Ciências Biológicas

Cadernos CB Virtual 2

❖Rafael Angel Torquemada Guerra (Org.)

❖Amélia Iaeca Kanagawa ❖Creusioni Figueredo dos Santos

❖Fabiana Sena da Silva ❖Frederico Barbosa de Sousa

❖Gilmara Alves Cavalcanti ❖Jorge Adriano Lubenow

❖Marcio Bernardino da Silva ❖Maria Alice Neves

❖Roberto Menezes

EDITORA UNIVERSITÁRIA UFPB

UFPB VIRTUAL

CIÊNCIAS BIOLÓGICAS VIRTUAL

UAB UNIVERSIDADE ABERTA DO BRASIL

Universidade Federal da Paraíba
Universidade Aberta do Brasil
UFPB VIRTUAL
COORDENAÇÃO DO CURSO DE LICENCIATURA EM CIÊNCIAS BIOLÓGICAS À DISTÂNCIA

Caixa Postal 5046– Campus Universitário - 58.051-900 – João Pessoa
Fone: 3216-7838 e 8832-6059
Home-page: portal.virtual.ufpb.br/biologia

C 569 Cadernos Cb Virtual 2 / Rafael Angel Torquemada Guerra ... [et al.].-
João Pessoa: Ed. Universitária, 2011.
610p. : Il.
ISBN: 978-85-7745-902-5
Educação a Distância. 2. Biologia
I. Guerra, Rafael Angel Torquemada.
UFPB/BC CDU: 37.018.43

Estatística Vital

Gilmara Alves Cavalcanti

ESTATÍSTICA VITAL

Gilmara Alves Cavalcanti

UNIDADE 1
ANÁLISE DE DADOS ESTATÍSTICOS

1. Situando a Temática

A Estatística é considerada por alguns autores como Ciência no sentido do estudo de uma população. É considerada como método quando utilizada como instrumento por outra Ciência.
A palavra estatística frequentemente está associada à imagem de aglomeração de números, dispostos em uma imensa variedade de tabelas e gráficos, representando informações tão diversas como nascimentos, mortes, taxas, populações, rendimentos, débitos, créditos, etc. Isto é devido ao uso comum da palavra estatística como sinônimo de dados, como, por exemplo, quando falamos das estatísticas de uma eleição, estatísticas da saúde, estatísticas de acidente de trânsito ou as estatísticas de acidentes de trabalho.

No sentido moderno da palavra, estatística lida com o desenvolvimento e aplicação de métodos para coletar, organizar, analisar e interpretar dados de tal modo que a segurança das conclusões baseada nos dados pode ser avaliada objetivamente por meio de proposições probabilísticas.

O propósito da estatística não é exclusivo de qualquer ciência isolada. Ao contrário, a estatística fornece um conjunto de métodos úteis em toda área científica onde haja a necessidade de se coletar, organizar, analisar e interpretar dados. Estes métodos podem ser usados tão eficazmente em farmacologia como em engenharia, biologia, em ciências sociais ou em física.

2. Problematizando a Temática

Ao estudarmos fenômenos naturais, econômicos ou biológicos tais como, a precipitação de chuvas em uma determinada região, a evolução da taxa de inflação em uma região metropolitana, a influência das marés no desenvolvimento de animais marinhos, etc., estamos lidando com experimentos cujos resultados não conhecemos e desejamos saber se as hipóteses que afirmamos são verdadeiras, isto é, se os fenômenos estão ocorrendo como esperávamos. Para isto, é necessário que os dados oriundos das observações possam nos dar informações claras e precisas. Estes dados devem ser organizados de forma adequada para podermos fazer uma análise crítica e fundamentada do fenômeno.

A partir de agora você está convidado a participar de uma experiência que consiste em obter um conjunto de dados, representá-lo em distribuições de frequências e apresentá-lo através de tabelas e gráficos. Verá como algumas medidas estatísticas podem nos auxiliar nesta análise e como utilizá-las.

3. Conhecendo a Temática

3.1. Conceitos Básicos de Estatística

"Podemos considerar a Estatística como um conjunto de métodos e processos quantitativos que serve para estudar e medir os fenômenos coletivos".

A estatística teve acelerado seu desenvolvimento a partir do século XVII, através dos estudos de Bernoulli, Fermat, Pascal, Laplace, Gauss, Galton, Pearson, Fisher, Poisson e outros que estabeleceram suas características essenciais. A Estatística tem como OBJETIVO o estudo dos fenômenos coletivos.

A Estatística é a ciência que trata da coleta, do processamento e da disposição dos dados.

Objetivando o estudo quantitativo e qualitativo dos dados (ou informações), obtidos nos vários campos da atividade científica, a Estatística manipula dois conjuntos de dados fundamentais: a ***"população"*** e a ***"amostra"***.

População (ou Universo)

É o conjunto dos seres, objetos ou informações que interessam ao estudo de um fenômeno coletivo segundo alguma(s) característica(s). É, portanto, um conjunto definido de informações relativas a qualquer área de interesse, podendo, quanto ao número de elementos, ser: finita (tamanho N) ou infinita. Na maioria das vezes não é conveniente, ou mesmo possível, realizar o levantamento dos dados referentes a todos os elementos de uma população. Portanto, analisamos parte da população, isto é, uma amostra.

Amostra

É um subconjunto não vazio ou parte da população. Duas considerações devem ser feitas sobre o estudo amostral dos fenômenos. Uma diz respeito aos cuidados que se deve tomar para assegurar que a amostra seja representativa da população. Para atender a essa exigência, deve-se selecionar os elementos de forma aleatória, de modo que todo e qualquer elemento da população tenha a mesma chance de participar da amostra, a outra diz respeito à precisão dos dados coletados, buscando minimizar os erros que poderiam induzir a conclusões equivocadas. O número de elementos de uma amostra é chamado o *tamanho* da amostra, e denotado por ***n***.

☑ Definição 1.1: Parâmetro

Uma característica numérica estabelecida para toda uma população é denominada ***parâmetro*****.** São valores, geralmente desconhecidos (e que, portanto, têm de ser estimados), que representam certas características da população.

☑ Definição 1.2: Estimador

É uma característica baseada em observações amostrais e usada para indicar o valor de um parâmetro populacional desconhecido.

☑ Definição 1.3: Estimativa

O valor numérico assumido pelo estimador numa determinada amostra é denominada ***estimativa***.

Exemplo 1.1:

No fenômeno coletivo eleição para reitor da UFPB, a população é o conjunto de todos os eleitores habilitados na Universidade. Um **parâmetro** é a proporção de votos do candidato A. Uma amostra pode ser um grupo de 300 eleitores selecionados em toda a UFPB. Um **estimador** é a proporção de votos do candidato A obtida na amostra. O valor resultante do estimador, a proporção amostral, é a **estimativa.**

Processos Estatísticos de Abordagem

Quando solicitados a estudar um fenômeno coletivo podemos optar entre os seguintes processos estatísticos:

a) **CENSO** – Avaliação direta de um parâmetro, utilizando-se todos os componentes da população. Entre as principais características de um Censo, podemos destacar: admite erro processual zero e tem confiabilidade 100%, caro, lento e quase sempre desatualizado. Nem sempre é viável.
b) **AMOSTRAGEM** (INFERÊNCIA) – Avaliação indireta de um parâmetro, com base em um estimador através do cálculo das probabilidades. Entre as principais características, podemos destacar: admite erro processual positivo e tem confiabilidade menor que 100%, é barata, rápida e atualizada. É sempre viável.

Dados Estatísticos

Normalmente, no trabalho estatístico, o pesquisador se vê obrigado a lidar com grande quantidade de valores numéricos resultantes de um censo ou de uma amostragem. Estes valores numéricos são chamados ***dados estatísticos.***

No sentido da disciplina, a Estatística ensina métodos racionais para a obtenção de informações a respeito de um fenômeno coletivo, além de obter conclusões válidas para o fenômeno e também permitir tomada de decisões, através dos dados estatísticos observados. Desta forma, a estatística pode ser dividida em duas áreas: **Estatística Descritiva** e **Estatística Inferencial**.

Estatística Descritiva

É a parte da Estatística que tem por objetivo ***descrever*** os dados observados. A Estatística Descritiva, na sua função de descrição dos dados, tem as seguintes atribuições:

- A obtenção dos dados estatísticos;
- A organização dos dados;
- A redução dos dados;
- A representação dos dados e
- A obtenção de algumas informações que auxiliam a descrição do fenômeno observado.

A obtenção ou coleta dos dados é normalmente feita através de um questionário ou de observação direta de uma população ou amostra. **A organização dos dados** consiste na ordenação e crítica quanto à correção dos valores observados, falhas humanas, omissões, abandono de dados duvidosos, etc. **A redução dos dados** envolve o entendimento e a compreensão de grande quantidade de dados através de simples leitura de seus valores individuais é uma tarefa extremamente árdua e difícil mesmo para o mais experimentado pesquisador. **A representação dos dados** compreende de técnicas para uma melhor visualização dos dados estatísticos, facilitando sua compreensão. Por exemplo, os gráficos, quando bem representativos, tornam-se importantes instrumentos de trabalho. É ainda atributo da Estatística Descritiva a obtenção de algumas informações que sumarizam os dados, facilitando a descrição dos fenômenos observados.

Estatística Inferencial (ou *Indutiva*)

É a parte da Estatística que tem por objetivo obter e generalizar conclusões para a população a partir de uma amostra. Complementando o processo descritivo, a Estatística Indutiva estuda parâmetros a partir do uso de estimadores usando o cálculo das probabilidades, elemento este que viabiliza a Inferência Estatística.

Dados ou Variáveis Estatísticas

As informações ou dados característicos dos fenômenos ou populações são denominados ***variáveis estatísticas*** ou simplesmente ***variáveis***. Conforme suas características particulares podem ser classificadas da seguinte forma:

- **Quantitativas:** São aquelas que podem ser expressas em termos numéricos. Em geral são as resultantes de medições, enumerações ou contagens. São subdivididas em ***contínuas*** e ***discretas*****,** conforme abaixo:

    - ***Contínuas*****:** São aquelas que podem assumir qualquer valor num certo intervalo de medida, podendo ser associados ao conjunto dos números reais, ou seja, é um conjunto não enumerável. Entre outras, enquadram-se nesta categoria as medidas de tempo, comprimento, espessura, área, volume, peso, velocidade, dosagem de hemoglobina no sangue, concentração de flúor na água oferecida à população, etc.
    - ***Discretas*****:** Quando só podem assumir determinados valores num certo intervalo, ou seja, é um conjunto finito ou enumerável. Em geral, representam números inteiros resultantes de processo de contagem, como o número de alunos por sala, de créditos por disciplinas, de pacientes atendidos diariamente num hospital, etc.

De modo geral, as **medições** dão origem a variáveis contínuas e as **contagens ou enumerações,** a variáveis discretas**.** Designamos estas variáveis por letras latinas, em geral, as últimas: X, Y, Z.

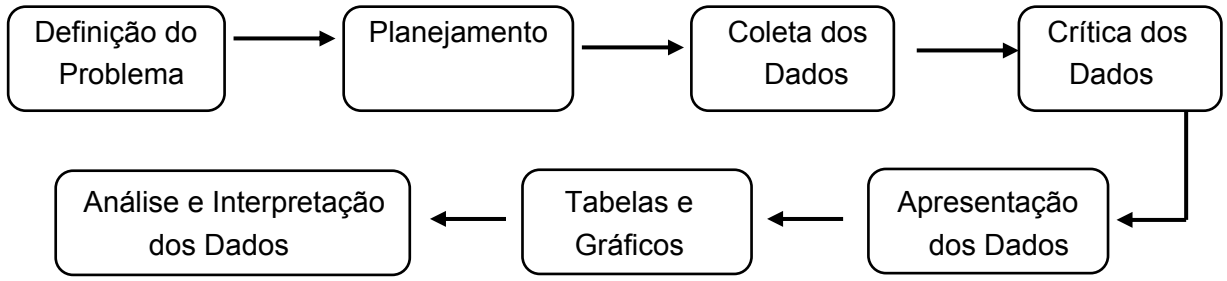
- **Qualitativas:** Nem sempre os elementos de uma população são exclusivamente contáveis. Muitas vezes, eles podem ser qualificados também segundo algumas de suas características típicas. Nesses casos, as variáveis podem ser agrupadas em ***nominais*** ou ***ordinais*** (por postos):

    - ***Nominais*****:** Quando puderem ser reunidas em categorias ou espécies com idênticos atributos. Aqui se incluem os agrupamentos por sexo, área de estudo, desempenho, cor, raça, nacionalidade e religião.
    - ***Ordinais*****:** Quando os elementos forem reunidos segundo a ordem em que aparecem dispostos numa lista ou rol. São típicas desta forma de agrupamento, variáveis como classe social, grau de instrução, entre outras.

Em geral, uma mesma população pode ser caracterizada por mais de um tipo de variável. Assim, os inscritos num vestibular, por exemplo, podem ser contados, medidos ou pesados, podem ser agrupados segundo o sexo ou área de estudo e podem ainda ser classificados segundo as notas obtidas nas provas prestadas.

3.2. Fases do Experimento Estatístico

Em linhas gerais, podemos distinguir no método estatístico as seguintes etapas:



3.2.1. Definição do Problema

Saber exatamente o que se pretende pesquisar, ou seja, definir corretamente o problema. Essa primeira fase consiste na formulação correta do problema a ser estudado.

3.2.2. Planejamento

É o trabalho inicial de coordenação no qual define-se a população a ser estudada estatisticamente, formulando-se o trabalho de pesquisa através da elaboração de questionário, entrevistas, etc.

A organização do plano geral implica em obter respostas para uma série tradicional de perguntas, antes mesmo do exame das informações disponíveis sobre o assunto, perguntas que procuram justificar a necessidade efetiva da pesquisa, a saber:

☺ *"quem", "o que", "sempre", "por que", "para que", "para quando".*

Imaginemos, por exemplo, que o Governo do Estado tenha necessidade de obter informações acerca do desempenho em Biologia dos estudantes matriculados na rede pública de ensino.

O primeiro trabalho da equipe encarregada da pesquisa será, evidentemente, o de obter respostas para aquelas perguntas. Seriam então:

* Quem deseja as informações?
* O que devemos perguntar no questionário?
* A pesquisa será periódica ou ocasional? Será executada sempre?
* Por que desejam as informações?
* Quando deverá estar concluída a pesquisa?
* Qual a época oportuna para a aplicação dos questionários?
* Para que desejam as informações?

Ainda na fase do planejamento, temos:

- O **Exame das Informações Disponíveis**: Trabalho inicial de coleta de trabalhos ou publicações sobre o assunto, obtendo-se relatórios sobre atividades semelhantes ou correlatas;
- A **Definição do Universo**: Isto é, saber qual o conjunto a ser pesquisado, distribuindo, classificando ou agrupando os elementos desse conjunto em subpopulações, para permitir um trabalho mais fácil, mais lógico, mais racional;
- O tipo de levantamento, **Censo** ou **Amostragem**: Deverá ser decidido com a devida antecedência e a necessária análise das vantagens e desvantagens de um e de outro, em virtude do custo financeiro e do prazo determinado para a conclusão do trabalho.

3.2.3. Coleta de Dados

Após cuidadoso planejamento e a devida determinação das características mensuráveis do fenômeno coletivamente típico que se quer pesquisar, damos início à **coleta dos dados** numéricos necessários à sua descrição.

A coleta dos dados poderá ser feita de diversas formas. A ideal é aquela que maximiza os recursos disponíveis, dados os objetivos e a precisão previamente estipulados. No seu planejamento, deve-se considerar o tipo de dado a ser coletado, o local onde este se manifestará, a freqüência de sua ocorrência, e outras particularidades julgadas importantes.

Quando os dados se referirem ou estiverem em poder de pessoas, sua coleta poderá ser realizada mediante respostas a questionários previamente elaborados. Esses questionários podem ser enviados aos entrevistados para devolução posterior ou podem ser aplicados pelos próprios pesquisadores ou por entrevistadores externos ou contratados.

Os dados ou informações representativas dos fenômenos ou problema em estudo podem ser obtidos de duas formas: ***por via direta*** ou *por* ***via indireta.***

➊ **Por Via Direta:** Quando feita sobre elementos informativos de registro obrigatório (Exemplo: fichas no serviço de ambulatório, nascimentos, casamentos, óbitos, matrículas de alunos etc.) ou, ainda, quando os dados são coletados pelo próprio pesquisador através de entrevistas ou questionários. A coleta direta de dados, com relação ao fator tempo, pode ser classificada em:

1.1. **Contínua:** Também denominada registro, é feita continuamente, tal como a de nascimentos e óbitos, etc. Também são do tipo contínuo o registro de certas doenças, como câncer, hanseníase, tuberculose e também algumas doenças infecciosas agudas com finalidade de controle.

1.2. **Periódica:** Quando feita em intervalos constantes de tempo, como os censos (de 10 em 10 anos), os balanços de uma farmácia, etc.;

1.3. **Ocasional:** Quando feita extemporaneamente, a fim de atender a uma conjuntura ou a uma emergência, como no caso de epidemias que assolam ou dizimam seres humanos

➋ **Por Via Indireta:** Quando é inferida de elementos conhecidos (coleta direta) e/ou conhecimento de outros fenômenos relacionados com o fenômeno estudado. Como exemplo, podemos citar a pesquisa sobre a mortalidade infantil, que é feita através de dados colhidos via coleta direta.

3.2.4. Crítica dos Dados

Os dados colhidos por qualquer via ou forma e não previamente organizados são chamados de **dados brutos**. Esses dados brutos, antes de serem submetidos ao processamento estatístico propriamente dito, devem ser ***"criticados"***, visando eliminar valores impróprios e erros grosseiros que possam interferir nos resultados finais do estudo. A crítica é ***externa*** quando visa às causas dos erros por parte do informante, por distração ou má interpretação das perguntas que lhe foram feitas; é ***interna*** quando se observa o material constituído pelos dados coletados. É o caso, por exemplo, da verificação de somas de valores anotados.

3.2.5. Apuração ou Processamento dos Dados (Apresentação dos Dados)

Uma vez assegurado que os dados brutos são consistentes, devemos submetê-los ao processamento adequado aos fins pretendidos. A apuração ou processamento dos dados pode ser **manual**, **eletromecânica** ou **eletrônica.** Os processos e métodos estatísticos aos quais os conjuntos de dados podem ser submetidos serão nosso objeto de estudo nas seções seguintes.

3.2.6. Exposição ou Apresentação dos Dados (Tabelas e Gráficos)

Por mais diversa que seja a finalidade que se tenha em vista, os dados devem ser apresentados sob forma adequada (**tabelas** ou **gráficos**), tornando mais fácil o exame daquilo que está sendo objeto de tratamento estatístico. No caso particular da estatística descritiva, o objetivo do estudo se limita, na maioria dos casos, à simples apresentação dos dados, assim entendida a exposição organizada e resumida das informações coletadas através de ***tabelas*** ou ***quadros*****,** bem como dos ***gráficos*** resultantes.

3.2.7. Análise e Interpretação dos Dados

Consiste em tirar conclusões que auxiliem o pesquisador a resolver seu problema, descrevendo o fenômeno através do cálculo de medidas estatísticas. O objetivo último da Estatística é tirar conclusões sobre o todo (população) a partir de informações fornecidas por parte representativa do todo (amostra). Realizadas as fases anteriores (**Estatística Descritiva**), fazemos uma análise dos resultados obtidos, através dos métodos da **Estatística Inferencial**, que tem por base a indução ou inferência, e tiramos desses resultados conclusões e previsões.

:: SAIBA MAIS... ::

Séries Estatísticas: São os dados organizados em forma de tabelas. De acordo com a época de ocorrência, o local e o fenômeno classificam-se, respectivamente, em:

a) **Série Temporal:** Os dados são observados segundo a época de sua ocorrência.
b) **Série Geográfica:** Os dados são observados segundo o local onde ocorreram.
c) **Série Especificativa:** Os dados são agrupados segundo a modalidade (espécie) de ocorrência.
d) **Série Mista ou de Dupla Entrada:** Fusão de duas ou mais séries simples. Pode ser visto como uma *Tabela de Contingência*, a qual ocorre quando os elementos da amostra ou da população são classificados de acordo com dois fatores.

Exemplos:

SÉRIE TEMPORAL

Tabela 1 – População brasileira no período de 1940 a 1970.

Anos	População
1940	41.236.315
1950	51.944.397
1960	70.119.071
1970	93.139.037

Fonte: Livro de Estatística.

SÉRIE GEOGRÁFICA

Tabela 2 – Região de origem de universitários. São Paulo, 2000.

Região	Quantidade
Urbana	240
Suburbana	1400
Rural	360
Total	2000

Fonte: Livro de Estatística.

SÉRIE ESPECIFICATIVA

Tabela 3 – Infecções por helmintos segundo a prevalência mundial. Brasil, 2001.

Helmintos	Percentual
Schistosoma	7
Enterobius	9
Necator	22
Ascaris	33
Total	

Fonte: Internet.

SÉRIE MISTA

Tabela 4 – Número de alunos segundo curso e sexo. Natal, 2003.

Curso	Sexo	
	M	F
Biologia	8	6
Ecologia	1	3
Total	9	9

Fonte: Livro de Estatística.

Entendi! Posso pensar na Estatística como números que resumem fatos e números puros dando-lhes algum significado. Ela apresenta ideias-chave que podem não estar imediatamente aparentes ao observar dados puros. Quando usamos a palavra “dados”, queremos dizer fatos ou números com base nos quais podemos tirar conclusões. Isso é legal! E essas tabelas? Ouvi dizer que existem normas técnicas para sua construção. Como são essas regras?

NORMAS TÉCNICAS PARA APRESENTAÇÃO TABULAR

De um modo geral tem-se a destacar em uma **tabela** (disposição escrita que se obtém referindo-se a uma coleção de dados numéricos a uma determinada ordem de classificação) os seguintes elementos essenciais (obrigatórios) e complementares (não-obrigatórios):

➔ **Elementos Essenciais:**

☑ ***Título:*** Indicação que precede a tabela e que contém a designação do fato observado, o local e a época em que foi registrado.

☑ ***Cabeçalho:*** Parte superior da tabela que especifica o conteúdo das colunas.

☑ ***Coluna Indicadora:*** Parte da tabela que especifica o conteúdo das linhas.

☑ ***Corpo da tabela:*** Conjunto de colunas e linhas que contêm as informações sobre a variável em estudo.

☑ ***Fonte:*** Entidade responsável pela informação.

➔ **Elementos Complementares:**

☑ ***Notas:*** Informações de natureza geral destinadas a conceituar ou esclarecer o conteúdo das tabelas ou a indicar a metodologia adotada no levantamento ou na elaboração dos dados.

☑ ***Chamadas*****:** Informações de natureza específica sobre determinada parte da tabela, destinada a conceituar ou a esclarecer dados***.***

☑ ***Sinais Convencionais*****:** Nenhuma casa da tabela deve ficar em branco, apresentando sempre um símbolo, a saber:

- – (hífen): quando o valor numérico é nulo;
- … (reticência): quando não se dispõe de dado;
- ? (ponto de interrogação): quando há dúvidas quanto à exatidão do valor numérico;
- 0,0: quando o valor numérico é muito pequeno para ser expresso pela unidade utilizada. Se os valores são expressos em números decimais, acrescenta-se o mesmo número de casas decimais ao valor zero;
- *x* (letra *x*): quando o dado for omitido a fim de evitar individualização da informação.

☺ As tabelas apresentadas oficialmente devem atender às normas da ABNT (resolução 886 de 20/10/60).

3.3. Estatística Descritiva

A **Estatística Descritiva** é a parte da estatística que se ocupa com a coleta, crítica, ordenação e apresentação das informações fundamentais à caracterização e descrição do fenômeno que se deseja estudar e interpretar. Aqui se trabalhará com alguma característica notável do objeto de estudo, a qual terá de ser coletada de alguma forma e em algum lugar. Na coleta das informações deve-se considerar, preferencialmente, toda a população; caso a obtenção de dados sobre toda a população (***censo***) seja difícil ou até mesmo impossível (dado o grande número de elementos ou a sua dispersão no tempo ou no espaço), o estudo poderá ser feito com base numa ***amostra representativa***.

3.3.1. Distribuições de Frequência

Os dados numéricos, após coletados, são colocados em série e apresentados em tabelas ou quadros. Quando se estuda uma variável (qualitativa ou quantitativa), o maior interesse do pesquisador é conhecer a distribuição dessa variável através das possíveis realizações (valores) da mesma. Iremos, pois, ver uma maneira de se dispor um conjunto de valores, de modo a se ter uma boa idéia global sobre esses valores, ou seja, de sua distribuição.

Uma distribuição de frequências pode ser apresentada nas seguintes maneiras:

- **Distribuição de Frequências por Valores** (variável qualitativa ou quantitativa discreta): É construída considerando-se todos os diferentes valores ou categorias, levando em consideração suas respectivas repetições.
- **Distribuição de Frequências por Intervalos ou Classes** (variável quantitativa): Constroem-se classes de valores, levando em consideração o número de valores que pertencem a cada classe e quando a variabilidade dos dados é grande. A construção de tabelas de freqüências para variáveis contínuas necessita de certos cuidados.

Exemplo 1.1: A Tabela 1 apresenta a distribuição de frequência da variável PROCEDÊNCIA, a partir dos dados do Quadro 1.

Tabela 1 – Freqüências e percentuais dos 46 estudantes de EV, segundo a região de procedência. João Pessoa, 1997.

Procedência	Nº Estudantes (F_i)	Percentual (f_i %)
Capital	20	43,5
Interior	16	34,8
Outra Região	10	21,7
Total	46	100,0

FONTE: Quadro 01.

Quadro 1 – Informações sobre sexo, curso, idade (anos), procedência, renda familiar, número de disciplinas matriculado(a), peso (kg) e altura (cm) de 46 alunos matriculados na disciplina Estatística Vital (EV).

ID	SEXO	CURSO	IDADE (Anos)	PROCEDÊNCIA	RENDA FAMILIAR	Nº. DISCIP. MATRIC.	PESO (kg)	ALTURA (cm)
1	Fem	Física	19	Interior	Média	6	47	156
2	Masc	Matem.	18	Capital	Média	6	75	167
3	Fem	Matem.	18	Outra Região	Média	6	61	169
4	Fem	Matem.	18	Capital	Média	6	56	163
5	Masc	Matem.	18	Capital	Média	6	80	178
6	Fem	Matem.	20	Interior	Média	6	44	158
7	Fem	Matem.	20	Interior	Média	6	52	158
8	Masc	Matem.	19	Capital	Média	6	67	174
9	Fem	Matem.	19	Outra Região	Média	3	48	167
10	Masc	Matem.	18	Capital	Média	6	83	180
11	Fem	Matem.	18	Capital	Média	6	53	163
12	Masc	Matem.	21	Outra Região	Média	5	66,5	175
13	Masc	Matem.	18	Interior	Média	6	78	180
14	Fem	Matem.	18	Interior	Não Info.	6	46	158
15	Fem	Matem.	18	Capital	Média	6	54	160
16	Fem	Matem.	19	Capital	Média	6	56	162
17	Fem	Matem.	19	Capital	Média	7	53	160
18	Fem	Matem.	18	Capital	Média	6	57	164
19	Fem	Física	23	Outra Região	Média	6	53	160
20	Masc	Matem.	18	Interior	Média	6	76	180
21	Masc	Matem.	21	Outra Região	Média	6	65	171
22	Masc	Matem.	19	Capital	Média	6	78,5	180
23	Masc	Matem.	19	Outra Região	Média	6	104	183
24	Fem	Matem.	17	Interior	Média	6	47,5	155
25	Masc	Matem.	18	Interior	Baixa	6	67,5	175
26	Masc	Matem.	19	Outra Região	Média	6	61	160
27	Masc	Matem.	17	Interior	Não Info.	6	68	169
28	Masc	Matem.	21	Interior	Média	5	75	178
29	Fem	Matem.	18	Interior	Média	5	58	154
30	Masc	Matem.	21	Outra Região	Média	6	65	165
31	Masc	Matem.	21	Capital	Média	6	67	178
32	Fem	Matem.	18	Capital	Alta	6	47	167
33	Masc	Matem.	21	Capital	Média	5	69	179
34	Fem	Matem.	19	Outra Região	Média	6	68	170
35	Masc	Matem.	18	Capital	Média	6	53	166
36	Fem	Matem.	17	Capital	Média	6	51	153
37	Fem	Matem.	19	Capital	Média	6	63	168
38	Masc	Matem.	19	Capital	Média	6	60	166

39	Masc	Matem.	18	Capital	Média	6	72	174
40	Masc	Matem.	21	Interior	Média	5	54	163
41	Masc	Matem.	18	Interior	Baixa	6	60	165
42	Masc	Matem.	19	Interior	Média	6	75	181
43	Fem	Matem.	18	Capital	Média	6	52	160
44	Masc	Matem.	18	Outra Região	Média	6	100	175
45	Masc	Matem.	22	Interior	Média	6	80	179
46	Masc	Matem.	21	Interior	Média	6	50	166

FONTE: Questionário aplicado (Aula 24/03/97).

Exemplo 1.2: A Tabela 2 apresenta a distribuição de frequência da variável Nº DE DISCIPLINAS MATRICULADO(A), a partir dos dados do Quadro 1 (*Dados Agrupados sem Intervalos*).

Tabela 2 – Frequências e percentuais do número de disciplinas matriculadas dos 46 estudantes de EV. João Pessoa, 1997.

Nº Disciplinas Matriculadas (X_i)	Nº Estudantes (F_i)	Percentual (f_i %)
3	1	2,2
5	5	10,9
6	39	84,8
7	1	2,2
Total	46	100,0

FONTE: Quadro 1.

CONTRUINDO UMA DISTRIBUIÇÃO DE FREQUÊNCIAS POR CLASSES

Regras Básicas para Elaboração de uma Distribuição de Frequências por Classes ou Intervalos:

1. Colete *n* dados referentes à variável cuja distribuição será analisada. É aconselhável que *n* seja superior a 50 para que possa ser obtido um padrão representativo da distribuição.

2. Efetua-se um **ROL ESTATÍSTICO** (organização dos dados de forma crescente) nos Dados Brutos (aqueles ainda não organizados numericamente).

3. Identifique o menor valor (*LI* ou $X_{mín}$) e o maior valor (*LS* ou $X_{máx}$) da amostra.

4. Calcule a **AMPLITUDE TOTAL** (*AT*) dos dados: $AT = X_{máx} - X_{mín}$.

5. Escolhe-se convenientemente o número de classes ***k*** (inteiro); $5 \le k \le 15$, onde podemos tomar:

$$k \cong \sqrt{n} \text{ ou } k \cong 1 + 3{,}3\log(n), \text{ se } n \ge 50$$

6. Calcule o **comprimento de cada classe** (*h*) dos dados: $h = \dfrac{AT}{k}$

É aconselhável construir classes de mesma amplitude.

7. Efetua-se o **AGRUPAMENTO EM CLASSES,** calculando os limites de cada classe:

1ª Classe:
Limite Inferior: $LI_1 = X_{\min}$ Limite Superior: $LS_1 = LI_1 + h$

2ª Classe:
Limite Inferior: $LI_2 = LS_1$ Limite Superior: $LS_2 = LI_2 + h$

$\vdots$

i-ésima Classe:
Limite Inferior: $LI_i = LS_{i-1}$ Limite Superior: $LS_i = LI_i + h$

Continue estes cálculos até que seja obtido um intervalo que contenha o maior valor da amostra ($X_{máx}$) entre seus limites.

8. Construa a **tabela de distribuição de frequências.**

Uma **tabela de distribuição de frequências (por classes ou valores)** deverá conter as seguintes colunas:

- Número de ordem de cada classe (*i*) ou valor.
- Limites de cada classe (no caso da distribuição de frequências por classes).
  - As classes são fechadas à esquerda e abertas a direita.
  - As observações iguais ao limite superior da classe (*i – 1*), o qual é igual ao limite inferior da classe *i*, pertencem à classe *i*. NOTAÇÃO: |------.
- Ponto Médio (PM_i) da *i*-ésima classe é denotado por: $PM_i = \frac{LI_i + LS_i}{2}$.
- Tabulação: Contagem dos dados pertencentes a cada classe ou a quantidade de vezes que o valor se repete.
- Frequência simples ou absoluta (F_i) da *i*-ésima classe ou do *i*-ésimo valor.
  - F_i = Número de observações da *i*-ésima classe (ou do *i*-ésimo valor).
  - Observe que: $\sum_{i=1}^{k} F_i = n$.
- Frequência Relativa (f_i) da *i*-ésima classe (ou do *i*-ésimo valor).
  - f_i = Número de observações da *i*-ésima classe (ou do *i*-ésimo valor) dividido pelo tamanho da amostra, isto é, $f_i = \frac{F_i}{n}$.
  - Observe que a soma de todos os valores de f_i deve ser igual a 1, ou seja, $\sum_{i=1}^{k} f_i = 1$.

    Multiplicando cada f_i por 100 obtém-se o percentual da classe (ou valor) correspondente, isto é, $f_i\% = f_i \times 100$.
- Existem outros tipos de frequências que também podem ser calculadas:
  - **Frequência Simples Acumulada (do tipo "abaixo de"):** Frequência simples acumulada da *i*-ésima classe ou valor:
    $$Fac_i = F_1 + F_2 + \cdots + F_i$$
  - **Frequência Relativa Acumulada:** Frequência relativa acumulada da *i*-ésima classe ou valor:
    $$fac_i = f_1 + f_2 + \cdots + f_i.$$

Exemplo 1.3: Elabore uma tabela de distribuição de frequências (dados agrupados em intervalos) da variável ALTURA, dos 46 estudantes de EV, usando-se os dados do Quadro 1.

☺ **Solução:**

Passo 1: Estabelecer o número de classes: $k \cong \sqrt{46} \cong 7$

Passo 2: Amplitude Total: $AT = 183 - 153 = 30$

Passo 3: Amplitude das Classes: $h = AT / k = 30/7 \cong 4{,}3$

Passo 4: Construção da Tabela de Distribuição de Frequências

Tabela 3 – Distribuição de frequências das alturas dos 46 estudantes de EV. João Pessoa, 1997.

Altura (X_i)	N^O Estudantes (F_i)	Percentual (f_i %)
153,0 |----- 157,3	4	8,7
157,3 |----- 161,6	8	17,4
161,6 |----- 165,9	7	15,2
165,9 |----- 170,2	10	21,7
170,2 |----- 174,5	3	6,5
174,5 |----- 178,8	6	13,0
178,8 |-----| 183,1	8	17,4
Total	46	100,0

FONTE: Quadro 1.

Exemplo 1.4: Elabore uma tabela de distribuição de frequências (dados agrupados em intervalos) da variável IDADE de 33 estudantes de CPE, conforme **Dados Brutos** abaixo:

DADOS BRUTOS

22	25	23	22	23	26	25	33	23	35
27	24	24	22	24	22	24	21	22	28
30	25	28	29	24	25	20	27	34	26
36	30	22							

ROL (DADOS ORDENADOS)

20	21	22	22	22	22	22	22	23	23
23	24	24	24	24	24	24	25	25	25
25	26	26	27	27	28	28	29	30	30
34	35	36							

☺ **Solução:**

Passo 1: Estabelecer o número de classes: $k \cong \sqrt{33} \cong 6$

Passo 2: Amplitude Total: $AT = 36 - 20 = 16$

Passo 3: Amplitude das Classes: $h = \frac{AT}{k} = \frac{16}{6} \cong 2{,}7$

Passo 4: Construção da Tabela de Distribuição de Frequências

Tabela 4 – Distribuição de frequências das idades de 33 estudantes de EV. João Pessoa, 1997.

Idade (X_i)	F_i
20,0 |----- 22,7	8
22,7 |----- 25,4	13
25,4 |----- 28,1	6
28,1 |----- 30,8	3
30,8 |----- 33,5	0
33,5 |-----| 36,2	3
Total	33

FONTE: Quadro 1.

:: ARREGAÇANDO AS MANGAS!! ::

DICA 1: Para construir a coluna indicadora (das classes) proceda da seguinte forma: o primeiro valor do intervalo 1 é o menor valor do rol, ou seja, 22. Esse é o LI (limite inferior) do primeiro intervalo. Para descobrir o seu LS (limite superior) some a 22 o valor de "*h*". Isto é, 20 + 2,7 = 22,7. Assim, o primeiro intervalo será: 22 |-- 22,7. Para o segundo intervalo repita o valor de 22,7 como sendo o LI do intervalo 2. Para descobrir o seu LS some ao valor de "*h*". Isto é, 22,7 + 2,7 = 25,4. Assim, o segundo intervalo será: 22,7 |-- 25,4. Para o terceiro intervalo repita o valor de 25,4 como sendo o LI do intervalo 3. Para descobrir o seu LS some ao valor de "h". Isto é, 25,4 + 2,7 = 28,1. Assim, o terceiro intervalo será: 25,4 |-- 28,1. Faça isso sucessivamente até que se encontre o número de classes.

DICA 2: Para construir a coluna da F_i perceba que no primeiro intervalo: **20,0** |--- **22,7**, o valor **20,0 é incluso** no intervalo (contabilizado no cálculo das frequências), no entanto, o valor de **22,7 não é**! Assim sendo, o valor de 22,7 só é incluso no intervalo seguinte. Volte ao rol e veja que existem 8 valores nesse intervalo (20 21 22 22 22 22 22 22). No segundo intervalo, **22,7** |--- **25,4**, o valor **22,7 é incluso** no intervalo (contabilizado no cálculo das frequências), no entanto, o valor de **25,4 não é**! Assim sendo, o valor de 25,4 só é incluso no intervalo seguinte. Volte ao rol e veja que existem 13 valores nesse intervalo (23 23 23 24 24 24 24 24 24 25 25 25 25). Siga o mesmo raciocínio para os demais intervalos!

➔ A Tabela 5, a seguir, é um exemplo de como calcular os **outros** tipos de frequências a partir da Tabela 3.

Exemplo 1.5: Elabore uma tabela de distribuição de frequências completa (dados agrupados em intervalos) da variável ALTURA de 46 estudantes de EV do Quadro 1.

☺ **Solução:**

Tabela 5 – Distribuição de frequências das alturas dos 46 Estudantes de EV. João Pessoa,1997.

Altura (X_i)	Frequência Absoluta (F_i)	Frequência Relativa (f_i)	Frequência Percentual (f_i %)	Freq. Abs. Acum. (Fac_i)	Freq. Relat. Acum. (fac_i)	Ponto Médio (pm_i)
153,0 |----- 157,3	4	0,087	8,7	4	0,087	155,15
157,3 |----- 161,6	8	0,174	17,4	12	0,261	159,45
161,6 |----- 165,9	7	0,152	15,2	19	0,413	163,75
165,9 |----- 170,2	10	0,217	21,7	29	0,630	168,05
170,2 |----- 174,5	3	0,065	6,5	32	0,695	172,35
174,5 |----- 178,8	6	0,130	13,0	38	0,825	176,65
178,8 |-----| 183,1	8	0,174	17,4	46	1,000	180,95
Total	46	1,000	100,0	---	---	---

FONTE: Quadro 1.

:: SAIBA MAIS... ::

☺ **Frequência Absoluta:** Corresponde a contagem dos valores do rol agrupados em cada intervalo. Não se preocupe, veja a dica 2 acima!

☺ **Frequência Relativa:** Calcule essa frequência pedindo ajuda a frequência absoluta da seguinte forma: divida cada valor da frequência absoluta pelo total. Assim, no intervalo 1 teremos, 4/46 = 0,087. No que se refere ao intervalo 2, faça 8/46 = 0,174. E assim por diante para cada intervalo! Lembre-se que a soma das frequências relativas é igual a 1.

☺ **Frequência Absoluta Acumulada "Abaixo de":** Repita o primeiro valor da frequência absoluta. Em seguida, some cada valor a seguir nessa mesma coluna. Isto é, 4 + 8 = 12, 12 + 7 = 19, 19 + 10 = 29, e assim por diante!

3.3.2. Representação Gráfica de Distribuições de Frequência

O **gráfico estatístico** é uma forma de apresentação dos dados estatísticos, cujo objetivo é produzir, no investigador ou no público em geral, uma impressão rápida e viva do fenômeno em estudo.

Para tornarmos possível uma representação gráfica, estabelecemos uma correspondência entre os termos da série estatística (tabela) e determinada figura geométrica, de tal modo que cada elemento da série seja representado por uma figura proporcional.

Requisitos:

A representação gráfica de um fenômeno deve obedecer aos seguintes requisitos primordiais:

- **Simplicidade** – Indispensável devido à necessidade de levar a uma rápida apreensão do sentido geral do fenômeno apresentado a fim de não nos perdermos na observação de minúcias de importância secundária;
- **Clareza** – O gráfico deve possibilitar uma correta interpretação dos valores representativos do fenômeno em estudo;
- **Veracidade** – Indispensável qualquer comentário, posto que, se não representa uma realidade, perde o gráfico sua finalidade.

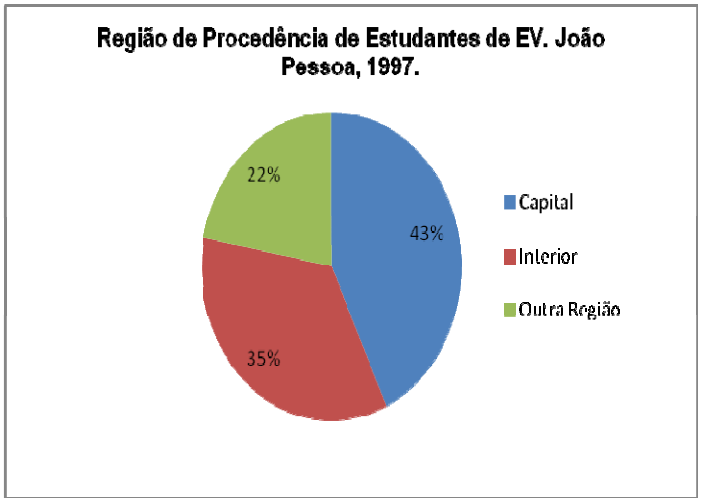
Os principais tipos de gráficos estatísticos para as distribuições de frequências são os **diagramas,** que são gráficos geométricos de, no máximo duas dimensões. Para sua construção, em geral, fazemos uso só do sistema cartesiano. Dentre os principais tipos de diagramas, destacamos:

☑ Diagrama de Setores (Gráfico de Pizza)

Funcionam dividindo seus dados em categorias ou grupos distintos. O gráfico consiste de um círculo dividido em fatias de pizza, cada qual representando um grupo. O tamanho de cada fatia é proporcional a quantidade de algo em cada grupo em comparação com os outros.

Quanto maior a fatia, maior a popularidade relativa daquele grupo. A quantidade de algo em cada grupo é chamada de ***frequência***.

Dividem seu conjunto inteiro de dados em grupos distintos. Isto é, se você somar a frequência de cada fatia, obterá 100%.



Fonte: Quadro 1.

:: SAIBA MAIS... ::

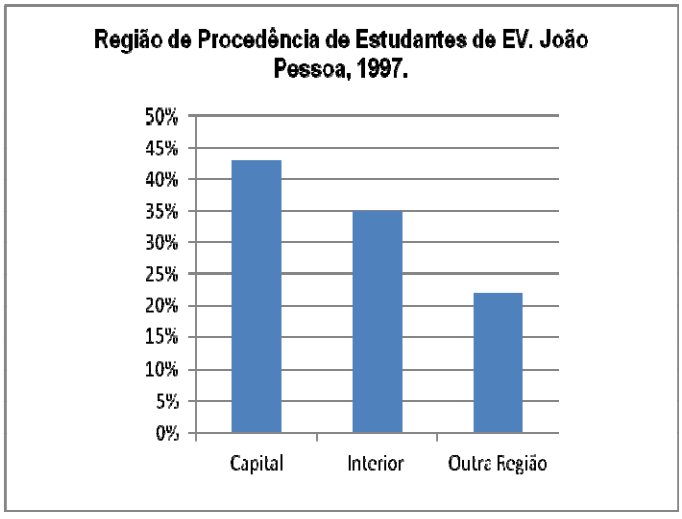
Os gráficos de setores podem ser úteis se você deseja comparar proporções básicas. Geralmente é fácil dizer à primeira vista quais grupos têm uma frequência alta em comparação aos outros. No entanto, essa forma gráfica é pouco útil se todas as fatias tiverem tamanhos semelhantes, pois se torna difícil visualizar diferenças sutis entre os tamanhos das fatias.

☑ Diagrama de Coluna/Barras

Permitem comparar tamanhos relativos, mas a vantagem de usar essa forma gráfica é que ela permite um maior grau de precisão.

São ideais em situações em que as categorias têm praticamente o mesmo tamanho, pois é possível identificar com muito mais precisão qual a categoria tem a frequência mais alta. Torna-se mais fácil enxergar as pequenas diferenças.

Cada coluna/barra representa uma determinada categoria, e o seu comprimento indica o valor. Todas as colunas/barras têm a mesma largura, o que facilita sua comparação, e quanto mais longa maior o valor.



Fonte: Quadro 1.

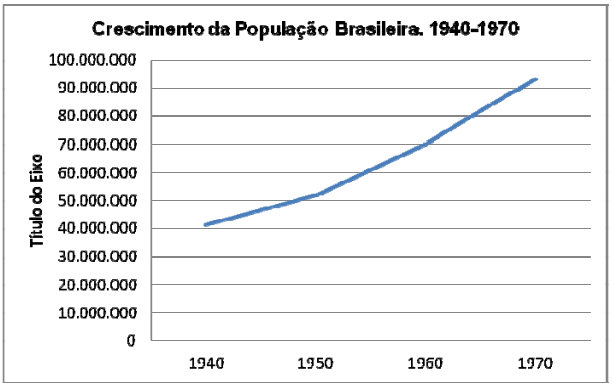
:: SAIBA MAIS... ::

O *"gráfico de colunas"* apresenta as categorias no eixo horizontal e a frequência ou porcentagem no eixo vertical. No *"gráfico de barras"* os eixos são invertidos. As categorias são mostradas no eixo vertical e a frequência no eixo horizontal. Portanto, as colunas são dispostas no sentido vertical e as barras no sentido horizontal. O gráfico de colunas tende a ser mais comum, mas gráficos de barras são úteis se os nomes de suas categorias forem muito longos, pois dão mais espaço para mostrar o nome de cada categoria.

☑ Diagrama de Linhas Simples/Em Faixa

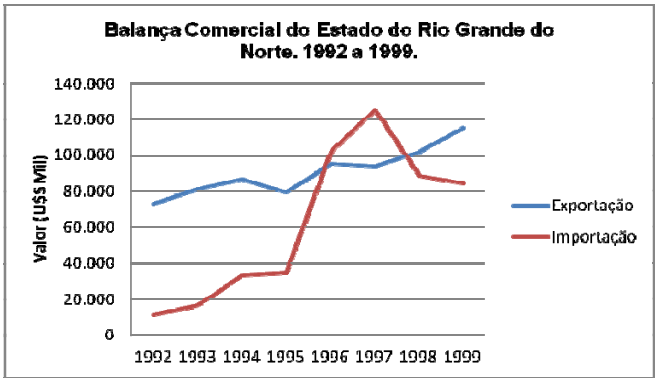
O diagrama de linhas simples é útil na representação de tabelas ou séries que evoluem ao longo do tempo (séries temporais), possibilitando a identificação de tendências. O diagrama de linhas em faixa é usado para comparar a evolução de duas variáveis e, ao mesmo tempo, a evolução de cada uma delas isoladamente.

Diagrama de Linhas Simples



Fonte: Apostila de Estatística.

Diagrama de Linhas em Faixa

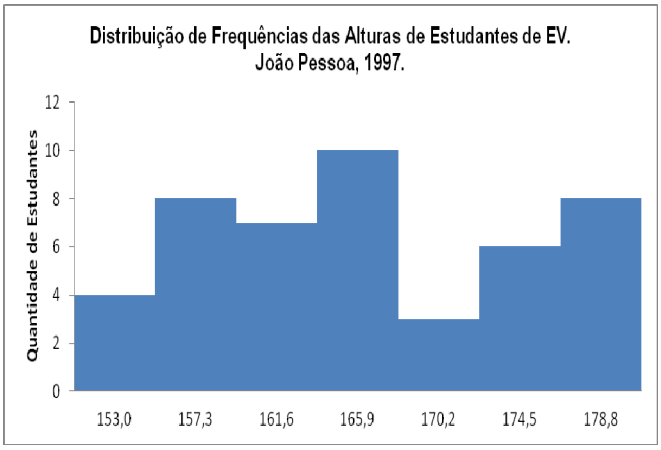


Fonte: Apostila de Estatística.

Perceba que todas essas formas gráficas estão associadas as variáveis qualitativas, ou seja, aquelas que expressam categorias. No caso das variáveis quantitativas discretas também podemos fazer uso de formas gráficas como diagrama de colunas/barras. No entanto, para representar as variáveis contínuas necessitamos de formas gráficas especificas como o ***histograma*** ou o ***polígono de frequências***.

☑ Histograma

Histogramas são como gráficos de colunas, mas com duas importantes diferenças. A primeira é que a área de cada coluna é proporcional à frequência, e a segunda é que não há espaço vazio entre as colunas no gráfico.

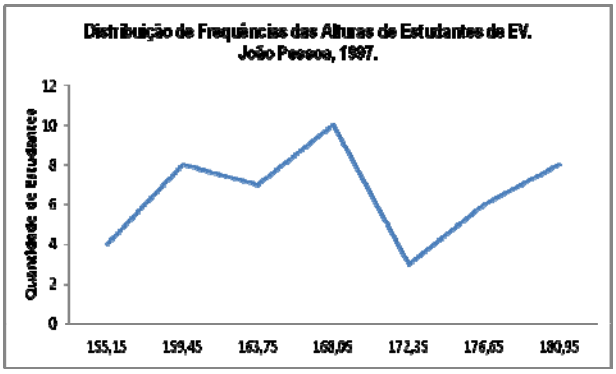


Fonte: Quadro 1.

É a representação gráfica de uma distribuição de frequências de variável quantitativa contínua (dados agrupados em intervalos) por meio de retângulos justapostos, centrados nos pontos médios das classes e cujas áreas são proporcionais às frequências das classes.

☑ Polígono de Frequência

É outra forma de representar graficamente uma distribuição de frequências de variável quantitativa contínua (dados agrupados em intervalos)



Fonte: Quadro 1.

Corresponde a uma linha poligonal traçada a partir do ponto médio de cada retângulo do histograma, cuja área total é igual a do histograma. Pode referir-se às frequências absolutas ou às frequências relativas, conforme a escala utilizada no eixo vertical.

3.4. Medidas Estatísticas

Vimos anteriormente a sintetização dos dados sob a forma de tabelas, gráficos e distribuições de frequências. Aqui, vamos aprender o cálculo de medidas que possibilitem representar um conjunto de dados (valores de uma variável quantitativa, isto é, informações numéricas), relativos à observação de determinado fenômeno de forma reduzida.

Os dados quantitativos, apresentados em tabelas e gráficos, constituem a informação básica do problema. É conveniente apresentar medidas que mostrem a informação de maneira resumida. Um conjunto de dados pode se reduzir a uma ou a algumas medidas numéricas que resumem todo o conjunto. Duas características importantes dos dados, que as medidas numéricas podem evidenciar são: *o valor central do conjunto* e a *dispersão dos números*.

Estes índices estatísticos são as **MEDIDAS DE POSIÇÃO** e, dentre as mais importantes, citamos as ***"Medidas de Tendência Central"***, que recebem tal denominação pelo fato dos dados observados tenderem, em geral, a se concentrar em torno de valores centrais. Dentre as medidas de tendência central, destacamos:

- **Média aritmética ou Média;**
- **Moda;**
- **Mediana.**

As outras medidas de posição são as ***"Separatrizes"***, que englobam:

- **Mediana;**
- **Quartis;**
- **Decis.**
- **Percentis.**

ÀS VEZES É PRECISO CHEGAR À RAIZ DO PROBLEMA!

Pode ser difícil identificar padrões e tendências em uma grande quantidade de números, e achar a **média** é geralmente o primeiro passo para conseguir enxergar o cenário mais geral. Com a média à sua disposição, é possível rapidamente achar os valores mais representativos dos seus dados e tirar importantes conclusões. Neste capítulo, vamos examinar várias formas de calcular alguns dos dados estatísticos mais importantes do pedaço (média, mediana e moda) e começar a entender como resumir dados com eficácia da forma mais concisa e útil possível.

3.4.1. Medidas de Tendência Central

São medidas que tendem para o centro da distribuição e tem a capacidade de representá-la como um todo. Dão o valor do ponto em torno do qual os dados se distribuem. As principais são: *Média Aritmética, Mediana, Moda*.

❶ Média Aritmética (ou simplesmente **MÉDIA)**

É a mais importante medida de tendência central, pois possui propriedades matemáticas convenientes.

☺ Notação: μ = Média da população ou média populacional.

$\overline{X}$ = Média da amostra ou média amostral.

É bem provável que já tenham lhe pedido para calcular a média antes. Uma forma de achar uma média de um conjunto de números é somar todos os números e, em seguida, dividir o resultado pela quantidade de números existentes.
Se você quer aprender Estatística será necessário se familiarizar com algumas notações estatísticas comuns. À primeira vista pode parecer um pouco estranho, mas você vai se acostumar.

LETRAS E NÚMEROS

Quase todos os cálculos estatísticos envolvem somar um conjunto de números. Como exemplo, se quisermos achar a idade média da turma, primeiro devemos somar as idades de todas as pessoas que frequentam essa turma.

O interessante é que se possa generalizar essa idéia para qualquer conjunto de números. Os estatísticos resolvem essa questão usando letras para representar números. Como exemplo, podemos usar a letra **X** para representar as idades dos alunos da turma da seguinte forma:

Idades específicas dos alunos da turma:		**Idades gerais dos alunos da turma:**
19 20 20 20 21	⟶	X_1 X_2 X_3 X_4 X_5

☺ Cada **X** representa a idade de uma pessoa da turma. É como identificar cada pessoa com um número específico.

Agora que temos uma forma geral de escrever as idades, podemos usar os nossos **X** para representá-las nos cálculos. Podemos escrever a soma das 5 idades da turma como:

$$\textbf{Soma} = X_1 + X_2 + X_3 + X_4 + X_5$$

☹ Mas e se não soubermos quantos números temos de somar? E se não soubermos quantas pessoas há na turma? Nesses casos não tem problema, basta chamar de ***n*** o número de valores. Assim, se não soubéssemos quantas pessoas havia na turma, diríamos que havia ***n*** pessoas e escreveríamos a soma de todas as idades como:

$$\textbf{Soma} = X_1 + X_2 + X_3 + X_4 + X_5 + \ldots + X_n = \Sigma X \text{ ou ainda } \Sigma X_i$$

☑ **Observação:** O símbolo **Σ** representa o somatório (Leia-se **ΣX** como somatório de **X**).

Podemos usar notação matemática para representar a média. Se juntarmos todas essa informações podemos escrever a média como:

$$\overline{X} = \frac{\sum X}{n}$$

Exemplo 1.6: Determinar a média do seguinte conjunto (amostra) de valores: 3, 7, 8, 10, 11.

☺ **Solução:** Perceba que, nesse caso, os dados estão em forma de rol.

$$\overline{X} = \frac{\sum X}{n} = \frac{3+7+8+10+11}{5} = 7{,}8$$

☑ **APLICAÇÃO:** Suponha que o exemplo acima represente o número de golfinhos vistos durante uma visita de 5 dias na localidade de Barra de Tabatinga (Litoral Sul de Natal/RN). Ao analisar a média pode-se observar que existe um número médio de 7,8 golfinhos na amostra.

TRABALHANDO COM FREQUÊNCIAS

Ao calcular a média de um conjunto de números, você irá, muitas vezes, perceber que alguns dos números são repetidos. Se você observar as idades da turma verá que na verdade temos 3 pessoas com 20 anos. É muito importante não deixar de incluir a *frequência* de cada número quando estiver calculando a média. Para termos certeza de não nos esquecermos disso, podemos incluí-la na fórmula. Usamos a letra ***f*** para representar a frequência.

$$\overline{X} = \frac{\sum (X_i . F_i)}{n}$$

Lembre-se que ao tratar com freqüências devemos nos lembrar que o conjunto de números pode estar agrupado em distribuições de frequências por valor e em classes. Assim podemos determinar a média seguindo o mesmo raciocínio. Acompanhe:

Exemplo 1.7 (Distribuição de Frequências por Valor): Determinar a média do seguinte conjunto (amostra) de valores: 2, 3, 8, 8, 5, 2, 2, 2, 8, 5, 3, 8, 2, 2, 5, 8, 2, 5, 8, 2

☺ **Solução:** Distribuição de Frequências por Valor

Dados Agrupados sem Intervalos

X	F	X.F
2	8	16
3	2	6
5	4	20
8	6	48
∑	20	90

$$\overline{X} = \frac{\sum (X_i . F_i)}{n} = \frac{90}{20} \Rightarrow \overline{X} = 4{,}5$$

☑ **APLICAÇÃO:** Suponha que o exemplo acima represente o número de filhotes nascidos vivos (*X*) observados em uma amostra de 20 ninhos de tartarugas marinhas (*F*). Ao analisar a média pode-se observar que existe um número médio de 4,5 filhotes nascidos vivos nessa amostra.

Lembre-se que no caso da distribuição de frequências por classes a coluna indicadora (X) corresponde aos intervalos. Para calcular a média é necessário fazer uso do ponto médio. Então:

$$\overline{X} = \frac{\sum (PM_i . F_i)}{n}$$

onde: PM_i: corresponde ao ponto médio de cada classe;

F_i: corresponde a freqüência absoluta de cada classe;

n: corresponde ao tamanho da amostra.

Exemplo 1.8 (Distribuição de Frequências em Classes): Utilizando os dados apresentados na Tabela 5, determine a altura média dos 46 estudantes de EV – Período 97.1 – Turma 01.

☺ **Solução:**

Altura (X)	Freqüência. Absoluta (F_i)	Ponto Médio PM_i	$PM_i.F_i$
153,0 |----- 157,3	4	155,15	620,60
157,3 |----- 161,6	8	159,45	1275,60
161,6 |----- 165,9	7	163,75	1146,25
165,9 |----- 170,2	10	168,05	1680,50
170,2 |----- 174,5	3	172,35	517,05
174,5 |----- 178,8	6	176,65	1059,90
178,8 |-----| 183,1	8	180,95	1447,60
Total ou $\sum$	46	---	7747,50

Então:

$$\overline{X} = \frac{\sum (PM_i . F_i)}{n} = \frac{7747,50}{46} = 168,42 \text{ cm}$$

➔ Análise: Na amostra observada de 46 alunos da disciplina de Estatística Vital verifica-se uma altura média de 168,42 cm.

☑ Vantagens e Desvantagens da Média:

É uma medida de tendência central que, por uniformizar os valores de um conjunto de dados, não representa bem os conjuntos que revelam tendências extremas. Ou seja, é **grandemente influenciada pelos valores extremos (grandes)** do conjunto. Além disso, não pode ser calculada para distribuições de freqüências com limites indeterminados (indefinidos).

☑ Propriedades:

1. A soma dos desvios tomados em relação à média é nula, isto é, $\sum_{i=1}^{n} \left(X_i - \overline{X} \right) = 0$.

2. 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, isto é, $Y_i = X_i \pm c \Rightarrow \overline{Y} = \overline{X} \pm c$.

3. 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, isto é, $Y_i = X_i \times c \Rightarrow \overline{Y} = \overline{X} \times c$ ou $Y_i = \frac{X_i}{c} \Rightarrow \overline{Y} = \frac{\overline{X}}{c}$, para $c \neq 0$.

:: SAIBA MAIS... ::

Se a média for enganosa devido a dados distorcidos e valores discrepantes então precisamos de alguma outra maneira de dizer qual é o valor típico. Isso pode ser feito tomando o valor do meio, ou central. Esse é um tipo diferente de média, ou medida de posição chamada de **mediana**.

Considere um conjunto de dados ***ordenados*** constituído de ***n*** valores. A mediana é o valor que divide o conjunto em duas partes iguais (isto é, em duas partes de 50% cada).

☺ Notação: *Med*

Para achar a mediana da turma, alinhe todas as idades em ordem crescente (rol) e, em seguida, escolha o valor central, da seguinte forma:

19 19 20 20 **20** 21 21 50 52

Este é o número central. É a mediana, 20.

E se o número de pessoas da turma fosse par?

19 20 20 20 21 21 50 52

Se houver número par de pessoas na turma, não haverá nenhum número exatamente no meio!

Nesse caso basta tirar a média dos dois números do meio (some-os e divida por 2), essa é a sua mediana. Neste caso, a mediana é 20,5.

☺ Se você tiver 9 números, a mediana é o número que está na posição 5.
☺ Se você tiver 8 números, ela é a média dos números que estão nas posições centrais: 4 e 5.
☹ E se você tiver *n* números? Nesse caso, observe o quadro abaixo:

COMO ACHAR A MEDIANA EM TRÊS PASSOS:
❶ Alinhe seus números em ordem crescente.
❷ Se você tiver um número ímpar de valores, a mediana é o valor que está no meio. Se tiver *n* números, o número do meio está na posição (n + 1)/2.
❸ Se você tiver um número par de valores, ache a mediana somando os dois valores do meio e dividindo por 2. Esses dois valores são os elementos que estão na posição: (n/2) e [(n/2) + 1].

Exemplo 1.9: Determinar a mediana da amostra a seguir: 3, 7, 8, 10, 11.

☺ **Solução**:

$$Med = \frac{n+1}{2} = \frac{5+1}{2} = \frac{6}{2} = 3^0 \; elemento \Rightarrow Me = 8$$

Exemplo 1.10: Determinar a mediana da amostra a seguir: 3, 7, 8, 10.

☺ **Solução**:

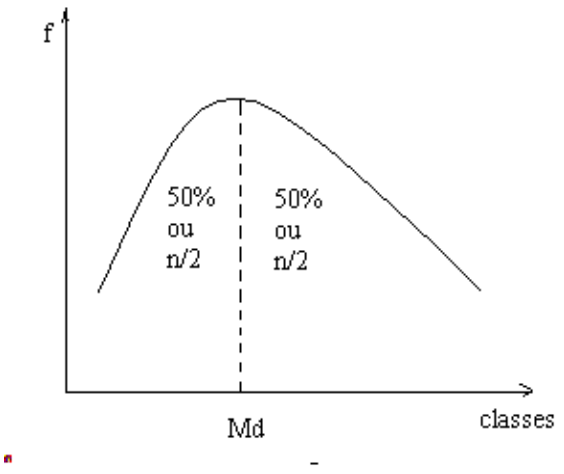
$$Med = \frac{(n/2)+[(n/2)+1]}{2} = \frac{(4/2)+[(4/2)+1]}{2} = \frac{2^0+3^0}{2} = \frac{7+8}{2} = 7{,}5$$

TRABALHANDO COM DISTRIBUIÇÃO DE FREQUÊNCIAS EM CLASSES

No caso de dados agrupados, relembramos que uma distribuição de frequências pode ser representada por meio de um Histograma. Dizemos então que a mediana será o valor de *X* (abscissa) cuja ordenada divide a área total do Histograma em duas partes iguais.

Em uma distribuição de frequências com dados agrupados em classes, denominamos **classe mediana** a classe que contém o elemento que está na posição (*n*/2) e, consequentemente, será esta a classe que conterá a **mediana**.

Figura 1: Cálculo da mediana para dados distribuídos em classes.



Cálculo da mediana:

$$Med = LI_{med} + \left(\frac{\frac{n}{2} - Fac_{ant}}{F_{med}} \right) \times h_{med}$$

onde:

* LI_{med} é o limite inferior da classe mediana;
* F_{med} é a frequência absoluta da classe mediana;
* Fac_{ant} é a frequência absoluta acumulada da classe anterior à classe mediana;
* h_{med} é a amplitude da classe mediana;
* n é o número de observações.

Assim, para dados agrupados em intervalos, a mediana é obtida através de interpolação de acordo com a fórmula dada na Figura 1.

Exemplo 1.11 (Distribuição de Frequências em Classes): Determinar a altura mediana dos 46 estudantes da turma de EV, - Período: 97.1, conforme os dados agrupados na Tabela 5.

☺ **Solução**:

- Classe mediana é a classe que contém o elemento que está na posição (*n*/2), ou seja, nesse caso, a classe mediana é a classe que contém o elemento que está na (46/2) = **23ª posição**.
- Logo, a classe mediana será a 4ª classe: 165,9 |----- 170,2 (**Classe mediana**: primeira classe que ultrapassar 50% (n/2) ou mais das observações).

➘ **Dica:** Para encontrar a classe mediana observe a coluna da Fac_{ant} perceba que no primeiro intervalo tem **4** números (assim teremos: o primeiro, o segundo, o terceiro e o quarto). No segundo intervalo teremos **4**+8 = 12 números (assim teremos: do quinto até o décimo segundo). No terceiro intervalo teremos 12+7 = 19 números (assim teremos: do décimo terceiro até o décimo nono). No quarto intervalo teremos 19+10 = 29 números (assim teremos: do vigésimo até o vigésimo nono). Pronto! Nessa classe temos o **23^0 elemento** essa é a classe que contém a mediana**!** Portanto,

- LI_{med} = 165,9
- F_{med} = 10
- h_{med} = 4,3
- Fac_{ant} = 19

Então: $$Med = LI_{med} + \left(\frac{\frac{n}{2} - F_{ant}}{F_{med}} \right) \times h_{med} = 165{,}9 + \left(\frac{\frac{46}{2} - 19}{10} \right) \times 4{,}3 = 165{,}9 + 1{,}72 = 167{,}62\ cm \cdot$$

☑ Propriedades da Mediana

1. A mediana não é influenciada por valores extremos (grandes) de uma série ou conjunto de dados;
2. A mediana de uma série de dados agrupados de classes extremas indefinidas pode ser calculada.

❸ Moda

Dado um conjunto ***ordenado*** de valores a moda é o valor mais popular. A moda é(são) o(s) valor(es) que ocorre(m) com maior frequência no conjunto de dados, ou seja é(são) o(s) valor(es) mais frequente(s). Ao contrário da média e da mediana, a moda tem de ser um valor existente no conjunto de dados.

As vezes os dados podem ter mais de uma moda. Se houver mais de um valor com a frequência mais alta, então cada um desses valores é uma moda. Acompanhe os tipos de séries modais a seguir:

☑ **Tipos de Séries Modais:** ☺ Notação: *Mo*

Um conjunto de valores pode ser classificado quanto a **moda** como:

* Unimodal (uma única moda) * Trimodal (três modas) * Amodal (não há moda)
* Bimodal (duas modas) * Polimodal (quatro modas)

Exemplo 1.12: Determine a moda dos seguintes conjuntos de dados abaixo:

a) 2, 2, 3, 3, 5, 5, 8, 8 $\Rightarrow$ Não existe moda (ou Amodal)
b) 2, 2, 3, 5, 5, 5, 8, 8 $\Rightarrow$ *Mo* = 5 (Unimodal)
c) 2, 2, 2, 3, 3, 5, 5, 5, 8 $\Rightarrow$ *Mo* = 2 e *Mo* = 5 (Bimodal)

TRABALHANDO COM DISTRIBUIÇÃO DE FREQUÊNCIAS

(a) Cálculo da Moda em uma Distribuição de Frequências por Valor

Basta olhar o elemento que ocorre com maior frequência, ou seja, aquele que mais se repete.

Exemplo 1.13: Suponha o **Exemplo 1.7**. Nesse caso, a maior frequência absoluta (F_i) é igual a 8 e o seu elemento correspondente (X) é 2. Portanto, *Mo* = 2. Ao analisar a moda pode-se observar que a maioria dos ninhos apresentaram 2 filhotes nascidos vivos.

(b) Cálculo da Moda em uma Distribuição de Frequências por Classes

Em uma distribuição de frequências em classes, denominamos **classe modal** a classe que possui a maior frequência, e, consequentemente, será esta classe que conterá a **moda**.

Exemplo 1.14: Utilizando os dados apresentados na Tabela 5, determine a altura modal (Moda) para dados agrupados em intervalos, a partir da fórmula de Czuber apresentada na Figura 2.

☺ **Solução**:

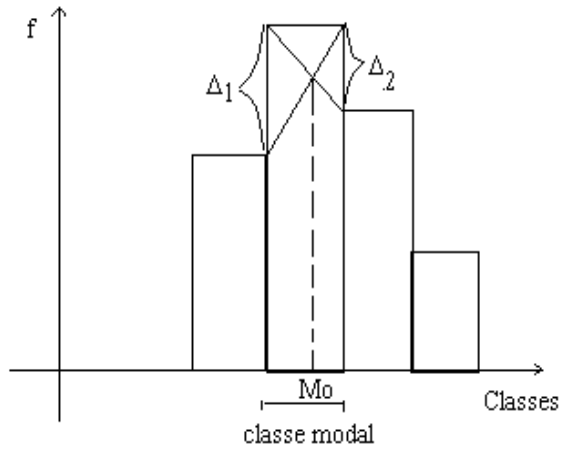
- A classe modal será o intervalo com maior frequência absoluta (*f*). Neste caso a classe modal será: 165,9 |----- 170,2 (4ª classe). Assim, L_{mo} = 165,9.
- h_{mo} = 170,2 – 165,9 = 4,3.
- $\Delta_1 = f_{modal} - f_{anterior}$ = 10 – 7 = 3, bem como, $\Delta_2 = f_{modal} - f_{posterior}$ = 10 – 3 = 7

Daí,

$$Mo = L_{mo} + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2} \right) \times h_{mo} = 165{,}9 + \left(\frac{3}{3+7} \right) \times 4{,}3 = 167{,}19 \text{ cm.}$$

➔ Análise: Na amostra observada de 46 alunos da disciplina de Estatística Vital verifica-se que a maioria deles medem 167,19 cm.

Figura 2: Cálculo da moda para dados distribuídos em classes.



FÓRMULA de CZUBER (Interpretação geométrica através de Histograma):

$$Mo = L_{mo} + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2} \right) \cdot h_{mo}$$

onde:

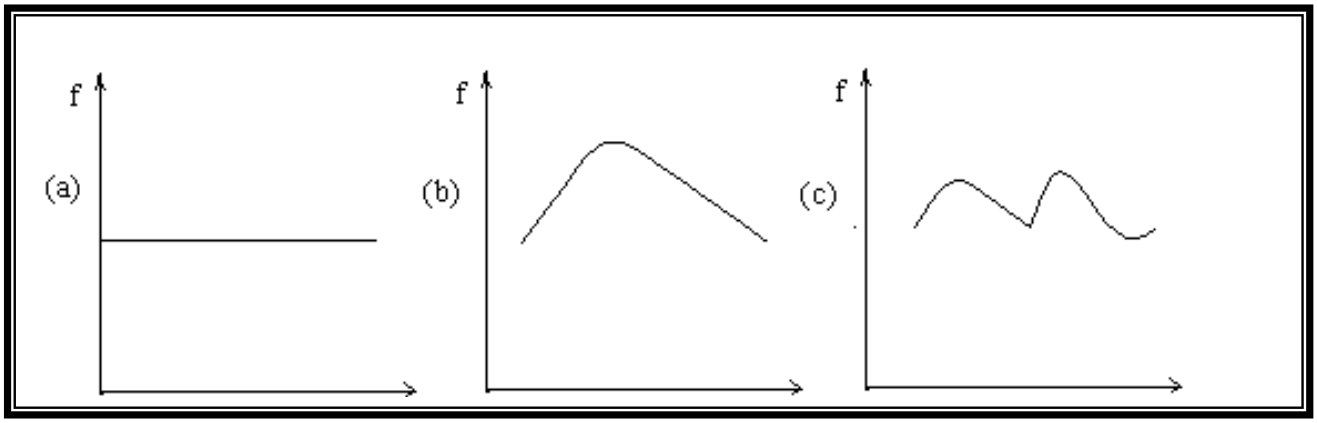
* L_{mo}: Limite inferior da classe modal
* L_{mo}: Amplitude da classe modal
* $\Delta_1 = f_{modal} - f_{anterior}$
* $\Delta_2 = f_{modal} - f_{posterior}$

☑ Observações:

(i) A moda de um conjunto de dados pode não existir (Figura 3 (a)).

(ii) A moda de um conjunto de dados pode não ser única (Figura 3 (c)).

Figura 3: Caracterização de dados quanto à moda.



☑ Vantagens e Desvantagens da Moda

- Não depende de todos os valores do conjunto de dados, podendo mesmo não se alterar com a modificação de alguns deles;
- Não é influenciada por valores extremos (grandes) do conjunto de dados
- Pode ser calculada para distribuições com limites indeterminados (indefinidos) na maioria dos casos.

3.4.2. Medidas de Dispersão

Não se preocupe com o jantar. Quando tiver um forno com desvio padrão mais baixo, você nunca mais vai queimar nada.

Nem tudo é confiável, mas como saber?

As medidas de posição fazem um excelente trabalho fornecendo-lhe um valor típico para seu conjunto de dados, mas elas **não lhe contam a história completa**. Tudo bem, você sabe onde está o centro dos seus dados, mas, muitas vezes, a média, a mediana e a moda sozinhas não são informações suficientes em situações em que você esteja resumindo um conjunto de dados. Nesta unidade, vamos mostrar como dar um passo a mais no seu nível de conhecimento de dados à medida que começamos a analisar **amplitudes** e **variação**.

O quadro a seguir apresenta as notas de 5 avaliações aplicadas em uma turma com 4 alunos. O professor deseja premiar o melhor aluno com uma bolsa de estudo. A questão é qual deles escolher? Cada aluno tem a mesma média de pontos, $\overline{X}_{Antônio} = \overline{X}_{João} = \overline{X}_{José} = \overline{X}_{Pedro} = 5$, mas há diferenças nítidas entre cada conjunto de dados. Precisamos encontrar uma forma de medir essas diferenças.

Alunos	Notas	Média
Antônio	5 5 5 5 5	5
João	6 4 5 4 6	5
José	10 5 5 5 0	5
Pedro	10 10 5 0 0	5

Observando-os detalhadamente, notamos que em cada grupo, os valores se distribuem diferentemente em relação à média. Necessitamos assim de uma medida estatística complementar para melhor caracterizar cada conjunto apresentado.

Podemos diferenciar cada conjunto de dados observando a forma em que os pontos se dispersam em relação a uma medida de posição. As pontuações de cada aluno são distribuídas de forma diferente, e, se medirmos como os pontos estão dispersos, o professor poderá tomar uma decisão mais embasada.

As medidas estatísticas responsáveis pela variação ou dispersão dos valores de um conjunto de dados são as medidas de dispersão ou de variabilidade, onde se destacam ***a amplitude total, a variância, o desvio padrão e o coeficiente de variação***. Em princípio, diremos que entre dois ou mais conjuntos de dados, o mais disperso (ou menos homogêneo) é aquele que tem a maior medida de dispersão.

❶ Amplitude Total

A amplitude nos diz quantos números os dados abrangem, como se estivéssemos medindo sua largura. Para calculá-la tomamos o maior número do conjunto de dados (chamado de ***limite superior*** – LS) e, em seguida, subtraímos do menor (chamado de ***limite inferior*** – LI):

$$AT = LS - LI$$

☺ Notação: AT

Exemplo 1.15: Com base no exemplo anterior, a amplitude de cada aluno é,

☺ **Solução:**

* $AT_{Antônio} = 5 - 5 = 0$
* $AT_{João} = 6 - 4 = 2$
* $AT_{José} = 10 - 0 = 10$
* $AT_{Pedro} = 10 - 0 = 10$

☹ As notas de José e Pedro têm a mesma amplitude, mas os valores são distribuídos de forma diferente. Será que a amplitude realmente mede bem a dispersão dos dados?

A amplitude:

☑ Só descreve a largura dos dados e não como eles são dispersos entre os limites.

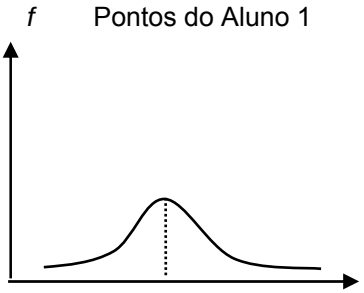
☑ Pode medir até que ponto os valores estão dispersos, mas é difícil ter uma idéia real de como os dados são distribuídos.

☑ É uma excelente forma rápida de ter uma idéia de como os valores são distribuídos, mas é um pouco limitada.

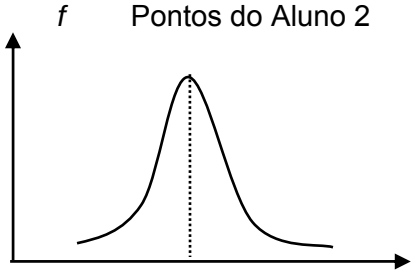
TRABALHANDO COM A VARIABILIDADE

Não queremos só medir a dispersão de cada conjunto de pontos. Queremos alguma forma de usar isso para ver se um aluno é merecedor da bolsa. Em outras palavras, queremos medir a variabilidade das notas dos alunos.

Uma maneira de fazer isso é observar a distância entre cada valor e a média aritmética. Se conseguirmos determinar um tipo de média para a distância entre os valores e a média aritmética, teremos uma forma de medir a variação e a dispersão. Quanto menor o resultado, mais próximos os valores estão da média.



Os valores estão dispersos a uma boa distância da média. Se o professor escolher este aluno, ele provavelmente não conseguirá prever se o aluno se sairá bem na prova. O aluno pode obter pontuação bastante alta se estiver em um bom dia. Mas, em um dia ruim, ele poderá não tirar uma nota tão alta assim, e isso significa comprometer a bolsa.

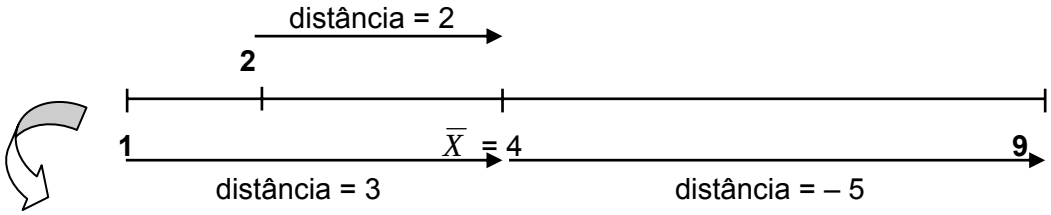


Os valores para este segundo conjunto de dados estão muito próximos da média e variam menos. Se o professor escolher esse aluno, ele terá uma boa ideia de como o aluno poderá se sair bem na prova.

☺ Então isso significa que basta calcular a média da distância entre os valores e a média aritmética!

CALCULANDO DISTÂNCIAS MÉDIAS

Suponha que você tenha três números: 1, 2 e 9. A média é igual a 4. O que acontece se acharmos a média da distância entre esses valores e a média aritmética?



Distância média = $\frac{(1\ a\ \overline{X})+(2\ a\ \overline{X})+(9\ a\ \overline{X})}{3}=\frac{3+2+(-5)}{3}=0$

A distância média entre os valores e a média aritmética é sempre 0, visto que as distâncias positivas e negativas se cancelam umas às outras. Como resolver esse problema? Precisamos achar uma maneira de fazer com que todas as distâncias sejam positivas!

Que tal elevarmos as distâncias ao quadrado? É provável que todos os números sejam positivos.

(Distância média)^2 = $\frac{3^2+2^2+(-5)^2}{3}=\frac{9+4+25}{3}=12{,}67$ (*até 2 casas decimais*)

Dessa vez, observa-se um número significativo, pois as distâncias não se cancelam! Ao somarmos os quadrados das distâncias obtemos um resultado não-negativo, sempre. Esse método de medir a dispersão é chamado de **variância**.

❷ Variância

☺ Notação: σ^2 é a variância da população ou variância populacional

S^2 é a variância da amostra ou variância amostral.

A variância de um conjunto de dados mede a variabilidade do conjunto em termos de desvios quadrados em relação à média aritmética. É uma quantidade sempre não negativa e expressa em unidades quadradas do conjunto de dados, sendo de difícil interpretação.

$$S^2 = \frac{\sum (X_i - \overline{X})}{n-1}$$

☑ **Observação:** Por se tratar de uma amostra o denominador da variância, S^2, é associado a um fator de correção: (n – 1).

TRABALHANDO COM FREQUÊNCIAS

(a) Distribuição de Frequências por Valor: Sejam X_1, X_2, ..., X_k as medidas da variável de interesse, realizadas para uma amostra de tamanho *n* extraída da população considerada. Definimos a variância da amostra (S^2) como:

$$S^2 = \frac{\sum (X_i - \overline{X}).F_i}{n-1}$$

onde:

X_i é o *i*-ésimo valor da variável de interesse; $\overline{X}$ é a média da amostra;
F_i é a freqüência absoluta do *i*-ésimo valor; *n* é o tamanho da amostra.

☑ **Observação:** A equação acima é utilizada quando nosso interesse não se restringe à descrição dos dados, mas, partindo da amostra, visamos tirar inferências válidas para uma respectiva população.

(b) Distribuição de Frequências por Classes Sejam pm_1, pm_2, ..., pm_k os pontos médios das classes, ocorrendo com frequências F_1, F_2, ..., F_k de modo que $\sum_{i=1}^{k} F_i = n$. A variância da amostra (S^2) é definida por como:

$$S^2 = \frac{\sum (pm_i - \overline{X}).F_i}{n-1}$$

onde:

pm_i é o ponto médio da *i*-ésima classe; $\overline{X}$ é a média da amostra;
F_i é a frequência absoluta da *i*-ésima classe; *n* é o tamanho da amostra.

:: SAIBA MAIS... ::

☺ Os estatísticos utilizam muito a variância para medir a dispersão dos dados. Ela é útil, pois utiliza cada valor para chegar ao resultado, e pode ser vista como uma média do quadrado das distâncias entre os valores e a média aritmética. No entanto, **o que realmente deseja-se é um número que nos dê a dispersão em termos da distância entre os valores e a média aritmética, e não a distância elevada a quadrado.**

☹ O problema com a variância é que pode ser difícil pensar em dispersão em termos de distâncias elevadas ao quadrado.	☺ Há uma maneira de corrigir isso. Basta tirar a raiz quadrada da variância. A isso chamamos de **desvio-padrão**.

❸ Desvio-Padrão

☺ Notação: σ é o desvio-padrão da população ou desvio-padrão populacional.
S é o desvio-padrão da amostra ou desvio-padrão amostral.

É outra medida de dispersão mais comumente empregada do que a variância, por ser expressa na mesma unidade do conjunto de dados. Mede a **"dispersão absoluta"** de um conjunto de valores e é obtida a partir da variância. É uma forma de dizer a que distância os valores típicos estão da média aritmética.

Desvio Padrão = $\sqrt{Variância}$ (Raiz quadrada da Variância).

Assim,

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

☑ Quanto menor o desvio-padrão, mais próximos os valores estão da média aritmética.
☑ O menor valor que o desvio-padrão pode assumir é 0.

❹ Coeficiente de Variação

É uma medida que expressa a variabilidade em termos **"relativos"**, comparando o desvio-padrão com a média:

$$CV = \frac{S}{\overline{X}} \times 100\% \qquad \text{OBS: } \overline{X} \neq 0 .$$

Observe também que o coeficiente de variação é adimensional e por este motivo permite a comparação das variabilidades de diferentes conjuntos de dados.

Exemplo 1.16: Utilizando os dados apresentados na Tabela 5, determine a variância, o desvio-padrão e o coeficiente de variação das alturas dos 46 estudantes de EV.

Altura (X_i)	Frequência Absoluta (F_i)	Ponto Médio (pm_i)	$(pm_i - \overline{X})^2 . F_i$
153,0 |----- 157,3	4	155,15	$(155,15 - 168,42)^2.4 = 704,3716$
157,3 |----- 161,6	8	159,45	$(159,45 - 168,42)^2.8 = 643,6872$
161,6 |----- 165,9	7	163,75	$(163,75 - 168,42)^2.7 = 152,6623$
165,9 |----- 170,2	10	168,05	$(168,05 - 168,42)^2.10 = 1,369$
170,2 |----- 174,5	3	172,35	$(172,35 - 168,42)^2.3 = 46,3347$
174,5 |----- 178,8	6	176,65	$(176,65 - 168,42)^2.6 = 406,3974$
178,8 |----- 183,1	8	180,95	$(180,95 - 168,42)^2.8 = 1256,0072$
Total ou $\sum$	46	---	3210,8294

☺ **Solução**:

* Variância: $S^2 = \frac{\sum (PM_i - \overline{X})^2 . f_i}{n-1} = \frac{3210,8294}{46-1} = 71,35\, cm^2$

* Desvio Padrão: $S = \sqrt{S^2} = \sqrt{71,35} = 8,45\, cm$

➘ Análise: Verifica-se uma variação de 8,45 cm nas alturas em torno da média observada.

* Coeficiente de Variação: $CV = \frac{S}{\overline{X}}.100 = \frac{8,45}{71,35}.100 = 5,02\%$

➘ Análise: Verifica-se uma variação relativa de 5,02% nas alturas em torno da média observada.

☺ É importante expressar a variabilidade em termos relativos porque, por exemplo, um desvio-padrão igual a 1 pode ser muito pequeno se a magnitude dos dados é da ordem de 1.000, mas pode ser considerado muito elevado se esta magnitude for da ordem de 10.

Exemplo 1.17: Uma fábrica classifica operários de acordo com os graus obtidos em testes de aptidão. Os dados são apresentados na distribuição de frequência abaixo:

Notas do Teste	F_i	Fac$_i$	pm_i	$pm_i F_i$	$(PM_i - \overline{X})$	$(PM_i - \overline{X})^2$	$(PM_i - \overline{X})^2 F_i$
0 |----- 2	6	6	1	6	-4,17	17,3889	104,3334
2 |----- 4	10	16	3	30	-2,17	4,7089	47,089
4 |----- 6	23	39	5	115	-0,17	0,0289	0,6647
6 |----- 8	11	50	7	77	1,83	3,3489	36,8379
8 |----| 10	8	58	9	72	3,83	14,6689	117,3512
Total ou $\sum$	58	-	-	300		40,149	306,2762

a) Calcule o grau médio obtido pelos operários;
b) O operário que tirar nota acima de $(\overline{X} + 2S)$ receberá um prêmio. Um operário para receber esta menção deverá ter tirado quanto?
c) Com base nos dados da tabela, a partir de que nota teremos 50% dos operários mais aptos?

☺ **Solução:**

a) O grau médio é dado por: $\overline{X} = \frac{\sum (PM.f)}{n} = \frac{300}{58} = 5,17$

b) A variância para os dados agrupados é dada pela fórmula:

$S^2 = \frac{\sum (PM_i - \overline{X})^2 . f_i}{n-1} = \frac{306,2762}{57} = 5,37$

➔ Desvio Padrão: $S = \sqrt{S^2} = \sqrt{5,37} = 2,32$

Análise: Verifica-se uma variação de 2,32 pontos nas notas em torno da média observada.

➔ Desta forma $(\overline{X} + 2S)$ = 9,81. Portanto, qualquer operário com nota maior que 9,81 receberá o prêmio.

c) A nota acima da qual estão 50% dos operários é chamada nota mediana, a qual é calculada para dados agrupados em intervalos por:

$$Med = LI_{Med} + \frac{\left(\frac{n}{2} - Fac_{ant}\right)}{F_{Med}}.h_{Med} = 4 + \frac{\left(\frac{58}{2} - 16\right)}{23}.2 = 4 + \left(\frac{26}{23}\right) = 4 + 1{,}13 = 5{,}13\,.$$

☑ **Observação:** A variância pode ser calculada também como,

$$S^2 = \frac{\sum_{i=1}^{k}\left(pm_i - \overline{X}\right)^2.F_i}{n-1} = \frac{\sum_{i=1}^{k} pm_i^2 F_i - \frac{\left(\sum_{i=1}^{k} pm_i F_i\right)^2}{n}}{n-1}\,.$$

☑ Comentários sobre as principais Medidas de Tendência Central e Dispersão:

1. O conjunto de todos os possíveis elementos de uma determinada pesquisa constitui uma **população estatística**. Sua média é a **média populacional**, usualmente representada pela letra grega μ. Na grande maioria das situações práticas, a média populacional é desconhecida e deve ser estimada a partir de dados amostrais. Se a amostra for extraída de forma adequada, a **média amostral** $\overline{X}$ é uma boa estimativa de μ.

2. Comparando a média e a mediana, temos que a mediana é pouco sensível à presença de valores muito altos ou muito baixos na amostra, enquanto a média já é muito sensível a esta situação. Para ilustrar o sentido desta afirmação, vamos considerar os dados abaixo:

   5 14 47 61 122 620

   A mediana deste conjunto de dados é: $Med = \frac{47+61}{2} = 54$

   A média é dada por: $\overline{X} = \frac{5+14+47+61+122+620}{6} = \frac{869}{6} = 144{,}8\,.$

   Observe que a maior observação (620) exerceu uma grande influência sobre a média (somente este dado é maior do que a média), o que então não sintetiza de forma adequada as informações contidas na massa de dados. Portanto, neste exemplo, a mediana parece ser a melhor medida para indicar a localização dos dados.

   De modo geral, quando o histograma construído para os dados da amostra é do tipo assimétrico, devemos preferir a mediana como medida de tendência central.

3. A amplitude, apesar de ser muito fácil de calcular, tem a desvantagem de levar em consideração apenas os dois valores extremos (máximo e mínimo) da massa de dados, desprezando os demais.

4. A **variância populacional** é representada por σ^2. Usualmente, a variância populacional é desconhecida e deve ser estimada a partir dos dados amostrais. Se a amostra foi extraída de forma adequada, a **variância amostral** S^2 é uma boa estimativa de σ^2.

5. As medidas $\overline{X}$, S^2 e S tomadas na amostra, denominadas **ESTATÍSTICAS**, são estimativas dos **PARÂMETROS POPULACIONAIS** μ, σ^2, σ (supostos desconhecidos).

4. Avaliando o que foi construído

Nesta unidade aprendemos a explorar dados estatísticos, onde estudamos desde a organização dos dados em tabelas e gráficos até o cálculo de medidas estatísticas importantes que serão utilizadas nas unidades subsequentes. Convidamos vocês a resolverem a lista de exercício anexa a este material, tentando descobrir no seu dia a dia a utilidade para o conteúdo aqui abordado. Este foi o início da convivência com a Estatística. Esperamos que tenha sido prazeroso. Procure seus tutores, use e abuse deste material.

UNIDADE 2
PROBABILIDADE

1. Situando a Temática

A teoria das probabilidades é o fundamento para a inferência estatística. O objetivo desta parte é que o aluno compreenda os conceitos mais importantes da probabilidade. O conceito de probabilidade faz parte do dia-a-dia dos trabalhadores das áreas das ciências exatas, ciências da saúde, ciências biológicas, ecologia, engenharia, etc., uma vez que seu conceito é frequentemente usado na comunicação diária. Por exemplo, podemos dizer que uma espécie tem 30% de chance de ser extinta. Um laboratório está 90% seguro de que um medicamento proporcione a cura de uma doença. Um biólogo afirma que nas marés altas o peso seco das ostras capturadas é 20% maior que as capturadas em marés baixas. Um aluno tem chance de 70% de ser aprovado em uma determinada disciplina. Um professor está 95% seguro de que um novo método de ensino proporcione uma melhor compreensão pelos alunos. Um engenheiro de produção afirma que uma nova máquina reduz em 20% o tempo de produção de um bem. Tal como mostram os exemplos, as pessoas expressam a probabilidade em porcentagem. Trabalhando com a probabilidade matemática é mais conveniente expressá-la como fração (as porcentagens resultam da multiplicação das frações por 100).

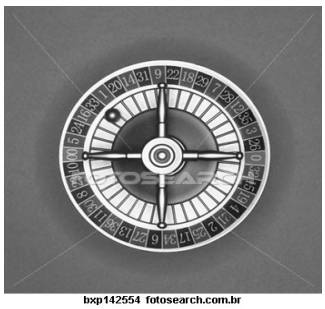
2. Problematizando a Temática

O conceito de probabilidade é fundamental para o estudo de situações onde os resultados são variáveis, mesmo quando mantidas inalteradas as condições de sua realização. Por exemplo, jogando-se um dado, temos seis resultados possíveis de cada vez; a observação do sexo dos candidatos inscritos num concurso público conduz a dois resultados possíveis – masculino ou feminino. Em ambos os casos, embora não sejamos capazes de afirmar de antemão que resultado particular ocorrerá, temos condições de descrever o conjunto de todos os resultados possíveis do experimento. A sua repetição continuada mostra certa regularidade nos resultados, o que nos permite estudar o experimento, apesar da incerteza nele presente.

A VIDA É CHEIA DE INCERTEZAS!

Às vezes, pode ser impossível dizer o que vai acontecer de um minuto para o outro. Mas certos eventos são mais prováveis de ocorrer do que outros, e é aí que entra em cena a teoria da probabilidade.

A probabilidade permite prever o futuro avaliando a probabilidade de os resultados ocorrerem. Saber o que pode acontecer ajuda você a tomar decisões mais embasadas.



PREPARE-SE PARA A ROLETA!

Provavelmente, você já viu pessoas jogando roleta em filmes. A crupiê gira a roleta e, em seguida, joga uma bolinha no sentido oposto. As apostas são feitas com base onde a pessoa acha que a bolinha vai parar. A roleta tem 38 casas em que a bolinha pode cair. As principais são numeradas de 1 a 36, e cada casa é colorida de vermelho ou de preto. Existem duas casas extras, de cor verde, com números 0 e 00.

Você pode fazer vários tipos de apostas. Por exemplo, pode apostar em um determinado número, se aquele número é par ou ímpar, ou na cor da casa. Lembrando que se a bolinha cair em uma casa verde, você perde! A roleta vai nos ajudar a navegar pelo mundo das probabilidades.

DETERMINANDO PROBABILIDADES NO JOGO DA ROLETA

Vamos tentar calcular a probabilidade de a bolinha parar no número 7. Veja como chegar à solução, passo a passo.

Perguntas	Respostas	Conclusão
1. Observe o tabuleiro da roleta. Em quantas casas a bolinha pode cair?	Há 38 casas	A probabilidade de a bolinha parar no número 7 é de 0,026. Não é impossível de acontecer, mas não é muito provável!
2. Quantas casas existem para o número 7?	Apenas 1	
3. Para calcular a probabilidade de obter um 7, pegue a resposta que você deu à questão 2 e divida-a pela sua resposta à questão 1.	1/38 = 0,026	

Para achar a probabilidade de ganhar, pegamos o número de formas de ganhar a aposta e a dividimos pelo número de resultados possíveis, da seguinte forma:

Probabilidade = (número de formas de ganhar)/(número de resultados possíveis)

Perguntas	Respostas
P(9) = ? Há 1 casa para o número 9 e 38 casas no total (Idem número 7).	1/38 = 0,026
P(Preto) = ? Há 18 casas pretas e 38 casas no total.	18/38 = 0,474
P(Verde) = ? Há 2 casas verdes e 38 casas no total.	2/38 = 0,053
P(38) = ? Esse evento é impossível, não há nenhuma casa com número 38.	0

Que tal agora conhecer os conceitos básicos que cercam a teoria das probabilidades mais formalmente?

3. Conhecendo a Temática

3.1. Espaços Amostrais e Eventos

Antes de passarmos à definição de probabilidade é necessário fixarmos os conceitos de *experimento aleatório*, *espaço amostral* e *evento*.

☑ **Experimento Aleatório** (Notação: E)
É o processo da coleta dos dados relativo a um fenômeno que acusa variabilidade em seus resultados.

☺ Características de um Experimento Aleatório:

1) Todo experimento aleatório poderá ser repetido indefinidamente sob condições, essencialmente inalteradas.
2) Embora não sejamos capazes de afirmar que resultado ″particular″ ocorrerá, seremos sempre capazes de descrever o conjunto de todos os possíveis resultados do experimento.

☑ **Espaço Amostral** (Notação: S ou Ω (ômega))
É o conjunto formado por todos os possíveis resultados de um experimento aleatório.

☑ **Eventos** (Notação: *A, B, C*, ...)
É qualquer subconjunto do espaço amostral.

3.1.1. Operações entre Eventos

Combinações de Eventos

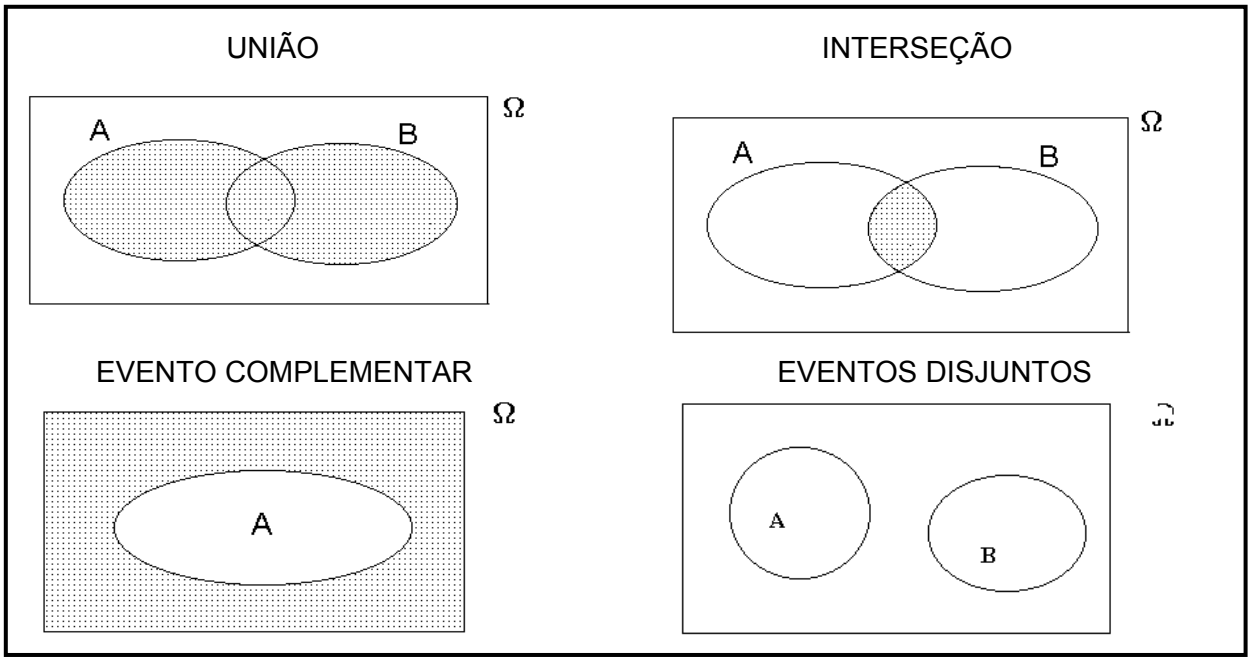
Sejam A e B eventos em um mesmo espaço amostral. Temos as definidas as seguintes operações entre conjuntos:

- **Evento União** $(A \cup B)$ (lê-se: *A* união *B*)**:**
  O evento união de *A* e *B* equivale à ocorrência de *A* **ou** de *B* **ou** de ambos. Contém os elementos do espaço amostral que estão em *A* **ou** em *B* **ou** em ambos.

- **Evento Interseção** $(A \cap B)$ (lê-se: *A* interseção *B*):
  O evento interseção de *A* e *B* equivale à ocorrência de *A* **e** de *B*, simultaneamente. Contém os elementos do espaço amostral que estão em *A* **e** em *B*.

- **Evento Complementar** $\bar{A}$ ou A^C (lê-se: $\bar{A}$ evento complementar de *A*):
  O evento complementar de *A* equivale **à não** ocorrência do evento *A*. Contém os elementos do espaço amostral que **não** estão em *A*.

- **Eventos Disjuntos ou Mutuamente Exclusivos:**
  Dois eventos *A* e *B* dizem-se mutuamente exclusivos ou mutuamente excludentes quando a ocorrência de um deles impossibilita a ocorrência do outro. Os dois eventos não têm nenhum elemento em comum. Exprime-se isto escrevendo: $(A \cap B) = \varnothing$.



3.2. O Conceito de Probabilidade

☑ **Definição 2.1:** Uma função *P*: Ω → *R* é dita uma "probabilidade" se satisfaz os seguintes axiomas:

(i) $P(\Omega) = 1$;

(ii) $0 \leq P(A) \leq 1$;

(iii) Sejam *A* e *B* eventos em um mesmo espaço amostral. Se *A* e *B* forem mutuamente exclusivos, então $P(A \cup B) = P(A) + P(B)$.

Vamos enunciar algumas propriedades relacionadas a P(*A*) que decorrem das condições acima e que não dependem da maneira pela qual calculamos P(*A*).

3.2.1. Propriedade de Probabilidade

Sejam *A* e *B* eventos em um mesmo espaço amostral:

1. Se ∅ é o evento impossível, então $P(\varnothing) = 0$;
2. Se A^C é o evento complementar de *A,* então $P(A^C) = 1 - P(A)$.
3. Se *A* e *B* são dois eventos quaisquer, então $P(A \cup B) = P(A) + P(B) - P(A \cap B)$;
4. Se o evento $A \subseteq B$, então $P(A) \leq P(B)$.

3.2.2. Probabilidade em Espaços Amostrais Finitos

Seja Ω um espaço amostral associado a um experimento aleatório constituído de *N* resultados igualmente prováveis (equiprováveis). Seja *A* um evento qualquer constituído de *r* resultados possíveis $(0 \leq r \leq N)$.

A probabilidade de ocorrência do evento *A,* denotada P(*A*), é dada por:

$$P(A) = \frac{\text{número de casos favoráveis a } A}{\text{número de casos possíveis}} = \frac{n(A)}{n(S)} = \frac{\#(A)}{\#(S)}$$

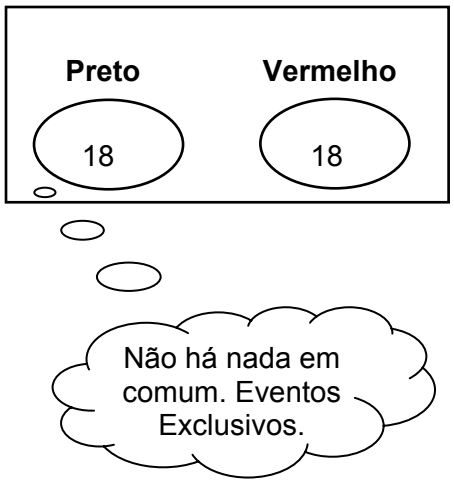
:: SAIBA MAIS... ::

Probabilidades apenas indicam a chance de algo ocorrer, não são garantias.

É importante lembrar que uma probabilidade indica apenas uma tendência a longo prazo. Ao jogar a roleta milhares de vezes você esperaria que a bolinha parasse em uma casa preta em aproximadamente 47% das vezes, e em uma casa verde em 5% das vezes. Embora você espere que a bolinha pare em uma casa verde com relativa pouca freqüência, isso não significa que este evento não possa acontecer.

VAMOS APOSTAR EM UM EVENTO AINDA MAIS PROVÁVEL!

Em vez de apostar que a bolinha vai parar em uma casa preta, vamos apostar que ela vai parar em uma casa preta ou vermelha.

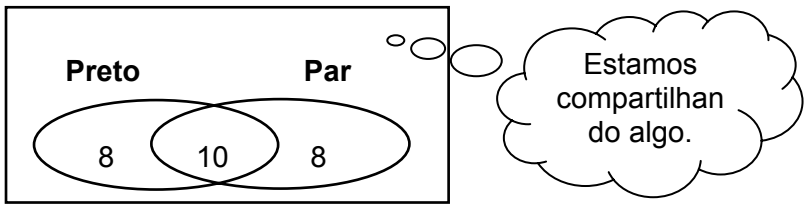


P(Preto ou Vermelho) = P(Preto) + P(Vermelho)
P(Preto ou Vermelho) = (18/38) + (18/38) = 0,947
Ou então, pelo complementar:
P(Preto ou Vermelho) = P(Verde^C)
P(Preto ou Vermelho) = 1 – P(Verde)
P(Preto ou Vermelho) = 1 – 0,053 = 0,947

Ao calcular a probabilidade de a bolinha parar em uma casa preta ou vermelha, estávamos trabalhando com dois eventos separados. Esses dois eventos são ***mutuamente exclusivos***, pois é impossível que a bolinha pare em uma casa que seja preta e vermelha.

E os eventos pretos e pares? Desta vez não são mutuamente exclusivos. É possível que a bolinha possa parar em uma casa que seja preta e par. Os dois eventos se interceptam, ou seja, não são exclusivos.

Se dois eventos são mutuamente exclusivos, apenas um dos dois pode ocorrer	Se dois eventos se interceptam, é possível que eles ocorram simultaneamente.



Para chegar a resposta correta precisamos subtrair a probabilidade de se obter preta e par (ou seja, da interseção), para que não seja contada duas vezes. Assim,

P(Preto ou Par) = P(Preto) + P(Par) – P(Preto e Par)
P(Preto ∪Par) = P(Preto) + P(Par) – P(Preto ∩Par)

☺ P(Preto ou Par) = (18/38) + (18/38) – (10/38) = 26/38 = 0,684

Exemplo 2.1: Em uma seleção para uma vaga de biólogo de uma grande empresa verificou-se que dos 100 candidatos 40 tinham experiência anterior e 30 possuíam curso de especialização. Vinte dos candidatos possuíam tanto experiência profissional como também algum curso de especialização. Escolhendo um candidato ao acaso, qual a probabilidade de que:

a) Ele tenha experiência ou algum curso de especialização?
b) Ele não tenha experiência anterior nem curso de especialização?

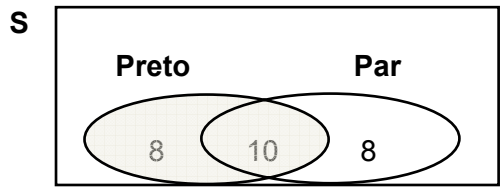
☺ **Solução:**
Vamos definir os seguintes eventos:
A = {O candidato possui experiência anterior}.
B = {O candidato possui especialização}.
Dados: P(A) = 40/100 = 0,40; P(B) = 30/100 = 0,30; P(A ∩ B) = 20/100 = 0,20.

(a) Ele tenha experiência **ou** algum curso de especialização?
$P(A \cup B) = P(A) + P(B) – P(A \cap B) = 0,40 + 0,30 – 0,20 = 0,50$

(b) Ele não tenha experiência anterior nem curso de especialização?
$P(A^C \cap B^C) = P((A \cup B)^C) = 1 – P(A \cup B) = 1 – [P(A) + P(B) – P(A \cap B)]$
$P(A^C \cap B^C) = 1 – [0,40 + 0,30 – 0,20] = 1 – 0,50 = 0,50.$

AS CONDIÇÕES SE APLICAM

A crupiê diz que a bolinha parou em uma casa preta. Qual é a probabilidade de que seja a casa também seja par? Este problema é um pouquinho diferente. Não queremos achar a probabilidade de se obter uma casa que seja preta e par, entre todas as casas possíveis. Queremos achar a probabilidade de que a casa seja par, uma vez que já sabíamos que ela é preta.



Em outras palavras, queremos descobrir quantas casas são pares, entre todas as casas pretas. Das 18 casas pretas, 10 delas são pares. Portanto,

P(Par sabendo que é preta) = 10/18 = 0,556

Perceba que nessa situação a probabilidade é calculada a partir de dois eventos condicionados. Expressa a probabilidade de um evento acontecer, uma vez ocorrido outro evento. Que tal agora conhecer os conceitos básicos que cercam a teoria da probabilidade condicional mais formalmente?

3.2.3. Probabilidade Condicional e Independência de Eventos

Dados dois eventos *A* e *B* contidos num espaço amostral *S* (ou Ω), muitas das vezes, estamos interessados na ocorrência de *A* dado que o evento *B* tenha ocorrido.

Para dar consistência à idéia de uma probabilidade condicional, suponhamos que uma organização de pesquisa junto a consumidores tenha estudado os serviços prestados dentro da garantia por 200 comerciantes de pneus em uma grande cidade, obtendo os resultados resumidos na tabela seguinte:

Vendedores de Pneus	Dentro da Garantia		Total
	Bom Serviço	Serviço Deficiente	
Com marca	64	16	80
Sem marca	42	78	120
Total	106	94	200

Selecionado aleatoriamente um desses vendedores de pneus, constatamos que as probabilidades de se escolher um vendedor de determinada marca (*M*), um vendedor que presta bons serviços dentro da garantia (*BS*), ou um vendedor de marca determinada e que presta bons serviços dentro da garantia ($M \cap BS$) são:

$$P(M) = 80/200 = 0{,}40 \quad , \quad P(BS) = 106/200 = 0{,}53 \quad \text{e} \quad P(M \cap BS) = 64/200 = 0{,}32 .$$

Todas essas probabilidades foram calculadas por meio da definição clássica de probabilidade. Como a segunda dessas probabilidades P(*BS*) é próxima a 0,50 (50%), vejamos o que acontece se limitamos a escolha a vendedores de uma marca determinada.

Isto reduz o espaço amostral as 80 escolhas, correspondentes à 1ª linha da tabela (*com marca*). Temos então, que a probabilidade de se escolher um vendedor que presta bons serviços (*BS*), sabendo (ou dado) que a marca de pneu vendido pelo mesmo é determinada será de $P(BS \mid M) = 64/80 = 0{,}80$, tendo-se uma melhora em relação a P(*BS*) = 0,53. Note que a probabilidade condicional que obtivemos aqui, $P(BS \mid M) = 0{,}80$ pode escrever-se como:

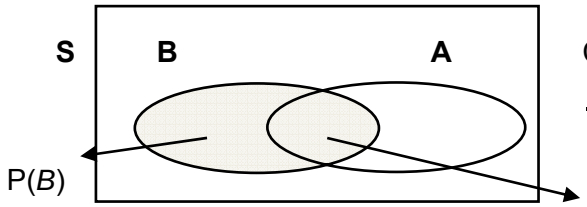
$$P(BS \mid M) = \frac{64/200}{80/200} = \frac{P(M \cap BS)}{P(M)} \quad .$$

Generalizando, formulamos a seguinte definição de probabilidade condicional, que se aplica a dois eventos quaisquer *A* e *B* pertencentes a um dado espaço amostral *S*:

Probabilidade Condicional

Se P(*B*) é diferente de zero, então a probabilidade condicional de *A* relativa a *B*, isto é, a probabilidade de *A* dado que *B* ocorreu é denotada por:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}, \quad \text{desde que} \quad P(B) > 0 \cdot$$



Como estamos tentando achar a probabilidade de *A* **uma vez ocorrido** *B*, estamos interessados apenas no conjunto de eventos onde *B* ocorre.

☑ Teorema da Multiplicação

O resultado a seguir, obtido a partir da definição de probabilidade condicional, fornece a probabilidade da ocorrência conjunta de dois eventos *A* e *B*, isto é, a probabilidade *P(A ∩ B)*:

$$P(A \cap B) = P(A) \cdot P(B \mid A) \quad \text{ou} \quad P(A \cap B) = P(B) \cdot P(A \mid B)$$

Dependendo da ordem de ocorrência dos eventos.

É HORA DE FAZER OUTRA APOSTA!

Antes de abandonar o jogo, a crupiê lhe faz uma oferta fantástica pela sua aposta final, o triplo ou nada. Apostando que a bolinha vai parar em uma casa preta duas vezes seguidas, você poderá reaver todas as suas fichas.

SE OS EVENTOS SE AFETAM, ELES SÃO DEPENDENTES

A probabilidade de se obter preto seguido de preto é um problema um pouquinho diferente da probabilidade de se obter uma casa par, uma vez que sabíamos que ela é preta. Observe:

A probabilidade de se obter uma casa par é afetada pelo evento de se obter uma casa preta. Já sabemos que a bolinha parou em uma casa preta.	P(Par | Preto) = 10/18 = 0,556
Se não soubéssemos que a bolinha havia parado em uma casa preta, a probabilidade seria diferente. Para calcular P(Par), examinamos quantas casas são pares.	P(Par) = 18/38 = 0,474

P(Par | Preto) dá um resultado diferente de P(Par). O fato de sabermos que a casa é preta muda a probabilidade.
Esses dois eventos são chamados de **dependentes**!

SE OS EVENTOS NÃO SE AFETAM, ELES SÃO INDEPENDENTES

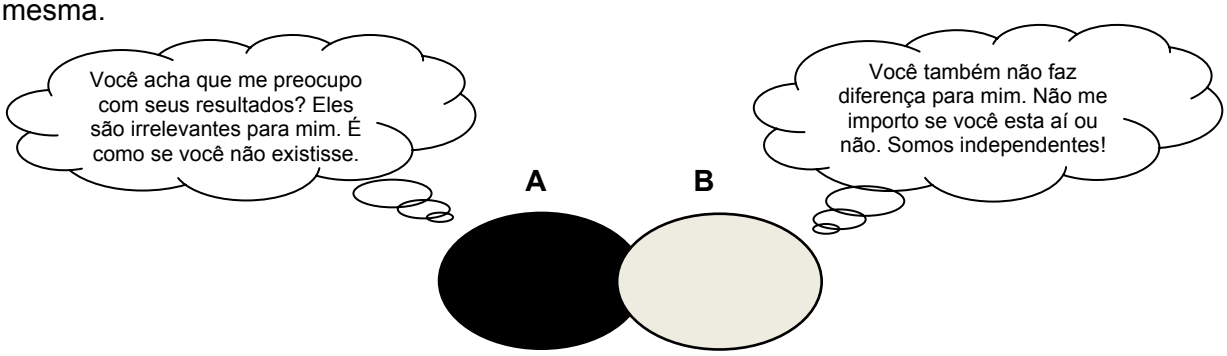
Nem todos os eventos são dependentes. Às vezes, os eventos não sofrem nenhuma influência um do outro, e a probabilidade de um evento permanece a mesma independentemente de outro evento acontecer ou não. Observe,

P(Preto) = 18/38 = 0,474
P(Preto | Preto) = 18/38 = 0,474

Essas duas probabilidades têm o mesmo valor! O evento de se obter uma casa preta nesse jogo não tem nenhuma influência sobre a probabilidade de se obter uma casa preta no próximo jogo. Esses eventos são chamados de **independentes**.

⬇

Eventos independentes não são afetados uns pelos outros. Eles de maneira alguma, influenciam as probabilidades do outro. Se um evento ocorre, a probabilidade do outro permanece exatamente a mesma.



Independência de Eventos

Dizemos que dois eventos *A* e *B* são ***independentes***, se as probabilidades condicionais P(*A* | *B*) = P(*A*) e P(*B* | *A*) = P(*B*). Isto equivale, a partir da regra da multiplicação, escrever a ocorrência simultânea de *A* e *B* como sendo:

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


É hora de calcular outra probabilidade! Qual é a probabilidade de a bolinha parar em uma casa preta duas vezes seguida? Precisamos achar:

P(Preto no jogo 1 ∩ Preto no jogo 2) = ?

Como os eventos são independentes o resultado é: (18/38).(18/38) = 324/1444 = 0,224

Exemplo 2.2: Uma caixa contém 4 lâmpadas boas e 2 queimadas. Retiram-se, ao acaso, 3 lâmpadas sem reposição. Calcule a probabilidade dessas 3 lâmpadas serem boas.

☺ **Solução:**
Seja o evento A_i = A i-ésima lâmpada é boa, então: A_1 = A primeira lâmpada é boa; A_2 = A segunda lâmpada é boa e A_3 = A terceira lâmpada é boa. A probabilidade dessas 3 lâmpadas serem boas é dada por:

$$P(A_1 \cap A_2 \cap A_3) = P(A_1) \times P(A_2 \mid A_1) \times P(A_3 \mid A_1 \cap A_2) = \frac{4}{6} \times \frac{3}{5} \times \frac{2}{4} = \frac{1}{5}$$

Exemplo 2.3: Sejam *A* e *B* dois eventos tais que P(*A*) = 0,4 e P($A \cup B$) = 0,7. Seja P(*B*) = *p*. Para que valor de *p*, *A* e *B* serão mutuamente exclusivos? Para que valor de *p* *A* e *B* serão independentes?

> ☺ **Solução:**
> Sabe-se de definições passadas que, *A* e *B* são **mutuamente exclusivos** se $(A \cap B) = \varnothing$. Logo, pelo teorema de probabilidade: $P(A \cap B) = 0$. Assim, a partir do teorema da soma tem-se que:
> $P(A \cup B) = P(A) + P(B) - P(A \cap B) \Rightarrow 0{,}7 = 0{,}4 + p + 0 \Rightarrow p = 0{,}7 - 0{,}4 \Rightarrow p = 0{,}3$.
>
> No caso de **independência** sabe-se que: Se *A* e *B* são independentes $P(A \cap B) = P(A) \times P(B) = 0{,}4\,p$. Como $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ temos que:
> $P(A \cup B) = P(A) + P(B) - P(A \cap B) \Rightarrow 0{,}7 = 0{,}4 + p - 0{,}4p \Rightarrow 0{,}7 - 0{,}4 = 0{,}6p \Rightarrow p = 0{,}5$.

4. Avaliando o que foi construído

Nesta unidade aprendemos a lidar com um conceito muito importante da estatística e que está presente quase diariamente nas nossas vidas, a probabilidade. Aprendemos nesta unidade que uma maneira de responder a pergunta “qual a probabilidade de chover hoje” seria observar, em um passado recente de dias, o número de dias que choveu e dividi-lo pelo total de dias. Aprendemos também os conceitos de probabilidade condicional e independência de eventos. Com isso, estamos nos preparando cada vez mais para as etapas futuras. Para você que está conosco, parabéns!

UNIDADE 3
VARIÁVEIS ALEATÓRIAS E DISTRIBUIÇÕES DE PROBABILIDADE

1. Situando a Temática

Na unidade anterior estudamos alguns fenômenos probabilísticos por meio de espaços amostrais mais simples. No entanto, em situações práticas mais gerais, é necessário ampliar esses conceitos para que tenhamos modelos probabilísticos que atendam as necessidades do problema. A definição do conceito de variável aleatória possibilitará uma maior flexibilidade e aplicabilidade dos conceitos de probabilidade em problemas diversos.

2. Problematizando a Temática

Ao estudarmos fenômenos aleatórios tais como, a renda de uma população, o desempenho escolar de um grupo de alunos, o impacto de uma dieta no peso de animais, etc., desejamos saber como controlar esses experimentos e tentar extrair conclusões sobre as respostas obtidas. Neste caso, usaremos uma ferramenta valiosa que são as variáveis aleatórias.

:: SAIBA MAIS... ::

Eventos improváveis acontecem, mas quais são as consequências?

Até então, examinamos como saber a probabilidade de certos eventos. O que a probabilidade não lhe diz é o impacto geral desses eventos, e o que isso significa para você. Sim, com certeza, você vai às vezes tirar a sorte grande na roleta, mas será que realmente vale a pena o risco com todo o dinheiro que você perde nesse meio tempo? Neste capítulo, vamos mostrar-lhe como usar a probabilidade para prever resultados a longo prazo e também medir a certeza dessas previsões.

3. Conhecendo a Temática

Quando na prática desejamos investigar algum fenômeno, estamos na realidade interessados em estudar a distribuição de uma ou mais variáveis relacionadas a este. Assim, por exemplo, podemos estar interessados em estudar a distribuição das notas de estudantes em uma determinada disciplina, do grau de instrução, da altura, etc.

O que pretendemos, nesta unidade, é apresentar alguns modelos teóricos de distribuição de probabilidade, aos quais um experimento aleatório estudado possa ser adaptado, o que permitirá a solução de um grande número de problemas práticos.

Imagine a seguinte situação:

Um jogo de frutas em um caça-níquel tem três visores. Se todos os visores ficarem alinhados de forma correta prepare-se para ganhar uma cascata de dinheiro! O valor é R$1,00 por jogo.

A quantia de dinheiro que você pode ganhar parece tentadora, mas gostaria de saber a probabilidade de obter qualquer uma dessas combinações antes de jogar.

Isso parece algo que podemos calcular. Veja as probabilidades de uma determinada figura aparecer em um visor:

R$	Cereja	Limão	Outros
1,00	0,2	0,2	0,5

A probabilidade de uma cereja aparecer é 0,2.

Os três visores são independentes um do outro, o que significa que a figura que aparece em um dos visores não tem efeito sobre as figuras que aparecerem em qualquer um dos outros.

PODEMOS CRIAR UMA DISTRIBUIÇÃO DE PROBABILIDADE EM DECORRÊNCIA DE SUA COMBINAÇÃO PARA O CAÇA-NÍQUEIS.

Veja as probabilidades das diferentes combinações de ganhar alguma coisa no caça-níqueis.

Combinação	Nenhuma	Cerejas	Limões	Reais/Cerejas	Reais
Probabilidade	0,977	0,008	0,008	0,006	0,001

Não queremos saber só a análise combinatória de ganhar, queremos saber quanto temos chance de ganhar.

As probabilidades estão atualmente escritas em termos de combinações de símbolos, o que dificulta ver rapidamente qual será nosso ganho.

Não precisamos escrevê-las assim. Em vez de escrever análise combinatória em termos de figuras, podemos escrevê-las em termos de quanto ganhamos ou perdemos em cada jogo. Basta pegar a quantia que vamos ganhar para cada combinação e subtrair a quantia que pagamos pelo jogo.

Combinação	Nenhuma	Cerejas	Limões	Reais/Cerejas	Reais
Ganho	– R$1,00	R$9,00	R$4,00	R$14,00	R$19,00
Probabilidade	0,977	0,008	0,008	0,006	0,001

Essa é a ***análise combinatória*** de ganhar, um conjunto das probabilidades para cada possível ganho ou perda.

Ao deduzir as combinações do caça-níqueis, você calculou as combinações de obter cada ganho ou perda. Isto é, calculou a distribuição de combinações de uma **variável aleatória**, que é uma variável que pode assumir um conjunto de valores onde cada valor está associado a uma combinação específica. Nesse caso, a variável aleatória representa a quantia que vamos ganhar em cada jogo.

☺ Quando queremos nos referir a uma incógnita aleatória, geralmente a representamos por uma letra maiúscula, como *X* ou *Y*.
☺ Os valores específicos que a incógnita pode assumir são representados por letras minúsculas, como *x* ou *y*.
☺ P(X = x) é uma forma de dizer "a probabilidade de que a variável X assuma um valor específico x".

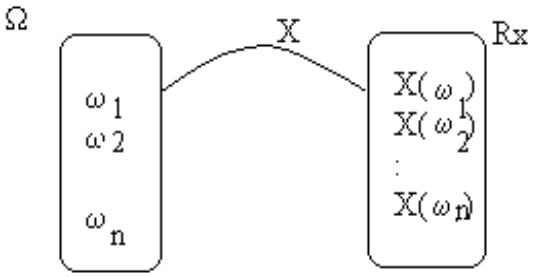
Combinação	Nenhuma	Cerejas	Limões	Reais/Cerejas	Reais
X	– 1	9	4	14	19
P(X = x)	0,977	0,008	0,008	0,006	0,001

A incógnita é **discreta**. Isso significa que ela só pode assumir valores exatos!

Que tal agora conhecer os conceitos básicos que cercam a teoria de variáveis aleatórias mais formalmente?

3.1. O Conceito de Variável Aleatória e Variáveis Aleatórias Discretas

☑ **Definição 3.1*:*** Seja E um experimento e *S* (ou Ω) um espaço amostral associado a E. Uma função X, que associe a cada elemento $\omega \in \Omega$ um número real, X(ω), é denominada ***variável aleatória***.



☑ **Observação:**
1. Cada elemento ω de Ω corresponderá a exatamente um valor;
2. Diferentes valores $\omega \in \Omega$ podem levar a um mesmo valor de X;
3. Nenhum elemento $\omega \in \Omega$ poderá ficar sem valor de X.

☑ **Definição 3.2*:*** Seja E um experimento e Ω seu espaço amostral. Seja X uma variável aleatória definida em Ω e seja R_x seu contradomínio. Seja *B* um evento definido em relação a R_x, isto é, $B \subset R_x$. Então, define-se o evento *A* como:

$$A = \{\omega \in \Omega \mid X(\omega) \in B\} = X^{-1}(B).$$

Assim, o evento *A* será constituído por todos os resultados em Ω para os quais $X(\omega) \in B$.

Exemplo 3.1: Suponha 2 moedas lançadas e observada a sequência de caras e coroas obtidas. Considere o espaço amostral associado a este experimento:

☺ **Solução:**

Espaço Amostral: Ω = {(Ca,Co), (Ca,Ca), (Co,Ca), (Co,Co)}

Agora, defina uma variável aleatória X = número de caras obtidas no lançamento de 2 moedas. Assim, temos que X = {0, 1, 2}, visto que:

X(Co, Co) = 0

X(Ca, Co) = X(Co, Ca) = 1

X(Ca, Ca) = 2.

Variáveis Aleatórias Discretas 

Denomina-se X uma variável aleatória discreta se o número de valores possíveis de X for um conjunto de pontos finito ou infinito enumerável. Digamos $R_X = \{x_1, x_2, \ldots, x_n, \ldots\}$.

☑ **Definição 3.3** (**Função de Probabilidade**): Seja X uma variável aleatória discreta. A cada possível resultado x_i de X está associado um número $p_i = P(X = x_i)$, denominado probabilidade da variável aleatória X assumir o valor x_i, satisfazendo as seguintes condições:

a) $p_i \geq 0$ para todo $x_i \in R_X$.

b) $\sum p_i = p_1 + p_2 + \ldots + p_n + \ldots = 1$ (a soma das probabilidades é igual a 1).

☑ **Definição 3.4** (**Função de Distribuição de Probabilidade**): Dada uma variável aleatória discreta X, definimos F(x) a *função de distribuição acumulada* ou, simplesmente, *função de distribuição* (f.d) de X, dada por:

$$F(x_i) = P(X \leq x_i) \Rightarrow F(x_i) = \sum_{i=1}^{n} P(X = x_i)$$

Exemplo 3.2: Considerando o exemplo 3.1, denote a função de probabilidade e a função de distribuição da variável aleatória X.

☺ **Solução:**

Seja X = número de caras obtidas no lançamento de 2 moedas, temos que a variável aleatória X assume os seguintes valores, X = {0, 1, 2}.

Temos que,

P(Co, Co) = P(X = 0) = 1/4

P(Ca, Co) = P(Co, Ca) = P(X = 1) =2/4 ou 1/2

P(Ca, Ca) = P(X = 2) =1/4

Denotamos a função de probabilidade de X por

x_i	0	1	2
$P(X = x_i)$	1/4	½	1/4

Por conseguinte, a função de distribuição acumulada de X é dada por

x_i	0	1	2
$F(x_i) = P(X \leq x_i)$	1/4	3/4	1

Exemplo 3.3: Um par de dados é lançado. Seja X a variável aleatória que associa a cada ponto (d_1, d_2) de Ω a soma desses números, isto é, $X(d_1, d_2) = d_1 + d_2$. Determine a função de probabilidade de X.

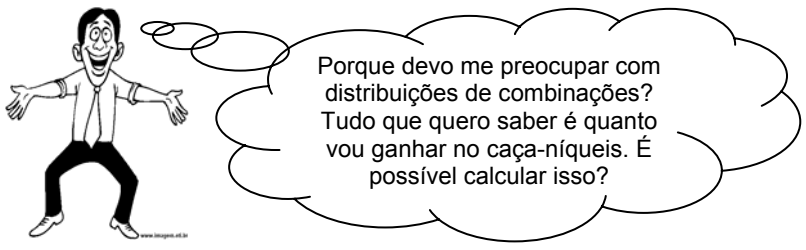
☺ **Solução:**
O espaço amostral Ω é formado de 36 pares ordenados, representando as possibilidades nos lançamentos de dois dados Ω = {(1, 1), (1, 2), (1, 3),..., (5, 6), ..., (6, 6)}.

Então, a variável aleatória $X = d_1 + d_2$ assume os seguintes valores X = {2, 3, 4, ..., 12}. Por conseguinte, a função de probabilidade de X obtida, calculando-se:

$P(X = 2) = P(d_1 = 1, d_2 = 1) = (1/6) \times (1/6) = 1/36$
$P(X = 3) = P(d_1 = 1, d_2 = 2) + P(d_1 = 2, d_2 = 1) = (1/36) + (1/36) = 2/36$
…..
$P(X = 12) = P(d_1 = 6, d_2 = 6) = 1/36$

Logo, a função de probabilidade de X será representada por

x_i	2	3	4	5	6	7	8	9	10	11	12
$P(X = x_i)$	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36



UMA VEZ CALCULADA A DISTRIBUIÇÃO DE PROBABILIDADES, VOCÊ PODE USAR ESSA INFORMAÇÃO PARA DETERMINAR O RESULTADO ESPERADO

OS VALORES ESPERADOS LHE DÃO UMA PREVISÃO DOS RESULTADOS...

Você tem uma distribuição de probabilidade referente a quantia que poderia ganhar no caça-níqueis, mas agora precisa saber quanto pode esperar ganhar ou perder a longo prazo. Isso pode ser feito calculando quanto você normalmente espera ganhar ou perder em cada jogo. Isto é, você pode calcular o **valor esperado**.

O valor esperado de uma variável aleatória X é um pouco semelhante à média aritmética, mas para distribuições de probabilidades. É representado por E(X) ou μ. Seu cálculo também é semelhante.

Para achar o valor esperado, multiplique cada valor de x pela sua respectiva probabilidade de ocorrência e, em seguida some os resultados. De forma bem simples:

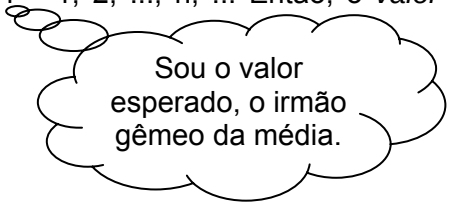
$$\mathbf{E(X) = \Sigma[x.P(X = x)]}$$

3.2. Valor Esperado e Variância de uma Variável Aleatória

Nos modelos probabilísticos que temos considerado, parâmetros podem ser empregados para caracterizar sua distribuição de probabilidade. Dada uma distribuição de probabilidade, é possível associar certos parâmetros, os quais fornecem informações valiosas sobre tal distribuição.

Um dos parâmetros mais importantes é o *valor esperado* (*esperança* ou *média*) de uma variável aleatória X, denotado por E(X) ou μ.

☑ **Definição 3.5** (**Valor Esperado ou Média**): Seja X uma variável aleatória **discreta** com possíveis valores $x_1, x_2, \ldots, x_n, \ldots$. Seja $p(x_i) = P(X = x_i)$, i = 1, 2, ..., n, ... Então, o *valor esperado* ou *média* da variável aleatória X é definido por:



$$\mu = E(X) = \sum_{i=1}^{\infty} x_i \cdot p(x_i),$$

se a série $\sum_{i=1}^{\infty} x_i \cdot p(x_i)$ convergir, ou seja, $\sum_{i=1}^{\infty} |x_i| \cdot p(x_i) < \infty$.

☑ **Observação**: E(X) mede o valor médio de X, sendo expressa na mesma unidade de X.

Vamos calcular o valor esperado dos ganhos no caça-níqueis. Recorda da nossa distribuição de probabilidade?

x	– 1	9	4	14	19
P(X = x)	0,977	0,008	0,008	0,006	0,001

E(X) = (– 1 x 0,977) + (9 x 0,008) + (4 x 0,008) + (14 x 0,006) + (19 x 0,001)
E(X) = – 0,77

Em outras palavras, em várias jogadas, você pode esperar perder R$0,77 para cada jogo. Isto significa que, se você jogasse 100 vezes no caça-níqueis poderia esperar perder R$77,00

Exemplo 3.4: Considere a variável aleatória definida no exemplo 3.2. Obtemos a E(X) por:

☺ **Solução:**

$$E(X)=\sum_{i=1}^{3} x_i\, p(x_i)=\left(0\times\frac{1}{4}\right)+\left(1\times\frac{1}{2}\right)+\left(2\times\frac{1}{4}\right)=1$$

Isto representa que, ao lançarmos a moeda 2 vezes esperamos que, em média, em um dos lançamentos apareça "Cara".

... E A VARIÂNCIA LHE DIZ SOBRE A DISPERSÃO DOS RESULTADOS

O valor esperado lhe diz quanto, em média, você pode esperar ganhar ou perder a cada jogo. Se você perdesse essa quantia todas as vezes, onde estaria a graça, e quem jogaria?

Só porque você pode esperar perder a cada vez que jogar não significa que não haja uma pequena chance de ganhar bastante dinheiro. Assim como a média, o valor esperado não conta a história por inteiro, pois a quantia que você tem chance de ganhar em cada jogo pode variar muito. Em sua opinião, como poderíamos medir isso?

AS DISTRIBUIÇÕES DE PROBABILIDADES TÊM VARIÂNCIA

O valor esperado fornece o valor típico, ou médio, de uma variável, mas não lhe diz nada sobre como os valores estão dispersos. Outro parâmetro importante que caracteriza uma variável aleatória é a *variância*, denotada V(X) ou σ^2. A variância de uma variável aleatória é uma medida que dá a idéia de dispersão dos valores da variável, em relação ao seu valor esperado (média).

☑ **Definição 3.6** (**Variância**): Seja uma variável aleatória X (discreta ou contínua) sua variância, denotada V(X) ou σ^2, é definida por:

$$\sigma^2 = V(X) = E\left[(X-\mu)^2\right],$$

onde μ = E(X) é a média de X.

☑ **Observações**:

- V(X) ≥ 0 e mede a variabilidade ou dispersão de X em torno da sua média μ;
- V(X) é expressa em unidades quadradas (o que torna difícil a sua interpretação);
- O **Desvio Padrão** $\sigma_X = \sqrt{V(X)}$ mede a dispersão absoluta de X, sendo expressa na mesma unidade da variável aleatória X.
- A definição de variância de uma variável aleatória (v.a.) X, pode ser re-escrita por:
-

$$\boxed{\sigma^2 = V(X) = E(X^2) - \left[E(X)\right]^2,}$$

onde: $E(X^2)=\sum_{i=1}^{\infty} x_i^2\, p(x_i)$.

x	– 1	9	4	14	19
x^2	$(-1)^2 = 1$	$(9)^2 = 81$	$(4)^2 = 16$	$(14)^2 = 196$	$(19)^2 = 361$
P(X = x)	0,977	0,008	0,008	0,006	0,001

Sabe-se que: E(X) = – 0,77

$E(X^2) = (1 \times 0{,}977) + (81 \times 0{,}008) + (16 \times 0{,}008) + (196 \times 0{,}006) + (361 \times 0{,}001)$

$E(X^2) = (0{,}977) + (0{,}648) + (0{,}128) + (1{,}176) + (0{,}361) = 3{,}29$

$V(X) = E(X^2) - [E(X)]^2 = 3{,}29 - (-0{,}77)^2 = 2{,}6971$

O desvio padrão também pode ser encontrado: $S = \sqrt{2{,}6971} = 1{,}642$

➘ Isso significa que, em média, nossos ganhos por jogo estarão a uma distância de 1,642 do valor esperado de – 0,77.

Propriedades Importantes do Valor Esperado

Sejam X uma v.a. e ***c*** = constante, então:

1. O valor esperado (média) de uma constante é a própria constante: $E(\boldsymbol{c}) = \boldsymbol{c}$.

2. Multiplicando-se uma constante por uma variável aleatória X, sua média fica multiplicada por esta constante:
$$E(\boldsymbol{c}.X) = \boldsymbol{c}.\ E(X).$$

3. Somando ou subtraindo uma constante de uma variável aleatória X, sua média fica somada ou subtraída desta constante:
$$E(X \pm \boldsymbol{c}) = E(X) \pm \boldsymbol{c}.$$

4. Sejam X e Y duas variáveis aleatórias, o valor esperado da soma/subtração de variáveis aleatórias equivale a soma/subtração dos valores esperados de X e Y:
$$E(X \pm Y) = E(X) \pm E(Y).$$

5. Sejam X e Y duas variáveis aleatórias independentes, temos que:
$$E(X.Y) = E(X).E(Y).$$

Propriedades Importantes da Variância

Sejam X uma v.a. e ***c*** = constante, então:

1. A variância de uma constante é zero: $V(\boldsymbol{c}) = 0$.

2. Multiplicando-se uma constante por uma variável aleatória X, sua variância fica multiplicada pelo quadrado da constante: $V(\boldsymbol{c}.X) = \boldsymbol{c}^2.\ V(X)$.

3. Sejam X e Y duas variáveis aleatórias independentes, a variância da soma/subtração de variáveis aleatórias equivale a soma das variâncias de X e Y: $V(X \pm Y) = V(X) + V(Y)$.

:: ARREGAÇANDO AS MANGAS!! ::

Até agora vimos como calcular e usar as distribuições de probabilidade, mas não seria bom se tivéssemos algo **mais fácil com que trabalhar** ou apenas mais rápido de calcular? A seguir vamos apresentar algumas distribuições de probabilidade especiais que seguem padrões muito bem definidos. Continue lendo para que possamos apresentá-lo a distribuição Binomial.

3.4. Experimentos Binomiais e a Distribuição Binomial

Dentre as funções de probabilidade, apresentaremos inicialmente uma distribuição discreta de grande importância, denominada ***Distribuição Binomial.*** Em seguida, faremos estudo de uma distribuição contínua de grande utilização na teoria da probabilidade, chamada a *Distribuição Normal*.

Para utilizar a teoria das probabilidades no estudo de um fenômeno concreto, devemos encontrar um modelo probabilístico adequado a tal fenômeno. Entendemos por *modelo probabilístico* para uma variável aleatória (v.a) X, uma forma específica de função de distribuição de probabilidade que reflita o comportamento de X. As propriedades básicas de um modelo probabilístico devem ser:

- **Adequação:** O modelo deve refletir adequadamente o mecanismo aleatório que ocasiona variação nas observações;
- **Simplicidade:** Utilização, sempre que possível, de hipóteses simplificadoras, de modo que o modelo se preste à análise estatística, sem sacrifício de adequação;
- **Parcimônia de Parâmetros:** Um número excessivo de parâmetros prejudicaria a análise estatística. Entre dois modelos que constituam aproximação adequada de um fenômeno, devemos preferir aquele que apresente o menor número de parâmetros.

☑ Distribuição de Bernoulli

Suponha que realizamos um experimento E, cujo resultado pode ser observado e classificado como *sucesso* ou *fracasso*, caso o evento que nos interessa ocorra ou não, respectivamente. Associe p, a probabilidade de sucesso, ao evento que nos interessa e $(1 - p) = q$, a probabilidade de fracasso. Definimos, então, a seguinte variável aleatória discreta:

$$X = \begin{cases} 0, & \text{se ocorrer fracasso} \\ 1, & \text{se ocorrer sucesso} \end{cases}.$$

Verifica-se que: $E(X) = p$ e $V(X) = p(1 - p)$, que são as principais características da variável aleatória X.

A distribuição de probabilidade de X é definida por:

x_i	0	1
$P(X = x_i)$	$(1 - p)$	p

☑ Distribuição Binomial (Experimentos Binomiais)

Um experimento binomial apresenta quatro propriedades:

❶ O experimento consiste em uma seqüência de *n* ensaios idênticos e independentes;

❷ Dois resultados são possíveis em cada ensaio. Um é denominado de **sucesso** e o outro de **fracasso**;

❸ A probabilidade de um sucesso é denotada por *p*, e não se modifica de ensaio para ensaio. (O mesmo se aplica à probabilidade de fracasso $q = (1 - p)$);

❹ Os ensaios são independentes;

Defina uma variável aleatória Y como sendo o número de sucessos nos *n* ensaios.

☑ **Definição 3.8:** Dizemos que uma variável aleatória discreta $X = X_1 + X_2 + ... + X_n$, onde cada X_i é um ensaio de Bernoulli, apresenta distribuição binomial com *n* provas (ensaios ou tentativas) e probabilidade *p* de sucesso, sendo sua função de probabilidade definida por:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \ k = 0,1,\ldots n,$$

pois, para $X = k$ teremos observado *k* sucessos, cada um com probabilidade *p* e conseqüentemente $(n - k)$ fracassos, cada um com probabilidade $q = (1 - p)$.

☺ **Notação:** X ~ Bin(*n*, *p*) equivale a dizer que *X* tem distribuição Binomial com parâmetros *n* e *p*.

OBS: $\binom{n}{k} = \frac{n!}{k!(n-k)!}$

Lembre-se que essa combinação representa o número de ensaios (n) em k tentativas.

Propriedades:

- E(X) = *np*
- V(X) = *npq*

Exemplo 3.5: Dois times de futebol, A e B, jogam entre si 6 vezes. Suponha que as probabilidades do time A ganhar, perder ou empatar sejam as mesmas e permaneçam constantes durante as 6 partidas. Encontre a probabilidade do time A ganhar 4 vezes e calcule a esperança e a variância.

☺ **Solução:**

Seja X = Número de vezes que o time A ganha.

X ~ Bin(*n*, *p*) ➔ X ~ Bin(6, 1/3)

Note que o time A pode ganhar, perder ou empatar, assim sendo: *p* = 1/3 (probabilidade de vencer, que corresponde ao **sucesso**) e, *q* = 2/3 (perder ou empatar, que corresponde ao **fracasso**). Além disso, eles jogam entre si 6 vezes, portanto, *n* = 6.

Logo, a probabilidade do time A ganhar 4 vezes é dada por:

$$P(X = 4) = \binom{6}{4}(1/3)^4(1-1/3)^{6-4} = 15 \times (1/3)^4 \times (2/3)^2 = \frac{20}{243} \cong 0{,}08.$$

$$OBS: \binom{6}{4} = \frac{6!}{4!(6-4)!} = \frac{6!}{4!2!} = \frac{6.5.4!}{4!2!} = \frac{6.5}{2.1} = \frac{30}{2} = 15$$

Esperança (média) de vitórias será: $E(X) = np = 6 \times \frac{1}{3} = 2$

Variância: $V(X) = npq = 6 \times \frac{1}{3} \times \frac{2}{3} = \frac{4}{3}$.

DADOS DISCRETOS TRABALHAM COM VALORES EXATOS...

Até agora estudamos distribuições de probabilidades em que os dados são ***discretos***. Isto é, os dados são compostos de valores numéricos distintos e podemos calcular a probabilidade de cada um desses valores. Se os dados forem discretos, eles podem assumir valores exatos. Geralmente, são dados que podem ser contados de alguma forma, tal como o número de balas em um baleiro, o número de perguntas respondidas corretamente em um programa de competição ou o número de vezes que a maré é cheia durante o período de um mês.

GUIA RÁPIDO PARA A DISTRIBUIÇÃO BINOMIAL

Veja um resumo de tudo que você precisa saber sobre distribuição Binomial:

❶ Quando usá-la?

Use a distribuição Binomial se estiver realizando um número fixo de tentativas independentes, onde cada tentativa pode ter um sucesso ou fracasso, e se você estiver interessado no número de sucessos ou fracassos.

❷ Como calcular as probabilidades?

Use, $P(X = k) = {}^{n}C_{k}.p^{k}.q^{(n-k)}$, onde, ${}^{n}C_{k}.= n!/[k!(n - k)!]$. Lembrando que, p = probabilidade de sucesso em uma tentativa, q = (1 – p), n = número de tentativas. Essa fórmula corresponde a apresentada anteriormente:

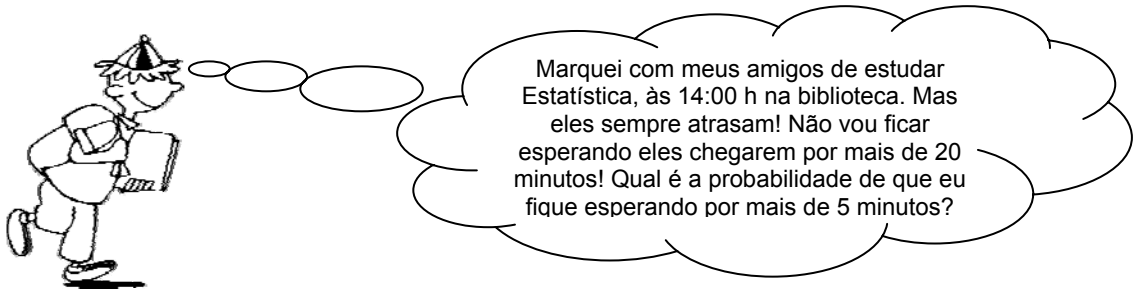
$P(X = k) = \binom{n}{k} p^{k}(1-p)^{n-k}$ lembrando que a combinação é dada por: $\binom{n}{k} = \frac{n!}{k!(n-k)!}$

❸ E o valor esperado e a variância?

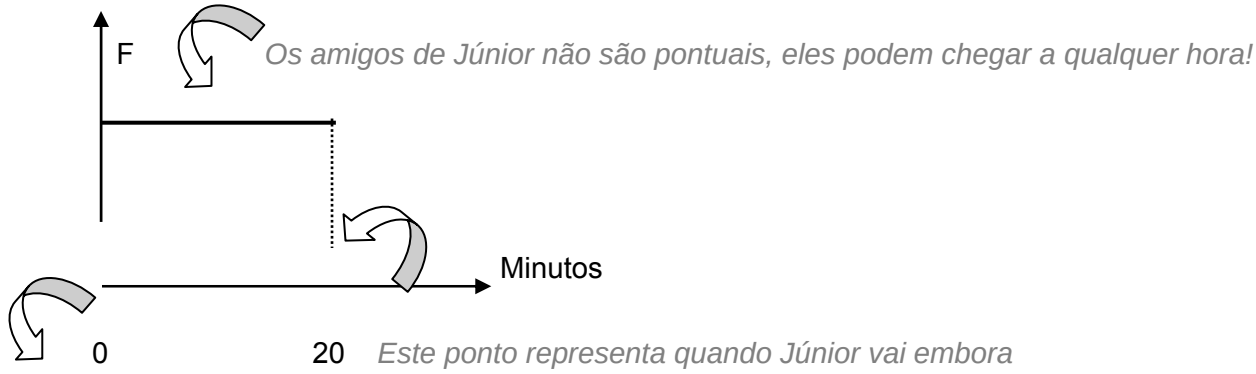
E(X) = np V(X) = npq

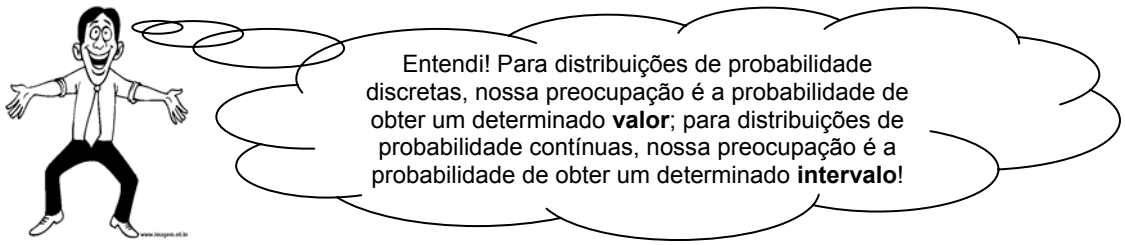
... MAS NEM TODOS OS DADOS NÚMERICOS SÃO DISCRETOS!

Nem sempre é possível dizer quais devem ser todos os valores de um conjunto de dados. Às vezes, os dados incluem um intervalo onde qualquer valor dentro daquele intervalo é possível. Como exemplo, suponha que você tivesse de medir com precisão pedaços de barbante com comprimento variando entre 10 e 11 polegadas. Poderia haver medidas 10 polegadas, 10,1 polegadas, 10,01 polegadas, e assim por diante, já que o comprimento poderia assumir qualquer valor dentro deste intervalo. Dados numéricos como estes são chamados de ***contínuos***. Geralmente, são dados que são medidos de alguma forma, em vez de contados, e tudo depende do grau de precisão da medição.



Veja o esboço da freqüência mostrando a quantidade de tempo que Júnior passa esperando seus amigos:





DISTRIBUIÇÕES DE PROBABILIDADE DISCRETAS NÃO RESOLVEM TODAS AS SITUAÇÕES

Até agora estudamos distribuições de probabilidades onde poderíamos especificar os valores exatos, mas nem sempre este é o caso para todos os conjuntos de dados. Alguns tipos de dados simplesmente não se enquadram nessas distribuições de probabilidades. Vamos examinar como funcionam as distribuições de probabilidade contínuas e introduzir você a uma das mais importantes distribuições de probabilidade que existem, a **distribuição Normal**.

3.5. Distribuição Normal

A distribuição normal é a mais importante das distribuições contínuas de probabilidade. Conhecida por alguns leitores como “a curva em forma de sino”, tem sua origem associada aos erros de mensuração. É sabido que, quando se efetuam repetidas mensurações de determinada grandeza com um aparelho equilibrado, não se chega ao resultado todas às vezes. Obtém-se, ao contrário, um conjunto de valores que oscilam, de modo aproximadamente simétrico, em torno do verdadeiro valor. Construindo um histograma desses valores e o correspondente polígono de

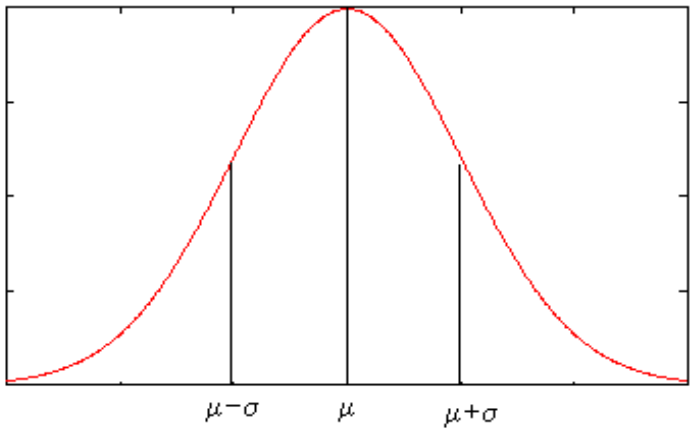
freqüências, obtém-se uma poligonal aproximadamente simétrica. A distribuição normal desempenha, não obstante, um papel preponderante na estatística e os processos de inferência, nela baseados, têm larga aplicação. Muitas das variáveis quantitativas analisadas em pesquisas nas diversas áreas de estudo correspondem ou se aproximam da distribuição normal.

AS FUNÇÕES DE DENSIDADE DE PROBABILIDADE PODEM SER USADAS PARA DADOS CONTÍNUOS

Podemos descrever a distribuição de probabilidade de uma variável aleatória contínua usando uma ***função de densidade de probabilidade***, *f*(*x*). É uma função que você pode usar para achar as probabilidades de uma variável contínua em um intervalo de valores.

Uma distribuição normal caracteriza-se por uma função real f(x) denominada de função densidade de probabilidade (f.d.p) da variável aleatória X, dado pelo modelo probabilístico abaixo e gráfico correspondente:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, \quad -\infty < x < +\infty, -\infty < \mu < +\infty, \sigma^2 > 0.$$



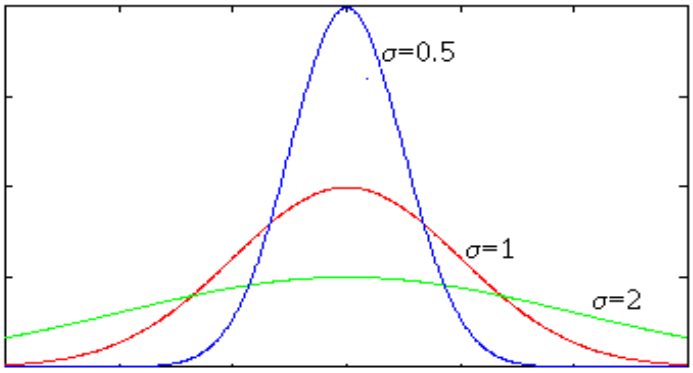
* A distribuição normal é uma curva em forma de sino.
* A curva é simétrica, com a densidade da probabilidade mais alta no centro da curva.
* A densidade da probabilidade diminui quanto mais você se afasta da média.
* Tanto a média quanto a mediana ficam no centro e têm a densidade de probabilidade mais alta.
* É definida por dois parâmetros: μ e σ^2. μ lhe diz onde está o centro da curva e σ lhe fornece a dispersão.

Propriedades da Curva Normal:

❶ É unimodal, isto é, f(x) tem um ponto de máximo cuja abscissa é x = μ. Esse ponto, situado no meio da distribuição, é aquele em que coincidem os valores da média, moda e mediana;
❷ f(x) é simétrica em relação à média μ;
❸ f(x) tem dois pontos de inflexão, cujas abscissas são x = ($\mu - \sigma$) e x = ($\mu + \sigma$);
❹ O desvio-padrão é dado por σ (a raiz quadrada positiva da variância σ^2);
❺ A área total sob a curva normal e acima do eixo horizontal equivale a 1 (o eixo das abscissas é o eixo dos valores da variável aleatória X);

❻ f(x) tem uma assíntota. A partir do topo, a curva cai gradativamente até formar as caudas que se estendem indefinidamente, aproximando-se cada vez mais da linha base sem, entretanto, jamais tocá-la.

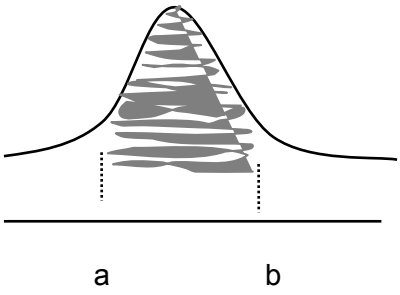
❼ Fixando-se a média, verifica-se que o achatamento da curva está diretamente ligado ao valor do desvio padrão σ, ou seja, quanto maior for o desvio padrão, mais achatada é a curva, como pode ser vista na figura abaixo.



☺ **Notação:** X ~ N (μ; σ^2), ou seja, X tem distribuição normal com média μ e variância σ^2.

ENTÃO COMO FAZEMOS PARA CALCULAR AS PROBABILIDADES NORMAIS?

Assim como qualquer outra distribuição de probabilidade contínua, você acha probabilidades calculando a área sobre a curva da distribuição. A curva dá a densidade da probabilidade, e a probabilidade é dada pela área entre determinados intervalos. Se, por exemplo, você quisesse achar a probabilidade de que uma variável aleatória X esteja entre *a* e *b*, seria preciso achar a área sob a curva entre os pontos *a* e *b*.



Parece complicado? Não se preocupe, é mais fácil do que você pensa. Calcular a área sob a curva normal seria difícil se você tivesse que fazer tudo isso sozinho, mas felizmente você tem uma mãozinha de ajuda na forma das tabelas de probabilidade. Basta calcular o intervalo da área que você deseja achar e, em seguida, consultar a probabilidade correspondente na tabela

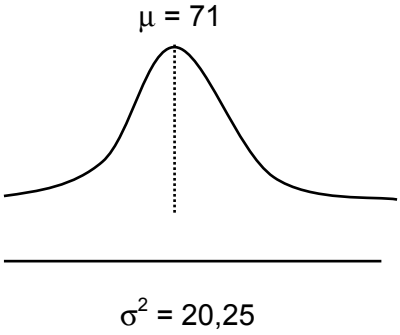
TRÊS PASSOS PARA CALCULAR PROBABILIDADES NORMAIS

❶ **Pegue sua distribuição e o intervalo** (você vai precisar da média e do desvio-padrão para achar suas probabilidades).

❷ **Padronize** (vamos mostrar como fazer isso em breve).

❸ **Consulte as probabilidades** (uma vez transformada sua curva normal, você pode consultar probabilidade usando tabelas. Pronto, tarefa cumprida!).

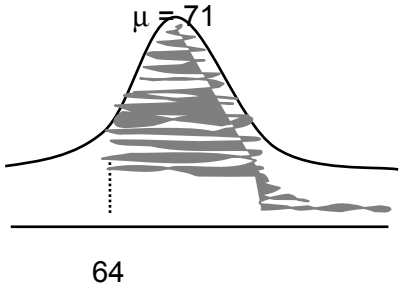
<u>PASSO 01:</u> DETERMINE SUA DISTRIBUIÇÃO



A primeira coisa que precisamos fazer é determinar a distribuição dos dados, ou seja, identificar os valores da média e do desvio-padrão. Suponha que Júnior e seus amigos foram pesquisados sobre as suas alturas. A altura média do grupo é de 71 polegadas, e a variância é de 20,25 $(\text{polegadas})^2$. Isso significa que, se X representa a altura dos alunos então, X ~ N(71; 20,25)

Também precisamos saber qual intervalo de valores nos dará a área de probabilidade correta. Necessitamos de um exemplo! Portanto, imagine que Júnior deseja achar a probabilidade de que algum dos seus amigos seja mais alto que ele. Júnior tem 64 polegadas de altura.

É fácil, Júnior quer saber se um dos seus amigos é mais alto do que ele. Por isso, podemos calcular as probabilidades com base na altura dele. A probabilidade de que haja alguém mais alto do que Júnior é dada pela expressão: **P(X > 64)**. Essa probabilidade é representada graficamente pela curva ao lado.



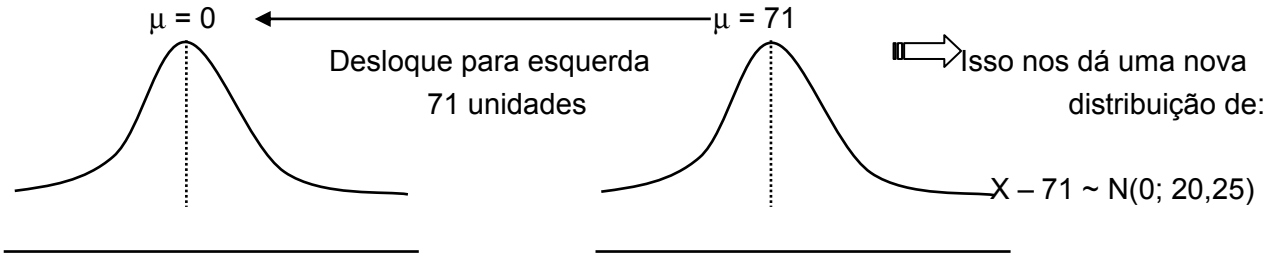
<u>PASSO 02:</u> PADRONIZE PARA N(0; 1)

O próximo passo é padronizar nossa variável X de forma que a média passe a ser 0 (zero)e o desvio-padrão 1. Isso nos dá uma **<u>variável normal padronizada</u>** Z, onde Z ~ N(0; 1)

Essa padronização se faz necessária porque as tabelas de probabilidade se concentram em dar as probabilidades para distribuições de N(0; 1), pois seria impossível gerar tabelas de probabilidade para cada curva de distribuição normal. Existe um número infinito de valores possíveis para μ e σ^2 e, como a curva normal utiliza esses valores como parâmetros para indicar o centro e a dispersão da curva, existe também um número infinito de curvas possíveis para a distribuição normal.

☺ Para padronizar, primeiro desloque a média....

Vamos começar transformando nossa distribuição normal para que a média passe a ser 0 em vez de 71. Para isso, deslocamos a curva para a esquerda em 71 unidades.



☺ ...em seguida, "comprima" a largura

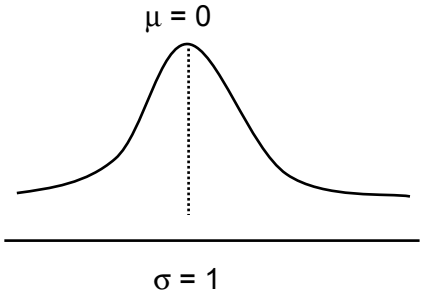
Também precisamos ajustar a variância. Para isso, "comprimimos" nossa distribuição dividindo pelo desvio-padrão. Sabemos que a variância é 20,25 e, por isso, o desvio-padrão é 4,5 (lembre-se da Unidade 1 que o desvio-padrão é a raiz quadrada da variância).

Ao fazer isso obtemos:

$$\frac{X-71}{4{,}5} \sim N(0;1)$$

Ou, Z ~N(0; 1) onde:

$$Z = \frac{X-71}{4{,}5}$$



"Comprima" a distribuição dividindo pelo desvio-padrão. Este é o ***escore padrão***. De modo geral você pode achar o escore padrão para qualquer variável Normal X usando:

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

☺ Agora ache Z para o valor específico para o qual você deseja achar a probabilidade

Até agora vimos como nossa distribuição de probabilidades pode ser padronizada para ir de X ~ N(μ; σ^2) a Z ~ N(0; 1). Aquilo que mais nos interessa são as probabilidades reais. O que precisamos fazer é tomar o intervalo de valores para o qual queremos achar as probabilidades e achar o escore padrão do limite desse intervalo. Em seguida, podemos consultar a probabilidade para o nosso escore padrão usando as tabelas de probabilidade normal.

Em nosso exemplo, deseja-se calcular a probabilidade de que tenha algum amigo mais alto do que Júnior. Precisamos achar **P(X > 64)**. O limite desse intervalo é 64 e, por isso, se calcularmos o escore padrão z de 64, poderemos usá-lo para achar nossa probabilidade. Portanto,

$$Z = \frac{X-\mu}{\sigma} = \frac{64-71}{4{,}5} = -1{,}56\,(com\ duas\ casas\ decimais)$$

Então – 1,56 é o escore padrão de 64, usando a média e o desvio padrão das alturas de Júnior e seus amigos. Agora que temos essa informação podemos passar para a etapa final, usando tabelas para consultar probabilidades.

Perceba que nos últimos parágrafos falamos sobre distribuição normal padronizada. Que tal pararmos um pouco agora para sermos apresentados formalmente a essa tão importante distribuição? Continue lendo e aprendendo! Não se esqueça que logo em seguida temos o último passo para encontrar nossa tão sonhada probabilidade!

☑ **Distribuição Normal Padrão:**

O cálculo direto de probabilidades envolvendo a distribuição normal não é um processo elementar. Notemos, entretanto, que a função de densidade normal depende de dois parâmetros, μ e σ, de modo que se tabelássemos as probabilidades diretamente a partir dessa função, seriam necessárias tabelas de dupla entrada para cada valor particular $\mu = \mu_0$ e $\sigma = \sigma_0$, complicando consideravelmente o problema. Recorre-se, por isso, a uma mudança de variável, transformando a variável aleatória X na variável aleatória Z assim definida:

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

Esta nova variável chama-se **variável normal padronizada**, ou **reduzida**, sendo sua média igual a **zero** ($\mu = 0$) e o seu desvio padrão é igual **um** ($\sigma = 1$).
Demonstração:

- Média: $E(Z) = E\left(\frac{X-\mu}{\sigma}\right) = \frac{E(X)-\mu}{\sigma} = \frac{\mu-\mu}{\sigma} = 0$
- Variância: $V(Z) = V\left(\frac{X-\mu}{\sigma}\right) = \frac{V(X)-0}{\sigma^2} = \frac{\sigma^2}{\sigma^2} = 1$

A curva normal padrão conserva as mesmas propriedades listadas anteriormente. Mediante tal transformação, basta construirmos uma única tabela, a da normal reduzida e, através dela, obtermos as probabilidades associadas a todas as distribuições $N(\mu; \sigma^2)$.

A utilidade notável da tabulação pela **variável normal padronizada** é devida ao fato de que, se X tiver ***qualquer*** distribuição normal $N(\mu; \sigma^2)$, a tabela da distribuição N(0; 1) pode ser empregada para calcular probabilidades associadas a X, simplesmente aplicando a transformada para a variável Z. Conseqüentemente, temos que:

$$P(a \le X \le b) = P\left(\frac{a-\mu}{\sigma} \le Z \le \frac{b-\mu}{\sigma}\right) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right),$$

onde $\Phi(z) = P(Z \le z)$, é a função de distribuição acumulada de N(0; 1).

PASSO 03: CONSULTE A PROBABILIDADE NA SUA TABELA

Agora que temos o escore padrão, podemos usar as tabelas de probabilidade para responder a pergunta do nosso amigo Júnior e achar a nossa probabilidade. As tabelas de probabilidade normal permitem consultar qualquer valor z e, depois, ler a probabilidade correspondente **P(Z < z)**.

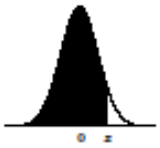
☺ **Então como usamos?**

Comece calculando z com 2 casas decimais. Este é o valor que você vai precisar consultar na tabela. Para consultar a probabilidade, é preciso usar a primeira coluna e a linha superior para achar seu valor z. A primeira coluna dá o valor de z com 1 casa decimal (sem arredondar) e a linha superior dá a segunda casa decimal. A probabilidade é onde as duas se interceptam!

Como exemplo, se quisesse achar P(Z < 0,86), você acharia 0,8 na primeira coluna, 0,06 na linha superior, e encontraria uma probabilidade de 0,8051.

Tabela 3: Função de distribuição Normal reduzida: $Z \sim N(0,1)$

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\,\pi}}\, e^{-\frac{1}{2}t^2}\, dt$$



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319

Acesse a plataforma MOODLE e faça o download da tabela da distribuição normal padronizada.

Vamos voltar ao exemplo das Alturas de Júnior e seus amigos. Relembrando, deseja-se calcular P(X > 64) = ? De acordo com o escore padrão queremos achar: P(Z > – 1,56). Portanto, vamos consultar – 1,56 na tabela. Obtemos uma probabilidade de 0,0594. Em outras palavras,

P(Z < – 1,56) = 0,0594.

Isso significa que: P(Z > – 1,56) = 1 – P(Z < – 1,56) = 1 – 0,0594 = 0,9406.

:: SAIBA MAIS... ::

As tabelas de probabilidades possibilitam que você consulte a probabilidade P(Z < z), onde z é algum valor. O problema é que você nem sempre deseja achar esse tipo de probabilidade; às vezes, você deseja achar a probabilidade de que uma variável aleatória contínua seja maior que z, isto é, P(Z > z), ou ainda, esteja entre dois valores, P(a < Z < b). Como usar as tabelas para achar a probabilidade que você procura?

*** Achando P(Z > z):** Nesse caso, diminua da probabilidade total a área onde Z < z. Em outras palavras, use: **P(Z > z) = 1 – P(Z < z)**.

*** Achando P(a< Z < b):** Nesse caso, calcule P(Z < b) e diminua a área de P(Z < a). Em outras palavras, use: **P(a< Z < b) = P(Z < b) – P(Z < a)**.

Exemplo 3.9: Os salários médios diário dos biólogos de uma empresa são distribuídos segundo uma distribuição normal com média de R$ 50,00 e desvio padrão de R$ 4,00. Encontre a probabilidade de um operário ter um salário diário abaixo de R$ 52,00?

☺ **Solução:**

Seja X = o salário diário dos funcionários. Interesse: Encontrar P (X < 52). Assim,

$$P(X < 52) = P\left(Z < \frac{52-\mu}{\sigma}\right) = P\left(Z < \frac{52-50}{4}\right) = P(Z < 0{,}50) = \Phi(0{,}50) = 0{,}6915\,.$$

Pode-se afirmar que a probabilidade de um biólogo apresentar um salário inferior a R$ 52,00 é de 69,15%.

:: SAIBA MAIS... ::

Dica:

Três importantes informações que irão facilitar o cálculo de probabilidades envolvendo a distribuição normal padrão, a partir da tabela que você baixou na plataforma MOODLE: (i) a tabela que você está utilizando apresenta as probabilidades de $P(Z \leq z_0) = F(z_0)$, ou seja, a função de distribuição acumulada. No entanto, esta tabela considera apenas valores positivos para Z. (ii) a área total sob a curva equivale a 1. Logo, a metade da curva representa probabilidade igual a 0,5; (iii) a curva da normal é simétrica. Essa propriedade será bastante útil no cálculo de probabilidades onde os valores de Z são negativos, ou seja, $P(X \leq -x_0) = 1 - P(X \leq +x_0)$.

4. Avaliando o que foi construído

Nesta unidade aprendemos o conceito de função de distribuição de probabilidade, o conceito de variável aleatória, além dos conceitos de esperança e variância de variáveis aleatórias. Conhecemos também duas distribuições importantíssimas na estatística que são as distribuições: Binomial e Normal. Particularmente, a distribuição normal será uma ferramenta essencial nas unidades seguintes. Faça todos os exercícios propostos, pois eles serão de grande valia. Aguardo você no MOODLE!

UNIDADE 4
TEORIA ELEMENTAR DA AMOSTRAGEM

1. Situando a Temática

Amostragem é uma área da Estatística que estuda técnicas de planejamento de pesquisa para possibilitar inferências sobre uma população a partir do estudo de uma pequena parte de seus componentes, uma amostra.

2. Problematizando a Temática

Ao fazermos uma jarra de suco e adicionamos açúcar desejamos saber se a quantidade de açúcar foi satisfatória. Para isto, não precisamos tomar toda a jarra de suco, uma colher basta. Da mesma forma, ao estudarmos um fenômeno probabilístico em uma população não precisamos investigar toda a população, e sim uma amostra dela. No entanto, algumas questões podem surgir: Como obter essa amostra? Qual deve ser o tamanho dessa amostra? Esta unidade tem como objetivo responder esta e mais algumas questões correlatas.

A ESTATÍSTICA TRABALHA COM DADOS, MAS DE ONDE VÊM OS DADOS?

Algumas vezes, é fácil colher dados, tais como as idades das pessoas que frequentam uma academia de ginástica ou os números das vendas de uma empresa fabricante de jogos. Mas e quando os dados são difíceis de colher? Às vezes, é difícil saber por onde começar a colher dados devido ao grande número de coisas envolvidas. Neste capítulo, vamos examinar como colher dados de forma eficaz na vida real, com eficácia, precisão, economia de tempo e dinheiro. Bem-vindo a mundo da amostragem!

COMO FUNCIONAM AS AMOSTRAGENS?

População
☺☺☺☺☺☺☺☺☺☺
☺☺☺☺☺☺☺☺☺☺
☺☺☺☺☺☺☺☺☺☺
☺☺☺☺☺☺☺☺☺☺
☺☺☺☺☺☺☺☺☺☺

A chave para se criar uma boa amostra é escolher uma amostra que combine o máximo possível com sua população. Se sua amostra for representativa, isto significa que ela tem características semelhantes à população. E isso, por sua vez, significa que você pode usar sua amostra para prever quais características a população terá.

Amostra
☺☺☺
☺☺☺
☺☺☺
☺☺☺
☺☺☺

Amostragem

3. Desenvolvendo a Temática

3.1. Conceitos Básicos

Muitas vezes faz-se necessária a coleta de dados diretamente na origem. Entretanto, quando é impossível se observar toda a população recorremos às técnicas de amostragem, onde

nos limitamos a uma amostra da população em estudo. Basicamente, nosso objetivo é coletar uma pequena fração da população de modo que as informações observadas na amostra possam ser generalizadas para a população. Para que esta generalização seja possível, os integrantes da amostra devem ser escolhidos adequadamente.

Antes de aprofundarmos nosso discurso, vamos definir alguns termos necessários:

- **População Objeto**: É a população de interesse sobre a qual desejamos obter informações (Exemplo: peças produzidas em uma fábrica);
- **População de Estudo**: É o conjunto de indivíduos de interesse específico (Exemplo: peças que permanecem em estoque);
- **Característica Populacional**: Aspectos da população que interessam serem medidos ou observados (Exemplo: diâmetro da peça);
- **Unidade Amostral**: Definida de acordo com o interesse do estudo, podendo ser uma peça, um indivíduo, uma fazenda, etc. Tal escolha deve ser feita no início do estudo;
- **Estrutura Amostral** ou **Amostra**: É o conjunto de unidades amostrais (Exemplo: o conjunto das peças selecionadas).

É importante ressaltar que existem dois tipos de amostragem, a saber:

- **Amostragem Probabilística**: É o procedimento através do qual existe uma probabilidade conhecida e diferente de zero (p) para cada elemento da população ser escolhido para constituir a amostra;
- **Amostragem Não-Probabilística**: Quando, no processo de seleção, não existe nenhum mecanismo probabilístico para selecionar os indivíduos da população para constituir a amostra.

:: ARREGAÇANDO AS MANGAS!! ::

Defina sua população alvo: A primeira coisa a se deixar bem claro é qual é a sua **população alvo** para que você saiba de onde está colhendo sua amostra. População alvo é o grupo que você está pesquisando e para o qual deseja colher resultados e, depende, em grande parte, da finalidade do seu estudo.

Defina as unidades amostrais: Uma vez definida a população alvo, é preciso decidir para que tipo de objeto você vai colher amostras.

Defina seu plano amostral: Por último, você precisa de uma lista de todas as unidade amostrais dentro da sua população alvo. É basicamente uma lista a partir da qual você pode escolher sua amostra.

De acordo com a definição de amostragem probabilística, existe a suposição de um sorteio com regras bem determinadas, cuja realização só será possível se a população for finita e totalmente acessível. **Esse tipo de amostragem é a melhor garantia para se obter uma representatividade da população pela amostra**. Os principais planos de amostragem probabilística são:

1. **Amostragem Aleatória (ou Casual) Simples:** Neste tipo de plano, supõe-se que todos os elementos da *população* têm igual probabilidade de pertencer à *amostra*, ou alternativamente, se todas as possíveis amostras, de mesmo tamanho, têm a mesma probabilidade de serem selecionadas. Normalmente, consideramos esse tipo de plano amostral quando a população é *homogênea*. Esse processo de amostragem pode ser feito com ou sem reposição do elemento amostrado. Uma técnica que garante esta igual probabilidade é a seleção aleatória de elementos, por exemplo, através de sorteio.
2. **Amostragem Sistemática:** Inicia com uma escolha aleatória de um elemento da população e, a partir deste, usa-se um sistema de seleção para compor o restante da amostra. Por exemplo, numa listagem de elementos da população, sorteamos um entre os dez primeiros da lista – o 5^o elemento. A partir do elemento sorteado, selecionamos um a cada quinze elementos (o 20^o, o 35^o e assim por diante). Este método de amostragem pode ser utilizado quando se quer planejar um período de tempo para execução da coleta de dados ou quando se deseja cobrir um determinado período de tempo com a amostra estudada. Também consideramos esse tipo de plano amostral quando a população é *homogênea*.
3. **Amostragem Estratificada:** Na amostragem estratificada a população é dividida em grupos internamente homogêneos (*estratos*) e em seguida é selecionada uma amostra aleatória de cada estrato. Este tipo de amostragem é usado quando o evento estudado numa população tem características distintas para diferentes categorias que dividem esta população, ou seja, dentro de cada estrato os elementos são bastante semelhantes entre si e, entre os estratos eles são *heterogêneos*. Assim, a estratificação é apropriada para agrupar os elementos por sexo, faixa etária, religião, escolaridade ou em populações heterogêneas como rendas, produções agrícolas, produções industriais, etc.
4. **Amostragem por Conglomerados:** A população é dividida em pequenas subpopulações, com elementos internamente heterogêneos, chamadas **conglomerados** (*clusters*). Seleciona-se uma amostra aleatória simples desses conglomerados, e deles selecionam-se aleatoriamente os elementos que irão compor a amostra. Assim, numa pesquisa sócio-econômica pode-se dividir a cidade em bairros (conglomerados), em seguida obter uma amostra aleatória de bairros e, então efetuar o levantamento estatístico nas residências dos bairros selecionados. Observe que, no caso da estratificação, indivíduos serão selecionados em cada estrato, enquanto no caso da divisão da população em conglomerados, selecionamos apenas parte dos conglomerados.
5. **Amostragem por Estágios Múltiplos:** Esta estratégia de amostragem pode ser vista como uma combinação de dois ou mais planos amostrais. Considere por exemplo uma população estratificada onde o número de estratos é muito grande. Ao invés de obter uma amostra aleatória de cada estrato, o que poderia ser inviável devido à quantidade de estratos, o pesquisador poderia optar por selecionar aleatoriamente alguns estratos e em seguida selecionar uma amostra de cada estrato selecionado. Neste caso, teríamos uma amostragem em dois estágios usando, nas duas vezes, a amostragem aleatória simples, sendo que no primeiro estágio as unidades amostrais são os estratos e no segundo são as componentes da população.

É importante ressaltar que certos cuidados dever ser tomados no processo de obtenção de uma amostra, ou seja, no processo de "amostragem", pois muitas vezes erros grosseiros e conclusões falsas ocorrem devido a falhas nesse processo.

:: SAIBA MAIS... ::

Exemplo de Amostragem Estratificada: Podemos dividir gomas de mascar nas diferentes cores (amarelo, verde, vermelho e cor-de-rosa). Cada cor representa um estrato diferente. Uma vez feito isso, você pode fazer a AAS em cada estrato para ter certeza de que cada grupo esteja representado na sua amostra global.

Exemplo de Amostragem por Conglomerados: Gomas de mascar podem ser vendidas em pacotes, onde cada pacote contém um número semelhante de gomas de mascar com cores diferentes. Cada pacote forma um conglomerado. Retira-se uma AAS de conglomerados e, em seguida, faz a pesquisa com tudo que está dentro de cada um desses conglomerados selecionados.

3.2. Distribuição Amostral da Média e da Proporção

Na Inferência Estatística, o principal problema é fazer uma afirmação sobre um parâmetro populacional (média, proporção, variância, etc.) baseado em informações coletadas de uma amostra, através de um Estimador (média amostral, variância amostral, etc.). No entanto, a validade de nossa afirmação seria melhor compreendida se soubéssemos o comportamento do Estimador ao retirarmos todas as amostras possíveis de tamanho *n* de uma população de tamanho *N*. Em outras palavras, estamos interessados em conhecer a **distribuição amostral** de um Estimador.

VAMOS COMEÇAR ESTIMANDO A MÉDIA DA POPULAÇÃO!

Suponha o exemplo das gomas de mascar. Como podemos usar os resultados do teste do sabor da amostra para saber o tempo médio de duração do sabor da goma de mascar na população total de gomas?

* A resposta é bastante intuitiva. Consideramos que a média de duração do sabor das gomas de mascar na amostra seja igual a da população. Em outras palavras, achamos a média amostral e a usamos como média da população também.

* A média amostral é a melhor estimativa que podemos fazer para a média populacional. É o valor mais provável para a média da população.

* A média amostral é chamada de **estimador pontual** para a média populacional.

☹ Então, como fazermos para achar a distribuição das médias amostrais?

Suponha a população de embalagens de gomas de mascar. Foram fornecidas a média (μ) e a variância (σ^2) da população. Seja X = número de gomas de mascar em uma embalagem.

➘

Cada embalagem escolhida aleatoriamente é uma **observação independente** de X, e por isso, cada embalagem de goma de mascar segue a mesma distribuição. Isto é, se X representa o número de gomas de mascar em uma embalagem escolhida aleatoriamente, então cada X tem um valor esperado μ e uma variância σ^2.

➘

Agora, vamos tomar uma amostra de *n* embalagens de gomas de mascar. É possível identificar o número de gomas de mascar nas embalagens de X_1 a X_n. Cada X_i é uma observação independente de X, o que significa que elas seguem a mesma distribuição. Cada X_i tem um valor esperado μ e uma variância σ^2.

➘

A média de gomas de mascar nessas *n* embalagens é representada por $\overline{X}$. O valor de $\overline{X}$ depende de quantas gomas de mascar existem em cada uma das *n* embalagens e, para calculá-lo, é preciso somar o número total de gomas de mascar e dividir por *n*.

$E(X_1) = \mu$ $\quad Var(X_1) = \sigma^2$ $E(X_2) = \mu$ $\quad Var(X_2) = \sigma^2$ ... $E(X_n) = \mu$ $\quad Var(X_n) = \sigma^2$	Esta é a média amostral, o número médio de gomas de mascar nas embalagens: $\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$

Existem ***muitas*** possíveis amostras de tamanho *n* que poderíamos ter tomado. Cada amostra possível é composta por *n* embalagens, o que significa que cada amostra é composta por *n* observações independentes de X. O número de gomas de mascar em cada embalagem escolhida aleatoriamente segue a mesma distribuição de todas as outras, e calculamos o número médio de gomas de mascar para cada amostra da mesma maneira.

Podemos formar uma distribuição a partir de todas as médias amostrais de todas as amostras possíveis. Isso é chamado de **distribuição amostral das médias**.

3.2.1. Distribuição da Média Amostral

A distribuição amostral das médias nos dá uma maneira de calcular probabilidades para a média de uma amostra. Para calcularmos a probabilidade de qualquer variável, primeiro é preciso saber sobre sua distribuição de probabilidades, isto significa que, se você quiser calcular probabilidades para a média amostral, é preciso saber como são distribuídas as médias amostrais.

A distribuição amostral da média $\overline{X}$ é uma distribuição que mostra as probabilidades de obter os possíveis valores das médias amostrais. Vamos supor uma população {1, 3, 5} com *N* = 3 elementos e a variável aleatória X assumido o valor do elemento na população, com a seguinte distribuição de probabilidade:

x_i	1	3	5
$P(X = x_i)$	1/3	1/3	1/3

Observe que a distribuição acima tem média (valor esperado) e variância dados por:

$$E(X) = \mu = \frac{1+3+5}{3} = 3 \text{ e } Var(X) = \sigma^2 = \frac{(1-3)^2 + (3-3)^2 + (5-3)^2}{3} = \frac{8}{3}.$$

Se retirarmos todas as amostras aleatórias de tamanho *n* = 2, com reposição, dessa população obtemos um total de $N^n = 3^2 = 9$ amostras com os seguintes resultados:

(1,1) (1,3) (1,5) (3,1) (3,3) (3,5) (5,1) (5,3) (5,5).

Considerando essas 9 possibilidades igualmente prováveis, podemos construir a *distribuição amostral da média* para uma amostra de tamanho 2. Para tanto, basta calcular a média de cada uma dessas amostras obtendo os seguintes valores $\bar{x}_i$: 1, 2, 3, 2, 3, 4, 3, 4, 5, respectivamente. Note que, a partir das médias amostrais obtidas nas 9 amostras possíveis, é possível obtermos a seguinte distribuição amostral para $\overline{X}$:

$\bar{x}_i$	1	2	3	4	5
$P(\overline{X} = \bar{x}_i)$	1/9	2/9	3/9	2/9	1/9

Ainda com respeito à distribuição amostral de $\overline{X}$, acima apresentada, observa-se que:

- A sua média (valor esperado) é igual à média da população, ou seja,
  $E(\overline{X}) = \mu_{\overline{X}} = = \left(1\times\frac{1}{9}\right) + \left(2\times\frac{2}{9}\right) + \left(3\times\frac{3}{9}\right) + \left(4\times\frac{2}{9}\right) + \left(5\times\frac{1}{9}\right) = \frac{27}{9} = 3 = \mu$;

- A sua variância é igual à variância da população dividida pelo tamanho da amostra. Temos que: $V(\overline{X}) = E(\overline{X}^2) - \left[E(\overline{X})\right]^2$.
  Logo, $E(\overline{X}^2) = \left(1^2\times\frac{1}{9}\right) + \left(2^2\times\frac{2}{9}\right) + \left(3^2\times\frac{3}{9}\right) + \left(4^2\times\frac{2}{9}\right) + \left(5^2\times\frac{1}{9}\right) = \frac{93}{9}$.
  Assim, $V(\overline{X}) = E(\overline{X}^2) - \left[E(\overline{X})\right]^2 = \frac{93}{9} - 3^2 = \frac{93}{9} - 9 = \frac{93-81}{9} = \frac{12}{9} = \frac{4}{3} = \frac{8/3}{2} = \frac{\sigma^2}{n}$. Tais relações entre μ e $\mu_{\overline{X}}$, e σ^2 e $\sigma^2_{\overline{X}}$, observadas no exemplo acima, podem ser generalizadas.

Usando a teoria das probabilidades é possível mostrar que os seguintes resultados gerais são válidos com relação à distribuição amostral da média. Seja X uma variável aleatória com valor esperado $E(X) = \mu$ e variância $V(X) = \sigma^2$ finita, isto é, $0 < \sigma^2 < \infty$. Seja $\overline{X}$ a média desta variável aleatória, obtida de amostra aleatória de tamanho *n*, selecionada *com reposição*. Então, temos que:

- $E(\overline{X}) = \mu_{\overline{X}} = \mu$;
- $V(\overline{X}) = \sigma^2_{\overline{X}} = \frac{\sigma^2}{n}$.

Além disso, tem-se o resultado conhecido como **Teorema Central do Limite**: Seja X uma variável aleatória com valor esperado E(X) = μ e variância V(X) = σ^2. Para *n* suficientemente grande,

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

apresenta, aproximadamente, uma distribuição normal com média μ e variância (σ^2/n), Logo,

$$\overline{X} \approx N\left(\mu, \frac{\sigma^2}{n}\right) \text{ e } Z = \frac{\overline{X} - \mu}{\left(\sigma/\sqrt{n}\right)} \approx N(0,1).$$

A distribuição da variável padronizada Z é conhecida por **Distribuição Normal Padrão**.

☑ Observações:

(1) O desvio padrão de $\overline{X}$, denotado por $\sigma_{\overline{X}} = \sigma/\sqrt{n}$, é chamado **erro padrão da média** e descreve a variabilidade das médias amostrais em torno da verdadeira média populacional μ. Assim, quanto maior o erro padrão da média, maior será a diferença entre parâmetro μ e sua estimativa $\overline{X}$, calculada a partir da amostra. Quando *n* é grande (σ^2/n) decresce, significando que a média amostral fornecerá uma estimativa mais segura para μ em grandes amostras.

(2) Para amostras *sem reposição,* de população finita, temos a média $\mu_{\overline{X}}$ = E($\overline{X}$) = μ e variância $\sigma^2_{\overline{X}}$ = V($\overline{X}$) = $\frac{\sigma^2}{n}.\frac{N-n}{N-1}$, onde N é o total de elementos da população.

(3) Para valores grandes de *n* ($n \geq 30$) a aproximação da distribuição amostral da média $\overline{X}$ pela distribuição Normal é considerada satisfatória.

Exemplo 4.1: Os registros de uma agência de turismo mostram que um turista gastou, durante o último ano, em média μ = US$ 800,00, sendo o desvio padrão dos gastos igual a σ = US$ 80,00. Ache a probabilidade de que uma amostra de 64 turistas apresente um gasto médio entre US$ 770,00 e US$ 825,00.

☺ Solução:

Considere a variável X = gastos (em US$). Embora a distribuição de X não seja conhecida, como o tamanho da amostra *n* = 64 é bastante grande, podemos admitir que a média amostral de $\overline{X}$ segue a distribuição Normal com parâmetros:

$\mu_{\overline{X}} = \mu =$ 800 e $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{80}{\sqrt{64}} = 10$.

Assim temos que

$$P(770 \leq \overline{X} \leq 825) = P\left(\frac{770-800}{10} \leq \frac{\overline{X}-\mu}{\left(\sigma/\sqrt{n}\right)} \leq \frac{825-800}{10}\right) = P(-3{,}0 \leq Z \leq 2{,}5) =$$

$$P(Z \leq 3{,}0) - P(Z \leq 2{,}5) = 0{,}9938 - 0{,}0013 = 0{,}9925.$$

Se considerarmos um grande número de amostras, cada uma com 64 turistas, em aproximadamente 99,25% delas o gasto médio estaria entre US$770,00 e US$825,00.

Ufa, entendi! Veja bem, de um modo geral, você espera que o número médio de gomas de mascar por embalagem de uma amostra seja igual ao número médio de gomas de mascar por embalagens da população. Em nossa situação, o número médio de gomas de mascar em cada embalagem da população é 10 e, portanto, isso é o que você espera para a amostra também!

☹ **Então, como fazermos para achar a distribuição das proporções amostrais?**

Continuando com a população de gomas de mascar: foi fornecida a proporção de gomas de mascar vermelhas na população, e podemos representá-la por *p*. Em outras palavras, p = 0,25 (25% das gomas de mascar são vermelhas).

⬇

Cada caixa jumbo de gomas de mascar é, na verdade, uma amostra de gomas de mascar colhida na população. Cada caixa contém 100 gomas de mascar; logo o tamanho da amostra é 100, e será representado por *n*.

Seja a variável aleatória X = Número de gomas de mascar vermelhas na amostra, então: X ~ Bin(*n*; p), onde *n* = 100 e p = 0,25. A proporção de gomas de mascar vermelhas da amostra depende de X. Isso significa que a proporção em si é uma variável aleatória. Podemos representá-la por $\hat{p}$, onde $\hat{p} = x/n$.

Não sabemos o número exato de gomas de mascar vermelhas na amostra, mas sabemos sua distribuição.

Há ***várias*** amostras possíveis com tamanho *n* que poderíamos ter colhido. Cada amostra possível compreenderia *n* gomas de mascar e o número de gomas de mascar vermelhas em cada uma delas seguiria e mesma distribuição. Para cada amostra, o número de gomas de mascar vermelhas é distribuído conforme Bin(*n*; p).

⬇

Podemos formar uma distribuição a partir de todas as proporções amostrais usando todas as amostras possíveis. Isso é chamado de **distribuição amostral de proporções**.

3.2.2. Distribuição Amostral da Proporção

Se o parâmetro de interesse *p* representa uma proporção (ou percentagem) de elementos com certa característica (atributo) na população, então chamamos a estatística correspondente na amostra de *proporção amostral*, denotando-a por:

$\hat{p} = x/n$, onde *x* = Nº de elementos da amostra que possuem a característica de interesse.

Pode-se mostrar que, sob certas condições, e se *n* é suficientemente grande, a distribuição da proporção amostral $\hat{p}$ é aproximadamente Normal, com média E($\hat{p}$) = *p*, e variância V($\hat{p}$) = *p.q*/*n* onde *q* =(1 – *p*). Dessa forma, temos que:

$$\hat{p} \approx N\left(p, \frac{p.q}{n}\right) \text{ e } Z = \frac{\hat{p} - p}{\sqrt{\frac{p.q}{n}}} \approx N(0,1).$$

No caso de uma população finita de tamanho N e uma amostra *sem reposição*, recomenda-se o uso do fator de correção populacional no cálculo da variância de $\hat{p}$, sendo expressa por:

$$V(\hat{p}) = \frac{p.q}{n}\frac{N-n}{N-1}.$$

Exemplo 4.2: Suponha que de um grande lote de produção, 10% dos itens produzidos apresentam algum tipo de defeito. Em uma amostra aleatória de tamanho 60, obtida do lote para inspeção de qualidade, calcule a probabilidade de ter mais de 15% dos itens defeituosos.

☺ **Solução:**

$$P(\hat{p} > 0{,}15) = P\left(\frac{\hat{p} - p}{\sqrt{\frac{p.q}{n}}} > \frac{0{,}15 - 0{,}10}{\sqrt{\frac{0{,}1 \times 0{,}9}{60}}}\right) = P(Z > 1{,}29) = 1 - 0{,}9015 = 0{,}0985.$$

Se considerarmos um grande número de amostras, cada uma contendo 60 itens, em aproximadamente 9,85% das amostras a proporção de itens defeituosos seria superior a 15%.

Entendi! A distribuição amostral da proporção é, na verdade, uma distribuição de probabilidade formada pelas proporções de todas as amostras possíveis de tamanho *n*. Se soubermos como são distribuídas as proporções poderemos usá-la para achar as probabilidades referentes à proporção de uma determinada amostra.

4. Avaliando o que foi construído

Nesta unidade aprendemos como coletar e determinar o tamanho de uma amostra. Agora já temos conhecimentos básicos para estudarmos alguns conceitos sobre de estimação de parâmetros. Portanto, programe-se. Planeje seus estudos. Já há muito o que estudar sobre distribuições amostrais.

UNIDADE 5
INTERVALOS DE CONFIANÇA E TESTES DE HIPÓTESES

1. Situando a Temática

Quando estudamos fenômenos probabilísticos estudamos também o comportamento de alguns parâmetros relacionados a este experimento. Tais parâmetros, muitas vezes, são impossíveis de serem determinados restando-nos apenas tentar estimá-los da melhor forma possível. Os procedimentos para tal estimação, juntamente com o fato de termos certeza que estamos obtendo uma boa estimativa para o parâmetro, será abordado nessa unidade quando estudaremos intervalos de confiança e testes de hipótese.

2. Problematizando a Temática

Qual a altura média do povo brasileiro? Qual a proporção de pessoas com nível superior em João Pessoa? A resposta para essas perguntas não são tão fáceis, mas para respondê-las com exatidão teríamos que medir todos os cidadãos brasileiros ou verificar quantos habitantes em João Pessoa possuem nível superior, o que é impossível. No entanto se coletarmos uma amostra e calcularmos a média e a proporção, respectivamente, será que essas estimativas estão próximas dos verdadeiros valores populacionais (parâmetros)? Outra pergunta seria a seguinte: Se a quantidade média de água ingerida por um ser humano é de 10 litros por semana, os brasileiros bebem muito ou pouca água? Como responderíamos a esta questão? A resposta para essas questões será vista nessa unidade.

ÀS VEZES, AS AMOSTRAS NÃO DÃO EXATAMENTE O RESULTADO CORRETO

Você já viu como usar estimadores pontuais para estimar o **valor exato** da média, variância ou proporção da população, mas a questão é como ter certeza de que sua estimativa está totalmente correta. Afinal de contas, suas considerações sobre a população dependem de apenas uma amostra, e se sua amostra fugir à regra? Neste capítulo, você vai ver **outra maneira de estimar as estatísticas de uma população**, uma maneira que **leva em conta as incertezas**. Pegue suas tabelas de probabilidade e vamos mostrar-lhes cada detalhe dos **intervalos de confiança**.

O PROBLEMA DA PRECISÃO

Estimadores pontuais são a melhor estimativa que podemos dar para as estatísticas de uma população. Você tomar uma amostra representativa dos dados e a usa para estimar estatísticas importantes sobre a população tais como: a média, a variância e a proporção. O problema ao deduzir estimadores pontuais é que dependemos dos resultados de uma única amostra para obtermos uma estimativa bastante precisa.

Os estimadores pontuais são de grande valor, mas podem gerar pequenos erros. Como não estamos trabalhando com a população inteira, estamos apenas dando uma melhor estimativa. Se a amostra que usamos for não-tendenciosa, é provável que a estimativa seja próxima ao verdadeiro valor da população. A questão é até que ponto próximo é próximo suficiente.

Em vez de dar um valor exato como estimativa para a média da população, podemos especificar alguns intervalos com uma estimativa. Por exemplo, poderíamos dizer que esperamos que o sabor da goma de mascar dure entre 55 e 65 minutos. Assim, continuamos tendo a impressão de que o sabor dura por aproximadamente 1 hora, mas temos uma certa margem de erro.

☹ A questão é com achar esse intervalo? Tudo depende da confiança que você deseja que seus resultados tenham...

3. Conhecendo a Temática

3.1. Estimação de Parâmetros

Há inúmeras situações reais em que se procura determinar valores para quantidades desconhecidas como médias e proporções. Certamente, é de interesse para muitos empresários saber a quantia média gasta por um turista em sua cidade; um produtor de televisão procura sempre saber qual o índice de audiência de determinados programas; um engenheiro de controle de qualidade procura determinar a proporção de itens produzidos com defeito em uma linha de produção.
A *estimação* consiste em determinar um valor amostral que substitua o respectivo valor real do parâmetro populacional desconhecido.

3.1.1. Conceitos Fundamentais

Para uma melhor compreensão dos temas mais importantes desta unidade, vamos definir alguns conceitos fundamentais dentro da inferência estatística:

- ***Estimador:*** É uma função matemática que leva em consideração os dados amostrais. Como tal função é calculada baseada em uma amostra, é considerada uma variável aleatória, caracterizada por uma distribuição de probabilidade. Assim, $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$, onde x_1, x_2, ... , x_n são *n* valores amostrais, é um estimador que representa a média populacional (parâmetro).

- ***Estimativa*****:** É um valor particular do estimador para uma dada amostra coletada. Assim, por exemplo, para uma dada amostra, $\overline{X} = 3{,}9kg$ pode ser uma estimativa para o verdadeiro peso médio, desconhecido, de recém-nascidos do sexo feminino em certa localidade.

- ***Estimação por ponto* (*Estimação Pontual*)****:** Chamamos de estimação pontual quando, a partir de uma amostra, um único valor é usado para estimar um parâmetro desconhecido.

Um estimador pontual para um parâmetro populacional θ, é geralmente representado por $\hat{\theta}$. Assim, $\overline{X}$, S^2, S e $\hat{p}$ são estimadores pontuais para os parâmetros μ, σ^2, σ e p respectivamente, isto é, $\hat{\mu} = \overline{X}$, $\hat{\sigma}^2 = S^2$, $\hat{\sigma} = S$ e $p = \hat{p} = x/n$, onde x = nº de elementos da amostra que possuem certa característica de interesse.

Quando achamos uma estimativa pontual, ela raramente coincide com o valor real do parâmetro. Uma desvantagem do uso de estimadores pontuais é que, se nenhuma informação adicional for dada, não há maneira de decidir o quão boa é a estimativa, pois não temos nenhuma idéia da sua precisão. Um procedimento mais desejável para estimação é, então, calcular um intervalo que tenha uma probabilidade pré-estabelecida de conter o parâmetro desconhecido.

A *Estimação por intervalo* ou *Intervalos de Confiança* é um método de estimação onde, a partir de uma amostra aleatória, determinamos um intervalo $[T_1, T_2]$ que contém o verdadeiro parâmetro com uma probabilidade conhecida $(1 - \alpha)$, chamada de **Grau** ou **Nível de Confiança**, onde α (**alfa**) é a probabilidade do intervalo não conter o verdadeiro valor do parâmetro desconhecido. Assim, se amostras aleatórias, do mesmo tamanho, são obtidas repetidamente da mesma população, uma certa percentagem de intervalos (nível de confiança) incluirá o parâmetro populacional desconhecido. Além disso, veremos que a partir das estimativas intervalares é possível inferir sobre o quão confiáveis são realmente as estimativas pontuais obtidas.

3.2. Intervalos de Confiança para Média Populacional

Um intervalo de confiança para uma média especifica um intervalo de valores dentro do qual o parâmetro populacional desconhecido, neste caso a média, pode estar. Estes intervalos podem ser usados, por exemplo, por um fabricante que deseja estimar sua produção média diária ou um pesquisador que deseja estimar o tempo de resposta média, por paciente, a uma nova droga.

De modo geral, estamos interessados em encontrar um intervalo na forma

$$[T_1 = \overline{X} - \varepsilon_0 ; T_2 = \overline{X} + \varepsilon_0] = [\overline{X} \pm \varepsilon_0],$$

onde ε_0 representa a semi-amplitude do intervalo de confiança, sendo chamado de **Erro de Precisão** em relação a μ. Portanto, o objetivo é encontrar ε_0, tal que:

$P(|\ \overline{X} - \mu\ | < \varepsilon_0) = 1 - \alpha$, que é equivalente a, $P(\mu - \varepsilon_0 < \overline{X} < \mu + \varepsilon_0) = 1 - \alpha$.

Note que essa afirmação probabilística pode ser reescrita por:

$P(\overline{X} - \varepsilon_0 < \mu < \overline{X} + \varepsilon_0) = 1 - \alpha$.

Em breve, entenderemos a necessidade destas duas últimas afirmações probabilísticas.

3.2.1. Intervalos de Confiança para Média Populacional μ Caso 01: σ^2 é Conhecida

Suponha que temos uma amostra aleatória de tamanho *n*, X_1, X_2, ... , X_n, de uma população cuja distribuição é normal com média μ e variância σ^2. Então:

$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$, apresenta distribuição: $\overline{X} \approx N\left(\mu, \frac{\sigma^2}{n}\right)$ e $Z = \frac{\overline{X} - \mu}{\left(\sigma/\sqrt{n}\right)} \approx N(0,1)$.

Sejam (1 – α) um nível de confiança qualquer, 0 < (1 – α) < 1. Temos que,

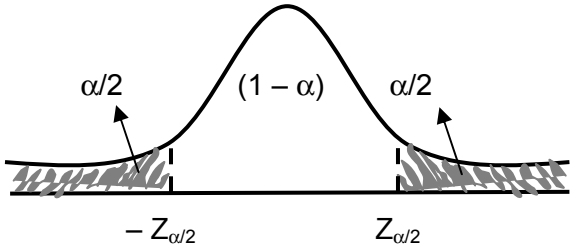
$$P(\mu - \varepsilon_0 < \overline{X} < \mu + \varepsilon_0) = 1 - \alpha$$

$$P\left(\frac{\mu - \varepsilon_0 - \mu}{\sigma/\sqrt{n}} < Z < \frac{\mu + \varepsilon_0 - \mu}{\sigma/\sqrt{n}}\right) = 1 - \alpha$$

$$P\left(\frac{-\varepsilon_0}{\sigma/\sqrt{n}} < Z < \frac{\varepsilon_0}{\sigma/\sqrt{n}}\right) = 1 - \alpha$$

$$P(-Z_{\alpha/2} < Z < +Z_{\alpha/2}) = 1 - \alpha$$

onde: $-Z_{\alpha/2} = \frac{-\varepsilon_0}{\sigma/\sqrt{n}}$ e $Z_{\alpha/2} = \frac{\varepsilon_0}{\sigma/\sqrt{n}}$.



Logo,

$$\varepsilon_0 = Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

A partir da expressão acima podemos também estimar, por exemplo, o **tamanho da amostra (*n*)** quando ε_0, *z* e σ são conhecidos:

$$n = \left(Z_{\alpha/2}\frac{\sigma}{\varepsilon_0}\right)^2$$

Como, $P(\mu - \varepsilon_0 < \overline{X} < \mu + \varepsilon_0) = 1 - \alpha \Rightarrow P(\overline{X} - \varepsilon_0 < \mu < \overline{X} + \varepsilon_0) = 1 - \alpha$, temos que

$$P\left(\overline{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha.$$

Em outras palavras, isso significa que a probabilidade de que o verdadeiro valor de μ pertença ao intervalo com grau de confiança igual a (1 – α) é:

$$IC[\mu;(1-\alpha)] = \left[\overline{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}; \overline{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

Exemplo 5.1: Para estimar gasto médio semanal no supermercado "A", coletou-se uma amostra aleatória de 16 consumidores, obtendo-se um gasto médio amostral de $\overline{X}$ = US$30,00. Supondo uma distribuição normal para a população, com desvio padrão σ = US$2,60, obtido de outros

estudos similares, calcule um intervalo de 95% de confiança para estimar o gasto médio semanal populacional no supermercado "A".

☺ **Solução:** Dados: σ = 2,6; *n* = 16 e $\overline{X}$ = 30. Para α = 5% ⇒ $z_{\alpha/2}$ = $P(Z \le z_{\alpha/2})$ = 1,96.

Logo, o intervalo de confiança será, então expresso por:

$$\text{IC } [\mu;\ 95\%] = \left[30 \pm 1{,}96\frac{2{,}6}{\sqrt{16}}\right] = [30 \pm 1{,}27] = [28{,}73;\ 31{,}27].$$

☑ **Observação**: No IC [μ; 95%] = [30 ± 1,27], o valor 1,27 é a estimativa do erro para a estimativa. Em outras palavras, há 95% de probabilidade da estimativa não diferir do verdadeiro valor da média (μ) por mais de 1,27.

☺ **Dica:** Abaixo, seguem os valores mais usados de $Z_{\alpha/2}$ tal que $P(Z \le Z_{\alpha/2}) = 1-(\alpha/2)$:

α	1%	5%	10%
$Z_{\alpha/2}$	2,57	1,96	1,64

3.2.2. Intervalos de Confiança para Média Populacional μ Caso 02: σ² NÃO é Conhecida

Quando a variância populacional é desconhecida, adota-se como estimador de σ^2 a **variância amostral (S^2)**, expressa por:

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}.$$

Agora, a estatística:

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{(n-1)},$$

terá distribuição **t-Student com "(n – 1)" graus de liberdade**, e não mais a distribuição normal padrão. No entanto, podemos re-escrever a estatística T como função da distribuição normal padrão (*Z*), da seguinte forma:

$$T = \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}} \times \frac{\sigma}{\sigma} = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \times \frac{\sigma}{S} = Z\frac{\sigma}{S}.$$

Logo,

$$t_{(n-1,\alpha/2)} = Z_{\alpha/2}\frac{\sigma}{S} \Rightarrow Z_{\alpha/2} = t_{(n-1,\alpha/2)}\frac{S}{\sigma}.$$

Substituindo $Z_{\alpha/2} = t_{(n-1,\alpha/2)}\frac{S}{\sigma}$ no intervalo de confiança do **Caso 01** teremos, quando a **variância populacional σ^2 é desconhecida**, o intervalo de confiança que contém o verdadeiro valor da média populacional μ com probabilidade (1 – α), expresso por:

$$IC[\mu;(1-\alpha)] = \left[\overline{X} - t_{(n-1,\alpha/2)}\frac{S}{\sqrt{n}};\ \overline{X} + t_{(n-1,\alpha/2)}\frac{S}{\sqrt{n}}\right]$$

Logo,

$$\varepsilon_0 = t_{(n-1,\alpha/2)} \frac{S}{\sqrt{n}}.$$

A partir da expressão acima podemos também estimar, por exemplo, o **tamanho da amostra (*n*)** quando ε_0, *Z* e S são conhecidos.

$$n = \left(t_{(n-1,\alpha/2)} \frac{S}{\varepsilon_0} \right)^2$$

Exemplo 5.2: Um fiscal de produtos alimentícios seleciona uma amostra aleatória de 16 pacotes de lanche marca "M" nas prateleiras de um supermercado. Pesa o conteúdo de cada pacote, encontrando um peso médio $\overline{X}$ = 170 g e um desvio padrão S = 5 g. O peso líquido indicado em cada pacote é 180 g. Verifique se um intervalo com 90% de confiança para o peso médio líquido verdadeiro abrange o peso líquido especificado na embalagem. Suponha distribuição normal para a população.

☺ **Solução:**

Dados: *n* = 16, $\overline{X}$ = 170 g e S = 5 g.
Para α = 10% e *n* = 16 $\Rightarrow$ $t_{(n-1;\,\alpha/2)} = t_{(15;\,0,05)} = 1,753$, obtido da tabela da distribuição t-Student, pois a informação que dispomos no problema diz respeito ao desvio padrão amostral.

Logo, o intervalo de confiança para o peso médio populacional será denotado por:

IC [μ; 90%] = $\left[170 \pm 1,753 \frac{5}{\sqrt{16}} \right]$ = [170 ± 2,19] = [167,81; 172,19].

☹ Note que o IC não abrange o peso líquido indicado na embalagem de 180g.

Exemplo 5.3: Em uma amostra de *n* = 9 testes de consumo, um motor experimental percorreu, respectivamente, 16, 14, 17, 15, 15, 14, 18, 17 e 18 km com l litro de gasolina (sob condições específicas). Supondo distribuição normal para a população, construa um intervalo de 99% de confiança para a distância média verdadeira do novo motor, com 1 litro de gasolina.

☺ **Solução**:

Seja X = quilômetros percorridos com 1 litro de gasolina.
Dados: *n* = 9, $\overline{X}$ = 16 km/l e S = 1,581 km/l.
Para α = 1% e *n* = 9 $\Rightarrow$ $t_{(n-1;\,\alpha/2)} = t_{(8;\,0,005)} = 3,355$, obtido da tabela da distribuição t-Student.

Logo, o intervalo de confiança será denotado por:

IC [μ; 99%] = $\left[16 \pm 3,355 \frac{1,581}{\sqrt{9}} \right]$ = [16 ± 1,77] ou [14,23; 17,77] km/l de gasolina.

Assim, podemos afirmar que com 99% de confiança, o intervalo [14,23 km/l; 17,77km/l] contém o verdadeiro valor para a distância percorrida pelo novo motor (em quilômetros) com um litro de gasolina.

Exemplo 5.4: Se um pesquisador sabe que uma população tem distribuição normal com desvio padrão σ=12. Considerando um nível de confiança de 95%, encontre o tamanho de amostra necessário para que a média amostral não se afaste em mais de 2 unidades do verdadeiro valor da média populacional.

☺ **Solução:**
Em nosso problema, observamos que o desvio padrão populacional é conhecido. Neste caso, usamos a expressão a seguir para o cálculo do tamanho de amostra:

Dados: $\sigma = 12$, $\varepsilon_0 = 2$ e que $\alpha = 5\% \Rightarrow Z_{\alpha/2} = 1{,}96$. Dessa forma,

$$n = \left(z_{\alpha/2} \frac{\sigma}{\varepsilon_0} \right)^2 = \left(1{,}96 \frac{12}{2} \right)^2 \cong 139.$$

3.3. Intervalos de Confiança para uma Proporção Populacional *p*

Estes intervalos podem ser usados para, por exemplo, informar sobre a proporção de alunos evadidos na UFPB, a proporção de itens defeituosos em uma linha de produção ou a proporção de uma população que é imune a certa enfermidade.

Seja X uma variável aleatória, representando o número de sucessos em *n* repetições independentes de um experimento, com dois possíveis resultados (sucesso e fracasso). Considere P(Sucesso) = *p* e a P(Fracasso) = *q* = (1 – *p*), constantes. Dessa forma, X ~ Bin(*n*, *p*), onde μ_X = E(X) = *np* e σ^2 = Var(X) = *np*(1 – *p*). Para *p* não muito próximo de 0 ou 1 e se *n* é suficientemente grande (*um critério usado na prática, é usar a aproximação quando np e np(1 – p) forem maiores do que 5*) temos, segundo o Teorema Central do Limite, que:

$$X \sim N\big(np, np(1-p)\big).$$

Logo, $\hat{p} = \frac{x}{n} \sim N\left(p, \frac{p(1-p)}{n} \right)$, visto que:

$$E(\hat{p}) = E\left(\frac{X}{n}\right) = \frac{1}{n}E(X) = \frac{1}{n}np = p \text{ e } Var(\hat{p}) = Var\left(\frac{X}{n}\right) = \frac{1}{n^2}Var(X) = \frac{1}{n^2}np(1-p) = \frac{p(1-p)}{n}.$$

Assim, $Z = \dfrac{\hat{p} - p}{\sqrt{\dfrac{p(1-p)}{n}}} \approx N(0,1)$.

O intervalo que estamos procurando, da forma $[\hat{p} \pm \varepsilon_0]$, será obtido por um caminho semelhante ao adotado no caso da média populacional μ chegando-se, facilmente, a,

$$\varepsilon_0 = z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}.$$

Note que a partir da expressão acima também podemos o **tamanho da amostra (*n*)** quando ε_0, Z e p são conhecidos.

No entanto, na prática p é desconhecido, sendo substituído pela proporção amostral $\hat{p}$. Tal substituição encontra justificativa no fato de que se n é suficientemente grande para garantir a aproximação para Normal, a estimativa deve ser razoavelmente próxima do valor real do parâmetro. Assim, o intervalo de confiança para p, ao nível de confiança $(1 - \alpha)$, é dado por:

$$IC[p;(1-\alpha)] = \left[\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}; \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right]$$

Exemplo 5.5: Para se avaliar a taxa de desemprego em uma cidade, coletou-se uma amostra aleatória de 1000 habitantes em idade de trabalho e observou-se que 87 eram desempregados. Estimar a percentagem de desempregados em toda a cidade (população) através de um intervalo de 95% de confiança.

☺ **Solução**:

Dados: n = 1000. A proporção amostral de desempregados é dada por:
$\hat{p} = \frac{87}{1000} = 0{,}087$. Logo, $\hat{q}$ = (1 – $\hat{p}$) = 0,913. Para α = 5% $\Rightarrow Z_{\alpha/2}$ = 1,96.

O intervalo de confiança será,

IC [p; 95%] = $\left[0{,}087 \pm 1{,}96\sqrt{\frac{(0{,}087)(0{,}913)}{1000}} \right]$ = [0,087 ± 0,0175] = [0,0695; 0,1045] ou, ainda, [6,95%; 10,45%].

Exemplo 5.6: Numa pesquisa de mercado, 57 das 150 pessoas entrevistadas preliminarmente afirmaram que seriam compradoras de certo produto a ser lançado. Essa amostra é suficiente para estimar a proporção real de futuros compradores, com um erro de 4% e confiança de 95%?

☺ **Solução**:

Dados: $\hat{p} = \frac{57}{150} = 0{,}38$ e $\hat{q}$ = (1 – $\hat{p}$) = 0,62; ε_0 = 0,04; $Z_{\alpha/2}$ = 1,96.

Logo, de modo a similar a média, o tamanho da amostra para proporção é calculado a partir da seguinte expressão:

$$n = \left(\frac{z_{\alpha/2}}{\varepsilon_0} \right)^2 \hat{p}(1-\hat{p}) \Rightarrow n = \left(\frac{1{,}96}{0{,}04} \right)^2 (0{,}38)(0{,}62) \cong 566 .$$

Como apenas 150 pessoas foram entrevistadas preliminarmente, a amostra não foi suficiente. Sendo necessário entrevistar mais (566 – 150) = 416 pessoas.

:: SAIBA MAIS... ::

NEM TUDO O QUE LHE DIZEM É ABSOLUTAMENTE CERTO
O problema é como saber quando o que lhe estão dizendo não é certo. **Testes de hipóteses** são ferramentas que usam amostras para testar se é provável ou não que afirmações estatísticas sejam verdadeiras. Eles são uma forma de **ponderar as evidências** e testar se resultados extremos podem ser explicados por **mera coincidência** ou se existem forças ocultas em ação. Embarque nesse passeio no próximo tópico e vamos mostrar como usar testes de hipótese para confirmar ou descartar suas mais profundas suspeitas!

3.3. Testes de Hipóteses

Na estimação de parâmetros, foram apresentados procedimentos que permitem definir estimadores pontuais ou por intervalos de parâmetros populacionais. Outro procedimento de inferência estatística – o Teste de Hipótese – tem como objetivo principal verificar, a partir de informações contidas em uma amostra aleatória, se hipóteses a respeito de parâmetros populacionais são ou não verdadeiras. Assim podemos estar interessados em: verificar uma especificação de qualidade de um produto, testar uma experiência de sucesso no passado, avaliar uma teoria ou decidir sobre suposições resultantes das observações. Logo, através dos testes de hipóteses podem-se eliminar, tanto quanto possível, falsas conclusões científicas.

3.3.1. Conceitos Fundamentais

O Teste de Hipótese se baseia numa situação experimental (amostra) e consiste na comparação de duas hipóteses chamadas *Hipótese Nula* e *Hipótese Alternativa*.

☑ *Hipótese Nula* (H_0) – É uma afirmação sobre o parâmetro, supostamente verdadeira, que vai ser posta à prova e na qual o teste é montado. Em geral, formula-se H_0 com o objetivo de rejeitá-la, isto é, formulamos H_0 contrária ao que suspeitamos que seja verdade. Por exemplo, se um cientista acha que uma nova droga é eficaz para certo tipo de paciente, então, por contradição, formulamos a hipótese H_0 de que a nova droga não é eficaz. Portanto, para provar que o cientista está certo, H_0 teria de ser rejeitada. Dessa forma, podemos pensar que o que estamos interessados deve ser alocado em H_1 (*Hipótese Alternativa*). Uma possível representação é: H_0: $\theta = \theta_0$, onde θ é qualquer parâmetro.

☑ *Hipótese Alternativa* (H_1) – Hipótese que vai ser comparada à hipótese nula, isto é, uma afirmação sobre o parâmetro que afirma "A hipótese nula H_0 é falsa".

Se usamos H_0: $\theta = \theta_0$ para representar a hipótese nula, então podemos usar as seguintes representações para as possíveis hipóteses alternativas:

- H_1: $\theta \neq \theta_0$.
- H_1: $\theta < \theta_0$.
- H_1: $\theta > \theta_0$.

3.3.2. Definição da Regra de Decisão, Erros e Nível de Significância

Quando testamos hipóteses estatísticas, qualquer que seja a decisão tomada, estamos sujeitos a cometer dois possíveis tipos de erros:

- **Erro do Tipo I**: Quando se rejeita a hipótese nula H_0 e a mesma é verdadeira. Denotamos por α a probabilidade de cometer este erro, isto é,

  $$\alpha = P(\text{Erro Tipo I}) = P(\text{Rejeitar } H_0 \mid H_0 \text{ é verdadeira}).$$

  O erro tipo I (α) também é conhecido como **nível de significância** de um teste de hipóteses.
- **Erro do Tipo II**: Não se rejeita a hipótese nula H_0, quando a mesma é falsa. Denotamos por β a probabilidade de cometer este erro, isto é,

  $$\beta = P(\text{Erro Tipo II}) = P(\text{Não rejeitar } H_0 \mid H_0 \text{ é falsa}).$$

O quadro abaixo resume as possibilidades das decisões envolvidas em um teste de hipótese, com as probabilidades de ocorrências dos erros tipo I (α) e II (β).

Quadro 1: Avaliação das Decisões em um Teste de Hipóteses

Decisão	Situação Real	
	H_0 é Verdadeira	H_0 é Falsa
Não Rejeitar H_0	Decisão Correta	Erro do Tipo II (β)
Rejeitar H_0	Erro do Tipo I (α)	Decisão Correta

Devido as dificuldades de se conseguir minimizar os dois tipos de erros ao mesmo tempo, em geral, nos preocupamos mais na possibilidade de rejeitar uma hipótese sendo ela verdadeira. Dessa forma, teremos uma maior atenção no controle do erro do tipo I. Por exemplo, se definimos as hipóteses:

- H_0: Uma nova droga não é eficaz para certos pacientes;
- H_1: Uma nova droga é eficaz para certos pacientes.

A aceitação de H_0, sendo esta hipótese falsa, possibilita a busca de outros meios de tratamentos, enquanto que a rejeição de H_0, sendo esta verdadeira, exclui a possibilidade de se prosseguir com outras opções para os pacientes. Logo, é desejável exercer um controle sobre α e mantê-lo pequeno. Dessa forma, os testes de hipóteses podem ser montados de maneira que, fixado o erro do tipo I, o erro do tipo II seja minimizado aumentando-se o tamanho da amostra.

☑ **Observação**: O significado de α usado nos Testes de Hipóteses é totalmente diferente de seu significado na Estimação por Intervalos. Nos Testes de Hipóteses, α representa a probabilidade de rejeitar uma hipótese nula suposta verdadeira, enquanto que na Estimação por Intervalos α representa a probabilidade de que os limites de confiança construídos não contenham o verdadeiro valor do parâmetro.

Estatística do Teste

A decisão de rejeitar ou não a hipótese nula (H_0) é baseada nos dados amostrais, que são usados para calcular o valor da **Estatística de Teste** e que servirá de referência para a tomada da decisão. Para isso, divide-se a curva da distribuição amostral da estatística em duas regiões, uma chamada **Região Crítica** (ou **Região de Rejeição de H_0**), e a outra **Região de Não Rejeição de H_0**. Temos, então, a seguinte **Regra de Decisão** do teste:

Se o valor calculado da estatística do teste pertencer à região crítica, rejeita-se H_0 em favor da hipótese alternativa; caso contrário, H_0 não será rejeitada em relação à hipótese alternativa.

Outras definições importantes, necessárias na formulação de um problema de Testes de Hipóteses são:

- ***Região Crítica do Teste*:** É a região de rejeição de H_0, isto é, o conjunto de valores de uma estatística que determina a rejeição de H_0. Rejeitamos a hipótese nula se a estatística de teste está na região crítica, porque isto indica uma discrepância significativa entre a hipótese nula e os dados amostrais.
- ***Valor Crítico do Teste*:** É o valor, ou valores, que separa(m) a região crítica (que levam a estatística do teste a rejeitar a hipótese nula) da região de não rejeição de H_0.

Dependendo da hipótese alternativa, temos os seguintes **Tipos de Testes de Hipóteses**:

- ***Teste Unilateral:*** Quando a região crítica do teste é localizada completamente em uma das extremidades da curva da distribuição amostral da estatística do teste.
    - ***Teste Unilateral à Esquerda:*** A região crítica (sombreada) localiza-se no extremo esquerdo da distribuição.

      Hipóteses: H_0: $\theta = \theta_0$ versus H_1: $\theta < \theta_0$
    - ***Teste Unilateral à Direita:*** A região crítica (sombreada) localiza-se no extremo direito da distribuição.

      Hipóteses: H_0: $\theta = \theta_0$ versus H_1: $\theta > \theta_0$
- ***Teste Bilateral*:** A região crítica (sombreada) localiza-se nas duas extremidades da distribuição.

  Hipóteses: H_0: $\theta = \theta_0$ versus H_1: $\theta \neq \theta_0$

A escolha entre usar um teste unilateral e um teste bilateral é determinada pelos objetivos do problema, no qual se deseja verificar uma afirmação a cerca do parâmetro populacional.

3.3.3. Fases na Realização de um Teste de Hipóteses

❶ Formular as hipóteses nula (H_0) e alternativa (H_1);
Esta é a afirmação que vamos testar.

❷ Decidir qual estatística de teste será usada para julgar a hipótese nula;

Precisamos escolher a melhor estatística que sirva para testar a afirmação.

❸ Fixar o nível de significância α;
Precisamos fixar um nível de significância que seja pequeno.

❹ Determinar a região crítica;
Precisamos de um determinado nível de certeza.

❺ Usar os valores amostrais para calcular o valor da estatística citada na fase 2;
Precisamos saber a raridade dos resultados, considerando que as afirmações sejam verdadeiras.

❻ Se o valor citado na fase anterior pertencer à região crítica, rejeitar H_0. Caso contrário, não rejeitar H_0.
Em seguida, vemos se ele está dentro dos limites de certeza, de modo a tomar a decisão.

3.3.4. Teste de Hipóteses para a Média Populacional μ Caso 01: σ^2 é Conhecida

O primeiro passo num Teste de Hipóteses consiste em formular a hipótese a ser testada. No Quadro 1, podemos observar que para cada possível hipótese existe uma região crítica e regra de decisão associada. No caso do teste de hipóteses para média populacional, supondo a variância populacional conhecida, utilizamos a seguinte estatística do teste:

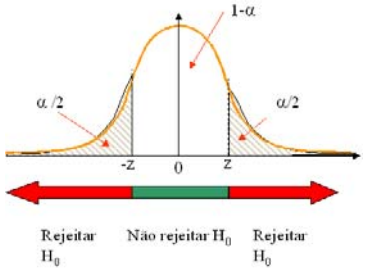
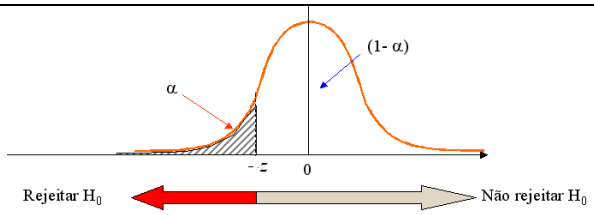
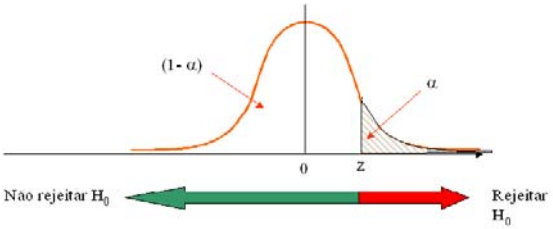
$$Z_C = \frac{\overline{X} - \mu_0}{\sigma / \sqrt{n}}$$

☺ Note que a estatística é calculada com base nas informações contidas na amostra.

O próximo passo consiste em fixar o nível de significância do teste (α). A seguir, apresentamos os valores mais usados para Z_α e $Z_{\alpha/2}$.

α	1%	5%	10%
Z_α	2,33	1,64	1,28
$Z_{\alpha/2}$	2,57	1,96	1,64

Quadro 2: Resumo das Hipóteses, Regiões Críticas e Regras de Decisão para a Média Populacional, considerando σ^2 conhecido.

Hipóteses	Região Crítica (sombreada)	Regra de Decisão (Rejeitar H_0)
H_0: $\mu = \mu_0$ H_1: $\mu \neq \mu_0$		$Z_c \leq -Z_{\alpha/2}$ ou $Z_c \geq Z_{\alpha/2}$
H_0: $\mu = \mu_0$ (*) H_1: $\mu < \mu_0$		$Z_c \leq -Z_{\alpha}$
H_0: $\mu = \mu_0$ (**) H_1: $\mu > \mu_0$		$Z_c \geq Z_{\alpha}$

(*) Por simplicidade, excluiu-se a possibilidade $\mu \geq \mu_0$ na hipótese nula H_0, com base no conhecimento de que tal fato levaria à mesma decisão que a aceitação simples de H_0: $\mu = \mu_0$.
(**) Por simplicidade, excluiu-se a possibilidade $\mu \leq \mu_0$ na hipótese nula H_0, com base no conhecimento de que tal fato levaria à mesma decisão que a aceitação simples de H_0: $\mu = \mu_0$.

Exemplo 5.7: O gerente de uma indústria de carnes enlatadas tem estabelecido a seguinte especificação: um novilho com 12 meses de vida resulta numa média de 250 kg de carne. A experiência passada indica que, mesmo com uma mudança na média, o desvio padrão permanece ligeiramente constante, em σ = 18 kg. Para determinar se a especificação está sendo observada, o gerente seleciona uma amostra aleatória com 100 novilhos e obteve uma média $\overline{X}$ = 253 kg de carne. Realize um teste de hipótese para verificar se houve mudança na especificação, a um nível de significância de 5%.

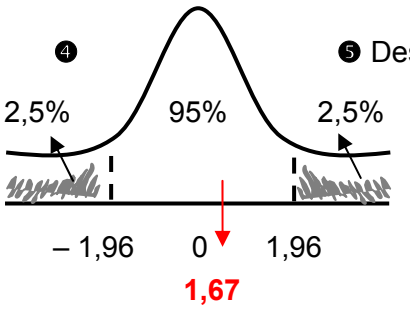
☺ **Solução:**

❶ H_0: $\mu = 250$ kg *versus* H_1: $\mu \neq 250$ kg (a especificação não está sendo observada)

❷ Temos que $\sigma = 18$ kg; $n = 100$, $\overline{X} = 253$ kg. Como σ é conhecido vamos fazer uso da variável Normal, Z.

❸ $\alpha = 5\%$.

❹



❺ Dessa forma, a estatística do teste é dada:

$$Z_c = \frac{253 - 250}{18/\sqrt{100}} = \mathbf{1,67}$$

Atenção: Como o teste é bilateral, o valor crítico ao nível $\alpha = 5\%$ será $Z_{\alpha/2} = 1,96$.

❻ Decisão: Como $-Z_{\alpha/2} < Z_c < Z_{\alpha/2} \Rightarrow$ Não existem evidências para rejeitar H_0. Logo, com base nos dados amostrais e com 5% de significância não podemos rejeitar a hipótese H_0, ou seja, não existe evidência para afirmar que a especificação está sendo violada.

3.3.5. Teste de Hipóteses para a Média Populacional μ Caso 02: σ^2 NÃO é Conhecida

Quando a variância populacional (σ^2) é desconhecida, precisamos estimá-la a partir das informações contidas na amostra, através da expressão

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}.$$

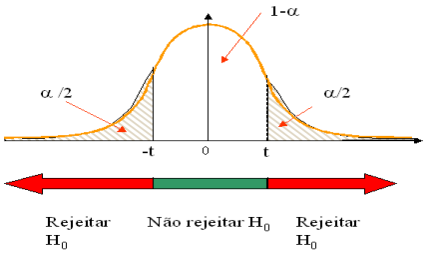
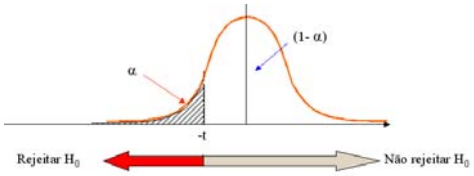
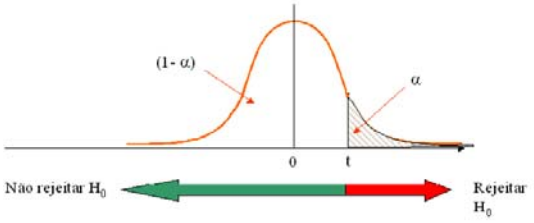
Dessa forma, a estatística do teste para média populacional μ quando σ^2 é desconhecida será expressa por:

$$T_C = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$$

, que segue uma distribuição **t-Student** com ($n - 1$) graus de liberdade.

O próximo passo consiste em fixar o nível de significância do teste (α). A seguir, apresentamos as regiões críticas e regras de decisão para as respectivas hipóteses.

Quadro 3: Resumo das Hipóteses, Regiões Críticas e Regras de Decisão para a Média Populacional, considerando σ^2 desconhecido.

Hipóteses	Região Crítica (sombreada)	Regra de Decisão (Rejeitar H_0)
H_0: $\mu = \mu_0$ H_1: $\mu \neq \mu_0$		$T_c \leq -t_{(n-1,\alpha/2)}$ ou $T_c \geq t_{(n-1,\alpha/2)}$
H_0: $\mu = \mu_0$ (*) H_1: $\mu < \mu_0$		$T_c \leq -t_{(n-1,\alpha)}$
H_0: $\mu = \mu_0$ (**) H_1: $\mu > \mu_0$		$T_c \geq t_{(n-1,\alpha)}$

(*) Por simplicidade, excluiu-se a possibilidade $\mu \geq \mu_0$ na hipótese nula H_0, com base no conhecimento de que tal fato levaria à mesma decisão que a aceitação simples de H_0: $\mu = \mu_0$.
(**) Por simplicidade, excluiu-se a possibilidade $\mu \leq \mu_0$ na hipótese nula H_0, com base no conhecimento de que tal fato levaria à mesma decisão que a aceitação simples de H_0: $\mu = \mu_0$.

Exemplo 5.8: O tempo médio necessário para completar uma tarefa era de 15 minutos. Obtém-se uma amostra aleatória de nove indivíduos e, durante o período de teste, seus tempos (X) para concluir a tarefa foram 11, 12, 15, 10, 12, 14, 15, 13 e 15. Assumindo que estes dados vêm de uma distribuição normal, teste a hipótese de que houve alteração no tempo médio para completar a tarefa. Use um nível de 5% de significância.

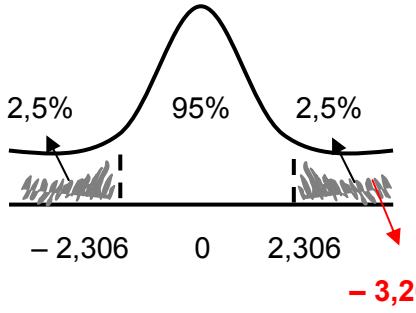
☺ **Solução:**

❶ H_0: μ = 15min *versus* H_1: $\mu \neq$ 15min (houve alteração no tempo médio)

❷ Com base nas informações amostrais, temos que n = 9; $\overline{X}$ = 13min e S = 1,871 min. Como o desvio-padrão da população é desconhecido, σ, a variável usada para a estatística de teste é a t-Student com (n – 1) graus de liberdade.

❸ α = 5%. Sendo n = 9, será $t_{(n-1;\ \alpha/2)} = t_{(8;\ 0{,}025)}$ = 2,306 (obtido da tabela da distribuição t-Student).

❹



❺ Dessa forma, a estatística do teste é dada:

$$T_c = \frac{13-15}{1{,}871/\sqrt{9}} = \mathbf{-3{,}207}$$

❻ Decisão: Como $T_c < -t_{n-1;\alpha/2}$, existem evidências para rejeitar H_0. Logo, com base nos dados amostrais e com 5% de significância, rejeita-se a hipótese H_0, ou seja, existem evidências para afirmar que os indivíduos apresentaram um tempo médio para executar a tarefa diferente do que era observado anteriormente.

3.3.6. Teste de Hipóteses para a uma Proporção Populacional *p*

Ao se fazer inferências sobre uma proporção populacional, *p*, tomamos nossas com base nas evidências sobre seu valor amostral, $\hat{p}$, de elementos com a característica de interesse.

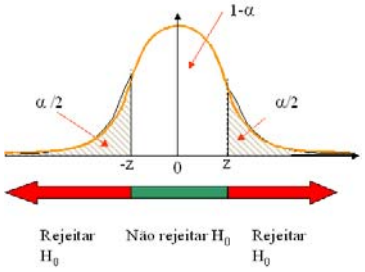
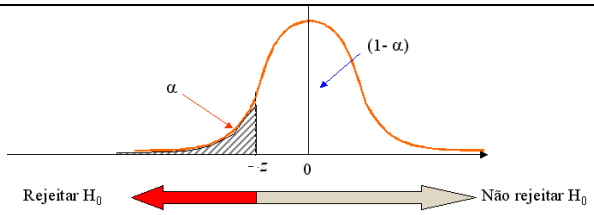
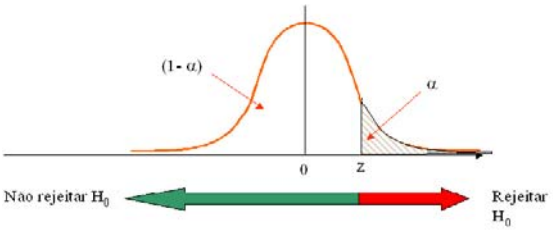
Pelo Teorema Central do Limite, sabe-se que, para *n* suficientemente grande, a proporção amostral, $\hat{p} = \frac{x}{n}$ segue, aproximadamente, uma distribuição $\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$.

Dessa forma, sob a hipótese H_0: p = p_0, a estatística do teste para a proporção populacional *p* será expressa por:

$$Z_C = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

Z_C ~ N(0; 1). Após fixar o nível de significância do teste (α), apresentamos a seguir as regiões críticas e regras de decisão para as respectivas hipóteses.

Quadro 4: Resumo das Hipóteses, Regiões Críticas e Regras de Decisão para a Proporção Populacional *p*.

Hipóteses	Região Crítica (sombreada)	Regra de Decisão (Rejeitar H_0)
H_0: $p = p_0$ H_1: $p \neq p_0$		$Z_c \leq -Z_{\alpha/2}$ ou $Z_c \geq Z_{\alpha/2}$
H_0: $p = p_0$ (*) H_1: $p < p_0$		$Z_c \leq -Z_{\alpha}$
H_0: $p = p_0$ (**) H_1: $p > p_0$		$Z_c \geq Z_{\alpha}$

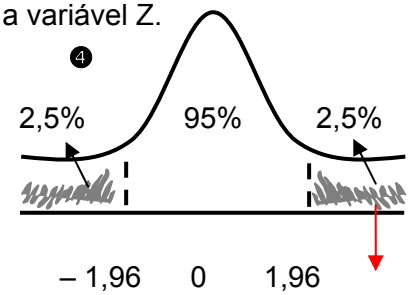
(*) Por simplicidade, excluiu-se a possibilidade $p \geq p_0$ na hipótese nula H_0, com base no conhecimento de que tal fato levaria à mesma decisão que a aceitação simples de H_0: $p = p_0$.
(**) Por simplicidade, excluiu-se a possibilidade $p \leq p_0$ na hipótese nula H_0, com base no conhecimento de que tal fato levaria à mesma decisão que a aceitação simples de H_0: $p = p_0$.

Exemplo 5.9: Afirma-se que em um alqueire de maçãs, 10% estão estragadas. De uma amostra aleatória de 150 maçãs examinadas, 30 estavam estragadas. O que você conclui sobre a proporção de maçãs estragadas em um alqueire a um nível de 5% de significância?

☺ **Solução:**

❶ H_0: p = 0,10 *versus* H_1: p ≠ 0,10.
❷ Com base nas informações amostrais, temos que *n* = 150 e $\hat{p}$ = 30/150 = 0,20. No caso da proporção use a variável Z.
❸ α = 5%.

❹



❺ Dessa forma, a estatística do teste é:

$$Z_c = \frac{0,2 - 0,1}{\sqrt{\frac{(0,1).(0,9)}{150}}} = \mathbf{4,08}$$

Atenção: Como o teste é bilateral o valor crítico ao nível α = 5% será $Z_{\alpha/2}$ = 1,96.

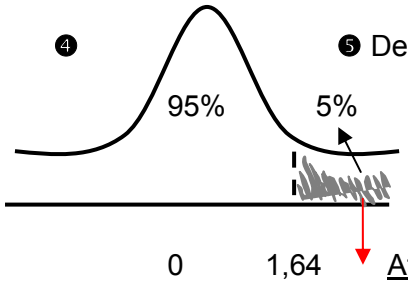
❻ Decisão: Como $Z_C > Z_{\alpha/2} \Rightarrow$ Existem evidências para rejeitar H_0. Logo, com base nos dados amostrais e ao nível de 5% de significância, podemos concluir que a porcentagem de maçãs estragadas é diferente de 10%.

Exemplo 5.10: De registros de vendas passadas sabe-se que 30% dos consumidores compram a pasta dental C. Uma nova propaganda desse produto é feita e, para testar sua eficácia, de uma amostra aleatória de 1000 consumidores que viram a propaganda, 334 responderam que compram a pasta dental C. Isso indica que a nova propaganda foi bem sucedida? Use um nível de 5% de significância para testar se a nova propaganda aumentou a proporção de consumidores da pasta dental C.

☺ **Solução:**

❶ H_0: p = 0,30 *versus* H_1: p > 0,30 (a nova propaganda aumentou as vendas da pasta C).
❷ Com base nas informações amostrais, temos que *n* = 1000 e $\hat{p}$ = 334/1000 = 0,334. No caso da proporção use a variável Z.
❸ α = 5%.

❹



❺ Dessa forma, a estatística do teste é dada:

$$Z_c = \frac{0,334 - 0,300}{\sqrt{\frac{(0,3).(0,7)}{1000}}} = \mathbf{2,35}$$

Atenção: Como o teste é unilateral, o valor crítico ao nível α = 5% será Z_α = 1, 64.

❻ Decisão: Como $Z_C > Z_\alpha \Rightarrow$ Existem evidências para rejeitar H_0. Logo, com base nos dados amostrais e ao nível de 5% de significância, podemos concluir a nova propaganda aumentou a proporção de consumidores que compram a pasta dental C.

4. Avaliando o que foi construído

Ao final desta unidade aprendemos duas importantes técnicas inferenciais: intervalos de confiança e testes de hipóteses. Ambas podem ser aplicadas no processo de tomada de decisão em inúmeros problemas práticos. Pratique tais conceitos resolvendo os exercícios propostos no MOODLE. Finalizamos essa viagem pelos Métodos Estatísticos.

REFERÊNCIAS

COSTA NETO, P.L., **Estatística**, Edgard Blucher, São Paulo, 1977.

FONSECA, J.S., MARTINS, G.A. & TOLEDO, G.L., **Estatística Aplicada**, Editora Atlas, 2ª ed., São Paulo, 1985.

GRIFFITHS, D. **Use a Cabeça. Estatística**, Editora Alta Books, Rio de Janeiro, 2009.

MEYER, P.L., **Probabilidade: Aplicações à Estatística**, Livros Técnicos e Científicos, Editora AS, Rio de Janeiro, 1983.

TRIOLA, M.F, **Introdução à Estatística**, Livros Técnicos e Científicos, 7ª ed,, Rio de Janeiro, 1999.

VIEIRA, S. **Introdução à Bioestatística**, Editora Campus, 1999.

Homenagem ao Pólo de Apoio Presencial de Pombal, Paraíba.