UNIVERSIDADE FEDERAL DE UBERLÂNDIA

Apostila de Estatística

PARA OS CURSOS DE ENGENHARIA DE AGRIMENSURA E CARTOGRÁFICA,
SITEMAS DE SISTEMAS DA INFORMACAO E GEOLOGIA

PROF: Vânia de F. L. Miranda **DISCIPLINA: Estatística**

Ementa

1 – ESTATISTICA DESCRITIVA

2 - MEDIDAS DE POSIÇÃO

- Média aritmética
- Mediana
- Moda

3 – MEDIDAS DE DISPERSÃO

- Amplitude total
- Desvio médio absoluto
- Variância e Desvio-padrão
- Coeficiente de variação
- Quantis: quartil, decil e percentil
- Medidas de posição e dispersão no Excel

4 – TEORIA DAS PROBABILIDADES

- Experimento aleatório
- Espaço amostral
- Eventos
- Conceito clássico de probabilidade
- Conceito axiomático de probabilidade
- Teorema do Produto e Teorema de Bayes

5 – VARIÁVEIS ALEATÓRIAS

- Conceito de variável aleatória
- Variável aleatória discreta
- Distribuição de probabilidade simples e acumulada
- Variável aleatória contínua
- Função densidade de probabilidade e função de distribuição de probabilidade

6 – DISTRIBUIÇÕES DE PROBABILIDADE

- Distribuição de Bernoulli
- Distribuição uniforme
- Distribuição binomial
- Distribuição de Poisson
- Distribuição hipergeométrica

Distribuição exponencial

Distribuição normal

Distribuições de probabilidade no Excel

7 – TEORIA DA AMOSTRAGEM

Conceito probabilístico de amostragem

Amostr agem com e sem reposição

Tipos de amostragem: amostragem aleatória simples, sistemática, estratificada e amostragem por conglomerados.

8 – ESTIMAÇÃO DE PARÂMETROS

Estimadores das características populacionais com base na amostra

Estimadores pontuais e por intervalos de confiança

Estimação da média populacional

Estimação da proporção populacional

Estimação da variância populacional

9 – TESTE DE HIPÓTESES

Conceitos iniciais de teste de hipótese

Erros de estimação: erro tipo I e erro tipo II

Teste de hipóteses para uma média

Teste de hipóteses para duas médias

Teste de hipóteses para a proporção

Teste de hipóteses para a variância

10 – CORRELAÇÃO E ANÁLISE DE REGRESSÃO

Diagrama de dispersão

Coeficiente de correlação de Pearson

Regressão linear simples: método dos mínimos quadrados

Testes de significância para os parâmetros de regressão

Análise de regressão no Excel

BIBLIOGRAFIA RECOMENDADA

BUSSAB, W. O. & MORETTIN, P. Estatística Básica. São Paulo: Atual Editora, 2002.

COSTA NETO, P. L. Estatística. São Paulo: Editora Edgard Blucher, 2002.

COSTA NETO, P.L. & CYBALISTA, M. Probabilidades, resumos teóricos exercícios resolvidos, exercícios propostos. São Paulo: Editora Edgard Blucher, 1974.

MEYER, P.L. Probabilidade - Aplicação à Estatística. Rio de Janeiro: LTC Editora, 1980.

MORETTIN, L. G. Estatística Básica – Probabilidade. Vol. 1. São Paulo: Makron Books, 1999.

MORETTIN, L. G. Estatística Básica – Inferência. Vol. 2. São Paulo: Makron Books, 1999.

TRIOLA, M. F. Introdução à Estatística. 7a. ed. Rio de Janeiro: LTC - LTC Editora,1999.

1 – Estatística Descritiva

Introdução

Todos nós temos um pouco de cientista. Quase que diariamente, temos “palpites” com relação a acontecimentos futuros em nossas vidas, a fim de prever o que acontecerá em novas situações ou experiências. À medida que essas situações ocorrem, podemos, às vezes, confirmar ou sustentar nossas ideias: outras vezes, entretanto, não temos tanta sorte e, por isso, acabamos experimentando consequências desagradáveis.

Tomemos alguns exemplos familiares: poderíamos investir na bolsa de valores, votar em algum candidato que prometesse resolver os problemas nacionais, jogar nos cavalos, tomar um remédio para reduzir os incômodos de um resfriado, jogar dados num cassino, tentar “adivinhar” o que nossos professores irão perguntar nas provas ou acertar (ás cegas) um encontro com uma garota desconhecida, marcado através do amigo.

Às vezes, ganhamos; às vezes perdemos. Desse modo, poderíamos fazer um belo investimento na bolsa, mas reconhecer que não fizemos a escolha do melhor candidato; poderíamos ganhar no cassino, mas descobrir que tomamos o remédio errado para nossa doença; conseguir aprovação nas provas, mas penar na companhia arranjada pelo amigo; e assim por diante. A verdade é que, infelizmente, nem todas as nossas previsões acabam-se tornando realidade.

O QUE E ESTATÍSTICA?

Estatística é o estudo das populações, das variações e dos métodos de redução de dados (R. A. Fisher)

DEFINICAO: *A estatística é uma parte da matemática aplicada que fornece métodos para planejar experimentos, obter dados e organiza-los, resumi-los, analisa-los, e interpreta-los e deles extrair conclusões.*

É dividida em duas partes: a **estatíst*ica descritiva e a inferência estatística***. A ***estatística descritiva*** se refere à maneira de como coletar, de apresentar um conjunto de dados em tabelas e gráficos e à maneira de resumir, através de certas medidas as informações contidas nesses dados; a ***inferência estatística*** se refere à maneira de estabelecer conclusões para toda uma população quando se observou apenas parte dessa população (amostra).

A estatística mantém com a matemática uma relação de dependência, solicitando-lhe auxílio, sem o qual não poderia desenvolver-se. Com as outras ciências mantém a relação de complemento, quando utilizada como instrumento de pesquisa. Em especial esta última é a relação que a estatística mantém com a Administração e Ciências Contábeis, Serviço Social, Marketing, Logística, etc. servindo como instrumento auxiliar na tomada de decisões.

Por que estudar Estatística?

O uso de técnicas computacionais pode parecer um problema para o pesquisador cujo treino e interesse não envolvam a matemática. Entretanto, a estatística tem aparecido, cada vez com maior freqüência, na literatura especializada. Então é razoável que os profissionais de área humanas adquiram um mínimo de conhecimento técnico sobre estatística. Além disso, é razoável que esse conhecimento seja combinado com um ponto de vista objetivo sobre a natureza da matéria, para que o profissional possa avaliar a importância do uso da estatística e ter segurança nas interpretações.

Outro resultado do estudo da estatística mais importante do que parece à primeira vista, é a familiarização com o "jargão" da estatística. A falta de conhecimento de certos termos pode resultar na total incompreensão de um artigo. A estatística utiliza termos que pertencem ao nosso vocabulário comum como amostra, população, média, variabilidade, correlação, regressão, mas dá-lhes um sentido técnico e específico. È claro que o conhecimento do significado comum é útil, mas pode conduzir à interpretação inadequada quando substitui o significado técnico.

Quando o pesquisador usa a estatística?

A estatística auxilia o pesquisador nas seguintes fases do trabalho:

a) na amostragem de dados ou no delineamento de um experimento;
b) na interpretação tabular e gráfica e no estudo descritivo de dados;
c) na análise de dados.

A estatística no dia-a-dia.

No mundo atual, a empresa é uma das vigas-mestras da Economia dos povos.

A direção de uma empresa de qualquer tipo, incluindo as estatais e governamentais, exige de seu administrador a importante tarefa de tomar decisões, e o conhecimento e o uso da Estatística facilitarão seu tríplice trabalho de organizar, dirigir e controlar a empresa.

Por meio de sondagem, de coleta de dados e de recenseamento de opiniões, podemos conhecer a realidade geográfica e social, os recursos naturais, humanos e financeiros disponíveis, as expectativas da comunidade sobre a empresa, e estabelecer suas metas, seus objetivos com maior possibilidade de serem alcançados a curto, médio ou longo prazo.

A Estatística ajudará em tal trabalho, como também na seleção e organização da estratégia a ser adotada no empreendimento e, ainda, na escolha técnica de verificação e avaliação da quantidade e da qualidade do produto e mesmo dos possíveis lucros e/ou persas.

Tudo isso que se pensou, que se planejou, precisa ficar registrado, documentado para evitar esquecimentos, a fim de garantir o bom uso do tempo, da energia e do material e, ainda, para um

controle eficiente do trabalho.

O esquema do planejamento é o **plano**, que pode ser resumido, com auxilio da Estatística em **tabelas** e **gráficos,** que facilitarão a compreensão visual dos cálculos matemático-estatísticos que lhes deram origem.

O homem de hoje, em suas múltiplas atividades, lança mão de processos e técnicas estatísticos, e só estudando-os evitaremos o erro das generalizações apressadas a respeito de tabelas e gráficos apresentados em jornais, revistas e televisão, frequentemente cometido quando se conhece apenas "por cima" um pouco de estatística.

Conceitos Fundamentais

População e amostra

- **População**: é o conjunto de elementos que têm, em comum, determinada característica (pessoas, coisas, objetos).
- **Amostra**: é todo subconjunto não vazio e com menor número de elementos do que o conjunto definido como população. .
- **Dados*:*** São informações obtidas, seja com base nos elementos que constituem a população, seja com base nos elementos que constituem a amostra.
- **Tendenciosidade:** todos os elementos da população tem que ter a mesma chance de fazer parte da amostra. Se existir elementos com maior ou menor possibilidade de participar da amostra então há tendenciosidade.

Variáveis (x_i):

É convencionalmente, o conjunto de resultados possíveis de um fenômeno.

As observações se constituem no material básico com que o pesquisador trabalha. Para que a estatística possa ser aplicada a essas observações, elas devem estar na forma de números. Exemplo: o tempo de percurso, para o trabalho dos empregados de um grande escritório, notas de um teste de coordenação física.

Estes números são os **dados** e a característica comum inerente aos mesmos é a **variabilidade ou variação** que apresentam. Essa característica, que pode assumir diferentes valores de indivíduo para indivíduo é chamada de **variável**.

Classificação das variáveis

- **Qualitativas:** são qualidades (ou atributos) podem ser separados em diferentes categorias que se distinguem por alguma característica não numérica.

Exemplos: sexo, religião, naturalidade, cor dos olhos, faixa etária, etc.

- **Quantitativas:** são números que representam contagens ou medidas, e podem ser;
  - **Contínuas:** variável que assume, teoricamente, qualquer valor entre dois limites (medidas por algum aparelho).

    Exemplo: peso, altura, etc.
  - **Discretas:** variável resultante de um conjunto finito de valores possíveis (contagens ou enumerações)

    Exemplos: quantidade de estudantes em uma disciplina, número de funcionários de uma empresa, número de filhos, etc.

Às vezes coletamos dados visando um fim especifico, ou obtemos dados não com uma finalidade especifica, mas porque desejamos explorá-los para ver o que pode se revelado.

Exercício (extraído de Bussab & Morettin (2003).

Um pesquisador está interessado em fazer um levantamento sobre alguns aspectos socioeconômicos dos empregados da seção de orçamentos da Companhia MB. Usando informações obtidas do departamento pessoal, ele elaborou a Tabela 1.1

Tabela 1: Informações sobre estado civil, grau de instrução, numero de filhos, salario (expresso como fração do salario mínimo), idade (medida em anos e meses) e procedência de 36 empregados da seção de orçamentos de uma Empresa.

$N^{\underline{0}}$	Estado civil	Grau de instrução	$N^{\underline{0}}$ de filhos	Salario	Idade		Região de procedência
					Anos	Meses	
1	Solteiro	Fundamental	0	4,00	26	3	Interior
2	Casado	Fundamental	1	4,56	32	10	Capital
...	...	...	...	...	...	...	...
35	Casado	Médio	2	19,40	48	11	Capital
36	Casado	Superior	3	23,30	42	2	Interior

Pode-se atribuir uma letra, digamos X, para representar tal variável. Observa-se na Tabela 1.1 que o pesquisador colheu informações sobre oito variáveis:

Tabela 1.1 - Variáveis de interesse do pesquisador.

Variável	Representação
Estado civil	X
Grau de instrução	Y
Número de filhos	Z
Salário	S
Idade	U
Região de procedência	V
Sexo	R
Classe social	T

a) Quais são variáveis qualitativas e quantitativas?

b) Classifique-as em nominais, ordinais, discretas e contínuas?

c) Agora, com base no que foi apresentado, elabore um exemplo análogo relacionado à sua área.

Coleta, Organização e Apresentação de dados

Os dados são coletados numa forma sem ordenação e sem nenhum tipo de arranjo sistemático. Nesse caso, eles são denominados de **dados brutos**. Então, esses dados sofrerão uma simples organização (ordenação) e serão denominados de **dados elaborados**.

Para ilustrar apresentaremos exemplo típico de dados **qualitativos nominais** na Tabela 2.1.

Tabela 2.1 - Dados brutos de marca de carros populares predominante em 25 cidades do triângulo, 1998.

Pálio	Corsa	Uno	Gol	Corsa
Uno	Gol	Uno	Pálio	Uno
Pálio	Uno	Gol	Corsa	Gol
Ka	Gol	Uno	Uno	Gol
Gol	Corsa	Gol	Uno	Uno

Um outro exemplo, agora de dados **quantitativos discretos** refere-se a contagem de ovos danificados no mercado municipal da cidade de Lavras, ao chegar um carregamento de ovos de uma cidade distante, os lojistas fizeram uma amostragem e inspecionaram 30 dúzias anotando o número de ovos danificados em cada uma delas. Os resultados do número de ovos danificados em cada dúzia

(embalagem) estão apresentados na Tabela 3.2 (Ferreira, 2005).

Tabela 3.1 - Dados brutos referentes ao número de ovos danificados em uma inspeção feita em 30 embalagens, de uma dúzia cada, em um carregamento para o mercado municipal de Lavras proveniente de uma cidade distante.

0	0	1	1	1	1	5	4	1	2
3	0	0	0	0	2	1	1	1	0
2	3	3	0	0	0	0	0	1	0

Essa representação dos dados nas Tabelas 3.1 e 3.2 é pouca informativa e para melhorá-la um pouco é possível ordenar os dados em uma sequência crescente ou decrescente ou agrupá-los quanto as suas categorias ou atributos. As Tabelas 3.3 e 3.4 contêm os dados das Tabelas 3.1 e 3.2, respectivamente, nessa nova organização. Na Tabela 3.3 são apresentados as marcas de carro de maior para menor frequência.

Tabela 4.1 - Dados elaborados de marca de carros populares predominante em 25 cidades do triângulo, 1998.

Uno	Uno	Gol	Gol	Corsa
Uno	Uno	Gol	Gol	Pálio
Uno	Uno	Gol	Corsa	Pálio
Uno	Uno	Gol	Corsa	Pálio
Uno	Gol	Gol	Corsa	Ka

Finalmente, na Tabela 3.4, estão apresentados os dados do número de ovos danificados na amostra de 30 dúzias do carregamento.

Tabela 5.1 - Dados elaborados referentes ao número de ovos danificados em uma inspeção feita em 30 embalagens, de uma dúzia cada, em um carregamento para o mercado municipal de Lavras proveniente de uma cidade distante.

0	0	0	1	2
0	0	1	1	3
0	0	1	1	3
0	0	1	1	3
0	0	1	2	4

Tabelas (ou Séries Estatísticas)

Os dados devem ser apresentados em tabelas construídas de acordo com as normas técnicas ditadas pela Fundação Instituto Brasileiro de Geografia e Estatística (Fundação IBGE)

Regras Gerais

Na construção de tabelas, os dados são apresentados em colunas verticais e linhas horizontais, conforme a classificação dos resultados da pesquisa.

Algumas recomendações preliminares são as seguintes:

a) A tabela deve ser simples. Tabelas simples são mais claras e objetivas. Desta forma, é conveniente que grandes volumes de informações sejam descritos em várias tabelas, em vez de em uma só.

b) A tabela deve ser auto-explicativa, isto é, sua compreensão deve estar desvinculada do texto.

c) Nenhuma casa da tabela deve ficar em branco, apresentando sempre um número ou um sinal.

d) Se houver duas ou mais tabelas em um texto, deverão receber um número, que será referido no texto.

e) As colunas externas de uma tabela não devem ser fechadas.

f) Na parte superior e inferior, as tabelas devem ser fechadas por linhas horizontais. O emprego de linhas verticais para a separação de colunas no corpo da tabela é opcional.

g) È conveniente que sejam evitados os arredondamentos. Quando for necessário, o arredondamento dos números que compõem a tabela deve ser efetuado segundo critérios de minimização de erros (com isto tenta-se evitar o acúmulo de erros de arredondamento decorrentes do processo de aproximação).

h) Deverá ser mantida uniformização quanto ao número de casas decimais.

i) Os totais e subtotais devem ser destacados.

j) A tabela deve ser maior no sentido vertical que no horizontal. Contudo, se uma tabela apresentar muitas linhas e poucas colunas (estreita demais), convém separá-la em uma maior quantidade de colunas. Neste caso, as colunas deverão ser separadas por linhas duplas.

Componentes das tabelas

Corpo: é o conjunto das informações que aparecem no sentido vertical e horizontal. Formado pelas linhas e colunas de dados

Cabeçalho: especifica o conteúdo das colunas

Coluna indicadora: é a divisão em sentido vertical, onde aparece a designação da natureza do conteúdo da linha. (Especifica o conteúdo das linhas).

Casa: São as divisões que aparecem no corpo da tabela.

Título: aparece sempre na parte superior da tabela, devendo ser sempre o mais claro e completo possível. Deve responder as perguntas: *o que? quando? onde?*, relativas ao fato estudado.

Rodapé: é um espaço na parte inferior da tabela utilizado para colocar informações necessárias referentes aos dados.

Fonte: é a indicação da entidade responsável pela elaboração da tabela. Deve ser colocada no rodapé, no final da tabela. Esse procedimento garante a honestidade científica e serve como indicativo para posteriores consultas.

Notas: também devem ser colocadas no rodapé, depois da fonte, de forma sintética. As notas têm caráter geral, referindo-se à totalidade da tabela. Devem ser enumeradas em algarismos romanos, quando existirem duas ou mais de duas (às vezes é usado o asterisco).

Quanto aos números, deve ser observado o seguinte:

d) Todo número inteiro constituído de mais de três algarismos deve ser agrupado de três em três, da direita para a esquerda, separando cada grupo por um ponto (ex.: 56.342.901) são exceções:

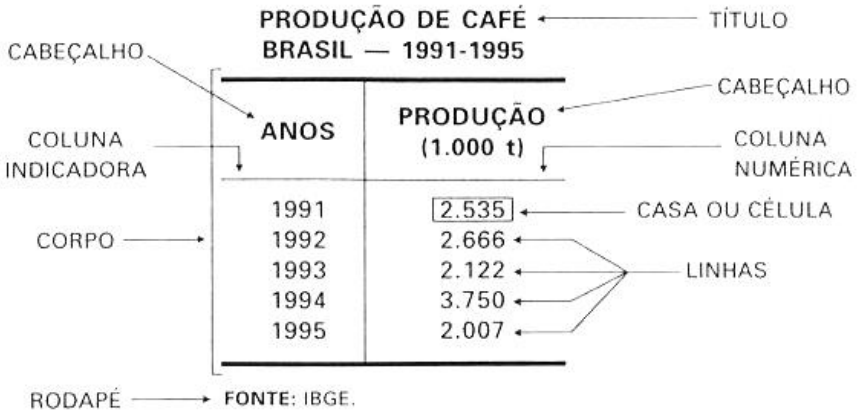
- *os algarismos que representam o ano (ex.: 2002)*
- *números de telefone (ex.: 622-9780)*
- *placas de veículos (ex.: GOX 3434)*

e) A parte decimal de um número deverá ser separada da parte inteira pela vírgula (ex.: 0,56)

f) A unidade de medida não leva o "s" do plural e nem o ponto final como abreviação (ex.: cm, m, kg, etc.).

g) Os símbolos de medida aparecem depois do número, sem espaço entre eles (ex.: 4,2m; 3h).

Exemplo:



Observações importantes:

- **Dados qualitativos** devem ser apresentados em tabelas com as frequências absolutas $\mu = \frac{\sum_{i=1}^{n} x_i}{N}$, frequências relativas$\sum_{i=1}^{n} x_i = x_1 + x_2 + x_3 + \ldots + x_n$ e frequências percentuais $\bar{x}_i$ e em gráficos em colunas ou em barras e de composição em setores (“pizza”).
- **Dados quantitativos** devem ser apresentados em tabelas com intervalos de classes juntamente com as frequências absolutas $Md = \frac{x_{\frac{n}{2}} + x_{\frac{n+2}{2}}}{2}$, frequências relativas$Md = x_{\frac{n+1}{2}}$, frequências percentuais $\bar{X} = \frac{\sum_{i=1}^{k} \bar{x}_i F_i}{n}$, ponto médio $\bar{x}_i$, e em gráficos chamados histograma e polígono de frequências.

  ***** Dados quantitativos discretos de menor variação devem ser organizados como os qualitativos.******

Tabelas de Distribuição de Frequências:

É uma tabela (série estatística) específica, onde os dados encontram-se dispostos em classes ou categorias juntamente com as frequências correspondentes. É importante ressaltar que essas representações não são, ainda, a melhor forma de apresentar os dados, pois se os tamanhos amostrais aqui apresentados fossem de ordem maior de dados (centenas ou milhares de dados), então essas representações consumiriam muito espaço e consequentemente seriam pouco funcionais para o propósito que se destinam.

Torna-se evidente a necessidade de resumir os dados, sem perda de muita informação contida neles. Dessa forma, para os dados qualitativos nominais e para os quantitativos discretos, percebe-se que eles poderiam ser resumidos agrupando suas categorias e apresentando-os em tabelas e gráficos.

Frequências absolutas, relativas e percentuais.

- **Frequências absolutas** (F_i)**:** são os dados estatísticos resultantes da coleta direta da fonte, sem outra manipulação senão a contagem ou medida.

  A leitura dos dados absolutos é sempre enfadonha e inexpressiva, embora esses dados traduzam um resultado exato e fiel, não têm a virtude de ressaltar de imediato as suas conclusões numéricas. Daí o uso imprescindível que a estatística faz dos dados relativos.
- **Frequências relativas** (F_{ri})**:** são os resultados de comparações por quociente (razões) que se estabeleça entre os dados absolutos e têm por finalidade realçar ou facilitar as comparações entre quantidades.

$$F_{ri} = \frac{F_i}{n}$$

em que *n* é o total de observações.

Traduzem-se os dados relativos, em geral, por meio de **percentagens**, **índices**, **coeficientes** e **taxas.**

- **Frequências percentuais** (F_{pi})**:** Traduzem-se os dados relativos, em geral, por meio de **percentagens**, **índices**, **coeficientes** e **taxas**.

$$F_{pi} = F_{ri} * 100$$

Tomamos 100 para base de comparação, também podemos tomar outro número qualquer, entre os quais destacamos o número 1. É claro que, supondo o total igual a 1, os dados relativos das parcelas serão todos menores que 1.

Regras para arredondamento de dados

a) quando o primeiro algarismo a ser abandonado for 0, 1, 2, 3 ou 4, fica inalterado o último algarismo a permanecer.

Ex: 48,23 = 48,2

b) quando o primeiro algarismo a ser abandonado for 6, 7, 8 ou 9, aumenta-se de uma unidade o último algarismo a permanecer.

Ex: 23,07 = 23,1 34,99 =35,0

Os dados qualitativos nominais da marca de carros populares predominantes em 25 cidades do triângulo em 1998 estão apresentados na Tabela 3.5.

Tabela 6.1 - Distribuição de frequências Absoluta, Relativa e Percentual da marca de carros populares predominante em 25 cidades do triângulo, 1998.

Marca	Freq. Abs. (f_i)	Freq. Rel. (f_r)	Freq. Perc. (fp(%))
Corsa	4	4/25 = 0,16	16
Gol	8	8/25 = 0,32	32
Ka	1	1/25 = 0,04	4
Pálio	3	3/25 = 0,12	12
Uno	9	9/25 =0,36	36
Σ	25	1,00	100

Na tabela 3.6, estão apresentados os dados referentes ao número de ovos danificados em uma inspeção feita em 30 embalagens de uma dúzia cada, em um carregamento para o mercado municipal de Lavras. Esses dados podem ser agrupados de modo análogo aos dados da marca de carros populares no triângulo.

Tabela 7.1 - Distribuição de frequências Absoluta, Relativa e Percentual referentes ao número de ovos danificados em uma inspeção feita em 30 embalagens, de uma dúzia cada, em um carregamento para o mercado municipal de Lavras proveniente de uma cidade distante.

Número de ovos quebrados (x_i)	Freq. Abs. (f_i)	Freq. Rel. (f_r)	Freq. Perc. (fp(%))
0	13	13/30 = 0,44	44
1	9	9/30 = 0,30	30
2	3	3/30 = 0,10	10
3	3	3/30 = 0,10	10
4	1	1/30 = 0,03	3
5	1	1/30 = 0,03	3
Σ	30	1,00	100

Gráficos

É uma forma de apresentação dos dados estatísticos, cujo objetivo é o de produzir, uma impressão mais rápida e viva do fenômeno em estudo, já que os gráficos falam mais rápido à compreensão que as tabelas.

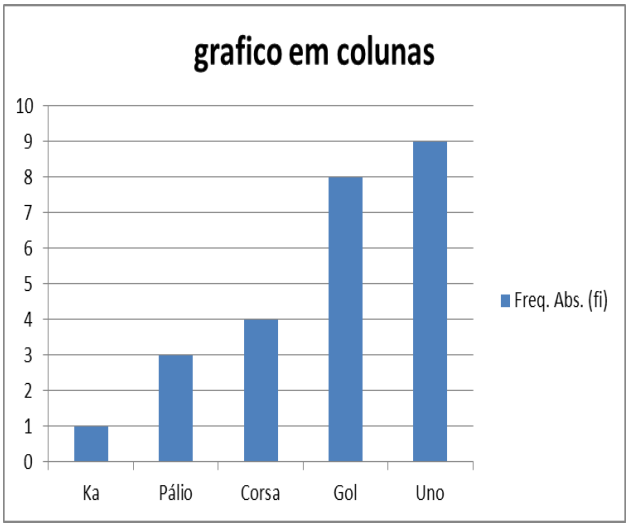
Gráfico em colunas ou em barras.

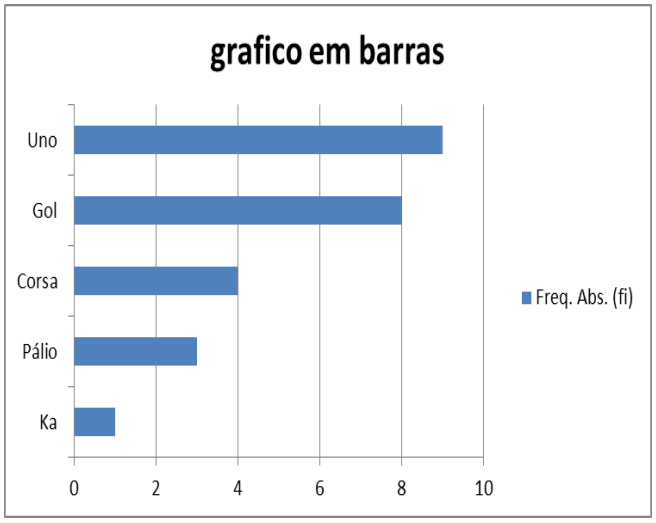
É a representação de uma tabela por meio de retângulos, dispostos verticalmente (em barras) ou horizontalmente (em colunas).

Quando em barras, os retângulos têm a mesma base e as alturas são proporcionais aos respectivos dados.

Quando em colunas, os retângulos têm a mesma altura e os comprimentos são proporcionais aos respectivos dados.

Gráficos em colunas e barras para representar o meio de transporte.





Outros exemplos encontrados em revistas e livros, em geral.

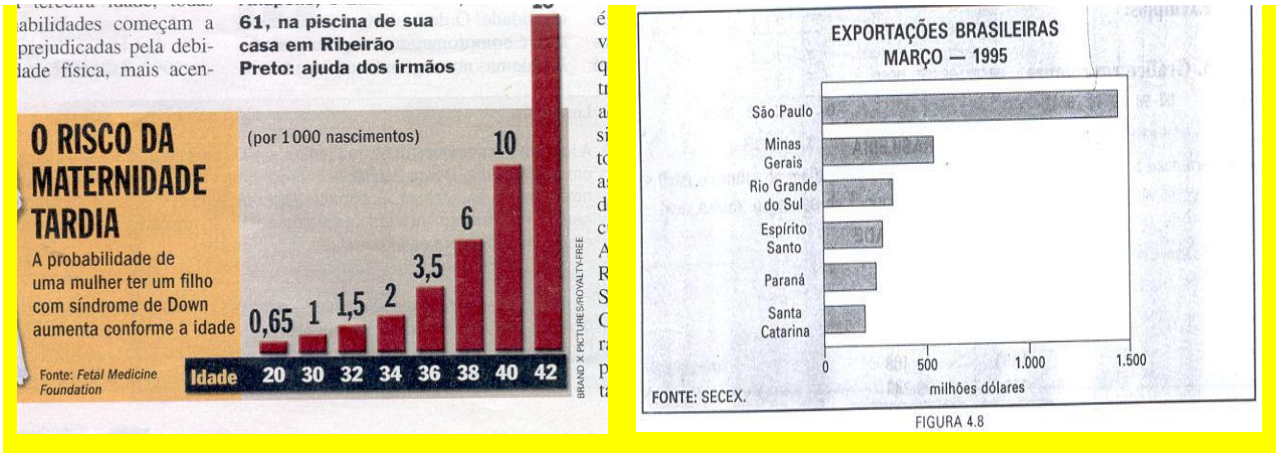
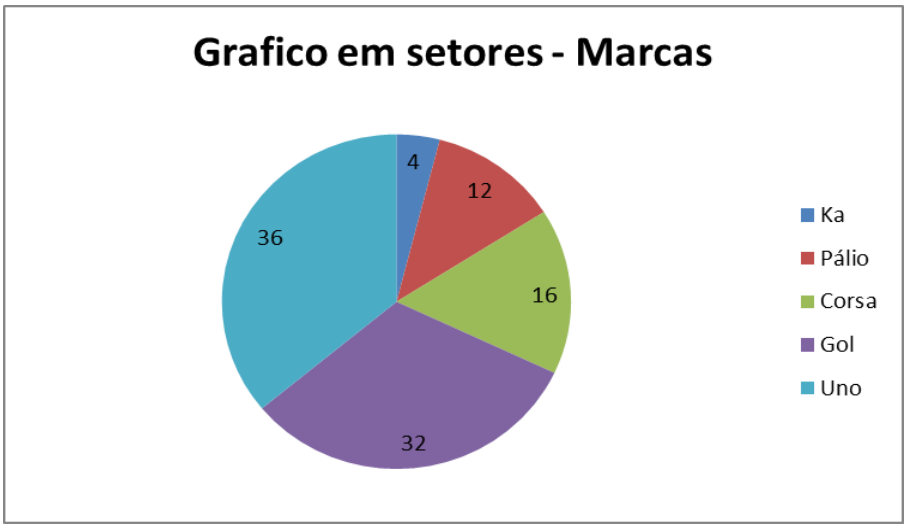


FIGURA 4.8

Gráfico em setores.

Este gráfico é constituído com base em um circulo. Pode ser construído através da formula:

$$\left.\begin{matrix} n \rightarrow 360^0 \\ F_i \rightarrow x \end{matrix}\right\} \Rightarrow nx = F_i * 360 \Rightarrow x = \frac{F_i * 360}{n}$$



Apresentação de dados quantitativos em tabelas distribuição de freqüências com intervalos de classes.

Os dados quantitativos são apresentados em distribuição de freqüências com intervalos de classes ou categorias, em que o numero de elementos pertencentes a cada classe é determinado e representa a freqüência de classe.

Algumas definições úteis:

1. **Dados brutos:** Dados originais na forma com que foram coletados (não foram numericamente organizados ou ordenados)
2. **Rol (Dados elaborados):** Dados numéricos arranjados em ordem crescente ou decrescente.

Algoritmo para a construção de tabelas com intervalos de classes:

I) Rol: organizar os dados coletados em ordem crescente ou decrescente.

II) Amplitude total (*A*): é a diferença entre o maior e menor valor da amostra (a partir do rol).

$$A = X_n - X_1 = maiorvalor - menorvalor$$

III) Número de classes (*k*): o numero de classes é escolhido por muitos autores como sendo um numero entre 5 e 20. A familiaridade do pesquisador com os dados é que deve indicar quantas classes devem ser construídas. Há 2 critérios propostos a seguir:

i) $k = \sqrt{n}$ para *n* ate 100$(n \leq 100)$

ii) $k = 5\,log_n$ se *n* for maior que 100 $(n > 100)$

IV) Amplitude de classe (c): é a diferença entre os limites superior e inferior de uma determinada classe.

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

V) Limite inferior da primeira classe (LI_{1^a})**:** o limite inferior da primeira classe deve ser um valor menor que o menor valor observado na amostra, uma vez que por mero acaso valores da população inferiores a X_1 podem não ter sido amostrados:

$$LI_{1^a} = X_1 - \frac{c}{2}$$

A forma de apresentação de uma classe adotada é dada por XX ├─YY, ou seja, a classe tem seu limite inferior XX incluído na classe e o seu limite superior YY excluído.

VI) Determinação das classes:

Para determinar as classes é preciso seguir os seguintes passos:

h) Somar ao valor inferior da primeira classe a amplitude de classe o obter-se o limite superior;

$$LS_{1^a} = LI_{1^a} + c$$

i) O limite superior da primeira classe será o limite inferior da segunda classe;

$$LI_{2^a} = LS_{1^a}$$

j) Repetem-se os passos (a) e (b) até completar as k classes, ou equivalentemente até que o maior valor esteja contido na ultima classe.

Tabelas de distribuição de frequências acumulas.

Outra possibilidade utilizada é fazer a tabela das distribuições de frequências acumulas.

Apresentação de dados quantitativos em gráficos

Histograma: Gráfico formado por retângulos cujas bases são proporcional ás amplitudes de classes e as alturas proporcionais às frequências $(F_i,\ F_{ri},\ F_{pi})$.

Polígono de frequências: Gráfico de linhas que une os pontos médios das classes no topo dos retângulos.

Ogivas: Gráficos de frequências acumuladas (“abaixo de” e “acima de”)

Exemplo: Conhecidas as notas de um teste aplicado a 50 funcionários de uma empresa após um curso de capacitação, faça:

84	68	33	52	47	73	68	61	73	77
74	71	81	91	65	55	57	35	85	88
59	80	41	50	53	65	76	85	73	60
67	41	78	56	94	35	45	55	64	74
65	94	66	48	39	69	89	98	42	54

a) Tabela de distribuição de frequências com intervalos de classes;
b) Histograma e um polígono de frequências num mesmo plano cartesiano:
c) Tabela de distribuição de frequências acumulas:
d) Ogivas

Resp.

a) Rol

33	35	35	39	41	41	42	45	47	48
50	52	53	54	55	55	56	57	59	60

61	64	65	65	65	66	67	68	68	69
71	73	73	73	74	74	76	77	78	80
81	84	85	85	88	89	91	94	94	98

Amplitude Total: $A = 98 - 33 = 65$

Número de classes: $k = \sqrt{n} = \sqrt{50} \cong 7{,}071 \cong 8$

Amplitude de classe: $c = \frac{A}{k-1} = \frac{65}{7} = 9{,}28 \cong 10$

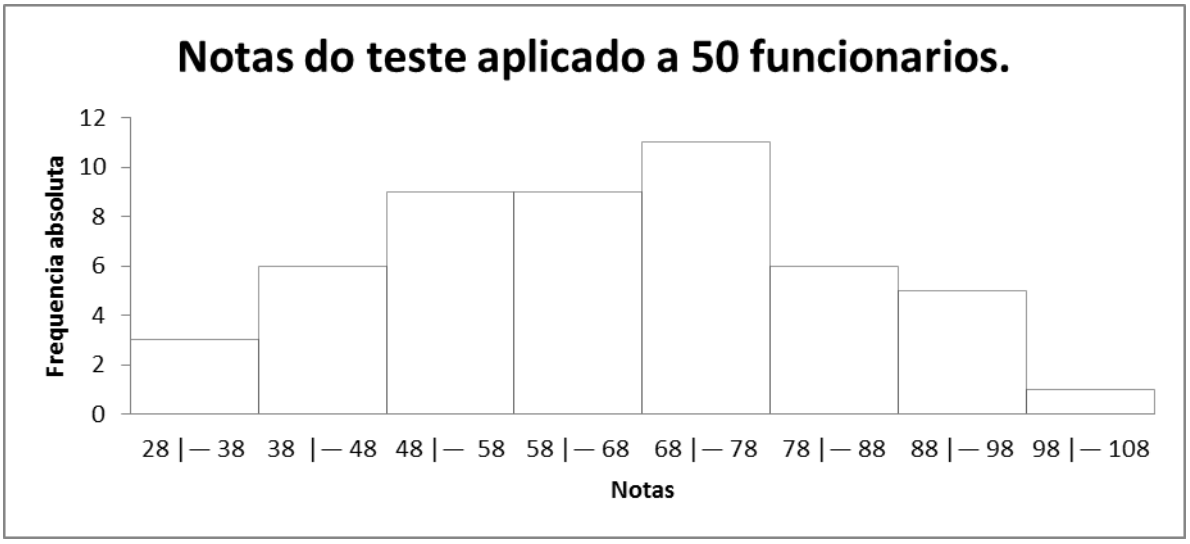
Limite Inferior da primeira classe: $LI_{1^a} = X_1 - \frac{c}{2} = 33 - \frac{10}{2} = 33 - 5 = 28$

Limite Superior da primeira classe: $LS_{1^a} = LI_{1^a} + c = 28 + 10 = 38$

Tabela com Intervalos de classes

Classes (i)	F_a	F_{ri}	F_{pi}	Ponto médio X(i)
28 |— 38	3	0.06	6	33
38 |— 48	6	0.12	12	43
48 |— 58	9	0.18	18	53
58 |— 68	9	0.18	18	63
68 |— 78	11	0.22	22	73
78 |— 88	6	0.12	12	83
88 |— 98	5	0.1	10	93
98 |— 108	1	0.02	2	103
Total	50	1	100	---

b)

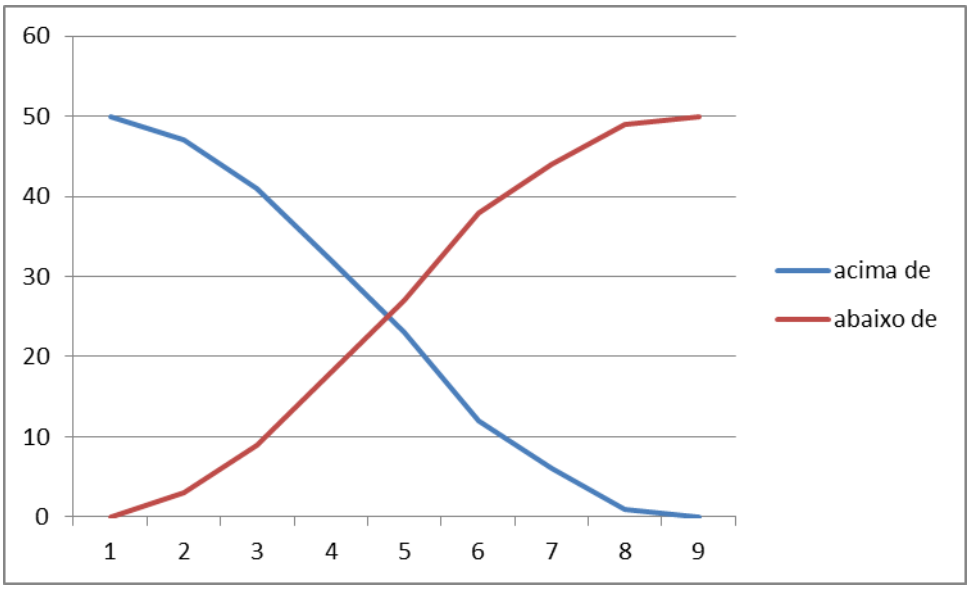


c)

Obser	acima de
28	50
38	47
48	41
58	32
68	23
78	12
88	6
98	1
108	0

obser	abaixo de
28	0
38	3
48	9
58	18
68	27
78	38
88	44
98	49
108	50

d)



INTERPOLAÇÃO EM DISTRIBUIÇÕES DE FREQUÊNCIA ACUMULADA

Exemplo: Dados da Tabela.

Limites (X_i)	$FC_i(X<X_i)$ = *fac*↓	$FC_i(X>X_i)$ = *fac*↑
-2,485	0	20
5,245	6	14
12,975	14	6
20,705	18	2
28,435	20	0

Qual a frequência acumulada abaixo de10?

$$7,73 \leftarrow \left\langle \begin{matrix} 5,245 & \rightarrow & 6 \\ 12,975 & \rightarrow & 14 \end{matrix} \right\rangle \rightarrow 8$$

Aplicando a regra de três simples temos:

$$\left.\begin{matrix} 7,73 & \rightarrow & 8 \\ 4,755 & \rightarrow & x \end{matrix}\right\} \Rightarrow x = \frac{8 * 4,755}{7,73} = 4,921$$

Então, abaixo de 10 tem-se: 4,921 + 6 = 10,921.

Qual a frequência acumulada acima de10?

$$7,73 \leftarrow \left\langle \begin{matrix} 5,245 & \rightarrow & 14 \\ 12,975 & \rightarrow & 6 \end{matrix} \right\rangle \rightarrow 8$$

Aplicando a regra de três simples temos:

$$\left.\begin{matrix} 7,73 & \rightarrow & 8 \\ 2,975 & \rightarrow & x \end{matrix}\right\} \Rightarrow x = \frac{8 * 2,975}{7,73} = 3,079$$

Então, acima de 10 tem-se: 3,079 + 6 = 9,079.

Exemplo: Dados fictícios.

X_i	f_i	$FC_i(X<X_i) = fac\downarrow$	$FC_i(X>X_i) = fac\uparrow$
0	5	5	80
4	10	15	75
8	45	60	65
12	12	72	20
16	5	77	8
20	3	80	3
	80		

Qual a frequência acumulada abaixo e acima de 7?

2 - MEDIDAS DE POSIÇÃO

Introdução:

Inúmeras vezes, nas mais diversas áreas do conhecimento, são necessárias comparações entre conjuntos de dados. Essas comparações visam sintetizar a informação e as decisões a serem tomadas a respeito de determinado conjunto de dados. Essas comparações podem ser realizadas por intermédio das medidas de posição e medidas de dispersão.

As **medidas de posição**, também, conhecidas como **medidas de tendência central** são valores obtidos a partir dos dados, que fornecem uma orientação quanto à posição da distribuição em relação ao eixo dos valores reais (eixo x), ou seja, o termo medida de posição é usado para indicar, ao longo da escala de medidas, onde a amostra ou a população está locada. Portanto, as medidas de posição mostram o valor representativo em torno do qual os dados tendem a agrupar-se, com maior ou menor frequência, isto é, são utilizadas para sintetizar em um único número o conjunto de dados observados. Entre vários tipos de medidas de posição destacam-se a média, a mediana e a moda. Esses parâmetros são úteis, pois descrevem propriedades da população, ou seja, caracterizam a população. A média aritmética é a medida de posição mais conhecida e aplicada. No entanto, nem sempre é a mais adequada.

As medidas de posição são usadas para representar (sintetizar) um único número típico de uma distribuição de dados. Porém, as medidas de posição nos dão uma informação incompleta a respeito de um conjunto de dados. Podendo assim nos confundir a ponto de tomarmos decisões ou escolhas não muito adequadas, ou seja, a média é uma medida de centro da distribuição, porém, nada informa com relação à dispersão dos valores em torno do centro. Portanto, torna-se necessário agregarmos mais informações sobre determinado conjunto de dados por intermédio das **medidas de dispersão**.

Logo, podemos estabelecer algumas relações: quanto maior a variabilidade (dispersão) dos dados menor a representatividade da média; quanto menor a dispersão, mais confiável é a média. Assim, dizemos que as medidas de dispersão servem para qualificar a média (LEVIN & FOX, 2004). De forma geral, as medidas de dispersão mostram o grau de afastamento dos valores observados em relação àquele valor representativo (que nem sempre é a média).

MEDIDAS DE POSIÇÃO – Definição:

As medidas de posição mais importantes são as medidas de tendência central, entre elas a **média**, a **mediana** e a **moda**.

a) Média Aritmética É uma medida de fácil compreensão, mais comum e simples de ser calculada. A média aritmética ou simplesmente média é, por definição, o resultado da divisão das somas de todos os valores da série pelo número de valores na série.

A média é utilizada quando:

- Deseja-se obter a medida de posição que possui a maior estabilidade;
- É base para outros procedimentos estatísticos.

a.1) Média Aritmética para dados ***não*** agrupados

A média de uma população ou **média populacional** é representada pela letra grega minúscula μ, sendo definida como:

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N} \qquad 1)$$

Em que μ é a média **populacional** da variável; $\sum X_i$ é a soma de todos os elementos da população e N é o número de elementos na população.

O estimador não viesado, mais eficiente e consistente da média populacional é a média **amostral**, denotada por $\bar{x}$ (leia-se X barra):

$$\bar{x} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n} \qquad 2)$$

Em que $\bar{X}$ é a média amostral da variável; $\sum X_i$ é a soma de todos os elementos da amostra e n é o número de elementos da amostra.

Exemplo 1: Sabendo-se que o número de peças defeituosas observados em **amostras** retiradas diariamente da linha de produção, durante uma semana foi de 10, 14, 13, 15, 16, 18 e 12 peças, têm, para número médio de peças defeituosas da semana:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{\sum_{i=1}^{7} X_i}{7} = \frac{x_1 + x_2 + \cdots + x_7}{7} = \frac{10+14+13+15+16+18+12}{7} = \frac{98}{7} = 14 \text{ peças/dia.}$$

a.2) Média Aritmética para dados agrupados

Média Aritmética para dados agrupados em tabelas sem intervalos de classe (variáveis discretas)

O cálculo da média amostral quando os dados estão agrupados, ou seja, estão em uma distribuição de frequências e quando a variável em questão é classificada como discreta, segue o mesmo princípio da fórmula básica da média aritmética, no entanto, as informações utilizadas não são todos os elementos da distribuição, mas sim cada classe (X_i) com sua frequência (f_i). A fórmula passa a ser:

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i f_i}{n} \qquad 3)$$

2Em que $\bar{X}$ é a média amostral da variável; $\sum_{i=1}^{k} x_i f_i$: é a somatória das multiplicações dos valores de cada classe por sua frequência; *k* é o número de classes e n é o número total de elementos da amostra (dados por $\sum f_i$).

Exemplo 2: Considere os números de gols por partida em um determinado campeonato de futebol, agrupados e apresentados na Tabela 2.1. Calcule o número médio de gols por partida.

Tabela 2.1 – Número de gols por partida em um total de 60 jogos.

Nº. de gols por partida (X_i)	f_i
0	7
1	12
2	16
3	12
4	9
5	2
6	2
Σ	60

Observe que cada "classe" ou atributo ou categorias da variável (nº. de gols por partida) apresenta sua frequência. Para calcular a média quando os dados estão agrupados, o modo mais prático é acrescentar na tabela uma coluna correspondente aos produtos $x_i f_i$ (em cada linha da

tabela, procede-se a multiplicação do valor de X_i por sua frequência f_i), e após a obtenção da somatória desses produtos ($\sum x_i f_i$) divide-se pelo total de observações.

Para o exemplo 2, esse procedimento é apresentado na tabela abaixo.

Tabela 2.2 – Número de gols por partida em um total de 60 jogos, com a coluna $x_i f_i$.

nº. de gols por partida (X_i)	f_i	$X_i f_i$
0	7	0
1	12	12
2	16	32
3	12	36
4	9	36
5	2	10
6	2	12
Σ	60	138

Logo, o cálculo da média amostral será realizado por intermédio da equação (3):

$$\bar{x} = \frac{\sum_{i=1}^{7} X_i f_i}{60} = \frac{x_1 f_1 + x_2 f_2 + \cdots + x_7 f_7}{60} = \frac{0+12+\cdots+12}{60} = \frac{138}{60} = 2{,}3 \text{ gols por partida.}$$

Observe que:

- a somatória dos produtos dos números de gols por suas frequências $(\sum X_i f_i)$ corresponde ao número total de gols durante o campeonato. Ao dividirmos esse total pelo número de jogos $(\sum f_i)$ estamos nos remetendo ao mesmo procedimento do cálculo da média aritmética simples. O que mudou, portanto, foi apenas a apresentação dos dados, mas não o conceito da medida;
- O valor encontrado ($\bar{x}$= 2,3 gols por partida) não é um resultado possível para qualquer jogo (nesse caso poderiam ser 2 gols, 3 gols, mas não 2,3 gols). No entanto, esse valor representa o todo e permite interpretar que a tendência geral foi de pouco mais de dois gols por partida nesse campeonato.

Média Aritmética para dados agrupados com intervalos de classes (variáveis discretas ou continuas)

Para o cálculo da média amostral quando os dados estão agrupados e a variável envolvida no processo é contínua, utiliza-se o raciocínio análogo ao cálculo da variável discreta, conforme a expressão abaixo:

$$\bar{X} = \frac{\sum_{i=1}^{k} f_i \bar{X}_i}{n} = \frac{f_1\bar{X}_1 + f_2\bar{X}_2 + \cdots + f_k\bar{X}_k}{n} \qquad 4)$$

Em que $\bar{X}_i$ é o ponto médio da classe e f_i é a frequência absoluta da classe i, para $i = 1,2,\ldots,k$ e k é o número de classes.

Exemplo 3: Em uma fábrica de pneus automotivos a matéria prima para a fabricação consiste em materiais derivados do petróleo, materiais sintéticos e borracha. As características dos diversos tipos de pneus fabricados são determinadas pela qualidade do material empregado em sua fabricação, e, neste sentido diversos testes são aplicados a estes produtos para a medição e verificação de sua qualidade. Considere que um bloco de borracha que deve ser submetido a testes para a verificação do coeficiente de atrito entre o bloco e uma superfície plana de cimento/asfalto. Uma força é aplicada ao bloco e este é arrastado por uma determinada distância permitindo que o coeficiente de atrito seja medido. Em uma sessão de testes foram realizadas 40 medições e o coeficiente de atrito medido foi dividido em quatro classes cujos resultados estão mostrados na Tabela 2.3, que indica a frequência absoluta (f_i) do coeficiente de atrito medido.

Tabela 2. 3 – Distribuição de frequências do coeficiente de atrito medido.

Classes de Coeficiente de Atrito Cinético	f_i
0,15 ├ 0,35	5
0,35 ├ 0,55	10
0,55 ├ 0,75	8
0,75 ├ 0,95	17
Σ	40

Analogamente ao procedimento das variáveis discretas será criada uma coluna com os pontos médios das classes $(\bar{X}_i)$ e a seguir outra coluna correspondente aos produtos $\bar{X}_i f_i$, conforme é apresentado na Tabela 2.4.

Tabela 2.4 – Distribuição de frequências, acrescentando-se as colunas $\bar{X}_i$ e $\bar{X}_i f_i$.

Classes de Coeficiente de Atrito Cinético	f_i	$\bar{X}_i$	$\bar{X}_i f_i$
0,15 ├ 0,35	5	0,25	1,25
0,35 ├ 0,55	10	0,45	4,50
0,55 ├ 0,75	8	0,65	5,20
0,75 ├ 0,95	17	0,85	14,45

Σ	40	-	25,40

O coeficiente de atrito cinético médio, ou seja, a média será determinada por meio da equação 4:

$$\bar{X} = \frac{\sum_{i=1}^{4} f_i \bar{X}_i}{n} = \frac{f_1\bar{X}_1 + f_2\bar{X}_2 + f_3\bar{X}_3 + f_4\bar{X}_4}{40} = \frac{5 * 0{,}25 + 10 * 0{,}45 + 8 * 0{,}65 + 17 * 0{,}85}{40}$$

$$\bar{X} = \frac{25{,}40}{40} = 0{,}635.$$

Observe que:

- A fórmula é exatamente a mesma para variáveis discretas ou contínuas;
- Todos os elementos de um determinado intervalo de classe são representados, no cálculo, pelo ponto médio da classe e não pelos seus valores reais (Hipótese Tabular Básica). Assim, para variáveis contínuas, o cálculo da média com dados agrupados gera um valor aproximado, e não idêntico ao cálculo com todos os elementos (dados não-agrupados);

PROPRIEDADES DA MÉDIA

- **A soma algébrica dos desvios em relação à média é nula.**
- **A soma de quadrados dos desvios de um conjunto de dados, em relação a uma constante qualquer K, será mínima se e somente se $k = \bar{X}$.**
- **Somando-se (ou subtraindo-se) uma constante (c) a todos os valores de uma variável, a média do conjunto fica aumentada (ou diminuída) dessa constante.**
- **Multiplicando-se (ou dividindo-se) todos os valores de uma variável por uma constante (c), a média do conjunto fica multiplicada (ou dividida) por essa constante.**

b) Mediana (Md)

A mediana é uma medida típica de tendência central, sendo definida em um conjunto de dados ordenados como o valor central, ou seja, o valor para o qual há tantas mensurações que o superam quanto são superados por ele. A mediana amostral (Md) é o melhor estimador da mediana populacional (μ_d) (FERREIRA, 2005). Para a estimação da mediana, é necessário ordenar os dados

(dados elaborados). A ordenação pode ser crescente ou decrescente, embora, no presente material, sejam consideradas as ordens crescentes.

b.1) Mediana para dados não agrupados

Para determinar mediana amostral para dados não agrupados é necessário que determine a posição em que se encontra a mediana:

i) Se o número de observações for **par**, a posição da mediana denotada por **E** será:

$$E = \frac{n}{2} \qquad 5)$$

e a mediana amostral será determinada por:

$$Md = \frac{X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n+2}{2}\right)}}{2} \qquad 6)$$

Exemplo.4: Considere a seguinte amostra de dados: 8, 9, 9, **11**, **12**, 13, 13, 14 que possui 8 elementos, portanto $n = 8$. Logo, n é par, então por meio da equação (5) tem-se que: $E = \frac{n}{2} = \frac{8}{2} = 4$,

ou seja, o elemento central apresenta ordem 4. Assim, a mediana será determinada por intermédio da equação (6):

$$Md = \frac{X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n+2}{2}\right)}}{2} = \frac{X_{\left(\frac{8}{2}\right)} + X_{\left(\frac{8+2}{2}\right)}}{2} = \frac{X_{(4)} + X_{(5)}}{2} = \frac{11+12}{2} = 11{,}5.$$

ii) Se o número de observações for **ímpar**, a posição da mediana denotada por **E** será:

$$E = \frac{n+1}{2} \qquad 7)$$

e a mediana amostra será determinada por:

$$Md = X_{\left(\frac{n+1}{2}\right)} \qquad 8)$$

Exemplo 5: Considere a seguinte amostra dados: 8, 9, 9, **11**, 12, 13, 13 que possui 7 elementos, isto é, $n = 7$. Logo, n é ímpar, então por meio da equação (7) tem-se que: $E = \frac{n+1}{2} = \frac{7+1}{2} = 4$, ou seja, o

elemento central apresenta ordem 4. Assim, a mediana será determinada por intermédio da equação (8):

$$Md = X_{\left(\frac{n+1}{2}\right)} = X_{\left(\frac{7+1}{2}\right)} = X_{(4)} = 11,$$

ou seja, o 4º elemento da amostra, que corresponde ao valor **11**, é a mediana do conjunto de dados.

b.2) Mediana para dados agrupados

Para dados agrupados, o cálculo da mediana segue o mesmo princípio usado para dados não-agrupados, ou seja, em um conjunto de valores dispostos de forma ordenada, a mediana é o valor que separa o conjunto em dois subconjuntos com mesmo número de elementos. Para se fazer essa determinação necessita-se de determinar as frequências acumuladas (ordenação dos dados).

Mediana para dados agrupados sem intervalos de classe (variável discreta)

Se a variável é discreta, o procedimento para determinar a mediana é o mesmo utilizado para dados não agrupados, em que o centro da amostra é diferente para os casos em que n é ímpar, ou n é par, isto é:

i) Determina-se a ordem do valor central com o uso das mesmas regras dos dados não agrupados:

ii) Determina-se a coluna de frequência acumulada (F_i) à distribuição com o objetivo de encontrar o valor central.

iii) Se n é ímpar, o valor encontrado no 2º passo já é a mediana. Se n é par, a média dos elementos encontrados no 2º passo é a mediana.

Exemplo 6: (n par): Utilizando os dados do exemplo 3.2 apresentados na Tabela 3.1, que contabilizou os números de gols por partida em um campeonato de futebol, vamos calcular a mediana desses valores.

O número de gols no campeonato foi 60, isto é, n é par. Então, por meio da equação (5) tem-se que a ordem do elemento central é: $E = \frac{n}{2} = \frac{60}{2} = 30$ (regra **i**).

A Tabela 3.1 foi reescrita, acrescendo-se a coluna de frequência acumulada para baixo (F_i) para formar a Tabela 3.5 (regra **ii**).

Tabela 2.5 – Nº. de gols por partida em um total de 60 jogos (f_i e F_i)

nº. de gols por partida (X_i)	f_i	F_i
0	7	7
1	12	19
2	16	35
3	12	47
4	9	56
5	2	58
6	2	60
Σ	60	-

Portanto, o elemento central é o **30º** elemento da amostra, ou seja, a "classe" (categoria ou atributo) cuja frequência acumulada é igual, ou imediatamente superior ao 30º elemento é a terceira "classe" ($F_3 = 35$). Logo, a mediana ou o número mediano de gols por partida será calculado por intermédio da equação (6) (regra **iii**):

$$Md = \frac{X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n+2}{2}\right)}}{2} = \frac{X_{\left(\frac{60}{2}\right)} + X_{\left(\frac{60+2}{2}\right)}}{2} = \frac{X_{(30)} + X_{(31)}}{2} = \frac{2+2}{2} = \frac{4}{2} = 2.$$

Exemplo 7: (n ímpar): Considere os dados amostrais de números de circuitos defeituosos em sistema composto por 4 circuitos. Uma amostra de 19 sistemas esta resumida na Tabela 2.6. Vamos determinar a mediana, ou seja, o número mediano de circuitos defeituosos por sistema.
A Tabela 2.6 apresenta uma coluna referente às frequências acumulada para baixo (F_i) (regra **ii**).

Tabela 2.6 – Distribuição dos números de circuitos defeituosos por sistema (f_i e F_i).

nº. de circuitos defeituosos (X_i)	f_i	F_i
1	**10**	**10**
2	7	17
3	1	18
4	1	19
Σ	19	-

Observe que o número de elementos (sistemas) é 19, isto é, n é ímpar. Então, por meio da equação (7) tem-se que a ordem do elemento central é:

$$E = \frac{n+1}{2} = \frac{19+1}{2} = 10 \text{ (regra } \mathbf{i}\text{)}.$$

Portanto, o elemento central é o **10º** elemento, ou seja, a "classe" cuja frequência acumulada é igual, ou imediatamente superior ao 10º elemento é a primeira "classe" ($F_1 = 10$). Logo, a mediana ou o

número mediano de circuitos defeituosos por sistema será determinado por meio da equação (8) (regra **iii**):

$$Md = X_{\left(\frac{n+1}{2}\right)} = X_{\left(\frac{19+1}{2}\right)} = X_{\left(\frac{20}{2}\right)} = X_{(10)} = 1 \text{ circuito defeituoso por sistema.}$$

Mediana para dados agrupados em tabelas com intervalos de classe para variável contínua

Se a variável é contínua é necessária uma interpolação dentro da classe que contém o centro da amostra para determinar o valor "exato" da mediana. O procedimento para determinar a mediana é:

i) Determinam-se as frequências acumuladas;

ii) Calcula-se a ordem por meio da equação (5) se n for par ou pela equação (7) se n for ímpar;

iii) Marca-se a classe correspondente à frequência acumulada imediatamente superior à ordem, que é a **classe mediana**, e aplica-se a fórmula de interpolação abaixo:

$$Md = LI_{Md} + \frac{\left(\frac{n}{2} - F_{acA}\right)}{f_{i_{Md}}} \times c_{Md}$$

Em que LI_{Md} é o limite inferior da classe mediana;

n é o número de elementos no conjunto de dados;

F_{acA} é a frequência acumulada da classe anterior à classe mediana;

cé a amplitude do intervalo da classe mediana;

$f_{i_{Md}}$ é a frequência absoluta da classe mediana;

Exemplo 8: Para ilustrar o exemplo 8 serão utilizados os dados do exemplo 3, que representa uma sessão de testes, ou seja, 40 medições referentes ao coeficiente de atrito. Na Tabela 3.7 é apresentado as frequências acumuladas das classes. Vamos calcular a mediana desses Coeficientes de Atrito Cinético.

Tabela 2.7 – Distribuição de frequências de 40 medições referente ao coeficiente de atrito.

Classes de Coeficiente de Atrito Cinético	f_i	F_{ac}
0,15 ├ 0,35	5	5
0,35 ├ 0,55	10	15
0,55 ├ 0,75	**8**	**23**
0,75 ├ 0,95	17	40
Σ	40	-

São 40 medições, ou seja, $n = 40$. Portanto a ordem é calculada por meio da equação (5):

$$E = \frac{n}{2} = \frac{40}{2} = 20.$$

A classe cuja frequência acumulada é imediatamente superior à ordem 20 é a terceira classe, portanto essa é a classe mediana **(0,55 ├ 0,75),** destacada na Tabela 3.7. Então, por intermédio da interpolação, equação (9), tem-se a mediana:

$$Md = LI_{Md} + \frac{\left(\frac{n}{2} - F_{acA}\right) \times c}{f_{i_{Md}}} = 0{,}55 + \frac{\left(\frac{40}{2} - 15\right) \times 0{,}20}{8} = 0{,}55 + \frac{(20 - 15) \times 0{,}20}{8}$$

$$Md = 0{,}55 + \frac{(5) \times 0{,}20}{8} = 0{,}55 + \frac{1}{8} = 0{,}55 + 0{,}125 = 0{,}675.$$

c) Moda (Mo)

A moda é o valor que ocorre com maior frequência em uma série de dados. Uma melhor definição poderia ser dada por aquele valor da variável em que há a mais densa concentração de valores na sua proximidade (FERREIRA, 2005). A moda amostral $(\boldsymbol{Mo})$ é o melhor estimador da moda populacional (μ_o). A moda não é afetada pelos extremos e também é uma medida muito utilizada na economia e quando:

- Desejamos obter uma medida rápida e aproximada de posição;
- A medida de posição deve ser o valor mais típico da distribuição.

c.1) Moda para dados não agrupados

Para determinar a moda em determinado conjunto de dados, procura-se o valor que mais se repete nesse conjunto de dados.

Exemplo 9: Considere a seguinte amostra: 8, 9, 9, 11, 13, 13, 13, 14. O valor que mais se repete é o 13, que aparece três vezes, portanto a moda é: $Mo = 13$.

c.2) Moda para dados agrupados

Moda para dados agrupados em tabelas sem intervalos de classe (variáveis discretas)

No caso de variáveis discretas, com os dados agrupados, torna-se muito simples a determinação da moda. Basta observar o valor (X_i) que apresenta maior frequência (f_i).

Exemplo 10: Para ilustrar o exemplo 10 serão considerados os dados do exemplo 7, que se refere ao número de circuitos defeituosos por sistema, observados em uma amostra de 19 sistemas.

Tabela 2.8 – Distribuição dos números de circuitos defeituosos por sistema.

nº. de circuitos defeituosos (X_i)	f_i
1	10
2	7
3	1
4	1
Σ	19

Observa-se que a maior frequência $(f_1 = 10)$ foi a da primeira "classe", cujo valor é 1 circuito defeituoso por sistema $(X_i = 1)$, por isso a moda da distribuição é: $Mo = 1$ circuito defeituoso/sistema.

Moda para variáveis contínuas agrupadas com intervalos de classes.

No caso de variáveis contínuas, a classe que apresenta maior frequência é denominada **classe modal**. Crespo (1999) afirma que a moda, nesse caso, é o valor dominante que está compreendido entre os limites da classe modal.

Depois que a classe modal está definida é necessário fazer a interpolação para determinação do valor da moda. Para esse fim existem diferentes métodos, sendo que nesse texto vamos aplicar o

método de Czuber (citado por FERREIRA, 2005) que permite encontrar o valor da moda de forma mais elaborada:

$$Mo = LI_{Mo} + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times c_{Mo} \qquad 9)$$

Em que:

LI_{Mo} é o limite inferior da classe modal;

Δ_1 é a diferença entre as frequências da classe modal e a imediatamente anterior;

Δ_2é a diferença entre as frequências da classe modal e a imediatamente posterior;

h_{Mo} é a amplitude da classe modal.

Exemplo 11: Os dados da Tabela 2.9 se referem às 40 medições do coeficiente de atrito. Vamos calcular a moda desses coeficientes de atrito cinético.

Tabela 2.9 – Distribuição de frequências do coeficiente de atrito medido.

Classes de Coeficiente de Atrito Cinético	f_i
0,15 ├ 0,35	5
0,35 ├ 0,55	10
0,55 ├ 0,75	8
0,75 ├ 0,95	**17**
Σ	40

A classe que apresentou maior frequência (f_i) foi a segunda classe (**0,75 ├ 0,95**), que apresentou dez elementos $(f_4 = 17)$. Esta é, então, a **classe modal**. Agora, será determinada a moda ou o coeficiente de atrito cinético modal por intermédio da equação (10), método de Czuber:

$$Mo = LI_{Mo} + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times c_{Mo} = 0{,}75 + \frac{(17-8)}{(17-8)+(17-0)} \times 0{,}20$$

$$Mo = 0{,}75 + \frac{(9)}{(9)+(17)} \times 0{,}20 = 0{,}75 + \frac{9}{26} \times 0{,}20 = 0{,}75 + \frac{1{,}8}{26}$$

$$Mo = 0{,}75 + 0{,}0692 = 0{,}8192$$

Observação: É possível encontrar séries de dados nas quais nenhum valor apareça mais do que os

outros, como por exemplo, a série: 8, 9, 10, 11, 13, 14 então, esta série é dita **amodal.** Em outros casos pode haver dois ou mais valores de concentração, como por exemplo, a série: 8, 9, 9, 11, 12, 13, 13, 14 então, os valores **9** e **13** ocorrem com maior frequência que os demais. Esta série apresenta duas modas, sendo dita **bimodal**.

Posição relativa da média, mediana e moda

Crespo (1999) cita que quando uma distribuição é simétrica, as três medidas coincidem. Porém, a assimetria as torna diferentes de modo que quanto maior a assimetria maior será essa diferença entre as três medidas. Assim, em uma distribuição em forma de sino, temos:

a) $\bar{X} = Md = Mo$, no caso de curva **simétrica**;

b) $\bar{X} > Md > Mo$, no caso de curva assimétrica positiva (**assimétrica à direita**);

c) $\bar{X} < Md < Mo$, no caso de curva assimétrica negativa (**assimétrica à esquerda**);

Assimetria: significa desvio ou afastamento da simetria (grau de deformação de uma curva), ou seja, existem valores elevados em uma das caudas.

Simétrica, se a média e a moda coincidem.

Assimétrica à esquerda ou negativa, se a média é menor que a moda.

Assimétrica à direita ou positiva, se a média é maior que a moda.

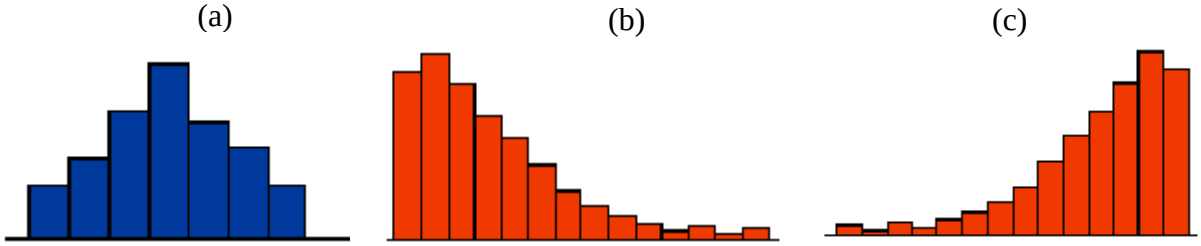


Figura 4.1 - Formas de distribuições em situações reais:

(a) distribuição em forma de sino simétrica;
(b) distribuição assimétrica à direita;
(c) distribuição assimétrica à esquerda.

Comparação entre média e a mediana

Suponha que se queira sintetizar em um único número os salários das pessoas que trabalham em determinado restaurante (cozinheiros, copeiros, garçons, recepcionistas etc.). Em uma situação hipotética, considerem os seguintes valores de salários: 200, 250, 250, 300, 450, 460, 510.

Sua média aritmética, isto é, o salário médio é: $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{200 + \cdots + 510}{7} = 345{,}7$.

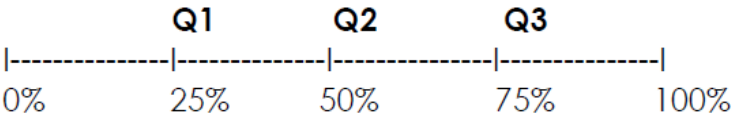
Esse valor representa, ou sintetiza razoavelmente, aquele conjunto de observações. Se incluirmos, entretanto, o salário de gerente do estabelecimento, os dados seriam: 200, 250, 250, 300, 450, 460, 510, 2300 e a média seria 601,4. Neste caso, não se pode dizer que a média sintetiza adequadamente o conjunto, pois apenas um valor é maior do que ela.

A mediana é a mesma, 300, em ambos os casos. O exemplo ilustra um fato de que a média é muito sensível a valores extremos de um conjunto de observações, enquanto, a mediana não sofre muito com a presença de alguns valores muito altos ou muito baixos. Costuma-se dizer que a mediana é mais robusta do que a média aritmética. Portanto, deve-se preferir a mediana como medida sintetizadora quando o histograma do conjunto de valores é assimétrico, isto é, quando há predominância de valores elevados em uma das caudas.

Quartis, decis e percentis.

a) ***Quartil:*** indicado por Q_r, e a separatriz que divide as observações (x) em quatro partes iguais. Logo: $q = 4e1 \leq r \leq 3$. O segundo quartil coincide com a Mediana ($Q_2 = M_d$). Em termos percentuais pode se dizer que 25% dos valores estão abaixo dos valores do primeiro quartil, 25% entre o primeiro e o segundo quartil, 25% entre o segundo e o terceiro quartis e 25% são os maiores que o terceiro quartil;

Divide a amostra em quatro partes iguais.



Para determinar Q_1:

1° Passo: Calcula-se $\frac{n}{4}$

2° Passo: Identifica-se a classe Q_1 pela F_{ac}

3° Passo: Aplica-se a fórmula:

$$Q_i = LI_i + \frac{\left(\frac{n}{4} - f_{acA}\right)}{f_{iclasse}}.c$$

Para determinar Q_2: igual a mediana

Para determinar Q_3:

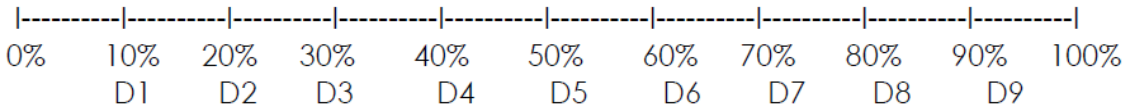
1° Passo: Calcula-se $\frac{3n}{4}$

2° Passo: Identifica-se a classe Q_3 pela F_{ac}

3° Passo: Aplica-se a mesma fórmula anterior, apenas substituindo $\frac{n}{4}$ por $\frac{3n}{4}$.

b) ***Decil:*** indicado por D_r, e a separatriz que divide os dados em dez partes, em décimos. Logo:

$q = 10e1 \leq r \leq 0$



Para determinar D_i:

1° Passo: Calcula-se $\frac{in}{10}$

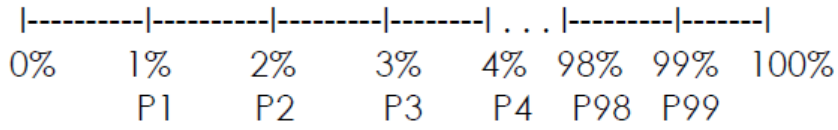
2° Passo: Identifica-se a classe D_i pela F_{ac}

3° Passo: Aplica-se a fórmula:

$$D_i = LI_i + \frac{\left(\frac{in}{10} - f_{acA}\right)}{f_{iclasse}}.c$$

c) ***Percentil:*** apontado por P_r, divide os dados em 100 partes, em centésimos. Logo,

$q = 100e1 \leq r \leq 99.$



Para determinar P_1:

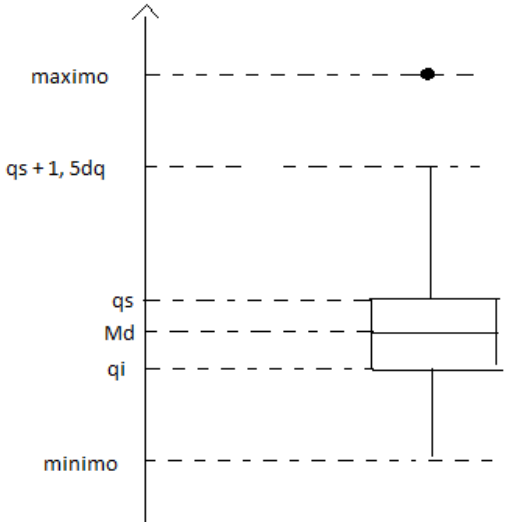
1° Passo: Calcula-se $\frac{in}{100}$

2° Passo: Aplica-se a fórmula:

$$P_i = LI_i + \frac{\left(\frac{in}{100} - f_{acA}\right)}{f_{iclasse}}.c$$

DIAGRAMA EM CAIXAS – (BOX PLOT)

O diagrama de caixa é uma apresentação gráfica que descreve simultaneamente varias características importantes de um conjunto de dados, tais como centro, dispersão, desvio de simetria e identificação de observações que estão, não geralmente, longe do seio dos dados *(outliers)*. Este apresenta três quartis, em uma caixa retangular, alinhado tanto horizontal como verticalmente. O retângulo e divido no valor correspondente a mediana, assim ele indica o quartil inferior, a mediana e o quartil superior.



Outras medias

a) **Media ponderada**: e uma media dos valores afetados pelos pesos diferentes, em que ela e calculada atribuindo pesos diferentes aos diversos valores.

$$\bar{x} = \frac{\sum(w_i x_i)}{\sum w_i} \qquad \textbf{10)}$$

Em que w_i e o peso associado a cada valor observado.

Exemplo 1: Considere as 5 notas de um teste 85, 90, 75, 80, 95, com os quatro primeiros testes valendo 15% cada um, e o ultimo valendo 40%.

$$\bar{x} = \frac{\sum(w_i x_i)}{\sum w_i} = \frac{(15.85) + (15.90) + (15.75) + (15.80) + (45.95)}{15 + 15 + 15 + 15 + 40} = \frac{8750}{100} = 87{,}5$$

b) **Media harmônica:** costuma ser usada como medida de tendência central para conjuntos de dados que consistem em taxas de variação, como por exemplo, velocidades.

$$\bar{x}_h = \frac{n}{\sum \frac{1}{x}} \qquad 11)$$

Obs.: nenhum valor pode ser zero.

Exemplo 2: Obtenha a media harmônica para: 2, 4 e 10.

$$\bar{x}_h = \frac{n}{\sum \frac{1}{x}} = \frac{3}{\frac{1}{2} + \frac{1}{4} + \frac{1}{10}} = \frac{3}{0{,}85} = 3{,}5$$

Exercício 1: A velocidade media, em mi/h, do percurso de ida e volta entre duas cidades e dado

abaixo. Qual e a velocidade media? (1mi = 1 609 Km)

42,6 41,3 38,2 42,9 43,4 43,7 40,8 34,2 40,1 41,2 40,5 41,7 39,8 39,6

c) Media geométrica: e usada em administração e economia para achar taxas medias de variação, de crescimento, ou razoes medias. Dados n valores (todos positivos), a media geométrica e a raiz n^{ma} do seu produto.

$$\bar{x}_g = \sqrt[n]{\prod n_i} \qquad 12)$$

***Exemplo 3**:* Obter a media geométrica de: 2, 4 10.

$$\bar{x}_g = \sqrt[n]{\prod n_i} = \sqrt[3]{2.4.10} = \sqrt[3]{80} = 4{,}3$$

Exercício 2: O fator de crescimento médio para o dinheiro, composto as tzas anuais de juro de 10%, 8%, 9%, 12% e 7% pode ser determinado calculando-se a media geométrica de 1,10 1,08 1,09 1,12 e 1,17. Calcule o fator médio de crescimento.

d) Media quadrática: e utilizada em geral em experimentos físicos, por exemplo, em sistemas de distribuição de energia em que as tensões e correntes são em geral dadas em termos de sua media quadrática. Obtém-se a media quadrática de um conjunto de valores elevando-se cada um ao quadrado, somando-se os resultados, dividindo-se o total pelo numero n de valores e tomando-se a raiz quadrada do resultado. Por exemplo, a média quadrática de 2, 4 e 10 e:

$$\sqrt{\frac{\sum x^2}{n}} = \sqrt{\frac{2+4+10}{n}} = \sqrt{40} = 6{,}3$$

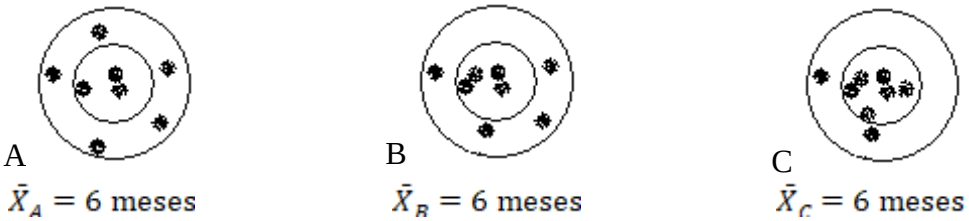
3) - MEDIDAS DE DISPERSÃO

As medidas de dispersão têm como objetivo apresentar um estudo descritivo de um conjunto de dados, isto é, determinar a variabilidade ou dispersão de um conjunto de dados em relação à medida de localização ou posição do centro da amostra.

As diferenças individuais em uma amostra ou população definem o que os estatísticos chamam de variabilidade ou dispersão do conjunto de mensurações, sendo que a variabilidade entre os elementos é vista pela perspectiva da dispersão em torno do centro da distribuição. As medidas de posição nem sempre é suficiente para sintetizar a informações contidas nos dados, ou seja, não são suficientes para caracterizarem completamente a distribuição dos dados. Portanto, são necessárias outras medidas para isso, e as medidas de dispersão pertencem a um conjunto de medidas que se

aplicam na caracterização de uma distribuição de mensurações (FERREIRA, 2005).

Vamos considerar um exemplo (diagrama abaixo) para discutir um pouco mais sobre a "deficiência" das medidas de posição. Suponha que queremos comparar o tempo de vida de 3 marcas (A, B e C) de lâmpadas em meses.



$\bar{X}_A = 6$ meses $\bar{X}_B = 6$ meses $\bar{X}_C = 6$ meses

As três marcas de lâmpadas apresentaram a mesma média (6 meses) para a variável tempo de vida. É notório que os conjuntos diferem razoavelmente um do outro. A lâmpada C apresentou uma menor dispersão de valores em torno do valor central (6 meses), sendo seguido pela lâmpada B e por último a lâmpada A. Se os conjuntos fossem representados apenas pelas respectivas médias eles seriam considerados iguais. Porém, analisando o diagrama acima vemos que a lâmpada C apresenta menor variabilidade consequentemente seria a melhor escolha.

1) Amplitude

A amplitude denotada por A, é a diferença entre o maior e o menor escore em uma distribuição, isto é, corresponde a diferença entre a maior (máximo) e a menor observação (mínimo) de um conjunto de dados. Essa medida é inconveniente (grosseira), apesar de ser facilmente calculada, pois não considera todas as observações, ou seja, leva em conta apenas os valores extremos: máximo e mínimo (LEVIN & FOX, 2004). Consequentemente, a amplitude é facilmente influenciada.

O estimador da amplitude para dados que não estão agrupados em classe é:

$$A = X_{(n)} - X_{(1)} = \text{maior valor - menor valor.} \qquad 1)$$

O estimador da amplitude para dados agrupados em classe é:

$$A = \bar{X}_k - \bar{X}_1 = \text{maior valor - menor valor.} \qquad 2)$$

Portanto, a amplitude para dados agrupados e para dados não-agrupados será:

$$A = \text{maior valor - menor valor.}$$

Exemplo 1: Uma **amostra** do tempo de vida de pneus de determinada marca apresentou os seguintes resultados: {40.000; 40.500; 35.600; 39.300; 37.200; 39.700; 35.000; 32.300 km}. Logo, o tempo de vida do pneu dessa marca varia de 32.300 a 40.500 km, ou seja, o tempo de vida apresenta uma amplitude de 8.200 km. Pois, por intermédio da equação (1) tem-se que

$A = 40.500 - 32.300 = 8.200$ km.

Exemplo 2: Os dados da Tabela abaixo representa uma sessão de testes, ou seja, 40 medições referentes ao coeficiente de atrito cinético de pneus automotivos. Na Tabela 3 é apresentado as frequências absolutas e os pontos médios de cada classe.

Tabela 3.1 – Distribuição de frequências referente a 40 medições do coeficiente de atrito cinético de pneus automotivos.

Classes de Coeficiente de Atrito Cinético	f_i	$\bar{X}_i$
0,15 ├ 0,35	5	0,25
0,35 ├ 0,55	10	0,45
0,55 ├ 0,75	8	0,65
0,75 ├ 0,95	17	0,85
Σ	40	-

Os dados na Tabela 3 estão agrupados em 4 (quatro) classes. Todos os pontos de uma classe podem ser representados por um único valor conhecido como ponto médio da classe. Observe que a primeira classe (0,15 ├ 0,35) é representada pelo valor 0,25, ou seja, esta classe que possui 5 pneus com coeficiente de atrito cinético entre 0,15 e 0,35 será representada pelo ponto médio $\bar{X}_1 = 0{,}25$. O ponto médio da classe é calculado pela média dos limites da classe. Esse critério é conhecido como hipótese tabular básica.

De acordo com a definição de amplitude é necessário, determinar o maior e menor valor dos coeficientes de atrito, tendo em vista que os coeficientes de atrito estão agrupados em classe e que cada classe será representada pelo seu respectivo ponto médio. Então, o menor e o maior valor correspondem ao ponto médio da primeira e da última classe respectivamente, ou seja, 0,25 e 0,85. Então, a amplitude será: $A = 0{,}85 - 0{,}25 = 0{,}60$, isto é, o coeficiente de atrito cinético varia entre 0,25 e 0,85.

3.2) Variância

A variância é uma boa medida, pois se baseia em todos os valores observados (dados) e é facilmente calculada e de fácil compreensão.

A variância **populacional** denotada por σ^2 é definida como sendo Soma de Quadrado dos Desvios (SQD) em relação à média dividida pelo tamanho da população (N). A variância pode ser considerada como um valor médio dos desvios ao quadrado, portanto, sendo conhecida, também, por quadrado médio (FERREIRA, 2005).

O estimador da variância **populacional** é:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N} = \frac{1}{N}\left[\sum x^2 - \frac{(\sum x)^2}{N}\right] \qquad 3)$$

A variância **amostral** denotada por s^2 poderia ser definida de forma análoga à variância populacional, ou seja, substituindo-se N por n e μ por $\bar{X}$. No entanto, isso não ocorre, devido a uma propriedade importante do estimador denominada de viés (tendenciosidade). Nesse caso, a soma de quadrado dos desvios é dividida por $n - 1$ ao invés de usar o n (FERREIRA, 2005).

A variância **amostral** é definida da seguinte forma:

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n - 1} \qquad 4)$$

em que, $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$.

Exemplo 3: Para a ilustração do cálculo da variância serão considerados os dados referentes ao tempo de vida de uma marca de pneu: {40.000; 40.500; 35.600; 39.300; 37.200; 39.700; 35.000; 32.300 km}. Primeiramente é preciso calcular o tempo de vida médio, isto é, a média de duração desse pneu, para posteriormente obtermos a variância por meio da fórmula ou equação 4:
O tempo médio de vida de uma marca de pneu é:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{40.000 + 40.500 + \cdots + 32.300}{8} = 37.450 \text{ km.}$$

Agora, temos condições de realizar o cálculo da variância:

$$s^2 = \frac{\sum_{i=1}^{8}\left(X_i - \bar{X}\right)^2}{n-1} = \frac{(40.000 - 37.450)^2 + (40.500 - 37.450)^2 + \cdots + (32.300 - 37.450)^2}{8-1} = \frac{60.300.000}{7}$$

$$s^2 = \frac{1}{7}[60.300.000] = 8.614.285{,}714 \ \ (\text{km})^2.$$

Nota-se que a unidade da variância corresponde à unidade de mensuração ao quadrado, isto é, o tempo de vida médio foi medido em km e sua variância foi expressa em $(\text{km})^2$.

Fórmula simplificada para cálculo da Variância

As fórmulas simplificadas para variâncias foram desenvolvidas com o objetivo de facilitar o cálculo e contornar problemas de arredondamento (precisão).

A fórmula simplificada para a variância **amostral** é (FERREIRA, 2005):

$$s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 - \frac{(\sum_{i=1}^{n} X_i)^2}{n}\right] \qquad 5)$$

Exemplo 4: Neste exemplo utilizaremos os dados do Exemplo 3 para calcular a variância por

intermédio da fórmula simplificada com o objetivo de mostrar que o resultado da variância será o mesmo obtido no Exemplo 3.

A **amostra** referente ao tempo de vida de uma marca de pneu é: {40.000; 40.500; 35.600; 39.300; 37.200; 39.700; 35.000; 32.300 km}

Utilizando a fórmula simplificada da variância **amostral**, tem-se:

$$s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 - \frac{(\sum_{i=1}^{n} X_i)^2}{n}\right]$$

$$= \frac{1}{8-1}\left[(40.000^2 + 40.500^2 + \cdots + 32.300^2) - \frac{(40.000 + 40.500 + \cdots + 32.300)^2}{8}\right] =$$

$$= \frac{1}{7}\left[1{,}128032 * 10^{10} - \frac{(299.600)^2}{8}\right] = \frac{1}{7}[1{,}128032 * 10^{10} - 1{,}122002 * 10^{10}]$$

$$= \frac{1}{7}[60.300.000] = 8.614.285{,}714(\text{km})^2$$

3.2.1) Variância amostral para dados agrupados

De acordo com Ferreira (2005), o estimador da variância para dados agrupados em classe é dado por:

$$s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{k} \bar{X}_i^2 f_i - \frac{(\sum_{i=1}^{k} \bar{X}_i f_i)^2}{n}\right] \qquad 6)$$

em que $\bar{X}_i$ é o ponto médio da classe *i* e f_i é a frequência absoluta da classe *i*.

Exemplo 5: Uma **amostra** de 40 medições do coeficiente de atrito cinético de pneus automotivos conforme a Tabela abaixo. Obter a variância amostral do coeficiente de atrito cinético dos 40 pneus testados:

Tabela 3.2: coeficiente de atrito cinético de pneus automotivos

Classes de Coeficiente de Atrito Cinético	f_i	$\bar{X}_i$
0,15 ├ 0,35	5	0,25
0,35 ├ 0,55	10	0,45
0,55 ├ 0,75	8	0,65
0,75 ├ 0,95	17	0,85
Σ	40	-

Utilizando a fórmula da variância **amostral para dados agrupados,** a equação (16) tem-se:

$$s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{k} \bar{X}_i^2 f_i - \frac{(\sum_{i=1}^{k} \bar{X}_i f_i)^2}{n}\right]$$

$$= \frac{1}{40-1}\left[(0{,}25^2 * 5 + \cdots + 0{,}85^2 * 17) - \frac{(0{,}25 * 5 + \cdots + 0{,}85 * 17)^2}{40}\right]$$

$$= \frac{1}{39}\left[18 - \frac{(25{,}4)^2}{40}\right]^2 = \frac{1}{39}[18 - 16{,}129] = 0{,}0480$$

O mesmo estimador pode ser usado substituindo $\bar{X}_i$, ponto médio da classe *i*, por X_i, valor da categoria ou atributo *i*, quando os dados são quantitativos discretos, isto é:

$$s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{k} X_i^2 f_i - \frac{\left(\sum_{i=1}^{k} X_i f_i\right)^2}{n}\right] \qquad 7)$$

Exemplo 6 (FERREIRA, 2005): Na Tabela 3.3, estão apresentados os dados referentes ao número de ovos danificados da inspeção feita em uma **amostra** de 30 embalagens de uma dúzia cada, de um carregamento para o mercado municipal de Lavras. Determine a variância.

Tabela 3.3 - Número de ovos danificados em uma inspeção feita em 30 embalagens, de uma dúzia cada, em um carregamento para o mercado municipal de Lavras proveniente de uma cidade distante.

Número de ovos quebrados (X_i)	f_i
0	13
1	9
2	3
3	3
4	1
5	1
Σ	30

Para calcular a variância será utilizada a equação acima:

$$s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{k} X_i^2 f_i - \frac{\left(\sum_{i=1}^{k} X_i f_i\right)^2}{n}\right] = \frac{1}{30-1}\left[\left(0^2 * 13 + 1^2 * 9 + \cdots + 5^2 * 1\right) - \frac{(0*13 + 1*9 + \cdots + 5*1)^2}{30}\right]$$

$$s^2 = \frac{1}{29}\left[89 - \frac{(33)^2}{30}\right] = \frac{1}{29}[89 - 36{,}3] = 1{,}8172 \text{ (ovos danificados)}^2.$$

3.3 - Desvio Padrão

O desvio padrão é a raiz quadrada da variância. Dessa forma o desvio padrão é expresso na mesma unidade dos dados (FERREIRA, 2005).

Desvio Padrão **Populacional**:

$$\sigma = \sqrt{\frac{1}{N}\left[\sum_{i=1}^{N} X_i^2 - \frac{(\sum_{i=1}^{N} X_i)^2}{N}\right]} \qquad 8)$$

Desvio Padrão **Amostral**:

$$s = \sqrt{\frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 - \frac{(\sum_{i=1}^{n} X_i)^2}{n}\right]} \qquad 9)$$

Para dados agrupados em classe o estimador do desvio padrão é:

$$s = \sqrt{\frac{1}{n-1}\left[\sum_{i=1}^{k} f_i \bar{X}_i^2 - \frac{(\sum_{i=1}^{k} f_i \bar{X}_i)^2}{n}\right]} \qquad 10)$$

O estimador acima pode ser usado substituindo $\bar{X}_i$, ponto médio da classe *i*, por X_i, valor da categoria ou atributo *i*, quando os dados são quantitativos discretos, isto é:

$$s = \sqrt{\frac{1}{n-1}\left[\sum_{i=1}^{k} X_i^2 f_i - \frac{(\sum_{i=1}^{k} X_i f_i)^2}{n}\right]} \qquad 11)$$

A **variância** e o **desvio padrão** são medidas que **só podem assumir valores não negativos** (positivo e igual a zero) e quanto maior for, maior será a dispersão dos dados, ou seja, maior será a variabilidade dos dados. Em outras palavras o desvio padrão e a variância medem a dispersão dos dados em torno da média.

Exemplo 7: Para apresentar o cálculo do desvio padrão utilizou-se os dados dos Exemplos 3.16 e 3.17, com o objetivo de enfatizar a relação entre desvio padrão e variância. Sabe-se por definição, que desvio padrão é a raiz quadrada da variância, e como já foram calculadas anteriormente nos exemplos 3.16 e 3.17, tem-se que o desvio padrão dos coeficientes de atrito cinético do pneu automotivo e o desvio padrão de ovos danificados são respectivamente:

$S = \sqrt{S^2} = \sqrt{0{,}0480} = 0{,}2190$ e $S = \sqrt{S^2} = \sqrt{1{,}8172} = 1{,}3480$ ovos danificados.

3.4 - Coeficiente de Variação

O desvio padrão e a variância são medidas da variabilidade absoluta dos dados. Essas medidas são dependentes da grandeza, escala ou unidade de medida empregada para mensurar os

dados. Conjuntos de dados com diferentes unidades de medidas não podem ter suas dispersões comparadas pela variância ou pelo desvio padrão. Mesmo para uma única unidade, se os conjuntos possuem médias de diferentes magnitudes, suas variabilidades não podem ser comparadas por essas medidas de dispersão apresentadas anteriormente. Para esta situação utiliza-se o coeficiente de variação (CV), pois ele não depende da grandeza, da escala ou unidade de medida empregada para mensurar os dados, ou seja, não possui unidade de medida (medida adimensional). Portanto, fica evidente que se deve usar o CV quando se tem diferentes unidades de medida e/ou médias de diferentes magnitudes (FERREIRA, 2005).

O coeficiente de variação **populacional** é:

$$CV = \frac{\sigma}{\mu} 100\%. \qquad 12)$$

O coeficiente de variação **amostral** é:

$$CV = \frac{S}{\bar{X}} 100\% \qquad 13)$$

Exemplo 8: A média e o desvio padrão do tempo de vida das lâmpadas de marca A e B são respectivamente: $\bar{X}_A = 4{,}0 \text{ meses}, S_A = 0{,}8 \text{ meses}$, $\bar{X}_B = 8{,}0 \text{ meses}$ e $S_B = 1{,}2 \text{ meses}$. Qual das lâmpadas possui maior uniformidade de tempo de vida?

Se, ao inspecionar as estatísticas, apresentadas você fosse induzido a responder que a lâmpada (A) seria a que possui maior uniformidade e que a razão seria o menor desvio padrão apresentado por ela (0,8 meses), você teria cometido um erro. O fundamento usado aqui para comparar a variabilidade das lâmpadas não foi correto, uma vez que o desvio padrão é uma medida de variabilidade absoluta. Embora as unidades não sejam diferentes, as médias das amostras o são. O procedimento adequado seria o de estimar o CV para ambas as lâmpadas e compará-los. De acordo com a equação (23), os coeficientes de variação são:

$$CV_A = \frac{S_A}{\bar{X}_A} x100 = \frac{0{,}8}{4{,}0} x100 = 20\% \text{ e } CV_B = \frac{S_B}{\bar{X}_B} x100 = \frac{1{,}2}{8{,}0} x100 = 15\%.$$

É fácil verificar que a lâmpada (B) é a mais uniforme, pois possui um menor CV que a lâmpada (A).

Exemplo 9: Testes de resistência à tração aplicada a dois tipos diferentes de aço produziram os seguintes resultados:

Tipo I: $\bar{X} = 27{,}45 \text{ kg/mm}^2$ e $S = 2{,}0 \text{ kg/mm}^2$

Tipo II: $\bar{X} = 147{,}0 \text{ kg/mm}^2$ e $S = 17{,}25 \text{ kg/mm}^2$

Os coeficientes de variação são, respectivamente, 7,29% e 11,73%. Conclui-se que, embora menos resistente, o tipo I se apresenta relativamente mais estável.

3.5 - Erro Padrão da Média

É uma medida da dispersão das médias amostrais em torno da media da população, ou seja, é uma medida que fornece uma ideia da precisão com que a média foi estimada (FERREIRA, 2005).

O erro padrão da média **populacional** é:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \qquad \text{ou } \sigma_{\bar{X}} = \sqrt{\frac{\sigma^2}{n}} \qquad 14)$$

em que σ é o desvio padrão populacional e n é o tamanho da amostra.

O erro padrão da média amostral é:

$$S_{\bar{X}} = \frac{S}{\sqrt{n}} \qquad \text{ou } S_{\bar{X}} = \sqrt{\frac{S^2}{n}}, \qquad 15)$$

em que *S* é o desvio padrão amostral e *n* é o tamanho da amostra.

REFERÊNCIAS

BUSSAB, W.O.; MORETTIN, P. A. **Estatística básica**. 5 ed. São Paulo: Saraiva, 2003.

CRESPO, A.A. **Estatística fácil**. 17.ed. São Paulo: Editora Saraiva, 1999.

FERREIRA, D. F. **Estatística básica**. Lavras: Editora UFLA, 2005.

LEVIN. J.; FOX, J. A. **Estatística para Ciências Humanas**. 9 ed. São Paulo: Prentice Hall, 2004.

4 - PROBABILIDADES

Neste capítulo e no próximo serão abordados os conceitos de probabilidade e serão considerados alguns modelos probabilísticos específicos que desempenham importante papel na estatística. Para o cálculo de probabilidades é necessário contar o número de vezes que um determinado evento de interesse ocorre, fazendo o uso de métodos de análise combinatória.

Probabilidades e espaço amostral

Antes de entrarmos no contexto de probabilidade é necessário entendermos alguns conceitos como: experimento, espaço amostral e eventos.

Denominamos de *experimento* a todo fenômeno ou ação que geralmente pode ser repetido e cujo resultado é aleatório.

Exemplo: Quando lançamos uma moeda, uma única vez estamos fazendo um experimento cujo resultado será cara ou coroa.

- ***Espaço amostral* (Ω)** e o conjunto de todos os possíveis resultados de um determinado experimento.

Exemplos: No lançamento de um dado, o espaço amostral é: Ω ={1, 2, 3, 4, 5, 6}. No lançamento de uma moeda, o espaço amostral é: Ω ={cara, coroa}.

- ***Evento*** e todo subconjunto do espaço amostral, geralmente representado por letra maiúscula (A, B, C, etc).

Outras definições importantes:

i) Evento certo → Ω (caracterizado pelo espaço amostral)

ii) Evento impossível→ Φ.

iii) Processo aleatório: Qualquer fenômeno que gere um resultado incerto ou casual.

Exemplo: lançamento de moeda, lançamento de dado, etc.

Características processo aleatório.

1) Pode ser repetido indefinidamente sob as mesmas condições.
2) Não se conhece a priori (inicialmente) o resultado, mas todos os resultados possíveis podem ser descritos.

Dentro deste contexto, ***Probabilidade*** pode ser definida como o número de eventos (pontos ou elementos) favoráveis divididos pelo número de elementos do espaço amostral:

$$P = \frac{X}{n}.$$

Em que X é o número de eventos favoráveis, e *n* número de eventos do espaço amostral.

OPERAÇÕES

A seguir apresentaremos o Diagrama de Venn para ilustrarmos algumas propriedades:

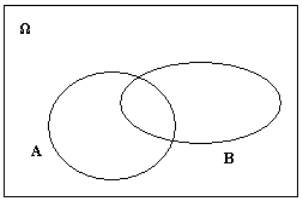
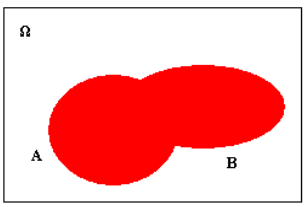
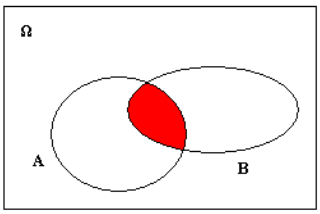


Figura1: Diagrama de Venn.

1) **União (∪):** $A \cup B = B \cup A$

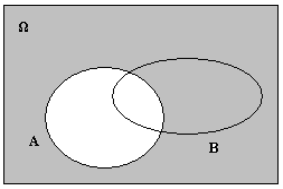


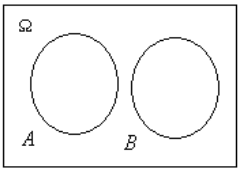
2) **Intersecção (∩):** $A \cap B = B \cap A$



3) **Complementar:** $A^C = \Omega - A$ (lê-se: complementar de *A*).

Observação Importante: Se *A* e *B* são conjuntos mutuamente exclusivos (disjuntos) então, $A \cap B = \Phi$.





Exercícios

1) Um casal pretende ter 3 filhos.

a) Determine o espaço amostral referente ao sexo dos filhos.

b) Qual o número de elementos (eventos) do espaço amostral?

c) Qual a probabilidade do casal ter exatamente 3 filhas?

d) Qual a probabilidade do casal ter exatamente dois filhos?

e) Qual a probabilidade do casal ter apenas um filho?

2) Jogando-se dois dados, calcular a probabilidade da soma dos pontos ser superior a nove.

Dessa forma podemos sintetizar a definição de ***Probabilidade de ocorrer um evento A*** $(\boldsymbol{P(A)})$ como a razão entre o número de possíveis resultados favoráveis ao evento *A* **(*n*(*A*))** e todos os possíveis resultados do experimento **(*n*(Ω))**, ou seja, número de elementos do espaço amostral.

$$\boldsymbol{P(A) = \frac{n(A)}{n(\Omega)}}.$$

Axiomas de Probabilidade

***Axioma* 1:** A probabilidade de um certo evento ocorrer corresponde a um número não negativo.

$$\boldsymbol{P(A) \geq 0}.$$

***Axioma* 2:** A probabilidade de ocorrer todo o espaço amostral é igual a um.

$$P(\Omega) = 1.$$

Teoremas

***Teorema* 1:** A probabilidade de um evento impossível ocorrer é $P(\Phi) = 0$.

Demonstração: Seja Ω o espaço amostral. Sabe-se que $\Omega = \Omega + \Phi$, então aplicando a função probabilidade de ambos os lados têm-se:

$$\Omega = \Omega + \Phi \Rightarrow P(\Omega) = P(\Omega) + P(\Phi) \Rightarrow 1 = 1 + P(\Phi) \Rightarrow P(\Phi) = 0$$

***Teorema* 2 (Probabilidade do complemento):** Seja Ω o espaço amostral. Então, a probabilidade de um evento *A* não ocorrer é:

$$P(A^C) = 1 - P(A).$$

Demonstração: Sabe-se que $A^C = \Omega - A$, então aplicando a função probabilidade de ambos os lados têm-se:

$$A^C = \Omega - A \Rightarrow P(A^C) = P(\Omega) - P(A) \Rightarrow P(A^C) = 1 - P(A)$$

***Teorema* 3 (Teorema da soma):** Se *A* e *B* são dois eventos do espaço amostral Ω a probabilidade que ocorra *A* ou *B* é:

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

Definição: Eventos mutuamente exclusivos

Dois ou mais eventos são mutuamente exclusivos quando a realização de um exclui a realização do outro.

Assim, no lançamento de uma moeda, o evento o evento "tirar cara" e o evento "tirar coroa" são mutuamente exclusivos, já que, ao se realizar um deles, o outro não se realiza.

***Corolário*:** Se dois eventos *A* e *B* são mutuamente exclusivos (disjuntos), isto é, $A \cap B = \Phi$, então:

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

Baseado no ***Axioma* 1** e no ***Corolário*** acima segue-se que $0 \leq P(A) \leq 1$.

Exercícios

1) Um lote é formado por 11 peças boas, 3 com defeitos leves, e 2 com defeitos graves. Considere como evento A defeito leve, evento B defeito grave, e evento C nenhum defeito.

Uma peça é retirada ao acaso desse lote. Qual a probabilidade que essa peça:

a) Seja boa?

b) Tenha defeito leve?

c) Tenha defeito grave?

d) Seja defeituosa?

Duas peças são retiradas ao acaso com reposição desse lote. Qual a probabilidade de:

e) Ambas serem boas?

f) Pelo menos uma boa?

Duas peças são retiradas ao acaso sem reposição desse lote. Qual a probabilidade de:

g) Ambas serem boas?

2) Se um dado é lançado duas vezes. Determine qual a probabilidade de ocorrer maior do que 3 no primeiro lance e menor do que 5 no segundo lance.

3) Em uma bolsa tem-se duas moedas de 1 centavo, três de 10 centavos e quatro de 1 real. Duas moedas são retiradas aleatoriamente da bolsa, determine as seguintes possibilidades (sem reposição).

a) Ambas as moedas serem de 1 centavo.

b) Uma moeda de 1 centavo e a outra moeda de 1 real.

c) Ambas do mesmo valor.

d) Pelo menos uma de 10 centavos.

e) Nenhuma moeda de 10 centavos.

Probabilidade Condicional e Independência

Probabilidade Condicional

A probabilidade condicional do evento A em relação ao evento B é denotada por:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \quad P(B) > 0.$$

A probabilidade condicional do evento B em relação ao evento A é denotada por:

$$P(B|A) = \frac{P(A \cap B)}{P(A)}, \quad P(A) > 0.$$

Exercício 1: Qual a probabilidade no lançamento de um dado, a face superior do dado ser maior ou igual a 4 sabendo que ela é par?

Exercício 2: Em uma urna tem-se 40 bolas, sendo10 pretas e 30 vermelhas (20 com manchas brancas e 10 sem manchas). Qual a probabilidade de se ter uma bola vermelha com mancha branca, sabendo que o evento bola vermelha já ocorreu.

Independência de Eventos

Dois eventos são independentes quando a realização ou a não-realização de um dos eventos não afeta a probabilidade da realização do outro e vice-versa, ou seja, dois eventos A e B são independentes se $P(A \cap B) = P(A)P(B)$.

Exercício 1: Considere o lançamento de uma moeda (não viciada) três vezes. Cujo evento A corresponde ao primeiro lançamento da moeda sair cara e o evento B corresponde ao segundo lançamento da moeda sair cara. Esses dois eventos são independentes?

Exercício 2: Distribuição de alunos matriculados em um determinado instituto de Matemática. Com base na Tabela abaixo, determine:

Curso	Sexo		Total
	Masculino	Feminino	
Agronomia	70	40	110

Matemática	15	15	30
Estatística	10	20	30
Sistemas	20	10	30
Total	115	85	200

a) Probabilidade do sexo masculino.

b) Probabilidade agronomia.

c) Probabilidade matemática.

d) Probabilidade agronomia e sexo feminino.

e) Probabilidade agronomia dado que ele é do sexo feminino.

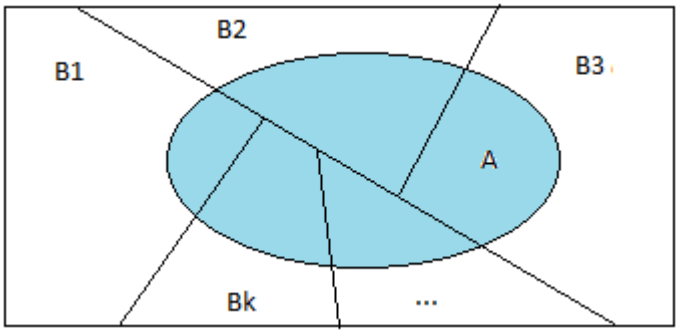
f) Verifique se sexo feminino e agronomia são eventos independentes.

TEOREMA DA PROBABILIDADE TOTOAL

Se os eventos B_1, B_2, B_3,..., B_k constituem uma partição do espaço amostral Ω, de modo que $P(B_i) \neq 0$ para i = 1,2,..., k, então para qualquer evento A de Ω,

$$P(A) = \sum_{i=1}^{k} P(B_i \cap A) = \sum_{i=1}^{k} P(B_i).P(A|B_i)$$

Prova:



$$A = (B_1 \cap A) \cup (B_2 \cap A) \cup \cdots \cup (B_K \cap A)$$

$$P(A) = P[(B_1 \cap A) \cup (B_2 \cap A) \cup \cdots \cup (B_K \cap A)]$$

$$P(A) = P(B_1 \cap A) + P(B_2 \cap A) + \cdots + P(B_K \cap A)$$

$$P(A) = \sum_{I=1}^{k} P(B_i \cap A)$$

Sabe-se que $P(A|B_i) = \frac{P(A \cap B_i)}{P(B_i)} \Leftrightarrow P(A|B_i)P(B_i) = P(A \cap B_i)$. Então,

$$P(A) = \sum_{I=1}^{k} P(B_i \cap A) = \sum_{I=1}^{k} P(A|B_i)P(B_i)$$

Nota: Se a união de n eventos mutuamente exclusivos é o próprio universo Ω, dizemos que tais eventos são mutuamente exclusivos e exaustivos, ou formam uma partição em Ω.

Exercício 1: Em certa linha de montagem, três máquinas B_1, B_2 e B_3 produzem 30%, 45% e 25% dos produtos, respectivamente. Sabe-se de experiências anteriores, que 2%, 3% e 2% dos produtos feitos por cada máquina são, respectivamente, defeituosos. Agora, suponha que um produto já acabado, seja selecionado aleatoriamente. Qual é a probabilidade de que tal produto apresente algum defeito?

TEOREMA DE BAYES

É um importante teorema que expressa o conceito de uma probabilidade condicional em função de outras probabilidades condicionais e marginais.

Teorema de Bayes: Se B_1, B_2,..., B_k são conjuntos mutuamente exclusivos cuja união resulta em Ω, então:

$$P(B_i|A) = \frac{P(B_i)P(A|B_i)}{\sum_{i=1}^{k} P(B_i)P(A|B_i)}.$$

Exercício 1: Considere cinco urnas cada uma com seis bolas. Duas dessas urnas (tipo C_1), tem três bolas brancas, duas outras urnas (tipo C_2), tem duas bolas brancas e a última (tipo C_3) tem seis bolas brancas. Escolhe-se uma urna ao acaso e retira-se uma bola desta. Qual a probabilidade de que a urna escolhida seja do tipo C_3, sabendo-se que a bola retirada á branca?

Exercício 2: Uma empresa produz circuitos integrados em três fábricas. A fábrica 1 produz 40% dos circuitos enquanto que as fábricas 2 e 3, produzem 30% cada. A probabilidade de que um circuito produzido por estas fábricas não funcione é de 0,01; 0,04 e 0,03 respectivamente. Qual a probabilidade de se pegar um circuito ao acaso da produção total da companhia, sendo ele da fábrica 1 e sabendo que ele não funciona?

Exercício 3: Redes bayeseanas são usadas nos sites da internet de fabricantes de alta tecnologia para permitir aos consumidores diagnosticar rapidamente problemas nos produtos. Como exemplo, temos que um fabricante de impressoras identificou que as falhas estão associadas a três tipos de problemas: máquina, programas e outros (tais como conectores) com probabilidades de 0,1 0,6 e 0,3, respectivamente. A probabilidade de uma falha na impressora devido a um problema de máquina é 0,9, devido a um problema de programa é 0,2 e devido a outros problemas é 0,5. Determine a probabilidade de uma impressora falhar. Se o consumidor entrar no site do fabricante para o diagnóstico da falha, qual será a causa mais provável do problema? Justifique com probabilidades. R: a) 0,36 ; b) 0,25; 0,333; **0,417**

5 - VARIAVEIS ALEATORIAS

Variável Aleatória Unidimensional

Para entendermos este conceito de variável aleatória (*v.a.*), imagine um lançamento de um dado. Tente dizer qual será o número resultante. É claro que, antes do lançamento, não podemos dizer com exatidão qual é o número que ocorrerá, pois o resultado depende do fator sorte e, por isso, é uma *variável aleatória*.

Variável Aleatória (v.a.) é uma variável cujos valores são determinados pelos resultados de experiências aleatórias, isto é, *uma função que associa valores reais aos eventos de um espaço amostral.*

Em vez de operar com o espaço amostral, agora utilizaremos *variável aleatória,* que adota valores de acordo com os resultados de um experimento aleatório.

Uma *v.a.* pode ser entendida como uma variável quantitativa, ou seja, uma *v.a.* pode ser classificada como ***discreta*** ou ***contínua***. As variáveis aleatórias dizem-se discretas, quando assumem um número determinado de valores contáveis (valores oriundos de um processo de contagem), ou contínuas, quando assumem qualquer valor num dado intervalo (valores oriundos de um processo de mensuração).

Variável Aleatória Discreta

O conceito de *v.a.* discreta será introduzido por meio de exemplos.

Exemplo 1: Se um experimento consiste no lançamento de dois dados, a função: X = “soma das faces dos dois dados”, define uma variável aleatória discreta, que pode assumir onze valores possíveis: 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 ou 12.

Exemplo 2: Se um experimento consiste em verificar o número de circuitos defeituosos num sistema formado por quatro circuitos, a função: X = “número de circuitos defeituosos”, define uma variável aleatória discreta, que pode assumir cinco valores possíveis: 0, 1, 2, 3 ou 4.

Com base nos exemplos acima fica claro que a variável aleatória discreta está vinculada a valores de uma contagem que resultam a números inteiros.

Variável Aleatória Contínua

A variável aleatória é dita contínua se corresponder a dados de medida, pertencentes aos $\mathbb{R}$. O conceito de *v.a.* contínua será mais bem entendido por meio do exemplo a seguir.

Exemplo 3: Se um experimento consiste em verificar as alturas de 30 universitários, a função: X = “Altura de um universitário”, define uma variável aleatória contínua, que pode assumir quaisquer valores entre 130 e 220 cm.

Exemplo 4: Se um experimento consiste em verificar (mensurar) os pesos dos 30 universitários, a função: Y = "Peso de um universitário", define uma variável aleatória contínua, que pode assumir quaisquer valores entre 60 e 130 kg.

Exemplo 5: Se um experimento consiste em verificar a durabilidade de um lote de 50 pneus, a função: Z: "tempo de vida útil de um pneu", define uma v.a. contínua, que pode assumir quaisquer valores entre 50.000 e 70.000 km.

Com base nos exemplos apresentados, a v.a. contínua está vinculada a dados oriundos de uma mensuração que resultam a um intervalo de números reais.

Distribuição de Probabilidades

Se uma variável aleatória X pode assumir os valores $x_1, x_2, \ldots, x_n$ com probabilidades respectivamente $P[X = x_1], P[X = x_2], \ldots, P[X = x_n]$, tais que $\sum_{i=1}^{n} P[X = x_i] = 1$, tem-se definida uma ***distribuição de probabilidade***.

No tocante a variáveis aleatórias discretas, a cada realização x_i corresponde uma probabilidade P[X = x_i]. Isso define uma função, chamada ***função de probabilidade***, a qual deve obedecer a algumas condições, quais sejam:

i) $P[X = x_i] \geq 0$, para todo i; ii) $\sum_{i=1}^{\infty} P[X = x_i] = 1$

em que o índice *i* é empregado para identificar os diferentes valores que a variável pode assumir. Essa função é denominada por inúmeros autores como ***função distribuição de probabilidade*** da variável aleatória discreta Y.

Nota: $\sum_{y_i > a}^{b} P[X = x_i] = P[a < x \leq b]$

Exemplo 6: Y: número de circuitos defeituosos num sistema formado por quatro circuitos tem-se:

X	0	1	2	3	4	
P[X = x]	1/8	2/8	2/8	2/8	1/8	$\sum_{i=1}^{5} P[X = x_i] = 1$

Observa-se que a distribuição de probabilidade acima é uma função de probabilidade pois, as condições (i) e (ii) foram satisfeitas, isto é, todas as probabilidades são maiores que zero e, a soma das probabilidades é igual a um.

Exercício 1: Experimento: jogar 2 moedas e observar o resultado (c = cara e r = coroa). Sendo *X*: número de caras em 2 lances de moeda.

Se, a variável Y for contínua, somente haverá interesse na probabilidade de que a variável assuma valores dentro de determinados intervalos, sendo sua distribuição de probabilidades caracterizada por uma ***função densidade de probabilidade*** (*f.d.p.*), *f(y)*, a qual deverá possuir as seguintes propriedades:

i) $f(y) \geq 0, \forall y \in \mathbb{R}$; ii) $\int_{-\infty}^{\infty} f(y)\ dy = 1$.

Nota: $P[a \leq y \leq b] = P[a < y \leq b] = P[a \leq y < b] = P[a < y < b] = \int_{a}^{b} f(y)\ dy, \forall a \text{ e } b.$

Exemplo 7: Para o caso das alturas dos universitários têm-se:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right],$$ que é a distribuição normal.

Exercício: Suponha que a v.a. X definida a seguir represente a produção de um determinado insumo

$$f(x) = \begin{cases} 6(x - x^2) & para \quad o \le X \le 1 \\ 0, & p.o.v. \ de \ X \end{cases}$$

a) Verificar se f(x) é uma f.d.p.

b) Representar graficamente f(x)

c) Calcular as seguintes probabilidades:

i) P(X<0,4) ii) P(0,5 < X< 0,8) iii) P(X=0,5)

d) Determinar F(X) e representá-la graficamente.

e) Calcular a média e o desvio padrão de X.

Função Repartição ou Função Distribuição Acumulada

A função de distribuição acumulada nos fornece a probabilidade de que a variável em questão esteja abaixo de um determinado valor. Em geral, ela é representada por $F(y)$ ou $\varphi(y)$. Assim,

$$F(y) = P[Y \le y].$$

i) Para uma variável aleatória discreta a função distribuição acumulada será definida como:

$$F(y_k) = P[Y \le y_k] = P[Y = y_1] + P[Y = y_2] + \cdots + P[Y = y_k] = \sum_{i=1}^{k} P[Y = y_i].$$

ii) Para uma variável aleatória contínua a função distribuição acumulada será definida como:

$$F(y_k) = P[Y \le y_k] = \int_{-\infty}^{y_k} f(y)dy.$$

Exemplo 8: Numa plantação de café, cujas folhas possuem um número Y variado de lesões provocadas pela praga bicho mineiro (*Perileucoptera coffeella*), obedecendo as seguintes proporções:

Nº lesões	0	1	2	3	4	5
proporção	0,32	0,28	0,20	0,12	0,06	0,02

Essas proporções podem ser interpretadas como probabilidades no sentido de que, se uma folha for tomada à plantação ao acaso, existe uma probabilidade, por exemplo, de 28% de que ela contenha apenas uma lesão. A probabilidade de que ela tenha 3 lesões, ou menos, é dada por:

$$F(Y=3)=F(3)=P[Y\leq 3]=P[Y=0 \text{ ou } Y=1 \text{ ou } Y=2 \text{ ou } Y=3]$$
$$F(Y=3)=F(3)=P[Y\leq 3]=P[Y=0]+P[Y=1]+P[Y=2]+P[Y=3]$$
$$F(Y=3)=F(3)=P[Y\leq 3]=0{,}32+0{,}28+0{,}20+0{,}12=0{,}92.$$

Exemplo 9: Seja a função densidade de probabilidade:

$$f(x)=\begin{cases}0, & x<0\\ kx, & 0\leq x\leq 2.\\ 0, & x>2\end{cases}$$

Encontre F(1).

Propriedades da Função Distribuição Acumulada ou Função Repartição

i) $0\leq F(y)\leq 1$; ***ii)*** se$y_1<y_2$, então, $F(y_1)\leq F(y_2)$; ***iii)*** $F(-\infty)=\lim_{y\to-\infty}F(y)=0$;

iv)$F(+\infty)=\lim_{y\to+\infty}F(y)=1$, corresponde ao evento certo;***v)*** $P[y_1<Y\leq y_2]=F(y_2)-F(y_1)$;

vi)$P[y_1\leq Y\leq y_2]=F(y_2)-F(y_1)+P[Y=y_1]$;

vii)$P[y_1<Y<y_2]=F(y_2)-F(y_1)-P[Y=y_2]$.

Gráfico da Função Distribuição Acumulada ou Função Repartição

Seja X a variável aleatória **discreta** com a seguinte função de probabilidade:

X	0	1	2	3	4	
P[X = *x*]	0,1	0,3	0,4	0,1	0,1	1,0

Então, sua função distribuição acumulada é:

X	0	1	2	3	4
F(*x*)	0,1	0,4	0,8	0,9	1,0

Portanto, o gráfico da função distribuição acumulada da variável aleatória X é:

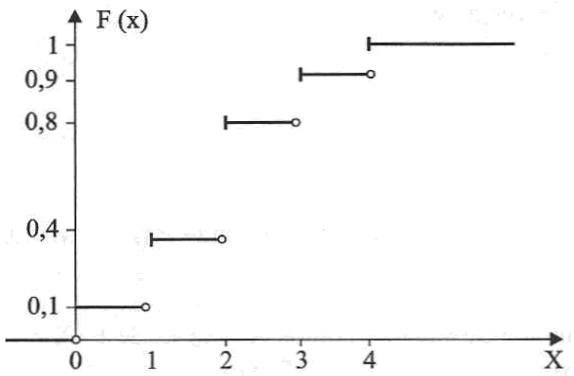


Figura 1 – Gráfico da função distribuição acumulada da variável aleatória X.

Seja X a variável aleatória **contínua**, então o gráfico genérico da função distribuição acumulada terá o seguinte comportamento:

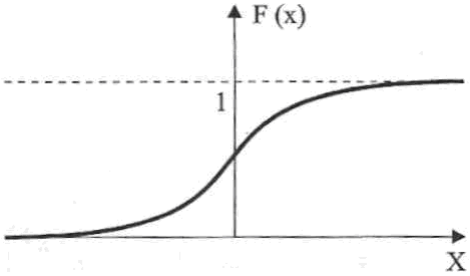


Figura 2 – Gráfico genérico da função distribuição acumulada de uma v.a. contínua X.

Exemplo10: Plote o gráfico da seguinte função distribuição acumulada:

$$F(x) = \begin{cases} 0, & x \leq 0 \\ x^2, & 0 < x < 1 \\ 1, & x \geq 1 \end{cases}$$

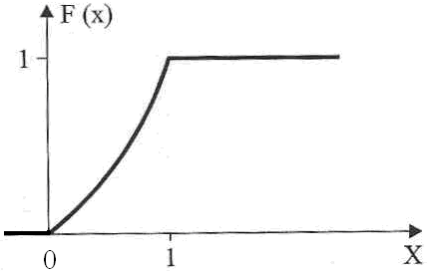


Figura 3 – Gráfico da função distribuição acumulada de uma v.a. contínua X.

Observação: Pode-se encontrar a função densidade de probabilidade, se existir, a partir de F(x), pois:

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

nos pontos onde F(x) é derivável.

Parâmetros característicos de uma Distribuição de Probabilidade

Esperança Matemática

Muitas vezes estamos interessados em estimar parâmetros característicos de uma distribuição de probabilidade de uma variável aleatória qualquer. Um primeiro parâmetro é a Esperança Matemática. A esperança matemática é uma média aritmética ponderada ou um valor esperado de uma variável aleatória. Na prática, a esperança pode ser entendida como um "centro de distribuição de probabilidade", isto é, a média de uma distribuição de probabilidade.

A Esperança Matemática é definida da seguinte forma:

i) Se X é uma variável aleatória discreta, então a esperança matemática é:

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

ii) Se X é uma variável aleatória contínua, então a esperança matemática é:

$$E(X) = \mu = \int_{-\infty}^{\infty} xf(x)dx.$$

Exemplo 11 - (MORETTIN, 1999): Uma seguradora paga R$ 30.000,00 em caso de acidente de carro e cobra uma taxa de R$ 1.000,00. Sabe-se que a probabilidade de que um carro sofra acidente é de 3%. Quanto espera a seguradora ganhar por carro segurado?

Propriedades da Esperança Matemática

As propriedades da esperança são:

1) $\boldsymbol{E(k) = k}$, sendo *k* uma constante.

Demonstração:

$$\boldsymbol{E(k) = \sum_{i=1}^{n} kP[X = x_i] = k\left[\sum_{i=1}^{n} P[X = x_i]\right]^1 = kx1 = k.}$$

2) $\boldsymbol{E(kX) = kE(X)}$, sendo *k* uma constante.

Demonstração:

$$\boldsymbol{E(kX) = \sum_{i=1}^{n} kx_iP[X = x_i] = k\sum_{i=1}^{n} x_iP[X = x_i] = kE(X).}$$

3) $\boldsymbol{E(aX \pm b) = aE(X) \pm b}$, sendo *a* e *b* constantes.

Demonstração:

$$\boldsymbol{E(aX \pm b) = E(aX) \pm E(b) = aE(X) \pm b.}$$

4) $\boldsymbol{E(X - \mu_x) = 0}$

Demonstração:

$$E(X - \mu_x) = E(X) - E(\mu_x) = \mu - \mu = 0.$$

5) $\boldsymbol{E(X \pm Y) = E(X) \pm E(Y)}$

Essa propriedade será demonstrada posteriormente, quando abordarmos o assunto de variáveis aleatórias bidimensionais.

6) $\boldsymbol{E(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} E(X_i)}$

Nota: Para demonstração das propriedades acima foi utilizada a definição de esperança matemática de uma variável aleatória discreta. Analogamente, é possível demonstrar as propriedades da esperança por meio da definição de esperança de uma variável aleatória contínua.

Variância

Já comentamos anteriormente que a esperança matemática nos fornece a média de uma distribuição de probabilidade. Porém, não temos informação a respeito do grau de dispersão das probabilidades em torno da média. Portanto, a medida que usaremos para estimar o grau de dispersão (ou de concentração) de probabilidade em torno da média será a variância.

A Variância é definida da seguinte forma:

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

i) Se X é uma variável aleatória discreta, então a esperança matemática é:

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

ii) Se X é uma variável aleatória contínua, então a esperança matemática é:

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

Variável Aleatória Bidimensional

Até o momento foi considerado para um determinado experimento a associação de um único valor *x*. No entanto, existem situações em que há interesse por dois ou mais resultados simultâneos, como por exemplo, o peso e a altura de uma pessoa, o sexo e peso de um recém nascido, etc. Para tanto, faz-se necessário a seguinte definição:

Definição: Sejam ***E*** um experimento aleatório, e **Ω** um espaço amostral referente a ***E***. Sejam X, Y duas variáveis aleatórias. Então, (X, Y) define uma ***variável aleatória bidimensional,*** que pode ser **discreta, contínua ou mista**.

Distribuição de Probabilidade Conjunta de duas variáveis aleatórias

Se a variável aleatória bidimensional (X, Y) pode assumir os valores $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ com as respectivas probabilidades $P[X = x_1, Y = y_1], P[X = x_2, Y = y_2], \dots, P[X = x_n, Y = y_n]$, tais que $\sum_{i=1}^{n} \sum_{j=1}^{n} P[X = x_i, Y = y_j] = 1$, tem-se definida uma ***distribuição de probabilidade conjunta*** de duas variáveis aleatórias.

Se a variável aleatória bidimensional (X, Y) é ***discreta***, sua distribuição de probabilidade conjunta será caracterizada por uma ***função de probabilidade conjunta*** $\boldsymbol{P[X = x_i, Y = y_i]}$, que devem obedecer a algumas condições:

i) $P[X = x_i, Y = y_i] \geq 0$, para todo i; ii) $\sum_{i=1}^{\infty} \sum_{i=1}^{\infty} P[X = x_i, Y = y_i] = 1$

Exemplo 12 - Seja o experimento de se lançar simultaneamente um dado e uma moeda, observando o resultado da face superior de ambos. Teremos então a seguinte função de probabilidade conjunta,

em que:

X: face superior do dado, e Y: face superior da moeda.

X	Y		$P[X = x_i]$
	Cara	Coroa	
1	1/12	1/12	2/12
2	1/12	1/12	2/12
3	1/12	1/12	2/12
4	1/12	1/12	2/12
5	1/12	1/12	2/12
6	1/12	1/12	2/12
$P[Y = y_i]$	6/12	6/12	1

$P[X = 1, Y = Cara]$= Probabilidade de sair face 1 no dado **e** sair cara na moeda

$P[X = 1, Y = Cara]$= P(sair face 1) x P(sair cara) = $\frac{1}{6} x \frac{1}{2} = \frac{1}{12}$

Se a variável aleatória bidimensional (X, Y) é ***contínua***, sua distribuição de probabilidade conjunta será caracterizada por uma ***função densidade de probabilidade conjunta*** $\boldsymbol{f(x, y)}$, que devem obedecer as seguintes condições:

i)$f(x, y) \geq 0, \forall (x, y) \in \mathbb{R}^2$; ii) $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\ dxdy = 1$.

Distribuição Marginal

Dada a v.a. bidimensional (X, Y), e sua distribuição de probabilidade conjunta, pode-se obter a distribuição de X conhecida como distribuição marginal de X desconsiderando-se Y e vice-versa.

Distribuição Marginal de v.a. Discretas

i) Distribuição Marginal de X: $\boldsymbol{P[X = x_i] = \sum_{j=1}^{\infty} P[X = x_i, Y = y_j]}$

ii) Distribuição Marginal de Y: $\boldsymbol{P[Y = y_j] = \sum_{i=1}^{\infty} P[X = x_i, Y = y_j]}$

Exemplo 13 - Para o caso do lançamento simultâneo do dado e da moeda tem-se:

X	Y		$P[X = x_i]$
	Cara	Coroa	
1	1/12	1/12	2/12
2	1/12	1/12	2/12
3	1/12	1/12	2/12
4	1/12	1/12	2/12
5	1/12	1/12	2/12
6	1/12	1/12	2/12
$P[Y = y_i]$	6/12	6/12	1

Determine:

a) Distribuição Marginal de X:

b) Distribuição Marginal de Y:

Distribuição Marginal de v.a. Contínuas.

i) Distribuição Marginal de X: $\boldsymbol{g(x) = \int_{-\infty}^{\infty} f(x,y)dy}$

ii) Distribuição Marginal de Y: $\boldsymbol{h(y) = \int_{-\infty}^{\infty} f(x,y)dx}$

Nota: $\boldsymbol{P(a < x < b, c < y < d) = \int_c^d \int_a^b f(x,y)dxdy}$

Função Distribuição Conjunta de X e Y:

$$\boldsymbol{F(x,y) = P[X \leq x, Y \leq y] = \int_{-\infty}^{x}\int_{-\infty}^{y} f(u,v)dudv}$$

Exemplo 14 - Se a seguinte função de probabilidade conjunta, $f(x,y) = 2 - x - y,\ 0 \leq x \leq 1$ e $0 \leq y \leq 1$, define uma distribuição bidimensional contínua, determine a probabilidade que X > ½ e Y > ½.

Esperança matemática:

No caso geral, temos que $E(XY) \neq E(X)E(Y)$.

No caso particular de *X* e *Y* forem variáveis aleatórias independentes, temos que: $E(XY) = E(X)E(Y)$

Em que: $E(XY) = \sum_{j=1}^{n}\sum_{i=1}^{m} x_i y_j P[X = x_i, Y = y_j]$ se X e Y forem discretas.

$E(XY) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} xyf(x,y)\,dxdy$ se X e Y forem continuas.

Covariância entre duas variáveis

E o valor médio do produto dos desvios de X e Y em relação as suas respetivas medias, mede o grau de dependências entre duas variáveis, em termo absoluto, ou seja, serve para verificar se duas variáveis aleatórias movimentam-se ou não no mesmo sentido.

$$Cov(X,Y) = E(XY) - E(X)E(Y)$$

Observação: Quando X e Y são independentes, $Cov(X,Y) = 0$. Porem se $Cov(X,Y) = 0$, não significa que as variáveis sejam independentes.

Matriz variância covariância

$$\boldsymbol{\sum x} = \begin{pmatrix} \boldsymbol{\sigma_{11}} & \cdots & \boldsymbol{\sigma_{1n}} \\ \vdots & \ddots & \vdots \\ \boldsymbol{\sigma_{n1}} & \cdots & \boldsymbol{\sigma_{nn}} \end{pmatrix}$$

Quando i = j temos a variância da variável unidimensional, e quando i ≠ j, temos a covariância entre as duas variáveis.

Coeficiente de Correlação (ρ_{xy}):

Mede o grau de dispersão linear entre duas variáveis X e Y, através de:

$$\rho_{xy} = \frac{Cov_{XY}}{\sigma_X \sigma_Y} = \frac{E(XY) - E(X)E(Y)}{\sigma_X \sigma_Y}$$

Exercícios:

1) Seja (X,Y) uma variável aleatória bidimensional discreta com a seguinte distribuição:

X	Y		
	-3	2	4
1	0,1	0,2	0,2
3	0,3	0,1	0,1

Calcular:

a) Cov_{XY} b) ρ_{xy}

2) A função densidade conjunta de duas variáveis aleatórias X e Y, continuas e:

$$f(x,y) = \begin{cases} cxy, 0 < x < 4, 1 < y < 5 \\ 0, cc \end{cases}$$

a) Determine a constante c;

b) Determine P(1 < x < 2, 2 < y < 3);

c) Determine as funções distribuição marginal de X e Y;

d) Determine a f.d conjunta de va X e Y;

e) Determine P(X + Y < 3).

f) Determine e interprete, Cov_{XY};

g) Determine e interprete ρ_{xy}:

6- DISTRIBUICOES DE PROBABILIDADES

MODELOS PROBABILISTICOS DISCRETOS

Vimos que o cálculo de probabilidades pode ser feito a *prori* ou a *posteriori*.

No cálculo a posteriori o pesquisador monta o experimento e com base no espaço amostral ele faz o calculo da probabilidade. Como exemplo, temos o caso de se jogar um moeda por 2 vezes e

calcular a probabilidade de não sair cara.

O cálculo a *priori* é feito a partir de hipóteses segundo um modelo matemático e sem experimentação, determinando as probabilidades de acontecimentos futuros.

Desta forma, se conhecemos as características da variável aleatória em estudo, podemos utilizar modelos teóricos de probabilidade para realizar os cálculos das possibilidades de acontecimentos de eventos de um espaço amostral.

Principais modelos discretos

- Uniforme Discreta
- Bernoulli
- Binomial
- Poisson
- Hipergeométrica

Distribuição Uniforme Discreta

É a mais simples das distribuições de variáveis aleatórias discretas

Definição: Considere uma v.a. *X* cujos valores vão de 1 a N, equiprováveis, ou seja, todos os valores têm igual probabilidade de ocorrência, então:

$$P(X = x) = \frac{1}{N} \quad \text{com: X= } x_1, x_2, \ldots, x_N$$

define a função de probabilidade da v. a. uniforme

Os parâmetros característicos dessa distribuição são:

Media $E(X) = \frac{N+1}{2}$ Variância $Var(X) = \frac{N^2-1}{12}$

Exemplo:

Lança-se um dado e define-se uma v.a. *X* como o valor obtido neste dado.

X: {1, 2, 3, 4, 5, 6}

P(*X* = 1) = 1/6

P(*X* = 2) = 1/6

P(*X* = 3) = 1/6

P(*X* = 4) = 1/6

P(*X* = 5) = 1/6

P(*X* = 6) = 1/6

$$P(X=x)=\begin{cases}\frac{1}{N}\text{se X=1,2,3,4,5,6}\\0,p.o.vdeX\end{cases}$$

$$E(X)=\frac{N+1}{2}=\frac{6+1}{2}=3{,}5 \qquad Var(X)=\frac{N^2-1}{12}=\frac{36-1}{12}=2{,}92$$

Distribuição Bernoulli

Considere uma única tentativa de um experimento aleatório.

Nessa tentativa podemos ter sucesso ou fracasso

Seja *p* a probabilidade de sucesso e *q* a probabilidade de fracasso (ou insucesso), com *p + q = 1*.

Seja a v.a. X = número de sucessos em uma única tentativa, então X assumira o valor 0, que corresponde ao fracasso, com probabilidade *q*, ou assumirá o valor 1, correspondente ao sucesso, com probabilidade *p*.

$$X=\begin{cases}0,fracassocomP(X=0)=q\\1,sucesssocomP(X=1)=p\end{cases}$$

Nessas condições a v. a. X tem distribuição de Bernoulli e sua função de probabilidade é dada por:

$$P(X=x)=p^x q^{1-x},emqueq=1-p$$

Os parâmetros da distribuição de Bernoulli são:

Média: $E(X)=\sum_{x=0}^{t}xP(X=x)\Rightarrow E(X)=0q+1p\Rightarrow E(X)=p$

Variância: $Var(X)=E(X^2)-[E(X)]^2$

em que $E(X^2)=\sum_{x=0}^{t}x^2P(X=x)\Rightarrow E(X^2)=0^0q+1^2p\Rightarrow E(X^2)=p$

logo $Var(X)=p-p^2\Rightarrow Var(X)=p(1-q)\Rightarrow Var(X)=pq$

Exemplo: Suponha um local com 50 pessoas sendo 30 do sexo feminino e 20 masculino. Seleciona-se uma pessoa desse local. Seja X: número de pessoas do sexo masculino.

a) Montar a função de probabilidade
b) Calcular E(X) e Var (X)

Distribuição Binomial

- É uma generalização da Bernoulli
- É a mais importante das distribuições teóricas de probabilidade para variáveis discretas
- É apropriada nas experiências onde ocorre somente duas situações (sucesso ou fracasso)
- Exemplos de experimentos binomiais: lançamento de moeda; aquisição de um produto; defeito de equipamento; opinião sim ou não, etc..

Um experimento se enquadra como um experimento binomial se as seguintes condições são satisfeitas:

1. A variável é discreta
2. Em cada experimento ocorre apenas o sucesso (p) ou fracasso (q)
3. Os experimentos repetidos são independentes
4. A probabilidade do sucesso (p) permanece constante de experimento para experimento
5. Um número fixo de n experiências são realizadas

Em resumo temos:

Notação

X~b(n,p)

Função de probabilidade binomial

$$P(X = x) = C_n^x p^x q^{n-x}, para x \leq n$$

em que: n é o número de experiências; p é a probabilidade do sucesso e q= 1 - p é a probabilidade do fracasso; $C_n^x = \frac{n!}{(n-x)!x!}$

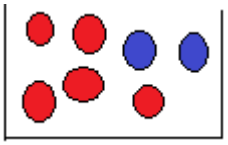
Os parâmetros da distribuição de Binomial são:

Média: $E(X) = \mu = np$ Variância: $Var(X) = \sigma^2 = npq$

Exemplos:

1) Considere o experimento: retiram-se 3 bolas da urna (com reposição) contendo 5 vermelhas e 2 azuis. Define-se uma v.a. *X* cujos valores representam o número total de bolas vermelhas dentre as 3 escolhidas.

X: {0, 1, 2, 3}



O experimento envolve 3 eventos independentes.

Para cada evento:

P(*vermelha*) = 5/7 = *p* (probabilidade de sucesso)

P(*azul*) = 2/7 = *q* (probabilidade de fracasso, *q* = 1 - *p*)

2) Sabe-se que, em certa região, 60% dos municípios possuem usinas de reciclagem de lixo. Cinco cidades são selecionadas nesta região.

a) Verifique se este experimento se enquadra como binomial

b) Construa a função de probabilidade para o número de cidades com usinas de reciclagem

c) Calcule a probabilidade de: i) que exatamente 3 cidades tenham usinas; ii) no máximo 1 tenha usina

d) Determinar o número esperado e o desvio padrão.

3) Um industrial fabrica peças das quais 1/5 são defeituosas. Dois compradores A e B classificaram as partidas adquiridas em categoria I e II, pagando R$1,20 e R$0,80, respectivamente do seguinte modo:

Comprador A: retira uma amostra de 5 peças; se encontrar mais que uma defeituosa classifica como II

Comprador B: retira uma amostra de 10 peças; se encontrar mais que duas defeituosas classifica como II.

Em média, qual comprador oferece maior lucro?

Distribuição de Poisson

Adequada às situações onde se deseja saber o número de sucessos (v. a. discreta) no domínio contínuo.

Exemplos de experimentos Poisson são:

- Número de ocorrências por mês;
- Número de veículos por hora;
- Número de defeitos/ m^2; etc..

A Distribuição de Poisson tem aplicação também nos casos em que os parâmetros n e p da distribuição binomial dificultam o cálculo por esta distribuição. (Eventos raros)

Geralmente adota-se que, quando n >50 e p< 0,10 na distribuição binomial, pode-se utilizar a Poisson para realizar o cálculo aproximado de probabilidade. Neste caso a media da Poisson será: $\lambda = \mu = n.p$

Notação:

$X \sim Po(\lambda)$

Função de probabilidade

$$P(X = x) = e^{-\lambda}\frac{\lambda^x}{x!} \qquad \text{com } X = 0, 1, 2, ...$$

Media e variância da Poisson

$$E(X) = \mu = \lambda \qquad Var(X) = \sigma^2 = \lambda$$

Exemplos de Aplicação

1) Em momentos de pico, a chegada de aviões a um aeroporto se dá segundo um modelo Poisson com taxa de 1 por minuto.

a) Determinar a probabilidade de 3 chegadas em um minuto qualquer do horário de pico.

b) Se o aeroporto pode atender a 2 aviões por minuto, qual a probabilidade de haver aviões sem atendimento imediato?

c) Previsões indicam que para o próximo ano o tráfego neste aeroporto deve dobrar, enquanto que

sua capacidade poderá ser aumentada em 50%. Como ficará a probabilidade de espera?

2) Engenheiros de uma Cia Telefônica estudam se o modelo Poisson se ajusta ao número de chamadas interestaduais, que chegam por hora, a uma central telefônica, durante o período noturno. Os dados a seguir referem-se a 650 períodos de uma hora:

Chamadas	0	1	2	3	4	5	6	7	8
Freq Obs	9	38	71	115	125	106	79	50	57

Verificar se a Poisson poderá ser utilizada para se fazer cálculos de probabilidades para esta variável. (Obs: verificação apenas visual)

3) Estima-se que 5% de um produto fabricado pela indústria X apresente defeitos. Um comerciante adquire um lote de 500 unidades desse produto.
   a) Qual a probabilidade se ter no máximo 20 unidades defeituosas?
   b) Se ele adquirir 60 unidades, qual a probabilidade de se ter exatamente 5 defeituosas?

Distribuição Hipergeométrica

Tanto a distribuição Binomial quanto a Hipergeométrica estão relacionadas com o numero de sucesso em uma amostra contendo n observações. Uma das diferenças nessas duas distribuições de probabilidade e a maneira pela qual as amostras são selecionadas. Para a distribuição binomial, os dados da amostra são selecionados *com reposição*, de uma população *finita*, ou *sem reposição* de uma população *infinita*. Para a distribuição hipergeométrica, os dados da amostra são selecionados *sem reposição* de uma população *finita*, assim o resultado de uma observação dependente dos resultados das observações anteriores.

Considere uma população de tamanho N. Seja A o numero total de sucessos na população, a probabilidade de X sucessos em uma amostra de tamanho *n*, selecionada sem reposição, e dada por:

$$P(X = x) = \frac{\binom{A}{x}\binom{N-A}{n-x}}{\binom{N}{n}}$$

Media e variância da Hipergeométrica

$$E(X) = \mu = \frac{nA}{N} \qquad Var(X) = \sigma^2 = \sqrt{\frac{nA(N-A)}{N^2}}\sqrt{\frac{N-n}{N-1}}$$

Exemplos:

1) Suponha que você esteja formando uma equipe de 80 alunos, de diferentes cursos. Na escola tem um total 300 alunos, 100 deles são do curso de agrimensura. Se os membros da equipe forem selecionados ao acaso, qual e a probabilidade de que a equipe conterá 20 alunos do curso de agrimensura?

$$P(X = 20) = \frac{\binom{100}{20}\binom{300-100}{80-20}}{\binom{300}{80}} = 0.298 = 29.8\%$$

MODELOS PROBABILISTICOS CONTINUOS

Principais modelos contínuos:

- Uniforme
- Exponencial
- Normal

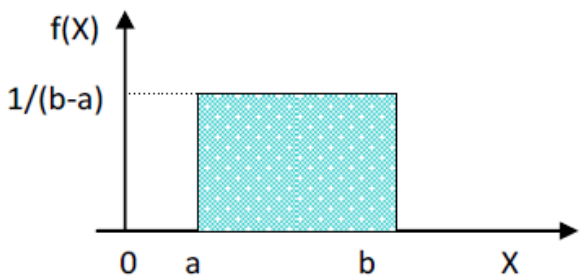
Distribuição Uniforme

Uma distribuição de variável aleatória contínua é a distribuição uniforme cuja função densidade de probabilidade é constante dentro de um intervalo de valores da variável aleatória X.

A variável aleatória X tem distribuição uniforme de probabilidades no intervalo (a, b) se a função densidade f(x) for dada por:

$$f(x) = \begin{cases} \frac{1}{b-a} para a \leq x \leq b \\ 0 para outros valores de x \end{cases}$$

A representação gráfica da distribuição uniforme é um retângulo com base definida pelos valores a e b que estabelecem os limites de valores possíveis da variável aleatória X, Figura abaixo.



Da definição da distribuição uniforme deduzimos:

 A área do retângulo é igual a 1, pois a base é (b – a) e a altura 1/(b – a).

 A probabilidade da variável aleatória X ser igual ou maior que a e, ao mesmo tempo, menor ou igual a b é igual a 1 ou 100%.

A média e a variância (os parâmetros) da variável aleatória X com distribuição uniforme de probabilidades no intervalo (a, b) são:

Media: $E(X) = \frac{a+b}{2}$ Variância: $Var(X) = \frac{(b-a)^2}{12}$

EXEMPLOS:

1) A variável aleatória X tem distribuição uniforme no intervalo (50, 200). Calcular:
   a) Média
   b) Desvio padrão.
   c) A probabilidade de um valor da variável X se encontrar entre 110 e 150.

2) Um ponto e escolhido ao acaso no intervalo [0, 2]. Qual a probabilidade de que esteja entre 1 e 1,5?

3) A dureza H de uma peca do aco pode ser pensada como sendo uma variável aleatória com distribuição uniforme [50, 70] da escala de Rockwel. Calcular a probabilidade de que uma peca tenha dureza entre 55 e 60.

Distribuição Exponencial

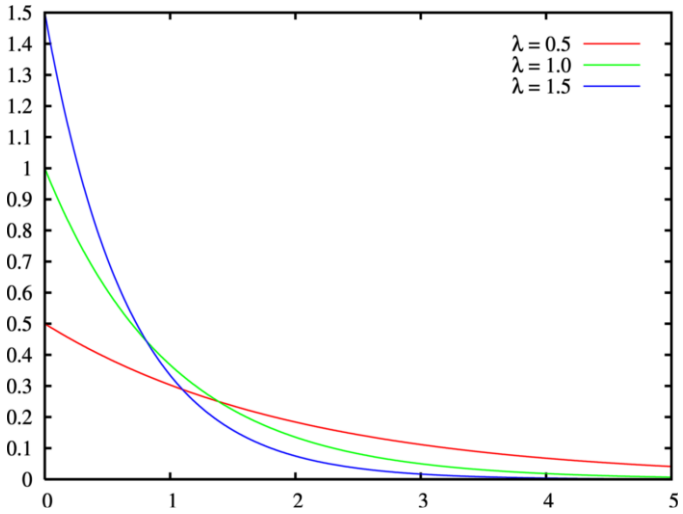
Em certas situações existe interesse no estudo de certas variáveis cujas realizações vão se tornando menos prováveis à medida que o tempo aumenta (Engenharia de confiabilidade).

A variável aleatória X tem distribuição exponencial de probabilidades se a função densidade f(x) for dada por:

$$f(x) = \lambda e^{-\lambda x}, x \geq 0$$

Media e variância da Exponencial

$$E(x) = \frac{1}{\lambda} \qquad V(x) = \frac{1}{\lambda^2}$$



Exemplos:

1) Tempo de vida de uma máquina de lavar roupa.

Quando nova → pouco provável que quebre (alta probabilidade de estar funcionando).

A medida que o tempo passa → baixa probabilidade de estar funcionando

2) Suponha que um componente eletrônico em um sistema de radar de aeronave tenha vida útil descrita por uma distribuição exponencial com taxa de falha 10^{-4}/h, isto é, $\lambda = 10^{-4}$. O tempo médio de falha para este componente é $1/\lambda = 10^4 = 10.000$ h. Se desejamos calcular a probabilidade de esse componente falhar antes de seu tempo esperado de vida, temos que calcular:

$$P\left[X < \frac{1}{\lambda}\right] = \int_0^{\frac{1}{2}} \lambda e^{-\lambda t}\, dt = 1 - e^{-1} = 0{,}63212$$

Esse resultado vale independente do valor de λ, isto é, a probabilidade de que o valor de uma variável aleatória exponencial seja menor que a sua média é 0,63212. Isso acontece, naturalmente, porque a distribuição não é simétrica.

3) Uma fabrica de tubos de TV determinou que a vida média de tubos de sua fabricação é de 800 horas de uso continuo e segue uma distribuição exponencial. Qual a probabilidade de que a fabrica tenha de substituir um tubo gratuitamente se oferece uma garantia de 300 horas de uso?

Distribuição Normal

Entre as distribuições teóricas de variável aleatória contínua, uma das mais empregadas é a distribuição normal. Sua importância em análise matemática resulta do fato de que muitas técnicas estatísticas, como análise de variância, de regressão e alguns testes de hipótese, assumem e exigem a normalidade dos dados. Além disso, a ampla aplicação dessa distribuição vem em parte devido ao teorema do limite central. Este teorema declara que na medida em que o tamanho da amostra aumenta, a distribuição amostral das médias amostrais tende para uma distribuição normal (Triola, 1998).

Função Densidade Probabilidade normal e dada por:

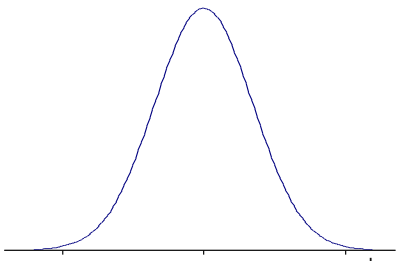
$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Notação: $X \sim N(\mu, \sigma^2)$

- É o modelo de distribuição de probabilidade mais utilizado na estatística
- Métodos e técnicas estatísticas paramétricas geralmente consideram o modelo de distribuição de probabilidade normal.
- Seu gráfico tem a forma campanular (sino)
- É uma distribuição simétrica em relação à média
- É duplamente assintótica em relação ao eixo das abscissas

$$\lim_{x\to-\infty} f(x) = 0 \qquad \lim_{x\to\infty} f(x) = 0$$

- Tem dois pontos de inflexão que correspondem à media ± desvio padrão



Para uma perfeita compreensão da distribuição normal, observe a Figura 01 e procure visualizar as seguintes propriedades:

1ª) A variável aleatória X pode assumir todo e qualquer valor real.

2ª) A representação gráfica da distribuição normal é uma curva em forma de sino, simétrica em torno da média μ, que recebe o nome de **curva normal** ou de **Gauss**.

3ª) A área total limitada pela curva e pelo eixo das abcissas é igual a 1, já que essa área

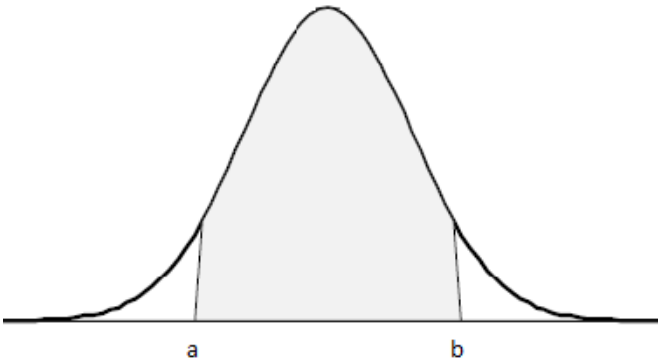
corresponde à probabilidade da variável aleatória X assumir qualquer valor real.

4ª) A curva normal é assintótica em relação ao eixo das abcissas, isto é, aproxima-se indefinidamente do eixo das abcissas sem, contudo, alcançá-lo.

5ª) Como a curva é simétrica em torno de μ, a probabilidade de ocorrer valor maior do que a média é igual à probabilidade de ocorrer valor menor do que a média, isto é, ambas as probabilidades são iguais a 0,5. Escrevemos: P(X > μ) = P(X < μ) = 0,5.

Quando temos em mãos uma variável aleatória com distribuição normal, nosso principal interesse é obter a probabilidade dessa variável aleatória assumir um valor em um determinado intervalo.

$$P(a < x < b) = \int_a^b \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$



Calcular esta integral toda vez, não seria fácil. A fim de ultrapassar este inconveniente, o Sr. Gauss (um dos estatísticos que inicialmente estudou esta função de distribuição) desenvolveu uma metodologia conducente à estandardização.

Distribuição Normal Reduzida ou Padronizada.

Em virtude da grande aplicação da distribuição normal, procurou-se tabelar os valores de probabilidade, que seria obtido por meio da integração da função densidade probabilidade normal num determinado intervalo. A dificuldade para se processar esse tabelamento se prendeu na infinidade de valores que μ e σ poderiam assumir. Nestas condições teria que se dispor de uma tabela para cada uma das infinitas combinações de μ e σ. Procurou-se, por isso, obter uma nova forma para a distribuição normal, que não sofresse a influência destes parâmetros (μ e σ). O problema foi solucionado mediante o emprego de uma nova variável Z definida por

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

que transforma todas as distribuições normais, em uma distribuição normal reduzida, ou padronizada, de media zero e desvio padrão um, Z∼N(0; 1). Assim, utilizamos apenas uma tabela para o calculo de probabilidades, para qualquer que seja a curva correspondente a uma distribuição normal. Desta forma, para um valor de x = μ numa distribuição normal qualquer, corresponde o

valor: z = 0, na distribuição normal reduzida. Para $x = \mu + \sigma$ tem-se z = 1, e assim por diante.

Exemplo de uso da tabela:

1) Calcular as seguintes probabilidades:

a) P(0 < Z < 1,28) b) P(Z>1,96) c) P(Z< 2,57)

d) P(Z< - 1,33) e) P(Z> -1,00) f) P(-1,00 < Z < 1,45)

g) P(1,00< Z < 1,96)

2) Determinar os valores de k para:

a) P(Z> k) = 0,0505 b) P(Z < k) = 0,9949 c) P(Z < k) = 0,0505

d) P(Z> K) = 0,8997 e) P(-1,00 < Z < k) = 0,6826

Exercícios:

1) Suponha que, em média, a empresa X transporte 10 t/dia de um determinado produto, com desvio padrão de 4 ton. Suponha que X tem distribuição normal. Qual a probabilidade que em um determinado dia a empresa transporte entre 8 e 11 ton.?

2) Suponha que o investimento em uma carteira de investimentos tenha distribuição normal com média de 100 mil reais e desvio padrão de 20 mil reais por dia.

a) Ao selecionar um dia aleatoriamente qual a probabilidade que o investimento: i) fique entre 70 e 110 mil reais? ii) seja superior a 105 mil? iii) seja inferior a 80 mil? iv) fique entre 110 e 120 mil?

b) Se o investimento for classificado em Baixo, Médio e Alto, de acordo com a seguinte regra:

Baixo ➔ 10% dos menores investimentos; Alto➔ 15% dos maiores investimentos, Médio➔ os demais 75%, quais serão os limites estimados para a classificação?

c) Se selecionarmos 1 dia, qual a probabilidade de que tenhamos investimentos superior a 130 mil?

3) A duração de certo tipo de pneu, em quilômetros rodados, e uma variável normal com duração media 60.000 km e desvio padrão 10.000 km.

a) Qual a probabilidade de um pneu aleatoriamente escolhido durar mais de 75.000 km?

b) Qual a probabilidade de um pneu aleatoriamente escolhido durar de 55.000 km?

7-Amostragem

Definições

População ou Universo: é o conjunto de indivíduos que apresentam uma ou mais características em comum (universo de estudo).

Amostra: é um subconjunto finito de uma população.

Amostragem: é o ato de tomar amostras da população.

Exemplos:

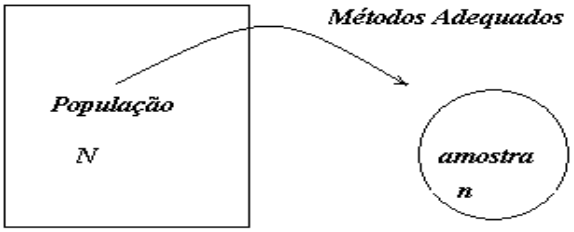
1) a escolha de passageiros para fazer a revista da bagagem;

2) o cozinheiro ao fazer a prova de um alimento;

3) a escolha de jogadores de futebol para fazer o exame de antidoping.

O objetivo da amostragem é determinar métodos para estudar as populações por meio de amostras. A amostragem nos possibilita concluir (inferir) sobre um todo a partir de apenas uma parte. Para isso é necessário sabermos como deve ser feito uma amostragem, ou seja, como coletar uma amostra.

Como selecionar uma amostra:



Representatividade: a amostra retirada dessa população tem por obrigação de preservar as características da população.

Por que amostrar?

i) Economia: menor custo;

ii) Rapidez: menor quantidade de trabalho, ou seja, menor tempo;

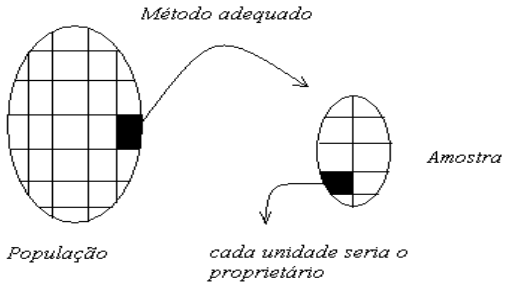
iii) Precisão: melhor qualidade no treinamento proporciona entrevistadores mais homogêneos possíveis, consequentemente maior precisão nos resultados.

Etapas num processo de amostragem

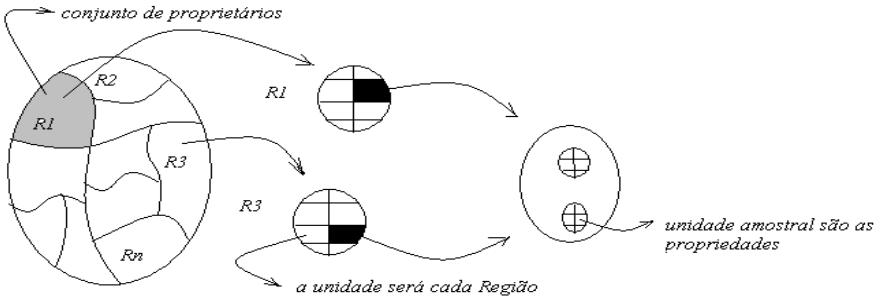
As etapas no planejamento e delineamento de uma pesquisa por amostragem são:

1) Objetivos da pesquisa: escrever ou estabelecer de forma clara os objetivos.
2) População que fornecerá a amostras: definição da população que se quer estudar.
3) Dados a serem coletados: decidir quais os dados serão coletados, ou seja, definir as variáveis baseando-se nos objetivos da pesquisa (atender ao item 1).
4) Definir o método de medição: entrevistador, a própria pessoa selecionada responde ao questionário, telefonemas, etc.
5) Grau de precisão desejado: dimensionar o tamanho da amostra, ou seja, definir *n*.
6) Listagem das unidades amostrais: escolha da unidade amostral.

Exemplo 1: População: Proprietários Rurais do Sul de Minas.



Exemplo 2: População: Proprietários Rurais do Sul de Minas.



7) Processo de amostragem mais adequado: a escolha da técnica de amostragem depende da característica da população.

8) Organização do trabalho de campo: questionário, estudo piloto (está ligado com a precisão).

9) Processamento e análise de dados: planejamento da tabulação dos dados e análise.

MÉTODOS AMOSTRAGEM

a) AMOSTRAGEM PROBABILÍSTICA:

Quando todos os elementos da população tiveram uma probabilidade conhecida e diferente de zero, de pertencer à amostra.

b) AMOSTRAGEM NÃO PROBABILÍSTICA:

Quando não se conhece a probabilidade de um elemento da população pertencer à amostra.

Principais Processos de Amostragem

Amostragem Simples ao Acaso ou Amostragem Aleatória Simples (ASA)

A ASA é o processo de amostragem mais simples e é utilizada quando se necessitam obter uma amostra representativa cujos elementos da população são todos homogêneos. Normalmente, este processo de amostragem é utilizado em associação com outros processos de amostragem, pois nem sempre é possível de forma imediata identificar todos os elementos da população como sendo homogêneos.

Propriedades importantes da ASA

i) qualquer amostra possível (*n*) tem igual chance de ser sorteada;

ii) cada elemento tem igual chance de pertencer à amostra.

A probabilidade de selecionar um indivíduo específico da população para uma amostra é 1/*N*.

A probabilidade de selecionar um indivíduo específico da população em *n* situações é *n*/*N*.

iii) o número de amostras possíveis de tamanho *n* que pode ser retirada de uma população de tamanho *N* é:

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

Exigências da ASA

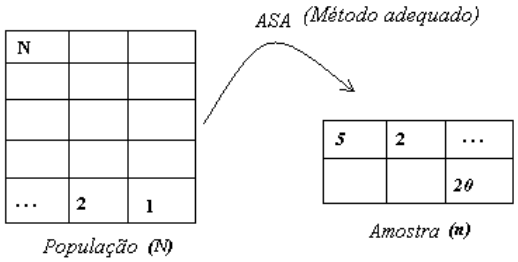
i) População finita;

ii) homogeneidade da população**.**

Procedimento para realizar a ASA

Enumeram-se todos os elementos da população (1, 2,..., *N*) e sorteiam-se *n* elementos mediante um dispositivo aleatório: computador, calculadora, tabela de números aleatórios, etc.

Esquema:



Exemplos de onde deve ou não aplicar a ASA:

1) Estudar a opinião de alunos de um determinado **curso** com relação a necessidade de acrescentar uma disciplina de física avançada na grade curricular; (ASA)
2) Estudar a opinião de alunos de uma determinada **universidade** com relação a necessidade de acrescentar uma disciplina de física avançada na grade curricular; (outro tipo de amostragem)

Amostragem Estratificada

O objetivo da amostragem estratificada é dividir a população heterogênea em subpopulações homogêneas (estratos), ou seja, na amostragem estratificada a população é dividida em grupos (estratos) mutuamente exclusivos e em seguida é feita a ASA em cada estrato. Suponha que uma população heterogênea seja dividida em *L* estratos com o objetivo de dividir a população heterogênea em *L* subpopulações homogêneos então, têm-se:

→*L* estratos de tamanho: $N_1, N_2, \dots, N_L$ e $N = N_1 + N_2 + \cdots + N_L = \sum_{h=1}^{L} N_h$.

→*L* amostras são retiradas (uma amostra de cada estrato): $n_1, n_2, \dots, n_L$ e $n = n_1 + n_2 + \cdots + n_L = \sum_{h=1}^{L} n_h$.

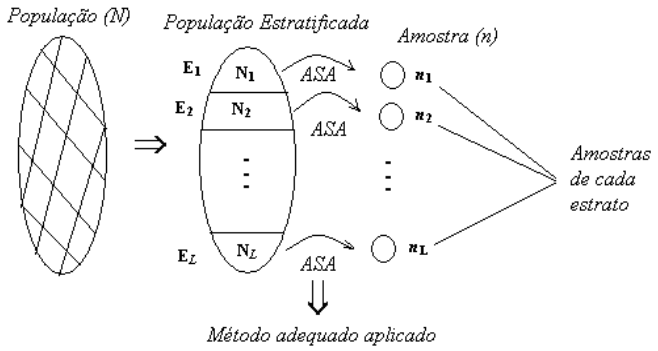
Portanto, para cada estrato é possível associarmos uma média e uma variância. A média e variância do estrato h são respectivamente:

$$\bar{X}_h = \frac{\sum_{i=1}^{n_h} x_{hi}}{n_h} \quad \text{e} \quad S_h^2 = \frac{1}{n_h - 1}\left[\sum_{i=1}^{n_h} X_{hi}^2 - \frac{\left(\sum_{i=1}^{n_h} x_{hi}\right)^2}{n_h}\right].$$

Procedimento para realizar a amostragem estratificada

A população é dividida em grupos ou estratos contendo elementos homogêneos e as amostras são retiradas separadamente de cada um desses grupos por meio da Amostragem Simples ao Acaso (ASA).

Esquema:



Classificação da amostragem estratificada quanto a retirada dos elementos em cada estrato

A amostragem estratificada pode ser classificada de acordo com a retirada dos elemento em cada estrato como Uniforme, Proporcional e Partilha Ótima.

10611 op 6 ou 8

Critério Uniforme

Para aplicar o critério Uniforme o tamanho das amostras de cada estrato deve ser o mesmo. Esse critério raramente é utilizado, devendo ser recomendado apenas para situações em que os estratos populacionais forem uniformes em tamanho, ou seja, os tamanhos dos estratos populacionais são iguais. Então, o tamanho da amostra de cada estrato pode ser obtido da seguinte forma:

$$n_h = \frac{n}{L}.$$

Em que: n_h é o tamanho da amostra retirada do estrato *h*, ou seja, o número de elementos que serão retirados do estrato *h*; *L* é a quantidade ou o número de estratos; *n* é o tamanho da amostra.

Recomendação: quando os tamanhos dos estratos são iguais $(N_1 = N_2 = N_3 = \cdots = N_h)$.

Critério Proporcional

No critério proporcional extrai-se de cada estrato uma quantidade de elementos n_h proporcional ao tamanho N_h do respectivo estrato. Este critério é recomendado quando o ***tamanho dos estratos são distintos e a variabilidade dos estratos é homogênea***, isto é:

Recomendação: $\begin{cases} N_1 \neq N_2 \neq N_3 \neq \cdots \neq N_h \\ \sigma_1 = \sigma_2 = \sigma_3 = \cdots = \sigma_h \Rightarrow \text{ homogeneidade de variância (homocedasticidade)} \end{cases}$

Para determinarmos a quantidade de elementos n_h que serão amostrados de cada estrato será utilizado a seguinte relação:

$$\frac{n_h}{n} = \frac{N_h}{\sum_{h=1}^{L} N_h} \text{ ou } \frac{n_h}{N_h} = \frac{n}{\sum_{h=1}^{L} N_h} \Rightarrow n_h = \frac{nN_h}{\sum_{h=1}^{L} N_h} \Rightarrow n_h = \frac{nN_h}{N}.$$

Em que: n_h é o tamanho da amostra do estrato *h*, ou seja, o número de elementos que serão retirados do estrato *h*; N_h é tamanho do estrato *h*; *L* é a quantidade de estrato; *n* o tamanho da amostra.

Exemplo: Suponha que uma empresa hoteleira deseja realizar uma pesquisa com os seus 84 funcionários, em que 25 pessoas são do sexo feminino (F) e as 59 restantes do sexo masculino (M). Estabelecendo $n = 9$ (10%

no mínimo), encontre o número de mulheres e de homens que deve ser entrevistados.

$$n_h = \frac{nN_h}{\sum_{h=1}^{L} N_h} = \frac{nN_h}{N_F + N_M} = \frac{nN_h}{N} \begin{cases} n_F = \dfrac{nN_F}{N} = \dfrac{9.25}{84} = 2{,}68 \Rightarrow n_F = 3 \text{ mulheres} \\ n_M = \dfrac{nN_M}{N} = \dfrac{9.59}{84} = 6{,}32 \Rightarrow n_M = 6 \text{ homens} \end{cases}$$

Critério Partilha Ótima (ou fração variável)

No critério partilha ótima extrai-se uma quantidade de elementos n_h proporcional ao tamanho N_h e ao desvio padrão σ_h do respectivo estrato. Este critério é recomendado quando os ***tamanhos dos estratos são distintos e a variabilidade dos estratos é heterogênea***, isto é:

Recomendação: $\begin{cases} N_1 \neq N_2 \neq N_3 \neq \cdots \neq N_h \\ \sigma_1 \neq \sigma_2 \neq \sigma_3 \neq \cdots \neq \sigma_h \Rightarrow \text{ heterogeneidade de variância (heterocedasticidade)} \end{cases}$

Para determinarmos a quantidade de elementos n_h que serão amostrados de cada estrato será utilizado a seguinte relação:

$$\frac{n_h}{n\sigma_h} = \frac{N_h}{\sum_{h=1}^{L} N_h\sigma_h} \quad \Rightarrow \quad \boldsymbol{n_h = \frac{N_h\sigma_h n}{\sum_{h=1}^{L} N_h\sigma_h}}.$$

Em que: n_h é o tamanho da amostra do estrato *h*, ou seja, o número de elementos que serão retirados do estrato *h*; N_h é tamanho do estrato *h*, ou seja, é o número de elementos do estrato *h*; σ_h é o desvio padrão do estrato *h*; *L* é a quantidade de estrato; *n* o tamanho da amostra.

Exemplo: Para ilustrar o procedimento da amostragem ótima considerou-se uma população fictícia de uma região, cujo interesse era obter informações sobre parâmetros de tecnologia dos produtores agrícolas da região.

Estratos (áreas em ha)	Nº de propriedades (N_h)	Desvio Padrão ($\boldsymbol{\sigma_h}$)	$\boldsymbol{N_h.\sigma_h}$	$\boldsymbol{n_h}$
0 |— 2	500	10	5.000	21
2 |— 5	320	11	3.520	15
5 |— 10	100	13	1.300	6
10 |— 20	50	20	1.000	4
20 |— 40	30	30	900	4
Totais	1.000	-	$\sum_{h=1}^{L=5} N_h\sigma_h = 11.720$	$n = 50$

Determine o tamanho amostral de cada estrato, ou seja, o número de propriedades que serão retiradas dos estratos para obtermos informações sobre parâmetros de tecnologia dos produtores agrícolas da região.

Nota-se que os tamanhos dos estratos são distintos e a variabilidade dos estratos é heterogênea então neste caso deve ser aplicado o critério de Partilha Ótima.

$$n_h = \frac{N_h \sigma_h n}{\sum_{h=1}^{L} N_h \sigma_h}$$

$$n_1 = \frac{N_1 \sigma_1 n}{\sum_{h=1}^{5} N_h \sigma_h} = \frac{500.10.50}{11720} = 21{,}33 = 21$$

$$n_2 = \frac{N_2 \sigma_2 n}{\sum_{h=1}^{5} N_h \sigma_h} = \frac{320.11.50}{11720} = 15{,}02 = 15$$

$$n_3 = \frac{N_3 \sigma_3 n}{\sum_{h=1}^{5} N_h \sigma_h} = \frac{100.13.50}{11720} = 5{,}54 = 6$$

$$n_4 = \frac{N_4 \sigma_4 n}{\sum_{h=1}^{5} N_h \sigma_h} = \frac{50.20.50}{11720} = 4{,}26 = 4$$

$$n_5 = \frac{N_5 \sigma_5 n}{\sum_{h=1}^{5} N_h \sigma_h} = \frac{30.30.50}{11720} = 3{,}84 = 4$$

Amostragem Sistemática

A amostragem sistemática é usada quando os elementos da população são heterogêneos e não podem ser agrupados em subpopulações homogêneas.

Procedimento para realizar a amostragem sistemática

Enumeram-se todos os elementos da população $(1,2,\ldots,N)$ e sorteia-se um primeiro elemento " i " para formar parte da amostra. Os demais são retirados em uma progressão aritmética, saltando " r " elementos, até completar o total da amostra (n). O valor " r " é chamado passos de amostragem e é determinado por:

$$\boldsymbol{r} = \frac{\boldsymbol{N}}{\boldsymbol{n}} \text{ elementos.}$$

O primeiro elemento deve ser sorteado entre os $\boldsymbol{r}$ primeiros

Esquema:

População enumerada: 1, 2,..., i,..., N.

A amostra sistemática será:

$$\begin{aligned}
&1^{o}\ elemento:\ i\\
&2^{o}\ elemento:\ i + r\\
&3^{o}\ elemento:\ i + 2r\\
&4^{o}\ elemento:\ i + 3r\\
&\quad\vdots\\
&n\text{-}ésimo\ elemento:\ i + (n-1)r
\end{aligned}$$

Exemplo: Um hotel mantém um arquivo contendo os registros de antigos hospedes, num total de 10.000 fichas das quais serão amostradas 1.000 fichas.

Vamos primeiramente determinar o valor "r " por intermédio de:

$$\boldsymbol{r} = \frac{\boldsymbol{N}}{\boldsymbol{n}} = \frac{\boldsymbol{10.000}}{\boldsymbol{1.000}} = \boldsymbol{10}.$$

Enumeram-se todos os elementos da população $(1,2,\ldots,10.000)$.

Sorteia-se um primeiro elemento, ou seja, a primeira ficha de hospede (um valor entre 1 e 10), por exemplo a ficha de número 5.

As fichas selecionadas serão:

$$1^{a}\ ficha:\ 5$$
$$2^{a}\ ficha:\ 5\ +\ 10=15$$
$$3^{a}\ ficha:\ 5\ +\ 2*10=25$$
$$\vdots$$
$$1.000^{a}\ ficha:\ 5\ +\ (1.000-1)*10=5+999*10=9.995$$

Amostragem por Conglomerados

Um Conglomerado é um subgrupo de elementos da população. O objetivo da amostragem por conglomerado é facilitar a coleta da informação. Cada conglomerado deve possuir a mesma heterogeneidade (mesmas características) que a população. Isto é, cada conglomerado deve representar bem toda a população.

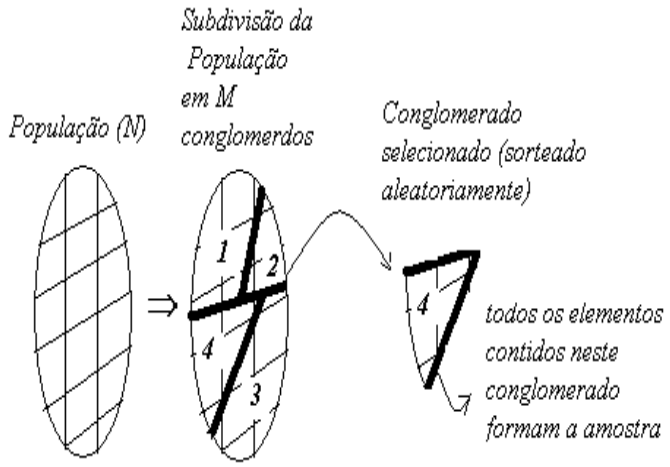
Motivação:

* Não tem a necessidade de cadastro de toda a população;
* Custo de locomoção (acesso aos elementos).

Procedimento para realizar a amostragem por conglomerado

Consiste em subdividir a população em conglomerados de elementos que reproduzam bem as características da população. Sorteia-se um determinado número de conglomerados. Todos os elementos contidos nos conglomerados selecionados vão compor a amostra.

Esquema:



Exemplo: Para estudar uma determinada característica da população de um país poderia sortear alguns estados, dentro de cada estado alguns municípios e dentro de cada município localidades.

População: População de um determinado país.

Conglomerados: “M” estados deste país em questão.

Amostra de conglomerados: “m” municípios.

Amostra: Bairros (localidades) dentro de cada município.

DISTRIBUIÇÕES AMOSTRAIS

Considerem-se todas as amostras possíveis de tamanho "n" que podem ser retiradas de uma população de tamanho "N" (com ou sem reposição). Para cada amostra pode-se calcular uma grandeza estatística, como a média, o desvio padrão etc., que varia de amostra para amostra. Com os valores obtidos para determinada grandeza, podemos construir uma distribuição de probabilidades, que será denominada de distribuição amostral. Para cada distribuição amostral é possível calcular a sua média, o seu desvio padrão, etc

Definições importantes:

- **- Parâmetro:** medida utilizada para descrever uma característica populacional. Ex: μ, σ
- **- Estimador:** é uma variável aleatória que é função dos dados amostrais. Ex: $\bar{X}$ é um estimador de μ
- **- Estimativa:** é o valor numérico assumido pelo estimador, quando são substituídos os dados amostrais. Ex: $\bar{X} = 170cm$
- **Inferência estatística**: objetivo de inferir propriedades de um agregado maior (a população) a partir de um
  conjunto menor (a amostra)

POPULAÇÃO x AMOSTRA

TIPO DE UTILIZACAO	POPULACAO	AMOSTRA
Variância	σ^2	s^2
Desvio padrão	σ	s
Media	μ	$\bar{x}$
Proporção	p	$\hat{p}$

Resumindo o processo:

- a) População com um parâmetro θ.
- b) Retira-se k amostras por um processo aleatório qualquer
- c) Calcula-se o valor $\hat{\theta}_i$ para cada amostra (1 = 1, 2, ... , k)
- d) Com os valores de $\hat{\theta}_i$ das k amostras constrói-se a distribuição amostral de θ.

DISTRIBUIÇÃO AMOSTRAL DAS MÉDIAS

Se os valores da média e do desvio padrão de uma população, de tamanho **N**, forem respectivamente **μ** e **σ,** e desta população são retiradas todas as possíveis amostras de tamanho **n**, sem reposição (população finita), os valores da média e do desvio padrão da distribuição amostral das médias correspondente serão:

$$E(\bar{X}) = \mu_{\bar{X}} = \mu \qquad \text{e} \qquad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

em que $\sqrt{\frac{N-n}{N-1}}$ é conhecido como fator de correção de população finita (populações pequenas).

Se a população for infinita (população grande), ou se amostragem for tomada com reposição, os valores acima ficarão:

$$E(\bar{X}) = \mu_{\bar{X}} = \mu \text{ e } \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

O fator de correção deve ser usado quando *n* exceder 5% do tamanho da população. Quando *N* é muito

maior em relação a *n*, a relação $\sqrt{\frac{N-n}{N-1}}$ tende a 1.

Exemplo 1: Numa urna tem-se 5 tiras de papel, numeradas 1, 3, 5, 5, 7. Uma tira é sorteada e recolocada na urna, então, uma segunda tira é sorteada. Sejam X_1 e X_2 o primeiro e o segundo números sorteados.

Amostra aleatória Simples

Pop{1, 3, 5, 5, 7} n = 2

X_1	X_2				$P(X_1=x)$
	1	3	5	7	
1	1/25	1/25	2/25	1/25	1/5
3	1/25	1/25	2/25	1/25	1/5
5	2/25	2/25	4/25	2/25	2/5
7	1/25	1/25	2/25	1/25	1/5
$P(X_2=x)$	1/5	1/5	2/5	1/5	1

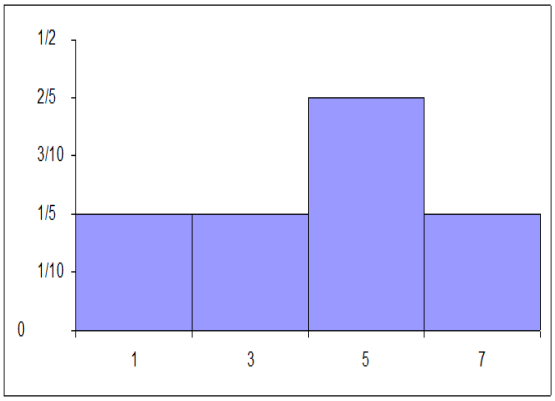
Distribuição amostral da média

$\bar{x}$	1	2	3	4	5	6	7	Total
$P(\bar{X} = \bar{x})$	1/25	2/25	5/25	6/25	6/25	4/25	1/25	1,00

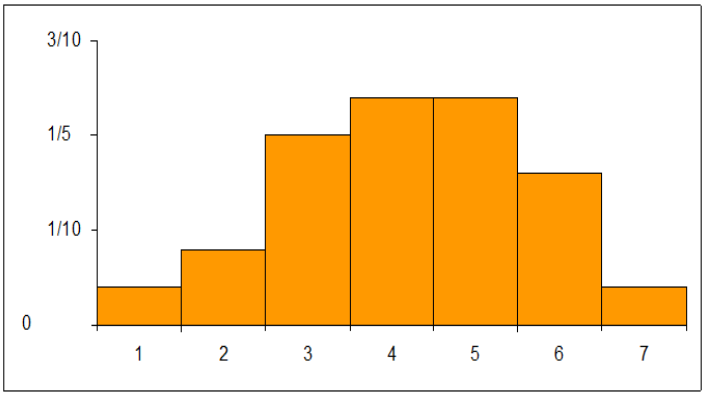
Distribuição amostral da variância

s^2	0	2	8	18	Total
$P(S^2 = s^2)$	7/25	10/25	6/25	2/25	1,00

Distribuição de X



Distribuição da media de X.



TEOREMA DO LIMITE CENTRAL

Por intermédio do Teorema do Limite Central, tem-se que quanto maior o tamanho da amostra, a distribuição de amostragem da média mais se aproxima da forma da distribuição normal, qualquer que seja a forma da distribuição da população. Na prática, a distribuição de amostragem da média pode se considerada como normal sempre que n ≥ 30.

Em síntese temos: "Se a variável aleatória X possui distribuição qualquer, com média μ e variância σ^2, a média amostral $(\bar{X})$, baseada em amostras aleatórias de tamanho n, possuirá distribuição normal aproximada com média das médias amostrais igual a média da população $(E(\bar{X}) = \mu_{\bar{X}} = \mu_X = \mu)$ e com a variância das médias amostrais igual a $\left(\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}\right)$.

Portanto, esse teorema permite aproximar a distribuição amostral de $\bar{X}$ por uma curva normal apropriada, independente da forma da distribuição da população.

Observações:

- Quanto maior o n (tamanho da amostra), melhor a aproximação normal.
- Se n ≥ 30 a aproximação normal é adequada, qualquer que seja a distribuição populacional. Amostragem sem reposição é recomendada quando (n/N > 0,05), então, deve-se fazer a correção para população finita e, portanto:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

Então, temos duas situações:

i) População Infinita: $\bar{X} \sim N\left(\mu_{\bar{X}} = \mu, \sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}\right)$

ii) População Finita: $\bar{X} \sim N\left(\mu_{\bar{X}} = \mu, \sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}\left(\frac{N-1}{N-n}\right)\right)$

Em função desses resultados temos:

i) $Z = \frac{\bar{X}-\mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$ se X tem distribuição Normal.

ii) $Z = \frac{\bar{X}-\mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \underset{a}{\sim} N(0,1)$ se X não tem distribuição Normal.

Exemplo 2: Seja XN (80, 26). Dessa população retiramos uma amostra de n=25. Calcular

a) $P(\bar{X} > 83)$ b) $P(\bar{X} < 82)$ c) $P(\bar{X} - 2\sigma_{\bar{X}} \leq \mu \leq \bar{X} - 2\sigma_{\bar{X}})$

Exemplo 3: Seja X: N (100, 85). Retiramos uma amostra de n=20. Determinar

a) $P(95 < \bar{X} < 105)$ b) $P(\bar{X} - Z_{\alpha}\sigma_{\bar{X}} \leq \mu \leq \bar{X} + Z_{\alpha}\sigma_{\bar{X}}) = 0{,}95$

Exemplo 3: Sabe-se que a média de tempo que candidatos a um determinado emprego gastam para responder um teste psicológico é de 30 minutos, com desvio padrão de 10 minutos.

a) Se selecionarmos um indivíduo qualquer dessa população, qual a probabilidade que ele gaste entre 25 e 35 minutos para responder ao teste? (revisão de distribuição normal). (0,383)

b) Se selecionarmos um grupo de 36 indivíduos dessa população, qual a probabilidade que a média do tempo gasto pelo grupo seja superior a 32 minutos?

DISTRIBUICAO AMOSTRAL DA MEDIA E DA DIFERENCA ENTRE MEDIAS

Distribuição Amostral de t (Student).

Sabe-se que $\bar{X} \sim N\left(\mu; \frac{\sigma^2}{n}\right)$, e sua distribuição padronizada é dada por: $Z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$

Em muitas situações não se conhece σ^2 ou σ, mas sim sua estimativa s^2 ou s

Precisamos substituir σ por seu estimador s na estatística $t = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}}$ a qual segue uma distribuição t de Student com (n-1) graus de liberdade. Esta estatística é utilizada quando se tem amostras pequenas $(n \leq 30)$, pois o valor de s^2 torna-se muito variável, ou seja, flutua muito de amostra para amostra. Nestas situações a distribuição deixa de ser normal padronizada.

Características da distribuição t

a) É simétrica em relação a média (semelhante a distribuição de z)

b) Tem forma campanular. Valores de t dependem da flutuação das estatísticas média e desvio padrão amostrais e z depende somente das mudanças da média das amostras

c) Quando n tende para infinito, a distribuição t tende para a distribuição normal. Na prática, a aproximação é considerada boa quando n >30.

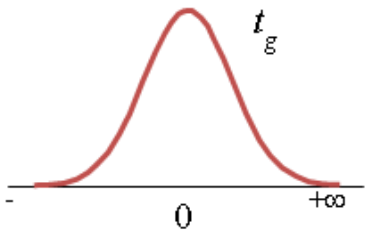
d) Possui n-1 graus de liberdade**.**

Condições para utilizar a distribuição de t de Student

a) O tamanho da amostra é pequeno $(n \leq 30)$

b) σ é desconhecido

c) A população tem distribuição essencialmente normal



Se $X_i \sim N(\mu, \sigma^2)$ e $Z = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$ então $t_{n-1} \sim \frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}}$

Exemplo: Obter os seguintes valores da distribuição t de Student

a) $P(T_{10} > 2{,}764)$
b) t / P (-t < t < t) = 0,95 com 10 g.l.
c) t / P (-t < t < t) = 0,90 com 20 g.l.
d) t / P (t > t) = 0,05 com 25 g.l.
e) t / P (t < t) = 0,10 com 10 g.l.
f) P (-1,753 < t < 1,753) com 15 g.l.

DISTRIBUICAO AMOSTRAL DA VARIANCIA E DA RELACAO ENTRE VARIANCIAS.

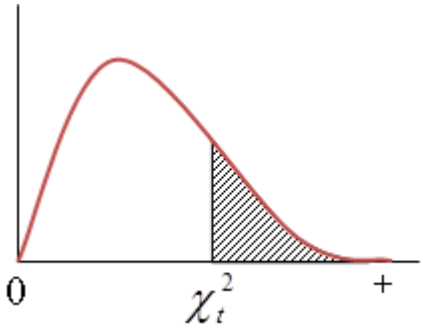
Distribuição Amostral de χ^2

É uma distribuição amostral de variâncias

Retira-se uma amostra de n elementos de uma população normal com média μ e variância σ^2, teremos a distribuição de uma $s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$, segue uma distribuição de χ2 com n-1 graus liberdade.

A variável $\chi^2 = \frac{(n-1)s^2}{\sigma^2}$ tem distribuição χ^2 com n-1 graus de liberdade.

- ✓ Os valores de χ^2 não podem ser negativos
- ✓ Não é simétrica em $\chi^2 = 0$
- ✓ Quanto maior o tamanho de n, a distribuição tende a normal.
- ✓ Como a curva não é simétrica, então se olha na tabela dois valores de χ^2, quando queremos saber se um valor está entre 2 limites**.**



Exemplo: Obter os seguintes valores da distribuição de χ^2 :

a) χ^2 / P ($\chi^2 > \chi^2$) = 0,025 com 17 g.l.

b) χ^2 / P ($\chi^2 < \chi^2$) = 0,025 com 17 g.l.

c) χ_1^2, χ_2^2 / P ($\chi_1^2 < \chi^2 < \chi_2^2$) = 0,90 com 10 g.l.

d) χ_1^2, χ_2^2 / P ($\chi_1^2 < \chi^2 < \chi_2^2$) = 0,95 com 15 g.l.

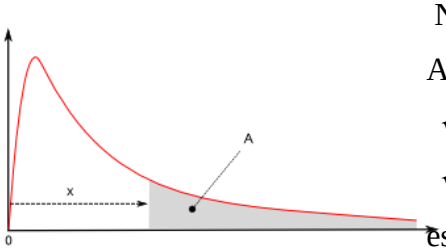
e) P (10,8508 < χ^2 < 31,4104) com 20 g.l.

Distribuição amostral de *F* (*de Snedecor*)

A distribuição F de Fisher-Snedecor, mais conhecida como distribuição F de Fisher (em honra a Ronald Fisher) ou distribuição F de Snedecor (em honra a George W. Snedecor) mede a razão entre duas distribuições qui-quadrado independentes.

Se $\frac{(n_1-1)s_1^2}{\sigma_1^2} \sim \chi^2_{n_1-1}$ e $\frac{(n_2-1)s_2^2}{\sigma_2^2} \sim \chi^2_{n_2-1}$ então $\frac{s_1^2}{\sigma_1^2}\frac{\sigma_2^2}{s_2^2} = \frac{s_1^2\sigma_2^2}{s_2^2\sigma_1^2} \sim F_{n_1-1,n_2-1}$

Tabelas da distribuição F



Na tabela abaixo, valores de x conforme Figura 01 para o valor de A.

ν_1: graus de liberdade da variável do numerador (primeira linha).

ν_2: graus de liberdade da variável do denominador (primeira coluna à esquerda da tabela).

Exemplo: Obter os seguintes valores da distribuição F de Snedecor:

a) F / P(F > F) = 0,10 com $v_1 = 8$ e $v_2 = 20$ g.l.

b) F / P(F < F) = 0,90 com $v_1 = 8$ e $v_2 = 20$ g.l.

c) F_1,F_2 / P(F_1 < F < F_2) = 0,95 com $v_1 = 10$ e $v_2 = 20$ g.l.

d) F / P(F > 1,84) com $v_1 = 10$ e $v_2 = 10$ g.l.

DISTRIBUIÇÃO AMOSTRAL DAS PROPORÇÕES

Se o valor da proporção de ocorrência de um evento em uma população, de tamanho **N**, for ***p,*** e desta população são retiradas todas as possíveis amostras de tamanho **n**, sem reposição, os valores da média e do desvio padrão da distribuição amostral das proporções correspondente serão:

$$E(\hat{p}) = p \text{ e } \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}\sqrt{\frac{N-n}{N-1}}$$

Se a população for infinita, ou se amostragem for tomada com reposição, os valores acima ficarão:

$$E(\hat{p}) = \mu_{\hat{p}} = p \text{ e } \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

Quando $n \to \infty$, a distribuição amostral de $\hat{p}$será aproximadamente Normal com média p e variância $p(1 - p)/n$, ou seja, $\hat{p} \underset{a}{\sim} N\left(p, \frac{p(1-p)}{n}\right)$.

Consequentemente, $\frac{\hat{p}-\mu_{\hat{p}}}{\sigma_{\hat{p}}} \underset{a}{\sim} N(0,1)$, ou seja, $Z = \frac{\hat{p}-p}{\sqrt{\frac{pq}{n}}} \underset{a}{\sim} N(0,1)$.

Quando p é desconhecida e a amostra é suficientemente grande, determinamos $\hat{p}_0 = \frac{X}{n}$, estimativa de p. Então, $\sigma_{\hat{p}} = \sqrt{\frac{\hat{p}_0(1-\hat{p}_0)}{n}}$.

8 - TEORIA DA ESTIMAÇÃO

INTRODUÇÃO

Antes de abordar a teoria da estimação vamos procurar entender o que vem a ser *estimador* e *estimativa*.

Um estimador, $\hat{\theta}$, do parâmetro θ é uma função qualquer dos elementos da *amostra*. Estimativa é o valor numérico assumido pelo estimador quando os valores observados são considerados.

Assim,

$\bar{X} = \frac{\sum_i^n X_i}{n}$, é um estimador da média populacional μ, e $\bar{X} = 150$ cm é uma estimativa da média populacional.

ESTIMAÇÃO POR PONTO E POR INTERVALO

Estimação Pontual

Quando a estimativa de um parâmetro populacional é dado por um único valor, tem-se uma estimativa pontual do parâmetro populacional, ou seja, na estimação pontual é fornecido como estimativa do parâmetro, apenas um valor numérico. Por exemplo, ao estimar μ (média populacional) podemos fazê-lo mediante o valor $\bar{X} = 20$ kg. Então, 20 kg é uma estimativa pontual para μ.

Desvantagem

A estimativa pontual não fornece nenhuma idéia de quão próximo é o valor dessa estimativa em relação ao valor do parâmetro. Uma maneira de se salvar essa desvantagem é usando estimadores por intervalo.

Estimação por Intervalo (Intervalos de Confiança)

A estimação por intervalo foi idealizada para procurar suprir essa desvantagem, ou seja, um intervalo é construído, de tal maneira que se possa atribuir probabilidades de que o valor real do parâmetro θ esteja ali contido. Na estimação por intervalo determina-se um intervalo baseando-se na distribuição amostral do estimador, considerando uma elevada probabilidade de conter o verdadeiro valor do parâmetro populacional desconhecido.

De modo geral as estimativas (pontual ou intervalar) devem ser bastante confiáveis, e para isso é necessário que os estimadores que as fornecerão apresentem boas propriedades, aliado ao fato de serem obtidas a partir de amostras representativas. A seguir serão apresentadas as propriedades de um bom estimador.

Propriedades dos Estimadores

1) *<u>Não tendenciosidade</u>*

Um estimador $\hat{\theta}$ é dito um estimador não tendencioso do parâmetro θse $E(\hat{\theta}) = \theta$.

Exemplo: $\bar{X} = \frac{\sum_i^n x_i}{n}$ é um estimador não tendencioso da média populacional μ.

Prova

$$E(\bar{X}) = E\left(\sum_i^n X_i \Big/ n\right)$$

$$E(\bar{X}) = E\left(\frac{1}{n}\sum_i^n X_i\right)$$

$$E(\bar{X}) = \frac{1}{n}E\left(\sum_i^n X_i\right)$$

$$E(\bar{X}) = \frac{1}{n}E(X_1 + X_2 + \text{L} + X_n)$$

$$E(\bar{X}) = \frac{1}{n}\{E(X_1) + E(X_2) + \text{L} + E(X_n)\}$$

$$E(\bar{X}) = \frac{1}{n}[m + m + \text{L} + m] = \frac{1}{n}nm$$

$$E(\bar{X}) = m$$

2) *Coerência ou Consistência*

Um estimador $\hat{\theta}$ é dito um estimador consistente do parâmetro θse:

i) $\lim_{n\to\infty} E(\hat{\theta}) = \theta$;

ii) $\lim_{n\to\infty} V(\hat{\theta}) = 0$.

Exemplo: $\bar{X} = \frac{\sum_i^n x_i}{n}$

i) $\lim_{n\to\infty} E(\bar{X}) = \lim_{n\to\infty} \mu = \mu$;

ii) $\lim_{n\to\infty} V(\bar{X}) = \lim_{n\to\infty} \frac{\sigma^2}{n} = 0$.

3) *Precisão ou Eficiência*

Se $\hat{\theta}_1$ e $\hat{\theta}_2$ são dois estimadores não tendenciosos de θ, então, $\hat{\theta}_1$ é mais eficiente que $\hat{\theta}_2$ se:

$$V(\hat{\theta}_1) < V(\hat{\theta}_2).$$

3.1) *Eficiência Relativa*: A eficiência relativa do estimador $\hat{\theta}_1$, em relação ao estimador $\hat{\theta}_2$ é dada por:

$$Ef_{\hat{\theta}_1,\hat{\theta}_2} = \frac{V(\hat{\theta}_2)}{V(\hat{\theta}_1)}.$$

Observação:

Quanto menor for a variância de um estimador maior será a sua eficiência.

Se $\hat{\theta}_1$ for menos eficiente que $\hat{\theta}_2$ então $Ef_{\hat{\theta}_1,\hat{\theta}_2} < 1$;

Se $\hat{\theta}_1$ for mais eficiente que $\hat{\theta}_2$ então $Ef_{\hat{\theta}_1,\hat{\theta}_2} > 1$.

Suficiência ou Precisão

Um estimador é suficiente se contém o máximo de informação com relação ao parâmetro por ele estimado.

$$\text{Quantidade de informação ou precisão} = \frac{1}{V(\hat{\theta})}.$$

MÉTODOS DE ESTIMAÇÃO

Na teoria de estimação os métodos de estimação são: Métodos dos Momentos; Métodos dos Quadrados Mínimos e Método da Máxima Verossimilhança. No presente material não será apresentada a metodologia dos métodos, pois este não é o objetivo da disciplina.

ESTIMAÇÃO POR INTERVALO

A estimação pontual não fornece idéia da margem de erro que é cometida ao se estimar um determinado parâmetro. A estimação por intervalo procura corrigir essa lacuna a partir da criação de um intervalo que garanta com alta probabilidade de conter o verdadeiro valor do parâmetro desconhecido.

Um conceito importante para a elaboração de intervalos de confiança é o de ***quantidade pivotal***. Seja $x' = [X_1, \ldots, X_n]$ uma amostra aleatória de densidade $f(.)$. Uma função $W(x, \theta)$, cuja distribuição não dependa de θ, é chamada de quantidade pivotal.

Um exemplo conhecido é a quantidade:

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}},$$

para $f(.)$ correspondente a uma normal $N(\mu, \sigma^2)$. Nesse caso, a distribuição de Z é uma normal $N(0,1)$, não depende de μ e σ^2.

Construção dos Intervalos de Confiança

Nesta seção será detalhada a metodologia utilizada para a obtenção de um Intervalo de Confiança.

Sabe-se que a função pivô Z tem distribuição normal padrão, ou seja, distribuição normal com média zero e variância 1. A distribuição de Z não depende da quantidade desconhecida μ, sendo possível obter os quantis inferior e superior $\frac{\alpha}{2\left(-Z_{\frac{\alpha}{2}} \text{ e } Z_{\frac{\alpha}{2}}\right)}$.

Observação:

Toda afirmação deve vir acompanhada de um grau de confiança, ou grau de certeza, ou seja, quanto se está certo ao comunicar aquela informação.

O nível ou grau de confiança é denotado por (1-α), onde α (alfa) é o nível de significância.

O estimador por intervalo para a média μ tem a forma:

$$\left[\bar{X} - \varepsilon \;;\; \bar{X} + \varepsilon\right]$$

A afirmativa probabilística seguinte, a definição Z anterior e os quantis inferior e superior da distribuição $N(0,1)$ permitem que se construa a regra de estimação de μ por intervalo. Assim,

$$P\left(-Z_{\frac{\alpha}{2}} \leq Z \leq Z_{\frac{\alpha}{2}}\right) = 1-\alpha.$$

Substituindo a definição de Z nessa expressão acima e isolando μ, obtém-se:

$$P\left(-Z_{\frac{\alpha}{2}} \leq Z \leq Z_{\frac{\alpha}{2}}\right) = 1-\alpha$$

$$P\left(-Z_{\frac{\alpha}{2}} \leq \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq Z_{\frac{\alpha}{2}}\right) = 1-\alpha$$

$$P\left(-Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \leq \bar{X}-\mu \leq Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

$$P\left(-\bar{X}-Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \leq -\mu \leq -\bar{X}+Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

$$P\left(\bar{X}+Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \geq \mu \geq \bar{X}-Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

$$P\left(\bar{X}-Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}+Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

E o intervalo de confiança para μ, com uma confiança de $1-\alpha$ pode então ser escrito como:

$$IC_{100(1-\alpha)\%}(\mu) : \bar{X} \pm Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$$ (variância conhecida).

Isto significa que o parâmetro μ apresenta uma probabilidade de $1-\alpha$ de estar entre os limites:

$$\bar{X}-Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \text{ e } \bar{X}+Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}.$$

INTERVALOS DE CONFIANÇA PARA MÉDIA

Intervalo de Confiança para μ de uma população Normal com ***Variância Populacional*** (σ^2) ***conhecida***:

i) Se ocorrer Amostragem com Reposição para População Finita (P.F.A.C.R.) ou para População Infinita (P.I.):

$$IC_{100(1-\alpha)\%}(\mu) : \bar{X} \pm Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \quad \text{ou} \quad \boxed{IC_{100(1-\alpha)\%}(\mu) : \left[\bar{X}-Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}; \bar{X}+Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]}$$

Interpretação: Existe 100(1-α)% de confiança que μ esteja contida no intervalo.

Observações:

a) 100(1-α)% é chamado de nível de confiança para o intervalo; $\bar{X}$ é a média amostral; *n* é o tamanho da amostra e α é o nível de significância.

b) $\bar{X} - Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$ e $\bar{X} + Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$, são chamados de limite inferior e superior de confiança, ou seja, limites inferior e superior do intervalo de confiança.

Nota: A estimativa por intervalo nos dá idéia do erro que podemos estar cometendo na estimação. Essa idéia de erro é dada em termos probabilísticos.

ii) Se ocorrer Amostragem Sem Reposição para População Finita (P.F.A.S.R.):

$$IC(\mu)_{100(1-\alpha)\%} : \bar{X} \pm Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \text{ ou } \boxed{IC(\mu)_{100(1-\alpha)\%}: \left[\bar{X} - Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}; \bar{X} + Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \right]}$$

Observações:

a) 100(1-α)% é chamado de nível de confiança para o intervalo; $\bar{X}$ é a média amostral; *n* é o tamanho da amostra; *N* é o tamanho da população, α é o nível de significância, $\sqrt{\frac{N-n}{N-1}}$ é o fator de correção.

b) $\bar{X} - Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$ e $\bar{X} + Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$, são chamados de limite inferior e superior de confiança, ou seja, limites inferior e superior do intervalo de confiança.

c) se $\frac{n}{N} > 0{,}05 \Rightarrow$ **P.F.A.S.R. (Intervalo de Confiança com o fator de correção).**

Exemplo 1: Uma máquina produz rolamentos que apresentam desvio padrão de 0,042 polegadas em seu diâmetro. Desejando-se conhecer o diâmetro médio dos rolamentos produzidos por esta máquina extraiu-se uma amostra de 100 rolamentos, observando-se uma média igual a 0,824 polegadas.

a) Obter o intervalo de confiança com 0,90 de confiança para o verdadeiro diâmetro médio dos rolamentos.

Solução: Primeiramente será retirada toda informação (dados) do problema. Tem-se: $\bar{X} = 0824$ polegadas, $\sigma = 0{,}042$ polegadas, $n = 100$ e $\alpha = 10{,}0\% = 0{,}10$.

Como nada foi informado a respeito do tamanho da população (*N*), será adotado o seguinte intervalo de confiança:

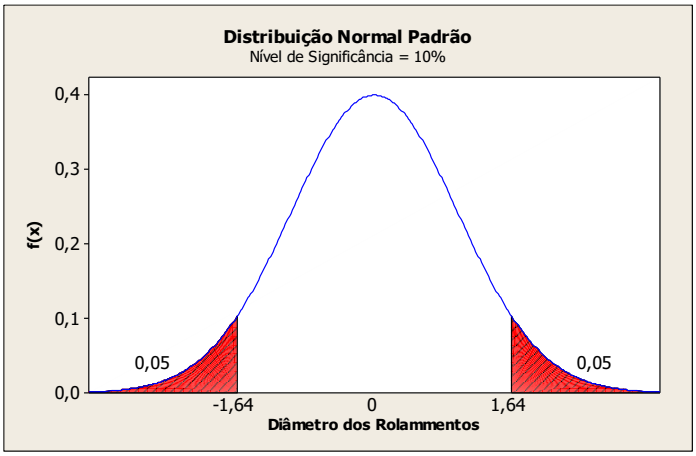


Figura 1 – Gráfico referente aos quantis inferior e superior $\frac{\alpha}{2}$ $\left(-Z_{\frac{\alpha}{2}} \text{ e } Z_{\frac{\alpha}{2}}\right)$, ao nível de significância de 10%.

$$IC(\mu)_{100(1-\alpha)\%}: \bar{X} \pm Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

$$IC(\mu)_{90,0\%}: 0,824 \pm Z_{\frac{0,10}{2}} \frac{0,042}{\sqrt{100}}$$

$$IC(\mu)_{90,0\%}: 0,824 \pm Z_{0,05} \frac{0,042}{\sqrt{100}}$$

$$IC(\mu)_{90,0\%}: 0,824 \pm 1,64 \frac{0,042}{\sqrt{100}}$$

$$IC(\mu)_{90,0\%}: 0,824 \pm 0,007$$

$$IC(\mu)_{90,0\%}: [0,817;\ 0,831]$$

Interpretação: Pode-se afirmar com 90,0% de confiança que diâmetro médio dos rolamentos (μ) produzidos por esta máquina está contido no intervalo [0,817; 0,831].

b) (Exercício) Obter o intervalo de com 0,95 de confiança para o verdadeiro diâmetro médio dos rolamentos.

Interpretação (Exercício):

c) (Exercício) Obter o intervalo de com 0,95 de confiança para o verdadeiro diâmetro médio dos rolamentos.

Interpretação (Exercício):

Exercício: Utilizando os dados do exercício anterior e supondo que a produção diária seja de 1.000 rolamentos.

a) Obter o intervalo de com 0,90 de confiança para o verdadeiro diâmetro médio dos rolamentos.

Interpretação (Exercício):

b) (Exercício) Obter o intervalo de com 0,95 de confiança para o verdadeiro diâmetro médio dos rolamentos.

Interpretação (Exercício):

c) (Exercício) Obter o intervalo de com 0,99 de confiança para o verdadeiro diâmetro médio dos rolamentos.

Interpretação (Exercício):

Exemplo 3: De uma população normal a variável aleatória X apresenta uma variância igual a 9. Retiramos uma amostra de 25 observações e obteve-se $\sum_{i=1}^{25} X_i = 152$.

a) Determine um intervalo de limites de 90,0% de confiança para μ.

Primeiramente será retirada toda informação (dados) do problema. Tem-se:

$$n = 25;\ \alpha = 10{,}0\% = 0{,}10;\ \sum_{i=1}^{25} X_i = 152 \Rightarrow \bar{X} = \frac{152}{25} = 6{,}08 \text{ e } \sigma^2 = 9 \Rightarrow \sigma = 3$$

Como nada foi informado a respeito do tamanho da população (*N*), será adotado o seguinte intervalo de confiança:

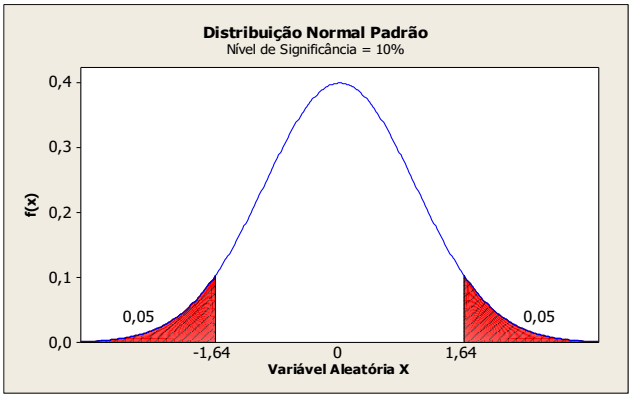


Figura 2 – Gráfico referente aos quantis inferior e superior $\frac{\alpha}{2}$ $\left(-Z_{\frac{\alpha}{2}} \text{ e } Z_{\frac{\alpha}{2}}\right)$, ao nível de significância de 10%.

$$IC_{90,0\%}(\mu): 6{,}08 \pm Z_{0,10/2}\frac{3}{\sqrt{25}}$$

$$IC_{90,0\%}(\mu): 6{,}08 \pm Z_{0,05}\frac{3}{\sqrt{25}}$$

$$IC_{90,0\%}(\mu): 6{,}08 \pm 1{,}64\frac{3}{\sqrt{25}}$$

$$IC_{90,0\%}(\mu): 6{,}08 \pm 0{,}99$$

$$IC_{90,0\%}(\mu): [5{,}096;\ \ 7{,}064]$$

Interpretação: Estatisticamente podemos afirmar com 90,0% de confiança que a média populacional μ da variável aleatória X se encontra no intervalo [5,096; 7,064].

b) (Exercício) Determine um intervalo de limites de 95,0% de confiança para μ.

Interpretação (Exercício):

c) (Exercício) Determine um intervalo de limites de 99,0% de confiança para μ.

Interpretação (Exercício):

Intervalo de Confiança para μ de uma população Normal com ***Variância Populacional*** (σ^2) desconhecida:

Como S^2 (variância amostral) é o estimador de $\sigma^2 \Rightarrow$ substituir σ^2 por S^2 ou σ por S.

i) **Amostras pequenas ($n \leq 30$):** usa-se distribuição *t*.

Para P.I. ou P.F.A.C.R.:

$$IC_{100(1-\alpha)\%}(\mu): \left[\bar{X} - t_{(n-1,\alpha/2)}\frac{S}{\sqrt{n}}; \bar{X} + t_{(n-1,\alpha/2)}\frac{S}{\sqrt{n}}\right]$$

Para P.F.A.S.R.:

$$IC_{100(1-\alpha)\%}(\mu): \left[\bar{X} - t_{(n-1,\alpha/2)} \frac{S}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}; \bar{X} + t_{(n-1,\alpha/2)} \frac{S}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \right]$$

ii) Amostras grandes ($n > 30$): usa-se distribuição *Z*.

Para P.I. ou P.F.A.C.R.:

$$IC_{100(1-\alpha)\%}(\mu): \left[\bar{X} - Z_{\alpha/2} \frac{S}{\sqrt{n}}; \bar{X} + Z_{\alpha/2} \frac{S}{\sqrt{n}} \right]$$

Para P.F.A.S.R.:

$$IC_{100(1-\alpha)\%}(\mu): \left[\bar{X} - Z_{\alpha/2} \frac{S}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}; \bar{X} + Z_{\alpha/2} \frac{S}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \right]$$

Observação: À medida que se aumenta o tamanho da amostra, a distribuição *t* de Student se aproxima da distribuição Normal, deste modo, quando se estiver trabalhando com amostras grandes ($n > 30$) pode-se utilizar a distribuição padronizada *Z*, em lugar da *t* na obtenção dos intervalos de confiança, mesmo que σ^2 seja desconhecida.

Exemplo: A altura nos homens de uma cidade apresenta distribuição normal, para se estimar a altura média dessa população levantou-se uma ***amostra*** de 150 indivíduos obtendo-se: $\sum_{i=1}^{150} X_i = 25.800$ cm e $\sum_{i=1}^{150} X_i^2 = 4.440.075$ cm^2.

a) Ao nível de 2% de significância, determine o intervalo de confiança para a altura média dos homens desta cidade.

Solução:

Como se trata de uma amostra, a variância que será determinada corresponde a variância amostral, ou seja, a variância populacional (σ^2) é desconhecida. Sabe-se que:

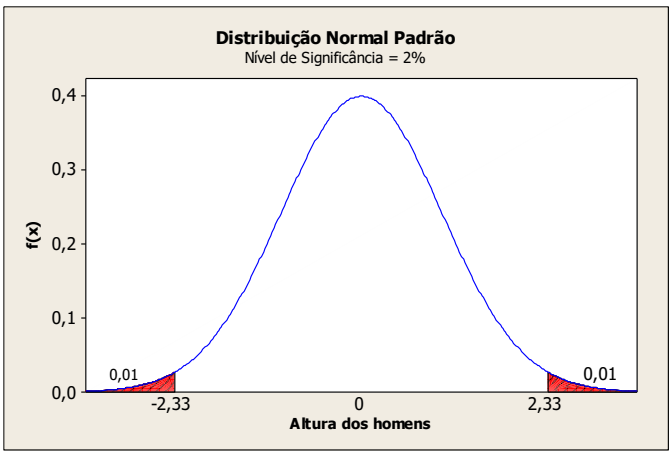
$$\sum_{i=1}^{150} X_i = 25.800 \text{ cm e } \sum_{i=1}^{150} X_i^2 = 4.440.075 \text{ cm}^2,$$

então a média e variância são respectivamente:

$$\bar{X} = \frac{\sum_{i=1}^{150} X_i}{150} = \frac{25.800}{150} = 172 \text{ cm.} S = \sqrt{\frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 - \frac{(\sum_{i=1}^{n} X_i)^2}{n}\right]}$$

$$= \sqrt{\frac{1}{149}\left[4.440.075 - \frac{(25.800)^2}{150}\right]} = \sqrt{16,61} = 4,07 \text{ cm.}$$

A amostra é de 150 indivíduos, isto é, $n = 150 > 30$ (amostra grande). Logo, o intervalo a ser adotado para determinar altura média dos homens desta cidade será:



$$IC_{100(1-\alpha)\%}(\mu): \bar{X} \pm Z_{\alpha/2}\frac{S}{\sqrt{n}}$$

$$IC_{98,0\%}(\mu): 172 \pm Z_{0,02/2}\frac{4,07}{\sqrt{150}}$$

$$IC_{98,0\%}(\mu): 172 \pm Z_{0,01}\frac{4,07}{\sqrt{150}}$$

$$IC_{98,0\%}(\mu): 172 \pm 2,33\frac{4,07}{\sqrt{150}}$$

$$IC_{98,0\%}(\mu): [171,22;\ \ 172,77]$$

Figura 3 – Gráfico referente aos quantis inferior e superior $\frac{\alpha}{2}$ $\left(-Z_{\frac{\alpha}{2}} \text{ e } Z_{\frac{\alpha}{2}}\right)$, ao nível de significância de 2%.

Interpretação: Pode-se afirmar com 98,0% de confiança que a estatura média dos homens desta cidade está contida no intervalo [171,22; 17277].

b) (Exercício) Ao nível de 5% de significância, determine o intervalo de confiança para a altura média dos homens desta cidade.

Interpretação (Exercício):

c) (Exercício) Ao nível de 1% de significância, determine o intervalo de confiança para a altura média dos homens desta cidade.

Interpretação (Exercício):

Exercício: Uma Cia adquiriu 500 cabos. Uma amostra de 30 deles ao acaso apresentou tensão de ruptura média igual a 2.400 kg com desvio padrão de 150 kg.

a) Obter o intervalo com 99% de confiança para a verdadeira tensão média de ruptura desses cabos.

Interpretação (Exercício):

b) Obter o intervalo com 95% de confiança para a verdadeira tensão média de ruptura desses cabos.

Interpretação (Exercício):

c) Obter o intervalo com 90% de confiança para a verdadeira tensão média de ruptura desses cabos.

Interpretação (Exercício):

Dimensionamento do tamanho da amostra

Nosso objetivo agora será determinar o tamanho da amostra *n*, de tal forma que o estimador obtido tenha um erro máximo de estimação igual a ε, com determinado grau de confiança (probabilidade). A precisão do intervalo, $IC(\mu)_{100(1-\alpha)\%}: \bar{X} \pm Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$, é $Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$. Isso significa que usando $\bar{X}$ para estimar μ, o erro $\varepsilon = |\bar{X} - \mu|$ é menor ou igual a $Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$, com 100(1-α)% de confiança. De maneira mais específica, o problema consiste em determinarmos *n*, de modo que

$$P[|\bar{X} - \mu| \leq \varepsilon] \cong 1 - \alpha,$$

isto é,

$$P\left[|\bar{X} - \mu| \leq Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right] \simeq 1 - \alpha.$$

Então, para um E fixo, a solução para o problema acima consiste em determinar *n* de tal forma que

$$\varepsilon = Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$$

ou equivalentemente,

$$\frac{\varepsilon^2}{Z^2_{\frac{\alpha}{2}}} = \frac{\sigma^2}{n}. \qquad (1)$$

Resolvendo (1) em *n*, obtém-se:

$$n = \frac{Z^2_{\frac{\alpha}{2}}\sigma^2}{\varepsilon^2} = \left(\frac{Z_{\frac{\alpha}{2}}\sigma}{\varepsilon}\right)^2. \qquad (2)$$

Para determinação da amostra, é preciso fixar o erro máximo desejado (*E*), com algum grau de confiança $1 - \alpha$ (traduzido pelo valor tabelado $Z_{\frac{\alpha}{2}}$) e possuir algum conhecimento a priori da variabilidade da população (σ^2). Ou seja, o erro máximo desejado e o nível de confiança são fixados pelo pesquisador. O uso de pesquisa passadas, estatísticas (informações), ou amostras piloto são os critérios mais usados. Em muitos casos, uma amostra piloto pode fornecer informação suficiente sobre a população, de tal forma que se pode obter um estimador inicial razoável para σ^2 (BOLFARINE & BUSSAB, 2005).

Exemplo: (MONTGOMERY & RUNGER, 2003): Os sistemas de escapamento de uma aeronave

funcionam devido a um propelente sólido. A taxa de queima desse propelente é uma característica importante do produto. As especificações requerem que a taxa média de queima seja de 50 cm/s. Sabemos que o desvio-padrão da taxa de queima é σ = 2 cm/s. Suponha que quiséssemos um erro na estimação da taxa média de queima do propelente do foguete menor do que 1,5 cm/s, com uma confiança de 95%. Qual deveria ser o tamanho apropriado da amostra?

Solução: Uma vez que σ = 2 cm/s e $Z_{\frac{\alpha}{2}}$= 1,96, pode-se determinar o tamanho da amostra da seguinte forma:

$$n = \left(\frac{Z_{\frac{\alpha}{2}}\sigma}{\varepsilon}\right)^2 = \left(\frac{1{,}96x2}{1{,}5}\right)^2 = 6{,}83 \cong 7$$

Exemplo: Suponha que uma amostra aleatória de tamanho 10 da variável renda familiar apresente os seguintes valores: 12, 18, 12, 18, 18, 30, 12, 12, 18, e 30. Determine o tamanho da amostra que apresente uma estimativa com erro máximo $E = \sqrt{2}$, com γ = 0,95.

Solução: No presente problema não se tem informação a respeito de σ^2. Mas, sabe-se que a partir de uma amostra piloto pode-se obter uma estimativa razoável para σ^2. Para esta amostra, $\bar{X} = 18$e S^2 = 48. Com S^2 = 48, para ter uma amostra que apresente uma estimativa com erro máximo $E = \sqrt{2}$, com $\gamma = 0{,}95$, é necessário que o tamanho da amostra seja

$$n = \frac{t^2_{\left(\frac{\alpha}{2};n-1\right)}S^2}{\varepsilon^2} = \frac{t^2_{(0{,}025;\,9)}x48}{(\sqrt{2})^2} = \frac{2{,}262^2x48}{(\sqrt{2})^2} = 122{,}80 = 123.$$

INTERVALOS DE CONFIANÇA PARA DIFERENÇA ENTRE DUAS MÉDIAS (AMOSTRAS INDEPENDENTES)

Intervalo de Confiança para diferença entre duas médias com *Variâncias Populacionais conhecidas* e independentes:

i) Para P.I. ou P.F.A.C.R.: $$IC_{100(1-\alpha)\%}\left(\mu_a-\mu_b\right):\left(\bar{X}_a-\bar{X}_b\right)\pm Z_{\alpha/2}\sqrt{\frac{\sigma_a^2}{n_a}+\frac{\sigma_b^2}{n_b}}$$

ii) Para P.F.A.S.R.: $$IC_{100(1-\alpha)\%}\left(\mu_a-\mu_b\right):\left(\bar{X}_a-\bar{X}_b\right)\pm Z_{\alpha/2}\sqrt{\frac{\sigma_a^2}{n_a}\frac{\left(N_a-n_a\right)}{\left(N_a-1\right)}+\frac{\sigma_b^2}{n_b}\frac{\left(N_b-n_b\right)}{\left(N_b-1\right)}}$$

Em que:

$\bar{X}_a$ e $\bar{X}_b$ são médias amostrais, isto é, são as estimativas pontuais das médias das populações *a* e *b*, respectivamente;

σ_a^2 e σ_b^2 as variâncias das populações *a* e *b*, respectivamente;

n_a e n_b tamanho das amostras retiradas das populações *a* e *b*, respectivamente; N_a e N_b tamanhos das populações *a* e *b*, respectivamente.

Regras de decisão envolvendo Intervalo de Confiança (IC) para diferença entre duas médias.

i) Se o IC incluir o zero, então, $\mu_a = \mu_b$.

ii) Se o IC não incluir o zero, então, $\mu_a \neq \mu_b$.

iii) Se os extremos do intervalo forem negativos, então, $\mu_a < \mu_b$.

iv) Se os extremos do intervalo forem positivos, então, $\mu_a > \mu_b$.

Exemplo 6: Um supermercado não sabe se deve comprar lâmpadas da marca A ou B de mesmo preço. Testa-se uma amostra de 100 lâmpadas de cada marca. Os resultados obtidos são apresentados a seguir:

Marca da lâmpada	$\bar{X}$	σ
A	1.160 h	90 h
B	1.140	80 h

a) Construa um intervalo de confiança com 2% de significância e indique qual lâmpada o supermercado deve comprar.

Solução: Sabe-se que a variâncias populacionais são conhecidas, pois se forneceu informações a respeito de σ para cada marca de lâmpada. O valor de α adotado foi de 2% (0,02), e o tamanho das amostras retiradas é de 100 lâmpadas cada, ou seja, $n_a = n_b = 100$lâmpadas. Na presente situação não se conhece o tamanho das populações, consequentemente não há necessidade de fazer o uso do fator de correção. Portanto, o intervalo adotado para indicar qual lâmpada o supermercado deve comprar será:

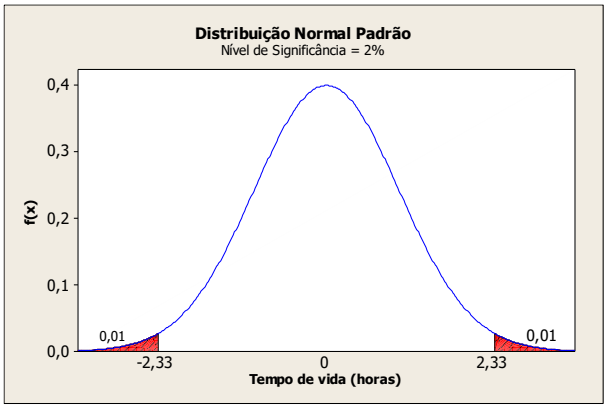


Figura 4 – Gráfico referente aos quantis inferior e superior $\frac{\alpha}{2}$ $\left(-Z_{\frac{\alpha}{2}} \text{ e } Z_{\frac{\alpha}{2}}\right)$, ao nível de significância de 2%.

$$IC_{100(1-\alpha)\%}(\mu_a-\mu_b):\left(\bar{X}_a-\bar{X}_b\right)\pm Z_{\alpha/2}\sqrt{\frac{\sigma_a^2}{n_a}+\frac{\sigma_b^2}{n_b}}$$

$$IC_{98\%}(\mu_a-\mu_b):(1.160-1.140)\pm Z_{0,02/2}\sqrt{\frac{90^2}{100}+\frac{80^2}{100}}$$

$$IC_{98\%}(\mu_a-\mu_b):(20)\pm Z_{0,01}\sqrt{\frac{90^2+80^2}{100}}$$

$$IC_{98\%}(\mu_a-\mu_b):20\pm 2,33\sqrt{\frac{90^2+80^2}{100}}$$

$$IC_{98\%}(\mu_a-\mu_b):20\pm 28,0569$$

$$IC_{98\%}(\mu_a-\mu_b):[-8,0569;\ \ 48,0569]$$

Interpretação: Portanto, pode-se afirmar com 98% de confiança que o tempo de vida médio das lâmpadas das marcas A e B são iguais. Ou seja, o cliente (proprietário do supermercado) pode comprar qualquer uma das lâmpadas, pois o zero está contido no intervalo.

b) Exercício: Construa um intervalo com 90% confiança e indique qual lâmpada o supermercado deve comprar.

Interpretação (Exercício):

Exercício: As empresas A e B produzem tubos de esgoto com variâncias em seus diâmetros iguais a 8 mm^2 e 10 mm^2, respectivamente. Uma amostra de 48 tubos da empresa A apresentou diâmetro médio igual a 40 mm, e uma amostra de 36 tubos da empresa B apresentou diâmetro médio de 42 mm. Verifique, por meio de um intervalo de confiança com 0,95 de probabilidade, se existe diferença entre os diâmetros médios dos tubos das marcas A e B.

Intervalo de Confiança para diferença entre duas médias com ***Variâncias Populacionais desconhecidas e amostras grandes e independentes*:**

i) Para P.I. ou P.F.A.C.R.: $$IC_{100(1-\alpha)\%}(\mu_a-\mu_b):\left(\bar{X}_a-\bar{X}_b\right)\pm Z_{\alpha/2}\sqrt{\frac{S_a^2}{n_a}+\frac{S_b^2}{n_b}}$$

ii) Para P.F.A.S.R.: $$IC_{100(1-\alpha)\%}(\mu_a-\mu_b):\left(\bar{X}_a-\bar{X}_b\right)\pm Z_{\alpha/2}\sqrt{\frac{S_a^2}{n_a}\frac{(N_a-n_a)}{(N_a-1)}+\frac{S_b^2}{n_b}\frac{(N_b-n_b)}{(N_b-1)}}$$

Em que:

- $\bar{X}_a$ e $\bar{X}_b$ são médias amostrais, isto é, são as estimativas pontuais das médias das populações *a* e *b*, respectivamente;
- S_a^2 e S_b^2 as variâncias das populações *a* e *b*, respectivamente;
- n_a e n_b tamanho das amostras retiradas das populações *a* e *b*, respectivamente; N_a e N_b tamanhos das populações *a* e *b*, respectivamente.

Intervalo de Confiança para diferença entre duas médias com ***Variâncias Populacionais desconhecidas e amostras pequenas e independentes*:**

Quando se desconhece as variâncias populacionais $(\sigma_a^2$ e $\sigma_b^2)$ torna-se necessário a substituição de seus valores paramétricos por suas estimativas amostrais $(S_a^2$ e $S_b^2)$. Neste caso, deve-se utilizar a distribuição t de Student, em lugar da normal. Além dessa alteração deve-se considerar ainda se as duas populações são homocedásticas ou heterocedásticas, isto é, se as variâncias populacionais desconhecidas são iguais ou diferentes, o que pode ser aferido por meio de um teste de hipótese para homogeneidade de variâncias.

i) Populações homocedásticas

Sendo as populações homocedásticas $(\boldsymbol{\sigma_a^2 = \sigma_b^2 = \sigma^2})$, assim, S_a^2 e S_b^2 são duas estimativas para um mesmo parâmetro (σ^2) então o intervalo de confiança para a diferença entre duas médias é dado por:

$$IC_{100(1-\alpha)\%}(\mu_a - \mu_b): (\bar{X}_a - \bar{X}_b) \pm t_{(\alpha/2;\, n_a+n_b-2)} S_p \sqrt{\frac{1}{n_a} + \frac{1}{n_b}}$$

com $t_{\frac{\alpha}{2}}$ com $v = n_a + n_b - 2$ graus de liberdade em que:

$$S_p = \sqrt{\frac{(n_a-1)S_a^2 + (n_b-1)S_b^2}{n_a+n_b-2}}.$$

ii) Populações heterocedásticas

Sendo as populações heterocedásticas $(\boldsymbol{\sigma_a^2 \neq \sigma_b^2})$, assim, S_a^2 e S_b^2 são duas estimativas de diferentes parâmetros, não podendo, pois serem combinadas em único valor. Então, o intervalo de confiança para a diferença entre duas médias é dado por:

$$IC_{100(1-\alpha)\%}(\mu_a-\mu_b):\left(\bar{X}_a-\bar{X}_b\right)\pm t_{(\alpha/2;\,v)}\sqrt{\frac{S_a^2}{n_a}+\frac{S_b^2}{n_b}}$$

com $t_{\frac{\alpha}{2}}$ com v graus de liberdade em que:

$$v=\frac{\left(\frac{S_a^2}{n_a}+\frac{S_b^2}{n_b}\right)^2}{\frac{\left(\frac{S_a^2}{n_a}\right)^2}{n_a-1}+\frac{\left(\frac{S_b^2}{n_b}\right)^2}{n_b-1}}.$$

Exemplo: Em uma prova de Estatística de 12 alunos de uma turma conseguiram média de 7,8 e desvio padrão de 0,6 ao passo que 15 alunos de outra turma do mesmo curso conseguiram média 7,4 com desvio de 0,8. Considerando distribuição normal para as notas ao nível de 5% de significância, determine o intervalo de confiança e indique se há diferença entre as turmas em termos de nota. Considere variâncias populacionais ***desconhecida***s, porém ***iguais***.

Solução: Diante do seguinte enunciado tem-se:

$\alpha = 5\% = 0{,}05;\ n_1 = 12;\ \bar{X}_1 = 7{,}8;\ S_1 = 0{,}6;\ n_2 = 15;\ \bar{X}_2 = 7{,}4$ e $S_2 = 0{,}8.$ Sendo as populações homocedásticas $\left(\boldsymbol{\sigma_a^2 = \sigma_b^2 = \sigma^2}\right)$ tem-se que:

$$S_p=\sqrt{\frac{(12-1)0{,}6^2+(15-1)0{,}8^2}{12+15-2}}=0{,}7189.$$

Logo, o intervalo de confiança é:

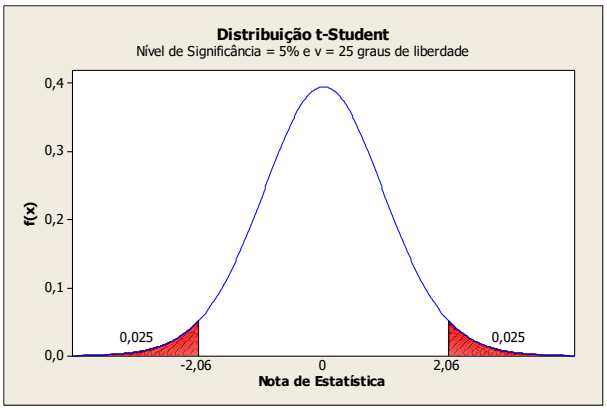


Figura 5 – Gráfico referente aos quantis inferior e superior $\frac{\alpha}{2\left(-t_{\left(\frac{v;\alpha}{2}\right)}\text{ e }t_{\left(\frac{v;\alpha}{2}\right)}\right)}$, ao nível de significância de 5% e 25 graus de liberdade.

$$IC_{95\%}(\mu_a-\mu_b):(\bar{X}_1-\bar{X}_2)\pm t_{\left(\frac{\alpha}{2;\,n_1+n_2-2}\right)}S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_1}}$$

$$IC(\mu_a - \mu_b): (\bar{X}_1 - \bar{X}_2) \pm t_{\left(\frac{0,05}{2;\, n_1+n_2-2}\right)} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_1}}$$

$$IC(\mu_a - \mu_b): (7,8 - 7,4) \pm t_{(0,025;\, 12+15-2)} S_p \sqrt{\frac{1}{12} + \frac{1}{15}}$$

$$IC(\mu_a - \mu_b): 0,4 \pm t_{(0,025;\, 25)} S_p \sqrt{\frac{1}{12} + \frac{1}{15}}$$

$$IC(\mu_a - \mu_b): 0,4 \pm 2,06 * 0,7189 * \sqrt{\frac{1}{12} + \frac{1}{15}}$$

$$IC(\mu_a - \mu_b): 0,4 \pm 0,5736$$

$$IC(\mu_a - \mu_b): [-0,1736;\ \ 0,9736]$$

Interpretação: Portanto, pode-se afirmar que estatisticamente as turmas são iguais, ou seja, que a média na prova de Estatística das turmas não diferem.

Exercício: O QI de 16 estudantes de uma zona pobre de certa cidade apresenta média de 107 pontos com desvio padrão de 10 pontos, enquanto que 14 estudantes de outra região rica da cidade apresentam média de 112 pontos, e desvio padrão de 8 pontos. O QI em ambas regiões tem distribuição normal. Determine O intervalo de confiança com uma certeza de 95%. Considere variâncias populacionais ***desconhecida***s, porém ***diferentes***.

Interpretação (Exercício):

INTERVALOS DE CONFIANÇA PARA DIFERENÇA ENTRE DUAS MÉDIAS (AMOSTRAS DEPENDENTES)

As amostras são consideradas dependentes quando as observações são correlacionadas.

Exemplo: Eficiência de uma dieta: Peso Inicial $\overrightarrow{dieta}$ Peso Final. Existe correlação entre o peso inicial e final.

O intervalo de confiança para diferença entre duas médias com amostras dependentes é dado por:

i) amostras pequenas (n≤30):

$$IC_{100(1-\alpha)\%}(\mu_D): \bar{D} \pm t_{(\alpha/2;\, v)} \frac{S_D}{\sqrt{n}}$$

com $t_{\frac{\alpha}{2}}$ com $v = n - 1$ graus de liberdade.

ii) amostras grandes (n>30):

$$IC_{100(1-\alpha)\%}(\mu_D): \bar{D} \pm Z_{(\alpha/2)} \frac{S_D}{\sqrt{n}}$$

em que $\bar{D} = \frac{\sum_{i=1}^{n} d_i}{n}$ e $S_D = \sqrt{\frac{1}{n-1}\left[\sum_{i=1}^{n} d_i^2 - \frac{\left(\sum_{i=1}^{n} d_i\right)^2}{n}\right]}$.

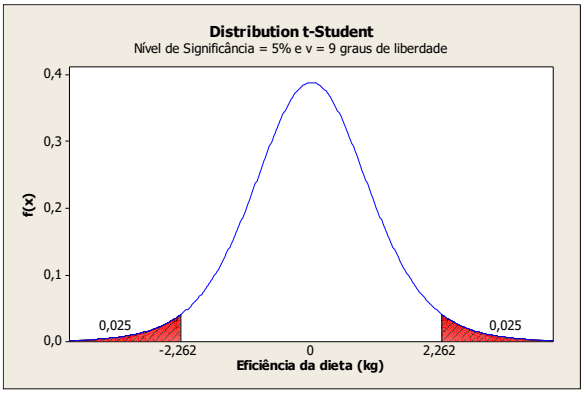
Exemplo Um grupo de 10 pessoas é submetido a um tipo de dieta por 10 dias, estando os pesos antes e depois marcados na tabela abaixo. Construa um intervalo de confiança ao nível de 5% de significância. Interprete os resultados.

Pessoas	Peso antes em kg	Peso depois em kg	d_i (1)	d_i^2
1	120	116	4	16
2	104	102	2	4
3	93	90	3	9
4	87	83	4	16
5	85	86	-1	1
6	98	97	1	1
7	102	98	4	16
8	106	108	-2	4
9	88	82	6	6
10	90	85	5	25

(1) d_i é a diferença de observações correlacionadas, ou seja, d_i = antes – depois; (2) d_i^2 é a diferença de observações correlacionadas ao quadrado.

Solução:

$\bar{D} = \frac{4+2+\cdots+5}{10} = 2{,}6$ kg e $S_D = \sqrt{\frac{1}{9}\left[(16+4+\cdots 25) - \frac{(26)^2}{10}\right]} = 2{,}59$ kg.



$$IC_{95\%}(\mu_D): \bar{D} \pm t_{\left(\frac{\alpha}{2;\nu}\right)} \frac{S_D}{\sqrt{n}}$$

$$IC_{95\%}(\mu_D): 2{,}6 \pm t_{\left(\frac{0{,}05}{2;9}\right)} \frac{2{,}59}{\sqrt{10}}$$

$$IC_{95\%}(\mu_D): 2{,}6 \pm 2{,}262 \frac{2{,}59}{\sqrt{10}}$$

$$IC_{95\%}(\mu_D): [0{,}747;\ 4{,}453]$$

Figura 6 – Gráfico referente aos quantis inferior e superior $\frac{\alpha}{2\left(-t_{\left(\frac{\nu;\alpha}{2}\right)} \text{ e } t_{\left(\frac{\nu;\alpha}{2}\right)}\right)}$, ao nível de significância de 5% e 9 graus de liberdade.

Interpretação: Pode-se afirmar com 95% de confiança que a dieta foi eficiente pois, $\mu_D > 0$.

INTERVALOS DE CONFIANÇA PARA PROPORÇÃO

Intervalo de Confiança para proporção de <u>amostras grandes</u> (n>30):

i) Para P.I. ou P.F.A.C.R.: $$IC_{100(1-\alpha)\%}(P): \hat{p} \pm z_{(\alpha/2)} \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

ii) Para P.F.A.S.R.: $$IC_{100(1-\alpha)\%}(P): \hat{p} \pm z_{(\alpha/2)} \sqrt{\frac{\hat{p}\hat{q}}{n}} \sqrt{\frac{N-n}{N-1}}, \text{ se } \frac{n}{N} > 0{,}05$$

Em que $\hat{p}$ é a proporção estimada na amostra; $\hat{q} = 1 - \hat{p}$; n é o tamanho da amostra e N é o tamanho da população.

Intervalo de Confiança para proporção de <u>amostras pequenas</u> (n≤30):

Quando a amostra for pequena deve-se utilizar a distribuição t de Student, em lugar da normal e o intervalo de confiança para a proporção será dado por:

i) Para P.I. ou P.F.A.C.R.: $$IC_{100(1-\alpha)\%}(P): \hat{p} \pm t_{\left(\frac{\alpha}{2};\ \nu=n-1\right)} \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

ii) Para P.F.A.S.R.:

$$IC_{100(1-\alpha)\%}(P): \hat{p} \pm t_{\left(\frac{\alpha}{2};\ v=n-1\right)} \sqrt{\frac{\hat{p}\hat{q}}{n}\frac{(N-n)}{(N-1)}},\ \text{se}\ \frac{n}{N} > 0,05$$

Em que $\hat{p}$ é a proporção estimada na amostra; $\hat{q} = 1 - \hat{p}$; n é o tamanho da amostra; N é o tamanho da população e *v* são os graus de liberdade (n-1).

Exemplo: Suponha que uma empresa de pesquisa eleitoral tenha entrevistado por telefone 400 eleitores, perguntando-lhes se votariam no candidato A ou no candidato B. Admita que 240 deles tenham respondido que votariam no candidato A.

a) Determine o intervalo de 95,0% de confiança para a proporção dos que indicam preferência pelo candidato A.

Solução: Sabe-se que o tamanho da amostra no presente exemplo é de 400 eleitores, ou seja, n = 400 (amostra grande, n>30). Desses 400 eleitores, 240 preferem o candidato A. Portanto, a proporção de eleitores que preferem o candidato A é:

$\hat{p} = \frac{240}{400} = 0,60 = 60\%$ então $\hat{q} = 1 - \hat{p} = 1 - 0,60 = 0,40 = 40\%$.

Como a amostra é grande e não se sabe o tamanho da população, então, utiliza-se o seguinte intervalo de proporção:

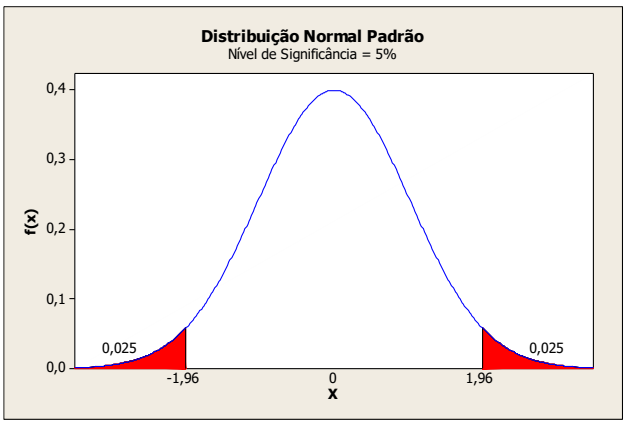


Figura 7 – Gráfico referente aos quantis inferior e superior $\frac{\alpha}{2\left(-Z\frac{\alpha}{2}\ \text{e}\ Z\frac{\alpha}{2}\right)}$, ao nível de significância de 5%.

$$IC_{100(1-\alpha)\%}(P): \hat{p} \pm z_{(\alpha/2)} \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

$$IC_{95\%}(P): \hat{p} \pm z_{(0,05/2)} \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

$$IC_{95\%}(P): 0,6 \pm z_{(0,025)} \sqrt{\frac{0,6.0,4}{400}}$$

$$IC_{95\%}(P): 0,6 \pm 1,96 \sqrt{\frac{0,24}{400}}$$

$$IC_{95\%}(P): 0,6 \pm 0,048$$

$$IC_{95\%}(P): [0,552;\ \ 0,648]$$

Interpretação: Assim, com uma amostra de tamanho 400, a pesquisa apresenta uma margem de erro de ± 4,8%, ou cerca de 5%. À vista do intervalo de confiança resultante (aproximadamente 55% a 65%), o candidato A pode sentir-se razoavelmente seguro quanto as suas perspectivas em relação à eleição.

b) (Exercício) Determine o intervalo de 99,0% de confiança para a proporção dos que indicam preferência pelo candidato A.

Interpretação (Exercício):

Exercício: Sabe-se por experiência que 5% da produção de um determinado artigo é defeituoso. Um novo empregado é contratado das 600 peças produzidas por ele, 82 são defeituosas. Se ele produzir mais artigos defeituosos do que o padrão da empresa ele é demitido. Determine um intervalo de 90% de confiança e verifique se você demitiria o empregado.

Interpretação (Exercício):

Dimensionamento do tamanho de amostras

Uma vez que $\hat{p}$ é o estimador de *p*, podemos definir o erro na estimação de *p* por meio de $\hat{p}$como $\varepsilon = |p - \hat{p}|$. Observe que estamos aproximadamente 100(1- α) confiantes de que esse erro seja menor do que $Z_{\frac{\alpha}{2}}\sqrt{p(1-p)/n}$. Ou seja, em situações em que o tamanho da amostra puder ser selecionado, podemos escolher *n* de modo a estarmos 100(1-α)% confiantes de que o erro seja menor do que algum valor especificado *E*. Se estabelecermos $\varepsilon = Z_{\frac{\alpha}{2}}\sqrt{p(1-p)/n}$ e resolvermos para n, o tamanho apropriado da amostra será (MONTGOMERY & RUNGER, 2003):

$$n = \left(\frac{Z_{\frac{\alpha}{2}}}{\varepsilon}\right)^2 \hat{p}(1-\hat{p}). \qquad (2)$$

Exemplo: Em uma amostra aleatória de 85 mancais de eixos de manivelas de motores de automóveis, 10 têm um acabamento de superfície mais rugoso do que as especificações permitidas. Quão grande deverá ser a amostra se quisermos estar 95% confiantes de que o erro em usar $\hat{p}$para estimar *p* seja menor do que 0,05?

Solução: Tem-se que $\hat{p} = 10/85 = 0{,}12$ é uma estimativa inicial de p. Dessa forma, o tamanho da amostra será:

$$n = \left(\frac{Z_{\frac{0{,}05}{2}}}{\varepsilon}\right)^2 \hat{p}(1-\hat{p}) = \left(\frac{1{,}96}{0{,}05}\right)^2 0{,}12(1-0{,}12) = \left(\frac{1{,}96}{0{,}05}\right)^2 0{,}12(0{,}88) \cong 163.$$

INTERVALOS DE CONFIANÇA PARA DIFERENÇA ENTRE DUAS PROPORÇÕES

Intervalo de Confiança para a diferença entre duas proporções de <u>amostras grandes</u> (n>30):

Dadas duas amostras independentes, de populações diferentes, o intervalo de confiança entre proporções nestas populações é dado por:

i) Para P.I. ou P.F.A.C.R.:

$$IC_{100(1-\alpha)\%}\left(P_a - P_b\right):\left(\hat{p}_a - \hat{p}_b\right) \pm z_{(\alpha/2)}\sqrt{\frac{\hat{p}_a\hat{q}_a}{n_a} + \frac{\hat{p}_b\hat{q}_b}{n_b}}$$

ii) Para P.F.A.S.R.: se $\frac{n_a}{N_a} > 0{,}05$ e se $\frac{n_b}{N_b} > 0{,}05$

$$IC_{100(1-\alpha)\%}\left(P_a - P_b\right):\left(\hat{p}_a - \hat{p}_b\right) \pm z_{(\alpha/2)}\sqrt{\frac{\hat{p}_a\hat{q}_a}{n_a}\frac{\left(N_a - n_a\right)}{\left(N_a - 1\right)} + \frac{\hat{p}_b\hat{q}_b}{n_b}\frac{\left(N_b - n_b\right)}{\left(N_b - 1\right)}}$$

Em que:

$\hat{p}_a$ é a proporção estimada na amostra retirada da população A;

$\hat{p}_b$ é a proporção estimada na amostra retirada da população B;

$\hat{q}_a = 1 - \hat{p}_a$ e $\hat{q}_b = 1 - \hat{p}_b$;

n_a e n_b são os tamanhos das amostras retiradas das populações A e B, respectivamente;

N_a e N_b são os tamanhos das populações A e B, respectivamente.

Nota: Se ocorrer P.F.A.S.R., o componente da variância referente a população na qual ocorreu P.F.A.S.R. deve ser multiplicado pelo seu respectivo fator de correção.

Intervalo de Confiança para a diferença entre duas proporções de <u>amostras pequenas</u> (n≤30):

Quando a amostra for pequena deve-se utilizar a distribuição t de Student, em lugar da normal e o intervalo de confiança para a diferença entre duas proporções será dado por:

i) Para P.I. ou P.F.A.C.R.:

$$IC_{100(1-\alpha)\%}\left(P_a - P_b\right):\left(\hat{p}_a - \hat{p}_b\right) \pm t_{\left(\frac{\alpha}{2};\ v=n_a+n_b-2\right)}\sqrt{\frac{\hat{p}_a\hat{q}_a}{n_a} + \frac{\hat{p}_b\hat{q}_b}{n_b}}$$

ii) Para P.F.A.S.R.: se $\frac{n_a}{N_a} > 0{,}05$ e se $\frac{n_b}{N_b} > 0{,}05$

$$IC_{100(1-\alpha)\%}\left(P_a - P_b\right): \left(\hat{p}_a - \hat{p}_b\right) \pm t_{\left(\frac{\alpha}{2};\ v=n_a+n_b-2\right)} \sqrt{\frac{\hat{p}_a\hat{q}_a}{n_a}\frac{\left(N_a-n_a\right)}{\left(N_a-n_a\right)} + \frac{\hat{p}_b\hat{q}_b}{n_b}\frac{\left(N_b-n_b\right)}{\left(N_b-n_b\right)}}$$

com $v = n_a + n_b - 2$ graus de liberdade.

Nota: Se ocorrer P.F.A.S.R., o componente da variância referente a população na qual ocorreu P.F.A.S.R. deve ser multiplicado pelo seu respectivo fator de correção.

Exercício: Dois setores de uma empresa querem saber se a proporção de funcionários que chegam atrasados ao trabalho é a mesma. Você como gerente da empresa precisa decidir qual setor receberá uma bonificação, ou seja, ganhará o setor que apresentar menor proporção de funcionários atrasados. Com base nos resultados abaixo verifique ao nível de 5% de significância se um ou os dois setores ganharão a bonificação.

Setor da Empresa	Administrativo	Financeiro
Proporção de funcionários atrasados	0,08	0,06
Tamanho da amostra	20	30

Interpretação (Exercício):

INTERVALO DE CONFIANÇA PARA VARIÂNCIA DE UMA POPULAÇÃO NORMAL

i) **COM MÉDIA CONHECIDA:**

$$IC_{100(1-\alpha)\%}\left(\sigma^2\right): \left[\frac{nS^2}{\chi^2_{\left(\frac{\alpha}{2};\ v=n\right)}};\ \frac{nS^2}{\chi^2_{\left(1-\frac{\alpha}{2};\ v=n\right)}}\right]$$

ii) COM MÉDIA DESCONHECIDA:

$$IC_{100(1-\alpha)\%}\left(\sigma^2\right): \left[\frac{(n-1)S^2}{\chi^2_{\left(\frac{\alpha}{2};\ v=n-1\right)}};\ \frac{(n-1)S^2}{\chi^2_{\left(1-\frac{\alpha}{2};\ v=n-1\right)}}\right]$$

em que S^2 é a variância amostral.

Exemplo: Sabe-se que o tempo de vida de um certo aparelho tem distribuição aproximadamente normal. Uma amostra de 25 aparelhos forneceu uma média de 500 horas e desvio padrão de 50 horas.

a) Construa um intervalo para σ^2 de 2% de probabilidade.

Solução: Sabe-se que n = 25, $\alpha = 2\%$, $\bar{X} = 500\ hs$ e $S = 50\ hs$ então:

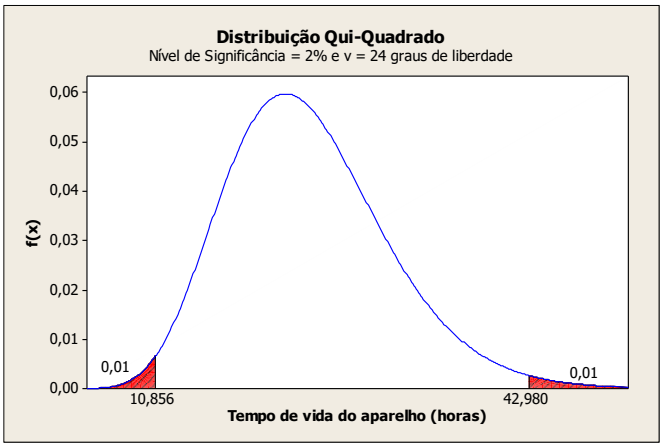


Figura 8 – Gráfico referente aos quantis inferior e superior $\frac{\alpha}{2}$ $\left(\chi^2_{\left(\frac{\alpha}{2};\ v=n-1\right)}\ \text{e}\ \chi^2_{\left(1-\frac{\alpha}{2};\ v=n-1\right)}\right)$, ao nível de significância de 2%.

$$IC_{100(1-\alpha)\%}(\sigma^2): \left[\frac{(n-1)S^2}{\chi^2_{\left(\frac{\alpha}{2};\ v=n-1\right)}};\ \frac{(n-1)S^2}{\chi^2_{\left(1-\frac{\alpha}{2};\ v=n-1\right)}}\right]$$

$$IC_{98\%}(\sigma^2): \left[\frac{(25-1)50^2}{\chi^2_{\left(\frac{0,02}{2};\ v=25-1\right)}};\ \frac{(25-1)50^2}{\chi^2_{\left(1-\frac{0,02}{2};\ v=25-1\right)}}\right]$$

$$IC_{98\%}(\sigma^2): \left[\frac{24*50^2}{\chi^2_{(0,01;\ v=24)}};\ \frac{24*50^2}{\chi^2_{(0,99;\ v=24)}}\right]$$

$$IC_{98\%}(\sigma^2): \left[\frac{24*50^2}{42,98};\ \frac{24*50^2}{10,856}\right]$$

$$IC_{98\%}(\sigma^2): [1396;\ 5527]$$

Interpretação: Pode-se afirmar com 98% de confiança que a variabilidade do tempo de vida do aparelho esta contida no intervalo [1.396; 5.527]

b) (Exercício) Construa um intervalo para σ^2 de 5% de probabilidade.

Interpretação (Exercício):

INTERVALO DE CONFIANÇA PARA O QUOCIENTE DE VARIÂNCIAS

Se S_1^2 e S_2^2 forem as variâncias de amostras aleatórias de tamanhos n_1 e n_2, respectivamente, provenientes de duas populações normais independentes, com variâncias desconhecidas σ_1^2 e σ_2^2, então um intervalo de confiança de 100(1 - α)% para o quociente (razão) de variâncias é dado por:

$$IC_{100(1-\alpha)\%}\left(\sigma_1^2/\sigma_2^2\right): \left[\frac{S_1^2}{S_2^2} F_{(1-\alpha/2;\ v_2=n_2-1,\ v_1=n_1-1)};\ \frac{S_1^2}{S_2^2} F_{(\alpha/2;\ v_2=n_2-1,\ v_1=n_1-1)}\right]$$

em que $F_{\left(\frac{\alpha}{2};\ v_2=n_2-1, v_1=n_1-1\right)}$ e $F_{\left(1-\frac{\alpha}{2};\ v_2=n_2-1, v_1=n_1-1\right)}$ são os pontos percentuais $\alpha/2$ superior e inferior da distribuição F, com $n_2 - 1$ graus de liberdade no numerador e $n_1 - 1$ graus de liberdade no denominador, respectivamente.

Nota: A Tabela da distribuição F contém somente pontos percentuais superiores, isto é, $F_{\left(\frac{\alpha}{2};\, u,v\right)}$. Os pontos percentuais inferiores $F_{\left(\frac{1-\alpha}{2};\, u,v\right)}$ podem ser encontrados como segue:

$$F_{\left(\frac{1-\alpha}{2};\, u,v\right)} = \frac{1}{F_{\left(\frac{\alpha}{2};\, v,u\right)}}$$

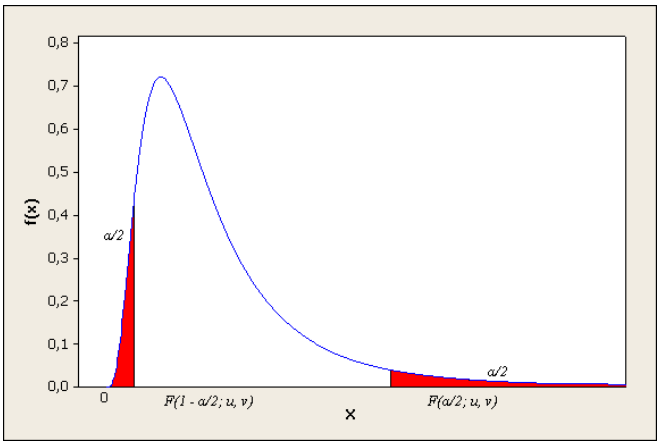


Figura 9 – Pontos percentuais superior e inferior da distribuição F, ao nível α de significância.

Dessa forma, $F_{\left(\frac{1-\alpha}{2};\, v_2=n_2-1, v_1=n_1-1\right)} = \frac{1}{F_{\left(\frac{\alpha}{2}; v_1=n_1-1, v_2=n_2-1\right)}}$, então o intervalo pode ser escrito como:

$$IC_{100(1-\alpha)\%}\left(\sigma_1^2/\sigma_2^2\right): \left[\frac{S_1^2}{S_2^2}\frac{1}{F_{(\alpha/2;\, v_1=n_1-1, v_2=n_2-1)}};\ \frac{S_1^2}{S_2^2}F_{(\alpha/2;\, v_2=n_2-1,\, v_1=n_1-1)}\right]$$

Regra de decisão para o intervalo de confiança (IC) para o quociente de variâncias:

1º) Se IC inclui 1 em sua extensão, então, estatisticamente $\sigma_1^2 = \sigma_2^2$.

2º) Se IC > 1, então, estatisticamente $\sigma_1^2 > \sigma_2^2$.

3º) Se IC < 1, então, estatisticamente $\sigma_1^2 < \sigma_2^2$.

Exemplo: De duas populações normais levantaram-se amostras de tamanho 9 e 11 respectivamente, obtendo-se $S_1^2 = 7{,}14$ e $S_2^2 = 3{,}21$. Construa um intervalo de confiança para o quociente das variâncias das duas populações ao nível de 10% e verifique se as variâncias populacionais podem ser consideradas relativamente iguais.

Solução: Diante das informações temos que $S_1^2 = 7{,}14$, $S_2^2 = 3{,}21$, $n_1 = 9$ e $n_2 = 11$, então, segue-se que: $\frac{S_1^2}{S_2^2} = \frac{7{,}14}{3{,}21} = 2{,}2243$, $v_1 = 8$ e $v_2 = 10$. Logo, o IC será:

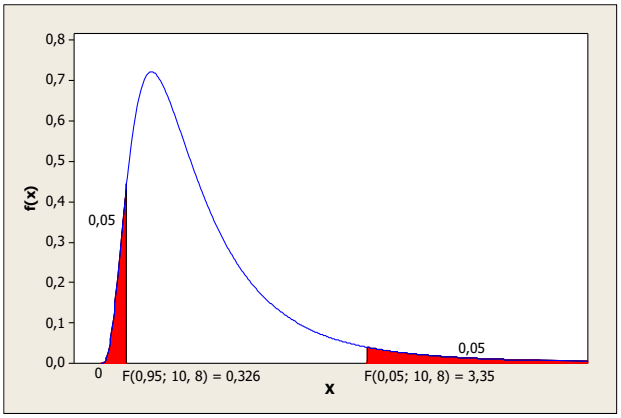


Figura 10 – Gráfico referente aos pontos percentuais superior e inferior da distribuição *F*, $F_{(0,05;\ 10,8)}$ e $F_{(0,95;\ 10,8)}$, ao nível de significância de 10%.

$$IC_{100(1-\alpha)\%}\left(\sigma_1^2/\sigma_2^2\right):\left[\frac{S_1^2}{S_2^2}F_{(1-\alpha/2;\ \nu_2=n_2-1,\ \nu_1=n_1-1)};\ \frac{S_1^2}{S_2^2}F_{(\alpha/2;\ \nu_2=n_2-1,\ \nu_1=n_1-1)}\right]$$

$$IC_{100(1-\alpha)\%}\left(\sigma_1^2/\sigma_2^2\right):\left[\frac{S_1^2}{S_2^2}\frac{1}{F_{(\alpha/2;\ \nu_1=n_1-1,\nu_2=n_2-1)}};\ \frac{S_1^2}{S_2^2}F_{(\alpha/2;\ \nu_2=n_2-1,\ \nu_1=n_1-1)}\right]$$

$$IC_{90\%}\left(\sigma_1^2/\sigma_2^2\right):\left[\frac{S_1^2}{S_2^2}\frac{1}{F_{(0,10/2;\ 8,10)}};\ \frac{S_1^2}{S_2^2}F_{(0,10/2;\ 10,8)}\right]$$

$$IC_{90\%}\left(\sigma_1^2/\sigma_2^2\right):\left[2,2243.\frac{1}{F_{(0,05;\ 8,10)}};\ 2,2243.F_{(0,05;\ 10,8)}\right]$$

$$IC_{90\%}\left(\sigma_1^2/\sigma_2^2\right):\left[2,2243.\frac{1}{3,0717};\ 2,2243.3,3472\right]$$

$$IC_{90\%}\left(\sigma_1^2/\sigma_2^2\right):[0,7241;\ 7,4452\ ]$$

Interpretação: Uma vez que esse intervalo de confiança inclui a unidade (um), não podemos afirmar que as variâncias para as duas populações sejam diferentes com um nível de 90% de confiança, ou seja, as populações são homocedásticas.

Exercício: Uma companhia fabrica propulsores para uso em motores de turbinas de avião. Uma das operações envolve esmerilhar o acabamento de uma superfície particular para um componente de liga de titânio. Dois processos diferentes para esmerilhar podem ser usados, podendo produzir peças com iguais rugosidades médias na superfície. Uma amostra aleatória de $n_1 = 11$peças, proveniente do primeiro processo, resulta em um desvio padrão de $S_1 = 5,1$ micro polegadas. Uma amostra aleatória de $n_2 = 16$peças, proveniente do segundo processo, resulta em um desvio padrão de $S_2 = 4,7$ micro polegadas. Considerando que os dois processos sejam independentes e que a rugosidade na superfície seja normalmente distribuída, encontre um intervalo de confiança de 90% para a razão de duas variâncias. Existe variabilidade da rugosidade da superfície para os dois processos?

Solução: temos que $S_1 = 5,1$, $S_2 = 4,7$, $n_1 = 11$ e $n_2 = 16$, então, segue-se que: $\frac{S_1^2}{S_2^2} = 1,177456$, $v_1 = 10$ e $v_2 = 15$. Logo, o IC será:

$$IC_{100(1-\alpha)\%}\left(\sigma_1^2/\sigma_2^2\right):\left[\frac{S_1^2}{S_2^2}F_{(1-\alpha/2;\ v_2=n_2-1,\ v_1=n_1-1)};\ \frac{S_1^2}{S_2^2}F_{(\alpha/2;\ v_2=n_2-1,\ v_1=n_1-1)}\right]$$

$$IC_{100(1-\alpha)\%}\left(\sigma_1^2/\sigma_2^2\right):\left[\frac{S_1^2}{S_2^2}\frac{1}{F_{(\alpha/2;\ v_1=n_1-1,v_2=n_2-1)}};\ \frac{S_1^2}{S_2^2}F_{(\alpha/2;\ v_2=n_2-1,\ v_1=n_1-1)}\right]$$

$$IC_{90\%}\left(\sigma_1^2/\sigma_2^2\right):\left[\frac{5,1^2}{4,7^2}\frac{1}{F_{(0,10/2;\ 10,\ 15)}};\ \frac{5,1^2}{4,7^2}F_{(0,10/2;\ 15,\ 10)}\right]$$

$$IC_{90\%}\left(\sigma_1^2/\sigma_2^2\right):\left[1,177456.\frac{1}{F_{(0,05;\ 10,\ 15)}};\ 1,177456.F_{(0,05;\ 15,\ 10)}\right]$$

$$IC_{90\%}\left(\sigma_1^2/\sigma_2^2\right):\left[1,177456.\frac{1}{2,5437};\ 1,177456.2,8450\right]$$

$$IC_{90\%}\left(\sigma_1^2/\sigma_2^2\right):\left[0,462891;\ 3,349862\ \right]$$

Interpretação: Uma vez que esse intervalo de confiança contém 1 em sua extensão, não podemos afirmar que as variâncias da rugosidade da superfície para os dois processos sejam diferentes com um nível de confiança de 90%. Ou seja, os dois processos diferentes para esmerilhar produzem peças com iguais rugosidades médias na superfície.

9-TEORIA DA DECISÃO – TESTES DE HIPÓTESES

O maior objetivo da inferência estatística é realizar inferências sobre os parâmetros desconhecidos a partir de amostras retiradas da população objeto de estudo. Uma das alternativas, muitas vezes utilizadas são os Testes de Hipóteses que consistem na tomada de decisões a partir da aceitação ou não de hipóteses, e por isso a teoria de testes de hipóteses também é chamada de Teoria da Decisão.

A Teoria da Decisão tem como objetivo de fornecer um processo de análise denominado de teste de hipóteses, que nos permite decidir por um valor do parâmetro θ ou por sua modificação com um grau de risco conhecido. Suponhamos que certa distribuição dependa de um parâmetro θ e que não se conheça θ ou, então, há razões para acreditar que θ variou, seja pelo passar do tempo ou, então, pela introdução de novas técnicas na produção (MORETTIN, 2005).

Hipótese Estatística

Uma hipótese, no contexto de inferência estatística, é definida como uma proposição acerca de um parâmetro populacional. Além disso, poder-se-ia dizer que é uma proposição cuja veracidade pode ser colocada em dúvida, ou que da qual não se tem total certeza. Em função da possibilidade de ela ser falsa, quase sempre pensa-se em uma hipótese complementar, a negação da primeira. A hipótese estatística é uma suposição quanto ao valor de um parâmetro que será verificado por

intermédio de um teste paramétrico ou uma informação quanto a natureza da população que seria verificado por meio de um teste não paramétrico (aderência). Portanto, pode-se definir teste de hipótese como a proposição de hipóteses H_0 e H_1.

De modo geral, as hipóteses irão se referir ao valor desconhecido do parâmetro em questão estar contido em subespaços do espaço paramétrico Θ (universo):

$$\begin{cases} H_0: \theta \in \Theta_0 \\ H_1: \theta \in \Theta_1 \end{cases}, \ \Theta = \Theta_0 \cup \Theta_1.$$

A rejeição de uma hipótese implica na aceitação da outra, e vice versa.

Tipos de Hipóteses

Formulam-se duas hipóteses básicas: a hipótese original de interesse, e sua complementar, que são respectivamente chamadas de hipótese de nulidade e hipótese alternativa, e são em geral simbolizadas por H_0 e H_1.

i) **H_0:** Hipótese nula ou de nulidade ou da existência → consiste na hipótese a ser testada.
ii) **H_1:** Hipótese alternativa → consiste na hipótese contrária a **H_0**.

Vejamos alguns exemplos de hipótese:

• Os pneus da marca A têm vida média $\mu = \mu_0$;
• O nível de inteligência de uma população de universitários é $\mu = \mu_0$;
• O equipamento A produz peças com variabilidade menor que a do equipamento B: $\sigma_A^2 < \sigma_B^2$;
• O pneu produzido pelo processo A é mais durável que o pneu produzido pelo processo B: $\mu_A > \mu_B$.

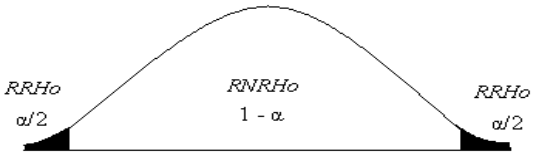
Tipos de Testes de Hipóteses

De acordo com o tipo de hipótese formulada pode-se ter os seguintes tipos de testes de hipóteses:

a) Teste Bilateral: Apresenta duas regiões de rejeição da hipótese H_0, situadas nos extremos da distribuição amostral, e é utilizado para testar as hipóteses do tipo:

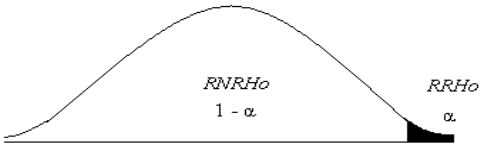
$$\begin{cases} H_0: \theta = \theta_0 \\ H_1: \theta \neq \theta_0 \end{cases} \Rightarrow$$

Testes bilaterais (duas regiões criticas).



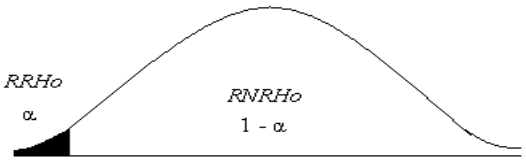
b) Teste Unilateral à Direita: Apresenta apenas uma única região de rejeição da hipótese H_0, situada no extremo superior da distribuição amostral, e é utilizado para testar hipóteses do tipo:

$\begin{cases} H_0: \theta = \theta_0 \\ H_1: \theta > \theta_0 \end{cases}$ $\Rightarrow$ Testes unilaterais à direita (uma região critica).



c) Teste Unilateral à Esquerda: Apresenta apenas uma única região de rejeição da hipótese H_0, situada no extremo inferior da distribuição amostral, e é utilizado para testar hipóteses do tipo:

$\begin{cases} H_0: \theta = \theta_0 \\ H_1: \theta < \theta_0 \end{cases}$ $\Rightarrow$ Testes unilaterais à esquerda (uma região critica)



d) $\begin{cases} H_0: \theta = \theta_0 \\ H_1: \theta = \theta_1 \end{cases}$ $\Rightarrow$ Testes aplicados a valores do parâmetro obtidos após a decisão tomada em um dos três testes anteriores.

Tipos de Erros

Ao realizar um teste de hipótese, dois tipos de erros são possíveis, rejeitar H_0 quando ela é verdadeira, ou aceitá-la quando ela é falsa. Esses erros são chamados, respectivamente, de erro tipo I e erro tipo II.

Decisão	**H_0 Verdadeira**	**H_0 Falsa**
Não Rejeitar H_0	Não há erro → $(1-\alpha) = \gamma$ [1] **Decisão correta**	Erro tipo II → β
Rejeitar H_0	Erro Tipo I → α [2]	Não há erro → $(1-\beta)$ [3] **Decisão correta**

[1] γ: nível de confiança; [2] α: nível de significância; [3] $(1-\beta)$: nível de significância;

Erro tipo I (α): ocorre quando rejeita-se H_0 e H_0 é verdadeira.

Erro tipo II (β): ocorre quando não rejeita-se H_0 e H_0 é falsa.

ALGORITMO PARA REALIZAÇÃO DE UM TESTE DE HIPÓTESE

i) Formular as hipóteses H_0 e H_1.

ii) Fixar o valor de α (nível de significância);

iii)Estabelecer a estatística do teste;

iv) Construir a regra de decisão: Região de Rejeição de H_0 (RRH_0) e Região de Não Rejeição de H_0 ($RNRH_0$).

v) Calcular a estatística adequada para o teste;

vi) Tomar a decisão;

vii) Conclusão.

TESTES DE HIPÓTESES PARA MÉDIA

Teste de hipótese para média μ de uma população Normal com *Variância Populacional* (σ^2) *conhecida*:

Estatística do teste: $Z_{calc} = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$

Hipóteses: $\begin{cases} H_0: \mu = \mu_0 \\ H_1: \mu \neq \mu_0 \end{cases}$ (Bilateral) ou $\begin{cases} H_0: \mu = \mu_0 \\ H_1: \mu > \mu_0 \end{cases}$ (unil. à direita) ou $\begin{cases} H_0: \mu = \mu_0 \\ H_1: \mu < \mu_0 \end{cases}$ (unil. à esquerda)

Exemplo 1: (MORETTIN, 2005): De uma população normal com variância 36, toma-se uma amostra casual de tamanho 16, obtendo-se $\bar{X} = 43$. Ao nível de 10%, testar as hipóteses: $\begin{cases} H_0: \mu = 45 \\ H_1: \mu \neq 45 \end{cases}$

Exemplo 2: Registros dos últimos anos de calouros de uma certa escola atestam que sua média num teste de QI foi μ = 115 e desvio padrão de σ = 20. Para saber se uma nova turma de calouros é típica desta escola, retirou-se uma amostra aleatória de 50 alunos desta nova classe, encontrado-se média de 118. Ao nível de 5% de significância, teste a hipótese de que esta nova turma apresenta a mesma característica das classes precedentes, com relação ao QI

Exemplo 3 (MORETTIN, 2005): Uma fábrica anuncia que o índice de nicotina dos cigarros da marca X apresenta-se abaixo de 26 mg por cigarro. Um laboratório realiza 10 análises do índice obtendo: 26, 24, 23, 22, 28, 25, 27, 26, 28, 24. Sabe-se que o índice de nicotina dos cigarros da marca X se distribui normalmente com variância 5,36 mg^2. Pode-se aceitar a afirmação do fabricante, ao nível de 5%?

Exemplo 4 (MORETTIN, 2005): Um fabricante de lajotas de cerâmicas introduz um novo material em sua fabricação e acredita que aumentará a resistência média, que é de 206 kg. A resistência das lajotas tem distribuição normal com desvio padrão de 12 kg. Retira-se uma amostra de 30 lajotas, obtendo-se $\bar{X} = 210$kg. Ao nível de 10%, pode o fabricante aceitar que a resistência média de suas lajotas tenha aumentado?

Teste de hipótese para média μ de uma população Normal com *Variância Populacional* (σ^2)

<u>*desconhecida*</u>:

i) amostras grandes (n>30):

Estatística do teste: $Z_{calc} = \frac{\bar{X} - \mu_0}{\frac{S}{\sqrt{n}}}$

ii) amostras pequenas (n≤30):

Estatística do teste: $t_{calc} = \frac{\bar{X} - \mu_0}{\frac{S}{\sqrt{n}}}$

Exemplo 1 (MORETTIN, 2005): Uma máquina é projetada para fazer esferas de aço de 1 cm de raio. Uma amostra de 10 esferas apresentou uma média de 1,004 cm e desvio padrão de 0,003 cm. Há razões para se suspeitar que a máquina esteja produzindo esferas com raio diferente de 1 cm ao nível de 10%?

Exemplo 2 (MORETTIN, 2005): Querendo determinar o peso médio de nicotina dos cigarros de sua produção, um fabricante recolheu uma amostra de 25 cigarros, obtendo:

$$\sum_{i=1}^{25} X_i = 950 \text{ mg} \quad \text{e} \quad \sum_{i=1}^{25} X_i^2 = 36106 \text{ mg}^2.$$

Supondo a distribuição norma para o peso de nicotina, testar se o peso médio de nicotina é inferior a 40mg.

TESTE DE HIPÓTESE PARA DIFERENÇA DE DUAS MÉDIAS (AMOSTRAS INDEPENDENTES)

Teste de hipótese para diferença de duas médias de populações Normal com <u>*Variâncias Populacionais conhecidas*</u>:

Estatística do teste: $Z_{calc} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$

Hipóteses: $\begin{cases} H_0: \mu_1 = \mu_2 \\ H_1: \mu_1 \neq \mu_2 \end{cases}$ (Bilateral) ou

$\begin{cases} H_0: \mu_1 = \mu_2 \\ H_1: \mu_1 > \mu_2 \end{cases}$ (unilateral à direita) ou

$\begin{cases} H_0: \mu_1 = \mu_2 \\ H_1: \mu_1 < \mu_2 \end{cases}$ (unilateral à esquerda)

Teste de hipótese para diferença de duas médias de populações Normal com *Variâncias Populacionais desconhecidas* com amostra grande ($n_1 + n_2 > 30$):

Estatística do teste: $Z_{calc} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$

Exemplo 1: Um supermercado não sabe se deve comprar lâmpadas da marca A ou B de mesmo preço. Testa-se uma amostra de 100 lâmpadas de cada marca e se quer saber se a marca A é melhor que a B ao nível de 2,5% de probabilidade.

Marca	$\bar{X}$	S
A	1160 horas	90 horas
B	1140 horas	80 horas

Exemplo 2: Retiradas amostras de aparelhos usados de duas marcas, encontrou-se os resultados apresentados no quadro a seguir. Verifique se existe diferença na durabilidade dos aparelhos A e B, ao nível de 5% de significância.

Marcas	A	B
Média	1160	1140
desvio padrão pop.	90	80
tamanho amostra	100	100

Teste de hipótese para diferença de duas médias de populações Normal com *Variâncias Populacionais desconhecidas e heterocedásticas* com amostra pequena ($n_1 + n_2 \leq 30$):

Estatística do teste: $t_{calc} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$, com *v* graus de liberdade

em que *v* é:

$$v = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 - 1}}.$$

Exemplo 1 (MORETTIN, 2005): O QI de 16 estudantes de uma zona pobre de certa cidade apresenta a média de 107 pontos com desvio padrão de 10 pontos, enquanto os 14 estudantes de outra região rica da cidade apresentam média de 112 pontos com desvio padrão de 8 pontos. O QI em ambas as regiões tem distribuição normal. Há uma diferença significativa entre os QIs médios dos dois grupos a 5% ?

Teste de hipótese para diferença de duas médias de populações Normal com *Variâncias Populacionais desconhecidas e homocedásticas* com amostra pequena ($n_1 + n_2 \leq 30$):

Estatística do teste: $t_{calc} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{Sp\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$, com $v = n_1 + n_2 - 2$ graus de liberdade,

em que *Sp* é dado por:

$$Sp = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}.$$

Exemplo 1 (MORETTIN, 2005): Em uma prova de Estatística, 12 alunos de uma classe conseguiram média 7,8 e desvio padrão 0,6, ao passo que 15 alunos de outra turma, do mesmo curso, conseguiram média 7,4 com desvio padrão de 0,8. Considerando distribuições normais para as notas, verificar se o primeiro grupo é superior ao segundo, ao nível de 5%. Como as populações são normais e com variâncias desconhecidas, podemos considerar que, apesar de desconhecidas, são iguais, já que são turmas de mesmo curso.

Teste de hipótese – Para uma proporção

Estatística do teste

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{\hat{p}\hat{q}}{n}}} \text{ ou } t = \frac{\hat{p} - p_0}{\sqrt{\frac{\hat{p}\hat{q}}{n}}} \text{ com v = n -1}$$

Exemplo 1: No departamento de controle de qualidade da empresa Lumiar (fabricante de lâmpadas), o gerente do departamento, com base em sua experiência, garante que 95% das lâmpadas fabricadas pela empresa não apresentam nenhum defeito. O presidente da empresa, um indivíduo preocupado com qualidade, verifica uma amostra aleatória de 225 lâmpadas e descobre que apenas 87% delas não apresentam defeitos. Ele, então, decide testar a hipótese (em um nível de significância de 0,05) de que 95% das lâmpadas fabricadas por sua empresa não apresentam defeitos.

Exemplo 2: Um investidor, deseja saber se a proporção de empresas que apresentam alto potencial para investimento é menor que 10%. Ele realiza uma amostragem de 122 empresas e encontrou 11 com alto potencial. Verifique a hipótese do investidor. Com um nível de 5% de significância

Teste de hipótese – Para a diferença entre proporções

Estatística do teste

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\left(\frac{\hat{p}_1\hat{q}_1}{n_1}\right) + \left(\frac{\hat{p}_2\hat{q}_2}{n_2}\right)}} \text{ ou } t = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\left(\frac{\hat{p}_1\hat{q}_1}{n_1}\right) + \left(\frac{\hat{p}_2\hat{q}_2}{n_2}\right)}} \textbf{ com } v = n_1 + n_2 - 2$$

Exemplo 1: John e Linda, executivos de vendas de uma grande empresa de computadores, são finalistas da competição anual entre os vendedores da empresa. Eles têm desempenhos idênticos, e o vencedor da competição será selecionado com base em seu "índice de conversão" (ou seja, o número de negócios potenciais convertidos em vendas). O gerente de vendas pegou aleatoriamente 100 dos contatos de John e descobriu que 84 deles haviam-se convertido em clientes. No caso de Linda, esse número foi de 82 para 100. O gerente de vendas precisa saber (com $\alpha = 0{,}05$) se existe diferença no índice de conversão, com base nas proporções das amostras.

TESTE DE QUI-QUADRADO (χ^2)

O teste de χ^2 mede a discrepância existente entre frequências observadas e frequências esperadas em um conjunto de dados, podendo ser utilizado como teste de aderência e de independência.

Passos:

1) Determinar modelo teórico (Independência)

2) Calcular as frequências esperadas (f_e) para cada classe da variável aleatória.

3) compara as f_e com as observadas (f_o) através do seguinte teste:

$$\chi_c^2 = \sum_{i=1}^{k} \frac{\left(f_{o_i} - f_{e_i}\right)^2}{f_{e_i}}$$

Em que k: número de classes

4) Se $\chi^2_{(\alpha,k-1)} < \chi^2_c$ então o modelo teórico não se ajusta a distribuição observada

Teste de independência

Em pesquisas ou levantamentos feitos por meio de entrevistas, questionários ou fichas, quando o pesquisador classifica os indivíduos amostrados segundo duas ou mais variáveis qualitativas categóricas ou ordinais, a apresentação tabular das frequências observadas pode ser feita através de uma tabela de contingência, isto é, uma tabela com duas ou mais entradas, cada entrada relativa a uma das variáveis. Com a tabela de contingência, conseguimos uma maneira conveniente de fazer descrição dos dados da amostra quando temos duas ou mais variáveis qualitativas a considerar.

Passamos agora à análise dos dados fornecidos pela tabela.

Uma indagação que pode ser objeto de um teste bastante simples é se as variáveis qualitativas envolvidas

são ou não independentes. Ou seja, podemos desejar testar as hipóteses:

H_0: variável linha independe da variável coluna (as variáveis são independentes)

H_1: variável linha depende da variável coluna (as variáveis não são independentes, ou seja, elas apresentam algum grau de associação entre si)

$$\chi_c^2 = \sum_{i=1}^{k} \frac{\left(f_{o_i} - f_{e_i}\right)^2}{f_{e_i}} \qquad f_e = \frac{(totallinha).(totalcoluna)}{totalgeral}$$

Em que k: número de classes

Neste caso os graus de liberdade são dados por:

a) GL = k-1 nas tabelas de simples entrada

b) GL = (h-1).(k-1) nas tabelas com h linhas e k colunas

A hipótese nula (H_0), será rejeitada, ao nível de significância estipulado, quando $\chi^2_{obs} > \chi^2_{(v,\alpha)}$, caso contrário, não rejeita-se H_0.

Exemplo 1: A tabela a seguir mostra os resultados de uma enquete com 200 assinantes de temporadas musicais, aos quais foi perguntado com que frequência costumavam assistir aos concertos em uma cidade vizinha. A frequência do comparecimento aos espetáculos foi dividida nas categorias nunca, ocasionalmente e frequentemente. Também foi perguntado aos respondentes se eles percebiam o local dos concertos como convenientes ou inconvenientes. Teste a hipótese de que a frequência de comparecimento aos concertos independe do local, com um nível de significância de 0,05.

Comparecimento Local	Frequentemente	Ocasionalmente	Nunca	Total
Conveniente	22 (16)	48 (40)	10 (24)	80
Inconveniente	18 (24)	52 (60)	50 (36)	120
Total	40	100	60	200

Teste de Hipótese para σ^2

Hipóteses

H_0 : σ^2 = a (hipótese nula)

H_1: $\sigma^2 \neq$ a (hipótese alternativa)

Estatística do teste: $F_{calc} = \frac{S_1^2}{S_2^2}$, em que S_1^2e a maior das variâncias amostrais)

com graus de liberdade do numerador: n_1 -1

e graus de liberdade do denominador: n_2 -1

Exemplo 1: Uma v.a. qualquer tem uma distribuição desconhecida com média *m* e variância s^2 desconhecidas. Retira-se uma amostra de 25 valores e calcula-se a variância amostral. Supondo que s^2 = 2,34, teste a hipótese de que a verdadeira variância s^2 seja de fato igual a 4, considerando 5% de significância.

10- CORRELACAO E REGRESSÃO LINEAR SIMPLES

INTODUÇÃO

Um dos objetivos da ciência é encontrar, descrever e predizer relações entre eventos que ocorrem na natureza. Um caminho para que isto aconteça é encontrar modelos que relacionem variáveis que descrevam a realidade. Pode-se atingir este objetivo por meio de modelos de regressão. A análise de regressão ocupa-se do estudo da dependência de uma variável, a variável dependente, em relação a uma ou mais variáveis, as variáveis explicativas, com o objetivo de estimar e/ou prever a média (da população) ou o valor médio da dependente em termos dos valores conhecidos ou fixos (em amostragem repetida) das explicativas. Ou seja, quando ajustamos um modelo que estabelece uma relação linear entre uma variável dependente e uma variável independente, estamos estimando um modelo de *regressão linear simples*. Quando existe uma relação linear entre uma variável dependente e duas ou mais variáveis independentes, ajusta-se um modelo de *regressão linear múltipla*.

Primeiramente, vamos entender o significado de *regressão linear simples*. Galton (1886), por meio de um famoso ensaio verificou que, embora houvesse uma tendência de pais altos terem filhos altos e pais baixos terem filhos baixos, a altura média dos filhos de pais de uma dada altura tendia a se deslocar ou "*regredir*" até a altura média da população como um odo. Daí, o nome de *regressão*, conhecida também como a lei de regressão universal de Francis Galton. Pearson & Lee (1903), coletou mais de mil registros das alturas dos membros de grupos de famílias, neste estudo verificou-se que tanto os filhos altos como baixos "*regrediram*" em direção à altura média de todos os homens. Nas palavras de Galton, tratava-se de uma regressão á "*mediocridade*".

DIAGRAMA DE DISPERSÃO

O diagrama de dispersão nos possibilita observar os dados graficamente e tirar conclusões prévias sobre a possível relação entre as variáveis. Para ilustrarmos a construção do diagrama de dispersão trabalharemos com o conjunto de dados do exemplo 1, que refere-se a volume de produção e custos totais de uma manufatura particular. Quais conclusões prévias se podem tirar da Figura 1? Observa-se na Figura 1 que maiores valores de custos tendem a se relacionar com maiores volumes de produção. Além disso, referente a esses dados, a relação entre o volume de produção e o custo total parece aproximar-se de uma linha reta, de fato, uma relação linear positiva é indicada entre *x* e *y*, como pode ser observado na Figura abaixo.

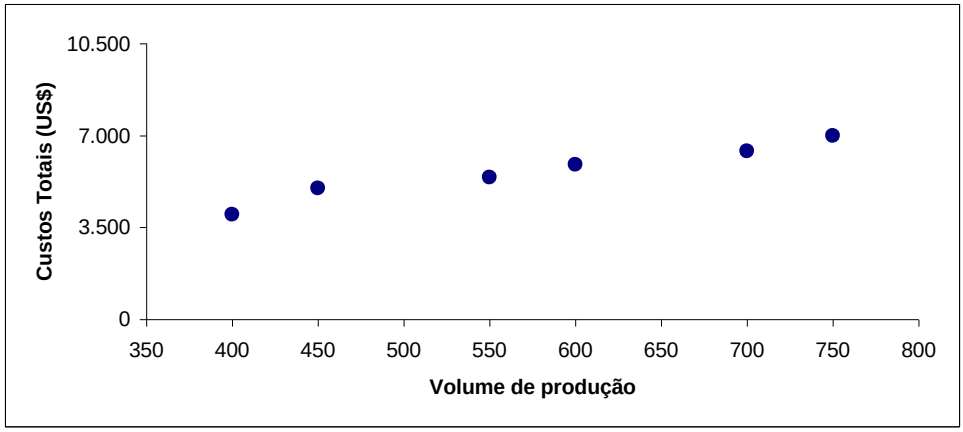


Figura – Diagrama de dispersão referente ao volume de produção e custo total de uma manufatura particular.

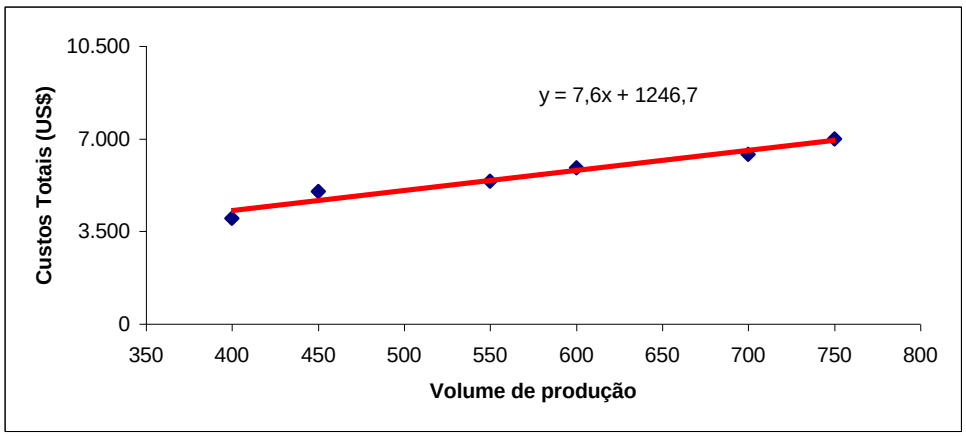


Figura – Gráfico da equação de regressão estimada para os dados de volume de produção e custo total de uma manufatura particular.

COEFICIENTE DE DETERMINAÇÃO

O coeficiente de determinação nos dá uma medida da eficiência (ou da qualidade) do ajuste do modelo, ou seja, indica quanto da variação de *y* (variação total) que é "explicada" pelo modelo de

regressão ajustado. Portanto, o coeficiente de determinação pode ser utilizado como um avaliador do modelo ajustado. O coeficiente de determinação é dado por:

$$R^2 = \frac{SPxy^2}{Sxx\ Syy}, \quad 0 \leq R^2 \leq 1, \quad (20)$$

em que $Sxx = \sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}$ é a soma de quadrados de *x* e $Syy = \sum_{i=1}^{n} y_i^2 - \frac{(\sum_{i=1}^{n} y_i)^2}{n}$ é a soma de quadrados de *y*.

COVARIANCIA E COEFICIENTE DE CORRELAÇÃO LINEAR DE PEARSON $(\boldsymbol{\rho})$

A covariância mede a força do relacionamento entre duas variáveis em termos absolutos através da seguinte equação (covariância amostral):

$$Cov(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}, \quad -\infty < Cov(x,y) < \infty.$$

Um coeficiente de correlação é a covariância dividida pelo produto do desvio padrão de cada variável. O coeficiente de correlação mede o grau de associação linear entre duas variáveis, *x* e *y*, ou seja, determina-se o grau de relacionamento ou a covariabilidade entre duas variáveis. Enquanto, que a regressão linear estabelece uma relação (função ou modelo) para as variáveis envolvidas. Outro aspecto importante é que na análise de regressão é necessário distinguir a variável dependente da variável independente, na análise de correlação tal distinção não é necessária. O coeficiente de correlação nada mais é do que uma covariância entre duas variáveis *x* e *y* que estão padronizadas, cujo objetivo de tal padronização é justamente para eliminar qualquer influência da escala. O estimador do coeficiente de correlação linear populacional de Pearson $(\boldsymbol{\rho})$é o coeficiente de correlação linear amostral, denotado por *r*:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{Cov(x,y)}{\sqrt{V(x)}\sqrt{V(y)}} = \frac{SPxy}{\sqrt{Sxx}\sqrt{Syy}}, \quad -1 \leq r \leq 1 \qquad (21)$$

em que $Sxx = \sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}$ é a soma de quadrados de *x* e $Syy = \sum_{i=1}^{n} y_i^2 - \frac{(\sum_{i=1}^{n} y_i)^2}{n}$ é a soma de quadrados de *y*.

Uma breve discussão é apresentada a respeito do coeficiente de correlação linear de Pearson: Se ρ = 0, tem-se que as variáveis *x* e *y* são *não correlacionadas linearmente*, ou seja, *ausência de*

correlação linear entre *x* e *y*. Dessa forma, pode-se dizer que não existem meios lineares acurados (precisos) para realizar previsões de valores de *y* conhecendo-se os valores de *x*, ou vice-versa (Figura 5 (c)). Se $\rho > 0$, indica que existe uma relação *linear positiva* entre *x* e *y*, o que significa que há uma tendência de pequenos valores de *x* estarem associados a pequenos valores de *y* e vice-versa, isto é, existe uma relação linear diretamente proporcional (Figura - 5 (a)). Se $\rho < 0$, indica que existe uma relação *linear negativa* entre *x* e *y*, o que significa que há uma tendência de pequenos valores de *x* estarem associados a pequenos valores de *y* e vice-versa, isto é, existe uma relação linear inversamente proporcional (Figura 5 (b)). Os diferentes tipos de correlação podem ser visualizados na Figura 5.

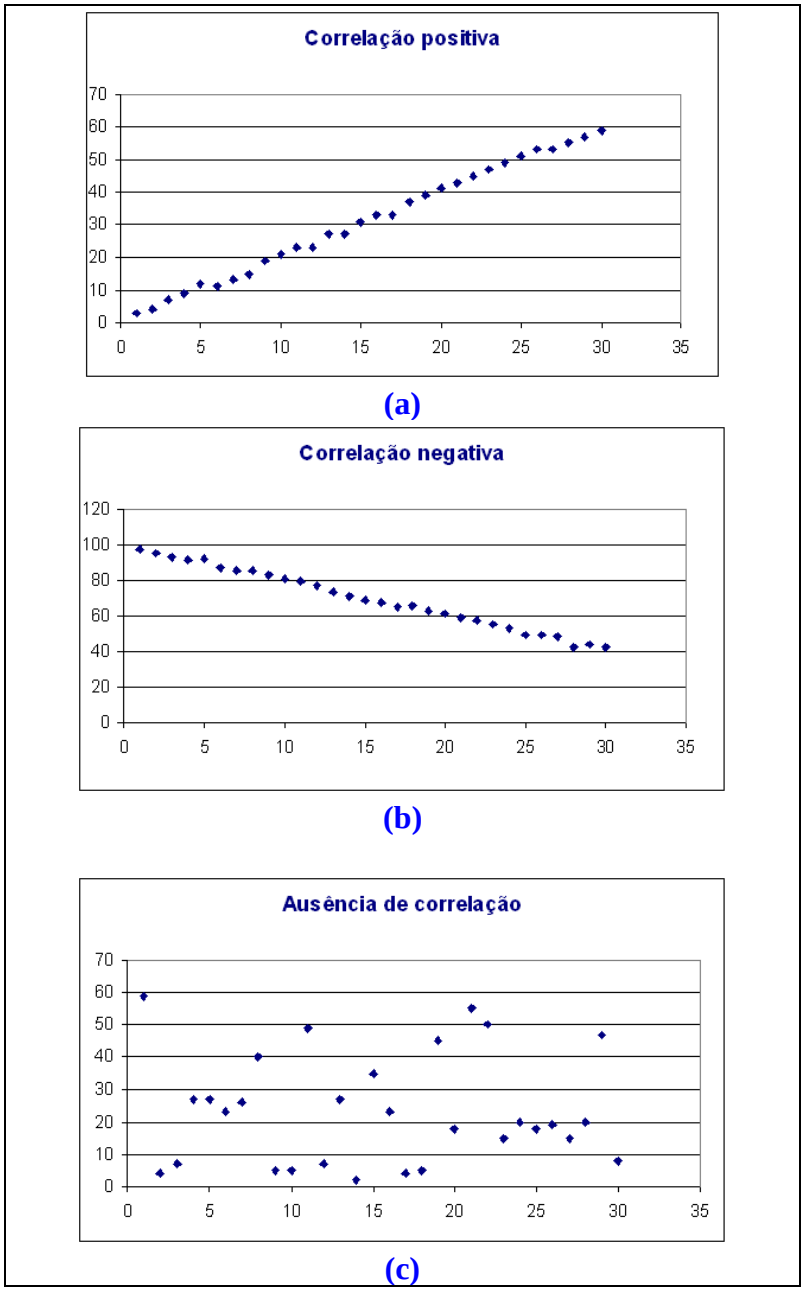


Figura 5 - Tipos de associação linear entre duas variáveis.

Para facilitar a interpretação do coeficiente de correlação vamos admitir as seguintes classificações para o coeficiente de correlação linear:

Coeficiente de Correlação	***Correlação***
$r = 1$	*Perfeita Positiva*
$0{,}8 \leq r < 1$	*Forte Positiva*
$0{,}5 \leq r < 0{,}8$	*Moderada Positiva*
$0{,}1 \leq r < 0{,}5$	*Fraca Positiva*
$0 < r < 0{,}1$	*Ínfima Positiva*
$r = 0$	*Nula*
$-0{,}1 < r < 0$	*Ínfima Negativa*
$-0{,}5 < r \leq -0{,}1$	*Fraca Negativa*
$-0{,}8 < r \leq -0{,}5$	*Moderada Negativa*
$-1 < r \leq -0{,}8$	*Forte Negativa*
$r = -1$	*Perfeita Negativa*

TESTE DE HIPÓTESE PARA COEFICIENTE DE CORRELAÇÃO

Hipóteses: $\begin{cases} \boldsymbol{H_0 : \rho = 0} \\ \boldsymbol{H_1 : \rho \neq 0} \end{cases}$

Estatística do Teste

$\boldsymbol{t = \frac{r\sqrt{n-2}}{\sqrt{1-R^2}}}$ com $v = n - 2$ graus de liberdade.

Nota: O procedimento para o teste de hipótese:

$$\begin{cases} \boldsymbol{H_0 : \rho = \rho_0} \\ \boldsymbol{H_1 : \rho \neq \rho_0} \end{cases}$$

em que $\rho_0 \neq 0$ é um pouco mais complicado. Para amostras moderadamente grandes ($n \geq 25$), a estatística do teste é:

$$Z = arctgh(r) = \frac{1}{2} ln\left(\frac{1+r}{1-r}\right) \qquad (22)$$

é distribuída de forma aproximadamente normal, com média $\mu_z = arctgh(\rho) = \frac{1}{2} ln\left(\frac{1+\rho}{1-\rho}\right)$ e variância $\sigma_z^2 = (n-3)^{-1}$.

Logo, para testar a hipótese $H_0: \rho = \rho_0$, podemos usar a estatística de teste:

$$Z_c = [arctgh(r) - arctgh(\rho_0)]\sqrt{n-3}, \qquad (23)$$

e rejeita-se $H_0: \rho = \rho_0$ se o valor da estatística do teste $|Z_c| > Z_{\frac{\alpha}{2}}$.

Exemplo 1: Relação entre o número de pessoas atendidas em um PS e o número de uma determinada cirurgia semanal para uma amostra de 10 hospitais.

Hospitais	Atendimentos	Cirurgias
1	907	11
2	926	11
3	506	7
4	741	9
5	789	9
6	889	10
7	874	9
8	510	6
9	529	7
10	420	6

a) Represente os dados em um gráfico.

Exemplo 2: Relação entre o número de clientes e as vendas semanais para uma amostra de 20 empresas de remessa de cargas.

Empresa	Clientes	Vendas	Empresa	Clientes	Vendas
1	907	11,2	11	679	7,63
2	926	11,05	12	872	9,43
3	506	6,84	13	924	9,46
4	741	9,21	14	607	7,64
5	789	9,42	15	452	6,92
6	889	10,08	16	729	8,95
7	874	9,45	17	794	9,33
8	510	6,73	18	844	10,23
9	529	7,24	19	1010	11,77
10	420	6,12	20	621	7,41

a) Represente os dados em um gráfico.

b) Calcule o coeficiente de correlação:

c) Faca o teste de hipótese sobre o coeficiente de correlação.

REGRESSAO LINEAR SIMPLES

O termo *linear* está relacionado à classificação do modelo. Os modelos de regressão são classificados como lineares, linearizáveis e não-lineares. Nos modelos não-lineares, não é possível encontrar uma forma analítica para a estimação dos parâmetros, isto é, as expressões dos estimadores não apresentam uma solução explícita, exigindo o uso de métodos numéricos iterativos.

Draper & Smith (1998) classificam os modelos de regressão como:

a) **modelos lineares:** aqueles que são lineares em relação aos parâmetros, ou seja:

$$\frac{\partial}{\partial \theta_i} f_j(X,\theta) = h(X),$$

para $i = 1,2,\dots,p;\ \ j = 1,2,\dots,n;$ "p" é o número de parâmetros do modelo e "n" o número de observações. Como ilustração, é apresentado o seguinte modelo de regressão:

$$Y = \beta_0 + \beta_1 X + \varepsilon,$$

em que o erro é aditivo e β_0 e β_1 são os parâmetros a serem estimados. O cálculo das derivadas parciais,

$$\frac{\partial Y}{\partial \beta_0} = 1 \quad \text{e} \quad \frac{\partial Y}{\partial \beta_1} = X,$$

mostra que nenhuma delas depende de algum parâmetro do modelo, portanto, o modelo é dito linear.

b) **modelos linearizáveis:** são modelos que por meio de alguma transformação se tornam lineares. Seja o modelo:

$$Y = \theta^X \varepsilon$$

em que, θ é um parâmetro a ser estimado e o erro é multiplicativo. Aplicando-se o logaritmo em ambos os lados da equação, tem-se:

$$\log(Y) = \log\left(\theta^X \varepsilon\right)$$
$$\log(Y) = X\log(\theta) + \log(\varepsilon).$$

Fazendo $G = log(Y)$; $c = log(\theta)$; $e = log(\varepsilon)$, a equação pode ser escrita como:

$$G = cX + e$$

sendo linear, pois

$\frac{\partial G}{\partial c} = X = h(X)$,

que independe do parâmetro, mostrando que o modelo original é linearizável.

c) **modelos não-lineares:** são modelos em que pelo menos uma das derivadas parciais depende de algum parâmetro do modelo. Seja o modelo:

$$Y = \theta_1 + \theta_2{}^X + \varepsilon$$

onde θ_1 e θ_2 são os parâmetros a serem estimados. O cálculo das derivadas parciais de y:

$$\frac{\partial Y}{\partial \theta_1} = 1 \quad \text{e} \quad \frac{\partial Y}{\partial \theta_2} = X\theta_2{}^{X-1},$$

mostra que a segunda delas depende do parâmetro θ_2, indicando que o modelo em questão é não-linear.

O termo *simples* e *múltipla* está relacionado ao número de variáveis independentes do modelo de regressão, isto é, quando existe uma relação linear entre uma variável dependente e uma variável independente, ajusta-se um modelo de *regressão linear simples*. Caso exista uma relação linear entre uma variável dependente e duas ou mais variáveis independentes, ajusta-se um modelo de *regressão linear múltipla*.

MODELO DE REGRESSÃO LINEAR SIMPLES

O modelo de regressão linear simples relata o estudo de como a variável dependente *y* se relaciona com uma variável independente *x*. O modelo estatístico de uma regressão linear simples é:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \text{ ou } y = \beta_0 + \beta_1 x + \varepsilon, \quad (1)$$

em que:

y_i: representa o i-ésimo valor observado;

x_i: representa a variável independente, *i* = 1, 2, ..., *n*;

ε_i: é o erro não observável associado a i-ésima observação;

β_0 e β_1: são os parâmetros do modelo (1), que são o intercepto ou coeficiente linear e o coeficiente

angular de regressão.

PRESSUPOSIÇÕES SOBRE O MODELO DE REGRESSÃO LINEAR SIMPLES

Ao estabelecer o modelo de regressão linear simples, pressupomos que:

i) A relação entre *x* e *y* é linear;

ii) Os valores de *x* são fixos, isto é, *x* não é uma variável aleatória;

iii) A média do erro é zero, isto é, $E(\varepsilon_i) = 0, \forall i = 1,2, \dots, n$;

iv) Para um dado valor de *x*, a variância do erro ε_i é sempre constante, isto é, $V(\varepsilon_i) = E(\varepsilon_i^2) = \sigma^2, \ \forall i = 1,2, \dots, n$. Diz-se, então, que o erro é homocedástico;

v) O erro de uma observação é não correlacionado com o erro de outra observação (os erros são independentes), ou seja, $E(\varepsilon_i \varepsilon_j) = 0 \text{ para } i \neq j$;

vi) O erro tem distribuição Normal com média zero e variância constante (σ^2), isto é, $\varepsilon_i \sim N(0, I\sigma^2)$.

Em síntese, temos que os erros são independentes e identicamente distribuídos (distribuição Normal com média zero e variância σ^2), ou seja,$\varepsilon_i \sim iidN(0, I\sigma^2)$. A quarta pressuposição se faz necessário para obter os intervalos de confiança e testes de hipóteses.

ESTIMADORES DE MÍNIMOS QUADRADOS

O objetivo na regressão é determinar estimadores de β_0 e β_1 de tal forma que as distâncias médias entre a reta de regressão e os valores observados sejam minimizadas, ou seja, o erro cometido deve ser o menor possível.

A partir do modelo (1) pode-se definir o erro da seguinte forma:

$$\varepsilon_i = y_i - \beta_0 - \beta_1 x_i. \qquad (2)$$

O método que será utilizado para determinar os estimadores de β_0 e β_1 é denominado de Método de Mínimos Quadrados. Esse método consiste em minimizar a soma de quadrados do erro ou resíduo do modelo (1) ao longo de todos os *n* pares (x_i, y_i). A partir da equação (2) pode-se definir a soma de quadrados dos resíduos (*Q*) como:

$$Q = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \qquad (3)$$

Para obter os estimadores de mínimos quadrados basta derivarmos a expressão (3) em relação aos

parâmetros β_0 e β_1 e posteriormente, igualarmos essas derivadas parciais a zero. Primeiramente, vamos obter as derivadas parciais (Sistema de Equações Normais, SEN):

$$(SEN)\begin{cases}\frac{\partial Q}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)\\ \frac{\partial Q}{\partial \beta_1} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)x_i\end{cases} \qquad (4)$$

Igualando essas derivadas a zero e substituindo β_0 e β_1, pelos respectivos estimadores $\hat{\beta}_0$ e $\hat{\beta}_1$ tem-se:

$$\begin{cases}-2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0\\ -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)x_i = 0\end{cases} \Leftrightarrow \begin{cases}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \quad (A)\\ \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)x_i = 0 \quad (B)\end{cases}$$

$$\Leftrightarrow \begin{cases}\sum_{i=1}^{n} y_i - n\hat{\beta}_0 - \hat{\beta}_1\sum_{i=1}^{n} x_i = 0\\ \sum_{i=1}^{n} x_i y_i - \hat{\beta}_0\sum_{i=1}^{n} x_i - \hat{\beta}_1\sum_{i=1}^{n} x_i^2 = 0\end{cases}$$

Isolando $\hat{\beta}_0$ na primeira equação segue-se que

$$\sum_{i=1}^{n} y_i - n\hat{\beta}_0 - \hat{\beta}_1\sum_{i=1}^{n} x_i = 0 \Leftrightarrow \hat{\beta}_0 = \frac{\sum_{i=1}^{n} y_i}{n} - \hat{\beta}_1\frac{\sum_{i=1}^{n} x_i}{n} = \bar{y} - \hat{\beta}_1\bar{x}.$$

Logo, o estimador de mínimos quadrados para $\hat{\beta}_0$ será:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}. \qquad (5)$$

Substituindo o resultado (5) na segunda equação, $\sum_{i=1}^{n} x_i y_i - \hat{\beta}_0\sum_{i=1}^{n} x_i - \hat{\beta}_1\sum_{i=1}^{n} x_i^2 = 0$, e resolvendo em relação a $\hat{\beta}_1$ tem-se:

$$\sum_{i=1}^{n} x_i y_i - \hat{\beta}_0\sum_{i=1}^{n} x_i - \hat{\beta}_1\sum_{i=1}^{n} x_i^2 = 0$$

$$\sum_{i=1}^{n} x_i y_i - (\bar{y} - \hat{\beta}_1\bar{x})\sum_{i=1}^{n} x_i - \hat{\beta}_1\sum_{i=1}^{n} x_i^2 = 0$$

$$\sum_{i=1}^{n} x_i y_i - \left(\frac{\sum_{i=1}^{n} y_i}{n} - \hat{\beta}_1\frac{\sum_{i=1}^{n} x_i}{n}\right)\sum_{i=1}^{n} x_i - \hat{\beta}_1\sum_{i=1}^{n} x_i^2 = 0$$

$$\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n} + \hat{\beta}_1 \frac{(\sum_{i=1}^{n} x_i)^2}{n} - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0$$

$$-\hat{\beta}_1 \frac{(\sum_{i=1}^{n} x_i)^2}{n} + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}$$

$$\hat{\beta}_1 \left[\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n} \right] = \sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}}$$

Logo, o estimador de mínimos quadrados para $\hat{\beta}_1$ é:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{SPxy}{Sxx}. \qquad (6)$$

Portanto, os estimadores de mínimos quadrados para β_0 e β_1 são, respectivamente:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \text{ e } \hat{\beta}_1 = \frac{SPxy}{Sxx}$$

em que: $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ é a média da variável independente *x*; $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$ é a média da variável dependente *y*; $SPxy = \sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i) \frac{(\sum_{i=1}^{n} y_i)}{n}$ é a soma de produtos entre *x* e *y* e $Sxx = \sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}$ é a soma de quadrados de *x*.

Exemplo 3: Uma importante aplicação da análise de regressão na contabilidade é a estimação do custo. Ao coletar dados sobre volume e custo e usar o método de mínimos quadrados para desenvolver uma equação de regressão estimada relacionando volume e custo, um contador pode estimar o custo associado a um volume de manufatura particular. Considere a seguinte amostra de volumes de produção e os dados de custos totais referentes a uma operação de manufatura.

Volume de produção (unidades)	Custos totais (US$)
400	4.000
450	5.000
550	5.400
600	5.900
700	6.400
750	7.000

Com esses dados desenvolva uma equação de regressão estimada que possa ser usada para prever o custo total de determinado volume de produção.

Solução: Primeiramente vamos calcular as informações necessárias:

$$n = 6$$

$$\left.\begin{array}{l} \sum_{i=1}^{6} y_i = 33.700 \\ \sum_{i=1}^{6} y_i^2 = 184.930.000 \end{array}\right\} \Rightarrow Syy = 5.648.333{,}333$$

$$\left.\begin{array}{l} \sum_{i=1}^{6} x_i^2 = 2.077.500 \\ \sum_{i=1}^{6} x_i = 3.450 \\ \sum_{i=1}^{6} x_i y_i = 20.090.000 \end{array}\right\} \Rightarrow Sxx = 93.750 \text{ e } SPxy = 712.500$$

$$\bar{x} = 575$$

$$\bar{y} = 5.616{,}67$$

Agora, temos condições de determinar $\hat{\beta}_0$ e $\hat{\beta}_1$,

$$\hat{\beta}_1 = \frac{SPxy}{Sxx} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}} = \frac{20.090.000 - \frac{(3.450)(33.700)}{6}}{2.077.500 - \frac{(3.450)^2}{6}} = 7{,}6$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 5.616{,}67 - 7{,}6x575 = 1.246{,}67$$

Logo, o modelo de regressão estimado é: $\hat{y}_i = 1.246{,}67 + 7{,}6x_i$.

INTERPRETAÇÃO DO COEFICIENTE DA REGRESSÃO LINEAR SIMPLES

Na regressão linear simples, interpreta-se $\hat{\beta}_1$ como uma estimativa da alteração em *y* correspondente à alteração de uma unidade na variável independente.

Então pode-se dizer que o incremento de cada unidade no volume de produção provoca um aumento médio de US$ 7,6 no custo por unidade produzida.

Exemplo 4: Uma importante aplicação da análise de regressão na contabilidade é a estimação do custo. Ao coletar dados sobre volume e custo e usar o método de mínimos quadrados para desenvolver uma equação de regressão estimada relacionando volume e custo, um contador pode estimar o custo associado a um volume de manufatura particular. Considere a seguinte amostra de volumes de produção e os dados de custos totais referentes a uma operação de manufatura.

Volume de produção (unidades)	Custos totais (US$)
400	4.000
450	5.000
550	5.400
600	5.900
700	6.400
750	7.000

Represente os dados em um gráfico.

a) Ajuste uma reta de regressão linear simples para os dados e esboce-a no gráfico.

Exemplo 3: Dados referentes a porcentagem de participação de mercado de 10 empresas atacadistas e seus rendimentos anual em milhões de reais**.**

Empresa	Participação Mercado (%)	Rendimento (milhões de reais)
1	26	23
2	25	21
3	31	28
4	29	27
5	27	23
6	31	28
7	32	27
8	28	22
9	30	26
10	30	25

a) Ajustar os dados a um intervalo linear.

b) Uma empresa com 24% de participação de mercado, qual deverá ser seu rendimento anual?

Exemplo 4: A propulsão de um motor (*y*) é uma função da temperatura de exaustão (*x*) em ºF

quando outras importantes variáveis são mantidas constantes. Considere os dados:

Propulsão	temperatura
4.300	1.760
4.650	1.652
3.200	1.485
3.150	1.390
4.950	1.820
4.010	1.665
3.810	1.550
4.500	1.700
3.008	1.270

a) Represente os dados em um gráfico.

b) Ajuste uma reta de regressão linear simples para os dados e esboce-a no gráfico.