,66	21,4	1,33	36,3	1,56
25,8	1,41	26,6	1,42	19,1	1,28
16,8	1,23	26,2	1,42	38,4	1,58
34,8	1,54	32,0	1,51	72,8	1,86
62,3	1,79	43,5	1,64	48,9	1,69
28,0	1,45	17,4	1,24	21,4	1,33
17,9	1,25	38,8	1,59	20,7	1,32
19,5	1,29	30,6	1,49	57,3	1,76
21,1	1,32	55,6	1,75	40,9	1,61
31,9	1,50	25,5	1,41		
28,9	1,46	52,1	1,72		

Use os intervalos de classes $[10; 20)$, $[20; 30)$, ... para construir um histograma dos dados originais. Use os intervalos $[1, 1; 1, 2)$, $[1, 2; 1, 3)$, ... para fazer o mesmo para os dados transformados. Qual é o efeito da transformação?

12) Em que situação você recomendaria utilizar a amostragem aleatória simples ou a amostragem sistemática?

13) Qual é a principal diferença entre amostragem estratificada e amostragem por conglomerados? Apresente um exemplo de cada uma delas.

14) Uma indústria do setor alimentício tem $N = 3.414$ empregados subdivididos nos seguintes setores:

Setores (h)	Número de funcionários (N_h)
Administrativo	314
Transporte	948
Produção	1.451
Outros	701
Total	3.414

Para se estudar o nível salarial médio da empresa, resolveu-se fazer uma amostra de $n = 180$ funcionários. Você julga que a amostragem aleatória simples seria adequada para este caso? Se não for, o que você recomendaria? Dê todos os detalhes do dimensionamento da amostra.

15) Pesquise quais técnicas de amostragem os principais institutos de pesquisa de opinião (Ibope, Datafolha, VoxPopuli, MDA) utilizam em suas pesquisas.

2.6 Medidas de Posição

Também conhecidas como medidas de tendência central, são uma maneira mais formal de representar os dados observados, pois resumem as informações em um valor central, em torno do qual os dados se concentram.

Média

É a medida de posição mais comum e mais utilizada. Também chamada de média amostral, definida por:

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

Exemplo: Sejam os dados observados: 5, 1, 3, 5, 6.

$$\bar{X} = \frac{\sum_{i=1}^5 x_i}{5} = \frac{20}{5} = 4u$$

Propriedades da média:

- i) A soma dos desvios dos dados em relação a média é igual a zero.

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

Do exemplo acima: $1 - 3 - 1 + 1 + 2 = 0$

- ii) Se somarmos ou multiplicarmos um valor por todos os elementos da amostra, a média também fica somada ou multiplicada por esse valor.

Exemplo: Somando 1 nos dados do exemplo acima: 6, 2, 4, 6, 7

$$\bar{X} = \frac{\sum_{i=1}^5 x_i}{5} = \frac{25}{5} = 5 = 4 + 1$$

Multiplicando os dados por 2 temos: 10, 2, 6, 10, 12

$$\bar{X} = \frac{\sum_{i=1}^5 x_i}{5} = \frac{40}{5} = 8 = 2 * 4$$

Média Ponderada

É uma média calculada com base em pesos, dada por:

$$\bar{X}_p = \frac{\sum_{i=1}^n a_i x_i}{\sum_{i=1}^n a_i} = \frac{a_1 x_1 + a_2 x_2 + \dots + a_n x_n}{a_1 + a_2 + \dots + a_n}$$

No caso de \bar{X} todas as observações tem peso 1, isto é $a_i = 1$.

Moda

É o valor mais frequente em um conjunto de dados, o que mais repete. Única das medidas de posição que pode ser usada para variáveis qualitativas.

Exemplo: $M_o = 5u$.

Mediana

A mediana é o valor central da amostra com os dados ordenados. Primeiro você deve calcular a posição da mediana e depois observar o seu valor.

A posição da mediana em uma amostra é dada por: $P = \frac{n+1}{2}$. A mediana é o valor que está nesta posição.

No caso de n ser par, a mediana é a média dos dois valores centrais.

Exemplo: Nos dados do exemplo: 1, 3, 5, 5, 6

$$P = \frac{n+1}{2} = \frac{5+1}{2} = 3$$

Logo a mediana é o número que está na 3^a posição. $M_d = 5u$

Exemplo: Em um conjunto de dados com 6 valores: 2, 3, 5, 6, 7, 8

$$P = \frac{n+1}{2} = \frac{6+1}{2} = 3.5$$

Logo a mediana é o número que está entre a 3^a e a 4^a posição. $M_d = \frac{5+6}{2} = 5.5u$

Média Aparada

A média amostral e a mediana amostral são influenciadas por valores extremos de uma forma bastante diferente: muito para a média e nada para a mediana. Como o comportamento extremo das duas medidas é indesejável, podemos considerar algumas alternativas que não sejam tão sensíveis quanto \bar{X} e nem tão insensíveis como M_d .

Observe que \bar{X} e M_d são extremidades opostas da mesma “família” de medidas.

Uma média aparada é algo intermediário entre estas duas medidas. Uma média aparada de 10% por exemplo é calculada eliminando-se os 10% inferiores e os 10% superiores da amostra, obtendo-se então a média dos dados restantes.

Medidas de posição relativa

A mediana divide o conjunto de dados em 2 partes. Mas pode ser de interesse do pesquisador dividir em mais partes. Os quartis por exemplo dividem o conjunto de dados em 4 partes, de modo que o 2º quartil é a própria mediana. E assim surgem decis, percentis, etc...

2.7 Medidas de dispersão

As medidas de posição são muito úteis mas sozinhas não transmitem toda a informação presente no conjunto de dados.

$$\text{Exemplo 1: } 5, 5, 5, 5 \rightarrow \bar{X} = 5u$$

$$\text{Exemplo 2: } 1, 1, 9, 9 \rightarrow \bar{X} = 5u$$

$$\text{Exemplo 3: } 2, 3, 7, 8 \rightarrow \bar{X} = 5u$$

Veja que 3 amostras completamente diferentes geram uma mesma medida de posição. Assim, torna-se necessário acrescentar uma medida que represente a variabilidade dos dados.

Amplitude

É a medida de dispersão mais simples, é a diferença entre a maior e a menor observação da amostra.

$$\text{Exemplo 1: } A = 0$$

$$\text{Exemplo 2: } A = 8$$

$$\text{Exemplo 3: } A = 6$$

Variância

É a principal medida de dispersão. É obtida pela soma dos quadrados dos desvios em relação a média \bar{X} .

$$S^2 = \frac{\sum_{i=1}^n (x_i - \bar{X})^2}{n-1}$$

Exemplo:

$$S^2 = \frac{\sum_{i=1}^5 (x_i - \bar{X})^2}{5-1} = \frac{(5-4)^2 + (1-4)^2 + (3-4)^2 + (5-4)^2 + (6-4)^2}{4} = \frac{16u^2}{4} = 4u^2$$

Propriedades da variância:

- i) Ao somar uma constante em todos os dados a variância não se altera.

Por exemplo, somando 1 nos dados do exemplo inicial: 6, 2, 4, 6, 7

$$S^2 = \frac{(6 - 5)^2 + (2 - 5)^2 + (4 - 5)^2 + (6 - 5)^2 + (7 - 5)^2}{5 - 1} = \frac{16}{4} = 4u^2$$

- ii) Se multiplicarmos um valor por todos os elementos da amostra, a variância fica multiplicada por esse valor ao quadrado.

Por exemplo, multiplicando por 2 os dados do exemplo inicial: 10, 2, 6, 10, 12

$$S^2 = \frac{(10 - 8)^2 + (2 - 8)^2 + (6 - 8)^2 + (10 - 8)^2 + (12 - 8)^2}{5 - 1} = \frac{64}{4} = 16u^2 = 2^2 * 4u^2$$

Desvio padrão

É utilizado para eliminar o inconveniente da unidade ao quadrado. Calculado pela raiz quadrada da variância.

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

Exemplo: $S = \sqrt{4} = 2u$

No entanto, este fica preso a unidade e depende da média.

Coeficiente de variação

Expressa a variabilidade dos dados sem o efeito da média e da unidade de medida. Logo pode ser utilizado para comparar diferentes conjuntos de dados. Dado por:

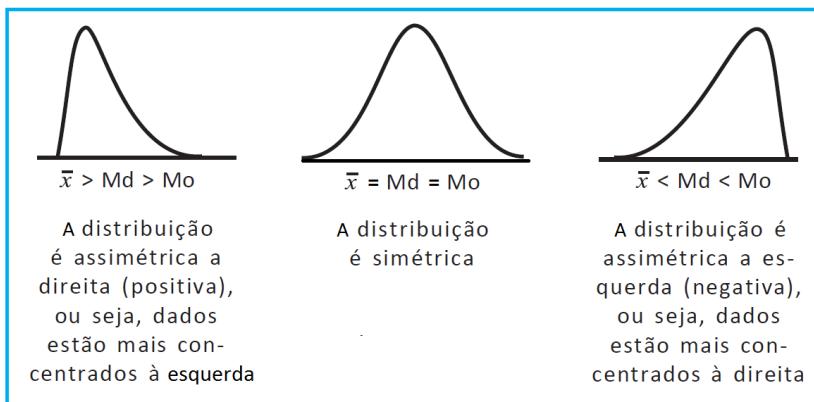
$$CV = \frac{S * 100}{\bar{X}}$$

Exemplo: $CV = \frac{2u}{4u} * 100 = 50\%$

2.8 Assimetria

As medidas de assimetria possibilitam analisar uma distribuição de acordo com as relações entre seus valores de moda, média e mediana, quando observadas graficamente.

Uma distribuição é dita simétrica quando apresenta o mesmo valor para a moda, a média e a mediana. Quando esta igualdade não acontece, temos uma distribuição assimétrica. Confira na figura abaixo as três situações possíveis:



Existem algumas maneiras de calcular a intensidade da assimetria presente em uma distribuição, geralmente estes cálculos são baseados na diferença entre a média e a mediana, sendo que o mais utilizado é o coeficiente de assimetria de Pearson, dado por:

$$As = \frac{3(\bar{X} - M_d)}{S}$$

E esta assimetria é classificada em relação a sua intensidade pelo seguinte critério:

- $|As| < 0,15 \rightarrow$ Assimetria Pequena
- $0,15 < |As| < 1 \rightarrow$ Assimetria Moderada
- $|As| > 1 \rightarrow$ Assimetria Elevada ou Forte

Curiosidade:

Quando a distribuição é simétrica (ou em forma de sino) existe uma regra conhecida como “regra empírica”, a qual pode ser explicada da seguinte maneira:

- i) Aproximadamente 68% das observações estarão distantes menos de um desvio padrão da média, isto é, dentro do intervalo $(\bar{X} - S; \bar{X} + S)$, ou para populações $(\mu - \sigma; \mu + \sigma)$;
- ii) Aproximadamente 95% das observações estarão distantes menos de um dois desvios padrão da média, isto é, dentro do intervalo $(\bar{X} - 2S; \bar{X} + 2S)$, ou para populações $(\mu - 2\sigma; \mu + 2\sigma)$;
- iii) Aproximadamente 99,7% das observações estarão distantes menos de um três desvios padrão da média, isto é, dentro do intervalo $(\bar{X} - 3S; \bar{X} + 3S)$, ou para populações $(\mu - 3\sigma; \mu + 3\sigma)$.

Boxplot

Atualmente um resumo esquemático denominado *Boxplot* vem sendo muito utilizado para descrever as principais características de um conjunto de dados. O *boxplot* é um gráfico no qual podemos identificar várias informações sobre os dados tais como: centro, dispersão, quartis, simetria, amplitude e a presença de possíveis “outliers”.

Outlier é um valor atípico no conjunto de dados, cujo valor destoa dos demais, ou seja, está distante da maioria dos dados.

Um *boxplot* é construído da seguinte maneira:

- Encontre os quartis da amostra. Sendo que a posição de cada um dos quartis pode ser calculada pela seguinte expressão:

$$PQ_i = \frac{i(n + 1)}{4}$$

em que $i=1,2,3$.

Após ser obtida a posição dos quartis, identifique-os na amostra, procedendo como na obtenção da mediana.

- Trace um retângulo de modo que uma das extremidades esteja sobre o quartil inferior e a outra sobre o quartil superior da amostra.
- Identifique o 2º quartil (mediana), que deverá estar localizado dentro do retângulo.
- Calcule a diferença interquartílica, isto é, o maior menos o menor: $DQ = Q_3 - Q_1$
- Agora calcule: $Q_1 - 1.5DQ$ e $Q_3 + 1.5DQ$.

Trace uma linha começando das extremidades do retângulo até cada um dos pontos encontrados.

Pronto, este é o seu gráfico *boxplot*. Se existir algum ponto além da linha tracejada no último passo então este ponto é um potencial *outlier*.

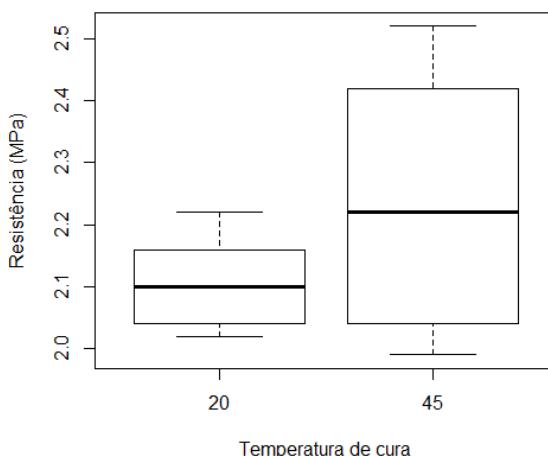


Figura 7: *Boxplot* da resistência à tensão da borracha nas temperaturas de cura 20°C e 45°C.

2.9 LISTA DE EXERCÍCIOS 2: Medidas de Posição e de Dispersão

1- Os dados a seguir são referentes ao tempo de oxidação-indução (em minutos) para diversos óleos comerciais. Calcule a média, a mediana, a moda, a amplitude, a variância e o desvio padrão para o tempo de oxidação destes óleos. Calcule os três quartis desta amostra e interprete-os.

87	103	145	160	180	195	132	145	211	105
145	153	152	138	87	99	93	119	129	

2- Os valores de pressão sanguínea, em mmHg, de nove indivíduos selecionados aleatoriamente

foram: 118,6 127,4 138,4 130,0 113,7 122,0 108,3 131,5 133,2

a) Qual a mediana e a média dos valores de pressão sanguínea?

b) Suponha que por um erro de anotação a pressão sanguínea do primeiro indivíduo seja 116,8 em vez de 118,6 (pequena alteração em um único valor). Calcule novamente a média e a mediana. O que o resultado obtido diz sobre a sensibilidade da mediana e da média?

c) Em quanto a maior observação da amostra pode ser diminuída sem afetar o valor da mediana?

3- Os salários de atletas profissionais recebem muita atenção da mídia. As principais estrelas dos times, em qualquer modalidade, recebem salários multimilionários. Raramente se passa uma temporada sem que haja uma negociação entre uma ou mais associações de jogadores e os presidentes dos clubes por salários adicionais ou mais benefícios para todos os atletas em seus respectivos esportes.

a) Se uma associação de atletas quisesse embasar seu argumento para maiores salários à todos os atletas, qual medida de tendência central deveria utilizar? Por quê?

b) Para negar este aumento, qual medida de tendência central os donos de times deveriam utilizar? Por quê?

4- A U.S. Environmental Protection Agency (EPA) define um limite para a concentração de chumbo em água potável. A concentração é de 0,015 miligramas por litro (mg/l) de água. Sob as diretrizes da EPA, se 90% das amostras de um estudo do sistema de águas tiverem uma concentração menor que 0,015mg/l, a água será considerada segura para ingestão. Foi feito recentemente um estudo sobre a concentração de chumbo na água nas residências de determinada cidade. O 90º percentil da amostra tem uma concentração de 0,00372mg/l. Os consumidores desta cidade estão sob risco de beber água com concentrações de chumbo não seguras para a saúde? Explique.

5- A U.S. Energy Information Administration monitora todas as usinas de energia nuclear em operação nos Estados Unidos. A tabela a seguir lista o número de usinas ativas operando em 19 estados amostrados (no ano 2000).

Alab.	Ariz.	Calif.	Flor.	Geor.	Illi.	Kan.	Loui.	Massac.	Missi.
5	3	4	5	4	13	1	2	1	1
Wisco.	New Y.	North C.	Ohio	Penn.	South C.	Tenne.	Tex.	Verm.	
3	6	5	2	9	7	3	4	1	

- a)** Calcule a média, a mediana e a moda para estes dados;
- b)** Calcule a *média 10% aparada*. Comente as possíveis vantagens e desvantagens que uma *média aparada* tem sobre a média aritmética regular.
- c)** Faça o gráfico *boxplot* para estes dados e discuta os resultados.

6- Muitas indústrias, utilizam partes moldadas como parte do processo de produção. O encolhimento das peças é frequentemente um grande problema. Então, um molde é construído para uma peça maior do que o nominal para permitir o encolhimento. Em um processo de moldagem por injeção, sabe-se que o encolhimento é influenciado por diversos fatores e, entre eles, está a velocidade da injeção, em pés por segundo, e a temperatura do molde, em graus Celsius. Os dois conjuntos de dados a seguir mostram os resultados de um experimento no qual a velocidade de injeção foi mantida em dois níveis (digamos “baixo” e “alto”) e a temperatura de molde foi mantida constante no nível “alta”. O encolhimento é medido em centímetros $\times 10^4$.

baixo	72,68	72,62	72,58	72,48	73,37	72,55	72,42	72,84	72,58	71,92
alto	93,25	93,19	92,87	93,29	93,37	92,98	93,47	93,75	93,89	92,62

- a)** Calcule a média para cada um dos níveis de injeção e diga em qual local ocorre o maior encolhimento médio;
- b)** Calcule a variância e o desvio-padrão e diga qual nível apresenta encolhimento mais homogêneo;
- c)** Calcule o coeficiente de variação para os dois locais e interprete os resultados. A conclusão é a mesma do item (b)? Qual das duas conclusões seria mais adequada? Por que?

7- Foi realizado um estudo sobre os efeitos do tabagismo nos padrões de sono. A medida observada é o tempo, em minutos, que se leva para dormir. Os dados obtidos são:

Fumantes	69,3	56,0	22,1	47,6	53,2	48,1
	52,7	34,4	60,2	43,8	23,2	13,8
Não-Fumantes	28,6	25,1	26,4	34,9	29,8	28,4
	38,5	30,2	30,6	31,8	21,1	13,9

- a)** Calcule as medidas de posição e dispersão para cada um dos grupos. Com base nos resultados obtidos diga qual dos grupos tende a apresentar distribuição mais simétrica (em forma de sino);
- b)** Comente o tipo de impacto que o tabagismo aparenta ter no tempo que se leva para dormir;
- c)** Compare adequadamente a variabilidade do tempo para dormir entre os dois grupos;
- d)** Construa o gráfico *boxplot*, para cada um dos grupos. Comente os possíveis motivos de dois valores muito próximos (13.8 e 13.9) ser considerado *outlier* em um grupo e no outro não.

8- Os seguintes dados são as medidas dos diâmetros de 35 cabeças de rebites em 1/100 de polegadas:

6,72	6,77	6,82	6,70	6,78	6,70	6,62	6,65	6,66
6,66	6,64	6,76	6,73	6,80	6,72	6,76	6,76	6,68
6,66	6,62	6,72	6,76	6,70	6,78	6,76	6,67	6,70
6,72	6,74	6,81	6,79	6,78	6,66	6,76	6,72	

- a)** Calcule a média, moda, mediana e desvio padrão amostrais;
- b)** Calcule a porcentagem de observações que estão entre $\bar{X} - S$ e $\bar{X} + S$. Qual a porcentagem de observações está entre $\bar{X} - 2S$ e $\bar{X} + 2S$? E entre $\bar{X} - 3S$ e $\bar{X} + 3S$?
- c)** Com base nos resultados da letra **a)** e da letra **b)** comente se há, ou não, indícios de que a amostra veio de uma população que apresenta distribuição simétrica.
- d)** Faça o gráfico *boxplot* para estes dados e veja se indica uma distribuição simétrica. DICA: utilize o aplicativo disponibilizado no link <https://talesufla.shinyapps.io/HistBox/>, basta digitar os dados em uma coluna do bloco de notas.
- e)** Utilizando o aplicativo disponibilizado no link <https://talesufla.shinyapps.io/TDFhist/>, faça uma tabela de distribuição de frequências e o histograma. Veja se histograma confirma as respostas anteriores sobre a simetria da distribuição dos dados.

9- Quando os dados originais se perderam ou foram destruídos, mas a tabela de distribuição de frequência foi preservada, as medidas de posição e dispersão ainda podem ser obtidas. Suponha que para cada um dos p valores distintos de x , digamos x_1, x_2, \dots, x_p , a frequência absoluta seja F_i (ou f_a), tal que $i = 1, 2, \dots, p$. Então a média e a variância podem ser calculadas por:

$$\bar{x} = \frac{\sum_{i=1}^p F_i x_i}{n} \quad s^2 = \frac{\sum_{i=1}^p F_i x_i^2 - \frac{1}{n} \left(\sum_{i=1}^p F_i x_i \right)^2}{n-1}$$

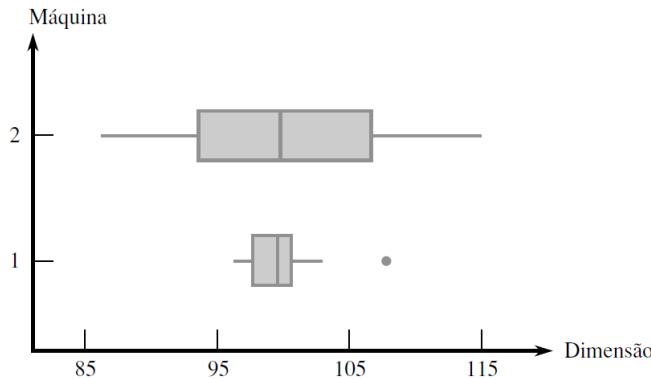
OBS: Quando a variável em estudo é quantitativa contínua e os dados estão agrupados em uma tabela de distribuição de frequências dividida em classes (ou intervalos), então o valor de x_i nas fórmulas acima pode ser substituído pelo ponto médio da respectiva classe.

Os dados abaixo referem-se ao número de empresas falidas por ano observadas em Varginha-MG.

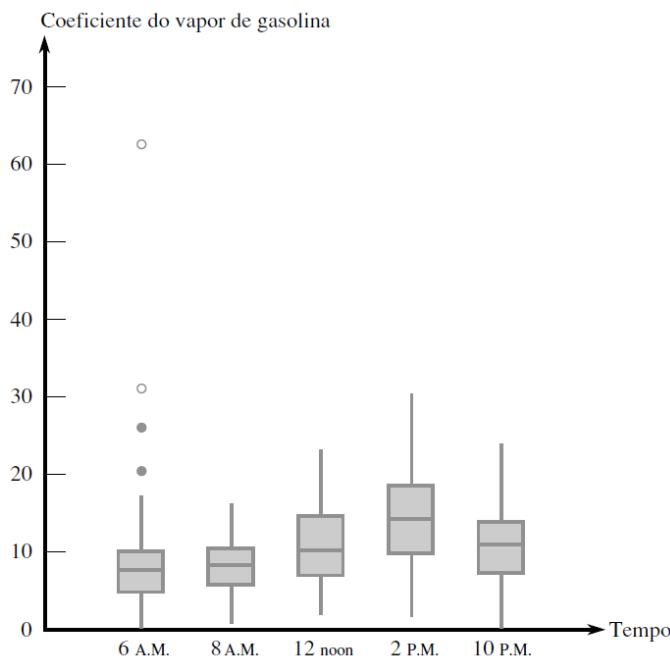
Emp. Falidas	0	1	2	3	4	5	6
Frequência	36	19	16	7	4	2	1

- a)** Calcule a média, a variância e o desvio padrão para estes dados;
- b)** Calcule a moda e a mediana destes dados.

10- Uma empresa usa duas máquinas diferentes para fabricar certo tipo de peça. Durante um turno, uma amostra de $n = 20$ peças produzidas por cada máquina é selecionada e o valor de uma importante dimensão de cada peça é determinado. O *boxplot* comparativo da figura a seguir foi construído a partir dos dados resultantes. Compare e destaque as diferenças entre as duas amostras.



11- O seguinte *boxplot* comparativo sobre coeficientes de vapor de gasolina para veículos em Detroit foi exibido no artigo “Receptor Modeling Approach to VOC Emission Inventory Validation” (J. of Envir. Engr., 1995, p. 483-490). Interprete os resultados.



3 Probabilidade

3.1 Noções de Probabilidade

Em estatística estamos interessados em lidar com resultados de experimentos aleatórios, cujos resultados possuem certa probabilidade de acontecer e não podem ser previstos com certeza.

O termo **probabilidade** se refere ao estudo da aleatoriedade e da incerteza. O desenvolvimento da probabilidade data de mais de 300 anos e teve sua origem relacionada a questões que envolviam “predizer” os resultados de jogos de azar, por matemáticos como: *Blaise Pascal [1623-1662]*, *Pierre de Fermat [1601-1665]*, *Abraham de Moivre [1667-1754]*, *Pierre Simon Laplace [1749-1827]* e *Jakob Bernoulli [1654-1705]*.

Algumas definições iniciais são necessárias:

- **Experimentos aleatórios:** são experimentos com resultado incerto ou casual, que pode ser repetido inúmeras vezes.
- **Espaço amostral:** é o conjunto de todos os resultados possíveis do experimento (Ω).
- **Evento:** é qualquer subconjunto do espaço amostral (E).

O objetivo da probabilidade é atribuir a cada evento (E), um número que exprime a chance de ocorrência deste evento.

Seja n_e a quantidade de elementos do evento E e, N a quantidade de elementos de Ω . Se os elementos de Ω são equiprováveis, então a definição clássica de probabilidades é dada por:

$$P[E] = \frac{n_e}{N} = \frac{\text{favoraveis}}{\text{possíveis}}$$

Exemplo: Três produtos são selecionados em uma linha de produção e classificados como defeituoso (D) ou não defeituoso (N).

Qual é o experimento? É aleatório?

$$\Omega = \{DDD, DDN, DND, NDD, DNN, NDN, NND, NNN\}$$

$$A : \text{escolher pelo menos 2 defeituosos} \Rightarrow \{DDD, DDN, DND, NDD\}$$

$$B : \text{nenhum produto seja defeituoso} \Rightarrow \{NNN\}$$

$$P[A] = \frac{n_e}{N} = \frac{4}{8} = \frac{1}{2}$$

$$P[B] = \frac{n_e}{N} = \frac{1}{8}$$

Propriedades

- i) $0 \leq P[E] \leq 1$
- ii) $P[\emptyset] = 0$
- iii) $P[\Omega] = 1$
- iv) $P[E^c] = 1 - P[E]$, em que E^c é o evento complementar
- v) $P[A \cup B] = P[A] + P[B] - P[A \cap B]$
- vi) $P[A \cap B] = 0$, se os eventos forem mutuamente exclusivos.

Exemplo: $P[A \cap B] = 0$...

Foi fácil encontrar as probabilidades do exemplo anterior. Por que preciso de fórmulas?

Técnicas de contagem

Suponha que seja adotado o seguinte critério de qualidade: Escolhe 5 peças e o lote é rejeitado se 3 delas forem defeituosas. Explorando as técnicas de contagem podemos calcular probabilidades como esta sem precisar relacionar todos os resultados.

Regra do produto

Se o primeiro objeto pode ser relacionado de n_1 maneiras e o segundo de n_2 , então o número de pares deles é dado por $n_1 \times n_2$. E assim, por diante para mais objetos.

Combinação

Dado um conjunto de n objetos diferentes, qualquer subconjunto não-ordenado de tamanho k é denominado combinação. O número de combinações de tamanho k que podem ser formadas a partir de n objetos distintos é representado por:

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

Exemplo: Resolvendo o exemplo do início da seção.

$$N = 2 \times 2 \times 2 \times 2 \times 2 = 2^5 = 32$$

$$n_e = C_{5,3} = \frac{5!}{3!(5-3)!} = \frac{5 \times 4}{2 \times 1} = \frac{20}{2} = 10$$

$$P[E] = \frac{n_e}{N} = \frac{10}{32} = 0,3125$$

Probabilidade Condisional

Sejam A e B dois eventos, denotamos $P[B/A]$ a probabilidade de ocorrência de B dado que A já tenha ocorrido.

Em situações simples, quando os resultados são igualmente prováveis, o cálculo de probabilidades condicionais pode se basear na intuição. No entanto, quando os experimentos são mais complicados, a intuição pode nos enganar, portanto uma definição geral de probabilidade condicional é dada por:

$$P[B/A] = \frac{P[A \cap B]}{P[A]}$$

Exemplo: Certo alimento industrializado pode ser produzido por duas maneiras diferentes: A1 e A2. A maneira A1 usa equipamentos mais antigos que A2, de forma que é mais lenta e um pouco menos confiável, porém mais barata. Suponha que em determinado dia, a maneira A1 gerou 8 produtos, dos quais 2 não ficaram com os aspectos sensoriais ideais (B1) e 6 ficaram bons (B2), ao passo que a linha A2 gerou 1 produto não ideal (B1) e 9 com as qualidades sensoriais ideais (B2).

Selecionando um produto aleatoriamente qual a probabilidade de ele ser produzido pela maneira 1 dado que não possui os aspectos sensoriais ideais?

	Asp. Sensorial		Total
	B1	B2	
Maneira	A1	2	6
	A2	1	9
Total	3		18

É possível perceber intuitivamente que a resposta é $\frac{2}{3}$.

Mas pela fórmula também funciona:

$$P[A1/B1] = \frac{P[A1 \cap B1]}{P[B1]} = \frac{\frac{2}{18}}{\frac{3}{18}} = \frac{2}{3}$$

no caso de união:

$$P[A1 \cup B1] = P[A1] + P[B1] - P[A1 \cap B1] = \frac{8}{18} + \frac{3}{18} - \frac{2}{18} = \frac{9}{18}$$

Eventos independentes

A probabilidade condicional nos permite entender melhor o conceito de independência entre eventos e do cálculo da probabilidade da interseção entre estes eventos. Relembre-se que:

Regra da Multiplicação: Se em um experimento ambos os eventos A e B podem ocorrer, então:

$$P[A \cap B] = P[A] \times P[B|A]$$

Mas se dois eventos forem independentes: $P[B|A] = P[B]$

Assim podemos compreender o seguinte teorema:

Teorema: Dois eventos são independentes se e só se $P[A \cap B] = P[A] \times P[B]$

Exemplo: Considere o experimento de lançar 2 moedas e observar a face virada para cima.

$$\Omega = \{CC, CR, RC, RR\}$$

A: sair cara na 1^a moeda

B: sair cara na 2^a moeda

$$P[A] = \frac{2}{4} = \frac{1}{2}$$

$$P[B] = \frac{2}{4} = \frac{1}{2}$$

$$P[A \cap B] = \frac{1}{4} = \frac{1}{2} \times \frac{1}{2} = P[A] \times P[B]$$

Logo o lançamento das duas moedas é independente.

Veja que não funciona para o caso do exemplo anterior, pois A1 e B1 não são independentes:

$$P[A1 \cap B1] = P[A1] \times P[B1|A1] = \frac{8}{18} \times \frac{2}{8} \neq P[A1] \times P[B1] = \frac{8}{18} \times \frac{3}{18}$$

OBS: Geralmente este teorema é utilizado ao contrário: afirma-se que os eventos são independentes para calcular a probabilidade da interseção pelo produto.

3.2 Teorema de Bayes

O teorema de Bayes é um dos mais famosos da teoria de probabilidades. A regra geral desses cálculos, que na verdade é uma aplicação simples da regra de multiplicação, remete ao reverendo Thomas Bayes, que viveu no século XVIII. Ao utilizar tal regra, foi criada uma metodologia estatística chamada de “inferência bayesiana” que tem atraído muita atenção nas aplicações, principalmente por ser capaz de resolver problemas complexos. Antes de expressá-la, precisamos relembrar algumas definições.

Pela regra da multiplicação temos que: $P[A \cap B] = P[A] \times P[B|A]$

Essa regra é importante porque freqüentemente deseja-se obter $P[A \cap B]$, ao passo que $P[A]$ e $P[B|A]$ podem ser especificados pela descrição do problema.

Definição: Os eventos A_1, A_2, \dots, A_k são mutuamente exclusivos se nenhum par deles tiver resultados comuns. Além disso, estes eventos são coletivamente exaustivos se constituem uma partição do espaço amostral, isto é, $A_1 \cup A_2 \dots \cup A_k = \Omega$.

Lei da Probabilidade Total: Sejam A_1, A_2, \dots, A_k eventos mutuamente exclusivos e coletivamente exaustivos. Então para qualquer outro evento B:

$$P[B] = \sum_{i=1}^k P[A_i \cap B] = \sum_{i=1}^k P[A_i] \times P[B|A_i]$$

Podemos então finalmente enunciar o teorema de Bayes.

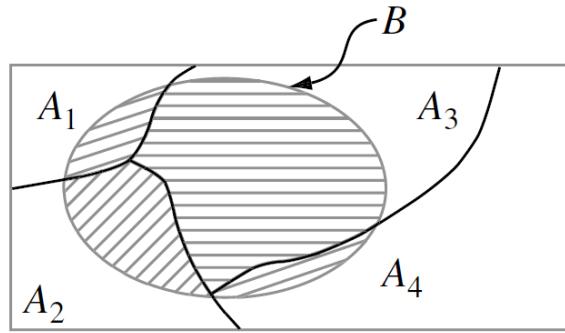


Figura 8: Partição do espaço amostral em eventos mutuamente exclusivos e coletivamente exaustivos A_i , acrescidos de um outro evento B .

Teorema de Bayes: Sejam A_1, A_2, \dots, A_k eventos mutuamente exclusivos e coletivamente exaustivos, com $P[A_i] > 0$ tal que $i = 1, \dots, k$. Então para qualquer outro evento B em que $P[B] > 0$, temos:

$$P[A_i|B] = \frac{P[A_i \cap B]}{P[B]} = \frac{P[A_i] \times P[B|A_i]}{P[B]} = \frac{P[A_i] \times P[B|A_i]}{\sum_{i=1}^k P[A_i] \times P[B|A_i]}$$

Ou em um caso mais particular:

$$P[A|B] = \frac{P[A \cap B]}{P[B]} = \frac{P[A] \times P[B|A]}{P[B]}$$

Veja que o teorema de Bayes permite calcular a probabilidade $P[A_i|B]$ utilizando as informações de $P[A_i]$ e $P[B|A_i]$, por este motivo ele é também chamado de teorema da probabilidade reversa, uma vez que para obter $P[A_i|B]$ utilizamos $P[B|A_i]$.

Nesse sentido, o teorema de Bayes geralmente é utilizado quando você está interessado em obter uma probabilidade condicional que é difícil de ser calculada inicialmente mas a outra condicional pode ser obtida diretamente.

Outro indício de que você tem um problema que pode ser resolvido pelo teorema de Bayes é quando existe uma probabilidade um tanto quanto subjetiva que não é possível de ser calculada usando a lei geral de probabilidades (favoráveis/possíveis). Esta probabilidade subjetiva é calculada pela lei da probabilidade total que é o denominador do teorema de Bayes.

Exemplo: Em uma vistoria os carros são testados para emissão excessiva de poluentes, de modo que se o teste for positivo o carro é reprovado e precisa acrescentar um filtro no escapamento. Sabe-se que 25% dos carros emitem quantidade considerada excessiva e o teste dá positivo para 99% destes, mas resulta positivo também para 17% dos carros que não emitem quantidade excessiva. Qual é a probabilidade de um carro que seja reprovado no teste realmente emitir quantidade excessiva de poluentes?

Eventos:

E : O carro emite poluentes em excesso.

N : O carro não emite poluentes em excesso.

R : O carro é reprovado no teste.

Quero saber: $P[E|R] = ???$

Quais probabilidades são conhecidas?

Perceba que, logicamente, é impossível saber $P[R]$ diretamente, pois é uma probabilidade subjetiva.

Mas ao mesmo tempo eu sei $P[E]$, que é a probabilidade do carro emitir poluentes excessivos, ou seja, $P[E] = 0,25$.

Consequentemente, temos que $P[N] = 0,75$.

Conhecemos também a probabilidade $P[R|E]$. Isto é, a probabilidade de ser reprovado dado que emite poluentes em excesso: $P[R|E] = 0,99$

Da mesma forma, conhecemos $P[R|N]$. Isto é, a probabilidade de ser reprovado dado que não emite poluentes em excesso, assim: $P[R|N] = 0,17$

Logo posso utilizar o teorema de Bayes.

$$P[E|R] = \frac{P[R|E]P[E]}{P[R|E]P[E] + P[R|N]P[N]}$$

Portanto:

$$P[E|R] = \frac{0,25 \times 0,99}{0,25 \times 0,99 + 0,75 \times 0,17} = \frac{0,2475}{0,375} = 0,66$$

Assim a probabilidade de um carro emitir poluentes em excesso dado que foi reprovado no teste é de 66%.

Veja que foi atualizado de 25% para 66% a probabilidade do carro emitir poluentes.

3.3 Diagrama de Árvore

Para os iniciantes em probabilidade, pode ser um pouco complicado identificar que algumas probabilidades fornecidas no problema são condicionais, e principalmente identificar quais são estas condicionais. Nesse sentido, o diagrama de árvore pode ser uma importante ferramenta para identificar estas probabilidades. Pois neste diagrama naturalmente os eventos da segunda camada são condicionais aos eventos da primeira camada, e assim por diante.

A figura abaixo ilustra como é construído o diagrama de árvore para o caso de dois eventos com dois níveis cada. Para mais eventos e principalmente, mais níveis, a idéia é a mesma.

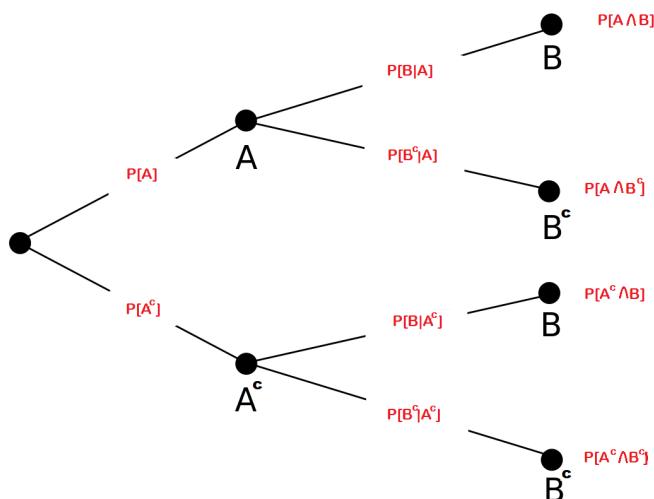


Figura 9: Exemplo de um diagrama de árvore para dois eventos com dois níveis cada.

Exercício: Uma rede de produtos alimentícios vende três marcas diferentes de um alimento. Dessas vendas, 50% são da marca 1 (a mais barata), 30% são da marca 2 e 20% são da marca 3. Dependendo das condições de armazenamento estes alimentos correm o risco de estragar antes da data de vencimento informada na embalagem. É sabido que 25% dos alimentos da marca 1 estragam antes da data de vencimento, enquanto os percentuais correspondentes para as marcas 2 e 3 são 20% e 10%, respectivamente.

Se um cliente voltar à loja com um produto estragado antes do vencimento, qual é a probabilidade de ele ser da marca 1? E da marca 2? E da marca 3?

Respostas: 0,61 0,29 0,10

Exercício: Os arquivos da polícia revelam que, das vítimas de acidente automobilístico que utilizam cinto de segurança, apenas 10% sofrem ferimentos graves, enquanto que a incidência é de 50% entre as vítimas que não utilizam cinto de segurança. Estima-se que a porcentagem de motoristas que frequentemente usam o cinto é de 60%. A polícia acaba de ser chamada para investigar um acidente em que houve um indivíduo gravemente ferido. Calcule a probabilidade de ele estar usando o cinto no momento do acidente. A pessoa que dirigia o outro carro não sofreu ferimentos graves. Calcule a probabilidade dela estar usando o cinto no momento do acidente. *Respostas:* 23,07% e 72,97%

DICA: tente resolver os exercícios acima usando o diagrama de árvore.

3.4 LISTA DE EXERCÍCIOS 3: Noções de probabilidade

1- Considere um experimento que consiste no lançamento de dois dados, um de cada vez, e que sejam definidos os seguintes eventos: $A = \text{soma dos números igual a } 9$ e $B = \text{número do primeiro dado maior ou igual a } 4$. Pede-se:

- a)** Enumere os elementos de A e B ;
- b)** Obtenha $A \cup B$, $A \cap B$ e A^c ;
- c)** Calcule as probabilidades dos eventos A, B e dos eventos definidos no item anterior;
- d)** Qual é a probabilidade de que o primeiro dado mostre a face 2 e o segundo a face 3?
- e)** Qual é a probabilidade de que ambos os dados mostrem a mesma face?
- f)** Qual é a probabilidade de que o segundo dado mostre um número par?

2- Quando Paulo vai ao futebol, a probabilidade de ele encontrar Ricardo é 0,4; a probabilidade de ele encontrar Fernando é igual a 0,10; a probabilidade de ele encontrar ambos, Ricardo e Fernando, é igual a 0,05. Qual é a probabilidade de Paulo encontrar Ricardo ou Fernando?

3- Sejam A e B eventos tais que $P[A] = 0,2$, $P[B] = p$, $P[A \cup B] = 0,6$. Calcular p considerando A e B :

- a)** mutuamente exclusivos;
- b)** independentes.

4- É comum, em muitas áreas industriais, o uso de máquinas envasadoras para colocar os produtos em caixas. Isso ocorre na indústria alimentícia, bem como em outras áreas nas quais os produtos têm uso doméstico, como o detergente. Tais máquinas não são perfeitas e podem: A, atender às especificações; B, encher as caixas menos do que o necessário; ou C, encher mais do que o necessário. Geralmente, o não enchimento das caixas é o que se deseja evitar. Seja $P[B] = 0,001$ enquanto $P[A] = 0,990$.

- a)** Forneça $P[C]$;
- b)** Qual é a probabilidade de a máquina não encher as caixas menos do que o necessário?
- c)** Qual é a probabilidade de a máquina encher as caixas mais do que o necessário ou encher menos do que o necessário?

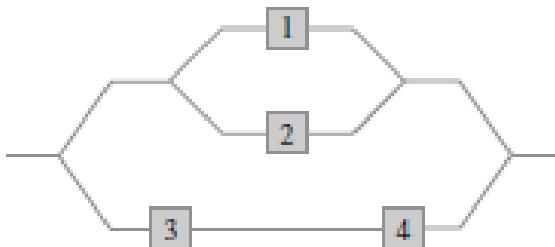
5- Num período de um mês, 100 pacientes sofrendo de determinada doença foram internados em um hospital. Informações sobre o método de tratamento aplicado (A, B) em cada paciente e o resultado final obtido estão no quadro a seguir. Sorteando aleatoriamente um desses pacientes, determinar a probabilidade de o paciente escolhido:

	A	B	Soma
Cura total	24	16	40
Cura parcial	24	16	40
Morte	12	8	20
Soma	60	40	100

- a) ter sido submetido ao tratamento A;
- b) ter sido totalmente curado;
- c) ter sido submetido ao tratamento A e ter sido parcialmente curado;
- d) ter sido submetido ao tratamento A ou ter sido parcialmente curado.
- e) Os eventos “morte” e “tratamento A” são independentes? Justifique.

6- O campo da engenharia de confiabilidade se desenvolveu rapidamente a partir do início da década de 1960. Um tipo de problema encontrado é o de se estimar a confiabilidade de um sistema a partir das confiabilidades dos subsistemas. A confiabilidade é definida, aqui, como a probabilidade do funcionamento apropriado durante um certo período de tempo.

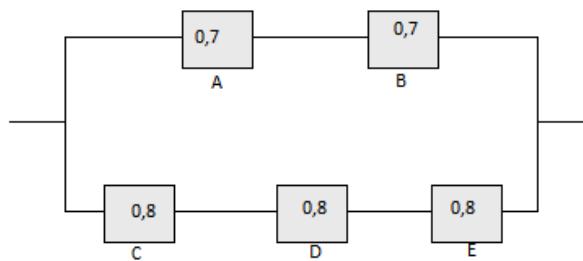
Considere o sistema de componentes ligados como na figura a seguir. Os componentes 1 e 2 estão ligados em paralelo, de forma que o subsistema funciona se, e somente se, 1 ou 2 funcionar. Como 3 e 4 estão ligados em série, o subsistema funcionará se, e somente se, 3 e 4 funcionarem. Se os componentes funcionarem independentemente um do outro e $P(\text{componente funciona})=0,8$, calcule $P(\text{sistema funciona})$.



7- Das 10 alunas de uma classe, 3 têm olhos azuis. Se duas delas são escolhidas ao acaso, em sequência, qual é a probabilidade de:

- a) ambas terem olhos azuis?
- b) nenhuma ter olhos azuis?
- c) pelo menos uma ter olhos azuis?

8- Um sistema de circuitos é apresentado na figura a seguir. Assuma que os componentes falham independentemente.



- a)** Qual é a probabilidade de que o sistema todo funcione?
- b)** Dado que o sistema funciona, qual é a probabilidade de que o componente *A* não esteja funcionando?
- 9-** Num certo colégio, 4% dos homens e 1% das mulheres têm mais de 1,75 de altura. 60% dos estudantes são mulheres. Um estudante é escolhido ao acaso e tem mais de 1,75. Qual a probabilidade de que seja homem?
- 10-** A probabilidade de um indivíduo da classe A comprar um carro é de $3/4$, da B é de $1/5$ e da C é de $1/20$. As probabilidades de os indivíduos comprarem um carro da marca *x* são $1/10$, $3/5$ e $3/10$, dado que sejam de A, B e C, respectivamente. Certa loja vendeu um carro da marca *x*. Qual a probabilidade de que o indivíduo que o comprou seja da classe B?
- 11-** Um certo programa pode ser usado com uma entre duas sub-rotinas A e B, dependendo do problema. A experiência tem mostrado que a sub-rotina A é usada 40% das vezes e a B é usada 60% das vezes. Se A é usada, existe 75% de chance de que o programa chegue a um resultado dentro do limite de tempo. Se B é usada, a chance é de 50%. Se o programa foi realizado dentro do limite de tempo, qual a probabilidade de que a sub-rotina A tenha sido a escolhida?
- 12-** A probabilidade de haver atraso no vôo diário que leva a mala postal a certa cidade é de 0,20. A probabilidade de haver atraso na distribuição local da correspondência é de 0,15, se não houve atraso no vôo, e 0,25 se houve atraso no vôo.
- a)** Qual é a probabilidade de a correspondência ser distribuída com atraso em certo dia?
- b)** Se em certo dia a correspondência foi distribuída com atraso, qual é a probabilidade de ter havido atraso no vôo?
- c)** Qual é a probabilidade de ter havido atraso no vôo, se a correspondência não foi distribuída com atraso.
- 13-** Suponha que quatro inspetores em uma fábrica de filmes tenham de estampar a data de validade em cada pacote de filme, ao final da linha de montagem. João, que estampa 20% dos pacotes, não estampa a data de validade em um de cada 200 pacotes; Tony, que estampa 60% dos pacotes, erra uma vez a cada 100 pacotes; Jefferson, que estampa 15% dos pacotes, erra uma vez a cada 90 pacotes; e Paulo, que estampa 5% dos pacotes, erra uma vez a cada 200 pacotes. Se um cliente reclama que sua embalagem de filme não contém a data de validade, qual é a probabilidade de que ela tenha sido inspecionada por João?

14- Uma indústria usa três planos analíticos para criar e desenvolver certo produto. Devido aos custos, os três planos são usados em momentos variados, de modo que os planos 1, 2 e 3 são usados para 30%, 20% e 50% dos produtos, respectivamente. O índice de defeitos é diferente para os três procedimentos: $P[D|P_1] = 0,01$, $P[D|P_2] = 0,03$ e $P[D|P_3] = 0,02$ em que $P[D|P_j]$ é a probabilidade de um produto apresentar defeito dado que veio do plano j.

a) Selecionando um produto aleatoriamente, qual a probabilidade de ele apresentar defeito?

b) Se o produto selecionado apresenta defeito, qual foi provavelmente o plano usado e, por consequência, responsável pelo defeito?

GABARITO

1- c) $P[A] = 4/36$, $P[B] = 18/36$, $P[A \cup B] = 19/36$, $P[A \cap B] = 3/36$ e $P[A^c] = 32/36$; **d)** $1/36$; **e)** $6/36$; **f)** $18/36$.

2- 0,45

3- a) $p = 0,4$; **b)** $p = 0,5$

4- a) 0,009; **b)** 0,999; **c)** 0,01;

5- a) $p = 0,6$; **b)** $p = 0,4$; **c)** 0,24; **d)** 0,76

6- 0,9856

7- a) 0,0667; **b)** 0,4667; **c)** 0,5333

8- a) 0,75112; **b)** 0,2045

9- 8/11

10- 4/7

11- 0,5

12- a) 0,170; **b)** 0,294; **c)** 0,181

13- 0,1124

14- a) $P[D] = 0,019$ **b)** $P[P_1|D] = 0,158$ $P[P_2|D] = 0,316$ $P[P_3|D] = 0,526$

A probabilidade condicional do plano 3, dado que deu defeito, é a maior das três. Portanto, um produto com defeito é, mais provavelmente, resultado do uso do plano 3.

4 Variáveis Aleatórias Discretas

Independente de um experimento gerar resultados qualitativos ou quantitativos, geralmente estamos interessados em aspectos numéricos dos resultados.

Uma variável aleatória é uma função que associa números reais com os resultados de experimentos aleatórios. Denota-se por uma letra maiúscula: X, Y, Z.

Uma **variável aleatória discreta** é aquela cujos valores possíveis formam um conjunto finito ou enumerável de valores.

Variáveis aleatórias são tão importantes que geralmente ignoramos o espaço amostral original do experimento e nos interessamos pela distribuição probabilidades da variável aleatória. Simplificando assim a descrição e a análise do experimento.

4.1 Distribuição de Probabilidades

A **função de probabilidades** de uma variável aleatória discreta fornece a probabilidade de ocorrência de cada um dos valores possíveis. Isto é, se $x_i, i = 1, 2, \dots, n$ são os valores assumidos pela variável X, estão:

$$p(x_i) = P[X = x_i]$$

Propriedades

- i) $0 \leq p(x_i) \leq 1 \quad \forall i = 1, 2, \dots, n$
- ii) $\sum_{i=1}^n p(x_i) = 1$

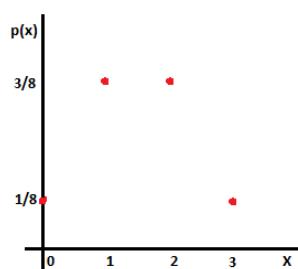
Ao conjunto $\{x_i, p(x_i); i = 1, 2, \dots, n\}$ damos o nome de **distribuição de probabilidades** da variável aleatória X. Que pode ser representada em uma tabela ou um gráfico.

Exemplo: Você observa se 3 alimentos estão estragados. Seja X o número de alimentos estragados. Determine a distribuição de probabilidades de X.

$$\Omega = \{EEE, EEB, EBE, BEE, EBB, BEB, BBE, BBB\} \quad \text{e} \quad X : \{0, 1, 2, 3\}$$

Distribuição de probabilidades da variável X.

X	P[X=x]
0	1/8
1	3/8
2	3/8
3	1/8



4.2 Função de Distribuição Acumulada

A função de distribuição acumulada $F(x)$ de uma variável aleatória discreta X é definida por:

$$F[x] = P[X \leq x] = \sum_{y \leq x} p(y)$$

Exemplo: Encontrar no máximo dois alimentos estragados.

$$F[2] = P[X \leq 2] = p(0) + p(1) + p(2) = \frac{1}{8} + \frac{3}{8} + \frac{3}{8} = \frac{7}{8}$$

Esperança

O valor esperado de uma variável aleatória (ou valor médio) é definido por:

$$E[X] = \sum_{i=1}^n x_i p(x_i)$$

Exemplo: Qual o valor esperado do número de produtos estragados?

$$E[X] = \sum_{i=1}^4 x_i p(x_i) = 0 * \frac{1}{8} + 1 * \frac{3}{8} + 2 * \frac{3}{8} + 3 * \frac{1}{8} = \frac{12}{8} = 1,5$$

Logo, assumindo que o alimento estar estragado, ou não estar estragado, são equiprováveis, o número médio de produtos com defeito em 3 alimentos é 1,5.

Variância

A variância de uma variável aleatória é obtida por:

$$VAR[X] = E[(X - E[X])^2] = E[X^2] - (E[X])^2,$$

em que:

$$E[X^2] = \sum_{i=1}^n x_i^2 p(x_i)$$

Covariância

A covariância mede o grau de dependência entre duas variáveis X e Y. É definida por:

$$COV[X, Y] = E[(X - E[X])(Y - E[Y])]$$

Exercício: Em determinada linha industrial a obtenção de certo solvente é trabalhosa e requer muita atenção, de modo que 1/3 dos solventes produzidos não ficam com as propriedades químicas desejáveis e consequentemente não podem ser vendidos e devem ser reciclados. Se forem produzidos 4 solventes em um turno de trabalho, calcule:

- a) A probabilidade de exatamente 2 solventes irem para a reciclagem.
- b) A distribuição de probabilidades da variável solventes separados para a reciclagem.
- c) A probabilidade de no máximo 3 solventes irem para a reciclagem.
- d) O valor esperado e o desvio padrão de solventes que voltam para a reciclagem.

a) Espaço amostral do experimento:

$$\Omega = \{VVVV, VVVR, VVRR, VRRL, VVRV, RVVV, RVRV, RVRR, VRVR, VRRV, VRVV, RVVR, RRVR, RRVV, RRRV, RRRR\}$$

Variável aleatória associada:

X: solvente vai para a reciclagem

$$X = \{0, 1, 2, 3, 4\}$$

$$X = 2 \Rightarrow VVRR, RVRR, VRVR, VRRV, RVVR, RRVV$$

$$p[VVRR] = \frac{2}{3} \times \frac{2}{3} \times \frac{1}{3} \times \frac{1}{3} = \frac{4}{81}$$

$$\text{Logo: } P[X = 2] = 6 \times \frac{4}{81} = 0,2963$$

b)

$$P[X = 0] = 1 \times \frac{16}{81} = \frac{16}{81} = 0,1975$$

$$P[X = 1] = 4 \times \frac{8}{81} = \frac{32}{81} = 0,3951$$

$$P[X = 2] = 6 \times \frac{4}{81} = \frac{24}{81} = 0,2963$$

$$P[X = 3] = 4 \times \frac{2}{81} = \frac{8}{81} = 0,0988$$

$$P[X = 4] = 1 \times \frac{1}{81} = \frac{1}{81} = 0,0123$$

Assim a distribuição de probabilidades fica:

X	P[X=x]
0	16/81
1	32/81
2	24/81
3	8/81
4	1/81

$$c) F[3] = P[X \leq 3] = p(0) + p(1) + p(2) + p(3) = \frac{16}{81} + \frac{32}{81} + \frac{24}{81} + \frac{8}{81} = \frac{80}{81}$$

d) O valor esperado é dado por:

$$E[X] = \sum_{i=1}^5 x_i p(x_i) = 0 * \frac{16}{81} + 1 * \frac{32}{81} + 2 * \frac{24}{81} + 3 * \frac{8}{81} + 4 * \frac{1}{81} = \frac{108}{81} = 1,33$$

E o desvio padrão?? É muito trabalhoso obtê-lo manualmente...

Visando facilitar estes cálculos vamos utilizar as distribuições de probabilidade.

Felizmente algumas variáveis aleatórias adaptam-se muito bem à uma série de problemas práticos e aparecem com frequência. Podemos então modelar seu comportamento para que, sob certas condições, não seja preciso recomeçar tudo a partir do espaço amostral.

4.3 Distribuição Binomial

Utiliza-se o termo “binomial” para designar situações em que os resultados de uma variável aleatória podem ser agrupados em 2 classes ou categorias. Estas categorias devem ser mutuamente exclusivas e coletivamente exaustivas.

Se uma Variável Aleatória X atende as seguintes condições:

- 1- tem n repetições independentes
- 2- é um termo binomial
- 3- a probabilidade “p” de ocorrência é constante.

Então X tem distribuição binomial com n repetições e probabilidade p de sucesso em cada ensaio. Notação: $X \sim B(n, p)$.

Calcula-se a probabilidade de ocorrência por:

$$P[X = x] = C_{n,x} p^x (1 - p)^{(n-x)} \quad x = 0, 1, 2, \dots, n$$

Esperança

O valor esperado de uma variável aleatória que possui distribuição Binomial pode ser facilmente obtido por:

$$E[X] = n \times p$$

Variância

A variância de uma variável aleatória que possui distribuição Binomial pode ser obtida por:

$$VAR[X] = n \times p \times (1 - p)$$

Exemplo: Considere o exemplo anterior sobre a produção de solventes. Calcule a probabilidade de 2 solventes serem separados para a reciclagem.

X: número de solventes separados para a reciclagem

$$n = 4 \quad p = \frac{1}{3} \quad X \sim B(4, \frac{1}{3}).$$

A variável X atende as condições 1 à 3?

$$P[X = 2] = C_{n,x} p^x (1 - p)^{(n-x)} = C_{4,2} \frac{1^2}{3} (1 - \frac{1}{3})^{(4-2)} = C_{4,2} \frac{1^2}{3} \frac{2^2}{3} = 6 \times \frac{4}{81} = 0,2963$$

Conforme solicitado na letra **d)** do exercício, precisamos obter o valor esperado da variável X.

$$E[X] = n \times p = 4 \times \frac{1}{3} = 1,33$$

E o desvio padrão de X (que não calculamos) também pode ser obtido facilmente:

$$DP[X] = \sqrt{VAR[X]} = \sqrt{n \times p \times (1-p)} = \sqrt{4 \times \frac{1}{3} \times \frac{2}{3}} = \sqrt{\frac{8}{9}} = 0,9428$$

Exercício 1: Quando as placas de circuito integrado usadas na fabricação de TV's são testadas, a porcentagem de placas com defeitos no longo prazo é igual a 5%. Em uma amostra aleatória de 25 peças calcule a probabilidade de:

- a) Exatamente 2 placas apresentarem defeito. 0,2305
- b) No máximo 3 placas apresentarem defeito. 0,9659
- c) Pelo menos uma placa apresentar defeito. 0,7226

Exercício 2: Uma certa doença pode ser curada através de procedimento cirúrgico em 80% dos casos. Dentre os que apresentam essa doença, sorteamos 15 pacientes que serão submetidos à cirurgia. Determine:

- (a) A probabilidade de todos serem curados. R = 0,0352
- (b) A probabilidade de, ao menos, 13 ficarem livres da doença. R = 0,3980
- (c) A probabilidade de, no máximo 13 ficarem livres da doença. R = 0,8329
- (d) A probabilidade de, pelo menos, 2 não serem curados. R = 0,8329

4.4 Distribuição de Poisson

É utilizada para descrever situações em que a variável aleatória Y está associada ao número médio (λ) de ocorrências por um período de tempo área ou volume específico.

Se a situação acima ocorre e as 3 condições da distribuição binomial são aceitas, então dizemos que Y tem distribuição de Poisson com média λ . Notação: $X \sim P(\lambda)$.

Calcula-se a probabilidade de ocorrência por:

$$P[Y = y] = \frac{e^{-\lambda} \times \lambda^y}{y!} \quad y = 0, 1, 2, \dots$$

Esperança e Variância

Uma característica importante da distribuição de Poisson é:

$$E[X] = VAR[X] = \lambda$$

Exemplo: Na preparação de determinado solvente é comum encontrar algumas partículas residuais do processo químico. São encontradas em média 6 partículas por litro. Qual a probabilidade de em um litro não encontrar nenhuma partícula?

Y: número de partículas por litro

$$\lambda = 6 \quad Y \sim P(6)$$

A variável Y atende as condições 1 à 3?

Assim a probabilidade de não encontrar nenhuma partícula é dada por:

$$P[Y = 0] = \frac{e^{-\lambda} \times \lambda^y}{y!} = \frac{e^{-6} \times 6^0}{0!} = 0,0025$$

Qual o valor esperado do número de partículas por litro?

$$E[Y] = \lambda = 6$$

E o desvio padrão?

$$DP[Y] = \sqrt{VAR[Y]} = \sqrt{6} = 2,4495$$

A distribuição de Poisson tem uma característica particular, na qual o λ pode ser transformado por regra de 3 simples. Por exemplo:

Qual é a probabilidade de encontrar 2 partículas em meio litro?

X: número de partículas em meio litro

$$\lambda = 3 \quad X \sim P(3)$$

Assim a probabilidade solicitada é:

$$P[X = 2] = \frac{e^{-\lambda} \times \lambda^x}{x!} = \frac{e^{-3} \times 3^2}{2!} = 0,2240$$

4.5 Relação entre as distribuições Binomial e Poisson.

A única diferença entre as distribuições Binomial e de Poisson é que na última você conhece a ocorrência média e não a probabilidade de ocorrência. As outras condições são iguais em ambas.

Na distribuição Binomial, quando n é grande mas a probabilidade de ocorrência de um evento é pequena, diz-se que temos um evento raro. Na prática considera-se um evento raro quando: $n > 50$ e $np < 5$.

Nesta situação, o cálculo pela distribuição Binomial pode se tornar complicado, inclusive estourando a capacidade de memória das calculadoras mais simples. Assim, a distribuição Binomial pode ser aproximada pela distribuição de Poisson, basta considerar $\lambda = np$. Quanto menor o valor de p e maior o valor de n , melhor a aproximação. Vem daí a denominação de “lei dos casos raros” para a distribuição de Poisson.

Exemplo: A probabilidade de um organismo rejeitar determinado medicamento é de 0,001. Foi feito uma pesquisa com 2000 pessoas utilizando este medicamento. Qual é a probabilidade de que exatamente 3 rejeitem o medicamento?

X: Organismo rejeitar o medicamento

$$n = 2000$$

$$p = 0,001$$

$$X \sim B(2000; 0,001)$$

$$P[X = 3] = C_{2000,3} \times 0,001^3 \times 0,999^{1997} = 0,1805$$

No entanto, temos: $n > 50$ e $n \times p = 2000 \times 0,001 = 2 < 5$

Assim, temos um caso raro, logo $X \sim P(2)$. Podemos calcular a probabilidade por:

$$P[X = 3] = \frac{e^{-2} \times 2^3}{3!} = 0,1804$$

Exercício 1: Na central de atendimento ao cliente de uma empresa chegam em média 6 reclamações por hora. Qual a probabilidade de que:

- a) Em uma hora não chegar nenhuma reclamação? R = 0,0025
- b) Chegar exatamente 3 reclamações em uma hora? R = 0,0892
- c) Chegar 5 reclamações em 2 horas? R = 0,0127
- d) Chegar pelo menos 1 reclamação em 20 minutos? R = 0,8646
- e) Chegar no máximo 2 reclamações em meia hora? R = 0,4232

Exercício 2: Em uma certa instalação industrial, acidentes acontecem com baixa frequência. Sabe-se que a probabilidade de ter acidentes em certo dia é de 0,005, e os acidentes são independentes uns dos outros. Qual é a probabilidade de que, em qualquer período de 400 dias, aconteça um acidente? 0,2707

Exercício 3: Uma plantação de tomate possui em média 2 pulgões por planta.

- a) Qual é a probabilidade de que uma planta amostrada desta população não possua pulgão? 0,1353
- b) Qual é a probabilidade de que em uma amostra de tamanho $n = 5$ plantas, duas não apresentem pulgão? 0,1183

4.6 LISTA DE EXERCÍCIOS 4: Distribuições de Probabilidade de Variáveis Aleatórias Discretas

1- Considere ninhadas de $n = 3$ filhotes de coelhos e que os eventos “nascer macho” e “nascer fêmea” sejam equiprováveis. Sendo X a variável aleatória que representa a ocorrência de filhotes fêmeas, construa a distribuição de probabilidades de X e o seu respectivo gráfico.

2- A função de probabilidade da variável aleatória X é: $P[X] = \frac{1}{5}$, para $X = 1, 2, 3, 4, 5$. Calcular $E[X]$. Usando esses resultados e as propriedades de esperança disponíveis no campus virtual, calcule:

- a) $E[X + 3]^2$
- b) $E[3X - 2]$

3- Uma pessoa vende um determinado tipo de máquina usado em construções. Visita semanalmente uma, duas ou três construtoras com probabilidades 0,2, 0,5 e 0,3, respectivamente. De cada contato pode conseguir a venda de 1 máquina por R\$ 120.000,00 com probabilidade de 0,3, ou nenhuma venda com probabilidade 0,7. Determinar o valor total esperado (médio) das vendas semanais.

4- Seja X o número de peças com defeito produzidas em um turno de determinada indústria. A distribuição de probabilidades de X é dada por:

X	0	1	2	3	4
$P[X = x_i]$	0,41	0,37	p	0,05	0,01

- a) Qual a probabilidade de serem encontradas duas peças defeituosas neste turno?
- b) Qual a probabilidade de ser encontrada no máximo uma peça defeituosa neste turno? E de serem encontradas mais de 3 peças defeituosas?
- c) Calcule o valor esperado, a variância e o desvio padrão de peças defeituosas neste turno.

5- Na produção de uma peça são empregadas duas máquinas. A primeira é utilizada para efetivamente produzir as peças, e o custo de produção é de R\$50,00 por unidade. Das peças produzidas nessa máquina, 90% são perfeitas. As peças defeituosas (produzidas na primeira máquina) são colocadas na segunda máquina para uma tentativa de recuperação (torná-las perfeitas). Nessa segunda máquina o custo por peça é de R\$25,00, mas apenas 60% das peças são de fato recuperadas. Sabendo que cada peça perfeita é vendida por R\$90,00, e que cada peça defeituosa é vendida por R\$20,00, calcule o lucro por peça esperado pelo fabricante.

6- De acordo com uma pesquisa da American Demographics (2005), entre jovens que compram online 20% possuem um aparelho celular com acesso a internet. Em uma amostra aleatória de 200 jovens que fazem este tipo de compra, considere X o número daqueles que possuem um telefone celular com acesso a internet.

- a) Explique porque X é uma variável aleatória Binomial.
- b) Qual o valor de p ? Interprete este valor.
- c) Qual o valor esperado de X e qual a variância de X ? Interprete estes valores.

7- Um empresário tem dois eventuais compradores de seu produto, que pagam preços em função da qualidade.

- O comprador A paga R\$150,00 por peça se em uma amostra de 100 peças não encontrar nenhuma peça defeituosa, mas paga somente R\$50,00 por peça, se em uma amostra de 100 peças encontrar uma ou mais peças defeituosas.
- O comprador B paga R\$200,00 por peça desde que encontre no máximo uma peça defeituosa em uma amostra de 120 peças, mas paga somente R\$30,00 por peça, se em uma amostra de 120 peças encontrar duas ou mais peças defeituosas.

Qual dos dois compradores deve ser escolhido pelo empresário, sabendo-se que a proporção de peças defeituosas pode ser considerada constante e igual a 3%?

8- Suponha que a probabilidade de qualquer peça produzida por uma determinada máquina ser defeituosa seja 0,2. Se 10 peças produzidas por essa máquina forem escolhidas ao acaso, qual é a probabilidade de não mais de uma peça defeituosa ser encontrada?

9- Um processo de produção que fabrica transistores opera, com fração de defeituosos de 2%. Todo dia extrai-se uma amostra aleatória de 50 transistores deste processo. Se a amostra contiver mais de dois defeituosos, o processo deve ser interrompido e a máquina regulada novamente. Qual é a probabilidade do processo ser interrompido?

10- Uma certa doença pode ser curada através de procedimento cirúrgico em 80% dos casos. Dentre os que apresentam essa doença, sorteamos 15 pacientes que serão submetidos à cirurgia. Determine:

(a) A probabilidade de todos serem curados. (b) A probabilidade de, ao menos, 13 ficarem livres da doença. (c) A probabilidade de, pelo menos, 2 não serem curados. (d) A média e o coeficiente de variação do número de pacientes curados.

11- Uma transportadora garante que consegue entregar 95% das encomendas sem nenhum dano. Em um total de 50 entregas feitas por esta empresa, qual a probabilidade de que pelo menos 2 encomendas tenham avarias no transporte?

12- Vinte por cento dos refrigeradores produzidos por uma empresa são defeituosos. Um comprador adotou o seguinte procedimento: de cada lote ele testa 20 aparelhos, e se houver pelo menos 2 defeituosos o lote é rejeitado. Admitindo-se que o comprador tenha aceitado o lote, qual a probabilidade de ter observado exatamente um aparelho defeituoso?

13- Na indústria têxtil, uma preocupação é com o número de pequenas avarias causadas pelas máquinas no processo de construção do tecido, estas avarias devem ser verificadas e concertadas antes de enviar o tecido para o estoque. Determinada máquina apresenta em média 1 defeito a cada dez metros quadrados de tecido. Qual é a probabilidade de que em 10 metros quadrados de tecido, produzidos por esta máquina, não ocorra nenhum defeito? Se esta máquina produz 30 metros quadrados de tecido por dia, qual a probabilidade de serem encontrados no máximo 2 defeitos na produção diária da máquina.

14- A aplicação de um fundo anti-corrosivo em chapas de aço de $1m^2$ é feita mecanicamente e pode produzir defeitos (pequenas bolhas nas pinturas) com média de 1,5 bolhas por m^2 . Pergunta-se a probabilidade de encontrar: **(a)** Pelo menos um defeito por m^2 ; **(b)** No máximo 2 defeitos; **(c)** Exatamente um defeito em $2m^2$; **(d)** Entre 2 e 4 defeitos em $2m^2$; **(e)** Não mais de um defeito em meio m^2 .

15- Nas competições de arco e flecha o competidor fica atirando flechas durante 1 minuto em cada alvo. Jhonny é competidor desta categoria e acerta em média 2 flechas por alvo.

(a) Qual a probabilidade de Jhonny acertar pelo menos uma flecha no alvo?

(b) Em determinada fase da competição, Jhonny deve atirar em 5 alvos e avança de fase se acertar pelo menos uma flecha em 4 desses alvos. Qual é a probabilidade de Jhonny passar de fase?

16- Um certo tipo de máquina utilizada na indústria alimentícia possui diversos mini motores que funcionam independentemente. Após 5 anos de uso encontra-se em média 1 destes mini motores com defeito por máquina. Sabendo que a máquina só funciona corretamente se todos os seus mini motores estiverem funcionando, qual é a probabilidade de uma máquina estar funcionando corretamente ao fim de 5 anos? Se você possui 3 destas máquinas, que foram adquiridas na mesma época, qual é a probabilidade de ao fim de 5 anos de uso você ter pelo menos 2 máquinas funcionando?

17- Suponha que em um recipiente existam 10.000 partículas. A probabilidade de que uma destas partículas escape do recipiente é igual a 0,0004. Admitindo-se que as partículas escapam de maneira independente, qual a probabilidade de que mais de quatro partículas escapem?

GABARITO

2: **a)** 36 **b)** 7 **3:** R\$ 75600,00

4: **a)** 0,16 **b)** 0,78 e 0,01 **c)** 0,88; 0,8456 e 0,91956

5: E[x]=R\$ 34,70 **6:** **b)** $p = 0,20$ **c)** 40 e 32

7: E[A] = R\$54,98; E[B] = R\$51,37 Logo deve ser escolhido o comprador A.

8: 0,3758 **9:** 0,0784

10: **a)** 0,0352 **b)** 0,3980 **c)** 0,8329 **d)** 12 e 12,9%

11: 0,7206 **12:** 0,8333 **13:** 0,3678 e 0,4232

14: **a)** 0,7768 **b)** 0,8088 **c)** 0,1493 **d)** 0,2240 **e)** 0,8266

15: **a)** 0,8646 **b)** 0,8611 **16:** 0,3679 e 0,3066 **17:** 0,3711

5 Variáveis Aleatórias Contínuas

Se a escala de medida de uma variável aleatória puder ser subdividida tanto quanto desejar, a variável será contínua. Neste tipo de variável é impossível enumerar todos os valores possíveis, assim não conseguimos montar uma tabela com X e $P[X = x]$.

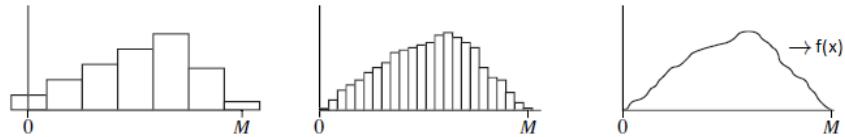


Figura 10: Exemplo de histograma para uma variável contínua, sendo medida cada vez com mais precisão.

Para uma variável aleatória contínua temos uma função de densidade de probabilidade (fdp). Que é uma curva, em função de x , cuja área abaixo corresponde as probabilidades.

Definição: A função $f(x)$ é a função de densidade de probabilidade para a variável aleatória contínua X , definida no conjunto dos reais se:

$$1) \quad f(x) \geq 0 \quad \forall x \in \mathbb{R}$$

$$2) \quad \int_{-\infty}^{+\infty} f(x)dx = 1$$

$$3) \quad P[a < X < b] = \int_a^b f(x)dx$$

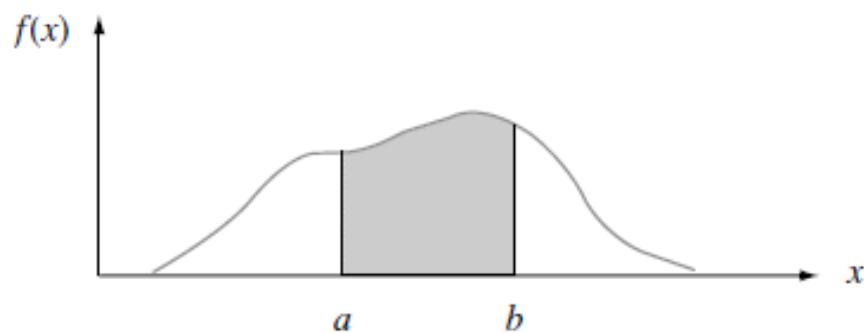


Figura 11: Em variáveis aleatórias contínuas a probabilidade de a variável X pertencer a um intervalo (a,b) é a área abaixo da curva neste intervalo.

OBS: Os itens 1 e 2 constituem condição necessária e suficiente para que $f(x)$ seja uma função de densidade de probabilidades.

Exemplo: Seja X a variável aleatória que representa o erro na temperatura de reação (em $^{\circ}\text{C}$), para um experimento realizado no laboratório. A fdp de X é dada por:

$$f(x) = \begin{cases} \frac{x^2}{3}, & -1 < x < 2; \\ 0, & \text{caso contrário.} \end{cases}$$

a) Verifique se $f(x)$ é uma fdp.

Como $\frac{x^2}{3} > 0$ sempre, basta verificar agora que a integral de $f(x)$ em todo o seu domínio é igual a 1.

$$\int_{-\infty}^{+\infty} f(x)dx = \int_{-1}^2 \frac{x^2}{3} dx = \left. \frac{x^3}{9} \right|_{-1}^2 = \frac{8}{9} + \frac{1}{9} = 1$$

Logo como os itens 1 e 2 da definição são atendidos, então $f(x)$ é uma fdp.

Um problema muito comum consiste em encontrar o valor de c para que $f(x)$ seja uma fdp, no caso do exemplo $c = \frac{1}{3}$.

b) Calcule $P[0 < X < 1]$

$$P[0 < X < 1] = \int_0^1 \frac{x^2}{3} dx = \left. \frac{x^3}{9} \right|_0^1 = \frac{1}{9}$$

5.1 Função de Distribuição Acumulada

A função de distribuição acumulada $F(x)$ de uma variável aleatória contínua X , com fdp $f(x)$, é dada por:

$$F(x) = P[X \leq x] = \int_{-\infty}^x f(t)dt \quad -\infty < x < +\infty$$

Exemplo: A função de distribuição acumulada do exemplo anterior é:

$$F(x) = P[X \leq x] = \int_{-\infty}^x f(t)dt = \int_{-\infty}^x \frac{t^2}{3} dt = \int_{-1}^x \frac{t^2}{3} dt = \left. \frac{t^3}{9} \right|_{-1}^x = \frac{x^3 + 1}{9}$$

Portanto:

$$F(x) = \begin{cases} 0, & x < -1; \\ \frac{x^3 + 1}{9}, & -1 < x < 2; \\ 1, & x \geq 2. \end{cases}$$

Naturalmente, $P[a < X < b] = F(b) - F(a)$.

Logo, no caso da letra **b**) do exemplo anterior:

$$P[0 < X < 1] = F(1) - F(0) = \frac{1^3 + 1}{9} - \frac{0^3 + 1}{9} = \frac{2}{9} - \frac{1}{9} = \frac{1}{9}$$

Esperança

O valor esperado (médio) de uma variável aleatória contínua é dado por:

$$E[X] = \int_{-\infty}^{+\infty} xf(x)dx$$

Variância

Assim como no caso discreto, a variância pode ser calculado por:

$$VAR[X] = E[(X - E[X])^2] = E[X^2] - (E[X])^2$$

em que:

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

Diferente do que ocorre com as distribuições discretas, a distribuição de uma variável aleatória contínua não pode ser deduzida por meio de argumentos probabilísticos simples.

A obtenção de uma fdp é feita com base na experiência de conhecimentos anteriores e dados disponíveis. A fdp funciona como um modelo para a distribuição de probabilidades dos valores da população contínua. É utilizada para obter todas as informações necessárias sobre a variável aleatória contínua am estudo.

Felizmente, há algumas famílias gerais de funções de distribuição de probabilidades que modelam muito bem as diversas situações experimentais.

5.2 A distribuição normal

É a distribuição de probabilidades mais importante em estatística, pois a maioria dos fenômenos da natureza apresentam tal comportamento.

Sua função de densidade de probabilidades é dada por:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \times e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad -\infty < x < +\infty$$

Diz-se que a variável aleatória X tem distribuição normal com parâmetros μ e σ^2 .
Notação: $X \sim N(\mu, \sigma^2)$

Esperança

A esperança de uma variável aleatória que possui distribuição normal é dada por:

$$E[X] = \mu$$

Variância

A variância de uma variável aleatória que possui distribuição normal é dada por:

$$VAR[X] = \sigma^2$$

Portanto os parâmetros da distribuição normal são a própria média e a variância.

Propriedades

- i) é simétrica em relação a μ (forma de sino);
- ii) a dispersão varia de acordo com σ ;

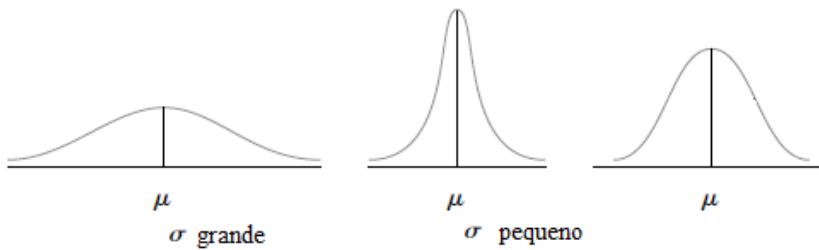


Figura 12: Distribuição normal e seus parâmetros.

iii) está totalmente definida conhecendo-se a média μ e a variância σ^2 .

Como destacado na propriedade iii), a distribuição normal fica completamente definida pelo conhecimento dos parâmetros μ e σ^2 , que são a média e a variância respectivamente.

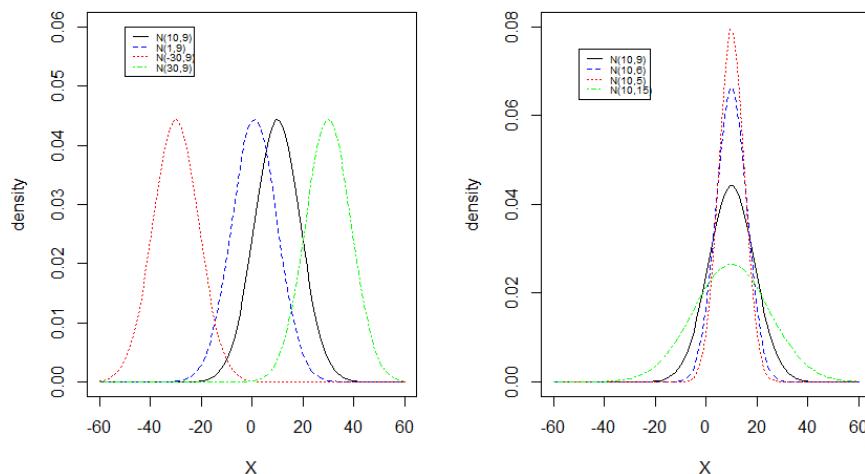
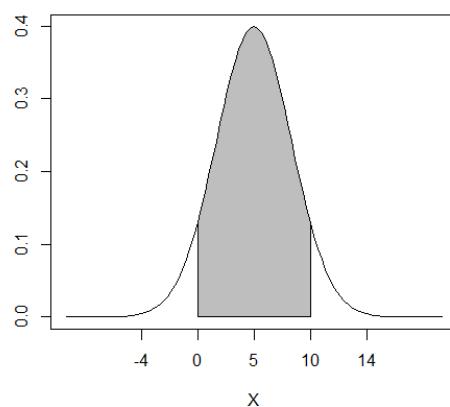


Figura 13: Distribuição normal variando as médias e mantendo a variância fixa e modificando a variância e deixando a média fixa.

DESAFIO: Seja uma variável aleatória que possui distribuição normal com média 5 e variância 9, isto é, $X \sim N(5, 9)$. Calcule a probabilidade de X estar entre 0 e 10.



5.3 A distribuição normal padrão

Conforme vimos para distribuições contínuas, o cálculo de probabilidades é feito utilizando a integral da função de densidade de probabilidade. Assim, a probabilidade solicitada no fim da seção anterior deve ser calculada resolvendo a integral:

$$P[0 < X < 10] = \int_0^{10} \frac{1}{3\sqrt{2\pi}} \times e^{-\frac{(x-5)^2}{2\times 9}} dx = ???$$

No entanto, nenhuma das técnicas de integração-padrão podem ser usadas para calcular esta integral. Desta forma, a saída é calcular esta integral computacionalmente e colocar os resultados em tabelas. Mas para cada combinação de média e variância, teria de ser construída uma tabela.

Pensando em evitar a elaboração de infinitas tabelas, foi criada a distribuição normal padrão. Todas as distribuições normais podem ser transformadas nela e assim utilizamos apenas uma tabela para calcular as probabilidades.

Seja $X \sim N(\mu, \sigma^2)$, a transformação é feita da seguinte maneira:

$$Z = \frac{X - \mu}{\sigma}$$

Assim, dizemos que a variável aleatória Z possui distribuição normal com média 0 e desvio padrão 1, que na notação fica $Z \sim N(0, 1)$. Em outras palavras, Z tem distribuição normal padrão.

Os valores assumidos pela distribuição normal padrão, variam de -4 a 4, pois a média é igual a 0 e o desvio padrão igual a 1. Já verificamos anteriormente que se a distribuição for simétrica cerca de 99,7% dos valores estão a três desvios padrões da média (entre -3 e 3 no caso da normal padrão).

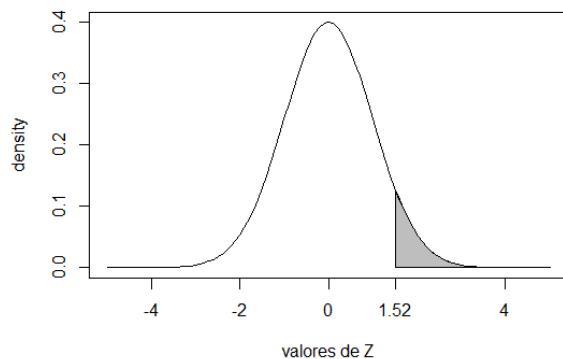
Existem diferentes versões para os valores tabelados da distribuição normal padrão. Em todas as tabelas, a forma como ela foi montada, isto é, as probabilidades que se encontram nela, está explícito em seu título. A tabela a ser utilizada nesta disciplina (veja na página 60) será a da distribuição acumulada da normal padrão denotada por $\phi[z]$, isto é: $\phi[z] = P[Z < z]$.

Para calcular a probabilidade, por exemplo, da variável normal padrão ser maior que 1,52, basta procurar na tabela a probabilidade correspondente ao valor de Z . Atente-se para o fato de que a parte inteira e a primeira casa decimal do valor de Z estão nas linhas da tabela e a segunda casa decimal está nas colunas da tabela.

Lembre-se que, por estarmos utilizando a tabela da distribuição acumulada, ao observar na tabela o valor encontrado é a probabilidade acumulada até 1,52, isto é $P[Z < 1,52]$. Para obter $P[Z > 1,52]$, fazemos:

$$P[Z > 1,52] = 1 - P[Z < 1,52] = 1 - 0,9357 = 0,0643 = 6,43\%$$

Graficamente a probabilidade que estamos procurando é:



Para um melhor entendimento do cálculo de probabilidades com a tabela da distribuição normal padrão reproduza os exercícios a seguir (utilize a tabela da página 60):

Exercício: Determine as seguintes áreas (probabilidades) e faça um esboço da figura para cada situação:

- (a) Acima de 2,3; (b) Entre 0,0 e 1,22; (c) Entre -2,3 e 0,0; (d) Entre -1,96 e 1,96;
- (e) Abaixo de -0,18; (f) Entre 0,27 e 1,18; (g) Abaixo de 1,38; (h) Acima de -1,0;
- (i) Entre -2,1 e 1,2

Repostas:

- a) 0,0107; b) 0,3888; c) 0,4893; d) 0,95; e) 0,4286; f) 0,2746; g) 0,9162; h) 0,8413; i) 0,8670

Exercícios da distribuição normal

Exemplo: Uma indústria elétrica fabrica lâmpadas que têm vida útil, antes de queimarem, normalmente distribuída com média igual a 800 horas e desvio padrão de 40 horas.

a) Qual a probabilidade de que uma lâmpada desta indústria dure mais de 834 horas?

Resposta: 0,1977 ou 19,77%

b) Qual a probabilidade de que uma lâmpada dure entre 778 e 834 horas?

Resposta: 0,5111 ou 51,11%

Exercício 1: Certa indústria produz latas de conservas de modo que o peso é uma variável aleatória normalmente distribuída com média de 990g e variância de 100g². Se uma lata for selecionada aleatoriamente, qual a probabilidade de que:

a) pese mais de 1Kg? Resposta: 0,1587 ou 15,87%

b) pese menos de 950g? Resposta: 0%

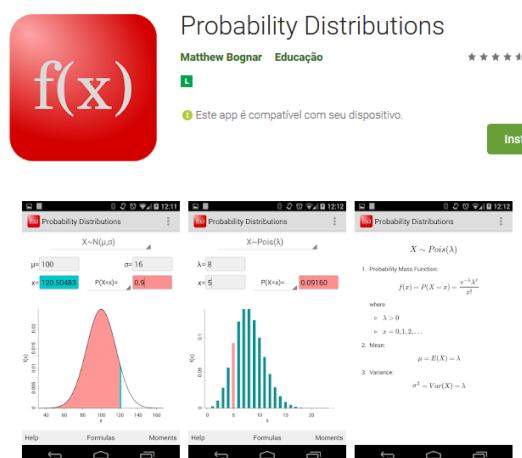
c) Uma lata é aceita pelo controle de qualidade desta indústria se seu peso diferir da média em no máximo ± 2 desvios-padrão. Qual a probabilidade de uma lata escolhida aleatoriamente ser aceita pelo controle de qualidade? Resposta: 0,9544 ou 95,44%

Exercício 2: Considere que as notas da primeira avaliação de estatística possuem distribuição normal com média 50 e desvio padrão de 16 pontos. Se o professor deseja aprovar os 10% melhores da turma já na P1, qual a nota corte para ser aprovado? Resposta: $X = 70,48\text{cm}$

Exercício 3: As árvores mais grossas de uma floresta deverão ser cortadas. Sabe-se que o diâmetro a altura do peito (DAP) desta floresta possui distribuição normal com média de 64cm e o desvio padrão é 3,2cm. Se o objetivo é cortar as 1,5% árvores mais grossas, a partir de qual DAP uma árvore deve ser cortada? Resposta: $X = 70,944\text{cm}$

Aplicativo para celular

Um aplicativo para *smartphones* chamado “Probability distributions” é muito útil para o entendimento do cálculo de probabilidades. Este aplicativo é gratuito e muito completo, com as principais distribuições de probabilidade. Ele apresenta as expressões das distribuições, calcula probabilidades e principalmente mostra o gráfico das distribuições de probabilidades.



5.4 LISTA DE EXERCÍCIOS 5: Variáveis aleatórias contínuas e a distribuição normal

1- Numa certa região, fósseis de pequenos animais são frequentemente encontrados e um arqueólogo estabeleceu o seguinte modelo de probabilidade para o comprimento, em centímetros, desses fósseis:

$$f(x) = \begin{cases} \frac{1}{40}x, & 4 \leq x < 8, \\ -\frac{1}{20}x + \frac{3}{5}, & 8 \leq x < 10, \\ \frac{1}{10}, & 10 \leq x \leq 11, \\ 0, & \text{caso contrário.} \end{cases}$$

- a)** Confirme que a função $f(x)$ é uma função densidade de probabilidade e construa seu gráfico.
b) Encontre o valor esperado para o comprimento dos fósseis da região.

2- O tempo de espera, **em horas**, entre sucessivos motoristas flagrados por um radar que ultrapassam o limite de velocidade, é uma variável aleatória contínua com função de distribuição acumulada

$$F(x) = \begin{cases} 0, & x < 0, \\ 1 - e^{-8x}, & x \geq 0. \end{cases}$$

Determine a probabilidade de o tempo de espera entre sucessivos motoristas ser menor que 12 minutos.

- a)** usando a função de distribuição acumulada de X ;
b) usando a função densidade de probabilidade de X .

3- Medições de sistemas científicos são sempre sujeitas à variação, algumas mais do que outras. Há muitas estruturas para se medir erros, e os estatísticos passam boa parte do tempo modelando esses erros. Suponha que o erro de medição X , de certa quantidade física, seja determinado pela função densidade

$$f(x) = \begin{cases} k(3 - x^2), & -1 \leq x \leq 1, \\ 0, & \text{caso contrário.} \end{cases}$$

- a)** Calcule o valor de k que torna $f(x)$ uma função densidade válida.
b) Determine a probabilidade de que um erro aleatório na medição seja menor que $1/2$.
c) Para essa medição em particular, não é desejável que a *magnitude* do erro (isto é, $|x|$) exceda $0,8$. Qual é a probabilidade de que esta *magnitude* seja excedida?

4- Em uma tarefa em um laboratório, se o equipamento estiver funcionando, a função densidade do resultado observado X , é:

$$f(x) = \begin{cases} 2(1-x), & 0 < x < 1, \\ 0, & \text{caso contrário.} \end{cases}$$

- a)** Determine a $F(x)$.
- b)** Qual é a mediana?
- c)** Qual é o 75° percentil da distribuição?
- d)** Calcule $P(X \leq 1/3)$.
- e)** Qual é a probabilidade de que X exceda 0,5?
- f)** Dado que $X \geq 0,5$, qual é a probabilidade de que X seja menor que 0,75?

5- Suponha que a força que age sobre uma coluna que ajuda a suportar um edifício tenha distribuição normal com média 15,0 kips e desvio padrão 1,25 kips. Qual é a probabilidade de a força:

- a)** Ser no máximo 18 kips?
- b)** Estar entre 10 e 12 kips?
- c)** Diferir de 15,0 kips por no máximo 2 desvios padrão?

6- A dureza Rockwell de um metal é determinada pela pressão de uma ponta rígida na superfície do metal e, sem seguida, pela medição da profundidade de penetração das pontas. Suponha que a dureza Rockwell de uma determinada liga tenha distribuição normal, com média de 70 e um desvio-padrão de 4.

- a)** Se um espécime é aceitável apenas se sua dureza estiver entre 62 e 72, qual é a probabilidade de que um espécime escolhido aleatoriamente tenha dureza aceitável?
- b)** Se o intervalo aceitável para a dureza fosse $(70 - c, 70 + c)$, para qual valor de c teríamos 95% de todos os espécimes com dureza aceitável?

7- Uma enchedora automática de garrafas de vinho está regulada para que o volume médio de líquido nas garrafas seja 750 cm^3 e o desvio padrão seja $7,5 \text{ cm}^3$. Pode-se admitir que a distribuição da variável “volume de líquido” é normal. Qual a probabilidade de uma garrafa selecionada ao acaso conter um volume de líquido menor do que 740 cm^3 ?

8- Um teste de aptidão para o exercício e certa profissão exige uma sequência de operações a serem executadas rapidamente, uma após a outra. Para passar no teste, o candidato deve completá-lo em 80 minutos no máximo. Admitindo que o tempo para completar o teste seja uma variável aleatória $N(90, 400)$, pede-se:

- a)** Qual a porcentagem dos candidatos que tem chance de ser aprovados?
- b)** Os 5% mais rápidos receberão um certificado especial. Qual o tempo máximo para fazer jus a tal certificado?

9- Um fabricante de alimentos enlatados sabe, por experiência passada, que o tempo de prateleira de seus alimentos é uma variável aleatória que tem distribuição normal com média de 600 dias e desvio padrão de 110 dias.

- a)** Qual é a probabilidade de um alimento, escolhido ao acaso, durar: i. mais de 750 dias? ii. entre 400 e 800 dias?
- b)** O comprador recebe uma garantia do fabricante de 320 dias, isto é, o fabricante substituirá todos os alimentos que estragarem antes de 320 dias. Sabendo-se que são fabricados 20000 enlatados por mês, quantos enlatados o fabricante terá que substituir mensalmente, devido à garantia dada?

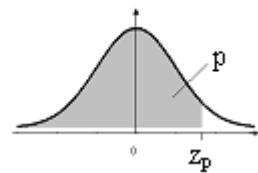
10- Uma máquina de empacotar determinado produto apresenta variações de peso com desvio padrão de 20g. Assumindo a distribuição normal para o peso dos pacotes, em quanto deve ser regulado o peso médio do pacote, para que apenas 10% dos pacotes tenham menos de 400g?

11- A distribuição da resistência de resistores de um tipo específico é normal. Sabe-se que 10% de todos os equipamentos apresentam resistência maior que 10,256 ohms e 5% tem resistência menor que 9,671 ohms. Quais são os valores da média e do desvio padrão da distribuição das resistências?

GABARITO

- 1- b)** 7,45 **2-** 0,7981 **3- a)** $k = 3/16$; **b)** 99/128; **c)** 0,164
4- b) 0,2928; **c)** 0,5 **d)** $5/9 = 0,55555$; **e)** $1/4 = 0,25$; **f)** $3/4$
5- a) 0,9918; **b)** 0,0082 **c)** 0,9544 **6- a)** 0,6687; **b)** 7,84 **7-** 0,0918
8- a) 31%; **b)** 57 min **9- a)** i. 0,0869, ii. 0,9312; **b)** 108 **10-** 425,7g.
11- $\mu = 10$; $\sigma = 0,2$

Tabela I: Distribuição Normal Padrão Acumulada



Fornece $\Phi(z) = P(-\infty < Z \leq z)$, para todo z , de 0,01 em 0,01, desde $z = 0,00$ até $z = 3,59$
A distribuição de Z é Normal(0;1)

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990
3,1	0,9990	0,9991	0,9991	0,9991	0,9992	0,9992	0,9992	0,9992	0,9993	0,9993
3,2	0,9993	0,9993	0,9994	0,9994	0,9994	0,9994	0,9994	0,9995	0,9995	0,9995
3,3	0,9995	0,9995	0,9995	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9997
3,4	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9998
3,5	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998

Obs.: Se $z < 0$, então $\Phi(z) = P(-\infty < Z \leq z) = 1 - \Phi(-z)$.

5.5 Aproximação de distribuições de probabilidade discretas à normal

A distribuição normal é, com frequência, uma boa aproximação para uma distribuição de probabilidades de uma variável aleatória discreta quando esta assume uma forma simétrica. Do ponto de vista teórico, algumas distribuições convergem para a normal conforme seus parâmetros se aproximam de certos limites. A distribuição normal é uma distribuição aproximada conveniente porque sua função de distribuição acumulada é facilmente tabulada. Vale destacar que algumas distribuições contínuas também podem ser aproximadas por uma normal.

5.5.1 Binomial → Normal

O cálculo de probabilidades em algumas distribuições binomiais pode ser extremamente trabalhoso. Por exemplo, considere $n = 135$ e $p = 0,6$. Qual a probabilidade de que $Y \geq 98$, por exemplo?

Para resolver este problema teríamos que calcular:

$$P[Y \geq 98] = P[Y = 98] + P[Y = 99] + \dots + P[Y = 135]$$

Na realidade, se considerássemos que a variável aleatória Y pudesse, apesar de discreta, ser razoavelmente bem aproximada por uma distribuição contínua normal, esse cálculo se tornaria bem mais simples usando a variável Z .

Exercício: Instale o aplicativo *probability distributions* em seu *smartphone* e selecione a distribuição binomial.

a) Depois coloque o valor $n = 10$ e $p = 0,6$ e aperte OK, veja a distribuição de probabilidades gerada. Agora aumente o valor de n , por exemplo coloque $n = 30$ e $p = 0,6$. Agora faça $n = 200$ e $p = 0,6$. O que aconteceu com a distribuição de probabilidades?

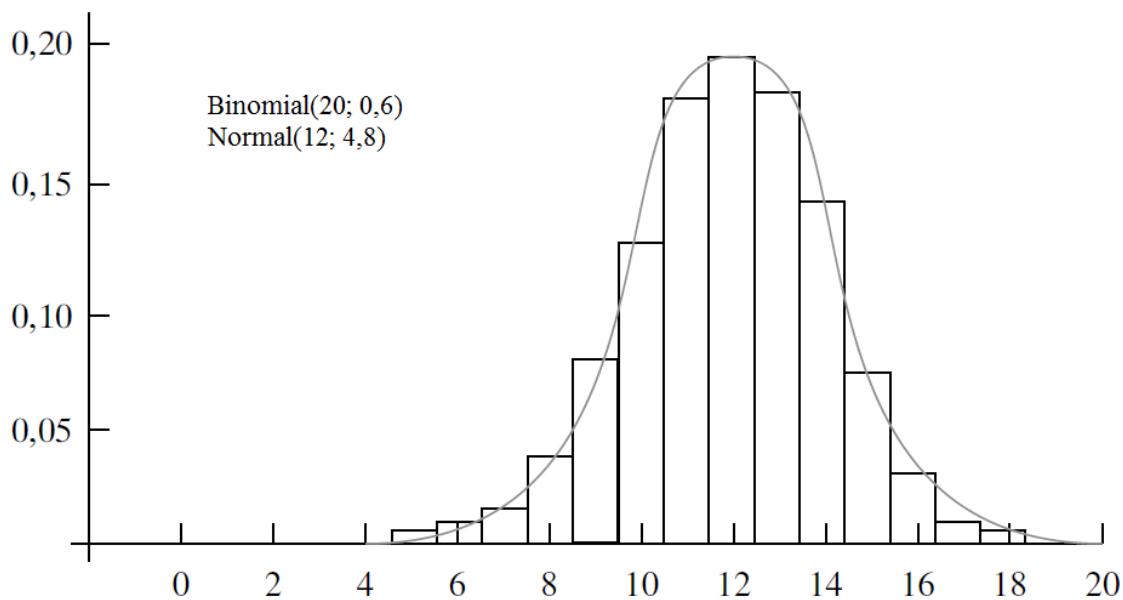
b) Varie agora os valores de p e veja o que acontece, por exemplo: $n = 20$ e $p = 0,6$, depois faça $n = 20$ e $p = 0,06$. O que aconteceu com a distribuição de probabilidades?

c) Com base nas respostas dos itens **a)** e **b)**, quando você pode afirmar que a distribuição normal é uma boa aproximação para um distribuição binomial?

Basta relembrar que se $X \sim B(n, p)$ então $E[X] = \mu = n \times p$ e $VAR[X] = \sigma^2 = n \times p \times (1 - p)$.

A figura a seguir exibe um gráfico de barras de probabilidade (distribuição de probabilidade) para a distribuição binomial com $n = 20$ e $p = 0,6$. Uma curva normal com valor médio e desvio padrão iguais aos valores correspondentes para a distribuição binomial ($\mu = np = 20(0,6) = 12$ e $\sigma^2 = np(1 - p) = 4,8$) foi sobreposta na distribuição de probabilidades da Binomial. A curva normal oferece uma aproximação muito boa, especialmente na parte central da figura.

A área de qualquer retângulo (probabilidade de qualquer valor X particular), exceto aqueles nas extremidades das caudas, pode ser aproximada com precisão pela área correspondente da curva normal.



Um aspecto que deve ser considerado é o da descontinuidade das variáveis discretas. Uma correção de continuidade deve ser realizada antes do cálculo das probabilidades requeridas. Para entender melhor essa correção, observe, na Figura acima, que os valores da binomial, diga-se Y , representam os pontos médios dos retângulos e que as áreas relativas a cada retângulo correspondem às probabilidades associadas aos valores de Y . Assim, $P(Y = 10)$, por exemplo, significa que a probabilidade poderia ser estimada pela área sob a curva normal que aproxima o gráfico de barras no intervalo de 9,5 a 10,5. Definindo X a variável normal com média $\mu = np$ e $\sigma^2 = np(1 - p)$, a probabilidade aproximada pela normal seria dada pela relação: $P(Y = y) \cong P(y - 0,5 < X < y + 0,5)$. A soma ou subtração do valor 0,5 é a comumente chamada de correção de continuidade.

Dizemos que a aproximação da distribuição de probabilidades de uma variável aleatória binomial pela distribuição normal é tão boa quanto maior for o n e mais próximo de 0,5 estiver o valor de p .

Na prática, a aproximação é adequada se $n \times p \geq 10$ e $n \times p \times (1 - p) \geq 10$. Para valores menores que esses, a distribuição terá muita inclinação para que a curva normal forneça uma aproximação precisa.

EXEMPLO 1: Um sistema é formado por 100 componentes, cada um dos quais com confiabilidade de 0,85 (probabilidade de funcionamento). Se esses componentes funcionam independentes uns dos outros e se o sistema completo funciona adequadamente quando pelo menos 80 componentes funcionam, qual a confiabilidade do sistema?

Y : número de componentes que funcionam

$$Y \sim B(n = 100, p = 0,85)$$

Queremos encontrar:

$$P[Y \geq 80] = P[Y = 80] + P[Y = 81] + P[Y = 82] + \dots + P[Y = 100]$$

Perceba que o valor esperado e a variância são dados por:

$$E[Y] = \mu = np = 100 \times 0,85 = 85$$

$$Var[Y] = \sigma^2 = n \times p \times (1 - p) = 100 \times 0,85 \times 0,15 = 12,75$$

Como $np = 85 \geq 10$ e $np(1 - p) = 12,75 \geq 10$, então podemos utilizar a distribuição normal para resolver este problema.

Agora basta considerar uma variável aleatória normal, X , com $E[X] = E[Y] = \mu = 85$ e $Var[X] = Var[Y] = \sigma^2 = 12,75$, que tem-se uma aproximação Normal para a variável Y .

$$P[Y \geq 80] = P[80 \leq Y \leq 100] \cong P[79,5 \leq X \leq 100,5] = P[-1,54 \leq Z \leq 4,34]$$

Portanto, a probabilidade de que o sistema funcione, ou seja, sua confiabilidade, é, aproximadamente:

$$P[79,5 \leq X \leq 100,5] = P[-1,54 \leq Z \leq 4,34] = 0,9382 \quad \text{ou} \quad 93,82\%$$

A probabilidade exata, calculada diretamente da distribuição binomial é 93,36%, ficando evidente a qualidade da aproximação da distribuição binomial pela normal.

5.5.2 Poisson → Normal

A aproximação normal à Poisson é realizada nos mesmos moldes da aproximação realizada à binomial. A variável Poisson possui média e variância definidas por $\mu = \sigma^2 = \lambda$. Controvérsias são encontradas na literatura para a definição de qual deve ser o valor mínimo de λ para que a aproximação seja considerada adequada. São sugeridos os valores de $\lambda > 7$, ou $\lambda > 15$, ou ainda, $\lambda > 25$ para se alcançarem boas aproximações.

Assim como no exercício 1, verifique também no aplicativo *probability distributions* o que acontece com a distribuição de probabilidades da variável aleatória conforme varia o parâmetro λ .

EXEMPLO 2: Suponha que a média estimada de um tipo de bactéria é igual a 27,6 por cm^2 . Para utilizar determinado produto o nível de contaminação da lâmina deve ser intenso, com mais de 35 bactérias por cm^2 . Determine a probabilidade de que o produto seja usado, isto é, de que sejam encontradas mais de 35 bactérias por cm^2 nesta lâmina.

Y : número de bactérias

$$Y \sim Poisson(\lambda = 27,6)$$

Queremos encontrar:

$$P[Y > 35] = P[Y = 36] + P[Y = 37] + P[Y = 38] + \dots$$

Ou utilizando a ideia do complemento por:

$$P[Y > 35] = 1 - P[Y = 0] + P[Y = 1] + P[Y = 2] + \dots + P[Y = 35]$$

Mas mesmo com o complemento, teríamos muito trabalho. A distribuição Normal pode também ser utilizada para o cálculo aproximado da distribuição Poisson. Para tanto, deve-se calcular a esperança e a variância desta variável aleatória:

$$\begin{aligned} E[Y] &= \mu &= \lambda = 27,6 \\ Var[Y] &= \sigma^2 &= \lambda = 27,6 \end{aligned}$$

Agora basta considerar uma variável aleatória normal, X , tal que $E[Y] = E[X] = \mu = 27,6$ e $Var[Y] = Var[X] = \sigma^2 = 27,6$, e tem-se uma aproximação Normal para a variável aleatória discreta Y .

$P[Y > 35]$ é equivalente a $P[X > 35,5]$ pela correção de continuidade. Assim, quando X vale 35,5, Z vale:

$$Z = \frac{35,5 - 27,6}{5,25} = 1,50.$$

Portanto a probabilidade de que sejam encontradas mais que 35 bactérias por cm^2 na lâmina é dada por:

$$P[Y > 35] \cong P[X > 35,5] = P[Z > 1,50] = 0,0668 \quad \text{ou} \quad 6,68\%$$

Pelo cálculo exato desta probabilidade na distribuição de poisson tem-se 0,0708, ficando clara a qualidade da aproximação.

EXERCICIO: Uma máquina produz parafusos, dos quais 10% são defeituosos. Usando a aproximação da distribuição binomial pela normal, determinar a probabilidade de uma amostra formada ao acaso de 400 parafusos produzidos pela máquina serem defeituosos:

- a) no máximo 30; 0,0571
- b) entre 30 e 50 (inclusive os extremos); 0,9198
- c) mais de 35 e menos de 45; 0,5467
- d) mais de 55. 0,0049

EXERCICIO: O número médio de aviões que pousam em um aeroporto movimentado de uma capital é de 3 a cada 2 minutos. Com base nestas informações calcule:

- a) a probabilidade aproximada de que numa hora, selecionada ao acaso, ocorram pelo menos 75 aterrizagens neste aeroporto. 0,9484
- b) a probabilidade do aeroporto receber entre 100 e 120 (inclusive) aviões em uma determinada hora. 0,1580

5.6 A distribuição Gama

Existem várias situações práticas nas quais a variável em estudo não está definida em todo o conjunto dos reais, ou principalmente não possui padrão simétrico. Nestes casos não pode ser utilizada a distribuição normal.

A distribuição Gama define uma família de distribuições, com comportamento assimétrico, definidas no intervalo $[0, +\infty]$.

As principais aplicações da a distribuição gama são: tempo de vida, tempo de espera, teoria da confiabilidade e teoria das filas.

A variável aleatória contínua X tem distribuição gama, com parâmetros $\alpha > 0$ e $\beta > 0$, se sua função de densidade de probabilidades for dada por:

$$f(x) = \begin{cases} \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-\frac{x}{\beta}}, & \text{se } x > 0; \\ 0, & \text{caso contrário.} \end{cases}$$

Notação: $X \sim Gama(\alpha, \beta)$.

A figura a seguir ilustra os gráficos da fdp gama para algumas combinações (α, β) . O parâmetro β é denominado parâmetro de escala porque os valores diferentes de 1 esticam ou comprimem a fdp na direção de x .

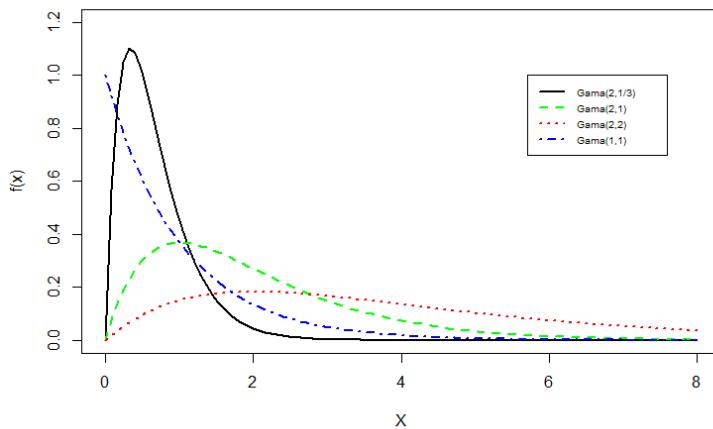


Figura 14: Distribuição gama variando os parâmetros.

A média e a variância de uma variável aleatória X com distribuição gama são:

$$E[X] = \alpha \times \beta \quad VAR[X] = \alpha \times \beta^2$$

A distribuição gama possui função de densidade de probabilidade muito complexa, tornando-se complicado trabalhar com ela sem o uso de softwares estatísticos. Nesse sentido iremos conhecer casos particulares dela.

5.7 A distribuição exponencial

A distribuição exponencial é um caso particular da distribuição gama, em que $\alpha = 1$ e $\beta = \frac{1}{\lambda}$.

Assim a variável aleatória contínua X possui distribuição exponencial, com $\lambda > 0$ se sua função de densidade de probabilidades for dada por:

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0; \\ 0, & \text{caso contrário.} \end{cases}$$

Notação: $X \sim E(\lambda)$.

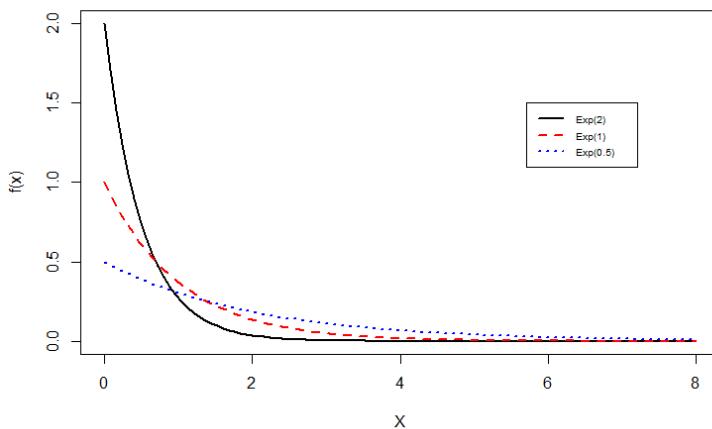


Figura 15: Distribuição exponencial variando o valor de λ .

A média e a variância de uma variável aleatória X com distribuição exponencial são:

$$E[X] = \frac{1}{\lambda} \quad VAR[X] = \frac{1}{\lambda^2}$$

A função de distribuição acumulada de uma variável aleatória exponencial é dada por:

$$F(x) = P[X \leq x] = \int_0^x \lambda e^{-\lambda t} dt = \frac{\lambda e^{-\lambda t}}{-\lambda} \Big|_0^x = -e^{-\lambda x} + 1$$

Portanto:

$$F(x) = \begin{cases} 1 - e^{-\lambda x}, & x \geq 0; \\ 0, & \text{caso contrário.} \end{cases}$$

A distribuição exponencial modela o tempo de espera entre ocorrências de processos Poisson. A distribuição gama modela o tempo de espera até a α -gésima ocorrência.

Exemplo: Dez minutos é o tempo médio entre as chamadas telefônicas para um escritório. Podemos considerar o tempo entre as chamadas como uma variável aleatória exponencial.

a) Qual a probabilidade de a primeira chamada ocorrer entre 7 e 12 minutos após abrir o escritório.

X: tempo entre chamadas
 $E[X] = 10 = \frac{1}{\lambda} \Rightarrow \lambda = 0,1$ $X \sim E(0,1)$

Assim:

$$P[7 < X < 12] = \int_7^{12} \lambda e^{-\lambda x} dx = \int_7^{12} 0,1 e^{-0,1x} dx = \frac{0,1 e^{-0,1x}}{-0,1} \Big|_7^{12} = -e^{-1,2} + e^{-0,7} = 0,1954$$

Naturalmente poderia ser feito também:

$$P[7 < X < 12] = F[12] - F[7] = 1 - e^{-0,1 \times 12} - (1 - e^{-0,1 \times 7}) = -e^{-1,2} + e^{-0,7} = 0,1954$$

Logo a probabilidade procurada é 19,54%.

b) Qual a probabilidade de haver 2 chamadas em 30 minutos?

Y: número de chamadas em 30 minutos
 $\lambda = 3$ $Y \sim P(3)$

$$P[Y = 2] = \frac{e^{-3} 3^2}{2!} = 0,2240 = 22,40\%$$

A falta de memória da exponencial

A distribuição exponencial possui a seguinte propriedade:

$$P[X \geq t + t_0 | X \geq t_0] = P[X \geq t]$$

Você deixou uma máquina funcionando por um tempo t_0 , saiu e depois voltou a observar a máquina por mais um tempo t . A probabilidade da máquina estragar nesse tempo t é a mesma independente se ela já esteve trabalhando por um tempo t_0 .

Em outras palavras, uma variável aleatória modelada pela distribuição exponencial não possui memória, isto é, não apresenta sinais de desgaste.

5.8 A distribuição Weibull

Se os componentes se deterioram, ou melhoram, ao longo do tempo, então a distribuição Weibull é a mais indicada para modelar esta variável.

A variável aleatória contínua X tem distribuição Weibull, com parâmetros $\alpha > 0$ e $\beta > 0$ se sua função de densidade de probabilidade for dada por:

$$f(x) = \begin{cases} \frac{\alpha}{\beta} \left(\frac{x}{\beta}\right)^{\alpha-1} e^{-(\frac{x}{\beta})^\alpha} & x > 0; \\ 0, & \text{caso contrário.} \end{cases}$$

Notação: $X \sim W(\alpha, \beta)$

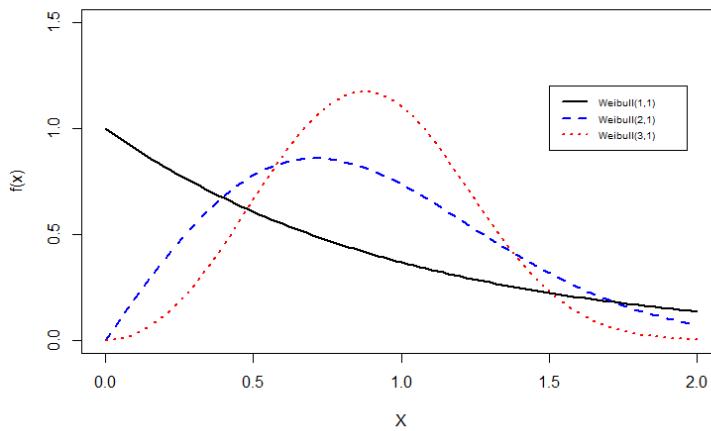


Figura 16: Distribuição Weibull variando o parâmetro de forma α .

Na distribuição Weibull α é conhecido como parâmetro de forma e β é chamado parâmetro de escala. Veja que se $\alpha = \beta = 1$ a distribuição Weibull se reduz a uma distribuição exponencial.

A média e a variância da distribuição Weibull são expressões complexas que não serão apresentadas neste material.

A função de distribuição acumulada da Weibull é dada por:

$$F(x) = P[X \leq x] = \begin{cases} 1 - e^{-(\frac{x}{\beta})^\alpha}, & x > 0; \\ 0, & \text{caso contrário.} \end{cases}$$

Exemplo: O tempo de prateleira (tempo até estragar, vida útil), em dias, de um certo tipo de alimento *in natura* possui distribuição Weibull com $\alpha = 2$ e $\beta = 10$. Qual é a probabilidade que este alimento estrague antes de 8 dias?

X: tempo de prateleira do alimento
 $\alpha = 2$ e $\beta = 10$ $X \sim W(10; 2)$

$$P[X \leq 8] = F(8) = 1 - e^{-(\frac{8}{10})^2} = 1 - 0,5273 = 0,4727 = 47,27\%$$

5.9 LISTA DE EXERCÍCIOS 6: Distribuições de Probabilidade Contínuas: Exponencial, Gama e Weibull.

1- Seja X o tempo entre duas chegadas sucessivas no guichê de atendimento rápido de um banco local. Se X possui distribuição exponencial com $\lambda = 1$ calcule os itens a seguir:

- a)** O tempo esperado entre duas chegadas sucessivas.
- b)** O desvio padrão do tempo entre chegadas sucessivas.
- c)** $P(X \leq 4)$
- d)** $P(2 \leq X \leq 5)$

2- Diversas experiências com determinado tipo de LED, utilizados em televisores, indicam que a distribuição exponencial sugere um bom modelo para cálculo do tempo de vida, ou tempo até a falha do LED. Suponha que o tempo médio seja 25.000 horas. Qual é a probabilidade de:

- a)** um televisor com este LED durar pelo menos 20.000 horas? No máximo 30.000 horas? Entre 20.000 e 30.000 horas?
- b)** o tempo de vida de um televisor exceder o valor médio em mais de 2 desvios padrão? Mais de 3 desvios padrão?
- c)** Qual a mediana do tempo de vida destes televisores?

3- Seja X uma variável aleatória contínua, exponencialmente distribuída, com função de densidade de probabilidade dada por:

$$f(x) = \left(\frac{c+44}{6}\right) e^{-2cx}, \quad x > 0$$

- a)** Determine o valor da constante c e o valor de λ .
- b)** Calcule $P(8\mu - 3\sigma < X < 10\mu + 6\sigma)$, onde $\mu = E[X]$ e $\sigma = \sqrt{VAR[X]}$.

4- O tempo entre a chegada de ônibus em certo ponto de uma avenida é representado por uma variável aleatória exponencial de média 10 minutos.

- a)** Determine a probabilidade de uma pessoa ter de esperar mais de uma hora por um ônibus.
- b)** Qual é a probabilidade de passar pelo menos 2 ônibus nesse ponto em 30 minutos?

5- Suponha que um sistema contém certo tipo de componente cujo tempo, em anos até a falha é dado pela variável aleatória T . Esta variável é muito bem modelada pela distribuição exponencial, com tempo médio até a falha de 5 anos. Se cinco destes componentes são instalados em sistemas diferentes, qual é a probabilidade de que pelo menos dois sistemas estejam funcionando ao final de 8 anos?

6- O tempo de resposta de computadores é uma importante aplicação das distribuições gama e exponencial. Suponha que um estudo sobre certo sistema de computador revele que o tempo de resposta, em segundos, tem uma distribuição exponencial com média de 3 segundos.

- a)** Qual a probabilidade de que o tempo de resposta exceda 5 segundos?
- b)** Qual a probabilidade de que o tempo de resposta exceda 10 segundos?
- c)** Dado que a resposta não aconteceu nos primeiros 10 segundos, qual a probabilidade de o computador não responder nos próximos 5 segundos?
- d)** Dado que a resposta não aconteceu nos primeiros 10 segundos, qual a probabilidade de o computador responder nos próximos 5 segundos?

7- A vida útil em horas, de uma broca de perfuração, em uma operação mecânica tem distribuição Weibull com $\alpha = 0,5$ e $\beta = 0,7071$. Determine a probabilidade de que a broca falhará em menos de 5 horas de uso. Explique porque a natureza desta variável aleatória não combina com uma distribuição exponencial.

8- Nos últimos anos, a distribuição de Weibull tem sido usada para modelar emissões de poluentes de vários motores. Seja X o valor da emissão de NO_x (g/gal) a partir de certo tipo de motor de quatro tempos selecionado aleatoriamente e suponha que X possua uma distribuição de Weibull com $\alpha = 1,2$ e $\beta = 8,95$. Calcule a probabilidade de o valor de emissão de NO_x deste motor ser inferior a 10 (g/gal). Qual a probabilidade do valor de emissão de NO_x deste motor estar entre 10 e 25 (g/gal)?

9- Seja X o limite de resistência à tração (ksi) de um corpo de prova de aço selecionado aleatoriamente que apresenta “fragilidade ao frio” em baixas temperaturas. Suponha que X tenha distribuição de Weibull com $\alpha = 1.3$ e $\beta = 10$.

- a)** Qual a probabilidade de X ser no máximo 15 ksi?
- b)** Qual a probabilidade de X estar entre 10 e 15 ksi?
- c)** Qual é o valor da mediana da distribuição da resistência?

10- O tempo de vida (em horas) de um aparelho de imagem por ressonância magnética é modelado por uma distribuição Weibull, com $\beta = 22.36$ e $\alpha = 2$. Dado que o aparelho já está funcionando há 10 horas, qual é a probabilidade dele estragar nas próximas 5 horas?

GABARITO

- 1: **a)** 1 **b)** 1 **c)** 0,982 **d)** 0,129
- 2: **a)** 0,449; 0,699 0,148 **b)** 0,05 0,018 **c)** aproximadamente 17329 horas
- 3: **a)** $c=4$, $\lambda = 8$ **b)** 0,0067 4: **a)** 0,0025 **b)** 0,8008 5: 0,2664
- 6: **a)** 0,1889 **b)** 0,0357 **c)** 0,1889 **d)** 0,8111 7: 0,9299 8: 0,6809 e 0,2867
- 9: **a)** 0,8162 **b)** 0,1841 **c)** 7,5432 10: 0,2212 ou 22,12%

5.10 Amostragem da distribuição normal: distribuições amostrais

Em estatística estamos interessados em fazer inferência acerca de uma população, baseando-se nas informações presentes em uma amostra aleatória da população. Em certas situações, não estamos interessados na própria variável em si, mas nas características dela.

Por exemplo um engenheiro de alimentos está interessado no volume de enchimento de uma lata de refrigerante de 350 ml. Para tanto, coleta uma amostra de 25 latas e nesta amostra calcula o volume médio, obtendo $\bar{X} = 348,7$. Provavelmente ele decidirá que a média da população é $\mu = 350$, embora a média amostral não tenha dado exatamente este valor.

O controle de qualidade da empresa de refrigerante deseja que seu produto não possua uma variabilidade muito grande. Por exemplo, o lote de refrigerantes só pode ser vendido se $\sigma < 5\text{ml}$. Nas 25 latas observadas o engenheiro encontra $S = 6$. Será um indício para rejeitar o lote?

Da mesma forma, se estivermos interessados em saber a proporção de pessoas que são favoráveis a determinado candidato em uma pesquisa de intenção de votos, iremos observar a proporção encontrada em uma amostra, $\hat{p} = 52\%$. É um indício de que não haverá 2º turno?

Obviamente os valores da população (μ , σ e p) não variam. Mas note que \bar{X} , S e \hat{p} são uma função dos valores observados e, naturalmente, variam de amostra para amostra. Assim estes valores também podem ser classificados como um variável aleatória, que por consequência, possuem uma distribuição de probabilidades, com base na qual serão tomadas as decisões para a realizar a inferência sobre a população.

Antes de prosseguir algumas definições são necessárias:

parâmetro: medida utilizada para descrever uma característica da população.

Ex: μ, σ^2, p

estimador: também chamado de estatística é a expressão algébrica utilizada para obter um valor aproximado do parâmetro.

$$\text{Ex: } \bar{X} = \frac{\sum_{i=1}^n x_i}{n} \quad S^2 = \frac{\sum_{i=1}^n (x_i - \bar{X})^2}{n-1} \quad \hat{p} = \frac{\text{favoraveis}}{\text{possíveis}}$$

estimativa: é o valor numérico determinado pelo estimador.

$$\text{Ex: } \bar{X} = 248,7 \quad S^2 = 36 \quad \hat{p} = 0,40$$

A distribuição de probabilidades de um estimador é chamada de **distribuição amostral**, pois o estimador depende apenas da amostra observada. A distribuição amostral de um estimador depende da distribuição da população, do tamanho da amostra e do método de escolha da amostra.

A distribuição amostral fundamenta-se na obtenção de todas as amostras possíveis de mesmo tamanho da população. É como se fossem realizadas todas as amostras possíveis da população e construída a distribuição a partir do histograma amostral por exemplo. Felizmente as distribuições dos principais estimadores são conhecidas.

5.10.1 A distribuição t de Student

A distribuição amostral da média, conforme veremos mais adiante, possui distribuição normal quando o valor de sigma é conhecido, ou quando a amostra for suficientemente grande, pois neste caso S é um bom estimador de σ e podemos considerar $\sigma = S$. Mas, em muitos cenários, σ não é conhecido e uma estimativa S deve ser calculada na mesma amostra que produziu a média amostral \bar{X} . Como resultado, na padronização tem-se a seguinte estatística:

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}.$$

Dizemos que T possui distribuição t de Student com $\nu = n - 1$ graus de liberdade. Notação: $T \sim t_{(\nu)}$. Essa distribuição tem grande aplicação em estatística e recebeu esse nome em homenagem ao pesquisador W. S. Gossett que realizou uma importante publicação a seu respeito em 1908 usando o pseudônimo de Student.

A figura a seguir ilustra diferentes densidades da t de Student, considerando diferentes valores de ν juntamente com a distribuição normal padrão. A distribuição t aproxima-se da normal padrão à medida que ν aumenta. A distribuição t tem caudas mais leves que a distribuição normal, e esse fato fica mais evidenciado à medida que os graus de liberdade diminuem.

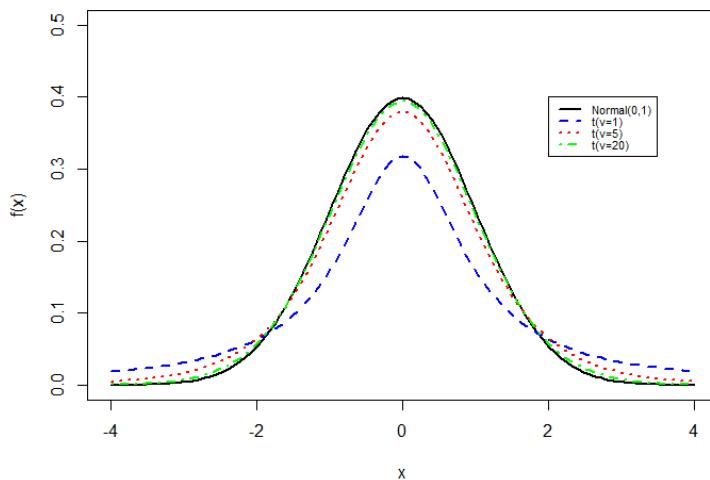


Figura 17: Distribuição t de Student variando o valor de ν .

Costuma-se denotar por $t_{\alpha;\nu}$ o valor t acima do qual encontramos uma área igual a α com ν graus de liberdade. Já que a distribuição é simétrica em torno da média 0, temos $t_{1-\alpha;\nu} = -t_{\alpha;\nu}$; ou seja, $t_{0,95;\nu} = -t_{0,05;\nu}$. Por exemplo: $t_{0,05;15} = 1,75$. Já $t_{0,95;15} = -1,75$.

A distribuição t é muito utilizada em problemas que lidam com inferência sobre uma média populacional ou mesmo em problemas que envolvem amostras comparativas, isto é, em casos em que deseja-se determinar se as médias de duas amostras são significativamente diferentes. O uso da distribuição t de Student requer que a amostra seja originada de uma população normal.

5.10.2 A distribuição χ^2

A distribuição qui-quadrado possui várias aplicações em estatística. Uma delas é a de propiciar mecanismos para a realização de inferências sobre o parâmetro σ^2 de uma população normal. Outra aplicação refere-se aos testes de falta de ajuste de um modelo teórico aos dados observados em um experimento ou levantamento amostral.

Se S^2 é a variância de uma amostra aleatória de tamanho n , retirada de uma população normal, com variância σ^2 , então a estatística:

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

tem distribuição qui-quadrado com $\nu = n - 1$ graus de liberdade. Notação: $\sim \chi_{(\nu)}^2$.

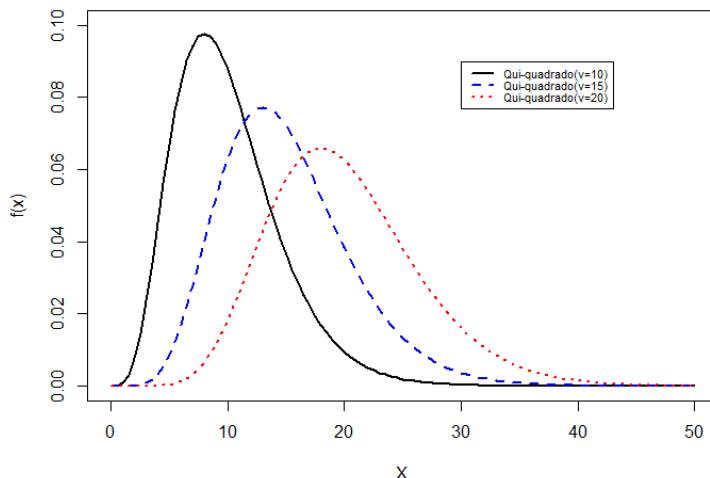


Figura 18: Distribuição χ^2 variando o valor de ν .

Costuma-se denotar por $\chi_{\alpha;\nu}^2$ o valor da distribuição χ^2 acima do qual encontramos uma área α com ν graus de liberdade.

Por exemplo para 7 graus de liberdade ($\nu = 7$), por exemplo, o valor de χ^2 que deixa uma área de 0,05 à direita, é $\chi_{0,05;7}^2 = 14,067$. Como pode ser observado na figura abaixo a distribuição χ^2 não é simétrica. Devido à falta de simetria a relação obtida nas distribuições t e normal não são válidas nesta distribuição. Assim temos, por exemplo $\chi_{0,95;7}^2 = 2,167$.

5.10.3 A distribuição F de Snedecor

De forma geral, a distribuição F é utilizada para comparar variâncias de duas populações normais. Esta comparação é feita dividindo uma pela outra, por este motivo, a distribuição F é também chamada de distribuição da razão de variâncias.

A mais importante aplicação da distribuição F é o seu emprego na análise de experimentos. Nesse caso, o investigador científico tem por objetivo comparar os efeitos de dois ou mais tratamentos sob determinadas condições. A hipótese de que a variabilidade entre os tratamentos é maior do que a variabilidade dentro dos tratamentos é testada usando a distribuição de probabilidade F .

Sejam duas populações normais, com amostras de tamanhos n_1 e n_2 . A estatística:

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{S_1^2}{S_2^2}$$

possui distribuição F sob $H_0 (\sigma_1^2 = \sigma_2^2)$, com $\nu_1 = n_1 - 1$ e $\nu_2 = n_2 - 1$ graus de liberdade.
Notação: $\sim F_{(\nu_1, \nu_2)}$.

A curva da distribuição F depende de dois parâmetros ν_1 e ν_2 , que por sua vez dependem do tamanho das amostras. Identificados estas duas informações, podemos caracterizar a curva.

Assim como nas demais distribuições, $F_{\alpha; \nu_1; \nu_2}$ é o valor da distribuição F que possui probabilidade α acima dele, com graus de liberdade ν_1 e ν_2 . Por exemplo, $F_{0,05;6;10} = 3,21$, como a distribuição F também não é simétrica $F_{0,95;6;10} = 0,24$.

A figura a seguir representa curvas da distribuição F para três diferentes combinações de parâmetros.

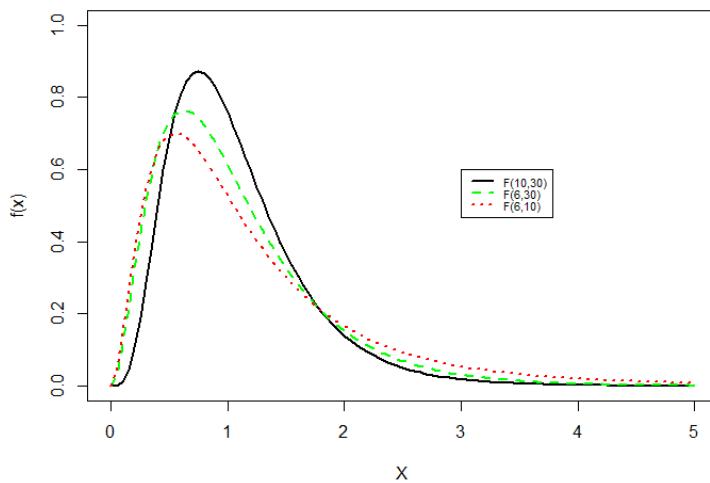


Figura 19: Distribuição F variando os valores de ν_1 e ν_2 .

5.10.4 Distribuição amostral de \bar{X}

Perceba que \bar{X} é uma variável aleatória contínua que pode assumir inúmeros valores. Precisamos então conhecer sua fdp.

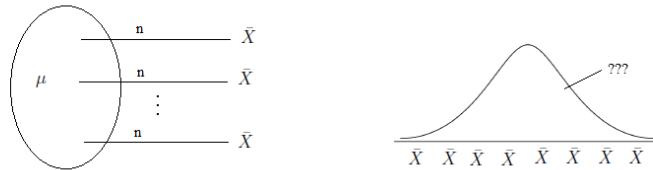


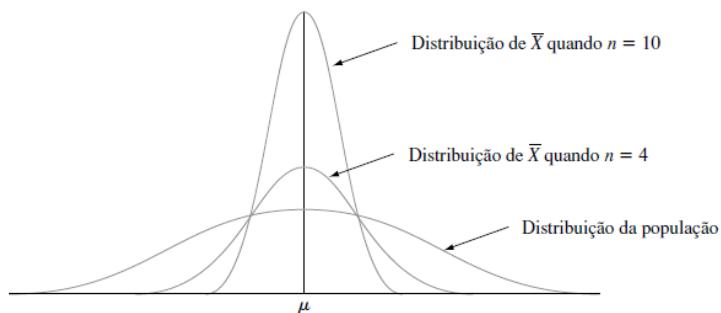
Figura 20: Ilustração da distribuição de \bar{X} .

É intuitivo de se pensar que a distribuição de \bar{X} dependa da distribuição de X . Assim temos dois cenários a serem considerados.

1º- X possui distribuição normal

Teorema: Se $X \sim N(\mu, \sigma^2)$, e podemos retirar amostras de tamanho n desta população (x_1, x_2, \dots, x_n) então, seja $\bar{X} = \frac{\sum_{i=1}^n x_i}{n}$, temos que $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$.

dem: basta verificar que $E[\bar{X}] = \mu$ e $VAR[\bar{X}] = \frac{\sigma^2}{n}$

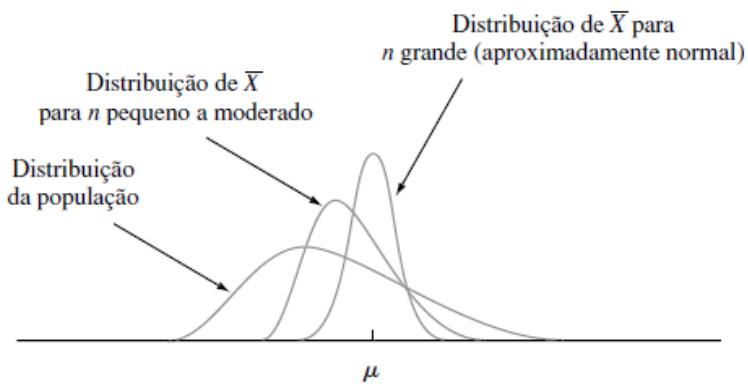


Mas e se a distribuição de probabilidades da população de origem for desconhecida? Nem sempre conhecemos tal distribuição.

2º- X não possui distribuição normal

Para identificar a distribuição de \bar{X} usaremos um dos mais importantes teoremas da estatística, o Teorema Central do Limite - TCL.

TCL: Seja x_1, x_2, \dots, x_n uma amostra de uma variável aleatória com distribuição de probabilidade QUALQUER, cuja média é μ e a variância é σ^2 . Então para n grande ($n \geq 30$), temos que a estatística $\bar{X} = \frac{\sum_{i=1}^n x_i}{n}$ possui distribuição aproximadamente normal, isto é, $\bar{X} \approx N(\mu, \frac{\sigma^2}{n})$.



Corolário: Se $\bar{X} \approx N(\mu, \frac{\sigma^2}{n})$, então a variável aleatória Z definida por $Z_n = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$ possui distribuição normal padrão, isto é, $Z_n \sim N(0, 1)$.

Exemplo: O tempo de prateleira de certo alimento perecível tem média de 800 horas e desvio padrão de 40 horas. Qual é a probabilidade de que uma amostra de 49 destes alimentos tenha tempo de prateleira médio inferior a 790 horas?

X: Tempo de prateleira do produto

$$\mu = 800 \quad \sigma = 40 \quad X \sim ?$$

\bar{X} : Tempo de prateleira médio de 49 produtos

$$\mu = 800 \quad \sigma = \frac{40}{\sqrt{49}} \quad \text{Pelo TCL } \bar{X} \sim N(800, \frac{40^2}{49})$$

$$P[\bar{X} < 790] = P[Z < -1,75] = 0,0401 = 4,01\%$$

$$Z_c = \frac{790 - 800}{\frac{40}{\sqrt{49}}} = -1,75$$

5.10.5 Distribuição amostral de \hat{p}

Assim como no caso da média, \hat{p} é uma variável aleatória contínua que precisamos identificar sua fdp.

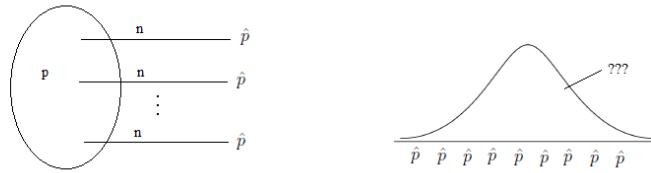


Figura 21: Ilustração da distribuição de \hat{p} .

Seja uma amostra na qual foi observada a característica A. Lembre-se que o estimador da proporção de pessoas com a característica A é calculado da seguinte maneira: $\hat{p} = \frac{\text{favoraveis}}{\text{possíveis}} = \frac{n_a}{n}$.

Se definirmos a variável aleatória X da seguinte maneira: **X: pessoa possuir a característica A**, e atribuirmos 1 para o indivíduo que possui a característica A e 0 para o indivíduo que não possui a característica, temos que $X \sim Bernoulli(p)$.

(OBS: A distribuição Bernoulli não foi abordada neste material, mas ela pode ser definida como uma realização da variável Binomial, em outras palavras, uma $Bernoulli(p)$ é o equivalente a uma $Binomial(1,p)$. A distribuição Binomial pode ser definida também como uma soma de n variáveis aleatórias Bernoulli.)

Podemos então reescrever o estimador por:

$$\hat{p} = \frac{0 + 1 + 0 + 1 + \dots + 0}{n} = \frac{\sum_{i=1}^n x_i}{n}, \quad \text{que é a mesma expressão de } \bar{X}$$

Pelas propriedades de esperança e variância temos então que:

$$E[\hat{p}] = \mu = p \qquad \qquad VAR[\hat{p}] = \sigma^2 = \frac{p(1-p)}{n}$$

Logo, com base nestas informações, perceba que podemos aplicar novamente o TCL e definir a distribuição de \hat{p} . De maneira resumida, podemos afirmar o seguinte:

Teorema: Quando o tamanho amostral é grande ($n \rightarrow \infty$), $\hat{p} \approx N(p, \frac{p(1-p)}{n})$ pelo TCL. Assim como no caso se \bar{X} , tem-se que:

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0, 1)$$

Exemplo: Sabe-se que 30% dos jovens tem problema de visão. Calcule a probabilidade de em uma classe com 40 jovens, a proporção de jovens com problema de visão seja maior que 32%.

X: jovens com problema de visão

$$E[X] = \mu = p = 0,30 \quad VAR[X] = \sigma^2 = p(1-p) = 0,30 \times 0,70 \quad X \sim ??$$

\hat{p} : proporção de jovens com problema de visão

$$E[\hat{p}] = \mu = p = 0,30 \quad VAR[\hat{p}] = \frac{\sigma^2}{n} = \frac{p(1-p)}{n} = \frac{0,30 \times 0,70}{40} = 0,00525$$

pelo TCL temos então que $\hat{p} \approx N(0,30; 0,00525)$

$$P[\hat{p} > 0,32] = P[Z > 0,28] = 0,3897 = 38,97\%$$

$$Z_c = \frac{0,32 - 0,30}{\sqrt{0,00525}} = 0,28$$

5.10.6 Distribuição amostral de S^2

Perceba que até aqui tratamos da distribuição de dois dos principais estimadores: média (\bar{X}) e proporção (\hat{p}).

Conforme já vimos, para o estimador da variância (S^2), as distribuições amostrais são χ^2 no caso de desejarmos realizar inferência sobre uma variância e F no caso de duas variâncias. No entanto, para a variância geralmente não estamos interessados em calcular probabilidades, faz mais sentido decidir se determinadas afirmações sobre σ^2 são ou não atendidas. Desta forma tais distribuições serão mais utilizadas em inferência estatística.

5.10.7 Gráficos Q-Q plots

Como sabemos se uma distribuição de probabilidades é um modelo razoável para os dados?

Essa questão é importante porque muitas das técnicas estatísticas estão baseadas na suposição de que a distribuição de probabilidades seja de um tipo específico (normal, na maioria dos casos). Em confiabilidade, por exemplo, a verificação se os dados de tempo de falha são provenientes de uma exponencial identifica o mecanismo de falha. Outra situação é a verificação da pressuposição de normalidade na obtenção de intervalos de confiança usando as distribuições t de Student (n pequeno), χ^2 e F .

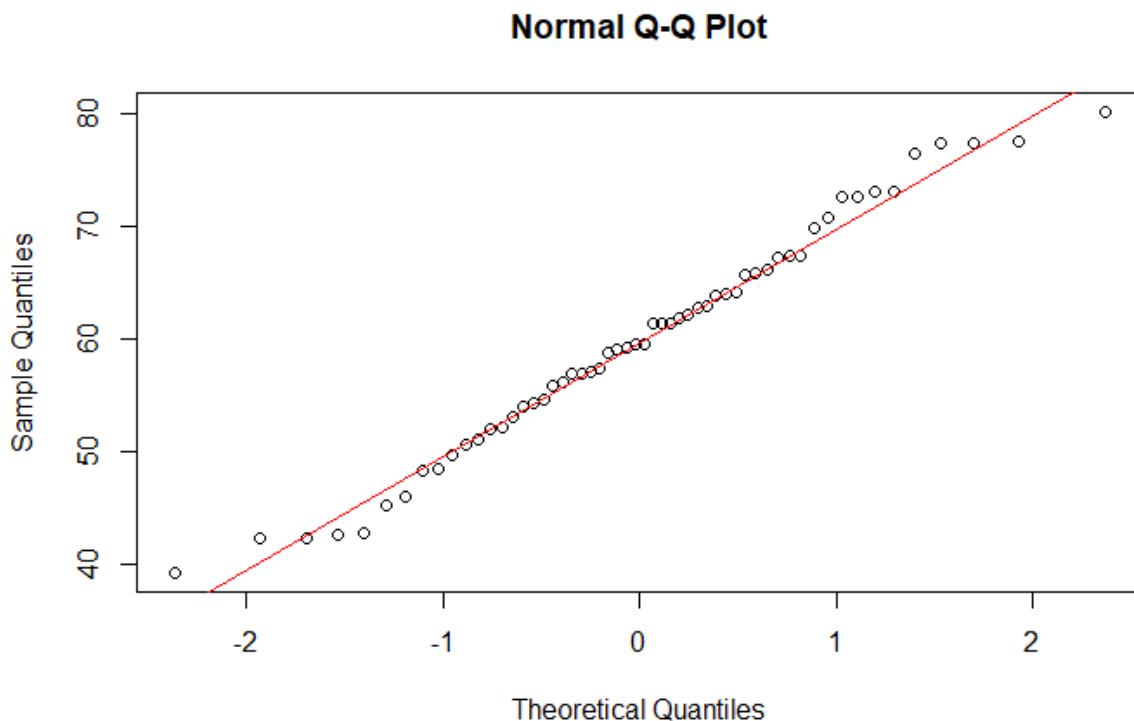
Histogramas podem fornecer uma ideia sobre a distribuição probabilística, mas não são, em geral, bons indicadores, a não ser que o tamanho da amostra seja bem grande. Um gráfico de probabilidade ou Q-Q plot, como é conhecido por alguns autores, é um método para determinar se os dados da amostra obedecem a uma distribuição hipotética, baseado no exame visual subjetivo dos dados. O procedimento é simples, pode ser feito rapidamente e é mais confiável que histogramas para pequenos a moderados tamanhos amostrais.

De maneira geral, o Q-Q plot é um gráfico de quantis ordenados de um conjunto de dados *versus* os quantis ordenados de outro conjunto de dados. Cada ponto (x,y) refere-se ao quantil de uma distribuição do eixo vertical (y) contra o quantil correspondente de outra distribuição ao longo do eixo horizontal (x). Se as duas distribuições são similares, os pontos situam-se na linha de identidade, $y = x$.

Os quantis dos dois conjuntos de dados podem ser observados ou teóricos. Quando os quantis de um banco de dados real são plotados com os quantis correspondentes de uma distribuição teórica, o Q-Q plot resultante serve como uma ferramenta visual para verificar o quanto o conjunto de dados pode ser ajustado pela distribuição teórica.

No R podemos usar a função `qqnorm()` para criar Q-Q plots e avaliar o ajuste de uma distribuição normal aos dados. De maneira mais geral, a função `qqplot()` cria Q-Q plots para qualquer distribuição teórica.

Veja abaixo um gráfico QQ-plot.



Nesta situação, como os quantis teóricos e o amostrais estão em torno da reta $y = x$ então podemos dizer que a amostra segue a distribuição teórica de interesse.

5.11 LISTA DE EXERCÍCIOS 7: Distribuições amostrais e Teorema Central do Limite

1- O que é uma distribuição amostral? Sua ideia é baseada em quê? Qual das três distribuições amostrais estudadas é simétrica? Qual distribuição amostral é utilizada para fazer inferência sobre a média de populações normais? Qual distribuição amostral deve ser utilizada para fazer inferência sobre a variância de uma população normal? E para fazer inferência sobre duas variâncias?

2- Qual a diferença entre os dois cenários relacionados ao cálculo de probabilidades para a média? Qual é a única exigência quando a distribuição da variável é desconhecida? Qual distribuição é utilizada para fazer inferência sobre a proporção? No caso da proporção tem alguma exigência? Qual?

3- Seja uma máquina que produz resistores elétricos com resistências média de 40 ohms e desvio padrão de 5 ohms. Calcule a probabilidade de que uma amostra aleatória de 36 desses resistores tenha uma resistência média de:

- a)b)c)** Entre 37 e 42 ohms.

4- Seja X uma variável aleatória distribuída normalmente com média de 1000 ml e variância de 4 ml^2 , representando o volume de recipiente de determinado produto químico. Sabe-se que, de acordo com o engenheiro de controle de qualidade, o volume médio dos recipientes deve estar compreendido entre 998 e 1002 ml. Caso contrário, multas severas são aplicadas. Determine a probabilidade de multas severas serem aplicadas, sabendo que são usadas amostras de 10 recipientes.

5- Um elevador tem suporte máximo de 700 kg para uma lotação de $n = 10$ pessoas. Sabendo que o peso médio de humanos é de 62 kg e cujo desvio padrão é igual a 10 kg, responder as seguintes questões, assumindo que o peso possui distribuição normal:

a) Qual é a probabilidade de uma pessoa pesar mais de 70 kg?

b) Qual é a probabilidade de o elevador ter sua carga máxima ultrapassada para um grupo aleatório de $n = 10$ pessoas que o utilizam?

c) Com base na resposta dada no item (b), você julga que a carga de suporte máximo está bem especificada para este elevador? Justifique.

6- Um catálogo de um fabricante indica para um determinado produto uma vida média de 1200 horas. Assuma o desvio padrão igual a 120 horas. Um cliente decide selecionar aleatoriamente 36 itens do referido produto e rejeitar a amostra se $\bar{X} < 1160$ horas. Se a indicação do fabricante for verdadeira, qual a probabilidade de rejeitar a amostra?

7- A taxa de glicemia em pessoas com boa saúde, X , tem distribuição normal com média 100 mg/dL e desvio padrão de 10 mg/dL. Se \bar{X} é a taxa média de glicemia de uma amostra de n elementos retirados dessa população, calcule $P(90 < \bar{X} < 110)$ para:

- a)** $n = 1$; **b)** $n = 4$; **c)** $n = 16$.

8- Sejam x_i ($i = 1, 2, \dots, n$) variáveis aleatórias (independentes e com mesma distribuição de probabilidade), com média μ e variância σ^2 . Considere a média amostral $\bar{X} = \frac{\sum_{i=1}^n x_i}{n}$. Mostre que $E(\bar{X}) = \mu$ e $\text{Var}(\bar{X}) = \frac{\sigma^2}{n}$.

DICA: Veja as propriedades de esperança e variância no campus virtual.

9- Se uma máquina produz resistores elétricos com resistência média de 40 ohms e desvio-padrão de 2 ohms, qual é a probabilidade de que uma amostra aleatória de 36 desses resistores tenha uma resistência combinada (total) de mais de 1458 ohms?

10- Um empresário afirma que apenas 25% de seus produtos precisam passar novamente por algumas etapas da linha de produção, a fim de reparar pequenos defeitos. Calcule a probabilidade de em uma caixa com 90 destes produtos tenha mais de 30% com defeitos.

11- Um laboratório alega que um produto seu cura 80% dos casos de certa doença. Qual a probabilidade de, em uma amostra de 45 pessoas:

- a) mais de 87% das pessoas serem curadas?
- b) menos de 63% das pessoas serem curadas?
- c) entre 85% e 95% das pessoas curadas?
- d) Exatamente 80% das pessoas serem curadas?

12- O consumo de energia solar nos Estados Unidos tem média mensal de 65 milhões de BTU. No último ano foram observados os seguintes consumos (em milhões de BTU):

55.2 59.7 62.6 63.8 66.4 68.5 69.8 70.8 70.2 69.7 68.7 66.3

Admitindo que o consumo de energia solar possui distribuição normal, qual é a probabilidade de o consumo médio no próximo mês ser superior a 67 milhões de BTU?

GABARITO

1- e 2- Teoria no caderno. **3-** a) 0,1151 b) 0,0082 c) 0,9916

4- 0,00158; **5-** a) 21,19%; b) 0,57%

6- 0,0228; **7-** a) 0,68; b) 0,9554 ; c) $\cong 1$

8- Dica: use as propriedades de esperança.

9- 0,0668; **10-** 0,1379 **11-** a) 0,1210; b) 0,0023 ; c) 0,1971; d) 0

12- Dica: $t_{0.0888,\nu=11} = 1,44$ $t_{0.0867,\nu=12} = 1,44$ $t_{0.0773,\nu=65} = 1,44$

6 Inferência Estatística

Estimação é o processo que consiste em utilizar estimadores e suas respectivas distribuições amostrais para estimar os valores de parâmetros populacionais desconhecidos. Esta é a ideia básica de inferência estatística.

Qualquer característica de uma população pode ser estimada a partir de uma amostra aleatória. Sendo que as mais comuns são: média, variância e proporção.

Para garantir a eficiência da inferência é necessário que a amostra seja coletada de maneira aleatória seguindo as técnicas de amostragem adequadas e que seja conhecida a distribuição amostral do estimador em questão.

6.1 Estimação Pontual

São estimadores que fornecem, para uma amostra, uma única estimativa do parâmetro de interesse.

Por exemplo:

- $\bar{X} = \frac{\sum_{i=1}^n x_i}{n}$ é o estimador pontual da média (μ).
- $S^2 = \frac{\sum_{i=1}^n (x_i - \bar{X})^2}{n-1}$ é o estimador pontual da variância (σ^2).
- $\hat{p} = \frac{\text{favoraveis}}{\text{possíveis}}$ é o estimador pontual da proporção (p).

No entanto, a estimação pontual, por ser um único número, é pouco informativa e não fornece a precisão ou confiabilidade da estimativa.

6.2 Estimação Intervalar

A estimação intervalar consiste na criação de um intervalo de valores possíveis, o qual admite-se que pode conter o parâmetro com certa probabilidade.

Esta probabilidade é denominada “nível de confiança” e é simbolizada por $1 - \alpha$. O valor de α pode ser entendido como a probabilidade de o verdadeiro valor do parâmetro não pertencer ao intervalo. Em geral, o valor de α é fixado pelo pesquisador em 1% ou 5%, gerando intervalos de 99% ou 95% de confiança respectivamente.

A obtenção dos limites dos intervalos de confiança (IC) é feita com base na distribuição amostral do estimador do parâmetro em questão.

A ideia de um IC é: (estimativa pontual \pm erro). Naturalmente este erro depende de quê?

6.2.1 IC para a média (μ)

Seja \bar{X} a estimativa pontual de μ , obtida a partir de uma amostra aleatória de tamanho n , de uma população com variância conhecida σ^2 . Temos pelo Teorema Central do Limite que:

$$Z_n = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \approx N(0, 1)$$

$$\pm Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \implies \pm Z \times \frac{\sigma}{\sqrt{n}} = \bar{X} - \mu \implies \mu : \bar{X} \pm Z \times \frac{\sigma}{\sqrt{n}}$$

Assim um intervalo de confiança de $100(1 - \alpha)\%$ para μ é dado por:

$$IC_{1-\alpha}(\mu) : \bar{X} \pm Z_{\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}$$

ou equivalentemente:

$$P \left[\bar{X} - Z_{\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + Z_{\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}} \right] \leq 1 - \alpha$$

Naturalmente, o que varia de amostra para amostra é o valor do intervalo e não o parâmetro. Assim a interpretação de um intervalo de confiança de 95% por exemplo é que ao realizar um certo número de amostras 95% dos intervalos obtidos conterão o verdadeiro valor do parâmetro. Veja na figura abaixo a ideia do intervalo de confiança para a média.

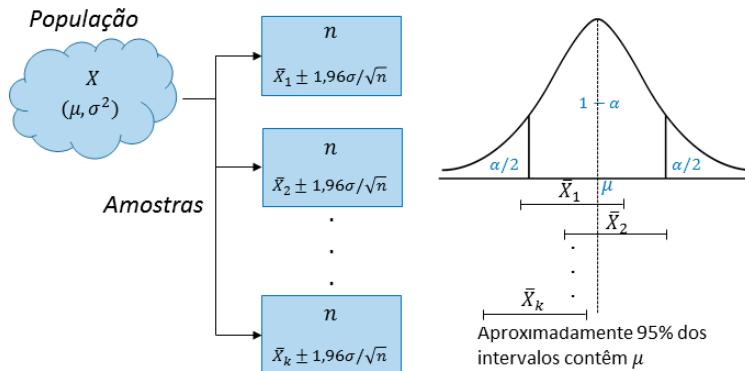


Figura 22: Ilustração da ideia de intervalos de confiança de 95% para a média μ .

Exemplo: A concentração média de sódio em 36 amostras de certo tipo de alimento foi de 2,6g/Kg. Sabendo que o desvio padrão da concentração de sódio é $\sigma = 0,3$ g/Kg obtenha um intervalo de confiança de 95% para a concentração média de sódio neste alimento.

$$n = 36 \quad \bar{X} = 2,6 \quad \sigma = 0,3 \quad \alpha = 0,05 \quad Z_{0,025} = 1,96$$

$$IC_{95\%}(\mu) : 2,6 \pm 1,96 \times \frac{0,3}{\sqrt{36}} \implies IC_{95\%}(\mu) : 2,6 \pm 0,1 \implies IC_{95\%}(\mu) : (2,5; 2,7)$$

Que significa que 95% dos intervalos obtidos em amostras de tamanho 36 conterão o verdadeiro valor da concentração média de sódio (μ).

OBS: Note que para resolver o exercício acima utilizamos o TCL, pois admitimos que a concentração média de sódio possui distribuição normal. Este intervalo só é válido porque $n > 30$, caso contrário o intervalo obtido só seria válido se a variável aleatória concentração de sódio tivesse distribuição normal.

No entanto, nem sempre conhecemos a variância populacional (σ^2).

Neste caso lembre-se que se temos uma amostra de uma distribuição normal, então a variável aleatória:

$$T = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}}$$

possui distribuição t de Student com $\nu = n - 1$ graus de liberdade.

Logo a distribuição t pode ser utilizada para obter o intervalo de confiança para μ , de modo análogo ao anterior.

O intervalo de confiança $100(1 - \alpha)\%$ para μ com σ desconhecido é dado por:

$$IC_{1-\alpha}(\mu) : \bar{X} \pm t_{(\frac{\alpha}{2}, \nu)} \times \frac{S}{\sqrt{n}}$$

ou equivalentemente:

$$P \left[\bar{X} - t_{(\frac{\alpha}{2}, \nu)} \times \frac{S}{\sqrt{n}} < \mu < \bar{X} + t_{(\frac{\alpha}{2}, \nu)} \times \frac{S}{\sqrt{n}} \right] \leq 1 - \alpha$$

Exemplo: A quantidade de proteínas foi medida em 7 amostras de um alimento, obtendo: 9, 8; 10, 2; 10, 4; 9, 8; 10, 0; 10, 2; 9, 6 gramas por porção. Assumindo que a distribuição da quantidade de proteínas é aproximadamente normal, determine um intervalo de confiança de 95% para a quantidade média de proteínas neste alimento.

$$n = 7 \quad \nu = 6 \quad \bar{X} = 10,0 \quad S = 0,283 \quad \alpha = 0,05 \quad t_{0,025;6} = 2,447$$

$$IC_{95\%}(\mu) : 10,0 \pm 2,447 \times \frac{0,283}{\sqrt{7}} \implies IC_{95\%}(\mu) : 10,0 \pm 0,26 \implies IC_{95\%}(\mu) : (9,74; 10,26)$$

Que significa que 95% dos intervalos obtidos de amostras de tamanho 7 conterão o verdadeiro valor médio de proteínas neste alimento.

Veja que é uma informação de o quanto confiável é o seu intervalo. Mas não significa necessariamente que a probabilidade do intervalo conter o parâmetro (μ) é de 95%.

6.2.2 IC para a proporção (p)

O intervalos de confiança para o parâmetro proporção (p) são obtidos de maneira similar aos intervalos para a média (μ).

Seja x o número de sucessos em uma amostra de tamanho n . O estimador da proporção de sucessos é:

$$\hat{p} = \frac{\text{favoráveis}}{\text{possíveis}} = \frac{x}{n}.$$

Já foi visto que se atribuirmos 1 para sucesso e 0 para fracasso, podemos reescrever o estimador pela seguinte expressão:

$$\hat{p} = \frac{\sum_{i=1}^n X_i}{n},$$

ou seja, a proporção amostral pode ser também entendida como uma média amostral. Assim, com n suficientemente grande ($n > 30$), podemos utilizar o Teorema Central do Limite e afirmar que \hat{p} tem distribuição aproximadamente normal com média p e variância $\frac{p(1-p)}{n}$, isto é:

$$\hat{p} \cong N\left(p, \frac{p(1-p)}{n}\right).$$

Consequentemente podemos obter a variável Z , por:

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}, \quad \Rightarrow \quad Z \sim N(0, 1)$$

Assim, partindo da afirmação probabilística:

$$P(-Z_{\frac{\alpha}{2}} < Z < Z_{\frac{\alpha}{2}}) = 1 - \alpha,$$

obtemos um intervalo de $100(1 - \alpha)\%$ de confiança para p . Ele é dado por:

$$IC_{1-\alpha}(p) : \hat{p} \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}}$$

ou equivalentemente:

$$P\left[\hat{p} - Z_{\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}} < p < \hat{p} + Z_{\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}}\right] \leq 1 - \alpha,$$

em que $Z_{\frac{\alpha}{2}}$ é o quantil da normal padrão que deixa acima (ou a direita) dele a probabilidade $\frac{\alpha}{2}$.

Mas perceba que os limites do intervalo dependem do parâmetro desconhecido p . Obviamente não desejamos encontrar um intervalo para um parâmetro, cujos limites dependem do próprio parâmetro. Como estamos trabalhando com amostras grandes, podemos substituir p pela estimativa pontual \hat{p} . Então para uma amostra particular de tamanho $n > 30$, o intervalo de confiança aproximado de $100(1 - \alpha)\%$ para a proporção fica:

$$IC_{1-\alpha}(p) : \hat{p} \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

ou equivalentemente:

$$P\left[\hat{p} - Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right] \leq 1 - \alpha$$

Exemplo: Uma amostra aleatória de 487 produtos de uma certa marca foi selecionada e analisado se possuíam todos os atributos sensoriais esperados. Um total de 35 produtos destes não possuíam as características esperadas. Calcule o intervalo de confiança de 95% para a proporção de produtos desta marca que não possuem todos os atributos sensoriais esperados.

$$\hat{p} = \frac{35}{487} = 0,0719 \quad n = 487 \quad \alpha = 5\% \\ Z_{0,025} = 1,96$$

$$IC_{95\%}(p) : \hat{p} \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \Rightarrow IC_{95\%}(p) : 0,0719 \pm 1,96 \sqrt{\frac{0,0719(1-0,0719)}{487}} \Rightarrow \\ IC_{95\%}(p) : 0,0719 \pm 0,0229 \Rightarrow IC_{95\%}(p) : (0,049; 0,0948)$$

6.2.3 IC para a variância (σ^2)

Os intervalos de confiança para a variância também podem ser obtidos pelo método pivotal como os anteriores. A distribuição amostral do estimador S^2 é a Qui-quadrado assim os intervalos de confiança são obtidos com base nos quantis desta distribuição.

No entanto, assim como comentado no capítulo 5, não nos preocuparemos com a obtenção destes intervalos, pois no caso da variância faz mais sentido decidir se determinadas afirmações sobre σ^2 são ou não atendidas. Desta forma, os testes de hipóteses serão mais utilizadas para fazer inferência sobre a variância.

6.3 Margem de Erro e dimensionamento de amostras

Qual deve ser o tamanho da amostra para se ter determinada precisão na estimativa de um parâmetro?

Uma das decisões mais importantes em um estudo estatístico é determinar o tamanho da amostra. É possível perceber que o aumento da amostra melhora a precisão da estimativa e diminui o comprimento do intervalo de confiança.

De modo geral, o tamanho da amostra depende da variabilidade da população, do coeficiente de confiança adotado e do erro de estimativa máximo admitido (*erro*), sendo esses dois últimos fixados pelo pesquisador.

Lembre-se que a padronização das variáveis é feita por:

$$\pm Z_{\frac{\alpha}{2}} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \Rightarrow \bar{X} - \mu = \pm Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} = ??? = \text{Margem de Erro}$$

$$\pm Z_{\frac{\alpha}{2}} = \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \Rightarrow \hat{p} - p = \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = ??? = \text{Margem de Erro}$$

Portanto:

$$IC_{1-\alpha}(\mu) : \bar{X} \pm ME$$

$$IC_{1-\alpha}(p) : \hat{p} \pm ME$$

Tamanho de n para estimar μ

No caso de estarmos interessados na média, se conhecermos a variância σ^2 (ou pelo menos sua estimativa), é possível estimar o tamanho amostral adequado para cada estudo.

Quando σ^2 é conhecida

$$\text{erro} = Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \Rightarrow \sqrt{n} = \frac{Z_{\frac{\alpha}{2}} \sigma}{\text{erro}} \Rightarrow n = \left(\frac{Z_{\frac{\alpha}{2}} \sigma}{\text{erro}} \right)^2$$

Quando σ^2 é desconhecida

$$\text{erro} = t_{\frac{\alpha}{2}, \nu} \frac{S}{\sqrt{n}} \Rightarrow n = \left(\frac{t_{\frac{\alpha}{2}, \nu} S}{\text{erro}} \right)^2$$

A segunda fórmula é mais utilizada na prática, pois nem sempre conhecemos o valor de σ . Mas existem 2 aspectos que devem ser destacados em relação ao seu uso:

- i) o valor tabelado de $t_{\frac{\alpha}{2}, \nu}$ depende dos graus de liberdade dado por $\nu = n - 1$.
- ii) S^2 ainda não é conhecido, pois a amostra ainda não foi realizada.

Então, como definir ν e S^2 se a amostra ainda não foi coletada?

Uma alternativa é obter uma amostra piloto de tamanho n' , calcular o quantil associado e S^2 . Se $n \leq n'$ significa que a amostra piloto já é suficiente para estimar μ com a confiança e precisão desejadas. Caso contrário obtem-se mais amostras até completar o valor de n .

Tamanho de n para estimar p

Caso o interesse esteja em obter uma amostra para estimar alguma proporção da população, temos:

$$\text{erro} = Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \Rightarrow \sqrt{n} = Z_{\frac{\alpha}{2}} \frac{\sqrt{\hat{p}(1 - \hat{p})}}{\text{erro}} \Rightarrow n = \frac{Z_{\frac{\alpha}{2}}^2 \hat{p}(1 - \hat{p})}{\text{erro}^2}$$

Novamente, para estimar n precisamos de um \hat{p} , mas este só é obtido a partir de uma amostra. Logo a ideia da amostra piloto deve ser utilizada também. Quando não for possível a obtenção desta amostra, podemos estabelecer um limite superior para n , basta notar que $\hat{p}(1 - \hat{p})$ será no máximo igual a $\frac{1}{4}$. Então se usarmos $\hat{p} = \frac{1}{2}$ obteremos o maior valor de n necessário, mesmo sem saber o \hat{p} da amostra.

Nesse caso pode ser utilizada a seguinte expressão:

$$n_{max} = \frac{Z_{\frac{\alpha}{2}}^2}{4erro^2}$$

Exemplo: Qual o tamanho da amostra necessário para verificar a proporção de pessoas de uma cidade a favor de que seja adicionado flúor no tratamento de água se desejamos estar 95% confiantes de que nossa estimativa estará a no máximo 1% da proporção real?

$$n = \frac{Z_{\frac{\alpha}{2}}^2}{4erro^2} \Rightarrow \frac{Z_{0,05}^2}{4 \times 0,01^2} \Rightarrow n = \frac{1,96^2}{4 \times 0,01^2} = 9604 \text{ pessoas.}$$

OBS: O valor de n deve ser sempre arredondado para cima.

6.3.1 Exercícios

Exercício 1: Sabe-se que a resistência à fraturas (em Mpa) de barras de cerâmicas possui desvio padrão $\sigma = 7,73$. Em uma amostra de $n = 62$ barras de cerâmicas encontrou-se uma resistência à fratura média de $\bar{X} = 89,10$.

- a) Calcule o intervalo de 95% de confiança para a resistência à fratura média real (μ).
- b) Calcule o intervalo de 99% de confiança para a resistência à fratura média real (μ). Qual dos dois intervalos é maior? Por quê?
- c) Quão grande deve ser uma amostra para estimar μ com um erro de no máximo 1,0 MPa, com confiança de 95%?

Exercício 2: O controle químico de uma doença será feito se a proporção de plantas doentes de uma região atingir 3%. Realizando uma amostra de tamanho $n = 500$ plantas, o pesquisador observou $x = 11$ plantas doentes. Se você fosse tomar a decisão com base na estimativa pontual, qual seria sua decisão? Se por outro lado você fizesse o intervalo de 95% de confiança, qual seria sua decisão? Justificar sua decisão realizando os cálculos apropriados.

Exercício 3: Determinar os valores de n necessários para se estimar o parâmetro p com confiança de 95% sendo admitidos erros de 3% e 5%. Considerar ainda diferentes amostras pilotos com as seguintes estimativas: $\hat{p} = 0,01$, $\hat{p} = 0,10$, $\hat{p} = 0,20$, $\hat{p} = 0,40$, $\hat{p} = 0,50$ e $\hat{p} = 0,60$.

Respostas:

Exercício 1: a) $IC_{95\%}(\mu) : (87, 18; 91, 02)$

b) $IC_{99\%}(\mu) : (86, 58; 91, 62)$. O intervalo de 99% de confiança será sempre maior que o de 95%, pois quanto maior a precisão menor será a probabilidade de estar fora do intervalo, isto é, menor será o valor de α , fazendo com que o quantil $Z_{\frac{\alpha}{2}}$ seja maior.

c) $n = 230$

Exercício 2: A estimativa pontual foi de $\hat{p} = \frac{x}{n} = \frac{11}{500} = 0,022$, o que nos levaria a tomar a decisão de não efetuar o controle, pois 2,2% é menor do que 3%. Devemos, no entanto, tomar a decisão considerando um erro da estimação, pois estamos observando uma amostra. Para isso construímos o intervalo de confiança de 95% para p resultando em: $IC_{95\%}(p) : (0,0091; 0,0349)$.

A verdadeira proporção de plantas doentes, com aproximadamente 95% de confiança, é um valor entre 0,91% e 3,49%. Como o intervalo de confiança abrangeu o nível de contaminação de 3%, então a recomendação é que se faça o controle, decisão diferente da que seria tomada somente com a estimativa pontual. Acontece que ao utilizar o IC, foi utilizado a distribuição amostral do estimador, que por definição leva em conta todas as possíveis amostras de mesmo tamanho da população.

Exercício 3:

$$e = 0,03 : n = 43, n = 385, n = 683, n = 1025, n = 1068 \text{ e } n = 1025.$$

$$e = 0,05 : n = 16, n = 139, n = 246, n = 369, n = 385, n = 369.$$

Portanto a estimativa que exige um maior tamanho amostral é $\hat{p} = 0,50$, conforme comentado na elaboração da fórmula a ser utilizada quando não é possível obter uma amostra piloto. Perceba que, além de ser um resultado matemático, esta conclusão também é natural, pois se a estimativa está próxima de 50% então ocorre uma certa indecisão (mistura) entre as partes, logo um tamanho amostral maior é necessário para um melhor discernimento sobre a verdadeira proporção p .

Em relação aos erros fixados (3% e 5%), é intuitivo que quanto menor a margem de erro admitida, maior será o tamanho de amostra necessário.

6.4 LISTA DE EXERCÍCIOS 8: Intervalos de Confiança, margem de erro e dimensionamento de amostras

1- A condutividade térmica do ferro Armco foi medida em um experimento considerando-se uma temperatura de 100 °F e uma entrada de potência de 550 W. Foram obtidos os 10 valores seguintes (em Btu/h-ft°F): 41,6; 41,48; 42,34; 41,95; 41,86; 42,18; 41,72; 42,26; 41,81; 42,04. Admitindo que a condutividade térmica é normalmente distribuída, obtenha intervalos de confiança de 95% e 99% para a condutividade média do ferro Armco. Qual dos intervalos é mais amplo? Por quê?

2- É importante que as máscaras usadas pelos bombeiros sejam capazes de resistir a altas temperaturas, pois esses profissionais trabalham com freqüência em temperaturas de 200-500°F. Em um teste de um tipo de máscara, 11 dos 55 equipamentos tiveram as lentes estouradas a 250°. Construa o IC de 90% para a proporção real de máscaras desse tipo, cujas lentes estourariam a 250°.

3- Em um experimento foram avaliados as 20 medidas seguintes da quantidade de proteínas (em gramas por porção) em certo tipo de alimento destinado a crianças:

$$\begin{array}{c} 9,85; 9,83; 9,75; 9,77; 9,67; 9,87; 9,67; 9,79; 9,85; 9,75 \\ 9,83; 9,78; 9,74; 9,99; 9,88; 9,95; 9,75; 9,93; 9,92; 9,89. \end{array}$$

Obtenha o intervalo de confiança de 95% para a quantidade média de proteínas por porção neste alimento.

OBS: Perceba que para construir este intervalo você precisa saber se a quantidade de proteínas no alimento possui distribuição normal. Verifique se tal pressuposição é aceitável por meio de um histograma, utilize o aplicativo disponibilizado no campus virtual.

4- Em uma amostra aleatória de 75 eixos dentre todos os produzidos no turno, 12 têm um acabamento de superfície que é mais áspero do que permitem as especificações. Obtenha um intervalo de confiança de 95% para a proporção real de eixos com defeito na superfície. Suponha que se a proporção real de eixos com defeito for superior a 25% o controle de qualidade da empresa recomenda paralisar o processo e reprogramar as máquinas. Com base no intervalo de confiança de 95%, o processo deve ser paralisado? Explique.

5- Perceba que a margem de erro no exercício acima foi de 8%. Qual o tamanho da amostra necessário, se desejarmos estar 95% confiantes de que este erro seja menor que 5%? E se exigirmos uma margem de erro menor que 2%, assim como em pesquisas eleitorais?

6- Um engenheiro está analisando a força de compressão de certo tipo de gás. Esta força tem distribuição aproximadamente normal, com desvio padrão de 25psi . Uma amostra aleatória de 12 espécimes tem uma força média de compressão de $\bar{X} = 3250\text{psi}$.

- a) Construa um intervalo de confiança de 95% para a força média de compressão deste gás.
- b) Construa um intervalo de confiança de 99% para a força média de compressão deste gás.
- c) Caso o engenheiro deseje estimar a força de compressão com uma margem de erro menor do que 10psi , qual o tamanho de amostra seria necessário?

7- Um engenheiro de alimentos está interessado em estimar o tempo médio necessário para completar determinada reação química. Sabendo que o desvio padrão da reação é de 0,45 minutos, qual deve ser o tamanho da amostra se o engenheiro deseja estar 95% confiante de que o erro na estimativa da média seja menor do que 0,25 minutos?

6.4 LISTA DE EXERCÍCIOS 8: Intervalos de Confiança, margem de erro e dimensionamento de amostras

8- Uma amostra aleatória das alturas de 50 estudantes universitários mostraram uma média de 174,5 cm. Sabendo que o desvio padrão das alturas é de 6,9 cm, obtenha um intervalo de confiança de 98% para a altura média de todos os estudantes. Porque não foi necessário assumir que as alturas tem distribuição normal neste caso?

9- O consumo regular de cereais pré-adoçados contribui para a decadência dos dentes, doenças cardíacas e outras doenças degenerativas de acordo com o Instituto Nacional de Saúde de Londres. Em uma amostra aleatória de 20 porções de determinada marca de cereal a quantidade média de açúcar foi de 11,3 gramas e o desvio padrão foi de 2,45 gramas. Assumindo que a quantidade de açúcar é distribuída normalmente, construa um intervalo de confiança de 95% para a média da quantidade de açúcar por porções.

10- Um estudo será realizado para estimar a proporção de residentes em certa cidade que é a favor da construção de uma usina nuclear. Qual é o tamanho da amostra necessário se desejarmos estar pelo menos 95% confiantes de que a estimativa está a 4% da real proporção de residentes dessa cidade a favor da construção da usina nuclear?

11- Uma indústria química criou um produto para acelerar o processo de secagem de tintas látex. Os valores a seguir registram o tempo de secagem, em horas, de certa marca de tinta látex.

3,4; 2,5; 4,8; 2,9; 3,6; 2,8; 3,3; 5,6; 3,7; 2,8; 4,4; 4,0; 5,2; 3,0; 4,8

Assumindo que o tempo de secagem de tintas látex tenha distribuição normal, determine os intervalos de confiança de 95% e 99% para o tempo médio até a secagem desta marca de tinta. Discuta o efeito ocasionado pelo aumento do nível de confiança.

12- O erro de estimação e , também conhecido como semi-amplitude do intervalo de confiança (IC) dado por

$$e = t_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}}$$

pode ser manipulado, aumentando ou diminuindo o comprimento do intervalo de confiança. De que maneira a variação de cada um dos três componentes da fórmula influenciam no comprimento do intervalo?

13- Suponha que uma amostra aleatória de 50 garrafas de uma marca específica de vinho seja selecionada e o teor alcoólico de cada garrafa seja determinado. Seja μ o teor médio de álcool da população de todas as garrafas da marca em estudo. Suponha que o intervalo de confiança de 95% resultante para a média seja $IC_{95\%}(\mu) : (7,8; 9,4)$. Diga se as afirmações a seguir são verdadeiras ou falsas:

() Um intervalo de confiança de 90% calculado dessa mesma amostra teria sido mais estreito.

() Existe 95% de chance de μ estar entre 7,8 e 9,4.

() podemos estar certos de que 95% de todas as garrafas desse vinho têm um conteúdo alcoólico que está entre 7,8 e 9,4.

() se forem retiradas 100 amostras de tamanho 50 e calculado o intervalo de confiança de 95% para cada amostra, 95 dos intervalos resultantes incluirão μ .

6.4 LISTA DE EXERCÍCIOS 8: Intervalos de Confiança, margem de erro e dimensionamento de amostras

14- O controle químico de uma doença será feito se a proporção de plantas doentes de uma região atingir 3%. Realizando uma amostra de tamanho $n = 500$ plantas, o pesquisador observou $x = 11$ plantas doentes. Se você fosse tomar a decisão com base na estimativa pontual, qual seria sua decisão? Se por outro lado você fizesse o intervalo de 95% de confiança, qual seria sua decisão? Justificar sua decisão realizando os cálculos apropriados. Qual o motivo da mudança de opinião sem mudar a amostra?

15- Em um estudo sobre a concentração de sódio em biscoitos foi realizada uma amostra de tamanho 60 obtendo uma média de 4,56 e desvio padrão de 0,75 gramas por Kg. Construa um intervalo de confiança de 95% para a concentração média de sódio. Foi preciso fazer alguma suposição para a obtenção deste intervalo? Qual?

16- Em um outro estudo sobre a concentração de sódio em biscoitos foi observada uma amostra de tamanho 20 obtendo uma média de 4,85 e desvio padrão de 0,82 gramas por Kg. Construa um intervalo de confiança de 95% para a concentração média de sódio. Foi preciso fazer alguma suposição para a obtenção deste intervalo? Qual?

Respostas:

1: $IC_{95\%}(\mu) : (41,7209; 42,1271)$ $IC_{99\%}(\mu) : (41,6322; 42,2158)$

2: $IC_{90\%}(p) : (0,11; 0,29)$ **3:** $IC_{95\%}(\mu) : (9,7839; 9,8621)$

4: $IC_{95\%}(p) : (0,08; 0,24)$. Não. **5:** 207 e 1291

6: a) $IC_{95\%}(\mu) : (3235,85; 3264,14)$ **b)** $IC_{99\%}(\mu) : (3231,41; 3268,59)$
c) $n_{95\%} = 24$ ou $n_{99\%} = 42$

7: 13 **8:** $IC_{98\%}(\mu) : (172,23; 176,77)$

9: $IC_{95\%}(\mu) : (10,15; 12,45)$ **10:** 601

11: $IC_{95\%}(\mu) : (3,2491; 4,3243)$ $IC_{99\%}(\mu) : (3,0405; 4,5329)$

12: Quanto maior o desvio padrão, maior é o IC. Quanto maior a confiança, maior é o IC. Aumentar o n considerando α fixo pode diminuir a amplitude do IC, desde que não aumente também o desvio padrão.

13: V, F, F, V.

14: $\hat{p} = \frac{x}{n} = \frac{11}{500} = 0,022$ $IC_{95\%}(p) : (0,0091; 0,0349)$

O IC leva em conta a distribuição amostral para generalizar a inferência para todas as outras amostras de mesmo tamanho (população).

15: $IC_{95\%}(\mu) : (4,3702; 4,7498)$

16: $IC_{95\%}(\mu) : (4,4663; 5,2337)$

6.5 Testes de Hipóteses

Um problema bastante comum consiste em decidir se determinada afirmação sobre o parâmetro populacional é, ou não, verdadeira. Naturalmente surgem 2 hipóteses:

- A que sugere que a afirmação é verdadeira denominada hipótese nula - H_0 . Geralmente o sinal de igual fica nesta hipótese.
- E a que sugere que a afirmação é falsa denominada hipótese alternativa - H_1 . Está sempre relacionada com o objetivo do estudo.

Dependendo do seu objetivo e de como é formulada a hipótese H_1 , o teste é chamado de bilateral (BL), unilateral à direita (UD) ou unilateral à esquerda (UE).

Exemplos:

$$\begin{cases} H_0 : \mu = 10 \\ H_1 : \mu \neq 10 \end{cases}$$

$$\begin{cases} H_0 : \mu \geq 500 \\ H_1 : \mu < 500 \end{cases}$$

$$\begin{cases} H_0 : p = 0,30 \\ H_1 : p \neq 0,30 \end{cases}$$

$$\begin{cases} H_0 : p \leq 0,70 \\ H_1 : p > 0,70 \end{cases}$$

As possíveis conclusões no teste são:

- rejeitar H_0 , pois existem evidências suficientes na amostra.
- não rejeitar H_0 , pois não existem evidências suficientes na amostra.

O critério para a rejeição, ou não de uma hipótese sobre o parâmetro deve ser elaborado com base em todas as amostras possíveis da população, logo depende naturalmente da amostra observada e da distribuição amostral do estimador.

É conveniente lembrar que mesmo tomando todos os cuidados necessários, as decisões tomadas podem não ser corretas, pois estaremos trabalhando com amostras. As situações possíveis são explicitadas na tabela a seguir.

Entre os dois possíveis erros, o erro tipo I é o mais grave. Assim a probabilidade de cometer o erro tipo I pode ser fixada pelo pesquisador e é denominada “nível de significância” do teste, simbolizada por α . Geralmente $\alpha = 5\%$ ou $\alpha = 1\%$.

Tabela 2: Possíveis situações envolvidas na escolha de uma hipótese.

		Decisão	
		Rejeitar H_0	Não rejeitar H_0
H_0 verdadeira	erro tipo I	✓	
	H_0 falsa	✓	erro tipo II

A mecânica dos testes de hipóteses

Um teste de hipóteses pode ser organizado nos seguintes passos:

1- Definir as hipóteses

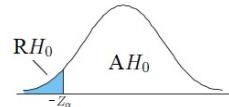
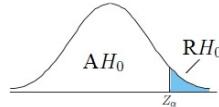
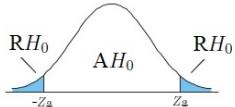
$$\begin{cases} H_0 : \\ H_1 : \end{cases}$$

2- Definir o nível de significância do teste

$$\alpha = ?$$

3- Definir a distribuição amostral do estimador.

4- Com base na distribuição amostral, obter o valor crítico do teste e criar a regra de decisão
 $(Z_{\frac{\alpha}{2}}, Z_\alpha, t_{\frac{\alpha}{2}, \nu}, t_{\alpha, \nu})$



Rejeito H_0 se...

5- Fazer os cálculos e tomar a decisão.

Conclusão:

A construção de um teste de hipóteses, até o passo 4, deve ser feita antes de ser coletada a amostra, evitando assim possíveis manipulações.

O passo mais difícil na elaboração de um teste de hipóteses, por incrível que pareça, é o primeiro. As hipóteses estão relacionadas com as questões a serem respondidas pela pesquisa, assim, principalmente as hipóteses, devem ser formuladas antes de ser realizado o experimento.

Os demais passos do teste são consequência direta das hipóteses, de modo que o único passo que muda nas diferentes situações é o passo 3. Naturalmente, ao mudar a distribuição amostral muda o valor crítico do passo 4 e a estatística de teste no passo 5. Mas a mecânica é sempre a mesma. Assim os testes para os parâmetros média (μ), proporção (p) e variância (σ^2) serão apresentados nas seções a seguir com exemplos.

6.5.1 Teste de Hipóteses para a média com σ conhecido

Exemplo: O tempo de secagem de certo tipo de tinta é normalmente distribuído com média 75 minutos e desvio padrão de 9 minutos. Alguns engenheiros de materiais propuseram um novo aditivo para diminuir o tempo médio de secagem. Para testar a eficiência do novo produto foi medido o tempo de secagem em 25 locais e encontrou-se uma média de 71 minutos. Verifique ao nível de significância de 5% se compensa lançar o produto no mercado.

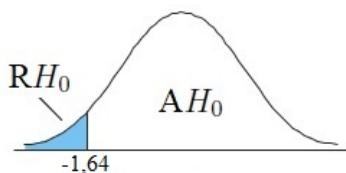
$$n = 25 \quad \bar{X} = 71 \quad \sigma = 9 \quad \mu = 75$$

$$1- \begin{cases} H_0 : \mu \geq 75 \\ H_1 : \mu < 75 \rightarrow UE \end{cases}$$

$$2- \alpha = 5\%$$

3- Distribuição normal.

$$4- Z_\alpha = Z_{0,025} = -1,64$$



Rejeito H_0 se $Z_c < -1,64$.

$$5- Z_c = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \quad Z_c = \frac{71-75}{9/\sqrt{25}} = -2,22$$

Conclusão: Como $Z_c < -1,64$ então rejeita-se H_0 ao nível de 5% de significância, ou seja, o tempo de secagem médio agora é menor que 75 minutos, logo compensa lançar o produto no mercado.

6.5.2 Teste de Hipóteses para a média com σ desconhecido

Em muitas situações o valor de σ não será conhecido e deverá ser substituído pelo seu estimador, obtendo assim a distribuição *t de Student*.

Exemplo: A ANVISA estabeleceu que a quantidade de açúcar em certo alimento deve ser em média 50 gramas por pacote. Desconfia-se que esta norma não esteja sendo atendida. Foi feita uma pesquisa sobre a quantidade de açúcar em 16 pacotes deste alimento, observando os seguintes dados:

53 42 62 52 58 42 58 58 47 53 59 45 52 49 47 47

Supondo normalidade dos dados, verifique ao nível de significância de 5% se a quantidade média de açúcar no alimento está fora das especificações da ANVISA.

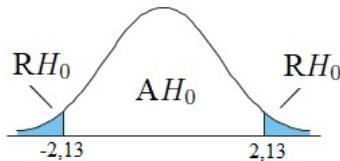
$$n = 16 \quad \bar{X} = 51,5 \quad S = 6,26 \quad \mu = 50$$

$$1- \begin{cases} H_0 : \mu = 50 \\ H_1 : \mu \neq 50 \rightarrow BL \end{cases}$$

$$2- \alpha = 5\%$$

3- Distribuição *t de Student.*

$$4- t_{(\frac{\alpha}{2}, \nu)} = t_{(0,025; 15)} = 2,13$$



Rejeito H_0 se $t_c < -2,13$ ou se $t_c > 2,13$.

$$5- t_c = \frac{\bar{X} - \mu}{S/\sqrt{n}} \quad t_c = \frac{51,5 - 50}{6,26/\sqrt{16}} = 0,9584$$

Conclusão: Como $-2,13 < t_c < 2,13$ então não se rejeita H_0 ao nível de 5% de significância, ou seja, a quantidade média de açúcar está dentro das especificações da ANVISA.

6.5.3 Teste de Hipóteses para uma proporção, p

Seja p a proporção de pessoas que apresentam determinada característica de uma população. Como já vimos, \hat{p} é o estimador de p e assim a estatística de teste é baseada na distribuição amostral deste estimador. Lembre-se que para uma amostra grande ($n > 30$), utilizamos o teorema central do limite para obter uma distribuição aproximadamente normal para \hat{p} .

Exemplo: Um relatório de uma companhia afirma que 40% de toda a água obtida através de poços artesianos no Nordeste é salobra. Há muitas controvérsias sobre essa afirmação, alguns dizem que a proporção é maior, outros que é menor. Para dirimir as dúvidas, 400 poços foram sorteados e observou-se, em 120 deles, água salobra. Qual seria a conclusão ao nível de 3%?

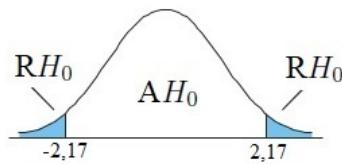
$$n = 400 \quad \hat{p} = \frac{120}{400} = 0,30 \quad p = 0,40$$

$$1- \begin{cases} H_0 : p = 0,40 \\ H_1 : p \neq 0,40 \rightarrow BL \end{cases}$$

$$2- \alpha = 3\%$$

3- Distribuição Normal.

$$4- Z_{\frac{\alpha}{2}} = Z_{\frac{0,03}{2}} = Z_{0,015} = 2,17$$



Rejeito H_0 se $t_c < -2,17$ ou se $t_c > 2,17$.

$$5- \quad Z_c = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \quad Z_c = \frac{0,30 - 0,40}{\sqrt{\frac{0,40(1-0,40)}{400}}} = -4,08$$

Conclusão: Como $Z_c < -2,17$ então rejeita-se H_0 ao nível de 3% de significância, ou seja, a proporção de água salobra no Nordeste é diferente de 40%.

6.5.4 Teste de Hipóteses para duas proporções, p_1 e p_2

Suponha que tenhamos uma amostra de m indivíduos da população 1 e n da população 2. Sejam p_1 e p_2 as proporções de indivíduos nas populações 1 e 2, respectivamente, que possuem a característica de interesse. Os estimadores naturais para estes parâmetros são: \hat{p}_1 e \hat{p}_2 .

Perceba que estamos interessados em testar a hipótese $H_0 : p_1 = p_2$ que é equivalente a testar a hipótese $H_0 : p_1 - p_2 = 0$. Considerando m e n , grandes (> 30), então pelo Teorema Central do Limite \hat{p}_1 e \hat{p}_2 tem individualmente distribuições aproximadamente normais. Assim, o estimador da diferença de proporções $\hat{p}_1 - \hat{p}_2$ também possui aproximadamente uma distribuição normal.

Neste caso, por se tratar de duas amostras as hipóteses são:

$$\begin{cases} H_0 : p_1 - p_2 = \Delta_0 \\ H_1 : p_1 - p_2 \neq \Delta_0, \quad p_1 - p_2 > \Delta_0, \quad p_1 - p_2 < \Delta_0 \end{cases}$$

Se $\Delta_0 = 0$, tem-se que:

$$\begin{cases} H_0 : p_1 = p_2 \\ H_1 : p_1 \neq p_2, \quad p_1 > p_2, \quad p_1 < p_2 \end{cases}$$

E consequentemente o valor de Z_c é dado por:

$$Z_c = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{m} + \frac{\hat{p}_2(1-\hat{p}_2)}{n}}} = \frac{(\hat{p}_1 - \hat{p}_2) - \Delta_0}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{m} + \frac{\hat{p}_2(1-\hat{p}_2)}{n}}} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{m} + \frac{\hat{p}_2(1-\hat{p}_2)}{n}}}$$

Exemplo: Uma Indústria deseja se instalar em uma cidade do sul de Minas. Embora a empresa desconfie que será mais bem aceita na cidade A do que na cidade B, foi feita uma pesquisa de modo a verificar onde as pessoas estariam mais interessadas na implantação da empresa. Se das 200 pessoas entrevistadas na cidade A, 120 são a favor da implantação da empresa e das 500 pessoas entrevistadas na cidade B, 240 são a favor da implantação da empresa. Teste a hipótese de que a verdadeira proporção de pessoas da cidade A a favor da empresa é maior que a da cidade B. Use um nível de significância de 5%.

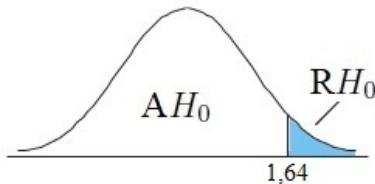
$$m = 200 \quad n = 500 \quad \hat{p}_A = \frac{120}{200} = 0,60 \quad \hat{p}_B = \frac{240}{500} = 0,48$$

$$1- \begin{cases} H_0 : p_A \leq p_B \\ H_1 : p_A > p_B \rightarrow UD \end{cases}$$

$$2- \alpha = 5\%$$

3- Distribuição Normal.

$$4- Z_\alpha = Z_{0,05} = 1,64$$



Rejeito H_0 se $Z_c > 1,64$.

$$5- Z_c = \frac{\hat{p}_A - \hat{p}_B}{\sqrt{\frac{\hat{p}_A(1-\hat{p}_A)}{m} + \frac{\hat{p}_B(1-\hat{p}_B)}{n}}} \quad Z_c = \frac{0,60 - 0,48}{\sqrt{\frac{0,60 \times (1-0,60)}{200} + \frac{0,48 \times (1-0,48)}{500}}} = 2,91$$

Conclusão: Como $Z_c > 1,64$ então rejeita-se H_0 ao nível de 5% de significância, ou seja, a proporção de pessoas favoráveis a implantação da empresa na cidade A é maior que na cidade B.

6.5.5 Teste de Hipóteses para a variância, σ^2 de uma população normal

Testes sobre a variância estão relacionados a variabilidade de processos e ao atendimento de certas especificações. Geralmente estas especificações são atendidas se a variância do processo é suficientemente pequena, menor que um valor pré-estabelecido σ_0^2 .

Seja σ^2 a variância populacional, seu estimador S^2 possui distribuição amostral de χ^2 com $\nu = n - 1$ graus de liberdade. A lógica de aplicação do teste de hipóteses é novamente semelhante as anteriores.

Ao contrário das distribuições anteriores a χ^2 não é simétrica. Assim os valores críticos para a regra de decisão do testes são obtidos por: $\chi^2_{(\alpha;\nu)}$ nos testes unilaterais a direita; $\chi^2_{(1-\alpha;\nu)}$ nos testes unilaterais a esquerda e nos testes bilaterais os dois valores são $\chi^2_{(\frac{\alpha}{2};\nu)}$ e $\chi^2_{(1-\frac{\alpha}{2};\nu)}$.

Exemplo: Sabe-se que o desvio padrão das tensões de ruptura de certos cabos produzidos por uma fábrica é de 300 kg. Depois de ter sido introduzida uma mudança no processo de fabricação desses cabos, as tensões de ruptura de uma amostra de 8 cabos apresentaram o desvio padrão de 240 kg. Assumindo que a tensão de ruptura tem distribuição aproximadamente normal, investigue a significância da diminuição aparente da variância, ao nível de 5%.

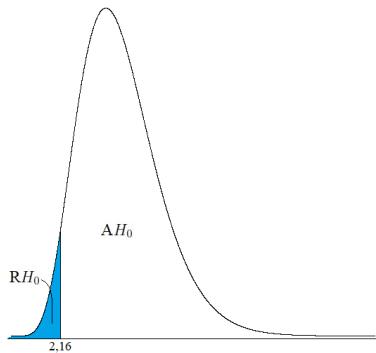
$$n = 8 \quad S = 240 \quad \sigma = 300$$

$$1- \begin{cases} H_0 : \sigma^2 \geq 300^2 \\ H_1 : \sigma^2 < 300^2 \rightarrow UE \end{cases}$$

$$2- \alpha = 5\%$$

3- Distribuição χ^2 .

$$4- \chi_{(1-\alpha;\nu)}^2 = \chi_{(0,95;7)}^2 = 2,16$$



Rejeito H_0 se $\chi_c^2 < 2,16$.

$$5- \chi_c^2 = \frac{(n-1)S^2}{\sigma_0^2} = \chi_c^2 = \frac{(7) \times 240^2}{300^2} = 4,48.$$

Conclusão: Como $\chi_c^2 > 2,16$ então não se rejeita H_0 ao nível de 5% de significância, ou seja, não houve diminuição na variância. Isto é, a diminuição aparente observada é meramente obra do acaso e não pode ser considerada significativa com esta amostra.

6.5.6 Teste de Hipóteses para a razão de variâncias, σ_1^2/σ_2^2 :

Consideremos agora o problema de testar se duas variâncias de duas populações normais σ_1^2 e σ_2^2 são iguais. Naturalmente as hipóteses devem ser de igualdade, no entanto, podemos transformá-la em uma razão da seguinte maneira:

$$\left\{ \begin{array}{l} H_0 : \sigma_1^2 = \sigma_2^2 \\ H_1 : \sigma_1^2 \neq \sigma_2^2, \quad \sigma_1^2 > \sigma_2^2, \quad \sigma_1^2 < \sigma_2^2 \end{array} \right. = \left\{ \begin{array}{l} H_0 : \sigma_1^2/\sigma_2^2 = 1 \\ H_1 : \sigma_1^2/\sigma_2^2 \neq 1, \quad \sigma_1^2/\sigma_2^2 > 1, \quad \sigma_1^2/\sigma_2^2 < 1 \end{array} \right.$$

Como já vimos, a distribuição amostral da razão de variâncias é a distribuição F_{ν_1, ν_2} , em que $\nu_1 = n_1 - 1$ é o grau de liberdade do numerador e $\nu_2 = n_2 - 1$ é o grau de liberdade do denominador e n_1, n_2 são os tamanhos das amostras obtidas das populações 1 e 2.

A distribuição F também não é simétrica em torno de zero e seus valores críticos são obtidos como na distribuição χ^2 .

Exemplo: A variabilidade no levantamento de impurezas de uma certa substância depende do processo utilizado. Usando dois processos, um engenheiro químico melhorou o segundo, esperando com isso reduzir essa variabilidade. Foram obtidas duas amostras, uma utilizando o primeiro processo e outra utilizando o segundo, de tamanhos 26 e 13, respectivamente, obtendo-se $S_1^2 = 1,34$ e $S_2^2 = 0,51$. Assumindo que o levantamento de impurezas possui distribuição normal teste a hipótese de igualdade das variâncias de ambos os processos ao nível de 5% de significância.

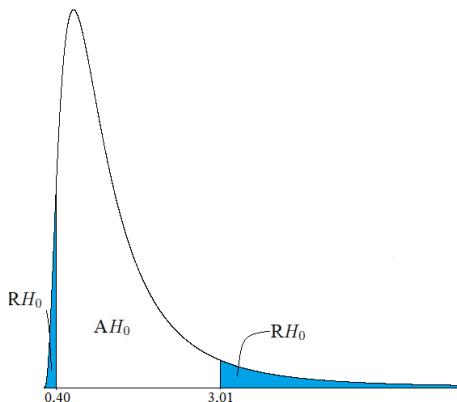
$$n_1 = 26 \quad n_2 = 13 \quad S_1^2 = 1,34 \quad S_2^2 = 0,51$$

$$1- \begin{cases} H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1 \\ H_1 : \frac{\sigma_1^2}{\sigma_2^2} \neq 1 \rightarrow BL \end{cases}$$

$$2- \alpha = 5\%$$

3- Distribuição F .

$$4- F_{(\frac{\alpha}{2}; \nu_1, \nu_2)} = F_{(0,025; 25, 12)} = 3,01 \\ F_{(1 - \frac{\alpha}{2}; \nu_1, \nu_2)} = F_{(0,975; 25, 12)} = 0,40$$



Rejeito H_0 se $F_c < 0,40$ ou se $F_c > 3,01$.

$$5- F_c = \frac{S_1^2}{S_2^2} = \frac{1,34}{0,51} = 2,63.$$

Conclusão: Como $0,40 < F_c < 3,01$ então não se rejeita H_0 ao nível de 5% de significância, ou seja, as variâncias dos dois processos são iguais. Portanto a correção não conseguiu acarretar uma diminuição significativa na variância.

OBS: Perceba que a natureza do teste F é bilateral, pois a hipótese alternativa mais natural é que a razão das variâncias é diferente de 1. No entanto, é comum que em alguns livros este teste seja ensinado colocando a maior variância amostral no numerador, obtendo-se, assim, apenas valores de $F_c \geq 1$, ficando portanto o teste unilateral à direita. É preciso estar atento para usar adequadamente os valores de ν_1 e ν_2 conforme as especificações de quais variâncias são usadas no numerador e denominador.

6.5.7 Teste de Hipóteses para duas médias μ_1 e μ_2

A necessidade de comparar 2 médias surge naturalmente na investigação científica. A ideia consiste em verificar se as diferenças encontradas entre as médias amostrais são casuais ou se realmente existe diferença entre as médias populacionais.

Ao comparar duas médias temos duas situações diferentes: quando as amostras forem independentes e quando forem dependentes.

Amostras independentes

Em amostras independentes temos 4 casos possíveis e em todos eles as hipóteses são:

$$\begin{cases} H_0 : \mu_1 - \mu_2 = \Delta_0 \\ H_1 : \mu_1 - \mu_2 \neq \Delta_0, \quad \mu_1 - \mu_2 > \Delta_0, \quad \mu_1 - \mu_2 < \Delta_0 \end{cases}$$

Geralmente $\Delta_0 = 0$, assim testaremos:

$$\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 \neq \mu_2, \quad \mu_1 > \mu_2, \quad \mu_1 < \mu_2 \end{cases}$$

Os 4 casos são divididos de acordo com a variância das 2 populações. O que fica diferente é a distribuição amostral e a estatística de teste.

- **1º Caso:** Populações normais com variâncias conhecidas e diferentes.

A estatística de teste será:

$$Z_c = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Os valores críticos ($Z_\alpha, Z_{\frac{\alpha}{2}}$) são obtidos da normal padrão.

- **2º Caso:** Populações normais com variâncias desconhecidas e diferentes.

A estatística de teste será:

$$t_c = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

Os valores críticos ($t_{\alpha,\nu}, t_{\frac{\alpha}{2},\nu}$) são obtidos da distribuição *t de Student* com os graus de liberdade determinados pela equação de Satterwaite:

$$\nu = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{(S_1^2/n_1)^2}{n_1-1} + \frac{(S_2^2/n_2)^2}{n_2-1}}$$

- **3º Caso:** Populações normais com variâncias conhecidas e iguais.

A estatística de teste será:

$$Z_c = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}} = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Os valores críticos ($Z_\alpha, Z_{\frac{\alpha}{2}}$) são obtidos da normal padrão.

- **4º Caso:** Populações normais com variâncias desconhecidas e iguais.

A estatística de teste será:

$$t_c = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

A primeira ideia para estimar σ^2 é fazer $\frac{S_1^2 + S_2^2}{2}$, mas como eventualmente $n_1 \neq n_2$ uma das amostras trará mais informações sobre σ^2 , assim faz-se uma média ponderada:

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

Os valores críticos $(t_{\alpha,\nu}, t_{\frac{\alpha}{2},\nu})$ são obtidos da distribuição *t de Student* com $\nu = n_1 + n_2 - 2$ graus de liberdade.

Obviamente quando as variâncias são conhecidas fica fácil decidir sobre o 1º e o 3º casos. Já quando elas são desconhecidas, para decidir entre o 2º e o 4º casos você pode fazer o teste F da razão de variâncias primeiro.

Exemplo: Em uma turma de estatística, 12 alunos de uma turma conseguiram média 78 e desvio padrão de 6 pontos, ao passo que 15 alunos de outra turma conseguiram média 74 e desvio padrão de 8 pontos. Considerando distribuições normais com variâncias iguais para as notas verifique se a primeira turma é melhor que a segunda ao nível de 5%.

$$n_1 = 12 \quad n_2 = 15 \quad \bar{X}_1 = 78 \quad S_1 = 6 \quad \bar{X}_2 = 74 \quad S_2 = 8 \quad \nu = 12 + 15 - 2 = 25$$

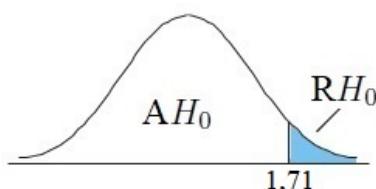
4º CASO

$$1- \begin{cases} H_0 : \mu_1 \leq \mu_2 \\ H_1 : \mu_1 > \mu_2 \rightarrow UD \end{cases}$$

$$2- \alpha = 5\%$$

3- Distribuição *t de Student*.

$$4- t_{\alpha,\nu} = t_{0,05;25} = 1,71$$



Rejeito H_0 se $t_c > 1,71$.

$$5- t_c = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} = \frac{11 \times 36 + 14 \times 64}{12 + 15 - 2} = 51,68$$

$$t_c = \frac{(78 - 74)}{\sqrt{51,68 \left(\frac{1}{12} + \frac{1}{15} \right)}} = 1,43$$

Conclusão: Como $t_c < 1,71$ então não se rejeita H_0 ao nível de 5% de significância, ou seja, não existem indícios para afirmar que a primeira turma seja melhor que a segunda.

Amostras dependentes ou pareadas

Neste tipo de situação geralmente toma-se duas observações em um mesmo conjunto de n indivíduos ou objetos. Como as amostras são dependentes não podemos usar as estatísticas anteriores, portanto usaremos o seguinte artifício.

$$\begin{cases} H_0 : \mu_d = 0 \\ H_1 : \mu_d \neq 0, \quad \mu_d > 0, \quad \mu_d < 0 \end{cases}$$

em que μ_d é a média das diferenças entre as observações, o qual possui distribuição t com $\nu = n - 1$ graus de liberdade. Assim ao invés de testar se as duas médias são iguais, estamos testando se a média das diferenças é igual a zero.

A estatística de teste é dada por:

$$t_c = \frac{\bar{d} - \mu_d}{S_d / \sqrt{n}}$$

em que \bar{d} e S_d são a média e o desvio padrão das diferenças respectivamente.

Exemplo: Um grupo de 10 pessoas é submetido a uma dieta, estando o peso antes do início (x_i) e no final (y_i) anotados abaixo. Ao nível de 5% podemos concluir que a dieta foi eficiente?

Pessoa	A	B	C	D	E	F	G	H	I	J
x_i	120	104	93	87	85	98	102	106	88	90
y_i	116	102	90	83	86	97	98	108	82	85
d_i	4	2	3	4	-1	1	4	-2	6	5

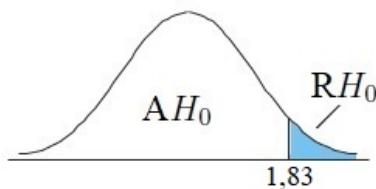
$$n = 10 \quad \bar{d} = 2,6 \quad S_d = 2,59$$

$$1- \begin{cases} H_0 : \mu_x \leq \mu_y \\ H_1 : \mu_x > \mu_y \end{cases} \Rightarrow \begin{cases} H_0 : \mu_d \leq 0 \\ H_1 : \mu_d > 0 \rightarrow UD \end{cases}$$

$$2- \alpha = 5\%$$

3- Distribuição t de Student.

$$4- t_{\alpha, \nu} = t_{0,05;9} = 1,83$$



Rejeito H_0 se $t_c > 1,83$.

$$5- t_c = \frac{\bar{d} - \mu_d}{S_d / \sqrt{n}} \quad t_c = \frac{2,6 - 0}{2,59 / \sqrt{10}} = 3,17$$

Conclusão: Como $t_c > 1,83$ então rejeita-se H_0 ao nível de 5% de significância, ou seja, a queda de peso foi significativa pois a média dos desvios é maior que zero, logo a dieta foi eficiente.

Mesmo quando as observações não são tomadas no mesmo indivíduo, o teste pareado pode fazer sentido desde que haja dependência dentro dos pares.

Por exemplo em um experimento para verificar a eficácia de um medicamento para controlar a pressão cardíaca. Utiliza-se 20 pacientes, tratando 10 com o remédio A e 10 com o remédio B. Mas a pressão cardíaca é muita influenciada pela idade e peso, então poderia agrupar as pessoas em pares com idade e peso semelhantes dando um medicamento para cada pessoa resultando em um pareamento natural. Sem esta combinação, um medicamento poderia parecer melhor que outro apenas porque seus pacientes eram mais leves ou mais jovens.

No entanto, para n pares, o pareamento leva a um menor grau de liberdade na distribuição t , sendo $n-1$ no caso de um teste pareado e $2n-2$ no caso de um teste para duas médias independentes. E um teste de hipóteses é mais poderoso quanto maior o grau de liberdade.

6.6 Relação entre testes de hipóteses bilaterais e intervalos de confiança

Perceba que existe uma relação direta entre a abordagem do teste de hipóteses e intervalos de confiança. Por exemplo no caso de realizarmos inferência baseados na distribuição normal, tanto no Intervalo de Confiança, quanto no Teste de Hipóteses, a estatística de teste vem da padronização ((variável - média)/desvio padrão) da variável em estudo.

Para o caso de uma média com σ conhecida, a estrutura de ambos, testes de hipóteses e estimativa via intervalos de confiança, baseia-se no seguinte cálculo:

$$Z_c = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

Como já vimos, no caso do intervalo de confiança, o valor do parâmetro (μ) é isolado na expressão acima a fim de obter limites dentro dos quais é “razoável” que o parâmetro em questão esteja.

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \Rightarrow \pm Z * \frac{\sigma}{\sqrt{n}} = \bar{X} - \mu \Rightarrow IC_{1-\alpha}(\mu) : \bar{X} \pm Z_{\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}}$$

Desta forma, testar as hipóteses:

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{cases}$$

considerando um nível de significância pré-estabelecido α , é equivalente a calcular o intervalo de confiança $100(1 - \alpha)\%$ para μ e rejeitar H_0 se μ_0 estiver fora do intervalo de confiança. Se μ_0 estiver dentro do intervalo de confiança, não se rejeita H_0 .

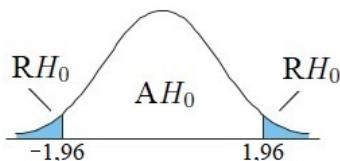
Exemplo: Um valor ideal para a concentração média de zinco recuperado é de 2,65 gramas por mililitro. Em uma amostra de 36 medições deste material obteve-se $\bar{X} = 2,6$ gramas por mililitro. Considerando que o desvio padrão da concentração média de zinco é $\sigma = 0,3$ g/ml, verifique se a concentração média de zinco está diferente da ideal ao nível de significância de 5%.

$$1- \begin{cases} H_0 : \mu = 2,65 \\ H_1 : \mu \neq 2,65 \end{cases}$$

$$2- \alpha = 0,05$$

3- Distribuição normal

$$4- Z_{0,025} = 1,96$$



Rejeito H_0 se $Z_c > 1,96$ ou se $Z_c < -1,96$.

5-

$$Z_c = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{2,6 - 2,65}{0,3/\sqrt{36}} \Rightarrow Z_c = -1$$

Conclusão: Como $-1,96 < Z_c < 1,96$, então não se rejeita H_0 ao nível de 5% de significância, ou seja, não existem evidências para afirmar que a concentração média de zinco seja diferente de 2,65 g/ml.

Obtenha o intervalo de confiança de 95% para a concentração média de zinco.

$$IC_{1-\alpha}(\mu) : \bar{X} \pm Z_{\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}} \Rightarrow IC_{95\%}(\mu) : 2,6 \pm \frac{1,96 * 0,3}{\sqrt{36}} \Rightarrow IC_{95\%}(\mu) : (2,5; 2,7).$$

Logo o valor ideal para a concentração média de zinco está dentro dos limites do intervalo com 95% de confiança, por isso aceitou-se H_0 . Com a mesma informação amostral, uma hipótese bilateral que envolva qualquer valor hipotético entre 2,5 e 2,7 não será rejeitada. A hipótese H_0 só será rejeitada se o valor hipotético estiver fora dos limites do intervalo de confiança, ficando assim, relacionadas estas duas formas de fazer inferência sobre os parâmetros de uma população.

A equivalência do intervalo de confiança com o teste de hipóteses se estende a diferença de duas médias, proporções, diferenças entre duas proporções, variâncias, razão de variâncias. Portanto, conhecendo a estatística de teste ou o intervalo de confiança você consegue migrar de um para o outro facilmente. E assim, intervalos de confiança que não foram abordados na disciplina podem ser obtidos.

Exercício: No caso do teste de hipóteses para a diferença entre duas médias, considerando populações normais com variâncias conhecidas e diferentes, a estatística de teste é a seguinte:

$$Z_c = \frac{(\bar{X}_A - \bar{X}_B) - (\mu_A - \mu_B)}{\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}}$$

A expressão do intervalo de confiança para a diferença entre duas médias não foi apresentada neste material. Mas, utilizando as idéias apresentadas acima obtenha a expressão para o intervalo de confiança $100(1 - \alpha)\%$ para a diferença de médias nesta situação.

Resposta:

$$IC_{100(1-\alpha)\%}(\mu_A - \mu_B) : (\bar{X}_A - \bar{X}_B) \pm \sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}$$

6.7 Nível descritivo de um teste de hipóteses (*p*-valor)

Até aqui, ao realizarmos um teste de hipóteses, partimos de um valor de α pré-fixado. Lembre-se que α é o nível de significância do teste, ou probabilidade de cometer o erro tipo I.

Os valores $\alpha = 5\%$ ou $\alpha = 1\%$ são utilizados até hoje por hábito das gerações anteriores. Mas uma boa alternativa seria deixar a cargo de quem vai utilizar as conclusões do teste a escolha do valor de α que não precisará ser fixado a priori, o usuário tem a liberdade de escolher o nível de significância que vai adotar (ou a taxa de erro tipo I que admite cometer).

A ideia consiste em calcular, supondo que a hipótese nula é verdadeira, a probabilidade de obter estimativas mais desfavoráveis, ou extremas (a luz da hipótese alternativa) do que está sendo observado na amostra. Esta probabilidade será o “nível de significância observado” (ou nível descritivo), denotado por *p*-valor.

Valores pequenos do *p*-valor evidenciam que a hipótese nula é falsa, pois a amostra fornece uma estimativa que tem pequena probabilidade de acontecer, se H_0 fosse verdadeira. O conceito de o que é “pequeno” fica a cargo do usuário, que assim decide qual α vai usar para comparar com o *p*-valor obtido.

Hipótese Unilateral

Consideremos, por exemplo o caso do teste de hipóteses unilateral para uma média com σ conhecido. Considerando a hipótese $H_0 : \mu = \mu_0$, o *p*-valor irá depender da hipótese alternativa isto é:

$$\begin{aligned} p\text{-valor} &= P(Z > Z_c | H_0 \text{ verdadeira}), \text{ se } H_1 : \mu > \mu_0 \\ p\text{-valor} &= P(Z < Z_c | H_0 \text{ verdadeira}), \text{ se } H_1 : \mu < \mu_0 \end{aligned}$$

Hipótese Bilateral

Para o teste de uma hipótese bilateral, ao calcularmos o nível descritivo, precisamos considerar que rejeitaremos valores de \bar{X} que se distanciam da hipótese nula, tanto pela direita quanto pela esquerda. Desta forma, um procedimento usual é multiplicar por dois a probabilidade obtida em um das caudas, de modo a preservar a ideia de afastamento bilateral. Neste caso, o *p*-valor leva também em conta a posição relativa entre a estimativa \bar{X} e o valor hipotético μ_0 .

Assim, se as hipóteses do teste são:

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{cases}$$

O nível descritivo (*p*-valor) é dado por:

$$\begin{aligned} p\text{-valor} &= 2 \times P(Z > Z_c | H_0 \text{ verdadeira}), \text{ se } \bar{X} > \mu_0 \\ p\text{-valor} &= 2 \times P(Z < Z_c | H_0 \text{ verdadeira}), \text{ se } \bar{X} < \mu_0 \end{aligned}$$

Regra de decisão

O que deve ficar claro é que o *p*-valor é o menor nível de significância em que H_0 seria rejeitada, para uma amostra observada. Uma vez que o *p*-valor tenha sido determinado, a conclusão, em qualquer nível específico α , resulta da comparação do *p*-valor com α :

- $p\text{-valor} \leq \alpha \Rightarrow$ Rejeição de H_0 ao nível α de significância.
- $p\text{-valor} > \alpha \Rightarrow$ Não-rejeição de H_0 ao nível α de significância.

Assim os passos para o procedimento de um teste de hipóteses com base no *p*-valor são um pouco diferentes do teste clássico e estão descritos abaixo:

- 1) Definir as hipóteses.
- 2) Definir e calcular a estatística de teste com base nas informações disponíveis.
- 3) Obter o *p*-valor com base na natureza do teste e no valor calculado da estatística de teste.
- 4) Julgue a significância das evidências fornecidas na amostra a favor, ou contra, H_0 , isto é, compare o *p*-valor com o nível de significância α desejado e conclua o teste.

A principal vantagem do uso de um *p*-valor é que os softwares estatísticos mais amplamente utilizados (SAS, R, SISVAR, etc...) incluem automaticamente um *p*-valor quando é feito algum teste de hipóteses.

Assim pode-se obter uma conclusão para o teste de hipóteses diretamente da saída do software, basta conhecer as hipóteses (H_0 e H_1) envolvidas. Não é necessário consultar uma tabela de valores críticos nem saber qual é a estatística de teste e nem saber qual a distribuição amostral do estimador em questão.

Exemplo: Uma associação de defesa do consumidor desconfia que embalagens de 450 gramas de um certo tipo de biscoito estão abaixo do peso. Para verificar tal afirmação, foram coletados ao acaso 40 pacotes em vários supermercados, obtendo-se uma média de peso de 447 gramas. Admitindo-se que o peso de pacotes tem distribuição normal com desvio padrão de 10 gramas, qual conclusão pode ser retirada com base no *p-valor*?

$$1- \begin{cases} H_0 : \mu \geq 450 \\ H_1 : \mu < 450 \end{cases}$$

$$2- Z_c = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \Rightarrow Z_c = \frac{447 - 450}{10/\sqrt{40}} = 1,89$$

3- Como o teste é unilateral à esquerda então:

$$p-valor = P(Z < Z_c | H_0 \text{ verdadeira}) = P(Z < 1,89) = 0,0294 \quad \text{ou } 2,94\%.$$

4- Assim o *p-valor*, ou “nível de significância observado” foi de 2,9%. O que significa que se a hipótese nula for verdadeira ($\mu = 450$) a probabilidade de obtermos uma amostra cuja média amostral seja menor ou igual a 447 é de apenas 2,9%, logo temos indícios de que H_0 não é verdadeira. Como já foi comentado, valores pequenos do *p-valor* evidenciam que devemos rejeitar H_0 .

A título de comparação, se tivéssemos fixado o nível de significância em $\alpha = 5\%$, então rejeitariíamos H_0 . Já se tivéssemos fixado em $\alpha = 1\%$, então aceitaríamos H_0 .

É este tipo de impasse que a interpretação de um teste de hipóteses com base no *p-valor* visa extinguir, pois, pode ser que o nível de significância adotado por uma pessoa não seja o mesmo que outros utilizariam. E fornecendo o *p-valor*, você dá o nível de significância observado na amostra deixando a cargo de quem vai utilizar os resultados escolher o valor de α que vai comparar.

6.8 Teste de Normalidade

Como vocês perceberam, vários testes dependem, pelo menos teoricamente, da suposição de normalidade dos dados. Mesmo que a distribuição amostral utilizada no teste não seja a normal. Portanto, antes de fazer o teste de χ^2 para uma variância, o teste F para razão de variâncias ou teste t para a média é preciso verificar se a suposição de normalidade da variável em estudo é atendida.

Uma maneira gráfica de fazer isso é utilizando o histograma da amostra. Outra maneira é o gráfico *q-qplot*. No software R, este gráfico pode ser facilmente obtido pelos comandos *qqnorm(dados)* e *qqline(dados)*.

6.8.1 O teste de Shapiro-Wilk

Mas a interpretação de gráficos pode se tornar um tanto quanto subjetiva, assim, agora que você já conhece a mecânica dos testes de hipótese, e principalmente a interpretação do *p-valor* pode aplicar o teste de Shapiro-Wilk.

O teste de Shapiro-Wilk é um dos testes mais poderosos na identificação de normalidade nos dados. As hipóteses envolvidas no teste são:

$$\begin{cases} H_0 : \text{A amostra vem de uma população normal;} \\ H_1 : \text{A amostra não vem de uma população normal.} \end{cases}$$

A maioria dos softwares de análise estatística fazem este teste. Por exemplo no software R a função que faz o teste de Shapiro-Wilk é a *shapiro.test()*.

Exemplo: Determinado tipo de freezer foi regulado para manter a temperatura em 2,5°C. Ao longo do dia a temperatura foi medida. Antes de realizar inferências a partir desses dados, a normalidade da variável “temperatura (°C) média no interior do freezer” foi investigada pela função *shapiro.test(temperatura)* do software R e o resultado é apresentado a seguir:

```
Shapiro-Wilk normality test

data: temperatura
W = 0.95969, p-value = 0.8058
```

Neste caso, como o *p-valor* foi maior que o nível de significância comumente adotado de 5% ($0.8058 > 0.05$), então aceita-se H_0 . Portanto podemos concluir que os dados de temperatura no interior do freezer possuem distribuição normal de acordo com o teste de Shapiro-Wilk.

6.9 LISTA DE EXERCÍCIOS 9: Testes de hipóteses

1- Construa hipóteses para os seguintes problemas e desenhe as regiões de rejeição e não rejeição de H_0 .

- a)** A prefeitura de uma cidade afirma que o tempo médio de espera na fila de atendimento é de 50 minutos. As pessoas acreditam que a espera demora mais tempo que o afirmado pela prefeitura.
- b)** Uma empresa que fabrica repelente de insetos garante que seu produto tem efeito médio de 10 horas. Nós acreditamos que, na verdade, este efeito dura bem menos.
- c)** Com receio da fiscalização, um fabricante de doces deseja evitar a deficiência e o excesso no enchimento dos potes de 250g de geléia.
- d)** Determinada vacina padrão possui eficiência de 90%. Uma empresa criou uma nova fórmula para a vacina, no entanto, antes de colocar no mercado, deseja saber se seu produto possui uma eficiência diferente do produto padrão.
- e)** Um grupo de consumidores deseja saber se um veículo apresenta um consumo maior do que o anunciado pelo fabricante que é de 10km/litro de combustível.

2- Teste a hipótese de que o conteúdo médio de recipientes de certo lubrificante é dez litros, se os conteúdos de uma amostra aleatória de dez recipientes são 10,2; 9,7; 10,1; 10,3; 10,1; 9,8; 9,9; 10,4; 10,3 e 9,8 litros. Use o nível de significância 0,01 e assuma que a distribuição dos conteúdos dos recipientes é normal.

3- Um comprador de blocos de cimento acredita que a qualidade dos produtos da marca A esteja se deteriorando. Sabe-se, por experiência passada, que a força média de esmagamento desses blocos era de 400 libras, com desvio padrão de 20 libras. Uma amostra de 100 blocos da marca A forneceu uma força média de esmagamento de 390 libras. Testar se a qualidade média dos blocos diminuiu ao nível de significância de 2,5%.

4- De acordo com um estudo sobre dietas, a alta ingestão de sódio pode estar relacionada a úlceras, câncer de estômago e enxaquecas. A necessidade humana de sal é de apenas 220 miligramas/dia, o que é ultrapassado na maioria das porções simples dos cereais prontos para servir. Se uma amostra aleatória de 20 porções similares de certo cereal tem média de conteúdo de sódio de 244 miligramas e desvio-padrão de 24,5 miligramas, isso sugere, ao nível de 0,05, que a média de sódio contido em uma porção de tal cereal é maior que 220 miligramas? Assuma uma distribuição normal para os conteúdos de sódio.

5- Uma empresa de óleo combustível afirma que 1/5 das casas de certa cidade é aquecido por óleo. Se, em uma amostra aleatória de mil casas dessa cidade, descobrimos que 136 são aquecidas por óleo, temos razão para acreditar que menos de 1/5 das casas é aquecido por óleo? Considere $\alpha = 1\%$.

6- Uma máquina de refrigerantes é considerada fora de controle se a variância dos conteúdos exceder 1,15 decilitros. Se uma amostra aleatória de 25 copos de bebida dessa máquina tem variância de 2,03 decilitros, isso indica, ao nível de significância de 0,05, que a máquina está fora de controle? Assuma que os conteúdos têm distribuição aproximadamente normal.

7- Em um problema no qual o pesquisador precisa comparar duas médias temos quantas situações (testes) possíveis? Quais são elas? Suponha que você quer comparar duas médias, quais informações você precisa para identificar qual teste de comparação de médias utilizar? Como você pode fazer para identificar cada uma destas informações?

8- Um experimento foi conduzido para comparar os conteúdos de álcool em um molho de soja em duas linhas de produção diferentes. A produção foi monitorada oito vezes por dia. Os dados de conteúdo de álcool são mostrados a seguir.

Linha 1	0,48	0,39	0,42	0,52	0,40	0,48	0,52	0,52
Linha 2	0,38	0,37	0,39	0,41	0,38	0,39	0,40	0,39

Assuma que ambas as populações são normais. Suspeita-se que a linha 1 não está produzindo tão constantemente (em termos de variabilidade) como a linha 2 em relação ao conteúdo de álcool. Teste a hipótese de que $\sigma_1^2 = \sigma_2^2$ com um nível de significância de 5%.

9- Camadas de óxidos em pastilhas de semicondutores são atacadas com uma mistura de gases, de modo a atingir a espessura apropriada. A variabilidade na espessura destas camadas de óxidos é uma característica crítica da pastilha de modo que uma baixa variabilidade é desejada para as próximas etapas do processo. Duas misturas diferentes de gases estão sendo estudadas por engenheiros de modo a determinar se uma delas é superior a outra no controle da variabilidade da espessura das camadas de óxido. Dezesseis pastilhas são atacadas com cada gás e os desvio-padrão obtidos são $S_1 = 1,96$ e $S_2 = 2,13$. Existem evidências ao nível e significância de 5% que indiquem que um gás é preferível em relação ao outro?

10- Duas companhias químicas podem fornecer uma matéria prima cuja concentração de determinado elemento é importante. A concentração média para ambos os fornecedores é a mesma, porém suspeita-se que a variabilidade na concentração pode diferir entre as companhias. O desvio-padrão da concentração de uma amostra de $n_1 = 10$ remessas produzidas pela companhia 1 é $S_1 = 9,7 \text{ g/l}$, enquanto que na companhia 2, uma amostra de $n_2 = 16$ remessas resulta em $S_2 = 5,8 \text{ g/l}$. Há evidências suficientes para concluir que a variância das duas populações são diferentes ao nível de 5%?

11- A gerente de uma indústria de suco de laranja enlatado está interessada em comparar o desempenho de duas linhas de produção diferentes de sua fábrica. Como a linha 1 é relativamente nova, ela suspeita que sua produção em número de caixas por dia seja maior do que o número de caixas produzidas pela linha mais velha, 2. Selecionam-se aleatoriamente dez dias de dados de cada linha, encontrando-se $\bar{x}_1 = 824,9$ e $\bar{x}_2 = 818,6$ caixas por dia. Devido à experiência com a operação desse tipo de equipamento, sabe-se que $\sigma_1^2 = 40$ e $\sigma_2^2 = 50$. Faça um teste de hipóteses apropriado, considerando $\alpha = 5\%$. Assuma que a quantidade de caixas produzidas por dia seja aproximadamente normal.

12- De acordo com a publicação *Chemical Engineering*, uma importante propriedade das fibras é a absorção de água. A porcentagem de absorção média de 25 pedaços de fibra de algodão selecionados aleatoriamente foi de 20 com desvio-padrão de 1,5. Uma amostra aleatória de 25 pedaços de acetato rendeu uma média de 12, com desvio-padrão de 1,25. Usando um nível de significância de 5%:

a) o que se pode dizer sobre as variâncias de ambas as populações (algodão e acetato)? Use um teste de hipóteses apropriado.

b) há fortes evidências de que a média populacional da porcentagem de absorção do algodão é significativamente maior que a do acetato? (Na escolha do teste considere o resultado obtido no item anterior).

13- Os dados a seguir referem-se a carga máxima (kN) de dois tipos diferentes de vigas.

Tipo	Tamanho amostral	\bar{x}	S
Grade de fibra de vidro	26	33,4	2,2
Grade de carbono comercial	26	42,8	4,3

Assumindo que as distribuições subjacentes sejam normais, teste a hipótese de que a média real das vigas de carbono é maior que a das vigas de fibra de vidro. Considere $\alpha = 1\%$. Obs: Para saber se as variâncias populacionais são iguais ou não, aplique um teste de hipóteses apropriado.

14- Em um estudo conduzido pelo Departamento de Nutrição Humana e Alimentos da Universidade da Virgínia, foram registrados os dados de comparação dos resíduos de ácido sórbico, em partes por milhão, em presunto imediatamente depois de mergulhado em uma solução de sorbato e após 60 dias de armazenamento.

Fatia	Antes do armazenamento	Após armazenamento
1	224	116
2	270	96
3	400	239
4	444	329
5	590	437
6	660	597
7	1400	689
8	680	576

Assumindo que as populações são normalmente distribuídas, há evidência suficiente, num nível de significância de 0,05, para dizermos que o tempo de armazenamento influencia as concentrações residuais de ácido sórbico?

15- Em um experimento designado a estudar os efeitos do nível de iluminação no desempenho da tarefa, as pessoas foram solicitadas a inserir uma sonda de ponta fina no furo de 10 agulhas em sucessão rápida para um nível de luz baixo com fundo preto e um nível mais alto com fundo branco. Cada valor é o tempo necessário para completar a tarefa.

Pessoas	fundo preto	fundo branco
1	25,85	18,23
2	28,84	20,84
3	32,05	22,96
4	25,74	19,68
5	20,89	19,50
6	41,05	24,98
7	25,01	16,61
8	24,96	16,07
9	27,47	24,59

Os dados indicam que o nível mais alto de iluminação produz uma diminuição de mais de 5s no tempo de conclusão médio real da tarefa? Considere o nível de significância de 5% e suponha que o tempo para completar a tarefa seja modelado por uma distribuição normal.

16- Dois tipos diferentes de ligas, A e B, foram usados para fabricar espécimes experimentais de uma junta de baixa tensão a ser usada em uma determinada aplicação de engenharia. A tensão de ruptura (ksi) de cada espécime foi determinada, e os resultados são resumidos na distribuição de frequências a seguir.

	A	B
26 – 30	6	4
30 – 34	12	9
34 – 38	15	19
38 – 42	7	10
Total	$m = 40$	$n = 42$

a) Teste a hipótese de que as proporções reais dos espécimes de ligas A e B que possuem uma tensão de ruptura de pelo menos 34 ksi são diferentes a um nível de 5%.

b) Obtenha um intervalo de 95% para a diferença entre as proporções reais dos espécimes de ligas A e B que possuem uma tensão de ruptura de pelo menos 34 ksi e compare o resultado com aquele do item anterior.

17- Durante uma epidemia de resfriado em um inverno, 2000 bebês foram pesquisados por uma renomada indústria farmacêutica para determinar se o medicamento da empresa é eficaz após dois dias. Entre 120 bebês que tiveram resfriado e receberam o medicamento, 29 foram curados dentro desse prazo. Entre 280 bebês que não receberam o medicamento, 56 foram curados em dois dias. Há alguma indicação significativa que apoia a afirmação da empresa sobre a eficácia do medicamento? Considere $\alpha = 3\%$.

18- Dor leve nas costas (DLC) é um problema de saúde sério em muitos ambientes industriais. Os dados a seguir referem-se ao intervalo de movimentos laterais (graus) para uma amostra de trabalhadores sem um histórico de DLC e para outra amostra com histórico dessa doença.

Condição	Tamanho amostral	\bar{x}	S
Sem DLC	28	91,5	5,5
DLC	31	88,3	7,8

a) Calcule um IC a 90% da diferença entre a extensão média de movimentos laterais da população para as duas condições. O intervalo sugere que o movimento lateral médio da população é diferente para as duas condições? Considere variâncias populacionais diferentes.

b) A conclusão é diferente se usarmos um nível de confiança de 95%?

19- São dados os pares de valores P e níveis de significância, α . Para cada par, afirme se o valor p observado levaria à rejeição de H_0 no nível de significância dado.

- a)** Valor $p = 0,084$ e $\alpha = 0,05$
- b)** Valor $p = 0,003$ e $\alpha = 0,001$
- c)** Valor $p = 0,498$ e $\alpha = 0,05$
- d)** Valor $p = 0,084$ e $\alpha = 0,10$
- e)** Valor $p = 0,039$ e $\alpha = 0,01$
- f)** Valor $p = 0,218$ e $\alpha = 0,10$

Respostas:

- 2:** Não se rejeita H_0 ($t_c = 0,77$).
- 3:** A qualidade dos blocos tem se deteriorado ($Z_c = -5$).
- 4:** Rejeita-se H_0 ($t_c = 4,38$).
- 5:** Rejeita-se H_0 ($Z_c = -5,06$).
- 6:** Rejeita-se H_0 ($\chi^2_c = 42,37$).
- 7:** Ver no caderno.
- 8:** Rejeita-se H_0 ($F_c = 19,76$).
- 9:** Aceita-se H_0 ($F_c = 0,85$).
- 10:** Aceita-se H_0 ($F_c = 2,79$).
- 11:** Rejeita-se H_0 ($Z_c = 2,1$).
- 12: a)** As variâncias são iguais ($F_c = 1,44$); **b)** Rejeita-se H_0 ($t_c = 20,48$).
- 13:** As variâncias são diferentes ($F_c = 3,82$). Rejeita-se H_0 ($t_c = 9,92$).
- 14:** Rejeita-se H_0 ($t_c = 2,67$).
- 15:** Rejeita-se H_0 ($t_c = 1,87$).
- 16: a)** Não se rejeita H_0 .
- 17:** Não se rejeita H_0 ($z_c = 0,93$).
- 18: a)** $IC(\mu_1 - \mu_2) : [0,3; 6,1]$.
- 19:** H_0 seria rejeitada apenas no item d.

7 Correlação e Regressão Linear Simples

Muitos problemas em engenharia e ciências envolvem explorar as relações entre duas ou mais variáveis. Quando estas variáveis são todas quantitativas podemos estabelecer uma relação matemática entre elas e estimar os parâmetros que modelam esta relação.

7.1 Correlação

Quantificar o nível de associação entre duas variáveis é uma ideia de suma importância na investigação científica nas mais diversas áreas do conhecimento. A “correlação” mede a força, ou o grau, de relacionamento entre duas variáveis.

A intensidade da associação linear simples entre duas variáveis quantitativas contínuas X e Y, pode ser mensurada pelo coeficiente de correlação de Pearson. Este coeficiente é estimado por:

$$r = \frac{SP_{XY}}{\sqrt{SQ_X SQ_Y}}$$

em que:

$$SQ_X = \sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i \right)^2}{n} \quad SQ_Y = \sum_{i=1}^n y_i^2 - \frac{\left(\sum_{i=1}^n y_i \right)^2}{n}$$

e

$$SP_{XY} = \sum_{i=1}^n x_i y_i - \frac{\left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n y_i \right)}{n}$$

O valor de “r” varia de -1 à 1, de modo que:

- i) Valores próximos de 1 indicam uma associação positiva entre as variáveis. Isto é, conforme aumenta-se uma, a outra também aumenta.
- ii) Valores próximos de -1 indicam um associação negativa entre as variáveis. Ou seja, conforme aumenta-se uma, a outra diminui.
- iii) Valores próximos de zero indicam não haver associação linear entre as variáveis.

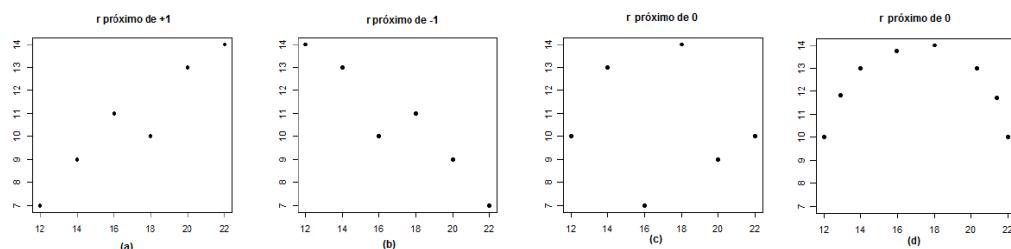


Figura 23: Exemplos de níveis de associação entre variáveis. Positivamente correlacionadas (a), negativamente correlacionadas (b) e falta de associação linear (c), (d).

7.2 Regressão

Geralmente o pesquisador deseja descrever o comportamento de uma variável chamada de “variável dependente” em função das demais chamadas de “variáveis independentes”. Tal relação entre as variáveis pode ser descrita por meio de funções matemáticas, as quais são chamadas modelos de regressão.

A palavra “regressão” foi empregada originalmente pelo estatístico inglês *Francis Galton* no século XIX, que, estudando a estatura das pessoas, elaborou a proposição (posteriormente confirmada) de que filhos de pais muito altos tendem a ser mais baixos do que seus pais, o oposto ocorrendo com os filhos de pais muito baixos. Daí o nome do termo, pois *Galton* percebeu que existe uma tendência dos dados se deslocarem ou regredirem à média da população.

De modo geral, um modelo de regressão pode ser escrito por:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \dots + \beta_p X_p + \varepsilon$$

em que: Y é a variável dependente, a qual deseja-se explicar o comportamento; X_1, X_2, \dots, X_p são as p variáveis independentes, com as quais deseja-se explicar o comportamento de Y , $\beta_0, \beta_1, \dots, \beta_p$ são os parâmetros do modelo, os quais devemos estimar para descrever a relação entre a variável Y e as variáveis X 's e ε é o erro associado ao modelo, que possui média zero e variância constante.

7.2.1 Regressão Linear Simples

O modelo de regressão linear simples é um caso particular do modelo apresentado acima, é o mais simples de todos. Ele é uma tentativa de estabelecer uma equação linear de uma reta para descrever o relacionamento entre duas variáveis X e Y .

A vantagem do uso de regressão comparado aos métodos de estimação anteriores é que ao invés de estimar um único parâmetro, identifica-se uma relação que possa existir entre as variáveis na população. Assim podemos predizer valores para Y , mesmo em doses não observadas de X .

Por exemplo, em um experimento para verificar se a temperatura do molde influencia no encolhimento das peças. Veja na figura abaixo a diferença entre fazer uma regressão linear simples do encolhimento em função da temperatura ou simplesmente calcular o encolhimento médio.

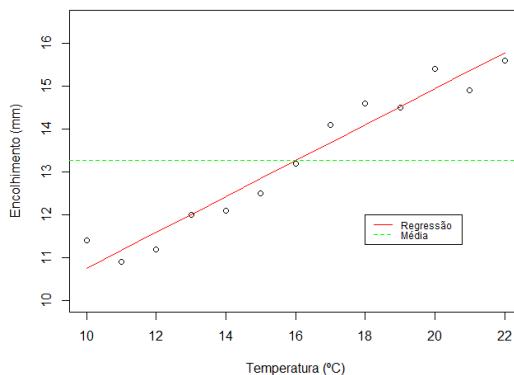


Figura 24: Relação entre uma equação de regressão linear simples e a média.

A equação de um modelo linear simples é dada por:

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

em que:

- Y é variável dependente;
- X é a variável independente;
- β_0 e β_1 são os parâmetros da equação;
- ε é o erro associado ao modelo, que possui média zero e variância constante.

Na regressão linear simples o β_0 é o intercepto da curva com o eixo Y e o β_1 é a inclinação da reta (coeficiente angular), se for positivo significa que a reta é crescente e se for negativo significa que a reta é decrescente.

OBS: O modelo de regressão estimado só é válido para realizar inferências dentro do intervalo observado na variável X . Fora deste intervalo estudos mais avançados são necessários.

Estimativas dos parâmetros

O objetivo de regressão é obter os estimadores de β_0 e β_1 , denominados \hat{a} e \hat{b} , respectivamente. Existem diferentes métodos de estimação destes parâmetros, sendo que um dos mais utilizados é o método de mínimos quadrados.

A ideia do método de mínimos quadrados consiste em obter os estimadores \hat{a} e \hat{b} de modo que as distâncias entre os pontos e a reta de regressão sejam mínimas. O método resulta nos seguintes estimadores:

$$\hat{b} = \frac{SP_{XY}}{SQ_X} \quad \hat{a} = \bar{Y} - \hat{b}\bar{X}$$

Coeficiente de determinação - R^2

Após obter as estimativas, uma maneira simples de medir a qualidade do ajuste obtido pela equação de regressão estimada é utilizando o coeficiente de determinação - R^2 .

O coeficiente de determinação varia de 0 à 100% e descreve o quanto da variabilidade ocorrida em Y é explicada pelo modelo de regressão estimado. É calculado por:

$$R^2 = \frac{SQ_{Regresso}}{SQ_{Total}}$$

No caso dos modelos de regressão linear simples temos que o coeficiente de determinação pode ser obtido calculando-se o quadrado do coeficiente de correlação de Pearson. **Mas cuidado**, tal relação só é válida no caso de modelos de regressão linear simples, caso contrário calcula-se o R^2 pela fórmula enunciada acima. A maioria dos softwares estatísticos estimam o R^2 ao obter as estimativas dos parâmetros do modelo de regressão.

Logo no nosso caso de regressão linear simples apenas fazemos: $R^2 = r^2$

EXEMPLO: Em um processo de moldagem por injeção, sabe-se que o encolhimento é influenciado por diversos fatores e, entre eles, está a velocidade da injeção, em pés por segundo, e a temperatura do molde, em graus Celsius. Os dados a seguir são referentes a um estudo onde observou-se a temperatura do molde, e o encolhimento em $cm \times 10^4$.

Temperatura ($^{\circ}\text{C}$)	10	12	14	16	18	20	22
Encolhimento ($cm \times 10^4$)	11,4	11,2	12,1	13,2	14,6	15,4	15,6

- a) Identifique qual é a variável independente (X) e qual é a variável dependente (Y).

X: temperatura; Y: encolhimento

- b) Verifique se existe evidências de que a temperatura influencia no encolhimento, isto é, estime a correlação entre as duas variáveis.

$$r = 0,9729$$

Como o coeficiente de correlação de Pearson está próximo de 1, indica que existe uma forte correlação positiva entre as variáveis. Sendo assim a temperatura pode ser utilizada para descrever o encolhimento por uma equação de regressão linear simples.

- c) Estime o modelo de regressão para prever o encolhimento em função da temperatura do molde. Interprete o modelo.

$$\hat{a} = 6,6428 \quad \hat{b} = 0,4196 \quad \text{Portanto: } \hat{Y} = 6,6428 + 0,4196X$$

No intervalo de temperatura de 10 à 22 $^{\circ}\text{C}$ existe um encolhimento médio de 6,6428 nas peças e para cada aumento de um grau $^{\circ}\text{C}$ na temperatura este encolhimento aumenta em 0,4196 $cm \times 10^4$.

- d) Qual é a porcentagem da variação no encolhimento do tamanho das peças é explicada pelo modelo de regressão?

Como é um modelo de regressão linear simples: $R^2 = 0,9729^2 = 0,9465$

Que significa que 94,65% da variação ocorrida no encolhimento das peças é explicada pelo modelo de regressão linear simples ajustado em função da temperatura.

- e) Estime o encolhimento médio caso a temperatura do molde seja de 15 $^{\circ}\text{C}$.

$$\hat{y} = 6,6428 + 0,4196X \Rightarrow \hat{y} = 6,6428 + 0,4196 \times 15 \Rightarrow \hat{y} = 12,9368cm \times 10^4$$

- f) É possível estimar o encolhimento médio caso a temperatura do molde seja de 29 $^{\circ}\text{C}$?

Não, pois a relação entre as duas variáveis (Temperatura, Encolhimento) só foi estudada no intervalo de temperatura de 10 à 22 $^{\circ}\text{C}$, assim só podemos realizar inferência para valores dentro do intervalo estudado.

7.3 LISTA DE EXERCÍCIOS 10: Correlação e Regressão Linear Simples

1- Um estudo de vida de prateleira do café torrado e moído foi realizado. Os testes sensoriais foram iniciados a partir do 9º dia de estocagem e depois a intervalos de mais ou menos 7 dias. Em cada época de avaliação sensorial uma amostra (pacote) foi obtida ao acaso. Seis provadores treinados avaliaram a amostra simultaneamente, julgando o produto quanto ao aroma em uma escala descritiva de 1 a 6 pontos: 6 = excelente; 5 = bom; 4 = aceitável; 3 = pouco aceitável; 2 = inaceitável e 1 = não bebível. Os resultados obtidos são apresentados na tabela a seguir.

Tempo de estocagem (dias)	Nota média para aroma
9	4,8
14	4,0
22	3,7
29	3,5
36	3,0
43	2,8

- a)** Identifique qual é a variável independente (X) e qual é a variável dependente (Y).
- b)** Calcule a correlação entre as duas variáveis.
- c)** Estime o modelo de regressão para prever a nota média dos avaliadores em função do tempo de estocagem. Interprete o modelo.
- d)** Qual é a porcentagem da variação das notas explicada pelo modelo de regressão?
- e)** Estime a nota média caso o tempo de armazenamento seja de 30 dias.
- f)** É possível estimar a nota média ao final de 2 meses de armazenamento? Porque?

2- Um estudo foi realizado para avaliar os efeitos da temperatura ambiente x no consumo de energia elétrica de uma indústria química, y . Outros fatores foram mantidos constantes e os dados foram coletados de uma fábrica experimental piloto.

y (BTU)	250	285	320	295	265	298	267	321
x (°F)	27	45	72	58	31	60	34	74

- a)** Faça um gráfico de dispersão (x,y).
- b)** Estime o coeficiente de correlação e interprete-o.
- c)** Obtenha as estimativas do modelo de regressão linear simples e interprete-o.
- d)** Preveja o consumo de energia para uma temperatura ambiente de 65°F.
- e)** Calcule o coeficiente de determinação e interprete-o.

3- Um professor em uma escola de negócios de uma universidade entrevistou uma dúzia de colegas sobre o número de reuniões profissionais de que eles participaram nos últimos cinco anos (X) e o número de trabalhos enviados por eles a revistas especializadas (Y) durante o mesmo período. Um resumo dos dados é fornecido a seguir:

$$n = 12, \bar{x} = 4, \bar{y} = 12, \\ \sum_{i=1}^n x_i^2 = 232, \quad \sum_{i=1}^n x_i y_i = 318.$$

Ajuste um modelo de regressão linear simples entre x e y determinando as estimativas do intercepto e da inclinação. Calcule o coeficiente de correlação entre estas variáveis e comente se o comparecimento em reuniões profissionais resultaria em mais trabalhos publicados.

OBS: Lembre-se que: $\hat{b} = \frac{SP_{XY}}{SQ_X}$ e $\hat{a} = \bar{Y} - \hat{b}\bar{X}$
e que:

$$SQ_X = \sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i \right)^2}{n} \quad SP_{XY} = \sum_{i=1}^n x_i y_i - \frac{\left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n y_i \right)}{n}$$

Respostas:

1: **b)** $r = -0,9658$; **c)** $\hat{y} = 5,0025 - 0,0537x$; **d)** $R^2 = 0,9327$
e) $\hat{y} = 4,32$ **f)** Não.

2: **b)** $r = 0,9896$ **c)** $\hat{y} = 218,25 + 1,3839x$; **d)** 308,20; **e)** $R^2 = 0,9793$ ou 97,93%

3: $\hat{y} = 37,8 - 6,45x$;

8 Referências Bibliográficas

- 1) MONTGOMERY, D. C. RUNGER, G. C. **Estatística Aplicada e Probabilidade para Engenheiros.** 6^a edição, John Wiley and Sons, 2016. 629p.
- 2) WALPOLE, R.E.; MYERS, R.H.; MYERS, S.L.; YE, K. **Probabilidade e Estatística para engenharia e ciências.** 8^a.ed.. São Paulo: Pearson Education do Brasil Editora, 2009, 491p.
- 3) HINES, W.W.; MONTGOMERY, D. C.; GOLDSMAN, D. M.; BORROR, C. M. **Probabilidade e Estatística na Engenharia.** 4^a ed.. Rio de Janeiro: LTC Editora, 2003, 604p.
- 4) DEVORE, J. L. **Probabilidade e Estatística para engenharia e Ciências,** 6^a ed. Tradução Joaquim Pinheiro Nunes da Silva. – São Paulo: Cengage Learning, 2006, 692p.
- 5) FERREIRA, D.F. **Estatística Básica.** 2^a ed.. Lavras: Editora da UFLA, 2009, 663p.
- 6) OLIVEIRA, M.S. de; BEARZOTI, E.; VILAS BOAS, F.L.; NOGUEIRA, D.A.; NICOLAU, L.A.; OLIVEIRA, H.S.S. de. **Introdução à Estatística.** 2^a ed.. Lavras: Editora da UFLA, 2014. 462p.
- 7) SPIEGEL. M.R. **Probabilidade e Estatística.** Tradução (de) Alfredo Alves de Farias. Makron Books, São Paulo, 2004, 518p.
- 8) MAGALHÃES, M.N.; LIMA, A.C.P. **Noções de Probabilidade e Estatística.** 4^a ed. São Paulo: Edusp, 2002. 392p.
- 9) MORETTIN, L.G. **Estatística Básica.** São Paulo, Pearson Prentice Hall.2010. 276p.
- 10) BUSSAB, W. O.; MORETTIN, P. A. **Estatística Básica.** 8a ed., Editora Saraiva, São Paulo. 2013, 548p.
- 11) ROSS, S. M. **Introduction to probability and Statistics for Engineers and Scientists.** 3^a edição. Elsevier Academic Press, 2004. 624p.

