



UNIVERSIDADE FEDERAL DE UBERLÂNDIA

Apostila de Estatística

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

PROF: Vânia de F. L. Miranda

DISCIPLINA: Estatística

Ementa

1 – ESTATÍSTICA DESCRIPTIVA

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
Amostragem 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 “adivinar” 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 É 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 **estatística 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. É 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 auxílio 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 específico, ou obtemos dados não com uma finalidade específica, 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.

Nº	Estado civil	Grau de instrução	Nº 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

- Quais são variáveis qualitativas e quantitativas?
- Classifique-as em nominais, ordinais, discretas e contínuas?
- 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) *É 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:

PRODUÇÃO DE CAFÉ BRASIL — 1991-1995	
CABEÇALHO	TÍTULO
COLUNA INDICADORA	CABEÇALHO
ANOS	PRODUÇÃO (1.000 t)
1991	2.535
1992	2.666
1993	2.122
1994	3.750
1995	2.007

CORPO → RODAPÉ → FONTE: IBGE.

CASA OU CÉLULA

LINHAS

Observações importantes:

- **Dados qualitativos** devem ser apresentados em tabelas com as frequências absolutas $\mu = \frac{\sum_{i=1}^n x_i}{N}$, frequências relativas $\sum_{i=1}^n x_i = x_1 + x_2 + x_3 + \dots + x_n$ 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 $Md = \frac{x_n + x_{n+1}}{2}$, frequências relativas $Md = \frac{x_{n+1}}{n}$, 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 ($F_{r,i}$):** 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.

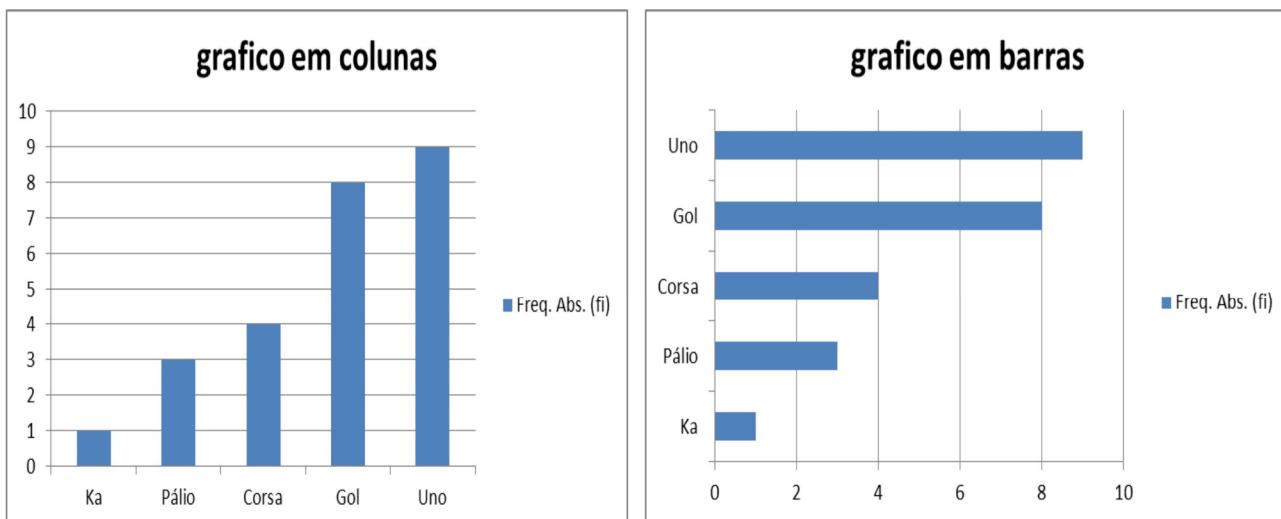
Gráfico em colunas ou em barras.

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.

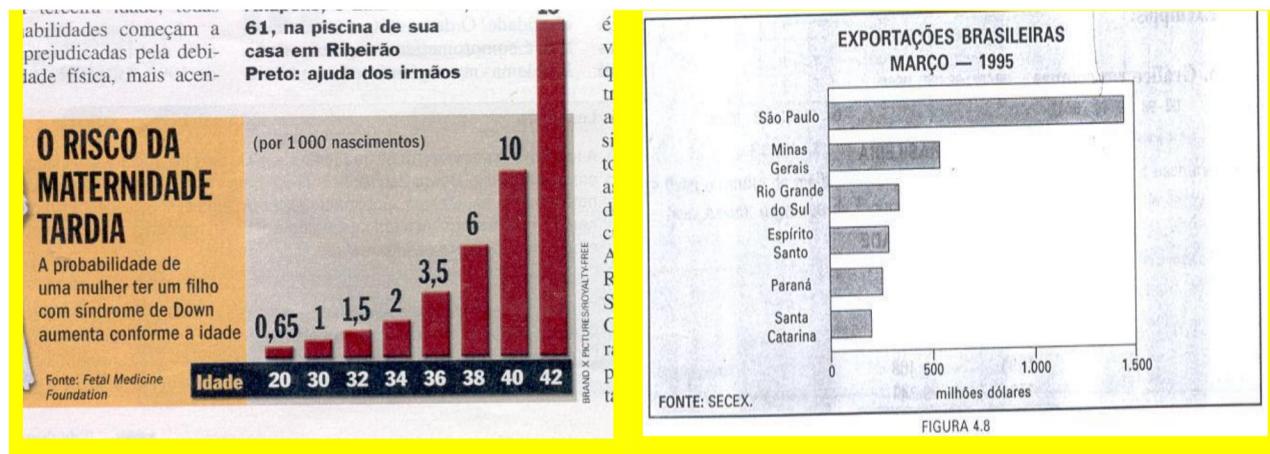


Gráfico em setores.

Este gráfico é constituído com base em um círculo. Pode ser construído através da fórmula:

$$n \rightarrow 360^\circ \quad F_i \rightarrow x \quad \Rightarrow nx = F_i * 360 \Rightarrow x = \frac{F_i * 360}{n}$$



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).

$$A = X_n - X_1 = \text{maior valor} - \text{menor valor}$$

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_{1a}): 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:

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

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;

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

- 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 ás 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: $LI_{1^a} = X_1 - \frac{c}{2} = 33 - \frac{10}{2} = 33 - 5 = 28$

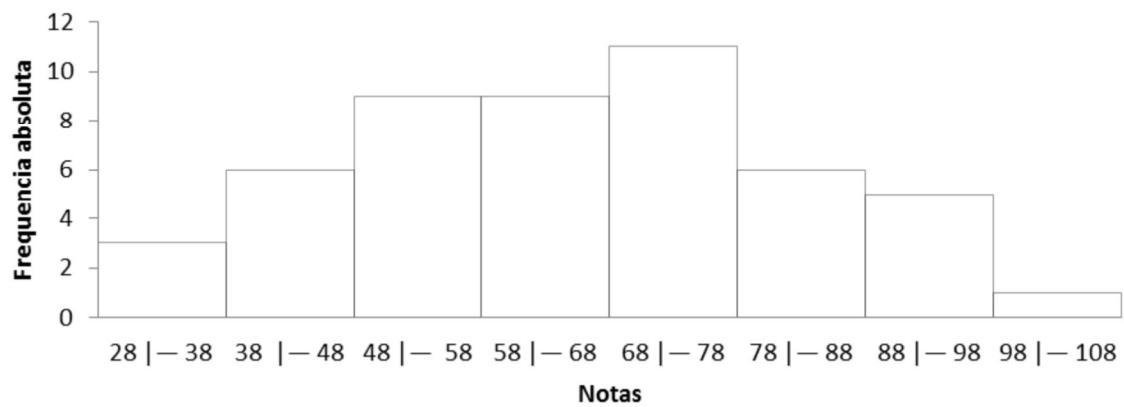
Limite Superior da primeira classe: $LS_{1^a} = LI_{1^a} + c = 28 + 10 = 38$

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)

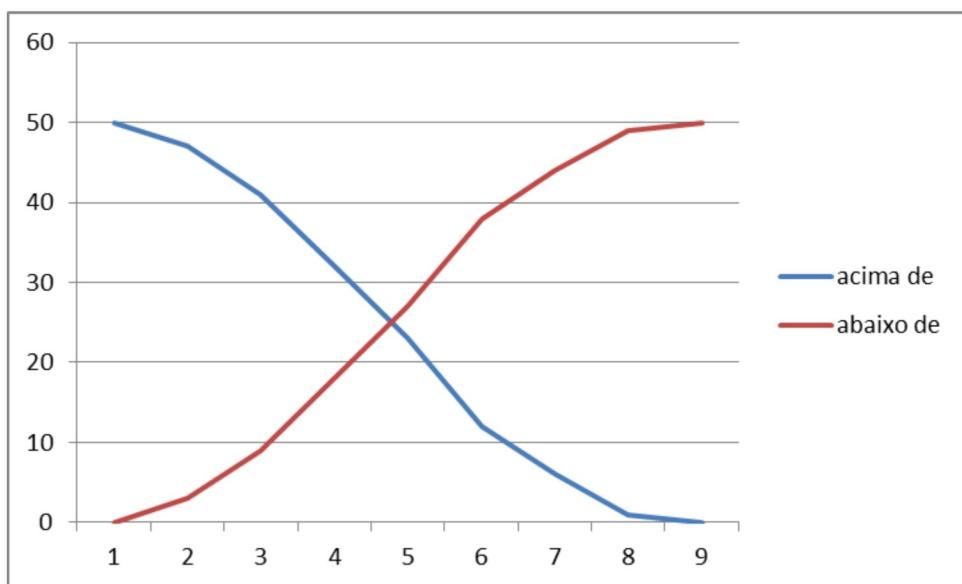
Notas do teste aplicado a 50 funcionários.



c)

Obser	acima de	obser	abaixo de
28	50	28	0
38	47	38	3
48	41	48	9
58	32	58	18
68	23	68	27
78	12	78	38
88	6	88	44
98	1	98	49
108	0	108	50

d)



INTERPOLAÇÃO EM DISTRIBUIÇÕES DE FREQUÊNCIA ACUMULADA

Exemplo: Dados da Tabela.

Limites (X_i)	$FC_i(X < X_i) = fac \downarrow$	$FC_i(X > X_i) = fac \uparrow$
-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 de 10?

$$7,73 \leftarrow \begin{cases} 5,245 \rightarrow 6 \\ 12,975 \rightarrow 14 \end{cases} \rightarrow 8$$

Aplicando a regra de três simples temos:

$$\begin{cases} 7,73 \rightarrow 8 \\ 4,755 \rightarrow x \end{cases} \Rightarrow x = \frac{8 * 4,755}{7,73} = 4,921$$

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

Qual a frequência acumulada acima de 10?

$$7,73 \leftarrow \begin{cases} 5,245 \rightarrow 14 \\ 12,975 \rightarrow 6 \end{cases} \rightarrow 8$$

Aplicando a regra de três simples temos:

$$\begin{cases} 7,73 \rightarrow 8 \\ 2,975 \rightarrow x \end{cases} \Rightarrow x = \frac{8 * 2,975}{7,73} = 3,079$$

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:

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

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):

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

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:

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

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:

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

Em 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 Σf_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):

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

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:

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

Em que \bar{x}_i é o ponto médio da classe e f_i é a frequência absoluta da classe i , para $i = 1, 2, \dots, k$ 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	1,2 5
0,35 0,55	10	0,45	4,5 0
0,55 0,75	8	0,65	5,2 0
0,75 0,95	17	0,85	14, 45

Σ	4	-	25,
0	40	-	40

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

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

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

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:

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

e a mediana amostral será determinada por:

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

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:

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

e a mediana amostra será determinada por:

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

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**):

$$Md = \frac{X\left(\frac{n}{2}\right) + X\left(\frac{n+1}{2}\right)}{2} = \frac{X\left(\frac{60}{2}\right) + X\left(\frac{60+1}{2}\right)}{2} = \frac{X_{(30)} + X_{(31)}}{2} = \frac{2+2}{2} = \frac{4}{2} = 2.$$

Exemplo 7: (n ímpar): Considere os dados amostrais de números de circuitos defeituosos em sistema composto por 4 circuitos. Uma amostra de 19 sistemas está 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:

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

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;

c é 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	4
		0

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 (**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, 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:

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

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_2 é 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.

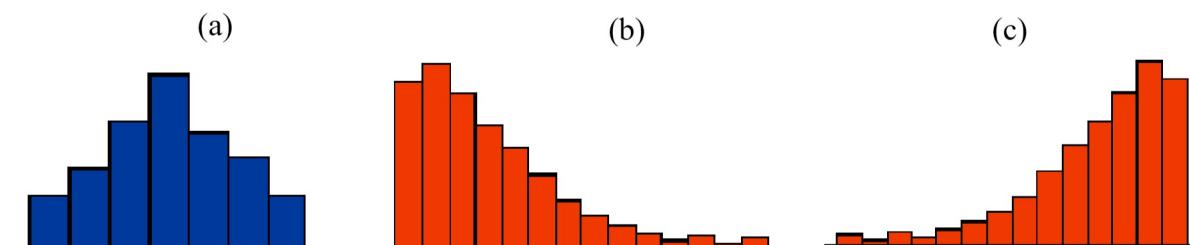


Figura 4.1 - Formas de distribuições em situações reais:

- (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: $\bar{x} = \frac{\sum_{i=1}^n x_i}{n} = \frac{200 + \dots + 510}{7} = 345,7$.

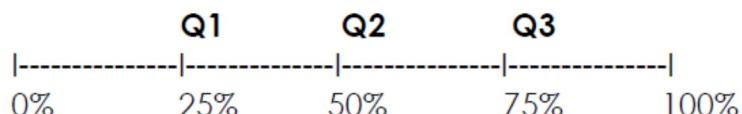
Esse valor representa, ou sintetiza razoavelmente, aquele conjunto de observações. Se incluirmos, entretanto, o salário de gerente do estabelecimento, os dados seriam: 200, 250, 250, 300, 450, 460, 510, 2300 e a média seria 601,4. Neste caso, não se pode dizer que a média sintetiza adequadamente o conjunto,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

1º Passo: Calcula-se $\frac{n}{4}$

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

3º Passo: Aplica-se a fórmula:

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

Para determinar Q_2 : igual a mediana

Para determinar Q_3 :

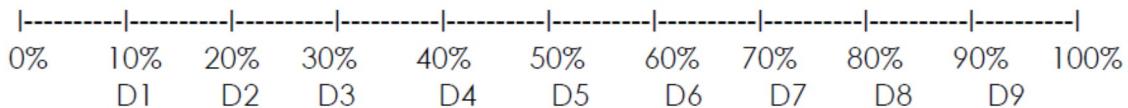
1º Passo: Calcula-se $\frac{3n}{4}$

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

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

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

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



Para determinar D_i :

1º Passo: Calcula-se $\frac{in}{10}$

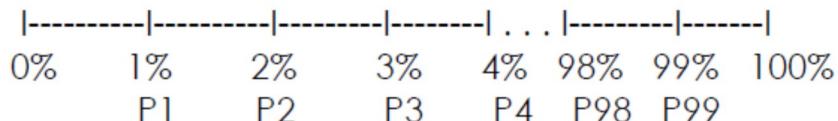
2º Passo: Identifica-se a classe D_i pela F_{ac}

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

3º Passo: Aplica-se a fórmula:

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 :

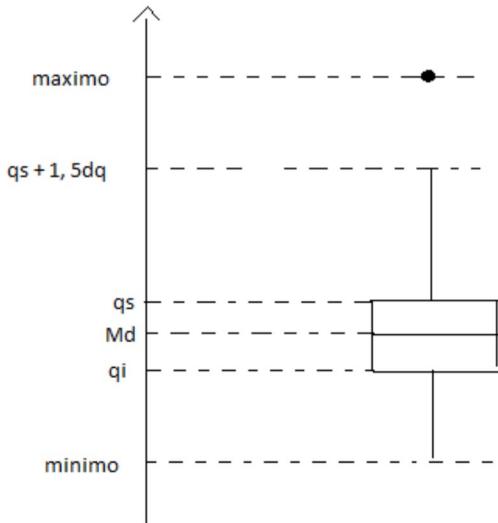
1º Passo: Calcula-se $\frac{in}{100}$

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

2º Passo: Aplica-se a fórmula:

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 é dividido no valor correspondente a mediana, assim ele indica o quartil inferior, a mediana e o quartil superior.



Outras medias

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

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

Em que w_i é 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%.

$$\bar{x} = \frac{\sum(w_i x_i)}{\sum w_i} = \frac{(15 \cdot 85) + (15 \cdot 90) + (15 \cdot 75) + (15 \cdot 80) + (40 \cdot 95)}{15 + 15 + 15 + 15 + 40} = \frac{8750}{100} = 87,5$$

- 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.

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

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 é dado

abaixo. Qual é a velocidade média? (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:** é usada em administração e economia para achar taxas médias de variação, de crescimento, ou razões médias. Dados n valores (todos positivos), a media geométrica é a raiz n^{ma} do seu produto.

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

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

$$\bar{x}_g = \sqrt[3]{2 \cdot 4 \cdot 10} = \sqrt[3]{80} = 4,3$$

Exercício 2: O fator de crescimento médio para o dinheiro, composto as taxas 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:** é 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 é:

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

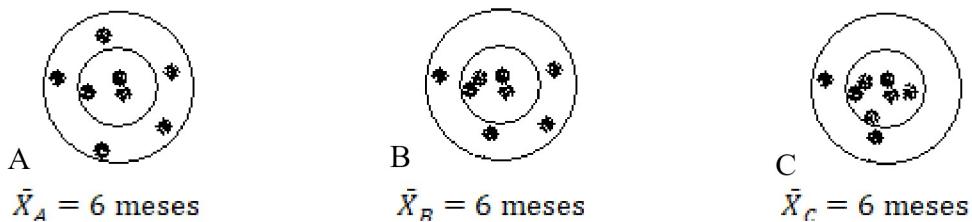
3)- MEDIDAS DE DISPERSÃO

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.



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:

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

O estimador da amplitude para dados agrupados em classe é:

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

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

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

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

$$A = 40.500 - 32.300 = 8.200 \text{ km.}$$

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
$0,15 \mid 0,35$	5	0,25
$0,35 \mid 0,55$	10	0,45
$0,55 \mid 0,75$	8	0,65
$0,75 \mid 0,95$	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 ($0,15 \mid 0,35$) 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:

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

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:

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

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:

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

Agora, temos condições de realizar o cálculo da variância:

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

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

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):

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

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:

$$\begin{aligned}
 s^2 &= \frac{1}{n-1} \left[\sum_{i=1}^n X_i^2 - \frac{(\sum_{i=1}^n X_i)^2}{n} \right] \\
 &= \frac{1}{8-1} \left[(40.000^2 + 40.500^2 + \dots + 32.300^2) \right. \\
 &\quad \left. - \frac{(40.000 + 40.500 + \dots + 32.300)^2}{8} \right] = \\
 &= \frac{1}{7} \left[1,128032 * 10^{10} - \frac{(299.600)^2}{8} \right] = \frac{1}{7} [1,128032 * 10^{10} - 1,122002 * 10^{10}] \\
 &= \frac{1}{7} [60.300.000] = 8.614.285,714 (\text{km})^2
 \end{aligned}$$

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:

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

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	-
	0	

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

$$\begin{aligned}
 s^2 &= \frac{1}{n-1} \left[\sum_{i=1}^k \bar{X}_i^2 f_i - \frac{(\sum_{i=1}^k \bar{X}_i f_i)^2}{n} \right] \\
 &= \frac{1}{40-1} \left[(0,25^2 * 5 + \dots + 0,85^2 * 17) - \frac{(0,25 * 5 + \dots + 0,85 * 17)^2}{40} \right]
 \end{aligned}$$

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

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:

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

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:

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

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

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:

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

Desvio Padrão Amostral:

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

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

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

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:

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

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:

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

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:

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

O coeficiente de variação **amostral** é:

$$CV = \frac{s}{\bar{x}} \times 100\% \quad 13)$$

Exemplo 8: A média e o desvio padrão do tempo de vida das lâmpadas de marca A e B são respectivamente: $\bar{X}_A = 4,0$ meses, $s_A = 0,8$ meses, $\bar{X}_B = 8,0$ meses e $s_B = 1,2$ 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:

$$CV_A = \frac{s_A}{\bar{X}_A} \times 100 = \frac{0,8}{4,0} \times 100 = 20\% \text{ e } CV_B = \frac{s_B}{\bar{X}_B} \times 100 = \frac{1,2}{8,0} \times 100 = 15\%.$$

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:

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

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

O erro padrão da média amostral é:

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

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)** é o conjunto de todos os possíveis resultados de um determinado experimento.

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

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

Outras definições importantes:

- i) Evento certo $\rightarrow \Omega$ (caracterizado pelo espaço amostral)
- ii) Evento impossível $\rightarrow \emptyset$.
- 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:

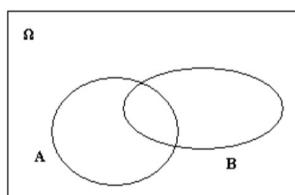
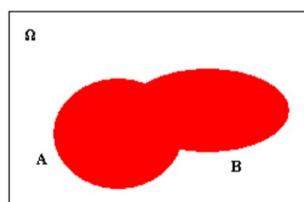
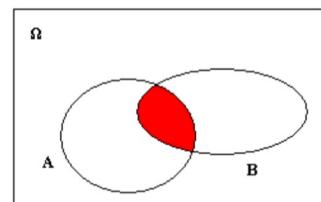


Figura1: Diagrama de Venn.

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

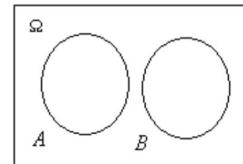
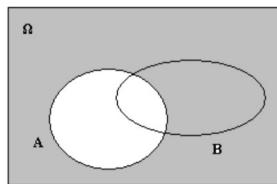


2) **Intersecção (I):** $A \cap B = B \cap A$



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

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



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** ($P(A)$) como a razão entre o número de possíveis resultados favoráveis ao evento A ($n(A)$) e todos os possíveis resultados do experimento ($n(\Omega)$), ou seja, número de elementos do espaço amostral.

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

Axiomas de Probabilidade

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

$$P(A) \geq 0.$$

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 “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, $A \cap B = \emptyset$, 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).