UNIVERSIDADE FEDERAL DE PELOTAS
INSTITUTO DE FÍSICA E MATEMÁTICA
DEPARTAMENTO DE MATEMÁTICA E ESTATÍSTICA

Estatística Básica

Versão Preliminar

Clause Fátima de Brum Piana
Amauri de Almeida Machado
Lisiane Priscila Roldão Selau

Pelotas, 2009.

Sumário

Unidade I. Introdução

Unidade II. Estatística Descritiva

Unidade III. Elementos de Probabilidade

Unidade IV. Inferência Estatística

Apêndice

Unidade I
Introdução

1.1. Considerações gerais

A coleta, o processamento, a interpretação e a apresentação de dados numéricos pertencem todos aos domínios da estatística. Essas atribuições compreendem desde o cálculo de pontos em esportes, a coleta de dados sobre nascimentos e mortes, a avaliação da eficiência de produtos comerciais, até a previsão do tempo. A informação estatística é apresentada constantemente em todos os meios de comunicação de massa: jornais, televisão, rádio e internet.

Observamos uma abordagem crescentemente quantitativa utilizada em todas as ciências, na administração e em muitas atividades que afetam diretamente nossas vidas. Isto inclui o uso de técnicas matemáticas nas decisões econômicas, públicas ou privadas; na avaliação de controles de poluição; na análise de problemas de tráfego; no estudo dos efeitos de vários medicamentos; na adoção de novas técnicas agrícolas e novas cultivares; em estudos demográficos como crescimento populacional e migração.

A partir destes poucos exemplos, podemos notar a importância da Estatística como ferramenta necessária para a compreensão dos fenômenos que ocorrem nas mais diferentes áreas.

1.2. População e amostra

É difícil encontrar duas coisas exatamente iguais. Há um pouco de variabilidade em quase tudo. De modo bem geral, podemos dizer que o objetivo da Estatística é fornecer métodos para se conviver, de modo racional, com a variabilidade. Isto é feito através da descoberta de regularidade nos dados relativos às situações em estudo. Para isso, duas ideias são de fundamental importância. Primeiramente, embora as observações sejam variáveis é sempre possível associar a elas a ideia de regularidade e expressar essa regularidade matematicamente. Por outro lado, devido à variabilidade inerente aos indivíduos, os pontos de interesse da Estatística são referentes aos grupos de indivíduos, ou seja, estudamos os indivíduos através dos grupos.

Quando estudamos uma determinada característica, geralmente, queremos obter conclusões para o conjunto de todos os indivíduos que apresentam tal característica. Chamamos de *população* o conjunto de todos os indivíduos ou objetos que apresentam uma característica em comum. Na maioria dos casos, ao estudarmos uma população, não temos acesso a todos os seus elementos. O estudo é feito, então, a partir de uma parte desta população, denominada *amostra*, que tem por objetivo representá-la.

1.3. Conceito e divisão

A Estatística, durante muitos séculos, esteve relacionada apenas com as informações a respeito do Estado. Hoje em dia, o conjunto de teorias, conceitos e métodos denominado Estatística está associado ao processo de descrição e inferência, debruçando-se, de modo particular, sobre questões relativas a sumarização eficiente de dados, planejamento e análise de experimentos e levantamentos e natureza de erros de medida e de outras causas de variação em um conjunto de dados.

A estatística pode ser dividida em duas partes principais: a *Estatística Descritiva* e a *Inferência Estatística* ou *Estatística Analítica*.

Enquanto a Estatística Descritiva cuida do resumo e da apresentação de dados de observação por meio de tabelas, gráficos e medidas, sem se preocupar com as populações de onde esses dados foram retirados, a Inferência Estatística tem como objetivo fornecer métodos que possibilitem a realização de inferência sobre populações a partir de amostras delas provenientes. A Inferência Estatística tem por base o cálculo de probabilidades e compreende dois grandes tópicos: a estimação de parâmetros e os testes de hipóteses.

Embora a Estatística Descritiva seja um ramo fundamental da Estatística, em muitos casos ela se torna insuficiente. Isto ocorre porque quase sempre as informações são obtidas de amostras e, consequentemente, sua análise exige generalizações que ultrapassam os

dados disponíveis. Essa necessidade, aliada ao desenvolvimento dos métodos probabilísticos, promoveu o crescimento da Estatística pela ênfase aos métodos generalizadores (Inferência Estatística), em acréscimo aos métodos puramente descritivos.

Alguns exemplos ilustram a necessidade dos métodos generalizadores:

– prever a duração média da vida útil de uma calculadora, com base no desempenho de muitas dessas calculadoras;

– comparar a eficiência de duas dietas para reduzir peso, com base nas perdas de peso de pessoas que se submeteram às dietas;

– determinar a dosagem ideal de um novo medicamento, com base em testes feitos em pacientes voluntários de hospitais selecionados aleatoriamente;

– prever o fluxo de tráfego de uma rodovia ainda em construção, com base no tráfego observado em rodovias alternativas.

Em todas essas situações existe incerteza porque dispomos apenas de informações parciais, incompletas ou indiretas. A Inferência Estatística trata de problemas onde a incerteza é inerente, utilizando métodos que se fundamentam na teoria das probabilidades. Os métodos de inferência tornam-se necessários para avaliar a confiabilidade dos resultados observados.

1.4. Informações históricas

Embora a palavra estatística ainda não existisse, existem indícios de que há 3000 anos a.C. já se faziam censos na Babilônia, China e Egito.

A própria Bíblia leva-nos a esse resgate histórico:

– o livro quarto do Velho Testamento, intitulado “Números”, começa com a seguinte instrução a Moisés: ”Fazer um levantamento dos homens de Israel que estivessem aptos para guerrear”;

– na época do Imperador César Augusto, saiu um edito para que se fizesse o censo em todo o Império Romano. Por isso Maria e José teriam viajado para Belém.

A Estatística teve origem na necessidade do Estado Político em conhecer os seus domínios. Sob a palavra estatística, provavelmente derivada da palavra “status” (estado, em latim), acumularam-se descrições e dados relativos ao Estado. Nas mãos dos governantes, a Estatística passou a constituir-se verdadeira ferramenta administrativa.

Em 1085, Guilherme, o Conquistador, ordenou que se fizesse um levantamento estatístico da Inglaterra, que deveria incluir informações sobre terras, proprietários, uso da terra, empregados, animais e que serviria também de base para o cálculo de impostos. Esse levantamento originou um volume intitulado “Domesday Book” (Livro do dia do juízo final).

Jonh Graunt
(1620 - 1674)

No século XVII, ganhou destaque na Inglaterra, a partir das Tábuas de mortalidade de Jonh Graunt e William Petty, a aritmética política que consistiu de exaustivas análises de nascimentos e mortes. Dessas análises resultou a conclusão, entre outras, de que a percentagem de nascimentos de crianças do sexo masculino era ligeiramente superior à de crianças do sexo feminino.

William Petty
(1623 - 1687)

Em 1708, foi organizado o primeiro curso de Estatística na Universidade de Yena, na Alemanha.

A palavra estatística foi cunhada pelo acadêmico alemão Gottfried Achenwall, em 1740. Também é ele quem estabelece as relações da Estatística com outras áreas, definindo-lhe o campo de ação.

Contudo, foi o casamento entre o cálculo das probabilidades e a Estatística, em meados do século XIX, que permitiu que a Estatística fosse organicamente estruturada e ampliasse largamente o seu campo de ação. O avanço na teoria das probabilidades possibilitou a descoberta das distribuições de probabilidade e, como consequência, a criação de técnicas de amostragem mais adequadas e de formas de relacionar as amostras com as populações de onde provieram.

Outro marco decisivo no desenvolvimento dos métodos estatísticos foi o advento da computação eletrônica, ferramenta valiosíssima que permitiu que a Estatística alargasse ainda mais os seus horizontes.

1.5. Conceitos fundamentais

1.5.1. Característica e variável

As unidades de uma população se distinguem e se caracterizam por um conjunto de particularidades, propriedades ou atributos. Cada uma dessas particularidades ou propriedades é uma *característica* ou *atributo* da população e de suas unidades. Cada característica pode manifestar-se nas unidades sob diferentes alternativas ou *níveis*. Por exemplo, sexo e grau de instrução são características de indivíduos de uma população. Os níveis (alternativas) para a característica sexo são dois: masculino e feminino, e para a característica grau de instrução poderiam ser quatro: fundamental, médio, graduação e pós-graduação.

Em geral, o conjunto de características das unidades de uma população é demasiadamente vasto e não totalmente conhecido para ser completamente descrito. Assim, apenas as características relevantes numa pesquisa específica é que são consideradas. O conjunto dessas características irá depender dos objetivos e das condições de realização da pesquisa. Desse modo, o interesse estará sempre focalizado não nas unidades em si, mas nas suas características relevantes.

O termo *variável* é utilizado genericamente para indicar aquilo que é sujeito à variação ou à inconstância. No contexto da pesquisa científica, uma variável é definida como a função que estabelece uma correspondência entre os níveis de uma característica e os valores de um conjunto numérico segundo uma escala de medida. Em outras palavras, uma variável é uma característica populacional que pode ser medida de acordo com alguma escala.

1.5.2. Escalas de medida

O termo *escala de medida* é usualmente relacionado com instrumentos como régua, balança, copos de medida, utilizados para determinar comprimento, peso, volume, etc. Ou seja, comumente tende-se a associar a mensuração com um processo de medida física com escala bem definida que possui uma origem ou ponto zero natural e uma unidade de medida constante. Frequentemente, entretanto, características devem ser representadas por escalas menos informativas, que não possuem as propriedades associadas com a maioria das medidas físicas.

Podemos classificar as escalas de medida em quatro categorias: escala nominal, escala ordinal, escala intervalar e escala de razão ou racional. Cada escala de medida possui seu próprio conjunto de pressuposições referentes à correspondência de números com entidades do mundo real e ao significado da realização das várias operações matemáticas sobre esses números. A complexidade e a informação aumentam conforme aumenta o nível da escala de medida.

Escala nominal

Uma variável de *escala nominal* classifica as unidades em classes ou categorias quanto à característica que representa, não estabelecendo qualquer relação de grandeza ou de ordem. É denominada nominal porque duas categorias quaisquer se diferenciam apenas pelo nome.

A escala nominal é a menos restritiva. A igualdade ou equivalência de classes é caracterizada pelas seguintes três propriedades:

- reflexividade: cada unidade em uma classe é igual a ela própria;
- simetria: para cada duas unidades em uma mesma classe, sejam A e B, A=B implica B=A;
- transitividade: para quaisquer três unidades em uma classe, sejam A, B e C, A=B e B=C implica A=C.

Os rótulos das categorias eventualmente podem ser numéricos, mas operações aritméticas sobre esses números não têm qualquer significado com respeito aos objetos do mundo real que eles identificam. A escala nominal permite apenas algumas operações aritméticas mais elementares. Pode-se contar o número de elementos de cada classe e determinar a classe mais numerosa ou efetuar testes de hipóteses estatísticas referentes à distribuição das unidades da população nas classes. Como uma escala nominal apenas classifica unidades, mas não infere grau ou quantidade, as várias classes não podem ser manipuladas matematicamente (por exemplo, por adição ou subtração de equivalentes numéricos daquelas classes). Consequentemente, a maioria das estatísticas usuais, como média e desvio padrão não têm sentido, pois as operações para sua determinação não são permitidas. Se tudo o que pode ser dito sobre um objeto é que ele é diferente de outros, então a escala de medida é nominal.

Escala ordinal

Uma variável de *escala ordinal* classifica as unidades em classes ou categorias quanto à característica que representa, estabelecendo uma relação de ordem entre as unidades pertencentes a categorias distintas. A escala ordinal é uma escala de ordenação, designando uma posição relativa das classes segundo uma direção. Qualquer conjunto de valores que preservem a ordem são válidos para essa variável tornando, dessa forma, a escala ordinal invariante sob transformações que preservem a ordem. Ou seja, uma escala ordinal pode ser transformada em outra escala ordinal. Isto implica que, se eventualmente forem empregados números para representar as classes, apenas a propriedade de ordem deve ser respeitada.

Assim como na escala nominal, operações aritméticas (somas, diferenças, etc.) entre esses valores não têm sentido. A escala ordinal mantém a propriedade da equivalência de classes da escala nominal, no sentido de que unidades equivalentes recebem a mesma ordem. Além da propriedade de simetria da escala nominal, a escala ordinal tem a propriedade de assimetria. Isso significa que classes podem ser designadas não apenas como equivalentes a outras classes, mas também como não equivalentes. Assim, por exemplo, uma escala ordinal pode designar que a classe A é maior do que a classe B e, portanto, que a classe B é menor que a classe A. A propriedade de transitividade é preservada na escala ordinal: se a classe A é maior ou mais elevada que a classe B, qualquer unidade particular da classe A é maior ou mais elevada que qualquer unidade específica da classe B.

Essas propriedades adicionais caracterizam a superioridade da escala ordinal em relação à escala nominal. Entretanto, as descrições estatísticas ainda são limitadas. As medidas descritivas restringem-se ao grupo das medidas de ordem (separatrizes) porque as operações aritméticas usuais não podem ser efetuadas com símbolos que caracterizam apenas ordem e designam quantidade vagamente. Alguns procedimentos estatísticos são especificamente apropriados para dados de ordem. Se o que pode ser dito sobre um objeto é que ele é maior, melhor, mais colorido, etc. que outro, então a escala é ordinal.

Escala intervalar

Uma variável de *escala intervalar*, além de ordenar as unidades quanto à característica mensurada, possui uma unidade de medida constante. A escala intervalar, ou escala de intervalo, aproxima-se da concepção comum de medida, mas não possui uma origem (ou ponto zero) única. O ponto zero dessa escala é arbitrário e não expressa ausência de quantidade.

Os exemplos mais comuns de escala de intervalo são as escalas *Celsius* e *Fahrenheit*, usadas para medir a temperatura. Cada uma dessas escalas assinala um zero arbitrário e diferenças de temperatura iguais são determinadas através da identificação de volumes iguais de expansão no líquido usado no termômetro. Dessa forma, a escala de intervalo permite inferências referentes a diferenças entre unidades a serem medidas, mas não se pode dizer que um valor em um intervalo específico da escala seja um múltiplo de outro. Por exemplo, a mensuração da temperatura de unidades permite determinar quanto uma é mais quente do que outra, mas não é correto dizer que um objeto com 30^oC está duas vezes mais quente que um com temperatura de 15^oC.

Segundo a fórmula de conversão de graus Celsius para graus Fahrenheit, $F = 9/5\,C + 32$, essas temperaturas, 30^oC e 15^oC, expressas em graus Fahrenheit são, respectivamente 86^oF e 59^oF, que não estão na razão 2:1. Pode-se dizer, entretanto, que uma diferença entre dois valores em uma escala é um múltiplo de uma diferença entre dois outros valores. Por exemplo, a diferença $30^o\,C - 0^o\,C$ é o dobro da diferença $15^o\,C - 0^o\,C$. As correspondentes diferenças na escala Fahrenheit são $86^o\,F - 32^o\,C$ e $59^o\,F - 32^o\,C$, que estão na mesma razão 2:1.

A escala intervalar é invariante sob transformações lineares positivas (ou seja, transformações da forma $y = a + bx,\ b > 0$). Isso significa que uma escala de intervalo pode ser transformada em outra por meio de uma transformação linear positiva. A transformação de graus Celsius em Fahrenheit é um exemplo de transformação linear.

A maioria das medidas descritivas, tais como média, desvio padrão, coeficiente de correlação, requer apenas escala de intervalo. Entretanto, algumas medidas, como o coeficiente de variação, podem ser enganosas quando aplicadas a dados de variável de escala intervalar. Se o que pode ser dito sobre um objeto é que ele é tantas unidades maior que outro, então a escala de medida é intervalar.

Escala de razão

Uma variável de *escala de razão* ou *racional* ordena as unidades quanto à característica mensurada, possui uma unidade de medida constante e sua origem (ou ponto zero) é única. Nessa escala o valor zero expressa ausência de quantidade. A escala de razão, ou escala racional, é a mais elaborada das escalas de medida, no sentido de que permite todas as operações aritméticas. É a escala de medida mais comum nas ciências físicas, tais como as escalas para a medida de comprimento, peso, etc.

Conforme a designação sugere, razões iguais entre valores da escala racional correspondem a razões iguais entre as unidades mensuradas. Dessa forma, escalas de razão são invariantes sob transformações de proporção positivas, ou seja, transformações da forma $y = cx,\ x > 0$. Por exemplo, se uma unidade tem 3m e a outra 1m, pode-se dizer que a primeira unidade tem altura 3 vezes superior a da segunda. Isso porque, se as alturas das duas unidades forem transformadas em centímetros, suas medidas serão, respectivamente, 300cm e 100cm, que estão na mesma razão 3:1. Pode-se efetuar a transformação das medidas de uma escala racional para outra escala racional meramente pela multiplicação por uma constante apropriada. Se puder ser dito que um objeto é tantas vezes maior, mais pesado, etc. que outro, então a escala de medida é de razão.

A escala racional contém toda a informação das escalas de nível mais baixo, ou seja, igualdade de classe, ordem e igualdade de diferenças, e mais ainda. Todas as medidas descritivas podem ser determinadas para dados de uma variável expressa em escala racional.

1.5.3. Classificação de variáveis

De modo geral, as variáveis podem ser divididas em dois grupos: variáveis categóricas e variáveis numéricas.

As *variáveis categóricas*, também denominadas fatores de classificação ou simplesmente fatores, são aquelas cujos valores representam categorias ou classes. Caracterizam-se por possuir um conjunto limitado de valores (níveis) que usualmente se repetem entre as unidades. As variáveis categóricas podem ser *qualitativas* ou *quantitativas*.

Variáveis categóricas qualitativas descrevem qualidades e, de acordo com a escala de medida, são classificadas em:

- *Nominais*: quando não houver um sentido de ordenação entre os seus possíveis valores. Exemplos: sexo (com os níveis masculino e feminino), raça de cavalos (com os níveis manga-larga, crioulo e árabe, por exemplo), região geográfica (com os níveis norte, sul, sudeste e leste), estado civil (com os níveis solteiro, casado e divorciado, por exemplo), linhagens de uma cultivar em um processo de melhoramento vegetal, etc.
- *Ordinais*: quando houver um sentido de ordenação entre os seus possíveis valores. Exemplos: faixas de idade (criança, adolescente, adulto, idoso), intensidade de cor (claro, escuro), intensidade de infestação (forte, média, fraca), grau de instrução (fundamental, médio, graduação, pós-graduação) etc.

Variáveis categóricas quantitativas descrevem quantidades. Possuem os mesmos atributos das variáveis qualitativas, mas, uma vez que seus níveis expressam quantidade, a cada nível está associado um valor, denominado valor do nível. Por exemplo, se uma variável exprime a quantidade de um tranquilizante utilizado contra a insônia, então os níveis poderão ser Dose 1, Dose 2 e Dose 3 e as quantidades (valores) associadas poderão ser 0, 2 e 4 mg.

As *variáveis numéricas* são aquelas cujos valores são números reais, de modo que cada valor representa um valor da variável e não uma categoria ou uma classe. De acordo com o processo de obtenção dos seus dados (valores), as variáveis numéricas são classificadas em:

- *Discretas*: descrevem dados discretos ou de enumeração, ou seja, obtidos por *processo de contagem*. As variáveis discretas só podem assumir valores do conjunto dos números inteiros não negativos (0, 1, 2, 3, ...). Exemplos: número de sementes germinadas, número de pacientes que se recuperam, número de frutos estragados, número de filhos de um casal, etc.
- *Contínuas*: descrevem dados contínuos ou de mensuração, ou seja, obtidos por *processo de medição*. As variáveis contínuas podem assumir qualquer valor do conjunto dos reais (-10, 0, $\sqrt{2}$, π). Exemplos: peso, altura, tempo de sono, teor de umidade, temperatura corporal, etc.

Observemos que variáveis categóricas quantitativas são, de certa forma, variáveis numéricas, mas, nesse caso, os valores representam quantidades associadas a categorias (níveis do fator).

A classificação correta de uma variável é fundamental, uma vez que esta discriminação é que irá indicar a possibilidade e a forma de utilização dos procedimentos estatísticos disponíveis.

1.5.4. Observação e conjunto de dados

Os números, taxas e outras informações coletados em experimentos ou levantamentos são denominamos *dados*. Todo dado é um valor de uma variável (numérico ou não numérico). A unidade da população em que são medidas as variáveis de interesse é chamada de *unidade de observação*. Uma planta, por exemplo, pode ser a unidade de

observação em uma determinada pesquisa. Os valores obtidos para a variável medida nas unidades de observação (nas plantas) são os dados.

Observação é o conjunto de valores referentes a todas as variáveis medidas em uma unidade de observação. Por exemplo, os valores referentes ao peso de matéria seca, à estatura e ao número de perfilhos de uma planta constituem uma observação. O conjunto de todas as observações, ou seja, todos os valores referentes a todas as unidades de observação, constituem o *conjunto de dados*.

As variáveis são representadas por letras maiúsculas (X, Y, Z, etc) e os seus valores (dados) por letras minúsculas (x, y, z, etc.). Assim, se uma variável é representada por X (xis maiúsculo), todos os seus valores serão representados por x (xis minúsculo).

Para diferenciar ou individualizar os valores de uma variável, acrescenta-se um índice i = 1, 2, ..., n, que representa a unidade ou a observação. Assim, um conjunto de n valores de uma variável X será representado por x_1, x_2, x_3, ..., x_n.

Como exemplo, tomemos o conjunto de dados apresentado na tabela abaixo. Esse conjunto é constituído por 19 unidades ou observações (i), uma variável identificadora (nome), uma variável do tipo fator (sexo) e três variáveis numéricas contínuas (idade, estatura e peso).

i	Nome	Sexo	Idade	Estatura	Peso
1	Alfredo	M	14	1,75	51,03
2	Carol	F	14	1,60	46,49
3	Jane	F	12	1,52	38,33
4	João	M	12	1,50	45,13
5	Luísa	F	12	1,43	34,93
6	Roberto	M	12	1,65	58,06
7	William	M	15	1,69	50,80
8	Bárbara	F	13	1,66	44,45
9	Juca	M	12	1,46	37,65
10	Joca	M	13	1,59	38,10
11	Judite	F	14	1,63	40,82
12	Felipe	M	16	1,83	68,04
13	Tomas	M	11	1,46	38,56
14	Alice	F	13	1,44	38,10
15	Henrique	M	14	1,61	46,49
16	Janete	F	15	1,59	51,03
17	Joice	F	11	1,30	22,91
18	Maria	F	15	1,69	50,80
19	Ronaldo	M	15	1,70	60,33

Este conjunto de dados é representado simbolicamente na tabela abaixo.

i	A	B	X	Y	Z
1	a_1	b_1	x_1	y_1	z_1
2	a_2	b_2	x_2	y_2	z_2
3	a_3	b_3	x_3	y_3	z_3
...	...	...	...	...	...
19	a_{19}	b_{19}	x_{19}	y_{19}	z_{19}

1.6. Bibliografia

COSTA, S.F. **Introdução Ilustrada à Estatística (com muito humor!)**. 2.ed., São Paulo: Harbra, 1992. 303p.

FARIA, E.S. de. **Estatística**. Edição 97/1. (Apostila)

FERREIRA, D.F. **Estatística Básica**. Lavras: Editora UFLA, 2005, 664p.

FREUND, J.E., SIMON, G.A. **Estatística Aplicada. Economia, Administração e Contabilidade**. 9.ed., Porto Alegre: Bookman, 2000. 404p.

PIMENTEL GOMES, F. **Iniciação à Estatística**. São Paulo: Nobel, 1978. 211p.

SILVA, J.G.C. da. **Estatística experimental:** análise estatística de experimentos. (Apostila) 2000. 318p.

SILVEIRA JÚNIOR, P., MACHADO, A.A., ZONTA, E.P., SILVA, J.B. da **Curso de Estatística**. v.1, Pelotas: Universidade Federal de Pelotas, 1989. 135p.

SPIEGEL, M.R. **Estatística**. São Paulo: McGraw-Hill, 1972. 520p.

Sistema Galileu de Educação Estatística. Disponível em: http://www.galileu.esalq.usp.br

Unidade II
Estatística Descritiva

2. Estatística Descritiva

O método científico, quando aplicado para solução de um problema científico, frequentemente gera dados em grande quantidade e de grande complexidade. Desse modo, a análise da massa de dados individuais, na maioria das vezes, não revela a informação subjacente, gerando a necessidade de algum tipo de condensação ou resumo dos dados.

A Estatística Descritiva é a parte da Estatística que desenvolve e disponibiliza métodos para resumo e apresentação de dados estatísticos com o objetivo de facilitar a compreensão e a utilização da informação ali contida.

Em resumo, a Estatística Descritiva tem por finalidade a utilização de tabelas, gráficos, diagramas, distribuições de frequência e medidas descritivas para:

- examinar o formato geral da distribuição dos dados;
- verificar a ocorrência de valores atípicos;
- identificar valores típicos que informem sobre o centro da distribuição;
- verificar o grau de variação presente nos dados.

Evidentemente, a validade do resumo dos dados está intimamente ligada à quantidade de informação disponível e à qualidade da obtenção dos dados. Pode-se pensar que todo método descritivo possui uma entrada, os dados, e uma saída, que pode ser uma medida descritiva ou um gráfico. Se a entrada é deficiente a saída também será de má qualidade.

2.1. Apresentação de dados

2.1.1. Séries Estatísticas

A reunião ou agrupamento de dados estatísticos, quando apresentados em tabelas ou em gráficos, para apreciação ou investigação, determina o surgimento das *séries estatísticas*.

As séries estatísticas resumem um conjunto ordenado de observações através de três fatores fundamentais:

a) tempo: refere-se a data ou a época em que o fenômeno foi investigado;

b) espaço: refere-se ao local ou região onde o fato ocorreu;

c) espécie: refere-se ao fato ou fenômeno que está sendo investigado e cujos valores numéricos estão sendo apresentados.

As séries estatísticas são classificadas de acordo com o fator que estiver variando, podendo ser simples ou mistas.

◆ **Séries simples:** são aquelas em que apenas um fator varia. Podem ser de três tipos:

– Série histórica (temporal ou cronológica ou evolutiva): onde varia o tempo permanecendo fixos o espaço e a espécie do fenômeno estudado.

Exemplo:

Tabela 2.1. Casos de sarampo notificados no Brasil de 1987 a 1992.

Ano	Número de casos
1987	65.459
1988	26.173
1989	55.556
1990	61.435
1991	45.532
1992	7.934

Fonte: Anuários estatísticos – IBGE.

– Série geográfica (territorial ou regional): onde varia o espaço permanecendo fixos o tempo e a espécie do fenômeno estudado.

Exemplo:

Tabela 2.2. Necessidades médias de energia em alguns países, em 1973.

País	kcal/per capita/dia
Brasil	2.174
Estados Unidos	2.397
Etiópia	2.120
Japão	1.125
México	2.114

Fonte: Necessidades Humanas de Energia – IBGE.

– Série especificativa (qualitativa ou categórica): onde varia a espécie do fenômeno estudado permanecendo fixos o tempo e espaço.

Exemplo:

Tabela 2.3. Abate de animais, por espécie, no Brasil, em 1993.

Espécie	Número de cabeças
Aves	1.232.978.796
Bovinos	14.951.359
Suínos	13.305.932
Ovinos	926.818
Caprinos	803.188
Equinos	165.691

Fonte: Anuário Estatístico do Brasil (1994).

◆ **Séries mistas:** são aquelas em que mais de um fator varia ou um fator varia mais de uma vez.

Exemplos:

– Série histórica geográfica (ou geográfica histórica)

Tabela 2.4. Taxa de atividade feminina urbana (em percentual) em três regiões do Brasil, 1981/90.

Região	Ano			
	1981	1984	1986	1990
Norte	28,9	30,3	34,0	37,1
Nordeste	30,2	32,6	34,3	37,8
Sudeste	34,9	37,2	40,1	40,7

Fonte: Anuário Estatístico do Brasil (1992).

– Série especificativa geográfica (ou geográfica especificativa)

Tabela 2.5. Consumo per capita anual de alguns tipos de alimentos, em algumas regiões metropolitanas do Brasil, no ano de 1988.

Cidade	Consumo (kg)		
	Hortaliças	Carnes	Pescado
Belo Horizonte	44,5	21,6	1,3
Rio de Janeiro	54,3	24,7	4,9
São Paulo	46,7	26,1	2,9
Curitiba	36,2	24,1	1,7
Porto Alegre	48,9	34,2	1,5

Fonte: Anuário Estatístico do Brasil (1992).

– Série especificativa histórica (ou histórica especificativa)

Tabela 2.6. Taxa de mortalidade (em percentual) de menores de um ano no Brasil, segundo as três principais causas, no período de 1984 a 1987.

Causa	1984	1985	1986	1987
Doenças infecciosas intestinais	20,6	17,3	17,9	16,8
Pneumonia	12,1	11,7	12,0	10,8
Perinatal	42,4	45,8	45,3	48,0

Fonte: Informe Epidemiológico SUS.

– Série especificativa histórica geográfica

Tabela 2.7. Número de vítimas em acidentes, segundo as grandes regiões do Brasil, nos anos de 1991 e 1992.

Região	Vítimas fatais		Vítimas não fatais	
	1991	1992	1991	1992
Norte	1.188	1.165	10.229	9.739
Nordeste	3.857	3.843	23.774	23.942
Sudeste	11.555	10.217	130.938	159.669
Sul	4.402	4.213	61.797	58.832
Centro-Oeste	2.220	1.949	22.147	22.086
Brasil	23.222	21.387	248.885	274.268

Fonte: Anuário Estatístico do Brasil (1994).

◆ **Série distribuição de frequências:** ocorre quando nenhum dos fatores varia. Nesta série os dados são agrupados em classes (intervalos com limites predeterminados) segundo suas respectivas frequências. Segundo a natureza dos dados, as distribuições de frequências, podem ser de dois tipos.

– Para dados de enumeração

Tabela 2.8. Número de alarmes falsos, acionados acidentalmente ou por mau funcionamento do equipamento, recebidos diariamente por uma empresa de segurança, na cidade de Pelotas, no mês de abril de 2003.

Classes (Número de alarmes falsos)	Frequência (Número de dias)
2	2
3	6
4	8
5	4
6	5
7	3
8	2
Total	30

Fonte: Dados fictícios.

– Para dados de mensuração

Tabela 2.9. Peso de 80 estudantes da Escola São José, em 1980.

Classes (Peso, em kg)	Frequência (Número de estudantes)
40 ├— 50	12
50 ├— 60	28
60 ├— 70	25
70 ├— 80	10
80 ├— 90	5
Total	80

Fonte: Dados fictícios.

A série distribuição de frequências será abordada com maiores detalhes na Seção 2.2 desta unidade.

2.1.2. Tabelas

A tabela é a forma não discursiva de apresentar informações, das quais o dado numérico se destaca como informação central. Sua finalidade é apresentar os dados de modo ordenado, simples e de fácil interpretação, fornecendo o máximo de informação num mínimo de espaço.

A construção de uma tabela, entretanto, deve obedecer a uma série de normas técnicas. Estas normas podem ser encontradas na publicação do IBGE intitulada "Normas de Apresentação Tabular" que tem como objetivo orientar a apresentação racional e uniforme de dados estatísticos na forma tabular.

Seguem abaixo algumas das principais normas e recomendações.

- **Elementos da tabela**

Uma tabela estatística é composta de elementos essenciais e elementos complementares. Os elementos essenciais são:

– Título: é a indicação que precede a tabela contendo a designação do fato observado, o local e a época em que foi estudado.
– Corpo: é o conjunto de linhas e colunas onde estão inseridos os dados.
– Cabeçalho: é a parte superior da tabela que indica o conteúdo das colunas.
– Coluna indicadora: é a parte da tabela que indica o conteúdo das linhas.

Os elementos complementares são:

– Fonte: entidade que fornece os dados ou elabora a tabela.
– Notas: informações de natureza geral, destinadas a esclarecer o conteúdo das tabelas.
– Chamadas: informações específicas destinadas a esclarecer ou conceituar dados numa parte da tabela. Deverão estar indicadas no corpo da tabela, em números arábicos entre parênteses, à esquerda nas casas e à direita na coluna indicadora.

Os elementos complementares devem situar-se no rodapé da tabela, na mesma ordem em que foram descritos.

- **Número da tabela**

Uma tabela deve ter número para identificá-la sempre que o documento apresentar uma ou mais tabelas, permitindo, assim, a sua localização. A identificação da tabela deve ser feita em números arábicos, de modo crescente, precedidos da palavra Tabela, podendo ou não ser subordinada a capítulos ou seções de um documento. Exemplos: Tabela 5, Tabela 10.4.

- **Apresentação de dados numéricos**

Toda tabela deve ter dado numérico para informar a quantificação de um fato específico observado, o qual deve ser apresentado em números arábicos.

A parte inteira dos dados numéricos deve ser separada por pontos ou espaços de três em três algarismos, da direita para a esquerda, por exemplo: 12.243.527 ou 12 243 527. A separação da parte inteira da decimal deve ser feita por vírgula, por exemplo: 25,67.

No sistema inglês, a separação da parte inteira é feita por vírgula, e a separação da parte inteira da decimal é feita por ponto, ou seja, é o inverso do sistema brasileiro.

- **Sinais convencionais**

Sempre que um dado numérico não puder ser apresentado, o mesmo deve ser substituído por um sinal convencional. A substituição de um dado numérico deve ser feita por um dos sinais abaixo, conforme o caso.

a) – (traço): indica dado numérico igual a zero não resultante de arredondamento;
b) .. (dois pontos): indica que não se aplica dado numérico;
c) ... (três pontos): indica dado numérico não disponível;
d) x (xis): indica dado numérico omitido a fim de evitar a individualização da informação;
e) 0, 0,0 ou 0,00: indica dado numérico igual a zero resultante de arredondamento.
f) ? (interrogação): quando há dúvida sobre a veracidade da informação.

Quando uma tabela contiver sinais convencionais, estes deverão ser apresentados em nota geral com seus respectivos significados.

- **Arredondamento**

Quando o primeiro algarismo a ser abandonado for 0, 1, 2, 3 ou 4, fica inalterado o último algarismo a permanecer. Exemplo: 48,23 $\rightarrow$ 48,2.
Quando o primeiro algarismo a ser abandonado for 5, 6, 7, 8 ou 9, aumenta-se de uma unidade o último algarismo a permanecer. Exemplo: 23,87 $\rightarrow$ 23,9.

- **Unidade de medida**

Uma tabela deve ter unidade de medida, inscrita no cabeçalho ou nas colunas indicadoras, sempre que houver necessidade de se indicar, complementarmente ao título, a expressão quantitativa ou metrológica a dos dados numéricos.
Esta indicação deve ser feita com símbolos ou palavras, entre parênteses. Exemplos: (m) ou (metros), (t) ou (toneladas), (R$) ou (reais).

uando os dados numéricos forem divididos por uma constante, esta deve ser indicada por algarismos arábicos, símbolos ou palavras, entre parênteses, precedendo a unidade de medida, quando for o caso. Exemplos:
(1.000 t): indica dados numéricos em toneladas que foram divididos por mil;
(1.000 R$): indica dados numéricos em reais que foram divididos por mil;
(%) ou (percentual): indica dados numéricos proporcionais a cem;
(1/1.000): indica dados numéricos divididos por 1/1.000, ou seja, multiplicados por mil.

- **Classe de frequência**

A classe de frequência é cada um dos intervalos não superpostos em que se divide uma distribuição de frequências. Toda classe deve ser apresentada, sem ambiguidade, por extenso ou com notação.
Toda a classe que inclui o extremo inferior do intervalo (EI) e exclui o extremo superior (ES), deve ser apresentada de uma destas duas formas:

$$EI \vdash ES \quad \text{ou} \quad [EI; ES)$$

- **Apresentação de tempo**

Toda a série histórica consecutiva deve ser apresentada por seus pontos inicial e final, ligados por hífen (–). Exemplos:
1892-912: quando varia o século;
1960-65: quando variam os anos dentro do século;
out 1991 - mar 1992: quando variam os meses dentro de anos.

Toda a série histórica não consecutiva deve ser apresentada por seus pontos inicial e final, ligados por barra (/). Exemplos:
1981/85: indica dados não apresentados para pelo menos um ano do intervalo;
out 1991 / mar 1992: indica dados não apresentados para pelo menos um mês do intervalo.

◆ **Apresentação da tabela**

– O corpo da tabela deve ser delimitado, no mínimo, por três traços horizontais.

– Recomenda-se não delimitar as tabelas à direita e à esquerda por traços verticais. É facultativo o uso de traços verticais para a separação de colunas no corpo da tabela.

– Quando, por excessiva altura, a tabela tiver que ocupar mais de uma página, não deve ser delimitada inferiormente, repetindo-se o cabeçalho na página seguinte. Deve-se usar no alto do cabeçalho a palavra continuação ou conclusão, conforme o caso.

– Se possuir muitas linhas e poucas colunas, poderá ser apresentada em duas ou mais partes dispostas lado a lado e separadas por traço duplo.

– A disposição da tabela deve estar na posição normal de leitura. Caso isso não seja possível, a apresentação será feita de forma que a rotação da página seja no sentido horário.

Exemplo:

Tabela 2.10. Total de estabelecimentos, pessoal ocupado, valor da produção e valor da transformação industrial das indústrias metalúrgicas, por Unidade da Federação do Brasil, 1982.

Unidade da Federação	Total de estabelecimentos	Pessoal ocupado (1)	Valor da produção (2) (1.000 Cr$)	Valor da transformação industrial (1.000 Cr$)
Rondônia	1	x	x	x
Acre	2	x	x	x
Amazonas	31	1.710	21.585	10.103
Roraima	2	x	x	x
Pará	43	1.675	6.492	3.287
Amapá	–	–	–	–
Maranhão	14	328	498	251
Piauí	12	193	454	159
Ceará	74	5.336	21.732	10.878
Rio Grande do Norte	11	343	1.267	383
Paraíba	30	794	2.089	1.265
Pernambuco	105	5.171	44.673	14.506
Alagoas	20	439	4.101	1.768
Sergipe	20	423	1.447	534
Bahia	116	5.527	89.072	27.679
Minas Gerais	736	54.264	954.258	306.856
Espírito Santo	42	2.281	22.923	6.297
Rio de Janeiro	847	40.768	635.731	177.358
São Paulo	4.699	272.983	2.531.363	939.032
Paraná	449	11.118	43.797	22.014
Santa Catarina	305	10.816	84.294	41.894
Rio Grande do Sul	706	30.103	156.680	74.316
Mato Grosso do Sul	29	485	1.643	623
Mato Grosso	13	528	884	686
Goiás	106	2.686	9.860	4.800
Distrito Federal	28	843	2.577	1.301
Brasil	8.452	448.932	4.637.512	1.646.043

Fonte: Pesquisa Industrial - 1982-1984. Dados gerais, Brasil. Rio de Janeiro: IBGE, v.9, 410p.
Nota: Sinais convencionais utilizados:
x Dado numérico omitido a fim de evitar a individualização da informação.
– Dado numérico igual a zero não resultante de arredondamento.
(1) Em 31.12.1982.
(2) Inclui o valor dos serviços prestados a terceiros e a estabelecimentos da mesma empresa.

2.1.3. Gráficos

Outro modo de apresentar dados estatísticos é sob uma forma ilustrada, comumente chamada de gráfico. Os gráficos constituem-se numa das mais eficientes formas de apresentação de dados.

Um gráfico é, essencialmente, uma figura construída a partir de uma tabela; mas, enquanto a tabela fornece uma ideia mais precisa e possibilita uma inspeção mais rigorosa aos dados, o gráfico é mais indicado para situações que visem proporcionar uma impressão mais rápida e maior facilidade de compreensão do comportamento do fenômeno em estudo.

Os gráficos e as tabelas se prestam, portanto, a objetivos distintos, de modo que a utilização de uma forma de apresentação não exclui a outra.

Para a confecção de um gráfico, algumas regras gerais devem ser observadas:

- **Normas para representação gráfica**

– Os gráficos, geralmente, são construídos num sistema de eixos chamado sistema cartesiano ortogonal. A variável independente é localizada no eixo horizontal (abscissas), enquanto a variável dependente é colocada no eixo vertical (ordenadas). No eixo vertical, o início da escala deverá ser sempre zero, ponto de encontro dos eixos.

– Iguais intervalos para as medidas deverão corresponder a iguais intervalos para as escalas. Exemplo: Se ao intervalo 10-15 kg corresponde 2 cm na escala, ao intervalo 40-45 kg também deverá corresponder 2 cm, enquanto ao intervalo 40-50 kg corresponderá 4 cm.

– O gráfico deverá possuir título, fonte, notas e legenda, ou seja, toda a informação necessária à sua compreensão, sem auxílio do texto.

– O gráfico deverá possuir formato aproximadamente quadrado para evitar que problemas de escala interfiram na sua correta interpretação.

- **Tipos de gráficos**

Podemos considerar quatro tipos principais de representação gráfica:

Estereogramas: são gráficos onde as grandezas são representadas por volumes. Geralmente são construídos num sistema de eixos bidimensional, mas podem ser construídos num sistema tridimensional para ilustrar a relação entre três variáveis. Exemplo:

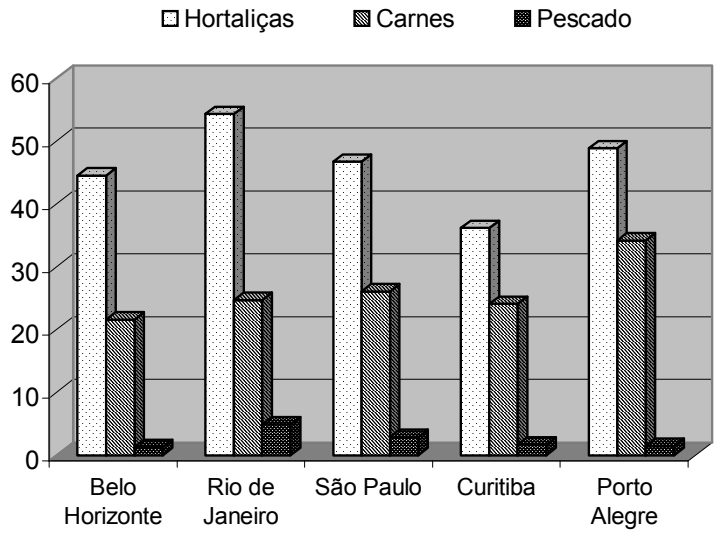


Figura 2.1. Consumo, em kg, de alguns tipos de alimentos “per capita” anual em algumas regiões metropolitanas do Brasil, em 1988.
Fonte: Anuário Estatístico do Brasil (1992).

Cartogramas: são representações em cartas geográficas (mapas).

Pictogramas ou gráficos pictóricos: são gráficos puramente ilustrativos, construídos de modo a ter grande apelo visual, dirigidos a um público muito grande e heterogêneo. Não devem ser utilizados em situações que exijam maior precisão. Exemplo:

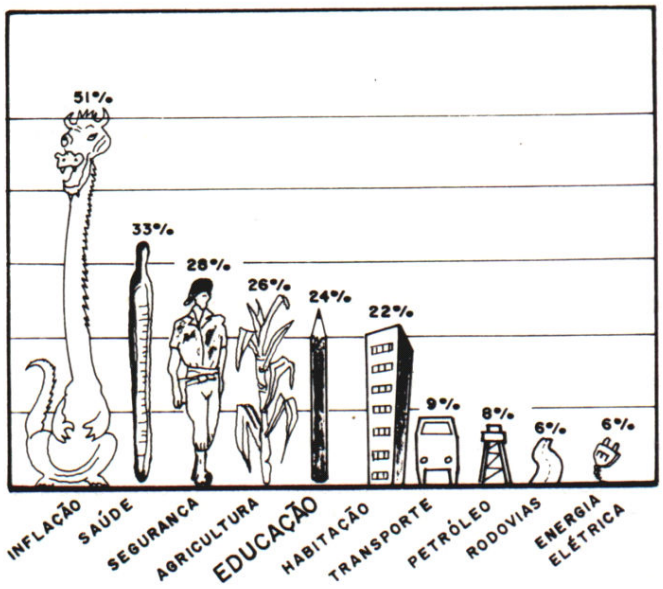


Figura 2.2. Problemas a serem solucionados pelo governo brasileiro de acordo com um levantamento encomendado pelo Ministério da Educação, em 1985.
Fonte: Silveira Júnior et al. (1989).

Diagramas: são gráficos geométricos de duas dimensões, de fácil elaboração e grande utilização. Podem ser ainda subdivididos em: gráficos de colunas, de barras, de linhas ou curvas e de setores.

a) *Gráfico de colunas:* neste gráfico as grandezas são comparadas através de retângulos de mesma largura, dispostos verticalmente e com alturas proporcionais às grandezas. A distância entre os retângulos deve ser, no mínimo, igual a 1/2 e, no máximo, 2/3 da largura da base dos mesmos. Exemplo:

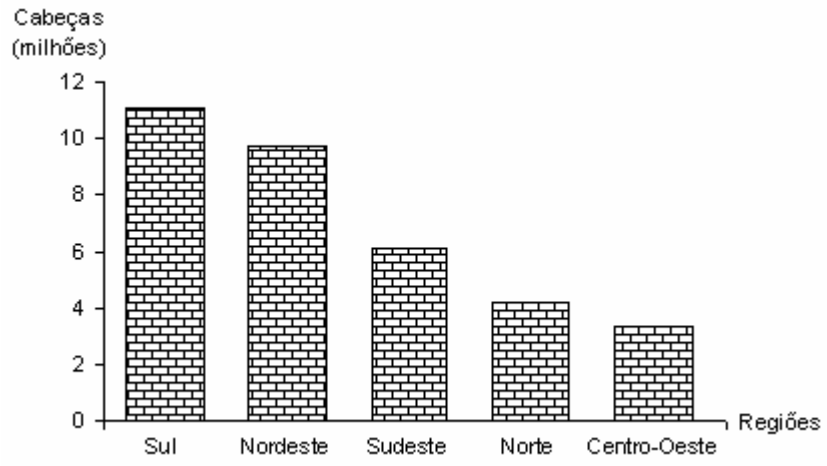


Figura 2.3. Efetivo do rebanho suíno no Brasil, segundo as grandes regiões em 1992.
Fonte: Anuário Estatístico do Brasil (1994).

b) *Gráfico de barras:* segue as mesmas instruções que o gráfico de colunas, tendo a única diferença que os retângulos são dispostos horizontalmente. É usado quando as inscrições dos retângulos forem maiores que a base dos mesmos. Exemplo:

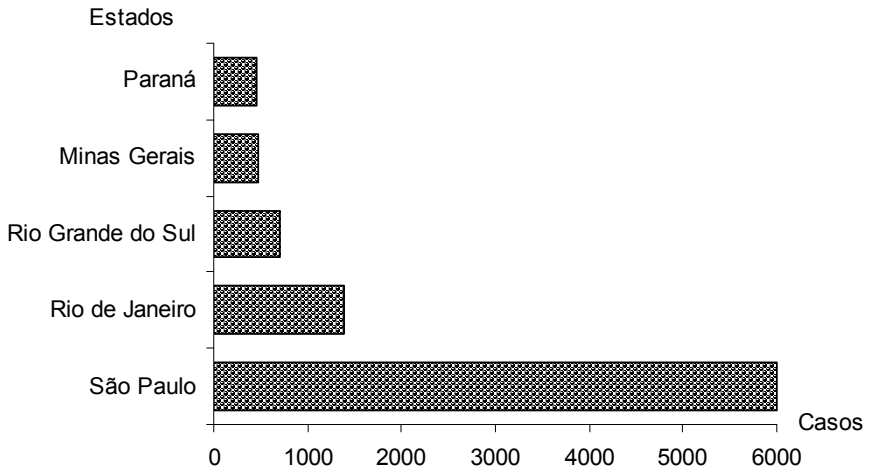


Figura 2.4. Casos notificados de AIDS nos cinco estados brasileiros de maior incidência em 1992.
Fonte: Anuário Estatístico do Brasil (1994).

c) *Gráfico de linhas ou curvas:* neste gráfico os pontos são dispostos no plano de acordo com suas coordenadas, e a seguir são ligados por segmentos de reta. É muito utilizado em séries históricas e em séries mistas quando um dos fatores de variação é o tempo, como instrumento de comparação. Exemplo:

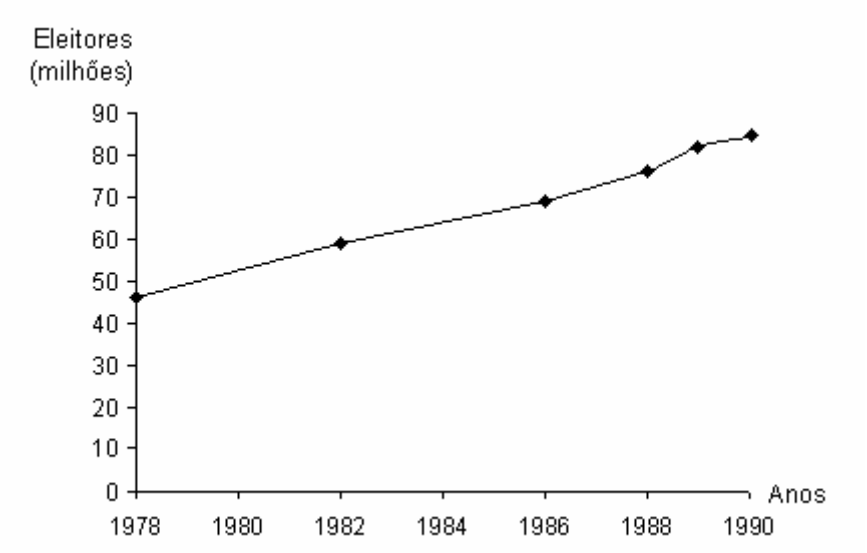


Figura 2.5. Eleitores inscritos para as eleições brasileiras - 1978/90.
Fonte: Anuário Estatístico do Brasil (1992).

d) *Gráfico em setores:* é recomendado para situações em que se deseja evidenciar o quanto cada informação representa do total. A figura consiste num círculo onde o total (100%) representa 360°, subdividido em tantas partes quanto for necessário à representação. Essa divisão se faz por meio de uma regra de três simples. Com o auxílio de um transferidor efetua-se a marcação dos ângulos correspondentes a cada divisão. Exemplo:

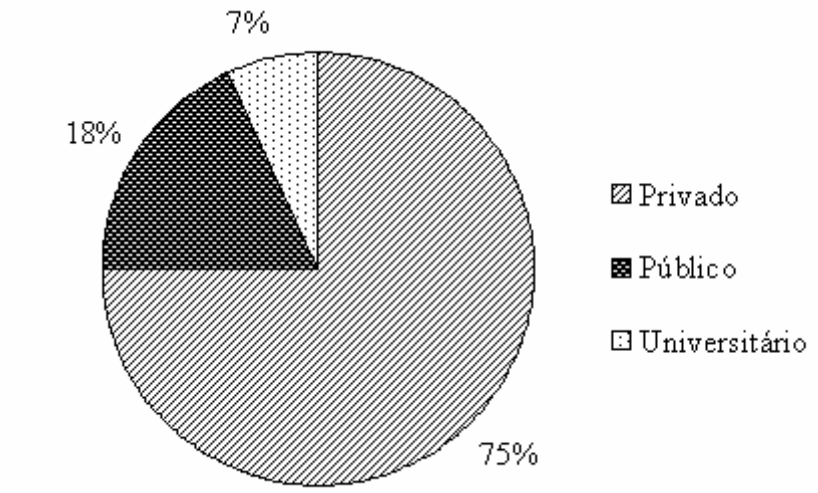


Figura 2.6. Hospitalizações pagas pelo SUS, segundo a natureza do prestador de serviço – 1993.
Fonte: Anuário Estatístico do Brasil (1994).

2.2. Distribuição de frequências e gráficos

Um grande número de dados necessita de uma forma eficiente de sumarização. Uma das formas mais comuns de resumir e apresentar dados é através de tabelas de distribuição de frequências. Estas tabelas podem ser de dois tipos: de *classificação simples* ou de *classificação cruzada*.

2.2.1. Tabelas de classificação simples

As tabelas de classificação simples são tabelas de frequências relativas a uma variável. As características dessas tabelas variam de acordo com o tipo de variável em estudo. Se a variável é do tipo categórica (fator), então são obtidas as frequências de ocorrência de cada nível dessa variável. Se a variável é do tipo numérica contínua, primeiro são obtidos intervalos de mesma amplitude e depois contados os valores que ocorrem em cada intervalo.

2.2.1.1. Distribuição de frequências de variáveis categóricas

Quando a variável em estudo for categórica ou, em alguns casos, numérica discreta, a tabela de distribuição de frequências apresentará a seguinte característica: cada valor da variável constituirá uma classe.

- **Construção da tabela**

A construção da tabela de distribuição de frequência para variáveis categóricas envolve apenas dois passos bastante simples:

1º passo: ordenar os níveis do fator, ou seja, colocá-los em ordem crescente de grandeza (rol). Cada nível constituirá uma classe. O número de cada classe da distribuição será representado por j, tal que j = 1, 2, ..., k.

2º passo: contar o número de elementos em cada classe, ou seja, contar quantas vezes o dado está repetido.

Veremos, por meio de exemplos, como construir uma tabela de distribuição de frequências para os dados de uma variável categórica (Exemplo 1) e de uma variável numérica discreta (Exemplo 2).

Exemplo 1:

Seja a variável em estudo o conceito obtido por 60 estudantes na disciplina de Estatística, para o qual os dados observados foram os seguintes:

ruim, médio, bom, médio, ruim, médio, ruim, médio, ruim, bom, médio, médio, bom, médio, médio, médio, ótimo, médio, bom, ótimo, bom, ótimo, médio, ótimo, médio, ruim, médio, ótimo, médio, médio, bom, ruim, bom, bom, médio, ruim, médio, médio, ótimo, médio, bom, ruim, ruim, bom, médio, médio, ruim, bom, médio, médio, bom, bom, bom, médio, ruim, bom, médio, médio, ruim, médio

Podemos observar que esta variável categórica qualitativa ordinal apresenta quatro níveis (ruim, médio, bom e ótimo). Como cada nível deve constituir uma classe da distribuição de frequências, já está determinado que o número total de classes (k) é quatro. O primeiro passo é a ordenação dos níveis da variável. Assim, temos

Número da classe (j)	Classe
1	Ruim
2	Médio
3	Bom
4	Ótimo

O passo seguinte é a contagem do número de estudantes em cada nível. Estes valores são denotados por F_j e chamados de *frequências absolutas* das classes. A partir da frequência absoluta podemos obter outras frequências de interesse numa distribuição, tais como:

- *frequência absoluta acumulada* na classe j, denotada por F_j', que expressa o número de elementos (observações) acumulados até a classe j;

- *frequência relativa* da classe j, denotada por f_j, que expressa a proporção de elementos (observações) na classe j;

- *frequência relativa acumulada* na classe j, denotada por f_j', que expressa a proporção de elementos (observações) acumulados até a classe j.

As frequências obtidas são então apresentadas na forma tabular.

Tabela 2.11. Frequência do conceito obtido por estudantes na disciplina de Estatística. UFPel, 2001.

j	Classe	F_j	F_j'	f_j	f_j'
1	Ruim	12	12	0,2	0,2
2	Médio	27	39	0,45	0,65
3	Bom	15	54	0,25	0,9
4	Ótimo	6	60	0,1	1
	Σ	60	-	1	-

Exemplo 2:

Muito frequentemente, as tabelas de distribuição de frequência de variáveis numéricas discretas são construídas da mesma forma que as das variáveis categóricas. Consideremos agora que a variável em estudo seja o número de animais portadores de brucelose em 350 propriedades rurais. Os valores observados para esta variável foram:

2, 5, 6, 0, 4, 4, 3, 4, 2, 2, 3, 3, 5, 3, 5, 1, 2, 4, 2, 3, 5, 4, 3, 3, 2, 3, 0, 4, 4, 3, 4, 0, 3, 1, 2, 4, 2, ...

Como cada valor da variável deve constituir uma classe e foram observados apenas sete valores diferentes para esta variável, a tabela de distribuição de frequências terá sete classes.

Número da classe (j)	Classe
1	0
2	1
3	2
4	3
5	4
6	5
7	6

Através da contagem do número de vezes que cada valor apareceu, ou seja, do número de observações em cada classe, obtemos as frequências absolutas, relativas e acumuladas, apresentadas na tabela a seguir.

Tabela 2.12. Frequência do número de animais portadores de brucelose em 350 propriedades rurais. UFPel, 2001.

j	Classe	F_j	F_j'	f_j	f_j'
1	0	55	55	0,1571	0,1571
2	1	60	115	0,1714	0,3286
3	2	112	227	0,32	0,6486
4	3	82	309	0,2343	0,8829
5	4	31	340	0,0886	0,9714
6	5	8	348	0,0229	0,9943
7	6	2	350	0,0057	1,0000
	Σ	350	-	1,0000	-

Devemos observar, ainda, que tão importante quanto saber construir uma tabela é saber interpretar os seus valores. Vejamos, como exemplo, o significado de alguns valores da tabela:

$F_4 = 82 \rightarrow$ significa que, das 350 propriedades rurais consultadas, 82 possuem três animais portadores de brucelose.

$F_3' = 227 \rightarrow$ significa que, das 350 propriedades rurais consultadas, 227 possuem menos de três animais portadores de brucelose.

$f_2 = 0,1714 \rightarrow$ significa que a proporção de propriedades rurais que possuem apenas um animal portador de brucelose é de 0,1714 (em percentual: 17,14).

$f_5' = 0,9714 \rightarrow$ significa que a proporção de propriedades rurais que possuem menos de quatro animais portadores de brucelose é de 0,9714 (em percentual: 97,14).

2.2.1.2. Distribuição de frequências de variáveis numéricas contínuas

Ao contrário das variáveis discretas, as variáveis contínuas assumem, em geral, muitos valores e, em sua grande maioria, diferentes uns dos outros. Para contornar problemas desse tipo, as tabelas de distribuição de frequências para variáveis contínuas são construídas de modo que cada classe seja constituída por um intervalo de valores da variável.

Devemos observar, no entanto, que em algumas situações uma variável discreta também poderá assumir tantos valores diferentes que a construção de uma tabela onde cada valor constitui uma classe seja impraticável. Em outras palavras, pode ocorrer que ela tenha tantas linhas que sua construção pouco auxilie na descrição resumida dos dados. Nesses casos, por uma questão de simplificação, é usual agrupar os dados discretos em intervalos de classe, da mesma forma que se agrupam os dados contínuos.

- **Construção da tabela**

O processo de construção da tabela de distribuição de frequência para variáveis numéricas segue os seguintes passos:

1º passo: ordenar o conjunto de dados, ou seja, colocar os dados brutos em ordem crescente de grandeza (rol).

2º passo: determinar o número de classes da tabela. De modo geral, este valor não deverá ser inferior a 5 e nem superior a 15. A definição do número de classes deverá ser orientada pelos objetivos do trabalho, mas existem algumas regras objetivas de determinação, como, por exemplo:

$$k = 1 + 3,32 \times \log n \text{ (Fórmula de Sturges)} \quad \text{ou} \quad k = \sqrt{n},$$

onde:

k = número de classes;
n = número de observações;
log = logaritmo de base 10.

3º passo: determinar a amplitude do intervalo. Para isto, podemos utilizar a seguinte expressão:

$$i = \frac{a_t}{k}$$

onde:

i = amplitude do intervalo;
$a_t = ES - EI$: amplitude total do conjunto de valores;
k = número de classes.

Convencionamos, também, que o arredondamento no número de classes (k) ou na amplitude do intervalo (i) é sempre feito para cima.

4º passo. Construir os intervalos de classe. O limite inferior da primeira classe será sempre o menor valor do conjunto de dados ($x_{(1)}$) e o limite superior será o limite inferior acrescido do valor da amplitude do intervalo de classe (i). Na sequência, o limite inferior da segunda classe será o limite superior da primeira e o limite superior da segunda classe será este limite inferior acrescido da amplitude do intervalo. Para todas as classes subsequentes, os intervalos deverão ser construídos da mesma forma que para a segunda:

j	Classe
1	$x_{(1)} \vdash\!\!- x_{(1)} + i$
2	$x_{(1)} + i \vdash\!\!- x_{(1)} + 2i$
...	...
k	$x_{(1)} + (k-1)i \vdash\!\!- x_{(1)} + ki$

Notamos, assim, que a amplitude do intervalo é constante para todas as classes. O intervalo fechado à esquerda e aberto à direita, representado pelo símbolo |— , garante a não superposição de classes.

Exemplo:

Tomemos a seguinte variável:

X = peso ao nascer (em kg) de 60 bovinos machos da raça Ibagé, para a qual os valores observados (e já ordenados) foram:

16, 17, 17, 18, 18, 18, 19, 20, 20, 20, 20, 20, 21, 21, 22, 22, 23, 23, 23, 23, 23, 23, 23, 23, 23, 25, 25, 25, 25, 25, 25, 26, 26, 27, 27, 27, 27, 28, 28, 28, 29, 29, 29, 30, 30, 30, 30, 30, 30, 30, 31, 32, 33, 33, 33, 34, 34, 35, 36, 39.

Sendo o peso uma variável contínua cujos valores poderiam ser todos diferentes entre si, não podemos considerar cada valor como sendo uma classe, de modo que não podemos saber de antemão o número de classes da distribuição de frequência. Este valor deverá ser determinado e, para isto, usaremos a fórmula de Sturges. Para n = 60, temos

$$k = 1 + 3,32 \times \log n$$

$$k = 1 + 3,32 \times \log 60$$

$$k = 1 + 3,32 \times 1,778 = 6,9$$

Como o número de classes tem que ser um número inteiro, teremos que arredondar o valor 6,9. Usaremos como regra o arredondamento para cima. Deste modo, o número de classes será k = 7.

Uma vez determinado o valor de k, temos que obter a amplitude dos intervalos. Sendo k = 7 e a amplitude total do conjunto de dados

$$a_t = ES - EI$$

$$a_t = 39 - 16 = 23,$$

temos

$$i = \frac{a_t}{k} = \frac{23}{7} = 3,2857.$$

Por uma questão de praticidade, vamos arredondar o valor da amplitude do intervalo para uma casa decimal, lembrando que o arredondamento, também neste caso, deverá ser sempre para cima. Assim, temos i = 3,3.

O próximo passo é a construção dos intervalos de classe. Tomamos como limite inferior da primeira classe o menor valor do conjunto de dados $x_{(1)}$ = 16. Somando ao 16 o valor da amplitude do intervalo i = 3,3, obtemos o limite superior deste intervalo. Todos os demais intervalos são construídos considerando como limite inferior o limite superior do intervalo de classe que o precede e como limite superior a soma do limite inferior com o valor 3,3. Assim, temos:

j	Classes
1	16,0 |— 19,3
2	19,3 |— 22,6
3	22,6 |— 25,9
4	25,9 |— 29,2
5	29,2 |— 32,5
6	32,5 |— 35,8
7	35,8 |—| 39,1

Para a obtenção das frequências absolutas das classes, contamos quantos valores (observações) do conjunto de dados pertencem a cada intervalo. As demais frequências, como já vimos anteriormente, derivam da frequência absoluta.

Em distribuições de frequências de variáveis contínuas, geralmente existe interesse em uma outra quantidade conhecida como ponto médio ou centro de classe, denotada por c_j. Os centros de classe são calculados da seguinte forma:

$$c_j = \frac{EI_j + ES_j}{2},$$

onde:

EI_j = extremo inferior da classe j

ES_j = extremo superior da classe j

No exemplo, temos:

$$c_1 = \frac{16 + 19,3}{2} = \frac{35,3}{2} = 17,65$$

$$c_2 = \frac{19,3 + 22,6}{2} = \frac{41,9}{2} = 20,95$$

$$\dots$$

$$c_7 = \frac{35,8 + 39,1}{2} = \frac{74,9}{2} = 37,45.$$

A tabela de frequências completa é apresentada a seguir.

Tabela 2.13. Frequência do peso ao nascer (em kg) de 60 bovinos machos da raça Ibagé. UFPel, 2001.

j	Classes	F_j	F'_j	f_j	f'_j	c_j
1	16 |— 19,3	7	7	0,1167	0,1167	17,65
2	19,3 |— 22,6	9	16	0,15	0,2667	20,95
3	22,6 |— 25,9	15	31	0,25	0,5167	24,25
4	25,9 |— 29,2	12	43	0,2	0,7167	27,55
5	29,2 |— 32,5	9	52	0,15	0,8667	30,85
6	32,5 |— 35,8	6	58	0,1	0,9667	34,15
7	35,8 |—| 39,1	2	60	0,0333	1,0000	37,45
	Σ	60	–	1,0000	–	–

A interpretação das frequências da tabela é exemplificada através de alguns valores:

$F_3 = 15 \rightarrow$ significa que 15 dos 60 bovinos nasceram com peso entre 22,6 e 25,9 kg (exclusive).

$F'_5 = 52 \rightarrow$ significa que 52 dos 60 bovinos nasceram com peso entre 16,0 e 32,5 kg (exclusive).

$f_2 = 0,15 \rightarrow$ significa que a proporção de bovinos que nasceram com peso entre 19,3 e 22,6 kg (exclusive) é de 0,15 (em percentual: 15).

$f'_6 = 0,9667 \rightarrow$ significa que a proporção de bovinos que nasceram com peso entre 16 e 35,8 kg (exclusive) é de 0,9667 (em percentual: 96,67).

Exercícios propostos:

2.1. Os dados a seguir se referem aos números de pães não vendidos em uma certa padaria até a hora do encerramento do expediente:

0	0	4	2	0	1	0	2	0	4
1	0	0	3	2	0	1	0	0	0
2	0	0	1	0	0	3	2	1	7
0	1	0	0	2	0	0	3	2	1

Construa a distribuição de frequências para esses dados.

2.2. Os dados em rol (ordenação horizontal) abaixo se referem aos valores gastos (em reais) pelas primeiras 50 pessoas que entraram em um determinado supermercado, no dia 01/01/2000.

3,11	8,88	9,26	10,81	12,69	13,78	15,23	15,62	17,00	17,39
18,36	18,43	19,27	19,50	19,54	20,16	20,59	22,22	23,04	24,47
24,58	25,13	26,24	26,26	27,65	28,06	28,08	28,38	32,03	36,37
38,98	38,64	39,16	41,02	42,97	44,08	44,67	45,40	46,69	48,65
50,39	52,75	54,80	59,07	61,22	70,32	82,70	85,76	86,37	93,34

Faça a distribuição de frequências desses dados.

2.2.1.3. Representação gráfica das distribuições de frequências

As distribuições de frequências podem ser representadas graficamente de duas formas distintas e exclusivas, são elas: o histograma e o polígono de frequências.

- **Histograma**

O histograma consiste de um conjunto de retângulos contíguos cuja base é igual à amplitude do intervalo e a altura proporcional à frequência das respectivas classes.

Na figura abaixo podemos observar o histograma da distribuição de frequências da Tabela 2.13.

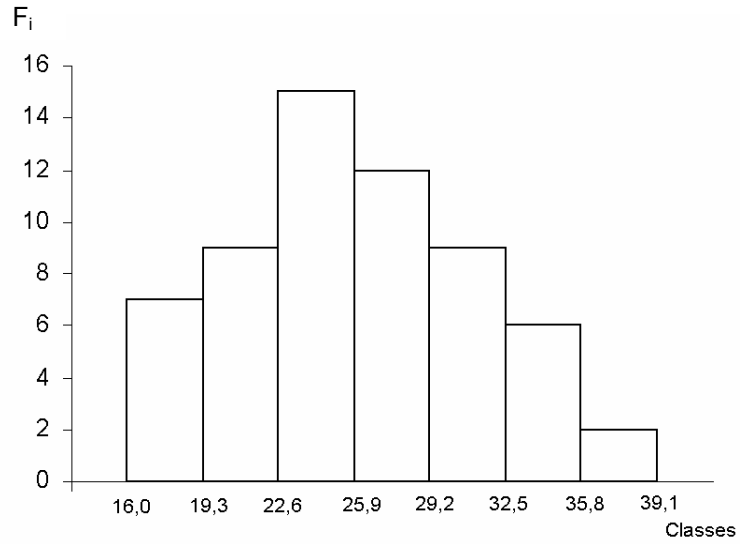


Figura 2.6. Peso ao nascer (em kg) de 60 bovinos machos da raça Ibagé. UFPel, 2001.

Quando trabalhamos com variáveis discretas, os retângulos dos histogramas se reduzem a retas e, consequentemente, deixam de ser contíguos. Vejamos um exemplo na figura a seguir que representa a distribuição da Tabela 2.12.

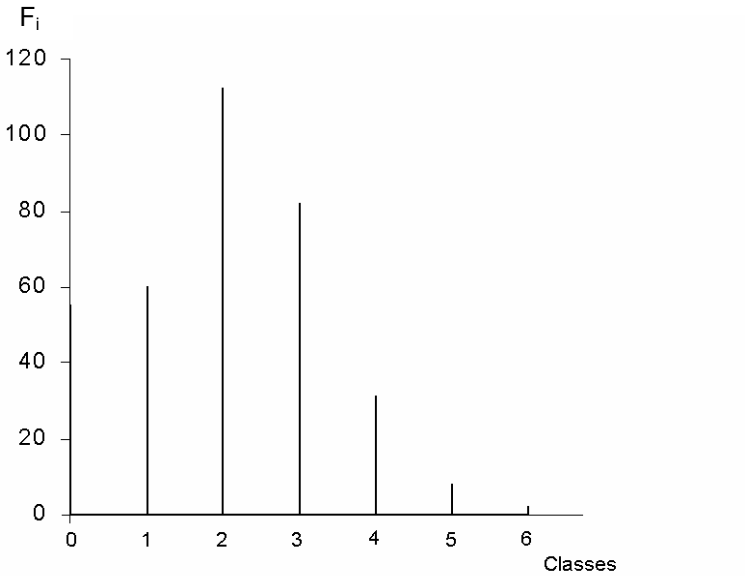


Figura 2.7. Número de animais portadores de brucelose em 350 propriedades rurais. UFPel, 2001.

◆ **Polígono de frequência**

O polígono de frequências é constituído por segmentos de retas que unem os pontos cujas coordenadas são o ponto médio e a frequência de cada classe. O polígono de frequências é fechado tomando-se uma classe anterior a primeira e uma posterior a última, uma vez que ambas possuem frequência zero.

Na Figura 2.8 podemos observar o polígono de frequências da distribuição da Tabela 2.13.

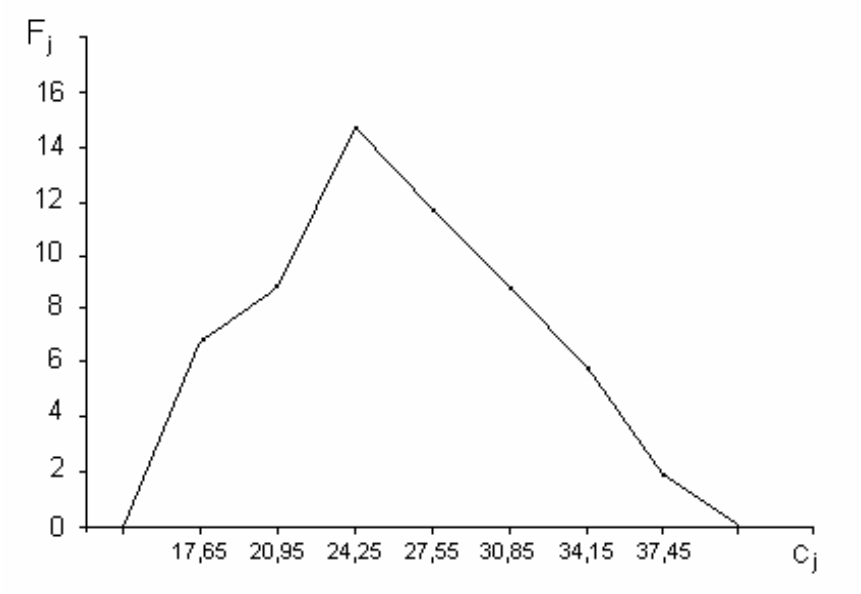


Figura 2.8. Peso ao nascer (em kg) de 60 bovinos machos da raça Ibagé. UFPel, 2001.

Exercício proposto:

2.3. Faça a representação gráfica (histograma e polígono de frequências, quando for o caso) das distribuições de frequências construídas nos Exercícios 2.1 e 2.2 da página 30.

2.2.2. Tabelas de classificação cruzada

Em algumas situações, pode haver interesse no estudo de duas ou mais variáveis simultaneamente. Daí surgem as distribuições conjuntas de frequências. As tabelas de classificação cruzada são tabelas de frequências relativas a duas variáveis, numéricas ou categóricas. Existe um número razoável de tipos de tabelas e gráficos para descrever esses casos.

2.2.2.1. Frequências cruzadas de variáveis categóricas

Quando um estudo envolve duas variáveis categóricas (fatores), a tabela de frequência cruzada dessas duas variáveis é conhecida também como tabela de dupla entrada, tabela de associação ou tabela de contingência. As regras básicas para sua construção são semelhantes às das tabelas de classificação simples. A diferença é que agora a tabela apresenta duas margens, cada qual com os totais referentes a um dos fatores.

Na Tabela 2.14, por exemplo, os 60 alunos da escola E foram classificados segundo duas variáveis categóricas: Conceito em Estatística e Hábito de fumar. Para isso, primeiramente, os alunos são classificados de acordo com o Conceito em Estatística e, posteriormente, dentro de cada nível deste fator, são classificados quanto ao Hábito de fumar.

Tabela 2.14. Distribuição dos alunos da escola E, segundo o hábito de fumar e conceito em Estatística.

Conceito	Hábito de fumar		Totais
	Sim	Não	
Ruim	5	8	13
Médio	10	16	26
Bom	5	10	15
Ótimo	2	4	6
Totais	22	38	60

Podemos observar que, com as frequências marginais (totais) da tabela cruzada, poderíamos resgatar a tabela de classificação simples de cada fator.

A representação gráfica de distribuições de frequências de variáveis categóricas pode ser feita através de dois tipos de gráficos:

- *Gráficos em duas dimensões (diagramas):* descrevendo a variação de um fator dentro dos níveis do outro.

Por exemplo, na Figura 2.9, observamos a variação do fator Hábito de fumar dentro de cada nível do fator Conceito em Estatística, enquanto que, na Figura 2.10, fica mais evidente a variação do fator Conceito em Estatística dentro de cada nível do fator Hábito de fumar.

Pode não ser necessário apresentar os dois gráficos simultaneamente. É mais comum apresentar apenas um deles, de acordo com o fato que desejamos ressaltar. Assim, no exemplo, se for mais importante ressaltar a distribuição de fumantes e não fumantes dentro de cada conceito, utilizamos a Figura 2.9. Se for mais importante ressaltar a distribuição do conceito em estatística dentro dos grupos de fumantes e não fumantes, utilizamos a Figura 2.10. Naturalmente, se ambas as situações forem relevantes podemos apresentar os dois diagramas.

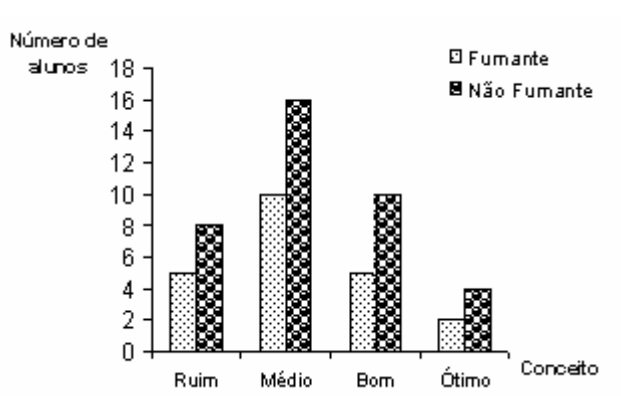


Figura 2.9. Distribuição dos alunos da escola E, segundo o hábito de fumar e conceito em Estatística.

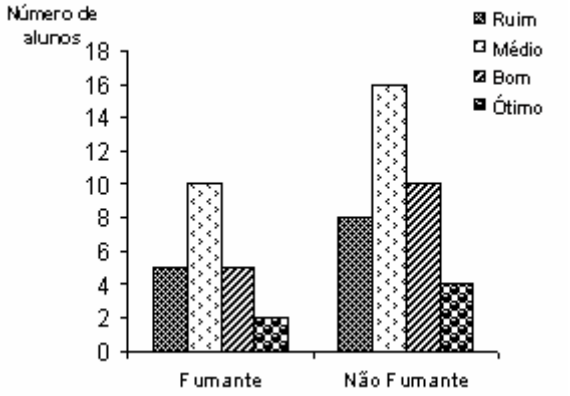


Figura 2.10. Distribuição dos alunos da escola E, segundo o hábito de fumar e conceito em Estatística.

A observação atenta destes gráficos já pode fornecer uma ideia da possível associação existente entre os fatores. Por exemplo, se o um fator apresenta o mesmo comportamento dentro de todos os níveis do outro, podemos supor que eles não estão associados, ou seja, comportam-se independentemente um do outro. Devemos observar, entretanto, que os gráficos fornecem apenas indicações, para verificar tais hipóteses (suposições) devemos utilizar os testes apropriados que serão vistos posteriormente.

– *Gráficos tridimensionais (estereogramas):* compostos por paralelogramos, dispostos em eixos tridimensionais, separados entre si, cujas bases são determinadas pelos níveis dos fatores e as alturas pelas suas respectivas frequências (Figura 2.11).

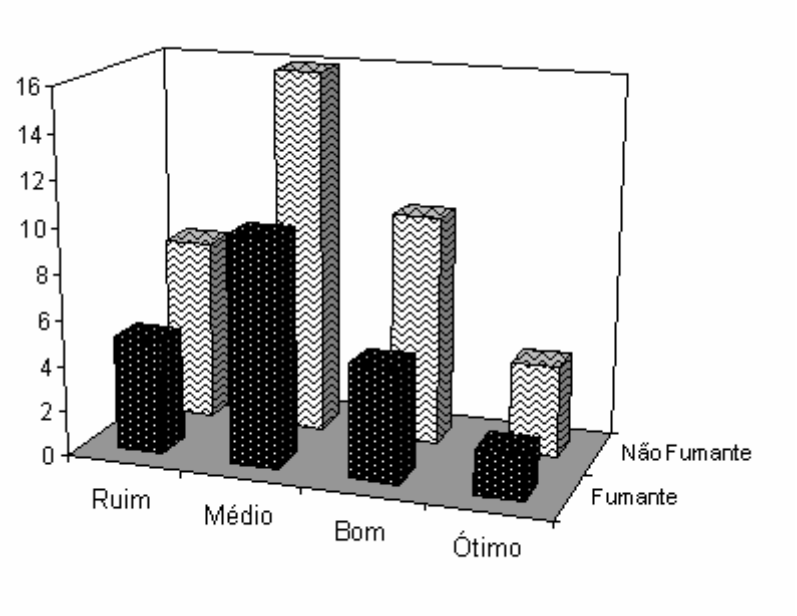


Figura 2.11. Distribuição dos alunos da escola E, segundo o hábito de fumar e conceito em Estatística.

2.2.2.2. Frequências cruzadas de variáveis numéricas

Ao estudarmos conjuntamente duas variáveis numéricas, as tabelas de classificação cruzada são, agora, denominadas tabelas de correlação. As ideias básicas sobre a construção dessas tabelas já foram vistas em seções anteriores.

As tabelas de frequências cruzadas de duas variáveis contínuas também são construídas de modo similar às de classificação simples, ou seja, seguindo todos os passos já descritos na Seção 2.2.1.2. Primeiramente, procedemos à classificação das observações segundo uma das variáveis, para em seguida, dentro de cada classe da primeira, classificá-las de acordo com a outra variável. Por exemplo, na Tabela 2.15, observamos a classificação dos 400 alunos do Colégio C, segundo duas variáveis contínuas: Nota em Estatística e Nota em Matemática.

Tabela 2.15. Distribuição dos alunos do Colégio C, segundo suas notas em Estatística e Matemática.

Estatística	Matemática			Totais
	0 ⊢ 4	4 ⊢ 7	7 ⊢ 10	
0 ⊢ 4	32	25	5	62
4 ⊢ 7	20	183	82	285
7 ⊢ 10	7	27	19	53
Totais	59	235	106	400

Os gráficos geralmente utilizados para descrever dados como estes são os histogramas em três dimensões (estereogramas), nos quais os retângulos cedem lugar aos paralelogramos. Agora, a base de cada paralelogramos é definida pelas amplitudes das classes das variáveis envolvidas. Este tipo de gráfico é pouco utilizado em trabalhos científicos pela dificuldade de execução e interpretação através dos meios disponíveis.

A relação entre duas variáveis contínuas também é comumente representada por diagramas de dispersão. Tomemos outro exemplo: para estudar o relacionamento entre as variáveis Peso do pai (X) e Peso do filho (Y), foram medidos os pesos (em kg) de dez alunos do Colégio C e de seus respectivos pais. Os resultados são apresentados numa tabela de correlação:

Observação (i)	1	2	3	4	5	6	7	8	9	10
Peso dos pais (x_i)	78	65	86	68	83	68	75	80	82	66
Peso dos filhos (y_i)	60	52	68	53	65	57	58	62	65	53

Esta tabela possibilita a construção do diagrama de dispersão de pontos (Figura 2.12). Este tipo de gráfico pode fornecer uma indicação do tipo de relacionamento que existe entre as duas variáveis. Por exemplo, se os pontos apresentarem a forma de elipse indicam a existência de uma relação linear (positiva ou negativa) entre as variáveis. A Figura 2.12 parece evidenciar um relacionamento linear positivo entre os pesos dos dez alunos e os pesos dos seus respectivos pais, sugerindo um estudo mais aprofundado desta correlação.

Através da análise de regressão linear, que será abordada mais adiante, é possível obter uma equação do tipo Y = a + bX, que descreve o peso dos filhos (Y) como uma função linear do peso dos pais (X).

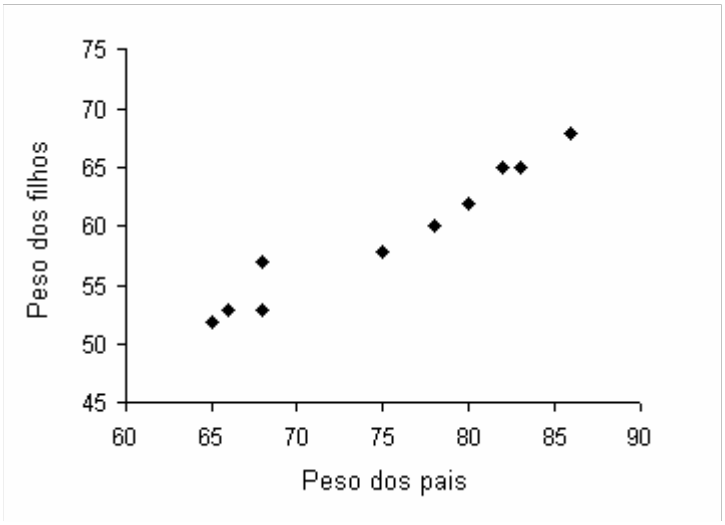


Figura 2.12. Dispersão dos pesos (em kg) de dez alunos do Colégio C e de seus respectivos pais.

2.3. Medidas Descritivas

As medidas descritivas têm o objetivo de reduzir um conjunto de dados observados (numéricos) a um pequeno grupo de valores que deve fornecer toda a informação relevante a respeito desses dados. Estas medidas são funções dos valores observados e podem ser classificadas em quatro grupos:

– *Medidas de localização*, também denominadas *medidas de tendência central* ou *medidas de posição*: indicam um ponto central onde, em muitas situações importantes, está localizada a maioria das observações.

– *Medidas separatrizes*: indicam limites para proporções de observações em um conjunto, podendo ser utilizadas para construir medidas de dispersão.

– *Medidas de variação* também denominadas *medidas de dispersão*: informam sobre a variabilidade dos dados.

– *Medidas de formato*: informam sobre o modo como os valores se distribuem. Compreendem as *medidas de assimetria*, que indicam se a maior proporção de valores está no centro ou nas extremidades, e as *medidas de curtose*, que descrevem grau de achatamento da distribuição.

Existe uma enorme variedade de medidas descritivas, muitas delas competidoras entre si. Um guia geral para escolha da medida mais adequada pode ser visto a seguir:

- Com que objetivo a medida está sendo obtida?
- A medida é fácil de interpretar? É intuitiva?
- Existem valores atípicos que podem afetá-la exageradamente?
- O propósito da análise é meramente descritivo ou planeja-se fazer inferências?

Uma medida descritiva deverá, sempre que possível, possuir as seguintes características: ser representativa, ser de fácil interpretação e prestar-se bem a tratamento matemático e/ou estatístico em etapas posteriores.

2.3.1. Medidas de localização ou tendência central

As medidas de localização ou tendência central têm o objetivo de representar o ponto de equilíbrio ou o centro de uma distribuição. Em muitos casos, podem ser considerados valores típicos ou representativos do conjunto.

As medidas mais utilizadas são a média aritmética, a mediana e a moda, embora outras também possam ser úteis em algumas situações.

◆ Média aritmética

A média aritmética, pela sua facilidade de cálculo e de compreensão aliada às suas propriedades matemáticas, é a medida de localização mais conhecida e utilizada. Pode ser de dois tipos: *simples* ou *ponderada*.

A *média aritmética simples*, representada por $\overline{x}$, é calculada considerando que todas as observações participam com o mesmo peso. Assim, para um conjunto de n observações (x_1, x_2, ..., x_n), a média aritmética simples ou simplesmente média é definida por

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

Exemplo:

Se X = tempo (h)
Para x_i = 9, 7, 5, 10, 4, temos

$$\overline{x} = \frac{\sum x_i}{n} = \frac{9+7+5+10+4}{5} = \frac{35}{5} = 7\ \text{h}$$

A *média aritmética ponderada*, representada por $\overline{x}_p$, é calculada considerando que pelo menos uma das observações deve participar com peso diferente das demais. Assim, se as observações x_1, x_2, ..., x_n forem associadas aos pesos p_1, p_2, ..., p_n, a média aritmética ponderada é dada por

$$\overline{x}_p = \frac{\sum_{i=1}^{n} x_i p_i}{\sum_{i=1}^{n} p_i}$$

Exemplo:

Para x_i = 7, 8, 6, 10, e
p_i = 10, 10, 8, 2, temos

$$\overline{x}_p = \frac{\sum x_i p_i}{\sum p_i} = \frac{7\times10+8\times10+6\times8+10\times2}{10+10+8+2} = \frac{218}{30} = 7{,}02$$

Propriedades matemáticas da média aritmética

1ª propriedade: A média de um conjunto de dados que não varia, ou seja, cujos valores são uma constante, é a própria constante.

2ª propriedade: Ao somar (ou subtrair) uma constante c por todos os valores de um conjunto de dados, sua média também é somada (ou subtraída) por esta constante.

Demonstração:

$$\begin{aligned}\overline{x}_{x+c} &= \frac{\sum(x_i + c)}{n} = \frac{x_1 + c + x_2 + c + \ldots + x_n + c}{n} \\ &= \frac{\sum x_i + \sum c}{n} \\ &= \frac{\sum x_i + nc}{n} \\ &= \frac{\sum x_i}{n} + \frac{nc}{n} = \frac{\sum x_i}{n} + c = \overline{x} + c\end{aligned}$$

Verificação numérica:

Ao somarmos a constante 2 a todos os valores do conjunto x_i = 9, 7, 5, 10, 4, cuja média é $\overline{x} = 7$, teremos um novo conjunto de valores x_i + 2 = 11, 9, 7, 12, 6, com uma nova média

$$\overline{x}_{x+2} = \frac{\sum x_i}{n} = \frac{11+9+7+12+6}{5} = \frac{45}{5} = 9 = 7 + 2\,,$$

logo, a média 7 sofreu a mesma operação que os valores x_i.

3ª propriedade: Ao multiplicar (ou dividir) uma constante c por todos os valores de um conjunto de dados, sua média também é multiplicada (ou dividida) por esta constante.

Demonstração:

$$\begin{aligned}\overline{x}_{cx} &= \frac{\sum cx_i}{n} = \frac{cx_1 + cx_2 + \ldots + cx_n}{n} \\ &= \frac{c\left(x_1 + x_2 + \ldots + x_n\right)}{n} \\ &= \frac{c\sum x_i}{n} = c\frac{\sum x_i}{n} = c\overline{x}\end{aligned}$$

Verificação numérica:

Ao multiplicarmos a constante 2 por todos os valores do conjunto de dados x_i = 9, 7, 5, 10, 4, cuja média é $\overline{x} = 7$, teremos um novo conjunto de valores $2x_i$ = 18, 14, 10, 20, 8, com uma nova média

$$\overline{x}_{2x} = \frac{\sum x_i}{n} = \frac{18+14+10+20+8}{5} = \frac{70}{5} = 14 = 2 \times 7\,,$$

logo, a média 7 sofreu a mesma operação que os valores x_i.

4ª propriedade: A soma de todos os desvios em relação à média de um conjunto de valores é nula, entendendo por desvio a diferença entre a observação e a média aritmética, ou seja,

$$\sum(x_i - \overline{x}) = 0\,.$$

É possível demonstrar esta propriedade aplicando as propriedades do somatório:

$$\begin{aligned}\sum(x_i-\overline{x}) &= \sum x_i - \sum \overline{x} \\ &= \sum x_i - n\overline{x}, \quad \text{sendo } \overline{x}=\frac{\sum x_i}{n}, \text{ temos} \\ &= \sum x_i - n\frac{\sum x_i}{n} \\ &= \sum x_i - \sum x_i = 0\end{aligned}$$

5ª propriedade: A soma dos quadrados dos desvios em relação a uma constante c, $\sum(x_i-c)^2$, é mínima quando $c=\overline{x}$.

Podemos demonstrar esta propriedade, somando e subtraindo do desvio uma constante de interesse ($\overline{x}$) e aplicando as propriedades do somatório:

$$\begin{aligned}\sum(x_i-c)^2 &= \sum(x_i-c+\overline{x}-\overline{x})^2 \\ &= \sum[(x_i-\overline{x})+(\overline{x}-c)]^2 \\ &= \sum[(x_i-\overline{x})^2+2(x_i-\overline{x})(\overline{x}-c)+(\overline{x}-c)^2] \\ &= \sum(x_i-\overline{x})^2+\sum 2(x_i-\overline{x})(\overline{x}-c)+\sum(\overline{x}-c)^2 \\ &= \sum(x_i-\overline{x})^2+2(\overline{x}-c)\sum(x_i-\overline{x})+n(\overline{x}-c)^2, \quad \text{sendo } \sum(x_i-\overline{x})=0, \text{ temos} \\ &= \sum(x_i-\overline{x})^2+n(\overline{x}-c)^2\end{aligned}$$

Observamos que $\sum(x_i-c)^2$ assumirá o menor valor quando $c=\overline{x}$, pois, neste caso, $n(\overline{x}-c)^2=0$.

Podemos verificar a 3ª e a 4ª propriedades da média aritmética no seguinte conjunto de dados:

i	x_i	$(x_i-\overline{x})$	$(x_i-\overline{x})^2$	$(x_i-6)^2$	$(x_i-9)^2$
1	9	2	4	9	0
2	7	0	0	1	4
3	5	-2	4	1	16
4	10	3	9	16	1
5	4	-3	9	4	25
Σ	35	0	26	31	46

Verificamos, assim, que:

- $\sum(x_i-\overline{x})=0$ (terceira propriedade da média)
- $\sum(x_i-\overline{x})^2=26 < \sum(x_i-6)^2=31$
- $\sum(x_i-\overline{x})^2=26 < \sum(x_i-9)^2=46$

(quarta propriedade da média)

◆ **Mediana**

A mediana, representada por Md, é a medida que divide um conjunto de dados *ordenado* em duas partes iguais: 50% dos valores ficam abaixo e 50% ficam acima da mediana.

Existem dois casos diferentes para o cálculo da mediana, mas em ambos o primeiro passo a ser tomado é a ordenação dos dados.

1º caso: quando n é ímpar

Determinamos, primeiramente, a posição mais central (p) do conjunto de dados ordenado

$$p = \frac{n+1}{2}.$$

A mediana será o valor do conjunto de dados que ocupa a posição p, ou seja,

$$Md = x_p.$$

Exemplo:

Se X = tempo (h)
Para x_i = 4, 5, 7, 9, 10, temos

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

logo,

$$Md = x_p = x_3 = 7\,h$$

2º caso: quando n é par

Neste caso, temos duas posições centrais no conjunto de dados ordenado, denotadas por p_1 e p_2. Ao utilizarmos a expressão $p = \frac{n+1}{2}$, obtemos um valor não inteiro. As posições p_1 e p_2 são os dois inteiros mais próximos do valor de p.

A mediana será a média aritmética simples dos valores do conjunto de dados que ocupam as posições p_1 e p_2, ou seja,

$$Md = \frac{x_{p_1} + x_{p_2}}{2}.$$

Exemplo:

Se X = tempo (h)
Para x_i = 4, 5, 7, 9, 10, 12, temos

$$p = \frac{n+1}{2} = \frac{6+1}{2} = 3,5 \begin{cases} p_1 = 3 \\ p_2 = 4 \end{cases}$$

logo,

$$Md = \frac{x_{p_1} + x_{p_2}}{2} = \frac{x_3 + x_4}{2} = \frac{7+9}{2} = 8\,h.$$

◆ **Moda**

A moda, representada por Mo, é o valor de maior ocorrência num conjunto de dados. É a única medida que pode não existir e, existindo, pode não ser única.

Exemplos:

X = peso (kg)

1. Para x_i = 2, 3, 7, 5, 7, 5, 8, 7, 9, temos Mo = 7 kg

2. Para x_i = 1, 3, 4, 5, 4, 8, 6, 8, temos Mo = 4 kg e 8 kg (conjunto bimodal)

3. Para x_i = 5, 7, 8, 3, 9, 1, 4, não existe Mo (conjunto amodal)

4. Para x_i = 1, 3, 4, 4, 5, 1, 3, 5, não existe Mo (conjunto amodal)

◆ **Características das principais medidas de localização ou tendência central**

O quadro abaixo apresenta as principais características da média, da mediana e da moda, destacando as vantagens e desvantagens de cada uma em relação às demais.

Média aritmética	**Mediana**	**Moda**
Vantagens - No cálculo da média participam todos os valores observados. - É uma medida de fácil interpretação e presta-se muito bem a tratamentos estatísticos adicionais. - É uma medida que sempre existe e é rígida e unicamente determinada. - É um valor típico de um conjunto de dados, podendo substituir todos os valores de um conjunto sem alterar o total. - É o ponto de equilíbrio de uma distribuição, sendo tão mais eficiente quanto mais simétrica for a distribuição dos valores ao seu redor.	*Vantagens* - Define exatamente o centro de uma distribuição, mesmo quando os valores se distribuem assimetricamente em torno da média. - Pode ser determinada mesmo quando não se conhece todos os valores do conjunto de dados. - É uma medida que sempre existe e é única. - Esta medida pode ser utilizada para definir o meio de um número de objetos, propriedades ou qualidades que possam de alguma forma ser ordenados. - É uma medida resistente, ou seja, não sofre influência de valores discrepantes.	*Vantagens* - É uma medida que têm existência real dentro do conjunto de dados e em grande número de vezes. - Não exige cálculo, apenas uma contagem. - Pode ser determinada também para variáveis qualitativas nominais.
Desvantagem - É uma medida altamente influenciada por valores discrepantes (não resistente).	*Desvantagem* - É uma medida que não se presta a cálculos matemáticos.	*Desvantagens* - É uma medida que não se presta a cálculos matemáticos. - Deixa sem representação todos os valores do conjunto de dados que não forem iguais a ela.

2.3.2. Medidas separatrizes

As medidas separatrizes delimitam proporções de observações de uma variável ordinal. Elas estabelecem limites para uma determinada proporção $0 \leq p \leq 1$ de observações. São medidas intuitivas, de fácil compreensão e frequentemente resistentes.

Para discutir medidas separatrizes, vamos considerar um conjunto de dados ordenado, representado como $y_{(1)}$, $y_{(2)}$, ..., $y_{(n)}$, pressupondo uma ordenação ascendente, de modo que $y_{(1)}$ é o menor valor e $y_{(n)}$ é o maior valor do conjunto.

Em todas as medidas separatrizes, é importante conhecer a posição que um valor ordenado ocupa em relação aos valores extremos, ou seja, a distância em relação ao extremo mais próximo. A posição ocupada por uma observação ordenada em relação à extremidade mais próxima é denominada profundidade.

Como a definição é feita em termos da extremidade mais próxima, a profundidade do mínimo e do máximo é igual a 1. O segundo menor e o segundo maior têm profundidade 2, o terceiro, 3 e assim por diante. Deste modo, têm profundidade i as observações $y_{(i)}$ e $y_{(n+1-i)}$. A profundidade de um valor ordenado é o menor valor entre i e n-i+1. Evidentemente, a profundidade cresce no sentido do centro até um certo ponto, decrescendo a seguir.

Se o número de observações é ímpar, então existe no conjunto um valor que tem a profundidade máxima. Dos n-1 valores que sobram, metade está à direita desse valor e metade está à esquerda. A *mediana* é o valor com a maior profundidade em qualquer conjunto de dados ordenado, sendo, portanto, a medida descritiva mais próxima do centro. Como é um indicador do centro do conjunto, a mediana é também uma medida de localização que compete com a média.

Como a mediana divide o conjunto em duas metades, é razoável pensar numa medida separatriz que efetue uma divisão adicional: dividir cada metade em duas metades. Essas medidas separatrizes são denominadas *quartis*. Todo o raciocínio relativo aos quartis e à mediana é facilmente estendidos para divisões adicionais. Cada quarta parte do conjunto de dados pode ser dividida em duas de modo que o conjunto fique dividido em oito partes. A medida separatriz dessas proporções é denominada *oitavo*.

De modo semelhante, é possível encontrar valores que delimitem porções expressas em percentagem de dados em um conjunto ordenado. Esses valores são denominados *percentis*. Entretanto, de todas essas medidas separatrizes, teremos interesse particular na mediana, já discutida na seção anterior, e nos quartis que serão tratados a seguir.

◆ Quartis

Os quartis, representados por Q_i, onde i = 1, 2 e 3, são três medidas que dividem um conjunto de dados *ordenado* em quatro partes iguais. São elas:

– *Primeiro quartil* (Q_1): 25% dos valores ficam abaixo e 75% ficam acima desta medida.

– *Segundo quartil* (Q_2): 50% dos valores ficam abaixo e 50% ficam acima desta medida. O segundo quartil de um conjunto de dados corresponde à mediana (Q_2 = Md).

– *Terceiro quartil* (Q_3): 75% dos valores ficam abaixo e 25% ficam acima desta medida.

Observa-se facilmente que o primeiro quartil é o percentil 0,25, a mediana é o percentil 0,5 e o terceiro quartil é o percentil 0,75.

O processo para obtenção dos quartis, da mesma forma que o da mediana, consiste em, primeiramente, ordenar os dados e, em seguida, determinar a posição (p) do quartil no conjunto de dados ordenado. Existem dois casos diferentes para a determinação de p:

1º caso: quando n é ímpar

– Para Q_1, temos: $p = \frac{n+1}{4}$;

– Para Q_2, temos: $p = \frac{2(n+1)}{4}$;

– Para Q_3, temos: $p = \frac{3(n+1)}{4}$.

2º caso: quando n é par

– Para Q_1, temos: $p = \frac{n+2}{4}$;

– Para Q_2, temos: $p = \frac{2n+2}{4}$;

– Para Q_3, temos: $p = \frac{3n+2}{4}$.

O quartil Q_i será o valor do conjunto de dados que ocupa a posição p, ou seja, $Q_i = x_p$. No caso de p não ser um número inteiro, o quartil será a média aritmética dos dois valores que ocupam as posições correspondentes ao menor e ao maior inteiros mais próximos de p. Por exemplo, se p=7,5, o quartil será a média aritmética dos valores que ocupam as posições 7 e 8.

Exemplos:

1º caso: quando n é ímpar

Seja X = peso (kg) e x_i = 3, 3, 4, 6, 7, 9, 9, 11 e 12

– Para Q_1, temos

$$p = \frac{n+1}{4} = \frac{9+1}{4} = 2,5$$

Como p não é um número inteiro, Q_1 será a média aritmética dos valores que ocupam as posições 2 e 3, ou seja,

$$Q_1 = \frac{x_2 + x_3}{2} = \frac{3+4}{2} = 3,5\,kg$$

– Para Q_2, temos

$$p = \frac{2(n+1)}{4} = \frac{2(9+1)}{4} = 5$$

Como p é inteiro, $Q_2 = x_p$, ou seja,

$$Q_2 = x_5 = 7\,kg$$

– Para Q_3, temos

$$p = \frac{3(n+1)}{4} = \frac{3(9+1)}{4} = 7,5$$

Como p não é inteiro, Q_3 será a média aritmética dos valores que ocupam as posições 7 e 8, ou seja,

$$Q_3 = \frac{x_7 + x_8}{2} = \frac{9+11}{2} = 10\,kg$$

2º caso: quando n é par

Seja X = Peso (kg) e x_i = 3, 3, 4, 6, 7, 9, 9, 11, 12 e 14

– Para Q_1, temos

$$p = \frac{n+2}{4} = \frac{10+2}{4} = 3$$

Sendo p um número inteiro, então,

$$Q_1 = x_3 = 4\,kg$$

– Para Q_2, temos

$$p = \frac{2n+2}{4} = \frac{2\times 10+2}{4} = 5{,}5$$

Como p não é inteiro, Q_2 será a média aritmética dos valores que ocupam as posições 5 e 6, ou seja,

$$Q_2 = \frac{x_5 + x_6}{2} = \frac{7+9}{2} = 8\,kg$$

– Para Q_3, temos

$$p = \frac{3n+2}{4} = \frac{3\times 10+2}{4} = 8$$

Sendo p um número inteiro, então,

$$Q_3 = x_8 = 11kg$$

2.3.3. Medidas de variação ou dispersão

As medidas de variação ou dispersão complementam as medidas de localização ou tendência central, indicando quanto as observações diferem entre si ou o grau de afastamento das observações em relação à média.

As medidas de variação mais utilizadas são: a amplitude total, a variância, o desvio padrão e o coeficiente de variação.

- **Amplitude total**

A amplitude total, denotada por a_t, fornece uma ideia de variação e consiste na diferença entre o maior valor e o menor valor de um conjunto de dados. Assim, temos

$$a_t = ES - EI,$$

onde:

ES: extremo superior do conjunto de dados ordenado;

EI: extremo inferior do conjunto de dados ordenado.

A amplitude total é uma medida pouco precisa, uma vez que utiliza apenas os dois valores mais extremos de um conjunto de dados. Também por esta razão é extremamente

influenciada por valores discrepantes. É utilizada quando apenas uma ideia rudimentar da variabilidade dos dados é suficiente.

Exemplo:

Se X = tempo (h)
Para x_i = 9, 7, 5, 10, 4, temos

$$a_t = ES - EI = 10 - 4 = 6\,h.$$

Significado: todos os valores do conjunto de dados diferem, no máximo, em 6h.

- **Amplitude interquartílica**

A amplitude interquartílica, denotada por a_q, é a diferença entre o terceiro quartil (Q_3) e o primeiro quartil (Q_1). Assim, temos

$$a_q = Q_3 - Q_1.$$

Apesar de ser uma medida pouco utilizada, a amplitude interquartílica apresenta uma característica interessante que é a resistência, ou seja, esta medida, ao contrário da amplitude total, não sofre nenhuma influência de valores discrepantes.

Exemplo:

Seja X = peso (kg) e x_i = 3, 3, 4, 6, 7, 9, 9, 11, 12,

onde: Q_3 = 10 kg e Q_1 = 3,5 kg., temos

$$a_q = Q_3 - Q_1 = 10 - 3{,}5 = 6{,}5\,kg$$

Significado: 50% dos valores (mais centrais) estão dentro deste intervalo, portanto, diferem, no máximo, em 6,5 kg.

- **Variância**

A variância, denotada por s^2, é a medida de dispersão mais utilizada, seja pela sua facilidade de compreensão e cálculo, seja pela possibilidade de emprego na inferência estatística. A variância é definida como sendo a média dos quadrados dos desvios em relação à média aritmética. Assim, temos

$$s^2 = \frac{\sum (x_i - \overline{x})^2}{n-1},$$

onde:

$n-1$: é o número de graus de liberdade ou desvios independentes.

A utilização do denominador $n-1$, em vez de n, tem duas razões fundamentais:

1. Como a soma dos desvios é nula, ou seja, $\sum (x_i - \overline{x}) = 0$, existe $n-1$ desvios independentes, isto é, conhecidos $n-1$ desvios o último está automaticamente determinado, pois a soma é zero.

2. O divisor $n-1$ faz com que a variância possua melhores propriedades estatísticas.

Nos casos em que a variância for utilizada apenas para descrever a variação de um conjunto de dados, então, ela poderá ser calculada utilizando o número de observações (n) como denominador e será denotada por s_n^2, ou seja,

$$s_n^2 = \frac{\sum (x_i - \overline{x})^2}{n}.$$

Mas se o objetivo for descrever a variação dos dados de uma amostra que será utilizada para inferir sobre a população, então a medida que deve ser utilizada é a variância com denominador $n-1$.

Exemplo:

Se X = tempo (h)
Para x_i = 9, 7, 5, 10, 4, onde $\overline{x} = 7\,h$, temos

$$\begin{aligned} s^2 &= \frac{\sum (x_i - \overline{x})^2}{n-1} \\ &= \frac{(9-7)^2 + (7-7)^2 + (5-7)^2 + (10-7)^2 + (4-7)^2}{5-1} \\ &= \frac{4+0+4+9+9}{4} = \frac{26}{4} = 6,5\,h^2. \end{aligned}$$

Propriedades matemáticas da variância

1ª propriedade: A variância de um conjunto de dados que não varia, ou seja, cujos valores são uma constante, é zero.

$$s_c^2 = \frac{\sum (c-c)^2}{n-1} = 0.$$

2ª propriedade: Se somarmos uma constante c a todos os valores de um conjunto de dados, a variância destes dados não se altera.

$$s_{x+c}^2 = \frac{\sum \left[(x_i + c) - (\overline{x} + c)\right]^2}{n-1} = \frac{\sum (x_i - \overline{x} + c - c)^2}{n-1} = \frac{\sum (x_i - \overline{x})^2}{n-1} = s^2.$$

3ª propriedade: Se multiplicarmos todos os valores de um conjunto de dados por uma constante c, a variância destes dados fica multiplicada pelo quadrado desta constante.

$$s_{cx}^2 = \frac{\sum (cx_i - c\overline{x})^2}{n-1} = \frac{\sum \left[c(x_i - \overline{x})\right]^2}{n-1} = \frac{\sum c^2 (x_i - \overline{x})^2}{n-1} = c^2 \frac{\sum (x_i - \overline{x})^2}{n-1} = c^2 s^2.$$

Desvantagens da variância:

– Como a variância é calculada a partir da média, é uma medida pouco resistente, ou seja, muito influenciada por valores discrepantes.

– Como a unidade de medida fica elevada ao quadrado, a interpretação da variância se torna mais difícil.

◆ **Desvio Padrão**

O desvio padrão, denotado por s, surge para solucionar o problema de interpretação da variância e é definido como a raiz quadrada positiva da variância. Assim, temos

$$s = \sqrt{s^2}\,.$$

Exemplo:

Se X = tempo (h)
Para x_i = 9, 7, 5, 10, 4, onde $s^2 = 6{,}5\,h^2$, temos

$$s = \sqrt{s^2} = \sqrt{6{,}5\,h^2} = 2{,}55\,h$$

Podemos observar que o desvio padrão é expresso na mesma unidade de medida que os dados, o que facilita a sua interpretação. Geralmente, o desvio padrão é apresentado junto com a média do conjunto de dados da seguinte forma: $\bar{x} \pm s$. Deste modo, temos a indicação da variação média dos dados em torno da média aritmética.

◆ **Coeficiente de Variação**

O coeficiente de variação, denotado por CV, é a medida mais utilizada quando existe interesse em comparar variabilidades de diferentes conjuntos de dados. Embora esta comparação possa ser feita através de outras medidas de variação, nas situações em que as médias dos conjuntos comparados são muito desiguais ou as unidades de medida são diferentes, devemos utilizar o CV.

O coeficiente de variação é definido como a proporção da média representada pelo desvio padrão e dado por

$$CV = \frac{s}{\bar{x}}\ 100\,.$$

Exemplo:

Se X = tempo (h)
Para x_i = 9, 7, 5, 10, 4, onde $\bar{x} = 7\,h$ e $s = 2{,}55\,h$, temos

$$CV = \frac{s}{\bar{x}}100 = \frac{2{,}55\,h}{7\,h}100 = 36{,}4\%$$

As vantagens do coeficiente de variação sobre as demais medidas de variação são as seguintes:

– O CV é desprovido de unidade de medida, uma vez que, é expresso em percentagem;
– O CV é uma medida relativa, ou seja, que relaciona o desvio padrão (s) com a sua respectiva média aritmética ($\bar{x}$). Deste modo, um desvio padrão maior pode, algumas vezes, representar uma variabilidade menor quando relacionado com a sua média.

A conveniência da utilização do CV para a comparação das variabilidades de conjuntos de dados com médias ou com unidades de medida diferentes pode ser verificada nos seguintes exemplos:

1. Consideremos que x_{1i} e x_{2i} são conjuntos de valores referentes a produção diária de leite (em kg) de vacas das raças Holandesa e Jersey, respectivamente, para os quais foram obtidas as seguintes medidas:

Holandesa: $\overline{x}_1 = 25\,\text{kg}$, $s_1 = 4,2\,\text{kg}$, $CV_1 = 16,8\%$

Jersey: $\overline{x}_2 = 13\,\text{kg}$, $s_2 = 3,4\,\text{kg}$, $CV_2 = 26,2\%$

Podemos observar que se utilizássemos o desvio padrão para comparar as variações dos conjuntos de dados, concluiríamos que o grupo das vacas holandesas é mais variável, pois apresenta o maior valor para esta medida. Entretanto, não podemos deixar de considerar que o desvio padrão 4,2, mesmo sendo o maior, quando relacionado à média 25, representa uma porção menor deste valor do que o desvio padrão 3,4 quando relacionado à média 13. Sendo assim, quando as médias são muito desiguais, devemos utilizar na comparação dos conjuntos de valores o CV que é uma medida relativa.

2. Consideremos, agora, que x_i e y_i são conjuntos de valores referentes a alturas (em cm) e pesos (em kg), respectivamente, de um grupo de estudantes, para os quais foram obtidas as seguintes medidas:

Altura: $\overline{x} = 165\,\text{cm}$, $s_X = 30\,\text{cm}$, $CV_X = 18,2\%$

Peso: $\overline{y} = 58\,\text{kg}$, $s_Y = 9\,\text{kg}$, $CV_Y = 15,5\%$

Verificamos, neste caso, que para a comparação de conjuntos de valores expressos em diferentes unidades de medida, o CV é a única medida que pode ser utilizada por ser desprovida de unidade de medida. Se utilizássemos qualquer outra medida de variação estaríamos comparando centímetros com quilogramas, o que não seria possível, uma vez que tais grandezas não são comparáveis.

2.3.4. Medidas de formato

O formato é um aspecto importante de uma distribuição. Embora mudanças em uma medida de variação também provoquem alterações no aspecto visual, o formato de uma distribuição se relaciona com as ideias de *simetria* e *curtose*.

Várias medidas têm o objetivo de informar sobre o formato de uma distribuição. Entre as mais precisas estão os coeficientes de assimetria e de curtose, que são calculados a partir dos momentos da distribuição.

◆ Momentos

Os momentos, denotados por m_r, são medidas calculadas com o propósito de estudar a distribuição. De um modo geral, tanto mais conhecemos uma distribuição quanto mais conhecermos sobre os seus momentos. O momento de ordem r centrado num valor a é dado por

$$m_r = \frac{\sum (x_i - a)^r}{n}.$$

Dois valores de a geram momentos importante num conjunto de dados:

– Quando $a = 0$, temos os momentos centrados na origem, denominados *momentos ordinários de ordem* r e representados por m'_r. Assim, temos

$$m'_r = \frac{\sum x_i^r}{n}.$$

Exemplos:

Para r = 1, temos: $m'_1 = \frac{\sum x_i}{n}$

Para r = 2, temos: $m'_2 = \frac{\sum x_i^2}{n}$

Para r = 3, temos: $m'_3 = \frac{\sum x_i^3}{n}$

Para r = 4, temos: $m'_4 = \frac{\sum x_i^4}{n}$

– Quando $a = \overline{x}$, temos os *momentos de ordem* r *centrados na média* e representados por m_r. Assim, temos

$$m_r = \frac{\sum (x_i - \overline{x})^r}{n}.$$

Exemplos:

Para r = 1, temos: $m_1 = \frac{\sum (x_i - \overline{x})}{n}$

Para r = 2, temos: $m_2 = \frac{\sum (x_i - \overline{x})^2}{n}$

Para r = 3, temos: $m_3 = \frac{\sum (x_i - \overline{x})^3}{n}$

Para r = 4, temos: $m_4 = \frac{\sum (x_i - \overline{x})^4}{n}$

◆ **Coeficiente de assimetria**

Entre as várias medidas de assimetria que devem informar se a maioria dos valores se localiza à esquerda, ou à direita, ou se estão uniformemente distribuídos em torno da média aritmética, temos o *coeficiente de assimetria*, denotado por a_3. Esta medida indica o grau e o sentido do afastamento da simetria e é obtida utilizando o segundo e o terceiro momentos centrados na média, através da seguinte expressão

$$a_3 = \frac{m_3}{m_2\sqrt{m_2}}.$$

A classificação da distribuição quanto a simetria é feita de acordo com o valor do a_3:

– Se $a_3 < 0$, a distribuição é classificada como *assimétrica negativa*, indicando que a maioria dos valores são maiores ou se localizam à direita da média aritmética.

– Se $a_3 = 0$, a distribuição é classificada como *simétrica*, indicando os valores estão uniformemente distribuídos em torno da média aritmética.

– Se $a_3 > 0$, a distribuição é classificada como *assimétrica positiva*, indicando que a maioria dos valores são menores ou se localizam à esquerda da média aritmética.

Assimétrica negativa Simétrica Assimétrica positiva

◆ **Coeficiente de curtose**

As medidas de curtose indicam o grau de achatamento de uma distribuição. O *coeficiente de curtose*, denotado por a_4, é calculado a partir do segundo e do quarto momentos centrados na média, através da seguinte expressão

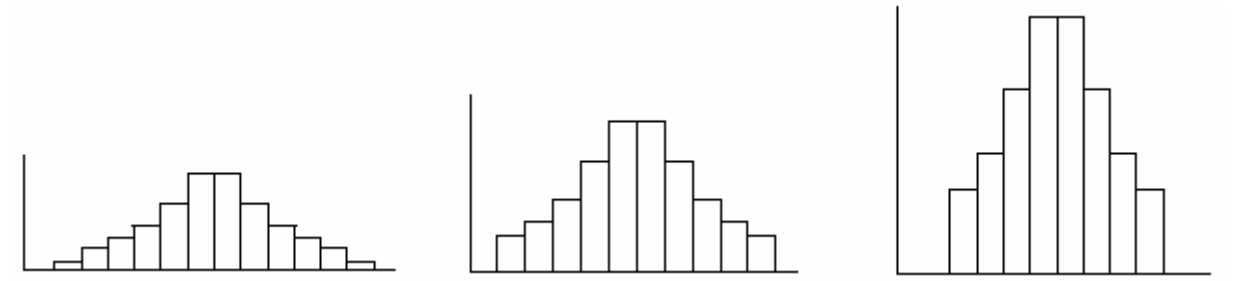
$$a_4 = \frac{m_4}{m_2^2}.$$

A curtose está relacionada com o grau de concentração das observações no centro e nas caudas da distribuição e não tem interpretação tão intuitiva quanto a simetria. A classificação da distribuição é feita de acordo com o valor do a_4, tendo por base a curtose que ocorre na distribuição normal, que é classificada como mesocúrtica.

– Se $a_4 < 3$, a distribuição é classificada como *platicúrtica*, indicando que ocorre baixa concentração de valores no centro, tornando a distribuição mais achatada que a distribuição normal. A concentração de valores nos eixos média mais ou menos o desvio padrão é maior que na distribuição normal.

– Se $a_4 = 3$, a distribuição é classificada como *mesocúrtica*, indicando que a concentração das observações ocorre de forma semelhante à da distribuição normal. A concentração de valores nos eixos média mais ou menos o desvio padrão é maior que na distribuição normal.

– Se $a_4 > 3$, a distribuição é classificada como *leptocúrtica*, indicando que ocorre alta concentração de valores no centro e nas caudas, o que provoca um pico maior que o da distribuição normal. A concentração de valores em torno dos eixos média mais ou menos o desvio padrão é menor do que na distribuição normal.



Platicúrtica Mesocúrtica Leptocúrtica

As medidas de curtose são muito pouco utilizadas, exceto em algumas áreas específicas como vendas, onde geralmente existe interesse no estudo da extensão dos picos das distribuições.

2.3.5. Medidas descritivas para dados agrupados em classe

As medidas descritivas podem ser calculadas a partir de dados agrupados em classe. Entretanto, quando calculadas a partir de tabelas de distribuição de frequências de variáveis contínuas, essas medidas, em geral, são apenas aproximações das medidas obtidas a partir dos dados não agrupados.

◆ **Medidas de localização ou tendência central**

– Média aritmética

Nas distribuições de frequências de variáveis contínuas inexistem valores individuais. Consideramos, então, que o melhor representante dos valores de uma classe é o centro de classe (c_j) e, a partir destes valores, determinamos a média da variável. Devemos observar, no entanto, que os centros de classe representam números diferentes de observações, não podendo participar da média com o mesmo peso. Assim, a média da distribuição será a média ponderada (pelas frequências absolutas) dos centros de classe, que é definida por

$$\overline{x}_p = \frac{\sum_{j=1}^{k} c_j F_j}{\sum_{j=1}^{k} F_j} = \frac{\sum_{j=1}^{k} c_j F_j}{n}.$$

O valor da média de uma distribuição é obtido com um erro provocado pelo agrupamento dos dados. Esse erro é tanto menor quanto maior for a simetria dos valores de cada classe em relação ao seu centro ou ponto médio (c_j). Entretanto, nas distribuições discretas, como a da Tabela 2.12, tal erro não é cometido, pois não existe representação pelo centro de classe.

– Classe mediana e classe modal

Embora existam expressões para o cálculo aproximado da mediana e da moda a partir de dados agrupados em classe, aqui nos interessará apenas determinar a classe mediana e a classe modal.

A *classe mediana* é aquela onde está compreendida a mediana. Esta classe é a primeira cuja frequência absoluta acumulada (F'_j) é maior ou igual ao valor de p (posição da mediana). A posição da mediana, como já vimos anteriormente, é obtida através da expressão

$$p = \frac{n+1}{2}.$$

A *classe modal* é aquela que possui a maior frequência absoluta, mas não é, necessariamente, a classe que compreende a moda do conjunto de valores.

◆ **Medidas de variação ou dispersão**

– Variância

Devido à inexistência de valores individuais na distribuição de frequências, devemos utilizar para o cálculo da variância a seguinte expressão

$$s^2 = \frac{\sum F_j \left(c_j - \overline{x}_p\right)^2}{n-1}.$$

A variância pode ser entendida como uma medida da extensão de um histograma ou de um polígono de frequências sobre o eixo horizontal.

– Desvio padrão e coeficiente de variação

O desvio padrão e o coeficiente de variação para dados agrupados são obtidos da mesma forma que para dados não agrupados. Assim, temos

$$s = \sqrt{s^2} \qquad \text{e} \qquad CV = \frac{s}{\overline{x}_p} 100 .$$

◆ **Medidas de formato**

As expressões que definem o *coeficiente de assimetria* e o *coeficiente de curtose*, também permanecem as mesmas que para os dados não agrupados, respectivamente,

$$a_3 = \frac{m_3}{m_2 \sqrt{m_2}} \qquad \text{e} \qquad a_4 = \frac{m_4}{m_2^2}.$$

Porém, os *momentos centrados da média*, utilizados no cálculo desses coeficientes, pelas mesmas razões já mencionadas para a variância e para a média, são assim definidos

$$m_2 = \frac{\sum F_j (c_j - \overline{x}_p)^2}{n}, \quad m_3 = \frac{\sum F_j (c_j - \overline{x}_p)^3}{n} \quad \text{e} \quad m_4 = \frac{\sum F_j (c_j - \overline{x}_p)^4}{n}.$$

Vamos utilizar a distribuição de frequências apresentada na Tabela 2.13, para exemplificar o cálculo das medidas descritivas a partir de dados agrupados em classe.

Para facilitar a obtenção destas medidas, convém utilizar a tabela auxiliar abaixo que inclui todos os cálculos intermediários necessários.

j	Classes	c_j	F_j	F_j'	$c_j F_j$	$F_j(c_j - \overline{x}_p)^2$	$F_j(c_j - \overline{x}_p)^3$	$F_j(c_j - \overline{x}_p)^4$
1	16 ⊢ 19,3	17,65	7	7	123,55	496,27	-4.178,63	35.184,10
2	19,3 ⊢ 22,6	20,95	9	16	188,55	235,93	-1.207,96	6.184,75
3	22,6 ⊢ 25,9	24,25	15	31	363,75	49,69	-90,43	164,58
4	25,9 ⊢ 29,3	27,55	12	43	330,60	26,28	38,90	57,57
5	29,2 ⊢ 32,5	30,85	9	52	277,65	205,64	982,94	4.698,44
6	32,5 ⊢ 35,8	34,15	6	58	204,90	391,72	3.165,08	25.573,88
7	35,8 ⊢⊣ 39,1	37,45	2	60	74,90	259,01	2.947,52	33.542,78
	Σ	–	60	–	1.563,9	1.664,54	1.657,42	105.406,11

A partir dos totais da última linha da tabela, podemos facilmente calcular as medidas. Assim temos:

Média aritmética: $\overline{x}_p = \frac{\sum_{j=1}^{k} c_j F_j}{n} = \frac{1563,9}{60} = 26,07\,\text{kg}$

Como n é par, existem duas posições centrais no conjunto. Sendo

$$p = \frac{n+1}{2} = \frac{60+1}{2} = 30,5\,,$$

as posições p_1 e p_2 são os inteiros mais próximos de 30,5, ou seja, 30 e 31, respectivamente. A primeira classe a apresentar freqüência absoluta acumulada igual à posição (de maior valor, no caso de n par) da mediana é a terceira, $F'_j = 31$, significando que os valores que ocupam as posições 30 e 31 pertencem a esta classe. Portanto, a classe mediana é a terceira.

A classe com maior frequência absoluta também é a terceira, $F_3 = 15$. Assim, a classe modal é a terceira.

Variância: $s^2 = \dfrac{\sum F_j \left(c_j - \overline{x}_p\right)^2}{n-1} = \dfrac{1664,54\,kg^2}{60-1} = 28,21kg^2$

Desvio padrão: $s = \sqrt{s^2} = \sqrt{28,21kg^2} = 5,331kg$

Coeficiente de variação: $CV = \dfrac{s}{\overline{x}_p}100 = \dfrac{5,331kg}{26,07\,kg}100 = 20,37\%$

Momentos centrados na média:

$$m_2 = \frac{\sum F_j (c_j - \overline{x}_p)^2}{n} = \frac{1664,54\,kg^2}{60} = 27,74\,kg^2$$

$$m_3 = \frac{\sum F_j (c_j - \overline{x}_p)^3}{n} = \frac{1657,42\,kg^3}{60} = 27,62\,kg^3$$

$$m_4 = \frac{\sum F_j (c_j - \overline{x}_p)^4}{n} = \frac{105406,11kg^4}{60} = 1756,77\,kg^4$$

Coeficientes de assimetria e curtose:

$a_3 = \dfrac{m_3}{m_2\sqrt{m_2}} = \dfrac{27,62\,kg^3}{27,74\,kg^2\sqrt{27,74\,kg^2}} = 0,189 \rightarrow$ indica que a distribuição é simétrica

$a_4 = \dfrac{m_4}{m_2^2} = \dfrac{1756,77\ kg^4}{\left(27,74\ kg^2\right)^2} = 2,283 \rightarrow$ indica que a distribuição é platicúrtica

Devemos salientar que as medidas para dados agrupados em classe vêm sendo cada vez menos utilizadas. A obtenção de medidas descritivas a partir de distribuições de frequências tem como principal objetivo facilitar o processo de cálculo, pois, quando se trata de conjuntos de dados muito grandes, essa tarefa é bastante trabalhosa. Outra razão que justifica o uso dessas medidas é a falta de acesso aos dados originais (não agrupados). Contudo, sabe-se que medidas obtidas a partir de dados agrupados em classe, na maioria das vezes,

não são exatas. Com o advento da computação e o desenvolvimento de programas estatísticos, o problema da dificuldade no processo de cálculo foi superado, uma vez que estes programas executam cálculos trabalhosos com rapidez e exatidão. Sendo assim, não havendo mais a dificuldade para a obtenção das medidas exatas, não há razão para continuarmos utilizando as medidas aproximadas.

No quadro abaixo podemos observar os valores das medidas calculadas a partir dos dados não agrupados e a partir da tabela de distribuição de frequências.

Dados não agrupados	Dados agrupados em classe
$\bar{x}$ = 25,78 kg	$\bar{x}$ = 26,07 kg
Md = 25 kg	Classe mediana: [22,6 ; 25,9)
Mo = 23 kg	Classe modal: [22,6 ; 25,9)
s^2 = 28,64 kg^2	s^2 = 28,21 kg^2
s = 5,352 kg	s = 5,331 kg
CV = 20,76%	CV = 20,37%
m_2 = 28,16 kg^2	m_2 = 27,74 kg^2
m_3 = 32,48 kg^3	m_3 = 27,62 kg^3
m_4 = 1.875,62 kg^4	m_4 = 1.756,77 kg^4
a_3 = 0,218	a_3 = 0,189
a_4 = 2,365	a_4 = 2,283

Comparando os valores das medidas calculadas pelos dois processos, podemos verificar que as medidas obtidas dos dados agrupados em classe são aproximações daquelas obtidas a partir dos dados não agrupados, que são as medidas exatas. Tanto maior será esta aproximação, quanto mais simétrica for a distribuição dos valores dentro dos intervalos de classe.

◆ Medida descritiva e escala de medida

Algumas medidas descritivas exigem uma escala de medida mínima para serem obtidas. A tabela relaciona a escala necessária para algumas medidas.

Medida	Escala mínima
Moda	Escala nominal
Percentis	Escala ordinal
Média aritmética	Escala intervalar

Exercícios propostos:

2.5. Os valores que seguem são os tempos (em segundos) de reação a um alarme de incêndio, após a liberação de fumaça de uma fonte fixa:

12 9 11 7 9 14 6 10

Calcule as medidas de localização (média, mediana e moda) e as medidas de variação (amplitude total, variância, desvio padrão e coeficiente de variação) para o conjunto de dados.

2.6. Foram registrados os tempos de frenagem para 21 motoristas que dirigiam a 30 milhas por hora. Os valores obtidos foram:

69	58	70	80	46	61	65	74	75	55	67
56	70	72	61	66	58	68	70	68	58	

Para este conjunto de valores, calcule os quartis e a amplitude interquartílica e interprete esses valores.

2.7. Calcule as medidas descritivas para o conjunto de dados referente ao número de pães não vendidos em uma certa padaria até a hora do encerramento do expediente

j	Classes	F_j	F_j'	c_jF_j	$F_j(c_j - \overline{x}_p)^2$	$F_j(c_j - \overline{x}_p)^3$	$F_j(c_j - \overline{x}_p)^4$
1	0	20					
2	1	7					
3	2	7					
4	3	3					
5	4	2					
6	5	1					
	Σ	40	–				

2.8. Calcule as medidas descritivas para o conjunto de dados agrupados em classes, apresentado na tabela abaixo.

Frequência do valor gasto (em reais) pelas primeiras 50 pessoas que entraram em um determinado supermercado, no dia 01/01/2000.

j	Classes	c_j	F_j	F_j'	c_jF_j	$F_j(c_j - \overline{x}_p)^2$	$F_j(c_j - \overline{x}_p)^3$	$F_j(c_j - \overline{x}_p)^4$
1	3,11 ├— 16,00		8					
2	16,00 ├— 28,89		20					
3	28,89 ├— 41,78		6					
4	41,78 ├— 54,67		8					
5	54,67 ├— 67,56		3					
6	67,56 ├— 80,45		1					
7	80,45 ├—┤ 93,34		4					
	Σ	–	50	–				

2.4. Análise exploratória de dados

Vimos que a média aritmética e a variância, por serem medidas de fácil compreensão e apresentarem boas propriedades matemáticas e estatísticas, são muito utilizadas para representar, respectivamente, a tendência central e a dispersão de um conjunto de valores. Entretanto, é importante destacar que essas medidas descrevem de forma ótima apenas as distribuições de frequências unimodais, simétricas e mesocúrticas. Podemos citar pelo menos uma limitação importante do uso indiscriminado da média e da variância na descrição de um conjunto de dados. Sabemos que essas duas medidas são pouco resistentes; portanto, numa distribuição assimétrica, seus valores seriam bastante afetados pelos valores discrepantes.

John Tukey, em 1970, propôs algumas técnicas que, dentre outras vantagens, contornavam esse problema advindo do uso da média e da variância para descrever distribuições assimétricas. O conjunto dessas técnicas, denominado Análise Exploratória de Dados, não só constituiu um complemento às técnicas estatísticas clássicas, como foi também uma valiosa alternativa para descrever dados que não seguem o modelo unimodal, simétrico e mesocúrtico. O enfoque proposto pela Análise Exploratória de Dados pretende obter medidas resistentes e robustas.

Vimos que medidas *resistentes* são aquelas que se mostram pouco sensíveis à presença de valores anômalos (discrepantes do núcleo central da distribuição). Uma medida resistente mostrará poucas variações diante da substituição dos valores originais por outros muito diferentes, devido a sua focalização na parte central ou relativamente agrupada da distribuição. Dentre as medidas resistentes, o enfoque clássico tem a mediana como principal exemplo. São denominadas medidas *robustas* aquelas que apresentam pouca sensibilidade diante dos desvios aos pressupostos básicos inerentes aos modelos probabilísticos, como acontece com relação à forma da distribuição, por exemplo.

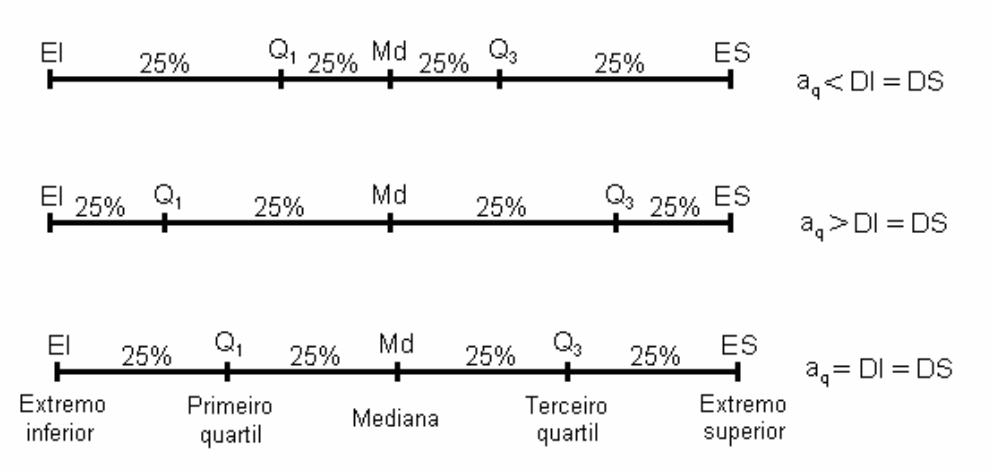
As técnicas exploratórias ajudam a comprovar as condições de aplicação dos testes de hipóteses (que serão vistos, mais adiante, na Inferência Estatística), a detectar erros ou valores discrepantes, a buscar a melhor transformação de dados quando houver necessidade, etc. Em geral, dão uma visão distinta, prévia, mas complementar às técnicas de Inferência, também chamadas de confirmatórias. Tudo isso repercute em melhor qualidade da análise de dados.

Nosso objetivo aqui é apresentar três dessas técnicas: o resumo de cinco números, o gráfico em caixa (“box plot”) e o diagrama de ramos e folhas. O gráfico em caixa, além de representar os dados dando uma ideia precisa do formato da distribuição, ainda permite a identificação de valores discrepantes.

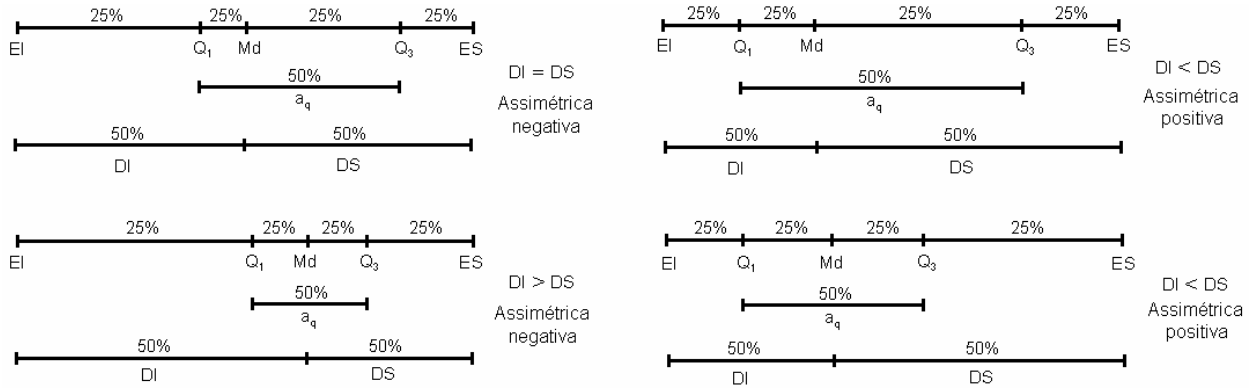
◆ Resumo de cinco números

O resumo de cinco números descreve o conjunto de dados através de cinco valores: a mediana (Md), os quartis, primeiro (Q_1) e terceiro (Q_3), e os extremos, inferior (EI) e superior (ES). A partir desses valores, podemos calcular: a amplitude interquartílica (a_q), obtida pela diferença entre os quartis; a dispersão inferior (DI), obtida pela diferença entre a mediana e o extremo inferior; e a dispersão superior (DS), diferença entre o extremo superior e a mediana. O resumo de cinco números fornece uma ideia acerca da simetria da distribuição porque o percentual de observações compreendido dentro de cada um desses intervalos é conhecido (25%).

Assim, se a diferença entre o primeiro quartil e extremo inferior é aproximadamente igual à diferença entre o extremo superior e o terceiro quartil ($Q_1 - EI \cong ES - Q_3$) e a diferença entre a mediana e o primeiro quartil é aproximadamente igual à diferença entre o terceiro quartil e a mediana ($Md - Q_1 \cong Q_3 - Md$), a distribuição é considerada simétrica. Na figura abaixo podemos observar alguns casos simétricos.



Se uma dessas duas condições não for atendida, então, a distribuição será assimétrica. Por exemplo, se a dispersão superior for muito maior que a dispersão inferior, teremos uma distribuição assimétrica positiva, indicando que a maior concentração de valores está entre o extremo inferior e a mediana, ou ainda que os valores que se localizam abaixo da mediana são mais homogêneos do que aqueles que se localizam acima dela. Se, em caso contrário, a dispersão inferior for maior que a dispersão superior, teremos uma distribuição assimétrica negativa. Na figura abaixo podemos observar alguns exemplos de distribuições assimétricas.

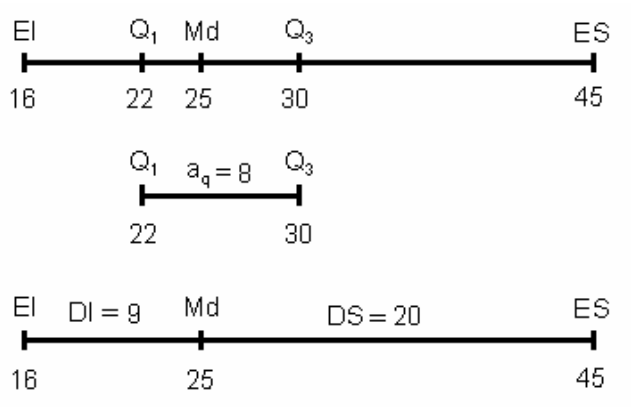


Consideremos agora um exemplo resolvido:

Os dados abaixo se referem aos pesos ao nascer (em kg) de 61 bovinos machos da raça Ibagé.

16, 17, 17, 18, 18, 18, 19, 20, 20, 20, 20,
20, 21, 21, 22, 22, 23, 23, 23, 23, 23, 23,
23, 23, 23, 25, 25, 25, 25, 25, 25, 26, 26,
27, 27, 27, 27, 28, 28, 28, 29, 29, 29, 30,
30, 30, 30, 30, 30, 30, 31, 32, 33, 33, 33,
34, 34, 35, 36, 39, 45

Para esses dados vamos obter o esquema de cinco números, a amplitude interquartílica e a dispersão inferior e a dispersão superior.



O resumo de cinco números permite verificar que a distribuição não é simétrica, pois as distâncias entre esses valores são diferentes.

Veremos a seguir que, através dos quartis e da amplitude interquartílica, também é possível identificar a presença de valores discrepantes no conjunto de dados.

Identificação de valores discrepantes

Um critério objetivo para a identificação de valores discrepantes num conjunto de dados utiliza duas medidas denominadas cerca inferior (CI) e cerca superior (CS). A cerca inferior é calculada subtraindo-se do primeiro quartil uma e meia amplitude interquartílica, e a cerca superior, somando-se esta mesma quantidade ao terceiro quartil. Assim, temos:

$$CI = Q_1 - 1{,}5a_q \quad \text{e} \quad CS = Q_3 + 1{,}5a_q$$

São considerados discrepantes os valores que estivem fora do seguinte intervalo:

$$\left[Q_1 - 1{,}5a_q ;\, Q_3 + 1{,}5a_q \right] .$$

Valores menores que a cerca inferior são denominados *discrepantes inferiores* e os valores maiores que a cerca superior são os *discrepantes superiores*.

No exemplo, serão considerados discrepantes os valores que estiverem fora dos limites da cerca superior e da cerca inferior:

$$CI = Q_1 - 1{,}5a_q = 22 - 1{,}5 \times 8 = 10$$

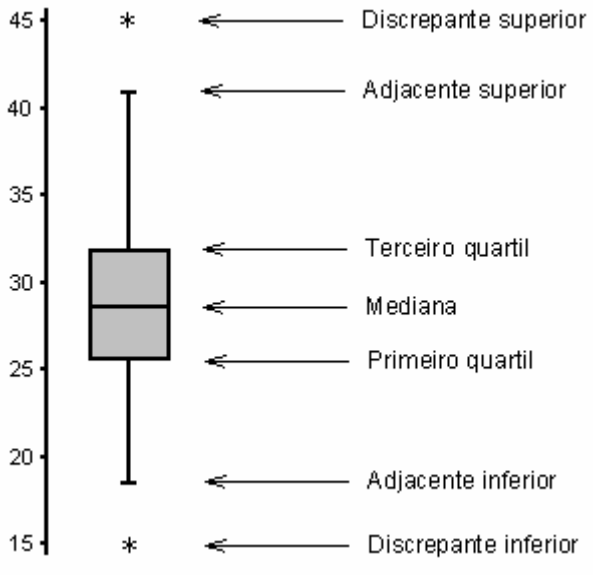
$$CS = Q_3 + 1{,}5a_q = 30 + 1{,}5 \times 8 = 42$$

Verificamos que o valor 45 ultrapassa a cerca superior, portanto, é classificado como discrepante superior.

- **Gráfico em caixa (box plot)**

A informação dada pelo resumo de cinco números pode ser apresentada em forma de um gráfico em caixa, que agrega uma série de informações a respeito da distribuição, tais como localização, dispersão, assimetria, caudas e dados discrepantes. Antes de construir o gráfico precisamos definir o que são valores adjacentes. São adjacentes o menor e o maior valores não discrepantes de um conjunto de dados, ou seja, o maior valor que não ultrapassa a cerca superior e o menor valor que não ultrapassa a cerca inferior. Se num conjunto de dados

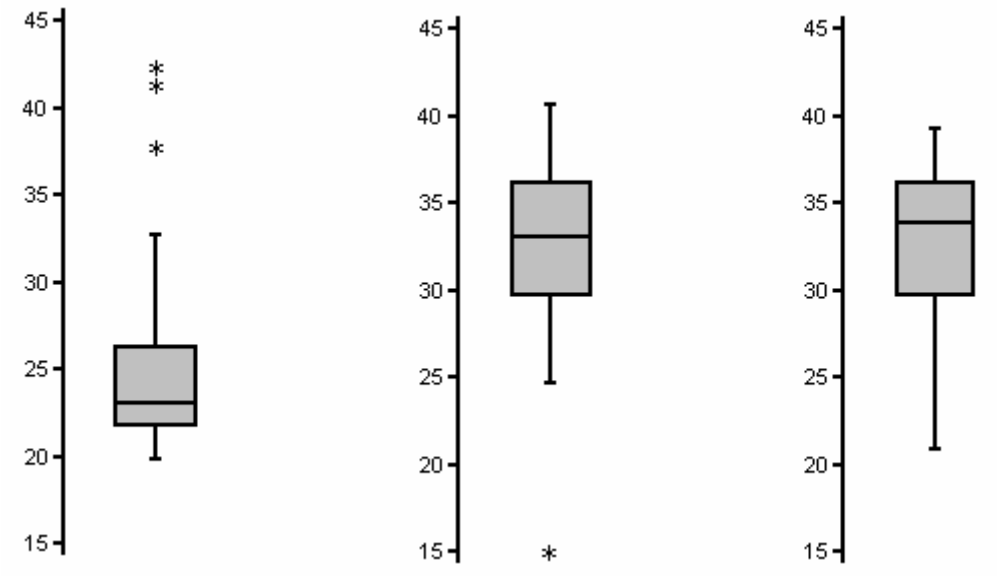
nenhum valor é considerado discrepante, os valores adjacentes são os próprios extremos. Para construir o gráfico em caixa, consideraremos um retângulo onde estarão representados os quartis e a mediana. A partir do retângulo, para cima e para baixo, seguem linhas, denominadas bigodes, que vão até os valores adjacentes. Os valores discrepantes recebem uma representação individual através de uma letra ou um símbolo. Assim, obtemos uma figura que representa muitos aspectos relevantes de um conjunto de dados, como podemos observar na ilustração abaixo.



A posição central dos valores é dada pela mediana e a dispersão pela amplitude interquartílica (a_q). As posições relativas da mediana e dos quartis e o formato dos bigodes dão uma noção da simetria e do tamanho das caudas da distribuição.

Na figura abaixo podemos observar o gráfico em caixa representando diferentes tipos de distribuições:

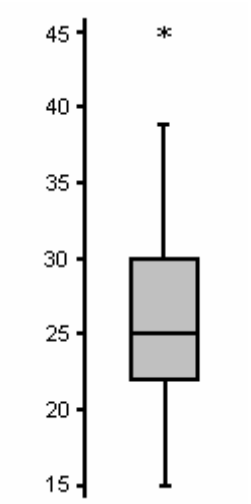
a) distribuição assimétrica positiva, com três valores discrepantes superiores;
b) distribuição simétrica, com um valor discrepante inferior;
c) distribuição assimétrica negativa, sem valores discrepantes.



a) assimétrica positiva b) simétrica c) assimétrica negativa

Vale lembrar que quando encontramos um valor discrepante num conjunto de dados, a sua origem deve ser investigada. Muitas vezes, os valores discrepantes, de fato, fazem parte do conjunto de dados, reforçando a característica assimétrica da distribuição. Mas, eventualmente, estes valores podem ser oriundos de erros na aferição ou no registro dos dados. Em geral, distribuições com caudas longas (indicadas por bigodes longos no gráfico), característica comum de distribuições assimétricas, apresentam uma tendência maior de produzir valores discrepantes. Nas figuras acima, os bigodes de diferentes tamanhos indicam distribuições assimétricas. O valor discrepante parece ser uma anomalia maior na figura b, pois se trata de uma distribuição simétrica e com caudas curtas. De qualquer modo, uma cuidadosa inspeção nos dados e nas eventuais causas da ocorrência desse(s) valor(es) é sempre uma providência necessária antes que qualquer atitude seja tomada em relação a esses dados.

A seguir temos o gráfico em caixa representando o conjunto de dados do exemplo, que se refere aos pesos ao nascer de bovinos machos da raça Ibagé.



- **Diagrama de ramo e folhas**

Trata-se de uma ferramenta exploratória útil para descrever pequenos conjuntos de dados. O método fornece uma boa visão geral dos dados sem que haja uma perda de informação detectável. Cada valor retém sua identidade e a única informação perdida é a ordem em que foram obtidos os dados. Eventualmente, alguns algarismos podem ser desprezados para facilitar a representação do conjunto.

O diagrama de ramos e folhas é um procedimento alternativo para resumir um conjunto de valores, que fornece uma ideia da forma de sua distribuição, semelhante a um histograma. Este gráfico é uma boa opção quando temos em mãos somente os dados, caneta e papel.

Para ilustrar a montagem do diagrama de ramo e folhas, consideremos os seguintes dados relativos às notas de 40 alunos em uma prova de Estatística.

78	59	86	94	43	56	78	84
57	49	96	68	67	65	75	73
67	87	84	45	56	94	87	56
85	76	86	79	78	77	59	76
68	49	86	87	83	94	85	96

O primeiro passo é a separação dos dados, combinando todos os valores que começam com 4, todos que começam com 5, todos que começam com 6, e assim por diante. Assim, temos

43 49 45
59 56 57 56 56 59
68 67 65 67 68
78 78 75 73 76 79 78 77 76
86 84 89 87 84 87 85 86 86 87 83 85
94 96 94 94 96

Esse arranjo já é bastante informativo, mas não é o tipo de diagrama utilizado na prática. Para simplificar ainda mais, mostramos o primeiro dígito uma vez para cada linha, à esquerda e separando dos outros dígitos por meio de uma linha vertical. Assim, temos

4	3 5 9
5	9 6 7 6 6 9
6	8 7 5 7 8
7	8 8 5 3 6 9 8 7 6
8	6 4 9 7 4 7 5 6 6 7 3 5
9	4 6 4 4 6

Isso é o que denominamos *diagrama de ramo e folhas*. Nesse arranjo, cada linha é denominada *ramo*, cada número no ramo à esquerda da linha vertical é chamado *rótulo do ramo* e cada número à direita da linha vertical é denominado *folha*. É bastante interessante que as folhas do diagrama sejam ordenadas facilitando ainda mais a interpretação. Dessa forma, nosso diagrama resulta assim:

4	3 5 9
5	6 6 6 7 9 9
6	5 7 7 8 8
7	3 5 6 6 7 8 8 8 9
8	3 4 4 5 5 6 6 6 7 7 7 9
9	4 4 4 6 6

Existem várias maneiras de organizar um diagrama de ramo e folhas. Por exemplo, os rótulos dos ramos ou as folhas poderiam ser de dois dígitos, como por exemplo, o conjunto 240, 242, 245, 248 e 249, sendo representado de duas formas:

24| 0 2 5 8 9

ou

2| 40 42 45 48 49.

Em casos de muitos valores pode ser necessário obter mais ramos, repetindo cada rótulo de ramo, por exemplo, duas vezes, sendo o primeiro com as folhas de 0 a 4 e o segundo com as folhas de 5 a 9. Esse tipo de diagrama é chamado diagrama de ramos duplos. Um diagrama de ramo e folhas pode ainda ser complementando com informações adicionais, como o número de observações em cada ramo.

Exercícios propostos:

2.9. Os dados abaixo se referem aos valores gastos (em reais) pelas primeiras 50 pessoas que entraram em um determinado Supermercado, no dia 01/01/2000.

3,11	8,88	9,26	10,81	12,69	13,78	15,23	15,62	17,00	17,39
18,36	18,43	19,27	19,50	19,54	20,16	20,59	22,22	23,04	24,47
24,58	25,13	26,24	26,26	27,65	28,06	28,08	28,38	32,03	36,37
38,98	38,64	39,16	41,02	42,97	44,08	44,67	45,40	46,69	48,65
50,39	52,75	54,80	59,07	61,22	70,32	82,70	85,76	86,37	93,34

Para esses dados:

a) Obtenha o resumo de cinco números.

b) Verifique se existem valores discrepantes.

c) Construa o gráfico em caixa.

d) Com base no gráfico, classifique a distribuição quanto à simetria. Justifique sua resposta.

2.10. As durações (em horas de uso contínuo) de 25 componentes eletrônicos selecionados de um lote de produção são:

834, 919, 784, 865, 839, 912, 888, 783, 655,
831, 886, 842, 760, 854, 939, 961, 826, 954,
866, 675, 760, 865, 901, 632, 718.

Construa um diagrama de ramo e folhas com rótulos de ramos com um dígito e folhas de dois dígitos. Use esse diagrama de ramo e folhas para decidir sobre a simetria desses dados.

2.5. Bibliografia

ANDRES, A.M., CASTILLO, J. de D.L. del **Bioestadistica para las Ciências de la Salud.** Madrid: Ediciones Norma, 1988. 614 p.

BOTELHO, E.M.D., MACIEL, A.J. **Estatística Descritiva (Um Curso Introdutório)** Viçosa: Universidade Federal de Viçosa, 1992. 65p.

COSTA, S.F. **Introdução Ilustrada à Estatística (com muito humor!)**. 2.ed., São Paulo: Harbra, 1992. 303p.

FARIA, E.S. de **Estatística** Edição 97/1. (Apostila)

FERREIRA, D.F. **Estatística Básica**. Lavras: Editora UFLA, 2005, 664p.

FREUND, J.E., SIMON, G.A. **Estatística Aplicada. Economia, Administração e Contabilidade**. 9.ed., Porto Alegre: Bookman, 2000. 404p.

PIMENTEL GOMES, F. **Iniciação à Estatística** São Paulo: Nobel, 1978. 211p.

SILVEIRA JÚNIOR, P., MACHADO, A.A., ZONTA, E.P., SILVA, J.B. da **Curso de Estatística** v.1, Pelotas: Universidade Federal de Pelotas, 1989. 135p.

SPIEGEL, M.R. **Estatística** São Paulo: McGraw-Hill, 1972. 520p.

Sistema Galileu de Educação Estatística. Disponível em: <http://www.galileu.esalq.usp.br>

Unidade III
Elementos de Probabilidade

3.1. Introdução à teoria das probabilidades

3.1.1. Introdução

A Estatística, desde as suas origens (antigo Egito – 2000 anos a.C.) até meados do século XIX, se preocupava apenas com a organização e a apresentação de dados de observação coletados empiricamente (Estatística Descritiva).

Somente com o desenvolvimento da teoria das probabilidades foi possível que a Estatística se estruturasse organicamente e ampliasse seu campo de ação, através da criação de técnicas de amostragem mais adequadas e de formas de relacionar as amostras com as populações de onde provieram (Inferência Estatística).

A probabilidade é uma área relativamente nova da matemática (considerando a idade da matemática) que tem como finalidade a modelagem de *fenômenos aleatórios*. Modelar significa conhecer matematicamente. Uma das funções da matemática é a criação de modelos que possibilitem o estudo dos fenômenos da natureza. Ao estudar um fenômeno, temos sempre o interesse de tornar a sua investigação mais precisa e, para isso, tentamos formular um modelo matemático que melhor o explique.

Na formulação do modelo matemático mais adequado deve-se levar em conta que certos pormenores sejam desprezados com o objetivo de simplificar o modelo. Deste modo, tanto maior será a representatividade do modelo quanto menor foi a importância destes detalhes na elucidação do fenômeno considerado.

A verificação da adequação do modelo escolhido não pode ser feita sem que alguns dados de observação sejam obtidos. Através da comparação dos resultados previstos pelo modelo com um determinado número de valores observados, poderemos concluir se o modelo é ou não adequado para explicar o fenômeno em estudo.

Dependendo do fenômeno que está sendo estudado, os modelos matemáticos podem ser de dois tipos:

a) Modelo determinístico: é aquele em que ao conhecer as variáveis de entrada, ou seja, as condições do experimento, é possível determinar as variáveis de saída, isto é, os seus resultados. Para os fenômenos determinísticos existe a *certeza* do resultado que ocorrerá. Na física clássica, a maioria dos fenômenos estudados são determinísticos.

Exemplo: Se o deslocamento de um objeto é definido pela expressão $s = v\,t$ e são conhecidos os valores de v (velocidade) e t (tempo), então o valor de s fica implicitamente determinado.

b) Modelo estocástico, probabilístico ou aleatório: é aquele em que, mesmo conhecendo as condições do experimento, não é possível determinar o seu resultado final. Neste modelo, é introduzido um componente aleatório e só é possível determinar a *chance* de ocorrência de um resultado. Na biologia, os fenômenos são probabilísticos.

Exemplo: O nascimento de um bovino. Não é possível determinar o sexo do recém nascido, somente a sua probabilidade de ocorrência: 0,5 para fêmea e 0,5 para macho.

A modelagem de um experimento aleatório implica em responder três questões fundamentais:

- Quais as possíveis formas de ocorrência?
- Quais são as chances de cada ocorrência?
- De que forma se pode calcular isso?

Um pouco de história...

Blaise Pascal
(1623 -1662)

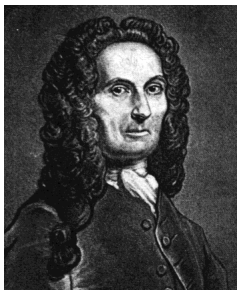
O estudo das probabilidades teve suas origens no século XVII, a partir do interesse de dois matemáticos franceses, *Pascal* e *Fermat*, em resolver problemas relacionados com jogos de azar, que lhes eram propostos pelo nobre francês Cavalheiro de Mère.

Data de 1713, entretanto, o primeiro grande tratado nesse campo escrito por *Jacques Bernoulli* denominado *Ars Conjectandi* (Arte das Conjecturas). Bernoulli exemplificou seu trabalho principalmente em termos de jogos de azar.

Pierre Fermat
(1601 -1665)

Na mesma linha seguiram alguns trabalhos subsequentes, com destaque para a obra de *Abraham de Moivre* iniciada, em 1718, com *Doctrine of Changes*. A famosa distribuição normal foi deduzida pelo francês de Moivre como resultado do limite da expansão do binômio $(a + b)^n$, embora ainda não tivesse sido colocada como uma distribuição de probabilidade. Mais tarde, essa relação veio a ter uma importância muito grande por estabelecer a aproximação da distribuição binomial em relação à normal.

O extraordinário avanço das probabilidades, entretanto, deu-se no início do século XIX, através do matemático *Laplace* com seu clássico *Theórie analytique des probabilités* (1812), onde mostra que a área sob a curva normal é $\sqrt{\pi}$ e fornece uma prova formal sobre o método dos quadrados mínimos.



Abraham de Moivre
(1667 -1754)

Praticamente no mesmo período, o matemático e astrônomo alemão *Gauss*, chegou aos mesmos resultados sobre a curva normal de probabilidades, estudando a distribuição dos erros de medida. Mas somente no século XX é que se desenvolveu uma teoria matemática rigorosa baseada em axiomas, definições e teoremas.

As distribuições de probabilidades são consideradas hoje a espinha dorsal da teoria estatística, pois todos os processos de inferência são aplicações de distribuições de probabilidades.

Assim, o conhecimento dos conceitos advindos da teoria das probabilidades é de grande importância para a correta utilização da estatística.

Carl Friedrich Gauss
(1777 -1855)

"Blaise Pascal nunca frequentou escola, foi educado pelo pai que era matemático. Quando Blaise estava com 11 anos descobriu sozinho que a soma dos ângulos de um triângulo era 180 graus! A 32ª proposição de Euclides! E ele jamais tinha ouvido falar em Euclides, pois seu pai lhe havia escondido, temendo que a geometria fosse lhe cansar a cabeça. Quando o pai ficou sabendo o que o filho acabava de (re)descobrir, chorou de alegria e ficou tão contente que lhe deu de presente os treze livros dos Elementos de Euclides." (Guedj, 2000)

"Com o intuito de ajudar o pai que, além de matemático, também era cobrador de impostos e tinha muitas contas a fazer, Blaise Pascal inventou uma máquina de calcular, a pascaline. O principal problema do cálculo mecânico é o que fazer quando, chegando a nove, se acrescenta um. Pascal criou um pequeno mecanismo em que ninguém tinha pensado antes dele, um 'transportador' que transportava automaticamente este número. Por conta da engenhoca, que na época era chamada de 'máquina aritmética', Pascal tinha se tornando um pequeno empresário. Havia montado uma empresa, feito o projeto de sua máquina, contratado operários, patenteado o processo e fabricado umas cinquenta pascalines. Produção em série, vendida a cem libras cada máquina. Ganhou um dinheirão!" (Guedj, 2000)

3.1.2. Conceitos fundamentais

◆ **Experimento probabilístico ou aleatório**: é toda experiência cujos resultados podem não ser os mesmos, ainda que sejam repetidos sob condições idênticas. Características desses experimentos:

– cada experimento pode ser repetido indefinidamente sob condições inalteradas;

– embora não possamos afirmar que resultado ocorrerá, é sempre possível descrever o conjunto de todos os possíveis resultados.

– quando o experimento for realizado repetidamente, os resultados individuais parecem ocorrer de forma acidental; mas se for repetido um grande número de vezes uma configuração definida ou regularidade surgirá.

Exemplos:

Experimento 1: Jogar um dado e observar a sua face superior.
Experimento 2: Lançar uma moeda até que apareça cara e contar o número de lançamentos.
Experimento 3: Selecionar uma carta do baralho e anotar o seu valor e naipe.
Experimento 4: Acender uma lâmpada e medir o tempo até que ela se apague.

◆ **Espaço amostral** (*S*)**:** é o conjunto de todos os possíveis resultados de um experimento aleatório, ou seja, é o conjunto universo relativo aos resultados de um experimento. A cada experimento aleatório está associado um conjunto de resultados possíveis ou espaço amostral.

Exemplos:

$S_1 = \{1, 2, 3, 4, 5, 6\} \rightarrow$ enumerável e finito
$S_2 = \{1, 2, 3, 4, ...\} \rightarrow$ enumerável e infinito
S_3 = {ás de ouro, ..., rei de ouro, ás de paus, ..., rei de paus, ..., ás de espada, ..., rei de espada, ás de copas, ..., rei de copas} $\rightarrow$ enumerável e finito
$S_4 = \{t; t \geq 0\} \rightarrow$ contínuo e infinito

◆ **Evento ou ocorrência**: é todo conjunto particular de resultados de *S* ou, ainda, todo subconjunto de *S*. Geralmente é designado por uma letra maiúscula (A, B, C). A todo evento será possível associar uma probabilidade.

Exemplo:

Se *S* = {1, 2, 3, 4, 5, 6}, então

A = {1, 2, 3},
B = Ocorrência de faces pares,
C = {5}, são eventos de *S*.

◆ **Operações com eventos**

Como o espaço amostral *S* e os eventos são conjuntos, as mesmas operações realizadas com conjuntos são válidas para eventos.

Exemplo: Se A e B, são eventos de *S*, então:

Ocorre $A \cup B$, se ocorrer A ou B (ou ambos).
Ocorre $A \cap B$, se ocorrer A e B.
Ocorre $\overline{A}$, se ocorrer *S*, mas não ocorrer A.
Ocorre $A - B$, se ocorrer A, mas não ocorrer B.

◆ **Ponto amostral**: é qualquer resultado particular de um experimento aleatório. Todo espaço amostral e todo evento são constituídos por pontos amostrais.

◆ **Eventos especiais**

Evento impossível: é aquele evento que nunca irá ocorrer, é também conhecido como o conjunto vazio ($\varnothing$). É um evento porque é subconjunto de qualquer conjunto, portanto é subconjunto de S ($\varnothing \subset S$).

Exemplo: $A_1 = \{(x, y); x^2 + y^2 < 0\}$

Evento certo: é aquele evento que ocorre toda vez que se realiza o experimento, portanto, esse evento é o próprio S. É evento porque todo conjunto é subconjunto de si mesmo ($S \subset S$).

Exemplo: $A_2 = \{(x, y); x^2 + y^2 \geq 0\}$

Eventos mutuamente exclusivos

Dois eventos A e B associados a um mesmo espaço amostral S, são mutuamente exclusivos quando a ocorrência de um impede a ocorrência do outro. Na teoria dos conjuntos, correspondem aos conjuntos disjuntos, que não possuem elementos comuns ($A \cap B = \varnothing$).

Exemplos:

Experimento 1: Lançamento de uma moeda e observação do resultado.

S = {c, k}, se definimos
A = Ocorrência de cara A = {c}
B = Ocorrência de coroa B = {k}, então A e B são mutuamente exclusivos.

Experimento 2: Lançamento de um dado e observação da face superior.

S = {1, 2, 3, 4, 5, 6}, se definimos
A = Ocorrência de número ímpar A = {1, 3, 5}
B = Ocorrência de maior do que 4 B = {5, 6}
$A \cap B = \{5\}$, logo, os eventos A e B não são mutuamente exclusivos.

3.1.3. Conceitos de probabilidade

3.1.3.1. Conceito clássico ou probabilidade "a priori"

Pierre-Simon Laplace
(1749 - 1827)

Como a teoria das probabilidades está historicamente ligada aos jogos de azar, esta associação gerou, inicialmente, um conceito chamado conceito clássico ou probabilidade "a priori", devido a Laplace.

Definição: Seja E um experimento aleatório e S o espaço amostral a ele associado, com n pontos amostrais, todos equiprováveis. Se existe, em S, m pontos favoráveis à realização de um evento A, então a probabilidade de A, indicada por P(A), será:

$$P(A) = \frac{m}{n} = \frac{\#A}{\#S}.$$

Notemos, entretanto, que, para que este conceito tenha validade, duas pressuposições básicas devem ser atendidas:

1. O espaço amostral *S* é enumerável e finito.
2. Os elementos do espaço amostral *S* são todos equiprováveis.

Exemplo:

Consideremos o seguinte experimento: lançamento de uma moeda honesta duas vezes e observação do lado superior.

O espaço amostral deste experimento é *S* = {cc, ck, kc, kk} e todos os seus pontos amostrais são equiprováveis:

$$p(cc) = p(kc) = p(ck) = p(kk) = \frac{1}{4}$$

Define-se o evento A = ocorrência de uma cara, então

A = {ck, kc}

e $P(A) = \frac{m}{n} = \frac{\#A}{\#S} = \frac{2}{4} = \frac{1}{2}$,

pois *S* possui quatro pontos amostrais, dos quais dois são favoráveis à ocorrência de A.

Consideremos, agora, outra situação onde o espaço amostral *S* refere-se ao *número de caras* obtido nos dois lançamentos da moeda honesta.

O espaço amostral passa a ser *S* = {0, 1, 2} e o evento A = ocorrência de uma cara, ou seja, A = {1}.

Observamos, nesta situação, que os pontos amostrais de *S* (0, 1 e 2) não são todos equiprováveis, pois

$$p(0) = p(kk) = \frac{1}{4},\ p(1) = p(kc) + p(ck) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2} \text{ e } p(2) = p(cc) = \frac{1}{4}.$$

Portanto, embora o evento A seja o mesmo, como uma das pressuposições básicas não foi atendida, o conceito clássico não pode ser imediatamente aplicado para calcular a sua probabilidade.

Podemos observar, a seguir, que a aplicação do conceito clássico, nesta situação, não conduz ao um resultado incorreto:

$$P(A) = \frac{m}{n} = \frac{\#A}{\#S} = \frac{1}{3}.$$

Recomendamos, então, partir sempre do espaço amostral original do experimento para aplicar o conceito clássico.

3.1.3.2. Frequência relativa ou probabilidade "a posteriori"

Richard Von Mises
(1883 - 1953)

O conceito de frequência relativa como estimativa de probabilidade ou probabilidade "a posteriori" surgiu através de Richard Von Mises.

Definição: Seja E um experimento aleatório e A um evento. Se após n realizações do experimento E (sendo n suficientemente grande), forem observados m resultados favoráveis ao evento A, então uma estimativa da probabilidade P(A) é dada pela frequência relativa

$$f = \frac{m}{n}.$$

Este conceito é baseado no princípio estatístico da estabilidade, ou seja, a medida que o número de repetições do experimento (n) aumenta, a frequência relativa $f = \frac{m}{n}$ se aproxima de P(A). O n deve ser suficientemente grande para que se possa obter um resultado com margem de erro razoável. Define-se o erro desta estimativa pela expressão

$$f - P(A) = \text{erro}.$$

A Figura 3.1 ilustra o princípio da estabilidade, tomando-se por base o número crescente de lançamentos de uma moeda e a probabilidade de se obter cara.

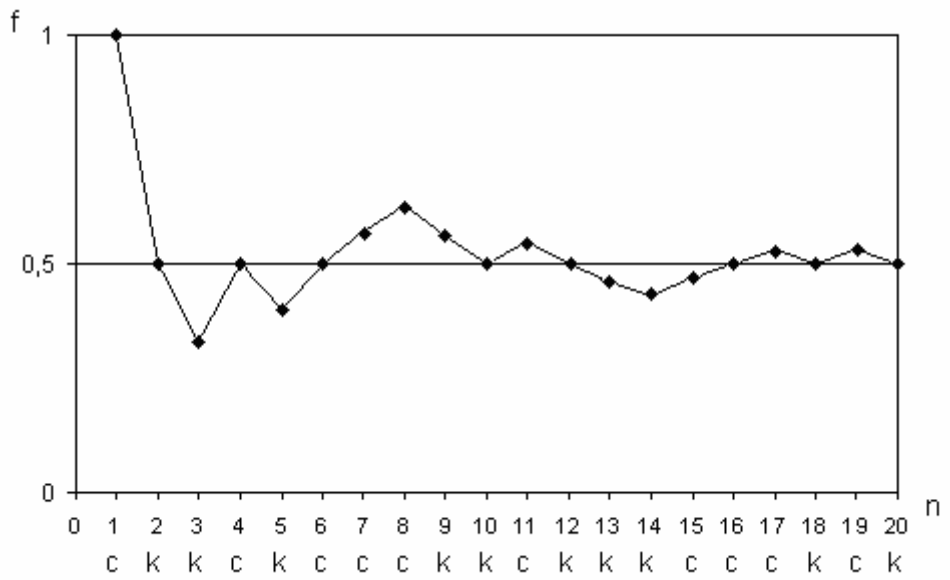


Figura 3.1. Estabilização da frequência relativa *f* quando n cresce.

Exemplo: Em Sobral (CE), observaram-se seis anos de seca no período de 1901-66 (66 anos). Qual é a probabilidade de ser seco o próximo ano?

A frequência relativa f será uma estimativa da probabilidade de ocorrer seca no próximo ano:

$$f = \frac{m}{n} = \frac{6}{66} = \frac{1}{11}$$

3.1.3.3. Conceito moderno ou axiomático

Andrei N. Kolmogorov

Já no século XX, como a conceituação até então não era apropriada a um tratamento matemático mais rigoroso, Andrei Nikolaevich Kolmogorov conceituou probabilidade através de axiomas rigorosos, tendo por base a teoria da medida.

Definição: Se A é um evento do espaço amostral S, então o número real P(A) será denominado probabilidade da ocorrência de A se satisfizer os seguintes axiomas:

Axioma 1. $0 \leq P(A) \leq 1$.

Axioma 2. $P(S) = 1$.

Axioma 3. Se A e B são eventos de S mutuamente exclusivos, então $P(A \cup B) = P(A) + P(B)$.

Notemos que A e B são mutuamente exclusivos se e somente se $A \cap B = \varnothing$.

O terceiro axioma pode ser generalizado para um número finito de eventos mutuamente exclusivos

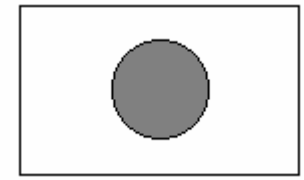
$$P(A_1 \cup A_2 \cup A_3 \cup \ldots \cup A_n) = \sum_{i=1}^{n} P(A_i)$$

O conceito axiomático não fornece formas e sim condições para o cálculo das probabilidades. Deste modo, os conceitos “a priori” e “a posteriori” se enquadram no conceito axiomático.

A principal vantagem do conceito axiomático é a possibilidade de extensão do estudo às variáveis contínuas, englobando eventos pertencentes a espaços amostrais infinitos não enumeráveis.

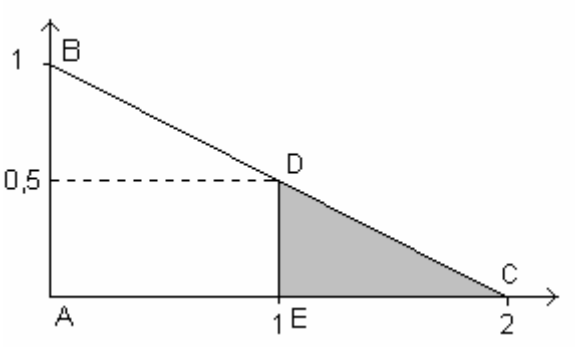
3.1.3.4. Probabilidade geométrica ou calculada como área

Seja *S* o espaço amostral associado a um experimento e A um evento de *S*. Definimos, então



$$P(A) = \frac{\text{área de A}}{\text{área de } S}$$

Exemplo: Seja o triângulo ABC um espaço amostral *S* e o triângulo CDE um evento A. A probabilidade P(A) é obtida da seguinte forma:



$$\text{área de } S = \frac{b \times h}{2} = \frac{2 \times 1}{2} = 1$$

$$\text{área de A} = \frac{b \times h}{2} = \frac{2 \times 1/2}{2} = \frac{1}{4}$$

$$P(A) = \frac{\text{área de A}}{\text{área de } S} = \frac{1/4}{1} = \frac{1}{4}$$

Vemos que a probabilidade de ocorrência de um evento é a *medida* do conjunto que representa o evento e pode ser calculada de diversas formas. Daí podemos fazer a seguinte generalização:

$$P(A) = \frac{\text{medida de A}}{\text{medida de S}} \rightarrow \text{medida} \begin{cases} \text{contagem} \\ \text{área} \\ \text{comprimento} \end{cases}$$

3.1.4. Teoremas para o cálculo de probabilidades

Teorema 1. Se $\varnothing$ é um evento impossível, então $P(\varnothing) = 0$.

Como $A \cup \varnothing = A$, então $P(A \cup \varnothing) = P(A)$ e
$A \cap \varnothing = \varnothing$, então A e $\varnothing$ são mutuamente exclusivos.

Utilizando então o terceiro axioma, temos

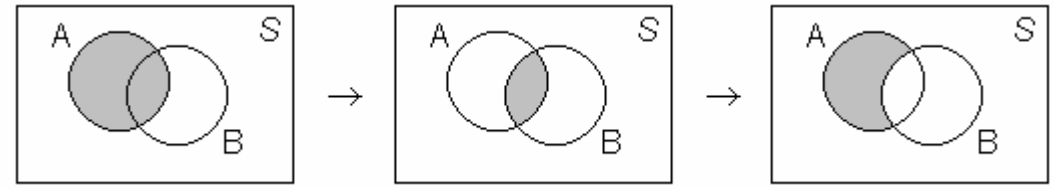
$$P(A \cup \varnothing) = P(A) + P(\varnothing)$$
$$P(A) = P(A) + P(\varnothing)$$
$$P(\varnothing) = P(A) - P(A) = 0$$

Teorema 2. Se $\overline{A}$ é o complemento de A, então $P(\overline{A}) = 1 - P(A)$.

Se $A \cup \overline{A} = S$, sendo A e $\overline{A}$ mutuamente exclusivos, então

$$P(S) = P(A \cup \overline{A}) = P(A) + P(\overline{A})$$
$$1 = P(A) + P(\overline{A})$$
$$P(\overline{A}) = 1 - P(A)$$

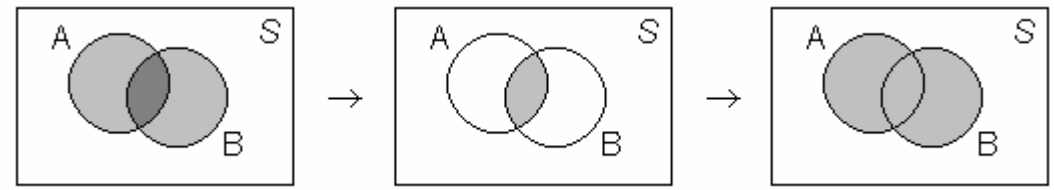
Teorema 3. Se A e B são dois eventos quaisquer, então $P(A - B) = P(A) - P(A \cap B)$.



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

Teorema da soma das probabilidades

Se A e B são dois eventos quaisquer, então $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.



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

3.1.5. Probabilidade condicional e independência

Sejam A e B dois eventos associados a um mesmo espaço amostral *S*. Se A e B não são eventos mutuamente exclusivos, ou seja, se $A \cap B \neq \varnothing$, então A e B poderão ser eventos independentes ou condicionados.

Para definir o que são eventos condicionados e eventos independentes, tomaremos, como exemplo, um experimento aleatório que será considerado em duas situações distintas.

Experimento: Uma caixa contém cinco bolas equiprováveis, sendo três azuis e duas brancas. Duas bolas são retiradas uma a uma e sua cor é observada. Definimos, então, dois eventos:

A_1: a primeira bola retirada é azul.
A_2: a segunda bola retirada é branca.

As probabilidades dos eventos A_1 e A_2 serão calculadas em duas situações.

Situação 1. Consideremos que a primeira bola retirada não é reposta (retirada sem reposição).

Sendo o espaço amostral enumerável, finito e equiprovável, podemos calcular probabilidade dos eventos através do conceito clássico. Deste modo,

$$P(A_1) = \frac{\#A_1}{\#S} = \frac{3}{5}.$$

Entretanto, a probabilidade do A_2 vai depender da ocorrência ou não do A_1.

Se ocorreu A_1, então $P(A_2/A_1) = \frac{\#A_2/A_1}{\#S} = \frac{2}{4}$

Se não ocorreu A_1, então $P(A_2) = \frac{\#A_2}{\#S} = \frac{1}{4}$.

Observamos, nesta situação, que, se a bola não for reposta, a probabilidade de ocorrência do A_2 fica alterada pela ocorrência ou não de A_1. Podemos definir, então:

◆ **Eventos condicionados**: dois eventos quaisquer, A e B, são condicionados quando a ocorrência de um altera a probabilidade de ocorrência do outro.
A probabilidade condicional de A é denotada por P(A/B) (lê-se probabilidade de A dado que ocorreu B).

Situação 2. Consideremos que a primeira bola retirada é reposta antes de tirar a segunda (retirada com reposição).

$$P(A_1) = \frac{\#A_1}{\#S} = \frac{3}{5}$$

Como a primeira bola é reposta, independente de ter ocorrido ou não A_1, a probabilidade de ocorrência de A_2 será a mesma.

Se ocorreu A_1, então $P(A_2/A_1) = \frac{\#A_2/A_1}{\#S} = \frac{2}{5}$

Se não ocorreu A_1, então $P(A_2) = \frac{\#A_2}{\#S} = \frac{2}{5}$.

Podemos verificar agora que, se a bola for reposta, a probabilidade de ocorrência do A_2 não é alterada pela ocorrência ou não do A_1, ou seja, $P(A_2) = P(A_2/A_1)$. Podemos definir, então:

◆ **Eventos independentes**: dois eventos quaisquer, A e B, são independentes quando a ocorrência de um não altera a probabilidade de ocorrência do outro, ou seja,

$$P(A) = P(A/B) \text{ e } P(B) = P(B/A).$$

Teorema do produto das probabilidades

Se A e B são dois eventos quaisquer, então

$$P(A \cap B) = P(A)\,P(B/A) = P(B)\,P(A/B).$$

Definimos, também

$$P(A/B) = \frac{P(A \cap B)}{P(B)} \quad \text{e} \quad P(B/A) = \frac{P(A \cap B)}{P(A)}$$

Se A e B são dois eventos independentes, então

$$P(A) = P(A/B) \text{ e } P(B) = P(B/A).$$

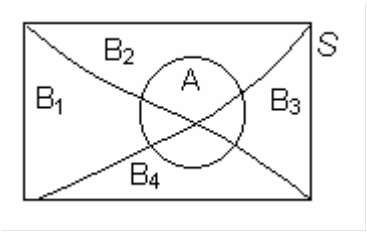
Logo,

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

Teorema de Bayes

Se *S* é um espaço amostral, com n=4 partições, onde está definido o evento A

Thomas Bayes
(1702 –1761)



$$B_1 \cup B_2 \cup B_3 \cup B_4 = S$$

$$\left.\begin{array}{l} B_1 \cap B_2 = \varnothing \\ B_1 \cap B_2 = \varnothing \\ \quad \ldots \\ B_1 \cap B_2 = \varnothing \end{array}\right\} B_i \cap B_j = \varnothing$$

podemos definir o evento A como

$$A = (A \cap B_1) \cup (A \cap B_2) \cup (A \cap B_3) \cup (A \cap B_4), \text{ logo,}$$
$$P(A) = \left[P(A \cap B_1) \cup P(A \cap B_2) \cup P(A \cap B_3) \cup P(A \cap B_4)\right].$$

Utilizando o terceiro axioma, temos

$$P(A) = P(A \cap B_1) + P(A \cap B_2) + P(A \cap B_3) + P(A \cap B_4).$$

Utilizando o teorema do produto, temos

$$P(A) = P(B_1)P(A/B_1) + P(B_2)P(A/B_2) + P(B_3)P(A/B_3) + P(B_4)P(A/B_4) = \sum_{i=1}^{4} P(B_i)P(A/B_i)$$

e

$$P(B_1/A) = \frac{P(A \cap B_1)}{P(A)} = \frac{P(B_1)\,P(A/B_1)}{\sum_{i=1}^{4} P(B_i)\,P(A/B_i)}.$$

Definimos, a partir desse exemplo, o teorema de Bayes:

Seja *S* um espaço amostral e B_1, B_2,..., B_n, uma de suas partições possíveis, tal que $B_i \cap B_j = \varnothing$ e $\bigcup_{i=1}^{n} B_i = S$. Se A é um evento de *S*, então:

$$P(A) = \sum_{i=1}^{n} P(B_i)P(A/B_i) \quad \text{e} \quad P(B_i/A) = \frac{P(B_i)\,P(A/B_i)}{\sum_{i=1}^{n} P(B_i)\,P(A/B_i)}.$$

Exercícios propostos:

3.1. Em 660 lançamentos de uma moeda, foram observadas 310 caras. Qual é a probabilidade de, num lançamento dessa moeda, obter-se coroa?

3.2. Se os registros indicam que 504, dentre 813 lavadoras automáticas de pratos vendidas por uma grande loja de varejo, exigiram reparos dentro da garantia de um ano, qual é a probabilidade de uma lavadora dessa loja não exigir reparo dentro da garantia?

3.3. Um grupo de pessoas é constituído de 60 homens e 40 mulheres. Sabe-se que 45 desses homens e 30 dessas mulheres votaram numa determinada eleição. Tomando-se, aleatoriamente, uma dessas pessoas, calcule a probabilidade de:

a) ser homem;
b) ser mulher;
c) ter votado;
d) não ter votado;
e) ser homem, sabendo-se que votou;
f) ser mulher, sabendo-se que não votou;
g) ter votado, sabendo-se que é mulher;
h) não ter votado, sabendo-se que é homem.

3.4. Em uma fábrica de parafusos, as máquinas A, B e C produzem 25 %, 35 % e 40 % do total produzido. Da produção de cada máquina, 5 %, 4 % e 2 %, respectivamente, são defeituosos. Escolhe-se ao acaso um parafuso e verifica-se que ele é defeituoso. Qual é a probabilidade de que seja da máquina A, da máquina B e da máquina C?

3.5. Em um estado (dos Estados Unidos) onde os automóveis devem ser testados quanto à emissão de poluentes, 25% de todos os carros emitem quantidades excessivas de poluentes. Ao serem testados, 99% de todos os carros que emitem excesso de poluentes são reprovados, mas 17% dos que não acusam emissão excessiva de poluentes também são reprovados. Qual é a probabilidade de um carro reprovado no teste acusar efetivamente excesso de emissão de poluentes?

3.6. Em uma certa comunidade, 6 % de todos os adultos com mais de 45 anos têm diabetes. Um novo teste diagnostica corretamente 84% das pessoas que têm diabetes e 98% das que não tem a doença.

a) Qual é a probabilidade de uma pessoa diagnosticada como diabética no teste, ter de fato a doença?
b) Qual é a probabilidade de uma pessoa que faça o teste, seja diagnosticada como não diabética?

3.2. Variáveis aleatórias

3.2.1. Introdução e conceito

Para facilitar a compreensão do conceito de variável aleatória, vamos tomar como exemplo o seguinte experimento aleatório.

Exemplo: Lançamento de uma moeda honesta três vezes e observação das faces que ocorrem.

O espaço amostral do experimento é

$$S = \{ccc, cck, ckc, kcc, kkc, kck, ckk, kkk\}.$$

Como a moeda é honesta, a probabilidade de ocorrer cara é igual à probabilidade de correr coroa: $P(c) = P(k) = \frac{1}{2}$.

Para que ocorra o resultado três caras (ccc), é necessário que ocorram, sucessivamente, os três eventos: cara no primeiro lançamento, cara no segundo lançamento e cara no terceiro lançamento, ou seja, deve ocorrer a *intersecção* destes três eventos. Como os lançamentos são independentes entre si, a probabilidade de ocorrer cara é a mesma em todos eles:

$$P(c) = \frac{1}{2}.$$

Logo, a probabilidade de ocorrer três caras P(ccc), é dada pelo *produto* das probabilidades de ocorrer cara em cada lançamento

$$P(ccc) = P(c) + P(c) + P(c) = \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{8}$$

De forma análoga, obtemos as probabilidades de todos os demais resultados possíveis.

$$P(cck) = P(c) + P(c) + P(k) = \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{8}$$

...

$$P(kkk) = P(k) + P(k) + P(k) = \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{8}.$$

Podemos observar, então, que

$$P(ccc) = P(cck) = P(ckc) = P(kcc) = P(kkc) = P(kck) = P(ckk) = P(kkk) = \frac{1}{8},$$

o que torna o espaço amostral equiprovável.

Observamos, também, que o espaço amostral é formado pela *união* dos eventos (ccc), (cck), (ckc), (kcc), (kkc), (kck), (ckk) e (kkk), que são todos mutuamente exclusivos. Sendo assim, a probabilidade do espaço amostral, P(*S*), é dada pela *soma* das probabilidades de cada evento

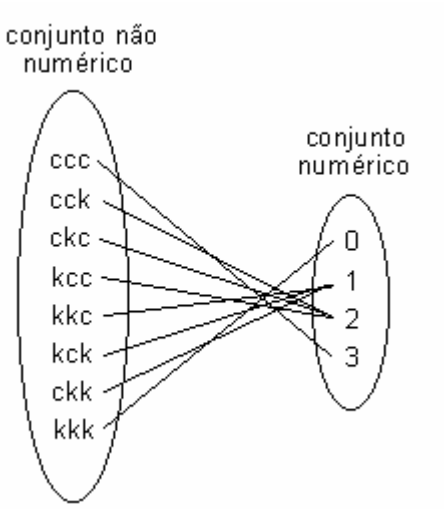
$$P(S) = P(ccc) + P(cck) + P(ckc) + P(kcc) + P(kkc) + P(kck) + P(ckk) + P(kkk)$$

$$P(S) = \frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} = 1.$$

Seja X a variável que representa o número de caras ocorrido nos três lançamentos, quais são os possíveis valores de X?

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

$X(ccc) = 3$
$X(cck) = 2$
$X(ckc) = 2$
$X(kcc) = 2$
$X(kkc) = 1$
$X(kck) = 1$
$X(ckk) = 1$
$X(kkk) = 0$



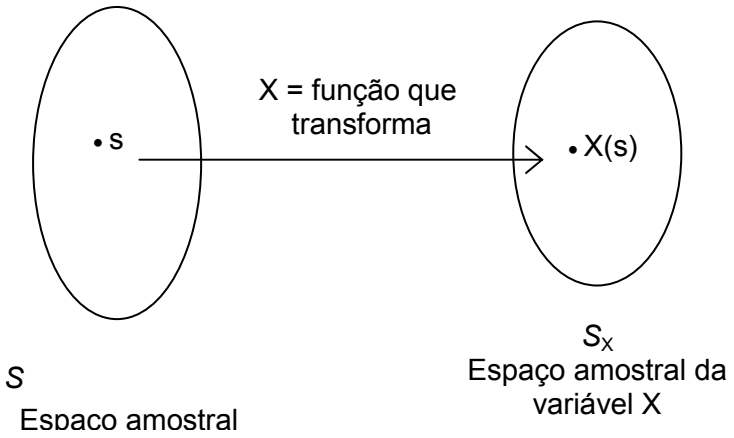
Através de X foi possível transformar um conjunto não numérico com oito pontos amostrais em um conjunto numérico com quatro pontos.

A partir deste exemplo podemos definir:

◆ **Variável aleatória** é uma função (ou regra) que transforma um espaço amostral qualquer em um espaço amostral numérico que será sempre um subconjunto do conjunto dos números reais.

No exemplo anterior, se X fosse a variável que representa o número de coroas, os conjuntos seriam os mesmos, mas a função seria outra, pois a correspondência é outra.

De modo geral, uma variável aleatória pode ser representada pelo esquema abaixo



As variáveis aleatórias podem ser classificadas como discretas ou contínuas. Por questões didáticas e de praticidade, vamos estudar cada tipo separadamente. Inicialmente, abordaremos as variáveis aleatórias discretas e suas principais distribuições de probabilidades e, mais adiante, as variáveis aleatórias contínuas.

3.2.2. Variáveis aleatórias discretas

Definição: São discretas todas as variáveis cujo espaço amostral S_X é *enumerável* finito ou infinito. Assim se X é uma variável aleatória discreta, então S_X é um subconjunto dos inteiros.

Tomemos como exemplo o seguinte experimento:

Lançamento de uma moeda até que ocorra face cara.

O espaço amostral básico deste experimento será

$$S = \{c, kc, kkc, kkkc, kkkkc, kkkkkc, ...\}.$$

Se definimos a variável aleatória X como o número de lançamentos até que ocorra cara, então, temos

$$S \xrightarrow{X} S_X = \{1, 2, 3, 4 ,5,...\}.$$

Se definimos outra variável aleatória Y como o número de coroas até que ocorra cara, então temos

$$S \xrightarrow{Y} S_Y = \{0, 1, 2, 3, 4,...\}.$$

Observamos que X e Y são variáveis aleatórias discretas, pois seus espaços amostrais são enumeráveis.

3.2.2.1. Função de probabilidade

Definição: Seja X uma variável aleatória discreta e S_X o seu espaço amostral. A função de probabilidade $P(X = x)$, ou simplesmente $p(x)$, será a função que associa a cada valor de X a sua probabilidade de ocorrência, desde que satisfaça duas condições:

1. $p(x) \geq 0, \forall\ x \in S_X$
2. $\sum_{x \in S_X} p(x) = 1$

Existem três formas distintas de representar uma função:

– *Representação tabular:* consiste em relacionar em uma tabela os valores da função de probabilidade.

– *Representação gráfica:* consiste em representar graficamente a relação entre os valores da variável e suas probabilidades.

– *Representação analítica:* estabelece uma expressão geral para representar o valor da função num ponto genérico da variável.

Para exemplificar as formas de representação de uma função de probabilidade, vamos considerar o seguinte experimento aleatório.

Exemplo: De uma urna com três bolas pretas e duas brancas, retiram-se, de uma vez, duas bolas. Se X é o número de bolas pretas retiradas, determine a função de probabilidade $P(X = x)$.

Observamos que o espaço amostral básico do experimento é um conjunto não numérico

$$S = \{P_1B_1,\ P_1B_2,\ P_2B_1,\ P_2B_2,\ P_3B_1,\ P_3B_2,\ P_1P_2,\ P_1P_3,\ P_2P_3,\ B_1B_2\}$$

e que a variável X transforma este espaço num conjunto numérico

$$S_X = \{0,\ 1,\ 2\}$$

Como o espaço amostral básico *S* é enumerável, finito e equiprovável, podemos obter as probabilidades associadas aos valores de X através do conceito clássico. Já vimos anteriormente que, neste tipo de experimento, o número de elementos do espaço e o número de pontos favoráveis à ocorrência do evento desejado podem ser obtidos através da combinação. Daí, temos:

$$P(X=0) = P(B_1B_2) = \frac{C_3^0\, C_2^2}{C_5^2} = \frac{1}{10} = 0{,}1 = 0{,}1$$

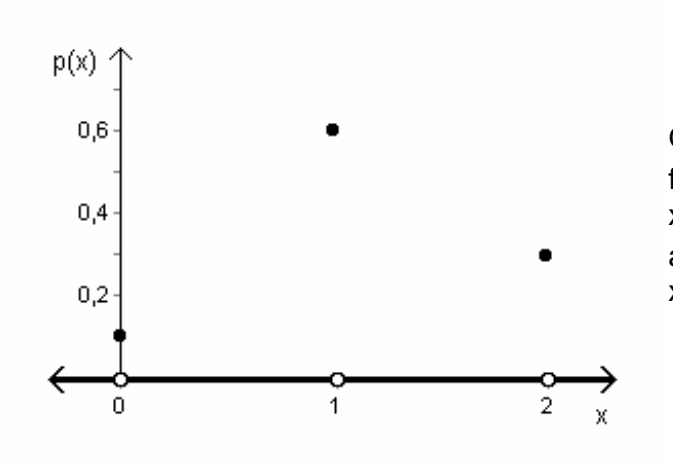
$$P(X=1) = P(P_1B_1) + P(P_1B_2) + P(P_2B_1) + P(P_2B_2) + P(P_3B_1) + P(P_3B_2) = \frac{C_3^1\, C_2^1}{C_5^2} = \frac{6}{10} = 0{,}6$$

$$P(X=2) = P(P_1P_2) + P(P_1P_3) + P(P_2P_3) = \frac{C_3^2\, C_2^0}{C_5^2} = \frac{3}{10} = 0{,}3$$

Obtidas as probabilidades, podemos fazer a representação tabular da função.

$X = x$	0	1	2	Σ
$P(X = x)$	0,1	0,6	0,3	1

Da mesma forma, é possível construir o gráfico para a função.



Observamos que $P(X = x)$ é uma função contínua para todo o $x \notin S_X$, ou seja, a função $P(X = x)$ assume o valor zero para todo o $x \notin S_X$.

A representação analítica da função é feita através da generalização da expressão utilizada para o cálculo da probabilidade de cada valor de X:

$$P(X = x) = \frac{C_3^x\, C_2^{2-x}}{C_5^2}, \quad \text{para } S_X = \{0,\ 1,\ 2\}$$

3.2.2.2. Função de distribuição ou probabilidade acumulada

Definição: Seja X uma variável aleatória discreta e S_X o seu espaço amostral. A função de distribuição, definida por F(x) ou $P(X \leq x)$ é a função que associa a cada valor de X a probabilidade $P(X \leq x)$. Desta forma, temos

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

Para o exemplo anterior, temos:

$$F(0) = P(X \leq 0) = \sum_{x \leq 0} P(X = x) = P(X = 0) = 0,1$$

$$F(1) = P(X \leq 1) = \sum_{x \leq 1} P(X = x) = P(X = 0) + P(X = 1) = 0,1 + 0,6 = 0,7$$

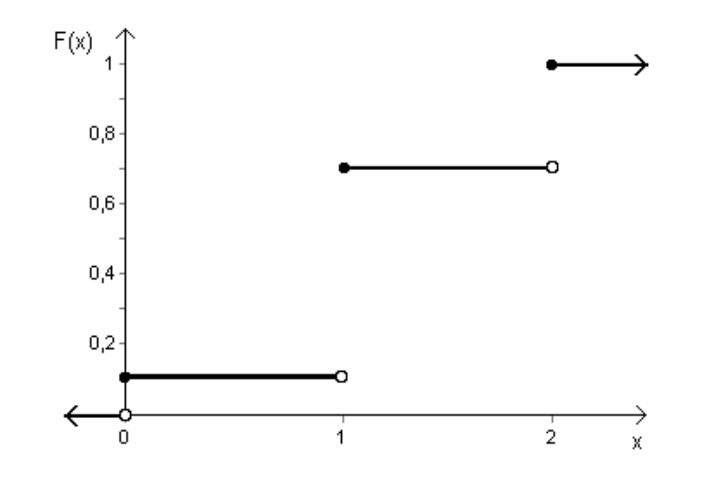
$$F(2) = P(X \leq 2) = \sum_{x \leq 2} P(X = x) = P(X = 0) + P(X = 1) + P(X = 2) = 0,1 + 0,6 + 0,3 = 1$$

Podemos também representar a função de distribuição acumulada de três formas:

– representação tabular

$X = x$	0	1	2	Σ
$P(X = x)$	0,1	0,6	0,3	1
F(x)	0,1	0,7	1	-

– representação gráfica



– representação analítica

$$F(x) = P(X \leq x) = \sum_{t \leq x} \frac{C_3^t \, C_2^{2-t}}{C_5^2}, \quad \text{para } S_X = \{0, 1, 2\}$$

3.2.2.3. Medidas descritivas

Visto que as medidas descritivas servem para descrever conjuntos de dados numéricos e que o espaço amostral de uma variável aleatória é sempre um conjunto numérico, podemos utilizar essas medidas para representar as distribuições de probabilidades de variáveis aleatórias.

- **Média ou valor esperado**

Definição: Seja X uma variável aleatória discreta e S_X o seu espaço amostral. O valor médio de X representado por E(X) ou μ_X ou simplesmente μ, é a média dos valores de X ponderada pelas suas respectivas probabilidades de ocorrência. Deste modo, temos

$$E(X) = \mu = \frac{\sum_{x \in S_X} x\,p(x)}{\sum_{x \in S_X} p(x) = 1} = \sum_{x \in S_X} x\,p(x)$$

Considerando o exemplo cuja distribuição de probabilidade é a seguinte

$X = x$	0	1	2	Σ
$P(X = x)$	0,1	0,6	0,3	1

o valor esperado para o número de bolas pretas retiradas será:

$$E(X) = \mu = \sum_{x \in S_X} x\,p(x) = 0 \times 0{,}1 + 1 \times 0{,}6 + 2 \times 0{,}3 = 1{,}2 \text{ bolas}$$

Devemos destacar que a média ou valor esperado possui propriedades matemáticas importantes, algumas já vistas na Unidade II, as quais são relacionadas a seguir.

Propriedades matemáticas da média

1ª propriedade: Se c é uma constante, então

$$E(c) = c$$

2ª propriedade: Se X é uma variável aleatória e c uma constante, ao somarmos a constante aos valores da variável, a média da variável também fica somada da constante.

$$E(c+X)=c+E(X)$$

3ª propriedade: Se X é uma variável aleatória e c uma constante, ao multiplicarmos a variável pela constante, a média da variável também fica multiplicada pela constante.

$$E(cX)=cE(X)$$

4ª propriedade: A média dos desvios é igual a zero.

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

5ª propriedade: A média dos desvios quadráticos é mínima.

$$E(X-\mu)^2 < E(X-c)^2\ 0$$

6ª propriedade: Se X e Y são duas variáveis aleatórias, a média da soma (ou diferença) das duas variáveis é igual à soma (ou diferença) de suas médias.

$$E(X \pm Y) = E(X) \pm E(Y)$$

7ª propriedade: Se X e Y são duas variáveis aleatórias *independentes*, a média do produto das duas variáveis é igual ao produto de suas médias.

$$E(XY) = E(X)E(Y), \text{ se X e Y são } independentes.$$

◆ **Variância**

Definição: Seja X uma variável aleatória discreta e S_X o seu espaço amostral. O grau médio de dispersão dos valores de X em relação a sua média é conhecido como variância que é representada por V(X), ou σ_X^2, ou simplesmente σ^2, e definida como a média dos quadrados dos desvios em relação á média. Sendo assim, temos

$$V(X) = \sigma^2 = E(X-\mu)^2 = \sum_{x \in S_X} (x-\mu)^2 p(x) \quad \text{(Fórmula de definição)}$$

ou

$$V(X) = \sigma^2 = E(X^2) - \mu^2, \text{ onde } E(X)^2 = \sum_{x \in S_X} x^2 p(x) \quad \text{(Fórmula prática)}$$

Para o exemplo cuja distribuição de probabilidade é

$X = x$	0	1	2	Σ
$P(X = x)$	0,1	0,6	0,3	1

e o valor esperado é μ = 1,2 bolas, a variância do número de bolas pretas retiradas será:

$$V(X) = \sigma^2 = \sum_{x \in S_X} (x-\mu)^2 p(x) = (0-1,2)^2 \times 0,1 + (1-1,2)^2 \times 0,6 + (2-1,2)^2 \times 0,3 = 0,36 \text{ bolas}^2.$$

A variância destaca-se entre as medidas de variação por apresentar algumas propriedades matemáticas.

Propriedades matemáticas da variância:

1ª propriedade: Se k é uma constante, então

$$V(k) = 0$$

2ª propriedade: Se X é uma variável aleatória e c uma constante, ao somarmos a constante aos valores da variável a variância da variável não se altera.

$$V(X+c)=V(X)$$

3ª propriedade: Se X é uma variável aleatória e k uma constante, ao multiplicarmos a variável pela constante a variância da variável fica multiplicada pelo quadrado constante.

$$V(kX) = k^2 V(X)$$

4ª propriedade: Se X e Y são duas variáveis aleatórias *independentes*, a variância da soma (ou diferença) das duas variáveis é igual à soma de suas variâncias.

$$V(X \pm Y) = V(X) + V(Y), \text{ se X e Y são } \textit{independentes}$$

◆ **Desvio padrão**

A partir da variância podemos obter o desvio padrão, denotado por σ e definido como a raiz quadrada da variância:

$$\sigma = \sqrt{\sigma^2} \ .$$

No exemplo: $\sigma = \sqrt{\sigma^2} = \sqrt{0{,}36} = 0{,}6$ bolas.

◆ **Momentos**

Já vimos anteriormente que os momentos são quantidades que auxiliam na descrição de um conjunto de valores. Da mesma forma, aqui, essas medidas são utilizadas para descrever as distribuições de probabilidade de variáveis aleatórias. A expressão geral do momento de ordem r de uma variável aleatória é a seguinte:

$$\mu_r = E(X - a)^r$$

Os tipos mais importantes de momentos são dois:

– Quando a = 0, temos os *momentos centrados na origem* ou *momentos ordinários de ordem* r:

$$\mu'_r = E(X - 0)^r = E(X^r)$$

Para r = 1, temos

$$\mu'_1 = E(X) = \sum_{x \in S_X} x\,p(x)$$

Para r = 2, temos

$$\mu'_2 = E(X^2) = \sum_{x \in S_X} x^2 p(x)$$

Para r = 3, temos

$$\mu'_3 = E(X^3) = \sum_{x \in S_X} x^3 p(x)$$

Para r = 4, temos

$$\mu'_4 = E(X^4) = \sum_{x \in S_X} x^4 p(x)$$

– Quando a = μ, temos os *momentos de ordem* r *centrados na média*

$$\mu_r = E(X-\mu)^r$$

Para r = 1, temos

$$\mu_1 = E(X-\mu)$$
$$\mu_1 = E(X) - E(\mu)$$
$$\mu_1 = \mu - \mu = 0$$

Para r = 2, temos

$$\mu_2 = E(X-\mu)^2 = \sum_{x \in S_X} (x-\mu)^2 p(x) \quad \text{(Fórmula de definição)}$$
$$\mu_2 = E(X-\mu)^2$$
$$\mu_2 = E(X^2 - 2X\mu + \mu^2)$$
$$\mu_2 = E(X^2) - E(2X\mu) + E(\mu^2)$$
$$\mu_2 = E(X^2) - 2\mu E(X) + \mu^2$$
$$\mu_2 = E(X^2) - 2\mu^2 + \mu^2$$
$$\mu_2 = E(X^2) - \mu^2 \quad \text{(Fórmula prática)}$$

Para r = 3, temos

$$\mu_3 = E(X-\mu)^3 = \sum_{x \in S_X} (x-\mu)^3 p(x) \quad \text{(Fórmula de definição)}$$
$$\mu_3 = E(X-\mu)^3$$
$$\mu_3 = E(X^3 - 3X^2\mu + 2X\mu^2 - \mu^3)$$
$$\mu_3 = E(X^3) - E(3X^2\mu) + E(3X\mu^2) - E(\mu^3)$$
$$\mu_3 = E(X^3) - 3\mu E(X^2) + 3\mu^2 E(X) - \mu^3$$
$$\mu_3 = E(X^3) - 3\mu E(X^2) + 3\mu^2\mu - \mu^3$$
$$\mu_3 = E(X^3) - 3\mu E(X^2) + 3\mu^3 - \mu^3$$
$$\mu_3 = E(X^3) - 3\mu E(X^2) + 2\mu^3 \quad \text{(Fórmula prática)}$$

Para r = 4, temos

$$\mu_4 = E(X-\mu)^4 = \sum_{x \in S_X} (x-\mu)^4 p(x) \quad \text{(Fórmula de definição)}$$
$$\cdots$$
$$\mu_4 = E(X^4) - 4\mu E(X^3) + 6\mu^2 E(X^2) - 3\mu^4 \quad \text{(Fórmula prática)}$$

- **Coeficiente de assimetria**

$$a_3 = \frac{\mu_3}{\mu_2\sqrt{\mu_2}} = \frac{\mu_3}{\mu_2^{3/2}}$$

- **Coeficiente de curtose**

$$a_4 = \frac{\mu_4}{\mu_2^2}$$

3.2.3. Variáveis aleatórias contínuas

Vamos considerar agora o seguinte experimento: tomar aleatoriamente uma peça de uma linha de fabricação, colocá-la em funcionamento e medir por quanto tempo ela funciona. Um possível espaço amostral básico para este experimento seria a anotação do dia e da hora em que a peça parou de funcionar. Um procedimento equivalente (e mais adequado do ponto de vista das aplicações) seria associar a cada ponto desse espaço amostral o tempo de funcionamento decorrido. Assim, teríamos uma variável aleatória X definida como o tempo de funcionamento da peça. Esta variável X seria uma variável aleatória contínua, visto que o conjunto dos seus valores não poderia ser enumerado, e o seu espaço amostral poderia ser representado como $\{x; x \geq 0\}$. Observamos também que, sendo X uma variável contínua, entre quaisquer dois valores distintos de X sempre existirão infinitos valores. A partir deste exemplo, podemos definir uma *variável aleatória contínua*.

Definição: São contínuas todas as variáveis cujo espaço amostral S_X é infinito *não enumerável*. Assim, se X é uma variável aleatória contínua, então X pode assumir qualquer valor num intervalo [a; b] ou no intervalo $(-\infty; +\infty)$ e o conjunto S_X será sempre definido como um intervalo.

São exemplos de variáveis aleatórias contínuas: o tempo de vida de um animal, a vida útil de um componente eletrônico, o peso de uma pessoa, a produção de leite de uma vaca, a quantidade de chuva que ocorre numa região.

3.2.3.1. Função densidade de probabilidade

Definição: Seja X uma variável aleatória contínua e S_X o seu espaço amostral. Uma função f associada à variável X é denominada função densidade de probabilidade se satisfizer duas condições:

1. $f(x) \geq 0, \forall\ x \in S_X$
2. $\int_{S_X} f(x)dx = 1 = P(X \in S_X)$ ⇦

A área sob a função f(x) no intervalo S_X é um, pois corresponde à probabilidade de a variável X pertencer ao espaço amostral S_X.

Consideremos agora dois exemplos resolvidos.

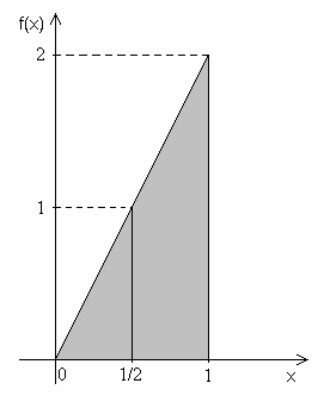
Exemplo 1. Seja a função f(x) = 2x, no intervalo S_X =[0,1]. Verifique se f(x) é uma função densidade de probabilidade.

Primeira condição: $f(x) \geq 0, \forall\ x \in S_X$

Como a função f(x) = 2x é linear, apenas dois pontos são suficientes para traçar a reta que representa a relação entre x e f(x). Podemos obter, então, os valores da função f(x) nos pontos 0 e 1 que são os limites do intervalo S_X:

- para x = 0, temos $f(0) = 2 \times 0 = 0$,
- para x = 1, temos $f(1) = 2 \times 1 = 2$.

A partir desses dois pontos é possível construir o gráfico da função



Podemos observar que todos os valores da função *f*(x) são não negativos no intervalo de 0 a 1; portanto, a primeira condição foi atendida.

Segunda condição: $\int_{S_X} f(x)dx = 1$

A integral é a ferramenta utilizada para se obter a área sob a função f(x) no intervalo S_X, que equivale a $P(X \in S_X)$ e deve ser igual a 1. Entretanto, na função f(x) = 2x essa área adquire a forma de um triângulo, podendo ser mais facilmente calculada através da expressão bh/2. Assim, temos

$$\text{Área} = \frac{bh}{2} = \frac{1 \times 2}{2} = 1.$$

Como a área sob a função f(x) é igual a 1, a segunda condição também foi atendida. Logo, a função f(x) = 2x no intervalo S_X =[0, 1] é uma função densidade de probabilidade.

Exemplo 2. Seja a função $f(x) = 6x - 6x^2$, no intervalo S_X =[0,1]. Verifique se f(x) é uma função densidade de probabilidade.

Primeira condição: $f(x) \geq 0, \ \forall \ x \in S_X$

Como $f(x) = 6x - 6x^2$ é uma função quadrática, são necessários, pelo menos, três pontos para traçar a parábola que representa a relação entre x e f(x). Devemos obter, então, os valores da função f(x) nos pontos 0 e 1, que são os limites do intervalo S_X, e no valor que corresponde ao ponto crítico da função. Para determinar este valor de x derivamos a função e igualamos a zero a primeira derivada. Deste modo, para

$f(x) = 6x - 6x^2$, temos

$f'(x) = 6 - 12x, \ \text{sendo} \ f'(x) = 0, \ \text{temos}$

$0 = 6 - 12x$

$x = \frac{6}{12} = \frac{1}{2} \ \rightarrow \ \text{valor de X que corresponde ao ponto crítico de } f'(x)$

Derivando a função pela segunda vez, é possível determinar se o ponto crítico é um ponto de máximo ou de mínimo.

$f''(x) = -12$

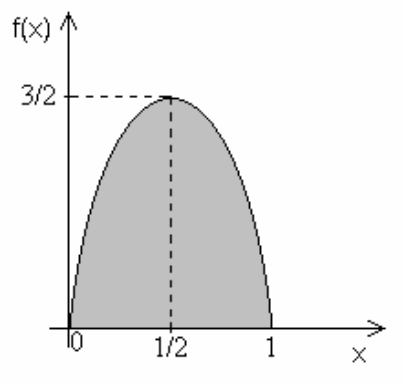
Sabemos que:

– se $f''(x) < 0$, a parábola tem concavidade para baixo, ou seja, tem ponto de máximo.
– se $f''(x) > 0$, a parábola tem concavidade para cima, ou seja, tem ponto de mínimo.

Como a segunda derivada resultou negativa, concluímos que a função tem ponto de máximo. Assim, obtemos os valores da função $f(x) = 6x – 6x^2$ nos pontos 0, 1/2 e 1:

– para x = 0, temos $f(0) = 6 \times 0 – 6 \times 0^2 = 0$,
– para x = 1/2, temos $f(1/2) = 6 \times 1/2 – 6 \times (1/2)^2 = 3/2$,
– para x = 1, temos $f(1) = 6 \times 1 – 6 \times 1^2 = 0$.

A partir desses três pontos é possível traçar o gráfico da função



Observamos que todos os valores da função f(x) são maiores que zero no intervalo de 0 a 1; portanto, a primeira condição foi atendida.

Segunda condição: $\int_{S_X} f(x)dx = 1$

Como a representação gráfica de f(x) no intervalo [0, 1] é uma parábola, a área sob a função pode ser obtida através da integração da diferencial da função (f(x)dx) neste intervalo. Daí, temos

$$\text{Área} = \int_0^1 f(x)dx = \int_0^1 \left(6x - 6x^2\right)dx = \int_0^1 6xdx - \int_0^1 6x^2dx = 6\int_0^1 xdx - 6\int_0^1 x^2dx = 6\left[\frac{x^2}{2}\right]_0^1 - 6\left[\frac{x^3}{3}\right]_0^1$$
$$= 3\left(1^2 - 0^2\right) - 2\left(1^3 - 0^3\right) = 3 - 2 = 1.$$

Como a área sob a função f(x) no intervalo $S_{\mathbf{x}}$, que equivale a $P(X \in S_{\mathbf{x}})$, é igual a 1, a segunda condição também foi atendida.

Portanto, a função $f(x) = 6x – 6x^2$, no intervalo $S_{\mathbf{x}} = [0, 1]$ é uma função densidade de probabilidade.

3.2.3.2. Função de distribuição ou probabilidade acumulada

Definição: Seja X uma variável aleatória contínua e S_X o seu espaço amostral. A função de distribuição, definida por F(x) ou $P(X \leq x)$, é a função que associa a cada valor de $x \in S_X$ a sua probabilidade acumulada $P(X \leq x)$. Desta forma, temos

$$F(x) = P(X \leq x) = \int_a^x f(t)dt \text{ , para } S_X = [a, b].$$

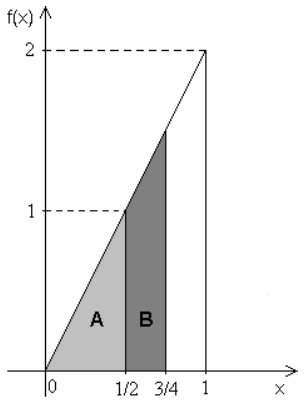
Sendo S_X =[a, b]., temos

$$F(a) = P(X \leq a) = 0$$
$$F(b) = P(X \leq b) = 1$$

Consideremos o Exemplo 1. Para a função densidade de probabilidade f(x) = 2x, no intervalo S_X =[0,1], definem-se os seguintes eventos:

A = {x; 0 < x ≤ 1/2}

B = {x; 1/2 ≤ x ≤ 3/4}



As probabilidades dos eventos A e B correspondem às suas respectivas áreas:

$$P(A) = \text{área de A} = \int_0^{1/2} 2x\,dx = 2\int_0^{1/2} x\,dx = 2\left[\frac{x^2}{2}\right]_0^{1/2} = \left(\frac{1}{2}\right)^2 - 0^2 = \frac{1}{4}$$

$$P(B) = \text{área de B} = \int_{1/2}^{3/4} 2x\,dx = 2\int_{1/2}^{3/4} x\,dx = 2\left[\frac{x^2}{2}\right]_{1/2}^{3/4} = \left(\frac{3}{4}\right)^2 - \left(\frac{1}{2}\right)^2 = \frac{9}{16} - \frac{1}{4} = \frac{9-4}{16} = \frac{5}{16}$$

Para f(x) = 2x, a função de distribuição acumulada F(x) será

$$F(x) = P(X \leq x) = \int_0^x 2t\,dt = 2\left[\frac{t^2}{2}\right]_0^x = x^2 - 0^2 = x^2.$$

As probabilidades dos eventos A e B podem ser obtidas de outra forma através da função de distribuição acumulada $F(x) = P(X \leq x) = x^2$. Assim, temos

$$F(1/2) = P(X \leq 1/2) = \left(\frac{1}{2}\right)^2 = \frac{1}{4} \quad \text{e}$$

$$F(3/4) = P(X \leq 3/4) = \left(\frac{3}{4}\right)^2 = \frac{9}{16},$$

donde resulta

$$P(A) = F(1/2) = \frac{1}{4}$$

e

$$P(B) = F(3/4) - F(1/2) = \frac{9}{16} - \frac{1}{4} = \frac{5}{16}$$

Para o Exemplo 2:

Seja $f(x) = 6x - 6x^2$ uma função densidade de probabilidade definida no intervalo S_X=[0,1]. Para f(x) a função de distribuição acumulada F(x) será

$$F(x) = \int_0^x \left(6t - 6t^2\right)dt = \int_0^x 6tdt - \int_0^x 6t^2dt = 6\left[\frac{t^2}{2}\right]_0^x - 6\left[\frac{t^3}{3}\right]_0^x = 3x^2 - 2x^3 .$$

3.2.3.3. Medidas descritivas

- **Média ou valor esperado**

Definição: Seja X uma variável aleatória contínua e S_X o seu espaço amostral. O valor médio de X, representado por E(X) ou μ , será dado por

$$E(X) = \mu = \int_{S_X} x\, f(x)dx$$

Sempre que a função for par e, portanto, simétrica, $F(\mu) = 1/2$.

- **Variância**

Definição: Seja X uma variável aleatória contínua e S_X o seu espaço amostral. A variância de X, representada por V(X) ou σ^2, será dada por

$$V(X) = \sigma^2 = E(X-\mu)^2 = \int_{S_X} (x-\mu)^2 f(x)dx \quad \text{(Fórmula de definição)}$$

ou

$$V(X) = \sigma^2 = E(X^2) - \mu^2 = \left[\int_{S_X} x^2 f(x)dx\right] - \mu^2 \quad \text{(Fórmula prática)}$$

Para o Exemplo 1: $f(x) = 2x$, $S_X = [0,1]$, temos:

$$E(X) = \mu = E(X) = \mu = \int_0^1 x\,2xdx = \int_0^1 2x^2dx = 2\int_0^1 x^2dx = 2\left[\frac{x^3}{3}\right]_0^1 = 2\left(\frac{1^3}{3} - \frac{0^3}{3}\right) = \frac{2}{3}$$

e

$$V(X) = \sigma^2 = \left[\int_0^1 x^2\,2xdx\right] - \mu^2 = \left[\int_0^1 2x^3dx\right] - \mu^2 = \left\{2\left[\frac{x^4}{4}\right]_0^1\right\} - \left(\frac{2}{3}\right)^2 = \frac{1}{2} - \frac{4}{9} = \frac{9-8}{18} = \frac{1}{18}$$

Para o Exemplo 2: $f(x) = 6x - 6x^2$, $S_X = [0,1]$, temos:

$$E(X) = \mu = \int_0^1 x\left(6x - 6x^2\right)dx = \int_0^1 \left(6x^2 - 6x^3\right)dx \int_0^1 6x^2dx - \int_0^1 6x^3dx$$

$$= 6\left[\frac{x^3}{3}\right]_0^1 - 6\left[\frac{x^4}{4}\right]_0^1 = 2 - \frac{6}{4} = \frac{8-6}{4} = \frac{1}{2}$$

e

$$V(X) = \sigma^2 = \left[\int_0^1 x^2\left(6x - 6x^2\right)dx\right] - \mu^2 = \left[\int_0^1 \left(6x^3 - 6x^4\right)dx\right] - \mu^2$$

$$= \left[\int_0^1 6x^3dx - \int_0^1 6x^4dx\right] - \mu^2 = \left\{6\left[\frac{x^4}{4}\right]_0^1 - 6\left[\frac{x^5}{5}\right]_0^1\right\} - \left(\frac{1}{2}\right)^2 = \frac{3}{2} - \frac{6}{5} - \frac{1}{4} = \frac{30-24-5}{20} = \frac{1}{20}$$

- **Momentos**

Segundo momento:

$$\mu_2 = E(X-\mu)^2 = \int_{S_X} (x-\mu)^2 f(x)dx \quad \text{(Fórmula de definição)}$$

$$\mu_2 = E(X^2) - \mu^2 = \left[\int_{S_X} x^2 f(x)dx\right] - \mu^2 \quad \text{(Fórmula prática)}$$

Terceiro momento:

$$\mu_3 = E(X-\mu)^3 = \int_{S_X} (x-\mu)^3 f(x)dx \quad \text{(Fórmula de definição)}$$

$$\mu_3 = E(X^3) - 3\mu E(X^2) + 2\mu^3 = \left[\int_{S_X} x^3 f(x)dx\right] - 3\mu\left[\int_{S_X} x^2 f(x)dx\right] + 2\mu^3 \quad \text{(Fórmula prática)}$$

Quarto momento:

$$\mu_4 = E(X-\mu)^4 = \int_{S_X} (x-\mu)^4 f(x)dx \quad \text{(Fórmula de definição)}$$

$$\mu_4 = E(X^4) - 4\mu E(X^3) + 6\mu^2 E(X^2) - 3\mu^4$$

$$= \left[\int_{S_X} x^4 f(x)dx\right] - 4\mu\left[\int_{S_X} x^3 f(x)dx\right] + 6\mu^2\left[\int_{S_X} x^2 f(x)dx\right] - 3\mu^4 \quad \text{(Fórmula prática)}$$

- **Coeficiente de assimetria**

$$a_3 = \frac{\mu_3}{\mu_2\sqrt{\mu_2}} = \frac{\mu_3}{\mu_2^{3/2}}$$

- **Coeficiente de curtose**

$$a_4 = \frac{\mu_4}{\mu_2^2}$$

3.3. Distribuições de probabilidade

Até o momento, as variáveis aleatórias consideradas não possuíam, necessariamente, qualquer sentido de aplicação. Entretanto, algumas variáveis aleatórias são muitos importantes e, devido a esta importância, surge o interesse em estudar suas distribuições de probabilidade.

Uma distribuição de probabilidade é essencialmente um modelo de descrição probabilística de uma população, entendendo por população o conjunto de todos os valores de uma variável aleatória. As ideias de população e distribuição de probabilidade são, deste modo, indissociáveis e serão, a partir de agora, tratadas como sinônimos. As distribuições de probabilidade formam a espinha dorsal da metodologia estatística, uma vez, que pela sua natureza, a estatística somente trabalha com variáveis cujos valores não ocorrem de modo determinístico.

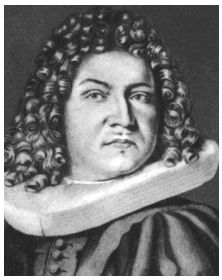
No estudo de uma variável aleatória é importante saber:
- o tipo de distribuição de probabilidade da variável;
- a função de probabilidade da variável;
- os parâmetros da distribuição;
- as medidas descritivas da distribuição (média, variância, assimetria).

Existem inúmeros modelos descrevendo o comportamento probabilístico de variáveis discretas e contínuas. Nas seções a seguir serão discutidos os principais tipos de distribuições discretas e contínuas.

3.3.1. Distribuições de probabilidade de variáveis discretas

As distribuições discretas mais importantes e utilizadas são: Bernoulli, Binomial, Hipergeométrica, Poisson, Uniforme, Multinomial, Geométrica, Binomial negativa e Hipergeométrica negativa. Todavia, nesta seção, trataremos apenas as cinco primeiras.

3.3.1.1. Distribuição de Bernoulli


Jacob Bernoulli
(1654 – 1705)

Esta distribuição foi deduzida no final do século XVII pelo matemático suíço Jacob Bernoulli.

Definição: modelo de descrição probabilística dos resultados de um experimento de Bernoulli.

O experimento (ou ensaio) de Bernoulli é definido como o experimento aleatório que possui apenas dois resultados possíveis.

Exemplos:

Experimento 1. Uma lâmpada é colocada numa luminária.

$$S = \{\text{acende, não acende}\}$$

Vamos considerar um dos resultados como sucesso, por exemplo, sucesso = acender. Definimos, então, a variável X como número de sucessos em uma repetição do experimento.

$$X = \text{número de sucessos}$$

A variável X só poderá assumir dois valores

$$X = \begin{cases} 0, & \text{se a lâmpada não acender} \\ 1, & \text{se a lâmpada acender} \end{cases}, \text{ sendo } S_X = \{0, 1\}.$$

Experimento 2. Uma semente é colocada para germinar.

S = {germina, não germina}

Se sucesso = germinar, então, a variável X = número de sucessos será

$$X = \begin{cases} 0, & \text{se a semente não germinar} \\ 1, & \text{se a semente germinar} \end{cases}, \text{ sendo } S_X = \{0, 1\}.$$

Se for conhecido o poder germinativo do lote de sementes, por exemplo, 87%, então, podemos concluir que a probabilidade de a semente germinar é 0,87. Como o evento {não germinar} é complemento do evento {germinar}, a probabilidade de não germinar será 1 – 0,87. Temos, então

$X = x$	0	1	Σ
$P(X = x)$	0,13	0,87	1

Experimento 3. O nascimento de um bovino.

S = {macho, fêmea}

Se sucesso = fêmea, então, a variável X = número de sucessos será

$$X = \begin{cases} 0, & \text{se nascer macho} \\ 1, & \text{se nascer fêmea} \end{cases}, \text{ sendo } S_X = \{0, 1\}$$

Sabe-se que a probabilidade de nascer fêmea é a mesma de nascer macho. Temos, então

$X = x$	0	1	Σ
$P(X = x)$	0,5	0,5	1

◆ **Função de probabilidade**

De modo geral, se X é uma variável que tem distribuição de Bernoulli, então a sua função de probabilidade será:

– Representação tabular

$X = x$	0	1	Σ
$P(X = x)$	$(1-\pi)$	π	1

onde:

π = probabilidade de sucesso

$(1-\pi)$ = probabilidade de fracasso

– Representação analítica

$$P(X = x) = \pi^x (1-\pi)^{1-x}, \text{ para } S_X = \{0, 1\}$$

◆ **Parâmetros**

A distribuição de Bernoulli tem apenas um parâmetro:

π = probabilidade de sucesso

Dizemos, então, que

$$X \sim \text{Ber}(\pi).$$

◆ **Medidas descritivas**

– Média ou valor esperado: $E(X) = \mu = \sum_{x \in S_X} x\,p(x)$

$$E(X) = 0 \times (1-\pi) + 1 \times \pi = \pi$$

Teorema: $E(X) = \mu = \pi$

– Variância: $V(X) = \sigma^2 = E(X^2) - \mu^2$

Como

$$E(X^2) = \sum_{x \in S_X} x^2 p(x) = 0^2 \times (1-\pi) + 1^2 \times \pi = \pi,$$

temos

$$V(X) = E(X^2) - \mu^2 = \pi - \pi^2 = \pi(1-\pi).$$

Teorema: $V(X) = \sigma^2 = \pi(1-\pi)$

– Coeficiente de Assimetria: $a_3 = \frac{\mu_3}{\mu_2 \sqrt{\mu_2}}$

onde:

$$\mu_2 = E(X-\mu)^2 = E(X^2) - \mu^2$$
$$\mu_3 = E(X-\mu)^3 = E(X^3) - 3\mu E(X^2) + 2\mu^3$$

Teorema: $a_3 = \frac{(1-\pi) - \pi}{\sqrt{\pi(1-\pi)}}$

3.3.1.2. Distribuição binomial

Definição: modelo que descreve probabilisticamente os resultados de uma sequência de experimentos de Bernoulli *independentes*, ou seja, onde a probabilidade de sucesso é sempre a mesma.

Podemos dizer que, se

$$X = Y_1 + Y_2 + \ldots + Y_n,$$

onde:

$Y_i \sim$ Ber (π) e $Y_{i's}$ são independentes, então X tem distribuição binomial.

Exemplo: Em uma estância 60% dos bovinos foram vacinados contra uma determinada doença. Se um bovino dessa estância for escolhido ao acaso, então, teremos um experimento de Bernoulli com

S = {vacinado, não vacinado},

onde:

P(vacinado) = 0,6 e P(não vacinado) = 0,4.

Se três bovinos forem escolhidos ao acaso, então teremos uma sequência de três experimentos de Bernoulli independentes uma vez que, a cada escolha, a probabilidade de sucesso permanecerá inalterada. O espaço amostral deste experimento será

S = {VVV, VVN, VNV, NVV, NNV, NVN, VNN, NNN},

onde:

V = vacinado e N = não vacinado.

Se a variável X é definida como o número de sucessos em n experimentos de Bernoulli independentes, com probabilidade de sucesso igual a π, então, no exemplo, onde n = 3 e π = 0,6 (se considerarmos sucesso = vacinado), o espaço amostral da variável X será S_X = {0, 1, 2, 3} e as probabilidades P(X = x) será:

$P(X = 0) = 1 \times \pi^0 \times (1 - \pi)^3 = 1 \times. 0{,}6^0 \times 0{,}4^3 = 0{,}064$

$P(X = 1) = 3 \times \pi^1 \times (1 - \pi)^2 = 3 \times. 0{,}6^1 \times 0{,}4^2 = 0{,}288$

$P(X = 2) = 3 \times \pi^2 \times (1 - \pi)^1 = 3 \times 0{,}6^2 \times 0{,}4^1 = 0{,}432$

$P(X = 3) = 1 \times \pi^3 \times (1 - \pi)^0 = 1 \times 0{,}6^3 \times 0{,}4^0 = 0{,}216$

Sendo assim, a distribuição de probabilidade da variável X será

$X = x$	0	1	2	3	Σ
$P(X = x)$	0,064	0,288	0,432	0,216	1

◆ **Função de probabilidade**

De modo geral, se X é uma variável que tem distribuição binomial, então a sua função de probabilidade será:

$$P(X = x) = P^{x,n-x}\, \pi^x (1-\pi)^{n-x}, \text{ para } S_X = \{0, 1, ..., n\}$$

◆ **Parâmetros**

A distribuição binomial tem dois parâmetros:

n = número de repetições do experimento de Bernoulli
π = probabilidade de sucesso

Dizemos, então, que

$X \sim$ Bin (n, π).

◆ **Medidas descritivas**

– Média ou valor esperado: $E(X) = \mu = \sum_{x \in S_X} x\,p(x)$

Considerando que uma variável X com distribuição binomial pode ser definida como a soma de variáveis de n variáveis Y independentes, podemos utilizar as propriedades da média.

Sendo $X = Y_1 + Y_2 + \ldots + Y_n$,

temos

$$E(X) = E(Y_1 + Y_2 + \ldots + Y_n)$$
$$E(X) = E(Y_1) + E(Y_2) + \ldots + E(Y_n)$$
$$E(X) = \pi + \pi + \ldots + \pi = n\pi$$

Teorema: $E(X) = \mu = n\pi$

– Variância: $V(X) = \sigma^2 = E(X^2) - \mu^2$

Sendo $X = Y_1 + Y_2 + \ldots + Y_n$,

temos

$$V(X) = V(Y_1 + Y_2 + \ldots + Y_n)$$
$$V(X) = V(Y_1) + V(Y_2) + \ldots + V(Y_n)$$
$$V(X) = \pi(1-\pi) + \pi(1-\pi) + \ldots + \pi(1-\pi) = n\pi(1-\pi)$$

Teorema: $V(X) = \sigma^2 = n\pi(1-\pi)$.

– Coeficiente de assimetria: $a_3 = \frac{\mu_3}{\mu_2\sqrt{\mu_2}}$

Teorema: $a_3 = \frac{(1-\pi)-\pi}{\sqrt{n\pi(1-\pi)}}$

Interpretação do coeficiente de assimetria:

Se $\pi > (1-\pi)$, a distribuição binomial é assimétrica negativa.
Se $\pi = (1-\pi) = 0{,}5$, a distribuição binomial é simétrica.
Se $\pi < (1-\pi)$, a distribuição binomial é assimétrica positiva.

– Coeficiente de curtose: $a_4 = \frac{\mu_4}{\mu_2^2}$

Fazendo $a_4' = a_4 - 3$, temos:

Teorema: $a_4' = \frac{1-6\pi(1-\pi)}{n\pi(1-\pi)}$

Interpretação do coeficiente de curtose a_4':

Se $a_4' < 0$, a distribuição binomial é platicúrtica.
Se $a_4' = 0$, a distribuição binomial é mesocúrtica.
Se $a_4' > 0$, a distribuição binomial é leptocúrtica.

No exemplo:

$E(X) = 3 \times 0,6 = 1,8$ bolas

Significado do valor esperado: Se o experimento (escolher três bovinos) for repetido um grande número de vezes, o valor esperado será o número médio de sucessos (bovinos vacinados) obtidos nesses experimentos.

$V(X) = 3 \times 0,6 \times 0,4 = 0,72$ bolas^2

Significado da variância: Se o experimento (escolher três bovinos) for repetido um grande número de vezes, a variância expressará a variação média do número de sucessos (bovinos vacinados) obtidos nesses experimentos em relação ao valor esperado.

$$a_3 = \frac{0,4 - 0,6}{\sqrt{3 \times 0,6 \times 0,4}} = -0,24 \rightarrow \text{ distribuição assimétrica negativa}$$

Significado do coeficiente de assimetria: A probabilidade de ocorrer valores maiores que a média (1,8) é maior que a probabilidade de ocorrer valores menores.

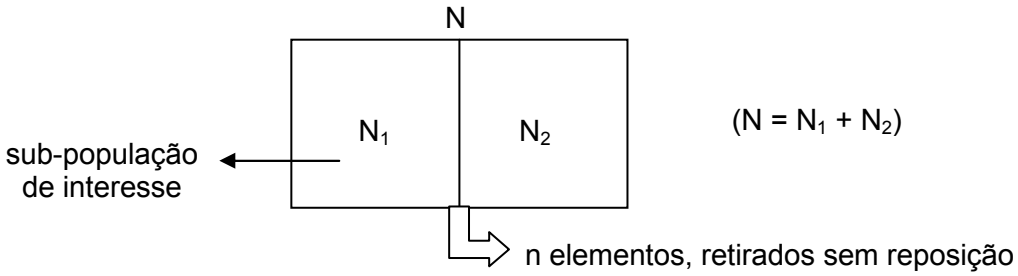
$$a'_4 = \frac{1 - 6 \times 0,6 \times (1 - 0,6)}{3 \times 0,6 \times (1 - 0,6)} = -0,61 \rightarrow \text{ distribuição platicúrtica}$$

3.3.1.3. Distribuição hipergeométrica

Definição: modelo que descreve probabilisticamente os resultados de uma sequência de experimentos de Bernoulli *dependentes*. Refere-se a experimentos que se caracterizam por retiradas *sem reposição*, ou seja, onde a probabilidade de sucesso *se altera* a cada retirada.

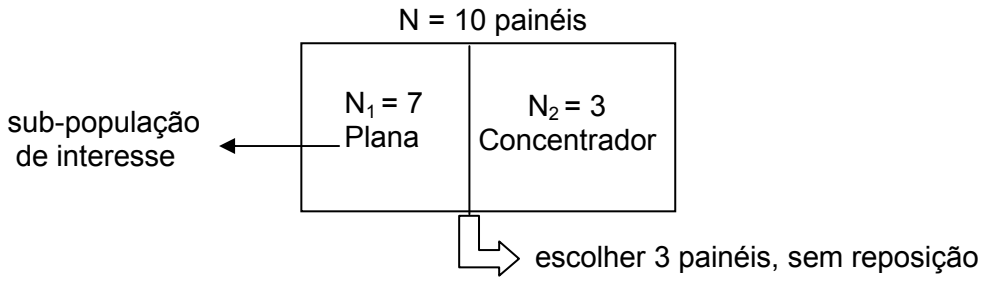
Tais experimentos podem ser descritos genericamente da seguinte forma:

Consideremos uma população de tamanho N, dividida em duas sub-populações de tamanho N_1 e N_2. Suponha que desejamos retirar dessa população um grupo de n elementos, um a um, sem reposição. Se a variável aleatória X é definida como o número de elementos da sub-população de tamanho N_1, observa-se uma relação de dependência entre os elementos retirados, pois, como não há reposição, a probabilidade de sucesso (retirar elemento da sub-população de tamanho N_1) muda a cada retirada.



X = número de elementos da sub-população (de interesse) de tamanho N_1

Exemplo: Dentre 10 painéis solares apresentados numa exposição, sete são do tipo placa plana e três são do tipo concentrador. Uma pessoa que visita a exposição escolhe, ao acaso, três painéis para observar. Se a variável aleatória X é definida como o número de painéis do tipo placa plana observados, construa a distribuição de probabilidade de X.



$$S = \{C_1C_2C_3,\ C_1C_2P_1,\ C_1C_2P_2,\ ...,\ P_5P_6P_7\}$$

$$\#S = C_{10}^3$$

X = número de painéis do tipo placa plana observados

$$S_X = \{0, 1, 2, 3\}$$

$$P(X=0) = \frac{C_7^0\, C_3^3}{C_{10}^3} = \frac{1\times 1}{120} = \frac{1}{120} = 0{,}008333$$

$$P(X=1) = \frac{C_7^1\, C_3^2}{C_{10}^3} = \frac{7\times 3}{120} = \frac{21}{120} = 0{,}175$$

$$P(X=2) = \frac{C_7^2\, C_3^1}{C_{10}^3} = \frac{21\times 3}{120} = \frac{63}{120} = 0{,}525$$

$$P(X=3) = \frac{C_7^3\, C_3^0}{C_{10}^3} = \frac{35\times 1}{120} = \frac{35}{120} = 0{,}2917$$

Sendo assim, a distribuição de probabilidade da variável X será

$X = x$	0	1	2	3	Σ
$P(X = x)$	1/120	21/120	63/120	35/120	1

◆ **Função de probabilidade**

De modo geral, se X é uma variável que tem distribuição hipergeométrica, então sua função de probabilidade será:

$$P(X=x) = \frac{C_{N_1}^x\, C_{N_2}^{n-x}}{C_N^n}, \text{ para } S_X = \{\max(0, n - N_2), ..., \min(n, N_1)\}$$

◆ **Parâmetros**

A distribuição hipergeométrica tem três parâmetros:
N = tamanho da população
N_1 = número de elementos da sub-população de interesse
n = número de elementos retirados (repetições do experimento de Bernoulli)

Dizemos, então, que

$$X \sim \text{Hip}(N, N_1, n).$$

◆ **Medidas descritivas**

– Média ou valor esperado: $E(X) = \mu = \sum_{x \in S_X} x\,p(x)$

Teorema: $E(X) = \mu = n\,\frac{N_1}{N}$

– Variância: $V(X) = \sigma^2 = E(X^2) - \mu^2$

Teorema: $V(X) = \sigma^2 = n\,\frac{N_1}{N}\,\frac{N_2}{N}\left(\frac{N-n}{N-1}\right)$,

onde:

$\frac{N-n}{N-1}$ é o fator de correção para populações finitas.

Agora torna-se necessário definirmos mais claramente quando uma população é considerada finita.

Entendemos por *população finita* aquela que pode ser esgotada por processo de amostragem. Uma população será considerada finita quando tiver um número finito de elementos e a amostragem for efetuada sem reposição. De maneira análoga, podemos definir uma *população infinita* como aquela que não se esgota por processo de amostragem. Assim, uma população será considerada infinita quando tiver um número infinito de elementos ou quando amostragem for efetuada com reposição.

Para o exemplo, as medidas descritivas são:

$$E(X) = 3 \times \frac{7}{10} = 2{,}1 \text{ painéis}$$

$$V(X) = 3 \times \frac{7}{10} \times \frac{3}{10} \times \left(\frac{10-3}{10-1}\right) = 0{,}49 \text{ painéis}^2$$

3.3.1.4. Distribuição de Poisson

Siméon Denis Poisson
(1781 - 1840)

A distribuição de Poisson, assim designada em homenagem ao matemático e físico francês Siméon Denis Poisson.

Definição: modelo que descreve probabilisticamente a sequência de um *grande número* de fenômenos *independentes* entre si, cada um com probabilidade de sucesso *muito pequena*.

Esta distribuição é importante no estudo de variáveis aleatórias de *ocorrência rara* em relação ao número total de ocorrências, como por exemplo:

- número de peças defeituosas observadas em uma linha de produção num determinado período de tempo;
- número de partículas radioativas emitidas numa unidade de tempo;
- número de cultivares selecionadas num processo de melhoramento;
- número de acidentes de trabalho ocorridos numa grande empresa num determinado período de tempo;
- número de ciclones ocorridos em certa região num determinado período de tempo

A distribuição de Poisson tem inúmeras aplicações na simulação de sistemas modelando o número de eventos ocorridos num intervalo de tempo, quando os eventos ocorrem a uma taxa constante.

◆ **Função de probabilidade**

De modo geral, se X é uma variável que tem distribuição de Poisson, então a sua função de probabilidade será:

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

onde:

X: número de sucessos;
e: número base dos logaritmos neperianos = 2,718 (constante);
λ: número médio de sucessos (sempre maior que zero).

Podemos demonstrar que a função $P(X = x) = e^{-\lambda}\frac{\lambda^x}{x!}$ é uma função de probabilidade provando que $\sum_{x \in S_x} p(x) = 1$. Como S_X = {0, 1, 2, ...}, temos

$$\begin{aligned}
\sum_{x=0}^{\infty} p(x) &= \sum_{x=0}^{\infty} e^{-\lambda}\frac{\lambda^x}{x!} \\
&= e^{-\lambda}\sum_{x=0}^{\infty}\frac{\lambda^x}{x!} \\
&= e^{-\lambda}\left(1 + \lambda + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \ldots\right) \\
&= e^{-\lambda}(e^{\lambda}) \\
&= e^{-\lambda+\lambda} = e^0 = 1
\end{aligned}$$

◆ **Parâmetros**

A distribuição de Poisson tem apenas um parâmetro:

λ: número médio de sucessos

Dizemos, então, que

$$X \sim \text{Poi}(\lambda).$$

◆ **Medidas descritivas**

– Média ou valor esperado: $E(X) = \mu = \sum_{x \in S_X} x\,p(x)$

$$\begin{aligned}
E(X) &= \sum_{x=0}^{\infty} x\,p(x) = \sum_{x=0}^{\infty} x\,e^{-\lambda}\frac{\lambda^x}{x!} \\
&= e^{-\lambda}\sum_{x=0}^{\infty} x\frac{\lambda^x}{x!} = e^{-\lambda}\sum_{x=0}^{\infty} x\frac{\lambda\lambda^{x-1}}{x(x-1)!} = e^{-\lambda}\sum_{x=1}^{\infty} x\frac{\lambda}{x}\frac{\lambda^{x-1}}{(x-1)!} \\
&= e^{-\lambda}\sum_{x=1}^{\infty}\lambda\frac{\lambda^{x-1}}{(x-1)!} = e^{-\lambda}\lambda\sum_{x=1}^{\infty}\frac{\lambda^{x-1}}{(x-1)!}, \text{ fazendo } y = x-1, \text{ temos} \\
&= e^{-\lambda}\lambda\sum_{y=0}^{\infty}\frac{\lambda^y}{y!} = e^{-\lambda}\lambda e^{\lambda} = e^{-\lambda+\lambda}\lambda = e^0\lambda = \lambda
\end{aligned}$$

Teorema: $E(X) = \mu = \lambda$

– Variância: $V(X) = \sigma^2 = E(X^2) - \mu^2$

$$E(X^2) = \sum_{x=0}^{\infty} x^2 p(x)$$

$$= \sum_{x=0}^{\infty} x^2 e^{-\lambda} \frac{\lambda^x}{x!} = e^{-\lambda} \sum_{x=0}^{\infty} x^2 \frac{\lambda^x}{x!} =$$

$$= e^{-\lambda} \sum_{x=1}^{\infty} x^2 \frac{\lambda}{x} \frac{\lambda^{x-1}}{(x-1)!} = e^{-\lambda}\lambda \sum_{x=1}^{\infty} x \frac{\lambda^{x-1}}{(x-1)!}, \text{ fazendo } y = x-1, \text{ temos}$$

$$E(X^2) = e^{-\lambda}\lambda \sum_{y=0}^{\infty} (y+1) \frac{\lambda^y}{y!} = e^{-\mu}\lambda \sum_{y=0}^{\infty} \left(y \frac{\lambda^y}{y!} + \frac{\lambda^y}{y!} \right)$$

$$= e^{-\lambda}\lambda \left(\sum_{y=0}^{\infty} y \frac{\lambda^y}{y!} + \sum_{y=0}^{\infty} \frac{\lambda^y}{y!} \right) = e^{-\mu}\lambda (\lambda e^{\lambda} + e^{\lambda})$$

$$= \lambda e^{\lambda} e^{-\lambda}\lambda + e^{\lambda} e^{-\lambda}\lambda = e^0\lambda^2 + e^0\lambda = \lambda^2 + \lambda$$

Assim, temos $V(X) = E(X^2) - \mu^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$.

Teorema: $V(X) = \sigma^2 = \lambda$

– Coeficiente de assimetria: $a_3 = \frac{\mu_3}{\mu_2 \sqrt{\mu_2}}$

Teorema: $a_3 = \frac{1}{\sqrt{\lambda}} \rightarrow$ Distribuição assimétrica positiva, tendendo para a simetria quando μ cresce.

– Coeficiente de curtose: $a_4 = \frac{\mu_4}{\mu_2^2}$

Teorema: $a_4 = \sqrt{\lambda} \rightarrow$ Distribuição platicúrtica, tendendo para mesocúrtica quando μ cresce.

3.3.1.5. Formas limite da distribuição binomial

Sob determinadas circunstâncias, uma distribuição de probabilidade pode tender para outra. Os casos mais importantes de aproximações entre distribuições e as circunstâncias em que ocorrem são os seguintes:

1. Hipergeométrica se aproxima da Binomial

Quando N (tamanho da população) é muito grande (ou tende para $+\infty$), a distribuição *hipergeométrica* se aproxima da distribuição *binomial*. Isso ocorre porque o fator de correção para populações finitas tende a 1.

$$\frac{N-n}{N-1} \cong \frac{N-n}{N} = \frac{N}{N} - \frac{n}{N} = 1 - \frac{n}{N} \cong 1$$

De modo geral, esta aproximação é considerada satisfatória quando o número de elementos retirados (n) não excede 5% da população (N), ou seja,

$$n \leq (0{,}05)\, N.$$

2. Binomial se aproxima da Poisson

Quando n (número de repetições do experimento) é muito grande (ou tende para + ∞) e π (probabilidade de sucesso) é muito pequena (ou tende para 0) a distribuição *binomial* se aproxima da distribuição de *Poisson*. Esta aproximação é considerada satisfatória quando

$$n\pi < 10 \quad \text{e} \quad n \geq 100.$$

3. Binomial se aproxima da Normal

Quando n (número de repetições do experimento) é muito grande (ou tende para + ∞) e π (probabilidade de sucesso) se aproxima de 0,5, a distribuição *binomial* se aproxima da distribuição *normal*.

Se $\pi = (1-\pi) = 0{,}5$, então a distribuição binomial será simétrica.

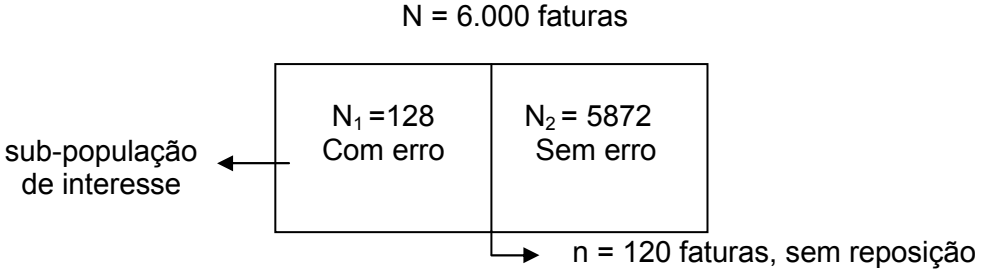
Em alguns casos, a distribuição hipergeométrica pode ser aproximada por uma distribuição binomial e esta binomial pode ser aproximada por uma distribuição de Poisson. Assim, existem situações em que a distribuição hipergeométrica pode ser aproximada por uma distribuição de Poisson.

Consideremos o exemplo seguinte:

Um auditor foi contratado para examinar uma coleção de 6.000 faturas, das quais 128 contêm erros. Se foi selecionada uma amostra de 120 faturas, qual é a probabilidade desta amostra conter exatamente duas faturas com erros?

Resolução:

As características do experimento evidenciam que a variável X = número de sucessos (faturas com erros) tem distribuição hipergeométrica com os seguintes parâmetros: N = 6.000, N_1 = 128 e n = 120.



Como N é suficientemente grande, pois

$$120 < 0{,}05 \times 6.000 = 300,$$

é razoável utilizarmos a aproximação binomial, cujos parâmetros são:

$$n = 120 \quad e \quad \pi = \frac{N_1}{N} = \frac{128}{6000} = 0{,}02133.$$

Como n pode ser considerado suficientemente grande e π suficientemente pequeno, uma vez que

$$n = 120 > 100 \quad e \quad n\pi = 120 \times 0{,}02133 = 2{,}56 < 10,$$

também é razoável utilizarmos a aproximação de Poisson, com parâmetro

$$\mu = E(X) = n\pi = 2{,}56.$$

Daí temos

$$P(X = 2) = e^{-\mu}\frac{\mu^x}{x!} = e^{-2{,}56}\frac{2{,}56^2}{2!} = 0{,}0773 \times \frac{6{,}5536}{2} \cong 0{,}2533.$$

Verificamos, assim, que nas situações que envolvem grandes valores de n e valores ainda maiores de N, a distribuição de Poisson torna-se bastante útil.

Exercícios propostos:

3.8. A probabilidade de um atirador acertar o alvo é de 0,25. Se quatro atiradores atiram, qual é a probabilidade do alvo ser atingido?

3.9. A taxa média de chegada de clientes em um posto de serviços é de 0,5 por minuto. Calcular a probabilidade de, em um dado minuto, chegarem dois clientes.

3.10. Sendo de 1% o percentual de canhotos numa população, qual é a probabilidade de haver apenas um canhoto numa classe de 30 alunos?

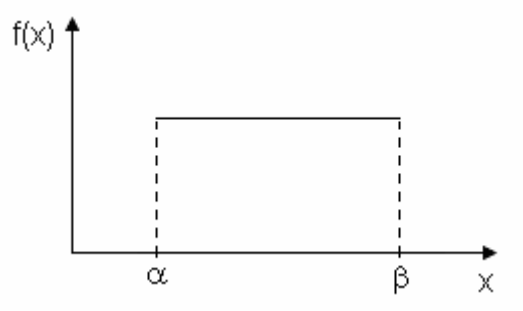
3.3.2. Distribuições de probabilidade de variáveis contínuas

É difícil identificar o tipo de distribuição de probabilidade de uma variável contínua. Geralmente é necessário fazer uma pesquisa bibliográfica para saber se a variável de interesse já foi estudada antes e a sua distribuição de probabilidade já foi identificada. Uma das formas de identificar o tipo de distribuição de uma variável contínua é observando o campo de variação desta variável.

Existem vários tipos de distribuições contínuas, dentre as quais podemos citar: Uniforme, Normal, Exponencial, Gama, Beta, Lognormal, Weibull, Gumbel. Aqui trataremos apenas das três primeiras, consideradas as mais importantes.

3.3.2.1. Distribuição uniforme

Definição: Seja X uma variável aleatória contínua que assume valores no intervalo $[\alpha, \beta]$. Se a probabilidade de X assumir valores num subintervalo é a mesma que para qualquer outro subintervalo de mesmo comprimento, então, esta variável tem distribuição uniforme.



- **Função densidade de probabilidade**

De modo geral, se X é uma variável aleatória contínua que tem distribuição uniforme, então sua função densidade de probabilidade será:

$$f(x) = \begin{cases} \dfrac{1}{\beta - \alpha}, & \text{para } \alpha \le x \le \beta \\ 0, & \text{em caso contrário} \end{cases}$$

Podemos demonstrar que é uma função densidade de probabilidade provando que $\int_{S_x} f(x)dx = 1.$

$$\int_{S_x} f(x)dx = \int_{\alpha}^{\beta} \left(\frac{1}{\beta - \alpha} \right) dx = \left[\frac{x}{\beta - \alpha} \right]_{\alpha}^{\beta} = \frac{\beta}{\beta - \alpha} - \frac{\alpha}{\beta - \alpha} = \frac{\beta - \alpha}{\beta - \alpha} = 1$$

- **Parâmetros**

A distribuição uniforme tem dois parâmetros:

α: menor valor para o qual a variável X está definida;
β: maior valor para o qual a variável X está definida.

Dizemos, então, que

$$X \sim U(\alpha, \beta).$$

◆ **Medidas descritivas**

– Média ou valor esperado: $E(X) = \mu = \int_{S_x} x f(x) dx$

$$E(X) = \mu = \int_{S_x} x f(x) dx$$
$$= \int_{\alpha}^{\beta} x \left(\frac{1}{\beta - \alpha} \right) dx = \frac{1}{\beta - \alpha} \left[\frac{x^2}{2} \right]_{\alpha}^{\beta}$$
$$= \frac{1}{\beta - \alpha} \left(\frac{\beta^2 - \alpha^2}{2} \right) = \frac{(\beta - \alpha)(\beta + \alpha)}{2(\beta - \alpha)} = \frac{(\beta + \alpha)}{2}$$

Teorema: $E(X) = \mu = \frac{(\beta + \alpha)}{2}$

– Variância: $V(X) = \sigma^2 = E(X^2) - \mu^2$

$$E(X^2) = \int_{S_x} x^2 f(x) dx$$
$$= \int_{\alpha}^{\beta} x^2 \left(\frac{1}{\beta - \alpha} \right) dx = \frac{1}{\beta - \alpha} \left[\frac{x^3}{3} \right]_{\alpha}^{\beta}$$
$$= \frac{1}{\beta - \alpha} \left(\frac{\beta^3 - \alpha^3}{3} \right) = \frac{\beta^3 - \alpha^3}{3(\beta - \alpha)}$$

Assim, temos

$$V(X) = \sigma^2 = E(X^2) - \mu^2 = \frac{\beta^3 - \alpha^3}{3(\beta - \alpha)} - \frac{(\beta + \alpha)^2}{4}$$
$$= \frac{4(\beta^3 - \alpha^3) - 3(\beta - \alpha)(\beta^3 + \alpha^3)^2}{12(\beta - \alpha)}$$
$$= \frac{4\beta^3 - 4\alpha^3 + 3\alpha^3 + 3\alpha^2\beta - 3\alpha\beta^2 - 3\beta^3}{12(\beta - \alpha)}$$
$$= \frac{\beta^3 - \alpha^3 + 3\alpha^3\beta - 3\alpha\beta^2}{12(\beta - \alpha)} = \frac{(\beta - \alpha)^3}{12(\beta - \alpha)} = \frac{(\beta - \alpha)^2}{12}$$

Teorema: $V(X) = \sigma^2 = \frac{(\beta - \alpha)^2}{12}$

◆ **Função de distribuição acumulada**

A função de distribuição acumulada da uniforme é facilmente encontrada:

$$F(x) = P(X \le x) = \int_{-\infty}^{x} f(x)\, dx = \begin{cases} 0, & \text{se } x < \alpha \\ \frac{x - \alpha}{\beta - \alpha}, & \text{se } \alpha \le x \le \beta \\ 1, & \text{se } x > \beta \end{cases}$$

Vejamos um exemplo resolvido.

Seja X uma variável aleatória contínua com distribuição uniforme no intervalo [5, 10].
Determinar as probabilidades:
a) P(X < 7)
b) P(X > 8,5)
c) P(8 < x < 9)
d) P(|x - 7,5| > 2)

Resolução:

Utilizando a função de distribuição acumulada:

a) $P(X < 7) = F(7) = \frac{7-5}{10-5} = \frac{2}{5} = 0,4$

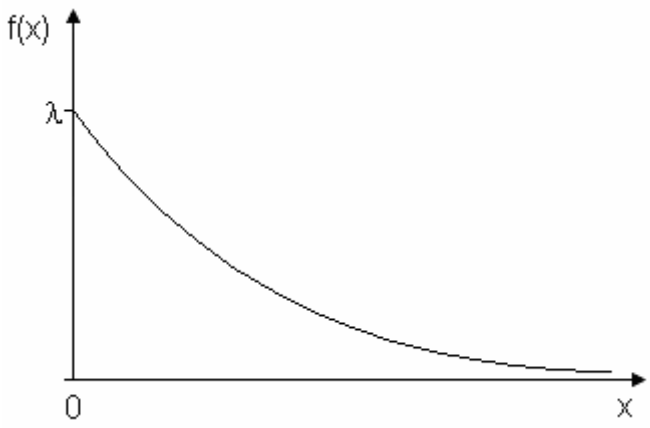
b) $P(X < 8,5) = 1 - F(8,5) = 1 - \frac{8,5-5}{10-5} = 1 - \frac{3,5}{5} = 1 - 0,7 = 0,3$

c) $P(8 < X < 9) = F(9) - F(8) = \frac{9-5}{10-5} - \frac{8-5}{10-5} = \frac{4-3}{5} = \frac{1}{5} = 0,2$

d) $P(|X-7| > 2) = P(X - 7,5 > 2 \text{ ou } X - 7,5 < -2)$
$= P(X > 9,5 \text{ ou } X < 5,5) = 1 - F(9,5) + F(5,5)$
$= 1 - \frac{9,5-5}{10-5} + \frac{5,5-5}{10-5} = 0,1 - 0,1 = 0,2$

3.3.2.2. Distribuição exponencial

Definição: Seja X uma variável aleatória contínua que só assume valores não negativos. Se esta variável é o tempo decorrido entre ocorrências sucessivas de um processo de Poisson, então ela tem distribuição exponencial.



Na distribuição de Poisson, a variável aleatória é definida como o número de ocorrências (sucessos) em determinado período de tempo, sendo a média das ocorrências no período definida como λ. Na distribuição exponencial, a variável aleatória é definida como o tempo entre duas ocorrências, sendo a média de tempo entre ocorrências igual a $1/\lambda$. Por exemplo, se a média de atendimentos no caixa de uma loja é de $\lambda = 6$ clientes/min, então o tempo médio entre atendimentos é $1/\lambda = 1/6$ de minuto ou 10 segundos.

A distribuição exponencial é muito utilizada no campo da confiabilidade para a modelagem do tempo até a ocorrência de falha em componentes eletrônicos, bem como do tempo de espera em sistemas de filas.

◆ **Função densidade de probabilidade**

De modo geral, se X é uma variável aleatória contínua que tem distribuição exponencial, então sua função densidade de probabilidade será:

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{para } x>0 \\ 0, & \text{em caso contrário} \end{cases}$$

Podemos demonstrar que é uma função densidade de probabilidade provando que $\int_{S_x} f(x)\,dx = 1$.

$$\int_{S_x} f(x)\,dx = \int_0^{\infty} \lambda\, e^{-\lambda x} dx = \left[e^{-\lambda x} \right]_0^{\infty} = e^{-\lambda\infty} - e^{-\lambda 0} = 0 - (-1) = 1$$

◆ **Parâmetros**

A distribuição exponencial tem apenas um parâmetro:

λ: número médio de ocorrências em determinado período de tempo ($\lambda>0$);

Dizemos, então, que

$$X \sim Exp(\lambda).$$

◆ **Medidas descritivas**

– Média ou valor esperado: $E(X) = \mu = \int_{S_x} x\, f(x) dx$

$$E(X) = \mu = \int_{S_x} x\, f(x)\,dx$$
$$= \int_0^{\infty} x\lambda\, e^{-\lambda x} dx = \frac{1}{\lambda}$$

Teorema: $E(X) = \mu = \frac{1}{\lambda}$

– Variância: $V(X) = \sigma^2 = E(X^2) - \mu^2$

$$V(X) = \sigma^2 = \left[\int_{S_x} x^2 f(x) dx \right] - \mu^2$$
$$= \left[\int_0^{\infty} x^2 \left(\lambda\, e^{-\lambda x} \right) dx \right] - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$$

Teorema: $V(X) = \sigma^2 = \frac{1}{\lambda^2}$

◆ **Função de distribuição acumulada**

A função de distribuição acumulada da exponencial pode ser facilmente encontrada:

$$F(x) = P(X \le x) = \int_{-\infty}^{x} f(x)\,dx = \begin{cases} 0, & \text{se } x < 0 \\ 1 - e^{-\lambda x}, & \text{se } x \ge 0 \end{cases}$$

Desta forma, $P(X > x) = 1 - F(x) = 1 - (1 - e^{-\lambda x}) = e^{-\lambda x}$.

◆ **Propriedade da distribuição exponencial**

A distribuição exponencial apresenta uma propriedade interessante que é denominada *falta de memória*. Isso significa que a probabilidade de ocorrência dos valores de X não é afetada pelo conhecimento da ocorrência de valores anteriores, ou seja,

$$P(X > s + t \mid X > s) = \frac{P(X > s + t \cap X > s)}{P(X > s)} = \frac{P(X > s + t)}{P(X > s)} = \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} = e^{-\lambda t} = P(X > t).$$

Consideremos o seguinte exemplo resolvido:

Suponha que um componente eletrônico tenha um tempo de vida X (em unidades de 1000 horas) que segue uma distribuição exponencial de parâmetro λ = 1. Suponha que o custo de fabricação do item seja 2 reais e que o preço de venda seja 5 reais. O fabricante garante devolução total se X < 0,90. Qual o lucro esperado por item?

Resolução:

Neste caso, temos $f(x) = e^{-x}$, para $x > 0$.

A probabilidade de um componente durar menos de 900 horas é dada por:

$$P(X < 0{,}9) = F(0{,}9) = 1 - e^{-0{,}9} = 0{,}5934$$

Assim, o lucro do fabricante será uma variável aleatória discreta Y com a seguinte distribuição:

$Y = y$	-2	3	Σ
$P(Y = y)$	0,5934	0,4066	1

Então o lucro esperado será:

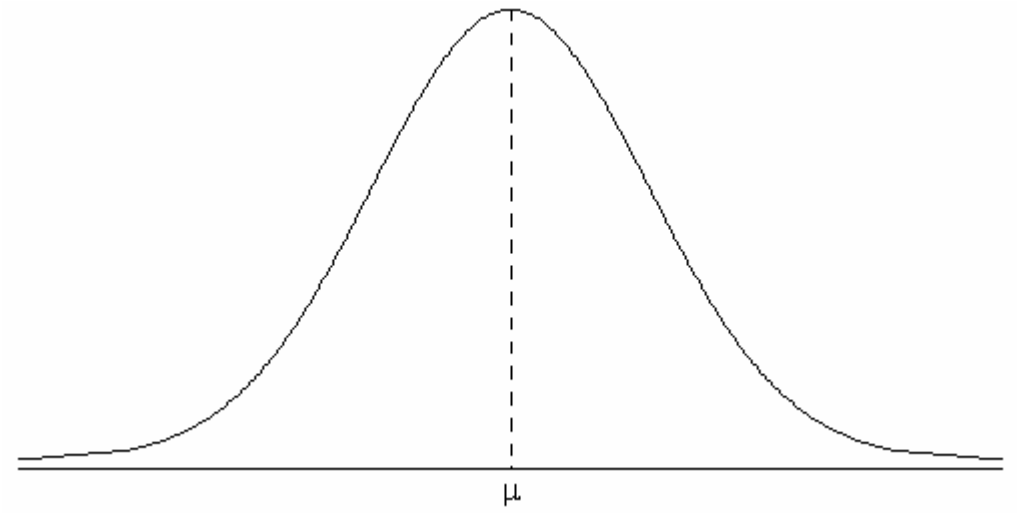
$$E(Y) = -2 \times 0{,}5934 + 3 \times 0{,}4066 = \text{R\$ } 0{,}03$$

3.3.2.3. Distribuição normal

A distribuição normal (ou distribuição de Gauss ou distribuição de Gauss-Laplace) é uma distribuição especialmente importante na metodologia estatística. Sua importância advém das suas propriedades, do número de fenômenos (variáveis) que podem, pelo menos aproximadamente, ser modelados através dela e da quantidade de métodos e técnicas que são derivados tendo-a como pressuposição básica. Esse conjunto de métodos e técnicas forma a chamada Estatística Clássica ou Estatística Paramétrica.

É uma distribuição teórica de frequências, onde a maioria das observações se situa em torno da média (centro da distribuição) e diminui gradual e simetricamente no sentido dos extremos. A distribuição normal é representada graficamente pela curva normal (também chamada curva de Gauss) que tem a forma de sino e é simétrica em relação ao centro, onde se localiza a média μ.

Curva normal



◆ **Função densidade de probabilidade**

De modo geral, se X é uma variável aleatória contínua, X possui distribuição normal se sua função densidade de probabilidade for

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

Parâmetros

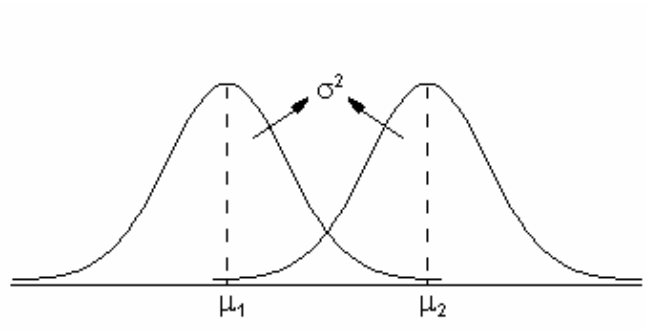
A distribuição normal tem dois parâmetros:

μ = média (determina o centro da distribuição)
σ^2 = variância (determina a dispersão da distribuição)

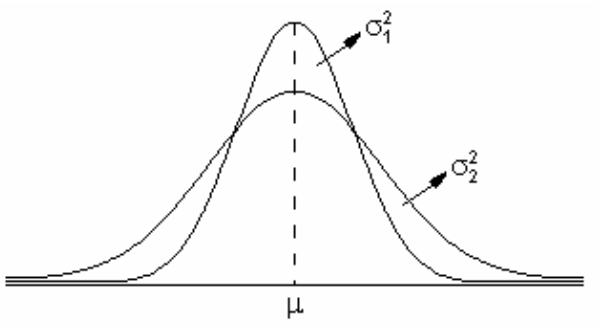
Dizemos, então, que

$$X \sim N(\mu, \sigma^2).$$

Cada vez que um dos parâmetros muda de valor, temos uma curva normal diferente.

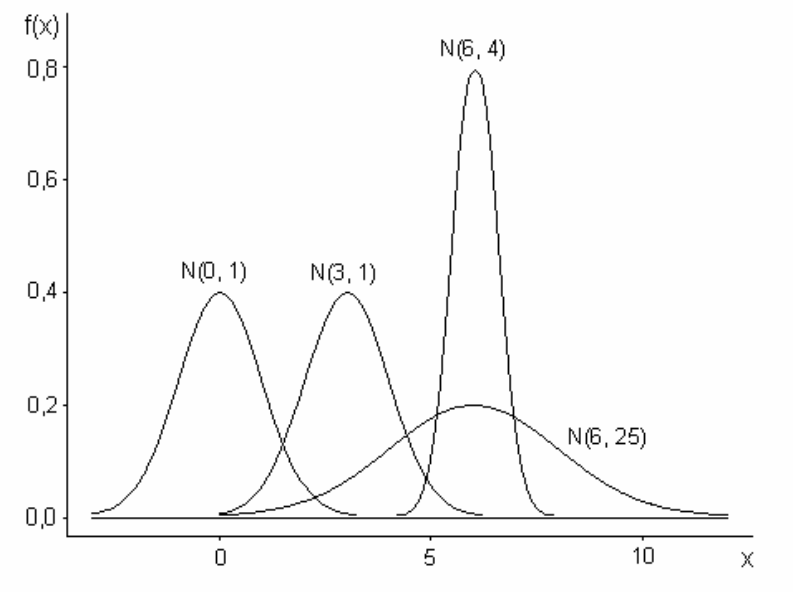


Populações normais com médias diferentes e mesma variância



Populações normais com variâncias diferentes e mesma média

Como consequência, existe um número infinito de curvas normais. Na figura abaixo, podemos observar alguns exemplos de curvas.



Medidas descritivas

– Média ou valor esperado: $E(X) = \mu = \int_{S_x} x\,f(x)dx$

$$E(X) = \mu = \int_{S_x} x\left(\frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}\right)dx$$

– Variância: $V(X) = \sigma^2 = \int_{S_x} (x-\mu)^2 f(x)dx$

$$V(X) = \sigma^2 = \int_{S_x} (x-\mu)^2\left(\frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}\right)dx$$

◆ **Propriedades da distribuição normal**

1. O máximo da função densidade de probabilidade se dá no ponto x = μ.

2. A distribuição é simétrica em relação ao centro onde coincidem a média, a moda e a mediana.

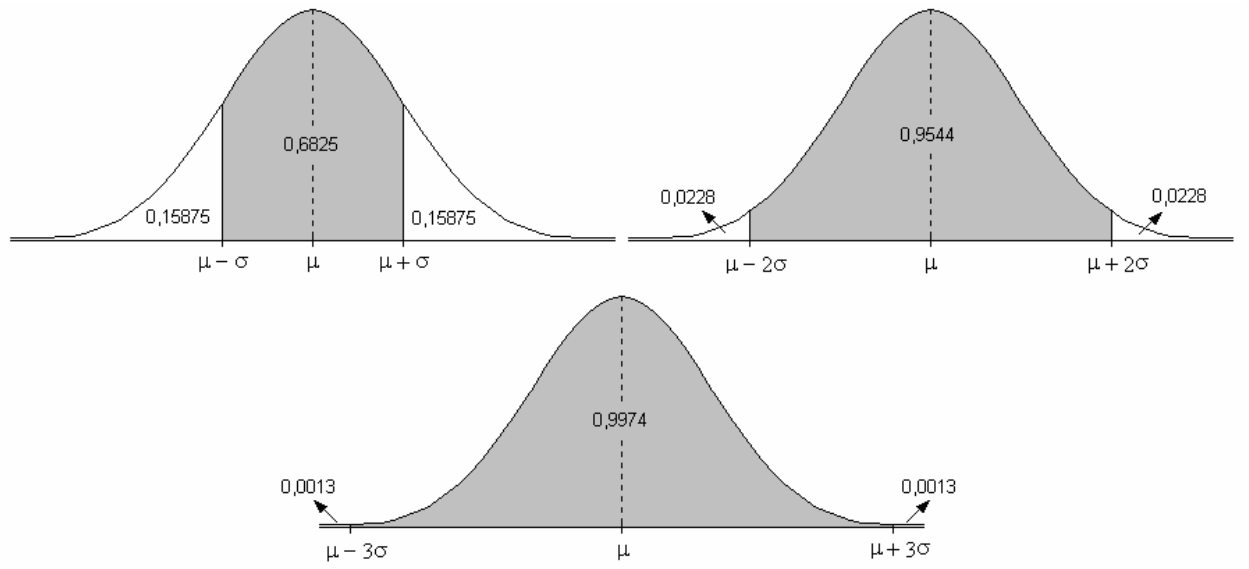
$$\mu = Mo = Md$$

3. Os pontos de inflexão (onde a curva passa de convexa para côncava) são exatamente $\mu-\sigma$ e $\mu+\sigma$.

4. Verifica-se na distribuição normal que:

$$P(\mu-\sigma < X < \mu+\sigma) = 0,6825$$
$$P(\mu-2\sigma < X < \mu+2\sigma) = 0,9544$$
$$P(\mu-3\sigma < X < \mu+3\sigma) = 0,9974$$

Considerando que a área sob a curva no intervalo de interesse é que corresponde a probabilidade, utilizamos as curvas abaixo para ilustrar esta propriedade.



Vimos que, para cada valor de μ e de σ, existe uma distribuição normal diferente. Daí existirem infinitas distribuições (e curvas) normais, pois basta que mude um dos parâmetros para termos outra distribuição. Deste modo, o cálculo de áreas sob a curva normal, frequentemente necessário, deverá ser feito sempre em função dos particulares valores de μ e σ. Para evitar a trabalhosa tarefa de calcular essas áreas todas as vezes que desejássemos obter as probabilidades associadas a uma certa variável X, foi determinada uma distribuição normal *padrão* ou *reduzida*. Através da distribuição normal padrão é possível estudar qualquer variável que tenha distribuição normal, com quaisquer valores para μ e σ.

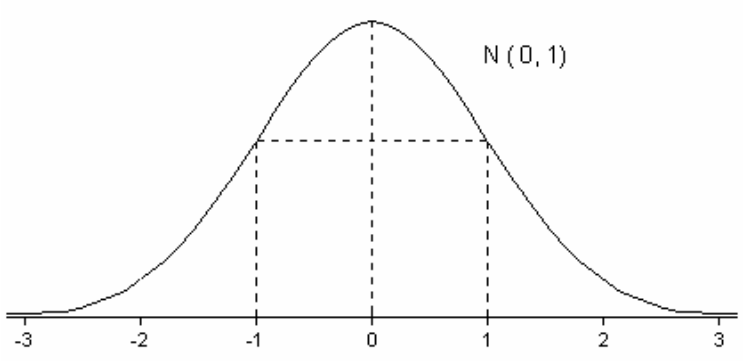
◆ Distribuição normal padrão

Definição: é a distribuição normal de uma variável Z que tem média igual a zero ($\mu = 0$) e desvio padrão igual a um ($\sigma = 1$). Para a variável Z, a função densidade de probabilidade resulta

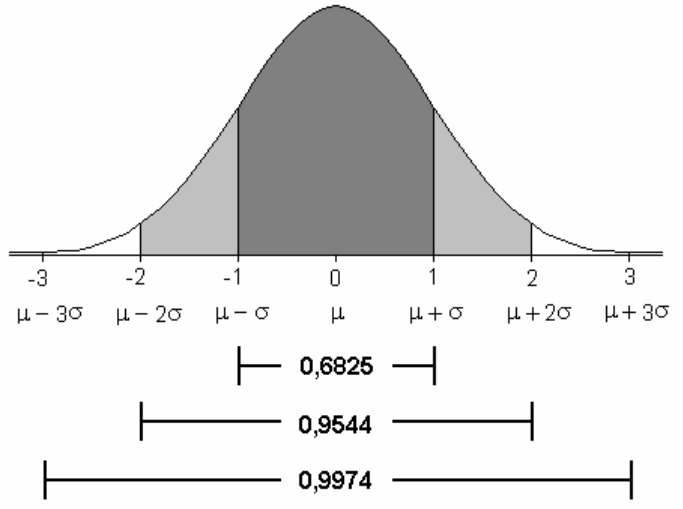
$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}, \quad -\infty < Z < +\infty.$$

A função densidade de probabilidade mais simplificada da distribuição normal padrão, facilitou o cálculo das áreas sob a sua curva. Assim, a curva normal padrão foi dividida em pequenas tiras, cujas áreas foram calculadas e apresentadas numa tabela. Na tabela da

distribuição normal padrão (Tabela I do Apêndice), podemos encontrar as áreas correspondentes aos intervalos de 0 a z.



Os valores negativos não são apresentados na tabela porque a curva é simétrica; portanto, as áreas correspondentes a estes valores são exatamente iguais às dos seus simétricos positivos, por exemplo, P(-1 < Z < 0) = P(0 < Z < 1). Podemos observar também, na tabela da distribuição normal padrão, que os valores de Z vão de 0 a 3,99. Este limite é estabelecido com base na quarta propriedade da distribuição normal, como podemos observar na figura abaixo.



Sabemos que no intervalo [μ-3σ; μ+3σ], que na normal padrão corresponde ao intervalo [-3; 3], temos 99,74 % dos valores de Z. Portanto, como podemos verificar na tabela, a área compreendida entre de 0 e 3,99 já é aproximadamente 0,5.

Veremos agora como a distribuição normal padrão e sua tabela podem ser utilizadas para a obtenção de probabilidades correspondentes a qualquer variável X que tenha distribuição normal.

A distribuição de uma variável X, com quaisquer valores para μ e σ, pode ser obtida pela transformação da variável X na variável Z, através da expressão

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

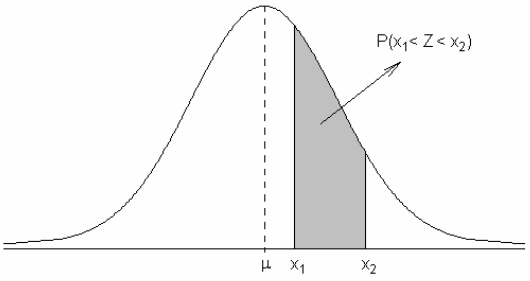
Assim, se x_1 e x_2 são valores de X com distribuição normal e z_1 e z_2 são valores de Z, tais que

$$z_1 = \frac{x_1 - \mu}{\sigma} \quad \text{e} \quad z_2 = \frac{x_2 - \mu}{\sigma},$$

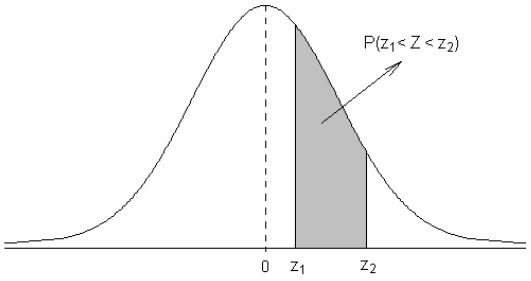
então, $P(x_1 < X < x_2) = P(z_1 < Z < z_2)$.

A relação é evidente, uma vez que a transformação muda as variáveis, mas não altera a área sob a curva, como podemos verificar na figura a seguir.

$X \sim N(\mu, \sigma^2)$



$Z \sim N(0, 1)$



$$P(x_1 < X < x_2) = P(z_1 < Z < z_2)$$

Sendo assim, para utilizar os valores da tabela, devemos transformar X em Z.

$$X \sim N(\mu, \sigma^2) \quad \downarrow \text{ transformar} \quad \rightarrow \quad Z = \frac{X - \mu}{\sigma} \quad Z \sim N(0, 1)$$

Após a transformação, podemos procurar na tabela a área compreendida entre 0 e z, que corresponderá à área entre μ e x.

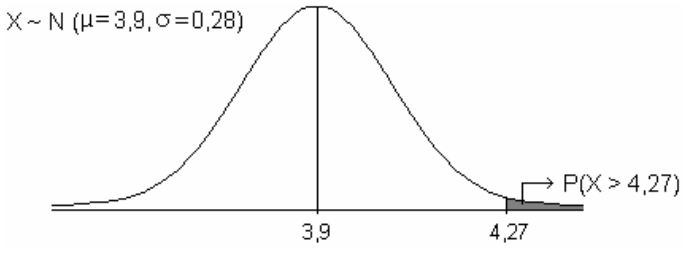
Consideremos o exemplo resolvido a seguir.

Sabendo que as notas de 450 alunos estão normalmente distribuídas, com média μ = 3,9 e desvio padrão σ = 0,28, determine:

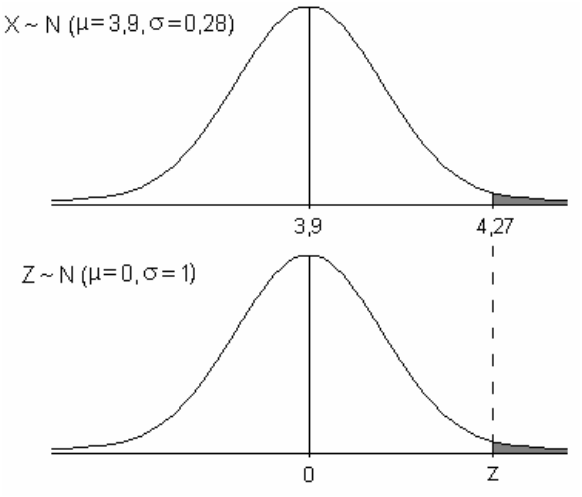
a) a probabilidade de um aluno ter nota maior que 4,27;
b) o número de alunos que têm nota superior a 4,27.

Resolução:

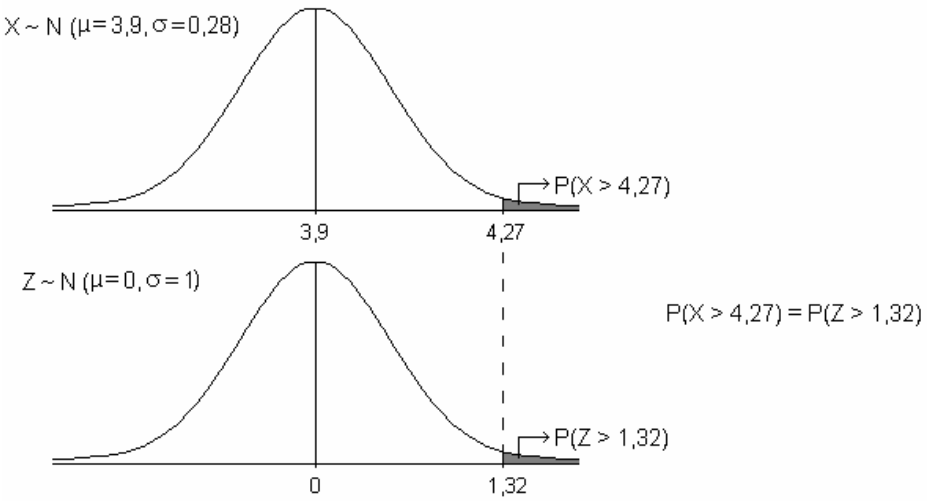
a) Sabemos que a probabilidade de ocorrer um valor dentro de um determinado intervalo corresponde à área sob a função densidade dentro deste intervalo. Sendo assim, para determinar a probabilidade de ocorrer uma nota maior do que 4,27, devemos encontrar a área localizada à direita de 4,27 na curva normal.



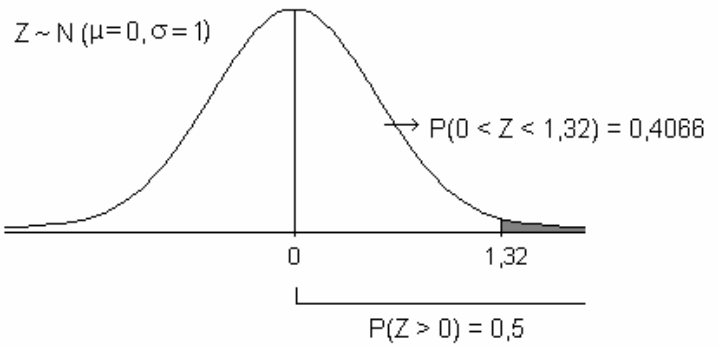
Para encontrar essa área, vamos utilizar a tabela da distribuição normal padrão. Inicialmente, fazemos a transformação da variável X para a variável Z, através da expressão $Z = \frac{X - \mu}{\sigma}$. Desta forma, determinamos o valor de z que corresponde ao valor x = 4,27.



Assim, temos $z = \frac{4,27 - 3,9}{0,28} = 1,32$

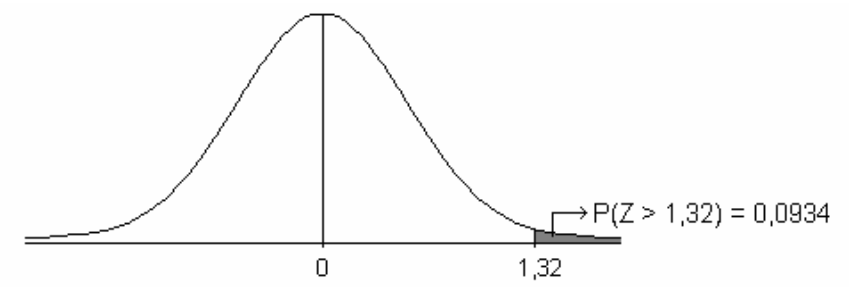


Sabemos que a tabela fornece a área entre 0 e z, portanto, o valor 0,4066, encontrado na tabela para z = 1,32, expressa a área compreendida entre 0 e 1,32. Como a área que nos interessa é a área à direita de 1,32 e sabemos que a área correspondente à metade da curva é 0,5, podemos encontrar a área de interesse calculando a diferença entre essas duas áreas.

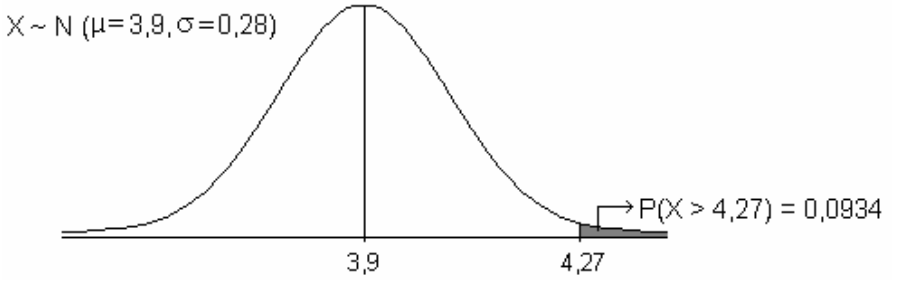


Assim, fazemos

$P(Z > 1,32) = P(Z > 0) – P(0<Z<1,32)$
$P(Z > 1,32) = 0,5 – 0,4066$
$P(Z > 1,32) = 0,0934$



Sabendo que a área à direita de z = 1,32 e igual à área à direita de x = 4,27, concluímos que a probabilidade de Z ser maior que 1,32 é igual à probabilidade de X ser maior que 4,27. Sendo assim, a probabilidade de um aluno tirar uma nota acima de 4,27 é de 0,0934 ou 9,34%, ou seja, $P(X > 4,27) = 0,0934$.



b) Para determinar o número de indivíduos que têm nota superior a 4,27, devemos saber qual é o percentual da população que têm nota acima de 4,27. No item a, vimos que este percentual é de 9,34%. Sendo assim, através de uma regra de três simples, podemos determinar quantos estudantes correspondem a 9,34% de uma população de 450 estudantes. Esse valor pode ser obtido facilmente multiplicando o tamanho da população pela probabilidade de ocorrer uma nota maior que 4,27. Assim, temos:

$$450 \times 0,0934 = 42,03$$

Concluímos, então, que, dos 450 estudantes, 42 têm nota superior a 4,27.

Exercícios propostos:

3.11. Uma variável X é uniformemente distribuída no intervalo [10, 20]. Determine:
a) valor esperado e variância de X;
b) P(12,31 < X < 16,50).

3.12. Os tempos até a falha de um dispositivo eletrônico seguem o modelo exponencial, com uma taxa de falha λ = 0,012 falhas/hora. Indique qual a probabilidade de um dispositivo escolhido ao acaso sobreviver a 50 horas? E a 100 horas?

3.13. Suponha que um mecanismo eletrônico tenha um tempo de vida X (em unidades de 1000 horas) que é considerado uma variável aleatória com função densidade de probabilidade dada por:

$f(x) = e^{-x}$, $x > 0$

$= 0$, em caso contrário.

Suponha ainda que o custo de fabricação de um item seja 2 reais e o preço de venda seja 5 reais. O fabricante garante total devolução se $X \leq 0,8$. Qual o lucro esperado por item?

3.14. Seja Z uma variável aleatória com distribuição normal padrão. Determine as seguintes probabilidades:

a) P(0 < Z < 1,73)
b) P(0,81 < Z < + ∞)
c) P(-1,25 $\leq$ Z $\leq$ -0,63)

3.15. Suponha que a estatura de recém-nascidos do sexo feminino é uma variável com distribuição normal de média μ = 48 cm e σ = 3 cm. Determine:

a) a probabilidade de um recém-nascido ter estatura entre 42 e 49 cm;
b) a probabilidade de um recém-nascido ter estatura superior a 52 cm;
c) o número que de recém-nascidos que têm estatura inferior à $\mu+\sigma$cm, dentre os 532 que nasceram numa determinada maternidade, no período de um mês.

3.16. Suponha que as notas de uma prova sejam normalmente distribuídas, com média μ=72 e desvio padrão σ=1,3. Considerando que 18% dos alunos mais adiantados receberam conceito "A" e 10% dos mais atrasados o conceito "R", encontre a nota mínima para receber "A" e a máxima para receber "R".

3.3. Bibliografia

COSTA, S.F. **Introdução Ilustrada à Estatística (com muito humor!)**. 2.ed., São Paulo: Harbra, 1992. 303p.

DEVORE, J. **Probability and statistics for engineering and the sciences** Brooks/Cole Publishing Companig. 1982. 640p.

FARIA, E.S. de **Estatística** Edição 97/1. (Apostila)

FERREIRA, D.F. **Estatística Básica**. Lavras: Editora UFLA, 2005, 664p.

FREUND, J.E., SIMON, G.A. **Estatística Aplicada. Economia, Administração e Contabilidade**. 9.ed., Porto Alegre: Bookman, 2000. 404p.

GUEDJ, D. **O teorema do papagaio**. São Paulo: Companhia das Letras, 2000. 501p.

MEYER, P. L. **Probabilidade: aplicações à estatística**. Rio de Janeiro: LTC, 1976.

PIMENTEL GOMES, F. **Iniciação à Estatística** São Paulo: Nobel, 1978. 211p.

SILVEIRA JÚNIOR, P., MACHADO, A.A., ZONTA, E.P., SILVA, J.B. da **Curso de Estatística**. v.1, Pelotas: Universidade Federal de Pelotas, 1989. 135p.

SILVEIRA JÚNIOR, P., MACHADO, A.A., ZONTA, E.P., SILVA, J.B. da. **Curso de Estatística**.v.2, Pelotas: Universidade Federal de Pelotas, 1992. 234p.

SPIEGEL, M.R. **Estatística**. São Paulo: McGraw-Hill, 1972. 520p.

VIEIRA, S. **Estatística Experimental**. 9.ed., São Paulo: Atlas, 1999. 185p.

Unidade IV
Inferência Estatística

Unidade IV. Inferência Estatística

4.1. Introdução e histórico

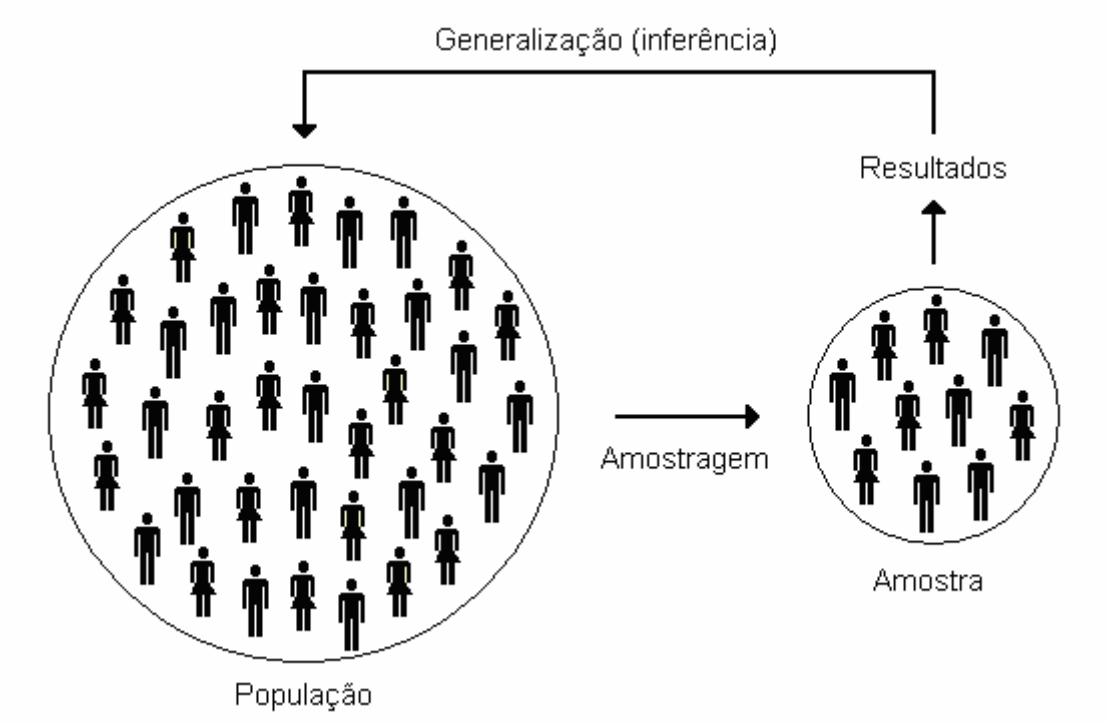
Estudamos na Unidade II as técnicas para resumir e descrever variáveis associadas a conjuntos de dados obtidos de populações inteiras ou de subconjuntos de populações. Na Unidade III, vimos como construir modelos probabilísticos, identificados por parâmetros, capazes de representar adequadamente o comportamento de algumas variáveis. Nesta unidade veremos os fundamentos teóricos para fazer afirmações sobre características de uma população com base em informações fornecidas por amostras.

"Não é preciso beber toda a garrafa para saber se o vinho é bom". Esta frase bastante popular ilustra melhor do que qualquer exemplo técnico o conceito de inferência estatística: dar informação sobre o *todo*, com base no conhecimento da *parte*. O uso de informações parciais para concluir sobre o todo faz parte do cotidiano da maioria das pessoas. Basta observar como a cozinheira prova a comida que está preparando ou como um comprador experimenta um pedaço de laranja na feira antes de decidir se vai comprar as laranjas ou não. Essas decisões são baseadas em procedimentos amostrais.

Na pesquisa científica, em geral, o processo também é esse. Levantamentos amostrais e experimentos são feitos com *amostras*, mas o pesquisador não quer suas conclusões restritas à amostra com a qual trabalhou, ao contrário, o ele quer estender os resultados que obteve para toda a *população*. Assim, o pesquisador quer fazer *inferência*.

Podemos conceituar a Inferência Estatística como o conjunto de procedimentos estatísticos que têm por finalidade generalizar conclusões de uma amostra para uma população.

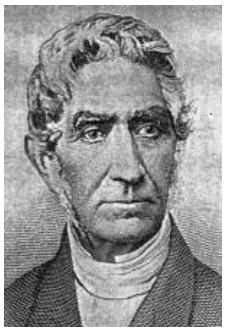
Outro exemplo da aplicação dos métodos de inferência estatística, presente em nosso dia a dia, são as pesquisas eleitorais. Vejamos o esquema da abaixo.



Para poder generalizar as conclusões obtidas da amostra para a população, não basta saber descrever convenientemente os dados da amostra, é preciso garantir que o processo de amostragem seja eficiente, ou seja, que a amostra seja representativa da população. Isto significa que a amostra deve possuir as mesmas características básicas da população no que diz respeito às variáveis que desejamos pesquisar.

A partir desta generalização surge o conceito fundamental de *erro provável*. A possibilidade de erro é inerente ao processo de inferência, ou seja, sempre que estudamos uma população a partir de uma amostra, existe a possibilidade de cometermos algum tipo de erro de conclusão. A grande aplicação da Inferência Estatística é fornecer métodos que permitam quantificar esse erro provável.

Um pouco de história...

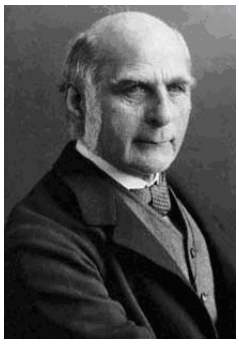


Lambert Quételet
(1796 - 1855)

O casamento entre a Estatística e o cálculo das probabilidades se deve ao astrônomo belga *Lambert Adolphe Jacques Quételet*, que, através de estudos na área social, mostrou que muitos fenômenos vivos apresentavam um comportamento regular. A expressão matemática dessa regularidade é conhecida, hoje, como distribuição de probabilidade.

Após Quételet, a Estatística teve um desenvolvimento sem precedentes, sendo o fenômeno da regularidade observado em muitos campos de pesquisa. As distribuições de probabilidade começaram a ser deduzidas, aumentando ainda mais o campo de aplicação.

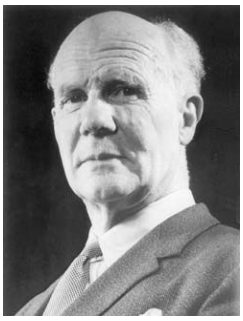
Karl Pearson
(1857 - 1936)



Francis Galton
(1822 - 1911)

O biólogo inglês *Francis Galton*, estudando a hereditariedade do caráter estatura na espécie humana, foi o primeiro a empregar o termo regressão para designar o fenômeno de retorno à média.

Em 1890, o inglês *Karl Pearson*, estimulado pelos trabalhos de Francis Galton, iniciou o estudo sobre relacionamento entre variáveis e, em 1900, deduziu a distribuição Qui-quadrado. Em 1908, o inglês *William Gosset*, aluno de Pearson, descobriu a distribuição t no intuito de resolver problemas relativos a pequenas amostras.



Egon Pearson
(1895 - 1980)

Alguns anos mais tarde, outro inglês, *Ronald Aylmer Fisher*, trouxe contribuições extremamente valiosas à Estatística. Fisher, com os resultados de Gosset, rapidamente descobriu as distribuições amostrais dos coeficientes de correlação, regressão, correlação múltipla e a distribuição da razão entre duas variâncias. Foi ele também quem estendeu e deu ideia mais precisa à técnica chamada Análise da Variação, até hoje uma das mais poderosas utilizadas na Estatística. Fisher trabalhou por quatorze anos na Estação Experimental de Rothamstead, Inglaterra, e, devido aos trabalhos que lá desenvolveu, é considerado o pai da Estatística Experimental.

Jerzy Neyman
(1894 - 1981)

Ronald Fisher
(1890 - 1962)

Após 1925, emergiram dois campos de extrema importância na Inferência Estatística, considerados os pilares da ciência: a teoria dos testes de hipóteses, sob inspiração de Egon Sharpe Pearson e Jerzy Neyman, e a teoria da estimação de parâmetros, desenvolvida por Fisher.

4.2. Conceitos fundamentais

Inicialmente, veremos alguns conceitos fundamentais da Inferência Estatística.

◆ **População** é o conjunto de todos os indivíduos ou elementos que atendem a determinadas características definidoras. Estas características dependem do objetivo do estudo.

Exemplos:

1. Pesquisa eleitoral no Rio Grande do Sul.
   Objetivo: Conhecer a preferência eleitoral no estado.
   População: Todos os eleitores que votam no RS.

2. Pesquisa sócio-econômica na Universidade Federal de Pelotas.
   Objetivo: Estimar a renda média das famílias dos estudantes da UFPel.
   População: Todos os estudantes da UFPel.

◆ **Amostra** é um subconjunto retirado da população com o objetivo de representá-la.

Exemplos:

1. Pesquisa eleitoral no Rio Grande do Sul.
   Amostra: Conjunto de 1.000 a 2.000 eleitores votantes no RS que serão entrevistados pelos pesquisadores.

2. Pesquisa sócio-econômica na Universidade Federal de Pelotas.
   Amostra: Conjunto de 200 estudantes da UFPel que serão entrevistados pelos pesquisadores.

◆ **Amostragem** é o método de seleção que empregamos para obtenção de amostras. Podemos distinguir dois tipos de amostragem: probabilística e não probabilística. A amostra será probabilística se todos os elementos da população tiverem probabilidade conhecida e diferente de zero de participarem da amostra. Caso contrário, a amostragem será não probabilística. A amostragem probabilística é a mais recomendável por garantir a imparcialidade da amostra. Assim, qualquer discrepância entre população e amostra é atribuída ao acaso.

Vejamos a seguir uma breve descrição dos principais tipos de amostragem.

Amostragem probabilística:

• *Amostragem aleatória simples*: considera a população homogênea e consiste num sorteio para a seleção dos elementos que comporão a amostra. Deste modo, todos os elementos da população têm a mesma probabilidade de fazer parte da amostra.

• *Amostragem aleatória estratificada*: é utilizada quando a população pode ser dividida em subgrupos cujos elementos são semelhantes entre si. A amostragem consiste em obter-se de cada grupo uma amostra aleatória. Esse processo pode gerar amostras bastante precisas, mas só é recomendado quando os grupos são homogêneos internamente e heterogêneos entre si, ou seja, a população é heterogênea e as diferenças constituem os estratos (grupos).

• *Amostragem aleatória por conglomerados*: neste caso, a população já é dividida em diferentes grupos (conglomerados) e extraem-se amostras apenas de conglomerados selecionados, e não de toda a população. O ideal é que cada conglomerado represente tanto quanto possível o total da população. Na prática, selecionam-se os conglomerados geograficamente. Escolhem-se aleatoriamente algumas regiões, em seguida algumas sub-regiões e finalmente, alguns lares. Esse processo possibilita ao pesquisador entrevistar apenas poucas pessoas.

- *Amostragem aleatória sistemática*: ocorre quando os elementos da população se apresentam ordenados e a retirada dos elementos da amostra é feita periodicamente. Por exemplo, em uma linha de produção, podemos, a cada dez itens produzidos, retirar um para avaliar a qualidade da produção. É um processo de amostragem mais preciso que a aleatória simples e tão preciso quanto à amostragem estratificada.

Amostragem não probabilística:

- *Amostragem de conveniência*: de acordo com determinado critério, é escolhido convenientemente um grupo de elementos que comporão a amostra. O pesquisador se dirige a grupos de elementos dos quais deseja saber a opinião. A amostra pesquisada muitas vezes está disponível no local e no momento onde a pesquisa estava sendo realizada.

- *Amostragem por julgamento*: enquadram-se aqui os diversos casos em que o pesquisador deliberadamente escolhe certos elementos para pertencer à amostra, por julgar tais elementos bem representativos. O perigo desse tipo de amostragem é grande, pois o pesquisador pode facilmente se enganar em seu pré-julgamento.

- *Amostragem por quota*: o pesquisador procura obter uma amostra que seja similar apenas em alguns aspectos à população. Há necessidade de conhecer características específicas da população para determinar a amostra.

- *Amostragem a esmo ou sem norma*: é a amostragem em que o pesquisador, para simplificar o processo, procura ser aleatório sem, no entanto, realizar propriamente o sorteio usando algum dispositivo aleatório confiável. Os resultados da amostragem a esmo são, em geral, equivalentes aos da amostragem probabilística se a população é homogênea e se não existe a possibilidade de o pesquisador ser inconscientemente influenciado por alguma característica dos elementos da população.

- *Amostragem acidental:* trata-se da formação de amostras por aqueles elementos que vão aparecendo. Este método é utilizado, geralmente, em pesquisas de opinião, em que os entrevistados são acidentalmente escolhidos.

Dos exemplos relacionados, centraremos nossa atenção na amostragem aleatória simples que é a maneira mais fácil de selecionarmos uma amostra probabilística de uma população. Além disso, esse procedimento será a base para o desenvolvimento de outros procedimentos amostrais.

- **Amostra aleatória simples**, também chamada de amostra casual simples, é aquela obtida de tal forma que todas as unidades da população tenham a mesma probabilidade de fazer parte da amostra, ou ainda, que todas as possíveis amostras tenham igual probabilidade de serem selecionadas.

O processo de obtenção das unidades que comporão a amostra aleatória simples é o sorteio (com ou sem reposição) de todos as unidades da população. Esse sorteio pode ser realizado utilizando-se pedacinhos de papel, tabelas de números aleatórios ou programas computacionais.

Outra maneira bastante prática de efetuar o sorteio é com a utilização de números aleatórios fornecidos pelas calculadoras científicas. Este procedimento envolve apenas três passos:

1) Identificar cada unidade da população por um número, numa seqüência de 1 a N ou de C a N+C−1.

2) Obter uma sequência de n números aleatórios utilizando a função “Random” da calculadora.

3) Efetuar para cada número aleatório a seguinte operação:

$$[U \times N] + C,$$

onde:

U: número aleatório, sendo $0 \leq U < 1$;
N: tamanho da população;
$[U \times N]$: parte inteira do produto $U \times N$;
C: número de ordem da primeira observação.

Por exemplo, se desejamos extrair uma amostra de tamanho n=10 de uma população de tamanho N=150, devemos enumerar de 1 a 150 as unidades da população e obter da calculadora uma seqüência de dez números aleatórios (U). Na tabela abaixo temos uma sequência de dez números aleatórios e, para cada um deles, os dois passos da operação $[U \times N] + 1$.

U	$U \times 150$	$[U \times 150] + 1$
0.301	45,15	46
0.938	140,70	141
0.574	86,10	87
0.205	30,75	31
0.720	108,00	109
0.702	105,30	106
0.152	22,80	23
0.505	75,75	76
0.633	94,95	95
0.566	84,90	85

Assim, as unidades da população que irão compor a amostra são aquelas identificadas pelos números 23, 31, 46, 76, 85, 87, 95, 106, 109 e 141.

A amostragem aleatória simples é o processo mais simples de amostragem, de modo que, dada uma população de N elementos, podemos extrair k amostras diferentes de tamanho n, onde

$$k = \begin{cases} N^n, & \text{se as retiradas são feitas com reposição} \\ C_N^n, & \text{se as retiradas são feitas sem reposição} \end{cases}$$

A probabilidade associada a cada uma das k amostras possíveis de tamanho n é assim definida

$$p(\text{amostra}) = \frac{1}{k} \quad .$$

Na prática, a amostra é sorteada da população unidade por unidade e, se o sorteio for com reposição, a probabilidade associada a cada unidade é

$$p(\text{unidade}) = \frac{1}{N} \quad .$$

Do ponto de vista da quantidade de informação contida na amostra, amostrar sem reposição é mais adequado. Entretanto, a amostragem com reposição conduz a um tratamento teórico mais simples, uma vez que implica na independência entre as unidades selecionadas. Essa independência facilita o desenvolvimento das propriedades dos estimadores que serão considerados.

4.3. Distribuições amostrais

Vamos adotar a partir de agora o conceito de população que é mais adequado para o tratamento teórico. Em vez de considerar a população como um conjunto de indivíduos ou objetos (população real), vamos trabalhar com a ideia de população apresentada na unidade III, ou seja, como o conjunto de todos os possíveis valores de uma variável aleatória, cuja distribuição de probabilidade é conhecida ou passível de ser obtida. Chamaremos este conjunto de valores de *população estatística*. Observemos que para utilizar os conceitos de probabilidade em Estatística é essencial saber qual é a distribuição de probabilidade da variável em estudo.

Vejamos também o significado mais preciso de uma amostra. Entendemos por *amostra aleatória* aquela amostra cujos elementos $[X_1, X_2, ..., X_n]$ são todos independentes entre si e têm a mesma distribuição de probabilidade da população (X), ou seja,

$$\begin{cases} - \text{os } X_{i's} \text{ são independentes} \\ - \text{os } X_{i's} \text{ têm a mesma distribuição de X} \end{cases}$$

Para garantir a independência entre os elementos da amostra, as escolhas devem ser feitas *com reposição*. Como os valores que compõem a amostra são aleatórios, qualquer função (total, média, variância, etc.) dos elementos da amostra será também uma variável aleatória. Denominamos *estatística* qualquer valor obtido em função da amostra. Como as estatísticas são funções de variáveis aleatórias, também são variáveis aleatórias e, como consequência, terão alguma distribuição de probabilidade com média, variância, etc. A distribuição de probabilidade de uma *estatística* é chamada de *distribuição amostral*.

O objetivo da inferência estatística é inferir para a população a partir da amostra. Assim, todas as informações que temos sobre a população são providas pela amostra, ou seja, trabalhamos efetivamente com estatísticas, que são variáveis aleatórias. Por essa razão, é fundamental que conheçamos as distribuições amostrais dessas estatísticas.

A média da amostra ($\bar{X}$) é a estatística mais utilizada porque apresenta propriedades interessantes. Vamos utilizar o exemplo a seguir para demonstrar as propriedades da distribuição amostral da média.

O mecânico de uma oficina de regulagem para carros com 4, 6 e 8 cilindros, cobra pelo serviço 40, 45 e 50 reais, respectivamente. Seja a variável X = valor cobrado pelo mecânico, com a seguinte distribuição de probabilidade:

X = x	40	45	50	Σ
P(X = x)	0,2	0,3	0,5	1

a) Determine a média e a variância da população.

b) Supondo a retirada de uma amostra aleatória de tamanho n, determine a distribuição de probabilidade, a média e a variância de cada elemento da amostra $[X_1, X_2, ..., X_n]$

c) Supondo a retirada de uma amostra de tamanho n = 2, com reposição, quantas e quais são as possíveis amostras retiradas da população e qual a probabilidade associada a cada uma? Determine a média e a variância da distribuição amostral da média $\bar{X}$.

d) Supondo a retirada de uma amostra de tamanho n = 3, com reposição, quantas e quais são as possíveis amostras retiradas da população e qual a probabilidade associada a cada uma? Determine a média e a variância da distribuição amostral da média.

Resolução:

a) Média e a variância da população:

$$E(X) = \mu = \sum_{x \in S_X} x\, p(x) = 40 \times 0{,}2 + 45 \times 0{,}3 + 50 \times 0{,}5 = 46{,}5$$

$$V(X) = \sigma^2 = E(X^2) - \mu^2 = (40^2 \times 0{,}2 + 45^2 \times 0{,}3 + 50^2 \times 0{,}5) - 46{,}5^2 = 15{,}25$$

b) Distribuição de probabilidade, média e variância de cada elemento da amostra $[X_1, X_2, ..., X_n]$

Amostra aleatória de tamanho n $[X_1, X_2, ..., X_n]$

Distribuição de probabilidade de X_1

$X_1 = x_1$	40	45	50	Σ
$P(X_1 = x_1)$	0,2	0,3	0,5	1

$E(X_1) = \mu = 46{,}5$
$V(X_1) = \sigma^2 = 15{,}25$

Distribuição de probabilidade de X_2

$X_2 = x_2$	40	45	50	Σ
$P(X_2 = x_2)$	0,2	0,3	0,5	1

$E(X_2) = \mu = 46{,}5$
$V(X_2) = \sigma^2 = 15{,}25$

Distribuição de probabilidade de X_n

$X_n = x_n$	40	45	50	Σ
$P(X_n = x_n)$	0,2	0,3	0,5	1

$E(X_n) = \mu = 46{,}5$
$V(X_n) = \sigma^2 = 15{,}25$

Verificamos, assim, que se $[X_1, X_2, ..., X_n]$ é uma amostra aleatória, o valor esperado de cada elemento da amostra é igual à média da população e a variância de cada elemento da amostra é igual à variância da população, ou seja,

$$E(X_i) = \mu \quad e \quad V(X_i) = \sigma^2$$

c) O número de amostras aleatórias possíveis é obtido por meio da expressão

$$k = N^n,$$

onde:

k = número de amostras possíveis de um mesmo tamanho
N = tamanho da população
n = tamanho da amostra

Assim, supondo uma amostra de tamanho dois, temos $k = N^n = 3^2 = 9$.

O conjunto de todas as possíveis amostras de tamanho dois, retiradas desta população de tamanho três, consiste no conjunto de todos os arranjos desses três elementos tomados dois a dois. A probabilidade associada a cada amostra é obtida pelo produto das probabilidades de cada elemento da amostra.

Na tabela abaixo temos todas as amostras possíveis, com suas respectivas probabilidades, e a média de cada amostra, obtida pela expressão $\bar{X} = \frac{\sum X_i}{n}$

Amostra	$[X_1, X_2]$	$P[X_1, X_2]$	$\bar{X}$
1	(40, 40)	$0,2 \times 0,2 = 0,04$	40
2	(40, 45)	$0,2 \times 0,3 = 0,06$	42,5
3	(40, 50)	$0,2 \times 0,5 = 0,10$	45
4	(45, 40)	$0,3 \times 0,2 = 0,06$	42,5
5	(45, 45)	$0,3 \times 0,3 = 0,09$	45
6	(45, 50)	$0,3 \times 0,5 = 0,15$	47,5
7	(50, 40)	$0,5 \times 0,2 = 0,10$	45
8	(50, 45)	$0,5 \times 0,3 = 0,15$	47,5
9	(50, 50)	$0,5 \times 0,5 = 0,25$	50

Para construir a *distrubuição amostral da média*, tomamos todos os diferentes valores que a estatística $\bar{X}$ assume e calculamos a probabilidade de ocorrência de cada um. A probabilidade associada a cada valor de $\bar{X}$ é obtida da seguinte maneira:

$P(\bar{X} = 40) = P(40, 40) = 0,04$

$P(\bar{X} = 42,5) = P(40, 45) + P(45, 40) = 0,06 + 0,06 = 0,12$

$P(\bar{X} = 45) = P(40, 50) + P(45, 45) + P(50, 40) = 0,10 + 0,09 + 0,10 = 0,29$

$P(\bar{X} = 47,5) = P(45, 50) + P(50, 45) = 0,15 + 0,15 = 0,30$

$P(\bar{X} = 50) = P(50, 50) = 0,25$

Assim, temos a distribuição amostral da média das amostras de tamanho dois. Essa distribuição de probabilidade pode ser apresentada na forma tabular abaixo.

$\bar{X} = \bar{x}$	40	42,5	45	47,5	50	Σ
$P(\bar{X} = \bar{x})$	0,04	0,12	0,29	0,3	0,25	1

Como todos os possíveis valores de $\bar{X}$ constituem uma população, também podemos obter o valor esperado a média $E(\bar{X})$ e a variância $V(\bar{X})$ desta população:

$$E(\bar{X}) = \mu_{\bar{X}} = \sum_{\bar{x} \in S_{\bar{x}}} \bar{x}\, p(\bar{x}) = 40 \times 0,04 + 42,5 \times 0,12 + 45 \times 0,29 + 47,5 \times 0,3 + 50 \times 0,25 = 46,5$$

$$V(\bar{X}) = \sigma^2_{\bar{X}} = E(\bar{X}^2) - \mu^2_{\bar{X}} = (40^2 \times 0,04 + 42,5^2 \times 0,12 + \ldots + 50^2 \times 0,25) - 46,5^2 = 7,625$$

d) Supondo uma amostra de tamanho três, temos

$k = N^n = 3^3 = 27$ amostras possíveis.

Da mesma forma como já visto no item c, podemos obter todas as possíveis amostras de tamanho três, suas probabilidades e suas médias. Essas valores são apresentados na tabela a seguir.

Amostra	$[X_1, X_2, X_3]$	$P[X_1, X_2, X_3]$	$\bar{X}$	Amostra	$[X_1, X_2, X_3]$	$P[X_1, X_2, X_3]$	$\bar{X}$
1	(40, 40, 40)	0,008	40	15	(45, 45, 50)	0,045	46,7
2	(40, 40, 45)	0,012	41,7	16	(45, 50, 40)	0,030	45
3	(40, 40, 50)	0,020	43,3	17	(45, 50, 45)	0,045	46,7
4	(40, 45, 40)	0,012	41,7	18	(45, 50, 50)	0,075	48,3
5	(40, 45, 45)	0,018	46,7	19	(50, 40, 40)	0,020	43,3
6	(40, 45, 50)	0,030	45	20	(50, 40, 45)	0,030	45
7	(40, 50, 40)	0,020	43,3	21	(50, 40, 50)	0,050	46,7
8	(40, 50, 45)	0,030	45	22	(50, 45, 40)	0,030	45
9	(40, 50, 50)	0,050	48,3	23	(50, 45, 45)	0,045	46,7
10	(45, 40, 40)	0,012	41,7	24	(50, 45, 50)	0,075	48,3
11	(45, 40, 45)	0,018	43,3	25	(50, 50, 40)	0,020	46,7
12	(45, 40, 50)	0,030	45	26	(50, 50, 45)	0,075	48,3
13	(45, 45, 40)	0,018	43,3	27	(50, 50, 50)	0,125	50
14	(45, 45, 45)	0,027	45				

A partir desses dados, construímos a distribuição amostral da média das amostras de tamanho três, apresentada na tabela abaixo.

$\bar{X} = \bar{x}$	40	41,7	43,3	45	46,7	48,3	50	Σ
$P(\bar{X} = \bar{x})$	0,008	0,036	0,114	0,207	0,285	0,225	0,125	1

Assim, obtemos também o valor esperado e a variância da média das amostras de tamanho três.

$$E(\bar{X}) = \mu_{\bar{X}} = \sum_{\bar{x} \in S_{\bar{X}}} \bar{x}\, p(\bar{x}) = 40 \times 0{,}008 + 41{,}7 \times 0{,}036 + \ldots + 50 \times 0{,}125 = 46{,}5$$

$$V(\bar{X}) = \sigma^2_{\bar{X}} = E(\bar{X}^2) - \mu^2_{\bar{X}} = (40^2 \times 0{,}008 + 41{,}7^2 \times 0{,}036 + \ldots + 50^2 \times 0{,}125) - 46{,}5^2 = 5{,}083$$

◆ **Resultados importantes**

Relacionando as medidas da distribuição amostral da média ($\bar{X}$) com as medidas da distribuição populacional (X), podemos verificar algumas propriedades importantes:

– A média das médias de todas as k amostras aleatórias possíveis, de mesmo tamanho n, extraídas de uma população, é igual à média da população, ou seja,

$$E(\bar{X}) = \mu .$$

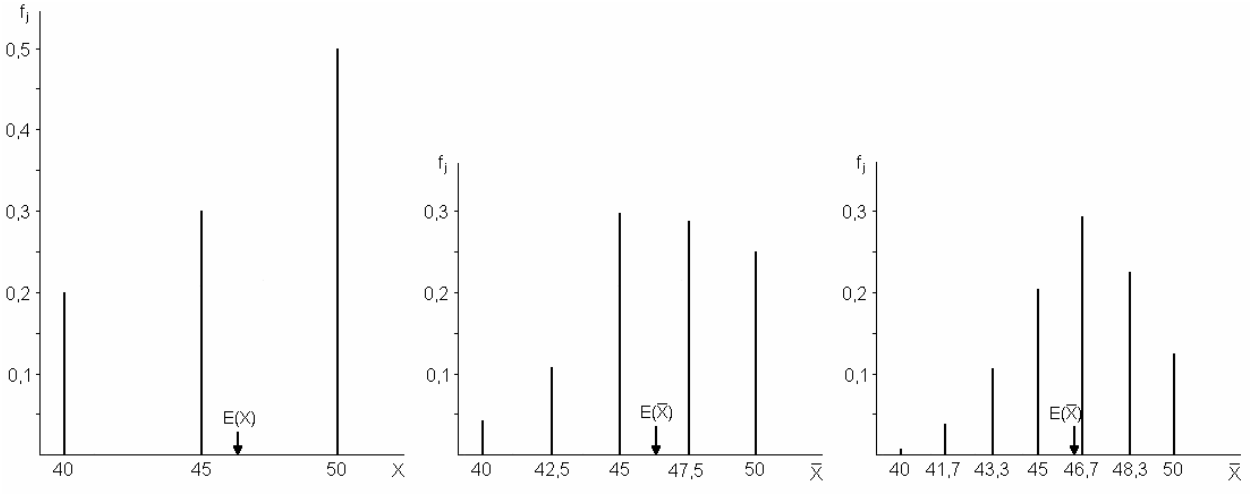
– A variância das médias de todas as k amostras aleatórias possíveis, de mesmo tamanho n, extraídas de uma população, é igual à variância da população dividida pelo tamanho da amostra, ou seja,

$$V(\bar{X}) = \frac{\sigma^2}{n} .$$

Deste resultado podemos obter também o desvio padrão da média que é igual ao desvio padrão da população dividido pela raiz do tamanho da amostra, ou seja,

$$\sigma_{\bar{X}} = \sqrt{V(\bar{X})} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sqrt{\sigma^2}}{\sqrt{n}} = \frac{\sigma}{\sqrt{n}}$$

Comparando o histograma da população X com os histogramas da média $\overline{X}$ para as amostras de tamanhos n = 2 e n = 3, observamos que, mesmo a distribuição da população não sendo simétrica, a distribuição amostral da média se aproxima da simetria à medida que o tamanho da amostra cresce. Podemos observar também que, conforme n vai aumentando, o histograma tende a se concentrar cada vez mais em torno de E(X) = E($\overline{X}$) = 46,5 e os valores extremos passam a ter pequena probabilidade de ocorrência.



População Média das amostras de tamanho 2 Média das amostras de tamanho 3

A tendência para a simetria e consequente aproximação para a normal pode ser verificada nos gráficos da figura 4.1, que mostram o comportamento do histograma para várias formas de distribuição da população e vários tamanhos da amostra.

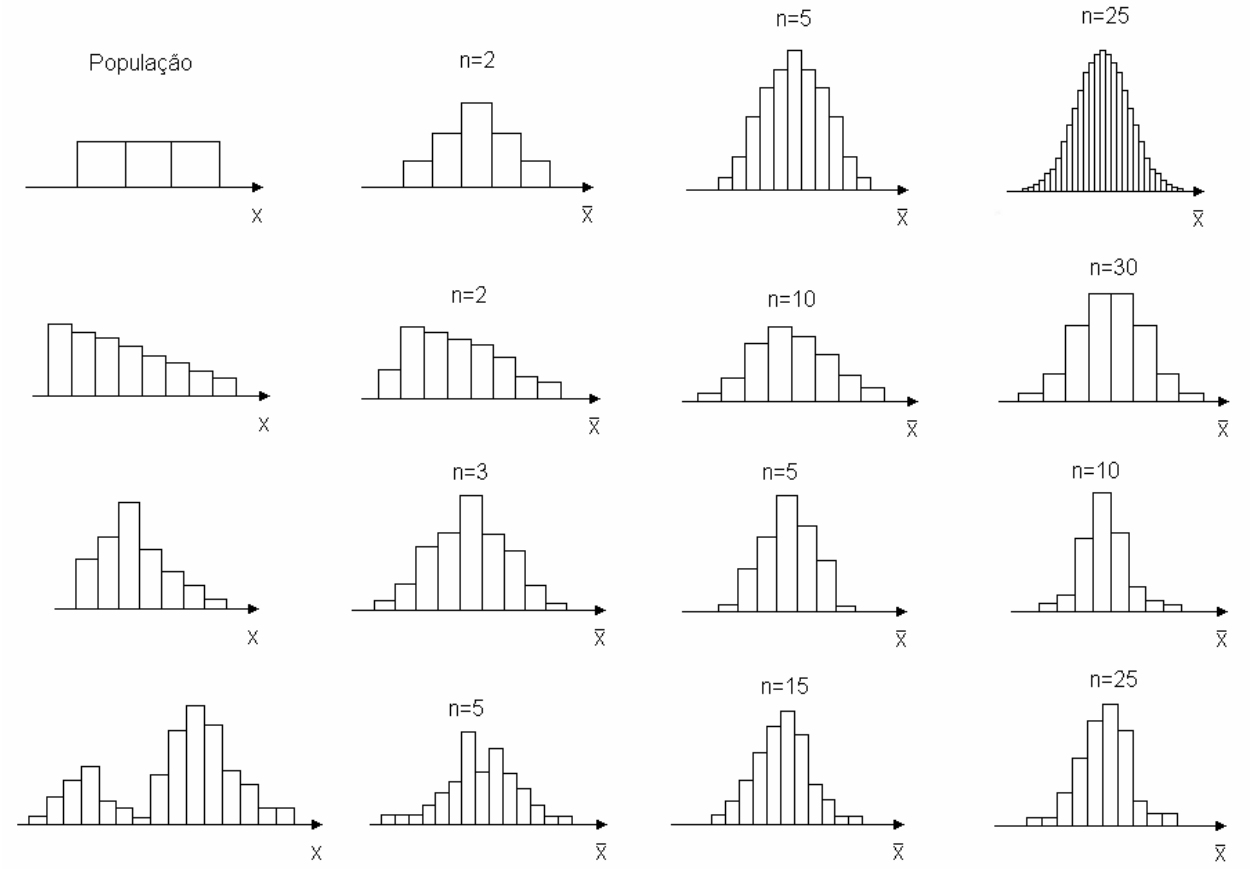


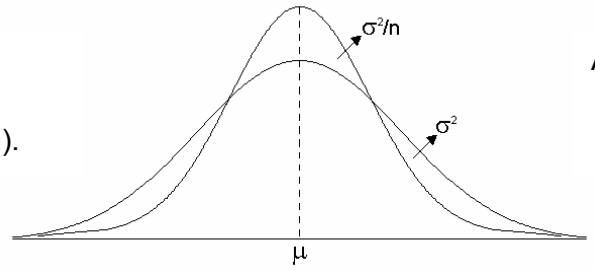
Figura 4.1. Histogramas correspondentes às distribuições amostrais de $\overline{X}$ para amostras extraídas de algumas populações.
Fonte: Bussab e Morettin, 2006.

Assim, para a pergunta "qual é a distribuição da média ($\bar{X}$)?" existem duas respostas diferentes:

1. Se a população (X) de onde foi extraída a amostra aleatória tiver distribuição normal, a distribuição amostral da média ($\bar{X}$) será normal. Pode-se dizer que:

se $X \sim N(\mu, \sigma^2)$,

então, $\bar{X} \sim N(\mu, \sigma^2/n)$.



As médias são iguais, mas a variância de $\bar{X}$ é n vezes menor.

2. Se a população (X) de onde foi extraída a amostra aleatória não tiver distribuição normal, a distribuição amostral da média ($\bar{X}$) se aproximará da normal à medida que o tamanho da amostra (n) cresce. Por exemplo, o número de insetos mortos na aplicação de um inseticida é uma variável que tem distribuição discreta, mas a distribuição do número *médio* de insetos mortos com a aplicação *pode* ser normal dependendo do tamanho da amostra.

Este resultado pode ser derivado do teorema fundamental da estatística paramétrica, denominado Teorema Central do Limite (TCL).

◆ **Teorema Central do Limite:** Se $(X_1, X_2, ..., X_n)$ é uma amostra aleatória de X, então a distribuição da soma de X ($X_+ = \sum X_i$) se aproxima da distribuição normal com média $n\mu$ e variância $n\sigma^2$. Assim, para n suficientemente grande, temos:

$$\frac{X_+ - n\mu}{\sqrt{n\sigma_x^2}} = Z \sim N(0,1).$$

Como consequência, a distribuição da média ($\bar{X}$) se aproxima da normal com média μ e variância σ^2/n. Assim, temos:

$$\frac{X_+ - n\mu}{\sqrt{n\sigma_x^2}} = \frac{\bar{X} - \mu}{\frac{\sigma_x}{\sqrt{n}}} = \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma_x} = Z \sim N(0,1).$$

A demonstração completa desse teorema não será dada porque exigiria recursos dos quais não dispomos, mas o importante é sabermos como esse resultado pode ser usado.

A importância da distribuição normal na estatística se deve em grande parte a este teorema. Observemos que se a população tem distribuição normal, então $\bar{X}$ terá distribuição normal exata. Se a população não tem distribuição normal, $\bar{X}$ poderá ter distribuição normal aproximada ou assintótica.

O TCL afirma que $\bar{X}$ aproxima-se de uma normal quando n tende para o infinito e a rapidez dessa convergência depende da distribuição da população da qual a amostra é retirada (Figura 4.1). Se a população tem uma distribuição próxima da normal, a convergência é rápida; mas se esta população se afasta muito da normal, a convergência é mais lenta, implicando numa amostra maior para que $\bar{X}$ tenha uma distribuição aproximadamente normal. Para a ordem de 30 a 50 elementos a aproximação pode ser considerada satisfatória.

Distribuições importantes como Binomial e Poisson (definidas como a soma de variáveis Bernoulli) se aproximam naturalmente da normal. Se a distribuição Binomial é simétrica (π=0,5), a aproximação (ou convergência) é mais rápida.

◆ **Combinação linear de variáveis**

Seja $[X_1, X_2, ..., X_n]$ uma amostra aleatória e seja $c_1, c_2, ..., c_n$ um conjunto de constantes, então,

$$Y = c_1X_1 + c_2X_2 + \ldots + c_nX_n$$

é uma combinação linear de variáveis.

Sendo Y função de uma variável aleatória, também será uma variável aleatória e, como consequência, terá uma distribuição de probabilidade, com valor esperado e variância. Se Y for a combinação linear de variáveis que têm distribuição normal, ou seja, $X_i \sim N(\mu, \sigma^2)$, então, a distribuição de Y também será normal, com os seguintes parâmetros:

– Valor esperado

$$\begin{aligned} E(Y) &= E(c_1X_1 + c_2X_2 + \ldots + c_nX_n) \\ &= E(c_1X_1) + E(c_2X_2) + \ldots + E(c_nX_n) \\ &= c_1E(X_1) + c_2E(X_2) + \ldots + c_nE(X_n) \end{aligned}$$

Como $E(X_i) = \mu$, temos

$$E(Y) = c_1\mu + c_2\mu + \ldots + c_n\mu = \mu\sum_{i=1}^{n} c_i$$

– Variância

$$V(Y) = V(c_1X_1 + c_2X_2 + \ldots + c_nX_n)$$

Como os $X_{i's}$ são todos independentes entre si, temos

$$\begin{aligned} V(Y) &= V(c_1X_1) + V(c_2X_2) + \ldots + V(c_nX_n) \\ &= c_1^2V(X_1) + c_2^2V(X_2) + \ldots + c_n^2V(X_n) \end{aligned}$$

Como $V(X_i) = \sigma^2$, temos

$$V(Y) = c_1^2\sigma^2 + c_2^2\sigma^2 + \ldots + c_n^2\sigma^2 = \sigma^2\sum_{i=1}^{n} c_i^2$$

Verificamos, assim, que se Y é a combinação linear de um conjunto de variáveis que têm distribuição normal, então $Y \sim N\left(\mu\sum_{i=1}^{n} c_i\ ,\ \sigma^2\sum_{i=1}^{n} c_i^2\right)$.

4.3.1. Distribuições amostrais de algumas estatísticas importantes

Nesta seção serão apresentadas com mais detalhe as distribuições de probabilidade das estatísticas mais utilizadas nos processos de inferência.

◆ **Distribuição qui-quadrado** (χ^2)

Seja uma variável aleatória $X \sim N(\mu, \sigma^2)$ e $[X_1, X_2, ..., X_n]$ uma amostra aleatória dela proveniente. Assim,

$$X_i \sim N(\mu, \sigma^2).$$

Padronizando, ou seja, transformando X_i em Z_i, temos

$$Z_i = \frac{X_i - \mu}{\sigma} \sim N(0, 1), \text{ sendo } -\infty < z_i < +\infty.$$

Seja Q uma nova variável definida como a soma dos quadrados de ν variáveis Z independentes. Então, dizemos que a variável Q tem distribuição qui-quadrado, denotada por χ^2, com parâmetro ν, ou seja,

$$Q = \sum_{i=1}^{\nu} Z_i^2 = Z_1^2 + Z_2^2 + ... + Z_\nu^2 \sim \chi^2(\nu),$$

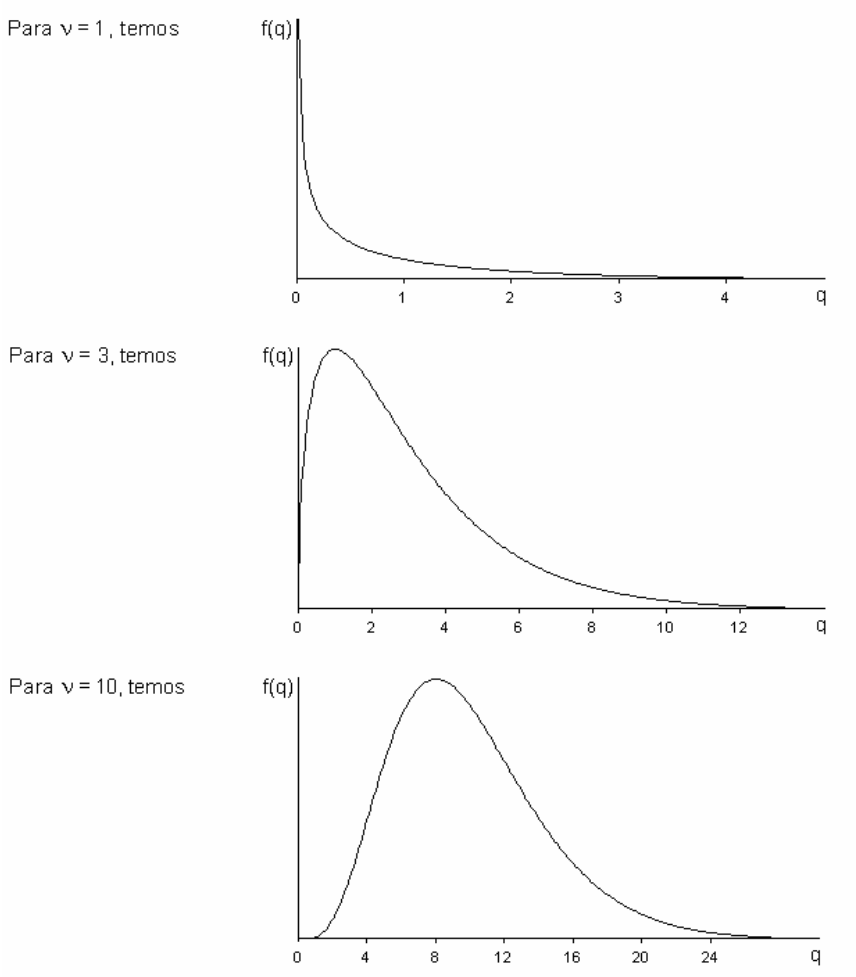
onde: ν = número de graus de liberdade ou variáveis independentes somadas.

A função densidade de probabilidade da distribuição χ^2 é dada por

$$f(q) = -\frac{1}{2^{\frac{\nu}{2}}\Gamma\left(\frac{\nu}{2}\right)} e^{-\frac{q}{2}} q^{\frac{\nu}{2}-1}, \text{ com } 0 \le q < +\infty.$$

Sendo a variável Q definida como uma soma de quadrados, seus valores nunca serão negativos. A curva da distribuição χ^2, representação gráfica da função densidade de probabilidade, muda o seu formato à medida que varia o número de graus de liberdade.

Exemplos:



A distribuição χ^2 tem média $\mu = \nu$ e variância $\sigma^2 = 2\nu$.

Uma variável importante na determinação de intervalos de confiança e testes de hipóteses a respeito da variância da população (σ^2) tem distribuição χ^2 e assim definida

$$Q = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(\nu), \text{ onde } \nu = n - 1.$$

Esta variável surge da seguinte situação: seja a variável $X \sim N(\mu, \sigma^2)$ e $[X_1, X_2, ..., X_n]$ uma amostra aleatória dela proveniente. A variância desta amostra será

$$S^2 = \frac{\sum(X_i - \bar{X})^2}{n-1} \quad \text{ou} \quad (n-1)S^2 = \sum(X_i - \bar{X})^2.$$

Dividindo os dois termos por uma constante de interesse (σ^2), não alteramos a igualdade. Assim, temos

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

Somando e subtraindo outra constante de interesse (μ), resolvendo o binômio e aplicando as propriedades da soma, temos

$$\frac{(n-1)S^2}{\sigma^2} = \frac{\sum(X_i - \bar{X})^2}{\sigma^2} = \frac{\sum(X_i - \bar{X} + \mu - \mu)^2}{\sigma^2} = \frac{\sum(X_i - \mu)^2}{\sigma^2} - \frac{n(\bar{X} - \mu)^2}{\sigma^2} = \sum\left(\frac{X_i - \mu}{\sigma}\right)^2 - \left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right)^2$$

Fazendo $\frac{X_i - \mu}{\sigma} = Z_i$ e $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = Z$, temos

$$\sum\left(\frac{X_i - \mu}{\sigma}\right)^2 = \sum Z_i^2 = Z_1^2 + Z_2^2 + ... + Z_n^2 = Q_n \sim \chi^2(n)$$

e

$$\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right)^2 = Z^2 = Q_1 \sim \chi^2(1),$$

então,

$$\frac{(n-1)S^2}{\sigma^2} = \frac{\sum(X_i - \bar{X})^2}{\sigma^2} = Q_n - Q_1,$$

donde resulta

$$Q_n - Q_1 = (Z_1^2 + Z_2^2 + \ldots + Z_n^2) - Z^2 = Q_{(n-1)} \sim \chi^2(n-1).$$

Outros exemplos de variáveis com distribuição χ^2 de ocorrência comum nos testes de hipóteses envolvendo dados de enumeração são:

$$Q = \sum_{i=1}^{k} \frac{(X_i - E_i)^2}{E_i} \sim \chi^2(\nu), \text{ onde } \nu = k - 1$$

$$Q = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(X_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2(\nu), \text{ onde } \nu = (r-1)(s-1)$$

◆ **Distribuição t de Student**

Seja uma variável Z, com distribuição normal padrão, e uma variável Q, com distribuição χ^2 independentes, então, dizemos que uma variável T definida como:

$$T = \frac{Z}{\sqrt{\frac{Q}{\nu}}}$$

tem distribuição t de Student com parâmetro ν.

Nesse contexto, uma variável Z com distribuição normal padrão e uma variável Q com distribuição χ^2 nas quais temos grande interesse nas aplicações são:

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

e

$$Q = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(\nu), \text{ onde } \nu = n-1.$$

Então, $\frac{Q}{\nu} = \frac{Q}{n-1} = \frac{\frac{(n-1)S^2}{\sigma^2}}{n-1} = \frac{S^2}{\sigma^2}$,

donde resulta $T = \frac{Z}{\sqrt{\frac{Q}{\nu}}} = \frac{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{S^2}{\sigma^2}}} = \frac{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\frac{S}{\sigma}} = \frac{\bar{X}-\mu}{S/\sqrt{n}}$

Assim, temos $T = \frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t(\nu)$, onde $\nu = n-1$

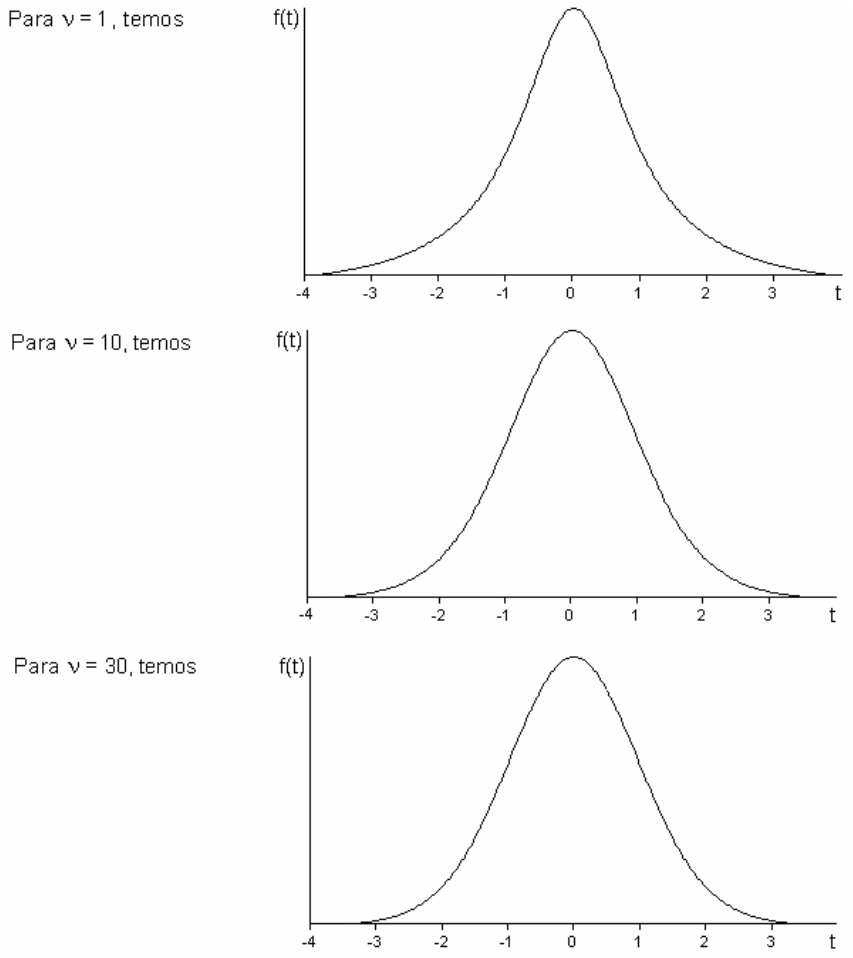
De uma maneira geral, uma variável com distribuição t é muito parecida com uma normal padrão, exceto que o desvio padrão, que aparece no denominador, é o desvio padrão amostral e não o populacional.

A função densidade de probabilidade da distribuição t de Student com ν graus de liberdade é dada por

$$f(t) = \frac{1}{\sqrt{\nu}\,B\left(\frac{1}{2},\frac{\nu}{2}\right)} \frac{1}{\left(1+\frac{t^2}{\nu}\right)^{\nu+\frac{1}{2}}}, \text{ com } -\infty < t < +\infty.$$

Podemos observar, nos exemplos a seguir, que a curva da distribuição t de Student, representação gráfica da função densidade de probabilidade, muda o seu formato à medida que varia o valor de ν.

Exemplos:



A distribuição t de Student tem média $\mu = 0$ e variância $\sigma^2 = \frac{\nu}{\nu - 2}$.

A variável $T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(\nu)$ é importante na determinação de intervalos de confiança e testes de hipóteses a respeito da média da população (μ).

De uma maneira geral, a razão $T = \frac{\hat{\theta} - E(\hat{\theta})}{S(\hat{\theta})}$ tem distribuição t, sendo $\hat{\theta}$ uma combinação linear de variáveis normais e $E(\hat{\theta})$ o seu valor esperado.

Por exemplo, $\hat{\theta} = c_1\bar{X}_1 + c_2\bar{X}_2$, onde $c_1 = 1$ e $c_2 = -1$,

então,

$$\theta = \mu_1 - \mu_2 \text{ e } S(\hat{\theta}) = S\left(\bar{X}_1 - \bar{X}_2\right),$$

resultando que

$$T = \frac{\left(\bar{X}_1 - \bar{X}_2\right) - (\mu_1 - \mu_2)}{S\left(\bar{X}_1 - \bar{X}_2\right)} \sim t(\nu).$$

◆ **Distribuição F de Snedecor**

Sejam duas variáveis Q_1 e Q_2 com distribuição χ^2 independentes, então, dizemos que uma variável F definida como:

$$F = \frac{\frac{Q_1}{\nu_1}}{\frac{Q_2}{\nu_2}},$$

tem distribuição F, com parâmetros ν_1 e ν_2.

A variável F é definida como a razão entre duas variáveis que têm distribuição χ^2. Vimos que uma variável com esta distribuição nunca assume valores negativos, portanto, os valores da variável F também não poderão ser negativos.

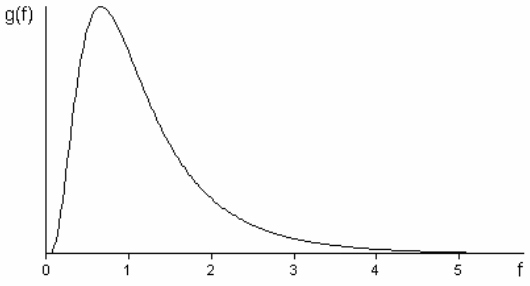
A função densidade de probabilidade da distribuição F é dada por

$$g(f) = \frac{1}{B\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right)} \left(\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}} \frac{f^{\frac{\nu_1}{2}-1}}{\left(1+\frac{\nu_1}{\nu_2} f\right)^{\frac{\nu_1+\nu_2}{2}}}, \text{ com } 0 \leq f < +\infty.$$

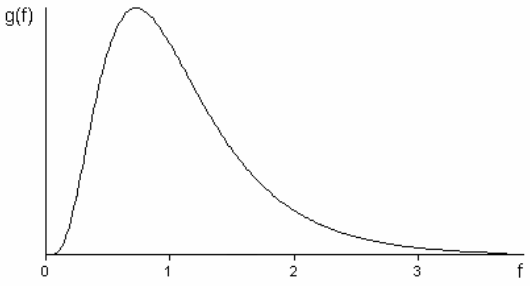
O gráfico desta função muda o seu formato à medida que os valores de ν_1 e ν_2 se alteram.

Exemplos:

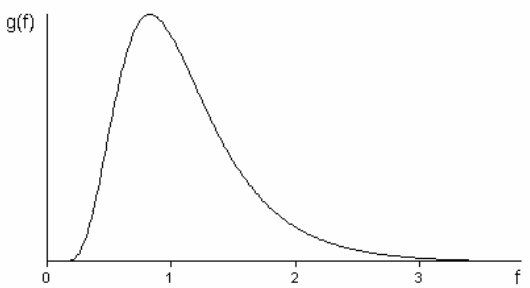
Para $\nu_1 = 10$ e $\nu_2 = 10$, temos



Para $\nu_1 = 10$ e $\nu_2 = 20$, temos



Para $\nu_1 = 20$ e $\nu_2 = 20$, temos



Outra variável comumente utilizada na determinação de intervalos de confiança e testes de hipóteses a respeito da variância da população (σ^2) tem distribuição F e surge da seguinte situação: sejam as variáveis aleatórias $X_1 \sim (\mu_1, \sigma_1^2)$ e $X_2 \sim (\mu_2, \sigma_2^2)$ e $\left[X_{11}, X_{12}, \ldots, X_{1n_1}\right]$ e $\left[X_{21}, X_{22}, \ldots, X_{2n_2}\right]$ amostras aleatórias delas provenientes. As respectivas variâncias dessas amostras serão

$$S_1^2 = \frac{\sum\left(X_{1i} - \bar{X}_1\right)^2}{n_1 - 1} \quad \text{e} \quad S_2^2 = \frac{\sum\left(X_{2i} - \bar{X}_2\right)^2}{n_2 - 1}$$

Podemos reescrever as expressões da seguinte forma

$$\frac{(n_1 - 1)S_1^2}{\sigma_1^2} = \frac{\sum\left(X_{1i} - \bar{X}_1\right)^2}{\sigma_1^2} = Q_1 \sim \chi^2(n_1 - 1)$$

e

$$\frac{(n_2 - 1)S_2^2}{\sigma_2^2} = \frac{\sum\left(X_{2i} - \bar{X}_2\right)^2}{\sigma_2^2} = Q_2 \sim \chi^2(n_2 - 1).$$

Se

$$F = \frac{\dfrac{Q_1}{\nu_1}}{\dfrac{Q_2}{\nu_2}} \sim F(\nu_1, \nu_2),$$

onde:

$$\frac{Q_1}{\nu_1} = \frac{\dfrac{(n_1 - 1)S_1^2}{\sigma_1^2}}{n_1 - 1} = \frac{S_1^2}{\sigma_1^2}$$

e

$$\frac{Q_2}{\nu_2} = \frac{\dfrac{(n_2 - 1)S_2^2}{\sigma_2^2}}{n_2 - 1} = \frac{S_2^2}{\sigma_2^2},$$

então, temos

$$F = \frac{\dfrac{Q_1}{\nu_1}}{\dfrac{Q_2}{\nu_2}} = \frac{\dfrac{S_1^2}{\sigma_1^2}}{\dfrac{S_2^2}{\sigma_2^2}} \sim F(\nu_1, \nu_2).$$

No caso de as variâncias populacionais serem iguais, ou seja, $\sigma_1^2 = \sigma_2^2 = \sigma^2$, temos

$$F = \frac{\dfrac{Q_1}{\nu_1}}{\dfrac{Q_2}{\nu_2}} = \frac{S_1^2}{S_2^2} \sim F(\nu_1, \nu_2),$$

onde:

$\nu_1 = n_1 - 1;$

$\nu_2 = n_2 - 1.$

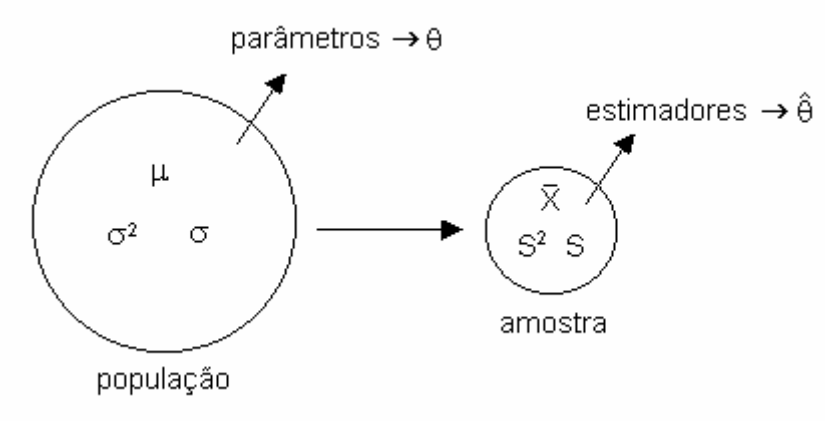
4.4. Estimação de parâmetros

4.4.1. Conceitos fundamentais

É de fundamental importância a compreensão e o domínio de alguns termos que serão usados com bastante frequência nos tópicos que seguem. Veremos, a seguir, os conceitos de parâmetro, estimador e estimativa.

Os *parâmetros* são valores (medidas) calculados diretamente da *população* e servem para caracterizá-la. Os parâmetros geralmente são valores desconhecidos, sempre são constantes, e são representados, genericamente, pela letra grega teta (θ). São exemplos de parâmetros: a média da população (μ) e a variância da população (σ^2).

Os *estimadores* são valores (medidas) calculados em uma *amostra* com objetivo de obter informação sobre os parâmetros e sobre a própria população. Todos os estimadores são estatísticas, uma vez que são valores amostrais. Sendo estatísticas, são também variáveis aleatórias, pois podem assumir diferentes valores dependendo da amostra. Os estimadores são representados, genericamente, pela letra teta com um acento circunflexo ($\hat{\theta}$), onde se lê teta chapéu. Dentre os exemplos de estimadores podemos citar a média da amostra ($\bar{X}$) e a variância da amostra (S^2).



Sendo o estimador uma variável que pode assumir diferentes valores, chamamos de *estimativa* um valor particular que o estimador assume.

Consideremos como exemplo a seguinte população constituída por quatro valores (N = 4):

X = x	1	2	3	4
P(X=x)	0,2	0,3	0,3	0,2

onde: $\mu = 2,5$ e $\sigma^2 = 1,05$.

Desta população, retiramos uma amostra aleatória de tamanho dois (n = 2), [X_1, X_2].

Assim, podemos calcular o número de diferentes amostras de tamanho dois que podem ser extraídas desta população de tamanho quatro:

$$k = N^n = 4^2 = 16 \text{ amostras.}$$

Sendo possível obter 16 amostras diferentes, para cada um dos parâmetros, μ e σ^2, será possível obter 16 estimativas. Na tabela a seguir temos todas as possíveis estimativas de cada um desses parâmetros.

Parâmetro		$\mu = 2,5$	$\sigma^2 = 1,05$
Estimador		$\bar{X} = \frac{\sum X_i}{n}$	$S^2 = \frac{\sum (X_i - \bar{X})^2}{n-1}$
Estimativas	Amostra 1: (1, 1)	$\bar{x}_1 = \frac{1+1}{2} = 1$	$s_1^2 = \frac{(1-1)^2 + (1-1)^2}{2-1} = 0$
	Amostra 2: (1, 2)	$\bar{x}_2 = \frac{1+2}{2} = 1,5$	$s_2^2 = \frac{(1-1,5)^2 + (2-1,5)^2}{2-1} = 0,5$
	Amostra 3: (1, 3)	$\bar{x}_3 = 2$	$s_3^2 = 2$
	Amostra 4: (1, 4)	$\bar{x}_4 = 2,5$	$s_4^2 = 4,5$
	Amostra 5: (2, 1)	$\bar{x}_5 = 1,5$	$s_5^2 = 0,5$
	Amostra 6: (2, 2)	$\bar{x}_6 = 2$	$s_6^2 = 0$
	Amostra 7: (2, 3)	$\bar{x}_7 = 2,5$	$s_7^2 = 0,5$
	Amostra 8: (2, 4)	$\bar{x}_8 = 3$	$s_8^2 = 2$
	Amostra 9: (3, 1)	$\bar{x}_9 = 2$	$s_9^2 = 2$
	Amostra 10: (3, 2)	$\bar{x}_{10} = 2,5$	$s_{10}^2 = 0,5$
	Amostra 11: (3, 3)	$\bar{x}_{11} = 3$	$s_{11}^2 = 0$
	Amostra 12: (3, 4)	$\bar{x}_{12} = 3,5$	$s_{12}^2 = 0,5$
	Amostra 13: (4, 1)	$\bar{x}_{13} = 2,5$	$s_{13}^2 = 4,5$
	Amostra 14: (4, 2)	$\bar{x}_{14} = 3$	$s_{14}^2 = 2$
	Amostra 15: (4, 3)	$\bar{x}_{15} = 3,5$	$s_{15}^2 = 0,5$
	Amostra 16: (4, 4)	$\bar{x}_{16} = 4$	$s_{16}^2 = 0$

Devemos considerar também que podem existir vários estimadores para um mesmo parâmetro. Por exemplo, a média aritmética simples ($\bar{X}$) e a média aritmética ponderada ($\bar{X}_p$), calculadas na amostra, bem como qualquer elemento em particular de uma amostra aleatória (X_i), são todos estimadores da média populacional (μ).

$$\left.\begin{array}{l} \bar{X} = \dfrac{\sum X_i}{n} \\ \bar{X}_p = \dfrac{\sum X_i p_i}{\sum p_i} \\ X_i \end{array}\right\} \text{ estimadores de } \mu$$

Da mesma forma, as variâncias S^2 (com denominador n – 1) e S_n^2 (com denominador n), calculadas na amostra, são dois estimadores da variância populacional (σ^2).

$$\left.\begin{array}{l} S^2 = \dfrac{\sum (X_i - \bar{X})^2}{n-1} \\ S_n^2 = \dfrac{\sum (X_i - \bar{X})^2}{n} \end{array}\right\} \text{ estimadores de } \sigma^2$$

Para escolher o melhor dentre todos os estimadores de um mesmo parâmetro, devemos optar pelo que tem melhores propriedades.

4.4.2. Propriedades dos estimadores

- **Imparcialidade ou não tendenciosidade**

Um estimador $\hat{\theta}$ é um estimador imparcial do parâmetro θ se o valor esperado de $\hat{\theta}$ for igual a θ.

$$E(\hat{\theta}) = \theta$$

Exemplos:

$\bar{X}$ é um estimador imparcial de μ, pois $E(\bar{X}) = \mu$.

$\bar{X}_p$ é um estimador imparcial de μ, pois $E(\bar{X}_p) = \mu$.

X_1 é um estimador imparcial de μ, pois $E(X_1) = \mu$.

S^2 é um estimador imparcial de σ^2, pois $E(S^2) = \sigma^2$.

S_n^2 não é um estimador imparcial de σ^2, pois $E(S_n^2) = \frac{n-1}{n}\sigma^2$

- **Eficiência ou variância mínima**

Se dois ou mais estimadores de um mesmo parâmetro são imparciais, é mais eficiente aquele que possui a menor variância.

Exemplo: Dentre todos os estimadores imparciais de μ ($\bar{X}$, $\bar{X}_p$ e X_1), a média simples ($\bar{X}$) é o mais eficiente porque tem a menor variância.

Demonstração: Considere uma amostra de tamanho n = 3 $[X_1, X_2, X_3]$

Média simples ($\bar{X}$)	Média ponderada ($\bar{X}_p$)
$\bar{X} = \frac{X_1 + X_2 + X_3}{3} = \frac{1}{3}(X_1 + X_2 + X_3)$	$\bar{X}_p = \frac{1X_1 + 2X_2 + 1X_3}{4} = \frac{1}{4}(X_1 + 2X_2 + X_3)$
$V(\bar{X}) = V\left[\frac{1}{3}(X_1 + X_2 + X_3)\right]$	$V(\bar{X}_p) = V\left[\frac{1}{4}(X_1 + 2X_2 + X_3)\right]$
$V(\bar{X}) = \frac{1}{9}\left[V(X_1 + X_2 + X_3)\right]$	$V(\bar{X}_p) = \frac{1}{16}\left[V(X_1 + 2X_2 + X_3)\right]$
$V(\bar{X}) = \frac{1}{9}\left[V(X_1) + V(X_2) + V(X_3)\right]$	$V(\bar{X}_p) = \frac{1}{16}\left[V(X_1) + 4V(X_2) + V(X_3)\right]$
$V(\bar{X}) = \frac{1}{9}(\sigma^2 + \sigma^2 + \sigma^2)$	$V(\bar{X}_p) = \frac{1}{16}(\sigma^2 + 4\sigma^2 + \sigma^2)$
$V(\bar{X}) = \frac{3\sigma^2}{9} = 0{,}33\sigma^2$	$V(\bar{X}_p) = \frac{6\sigma^2}{16} = 0{,}38\sigma^2$

Comparando as variâncias dos três estimadores: $V(\bar{X}) = 0{,}33\sigma^2 < V(\bar{X}_p) = 0{,}38\sigma^2 < V(X_1) = \sigma^2$, verificamos que a média aritmética simples tem a menor variação, portanto, é o estimador mais eficiente.

- **Consistência**

Um estimador é consistente se à medida que o tamanho da amostra aumenta o valor do estimador se aproxima do parâmetro.

$$\hat{\theta} \xrightarrow{n \to N} \theta$$

Exemplo:

S_n^2 é um estimador consistente de σ^2.

Com base nessa propriedade, podemos concluir que:

- Se a amostra for pequena, devemos utilizar S^2 para estimar σ^2.
- Se a amostra for grande, podemos utilizar S^2 ou S_n^2 para estimar σ^2.

4.4.3. Processos de estimação

Um parâmetro pode ser estimado de duas formas: por ponto ou por intervalo.

- **Estimação por ponto**

É o processo através do qual obtemos um único ponto, ou seja, um único valor para estimar o parâmetro.

Exemplo: Amostra (1, 3, 2)

$$\bar{x} = \frac{\sum x_i}{n} = \frac{1+3+2}{3} = 2 \leftarrow \text{estimativa pontual ou por ponto de } \mu$$

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} = \frac{(1-2)^2 + (3-2)^2 + (2-2)^2}{3-1} = 1 \leftarrow \text{estimativa pontual ou por ponto de } \sigma^2$$

- **Estimação por intervalo**

É um processo que permite obter um intervalo onde, com uma determinada probabilidade (nível de confiança), podemos esperar encontrar o verdadeiro valor do parâmetro.

$$LI < \theta < LS$$

As estimativas por intervalo são preferíveis às estimativas por ponto porque indicam a precisão, ou seja, sabemos a probabilidade de o intervalo conter o parâmetro.

4.4.3.1. Intervalos de confiança para a média

- **Intervalo de confiança para a média de uma população** (μ)

Para a construção do intervalo de confiança devemos levar em conta se conhecemos a variância populacional. Sendo assim, duas situações serão consideradas:

Situação 1. Quando a variância da população (σ^2) é conhecida

Considere que desejamos estimar a média μ de uma população X.

Para determinar o intervalo de confiança (IC) para μ, utilizamos o estimador $\bar{X}$ que, como já foi demonstrado, é o melhor estimador de μ.

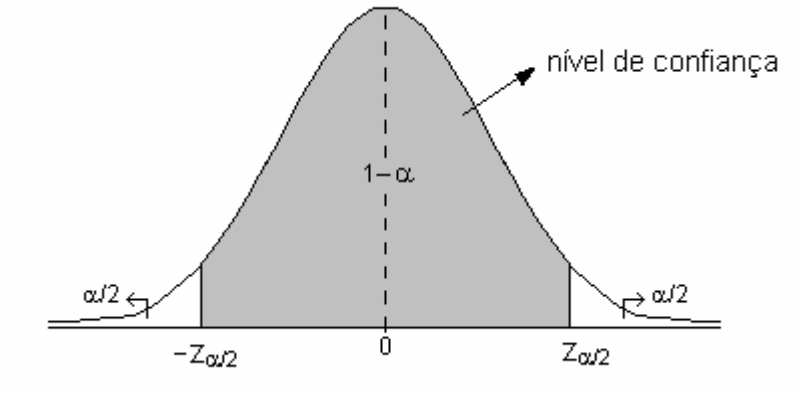
De acordo com o TCL, se $X \sim N(\mu, \sigma^2)$, então, $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$.

Padronizando a variável $\bar{X}$, temos $Z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu}{\sqrt{V(\bar{X})}} = \frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$,

sendo que Z tem distribuição normal com média igual a zero e variância igual a um, ou seja,

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

De acordo com a figura abaixo, vemos que $1-\alpha$ é a probabilidade de que a variável Z assuma um valor entre $-z_{\alpha/2}$ e $z_{\alpha/2}$ e α é a probabilidade de Z não estar entre $-z_{\alpha/2}$ e $z_{\alpha/2}$.



Daí, temos

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

O valor α é denominado *nível de significância* ou *taxa de erro* (usualmente com valor 0,05 ou 0,01), enquanto o valor $1-\alpha$ representa o *nível de confiança* do intervalo.

Sabendo que $Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$ e fazendo a substituição, temos

$$P(-z_{\alpha/2} < \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} < z_{\alpha/2}) = 1 - \alpha.$$

Como o objetivo é construir um intervalo de confiança para a média da população, devemos isolar μ na expressão. Podemos alcançar este objetivo manipulando a expressão:

$$P(-z_{\alpha/2} < \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} < z_{\alpha/2}) = 1-\alpha$$

$$P(-z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \bar{X} - \mu < z_{\alpha/2} \frac{\sigma}{\sqrt{n}}) = 1-\alpha$$

$$P(-\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < -\mu < -\bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}) = 1-\alpha$$

$$P[(-\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < -\mu < -\bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}) \times (-1)] = 1-\alpha$$

$$P(\bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} > \mu > \bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}) = 1-\alpha$$

Daí resulta a expressão do intervalo de confiança para a média de uma população:

$$P(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) = 1-\alpha$$

Este intervalo de confiança também pode ser expresso da seguinte forma:

$$IC\ (\mu;\ 1-\alpha):\ \bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

onde:

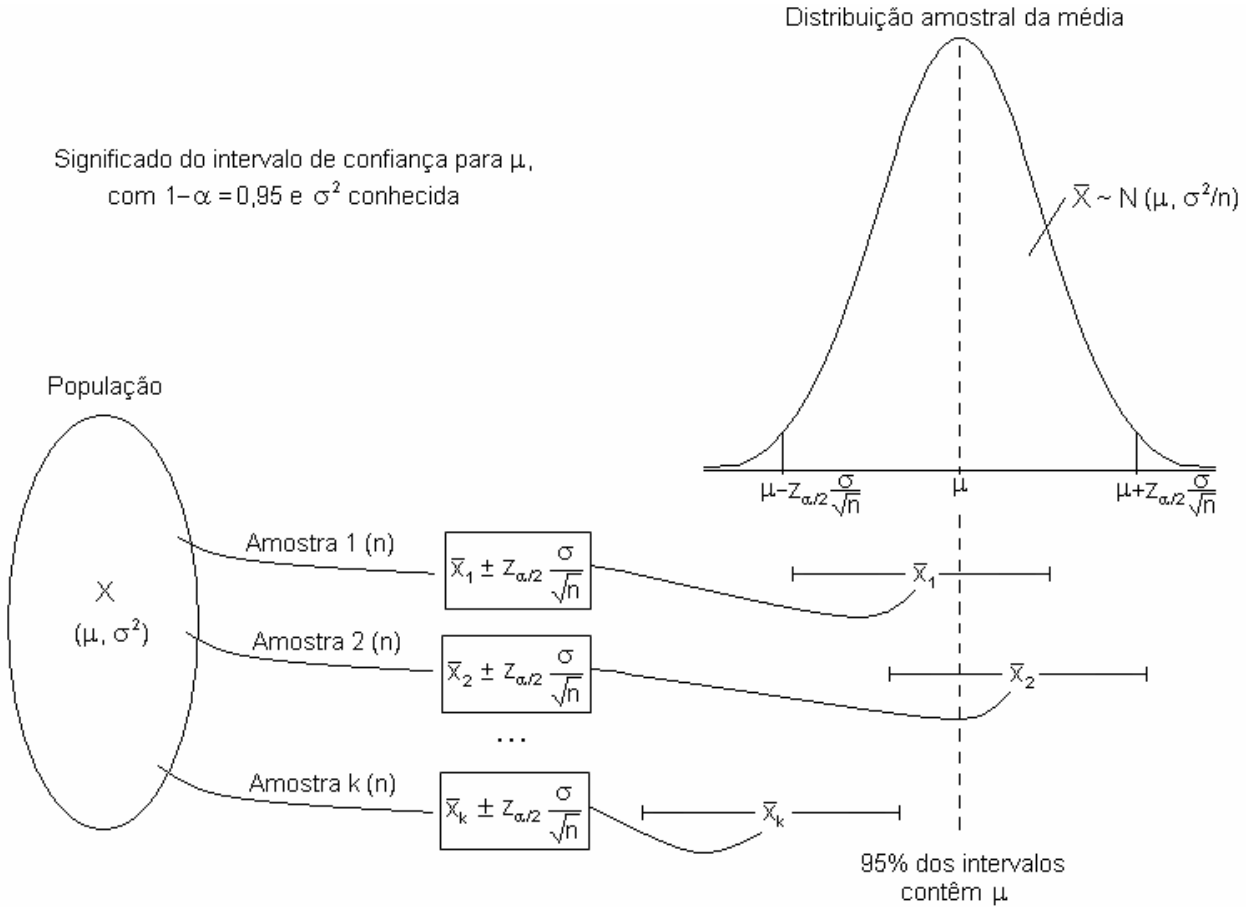
$\bar{X}$: é o estimador de μ;

$z_{\alpha/2}$: é o valor da variável Z que delimita a área $\alpha/2$ (Tabela I do Apêndice);

n: é o tamanho da amostra;

σ: é desvio padrão da população (parâmetro).

É importante salientar que μ é um parâmetro (constante) e os limites do intervalo é que são aleatórios. Assim, a interpretação do intervalo ao nível de 95% de confiança, por exemplo, deve ser da seguinte maneira: se pudéssemos construir uma quantidade grande de intervalos, todos baseados em amostras de tamanho n, 95% deles conteriam o parâmetro μ, como ilustra a figura abaixo.



Observemos que, escolhida uma amostra e encontrada sua média $\bar{x}_0$, podemos construir o intervalo $\left(\bar{x}_0 - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}; \bar{x}_0 + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$, mas este intervalo pode ou não conter o parâmetro μ. A probabilidade de que contenha o parâmetro μ é $1-\alpha$.

Podemos verificar também que todos os intervalos com mesmo nível de confiança têm a mesma amplitude: $2z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$.

Na maioria dos casos não conhecemos, de fato, o parâmetro σ, pois não estudamos a população inteira. Entretanto, com base na propriedade de consistência dos estimadores, quando a amostra tem tamanho grande, a estimativa de um parâmetro é considerada suficientemente próxima do parâmetro. Assim, quando trabalhamos com grandes amostras a estimativa de σ, que é s (desvio padrão da amostra), pode ser usada no lugar do parâmetro. Consideramos a amostra suficientemente grande para utilizar a variável Z quando n é maior que 30.

Duas pressuposições devem ser atendidas para a utilização desta metodologia:

1. A variável em estudo tem distribuição normal, X ~ N (μ, σ^2).
2. A variância populacional é conhecida ou o tamanho da amostra é suficientemente grande para obtenção de uma estimativa aproximada da variação populacional (σ).

Consideremos o seguinte o exemplo resolvido.

Uma amostra de 100 terneiros de dois meses de idade da raça Ibagé apresentou peso médio de 65,5kg e desvio padrão de 4,8kg. Obtenha o intervalo de confiança, ao nível de 95%, para o verdadeiro peso médio de terneiros e redija a conclusão.

Variável em estudo: X = peso de terneiros (kg)

Pressuposições: 1. A variável em estudo tem distribuição normal.
2. A amostra tem tamanho suficiente para estimar σ.

Estimativas:
$\bar{x}$ = 65,5 kg
s = 4,8 kg $\cong \sigma$
n = 100 terneiros

$z_{\alpha/2} = z_{0,025} = 1,96$

$$\text{IC}(\mu; 1-\alpha): \bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$$\text{IC}(\mu; 0,95): 65,5 \pm 1,96 \times \frac{4,8}{\sqrt{100}}$$

IC (μ; 0,95): 65,5 $\pm$ 0,941
Limite inferior = 65,5 – 0,941 = 64,56
Limite superior = 65,5 + 0,941 = 66,44

P(64,56 < μ < 66,44) = 0,95

Concluímos que o intervalo de confiança, ao nível de 95%, para o verdadeiro peso médio de terneiros de dois meses de idade da raça Ibagé é de 64,56 a 66,44 kg.

Situação 2. Quando a variância da população (σ^2) é desconhecida

Quando a amostra é pequena, não podemos supor que o desvio padrão da amostra (s) seja uma estimativa suficientemente aproximada do parâmetro σ. Como não conhecemos a variância populacional, não podemos utilizar a variável Z, que tem distribuição normal padrão, para construir o intervalo de confiança para μ.

Nesse caso, em vez de Z, utilizamos a estatística T que não tem distribuição normal e sim distribuição t de Student, com parâmetro ν:

$$T = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \sim t\ (\nu),$$

onde:

$\bar{X}$: é a média da amostra (estimador de μ);
S: é o desvio padrão da amostra (estimador de σ);
n: tamanho da amostra;
$\nu = n - 1$: é o número de graus de liberdade associado à variância da amostra S^2.

Sob o ponto de vista das aplicações, podemos definir a estatística T de uma forma mais genérica:

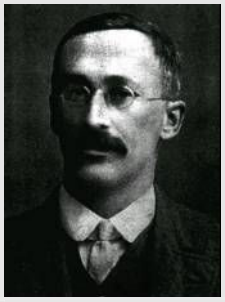
$$T = \frac{\hat{\theta} - \theta}{S(\hat{\theta})} \sim t\ (\nu),$$

onde:

θ : é o parâmetro que está sendo estimado;
$\hat{\theta}$: é o estimador do parâmetro;
$S(\hat{\theta})$: é o estimador do desvio (ou erro) padrão de $\hat{\theta}$.

Distribuição t

Em 1908, o pesquisador inglês William Gosset, ao tentar resolver problemas relativos a pequenas amostras, descobriu a *distribuição t*.



William Gosset
(1876 - 1937)

Gosset trabalhava, na época, numa cervejaria na Irlanda e estava ciente de que seus empregadores não queriam que funcionários publicassem o que quer que fosse, com receio de que segredos industriais caíssem no domínio público e, principalmente, nas mãos da concorrência. Por isso Gosset ao descobrir uma nova distribuição de probabilidades (distribuição t), publicou seus trabalhos sob o pseudônimo de Student.

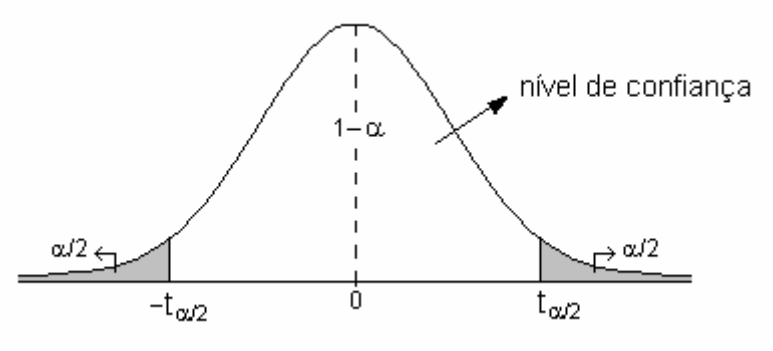
A distribuição t tem formato de campânula, é simétrica em torno da média ($\mu = 0$) que está localizada no centro da distribuição e varia de $-\infty$ a $+\infty$. Sua curva se assemelha à da distribuição normal padrão, sendo um pouco mais achatada no centro.

Como o parâmetro da distribuição t é o número de graus de liberdade ($\nu = n - 1$), o formato da curva se altera toda vez que muda o tamanho da amostra (n).

A distribuição t se aproxima da normal padrão à medida que o n cresce. Isto ocorre porque quando o tamanho da amostra se aproxima do tamanho da população ($n \rightarrow N$), o estimador S se aproxima do parâmetro σ ($S \rightarrow \sigma$) e, consequentemente, a estatística T se aproxima da variável Z ($T \rightarrow Z$).

Na prática, com 30 graus de liberdade a distribuição t é aproximadamente igual à distribuição normal padrão e com 120 graus de liberdade é exatamente igual, ou seja, as curvas se sobrepõem. Por essa razão, o tamanho 30 é adotado como referência para considerarmos uma amostra grande ou pequena. Quando n é menor ou igual a 30, a amostra é considerada pequena para utilizarmos a variável Z, devemos, portanto, utilizar a distribuição t para construir o intervalo.

Como já foi visto para a variável Z, na figura a seguir podemos observar que $1-\alpha$ é a probabilidade de que a variável T assuma um valor entre $-t_{\alpha/2}$ e $t_{\alpha/2}$ e α é a probabilidade de T não estar entre $-t_{\alpha/2}$ e $t_{\alpha/2}$.



Daí, temos

$$P(-t_{\alpha/2} < T < t_{\alpha/2}) = 1-\alpha.$$

Sabendo que $T = \dfrac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}}$ e fazendo a substituição, temos

$$P(-t_{\alpha/2} < \frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}} < t_{\alpha/2}) = 1-\alpha.$$

Isolando o parâmetro μ na expressão, temos

$$P(-t_{\alpha/2} < \frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}} < t_{\alpha/2}) = 1-\alpha$$

$$P(-t_{\alpha/2}\frac{S}{\sqrt{n}} < \bar{X}-\mu < t_{\alpha/2}\frac{S}{\sqrt{n}}) = 1-\alpha$$

$$P(-\bar{X}-t_{\alpha/2}\frac{S}{\sqrt{n}} < -\mu < -\bar{X}+t_{\alpha/2}\frac{S}{\sqrt{n}}) = 1-\alpha$$

$$P[(-\bar{X}-t_{\alpha/2}\frac{S}{\sqrt{n}} < -\mu < -\bar{X}+t_{\alpha/2}\frac{S}{\sqrt{n}}) \times (-1)] = 1-\alpha$$

$$P(\bar{X}+t_{\alpha/2}\frac{S}{\sqrt{n}} > \mu > \bar{X}-t_{\alpha/2}\frac{S}{\sqrt{n}}) = 1-\alpha$$

$$P(\bar{X}-t_{\alpha/2}\frac{S}{\sqrt{n}} < \mu < \bar{X}+t_{\alpha/2}\frac{S}{\sqrt{n}}) = 1-\alpha$$

donde resulta

$$IC\ (\mu;1-\alpha):\ \bar{X} \pm t_{\alpha/2}\frac{S}{\sqrt{n}},$$

sendo $t_{\alpha/2}$ o valor da estatística T que delimita a área $\alpha/2$. Este valor é encontrado na tabela da distribuição t de Student (Tabela II do Apêndice), a partir dos valores de ν e de α.

Generalizando a expressão, temos

$$IC\ (\theta\,;\ 1-\alpha):\ \hat{\theta} \pm t_{\alpha/2}\ S(\hat{\theta}).$$

Assim, para o caso particular de $\theta = \mu$, temos: $\hat{\theta} = \bar{X}$ e $S(\hat{\theta}) = S(\bar{X}) = \frac{S}{\sqrt{n}}$.

Para a utilização desta metodologia a seguinte pressuposição deve ser atendida:

A variável em estudo tem distribuição normal: $X \sim N(\mu, \sigma^2)$.

Devido à aproximação com a distribuição normal padrão a partir de ν=30, a estatística T, que tem distribuição t de Student, poderá ser utilizada para construir intervalos de confiança para a média, também quando a amostra for grande.

Consideremos o seguinte exemplo resolvido.

Através da amostra de tamanho 15 que segue, procura-se estimar a verdadeira potência média de aparelhos eletrônicos de alta sensibilidade medida em microwatts:

26,7; 25,8; 24,0; 24,9; 26,4; 25,9; 24,4; 21,7; 24,1; 25,9; 27,3; 26,9; 27,3; 24,8; 23,6.

Resolução:

Variável em estudo: X = potência de aparelhos eletrônicos de alta sensibilidade (μw)
Pressuposição: A variável em estudo tem distribuição normal.

Obtenção das estimativas:

$$\bar{x} = \frac{26,7 + 25,8 + \ldots + 23,6}{15} = 25,31\ \mu w$$

$$s^2 = \frac{(26,7 - 25,31)^2 + (25,8 - 25,31)^2 + \ldots + (23,6 - 25,31)^2}{15 - 1} = 2,493\ \mu w^2$$

$$s = \sqrt{2,493} = 1,579\ \mu w$$

Sendo $\theta = \mu$, temos

$$\hat{\theta} = \bar{X} = 25,31$$

$$S(\hat{\theta}) = S(\bar{X}) = \frac{S}{\sqrt{n}} = \frac{1,579}{\sqrt{15}} = 0,4076$$

$\nu = n - 1 = 15 - 1 = 14$

$t_{\alpha/2(\nu)} = 2,145$

$IC\ (\theta\,;\ 1-\alpha):\ \hat{\theta} \pm t_{\alpha/2}\ S(\hat{\theta})$

$IC\ (\mu;\ 0,95):\ 25,31 \pm 2,145 \times 0,4076$

$IC\ (\mu;\ 0,95):\ 25,31 \pm 0,874$

Limite inferior = 25,31 – 0,874 = 24,44

Limite superior = 25,31 + 0,874 = 26,18

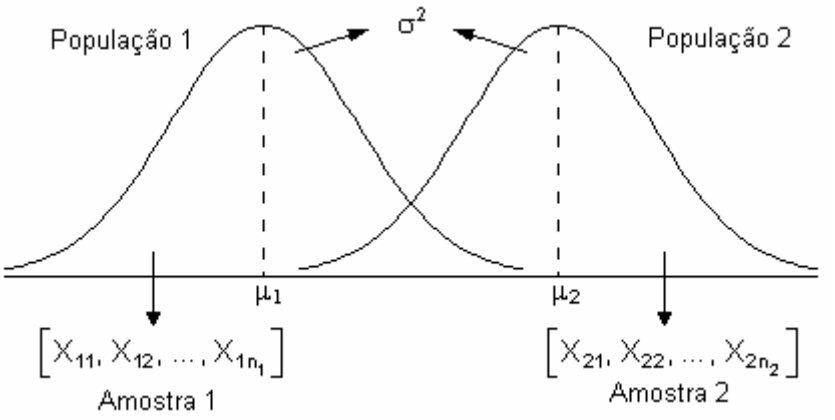
$P(24,44 < \mu < 26,18) = 0,95$

Concluímos que a probabilidade de o intervalo de 24,44μw a 26,18μw conter a verdadeira potência média de aparelhos eletrônicos de alta sensibilidade é de 0,95.

◆ **Intervalo de confiança para diferença entre médias de duas populações** (μ_1 - μ_2)

Para utilizar a estatística T no estudo de uma variável X em duas populações distintas, três pressuposições devem ser atendidas:

1. A variável em estudo tem distribuição normal.
   $X \sim N(\mu, \sigma^2)$
2. As variâncias das populações são iguais ($\sigma_1^2 = \sigma_2^2$).
3. As amostras retiradas das populações são independentes.



Atendidas as pressuposições, desejamos comparar as médias das populações, estimando por intervalo, o parâmetro $\theta = \mu_1 - \mu_2$. Utilizamos, então, a variável aleatória T.

$$T = \frac{\hat{\theta} - \theta}{S(\hat{\theta})} \sim t\ (\nu),$$

onde:

$\theta = \mu_1 - \mu_2$

$\hat{\theta} = \bar{X}_1 - \bar{X}_2$

$S(\hat{\theta}) = S(\bar{X}_1 - \bar{X}_2)$

$\nu = (n_1 - 1) + (n_2 - 1) = n_1 + n_2 - 2$

Vejamos como se obtém a estimativa do desvio (ou erro) padrão de $\hat{\theta}$. Sendo $\hat{\theta} = \bar{X}_1 - \bar{X}_2$, o desvio padrão $\hat{\theta}$ é obtido extraindo a raiz quadrada da variância da diferença entre as médias, ou seja,

$$S(\hat{\theta}) = S(\bar{X}_1 - \bar{X}_2) = \sqrt{S^2(\bar{X}_1 - \bar{X}_2)}\ .$$

Como as variáveis $\bar{X}_1$ e $\bar{X}_2$ são independentes, podemos utilizar a propriedade de que a variância da soma ou diferença de variáveis é igual à soma das variâncias dessas variáveis. Daí, temos

$$\sqrt{S^2(\bar{X}_1 - \bar{X}_2)} = \sqrt{S^2(\bar{X}_1) - S^2(\bar{X}_2)}\ .$$

Como a variância da média é $V(\bar{X}) = \frac{\sigma^2}{n}$, então, o estimador desta variância será $S^2(\bar{X}) = \frac{S^2}{n}$. Como consequência, temos

$$\sqrt{S^2\left(\bar{X}_1\right) + S^2\left(\bar{X}_2\right)} = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}\ .$$

Sendo as variâncias amostrais, S_1^2 e S_2^2, estimativas da mesma variância (σ^2), é possível combiná-las através da média. Assim, em vez de duas estimativas (S_1^2 e S_2^2), utilizamos S^2 que é a média das variâncias das amostras, ponderada pelos seus respectivos graus de liberdade, ou seja,

$$S^2 = \frac{S_1^2(n_1-1)+S_2^2(n_2-1)}{(n_1-1)+(n_2-1)}.$$

Daí resulta que

$$\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}} = \sqrt{\frac{S^2}{n_1}+\frac{S^2}{n_2}} = \sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)S^2}\ .$$

Assim, o estimador do erro padrão de $\hat{\theta}$ é dado por

$$S(\hat{\theta}) = S(\bar{X}_1-\bar{X}_2) = \sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)S^2}\ \text{, onde } S^2 = \frac{S_1^2(n_1-1)+S_2^2(n_2-1)}{(n_1-1)+(n_2-1)}.$$

Sabemos que, de modo geral, o intervalo de confiança para um parâmetro θ é assim definido

$$IC(\theta;1-\alpha): \hat{\theta} \pm t_{\alpha/2}\, S(\hat{\theta}).$$

Para o caso particular de $\theta = \mu_1 - \mu_2$, temos

$$IC(\mu_1-\mu_2;1-\alpha) = \bar{X}_1-\bar{X}_2 \pm t_{\alpha/2}\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)S^2}\ .$$

Vejamos um exemplo resolvido.

Dez cobaias adultas criadas em laboratório, foram separadas, aleatoriamente, em dois grupos: um foi tratado com ração normalmente usada no laboratório (padrão) e o outro grupo foi submetido a uma nova ração (experimental). As cobaias foram pesadas no início e no final do período de duração do experimento. Os ganhos de peso (em gramas) observados foram os seguintes:

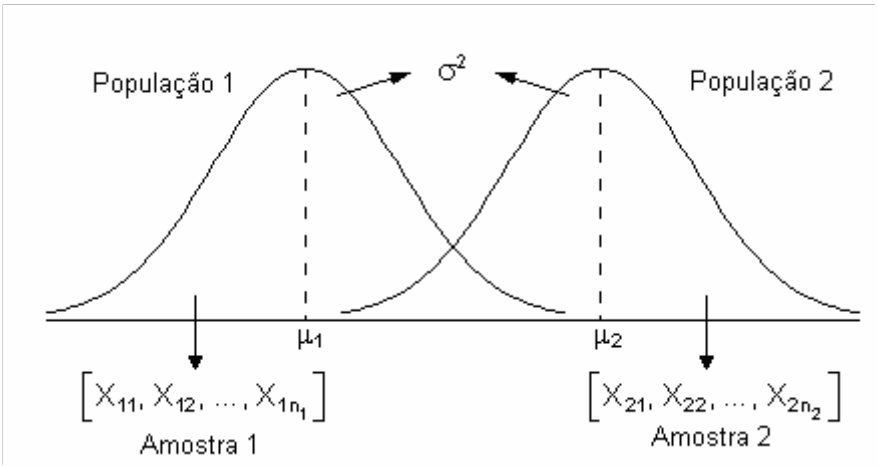
Ração padrão	200	180	190	190	180
Ração experimental	220	200	210	220	210

Construa o intervalo de confiança, ao nível de 99%, para a diferença entre as médias das duas populações.

Resolução:

Variável em estudo: X = ganho de peso (g)
Pressuposições:

- A variável em estudo tem distribuição aproximadamente normal, X ~ N (μ, σ^2).
- As variâncias das populações são iguais ($\sigma_1^2 = \sigma_2^2 = \sigma^2$).
- As amostras retiradas das populações são independentes.



Estimativas:

Amostra 1: $n_1 = 5$ $\bar{x}_1 = 188$ $s_1^2 = 70$

Amostra 2: $n_2 = 5$ $\bar{x}_2 = 212$ $s_2^2 = 70$

Sendo $\theta = \mu_1 - \mu_2$, temos:

$$\hat{\theta} = \bar{X}_1 - \bar{X}_2 = 188 - 212 = -24$$

$$S^2 = \frac{S_1^2(n_1 - 1) + S_2^2(n_2 - 1)}{(n_1 - 1) + (n_2 - 1)} = \frac{70 \times 4 + 70 \times 4}{4 + 4} = 70$$

$$S(\hat{\theta}) = S(\bar{X}_1 - \bar{X}_2) = \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) S^2} = \sqrt{\left(\frac{1}{5} + \frac{1}{5}\right) 70} = 5{,}292$$

$$\nu = (n_1 - 1) + (n_2 - 1) = 4 + 4 = 8$$

$t_{\alpha/2(\nu)} = 3{,}36$

$$IC(\theta; 1-\alpha): \hat{\theta} \pm t_{\alpha/2}\, S(\hat{\theta})$$

IC ($\mu_1 - \mu_2$; 0,99): -24 $\pm$ 3,36 $\times$ 5,292

IC ($\mu_1 - \mu_2$; 0,99): -24 $\pm$ 17,78

Limite inferior = -24 - 17,78 = -41,78
Limite superior = -24 + 17,78 = -6,22

P(-41,78 < $\mu_1 - \mu_2$ < -6,22) = 0,99

Concluímos que a probabilidade de o intervalo de -41,78 a -6,22 conter a verdadeira diferença entre o ganho de peso médio da população que recebeu ração padrão e o ganho de peso médio da população que recebeu a ração experimental é de 0,99. Como o valor zero está fora do intervalo podemos concluir que as médias não são iguais.

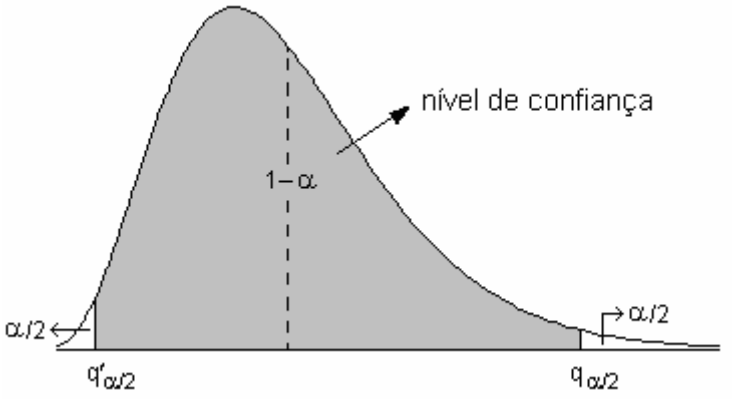
4.4.3.2. Intervalo de confiança para a variância de uma população (σ^2)

Sabemos que o estimador não-tendencioso de σ^2 é S^2. No entanto, para se construir um intervalo de confiança para σ^2 é necessário ainda conhecer como este estimador S^2 se comporta, ou seja, qual é a sua distribuição de probabilidade. Considerando uma população com distribuição normal, com média μ e variância σ^2, e que desta população seja selecionada uma amostra aleatória de tamanho n, então:

$$Q = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}{\sigma^2} = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(\nu),$$

ou seja, a variância da amostra (S^2) é uma variável aleatória que tem distribuição χ^2 com parâmetro ν=n−1 graus de liberdade. Assim, a distribuição χ^2 é a base para inferências a respeito da variância σ^2.

De acordo com a figura abaixo, vemos que $1-\alpha$ é a probabilidade de que a variável Q assuma um valor entre $q'_{\alpha/2}$ e $q_{\alpha/2}$ e α é a probabilidade de Q não estar entre $q'_{\alpha/2}$ e $q_{\alpha/2}$



Daí, temos

$$P(q'_{\alpha/2} < Q < q_{\alpha/2}) = 1-\alpha.$$

Sendo $Q = \frac{(n-1)S^2}{\sigma^2}$, ao substituirmos Q na expressão acima, obtemos:

$$P(q'_{\alpha/2} < \frac{(n-1)S^2}{\sigma^2} < q_{\alpha/2}) = 1-\alpha.$$

A manipulação algébrica desta desigualdade resulta no intervalo de confiança para σ^2:

$$P\left(\frac{(n-1)S^2}{q_{\alpha/2}} < \sigma^2 < \frac{(n-1)S^2}{q'_{\alpha/2}}\right) = 1-\alpha.$$

onde:

S^2 é o estimador da variância populacional σ^2;

n é o tamanho da amostra;

$\nu = n-1$ é o número de graus de liberdade associado à variância;

$q'_{\alpha/2}$ é o valor da distribuição qui-quadrado, com ν graus de liberdade, que delimita a área $\alpha/2$ à esquerda (Tabela III do Apêndice);

$q_{\alpha/2}$ é o valor da distribuição qui-quadrado com ν graus de liberdade que delimita a área $\alpha/2$ à direita (Tabela III do Apêndice).

Assim, os limites do intervalo de confiança para a variância populacional (σ^2) são dados por:

$$\left[\frac{(n-1)S^2}{q_{\alpha/2}}; \frac{(n-1)S^2}{q'_{\alpha/2}}\right].$$

Para determinar um intervalo de confiança para o desvio padrão populacional (σ) basta tomar a raiz quadrada positiva dos limites do intervalo para a variância populacional:

$$\left[\sqrt{\frac{(n-1)S^2}{q_{\alpha/2}}}; \sqrt{\frac{(n-1)S^2}{q'_{\alpha/2}}}\right]$$

Consideremos um exemplo resolvido:

Uma das maneiras de manter sob controle a qualidade de um produto é controlar sua variabilidade. Uma máquina de encher garrafas de refrigerante está regulada para enchê-las conforme uma distribuição normal com média de 200ml. Colheu-se uma amostra de 11 garrafas e observou-se uma variância de 8,38ml^2. Construa o intervalo, ao nível de 90% de confiança, para a variância populacional e um intervalo de mesma confiabilidade para o desvio padrão da população.

Resolução:

Devemos, inicialmente, determinar os valores da distribuição χ^2 com 10 graus de liberdade, de modo que $q'_{\alpha/2}$ e $q_{\alpha/2}$ tenham uma área igual a 0,05 à sua esquerda e à sua direita, respectivamente. Estes valores são: $q'_{\alpha/2(\nu)} = 3,94$ e $q_{\alpha/2(\nu)} = 18,31$.

Assim, o intervalo de confiança para a variância será:

$$\left[\frac{(n-1)s^2}{q_{\alpha/2}}; \frac{(n-1)s^2}{q'_{\alpha/2}}\right] = \left[\frac{(11-1)\times 8,38}{18,31}; \frac{(11-1)\times 8,38}{3,94}\right] = [4,58; 21,27].$$

E o intervalo de confiança para o desvio padrão será:

$$\left[\sqrt{\frac{(n-1)s^2}{q_{\alpha/2}}}; \sqrt{\frac{(n-1)s^2}{q'_{\alpha/2}}}\right] = \left[\sqrt{\frac{(11-1)\times 8,38}{18,31}}; \sqrt{\frac{(11-1)\times 8,38}{3,94}}\right] = [2,14; 4,61].$$

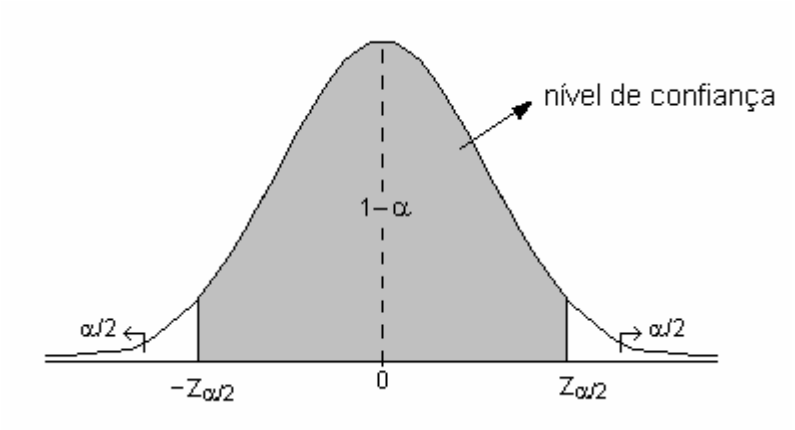
Concluímos, com uma confiança de 90%, que os intervalos 4,58 a 21,27 e 2,14 a 4,61 cobrem, respectivamente, a verdadeira variância e o verdadeiro desvio padrão da população.

4.4.3.3. Intervalo de confiança para a proporção de uma população (π)

Se o objetivo é estimar a proporção populacional (π), através de uma amostra aleatória desta população, utilizamos como estimador a proporção da amostra (P).

De acordo com o teorema central do limite, quando np>5 e n(1-p)>5, a distribuição amostral de P se aproxima da distribuição normal com média $\mu_P = \pi$ e desvio padrão $\sigma_P = \sqrt{\frac{\pi(1-\pi)}{n}}$.

Deste modo, podemos utilizar a distribuição normal para construir o intervalo de confiança para a proporção populacional. Lembramos que $1-\alpha$ é a probabilidade de que a variável Z assuma um valor entre $-z_{\alpha/2}$ e $z_{\alpha/2}$ e α é a probabilidade de Z não estar entre $-z_{\alpha/2}$ e $z_{\alpha/2}$.



Assim, temos que:

$$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1-\alpha.$$

Como $Z = \frac{P-\mu_P}{\sigma_P} = \frac{P-\pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}$, substituímos Z na expressão acima e obtemos:

$$P(-z_{\alpha/2} < \frac{P-\pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} < z_{\alpha/2}) = 1-\alpha$$

A manipulação algébrica desta desigualdade resulta no intervalo de confiança para π:

$$P(P - z_{\alpha/2}\sqrt{\frac{\pi(1-\pi)}{n}} < \pi < P + z_{\alpha/2}\sqrt{\frac{\pi(1-\pi)}{n}}) = 1-\alpha.$$

Como podemos verificar na expressão, o erro padrão $\sigma_P = \sqrt{\frac{\pi(1-\pi)}{n}}$ é um valor desconhecido, uma vez que não conhecemos π. Entretanto, com base na propriedade de consistência dos estimadores, quando o tamanho da amostra é grande, podemos considerar o valor do estimador (P) suficientemente próximo do parâmetro (π), o que possibilita a substituição de π por P na expressão do intervalo de confiança. Assim, temos:

$$P(P - z_{\alpha/2}\sqrt{\frac{P(1-P)}{n}} < \pi < P + z_{\alpha/2}\sqrt{\frac{P(1-P)}{n}}) = 1-\alpha.$$

onde:

P é o estimador da proporção populacional π;
n é o tamanho da amostra e
$z_{\alpha/2}$ é o valor da variável Z que delimita a área $\alpha/2$ (Tabela I do Apêndice).

Vejamos um exemplo resolvido:

Foi realizada uma pesquisa de mercado para verificar a preferência da população de em relação ao consumo de determinado produto. Para isso, foi colhida uma amostra de 300 consumidores, dos quais 180 disseram consumir o produto. Encontre o intervalo ao nível de 99% de confiança para a proporção de consumidores do produto na população.

Resolução:

A estimativa por ponto para a proporção populacional será: p = 180/300 = 0,60.

Como o nível de confiança adotado é de 99%, temos α = 0,01. Assim, o valor de Z que delimita a área $\alpha/2$ = 0,005 à direita é 2,575.

Então, o intervalo de confiança de 99% para a proporção populacional será:

$$IC(\pi;0,99): P \pm z_{\alpha/2}\sqrt{\frac{P(1-P)}{n}}$$

$$IC(\pi;0,99): 0,60 \pm 2,575\sqrt{\frac{0,60(1-0,60)}{300}}$$

$$IC(\pi;0,99): 0,60 \pm 2,575 \times 0,0283$$

$$IC(\pi;0,99): 0,60 \pm 0,0728$$

Limite inferior = 0,60 – 0,0728 = 0,5272
Limite superior = 0,60 + 0,0728 = 0,6728

$$P(0,5272 < \pi < 0,6728) = 0,99$$

Concluindo, pode-se afirmar, com uma confiança de 99%, que o intervalo de 0,53 a 0,67 contém a proporção populacional de consumidores que preferem o produto pesquisado.

Exercícios propostos:

4.1. Um engenheiro de desenvolvimento de um fabricante de pneus está investigando a vida do pneu em relação a um novo componente de borracha. Ele fabricou 40 pneus e testou-os até o fim da vida em um teste na estrada. A média e o desvio padrão da amostra são 61.492 km e 6.085 km, respectivamente. O engenheiro acredita que a vida média desse novo pneu está em excesso em relação a 60.000 km. Obtenha o intervalo de confiança, ao nível de 95%, para a vida média do pneu e conclua a respeito da suposição do engenheiro.

4.2. Um agrônomo realizou um levantamento para estudar o desenvolvimento de duas espécies de árvores, a Bracatinga e a Canafístula. Para esta finalidade foram coletadas duas amostras de tamanhos igual a 10 árvores. Os resultados para altura, em metros, estão descritos abaixo para as duas amostras:

Bracatinga	6,5	6,9	6,9	8,6	8,7	8,2	10,0	10,3	13,4	14,4
Canafístula	9,3	10,1	11,4	15,2	17,2	14,8	15,9	20,6	21,9	23,8

Para verificar a hipótese de que as alturas das duas espécies são diferentes, o agrônomo adotou o seguinte critério. Construir os intervalos com 95% de confiança, para cada uma das espécies. Se os intervalos se sobrepõem (se interceptam) concluir que não há diferenças significativas entre as duas alturas medias, caso contrário, concluir que há diferenças entre as mesmas. Baseado neste critério qual a conclusão do agrônomo?

4.3. Na fabricação de semicondutores o ataque químico por via úmida é frequentemente usado para remover silicone da parte posterior das pastilhas antes da metalização. A taxa de ataque é uma característica importante nesse processo e é sabido que ela segue uma distribuição normal. Duas soluções diferentes para ataque químico são comparadas, usando duas amostras aleatórias de pastilhas. As taxas observadas de ataque (10-3 polegadas/min) são dadas a seguir:

Solução 1	9,9	9,4	9,3	9,6	10,2	10,6	10,3	10,0	10,3	10,1
Solução 2	10,2	10,6	10,7	10,4	10,5	10,0	10,7	10,4	10,3	-

Os dados justificam a afirmação de que a taxa média de ataque seja a mesma para as duas soluções? Considere que ambas as populações têm variâncias iguais, construa o intervalo de confiança, ao nível de 95%, para a diferença entre as médias e conclua.

4.4. Considere os dados do exercício 4.1. Construa um intervalo de 90% para a variância da vida do pneu. Depois converta esse intervalo apresentando-o em termos de desvio padrão.

4.5. Uma amostra aleatória de 250 dispositivos eletrônicos apresentou 27 unidades defeituosas. Estime a fração de não conformes e construa um intervalo de 95% de confiança para o verdadeiro valor da fração de não conformes.

4.5. Testes de hipóteses

O teste de hipótese é um procedimento estatístico em que se busca verificar uma hipótese a respeito da população, no sentido de aceitá-la ou rejeitá-la, a partir de dados amostrais, tendo por base a teoria das probabilidades.

Em geral, um problema científico (expresso na forma de pergunta) conduz a uma hipótese científica (resposta provisória a esta pergunta) que requer uma pesquisa científica para a sua verificação. O teste de hipótese é um dos procedimentos mais utilizados na pesquisa científica, sobretudo na pesquisa experimental.

De modo geral, podemos definir cinco passos para construção de um teste de hipóteses:

1. Definir as hipóteses estatísticas.
2. Fixar a taxa de erro aceitável.
3. Escolher a estatística para testar a hipótese e verificar as pressuposições para o seu uso.
4. Usar as observações da amostra para calcular o valor da estatística do teste.
5. Decidir sobre a hipótese testada e concluir.

4.5.1. Testes para a média populacional

- **Hipóteses estatísticas**

A hipótese estatística é uma suposição feita a respeito de um ou mais *parâmetros*, tais como, médias de populações (μ), variâncias de populações (σ^2), etc. As hipóteses estatísticas surgem de problemas científicos.

Existem dois tipos básicos de hipóteses estatísticas:

Hipótese de nulidade (H_0): é a hipótese que está sob verificação. Esta hipótese supõe a igualdade dos parâmetros que estão sendo testados.

Hipótese alternativa (H_A): é a hipótese que será considerada caso a hipótese de nulidade seja rejeitada. Esta hipótese supõe que os parâmetros testados são diferentes.

Duas situações são comuns em testes de hipóteses a respeito da média da população (μ):

1. Comparação de uma média (μ) com um valor padrão (μ_0)

Nesta situação, temos uma população da qual é extraída uma amostra e a média desta amostra é comparada com um valor já conhecido (valor padrão) que serve como referência.

$$
\begin{array}{l}
H_0 : \mu = \mu_0 \text{ ou } \mu - \mu_0 = 0 \\
\left.\begin{array}{l}
H_A : \mu \neq \mu_0 \text{ ou } \mu - \mu_0 \neq 0 \quad \text{hipótese bilateral} \\
\quad \mu > \mu_0 \text{ ou } \mu - \mu_0 > 0 \quad \text{hipótese unilateral direita} \\
\quad \mu < \mu_0 \text{ ou } \mu - \mu_0 < 0 \quad \text{hipótese unilateral esquerda}
\end{array}\right\} \text{Devemos escolher a } H_A \text{ mais apropriada}
\end{array}
$$

2. Comparação entre duas médias (μ_1 e μ_2)

Nesta situação, temos duas populações, de cada uma é extraída uma amostra, e as médias das duas amostras são comparadas.

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

$H_A : \mu_1 \neq \mu_2$ ou $\mu_1 - \mu_2 \neq 0$ hipótese bilateral

$\mu_1 > \mu_2$ ou $\mu_1 - \mu_2 > 0$ hipótese unilateral direita

$\mu_1 < \mu_2$ ou $\mu_1 - \mu_2 < 0$ hipótese unilateral esquerda

} Devemos escolher a H_A mais apropriada

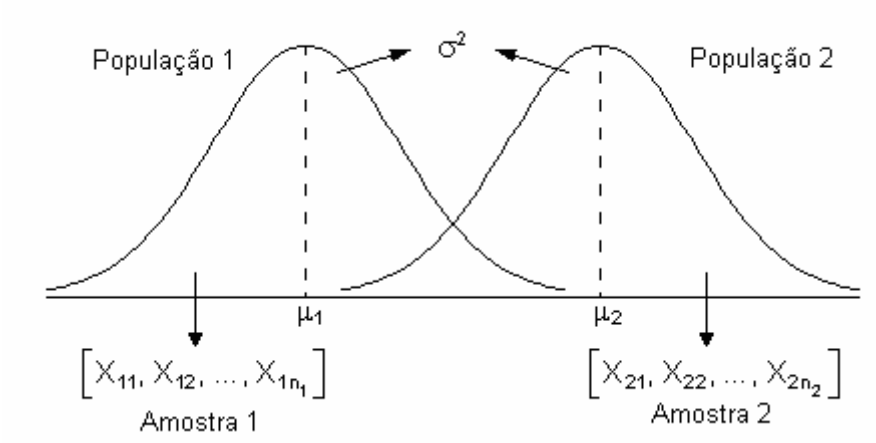
Consideremos o exemplo a seguir.

Vamos supor que nosso problema científico seja a pergunta “As raças bovinas Holandesa e Jersey diferem quanto à produção de leite?” e nossa hipótese científica seja a afirmação “A raça Holandesa produz mais leite que a raça Jersey”. Esta hipótese pode ser verificada de duas formas: pela avaliação das populações inteiras de vacas das duas raças, ou seja, todas as vacas das raças Holandesa e Jersey, ou por meio da avaliação de amostras que serão retiradas dessas populações. Obviamente, seria impossível avaliar todas as vacas das duas raças. E ainda que fosse possível, sabemos que o processo de amostragem pode fornecer precisão suficiente; portanto, será muito mais econômico e menos trabalhoso utilizar uma amostra.

Ao utilizarmos amostras, consideremos que temos duas populações:

População 1 – vacas da raça Holandesa
População 2 – vacas da raça Jersey

Nestas populações vamos estudar a variável contínua X = produção de leite, supondo que $X \sim N(\mu, \sigma^2)$ e que $\sigma_1^2 = \sigma_2^2$, conforme figura abaixo



Assim nossos parâmetros de interesse são:

$E(X_1) = \mu_1$ = produção média da população 1
$E(X_2) = \mu_2$ = produção média da população 2

Assim, devemos considerar as seguintes hipóteses estatísticas.

$$\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_A : \mu_1 \neq \mu_2 \end{cases}$$

Observemos que a hipótese alternativa não corresponde necessariamente à expectativa do pesquisador, ou seja, à hipótese científica. A hipótese a ser testada em um teste é sempre a hipótese de igualdade entre os parâmetros, enquanto a hipótese alternativa deve ser definida pelo pesquisador, podendo ser bilateral ou unilateral. A escolha entre uma e outra, entretanto, jamais deve ser feita com base nos dados da amostra, tampouco na expectativa do pesquisador. A hipótese bilateral é mais genérica e deve ser utilizada quando não temos motivos suficientes para esperar que um dos parâmetros seja maior ou menor que o outro. Assim, supomos apenas que os parâmetros serão diferentes, caso a hipótese de igualdade seja rejeitada.

As situações de aplicação da hipótese unilateral são mais restritas e nem sempre são muito claras. A opção por uma hipótese unilateral exige que tenhamos mais informações sobre o comportamento da variável de interesse na situação da pesquisa. Estudos anteriores, por exemplo, podem prover evidências que suportem uma hipótese unilateral.

A hipótese unilateral pode ser também uma decorrência lógica da situação de pesquisa, como, por exemplo, quando comparamos a média de um grupo tratado (que recebe determinado tratamento) com a média de um grupo controle ou testemunha (que não recebe o tratamento). Neste caso, se o tratamento não tem efeito, esperamos que as médias dos dois grupos sejam iguais; mas se o tratamento tem efeito significativo é bastante razoável esperar que a média do grupo tratado (que expressa este efeito) seja maior (e nunca menor do que a média do grupo controle. Outra situação típica da aplicação da hipótese unilateral será apresentada e discutida na seção 4.9.

Um teste de hipótese também pode ser classificado de acordo com o tipo de hipótese alternativa que adota: se a hipótese alternativa é bilateral, dizemos que o teste é bilateral; se a hipótese é unilateral, o teste é unilateral.

◆ Erros de Conclusão

Já vimos que um elemento intrínseco ao processo de inferência é o erro. Num teste de hipóteses, devemos considerar que as hipóteses estatísticas são estabelecidas a respeito de valores populacionais (parâmetros) e as conclusões são obtidas a partir de dados amostrais (estimativas), ou seja, baseamos nossas conclusões em apenas uma parte da informação (amostra) que, eventualmente, pode não representar o todo (população), portanto, existe a possibilidade de estarmos cometendo um erro de conclusão. Como a hipótese sob verificação é H_0, dois tipos de erro estão associados à decisão a respeito dela, são eles:

Erro Tipo I: rejeitar H_0 quando ela é verdadeira

α = P(erro tipo I) $\rightarrow$ probabilidade de cometer o erro tipo I

Erro Tipo II: não reijeitar H_0 quando ela é falsa

β = P(erro tipo II) $\rightarrow$ probabilidade de cometer o erro tipo II

A tabela a seguir ilustra os dois tipos de erro.

Decisão	Situação de H_0	
	Verdadeira	Falsa
Não rejeitar	Acerto	Erro Tipo II
Rejeitar	Erro Tipo I	Acerto

Como consequência, temos que: $1-\alpha$ é a probabilidade de não cometer o erro tipo I, ou seja, é a capacidade de não rejeitar H_0 verdadeira, e $1-\beta$ é a probabilidade de não cometer o erro tipo II, ou seja, é a capacidade de rejeitar H_0 falsa. A probabilidade $1-\beta$ é denominada

poder do teste. Podemos dizer, então, que o poder do teste é a probabilidade de declarar diferenças quando elas, de fato, existem. O poder de um teste está relacionado com os seguintes fatores: tamanho da amostra, variabilidade da variável e magnitude da diferença existente entre as médias.

É importante ressaltar ainda que as duas taxas de erro (α e β) estão relacionadas negativamente, de modo que a redução de α implica no aumento de β e vice-versa. Para que os testes de hipóteses tenham validade, é necessário que sejam delineados de modo a minimizar os erros de conclusão. Entretanto, o único meio de reduzir ambos os tipos de erro é aumentando o tamanho da amostra, o que nem sempre é viável. Na prática, devemos definir qual dos dois erros é mais grave e, então, minimizá-lo. Podemos adiantar, contudo, que, via de regra, a preocupação está voltada para o erro tipo I, pois na maioria dos casos ele é considerado o mais grave. A probabilidade de ocorrência do erro tipo I (α) é chamada de *nível de significância* do teste.

◆ **Estatística do teste**

Para testar hipóteses a respeito do parâmetro μ, utilizamos a variável aleatória T,

$$T = \frac{\hat{\theta} - \theta}{S(\hat{\theta})}.$$

Vejamos agora como o valor da variável T é obtido nas duas situações mais comuns de testes de hipóteses a respeito de μ.

Situação 1. Comparação de uma média (μ) com um valor padrão (μ_0).

Inicialmente, devemos lembrar que para compararmos uma média com um valor padrão utilizando a variável aleatória T, a seguinte pressuposição deve ser verdadeira:

A variável em estudo tem distribuição normal, ou seja, $X \sim N(\mu, \sigma^2)$.

Neste caso, a hipótese estatística sob verificação será:

$$H_0 : \mu = \mu_0 .$$

Sob H_0 verdadeira, a variável aleatória $T = \frac{\hat{\theta} - \theta}{S(\hat{\theta})}$ tem distribuição t com parâmetro ν, onde:

$\theta = \mu = \mu_0$

$\hat{\theta} = \bar{X}$

$S(\hat{\theta}) = S(\bar{X}) = \frac{S}{\sqrt{n}}$

$\nu = n - 1$

Daí resulta que $T = \frac{\hat{\theta} - \theta}{S(\hat{\theta})} = \frac{\bar{X} - \mu_0}{\frac{S}{\sqrt{n}}}$

Podemos verificar que, sendo μ_0 um valor conhecido, a variável aleatória T torna-se uma estatística, pois passa a ser função apenas da amostra.

Situação 2. Comparação entre duas médias (μ_1 e μ_2)

Ao compararmos duas médias populacionais utilizando a estatística T, três pressuposições devem ser verdadeiras:

1. A variável em estudo tem distribuição normal, ou seja, X ~ N (μ, σ^2);
2. As variâncias das populações são iguais ($\sigma_1^2 = \sigma_2^2$);
3. As amostras retiradas das populações são independentes.

Atendidas as pressuposições, a hipótese estatística sob verificação será:

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

Quando H_0 supõe que o parâmetro estimado é igual a zero, ou seja, $\theta = \mu_1 - \mu_2 = 0$, temos:

$$T = \frac{\hat{\theta} - \theta}{S(\hat{\theta})} = \frac{\hat{\theta} - 0}{S(\hat{\theta})} = \frac{\hat{\theta}}{S(\hat{\theta})} \sim t\ (\nu).$$

Daí resulta que $T = \dfrac{\hat{\theta}}{S(\hat{\theta})} = \dfrac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right) S^2}}$,

onde:

$\hat{\theta} = \bar{X}_1 - \bar{X}_2$

$S(\hat{\theta}) = \bar{X}_1 - \bar{X}_2 = \sqrt{\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right) S^2}$, sendo $S^2 = \dfrac{S_1^2 (n_1 - 1) + S_2^2 (n_2 - 1)}{(n_1 - 1) + (n_2 - 1)}$

$\nu = (n_1 - 1) + (n_2 - 1)$

Observemos que, também nesta situação, a variável aleatória T passa a ser uma estatística.

◆ **Critério de decisão**

A regra de decisão a respeito de H_0 pode ser estabelecida com base num valor crítico:

Teste bilateral: se a hipótese alternativa for bilateral, o valor crítico será:

$t_{\alpha/2(\nu)}$: valor da estatística T, para ν graus de liberdade, que delimita a área $\alpha/2$, encontrado na tabela da distribuição t (limites bilaterais da Tabela II do Apêndice).

Teste unilateral: se a hipótese alternativa for unilateral, o valor crítico será:

$t_{\alpha(\nu)}$: valor da estatística T, para ν graus de liberdade, que delimita a área α, encontrado na tabela da distribuição t (limites unilaterais da Tabela II do Apêndice).

Para decidir comparamos o valor da estatística $T = \dfrac{\hat{\theta}}{S(\hat{\theta})}$ com o valor crítico:

– Rejeitamos H_0, ao nível α, se o valor da estatística, em módulo, for maior que o valor crítico:

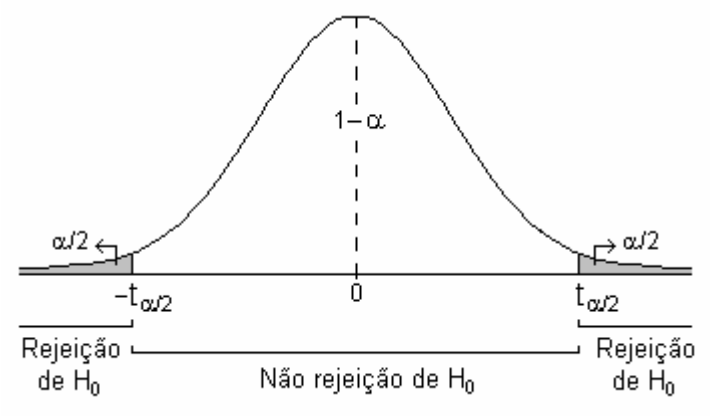
$$| t | > \text{valor crítico}$$

– Não temos motivos suficientes para rejeitar H_0, ao nível α, se o valor da estatística, em módulo, for menor que o valor crítico:

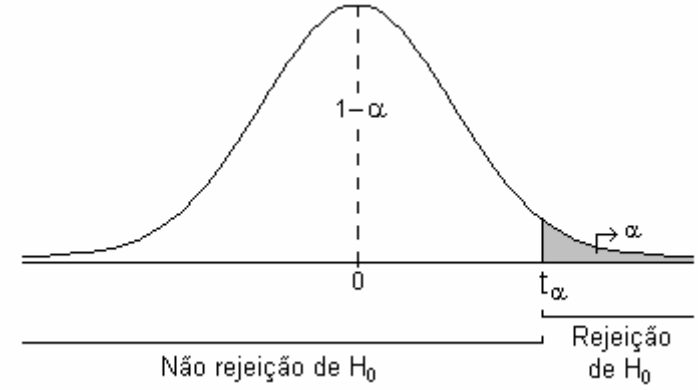
$$| t | < \text{ valor crítico}$$

Podemos observar a seguir as regiões de rejeição H_0 na curva da distribuição t para cada uma das três possibilidades de hipótese alternativa:

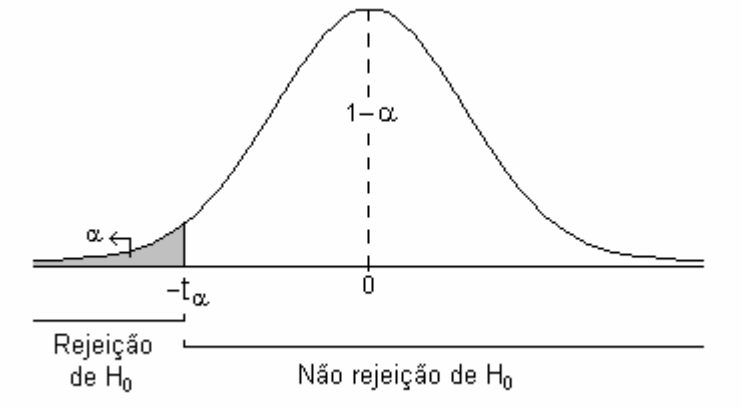
– Para hipótese alternativa bilateral, ou seja, $H_A : \mu_1 - \mu_2 \neq 0$



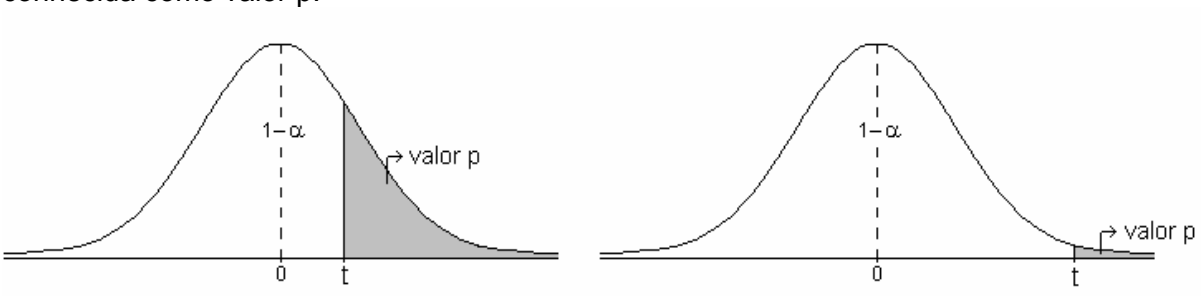
– Para hipótese alternativa unilateral direita, ou seja, $H_A : \mu_1 - \mu_2 > 0$, temos:



– Para hipótese alternativa unilateral esquerda, ou seja, $H_A : \mu_1 - \mu_2 < 0$, temos:

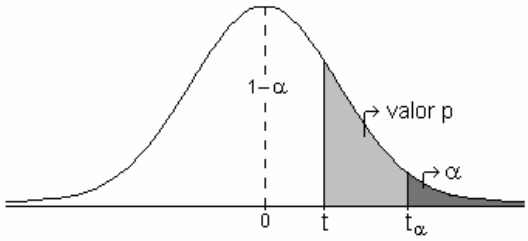


Outro critério tem sido frequentemente utilizado para decidir sobre H_0. Essa decisão também pode ser baseada em um valor que expressa a probabilidade de que seja obtido um valor t mais extremo que o valor observado, dado que H_0 é verdadeira. Esta probabilidade é conhecida como valor p.

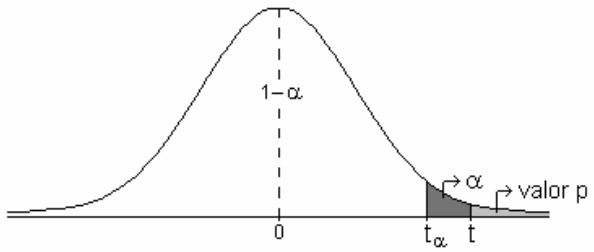


A decisão a respeito de H_0 é tomada da seguinte forma:

Se o valor p for maior ou igual a α, não rejeitamos a hipótese nula, pois t é típico ou está em uma região de alta probabilidade.



Se o valor p for menor que α, rejeitamos a hipótese nula, pois t é atípico ou está em uma região de baixa probabilidade.



- **Considerações finais**

Os intervalos de confiança e os testes de hipóteses bilaterais são procedimentos estatísticos equivalentes. Portanto, se forem utilizados para analisar os mesmos dados, ao mesmo nível de significância, devem conduzir aos mesmos resultados.

O intervalo de confiança para uma média equivale ao teste de hipóteses que compara uma média com um padrão. Observe as expressões:

Intervalo de confiança para uma média (μ)

$$IC(\mu;1-\alpha): \bar{X} \pm t_{\alpha/2} \frac{S}{\sqrt{n}}.$$

Estatística T para a comparação de uma média (μ) com um valor padrão (μ_0)

$$T = \frac{\bar{X} - \mu_0}{\frac{S}{\sqrt{n}}}.$$

Da mesma forma, o intervalo de confiança para a diferença entre duas médias equivale ao teste de hipóteses que compara duas médias. Observe as expressões:

Intervalo de confiança para a diferença entre duas médias ($\mu_1 - \mu_2$)

$$IC(\mu_1 - \mu_2; 1-\alpha): \bar{X}_1 - \bar{X}_2 \pm t_{\alpha/2}\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)S^2}$$

Estatística T para a comparação entre duas médias (μ_1 e μ_2)

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)S^2}}$$

Exemplos:

1. Se no teste de hipóteses, ao nível de 1% de significância, rejeitamos $H_0 : \mu - \mu_0 = 0$, significa que a diferença entre a média e o valor padrão é diferente de zero, ou seja, a média é diferente do valor padrão. Construindo o intervalo de confiança para μ, ao nível de 99%, devemos esperar que o valor padrão (μ_0) esteja fora do intervalo. Caso contrário, os resultados seriam contraditórios.

2. Se no teste de hipóteses, ao nível de 5% de significância, não rejeitamos $H_0 : \mu_1 - \mu_2 = 0$ significa que a diferença entre as duas médias deve ser zero, ou seja, as médias podem ser consideradas iguais. Construindo o intervalo de confiança, ao nível de 95%, para a diferença entre essas médias ($\mu_1 - \mu_2$), devemos esperar que o valor zero esteja dentro do intervalo. Caso contrário, os resultados seriam contraditórios.

Vejamos a seguir dois exemplos resolvidos:

Exemplo 1. Um botânico recebeu a informação de que o diâmetro médio de flores de uma determinada planta é de 9,6cm. Para testar a veracidade da informação, tomou uma amostra aleatória de 30 plantas, cujo diâmetro médio de flores observado foi 9,3cm, com desvio padrão de 3,2cm.

a) Verifique, utilizando teste de hipóteses ao nível de 5% de significância, se a informação recebida pelo botânico é verdadeira.
b) Verifique se a informação é verdadeira, utilizando intervalo de confiança ao nível de 95%.
c) Houve coerência entre os resultados do teste de hipóteses e do intervalo de confiança?

Resolução:

a) Teste de hipóteses

Variável em estudo: X = diâmetro de flores (cm)

1. Pressuposição: A variável em estudo tem distribuição normal.

2. Hipóteses estatísticas: $\begin{cases} H_0 : \mu - \mu_0 = 0 \\ H_A : \mu - \mu_0 \neq 0 \end{cases}$

A hipótese de nulidade supõe que o diâmetro médio de flores da população desta espécie de planta é igual ao valor padrão 9,6cm.

3. Estatística do teste

$\theta = \mu - \mu_0 = 0$

$\hat{\theta} = \bar{X} - \mu_0$

$S(\hat{\theta}) = S(\bar{X}) = \frac{S}{\sqrt{n}}$

$\nu = n - 1$

$T = \frac{\hat{\theta}}{S(\hat{\theta})} = \frac{\bar{X} - \mu_0}{\frac{S}{\sqrt{n}}}$

Assim, temos:

$\bar{x} - \mu_0 = 9,3 - 9,6 = -0,3$

$\frac{s}{\sqrt{n}} = \frac{3,2}{\sqrt{30}} = 0,5842$

$\nu = 30 - 1 = 29$

$t_{\alpha/2(29)} = 2,045$

$t = \frac{-0,3}{0,5842} = -0,5135$

4. Decisão e conclusão

Como $|t = -0,5135| < t_{\alpha/2(29)} = 2,045$, não temos motivos para rejeitar H_0. Concluímos, ao nível de 5% de significância, que o diâmetro médio de flores desta planta não difere significativamente do valor padrão $\mu_0 = 9,6$. Portanto, a informação recebida pelo botânico é verdadeira.

b) Intervalo de confiança

Pressuposição: A variável em estudo tem distribuição normal.

Estimativas:

$\bar{x} = 9,3$

$\frac{s}{\sqrt{n}} = \frac{3,2}{\sqrt{30}} = 0,5842$

$\nu = 30 - 1 = 29$

$t_{\alpha/2(29)} = 2,045$

Sendo $IC(\theta; 1-\alpha): \hat{\theta} \pm t_{\alpha/2} S(\hat{\theta})$

$$IC(\mu; 1-\alpha): \bar{X} \pm t_{\alpha/2} \frac{S}{\sqrt{n}}$$

temos:

IC (μ; 0,95): $9,3 \pm 2,045 \times 0,5842$

Limite inferior = 9,3 – 1,195 = 8,11
Limite superior = 9,3 + 1,195 = 10,50

P(8,11 < μ < 10,50) = 0,95

Concluímos que o intervalo de confiança, ao nível de 95%, para o verdadeiro diâmetro médio de flores desta planta é de 8,11 a 10,50cm.

c) Sim, o resultado do teste de hipóteses esta coerente com o do intervalo de confiança, pois o valor padrão 9,6, que segundo o teste de hipótese não difere de μ, está dentro do intervalo de confiança, ou seja, é um valor possível para μ.

Exemplo 2. Para investigar se o treinamento é ou não transferido pelo ácido nucléico, 10 ratos foram treinados em discriminar se havia luz ou escuridão. Posteriormente, esses ratos foram mortos, o ácido nucléico dos mesmos foi extraído e injetado em 10 ratos. Simultaneamente o ácido nucléico de 10 ratos não treinados foi injetado em outros 10. Os 20 ratos injetados com ácido nucléico foram observados durante um período de tempo quanto à capacidade de discriminar luz e escuridão. O número de erros relativos a cada rato está na tabela abaixo.

Treinados	7	9	6	11	13	8	7	13	12	9
Não treinados	12	8	9	13	14	9	8	10	7	15

a) Verifique, utilizando teste de hipóteses ao nível de 5% de significância, se o treinamento é ou não transferido pelo ácido nucléico.
b) Construa o intervalo de confiança, ao nível de 95%, para a diferença entre as médias das duas populações.
c) Houve coerência entre os resultados do teste de hipóteses e do intervalo de confiança?

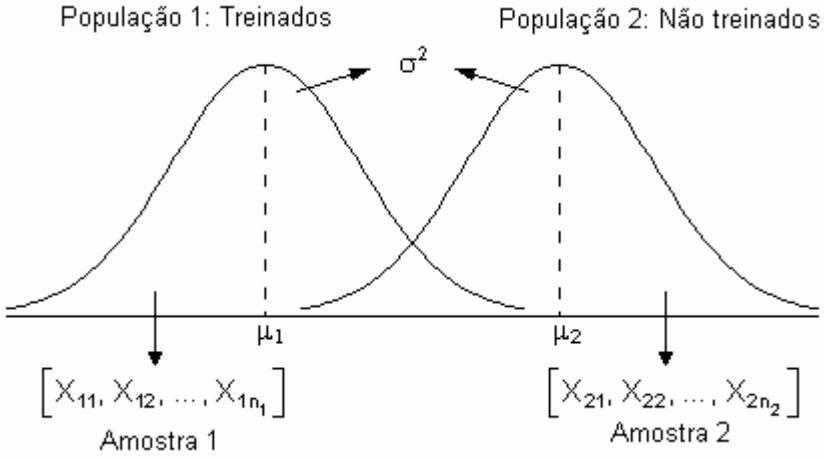
Resolução:

a) Teste de hipóteses

Variável em estudo: X = número de erros ao discriminar luz e escuridão

1. Pressuposições:

– A variável em estudo tem distribuição normal, ou seja, X ~ N (μ, σ^2);
– As variâncias das populações são iguais ($\sigma_1^2 = \sigma_2^2 = \sigma$);
– As amostras retiradas das populações são independentes.



2. Hipóteses estatísticas: $\begin{cases} H_0 : \mu_1 - \mu_2 = 0 \\ H_A : \mu_1 - \mu_2 \neq 0 \end{cases}$

A hipótese de nulidade supõe a igualdade entre as médias das duas populações.

3. Estatística do teste

$$\theta = \mu_1 - \mu_2 = 0$$
$$\hat{\theta} = \bar{X}_1 - \bar{X}_2$$
$$S(\hat{\theta}) = S(\bar{X}_1 - \bar{X}_2) = \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)S^2}$$
$$\nu = (n_1 - 1) + (n_2 - 1)$$
$$T = \frac{\hat{\theta}}{S(\hat{\theta})} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)S^2}}$$

Amostra 1: $n_1 = 10$ $\bar{x}_1 = 9,5$ $s_1^2 = 6,722$

Amostra 2: $n_2 = 10$ $\bar{x}_2 = 10,5$ $s_2^2 = 7,833$

$$\bar{x}_1 - \bar{x}_2 = 9,5 - 10,5 = -1$$
$$s^2 = \frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{(n_1 - 1) + (n_2 - 1)} = \frac{6,722 \times 9 + 7,833 \times 9}{9 + 9} = 7,278$$
$$\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)s^2} = \sqrt{\left(\frac{1}{10} + \frac{1}{10}\right)7,278} = 1,206$$

$$\nu = 9 + 9 = 18$$
$t_{\alpha/2(18)} = 2,101$

$$t = \frac{-1}{1,206} = -0,8292$$

4. Decisão e conclusão

Como $|t = -0,8292| < t_{\alpha/2(18)} = 2,101$, não temos motivos para rejeitar H_0. Concluímos, então, ao nível de 5% de significância, que a média de erros do grupo que recebeu ácido nucléico de ratos treinados não diferiu significativamente da média de erros do grupo que recebeu ácido nucléico de ratos não treinados. Se o treinamento fosse transferido pelo acido nucléico, a média de erros da população 1 deveria ser menor que a média de erros da população 2. Portanto, há evidências de que o treinamento não é transferido pelo ácido nucléico.

b) Intervalo de confiança para a diferença entre as médias

Variável em estudo: X = número de erros ao discriminar luz e escuridão

Pressuposições:
– A variável em estudo tem distribuição normal, ou seja, $X \sim N(\mu, \sigma^2)$;
– As variâncias das populações são iguais ($\sigma_1^2 = \sigma_2^2 = \sigma^2$);
– As amostras retiradas das populações são independentes.

Estimativas:

$$\bar{x}_1 - \bar{x}_2 = 9,5 - 10,5 = -1$$

$$s^2 = \frac{s_1^2(n_1-1)+s_2^2(n_2-1)}{(n_1-1)+(n_2-1)} = \frac{6{,}722\times 9 + 7{,}833\times 9}{9+9} = 7{,}278$$

$$\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)s^2} = \sqrt{\left(\frac{1}{10}+\frac{1}{10}\right)7{,}278} = 1{,}206$$

$\nu = 9+9 = 18$
$t_{\alpha/2(18)} = 2{,}101$

Sendo $IC(\mu_1-\mu_2;1-\alpha): \bar{X}_1 - \bar{X}_2 \pm t_{\alpha/2}\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)S^2}$,

temos:

IC (μ_1 - μ_2; 0,95): -1 $\pm$ 2,101 $\times$ 1,206
IC (μ_1 - μ_2; 0,95): -1 $\pm$ 2,533
Limite inferior = -1 - 2,533 = -3,533
Limite superior = -1 + 2,533 = 1,533

P(-3,533 < μ_1 - μ_2 < 1,533) = 0,95

Concluímos que a probabilidade de a verdadeira diferença entre a média de erros da população que recebeu ácido nucléico de ratos treinados e a média de erros da população que recebeu ácido nucléico de ratos não treinados estar entre -3,533 e 1,533 é de 0,95.

c) Pelo teste de hipóteses, concluímos que a verdadeira diferença entre as médias deve ser zero e, pelo intervalo de confiança, concluímos que zero é um valor possível para a verdadeira diferença entre as médias, uma vez que se encontra dentro do intervalo. Portanto, o resultado do teste de hipóteses está de acordo com o do intervalo de confiança.

4.5.2. Testes para a variância populacional

4.5.2.1. Teste para a variância de uma população

Para aplicar o teste para a variância é necessário supor a normalidade da população de onde será extraída a amostra. Se essa suposição é violada, o teste deixa de ser exato. Uma hipótese testada com frequência é que a variância tenha um valor especificado σ_0^2. Assim, as hipóteses a serem testadas são:

$H_0 : \sigma^2 = \sigma_0^2$
$H_A : \sigma^2 \neq \sigma_0^2 \leftarrow$ hipótese bilateral
$\sigma^2 > \sigma_0^2 \leftarrow$ hipótese unilateral direita
$\sigma^2 < \sigma_0^2 \leftarrow$ hipótese unilateral esquerda

A estatística do teste é Q que tem distribuição qui-quadrado com parâmetro ν = n$-$1 e é assim definida:

$$Q = \frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2(\nu),$$

onde:

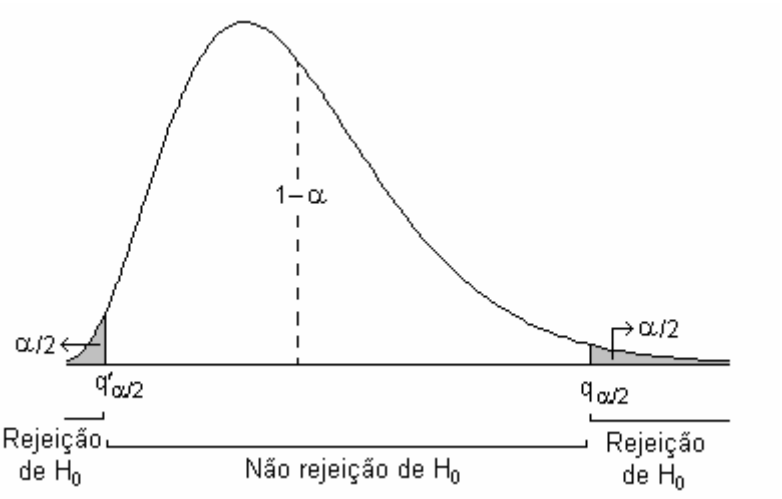
S^2 é o estimador da variância populacional σ^2;
n é o tamanho da amostra;
ν = n$-$1 é o número de graus de liberdade associado à variância.

A região de rejeição de H_0 é definida em função do tipo de hipótese alternativa. Logo, fixado um nível de significância α, a hipótese nula é rejeitada se o valor da estatística do teste ultrapassar o valor crítico (inferior ou superior) da distribuição qui-quadrado (Tabela III do Apêndice):

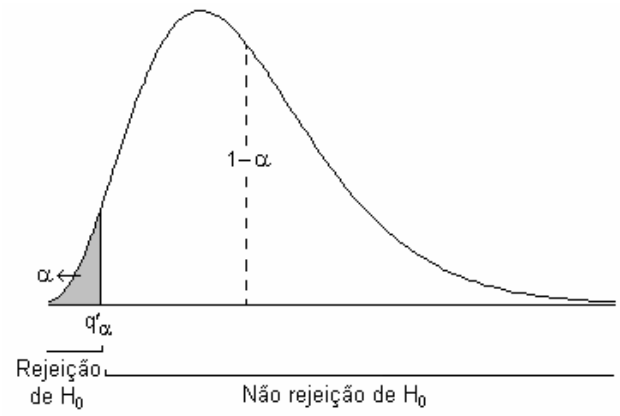
– se $q > q_{\alpha/2(\nu)}$ ou $q > q'_{\alpha/2(\nu)}$, rejeitamos H_0;

– se $q < q_{\alpha/2(\nu)}$ e $q < q'_{\alpha/2(\nu)}$, não rejeitamos H_0.

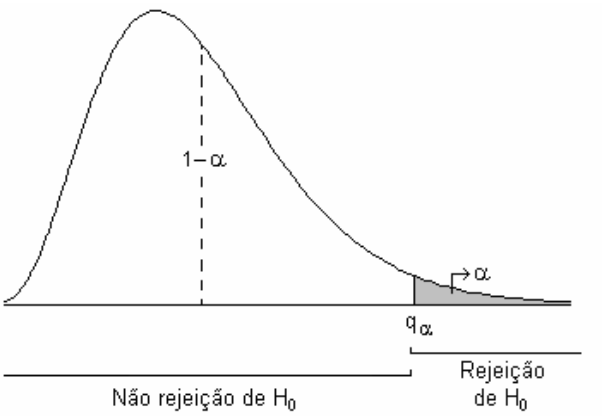
Para $H_A : \sigma^2 \neq \sigma_0^2$ (bilateral)



Para $H_A : \sigma^2 < \sigma_0^2$ (unilateral esquerda)



Para $H_A : \sigma^2 > \sigma_0^2$ (unilateral direita)



Esse teste tem larga aplicação no controle da qualidade, uma vez que o monitoramento da variabilidade é essencial para a garantia de qualidade.

Consideremos o exemplo resolvido.

Uma máquina de empacotar café está regulada para encher os pacotes com desvio padrão de 10g e média de 500g e onde o peso de cada pacote distribui-se normalmente. Colhida uma amostra de n = 16, observou-se uma variância de 169g^2. É possível afirmar com este resultado que a máquina está desregulada quanto à variabilidade, supondo uma significância de 5%?

Resolução:

Hipóteses estatísticas: $\begin{cases} H_0 : \sigma^2 = \sigma_0^2 \\ H_A : \sigma^2 \neq \sigma_0^2 \end{cases}$

Sendo $\sigma_0^2 = 100$, $s^2 = 169$ e n = 16, temos:

$$q = \frac{(n-1)s^2}{\sigma_0^2} = \frac{(16-1)\times 169}{100} = 25{,}35 .$$

Como α = 0,05 e ν=15, os valores críticos são $q'_{0{,}025(15)} = 6{,}26$ e $q_{0{,}025(15)} = 27{,}49$. O valor calculado está contido neste intervalo, portanto, não rejeitamos H_0. Concluímos, ao nível de 5% de significância, que não há evidência de que e a máquina esteja desregulada.

4.5.2.2. Teste de homogeneidade de variâncias (teste F)

Considere duas estimativas distintas e independentes de σ^2, representadas por s_1^2 e s_2^2, com ν_1 e ν_2 graus de liberdade, respectivamente. Frequentemente, temos interesse em verificar se tais estimativas são homogêneas, ou seja, se são de fato estimativas de um mesmo parâmetro. Um teste apropriado para essa finalidade é o teste F. Neste teste, verificamos a hipótese de nulidade (H_0) por meio de uma estatística que tem distribuição F, com parâmetros ν_1 e ν_2.

- **Hipóteses estatísticas**

A hipótese que está sob verificação (H_0) é a hipótese de igualdade entre as variâncias populacionais. O conjunto das hipóteses, incluindo todas as possíveis hipóteses alternativas, é:

$$\begin{aligned} H_0 &: \sigma_1^2 = \sigma_2^2 \text{ ou } \sigma_1^2/\sigma_2^2 = 1 \\ H_A &: \sigma_1^2 \neq \sigma_2^2 \text{ ou } \sigma_1^2/\sigma_2^2 \neq 1 \leftarrow \text{hipótese bilateral} \\ &\sigma_1^2 > \sigma_2^2 \text{ ou } \sigma_1^2/\sigma_2^2 > 1 \leftarrow \text{hipótese unilateral direita} \\ &\sigma_1^2 < \sigma_2^2 \text{ ou } \sigma_1^2/\sigma_2^2 < 1 \leftarrow \text{hipótese unilateral esquerda} \end{aligned}$$

- **Estatística do teste**

Como vimos na seção 4.4.3, dadas duas amostras independentes, de tamanhos n_1 e n_2, retiradas de duas populações normais, a variável aleatória F tem distribuição F, com parâmetros ν_1 e ν_2

$$F = \frac{\frac{S_1^2}{\sigma_1^2}}{\frac{S_2^2}{\sigma_2^2}} \sim F(\nu_1, \nu_2) ,$$

onde:

S_1^2 e S_2^2: são as variâncias das amostras retiradas das populações 1 e 2, respectivamente;

σ_1^2 e σ_2^2: são as variâncias das populações 1 e 2, respectivamente;

$\nu_1 = (n_1 - 1)$ e $\nu_2 = (n_2 - 1)$: são os graus de liberdade de S_1^2 e S_2^2, respectivamente.

Se a hipótese de nulidade é verdadeira, ou seja, $\sigma_1^2 = \sigma_2^2 = \sigma^2$, a variável F torna-se

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

Deste modo, dizemos que, sob H_0 verdadeira, a estatística $F = \frac{S_1^2}{S_2^2} \sim F(\nu_1, \nu_2)$, ou seja, a estatística F, obtida através da razão de duas variâncias amostrais independentes, tem distribuição F com parâmetros ν_1 e ν_2, graus de liberdade do numerador e denominador, respectivamente.

Assim, para que a estatística F possa ser utilizada na comparação das variâncias populacionais, duas pressuposições devem ser atendidas:

1. A variável em estudo tem distribuição normal, X ~ N (μ, σ^2).
2. As amostras são independentes.

◆ **Critério de decisão**

Para efetuar o teste F, convencionamos que a estatística F é obtida fixando a maior variância no numerador. Isto garante que o valor da estatística nunca será menor que 1 e possibilita a utilização das tabelas mais comumente disponíveis para consulta. Por convenção, vamos considerar que $S_1^2 \geq S_2^2$.

A regra de decisão a respeito de H_0 pode ser estabelecida com base no valor crítico $f_{\alpha/2(\nu_1,\nu_2)}$, que, para os graus de liberdade ν_1 e ν_2, delimita a área $\alpha/2$ (Tabela IV do Apêndice).

– Rejeitamos H_0, ao nível α, se o valor da estatística F for maior que o valor crítico:

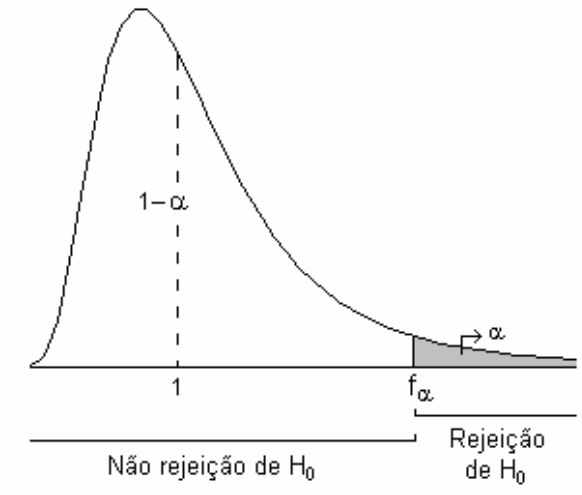
$$f = \frac{s_1^2}{s_2^2} > f_{\alpha/2(\nu_1,\nu_2)}.$$

– Não rejeitamos H_0, ao nível α, se o valor da estatística F for menor que o valor crítico:

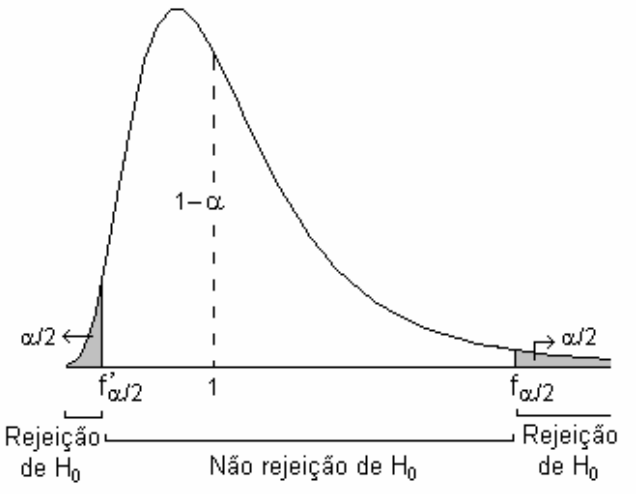
$$f = \frac{s_1^2}{s_2^2} < f_{\alpha/2(\nu_1,\nu_2)}.$$

Podemos observar a seguir as regiões de rejeição de H_0 na curva da distribuição F, para os casos de hipótese alternativa unilateral e de hipótese alternativa bilateral.

Para $H_A: \sigma_1^2 > \sigma_2^2$ (unilateral direita)



Para $H_A: \sigma_1^2 \neq \sigma_2^2$ (bilateral)



Neste caso, quando a hipótese alternativa é bilateral, dizemos que o teste é bilateral condicionado, porque somente um lado da curva será considerado. Como a tabela usualmente utilizada é adequada para testes F unilaterais, a área de rejeição à direita do valor crítico será representada por $\alpha/2$.

Consideremos dois exemplos resolvidos:

Exemplo 1. Os valores abaixo se referem aos pesos ao nascer (em kg) de bovinos da raça Ibagé, em duas épocas distintas:

Agosto	18	25	16	30	35	23	21	33	32	22	-	-	-
Setembro	27	30	20	30	33	34	17	33	20	23	39	23	28

Efetue o teste de homogeneidade de variâncias, ao nível α = 0,05.

Resolução:

Variável em estudo: X = peso ao nascer (em kg) de bovinos da raça Ibagé

1. Pressuposições
– A variável em estudo tem distribuição normal, X ~ N (μ, σ^2).
– As amostras retiradas das populações são independentes.

2. Hipóteses estatísticas: $\begin{cases} H_0 : \sigma_1^2 = \sigma_2^2 \\ H_A : \sigma_1^2 \neq \sigma_2^2 \end{cases}$

3. Estatística do teste

$$F = \frac{S_1^2}{S_2^2}$$

Amostra 1 (agosto): $\nu_1 = n_1 - 1 = 9$ $\bar{x}_1 = 25,5$ $s_1^2 = 43,83$

Amostra 2 (setembro): $\nu_2 = n_2 - 1 = 12$ $\bar{x}_2 = 27,46$ $s_2^2 = 42,60$

$$f = \frac{s_1^2}{s_2^2} = \frac{43,83}{42,60} = 1,029$$

4. Decisão e conclusão

Como f = 1,029 < $f_{0,025(9,\ 12)}$ = 3,44, não temos informações suficientes para rejeitar H_0. Assim, concluímos, ao nível de 5% de significância, que as variâncias de pesos ao nascer de bovinos da raça Ibagé nas diferentes épocas são homogêneas.

Exemplo 2. Um experimento foi conduzido para comparar duas cultivares de soja (A e B) quanto ao rendimento médio por hectare. Os resultados obtidos foram os seguintes:

Cultivar A: $n_1 = 8$ $\bar{x}_1 = 3,8\,t.ha^{-1}$ $s_1^2 = 0,04\,(t.ha^{-1})^2$

Cultivar B: $n_2 = 10$ $\bar{x}_2 = 4,6\,t.ha^{-1}$ $s_2^2 = 0,36\,(t.ha^{-1})^2$

Verifique, utilizando α=0,05, se a pressuposição de homogeneidade de variâncias foi atendida.

Resolução:

Variável em estudo: X = rendimento de soja (por hectare)

1. Pressuposições
– A variável em estudo tem distribuição normal, $X \sim N(\mu, \sigma^2)$.
– As amostras retiradas das populações são independentes.

2. Hipóteses estatísticas: $\begin{cases} H_0 : \sigma_1^2 = \sigma_2^2 \\ H_A : \sigma_1^2 \neq \sigma_2^2 \end{cases}$

3. Estatística do teste

$$F = \frac{S_1^2}{S_2^2}$$

Amostra 1 (cultivar A): $\nu_1 = n_1 - 1 = 7$ $\quad \bar{x}_1 = 3{,}8\,t.ha^{-1}$ $\quad s_1^2 = 0{,}04\,(t.ha^{-1})^2$

Amostra 2 (cultivar B): $\nu_2 = n_2 - 1 = 9$ $\quad \bar{x}_2 = 4{,}6\,t.ha^{-1}$ $\quad s_2^2 = 0{,}36\,(t.ha^{-1})^2$

Por convenção colocamos a maior estimativa no numerador, resultando assim

$$f = \frac{s_2^2}{s_1^2} = \frac{0{,}36}{0{,}04} = 9\,.$$

4. Decisão e conclusão

Como $f = 9 > f_{0,025(9,\ 7)} = 4{,}82$, temos evidências suficientes para a rejeição de H_0. Assim, podemos concluir, ao nível de 5% de significância, que as variâncias de rendimento de grãos (em t/ha) das cultivares A e B não são homogêneas.

4.5.3. Testes para a proporção populacional

4.5.3.1. Teste para a proporção de uma população

O teste para a proporção populacional, em geral, é utilizado para verificar se a proporção π de elementos da população que possuem uma determinada característica é igual a um determinado valor π_0. Assim as hipóteses estatísticas são:

$H_0 : \pi = \pi_0$ ou $\pi - \pi_0 = 0$

$H_A : \pi \neq \pi_0$ ou $\pi - \pi_0 \neq 0 \leftarrow$ hipótese bilateral

$\pi > \pi_0$ ou $\pi - \pi_0 > 0 \leftarrow$ hipótese unilateral direita

$\pi < \pi_0$ ou $\pi - \pi_0 < 0 \leftarrow$ hipótese unilateral esquerda

O estimador da proporção π é a proporção amostral P, que tem uma distribuição aproximadamente normal, com média $\mu_P = \pi$ e desvio padrão $\sigma_P = \sqrt{\frac{\pi(1-\pi)}{n}}$, quando a amostra é grande, ou seja, quando np>5 e n(1-p)>5.

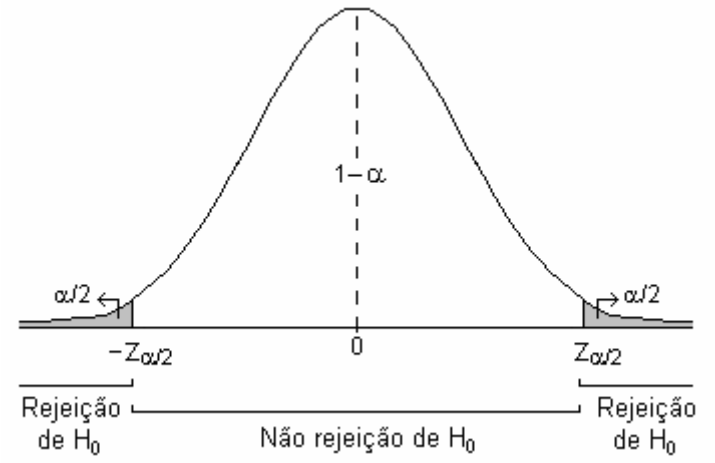
Assim, utilizamos a variável Z para testar H_0:

$$Z = \frac{P - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}\,.$$

Sob $H_0 : \pi = \pi_0$ verdadeira, temos

$$Z = \frac{P - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$$

A decisão sobre H_0 é baseada nos valores críticos $z_{\alpha/2}$, para o teste bilateral, ou z_α, para o teste unilateral, encontrados na tabela da distribuição Z (Tabela I do Apêndice).



Assim, fixando o nível de significância α, a hipótese nula será rejeitada se:

$|z| > z_{\alpha/2}$, no teste bilateral;
$z > z_\alpha$, no teste unilateral à direita;
$z < z_\alpha$, no teste unilateral à esquerda.

Vamos considerar um exemplo resolvido:

As condições de mortalidade de uma região são tais que a proporção de nascidos que sobrevivem até 70 anos é de 0,60. Testar esta hipótese ao nível de 5% de significância se em 1000 nascimentos amostrados aleatoriamente, verificou-se 530 sobreviventes até os 70 anos.

Resolução:

Hipóteses estatísticas: $\begin{cases} H_0 : \pi = \pi_0 \\ H_A : \pi \neq \pi_0 \end{cases}$

Sendo $\pi_0 = 0,60$ e $p = 0,53$, temos:

$$z = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} = \frac{0,53 - 0,60}{\sqrt{\frac{0,60(1-0,60)}{1000}}} = -4,52 .$$

Considerando que o teste é bilateral e tendo $\alpha = 0,05$, os valores críticos são $-z_{\alpha/2} = -1,96$ e $z_{\alpha/2} = 1,96$.

Como este valor pertence à região de rejeição, rejeitamos a hipótese nula, ao nível de 5% de significância. Concluímos que a taxa dos que sobrevivem até os 70 anos é menor do que 60%.

4.5.3.2. Teste para a diferença entre duas proporções

A aproximação da distribuição normal também pode ser usada para testar hipóteses sobre diferenças entre proporções de duas populações, ou seja, para testar as hipóteses:

$$H_0 : \pi_1 = \pi_2 \text{ ou } \pi_1 - \pi_2 = 0$$
$$H_A : \pi_1 \neq \pi_2 \text{ ou } \pi_1 - \pi_2 \neq 0 \leftarrow \text{hipótese bilateral}$$
$$\pi_1 > \pi_2 \text{ ou } \pi_1 - \pi_2 > 0 \leftarrow \text{hipótese unilateral direita}$$
$$\pi_1 < \pi_2 \text{ ou } \pi_1 - \pi_2 < 0 \leftarrow \text{hipótese unilateral esquerda}$$

Nesse caso, duas amostras aleatórias de tamanho n_1 e n_2 são retiradas das populações, gerando x_1 e x_2 itens pertencentes às classes da característica de interesse. Então $P_1 = X_1/n_1$ e $P_2 = X_2/n_2$ são os estimadores das proporções populacionais π_1 e π_2.

A variável $P_1 - P_2$ terá uma distribuição aproximadamente normal com média $\pi_1 - \pi_2$ e variância $\sigma^2_{P_1-P_2} = \dfrac{\pi_1(1-\pi_1)}{n_1} + \dfrac{\pi_2(1-\pi_2)}{n_2}$.

Assim, a variável Z resulta:

$$Z = \frac{(P_1 - P_2) - (\pi_1 - \pi_2)}{\sqrt{\dfrac{\pi_1(1-\pi_1)}{n_1} + \dfrac{\pi_2(1-\pi_2)}{n_2}}}.$$

Sob $H_0 : \pi_1 = \pi_2$ verdadeira, temos $\pi_1 - \pi_2 = 0$ e como consequência:

$$Z = \frac{(P_1 - P_2)}{\sqrt{\dfrac{\pi_1(1-\pi_1)}{n_1} + \dfrac{\pi_2(1-\pi_2)}{n_2}}}.$$

Como os valores de π_1 e π_2 não são conhecidos, devemos utilizar suas estimativas p_1 e p_2. Desta forma, o valor de Z é:

$$z = \frac{p_1 - p_2}{\sqrt{\dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}}}.$$

A decisão sobre H_0 é baseada nos valores críticos $z_{\alpha/2}$, para o teste bilateral, ou z_α, para o teste unilateral, encontrados na tabela da distribuição normal padrão.

Vejamos o exemplo resolvido:

Em uma pesquisa de opinião, 32 entre 80 homens declararam apreciar certa revista, acontecendo o mesmo com 26 entre 50 mulheres. Ao nível de 5% de significância os homens e as mulheres apreciam igualmente a revista?
Resolução:

Hipóteses estatísticas: $\begin{cases} H_0 : \pi_1 = \pi_2 \\ H_A : \pi_1 \neq \pi_2 \end{cases}$

Sendo $p_1 = 32/80 = 0,40$ e $p_2 = 26/50 = 0,52$, temos:

$$z = \frac{p_1 - p_2}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} = \frac{0,40 - 0,52}{\sqrt{\frac{0,40 \times 0,60}{80} + \frac{0,52 \times 0,48}{50}}} = -1,34$$

Como $\alpha = 0,05$, então, $z_{\alpha/2} = 1,96$.

Sendo o |z| calculado menor que o valor crítico, não rejeitamos a hipótese de igualdade entre as preferências de homens e mulheres. Concluímos, ao nível de 5% de significância, que não há diferença significativa entre as preferências de homens e mulheres quanto à revista.

4.6. Quebras nas pressuposições adotadas no processo de inferência

O teste t somente pode ser aplicado com resultados exatos se as pressuposições de normalidade, homogeneidade de variâncias e independência entre amostras forem atendidas. Das três, a pressuposição de normalidade é a mais robusta, ou seja, mesmo com normalidade aproximada o teste poderá ser realizado. Ademais, o teorema central do limite é um poderoso auxiliar nestes casos, uma vez que a distribuição necessita ser considerada em termos da média da variável. As outras duas deverão ser consideradas com maior cuidado. No caso de variâncias heterogêneas, apenas procedimentos aproximados poderão ser utilizados. Veremos a seguir dois procedimentos alternativos para os casos de violação dessas pressuposições.

4.6.1. Heterogeneidade de variâncias

Se as variâncias das populações não são iguais ($\sigma_1^2 \neq \sigma_2^2$), não podemos combinar as variâncias amostrais, S_1^2 e S_2^2, pois estas não são mais estimativas de um mesmo parâmetro (σ^2). Nesse caso, o erro padrão do estimador resulta em:

$$S(\hat{\theta}) = S\left(\bar{X}_1 - \bar{X}_2\right) = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

Diversas modificações do teste são propostas na literatura, entretanto, nenhuma é considerada completamente satisfatória. Aqui será mencionada uma delas, bastante utilizada, proposta por Satterthwaite, que estima o número de graus de liberdade associado à estatística, uma vez que não existe uma variância ponderada com número único de graus de liberdade, como no caso de variâncias homogêneas. Para a obtenção do valor crítico, utilizamos, então, o número de graus de liberdade estimado (ν'), através da Fórmula de Satterthwaite:

$$\nu' = \frac{\left(S_1^2 + S_2^2\right)^2}{\frac{\left(S_1^2\right)^2}{n_1 - 1} + \frac{\left(S_2^2\right)^2}{n_2 - 1}}$$

Nesse caso, rejeitamos H_0, ao nível α, se $| t | > t(\nu')$.

Vamos considerar o exemplo resolvido:

Com referência ao Exemplo 2 da seção 4.7.2.2, verifique se os rendimentos médios das cultivares A e B diferem significativamente entre si, utilizando $\alpha = 0,05$.

Resolução:

1. Pressuposições:

- A variável em estudo tem distribuição aproximadamente normal, X ~ N (μ, σ^2).
- As amostras retiradas das populações são independentes.

2. Hipóteses estatísticas: $\begin{cases} H_0 : \mu_1 - \mu_2 = 0 \\ H_A : \mu_1 - \mu_2 \neq 0 \end{cases}$

A hipótese de nulidade supõe a igualdade entre as médias das duas populações.

3. Estatística do teste

$$\theta = \mu_1 - \mu_2 = 0$$
$$\hat{\theta} = \bar{X}_1 - \bar{X}_2 = 3{,}8 - 4{,}6 = -0{,}8$$

Como verificamos na resolução do Exemplo 2, a pressuposição de homogeneidade de variâncias não foi atendida, portanto, não podemos combinar as variâncias das amostras. Temos, então

$$S(\hat{\theta}) = S(\bar{X}_1 - \bar{X}_2) = \sqrt{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)} = \sqrt{\left(\frac{0{,}04}{8} + \frac{0{,}36}{10}\right)} = \sqrt{0{,}041} = 0{,}2025$$

$$T = \frac{\hat{\theta}}{S(\hat{\theta})} = \frac{-0{,}8}{0{,}2025} = -3{,}951$$

4. Decisão e conclusão

Devido à heterogeneidade das variâncias, o número de graus de liberdade deve-se ser estimado pela fórmula de Satterthwaite:

$$\nu' = \frac{\left(s_1^2 + s_2^2\right)^2}{\frac{\left(s_1^2\right)^2}{n_1 - 1} + \frac{\left(s_2^2\right)^2}{n_2 - 1}} = \frac{(0{,}04 + 0{,}36)^2}{\frac{(0{,}04)^2}{8-1} + \frac{(0{,}36)^2}{10-1}} = \frac{0{,}16}{0{,}0002286 + 0{,}0144} = 10{,}94$$

Para consulta à tabela, o número de graus de liberdade pode ser arredondado. No exemplo, o valor crítico será $t_{0{,}025(11)} = 2{,}201$.

Como $|t = -3{,}951| > t_{\alpha/2(\nu')} = 2{,}201$, temos motivos para rejeitar H_0. Concluímos, então, ao nível de 5% de significância, que o rendimento médio de grãos da cultivar A diferiu significativamente do rendimento médio de grãos da cultivar B. Portanto, há evidências de que a cultivar B é mais produtiva.

4.6.2. Dependência entre as amostras

Quando comparamos as médias de duas populações, pode ocorrer uma diferença significativa devido a fatores externos não-controláveis que inflacionam as estimativas das variâncias. Um modo de contornar este problema é coletar observações aos pares, de modo que os dois elementos de cada par sejam homogêneos em todos os sentidos, exceto naquele que se quer comparar.

Por exemplo, para testar dois métodos de ensino A e B, pode-se usar pares de gêmeos, sendo que um recebe o método de ensino A e o outro o método de ensino B. Este procedimento controla a maioria dos fatores externos que afetam a aprendizagem. Se houver diferença, ela realmente se deve ao método.

Outra forma é fazer as observações das duas amostras no mesmo indivíduo. Por exemplo, medindo uma característica do indivíduo antes e depois dele ser submetido a um tratamento.

Nesses casos, temos duas amostras, mas as observações estão emparelhadas, isto é, a amostra é formada pelos pares:

$$(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n).$$

Ao usarmos esse procedimento, passamos a ter uma dependência entre as observações de um mesmo par. Por essa razão, não podemos preceder o teste de comparações de médias da mesma forma como é realizado no caso de independência entre as amostras.

Assim, para verificar se existe diferença entre μ_X e μ_Y, definimos a variável $D= X_i - Y_i$. Como resultado, temos a amostra: $D_1 = X_1 - Y_1$, $D_2 = X_2 - Y_2$, ..., $D_n = X_n - Y_n$. Desta forma, reduz-se o problema para análise de uma única população.

Supondo que D tem distribuição $N(\mu_D, \sigma_D)$, temos $\bar{D} = \frac{1}{n}\sum D_i = \frac{1}{n}\sum (X_i - Y_i) = \bar{X} - \bar{Y}$, e assim $\bar{D}$ terá distribuição $N(\mu_D, \frac{\sigma_D}{\sqrt{n}})$. Definindo: $S_D^2 = \frac{1}{n-1}\sum (D_i - \bar{D})^2 = \frac{\sum D_i^2 - n\bar{D}^2}{n-1}$, tem-se que a estatística:

$$T = \frac{\bar{D} - \mu_D}{\frac{S_D}{\sqrt{n}}} \sim t(\nu), \text{ onde } \nu = n-1.$$

As hipóteses de interesse são:

$H_0 : \mu_X = \mu_Y$ ou $\mu_D = 0$

$H_A : \mu_X \neq \mu_Y$ ou $\mu_D \neq 0 \leftarrow$ hipótese bilateral

$\mu_X > \mu_Y$ ou $\mu_D > 0 \leftarrow$ hipótese unilateral direita

$\mu_X < \mu_Y$ ou $\mu_D < 0 \leftarrow$ hipótese unilateral esquerda

Sob $H_0 : \mu_X = \mu_Y$ verdadeira, temos $\mu_D = 0$ e como consequência:

$$T = \frac{\bar{D} - \mu_D}{\frac{S_D}{\sqrt{n}}} = \frac{\bar{D} - 0}{\frac{S_D}{\sqrt{n}}} = \frac{\bar{D}}{\frac{S_D}{\sqrt{n}}} \sim t(\nu).$$

A decisão sobre H_0 é tomada comparando o valor calculado para t com o valor crítico da distribuição t (Tabela II do Apêndice).

Vejamos um exemplo resolvido:

Cinco operadores de máquinas são treinados em duas máquinas de diferentes fabricantes, para verificar qual delas apresentava maior facilidade de aprendizagem. Mediu-se o tempo que cada um dos operadores gastou na realização de uma mesma tarefa com cada um dos dois tipos de máquinas. Os resultados estão na tabela a seguir.

Operador	Máquina X	Máquina Y	Diferença
1	80	75	5
2	72	70	2
3	65	60	5
4	78	72	6
5	85	78	7

Ao nível de 5% é possível afirmar que há diferença no tempo médio da tarefa realizada na máquina X e na máquina Y?

Resolução:

Hipóteses estatísticas: $\begin{cases} H_0 : \mu_X = \mu_Y \\ H_A : \mu_X \neq \mu_Y \end{cases}$

Pela tabela vemos que:

d_i = 5, 2, 5, 6 e 7,

logo,

$$\bar{d} = 5 \quad \text{e} \quad s_D^2 = \frac{\sum \left(d_i - \bar{d}\right)^2}{n-1} = 28{,}5 .$$

Assim, obtemos

$$t = \frac{\bar{d}}{\frac{s_D}{\sqrt{n}}} = \frac{5}{\frac{5{,}339}{\sqrt{5}}} = 2{,}094 .$$

Sendo α = 0,05 e ν = 4, temos $t_{0{,}025(4)}$ = 2,778.

Como o t calculado é menor que o valor crítico, não rejeitamos a hipótese nula. Concluímos, a 5% de significância, que não há diferença no tempo médio da tarefa realizada na máquina X e na máquina Y.

Exercícios propostos:

4.6. Cinco medidas do conteúdo de alcatrão em um cigarro X acusaram: 14,5, 14,2, 14,4, 14,8, e 14,1 miligramas por cigarro. Este conjunto de cinco valores tem média 14,4 e desvio padrão 0,274. O leitor pretende testar a hipótese nula H_0: μ = 14,1 (conforme declarado no maço) ao nível de 0,05 de significância.

a) H_0 seria aceita, contra a alternativa H_A: $\mu \neq 14{,}1$?
b) H_0 seria aceita, contra a alternativa H_A: $\mu < 14{,}1$?
c) H_0 seria aceita, contra a alternativa H_A: $\mu > 14{,}1$?
d) Que suposições são necessárias para fazer o teste de hipóteses?

4.7. Suponha que um fabricante sem escrúpulos deseje uma “prova científica” de que um aditivo químico totalmente inócuo melhora o rendimento.

a) Se um grupo de pesquisa analisa esse aditivo com um experimento, qual é a probabilidade de chegar a um “resultado significativo” com α = 0,05 (para promover o aditivo com “afirmações científicas”) mesmo que o aditivo seja totalmente inócuo?

b) Se dois grupos independentes de pesquisa analisam o aditivo, qual é a probabilidade de que pelo menos um deles chegue a um "resultado significativo", mesmo que o aditivo seja totalmente inócuo?

c) Se 32 grupos independentes de pesquisa analisam o aditivo, qual é a probabilidade de que pelo menos um deles chegue a um "resultado significativo", mesmo que o aditivo seja totalmente inócuo?

4.8. Suponha que um farmacêutico pretenda achar um novo unguento para reduzir inchação. Para tanto, ele fabrica 20 medicamentos diferentes e testa cada um deles, ao nível de 0,10 de significância, quanto a finalidade em vista.

a) Qual a probabilidade de ao menos um deles "se revelar" eficaz mesmo que todos sejam totalmente inócuos?

b) Qual a probabilidade de mais de um deles "se revelarem" eficazes, mesmo que todos sejam totalmente inócuos?

4.9. O fabricante de uma certa marca de aparelhos eletrônicos informou que a potência média dos seus aparelhos é de 27 microwatts. O gerente de uma loja que vende os aparelhos utiliza uma amostra de 15 aparelhos para checar se a informação do fabricante é verdadeira. Os valores (em microwatts) obtidos para a amostra foram os seguintes:

26,7; 25,8; 24,0; 24,9; 26,4; 25,9; 24,4; 21,7; 24,1; 25,9; 27,3; 26,9; 27,3; 24,8; 23,6

Utilize um teste de hipótese, ao nível de 5% de significância, e verifique qual foi a conclusão do gerente.

4.10. Dez cobaias adultas criadas em laboratório, foram separadas, aleatoriamente, em dois grupos: um foi tratado com ração normalmente usada no laboratório (padrão) e o outro grupo foi submetido a uma nova ração (experimental). As cobaias foram pesadas no início e no final do período de duração do experimento. Os ganhos de peso (em gramas) observados foram os seguintes:

Ração padrão	200	180	190	190	180
Ração experimental	220	200	210	220	210

Utilize um teste de hipótese, ao nível α=0,01, para verificar se as duas rações diferem entre si.

4.11. Os valores abaixo se referem aos pesos ao nascer (em kg) de bovinos da raça Ibagé, em duas épocas distintas:

Agosto: 18 25 16 30 35 23 21 33 32 22
Setembro: 27 30 20 30 33 34 17 33 20 23 39 23 28

Efetue o teste de homogeneidade de variâncias, ao nível α = 0,05.

4.12. Um engenheiro deseja testar a hipótese de que o percentual de peças defeituosas é inferior a 10%. Uma amostra aleatória com 75 peças revelou 6 peças defeituosas. Use α = 0,05 e conclua a respeito.

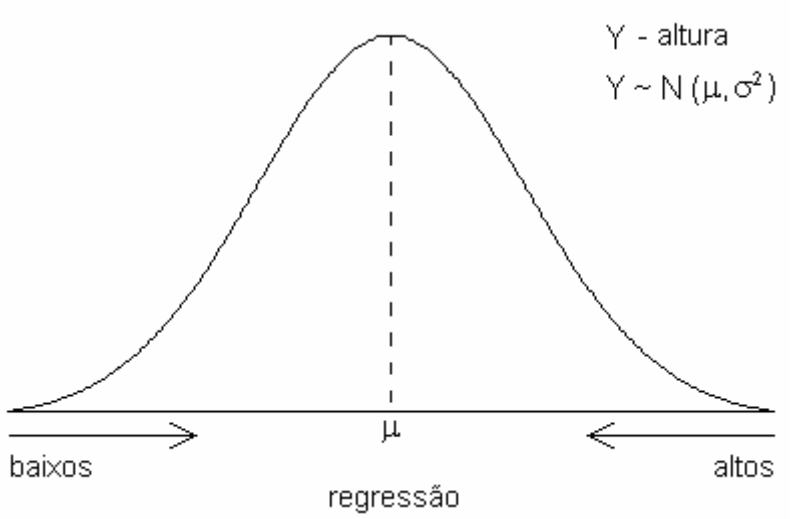
4.7. Regressão linear simples

4.7.1. Introdução

Em muitos estudos estatísticos, o objetivo do pesquisador é estabelecer relações que possibilitem predizer uma ou mais variáveis em termos de outras. Assim é que se fazem estudos para predizer as vendas futuras de um produto em função do seu preço, a perda de peso de uma pessoa em decorrência do número de dias que se submete a uma determinada dieta, a despesa de uma família com médico e remédios em função da renda, o consumo per capita de certos alimentos em função do seu valor nutritivo e do gasto com propaganda na TV, a produção de uma determinada cultura em função da quantidade de nutriente aplicada no solo, etc.

Naturalmente, o ideal seria que pudéssemos predizer uma quantidade exatamente em termos de outra, mas isso raramente é possível. Na maioria dos casos devemos contentar-nos com a predição de médias, ou valores esperados. Por exemplo, não podemos predizer exatamente quanto ganhará um bacharel nos 10 anos após a formatura, mas com base em dados adequados, é possível predizer o ganho médio de todos os bacharéis nos 10 anos após a formatura. Analogamente, podemos predizer a safra média de certa variedade de trigo em termos do índice pluviométrico de julho, e a nota média de um estudante em função do seu QI. Sendo assim, podemos dizer que a predição do valor médio de uma variável em função dos valores de outra constitui o problema principal da regressão.

A origem desse termo remonta a *Francis Galton* (1822-1911), que o empregou pela primeira vez em um estudo da relação entre as alturas de pais e filhos. Galton observou, nesse estudo, que filhos de pais muito altos, em média, não eram tão altos quanto os seus pais, da mesma forma que filhos de pais muito baixos, em média, não eram tão baixos quanto os seus pais. A partir dessas observações, concluiu que a altura dos filhos "tendia" para a média (μ) da espécie, ou seja, a cada geração a altura dos filhos convergia ou "regredia" para a média. Esse fenômeno de retorno à média foi, então, denominado *regressão*.



Por questões históricas o termo é utilizado até hoje, mas abriga uma série de técnicas estatísticas.

A expressão *regressão linear simples* é utilizada por duas razões: a regressão é *linear* porque a relação entre X e Y é expressa por uma equação de primeiro grau, representada graficamente por uma reta, e é *simples* porque envolve apenas duas variáveis.

- **Ajustamento de curvas**

Sempre que possível, procuramos expressar em termos de uma equação matemática as relações entre grandezas conhecidas e grandezas que devem ser determinadas. Isso ocorre

com frequência nas ciências naturais, onde, por exemplo, a relação entre o volume (y) e pressão (x) de um gás, a uma temperatura constante, é dada pela expressão

$$y = \frac{k}{x},$$

sendo k uma constante numérica. Outro exemplo pode ser a relação entre uma cultura de bactérias (y) e o tempo (x) em que esteve exposta a certas condições ambientais, que é dada por

$$y = ab^x,$$

onde a e b são constantes numéricas. Mais recentemente, equações como essas têm sido usadas para descrever relações também no campo das ciências do comportamento, das ciências sociais e outros.

Essa representação matemática dos fenômenos é feita "ajustando-se" uma curva aos dados observados, de tal forma que, a partir dessa "curva ajustada", possamos representar, gráfica ou analiticamente, a relação entre as variáveis. Então, ajustar uma curva é determinar uma função matemática que possa representar um conjunto de observações. Sempre que utilizamos dados observados para chegar a uma equação matemática encontramos três tipos de problema:

1. Devemos decidir que tipo de curva e, daí, que tipo de equação de predição queremos utilizar.

2. Devemos achar a equação particular que é a melhor em determinado sentido.

3. Devemos investigar possíveis problemas relativos ao mérito da equação escolhida e da predição feita a partir dela.

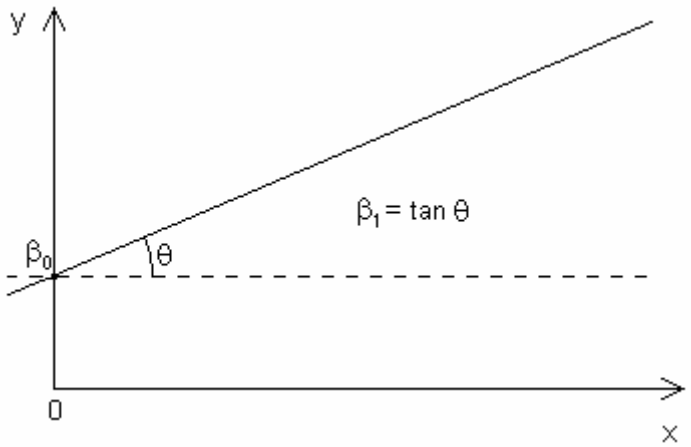
Aqui vamos restringir nosso estudo às equações lineares com duas incógnitas. Estas equações lineares são úteis e importantes não só porque muitas relações têm efetivamente esta forma, mas também porque em geral constituem boas aproximações de relações que, de outro modo, seriam difíceis de descrever em termos matemáticos.

- **Modelo estatístico**

Sendo x e y duas variáveis que se relacionam de forma linear, esta relação é expressa pela seguinte equação:

$$y = \beta_0 + \beta_1 x,$$

Na figura a seguir podemos observar a representação gráfica desta equação.



Se Y é uma variável aleatória, então, está sujeita a um erro de observação. Este erro (e_i) deverá ser adicionado ao modelo, desde que se admitam como verdadeiras as seguintes pressuposições:

1. Os erros são aleatórios, têm média zero e variância constante, ou seja, $E(e_i) = 0$ e $V(e_i) = \sigma^2$.

2. Os erros têm distribuição normal e são independentes entre si.

3. O modelo é adequado para todas as observações, não podendo haver nenhum valor de X que produza um valor de Y discrepante dos demais.

4. A variável X é fixa (não aleatória).

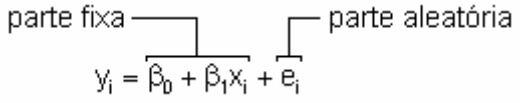
Assim, o modelo de regressão linear simples será:

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

onde:

y_i : é a variável resposta (dependente)
x_i : é a variável preditora (independente)
β_0 : é o intercepto ou coeficiente linear
β_1 : é o coeficiente angular ou de regressão
e_i : erro (variação aleatória não controlável)

Sendo assim, verificamos que este modelo é composto por uma parte fixa e uma parte aleatória:



A parte fixa do modelo informa como X influencia Y e a parte aleatória mostra que Y possui uma variabilidade inerente, significando que X não é a única variável que influencia Y, embora consideremos que sua influência seja preponderante. Aliás, devemos ressaltar que este modelo será adequado quando a parte fixa for preponderante sobre a aleatória.

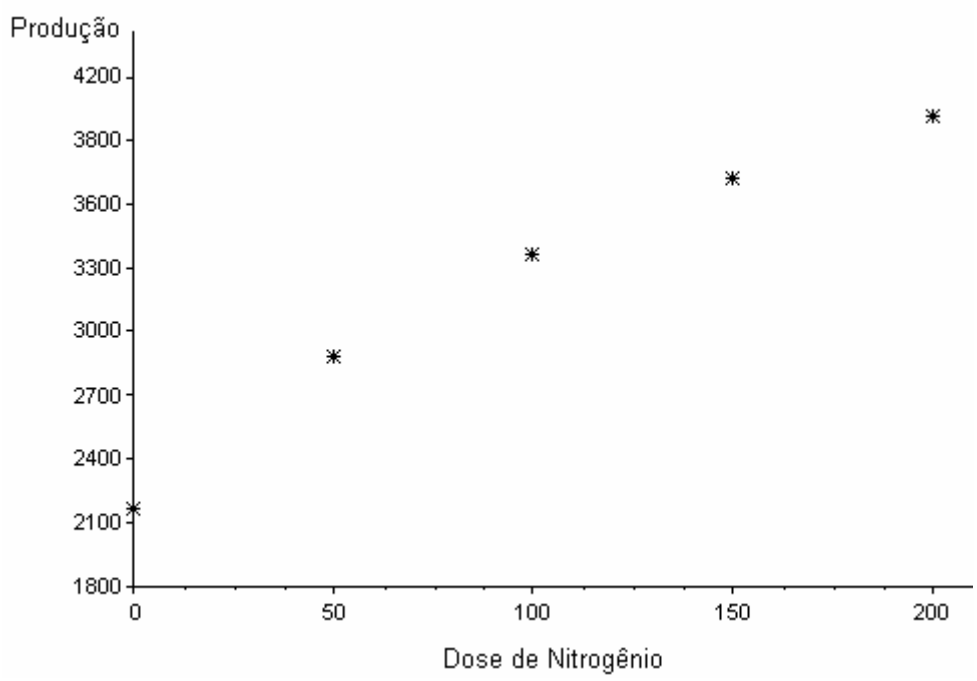
A título de ilustração, consideremos o exemplo a seguir.

Exemplo 1. Um experimento foi conduzido para estudar o efeito da dose de Nitrogênio aplicada no solo sobre a produção de uma espécie de forrageira. Para as cinco doses utilizadas, foram observados os seguintes resultados:

Parcela	Dose de Nitrogênio (kg/ha)	Produção de forragem (kg/ha)
1	0	2.160
2	50	2.880
3	100	3.360
4	150	3.720
5	200	4.020

De modo geral, um gráfico de dispersão de valores observados para a variável resposta já é suficiente para indicar o tipo de curva (reta, parábola, etc) que melhor descreve o padrão geral dos dados. A figura a seguir mostra a dispersão dos valores observados para a variável produção de forragem quando diferentes doses de Nitrogênio foram aplicadas.

Podemos observar uma tendência linear dos dados, o que nos permite supor que a relação entre dose de Nitrogênio e produção de forragem seja linear.



Admitindo, então, o relacionamento linear entre as variáveis, vamos adotar o modelo de regressão linear simples:

$$y_i = \beta_0 + \beta_1 x_i + e_i ,$$

onde:

y_i é a produção de forragem (variável resposta), em kg;

x_i é a dose de Nitrogênio (variável preditora), em kg;

β_0 é a produção de forragem quando a dose de Nitrogênio aplicada for nula (intercepto), em kg;

β_1 é a quantidade que varia na produção de forragem para cada unidade (kg) aplicada de Nitrogênio (coeficiente de regressão), em kg/kg.

e_i é o erro (variação aleatória não controlável)

4.7.2. Análise de regressão

A análise de regressão tem por objetivo determinar a equação que melhor representa a relação existente entre duas variáveis e, a partir desta equação, fazer predições para a variável resposta. Para isso, é necessário que uma sequência de passos seja seguida:

1. Obtenção das estimativas (por ponto) dos coeficientes β_0 e β_1 para ajustar a equação da regressão.
2. Aplicação dos testes de significância para as estimativas obtidas, a fim de verificar se a equação de regressão é adequada.
3. Cálculo dos intervalos de confiança para os valores estimados pela equação de regressão.

4.7.2.1. Estimação dos parâmetros do modelo

Quando temos n observações, temos n pares de valores, (x_1, y_1); (x_2, y_2); ... ; (x_n, y_n), onde os valores observados para a variável resposta (y_i) são representados pela equação da regressão:

$$y_i = \beta_0 + \beta_1 x_i + e_i \quad i = 1,2,\ldots,n \qquad \begin{cases} y_1 = \beta_0 + \beta_1 x_1 + e_1 \\ y_2 = \beta_0 + \beta_1 x_2 + e_2 \\ \quad\quad \ldots \\ y_n = \beta_0 + \beta_1 x_n + e_n \end{cases}$$

Os coeficientes β_0 e β_1 são os parâmetros do modelo, e, portanto, constantes desconhecidas, que serão estimados a partir dos valores da amostra.

Se $y_i = \beta_0 + \beta_1 x_i + e_i$, então

$$E(y_i) = E(\beta_0 + \beta_1 x_i + e_i)$$
$$E(y_i) = E(\beta_0) + E(\beta_1 x_i) + E(e_i)$$
$$E(y_i) = \beta_0 + \beta_1 x_i$$

Sendo assim, se $y_i = \beta_0 + \beta_1 x_i + e_i$, então

$$y_i = E(y_i) + e_i,$$

logo,

$$e_i = y_i - E(y_i).$$

A estimação dos parâmetros β_0 e β_1 é efetuada através do método dos mínimos quadrados.

◆ **Método dos mínimos quadrados**

Este método tem como objetivo obter as estimativas dos parâmetros β_0 e β_1 de tal forma que a soma dos quadrados dos erros ($\sum e_i^2$) seja o menor valor possível.

Vimos que $e_i = y_i - E(y_i)$ e $E(y_i) = \beta_0 + \beta_1 x_i$,

logo,

$$\sum e_i^2 = \sum [y_i - E(y_i)]^2 = \sum [y_i - (\beta_0 + \beta_1 x_i)]^2.$$

Para encontrar os valores de β_0 e β_1 que tornam mínima a soma de quadrados dos erros, devemos, inicialmente, encontrar para a expressão acima as derivadas parciais em relação a β_0 e β_1.

$$\frac{\partial \sum e_i^2}{\partial \beta_0} = 2\sum (y_i - \beta_0 - \beta_1 x_i).(-1)$$
$$\frac{\partial \sum e_i^2}{\partial \beta_1} = 2\sum (y_i - \beta_0 - \beta_1 x_i).(-x_i)$$

Observamos que os valores de β_0 e β_1 das duas expressões acima variam de acordo com os valores que se atribui às derivadas parciais. Entretanto, para obter os pontos críticos (máximos ou mínimos), devemos igualar essas derivadas a zero, onde β_0 e β_1 assumem um valor particular, ou seja, representam as estimativas dos parâmetros de forma que a soma dos quadrados dos erros seja mínima. Deste modo, igualando a zero as derivadas parciais, temos

$$-2\sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \text{, sendo } \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

e

$-2\sum(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)x_i$, sendo $\sum(y_i x_i - \hat{\beta}_0 x_i - \hat{\beta}_1 x_i^2) = 0$.

Podemos, então determinar os valores de $\hat{\beta}_0$ e $\hat{\beta}_1$, através de um sistema de equações normais.

$$\begin{cases} \sum(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \\ \sum(y_i x_i - \hat{\beta}_0 x_i - \hat{\beta}_1 x_i^2) = 0 \end{cases}$$

Aplicando as propriedades da soma, temos

$$\begin{cases} \sum y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum x_i = 0 \\ \sum y_i x_i - \hat{\beta}_0 \sum x_i - \hat{\beta}_1 \sum x_i^2) = 0 \end{cases}$$

e arrumando a expressão para que os termos fiquem positivos, temos

$$\begin{cases} n\hat{\beta}_0 + \hat{\beta}_1 \sum x_i = \sum y_i \\ \hat{\beta}_0 \sum x_i + \hat{\beta}_1 \sum x_i^2 = \sum y_i x_i \end{cases}$$

A resolução do sistema pode ser feita por substituição. Começamos por isolar o $\hat{\beta}_0$ na primeira equação:

$$\begin{aligned} & n\hat{\beta}_0 + \hat{\beta}_1 \sum x_i = \sum y_i \\ & \frac{n\hat{\beta}_0}{n} + \frac{\hat{\beta}_1 \sum x_i}{n} = \frac{\sum y_i}{n} \\ & \frac{n\hat{\beta}_0}{n} + \hat{\beta}_1 \frac{\sum x_i}{n} = \frac{\sum y_i}{n} \\ & \hat{\beta}_0 + \hat{\beta}_1 \bar{x} = \bar{y} \\ & \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \end{aligned}$$

Determinado o valor de $\hat{\beta}_0$, isolamos o $\hat{\beta}_1$ na segunda equação:

$$\begin{aligned} & \hat{\beta}_0 \sum x_i + \hat{\beta}_1 \sum x_i^2 = \sum y_i x_i \\ & (\bar{y} - \hat{\beta}_1 \bar{x}) \sum x_i + \hat{\beta}_1 \sum x_i^2 = \sum y_i x_i \\ & \sum x_i \bar{y} - \sum x_i \hat{\beta}_1 \bar{x} + \hat{\beta}_1 \sum x_i^2 = \sum y_i x_i \\ & \sum x_i \bar{y} - \hat{\beta}_1 \bar{x} \sum x_i + \hat{\beta}_1 \sum x_i^2 = \sum y_i x_i \\ & \sum x_i \bar{y} + \hat{\beta}_1 \sum x_i^2 - \hat{\beta}_1 \bar{x} \sum x_i = \sum y_i x_i \\ & \hat{\beta}_1 \left(\sum x_i^2 - \bar{x} \sum x_i \right) = \sum y_i x_i - \sum x_i \bar{y} \end{aligned}$$

$$\hat{\beta}_1 = \frac{\sum y_i x_i - \sum x_i \bar{y}}{\sum x_i^2 - \bar{x} \sum x_i} = \frac{\sum y_i x_i - \sum x_i \frac{\sum y_i}{n}}{\sum x_i^2 - \frac{\sum x_i}{n} \sum x_i} = \frac{\sum y_i x_i - \frac{\sum x_i \sum y_i}{n}}{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}}.$$

Os estimadores $\hat{\beta}_0$ e $\hat{\beta}_1$ são os pontos críticos das raízes das equações $\frac{\partial \sum e_i^2}{\partial \beta_0} = 0$ e $\frac{\partial \sum e_i^2}{\partial \beta_1} = 0$, podendo ser pontos de mínimo ou de máximo. Entretanto, demonstra-se que os pontos críticos de qualquer função que seja uma soma de quadrados serão sempre pontos de mínimo. Daí podemos concluir que de $\hat{\beta}_0$ e $\hat{\beta}_1$ são pontos de mínimo, ou seja, a soma de quadrados dos erros é mínima.

Consideremos agora o experimento descrito no Exemplo 1. É importante lembrar que, sendo uma técnica de inferência, a análise de regressão linear simples tem o objetivo de determinar a equação que melhor represente o relacionamento entre as variáveis na população. No exemplo em questão, busca modelar a resposta média desta espécie de forrageira quando diferentes doses de Nitrogênio são aplicadas no solo. Sendo assim, cada parcela do experimento constitui uma amostra da população para uma determinada dose de Nitrogênio. Através da equação da reta ajustada podemos obter as estimativas dos valores médios das populações, denotados por $E(y/x_i)$ ou μ_i, para qualquer quantidade de Nitrogênio que pertença ao intervalo estudado, no exemplo, de 0 a 200 kg/ha. Vejamos agora como essas estimativas são obtidas.

Vimos que os valores observados são expressos por $y_i = \beta_0 + \beta_1 x_i + e_i$ e os valores esperados por $E(y_i/x_i) = \beta_0 + \beta_1 x_i$. As estimativas destes valores esperados são denotadas por $\hat{\mu}_i$ e podem ser obtidas através da equação ajustada:

$$\hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

A partir daí podemos obter também as estimativas dos erros. Sendo $e_i = y_i - E(y_i/x_i)$, as estimativas dos erros são obtidas por

$$\hat{e}_i = y_i - \hat{\mu}_i .$$

Utilizando os dados do Exemplo 1, vamos estimar os parâmetros do modelo de regressão linear simples. Inicialmente, devemos construir uma tabela auxiliar que inclua todos os cálculos intermediários para a obtenção das estimativas dos parâmetros, através do modelo $y_i = \beta_0 + \beta_1 x_i + e_i$.

Tabela auxiliar:

i	Dose de Nitrogênio (x_i)	Produção de forragem (y_i)	x_i^2	$x_i y_i$
1	0	2.160	0	0
2	50	2.880	2.500	144.000
3	100	3.360	10.000	336.000
4	150	3.720	22.500	558.000
5	200	4.020	40.000	804.000
Σ	500	16.140	75.000	1.842.000
Média	100	3.228	-	-

Obtidas a soma de quadrados X e a soma de produtos de X e Y, podemos calcular as estimativas de β_1 e β_0.

$$\hat{\beta}_1 = \frac{\sum y_i x_i - \frac{\sum x_i \sum y_i}{n}}{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}} = \frac{1842000 - \frac{500 \times 16140}{5}}{75000 - \frac{500^2}{5}} = \frac{228000}{25000} = 9{,}12$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 3228 - 9{,}12 \times 100 = 2316$$

Podemos obter agora as estimativas das médias de produção de forragem e dos erros para cada dose de Nitrogênio.

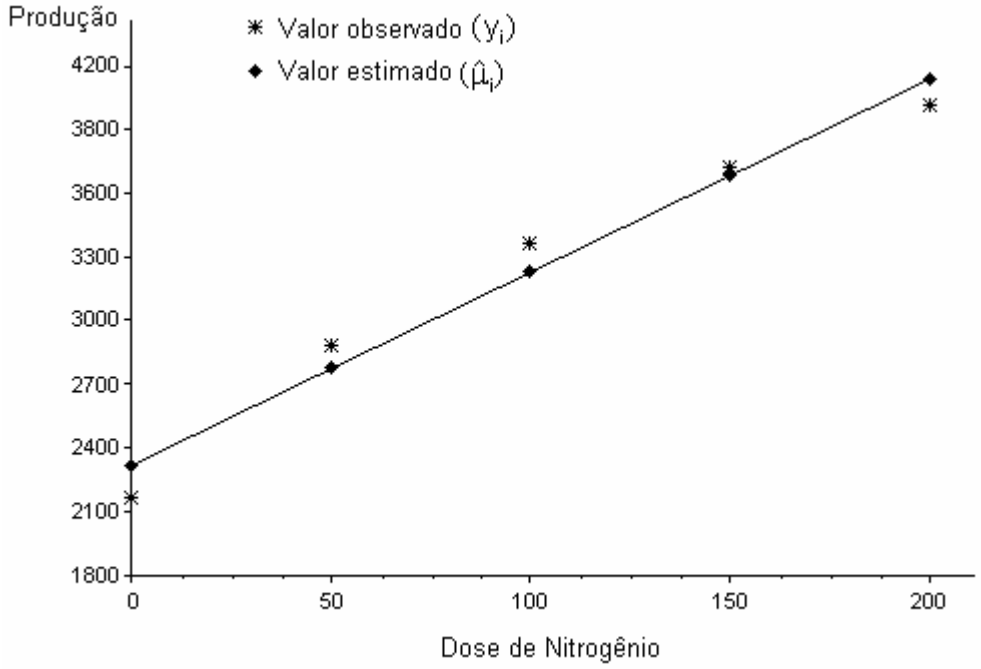
Sendo $\hat{\mu}_i = 2316 + 9{,}12 x_i$, temos

$\hat{\mu}_1 = 2316 + 9{,}12\, x_1 = 2316 + 9{,}12 \times 0 = 2.316$ kg/ha;
$\hat{\mu}_2 = 2316 + 9{,}12\, x_2 = 2316 + 9{,}12 \times 50 = 2.772$ kg/ha;
$\hat{\mu}_3 = 2316 + 9{,}12\, x_3 = 2316 + 9{,}12 \times 100 = 3.228$ kg/ha;
$\hat{\mu}_4 = 2316 + 9{,}12\, x_4 = 2316 + 9{,}12 \times 150 = 3.684$ kg/ha;
$\hat{\mu}_5 = 2316 + 9{,}12\, x_5 = 2316 + 9{,}12 \times 200 = 4.140$ kg/ha;

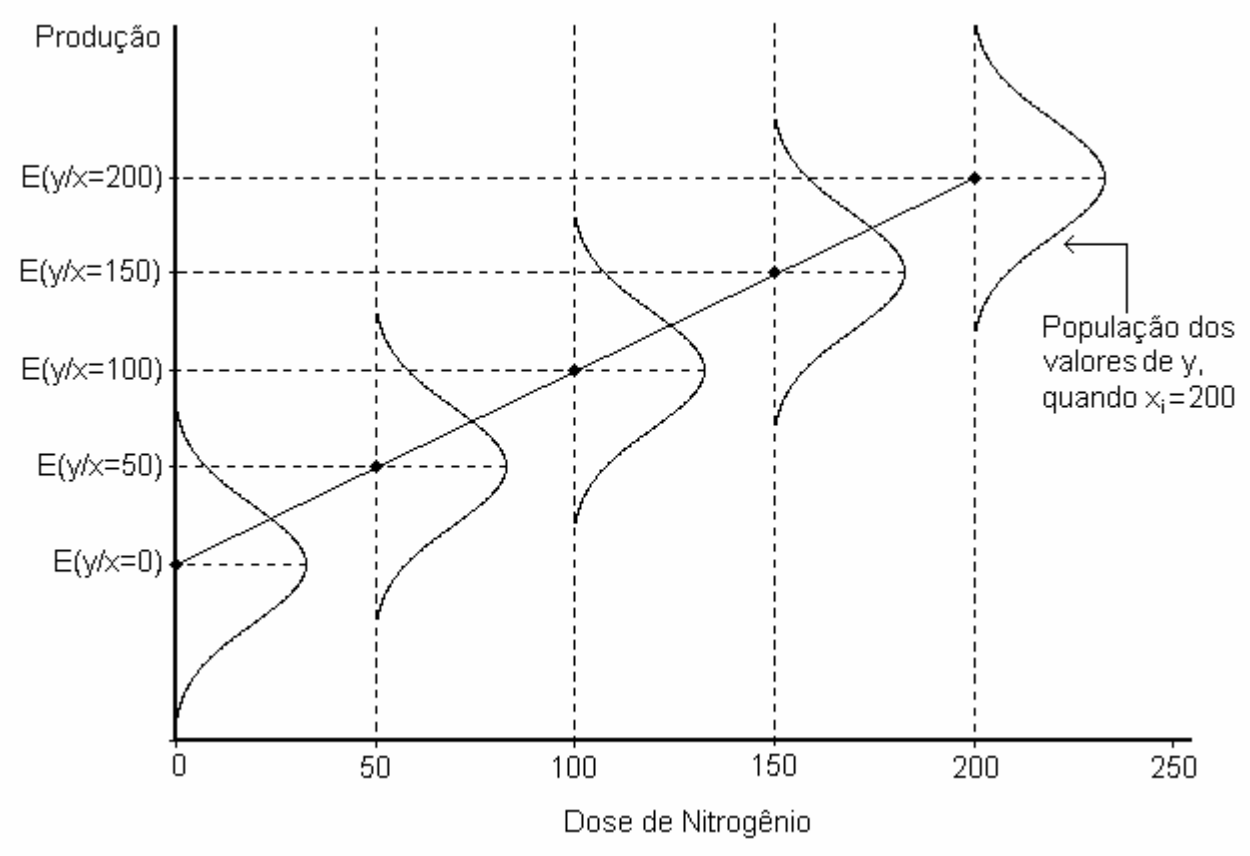
Sendo $\hat{e}_i = y_i - \hat{\mu}_i$, temos

$\hat{e}_1 = y_1 - \hat{\mu}_1 = 2160 - 2316 = -156$ kg/ha;
$\hat{e}_2 = y_2 - \hat{\mu}_2 = 2880 - 2772 = 108$ kg/ha;
$\hat{e}_3 = y_3 - \hat{\mu}_3 = 3360 - 3228 = 132$ kg/ha;
$\hat{e}_4 = y_4 - \hat{\mu}_4 = 3720 - 3684 = 36$ kg/ha;
$\hat{e}_5 = y_5 - \hat{\mu}_5 = 4020 - 4140 = -120$ kg/ha.

Na figura abaixo podemos observar o gráfico de dispersão dos valores de Y com a reta ajustada.



Admitindo que a variável resposta tem distribuição normal, os valores $\hat{\mu}_i$ estimam a produções médias populacionais $E(y/x_i)$ correspondentes às cinco doses de Nitrogênio aplicadas. O valor $y_5 = 4.020$ kg/ha , por exemplo, é o valor observado na parcela que recebeu 200 kg/ha de Nitrogênio e que constitui uma amostra aleatória da população que recebe esta dose, enquanto o valor $\hat{y}_5 = 4.140$ kg/ha é a estimativa da média desta população $E(y/x_i) = 200$, conforme podemos observar na figura a seguir.



É importante destacar também que o modelo de regressão linear simples pressupõe que as variâncias das populações de valores de Y são iguais para quaisquer valores de X. Essa homogeneidade de variâncias é representada na figura 4.4 pelas curvas de mesmo formato.

4.7.2.2. Testes de significância para a estimativa de β_1

Devemos considerar que as estimativas de β_0 e β_1, obtidas até agora, são estimativas por ponto, de modo que não sabemos o quão próximas elas estão dos parâmetros. Dentre os parâmetros do modelo de regressão linear simples, o coeficiente de regressão (β_1) é considerado o mais importante, pois é ele quem define a declividade da reta. Sendo assim, quando estimamos o β_1, devemos verificar se esta estimativa difere significativamente de zero. Esta verificação é feita através de um teste de hipóteses, cujas hipóteses de interesse são:

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_A : \beta_1 \neq 0 \end{cases}$$

Se o β_1 não diferir estatisticamente de zero significa que o efeito linear de X sobre Y não é significativo. Para testar H_0 podemos utilizar dois procedimentos: a *análise da variação* e o *teste t*, já estudado anteriormente.

◆ Análise da variação

A análise da variação consiste em decompor a variação total das observações, representada pelos desvios $(y_i - \bar{y})$, em duas partes:

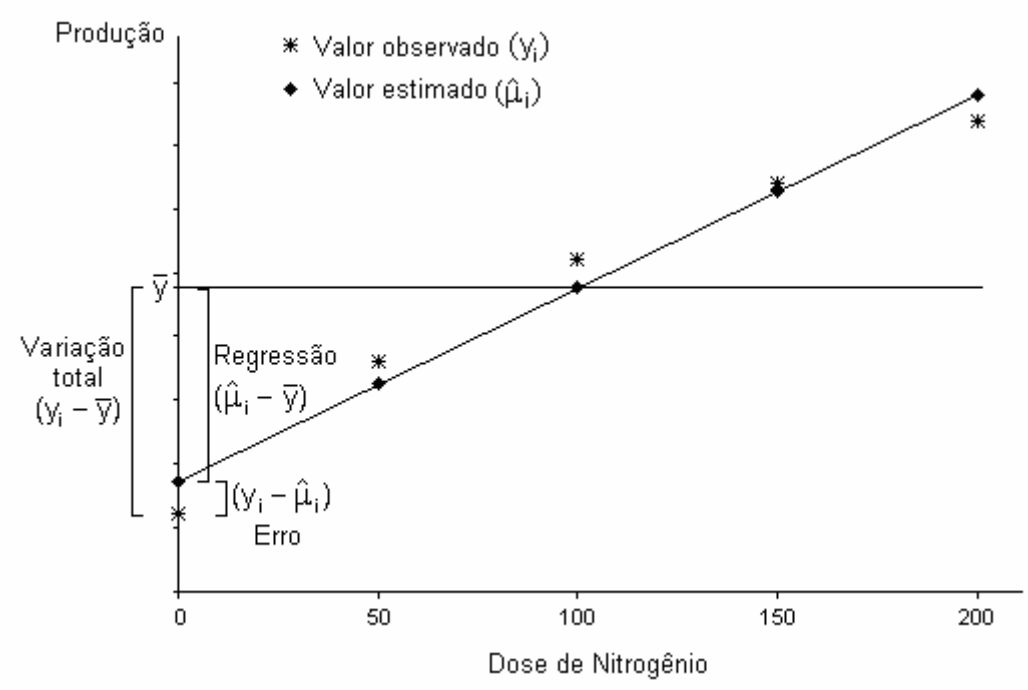
- a variação explicada pela reta da regressão, representada pelos desvios $(\hat{\mu}_i - \bar{y})$.

- a variação aleatória, não explicada pela reta, representada pelos desvios $(y_i - \hat{\mu}_i)$.

Assim, a variação de cada observação pode ser representada pela seguinte expressão:

$$(y_i - \bar{y}) = (\hat{\mu}_i - \bar{y}) + (y_i - \hat{\mu}_i)$$
$$(y_i - \bar{y}) = (\hat{\mu}_i - \bar{y}) + \hat{e}_i$$

Esses desvios podem ser observados na figura abaixo, onde temos o gráfico de dispersão dos pontos e a reta ajustada para os dados do experimento com Nitrogênio (Exemplo 1).



Considerando que a soma de desvios em relação à média é sempre zero, para obtermos a variação total das observações, devemos somar os quadrados dos desvios, o que resulta

$$\sum (y_i - \bar{y})^2 = \sum (\hat{\mu}_i - \bar{y})^2 + \sum (y_i - \hat{\mu}_i)^2$$

variação total | desvio explicado pela reta | desvio não explicado pela reta (erro)

Ao dividirmos as somas de quadrados (Q) pelos graus de liberdade obtemos as variâncias (V), também denominadas quadrados médios.

Os graus de liberdade e as variâncias (quadrados médios) são obtidos da seguinte forma:

– Grau de liberdade total: ν_{Total} = n–1, onde n é o número de observações.

– Grau de liberdade da regressão: ν_{Reg} = p–1, onde p é o número de parâmetros do modelo.

– Grau de liberdade do erro: ν_{Erro} = n–p

– Variância da regressão: $V_{Reg} = \frac{Q_{Reg}}{\nu_{Reg}}$

– Variância do erro: $V_{Erro} = \frac{Q_{Erro}}{\nu_{Erro}}$

A variância do erro (V_{Erro}) e a variância da regressão (V_{Reg}) são utilizados para testar a hipótese de interesse ($H_0 : \beta_1 = 0$). A V_{Erro} estima a variação aleatória (σ^2), enquanto a V_{Reg} estima a variação da regressão (σ^2_{Reg}) que é composta pela variação aleatória (σ^2) mais o efeito linear de X sobre Y (ϕ_{Reg}), ou seja, $\sigma^2_{Reg} = \sigma^2 + \phi_{Reg}$. Assim, temos um conjunto de hipóteses a respeito das variâncias que corresponde ao conjunto de hipóteses a respeito do β_1:

$$\begin{cases} H_0 : \sigma^2_{Reg} = \sigma^2 \rightarrow \text{efeito linear de X sobre Y não é significativo} \\ H_A : \sigma^2_{Reg} > \sigma^2 \end{cases}$$

$$\begin{cases} H_0 : \beta_1 = 0 \rightarrow \text{efeito linear de X sobre Y não é significativo} \\ H_A : \beta_1 \neq 0 \end{cases}$$

Para testar H_0, utilizamos a estatística F, que é definida como a razão entre duas variâncias e tem distribuição F, com parâmetros ν_1 e ν_2:

$$F = \frac{V_{Reg}}{V_{Erro}}$$

Se esta razão for significativamente maior do que 1 (um), concluímos que a variação da regressão é significativamente maior que a variação do erro e que, portanto, esta diferença se deve ao efeito linear de X sobre Y. Vale lembrar que o modelo só é adequado para explicar o relacionamento entre as duas variáveis quando a parte fixa do modelo (Regressão) é preponderante sobre a parte aleatória (Erro).

Em geral, a análise da variação é apresentada na forma de tabela, conforme o esquema abaixo.

Tabela da análise da variação:

Fonte de variação	ν	Q	E(V)	F
Regressão	p – 1	$\sum(\hat{\mu}_i - \bar{y})^2$	$\sigma^2 + \phi_{Reg}$	$\frac{V_{Reg}}{V_{Erro}}$
Erro	n – p	$\sum e_i^2 = \sum(y_i - \hat{\mu}_i)^2$	σ^2	
Total	n – 1	$\sum(y_i - \bar{y})^2$		

Para facilitar o processo de cálculo na obtenção das somas de quadrados, as seguintes fórmulas práticas podem ser utilizadas:

$$Q_{Total} = \sum y_i^2 - \frac{(\sum y_i)^2}{n};$$

$$Q_{Reg} = \hat{\beta}_1^2 \sum (x_i - \overline{x})^2;$$

$Q_{Erro} = Q_{Total} - Q_{Reg}$ (por diferença).

A decisão a respeito de H_0 será tomada comparando o valor da estatística F com o valor crítico encontrado na tabela de F.

Rejeitamos H_0, ao nível α de significância, se $f = \frac{V_{Reg}}{V_{Erro}} > f_{\alpha(\nu_1,\nu_2)}$.

Não rejeitamos H_0, ao nível α de significância, se $f = \frac{V_{Reg}}{V_{Erro}} < f_{\alpha(\nu_1,\nu_2)}$.

Para o Exemplo 1 vamos testar a hipótese de interesse a respeito do β_1. Inicialmente, obtemos as somas de quadrados, através das fórmulas práticas. Temos então:

$$Q_{Total} = \sum y_i^2 - \frac{(\sum y_i)^2}{n} = 54248400 - \frac{260499600}{5} = 2148480$$

$$Q_{Reg} = \hat{\beta}_1^2 \sum (x_i - \overline{x})^2 = 9{,}12^2 \times 25000 = 2079360$$

$Q_{Erro} = Q_{Total} - Q_{Reg} = 2148480 - 2079360 = 69120$

Obtidas as somas de quadrados, os demais resultados podem ser apresentados diretamente na tabela da análise da variação.

Tabela da análise da variação:

Fonte de variação	ν	Q	V	F
Regressão	1	2.079.360	2.079.360	90,25
Erro	3	69.120	23.040	
Total	4	2.148.480		

Como o valor calculado f = 90,25 foi maior que o valor crítico $f_{0,01(1,3)} = 34{,}12$, concluímos, ao nível α = 0,01, que o efeito linear da dose de Nitrogênio sobre a produção desta forrageira é significativo, sendo que essa relação pode ser expressa pela equação $\hat{\mu}_i = 2316 + 9{,}12x_i$. Isto significa que para cada kg/ha de Nitrogênio aplicado no solo a produção de forragem aumenta, em média, 9,12 kg/ha.

◆ **Teste t**

Outro procedimento que pode ser utilizado para testar $H_0 : \beta_1 = 0$ é o teste t. Como já visto em seções anteriores, utilizamos a estatística T que tem distribuição t de Student quando H_0 é verdadeira. Nesse caso, temos $\theta = \beta_1 = 0$, resultando:

$$T = \frac{\hat{\theta} - \theta}{S(\hat{\theta})} = \frac{\hat{\theta} - 0}{S(\hat{\theta})} = \frac{\hat{\theta}}{S(\hat{\theta})} \sim t(\nu),$$

onde:

$\hat{\theta} = \hat{\beta}_1$;

$S(\hat{\theta}) = S(\hat{\beta}_1)$;

$\nu = n - 2$;

A estimativa do erro padrão do estimador do coeficiente de regressão, $S(\hat{\beta}_1)$, é obtida da seguinte forma:

$$V(\hat{\beta}_1) = V\left(\frac{\sum y_i(x_i - \bar{x})}{\sum (x_i - \bar{x})^2}\right)$$

$$V(\hat{\beta}_1) = \left(\frac{1}{\sum (x_i - \bar{x})^2}\right)^2 V\left[\sum y_i(x_i - \bar{x})\right]$$

$$V(\hat{\beta}_1) = \frac{1}{\left[\sum (x_i - \bar{x})^2\right]^2} (\sum x_i - \bar{x})^2 V(y_i)$$

$$V(\hat{\beta}_1) = \frac{V(y_i)}{\sum (x_i - \bar{x})^2}$$

$$V(\hat{\beta}_1) = \frac{\sigma^2}{\sum (x_i - \bar{x})^2}$$

Sendo σ^2 um parâmetro desconhecido, utilizamos o seu estimador

$$S^2 = \frac{\sum (y_i - \hat{\mu}_i)^2}{n - 2} = \frac{\sum \hat{e}_i^2}{n - 2}$$

para obter a estimativa da variância do estimador do coeficiente de regressão

$$S^2(\hat{\beta}_1) = \frac{S^2}{\sum (x_i - \bar{x})^2} = \frac{\frac{\sum \hat{e}_i^2}{n - 2}}{\sum (x_i - \bar{x})^2}.$$

Daí resulta que

$$S(\hat{\beta}_1) = \sqrt{S^2(\hat{\beta}_1)} = \sqrt{\frac{\frac{\sum \hat{e}_i^2}{n - 2}}{\sum (x_i - \bar{x})^2}}.$$

Assim, sob H_0 verdadeira, temos

$$T = \frac{\hat{\beta}_1}{S(\hat{\beta}_1)} = \frac{\hat{\beta}_1}{\sqrt{\frac{\sum \hat{e}_i^2}{\frac{n-2}{\sum (x_i - \bar{x})^2}}}} \sim t(\nu)$$

No exemplo, temos

$$t = \frac{\hat{\beta}_1}{s(\hat{\beta}_1)} = \frac{9,12}{\sqrt{\frac{23040}{25000}}} = \frac{9,12}{0,96} = 9,5$$

Como o valor calculado t = 9,5 foi maior que o valor crítico valor $t_{\alpha/2(3)}$ = 5,841, concluímos, ao nível α = 0,01, que o efeito linear da dose de Nitrogênio sobre a produção desta forrageira é significativo. Podemos verificar também a correspondência entre os valores das estatísticas F e T. O valor da estatística F deve ser igual ao quadrado do valor da estatística T $(f = t^2)$. Para esse exemplo temos $f = 90,25 = 9,5^2 = t^2$.

Vimos em seções anteriores que o teste t bilateral e o intervalo de confiança, para um mesmo nível α, são procedimentos estatísticos equivalentes de modo que conduzem aos mesmos resultados. Sendo assim, o intervalo de confiança também pode ser utilizado para verificar se β_1 difere significativamente de zero ou não. Utilizando as mesmas expressões acima deduzidas, podemos obter o intervalo de confiança para o β_1. Partindo da expressão geral para intervalos de confiança

$$IC(\theta; 1-\alpha): \hat{\theta} \pm t_{\alpha/2} S(\hat{\theta}),$$

e fazendo as substituições referentes ao parâmetro em questão, temos

$$IC(\beta_1; 1-\alpha): \hat{\beta}_1 \pm t_{\alpha/2} S(\hat{\beta}_1)$$

$$IC(\beta_1; 1-\alpha): \hat{\beta}_1 \pm t_{\alpha/2} \sqrt{\frac{\sum \hat{e}_i^2}{\frac{n-2}{\sum (x_i - \bar{x})^2}}}$$

No exemplo, temos

$$IC(\beta_1; 1-\alpha): \hat{\beta}_1 \pm t_{\alpha/2} \sqrt{\frac{\sum \hat{e}_i^2}{\frac{n-2}{\sum (x_i - \bar{x})^2}}}$$

$$IC(\beta_1; 0,99): 9,12 \pm 5,841 \sqrt{\frac{23040}{25000}}$$

$$IC(\beta_1; 0,99): 9,12 \pm 5,61$$

$$\text{Limite inferior}: 9,12 - 5,61 = 3,51$$

$$\text{Limite superior}: 9,12 + 5,61 = 14,63$$

$$P(3,51 < \beta_1 < 14,63) = 0,99$$

Assim, concluímos que probabilidade de os limites 3,51 e 14,63 conterem o verdadeiro valor do coeficiente de regressão β_1 é de 0,99. Portanto, o efeito linear da dose de Nitrogênio sobre a produção da forrageira é significativo.

O teste de significância e o intervalo de confiança para o parâmetro β_0 são feitos de maneira análoga. Nesse caso, a estatística

$$T = \frac{\hat{\theta}}{S(\hat{\theta})} \sim t(\nu)$$

é utilizada considerando o seguinte:

$\theta = \beta_0$;

$\hat{\theta} = \hat{\beta}_0$;

$$S(\hat{\theta}) = S(\hat{\beta}_0) = \sqrt{\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}\right] S^2} = \sqrt{\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}\right] \frac{\sum e_i^2}{n-2}};$$

$\nu = n - 2$.

4.7.2.3. Intervalos de confiança para as médias das populações μ_i

Como vimos anteriormente, μ_i é um parâmetro e $\hat{\mu}_i$ é a estimativa pontual desse parâmetro. Vejamos agora como construir um intervalo de confiança para μ_i. Consideremos a expressão geral do intervalo de confiança:

$$IC(\theta; 1-\alpha): \hat{\theta} \pm t_{\alpha/2}\, S(\hat{\theta}),$$

onde:

$\theta = \mu_i$

$\hat{\theta} = \hat{\mu}_i$

$S(\hat{\theta}) = S(\hat{\mu}_i)$

$\nu = n - 2$.

Para obter o erro padrão do estimador $\hat{\mu}_i$, partimos do modelo

$$\hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i,$$

Sendo $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$, temos

$$\hat{\mu}_i = \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i$$

$$\hat{\mu}_i = \bar{y} + \hat{\beta}_1 (x_i - \bar{x}).$$

Para este modelo temos

$$V(\hat{\mu}_i) = V\left[\bar{y} + \hat{\beta}_1 (x_i - \bar{x})\right]$$

$$V(\hat{\mu}_i) = V(\bar{y}) + V\left[\hat{\beta}_1 (x_i - \bar{x})\right]$$

$$V(\hat{\mu}_i) = V(\bar{y}) + (x_i - \bar{x})^2\, V(\hat{\beta}_1)$$

Sendo $\hat{\beta}_1 = \frac{\sum y_i (x_i - \bar{x})}{\sum (x_i - \bar{x})^2}$, $V(\bar{y}) = \frac{\sigma^2}{n}$ e $\sigma^2 = \frac{\sum e_i^2}{n-2}$, temos

$$V(\hat{\mu}_i) = \frac{\sigma^2}{n} + (x_i - \bar{x})^2 \, V\left(\frac{\sum y_i (x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \right)$$

$$V(\hat{\mu}_i) = \frac{\sigma^2}{n} + (x_i - \bar{x})^2 \frac{\sigma^2}{\sum (x_i - \bar{x})^2}$$

$$V(\hat{\mu}_i) = \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right) \sigma^2$$

$$V(\hat{\mu}_i) = \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right) \left(\frac{\sum e_i^2}{n-2} \right)$$

Sendo σ^2 um valor desconhecido, utilizamos o seu estimador

$$S^2 = \frac{\sum \hat{e}_i^2}{n-2}$$

para obter a estimativa da variância do estimador $\hat{\mu}_i$

$$S^2(\hat{\mu}_i) = \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right) S^2 = \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right) \left(\frac{\sum \hat{e}_i^2}{n-2} \right).$$

Daí resulta que

$$S(\hat{\mu}_i) = \sqrt{S^2(\hat{\mu}_i)} = \sqrt{\left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right) \left(\frac{\sum \hat{e}_i^2}{n-2} \right)}.$$

O intervalo de confiança para μ_i é obtido pela expressão

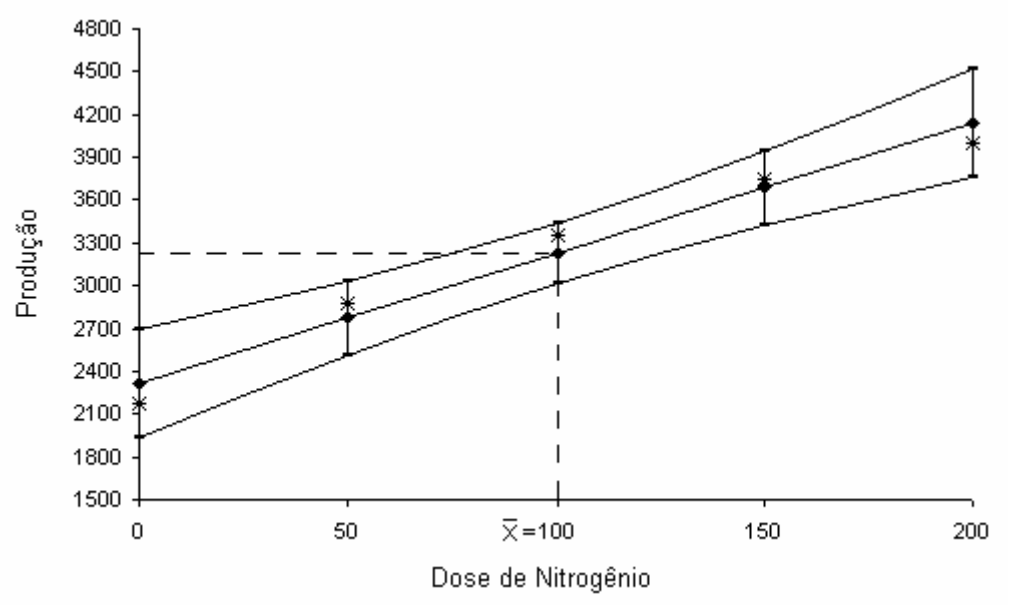
$$IC(\mu_i; 1-\alpha) : \hat{\mu}_i \pm t_{\alpha/2} \, S(\hat{\mu}_i)$$

$$IC(\mu_i; 1-\alpha) : \hat{\mu}_i \pm t_{\alpha/2} \sqrt{\left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right) \left(\frac{\sum \hat{e}_i^2}{n-2} \right)}.$$

Utilizando a expressão acima, vamos construir os intervalos de confiança para as médias do Exemplo 1. Na tabela auxiliar abaixo temos os cálculos intermediários e os valores obtidos para os limites dos intervalos, considerando α = 0,05 e o valor $t_{\alpha/2(3)}$ = 3,183.

i	x_i	y_i	$(x_i - \bar{x})^2$	$\hat{\mu}_i$	e_i^2	$s(\hat{\mu}_i)$	$t_{\alpha/2}\, s(\hat{\mu}_i)$	Limite inferior	Limite superior
1	0	2160	10000	2.316	24.336	117,58	374,26	1.941,76	2.690,24
2	50	2880	2500	2.772	11.664	83,14	264,63	2.507,37	3.036,63
3	100	3360	0	3.228	17.424	67,88	216,06	3.011,93	3.444,07
4	150	3720	2500	3.684	1.296	83,14	264,63	3.419,37	3.948,63
5	200	4020	10000	4.140	14.400	117,58	374,26	3.765,76	4.514,24
Σ	500	16.140	25.000	16.140	69.120	-	-	-	-

A figura a seguir apresenta o gráfico de dispersão dos valores de Y com os intervalos ao nível de 95% de confiança estimados para as médias μ_i. Podemos observar que o intervalo de confiança tem maior precisão no ponto $x_i = \bar{x}$, onde o desvio $(x_i - \bar{x})$ é igual a zero. À medida que se distancia da média, o intervalo de confiança aumenta sua amplitude, ou seja, diminui a precisão.



Dispersão dos valores de Y com os intervalos ao nível de 95% de confiança estimados para as médias μ_i.

4.8. Testes de qui-quadrado (χ^2)

4.8.1. Considerações gerais

Até agora tratamos da análise dos chamados dados de *medição* ou *mensuração*, que são valores referentes a variáveis numéricas de variação *contínua,* tais como peso, altura, temperatura, etc.

Em muitos casos, entretanto, é comum o pesquisador defrontar-se com problemas em que necessita verificar, a partir de um grupo de indivíduos (amostra), se frequências observadas em classes de uma variável *qualitativa* (cor, forma, estado, opinião, etc.) estão de acordo com frequências resultantes de uma teoria.

As classes, também denominadas categorias, são as alternativas das variáveis qualitativas em estudo. Os indivíduos que constituem a amostra são enquadrados nessas classes e contados. As observações numéricas resultantes dessa contagem são dados de enumeração e representam as *frequências observadas* nessas classes. Os dados de enumeração provenientes de uma teoria são denominados *frequências esperadas*.

O teste que permite verificar se frequências observadas estão de acordo com frequências esperadas é denominado teste qui-quadrado por utilizar a estatística Q que tem distribuição qui-quadrado. A seguir são relacionados alguns exemplos de proporções que podem ser verificadas através do teste qui-quadrado:

- proporções de germinação de sementes;
- proporção de pacientes curados após a aplicação de uma vacina ou medicamento;
- proporção de peças defeituosas que saem de uma linha de montagem;
- em estudos no campo da genética, proporções de fenótipos resultantes de cruzamentos de indivíduos.

4.8.2. Estatística do teste

Quando queremos verificar se as diferenças entre as frequências observadas e esperadas nas classes de uma variável qualitativa são reais ou casuais, utilizaremos o teste qui-quadrado dado por uma estatística Q, que tem distribuição qui-quadrado com parâmetro ν. Esta estatística é assim definida:

$$Q = \sum_{i=1}^{k} \frac{(X_i - E_i)^2}{E_i} \sim \chi^2(\nu),$$

onde:

X_i: é a frequência observada da classe i;
E_i: é a frequência esperada da classe i;
k: é número total de classes da variável;
$\nu = k - 1$: é o número de graus de liberdade ou classes independentes.

O teste qui-quadrado está sujeito a algumas restrições que devem ser observadas sempre que ele for utilizado.

1. O teste é válido apenas para *frequências absolutas*. Percentagens e proporções devem ser transformadas em frequências absolutas antes da realização do teste.

2. A distribuição qui-quadrado é uma distribuição derivada da distribuição normal, sendo, portanto, uma distribuição contínua. Como os dados analisados através de procedimentos qui-quadrado são provenientes de processos de contagem, algumas considerações deverão ser feitas para garantir uma boa aproximação.

a) Usar uma correção, chamada de "correção de continuidade", sempre que se trabalha com apenas um grau de liberdade. Essa correção consiste em subtrair 0,5 do módulo da diferença entre as frequências observada e esperada, ou seja,

$$Q = \sum_{i=1}^{k} \frac{\left(|X_i - E_i| - 0,5 \right)^2}{E_i}.$$

b) A aproximação de distribuições discretas para contínuas só é razoável quando se assegura que nenhuma frequência esperada seja inferior a 5, de forma que a aproximação melhora para valores maiores. Assim, quando há frequências esperadas menores que 5 é conveniente agrupá-las.

4.8.3. Classificação simples

Quando o objetivo for verificar se as frequências observadas concordam com as frequências esperadas dadas por uma teoria, teremos uma tabela de *classificação simples*, ou seja, onde os indivíduos são classificados segundo um único fator qualitativo. De modo geral, podemos representar esta tabela da seguinte forma

A	A_1	A_2	...	A_k
Frequência observada	X_1	X_2	...	X_k
Frequência esperada	E_1	E_2	...	E_k

onde:

A é a variável categórica;
A_i são as classes (categorias) da variável;
k é o número total de classes da variável;
X_i são as frequências observadas nas classes da variável;
E_i são as frequências esperadas para as classes da variável segundo uma determinada teoria.

Neste caso, a hipótese de nulidade a ser testada é a *hipótese de aderência* ou *concordância* que supõe que os dados observados se ajustam a teoria dada pelas frequências esperadas. A hipótese alternativa, naturalmente, deve supor o contrário. Assim temos

H_0: as frequências observadas *concordam* com as frequências esperadas
H_A: as frequências observadas *não concordam* com as frequências esperadas

4.8.4. Classificação dupla

Em muitos casos, o objetivo não é apenas verificar se as frequências observadas concordam com as esperadas, mas sim verificar se *duas variáveis qualitativas* (A e B) inerentes de um mesmo indivíduo são ou não *independentes* entre si.

Neste caso, os indivíduos são classificados segundo essas duas variáveis e dispostos em uma tabela de dupla entrada denominada *tabela de contingência*. A mais simples dessas tabelas é a 2×2, na qual cada variável tem apenas duas alternativas (classes). Entretanto, numa tabela $r \times s$ (linha por coluna) pode haver mais de duas alternativas para um ou ambas as variáveis.

A	B				Totais
	B_1	B_2	…	B_s	
A_1	X_{11} (E_{11})	X_{12} (E_{12})	…	X_{1s} (E_{1s})	X_{1+}
A_2	X_{21} (E_{21})	X_{22} (E_{22})	…	X_{2s} (E_{1s})	X_{2+}
…	…	…	…	…	…
A_r	X_{r1} (E_{r1})	X_{r2} (E_{r2})	…	X_{rs} (E_{rs})	X_{r+}
Totais	X_{+1}	X_{+2}	…	X_{+s}	X_{++}

O teste é efetuado através da variável Q, assim definida:

$$Q = \sum_i \sum_j \frac{(X_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2(\nu),$$

onde:

X_{ij}: frequência observada da linha i e coluna j, sendo i = 1, 2,..., r e j = 1, 2,..., s;
E_{ij}: frequência esperada da linha i e coluna j;
r: número total de linhas (classes da variável A);
s: número total de colunas (classes da variável B);
$\nu = (r-1).(s-1)$: número de graus de liberdade ou classes independentes.

As frequências esperadas (E_{ij}) são obtidas através da seguinte expressão

$$E_{ij} = X_{i+} \times \frac{X_{+j}}{X_{++}} = \frac{X_{i+} \times X_{+j}}{X_{++}},$$

onde:

X_{i+} : somatório da linha i

X_{+j} : somatório da coluna j

X_{++} : somatório de todas as linhas e todas as colunas

A hipótese a ser testada é a *hipótese de independência*, que supõem que as variáveis A e B independem entre si, ou seja,

H_0: a variável A *independe* da variável B
H_A: a variável A *depende* da variável B

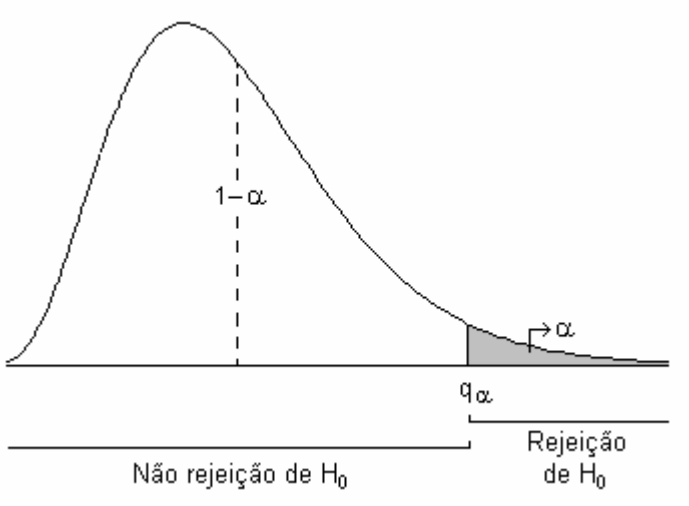
4.8.5. Critério de decisão

A regra de decisão a respeito de H_0 pode ser estabelecida com base no valor crítico $q_{\alpha(\nu)}$ que, para o número de graus de liberdade ν, delimita a área α (Tabela III do Apêndice), ou seja,

- Rejeitamos H_0, ao nível α, se $q = \sum_{i=1}^{k} \frac{(x_i - e_i)^2}{e_i} > q_{\alpha(\nu)}$.

- Não temos motivos suficientes para rejeitar H_0, se $q = \sum_{i=1}^{k} \frac{(x_i - e_i)^2}{e_i} < q_{\alpha(\nu)}$.

Podemos observar na figura abaixo a região de rejeição de H_0 na curva da distribuição qui-quadrado:



Consideremos os exemplos resolvidos:

Exemplo 1. Num determinado cruzamento, os indivíduos resultantes foram classificados em quatro fenótipos e contados, sendo observado o seguinte:

Fenótipo	Número de indivíduos (X_i)
A	103
B	37
C	28
D	8

Verifique se esse resultado concorda com as respectivas proporções de 9/16; 3/16; 3/16 e 1/16 dadas pelas leis de Mendel, usando α = 0,05.

Resolução:

1. Hipóteses estatísticas

H_0: as frequências observadas *concordam* com as frequências esperadas
H_A: as frequências observadas *não concordam* com as frequências esperadas

2. Estatística do teste

Obtenção das frequências esperadas (E_i)

X_i	E_i
103	$176 \times \frac{9}{16} = 99$
37	$176 \times \frac{3}{16} = 33$
28	$176 \times \frac{3}{16} = 33$
8	$176 \times \frac{1}{16} = 11$
176	176

Grau de liberdade: ν = k - 1 = 4 - 1 = 3

Taxa de erro: α = 0,05

$$q = \sum_{i=1}^{k} \frac{(x_i - e_i)^2}{e_i} = \frac{(103-99)^2}{99} + \frac{(37-33)^2}{33} + \frac{(28-33)^2}{33} + \frac{(8-11)^2}{11}$$
$$q = 0,1616 + 0,4848 + 0,7576 + 0,8182 = 2,222$$

3. Decisão e conclusão

Como q = 2,222 < $q_{\alpha(\nu)}$ = 7,81, não temos motivos para rejeitar H_0. Assim, concluímos, ao nível de 5% de significância, que as frequências observadas não diferem significativamente das frequências esperadas segundo as leis de Mendel.

Exemplo 2. Nos resultados de uma determinada pesquisa de opinião, foram obtidas 35 respostas favoráveis à opção 1 e 46 respostas favoráveis à opção 2. Verifique se esses resultados concordam com as proporções esperadas de 1/2 para a opção 1 e 1/2 para a opção 2. Use α=0,01.

Resolução:

1. Hipótese estatística

H_0: as frequências observadas *concordam* com as frequências esperadas
H_A: as frequências observadas *não concordam* com as frequências esperadas

2. Estatística do teste

Obtenção das frequências esperadas (E_i)

X_i	E_i
35	$81 \times \frac{1}{2} = 40,5$
46	$81 \times \frac{1}{2} = 40,5$
81	81

Grau de liberdade: $\nu = k - 1 = 2 - 1 = 1$

Taxa de erro: α = 0,01

Como o número de graus de liberdade é igual a 1, obtém-se o valor da estatística Q procedendo à correção de continuidade:

$$q = \sum_{i=1}^{k} \frac{(|x_i - e_i| - 0,5)^2}{e_i} = \frac{(|35 - 40,5| - 0,5)^2}{40,5} + \frac{(|46 - 40,5| - 0,5)^2}{40,5} = 0,7469 + 0,7469 = 1,494$$

3. Decisão e conclusão

Como q = 1,494 < $q_{\alpha(\nu)}$ = 6,63, não temos motivos para rejeitar H_0. Concluímos, então, ao nível de 1% de significância, que as frequências observadas não diferem significativamente das frequências esperadas. Portanto, as duas opções têm a mesma frequência.

Exemplo 3. Trezentos proprietários de uma certa marca de carro foram entrevistados sobre o desempenho e o consumo de combustível de seus carros. Os resultados obtidos na pesquisa de opinião foram os seguintes:

Consumo	Desempenho		Totais
	Regular	Bom	
Alto	152	48	200
Baixo	88	12	100
Totais	240	60	300

Verifique, com α = 0,05, se os atributos consumo e desempenho são independentes.

Resolução:

1. Hipótese estatística

H_0: o consumo *independe* do desempenho
H_A: o consumo *depende* do desempenho

2. Estatística do teste

Obtenção das frequências esperadas (E_i)

$$E_{11} = 200 \times \frac{240}{300} = 160, \qquad E_{12} = 200 \times \frac{60}{300} = 40$$

$$E_{21} = 100 \times \frac{240}{300} = 80, \qquad E_{22} = 100 \times \frac{60}{300} = 20$$

Consumo	Desempenho		Totais
	Regular	Bom	
Alto	152 (160)	48 (40)	200
Baixo	88 (80)	12 (20)	100
Totais	240	60	300

Grau de liberdade: $\nu = (r-1).(s-1) = (2-1) \cdot (2-1) = 1$

Taxa de erro: α = 0,05

Como o número de graus de liberdade é igual a 1, obtém-se o valor da estatística Q procedendo à correção de continuidade:

$$q = \sum_{i,j} \frac{\left(\left|x_{ij} - e_{ij}\right| - 0,5\right)^2}{e_i} = \frac{\left(\left|152-160\right| - 0,5\right)^2}{160} + \frac{\left(\left|48-40\right| - 0,5\right)^2}{40} + \frac{\left(\left|88-80\right| - 0,5\right)^2}{80} + \frac{\left(\left|12-20\right| - 0,5\right)^2}{20}$$

$$q = 0,3516 + 1,406 + 0,7031 + 2,813 = 5,274$$

3. Decisão e conclusão

Como q = 5,274 > $q_{\alpha(\nu)}$ = 3,84, temos motivos suficientes para rejeitar H_0. Concluímos, então, ao nível de 5% de significância, que as frequências observadas diferem significativamente das frequências esperadas. Portanto, o consumo de combustível depende do desempenho do carro.

Exemplo 4. O efeito de diversos tratamentos no controle de certa doença está apresentado na tabela abaixo

Tratamento	Evolução da doença		Totais
	Regrediu	Não regrediu	
A	92	13	105
B	62	12	74
C	35	14	49
D	19	13	32
Totais	208	52	260

Verifique qual a relação existente entre os diversos tratamentos e a evolução da doença, utilizando α = 0,05.

Resolução:

1. Hipótese estatística

H_0: a evolução da doença *independe* do tratamento
H_A: a evolução da doença *depende* do tratamento

2. Estatística do teste

Obtenção das frequências esperadas (E_i)

$$E_{11} = 105 \times \frac{208}{260} = 84 \quad E_{21} = 74 \times \frac{208}{260} = 59{,}2 \quad E_{31} = 49 \times \frac{208}{260} = 39{,}2 \quad E_{41} = 32 \times \frac{208}{260} = 25{,}6$$

$$E_{12} = 105 \times \frac{52}{260} = 21 \quad E_{22} = 74 \times \frac{52}{260} = 14{,}8 \quad E_{32} = 49 \times \frac{52}{260} = 9{,}8 \quad E_{42} = 32 \times \frac{52}{260} = 6{,}4$$

Tratamento	Evolução da doença		Totais
	Regrediu	Não regrediu	
A	92 (84)	13 (21)	105
B	62 (59,2)	12 (14,8)	74
C	35 (39,2)	14 (9,8)	49
D	19 (25,6)	13 (6,4)	32
Totais	208	52	260

Grau de liberdade: $\nu = (r-1).(s-1) = (4-1) \,.\, (2-1) = 3$

Taxa de erro: α = 0,05

$$q = \sum_{i,j} \frac{\left(x_{ij} - e_{ij}\right)^2}{e_i} = \frac{(92-84)^2}{84} + \frac{(13-21)^2}{21} + \ldots + \frac{(13-6{,}4)^2}{6{,}4}$$

$$q = 0{,}7619 + 3{,}048 + \ldots + 6{,}806 = 15{,}23$$

3. Decisão e conclusão

Como q = 15,23 > $q_{\alpha(\nu)}$ = 7,81, temos motivos suficientes para rejeitar H_0. Assim, concluímos, ao nível de 5% de significância, que as frequências observadas diferem significativamente das frequências esperadas. Portanto, a evolução da doença depende do tratamento utilizado.

4.9. Bibliografia

COSTA, S.F. **Introdução Ilustrada à Estatística (com muito humor!)**. 2.ed., São Paulo: Harbra, 1992. 303p.

DEVORE, J. **Probability and statistics for engineering and the sciences**. Brooks/Cole Publishing Companig. 1982. 640p.

FARIA, E.S. de **Estatística** Edição 97/1. (Apostila)

FERREIRA, D.F. **Estatística Básica**. Lavras: Editora UFLA, 2005, 664p.

FREUND, J.E., SIMON, G.A. **Estatística Aplicada. Economia, Administração e Contabilidade**. 9.ed., Porto Alegre: Bookman, 2000. 404p.

MEYER, P. L. **Probabilidade: aplicações à estatística**. Rio de Janeiro: LTC, 1976.

RIBEIRO, J.L.D.; TEN CATEN, C.S. **Estatística Industrial**. Porto Alegre, Universidade Federal do Rio Grande do Sul, 2000. 135p.

SILVA, J.G.C. da. **Estatística Experimental**. 1. Planejamento de Experimentos. 1. ed. Pelotas, RS: Instituto de Física e Matemática, Universidade Federal de Pelotas, 1997. v.1. 216p.

SILVA, J.G.C. da **Estatística experimental:** análise estatística de experimentos. Pelotas, RS: Instituto de Física e Matemática, Universidade Federal de Pelotas, 2000. 318p.

SPIEGEL, M.R. **Estatística** São Paulo: McGraw-Hill, 1972. 520p.

VIEIRA, S. **Estatística Experimental**. 9.ed., São Paulo: Atlas, 1999. 185p.

Apêndice

1. Notação somatório

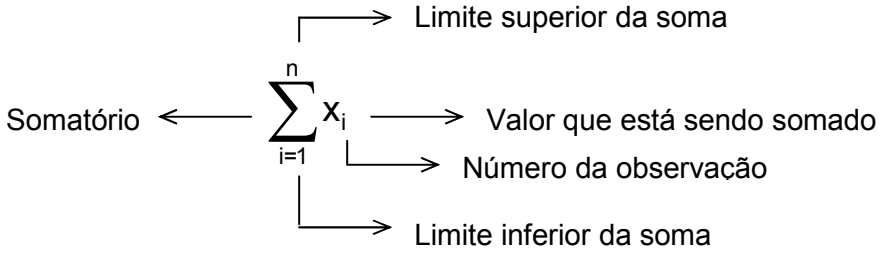
Consideremos a seguinte tabela de valores:

i	x_i	y_i
1	1	2
2	0	1
3	2	-2
4	-1	1
5	4	0

onde:
i é o número da observação, tal que i = 1, 2, ..., n
n é o número total de observações
x_i é o valor da variável X para a observação i, tal que $x_1 = 1$, $x_2 = 0$,...,$x_5 = 4$
y_i é o valor da variável Y para a observação i, tal que $y_1 = 2$, $y_2 = 1$,...,$y_5 = 0$
$x_{(i)}$ é o valor da variável X para a observação i, tal que $x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)}$
$x_{(1)}$ é o menor valor da variável X
$x_{(n)}$ é o maior valor da variável X

Somatório (Σ)

A notação Σ indica a soma seqüencial de um conjunto de valores. De modo geral, temos



A notação $\sum_{i=1}^{n}$ inclui todos os valores do intervalo e pode ser simplificada por $\sum$, onde omitimos os índices, ou seja, $\sum_{i=1}^{n} = \sum$

Exemplos:

1. $x_1 + x_2 + x_3 + x_4 + x_5 = \sum_{i=1}^{5} x_i$
2. $y_2 + y_3 + y_4 = \sum_{i=2}^{4} y_i$

Outras quantidades de interesse:

1. $x_1^2 + x_2^2 + x_3^2 + x_4^2 + x_5^2 = \sum_{i=1}^{5} x_i^2$ (soma de quadrados)
2. $(x_1 + x_2 + x_3 + x_4 + x_5)^2 = \left(\sum_{i=1}^{5} x_i\right)^2$ (quadrado da soma)
3. $x_1y_1 + x_2y_2 + x_3y_3 + x_4y_4 + x_5y_5 = \sum_{i=1}^{5} x_iy_i$ (soma de produtos)
4. $(x_1 + x_2 + x_3 + x_4 + x_5) \times (y_1 + y_2 + y_3 + y_4 + y_5) = \sum_{i=1}^{5} x_i \sum_{i=1}^{5} y_i$ (produto da soma)

◆ **Propriedades da soma**

1ª propriedade: A soma é distributiva, ou seja, se cada termo da soma é multiplicado por uma constante c, os termos podem ser somados e a soma multiplicada pela constante.

$$\sum_{i=1}^{n} c\,x_i = c\,x_1 + c\,x_2 + \ldots + c\,x_n = c\left(x_1 + x_2 + \ldots + x_n\right) = c\sum_{i=1}^{n} x_i$$

2ª propriedade: A soma de uma constante c sobre n termos é igual a n vezes esta constante.

$$\sum_{i=1}^{n} c = c + c + \ldots + c = c\left(1 + 1 + \ldots + 1\right) = nc$$

3ª propriedade: A soma é associativa, ou seja, o somatório da soma é igual a soma de somatórios.

$$\sum_{i=1}^{n}\left(x_i + y_i\right) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i$$

$$\left(x_1 + y_1\right) + \left(x_2 + y_2\right) + \ldots + \left(x_n + y_n\right) = \left(x_1 + x_2 + \ldots + x_n\right) + \left(y_1 + y_2 + \ldots + y_n\right)$$

As propriedades devem ser usadas no sentido de simplificar as operações. Sempre que houver uma operação que precede a soma, devemos desenvolvê-la antes de aplicar as propriedades.

$$\begin{aligned}\sum_{i=1}^{n}\left(x_i - c\right)^2 &= \sum_{i=1}^{n}\left(x_i^2 - 2cx_i + c^2\right)\\ &= \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} 2cx_i + \sum_{i=1}^{n} c^2\\ &= \sum_{i=1}^{n} x_i^2 - 2c\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} c^2\\ &= \sum_{i=1}^{n} x_i^2 - 2c\sum_{i=1}^{n} x_i + nc^2\end{aligned}$$

2. Noções sobre conjuntos

◆ **Conjunto**

Um conjunto é uma coleção bem definida de objetos chamados membros ou elementos. Geralmente, um conjunto é denotado por letra maiúscula (A, B, C) e os seus elementos por letras minúsculas (a, b, c).

Um conjunto pode ser definido de duas formas:

– *Método da listagem:* relacionando todos os elementos do conjunto.

Exemplo: A = {a, e, i, o, u} conjunto das vogais do alfabeto

– *Método da propriedade:* indicando uma propriedade que seja válida para todos os elementos do conjunto e só para eles.

Exemplo: A = {x; x é uma vogal} conjunto das vogais do alfabeto

◆ **Conjunto universal ou universo**

Quando restringimos nosso estudo a subconjuntos de um determinado conjunto, então este conjunto é chamado de conjunto universal, ou universo, ou espaço e é denotado por U.

Exemplo: Conjunto dos números reais R

$$A = \{x \in R;\ a \leq x \leq b\}$$
$$B = \{x \in R;\ 0 < x < b\}$$

◆ **Conjunto vazio**

O conjunto vazio, denotado por $\varnothing$ ou { }, é um conjunto desprovido de elementos. O conjunto vazio é subconjunto de qualquer conjunto.

Exemplo: $A = \{x \in R;\ x^2 < 0\} = \varnothing$

◆ **Representação geométrica de conjuntos**

Um universo pode ser representado geometricamente pelo conjunto de pontos interiores de um retângulo e os seus subconjuntos, tais como A e B, são representados pelos pontos interiores de círculos. Tais representações, denominadas *diagramas de Venn*, são úteis para dar intuição geométrica sobre a relação entre conjuntos.

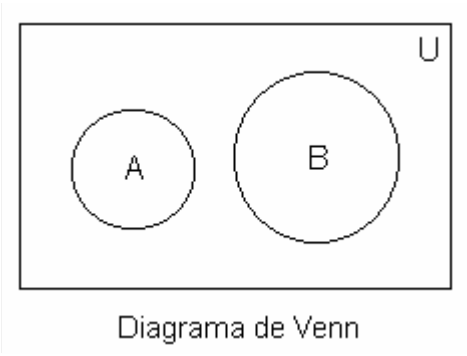
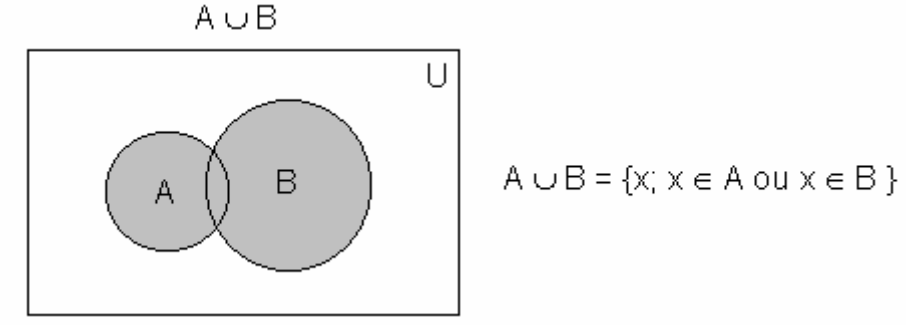


Diagrama de Venn

◆ **Operações com conjuntos**

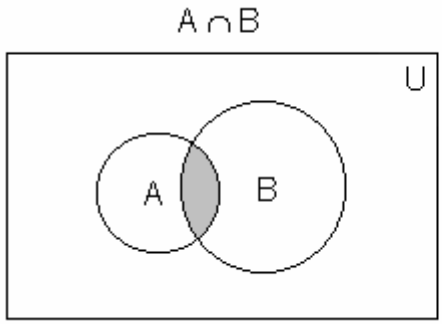
1. União ($\cup$)

A união de dois conjuntos A e B, representada por $A \cup B$, é o conjunto de todos os elementos que pertencem a A ou a B ou a ambos.



2. Intersecção ($\cap$)

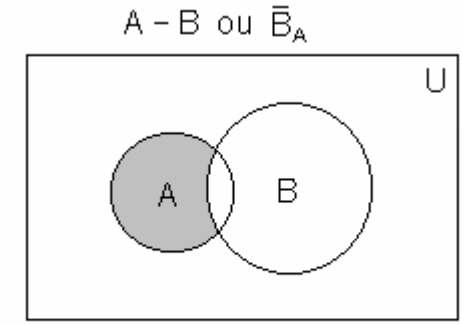
A intersecção de dois conjuntos A e B, representada por $A \cap B$, é o conjunto de todos os elementos que pertencem a e a B.



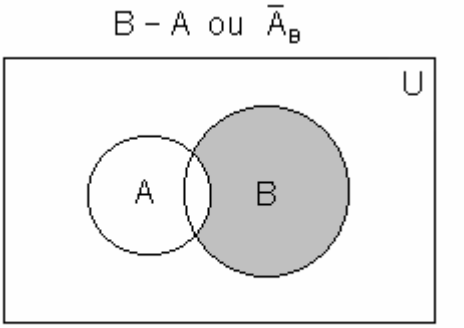
$$A \cap B = \{x;\ x \in A \text{ e } x \in B\}$$

3. Diferença ou complemento relativo

A diferença de dois conjuntos A e B, denotada por $A - B$, ou o complemento de B em relação a A, denotado por $\bar{B}_A$, é o conjunto de todos os elementos que pertencem a A e não pertencem a B.



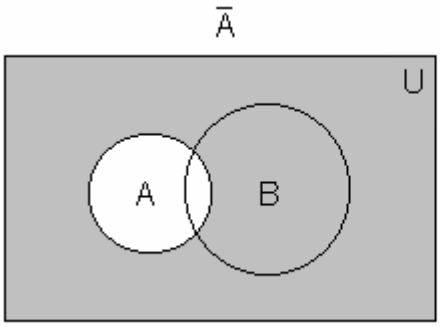
$$A - B = \{x;\ x \in A \text{ e } x \notin B\}$$



$$B - A = \{x;\ x \notin A \text{ e } x \in B\}$$

4. Complemento absoluto

O complemento absoluto ou simplesmente complemento de A, denotado por $\bar{A}$, é o conjunto de todos os elementos que não pertencem a A.



$$\bar{A} = \{x;\ x \notin A\}$$

Principais propriedades da união:

- $A \cup \varnothing = A$
- $A \cup A = A$
- $A \cup B = B \cup A$
- $(A \cup B) \cup C = A \cup (B \cup C)$
- $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$

Principais propriedades da intersecção:

- $A \cap \varnothing = \varnothing$
- $A \cap A = A$
- $A \cap B = B \cap A$
- $(A \cap B) \cap C = A \cap (B \cap C)$
- $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$

3. Notação fatorial

O produto dos inteiros positivos de 1 a n é representado pelo símbolo especial n! (lê-se “n fatorial”). Assim, temos

$$n! = 1 \times 2 \times 3 \times \ldots \times (n-2) \times (n-1) \times n$$

Define-se, também, que 0! = 1.

Exemplos:

$$2! = 2 \times 1 = 2$$
$$4! = 4 \times 3 \times 2 \times 1 = 24$$
$$5! = 5 \times 4! = 5 \times 24 = 120$$
$$\frac{8!}{6!} = \frac{8 \times 7 \times 6!}{6!} = 8 \times 7 = 56$$

4. Análise combinatória

Algumas técnicas de contagem foram desenvolvidas para determinar, sem enumeração direta, o número de elementos de certo conjunto, ou o número de resultados possíveis de um certo experimento. Essas técnicas são chamadas de análise combinatória.

Seja um conjunto A com n elementos distintos entre si. Se x elementos são retirados de A é possível formar grupos de três tipos:

Permutações: Grupos que se distinguem apenas pela *ordem* dos seus elementos. Se x = n, então, o número de possíveis permutações de n é dado por

$$P_n = n! \text{ grupos}$$

Arranjos: Grupos que se distinguem pela *ordem* e pela *natureza* dos seus elementos. Se x < n, então, o número de possíveis arranjos de n, tomados x a x, é dado por

$$A_n^x = \frac{n!}{(n-x)!} \text{ grupos}$$

Combinações: Grupos que se distinguem apenas pela *natureza* dos seus elementos. Se x < n, então, o número de possíveis combinações de n, tomados x a x, é dado por

$$C_n^x = \frac{n!}{x!(n-x)!} \text{ grupos}$$

Exemplo:

Seja A = {a, b, c, d}, onde n = 4.

1. Se são retirados quatro elementos, quantos grupos é possível formar?

Se x = n, então, os grupos formados serão permutações:

$$P_4 = 4! = 24 \text{ grupos}$$

{(a, b, c, d), (a, b, d, c), (a, c, b, d), (a, c, d, b), (a, d, b, c), (a, d, c, b), (b, a, c, d), (b, a, d, c), (b, c, a, d), (b, c, d, a), (b, d, a, c) (b, d, c, a), (c, a, b, d), (c, a, d, b), (c, b, a, d), (c, b, d, a), (c, d, a, b), (c, d, b, a), (d, a, b, c), (d, a, c, b), (d, b, a, c), (d, b, c, a), (d, b, a, c)}

2. Se são retirados dois elementos, quantos grupos que diferem pela ordem e pela natureza é possível formar?

Se os grupos formados devem diferir pela ordem e pela natureza, então serão arranjos:

$$A_4^2 = \frac{4!}{(4-2)!} = \frac{24}{2} = 12 \text{ grupos}$$

{(a, b), (b, a), (a, c), (c, a), (a, d), (d, a), (b, c), (c, b), (b, d), (d, b), (c, d), (d, c)}

3. Se são retirados dois elementos, quantos grupos que diferem apenas pela natureza é possível formar?

Se os grupos formados devem diferir apenas pela natureza, então serão combinações:

$$C_4^2 = \frac{4!}{2!(4-2)!} = \frac{24}{4} = 6 \text{ grupos}$$

{(a, b), (a, c), (a, d), (b, c), (b, d), (c, d)}

Permutações com repetição: Grupos com elementos repetidos que se distinguem apenas pela *ordem* dos seus elementos. Neste caso, n passa a ser o número de elementos retirados e x é o número de repetições de um dado elemento. O número de possíveis permutações de n, com x repetições um dado elemento, é dado por

$$P_n^{x,n-x} = \frac{n!}{x!(n-x)!} \text{ grupos}$$

Exemplo:

Seja um conjunto A formado por três moedas de ouro e quatro de prata.

A = {o, o, o, p, p, p, p}

1. Se quatro moedas são retiradas, de quantas maneiras diferentes podemos retirar duas moedas de prata?

Se n = 4 e x =2, então,

$$P_4^{2,4-2} = \frac{4!}{2!(4-2)!} = \frac{24}{4} = 6$$

{(p, p, o, o), (p, o, p, o), (p, o, o, p), (o, o, p, p), (o, p, o, p), (o, p, p, o)}

2. Se quatro moedas são retiradas, de quantas maneiras diferentes podemos retirar Três moedas de prata?

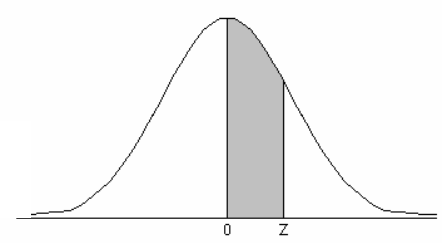
Se n = 4 e x =3, então,

$$P_4^{3,4-3} = \frac{4!}{3!(4-3)!} = \frac{24}{6} = 4$$

{(p, p, p, o), (p, p, o, p), (p, o, p, p), (o, p, p, p)}

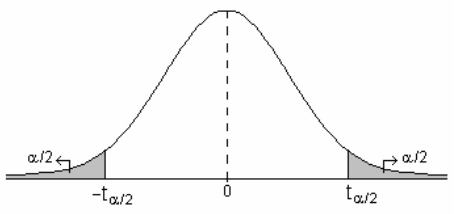
5. Tabelas estatísticas

Tabela I. Área sob a curva normal padrão de 0 a z, $P(0 \leq Z \leq z)$.



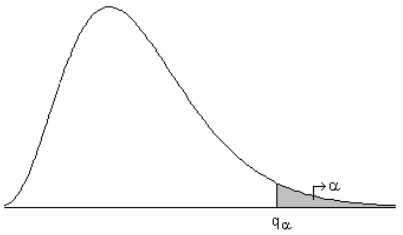
z	0	1	2	3	4	5	6	7	8	9
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0754
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2133	0,2157	0,2190	0,2224
0,6	0,2258	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2518	0,2549
0,7	0,2580	0,2612	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2996	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1,0	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319
1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441
1,6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545
1,7	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633
1,8	0,4641	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699	0,4706
1,9	0,4713	0,4719	0,4726	0,4732	0,4738	0,4744	0,4750	0,4756	0,4761	0,4767
2,0	0,4772	0,4778	0,4783	0,4788	0,4793	0,4798	0,4803	0,4808	0,4812	0,4817
2,1	0,4821	0,4826	0,4830	0,4834	0,4838	0,4842	0,4846	0,4850	0,4854	0,4857
2,2	0,4861	0,4864	0,4868	0,4871	0,4875	0,4878	0,4881	0,4884	0,4887	0,4890
2,3	0,4893	0,4896	0,4898	0,4901	0,4904	0,4906	0,4909	0,4911	0,4913	0,4916
2,4	0,4918	0,4920	0,4922	0,4925	0,4927	0,4929	0,4931	0,4932	0,4934	0,4936
2,5	0,4938	0,4940	0,4941	0,4943	0,4945	0,4946	0,4948	0,4949	0,4951	0,4952
2,6	0,4953	0,4955	0,4956	0,4957	0,4959	0,4960	0,4961	0,4962	0,4963	0,4964
2,7	0,4965	0,4966	0,4967	0,4968	0,4969	0,4970	0,4971	0,4972	0,4973	0,4974
2,8	0,4974	0,4975	0,4976	0,4977	0,4977	0,4978	0,4979	0,4979	0,4980	0,4981
2,9	0,4981	0,4982	0,4982	0,4983	0,4984	0,4984	0,4985	0,4985	0,4986	0,4986
3,0	0,4987	0,4987	0,4987	0,4988	0,4988	0,4989	0,4989	0,4989	0,4990	0,4990
3,1	0,4990	0,4991	0,4991	0,4991	0,4992	0,4992	0,4992	0,4992	0,4993	0,4993
3,2	0,4993	0,4993	0,4994	0,4994	0,4994	0,4994	0,4994	0,4995	0,4995	0,4995
3,3	0,4995	0,4995	0,4995	0,4996	0,4996	0,4996	0,4996	0,4996	0,4996	0,4997
3,4	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4998
3,5	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998
3,6	0,4998	0,4998	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999
3,7	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999
3,8	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999
3,9	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000

Tabela II. Limites da distribuição t de Student.



Graus de Liberdade (ν)	**Limites bilaterais**: $P(\|t\| > t_{\alpha/2})$							
	Nível de Significância (α)							
	0,50	0,20	0,10	0,05	0,025	0,02	0,01	0,005
1	1,000	3,078	6,314	12,706	25,542	31,821	63,657	127,320
2	0,816	1,886	2,920	4,303	6,205	6,965	9,925	14,089
3	0,715	1,638	2,353	3,183	4,177	4,541	5,841	7,453
4	0,741	1,533	2,132	2,776	3,495	3,747	4,604	5,598
5	0,727	1,476	2,015	2,571	3,163	3,365	4,032	4,773
6	0,718	1,440	1,943	2,447	2,969	3,143	3,707	4,317
7	0,711	1,415	1,895	2,365	2,841	2,998	3,500	4,029
8	0,706	1,397	1,860	2,306	2,752	2,896	3,355	3,833
9	0,703	1,383	1,833	2,262	2,685	2,821	3,250	3,690
10	0,700	1,372	1,813	2,228	2,634	2,764	3,169	3,581
11	0,697	1,363	1,796	2,201	2,503	2,718	3,106	3,497
12	0,695	1,356	1,782	2,179	2,560	2,681	3,055	3,428
13	0,694	1,350	1,771	2,160	2,533	2,650	3,012	3,373
14	0,692	1,345	1,761	2,145	2,510	2,624	2,977	3,326
15	0,691	1,341	1,753	2,132	2,490	2,602	2,947	3,286
16	0,690	1,337	1,746	2,120	2,473	2,583	2,921	3,252
17	0,689	1,333	1,740	2,110	2,458	2,567	2,898	3,223
18	0,688	1,330	1,734	2,101	2,445	2,552	2,878	3,197
19	0,688	1,328	1,729	2,093	2,433	2,539	2,861	3,174
20	0,687	1,325	1,725	2,086	2,423	2,528	2,845	3,153
21	0,686	1,323	1,721	2,080	2,414	2,518	2,831	3,135
22	0,686	1,321	1,717	2,074	2,406	2,508	2,819	3,119
23	0,685	1,319	1,714	2,069	2,398	2,500	2,807	3,104
24	0,685	1,318	1,711	2,064	2,391	2,492	2,797	3,091
25	0,684	1,316	1,708	2,060	2,385	2,485	2,787	3,078
26	0,684	1,315	1,706	2,056	2,379	2,479	2,779	3,067
27	0,684	1,314	1,703	2,052	2,373	2,473	2,771	3,057
28	0,683	1,313	1,701	2,048	2,369	2,467	2,763	3,047
29	0,683	1,311	1,699	2,045	2,364	2,462	2,756	3,038
30	0,683	1,310	1,697	2,042	2,360	2,457	2,750	3,030
40	0,681	1,303	1,684	2,021	2,329	2,423	2,705	2,971
60	0,679	1,296	1,671	2,000	2,299	2,390	2,660	2,915
120	0,677	1,289	1,658	1,980	2,270	2,358	2,617	2,860
...	0,674	1,282	1,645	1,960	2,241	2,326	2,576	2,807
	0,25	0,10	0,05	0,025	0,0125	0,01	0,005	0,0025
Graus de Liberdade (ν)	Nível de Significância (α)							
	Limites unilaterais: $P(\|t\| > t_{\alpha})$							

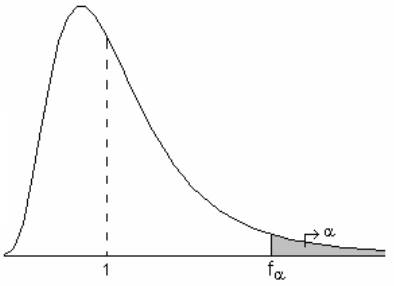
Tabela III. Limites unilaterais da distribuição qui-quadrado (χ^2).



Graus de Liberdade	Nível de significância (α)									
	Esquerda (q')					Direita (q)				
(ν)	0,005	0,01	0,025	0,05	0,1	0,1	0,05	0,025	0,01	0,005
1	0,00	0,00	0,00	0,00	0,02	2,71	3,84	5,02	6,63	7,88
2	0,01	0,02	0,05	0,10	0,21	4,61	5,99	7,38	9,21	10,60
3	0,07	0,11	0,22	0,35	0,58	6,25	7,81	9,35	11,34	12,84
4	0,21	0,30	0,48	0,71	1,06	7,78	9,49	11,14	13,28	14,86
5	0,41	0,55	0,83	1,15	1,61	9,24	11,07	12,83	15,09	16,75
6	0,68	0,87	1,24	1,64	2,20	10,64	12,59	14,45	16,81	18,55
7	0,99	1,24	1,69	2,17	2,83	12,02	14,07	16,01	18,48	20,28
8	1,34	1,65	2,18	2,73	3,49	13,36	15,51	17,53	20,09	21,95
9	1,73	2,09	2,70	3,33	4,17	14,68	16,92	19,02	21,67	23,59
10	2,16	2,56	3,25	3,94	4,87	15,99	18,31	20,48	23,21	25,19
11	2,60	3,05	3,82	4,57	5,58	17,28	19,68	21,92	24,72	26,76
12	3,07	3,57	4,40	5,23	6,30	18,55	21,03	23,34	26,22	28,30
13	3,57	4,11	5,01	5,89	7,04	19,81	22,36	24,74	27,69	29,82
14	4,07	4,66	5,63	6,57	7,79	21,06	23,68	26,12	29,14	31,32
15	4,60	5,23	6,26	7,26	8,55	22,31	25,00	27,49	30,58	32,80
16	5,14	5,81	6,91	7,96	9,31	23,54	26,30	28,85	32,00	34,27
17	5,70	6,41	7,56	8,67	10,09	24,77	27,59	30,19	33,41	35,72
18	6,26	7,01	8,23	9,39	10,86	25,99	28,87	31,53	34,81	37,16
19	6,84	7,63	8,91	10,12	11,65	27,20	30,14	32,85	36,19	38,58
20	7,43	8,26	9,59	10,85	12,44	28,41	31,41	34,17	37,57	40,00
21	8,03	8,90	10,28	11,59	13,24	29,62	32,67	35,48	38,93	41,40
22	8,64	9,54	10,98	12,34	14,04	30,81	33,92	36,78	40,29	42,80
23	9,26	10,20	11,69	13,09	14,85	32,01	35,17	38,08	41,64	44,18
24	9,89	10,86	12,40	13,85	15,66	33,20	36,42	39,36	42,98	45,56
25	10,52	11,52	13,12	14,61	16,47	34,38	37,65	40,65	44,31	46,93
26	11,16	12,20	13,84	15,38	17,29	35,56	38,89	41,92	45,64	48,29
27	11,81	12,88	14,57	16,15	18,11	36,74	40,11	43,19	46,96	49,64
28	12,46	13,56	15,31	16,93	18,94	37,92	41,34	44,46	48,28	50,99
29	13,12	14,26	16,05	17,71	19,77	39,09	42,56	45,72	49,59	52,34
30	13,79	14,95	16,79	18,49	20,60	40,26	43,77	46,98	50,89	53,67
40	20,71	22,16	24,43	26,51	29,05	51,81	55,76	59,34	63,69	66,77
50	27,99	29,71	32,36	34,76	37,69	63,17	67,50	71,42	76,15	79,49
60	35,53	37,48	40,48	43,19	46,46	74,40	79,08	83,30	88,38	91,95
70	43,28	45,44	48,76	51,74	55,33	85,53	90,53	95,02	100,43	104,21
80	51,17	53,54	57,15	60,39	64,28	96,58	101,88	106,63	112,33	116,32
90	59,20	61,75	65,65	69,13	73,29	107,57	113,15	118,14	124,12	128,30
100	67,33	70,06	74,22	77,93	82,36	118,50	124,34	129,56	135,81	140,17

Nota: Se o teste for bilateral, o valor de α deve ser dividido por dois.

Tabela IV. Limites unilaterais superiores da distribuição F: $P[F > f_\alpha]$



		ν_1																			
ν_2	α	1	2	3	4	5	6	7	8	9	10	11	12	15	20	24	30	40	60	120	Inf.
	0,05	161,4	199,5	215,7	224,6	230,2	234,0	236,8	238,9	240,5	241,9	243,0	243,9	245,9	248,0	249,1	250,1	251,1	252,2	253,3	254,3
1	0,025	647,8	799,5	864,2	899,6	921,8	937,1	948,2	956,7	963,3	968,6	976,7	984,9	984,9	993,1	997,2	1001,	1006,	1010,	1014,	1018,
	0,01	4052,	5000,	5403,	5625,	5764,	5859,	5928,	5982,	6022,	6056,	6082,	6106,	6157,	6209,	6235,	6261,	6287,	6313,	6339,	6366,
	0,001	4053*	5000*	5404*	5625*	5764*	5859*	5929*	5981*	6023*	6056*	6084*	6107*	6158*	6209*	6235*	6261*	6287*	6313*	6340*	6366*
	0,05	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40	19,40	19,41	19,43	19,45	19,45	19,46	19,47	19,48	19,49	19,50
2	0,025	38,51	39,00	39,17	39,25	39,30	39,33	39,36	39,37	39,39	39,40	39,41	39,41	39,43	39,45	39,46	39,46	39,47	39,48	39,49	39,50
	0,01	98,50	99,00	99,17	99,25	99,30	99,33	99,36	99,37	99,39	99,40	99,41	99,42	99,43	99,45	99,46	99,47	99,47	99,48	99,49	99,50
	0,001	998,5	999,0	999,2	999,2	999,3	999,3	999,4	999,4	999,4	999,4	999,4	999,4	999,4	999,4	999,5	999,5	999,5	999,5	999,5	999,5
	0,05	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,76	8,74	8,70	8,66	8,64	8,62	8,59	8,57	8,55	8,53
3	0,025	17,44	16,04	15,44	15,10	14,88	14,73	14,62	14,54	14,47	14,42	14,34	14,25	39,43	14,17	14,12	14,08	14,04	13,99	13,95	13,90
	0,01	34,12	30,82	29,46	28,71	28,24	27,91	27,67	27,49	27,35	27,23	27,13	27,05	26,87	26,69	26,60	26,50	26,41	26,32	26,22	26,13
	0,001	167,0	148,5	141,1	137,1	134,6	132,8	131,6	130,6	129,9	129,2	128,8	128,3	127,4	126,4	125,9	125,4	125,0	124,5	124,0	123,5
	0,05	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,93	5,91	5,86	5,80	5,77	5,75	5,72	5,69	5,66	5,63
4	0,025	12,22	10,65	9,98	9,60	9,36	9,20	9,07	8,98	8,90	8,84	8,75	8,66	8,66	8,56	8,51	8,46	8,41	8,36	8,31	8,26
	0,01	21,20	18,00	16,69	15,98	15,52	15,21	14,98	14,80	14,66	14,55	14,45	14,37	14,20	14,02	13,93	13,84	13,75	13,65	13,56	13,46
	0,001	74,14	61,25	56,18	53,44	51,71	50,53	49,66	49,00	48,47	48,05	47,70	47,41	46,76	46,10	45,77	45,43	45,09	44,75	44,40	44,05
	0,05	6,61	5,79	5,41	5,19	5,05	4,95	4,88	8,82	4,77	4,74	4,70	4,68	4,62	4,56	4,53	4,50	4,46	4,43	4,40	4,36
5	0,025	10,01	8,43	7,76	7,39	7,15	6,98	6,85	6,76	6,68	6,62	6,52	6,43	6,46	6,33	6,28	6,23	6,18	6,12	6,07	6,02
	0,01	16,26	13,27	12,06	11,39	10,97	10,67	10,46	10,29	10,16	10,05	9,96	9,89	9,72	9,55	9,47	9,38	9,29	9,20	9,11	9,02
	0,001	47,18	37,12	33,20	31,09	29,75	28,84	28,16	27,64	27,24	26,92	26,64	26,42	25,91	25,39	25,14	24,87	24,60	24,33	24,06	23,79
	0,05	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,03	4,00	3,94	3,87	3,84	3,81	3,77	3,74	3,70	3,67
6	0,025	8,81	7,26	6,60	6,23	5,99	5,82	5,70	5,60	5,52	5,46	5,37	5,27	5,27	5,17	5,12	5,07	5,01	4,96	4,90	4,85
	0,01	13,75	10,92	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87	7,79	7,72	7,56	7,40	7,31	7,23	7,14	7,06	6,97	6,88
	0,001	35,51	27,00	23,70	21,92	20,81	20,03	19,46	19,03	18,69	18,41	18,18	17,99	17,56	17,12	16,89	16,67	16,44	16,21	15,99	15,75
	0,05	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	6,60	3,57	3,51	3,44	3,41	3,38	3,34	3,30	3,27	3,23
7	0,025	8,07	6,54	5,89	5,52	5,29	5,12	4,99	4,90	4,82	4,76	4,67	4,57	3,51	4,47	4,42	4,36	4,31	4,25	4,20	4,14
	0,01	12,25	9,55	8,45	7,85	4,46	7,19	6,99	6,84	6,72	6,62	6,54	6,47	6,31	6,16	6,07	5,99	5,91	5,82	5,74	5,65
	0,001	29,25	21,69	18,77	17,19	16,21	15,52	15,02	14,63	14,33	14,08	13,88	13,71	13,32	12,93	12,73	12,53	12,33	12,12	11,91	11,70
	0,05	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,31	3,28	3,22	3,15	3,12	3,08	3,04	3,01	2,97	2,93
8	0,025	7,57	6,06	5,42	5,05	4,82	4,65	4,53	4,43	4,36	4,30	4,20	4,10	4,10	4,00	3,95	3,89	3,84	3,78	3,73	3,67
	0,01	11,26	8,65	7,59	7,01	6,63	6,37	6,18	6,03	5,91	5,81	5,74	5,67	5,52	5,36	5,28	5,20	5,12	5,03	4,95	4,86
	0,001	25,42	18,49	15,83	14,39	13,49	12,86	12,40	12,04	11,77	11,54	11,35	11,19	10,84	10,48	10,30	10,11	9,92	9,73	9,53	9,33
	0,05	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,10	3,07	3,01	2,94	2,90	2,86	2,83	2,79	2,75	2,71
9	0,025	7,21	5,71	5,08	4,72	4,48	4,32	4,20	4,10	4,03	3,96	3,87	3,77	3,77	3,67	3,61	3,56	3,51	3,45	3,39	3,33
	0,01	10,56	8,02	6,99	6,42	6,06	5,80	5,61	5,47	5,35	5,26	5,18	5,11	4,96	4,81	4,73	4,65	4,57	4,48	4,40	4,31
	0,001	22,86	16,39	13,90	12,56	11,71	11,13	10,70	10,37	10,11	9,89	9,72	9,57	9,24	8,90	8,72	8,55	8,37	8,19	8,00	7,81
	0,05	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,94	2,91	2,85	2,77	2,74	2,70	2,66	2,62	2,58	2,54
10	0,025	6,94	5,46	4,83	4,47	4,24	4,07	3,95	3,85	3,78	3,72	3,62	3,52	3,52	3,42	3,37	3,31	3,26	3,20	3,14	3,08
	0,01	10,04	7,56	6,55	5,99	5,64	5,39	5,20	5,06	4,94	4,85	4,78	4,71	4,56	4,41	4,33	4,25	4,17	4,08	4,00	3,91
	0,001	21,04	14,91	12,55	11,28	10,48	9,92	9,52	9,20	8,96	8,75	8,59	8,45	8,13	7,80	7,64	7,47	7,30	7,12	6,94	6,76
	0,05	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85	2,82	2,79	2,72	2,65	2,61	2,57	2,53	2,49	2,45	2,40
11	0,025	6,72	5,26	4,63	4,28	4,04	3,88	3,76	3,66	3,59	3,53	3,43	3,33	3,33	3,23	3,17	3,12	3,06	3,00	2,94	2,88
	0,01	9,65	7,21	6,22	5,67	5,32	5,07	4,89	4,74	4,63	4,54	4,46	4,40	4,25	4,10	4,02	3,94	3,86	3,78	3,69	3,60
	0,001	19,69	13,81	11,56	10,35	9,58	9,05	8,66	8,35	8,12	7,92	7,76	7,63	7,32	7,01	6,85	6,68	6,52	6,35	6,17	6,00
	0,05	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75	2,72	2,69	2,62	2,54	2,51	2,47	2,43	2,38	2,34	2,30
12	0,025	6,55	5,10	4,47	4,12	3,89	3,73	3,61	3,51	3,44	3,37	3,28	3,18	3,18	3,07	3,02	2,96	2,91	2,85	2,79	2,72
	0,01	9,33	6,93	5,95	5,41	5,06	4,82	4,64	4,50	4,39	4,30	4,22	4,16	4,01	3,86	3,78	3,70	3,62	3,54	3,45	3,36
	0,001	18,64	12,97	10,80	9,63	8,89	9,38	8,00	7,71	7,48	7,29	7,14	7,00	6,71	6,40	6,25	6,09	5,93	5,76	5,59	5,42

* Estes valores devem ser multiplicados por 100.

Continua

Continuação

		ν_1																			
ν_2	α	1	2	3	4	5	6	7	8	9	10	11	12	15	20	24	30	40	60	120	Inf.
	0,05	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67	2,63	2,60	2,53	2,46	2,42	2,38	2,34	2,30	2,25	2,21
13	0,025	6,41	4,97	4,35	4,00	3,77	3,60	3,48	3,39	3,31	3,25	3,15	3,05	3,05	2,95	2,89	2,84	2,78	2,72	2,66	2,60
	0,01	9,07	6,70	5,74	5,21	4,86	4,62	4,44	4,30	4,19	4,10	4,02	3,96	3,82	3,66	3,59	3,51	3,43	3,34	3,25	3,17
	0,001	17,81	12,31	10,21	9,07	8,35	7,86	7,49	7,21	6,98	6,80	6,65	6,52	6,23	5,93	5,78	5,63	5,47	5,30	5,14	4,97
	0,05	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60	2,56	2,53	2,46	2,39	2,35	2,31	2,27	2,22	2,18	2,13
14	0,025	6,30	4,86	4,24	3,89	3,66	3,50	3,38	3,29	3,21	3,15	3,05	2,95	2,95	2,84	2,79	2,73	2,67	2,61	2,55	2,49
	0,01	8,86	6,51	5,56	5,04	4,69	4,46	4,28	4,14	4,03	3,94	3,86	3,80	3,66	3,51	3,43	3,35	3,27	3,18	3,09	3,00
	0,001	17,14	11,78	9,73	8,62	7,92	7,43	7,08	6,80	6,58	6,40	6,26	6,13	5,85	5,56	5,41	5,25	5,10	4,94	4,77	4,60
	0,05	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54	2,51	2,48	2,40	2,33	2,29	2,25	2,20	2,16	2,11	2,07
15	0,025	6,20	4,77	4,15	3,80	3,58	3,41	3,29	3,20	3,12	3,06	2,96	2,86	2,86	2,76	2,70	2,64	2,59	2,52	2,46	2,40
	0,01	8,68	6,36	5,42	4,89	4,56	4,32	4,14	4,00	3,89	3,80	3,73	3,67	3,52	3,37	3,29	3,21	3,13	3,05	2,96	2,87
	0,001	16,59	11,34	9,34	8,25	7,57	7,09	6,74	6,47	6,26	6,08	5,94	5,81	5,54	5,25	5,10	4,95	4,80	4,64	4,47	4,31
	0,05	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,45	2,42	2,35	2,28	2,24	2,19	2,15	2,11	2,06	2,01
16	0,025	6,12	4,69	4,08	3,73	3,50	3,34	3,22	3,12	3,05	2,99	2,89	2,79	2,79	2,68	2,63	2,57	2,51	2,45	2,38	2,32
	0,01	8,53	6,23	5,29	4,77	4,44	4,20	4,03	3,89	3,78	3,69	3,61	3,55	3,41	3,26	3,18	3,10	3,02	2,93	2,84	2,75
	0,001	16,12	10,97	9,00	7,94	7,27	6,81	6,46	6,19	5,98	5,81	5,67	5,55	5,27	4,99	4,85	4,70	4,54	4,39	4,23	4,06
	0,05	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45	2,41	2,38	2,31	2,23	2,19	2,15	2,10	2,06	2,01	1,96
17	0,025	6,04	4,62	4,01	3,66	3,44	3,28	3,16	3,06	2,98	2,92	2,82	2,72	2,72	2,62	2,56	2,50	2,44	2,38	2,32	2,25
	0,01	8,40	6,11	5,18	4,67	4,34	4,10	3,93	3,79	3,68	3,59	3,52	3,46	3,31	3,16	3,08	3,00	2,92	2,83	2,75	2,65
	0,001	15,72	10,66	8,73	7,68	7,02	6,56	6,22	5,96	5,75	5,58	5,44	5,32	5,05	4,78	4,63	4,48	4,33	4,18	4,02	3,85
	0,05	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41	2,37	2,34	2,27	2,19	2,15	2,11	2,06	2,02	1,97	1,92
18	0,025	5,98	4,56	3,95	3,61	3,38	3,22	3,10	3,01	2,93	2,87	2,77	2,67	2,67	2,56	2,50	2,44	2,38	2,32	2,26	2,19
	0,01	8,29	6,01	5,09	4,58	4,25	4,01	3,84	3,71	3,60	3,51	3,44	3,37	3,23	3,08	3,00	2,92	2,84	2,75	2,66	2,57
	0,001	15,38	10,39	8,49	7,46	6,81	6,35	6,02	5,76	5,56	5,39	5,25	5,13	4,87	4,59	4,45	4,30	4,15	4,00	3,84	3,67
	0,05	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38	2,34	2,31	2,23	2,16	2,11	2,07	2,03	1,98	1,93	1,88
19	0,025	5,92	4,51	3,90	3,36	3,33	3,17	3,05	2,96	2,88	2,82	2,72	2,62	2,62	2,51	2,45	2,39	2,33	2,27	2,20	2,13
	0,01	8,18	5,93	5,01	4,50	4,17	3,94	3,77	3,63	3,52	3,43	3,36	3,30	3,15	3,00	2,92	2,84	2,76	2,67	2,58	2,49
	0,001	15,08	10,16	8,28	7,26	6,62	6,18	5,85	5,59	5,39	5,22	5,08	4,97	4,70	4,43	4,29	4,14	3,99	3,84	3,68	3,51
	0,05	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,31	2,28	2,20	2,12	2,08	2,04	1,99	1,95	1,90	1,84
20	0,025	5,87	4,46	3,86	3,51	3,29	3,13	3,01	2,91	2,84	2,77	2,68	2,57	2,57	2,46	2,41	2,35	2,29	2,22	2,16	2,09
	0,01	8,10	5,85	4,94	4,43	4,10	3,87	3,70	3,56	3,46	3,37	3,30	3,23	3,09	2,94	2,86	2,78	2,69	2,61	2,52	2,42
	0,001	14,82	9,95	8,10	7,10	6,46	6,02	5,69	5,44	5,24	5,08	4,94	4,82	4,56	4,29	4,15	4,00	3,86	3,70	3,54	3,38
	0,05	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32	2,28	2,25	2,18	2,10	2,05	2,01	1,96	1,92	1,87	1,81
21	0,025	5,83	4,42	3,82	3,48	3,25	3,09	2,97	2,87	2,80	2,73	2,64	2,53	2,53	2,42	2,37	2,31	2,25	2,18	2,11	2,04
	0,01	8,02	5,78	4,87	4,37	4,04	3,81	3,64	3,51	3,40	3,31	3,24	3,17	3,03	2,88	2,80	2,72	2,64	2,55	2,46	2,36
	0,001	14,59	9,77	7,94	6,95	6,32	5,88	5,56	5,31	5,11	4,95	4,81	4,70	4,44	4,17	4,03	3,88	3,74	3,58	3,42	3,26
	0,05	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30	2,26	2,23	2,15	2,07	2,03	1,98	1,94	1,89	1,84	1,78
22	0,025	5,79	4,38	3,78	3,44	3,22	3,05	2,93	2,84	2,76	2,70	2,60	2,50	2,50	2,39	2,33	2,27	2,21	2,14	2,08	2,00
	0,01	7,95	5,72	4,82	4,31	3,99	3,76	3,59	3,45	3,35	3,26	3,18	3,12	2,98	2,83	2,75	2,67	2,58	2,50	2,40	2,31
	0,001	14,38	9,61	7,80	6,81	6,19	5,76	5,44	5,19	4,99	4,83	4,70	4,58	4,33	4,06	3,92	3,78	3,63	3,48	3,32	3,15
	0,05	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,27	2,24	2,20	2,13	2,05	2,01	1,96	1,91	1,86	1,81	1,76
23	0,025	5,75	4,35	3,75	3,41	3,18	3,02	2,90	2,81	2,73	2,67	2,57	2,47	2,47	2,36	2,30	2,24	2,18	2,11	2,04	1,97
	0,01	7,88	5,66	4,76	4,26	3,94	9,71	3,54	3,41	3,30	3,21	3,14	3,07	2,93	2,78	2,70	2,62	2,54	2,45	2,35	2,26
	0,001	14,19	9,47	7,67	6,69	6,08	5,65	5,33	5,09	4,89	4,73	4,60	4,48	4,23	3,96	3,82	3,68	3,53	3,38	3,22	3,05
	0,05	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25	2,22	2,18	2,11	2,03	1,98	1,94	1,89	1,84	1,79	1,73
24	0,025	5,72	4,32	3,72	3,38	3,15	2,99	2,87	2,78	2,70	2,64	2,54	2,44	2,44	2,33	2,27	2,21	2,15	2,08	2,01	1,94
	0,01	7,82	5,61	4,72	4,22	3,90	3,67	3,50	3,36	3,26	3,17	3,09	3,03	2,89	2,74	2,66	2,58	2,49	2,40	2,31	2,21
	0,001	14,03	9,34	7,55	6,59	5,98	5,55	5,23	4,99	4,80	4,64	4,51	4,39	4,14	3,87	3,74	3,59	3,45	3,29	3,14	2,97
	0,05	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24	2,20	2,16	2,09	2,01	1,96	1,92	1,87	1,82	1,77	1,71
25	0,025	5,69	4,29	3,69	3,35	3,13	2,97	2,85	2,75	2,68	2,61	2,51	2,41	2,41	2,30	2,24	2,18	2,12	2,05	1,98	1,91
	0,01	7,77	5,57	4,68	4,18	3,85	3,63	3,46	3,32	3,22	3,13	3,05	2,99	2,85	2,70	2,62	2,54	2,45	2,36	2,27	2,17
	0,001	13,88	9,22	7,45	6,49	5,88	5,46	5,15	4,91	4,71	4,56	4,42	4,31	4,06	3,79	3,66	3,52	3,37	3,22	3,06	2,89
	0,05	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22	2,18	2,15	2,07	1,99	1,95	1,90	1,85	1,80	1,75	1,69
26	0,025	5,66	4,27	3,67	3,33	3,10	2,94	2,82	2,73	2,65	2,59	2,49	2,39	2,39	2,28	2,22	2,16	2,09	2,03	1,95	1,88
	0,01	7,72	5,53	4,64	4,14	3,82	3,59	3,42	3,29	3,18	3,09	3,02	2,96	2,81	2,66	2,58	2,50	2,42	2,33	2,23	2,13
	0,001	13,74	9,12	7,36	6,41	5,80	5,38	5,07	4,83	4,64	4,48	4,35	4,24	3,99	3,72	3,59	3,44	3,30	3,15	2,99	2,82
	0,05	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31	2,25	2,20	2,16	2,13	2,06	1,97	1,93	1,88	1,84	1,79	1,73	1,67
27	0,025	5,63	4,24	3,65	3,31	3,08	2,92	2,80	2,71	2,63	2,57	2,47	2,36	2,36	2,25	2,19	2,13	2,07	2,00	1,93	1,85
	0,01	7,68	5,49	4,60	4,11	3,78	3,56	3,39	3,26	3,15	3,06	2,98	2,93	2,78	2,63	2,55	2,47	2,38	2,29	2,20	2,10
	0,001	13,61	9,02	7,27	6,33	5,73	5,31	5,00	4,76	4,57	4,41	4,28	4,17	3,92	3,66	3,52	3,38	3,23	3,08	2,92	2,75

Continua

Continuação

		ν_1																			
ν_2	α	1	2	3	4	5	6	7	8	9	10	11	12	15	20	24	30	40	60	120	Inf.
	0,05	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19	2,15	2,12	2,04	1,96	1,91	1,87	1,82	1,77	1,71	1,65
28	0,025	5,61	4,22	3,63	3,29	3,06	2,90	2,78	2,69	2,61	2,55	2,45	2,34	2,34	2,23	2,17	2,11	2,05	1,98	1,91	1,83
	0,01	7,64	5,45	4,57	4,07	3,75	3,53	3,36	3,23	3,12	3,03	2,95	2,90	2,75	2,60	2,52	2,44	2,35	2,26	2,17	2,06
	0,001	13,50	8,93	7,19	6,25	5,66	5,24	4,93	4,69	4,50	4,35	4,22	4,11	3,86	3,60	3,46	3,32	3,18	3,02	2,86	2,69
	0,05	4,18	3,33	2,93	2,70	2,55	2,43	2,35	2,28	2,22	2,18	2,14	2,10	2,03	1,94	1,90	1,85	1,81	1,75	1,70	1,64
29	0,025	5,59	4,20	3,61	3,27	3,04	2,88	2,76	2,67	2,59	2,53	2,43	2,32	2,32	2,21	2,15	2,09	2,03	1,96	1,89	1,81
	0,01	7,60	5,42	4,54	4,04	3,73	3,50	3,33	3,20	3,09	3,00	2,92	2,87	2,73	2,57	2,49	2,41	2,33	2,23	2,14	2,03
	0,001	13,39	8,85	7,12	6,19	5,59	5,18	4,87	4,64	4,45	4,29	4,16	4,05	3,80	3,54	3,41	3,27	3,12	2,97	2,81	2,64
	0,05	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16	2,12	2,09	2,01	1,93	1,89	1,84	1,79	1,74	1,68	1,62
30	0,025	5,57	4,18	3,59	3,25	3,03	2,87	2,75	2,65	2,57	2,51	2,41	2,31	2,31	2,20	2,14	2,07	2,01	1,94	1,87	1,79
	0,01	7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17	3,07	2,98	2,90	2,84	2,70	2,55	2,47	2,39	2,30	2,21	2,11	2,01
	0,001	13,29	8,77	7,05	6,12	5,53	5,12	4,82	4,58	4,39	4,24	4,11	4,00	3,75	3,49	3,36	3,22	3,07	2,92	2,76	2,59
	0,05	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08	2,04	2,00	1,92	1,84	1,79	1,74	1,69	1,64	1,58	1,51
40	0,025	5,42	4,05	3,46	3,13	2,90	2,74	2,62	2,53	2,45	2,39	2,29	2,18	2,18	2,07	2,01	1,94	1,88	1,80	1,72	1,64
	0,01	7,31	5,18	4,31	3,83	3,51	3,29	3,12	2,99	2,89	2,80	2,73	2,66	2,52	2,37	2,29	2,20	2,11	2,02	1,92	1,80
	0,001	12,61	8,25	6,60	5,70	5,13	4,73	4,44	4,21	4,02	3,87	3,75	3,64	3,40	3,15	3,01	2,87	2,73	2,57	2,41	2,23
	0,05	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99	1,95	1,92	1,84	1,75	1,70	1,65	1,59	1,53	1,47	1,39
60	0,025	5,29	3,93	3,34	3,01	2,79	2,63	2,51	2,41	2,33	2,27	2,17	2,06	2,06	1,94	1,88	1,82	1,74	1,67	1,58	1,48
	0,01	7,08	4,98	4,13	3,65	3,34	3,12	2,95	2,82	2,72	2,63	2,56	2,50	2,35	2,20	2,12	2,03	1,94	1,84	1,73	1,60
	0,001	11,97	7,76	6,17	5,31	4,76	4,37	4,09	3,87	3,69	3,54	3,42	3,31	3,08	2,83	2,69	2,55	2,41	2,25	2,08	1,89
	0,05	3,92	3,07	2,68	2,45	2,29	2,17	2,09	2,02	1,96	1,91	1,86	1,83	1,75	1,66	1,61	1,55	1,50	1,43	1,35	1,25
120	0,025	5,15	3,80	3,23	2,89	2,67	2,52	2,39	2,30	2,22	2,16	2,05	1,94	1,94	1,82	1,76	1,69	1,61	1,53	1,43	1,31
	0,01	6,85	4,79	3,95	3,48	3,17	2,96	2,79	2,66	2,56	2,47	2,40	2,34	2,19	2,03	1,95	1,86	1,76	1,66	1,53	1,38
	0,001	11,38	7,32	5,79	4,95	4,42	4,04	3,77	3,55	3,38	3,24	3,12	3,02	2,78	2,53	2,40	2,26	2,11	1,95	1,76	1,54
	0,05	3,84	3,00	2,60	2,37	2,21	2,10	2,01	1,94	1,88	1,83	1,79	1,75	1,67	1,57	1,52	1,46	1,39	1,32	1,22	1,00
Inf.	0,025	5,02	3,69	3,12	2,79	2,57	2,41	2,29	2,19	2,11	2,05	1,94	1,83	1,83	1,71	1,64	1,57	1,48	1,39	1,27	1,00
	0,01	6,63	4,61	3,78	3,32	3,02	2,80	2,64	2,51	2,41	2,32	2,24	2,18	2,04	1,88	1,79	1,70	1,59	1,47	1,32	1,00
	0,001	10,83	6,91	5,42	4,62	4,10	3,74	3,47	3,27	3,10	2,96	2,84	2,74	2,51	2,27	2,13	1,99	1,84	1,66	1,45	1,00

Fonte: Silva, 2000.

6. Lista de respostas dos exercícios propostos

Unidade II - Estatística Descritiva

2.1

j	Classe	F_j	F_j'	f_j	f_j'
1	0	20	20	0,50	0,50
2	1	7	27	0,18	0,68
3	2	7	34	0,18	0,85
4	3	3	37	0,08	0,93
5	4	2	39	0,05	0,98
6	5	0	39	0,00	0,98
7	6	0	39	0,00	0,98
8	7	1	40	0,03	1,00
	Total	40	-	1	-

2.2. k = 7 i = 12,89

j	Classe	F_j	F_j'	f_j	f_j'	c_j
1	3,11|— 16,00	8	8	0,16	0,16	9,555
2	16,00|— 28,89	20	28	0,4	0,56	22,445
3	28,89|— 41,78	6	34	0,12	0,68	35,335
4	41,78|— 54,67	8	42	0,16	0,84	48,225
5	54,67|— 67,56	3	45	0,06	0,9	61,115
6	67,56|— 80,45	1	46	0,02	0,92	74,005
7	80,45|—|93,34	4	50	0,08	1	86,895
	Σ	50	-	1	-	-

2.3. Gráfico

2.4. r = -0,7732

2.5. $\bar{x}$ = 9,75 Mo = 9 Md = 9,5 a_t = 8 s^2 = 6,79 s = 2,61 CV = 26,72%

2.6. Q_1 = 58 Q_2 = 67 Q_3 = 70 a_q = 12

2.7. $\bar{x}$ = 1,075 Mo = 0 Md = 0,5 s^2 = 1,87 s = 1,31 CV = 127,07%
m_2 = 1,819 m_3 = 2,815 m_4 = 11,419 a_3 = 1,147 a_4 = 3,45
Classificação: assimétrica positiva e leptocúrtica.

2.8. $\bar{x}$ = 34,56 Mo = 16,00 |-- 28,89 Md = 16,00 |-- 28,89 s^2 = 491,06 s = 22,16
CV = 64,12% m_2 =481,24 m_3 =11011,7 m_4 =755077 a_3 =1,043 a_4 =3,260
Classificação: assimétrica positiva e leptocúrtica.

2.9.
a) EI = 3,1 Q_1 = 19,27 Md = 27,86 Q_3 = 45,4 ES = 93,3
b) Os valores 85,76 ; 86,37 e 93,34 são considerados discrepantes
c) Gráfico
d) Distribuição assimétrica negativa

2.10. Assimetria negativa
6| 32 55 75
7| 18 60 60 83 84
8| 26 31 34 39 42 54 65 65 66 86 88
9| 01 12 19 39 54 61

Unidade III - Elementos de probabilidade

3.1. 0,5303

3.2. 0,3801

3.3. a) 0,6 b) 0,4 c) 0,75 d) 0,25
e) 0,6 f) 0,4 g) 0,75 h) 0,25

3.4. P(A|D) = 0,3623 P(B|D) = 0,4058 P(C|D) = 0,2319

3.5. 0,66

3.6. a) 0,7283 b) 0,2092

3.7. a) 1,20 b) 0,90 c) 2,10 d) 0,30 e) 0,70 f) 0,72
g) 0,63 h) -0,36 i) 0,63 j) 2,07 k) -0,5345

3.8. 0,6836

3.9. 0,0758

3.10. 0,2242

3.11. a) E(X) = 15 V(X) = 8,33
b) 0,4190

3.12. P(X > 50) = 0,5488 P(X > 100) = 0,5488

3.13. R$ 0,25

3.14. a) 0,4582 b) 0,2090 c) 0,1587

3.15. a) 0,6078 b) 0,0912 c) 447,6

3.16. nota mínima para A = 73,83
nota máxima para R = 70,33

Unidade IV - Inferência Estatística

4.1. (59.606; 63.378)

4.2.

4.3. a)

b) (-0,785; -0,115)

4.4. (;)

4.5. (0,0695; 0,1465)

4.6. a) $t_{0,05} = 2,776 \Rightarrow H_0$ não é rejeitada
b) $t_{0,025} = 3,495 \Rightarrow H_0$ não é rejeitada
c) $t_{0,025} = 3,495 \Rightarrow H_0$ não é rejeitada
d) A variável em estudo tem distribuição normal

4.7. a) 5% = P (erro Tipo I)
b) 0,0975
c) 0,8063

4.8. a) 0,6513
b) 0,6126

4.9. $t_{0,025}$ =
t_c = (teste bilateral)
Não se rejeita H_0.

4.10. $t_{0,005}$ =
z_c = (teste bilateral)
Não se rejeita H_0.

4.11. $f_{0,05}$ =
f_c = (teste bilateral)
Não se rejeita H_0.

4.12. $z_{0,05}$ = -1,645
z_c = -0,5774 (teste unilateral)
Não se rejeita H_0.