Instituto Federal do Norte de Minas Gerais (IFNMG) - Campus Januária

Curso de Bacharelado em Sistemas de Informação

Notas de Aula

Estatística e probabilidade

Prof. Gustavo Pereira Gomes
E-mail: gustavo.gomes@ifnmg.edu.br

Januária
2023

Sumário

Capítulo 1

Introdução

A noção de "Estatística" foi originalmente derivada da mesma raiz da palavra "Estado", já que foi a função tradicional de governos centrais no sentido de armazenar registros da população, nascimentos e mortes, produção das lavouras, taxas e muitas outras espécies de informação e atividades. A contagem e mensuração dessas quantidades gera todos os tipos de dados numéricos que são úteis para o desenvolvimento de muitos tipos de funções governamentais e formulação de políticas públicas.

Estatística é uma ciência definida como o conjunto de técnicas que permite, de forma sistemática, organizar, descrever, analisar e interpretar dados oriundos de estudos ou experimento, realizados em qualquer área do conhecimento. Entende-se por dados um (ou mais) conjunto de valores, numéricos ou não. A grosso modo pode-se dividir a Estatística em três grandes áreas: Estatística Descritiva, Probabilidade e Inferência Estatística.

A Estatística Descritiva pode ser definida como um conjunto de técnicas destinadas a descrever e resumir os dados, afim de que possamos tirar informações e conclusões a respeito de características de interesse. Já a Probabilidade é a base matemática sob a qual a Estatística é construída. Fornece métodos para quantificar a incerteza existente em determinada situação, usando ora um número ora uma função matemática. Por fim, a Inferência Estatística é o estudo de técnicas que possibilitam a realização de conclusões à respeito de uma população a partir do estudo de amostras, tendo por base o cálculo das probabilidades.

1.1 Conceitos básicos

1.1.1 População e amostra

Definição 1.1.1. População é o conjunto de todos os elementos relativos a um determinado fenômeno que possuem pelo menos uma característica em comum, podendo ser finita ou infinita.

(a) Finita: apresenta um número limitado de elementos, dessa forma, é possível enumerar todos os elementos componentes;

(b) Infinita: apresenta um número ilimitado de elementos, o que torna impossível a enumeração de todos os elementos componentes. Entretanto, tal definição existe apenas no campo teórico, porque na prática, nunca encontraremos populações com infinitos elementos, mas somente, populações com grande número de componentes. Nessas circunstâncias, tais populações são tratadas como se fossem infinitas.

Exemplo 1.1.1. (a) As indústrias localizadas na cidade de Montes Claros é uma população finita.

(b) As aves que contraíram a febre aviária é um exemplo de população infinita.

Definição 1.1.2. Amostra é um conjunto de elementos extraídos da população.

Exemplo 1.1.2. Os estudantes do IFNMG constituem uma população com uma característica em comum: matriculados em um determinado curso do IFNMG. Os alunos do primeiro período do IFNMG do curso de Administração noturno é uma amostra dessa população.

$$\begin{array}{ccccc} \text{População} & \rightarrow & \text{Amostra} & \rightarrow & \text{Estatística Descritiva} \\ \uparrow & & & & \downarrow \\ \text{Inferência Estatística} & \leftarrow & \text{Erro} & \leftarrow & \text{Probabilidade} \end{array}$$

1.1.2 Censo × Amostragem

Na pesquisa estatística a forma de coleta dos dados pode ser feita através de censo ou amostragem.

Definição 1.1.3. Censo é o exame completo de toda população e a amostragem é o estudo da amostra de indivíduos de uma determinada população.

Observação 1.1.1. Vantagens da pesquisa por amostragem em relação ao censo:

(a) é mais barata;

(b) é mais rápida;

(c) resultados próximos ao censo;

(d) é mais fácil de ser controlada por envolver operações menores.

Desvantagens da pesquisa por amostragem em relação ao censo:

(a) o censo pode ser mais vantajoso quando a população é pequena e/ou as informações são de fácil obtenção.

(b) os resultados da pesquisa por amostragem carregam erro;

(c) se a população for muito heterogênea o erro pode ser muito grande (e a precisão muita baixa), nesse caso pode ser necessária uma amostra muito grande.

O pesquisador busca generalizar conclusões referentes à amostra, estendendo-as para toda a população da qual essa amostra foi extraída. Existem técnicas adequadas para recolher amostras, de forma a garantir (tanto quanto possível) o sucesso da pesquisa e dos resultados. Os tipos de amostragem mais comuns são:

(a) Amostragem aleatória simples: este tipo de amostragem é baseado no sorteio da amostra. Numera-se a população de 1 a n e, utilizando um dispositivo aleatório qualquer, por exemplo sorteio (geralmente sem reposição), escolhem-se k números dessa sequência, os quais corresponderão aos elementos da amostra.

Exemplo 1.1.3. Imagine que você queira amostrar um número de pessoas que estão fazendo um determinado concurso com n inscritos. Devemos enumerar cada um dos n candidatos e sortear k deles.

(b) Amostragem proporcional estratificada: quando as populações é muito heterogênea (quando as características observadas variam muito de um indivíduo para outro) é aconselhável subdividir a população em subgrupos (estratos) homogêneos, pois pode ser razoável supor que a variável de interesse apresente comportamento distinto nos diferentes estratos. Assim, para que uma amostra seja representativa, é necessário utilizar-se uma amostragem proporcional estratificada, que considera os estratos (subgrupos) e obtém a amostragem proporcional a estes.

Exemplo 1.1.4. Considere como população os professores de determinada universidade. Queremos determinar uma amostra de tamanho 48 para realizar uma pesquisa sobre o gosto musical. Entretanto dividimos a população em estratos por idade: (1) de 20 anos à 30 anos; (2) de 31 à 41; (3) acima de 41. Se os estratos possuem 40, 80 e 120 professores, respectivamente, então utilizando amostragem proporcional estratificada, devemos tomar 8, 16 e 24 professores, respectivamente, de cada estrato.

(c) Amostragem estratificada uniforme: não utiliza o critério de proporcionalidade, pois se seleciona a mesma quantidade de elementos de cada estrato, devendo ser usada para comparar os estratos ou obter estimativas separadas para cada estrato.

Exemplo 1.1.5. Uma empresa de automação conta com 480 funcionários, dos quais 288 são do sexo feminino e os 192 restantes do sexo masculino. Considerando a variável "sexo" para estratificar essa população, vamos obter uma amostra estratificada uniforme de 50 funcionários. Supondo que haja homogeneidade dentro de cada categoria, pode-se obter amostra estratificada uniforme de 50 funcionários com a seleção de 25 elementos de cada estrato.

(d) Amostragem sistemática: é um procedimento para a amostragem aleatória simples aplicada quando os elementos da população já estão ordenados. Assim, não é necessário construir um sistema de referência ou de amostragem. Inicialmente obtém-se o tamanho da população (N), calcula-se o tamanho da amostra (n) e encontra-se o intervalo de retirada $k = N/n$. Em seguida, sorteia-se o ponto de partida e a cada k elementos da população, retira-se um para fazer parte da amostra, até completar o valor de n.

Exemplo 1.1.6. Num estoque de 100 peças numeradas, para obtermos 13 amostras sistemáticas podemos retirar as peças de número 2, 10, 18, 26, ... , 98 ($N = 100$, $n = 13$, $k = 7,7 \approx 8$ e o ponto de partida escolhido foi o 2).

(e) Amostragem por meio de conglomerados: é aplicada quando a população apresenta uma subdivisão em pequenos grupos, chamados conglomerados. É possível e, muitas vezes, conveniente fazer a amostragem por meio desses conglomerados, cujos elementos constituirão a amostra, ou seja, as unidades de amostragem sobre as quais é feito o sorteio, passam a ser conglomerados e não mais elementos individuais da população. A amostragem por meio de conglomerados é adotada por motivos de ordem prática e econômica.

Exemplo 1.1.7. Quarteirões de um bairro.

(f) Amostragem intencional: a amostra pesquisada muitas vezes está disponível no local e no momento onde a pesquisa está sendo realizada. Com base em seu julgamento, o pesquisador seleciona os elementos que julga mais representativos da população.

Exemplo 1.1.8. Para saber a preferência por determinado cosmético, o pesquisador entrevista os frequentadores de um grande salão de beleza.

Observação 1.1.2. Não há dúvida de que uma amostra não representa perfeitamente uma população. Ou seja, a utilização de uma amostra implica na aceitação de uma margem de erro que denominaremos erro amostral, porém podemos limitar seu valor através da escolha de uma amostra de tamanho adequado (não vamos estudar como determinar o tamanho de uma amostra, pois para isso, é necessário ter uma bagagem maior dos conceitos estatísticos).

Exercícios

1. Um estudo sobre o desempenho dos vendedores de uma grande cadeia de lojas de varejo está sendo planejado. Para tanto, deve ser colhida uma amostra probabilística dos vendedores. Classifique cada uma das amostras abaixo conforme a seguinte tipologia:

   (A) Amostragem aleatória simples

   (B) Amostragem sistemática

   (C) Amostragem estratificada

   (D) Amostragem por meio de conglomerados

   (a) ( ) Lista de todos os vendedores (que atuam em todas as lojas da rede). Selecionei todos vendedores que ocupavam posições múltiplas de 15 (15ª posição, 30ª posição, 45ª posição, 60ª posição, 75ª posição, 90ª posição, 105ª posição, etc)

   (b) ( ) Escolhi casualmente 3 lojas da rede. A amostra foi composta de todos os vendedores que atuam em cada uma destas 3 lojas.

   (c) ( ) Em cada uma das lojas, identifiquei todos os vendedores (lista de vendedores por loja). Selecionei aleatoriamente k vendedores da loja, onde k é um número inteiro proporcional à quantidade de vendedores da loja.

   (d) ( ) Lista de todos os vendedores (que atuam em todas as lojas da rede). Selecionei aleatoriamente n vendedores.

1.1.3 Tipos de variáveis estatísticas

Definição 1.1.4. Variável é uma característica de interesse do estudo. Existem dois tipos de variáveis:

(a) Qualitativas: descrevem qualidades (rótulos ou classes). Elas podem ser:

   (i) Ordinais: variáveis que têm uma ordenação natural, indicando intensidades crescentes de realização;

   (ii) Nominais: variáveis em que não é possível estabelecer uma ordem natural entre seus valores.

(b) Quantitativas: indicam a quantidade de algo e é sempre numérica. Elas podem ser:

   (i) Discretas: variáveis resultantes de contagens, assumindo assim, em geral, valores inteiros não negativos;

   (ii) Contínuas: assumem valores em intervalos dos números reais e, geralmente, são provenientes de uma mensuração.

Exemplo 1.1.9. Variáveis tais como turma (A ou B), sexo (feminino ou masculino), profissão, região geográfica são variáveis qualitativas nominais. Por outro lado, variáveis como tamanho (pequeno, médio ou grande), classe social (baixa, média ou alta), faixa de idade (criança, adolescente, adulto e idoso) são variáveis qualitativas ordinais. O número de irmãos $(0, 1, 2, \ldots)$ e o número de defeitos $(0, 1, 2, \ldots)$ são quantitativas discretas, enquanto que peso e altura são quantitativas contínuas.

Exercícios

1. Determine a população e a característica em comum na situação: Deseja-se conhecer o consumo total de energia elétrica em Mwh nas residências da cidade de Santa Maria (RS) no ano de 2022.

2. Classifique as variáveis abaixo:

   (a) Cor dos olhos das alunas.
   (b) Produção de café no Brasil (em toneladas).
   (c) Número de defeitos em aparelhos de TV.
   (d) Comprimento dos pregos produzidos por uma empresa.
   (e) Religião.
   (f) Velocidade de um carro.
   (g) O grau de escolaridade de uma pessoa.

1.2 Fases do método estatístico

Ao desenvolver um estudo estatístico completo, existem algumas fases do seu método que devem ser desenvolvidas em sequência, para chegar aos resultados finais do trabalho. As principais fases do método estatístico são:

(a) Definição do problema: saber exatamente aquilo que se pretende pesquisar.

(b) Planejamento: consiste em planejar o modo como serão realizadas as fases seguintes, determinando o objetivo da pesquisa e os métodos que serão utilizados.

(c) Coleta de dados: fase operacional. É o registro sistemático de dados, com um objetivo determinado. Após a definição do problema a ser estudado e o estabelecimento do planejamento da pesquisa, o passo seguinte é a coleta de dados.

(d) Apuração dos dados: resumo dos dados através de sua contagem e agrupamento. É a condensação e a tabulação de dados.

(e) Apresentação dos dados: há duas formas de apresentação que não se excluem mutuamente. Primeiro, a apresentação tabular, que é uma apresentação numérica dos dados em linhas e colunas distribuídas de modo ordenado, segundo regras práticas fixadas pelo Conselho Nacional de Estatística; e, segundo, a apresentação gráfica dos dados numéricos, que constitui uma apresentação geométrica que permite uma visão rápida e clara do fenômeno.

(f) Análise e interpretação dos dados: a última fase do trabalho estatístico é a mais importante e a mais delicada. Está ligada essencialmente ao cálculo de medidas e coeficientes cuja finalidade principal é descrever o fenômeno.

1.3 Tabela de frequência

A tabela é uma forma muito útil de resumir e organizar a informação observada sobre uma variável de forma precisa. Nessa seção, vamos fazer o estudo detalhado das tabelas de frequências de variáveis quantitativas (discreta ou contínuas), sendo que quando a variável é contínua (ou quando muitos dados possuem frequências baixas), devemos primeiro construir intervalos e depois obter as frequências para cada intervalo (onde ocorre perda de informação), como veremos a seguir. Desta forma, vamos mostrar através de exemplos como esse tipo de tabela é construído.

Exemplo 1.3.1. Os dados abaixo indicam o número de filhos por casal investigado em 20 famílias de uma região.

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

Primeiramente, contamos quantas vezes o dado está repetido. Os valores provenientes desta contagem, denotados por f_i, são denominados frequências absolutas.

nº de filhos	f_i
0	2
1	5
2	2
3	6
4	3
5	2
$\sum$	20

A frequência absoluta acumulada, denotada por F_i, expressa o número de elementos acumulados em cada categoria.

nº de filhos	f_i	F_i
0	2	2
1	5	7
2	2	9
3	6	15
4	3	18
5	2	20
$\sum$	20	

A frequência relativa, denotada por fr_i, expressa a proporção de elementos em cada categoria.

nº de filhos	f_i	F_i	fr_i
0	2	2	0,1
1	5	7	0,25
2	2	9	0,1
3	6	15	0,3
4	3	18	0,15
5	2	20	0,1
$\sum$	20		1,0

Por último, temos a frequência relativa acumulada, denotada por Fr_i, que expressa a proporção de elementos acumulada em cada categoria.

nº de filhos	f_i	F_i	fr_i	Fr_i
0	2	2	0,1	0,1
1	5	7	0,25	0,35
2	2	9	0,1	0,45
3	6	15	0,3	0,75
4	3	18	0,15	0,9
5	2	20	0,1	1,0
$\sum$	20		1,0	

Sendo assim, interprete os dados em destaque:

nº de filhos	f_i	F_i	fr_i	Fr_i
0	2	2	0,1	0,1
1	5	7	0,25	0,35
2	2	9	0,1	0,45
3	6	15	0,3	0,75
4	3	18	0,15	0,9
5	2	20	0,1	1,0
$\sum$	20		1,0	

Exemplo 1.3.2. Suponhamos termos feito uma coleta de dados relativos às idades de 30 alunos que compõe uma amostra dos alunos do curso de BSI do IFNMG/Januária (dado fictício).

24	23	22	28	35	21	23	23	33	34
24	21	25	36	26	22	30	32	25	26
33	34	21	31	25	31	26	25	35	33

Observe que, neste caso, a tabela com as frequências absolutas não é conveniente, pois muitos dados possuem frequência 1.

IDADE	FREQ.
21	3
22	2
23	3
24	2
25	4
26	3
28	1
30	1
31	2
32	1
33	3
34	2
35	2
36	1
$\sum$	30

Logo, o ideal é agrupar os valores em intervalos/classes. Entretanto, ganhamos em simplicidade e perdemos em pormenores.

IDADES (ANOS)			FREQUÊNCIA
21	$\vdash$	24	8
24	$\vdash$	27	9
27	$\vdash$	30	1
30	$\vdash$	33	4
33	$\vdash$	36	7
36	$\vdash$	39	1
TOTAL			30

onde, por exemplo, o intervalo (ou classe) $21 \vdash 24$ é um intervalo fechado à esquerda e aberto à direita ($21 \leq x < 24$). Para construir tabelas de distribuição em classes, devemos seguir o seguinte passo-a-passo:

(i) Calcular a amplitude amostral (AA): diferença entre o valor máximo e mínimo da amostra.

(ii) Escolher o número de classes conveniente (k): podemos utilizar a regra de Sturges ($k = 1 + 3,3 \cdot \log n$, onde n é o número total de dados) ou um outro método (por exemplo, $k = \sqrt{n}$).

(iii) Determinar a amplitude de cada classe: basta dividir a amplitude amostral pelo número de classes.

(iv) Tome o menor dado e adicione a amplitude calculada no item anterior: $21 \vdash 24$.

(v) Repita este processo para obter as classes: $24 \vdash 27$, $27 \vdash 30$, $30 \vdash 33$, $33 \vdash 36$ e $36 \vdash 39$.

(vi) Calcule as frequências f_i, F_i, fr_i e Fr_i e, em seguida, construa a tabela.

Exemplo 1.3.3. 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

Faça a distribuição de frequências desses dados.

Solução:

Peso	f_i	F_i	fr_i	Fr_i
16 ⊢ 19,3	7	7	0,1167	0,1167
19,3 ⊢ 22,6	9	16	0,1500	0,2667
22,6 ⊢ 25,9	15	31	0,2500	0,5167
25,9 ⊢ 29,2	12	43	0,2000	0,7167
29,2 ⊢ 32,5	9	52	0,1500	0,8667
32,5 ⊢ 35,8	6	58	0,1000	0,9667
35,8 ⊢ 39,1	2	60	0,0333	1,0000
Σ	60	–	1,0000	–

Exercícios

1. Considere os dados: $21, 21, 21, 22, 22, 23, 23, 24, 25, 25, 25, 25, 26, 26, 26, 28, 30$. Construa uma tabela de frequências sem intervalos e com intervalos.

2. Os dados abaixo representam o preço, em reais, do produto A vendido em 25 diferentes estabelecimentos.

20,5 19,5 15,6 24,1 9,9 15,4 12,7 5,4 17,0 28,6 16,9 7,8 23,3

11,8 18,4 13,4 14,3 19,2 9,2 16,8 8,8 22,1 20,8 12,6 15,9

   (a) Calcule a amplitude amostral.
   (b) Calcule o número de intervalos utilizando a fórmula $k = \sqrt{n}$.
   (c) Calcule a amplitude dos intervalos.

(d) Construa a tabela de frequências em classes.

3. As informações abaixo indicam o número de acidentes ocorridos com 70 motoristas de uma empresa de ônibus nos últimos 5 anos:

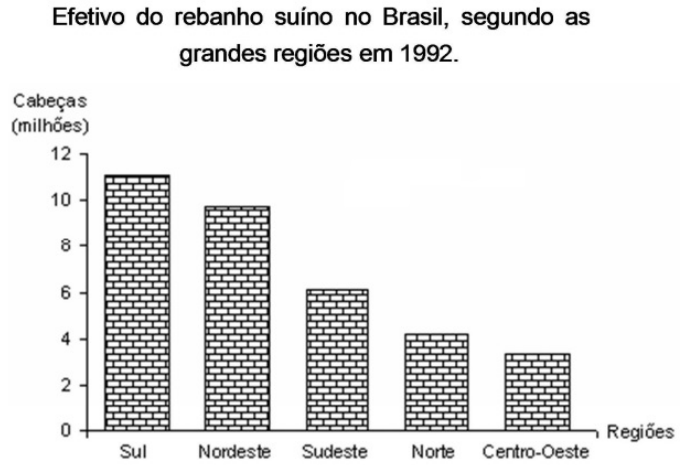
Nº DE ACIDENTES	0	1	2	3	4	5	6	7
Nº DE MOTORISTAS	15	11	20	9	6	5	3	1

(a) Determine o número de motoristas com menos de 1 acidente.
(b) Determine o percentual de motoristas com pelo menos 3 acidentes.
(c) Determine o percentual de motoristas com no máximo 2 acidentes.
(d) Qual o número total de acidentes ocorrido no período?

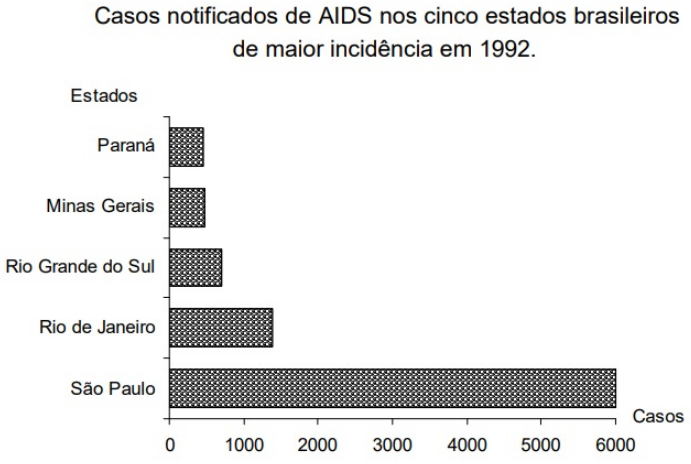
1.4 Representações gráficas

Os gráficos são recursos utilizados pela estatística com o objetivo de apresentar os dados de forma rápida e de fácil compreensão com enfoque no visual.

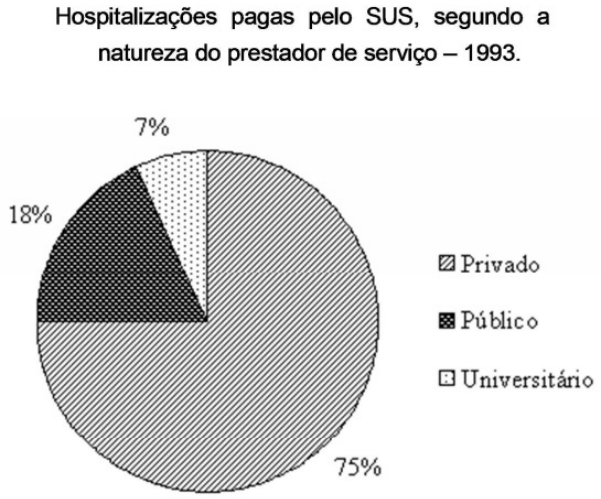
Exemplo 1.4.1. Gráfico de colunas: retângulos verticais com alturas proporcionais às grandezas.



Exemplo 1.4.2. Gráfico de barras: retângulos horizontais, recomendados quando os nomes das categorias são maiores que a base do retângulos.

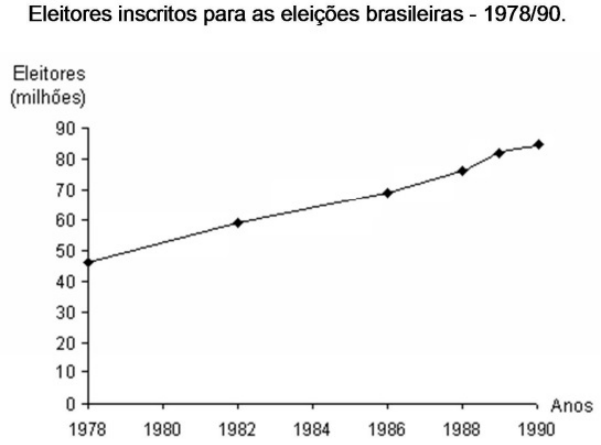


Exemplo 1.4.3. Gráfico de setores: recomendados para situações em que se deseja evidenciar o quanto cada informação representa no total.



Fonte: Anuário Estatístico do Brasil (1994).

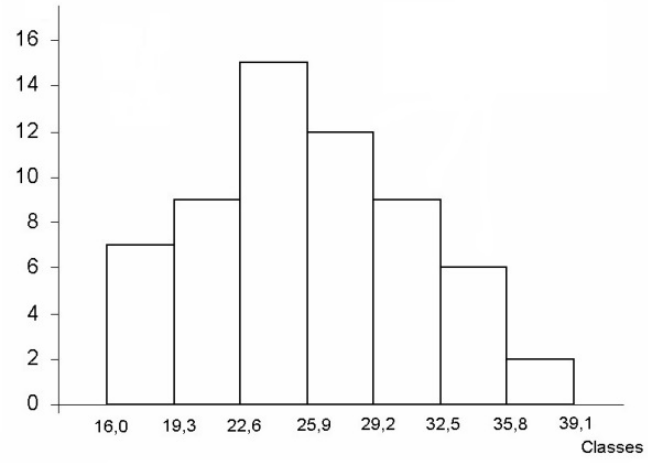
Exemplo 1.4.4. Gráfico de linhas: recomendados para representar conjuntos dedados em que uma das variáveis é contínua, como o tempo, sempre representado no eixo das abscissas.



Fonte: Anuário Estatístico do Brasil (1992).

Exemplo 1.4.5. 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.

Figura - Histograma para o peso ao nascer de 60 bovinos da raça Ibagé.

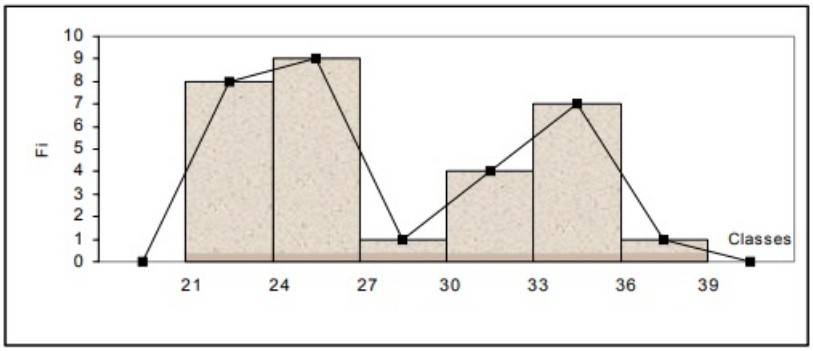


UFPel, 2001.

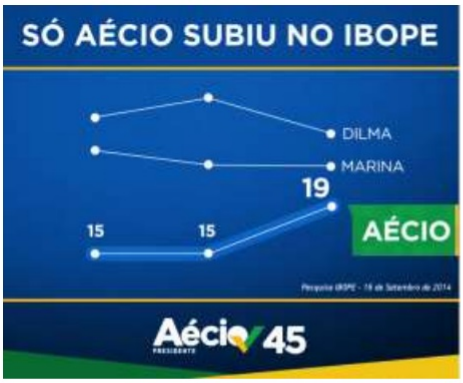
Exemplo 1.4.6. Polígono de frequências: é um gráfico em linha, sendo as frequências marcadas sobre perpendiculares ao eixo horizontal, levando em consideração os pontos médios dos intervalos de classe.

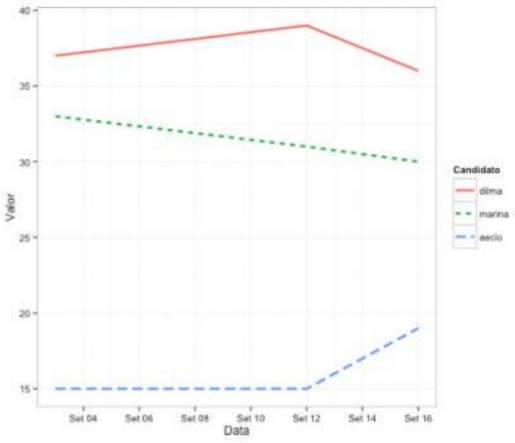
Classe	F_i	x_i	F_{ac}	f_i	$\mathbf{f_{ac}}$
21 |- 24	8	22,5	8	0,267	**0,267**
24 |- 27	9	25,5	17	0,300	**0,567**
27 |- 30	1	28,5	18	0,033	**0,600**
30 |- 33	4	31,5	22	0,133	**0,733**
33 |- 36	7	34,5	29	0,233	**0,966**
36 |- 39	1	37,5	30	0,033	**1,000**
TOTAL	30	-	-	1,000	-

HISTOGRAMA E POLÍGONO DE FREQUÊNCIA SIMPLES DA TABELA ACIMA

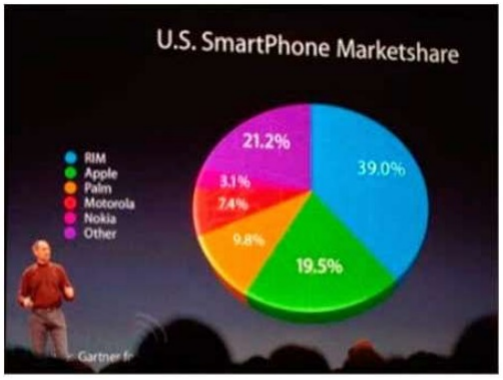


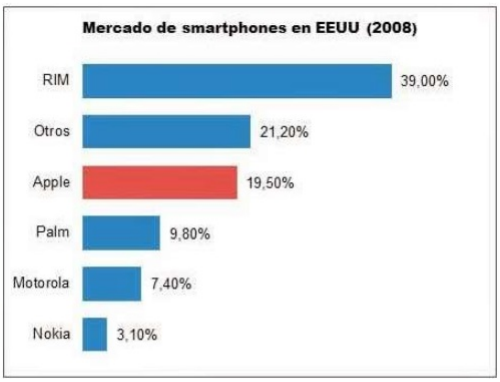
Exemplo 1.4.7. Exemplos de manipulação de gráficos:





Note que a distância de Marina para Aécio (11 pontos percentuais) está menor do que a distância de 15 para 19 (4 pontos percentuais) do próprio Aécio .





Steve Jobs usa um gráfico de pizza em 3D onde a fatia da Apple está voltada para o público (o que faz com que ela pareça maior) e a fatia outros está do lado oposto. Assim pode parecer que a Apple estaria em segundo (em relação à categoria outros). O gráfico de barra 2D desaparece com a ilusão de ótica.

Exercícios

1. Um analista deseja conhecer a evolução da situação financeira de uma empresa. Para tanto, coletou uma série de dados referentes ao lucro e despesa, conforme a tabela abaixo.

Faturamento e despesa de uma empresa de Lajeado – 2009/ 2014

Ano	Faturamento (R$)	Despesa (R$)
2009	230.500,00	195.100,00
2010	297.200,00	252.700,00
2011	361.900,00	340.800,00
2012	455.700,00	405.400,00
2013	560.000,00	497.200,00
2014	610.600,00	560.800,00

Fonte: Dados fictícios, apenas para fins ilustrativos.

   (a) Faça um gráfico de colunas.

   (b) Faça um gráfico de linhas colocando contendo as informações referentes ao faturamento e a despesa no mesmo gráfico.

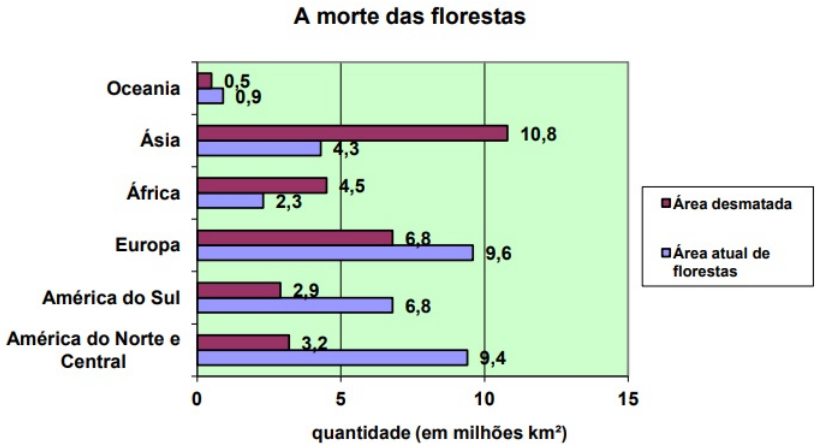
2. A dona de um restaurante registrou durante 6 meses quantos clientes ele recebia a cada semana. Os dados estão abaixo em ordem crescente.

501	512	516	525	528	536	546	556	564
567	589	597	601	603	605	612	615	624
629	635	642	645	648	651			

   (a) Faça a tabela de frequência da distribuição com 6 intervalos de classe.

   (b) Construa o histograma e o polígono de frequência da distribuição.

3. O estado das florestas do planeta e o que foi devastado pela ocupação humana, são os dados que estão representados no gráfico a seguir. Observe estes dados que foram publicados na revista Época de 08/02/1999 e depois responda:



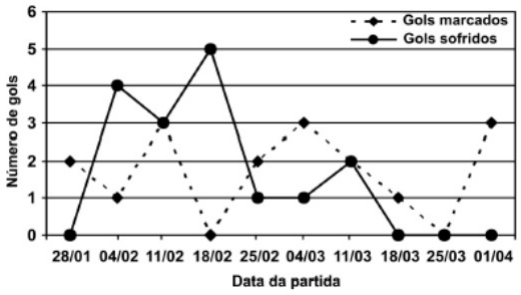
   (a) Em quais continentes mais da metade das florestas foi devastada pela ocupação humana?

   (b) Qual a área atual de florestas no mundo todo?

   (c) Qual a área desmatada no mundo todo?

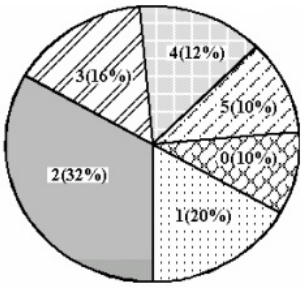
   (d) É verdade que a área desmatada na Ásia é o dobro da área desmatada em todo o continente americano?

4. No gráfico estão representados os gols marcados e os gols sofridos por uma equipe de futebol nas dez primeiras partidas de um determinado campeonato.



Considerando que, neste campeonato, as equipes ganham 3 pontos para cada vitória, 1 ponto por empate e 0 ponto em caso de derrota, a equipe em questão, ao final da décima partida, terá acumulado quantos pontos?

5. O gráfico , em forma de pizza, representa as notas obtidas em uma questão pelos 32.000 candidatos presentes à primeira fase de uma prova de vestibular. Ele mostra, por exemplo, que 32% desses candidatos tiveram nota 2 nessa questão.



(a) Quantos candidatos tiveram nota 3?

(b) É possível afirmar que a nota média, nessa questão, foi menor ou igual 2? Justifique sua resposta.

Capítulo 2

Estatística descritiva

Na maior parte das vezes em que os dados estatísticos são analisados, procuramos obter um valor para representar um conjunto de dados a fim de sintetizar, da melhor maneira possível, o comportamento do conjunto no qual o valor é originário. Com isso, a Estatística Descritiva visa descrever os dados de uma variável quantitativa a fim de resumi-los na forma de valores numéricos. Tais valores são chamados de medidas descritivas e se calculados a partir de dados populacionais, são denominadas parâmetros e se calculados a partir de dados amostrais são denominadas estatísticas. Classificam-se as medidas descritivas como: medidas de posição (tendência central e separatrizes), medidas de dispersão, medidas de assimetria e de curtose.

2.1 Medidas de posição

Já estudamos que é possível sintetizar dados estatísticos utilizando tabelas e gráficos. Mas, também, pode ser de interesse apresentar esses dados através de medidas que sintetizam as informações observadas por valores "representativos" de todo o conjunto. Essas medidas são chamadas de medidas de posição que se dividem em: medidas de tendência central (média, moda e mediana) e medidas separatrizes (quartis, decis e percentis). Na disciplina, vamos abordar somente as medidas de tendência central.

2.1.1 Média aritmética

A média aritmética, simbolizada por $\overline{x}$ (x barra) no caso de amostra e μ (mi) para população, é a medida de tendência central mais utilizada para descrever um conjunto de dados, pois é de fácil cálculo e compreensão.

Definição 2.1.1. A média aritmética é definida como a soma das observações dividida pelo número de observações.

(a) Média aritmética para dados não tabelados: consiste na soma de todas as observações $x_1, \ldots, x_n$ dividida pelo número n de observações de um determinado grupo.

$$\overline{x} = \frac{x_1 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

(b) Média aritmética para dados tabelados: quando os dados estiverem agrupados em uma tabela de frequências, pode-se obter a média aritmética da distribuição fazendo

$$\overline{x} = \frac{x_1 \cdot f_1 + \cdots + x_n \cdot f_n}{n} = \frac{\sum_{i=1}^{n} x_i \cdot f_i}{n},$$

onde os elementos $x_1, \ldots, x_n$ apresentarem as frequências absolutas $f_1, \ldots, f_n$, respectivamente.

Exemplo 2.1.1. Sabe-se que a produção leiteira diária da vaca A, durante uma semana, foi de 10, 14, 13, 15, 18 e 12 litros. Calcule a produção média de leite da semana.

Exemplo 2.1.2. A seguinte tabela de frequência dá as notas de estudantes em determinada escola. Encontre a média aritmética das notas.

notas (x_i)	frequência (f_i)
3	3
4	5
5	6
6	7
7	6

Exemplo 2.1.3. Encontre a média aritmética das estaturas destas 40 pessoas.

Estaturas (cm)	f_i
150 ⊢ 154	4
154 ⊢ 158	9
158 ⊢ 162	11
162 ⊢ 166	8
166 ⊢ 170	5
170 ⊢ 174	3

Definição 2.1.2. O desvio em relação à média (d_i) é a diferença entre cada elemento de um conjunto de valores e a média aritmética, ou seja,

$$d_i = x_i - \overline{x}.$$

Exemplo 2.1.4. Calcule o desvio em relação à média dos termos 10, 14, 13, 15, 18 e 12.

Proposição 2.1.1. *Propriedades da média aritmética:*

1. *Somando-se (ou subtraindo-se) um mesmo valor (uma constante) a cada um dos valores de uma variável, a média da variável fica somada (ou subtraída) a mesma constante.*

2. *Multiplicando-se (ou dividindo-se) um mesmo valor (uma constante) a cada um dos valores de uma variável, a média da variável fica multiplicada (ou dividida) pela mesma constante.*

3. *A soma dos desvios em relação à media é nula.*

Observação 2.1.1. (i) Vantagens: fácil interpretação, sempre existe e é única.

(ii) Desvantagem: a média não será uma medida apropriada para representar os dado com extremos muito discrepantes, pois a média aritmética é afetada por todos os dados observados.

2.1.2 Moda

Definição 2.1.3. Denominamos moda (M_o) o valor que ocorre com maior frequência em um conjunto de dados (séries estatísticas ou, simplesmente, série).

Exemplo 2.1.5. Encontre a moda da série: 7, 8, 9, 10, 10, 11, 12, 13 e 15.

Exemplo 2.1.6. Considerando a tabela de frequência referente ao número de filhos homens de 34 famílias, calcule a moda.

Nº de meninos	Frequência
0	2
1	6
2	10
3	12
4	4
	Σ=34

Como 3 aparece com maior frequência, temos que $M_o = 3$.

Exemplo 2.1.7. Calcule a moda das estaturas.

Estaturas (cm)	f_i
150 ⊢ 154	4
154 ⊢ 158	9
158 ⊢ 162	11
162 ⊢ 166	8
166 ⊢ 170	5
170 ⊢ 174	3

Neste caso, a moda será a média aritmética dos extremos do terceiro intervalo (existem vários outros tipos de moda para dados tabelados, mas nessa disciplina, vamos empregar este método).

Observação 2.1.2. Podemos encontrar séries que não possuem moda (série amodal), por exemplo, 3,5,8,10,12,13. Em outros casos, podem haver mais de uma moda: 2,3,4,4,4,5,6,7,7,7,8,9 (série bimodal).

Observação 2.1.3. (i) Vantagens: não exige cálculo e sempre tem existência real (sempre é representada por um elemento do conjunto de dados).

(ii) Desvantagem: não se presta a cálculos matemáticos, pode não existir, pode não ser única.

2.1.3 Mediana

Definição 2.1.4. A mediana (M_d) é definida como o número que se encontra no centro de uma série estatística dispostos segundo uma ordem.

Exemplo 2.1.8. A mediana do conjunto de dados 5, 13, 10, 2, 18, 15, 6, 16 e 9 (número ímpar de observações) é $M_d = 10$.

Exemplo 2.1.9. A série estatística: 2, 6, 7, 10, 12, 13, 18 e 21 (número par de observações) possui mediana 11 (ponto médio entre os dois valores centrais da série).

Observação 2.1.4. Para dados agrupados em tabelas de frequência, calcula-se $\frac{\sum f_i}{2}$ e verificamos se existe algum elemento com essa frequência acumulada, caso não tenha, a mediana será o elemento cuja frequência absoluta acumulada (F_i) seja exatamente maior que esse valor. Caso exista uma frequência absoluta acumulada em que $F_i = \frac{\sum f_i}{2}$, então a mediana será dada pela média aritmética entre o valor correspondente à essa frequência acumulada e o valor seguinte.

Exemplo 2.1.10. Calcule a mediana dos dados:

(a)

Nº de meninos	Frequência
0	2
1	6
2	10
3	12
4	4
	Σ=34

(b)

Elementos	f_i	F_i
12	1	1
14	2	3
15	1	4
16	2	6
17	1	7
20	1	8
Total	8	

Exemplo 2.1.11. (Calculo de mediana para dados tabulados com intervalos de classes) Calcule a mediana.

Estaturas (cm)	f_i
150 ⊢ 154	4
154 ⊢ 158	9
158 ⊢ 162	11
162 ⊢ 166	8
166 ⊢ 170	5
170 ⊢ 174	3

Inicialmente, fazemos $\frac{\sum f_i}{2}$ e localizamos a classe mediana: i) caso exista uma frequência acumulada igual a $\frac{\sum f_i}{2}$, a mediana será o limite superior da classe correspondente; ii) caso contrário, a classe mediana será a primeira classe que possui frequência acumulada simples maior que $\frac{\sum f_i}{2}$.

Estaturas (cm)	f_i	F_i
150 ⊢ 154	4	4
154 ⊢ 158	9	13
158 ⊢ 162	11	24
162 ⊢ 166	8	32
166 ⊢ 170	5	37
170 ⊢ 174	3	40

Em seguida, aplicamos a fórmula

$$M_d = \ell^* + \frac{\left(\frac{\sum f_i}{2} - F_{ant}\right) \cdot h^*}{f^*},$$

onde

ℓ^*:	limite inferior da classe mediana
F_{ant}:	frequência acumulada da classe anterior à classe mediana
f^*:	frequência simples da classe mediana
h^*:	amplitude do intervalo da classe mediana

Logo,

$$M_d = 158 + \frac{(20-13)\cdot 4}{11} = 160,54.$$

Observação 2.1.5. (i) Vantagens: pode ser determinada mesmo quando não se conhece todos os valores do conjunto de dados, sempre existe e é única, não sofre influência de valores discrepantes.

(ii) Desvantagem: não se presta a cálculos matemáticos.

Exercícios

1. 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: $12, 9, 11, 7, 9, 14, 6, 10$. Para este conjunto de valores, calcule as medidas de posição (média, mediana e moda).

2. Uma escola avalia o seu curso através de um questionário com 50 perguntas sobre diversos aspectos de interesse. Cada pergunta tem uma resposta numa escala de 1 a 5, onde a maior nota significa melhor desempenho. Para cada aluno é então encontrada a nota média. Na última avaliação recorreu-se a uma amostra de 42 alunos, e os resultados estão em baixo.

4.2	2.7	4.6	2.5	3.3	4.7
4.0	2.4	3.9	1.2	4.1	4.0
3.1	2.4	3.8	3.8	1.8	4.5
2.7	2.2	3.7	2.2	4.4	2.8
2.3	1.9	3.6	3.9	2.3	3.4
3.3	1.8	3.5	4.1	2.2	3.0
4.1	3.4	3.2	2.2	3.0	2.8

(a) Proceda à organização dos dados construindo uma tabela de frequências onde figurem as frequências absolutas, relativas, absolutas acumuladas e relativas acumuladas.

(b) Calcule a média, moda e mediana usando os dados agrupados e também usando os dados não agrupados. Compare os resultados.

3. A média harmônica é representada pelo inverso da média aritmética dos inversos dos valores das observações, ou seja,

$$\overline{x}_h = \frac{n}{\sum \frac{1}{x_i}},$$

onde x_i é cada dado da amostra e n é o número de dados da amostra. Calcule a média harmônica para o conjunto de dados: 8; 4; 3; 10; 23; 3; 9; 11; 17; 22; 1.

2.2 Medidas de dispersão

Considere os seguintes conjuntos de dados:

$$X = \{70, 70, 70, 70, 70\}, Y = \{68, 69, 70, 71, 72\} \text{ e } Z = \{5, 15, 50, 120, 160\}.$$

Note que a média de cada conjunto é igual a 70, entretanto eles diferem grandemente em variabilidade. Por esta razão torna-se necessário estabelecer medidas que indiquem o grau de dispersão, ou variabilidade em relação ao valor central, medidas essas chamadas de medidas de dispersão. As principais medidas de dispersão são: amplitude total, variância, desvio padrão e o coeficiente de variação.

2.2.1 Amplitude total

Definição 2.2.1. A amplitude total é(a_t) a diferença entre o maior e menor valor observado.

Exemplo 2.2.1. Para os conjuntos

$$X = \{70, 70, 70, 70, 70\}, Y = \{68, 69, 70, 71, 72\} \text{ e } Z = \{5, 15, 50, 120, 160\},$$

obtemos as amplitudes totais 0, 4 e 155, respectivamente.

Observação 2.2.1. (i) Vantagem: é obtida de forma fácil e simples.

(ii) Desvantagem: só leva em conta os dois valores extremos da série, sem abranger os valores intermediários, ou seja, é pouco precisa.

Observação 2.2.2. Quando os dados estão agrupados em classes, a amplitude total é a diferença entre o limite superior da última classe e o limite inferior da primeira classe.

2.2.2 Variância

Definição 2.2.2. A variância (s^2) é dada por

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

Observação 2.2.3. Quando calculamos a variância de uma população, adotamos o denominador n em vez de $n-1$.

Exemplo 2.2.2. As idades de 10 alunos do curso de matemática do IFNMG são: 18, 22, 20, 24, 19, 19, 23, 23, 22 e 20. Qual a variância das idades desses alunos?

Observação 2.2.4. Quando os dados estão agrupados em tabelas de frequências sem intervalos, basta utilizar a fórmula:

$$s^2 = \frac{\sum f_i \cdot x_i^2 - \frac{(\sum x_i \cdot f_i)^2}{n}}{n-1}.$$

Caso os dados estejam agrupados em intervalos, substitua cada x_i na fórmula acima pelo ponto médio de cada intervalo.

Exemplo 2.2.3. Calcule a variância do seguinte conjunto de dados:

CLASSES	f_i
2 ⊢ 6	5
6 ⊢ 10	12
10 ⊢ 14	21
14 ⊢ 18	15
18 ⊢ 22	7

Solução: Para facilitar os cálculos, vamos adicionar as três colunas:

CLASSES	f_i	Ponto médio (pm)	$f_i \cdot pm$	$f_i \cdot pm^2$
2 ⊢ 6	5	4	20	80
6 ⊢ 10	12	8	96	768
10 ⊢ 14	21	12	252	3024
14 ⊢ 18	15	16	240	3840
18 ⊢ 22	7	20	140	2800
	$\sum = 60$		$\sum = 748$	$\sum = 10512$

Logo, $s^2 = 19,782$.

Proposição 2.2.1. *Propriedades da variância:*

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

2. *Ao somar uma constante a todos os valores de um conjunto de dados, a variância destes dados não se altera.*

3. *Ao multiplicar todos os valores de um conjunto de dados por uma constante, a variância destes dados fica multiplicada pelo quadrado desta constante.*

Observação 2.2.5. (i) Vantagem: fácil compreensão, leva em conta todos os valores do conjunto de dados e possui propriedades estatísticas importantes para a inferência.

(ii) Desvantagem: como a variância é calculada a partir da média, é uma medida pouco resistente (muito influenciada por valores atípicos); como a unidade de medida fica elevada ao quadrado, a interpretação da variância se torna mais difícil.

2.2.3 Desvio padrão

Como medida de dispersão, a variância tem a desvantagem de apresentar unidade de medida igual ao quadrado da unidade dos dados. Assim, por exemplo, se os dados são medidos em metros, a variância é dada em metros ao quadrado. Por esse motivo, definiu-se uma nova medida, denominada desvio padrão, que indica a dispersão dos dados dentro da amostra, isto é, o quanto os dados em geral diferem da média.

Definição 2.2.3. O desvio padrão (s) é calculado como a raiz quadrada da variância, ou seja,

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

Exemplo 2.2.4. Sabemos que a variância das seguintes idades: $18, 22, 20, 24, 19, 19, 23, 23, 22$ e 20 é $s^2 = 3,8$ anos^2. Logo, $s = 1,949$ anos.

Observação 2.2.6. (i) Quanto maior for o desvio padrão, maior será a dispersão dos dados.

(ii) O desvio padrão possui as mesmas propriedades da variância.

2.2.4 Coeficiente de variação

Em alguns casos, se tem o interesse de comparar variabilidades de diferentes conjuntos de valores. A comparação, a partir do desvio padrão, se torna difícil em situações em que as médias são muito desiguais ou as unidades de medidas são diferentes. Uma medida que supre essa necessidade é o coeficiente de variação.

Definição 2.2.4. O coeficiente de variação (CV), que é dado por:

$$CV = \frac{s}{\overline{x}},$$

onde, s é o desvio padrão e $\overline{x}$ é a média aritmética dos dados.

Exemplo 2.2.5. Considere a série $4, 5, 7, 9$ e 10 referente ao tempo (min) para realizar um determinada tarefa. Assim, $\overline{x} = 7$ min, $s = 2,55$ min e $CV = 0,364 = 36,4\%$.

Observação 2.2.7. O CV é utilizado para analisar qual amostra é mais homogênea (menor variabilidade). Aquela que apresentar o menor CV é a mais homogênea.

Exemplo 2.2.6. Uma empresa avaliou 30 lotes de peças de uma indústria A. O número de peças defeituosas por lote é apresentado na tabela a seguir.

Peças defeituosas	lotes (f_i)
0	9
1	9
2	5
3	4
4	2
5	1
Total	30

(a) Calcule o CV destes dados.

(b) A mesma empresa avaliou também 30 lotes de peças de uma indústria B, cuja média e desvio padrão foram, respectivamente, $\overline{x_B} = 0,7$ peças e $s_B = 1,022$ peças. Calcule o CV dessa amostra e diga qual da duas indústrias, A ou B, produziram peças mais homogêneas.

Observação 2.2.8. Vantagens do CV: utilizada para comparar variabilidades de diferentes conjuntos de dados e é desprovido de unidade de medida.

Exercícios

1. Considerando os dados da seguinte distribuição de frequências, calcule: a) a média aritmética; b) a amplitude total; c) a variância; d) desvio padrão.

Peças defeituosas	lotes (f_i)
0	9
1	9
2	5
3	4
4	2
5	1
Total	30

2. Um centro de saúde A registrou na tabela seguinte as idades dos 40 pacientes atendidos em uma semana do mês de março do ano passado.

Idades (anos)	Nº de pacientes (f_i)
5	5
9	8
13	17
17	6
21	4
Total	40

   (a) Calcule a média aritmética, a variância, o desvio padrão e o coeficiente de variação das idades.

   (b) Um outro centro de saúde B, registrou as idades dos 50 pacientes atendidos em uma semana do mês de março do ano passado, cuja média e desvio padrão foram, respectivamente, $\overline{x_B} = 15,92$ anos e $s_B = 5,25$ anos. Calcule o CV desses dados e diga qual dos dois centros de saúde, A ou B, são mais homogêneos.

Capítulo 3

Probabilidade

O conhecimento de probabilidade constrói a base que nos permite entender como a interferência estatística e as técnicas de auxílio de decisão são desenvolvidas, porque elas funcionam, e como as conclusões obtidas a partir desses procedimentos podem ser apresentados e interpretados corretamente. Para o entendimento de probabilidade, alguns conceitos são necessários.

3.1 Definições básicos

Definição 3.1.1. Um experimento é qualquer processo de observação.

Exemplo 3.1.1. São exemplos de experimentos:

(a) Uma observação meteorológica ou sísmica.

(b) Uma pesquisa de opinião para saber quantos eleitores votarão no candidato x ou y na última eleição.

(c) A observação dos pontos da face superior no lançamento de um dado.

(d) Suponha que estejamos na janela de um apartamento situado no 3^{o} andar. Então, soltamos uma pedra e medimos o tempo que ela demora a cair

Observação 3.1.1. No item (d), se repetirmos o experimento tantas vezes quanto quisermos, iremos medir o mesmo tempo. Só mediremos outro valor de tempo, se as condições do experimento mudarem (por exemplo, se formos para local mais alto ou mais baixo e soltarmos a pedra ou se as condições climáticas são diferentes). Neste caso, dizemos que o experimento é determinístico. Entretanto, o mesmo não ocorre para as outras situações descritas nos itens (a), (b) e (c).

Definição 3.1.2. Um experimento é dito aleatório quando, mesmo repetidos várias vezes sob condições semelhantes, apresentam resultados imprevistos.

Exemplo 3.1.2. (a) Os itens (a), (b) e (c) do Exemplo 3.1.1.

(b) Seleção de três itens ao acaso de uma linha de fabricação, classificando cada um como defeituoso ou não defeituoso.

(c) Registrar as vazões num certo rio, no mesmo mês, dia e hora em anos sucessivos.

Definição 3.1.3. (i) Espaço amostral, denotado por S, é o conjunto de todos os possíveis resultados de um experimento aleatório.

(ii) Um evento é qualquer subconjunto do espaço amostral de um experimento aleatório.

Exemplo 3.1.3. (a) Em relação ao experimento aleatório: "observação dos pontos da face superior no lançamento de um dado" temos que o espaço amostral é $S = \{1, 2, 3, 4, 5, 6\}$. Além disso, "sair um número par" e "sair número menor que 4" são exemplos de eventos que podem ser representados por $A = \{2, 4, 6\}$ e $B = \{1, 2, 3\}$, respectivamente.

(b) Em relação ao experimento aleatório: "seleção de três itens ao acaso de uma linha de fabricação, classificando cada um como defeituoso ou não defeituoso" temos que

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

sendo D defeituoso e N não defeituoso. Se A é o evento "selecionar pelo menos uma peça defeituosa", então $A = \{DDD, DDN, DND, DNN, NDD, NDN, NND\}$.

(c) Em relação ao experimento aleatório: "registrar as vazões num certo rio, no mesmo mês, dia e hora em anos sucessivos" temos que $S = \{q \,|\, 0 \leq q \leq q_{\max}\}$, onde q é a vazão. Sendo assim, $A = \{q \,|\, q > 1\}$ é um evento.

Observação 3.1.2. Novos eventos podem ser originados da combinação de eventos já existentes, por exemplo, se A e B são dois eventos de um espaço amostral S, então o evento:

(i) A união B ($A \cup B$) é o evento que ocorre A ou B;

(ii) A interseção B ($A \cap B$) é o evento que ocorre A e B;

(iii) complementar de A (A^c) é o evento não ocorrer A;

(iv) diferença de A e B ($A - B$) é o evento ocorrer A, mas não ocorrer B.

Exemplo 3.1.4. No lançamento de um dado, considere os seguintes eventos: A: "sair face menor que 4" e B: "sair face ímpar". Assim, $A = \{1, 2, 3\}$ e $B = \{1, 3, 5\}$. Logo,

(a) $A \cup B = \{1, 2, 3, 5\}$ (sair face menor que 4 ou face ímpar)

(b) $A \cap B = \{1, 3\}$ (sair face menor que 4 e face ímpar)

(c) $A^c = \{4, 5, 6\}$ (não sair face menor que 4, ou seja, sair face maior ou igual a 4)

(d) $A - B = \{2\}$ (sair face menor que 4, mas não sair face ímpar)

Definição 3.1.4. Dois eventos A e B (de um mesmo espaço amostral) são ditos mutuamente exclusivos se eles não puderem ocorrer simultaneamente, isto é, $A \cap B = \emptyset$.

Exemplo 3.1.5. Considere o evento aleatório: "observação dos pontos da face superior no lançamento de um dado". Se $A = \{1\}$, $B = \{5, 6\}$ e $C = \{2, 4, 6\}$, então A e B são mutuamente exclusivos, entretanto, B e C não são.

Exercícios

1. Descreva o espaço amostral S dos seguintes experimentos aleatórios:
   (a) Observação da face superior no lançamento de dois dados.
   (b) São lançados dois dados e registra-se a soma dos números das faces dos dados voltadas para cima.
   (c) Escolhe-se ao acaso uma família com três crianças de um município e são registrados os sexos dos filhos.
   (d) No item (c), observa-se apenas o número de meninas na família selecionada.
   (e) Observação das faces superiores no lançamento de um dado e uma moeda.
   (f) Uma árvore é selecionada em um parque e sua altura em centímetros é medida.
   (g) A partir de certo momento, registra-se a quantidade de veículos que passam por um pedágio até que passe a primeira motocicleta.

3.2 Conceito clássico de probabilidades

Existem três formas de se definir a probabilidade de um evento para um experimento aleatório: a forma clássica (a priori), a forma baseada na frequência (posteriori) e a forma axiomática. Nesse material vamos utilizar a forma clássica e vamos assumir que: (i) todos os espaços amostrais são enumeráveis e finitos; (ii) todos os elementos de um espaço amostral possuem a mesma chance de ocorrer (S é um conjunto equiprovável).

Observação 3.2.1. (a) Atualmente, o conceito mais aceito é a forma axiomática.

(b) O conceito axiomático não fornece formas e sim condições para o cálculo das probabilidades.

(c) Os conceitos a priori e a posteriori se enquadram no conceito axiomático.

Definição 3.2.1. Num experimento aleatório, a probabilidade de ocorrer um evento A de um espaço amostral S, denotada $P(A)$, é o número real

$$P(A) = \frac{n(A)}{n(S)},$$

onde $n(A)$ é o número de elementos de A e $n(S)$ é o número de elementos de S.

Exemplo 3.2.1. No lançamento de um dado, qual a probabilidade de:

(a) Sair um número par?

(b) Não sair um número ímpar?

(c) Sair um número primo e ímpar?

(d) Sair um número primo ou ímpar?

(e) (Evento impossível) Sair um número que não é par e nem é ímpar?

(f) (Evento certo) Sair um número menor que 100?

Exemplo 3.2.2. Numa classe com 50 alunos, 35 estudam inglês, 22 estudam espanhol e 10 estudam inglês e espanhol.

(a) Qual a probabilidade de um aluno que estuda inglês também estudar espanhol?

(b) Ao apontarmos, ao acaso, um desses jovens, qual é a probabilidade de ele não estudar inglês nem espanhol?

Exemplo 3.2.3. No lançamento de 3 moedas distintas, qual a probabilidade de:

(a) Ocorrer coroa pelo menos uma vez?

(b) Não ocorrer cara?

Proposição 3.2.1. *As seguintes afirmações são válidas:*

1. $0 \leq P(A) \leq 1$
2. $P(S) = 1$ *e* $P(\emptyset) = 0$, *onde* S *é o evento certo e* $\emptyset$ *é o evento impossível.*
3. *Se* A *e* B *são mutuamente exclusivos* ($A \cap B = \emptyset$), *então* $P(A \cup B) = P(A) + P(B)$.
4. *Se* A *e* B *forem eventos quaisquer, então* $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

5. $P(A^c) = 1 - P(A)$.

Exemplo 3.2.4. A probabilidade de ocorrer um acidente em uma competição de carros é 0,18; a probabilidade de chover em um dia de competição é 0,28; e a probabilidade de ocorrer acidente e chuva em um dia de competição é 0,08. Determine a probabilidade de:

(a) não ocorrer acidente na próxima competição;

(b) chover ou ocorrer um acidente na próxima competição;

(c) não chover e não ocorrer acidente na próxima competição;
Dica: Utilize a igualdade $A^c \cup B^c = (A \cap B)^c$

(d) chover, mas não ocorrer acidente na próxima competição.

3.3 Probabilidade condicional

Definição 3.3.1. Se A e B são dois eventos de um espaço amostral S e $P(B) > 0$, então a probabilidade de ocorrer A, sendo que B ocorreu é denotada por $P(A|B)$ e calculada por

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{n(A \cap B)}{n(B)}.$$

Exemplo 3.3.1. Dois dados são lançados simultaneamente. Qual a probabilidade da soma ser igual a 6, dado que o primeiro dado saiu um nº menor que 3? Vamos resolver esse exemplo de dois modos:

1º modo: Utilizando a fórmula da Definição 3.3.1: Neste caso, $A = \{$soma ser igual a 6$\}$ e $B = \{$o primeiro dado saiu um nº menor que 3$\}$. Assim, $P(A \cap B) = \frac{2}{36}$ e $P(B) = \frac{12}{36}$, logo

$$P(A|B) = \frac{2}{12} = \frac{1}{6}.$$

2º modo: Sem utilizar a fórmula: Neste caso, o nosso espaço amostral é reduzido à

$$\{(1,1),(1,2),\ldots,(1,6),(2,1),(2,2),\ldots,(2,6)\},$$

pois sabemos que B já ocorreu. Assim, a probabilidade procurada é $\frac{2}{12} = \frac{1}{6}$.

Exemplo 3.3.2. Em uma urna há 10 bolas numeradas de 1 a 10. Determine a probabilidade de retirarmos uma bola numerada com um número par, sabendo que saiu uma bola com numeração maior que 4.

Observação 3.3.1. Quando o evento A independe do evento B, ou seja, a ocorrência de B não interfere a ocorrência de A, temos que

$$P(A|B) = P(A) \Rightarrow P(A \cap B) = P(A) \cdot P(B).$$

Exemplo 3.3.3. (a) Lançam-se sucessivamente um dado e uma moeda. Qual é a probabilidade de se obter cara e um número ímpar?

(b) Em uma urna há 4 fichas brancas e 6 azuis. Qual a probabilidade de retirarmos, sucessivamente, uma ficha branca e outra azul com reposição? E sem reposição?

Exercícios

1. Sejam A e B dois eventos em um espaço amostral S tais que $P(A) = 1/2$, $P(B) = 1/4$ e $P(A \cap B) = 1/5$. Calcule as probabilidades dos seguintes eventos:

   (a) B não ocorre.

   (b) Não ocorre A e B.

   (c) A ou B ocorre.

   (d) Não ocorre B e A sim.

2. Suponha que 60% das pessoas assinem o jornal A, 40% o jornal B e 30% ambos.

   (a) Se selecionarmos ao acaso uma pessoa que assina ao menos um dos jornais, qual a probabilidade de que assine o jornal A?

   (b) Dado que um pessoa não assina o jornal A, qual a probabilidade de que também não assine B?

3. Em um curso secundário, 1/3 dos estudantes são do sexo masculino. A proporção de rapazes que estudam estatística é 20% e apenas 10% das moças dedicam-se à disciplina. Obtenha as probabilidade de que

   (a) Um estudante escolhido ao acaso estude estatística;

   (b) Um estudante de estatística selecionado ao acaso seja do sexo feminino.

4. Dois dados são lançados conjuntamente. Determine a probabilidade da soma ser 10 ou maior que 10.

5. Uma bola é retirada ao acaso de uma urna que contém 6 vermelhas, 4 brancas e 5 azuis. Determinar a probabilidade dela:

   (a) ser vermelha;

   (b) ser branca;

   (c) não ser vermelha;

   (d) ser vermelha ou branca;

   (e) de que 3 bolas sejam retiradas na ordem vermelha, branca e azul, com reposição;

   (f) de que 3 bolas sejam retiradas na ordem vermelha, branca e azul, sem reposição.

6. Um experimento consiste em observar a soma dos números de 2 dados quando eles são jogados.

   (a) Descreva o espaço amostral.

   (b) Assumido todos os resultados equiprováveis, encontre a probabilidade da soma ser 7 e a probabilidade da soma ser maior que 10.

7. Teresa tem três irmãs: Maria, Inês e Joana. Ela vai escolher, ao acaso, uma das irmãs para ir a um arrail no próximo fim de semana. Teresa vai escolher, também ao acaso, se vai no arraial no próximo sábado ou no próximo domingo. Qual é a probabilidade de Teresa escolher ir no arraial no sábado com Maria?

8. Num lançamento de um dado viciado, a probabilidade de ocorrer cada número ímpar é o dobro da probabilidade de ocorrer cada número par.

   (a) Calcule a probabilidade de sair face 2.

   (b) Calcule a probabilidade de que o número de pontos obtidos seja superior a 3.

9. Um submarino atira 3 torpedos contra um porta-aviões. O porta-aviões só será afundado de 2 ou mais torpedos o atingirem. Sabendo que a probabilidade de um torpedo acertar o porta-aviões é de $0,4$, qual é a probabilidade de afundar o porta-aviões.

10. As preferências de homens e mulheres por cada gênero de filme alugado em uma locadora de vídeos estão apresentadas na tabela a seguir.

Sexo	Tipo de filme		
	Comédia	Romance	Policial
Masculino	136	92	248
Feminino	102	195	62

Sorteando-se ao acaso um registro de locação, pede-se a probabilidade de:

(a) ser um filme policial alugado por uma mulher;

(b) ser uma comédia;

(c) ser de um homem ou de um romance;

(d) ser de um filme policial dado que foi alugado por um homem.

11. Em uma urna serão colocadas 4 bolas azuis, numeradas de 1 a 4, e 5 bolas amarelas, numeradas de 1 a 5. Sorteando uma bola dessa urna, qual é a probabilidade dela

(a) ser azul ou ter número ímpar?

(b) ser azul e ter número ímpar?

12. Três cavalos (A, B e C) estão numa corrida; A é duas vezes mais provável de ganhar que B e B é duas vezes mais do que C. Quais são as probabilidades de vitória de cada um, isto é, $P(A)$, $P(B)$ e $P(C)$? Qual é a probabilidade de que B ou C ganhe?

13. Quatro estudantes afirmam que os pneus de seus carros furaram e, por esta razão, não puderam comparecer à prova. Para confirmar as alegações, o professor pede que os estudantes identifiquem o pneu furado. Se nenhum pneu furou e eles escolheram aleatoriamente um pneu que supostamente teria furado, qual é a probabilidade de que escolham o mesmo pneu?

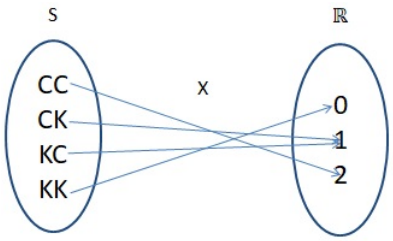
Capítulo 4

Variáveis aleatórias

4.1 Definições básicas

Definição 4.1.1. Variável aleatória $(X, Y, Z, \ldots)$ é uma função que associa cada elemento de um espaço amostral S a um número real, ou seja, é uma função X com domínio S e contra domínio $\mathbb{R}$.

Exemplo 4.1.1. Considere o seguinte experimento aleatório: "lançamento de duas moedas e a observação de suas faces" e X o número de caras obtidas. Se C: cara e K: coroa, obtemos que



Assim, $X(CC) = 2$, $X(CK) = X(KC) = 1$ e $X(KK) = 0$.

Exemplo 4.1.2. Considere o experimento aleatório: "observar o tempo (em min) de reação a um certo medicamento" e defina Y como o tempo de reação ao medicamento. Assim, Y é uma variável aleatória que pode assumir qualquer valor real positivo, ou seja, $Y(t) = 2$ ou 5 ou 1 ou 15 ou $16,2349$ ou $7,1232847$ ou 3 ou $3,1$ ou $3,11$ ou $10,4$ ou $0,09$ ou $30,01$ ou $30,0001$ (em min).

Note que as variáveis aleatórias dos dois exemplos são diferentes, pois no primeiro, o espaço amostral é finito e a função assume valores inteiros, diferentemente do segundo exemplo. Vamos estudar dois tipos de variáveis aleatórias: discretas e contínuas.

4.2 Variáveis aleatórias discretas

Definição 4.2.1. Uma variável aleatória X é dita discreta quando o número de valores assumidos por X for finito ou infinito enumerável, ou seja, seus possíveis valores podem ser dispostos em uma lista finita ou infinita.

Exemplo 4.2.1. São exemplos de variáveis aleatórias discretas:

(a) X: o número de caras em um lançamento de uma única moeda.

(b) Y: o número de coroas até que ocorra cara.

(c) Z: o número de lançamentos até que ocorra cara.

Observação 4.2.1. Note que, em relação ao exemplo anterior, adotando C para cara e K para coroa, tem-se: no item (a) $S = \{C, K\}$, $X(K) = 0$ e $X(C) = 1$; nos itens (b) e (c): $S = \{C, KC, KKC, KKKC, KKKKC, \ldots\}$, sendo $Y(C) = 0$, $Y(KC) = 1$, $Y(KKC) = 2$, ... e $Z(C) = 1$, $Z(KC) = 2$, $Z(KKC) = 3$, ...

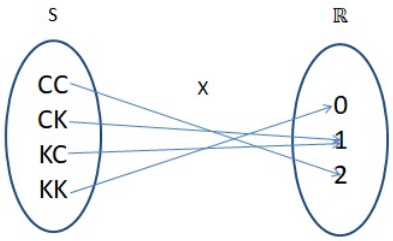
Definição 4.2.2. Seja X uma variável aleatória discreta que pode assumir os valores $x_1, x_2, x_3, \ldots$. Chama-se função de probabilidade (ou função de distribuição de probabilidade) da variável aleatória discreta X a função P que a cada x_i $(i = 1, 2, 3, \ldots)$ associa sua probabilidade de ocorrência, denotada por $P(X = x_i)$, satisfazendo as seguintes condições:

(i) $0 \leq P(X = x_i) \leq 1$, para todo $i = 1, 2, 3, \ldots$

(ii) $\sum P(X = x_i) = 1$

Esta função P pode ser expressa por uma tabela, gráfico ou fórmula como veremos a seguir.

Exemplo 4.2.2. Considere o experimento aleatório: "lançamento de duas moedas" e a variável aleatória X: número de caras obtidas (variável aleatória discreta). Assim,

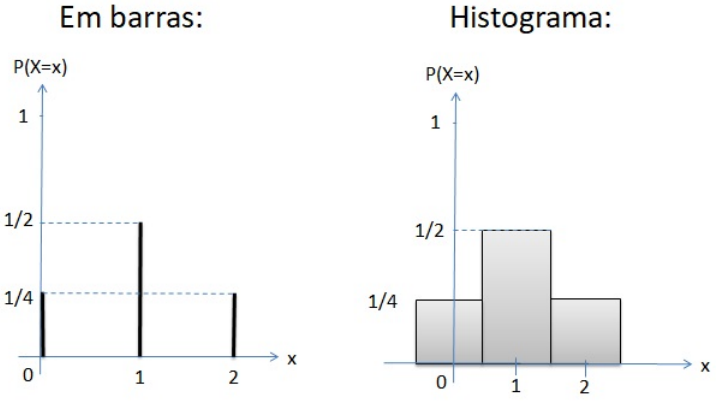


sendo $x_1 = 0$, $x_2 = 1$ e $x_3 = 2$. Agora, vamos expressar a função de probabilidade P de três formas diferentes:

(i) Tabela (distribuição de probabilidade):

x_i	0	1	2	$\sum$
$P(X = x_i)$	1/4	2/4	1/4	1

(ii) Gráfico:



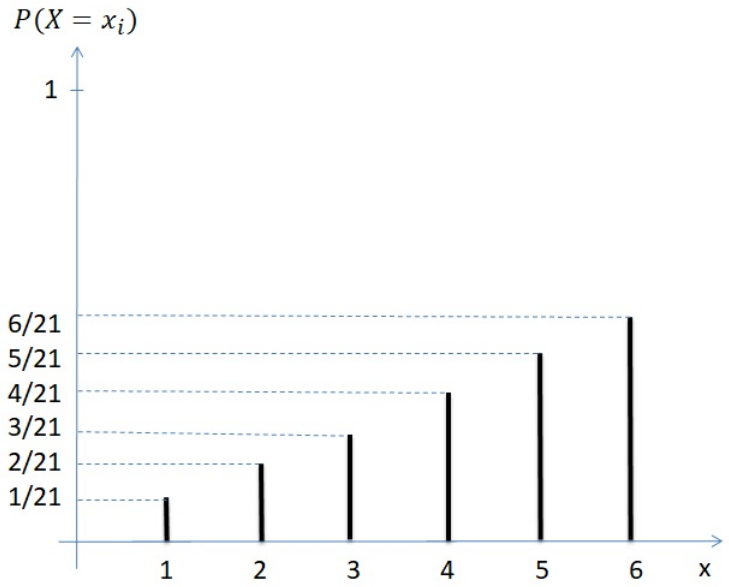
(iii) Fórmula: $P(X = x_i) = \frac{1}{4} \cdot \binom{2}{x_i}$ para $i = 1, 2, 3$, onde $\binom{n}{p} = \frac{n!}{p! \cdot (n-p)!}$.

Exemplo 4.2.3. Considere o lançamento de um dado viciado de tal forma que a probabilidade é proporcional ao valor obtido no lançamento e a variável aleatória X: número de pontos obtidos num lançamento. Faça uma tabela e um gráfico para representar a função de probabilidade que representa a situação. Em seguida, encontre sua fórmula.

Solução:

x_i	1	2	3	4	5	6	$\sum$
$P(X = x_i)$	1/21	2/21	3/21	4/21	5/21	6/21	1

$$P(X = x_i) = \frac{x_i}{21}$$



Definição 4.2.3. Dada a variável aleatória discreta X, chama-se função de distribuição a função $F(x)$ dada por

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

onde x é um número real.

Observação 4.2.2. Note que F é uma função de domínio $\mathbb{R}$ e contra domínio $[0, 1]$.

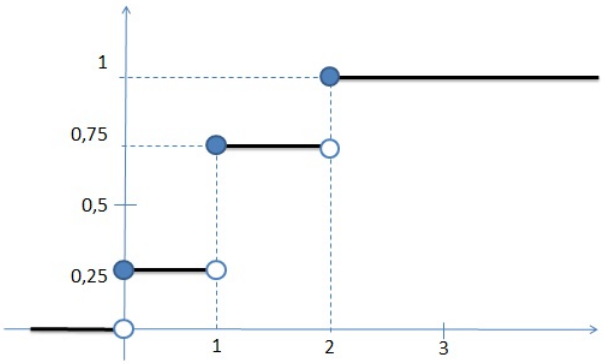
Exemplo 4.2.4. Considere o experimento que consiste no lançamento independente de uma moeda duas vezes e seja X: número de caras obtidas. Obtemos que $P(X = 0) = P(X = 2) = 1/4$ e $P(X = 1) = 1/2$. Logo,

$$\begin{array}{lcllcl} F(-2) & = & P(X \leq -2) & = & 0 & \\ F(0) & = & P(X \leq 0) & = & P(0) & = 1/4 \\ F(1) & = & P(X \leq 1) & = & P(1) + P(0) & = 3/4 \\ F(1,6) & = & P(X \leq 1,6) & = & P(1) + P(0) & = 3/4 \\ F(5) & = & P(X \leq 5) & = & P(2) + P(1) + P(0) & = 1 \end{array}$$

Portanto,

$$F(x) = \begin{cases} 0 & \text{se} \quad x < 0 \\ 0,25 & \text{se} \quad 0 \leq x < 1 \\ 0,75 & \text{se} \quad 1 \leq x < 2 \\ 1 & \text{se} \quad x \geq 2 \end{cases}$$

e o gráfico de F fica:



Exemplo 4.2.5. De uma urna com três bolas pretas e duas brancas, retiram-se duas bolas juntas (aqui deve-se considerar todas as bolas distintas). Se X é o número de bolas pretas retiradas, determine a função de probabilidade $P(X = x_i)$ e a função de distribuição F.

Solução: O espaço amostral é $S = \{B_1B_2, P_1B_1, P_1B_2, P_2B_1, P_2B_2, P_3B_1, P_3B_2, P_1P_2, P_1P_3, P_2P_3\}$ (possui $C_{5,2}$ elementos). Assim, $X(B_1B_2) = 0$, $X(P_1B_1) = X(P_1B_2) = X(P_2B_1) = X(P_2B_2) = X(P_3B_1) = X(P_3B_2) = 1$ e $X(P_1P_2) = X(P_1P_3) = X(P_2P_3) = 2$. Logo, $P(X = 0) = 1/10$, $P(X = 1) = 6/10$ e $P(X = 2) = 3/10$. Por último, observe que $F(0) = 1/10$, $F(1) = 7/10$ e $F(2) = 1$, ou seja,

$$F(x) = \begin{cases} 0 & \text{se} \quad x < 0 \\ 0,1 & \text{se} \quad 0 \leq x < 1 \\ 0,7 & \text{se} \quad 1 \leq x < 2 \\ 1 & \text{se} \quad x \geq 2 \end{cases}$$

Exercícios

1. Seja X uma variável aleatória discreta com a seguinte distribuição de probabilidade:

x_i	-2	-1	2	4	$\sum$
$P(X = x_i)$	1/4	1/8	1/2	1/8	1

   Pede-se:

   (a) Traçar o gráfico da função de probabilidade de X.

   (b) Obter a função de distribuição e traçar seu gráfico.

2. Consideremos a jogada de um par de dados e seja X a variável aleatória "soma dos pontos".

   (a) Esta variável aleatória é discreta ou contínua?

   (b) Determine a função de probabilidade de X e construa seu gráfico.

   (c) Calcule $P(X = 6)$, $P(X > 8)$ e $P(5 < x \leq 7)$.

3. Um homem possui 4 chaves em seu bolso. Como está escuro, ele não consegue ver qual a chave correta para abrir a porta de sua casa. Ele testa cada uma das chaves até encontrar a correta.

   (a) Determine o espaço amostral desse experimento.

   (b) Defina a variável aleatória X: número de chaves experimentadas até conseguir abrir a porta (inclusive a chave correta). Quais são os valores de X?

   (c) Faça a tabela da função de probabilidade de X.

4.3 Variáveis aleatórias contínuas

Já sabemos que X é dita uma variável aleatória contínua se ela pode assumir qualquer valor em um determinado intervalo de números reais, consequentemente, não podemos listar todos os resultados possíveis. No caso das variáveis aleatórias discretas, definimos a função de probabilidade $P(X = x_i)$ que associa cada x_i $(i = 1, 2, 3, \ldots)$ a sua probabilidade. O mesmo conceito não se aplica as variáveis aleatórias contínuas, pois não podemos enumerar (listar) os valores da variável aleatória e, portanto, não se pode indagar qual a probabilidade do i-ésimo valor de X. Deste modo, vamos definir um novo conceito para as variáveis aleatórias contínuas "equivalente" a função de probabilidade das variáveis aleatórias discretas.

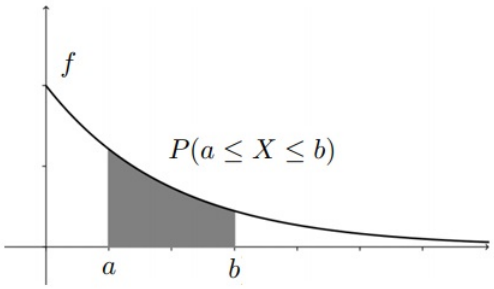
Definição 4.3.1. Seja X uma variável aleatória contínua. A função densidade de probabilidade $f(x)$ é uma função que satisfaz as seguintes condições:

(i) $f(x) \geq 0$ para todo valor de x no contra domínio de X.

(ii) A área sob a curva da função densidade é igual a 1, isto é,

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

(iii) $P(a \leq X \leq b)$ é a área sob a curca limitada pelos pontos a e b.



Observação 4.3.1. (i) Se f é a função densidade de probabilidade, então $f(x)$ não é a probabilidade. A probabilidade é calculada quando encontramos a área sob a curva de f entre a e b.

(ii) Se X é uma variável aleatória contínua, a probabilidade de X tomar um determinado valor é zero, ou seja, $P(X = a) = 0$. Nesse caso, o que faz sentido é falar da probabilidade de X estar compreendido entre dois valores diferentes.

(iii) $P(a \leq X \leq b) = P(a < X \leq b) = P(a \leq X < b) = P(a < X < b)$.

Exemplo 4.3.1. Seja X uma variável aleatória contínua com a seguinte função densidade de probabilidade:

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

(a) Verifique se f é realmente uma função de densidade de probabilidade.

(b) Se sim, determine o valor de $P(0 < X < 1)$, $P(X < 1/2)$ e $P(X \leq 1/2 \,|\, 1/3 \leq X < 2/3)$.

Exemplo 4.3.2. Determine a constante c de modo que a função

$$f(x) = \begin{cases} cx^2, & \text{se } 0 < x < 3 \\ 0, & \text{caso contrário} \end{cases}$$

seja uma função de densidade de probabilidade e calcule $P(1 < X < 2)$.

Observação 4.3.2. A função de distribuição para uma variável aleatória contínua é dada por

$$F(x) = \int_{-\infty}^{x} f(t)\,dt,$$

mas não vamos entrar em detalhes.

Exercícios

1. Classifique cada uma das variáveis aleatórias como discretas ou contínuas.

   (a) peso de uma pessoa;

   (b) número de acidentes ocorridos em uma semana;

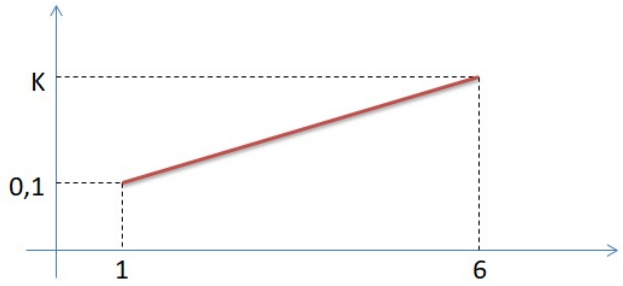
   (c) número de filhos;

   (d) vida útil de um componente eletrônico;

   (e) quantidade de chuva que ocorre numa região;

   (f) número de funcionários de uma empresa.

2. Considere a função f abaixo:



   (a) Encontre o valor de k para que f seja uma função densidade de probabilidade de uma variável aleatória contínua X e determine a equação que define f.

   (b) Calcule $P(2 \leq X \leq 3)$.

3. Considere a seguinte função densidade de probabilidade da variável aleatória contínua X:

$$f(x) = \begin{cases} 2e^{-2x} & \text{se} \quad x \geq 0 \\ 0 & \text{se} \quad x < 0 \end{cases}$$

   Calcule $P(0 < X \leq 2)$ e $P(X \leq 1)$.

4. O tempo (em anos) adequado de troca de uma peça de certa marca de computador é uma v.a. com a seguinte função densidade:

$$f(x) = \begin{cases} x/8, & \text{se } 0 \leq x \leq 4 \\ 0, & \text{caso contrário} \end{cases}$$

   Qual a probabilidade de um computador necessitar da troca da peça antes de um ano e três meses de uso?

4.4 Medidas descritivas

Observe que o espaço amostral pode não ser um conjunto numérico, entretanto, se X é a variável aleatória que descreve a situação, então o conjunto imagem de X é sempre um conjunto numérico, logo pode-se calcular as medidas descritivas em relação à esse conjunto.

Nos modelos probabilísticos, parâmetros podem ser empregados para caracterizar sua distribuição de probabilidade. Dada uma distribuição de probabilidade é possível associar certos parâmetros, os quais fornecem informação valiosa sobre tal distribuição. Os parâmetros mais importantes são: valor esperado (esperança ou média), variância e desvio padrão.

Definição 4.4.1. O valor esperado (esperança ou média) de uma variável aleatória X, denotado por $E(X)$ ou μ, é definida por:

(i) $E(X) = \sum_{i=1}^{\infty} x_i P(X = x_i)$, se X é discreta.

(ii) $E(X) = \int_{-\infty}^{+\infty} x f(x)\, dx$, se X é contínua.

Definição 4.4.2. Seja X uma variável aleatória discreta ou contínua com esperança dada por $E(X)$, a variância de X, denotado $\mathrm{Var}(X)$ ou σ^2, é definida por

$$\mathrm{Var}(X) = E(X-\mu)^2 = E(X^2) - [E(X)]^2.$$

Definição 4.4.3. Se X é uma variável aleatória discreta ou contínua, então o desvio padrão de X, denotado $\mathrm{DP}(X)$ ou σ, é definido como

$$\mathrm{DP}(X) = \sqrt{\mathrm{Var}(X)}.$$

Observação 4.4.1. Não confundir μ (média de todos os valores de X para os quais a probabilidade é conhecida) com $\overline{x}$ (média de alguns valores de X, geralmente em relação à uma amostra de valores). Da mesma forma, não confundir σ^2 com s^2 e σ com s.

Exemplo 4.4.1. De uma urna com três bolas pretas e duas brancas, retiram-se duas bolas juntas (aqui deve-se considerar todas as bolas distintas). Vimos que se X é o número de bolas pretas retiradas, então $S = \{B_1B_2, P_1B_1, P_1B_2, P_2B_1, P_2B_2, P_3B_1, P_3B_2, P_1P_2, P_1P_3, P_2P_3\}$ e $P(X=0) = 1/10$, $P(X=1) = 6/10$ e $P(X=2) = 3/10$. Além disso,

$$F(x) = \begin{cases} 0 & \text{se} \quad x < 0 \\ 0,1 & \text{se} \quad 0 \le x < 1 \\ 0,7 & \text{se} \quad 1 \le x < 2 \\ 1 & \text{se} \quad x \ge 2 \end{cases}$$

Sendo assim, como a variável aleatória é discreta, temos que $\mathrm{E}(X) = \sum x \cdot P(X=x)$, $\mathrm{Var}(X) = \sum (x-\mu)^2 \cdot P(X=x)$ e $\mathrm{DP}(X) = \sqrt{\mathrm{Var}(X)}$. Daí,

$$\begin{aligned} \mathrm{E}(X) &= 0 \cdot \frac{1}{10} + 1 \cdot \frac{6}{10} + 2 \cdot \frac{3}{10} = 1,2 \\ \mathrm{Var}(X) &= (0-1,2)^2 \cdot \frac{1}{10} + (1-1,2)^2 \cdot \frac{6}{10} + (2-1,2)^2 \cdot \frac{3}{10} = 0,36 \\ \mathrm{DP}(X) &= 0,6 \end{aligned}$$

Interpretamos a média/valor esperado e o desvio-padrão, respectivamente, do seguinte modo:

- Se o experimento fosse repetido um grande número de vezes, esperaríamos que o número médio de bolas pretas escolhidas fosse 1,2.
- Se o experimento fosse repetido um grande número de vezes, a variação média do número de bolas pretas escolhidas em torno do valor esperado seria 0,6.

Capítulo 5

Distribuições de probabilidade

Uma distribuição de probabilidade é um modelo matemático que relaciona um certo valor da variável em estudo com a sua probabilidade de ocorrência. Existem vários tipos de distribuição de probabilidade e algumas delas estão listadas abaixo:

(i) Para variáveis aleatórias discretas: Binomial, Poisson e Geométrica.

(ii) Para variáveis aleatórias contínuas: Normal, Gama, Valores Extremos e Exponencial.

Iremos estudar a distribuição binomial e a normal.

5.1 Distribuição binomial

Vamos caracterizar esta distribuição a partir da seguinte situação: Considere um experimento aleatório (discreto) consistindo em n tentativas independentes e a probabilidade de ocorrer "sucesso" em cada uma das n tentativas é sempre igual a p, consequentemente, a probabilidade de ocorrer "fracasso" é $q = 1 - p$. Seja

$$X\text{: número de sucessos nas } n \text{ tentativas,}$$

então X pode assumir os valores $0, 1, 2, \ldots, n$. Nestas condições dizemos que a variável aleatória discreta X tem distribuição binomial com parâmetros n e p e escrevemos $X \sim B(n, p)$ (lê-se: X possui distribuição binomial com parâmetros n e p). Assim, se $X \sim B(n, p)$, então sua função de probabilidade é

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

onde n é o número de testes, x é o número de sucessos, p é a probabilidade de sucesso, q é a probabilidade de fracasso ($q = 1 - p$) e $\binom{n}{x} = \frac{n!}{x!(n-x)!}$.

Exemplo 5.1.1. Uma moeda não viciada é lançada 8 vezes. Encontre a probabilidade de:

(a) Sair 5 caras.

(b) Sair pelo menos uma cara.

(c) Sair no máximo 2 caras.

Solução: Observe que essa situação representa uma distribuição binomial considerando X: "número de caras" (sucesso), $n = 8$, $p = 1/2$ e $q = 1/2$.

(a) $P(X = 8) = \binom{8}{5}(1/2)^5(1/2)^3 = 7/22 = 0,22 = 22\%$

(b) $P(X \geq 1) = 1 - P(X = 0) = 1 - \binom{8}{0}(1/2)^0(1/2)^8 = 255/256 = 0,996 = 99,6\%$

(c) $P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2) = 1/256 + 8/256 + 28/256 = 37/256 = 0,14 = 14\%$

Exemplo 5.1.2. O Cruzeiro, no campeonato brasileiro, tem 4/5 de probabilidade de vitória sempre que joga dentro do Mineirão. Se o Cruzeiro jogar 20 partidas no Mineirão, calcule a probabilidade de vencer exatamente 15 partidas.

Solução: X: "número de vitórias do Cruzeiro no Mineirão", $n = 20$, $p = 4/5 = 0,8$, $q = 1 - p = 0,2$ e $x = 15$. Logo,

$$P(X = 15) = \binom{20}{15} 0,8^{15} \cdot 0,2^5 = 0,1746 = 17,46\%$$

Observação 5.1.1. A distribuição binomial é uma sequência de experimentos de Bernoulli independentes entre si, onde a probabilidade de sucesso é constante em todas as repetições do experimento.

Exercícios

1. Suponha que a probabilidade dos pais terem um filho(a) com cabelos loiros seja 1/4. Se houverem 6 crianças na família, qual é a probabilidade de que:
   (a) Metade delas terem cabelos loiros?
   (b) Pelo menos duas tenham cabelos loiros?
   (c) Nenhuma tenha cabelo loiro?

2. Determine a probabilidade de ocorrer exatamente duas caras no lançamento de três moeda. E se fossem lançadas dez moedas?

3. Numa empresa, 40% dos contratos são pagos em dia. Qual a probabilidade de que, entre 12 contratos, três ou menos sejam pagos em dia?

4. Se a probabilidade de atingir um alvo num único disparo é $0,3$, qual é a probabilidade de que em 4 disparos o alvo seja atingido no mínimo 3 vezes?

5. Um sistema de segurança consiste em 4 alarmes (idênticos) de pressão alta, com probabilidade de sucesso $p = 0,8$ (cada um). Qual a probabilidade de se ter:
   (a) Exatamente 3 alarmes soando quando a pressão atingir o valor limite?
   (b) Pelo menos 3 em 4 alarmes soando quando houver uma invasão?

6. Um inspetor de qualidade extrai uma amostra de 10 tubos aleatoriamente de uma carga muito grande de tubos que se sabe que contém 20% de tubos defeituosos. Qual é a probabilidade de que não mais do que 2 dos tubos extraídos sejam defeituosos?

7. Um engenheiro de inspeção extrai uma amostra de 15 itens aleatoriamente de um processo de fabricação sabido produzir 85% de itens aceitáveis. Qual a probabilidade de que 10 dos itens extraídos sejam aceitáveis?

8. Num determinado processo de fabricação, 10% das peças são consideradas defeituosas. As peças são acondicionadas em caixas com 5 unidades cada uma. Então:
   (a) Qual a probabilidade de haver exatamente 3 peças defeituosas numa caixa?
   (b) Qual a probabilidade de haver duas ou mais peças defeituosas numa caixa?

9. Se a probabilidade de um certo gado sofrer uma dada reação nociva, resultante da injeção de um determinado soro, é 0,001. Determinar a probabilidade de, entre 2000 animais:

   (a) exatamente 3 sofrerem aquela reação;

   (b) mais do que 2 sofrerem aquela reação.

10. Placas de vídeo são expelidas em lotes de 30 unidades. Antes que a remessa seja aprovada, um inspetor escolhe aleatoriamente cinco placas do lote e as inspeciona. Se uma ou mais forem defeituosas, todo o lote é inspecionado. Suponha que haja três placas defeituosas nesse lote. Qual a probabilidade de que o controle de qualidade aponte para a inspeção total?

11. Em um sistema de transmissão de dados existe uma probabilidade igual a $0,05$ de um dado ser transmitido erroneamente. Ao se realizar um teste para analisar a confiabilidade do sistema foram transmitidos 4 dados.

   (a) Qual é a probabilidade de que tenha havido erro na transmissão?

   (b) Qual é a probabilidade de que tenha havido erro na transmissão de exatamente 2 dados?

12. Jogando-se uma moeda honesta cinco vezes e observando a face voltada para cima. Há interesse em calcular a probabilidade de ocorrência de uma, duas, ... , cinco caras. Qual é a probabilidade de obter ao menos quatro caras?

13. Suponha que você vai fazer uma prova com 10 questões do tipo verdadeiro-falso. Você nada sabe sobre o assunto e vai responder as questões por adivinhação.

   (a) Qual é a probabilidade de acertar exatamente 5 questões?

   (b) Qual é a probabilidade de acertar pelo menos 8 questões?

14. Suponha que 10% da população seja canhota. São escolhidas 3 pessoas ao acaso, com o objetivo de calcular a probabilidade de que o número de canhotos entre eles seja 0, 1, 2 ou 3. Qual é a probabilidade de ao menos uma das pessoas ser canhota?

15. A probabilidade de ocorrência de turbulência em um determinado percurso a ser feito por uma aeronave é de 0,4 em um circuito diário. Seja X o número de vôos com turbulência em um total de 7 desses vôos (ou seja, uma semana de trabalho). Qual a probabilidade de que:

   (a) Não haja turbulência em nenhum dos 7 vôos?

   (b) Haja turbulência em pelo menos 3 deles?

5.2 Distribuição normal

É uma das mais importantes distribuições de probabilidades, pois muitas variáveis encontradas na natureza se distribuem de acordo com o modelo normal. Também, é conhecida como distribuição de Gauss, Laplace ou Laplace-Gauss.

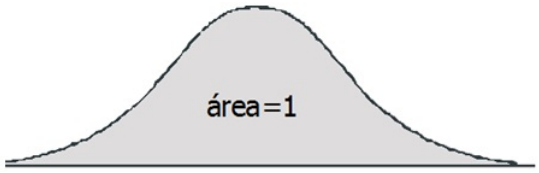
Uma variável aleatória contínua X tem uma distribuição normal se sua função densidade de probabilidade for do tipo:

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

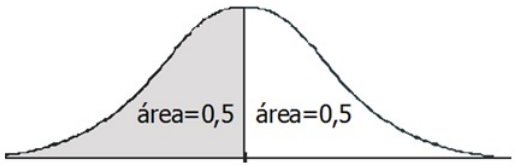
para $-\infty < x < \infty$, onde μ é a média da distribuição, σ é o desvio padrão da distribuição, $\pi \approx 3,14$ e $e \approx 2,71$. Nesse caso, escrevemos que $X \sim N(\mu, \sigma^2)$ (lê-se: X tem distribuição normal com parâmetros μ e σ^2).

Proposição 5.2.1. *A função densidade de probabilidade acima possui as seguintes propriedades:*

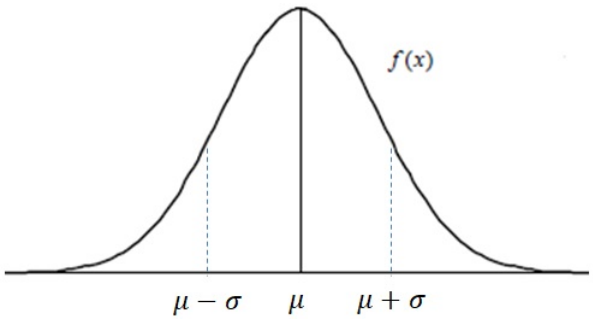
1. *Seu gráfico é em forma de sino e a área abaixo da curva é 1.*



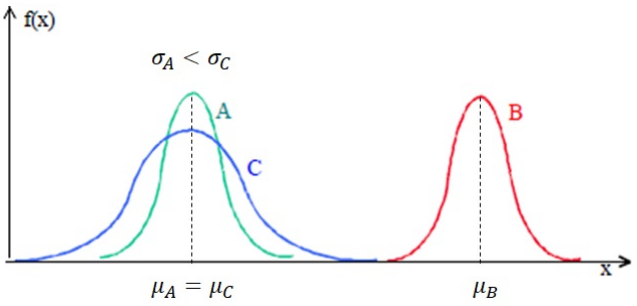
2. *A curva é simétrica em relação a μ.*



3. *É uni modal.*
4. *Quanto maior (ou menor) o valor de x, a função f fica cada vez mais próxima do zero, ou seja, $f(x)$ tende a zero quando x tende para $-\infty$ ou $+\infty$.*
5. *Os pontos de abscissas $\mu - \sigma$ e $\mu + \sigma$ são pontos de inflexão.*



Observação 5.2.1.



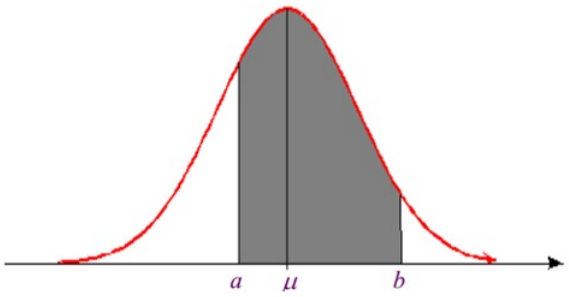
Em relação à essas distribuições normais, temos que:

(i) da distribuição A para B, muda a tendência central, mas a variabilidade é constante;

(ii) da distribuição A para C, muda a variabilidade, mas a tendência central é constante;

(iii) da distribuição B para C, muda a tendência central e a variabilidade.

Observação 5.2.2. Já vimos que para calcular a probabilidade $P(a \leq X \leq b)$ devemos calcular a área sob a curva da função densidade de probabilidade f.



Porém, o cálculo dessa área é muito complexo, então veremos a seguir como calcular essas probabilidades de forma mais simples com o auxílio de uma tabela.

Observação 5.2.3.

1. Se $X \sim N(\mu, \sigma^2)$, então utilizando a mudança de variável $Z = \frac{X-\mu}{\sigma}$, temos que $Z \sim N(0,1)$. A distribuição normal cuja média é zero e o desvio padrão é um é denominada distribuição normal padrão (ou reduzida).

2. Existe uma tabela que apresenta as probabilidades para vários possíveis valores de Z e no exemplo a seguir veremos como utilizá-la.

Tabela – Distribuição normal – valores de $P(0 \leqslant Z \leqslant z_0)$

z_0	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	0,0753
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	0,2123	0,2157	0,2190	0,2224
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,7	0,2580	0,2611	0,2642	0,2673	0,2703	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	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

Exemplo 5.2.1. Seja uma variável aleatória $X \sim N(20, 16)$, ou seja, X é uma variável aleatória contínua com distribuição normal de média $\mu = 20$ e variância $\sigma^2 = 16$. Calcule:

(a) $P(20 \leq X \leq 23)$

(b) $P(16 \leq X \leq 20)$

(c) $P(16 \leq X \leq 23)$

(d) $P(X \leq 16)$

(e) $P(X \geq 23)$

Solução:

(a) $P(20 \leq X \leq 23) = P(0 \leq Z \leq 0,75) = 0,2734$

(b) $P(16 \leq X \leq 20) = P(-1 \leq Z \leq 0) = P(0 \leq Z \leq 1) = 0,3413$

(c) $P(16 \leq X \leq 23) = P(-1 \leq Z \leq 0,75) = P(-1 \leq Z \leq 0) + P(0 \leq Z \leq 0,75) = 0,6147$

(d) $P(X \leq 16) = P(Z \leq -1) = 0,5 - P(-1 \leq Z \leq 0) = 0,1587$

(e) $P(X \geq 23) = P(Z \geq 0,75) = 0,5 - P(0 \leq Z \leq 0,75) = 0,2266$

Exemplo 5.2.2. As alturas dos alunos de uma determinada escola são normalmente distribuídas com média $1,60$ m e desvio padrão $0,30$ m. Encontre a probabilidade de um aluno medir:

(a) Entre $1,50$ e $1,80$ m.

(b) Mais de $1,75$ m.

(c) Menos de $1,48$ m.

(d) Qual deve ser a medida mínima para escolhermos 10% dos mais altos?

Exercícios

1. Calcule as probabilidades abaixo:

   (a) $P(0 \leq Z \leq 1)$
   (b) $P(-2,55 \leq Z \leq 1,2)$
   (c) $P(Z \geq 1,93)$
   (d) $P(Z > 1)$
   (e) $P(Z < 1)$
   (f) $P(Z > -0,34)$
   (g) $P(0 < Z < 1,5)$
   (h) $P(-2,88 < Z < 0)$
   (i) $P(-0,56 < Z < -0,20)$
   (j) $P(-0,49 < Z < 0,49)$

2. Uma variável aleatória contínua X apresenta distribuição normal com média 40 e desvio padrão igual a 3. Determine os valores de X para os seguintes valores de Z:

   (a) $0,10$
   (b) $2,00$
   (c) $0,75$
   (d) $-2,53$
   (e) $-3,00$
   (f) $-3,20$

3. Supondo que a altura X de um estudante do sexo masculino, tomado ao acaso de uma universidade, tenha distribuição normal com média 170 cm e desvio padrão 10 cm. Calcule:

   (a) $P(X > 190)$
   (b) $P(150 < X < 190)$
   (c) $P(X \leq 160)$

4. Suponha que em certa região, o peso dos homens adultos tenha distribuição normal com média 70 kg e desvio padrão 16 kg. E o peso das mulheres adultas tenha distribuição normal com média 60 kg e desvio padrão 12 kg. Ao selecionar uma pessoa ao acaso, o que é mais provável: uma mulher com mais de 75 kg, ou um homem com mais de 90 kg?

5. A concentração de um poluente em água liberada por uma fábrica tem distribuição normal de média 8 ppm e desvio padrão $1,5$ ppm. Qual a chance, de que num dado dia, a concentração do poluente exceda o limite regulatório de 10 ppm?

6. O salário semanal dos operários industriais são distribuídos normalmente em torno de uma média de R$ 180,00 com desvio padrão de R$ 25,00. Pede-se:

(a) Encontre a probabilidade de um operário ter salário semanal situado entre R$ 150,00 e R$ 178,00.

(b) Dentro de que desvio de ambos os lados da média cairão 96% dos salários?

7. A duração de um pneu de automóvel, em quilômetros rodados, apresenta distribuição normal com média 70000 km, e desvio-padrão de 10000 km. Com isso:

   (a) Qual a probabilidade de um pneu escolhido ao acaso durar mais de 85000 km?

   (b) Qual a probabilidade de um pneu durar entre 68500 km e 75000 km?

   (c) Qual a probabilidade de um pneu durar entre 55000 km e 65000 km?

   (d) O fabricante deseja fixar uma garantia de quilometragem, de tal forma que, se a duração de um pneu for inferior à da garantia, o pneu será trocado. De quanto deve ser essa garantia para que somente 1% dos pneus sejam trocados?

8. Uma empresa produz televisores de dois tipos, tipo A (comum) e tipo B (luxo), e garante a restituição da quantia paga se qualquer televisor apresentar defeito grave no prazo de seis meses. O tempo para ocorrência de algum defeito grave nos televisores tem distribuição normal sendo que, no tipo A, com média de 10 meses e desvio padrão de 2 meses e no tipo B, com média de 11 meses e desvio padrão de 3 meses. Os televisores de tipo A e B são produzidos com lucro de 1200 u.m. e 2100 u.m. respectivamente e, caso haja restituição, com prejuízo de 2500 u.m. e 7000 u.m., respectivamente. Calcule as probabilidades de haver restituição nos televisores do tipo A e do tipo B.

9. Suponha que as medidas da corrente elétrica em pedaço de fio sigam a distribuição Normal, com uma média de 10 miliamperes e uma variância de 4 miliamperes.

   (a) Qual a probabilidade de a medida exceder 13 miliamperes?

   (b) Qual a probabilidade de a medida da corrente estar entre 9 e 11 miliamperes?

   (c) Determine o valor para o qual a probabilidade de uma medida da corrente estar abaixo desse valor seja 0,98.

Capítulo 6

Introdução a Inferência Estatística

Sabemos que a Inferência Estatística é o estudo de técnicas que possibilitam a realização de conclusões à respeito de uma população a partir do estudo de amostras, tendo por base o cálculo das probabilidades. Esse procedimento é realizado a todo momento no dia-a-dia das pessoas, por exemplo, quando vamos comprar uma melancia, geralmente, o vendedor disponibiliza uma fatia para que as pessoas possam experimentar o produto e, baseado no sabor da melancia à amostra, o comprador decide se compra ou não.

Além disso, quando se realiza uma pesquisa estatística, na prática, temos alguma ideia sobre a distribuição em estudo, mas não o valor exato. Por exemplo, é razoável supor que as alturas dos brasileiros adultos pode ser representada por um modelo normal, na qual precisamos da média e da variância para determinar a distribuição. Assim, se conseguíssemos medir as alturas de todos os brasileiros adultos, não seria necessário a inferência, entretanto, isso é inviável e o propósito da pesquisa seria descobrir/estimar esses parâmetros.

Por outro lado, mesmo sendo "possível" fazer a medição de um certo parâmetro de todos os elementos da população, esse processo pode ser destrutivo: para medir a duração de uma certa lâmpada, poderíamos tomar todas as lâmpadas (do tipo específico em estudo) do Brasil e fazer a medição do tempo de duração de cada uma delas até que elas parem de funcionar, ou seja, ao final do processo não sobraria lâmpadas no Brasil. Logo, a solução seria tomar uma amostra aleatória, analisá-la e depois inferir propriedades para todas.

Também, podemos aplicar as ideias da inferência em outras ocasiões, onde a população é a função de probabilidade (ou função densidade de probabilidade) de uma variável aleatória. Por exemplo, suponha que temos uma moeda e desconfiamos da honestidade dela. Assim, vamos lançar essa moeda 50 vezes e atribuir a variável aleatória X: "número de caras depois dos 50 lançamentos". Assumimos que $X \sim B(50, p)$. Se ao final dos lançamentos foram obtidas 36 caras, então existe alguma evidência de que a moeda é honesta? Inicialmente, partimos do princípio de que a moeda é honesta, ou seja, $p = q = \frac{1}{2}$. Logo, podemos obter as probabilidades e esses valores podem ajudar a tomar a decisão a respeito da moeda. Se a decisão fosse rejeitar a honestidade da moeda, isto é, há indícios de que a moeda não é honesta, qual seria uma estimativa para o valor de p (baseado nos resultados obtidos)?

O exemplo anterior da moeda indica duas situações que podem ocorrer em relação à Inferência Estatística: (i) testar se uma determinada hipótese é aceita ou rejeitada (Teste de Hipótese); (ii) estimar parâmetros (Estimação). Nessas notas de aula vamos dar maior enfoque nos Testes de Hipóteses em certas situações.

6.1 Conceitos básicos dos Testes de hipóteses

Um dos principais objetivos da estatística é a tomada de decisões a respeito da população, com base na observação de amostras, ou seja, obtenção de conclusões válidas para toda a população com base em amostras retiradas dessas populações. Para tanto, torna-se necessário a formulação de suposições relativas às populações. Essas suposições, que podem ou não ser verdadeiras, são chamadas de hipóteses estatísticas e constituem, geralmente, em considerações a respeito das distribuições de probabilidade das populações. É muito comum a formulação de uma hipótese estatística com o objetivo de rejeitá-la.

Definição 6.1.1. Hipótese estatística é uma suposição feita a respeito de um ou mais parâmetros (μ, σ^2, σ, etc.) da população.

Para exemplificar a ideia que iremos empregar ao longo dessa seção, imagine que um agrônomo deseja realizar um experimento com o objetivo de verificar qual dos métodos A ou B de cultivo do milho resulta na maior qualidade do mesmo. O pesquisador formula a hipótese de que não existem diferenças entre os métodos em relação à qualidade (a qualidade do milho com o método A é igual a método B). Essa hipótese inicial formulada é denominada hipótese de nulidade (ou hipótese nula), representada por H_0.

Para testar essa hipótese é coletada uma amostra aleatória representativa de cada população, sendo calculadas as estatísticas necessárias para o teste. Naturalmente, devido ao fato de ser utilizada uma amostra aleatória, haverá diferenças entre o que se esperava e o que realmente foi obtido na amostra. A questão a ser respondida é: as diferenças são significativas o bastante para que a hipótese H_0 seja rejeitada? Esta não é uma pergunta simples de responder: dependerá do teste aplicado, da confiabilidade desejada para o resultado, entre outros fatores.

Se o agrônomo verificar que os resultados obtidos na pesquisa com o milho diferem acentuadamente dos resultados esperados para a hipótese H_0, deve-se concluir que as diferenças observadas são significativas e rejeita-se a hipótese H_0. Ao rejeitá-la, aceitamos outra hipótese, denominada hipótese alternativa e é representada por H_1. No nosso exemplo, H_1: os métodos de cultivos A e B testados diferem entre si em relação à produção de milho.

Definição 6.1.2. Existem dois tipos de hipóteses estatísticas:

(i) Hipótese de nulidade (H_0): é a hipótese a ser testada e refere-se sempre a um valor de um parâmetro da população (contém um dos símbolos $=$, $\leq$ ou $\geq$).

(ii) Hipótese alternativa (H_1): é o oposto da hipótese nula, ou seja, é a conclusão que se chegaria quando rejeita-se a hipótese nula (nunca contém os símbolos $=$, $\leq$ ou $\geq$).

Exemplo 6.1.1. (a) Suponha que a variável de estudo é "a pressão arterial" e queremos determinar se um medicamento é eficaz no controle da pressão arterial. Além disso, temos duas populações: A: hipertensos com uso de medicamento e a população B: hipertensos sem o uso de medicamento. Podemos considerar H_0: $\mu_A = \mu_B$ e H_1: $\mu_A < \mu_B$ (teste unilateral).

(b) Considere a variável de estudo "nota dos alunos" e queremos determinar se o método de ensino A é melhor que o método de ensino B. Dadas as populações: A: alunos ensinados pelo método A e B: alunos ensinados pelo método B. Podemos considerar H_0: $\mu_A = \mu_B$ e H_1: $\mu_A \neq \mu_B$ (teste bilateral).

Os métodos que permitem decidir se uma hipótese deve ser aceita ou rejeitada são denominados testes de hipóteses ou testes de significância. Porém, ao tomarmos decisões de rejeitar ou aceitar uma determinada hipótese, estamos sujeitos a cometer dois tipos de erro: rejeitar uma hipótese nula verdadeira (erro tipo I) ou aceitar uma hipótese nula falsa (erro tipo II).

Decisão	Situação real	
	H_0 verdadeira	H_0 falsa
Aceitar H_0	Decisão correta	Erro tipo II
Rejeitar H_0	Erro tipo I	Decisão correta

Definição 6.1.3. Teste de hipóteses é uma regra de decisão para aceitar, ou rejeitar, uma hipótese estatística com base nos elementos de uma amostra.

Observação 6.1.1. Vamos denotar por α o erro tipo I e por β o erro tipo II. As duas taxas de erro α e β estão relacionadas de modo que a redução de α implica no aumento de β e vice-versa. O único meio de reduzir ambos os tipos de erro é aumentando o tamanho da amostra, o que nem sempre é viável. Em geral, a preocupação está voltada para o erro tipo I, pois na maioria dos casos ele é considerado o mais grave.

Devemos tomar como H_0 aquela hipótese, que, rejeitada, conduza a um erro de tipo I mais importante de evitar. Vejamos um exemplo (Neyman): suponha um experimento para se determinar se um produto A é ou não cancerígeno. Após realizado o teste, podemos concluir: (i) A é cancerígeno ou (ii) A não é cancerígeno. Cada uma dessas conclusões pode estar errada e temos os dois tipos de erro já mencionados, dependendo de qual hipótese seja H_0. Do ponto de vista do usuário do produto, a hipótese a ser testada deve ser

$$H_0: \ A \text{ é cancerígeno}$$

pois a probabilidade de erro na rejeição dessa hipótese, se ela for verdadeira, deve ser um valor muito pequeno.

Definição 6.1.4. O nível de significância do teste, representado por α, é a probabilidade máxima aceitável de cometer um erro do tipo I. É estabelecido antes do teste por quem o realiza e, geralmente, é fixamos em 5% ($\alpha = 0,05$) ou em 1% ($\alpha = 0,01$). Assim,

$$\alpha = P(\text{Erro do tipo I}) = P(\text{Rejeitar } H_0 \,|\, H_0 \text{ é verdadeira}).$$

Observação 6.1.2. Se utilizarmos o nível de significância de 5%, temos 5 chances em 100 de rejeitarmos uma hipótese que deveria ser aceita, isto é, há uma confiança de 95% de que tenhamos tomado uma decisão correta.

Para aplicar um teste de hipótese, vamos sempre seguir os seis passos:

(i) Definir as hipóteses H_0 e H_1.

(ii) Fixar o nível de significância α (taxa de erro aceitável).

(iii) Use a teoria e as informações disponíveis para decidir qual estatística (estimador) será usada para testar a hipótese H_0.

(iv) Determinar a região crítica e a região de aceitação em função de α pelas tabelas estatísticas apropriadas.

(v) Baseando na amostra, calcular o valor da estatística de teste (valor calculado a partir da amostra e que é usado para tomar a decisão acerca de rejeitar ou não a hipótese nula).

(vi) Concluir: se a estatística de teste pertencer a região crítica, rejeita H_0. Caso contrário, aceita H_0.

Vamos estudar duas situações comuns em testes de hipóteses com respeito a média populacional: comparação de uma média (μ) com um valor padrão (μ_0) quando σ^2 é conhecida ou $n > 30$ e quando σ^2 é desconhecida e $n \leq 30$.

6.2 Testes de hipóteses com respeito a μ

6.2.1 σ^2 é conhecida ou $n > 30$

Vamos considerar que a variável em estudo tem distribuição normal e a variância σ^2 é conhecida ou $n > 30$. A hipótese sob verificação é H_0: $\mu = \mu_0$ e a estatística do teste é

$$Z = \frac{\overline{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}},$$

onde Z possui distribuição normal padrão.

Exemplo 6.2.1. A associação dos proprietários de indústrias metalúrgicas está preocupada com o tempo perdido em acidentes de trabalho, cuja média, nos últimos tempos, tem sido da ordem de 60 horas/homem por ano com desvio padrão de 20 horas/homem, segundo a distribuição normal. Tentou-se um programa de prevenção de acidentes e, após o mesmo, tomou-se uma amostra de 9 indústrias e mediu-se o número de horas/homem perdidas por acidente, que foi de 50 horas. Você diria, ao nível de 5%, que há evidência de melhora?

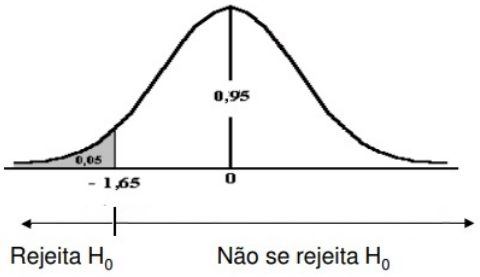
Solução:

(i) H_0: $\mu = 60$ e H_1: $\mu < 60$

(ii) $\alpha = 0,05$

(iii) A variável em estudo tem distribuição normal.

(iv)



(v) $z = \dfrac{50 - 60}{20/\sqrt{9}} = -1,50$

(vi) Conclusão: Não é possível, ao nível de 5% de significância, afirmar que a campanha deu resultado.

Exemplo 6.2.2. Um fabricante informa que a duração média da vida de um equipamento é 500 horas com desvio padrão de 5 horas, segundo a distribuição normal. Foram amostradas 100 desses equipamentos, obtendo-se média de 498 horas. Há evidências suficientes para rejeitar a afirmação do vendedor, com um nível de confiança de 95%?

Solução:

(i) H_0: $\mu = 500$ e H_1: $\mu \neq 500$

(ii) $\alpha = 0,05$

(iii) A variável em estudo tem distribuição normal.

(iv)

(v) $z = \dfrac{498 - 500}{5/\sqrt{100}} = -4$

(vi) Conclusão: rejeitamos H_0 a 5% de significância. Ou seja, há evidências suficientes para rejeitar a afirmação do vendedor.

As situações acima não são muito realistas, pois em geral, não é conhecido o valor da variância da população.

6.2.2 σ^2 é desconhecida e $n \leq 30$

Nesse caso, assumimos que a variável de estudo tem distribuição normal, variância σ^2 desconhecida e $n \leq 30$. A hipótese sob verificação é H_0: $\mu = \mu_0$ e a estatística do teste é

$$T = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} \sim t_v,$$

onde $v = n - 1$ é o grau de liberdade[1] e t possui uma distribuição t-Student[2]. Utilizamos a distribuição t para determinar a região de rejeição e aceitação.

Unicaudal	75%	80%	85%	90%	95%	97,50%	99%	99,50%
Bicaudal	50%	60%	70%	80%	90%	95%	98%	99%
gl								
1	1,000	1,376	1,963	3,078	6,314	12,710	31,820	63,660
2	0,816	1,061	1,386	1,886	2,920	4,303	6,965	9,925
3	0,765	0,978	1,250	1,638	2,353	3,182	4,541	5,841
4	0,741	0,941	1,190	1,533	2,132	2,776	3,747	4,604
5	0,727	0,920	1,156	1,476	2,015	2,571	3,365	4,032
6	0,718	0,906	1,134	1,440	1,943	2,447	3,143	3,707
7	0,711	0,896	1,119	1,415	1,895	2,365	2,998	3,499
8	0,706	0,889	1,108	1,397	1,860	2,306	2,896	3,355
9	0,703	0,883	1,100	1,383	1,833	2,262	2,821	3,250
10	0,700	0,879	1,093	1,372	1,812	2,228	2,764	3,169

Exemplo 6.2.3. Um processo deveria produzir bancadas com 0,85 m de altura. O engenheiro desconfia que as bancadas que estão sendo produzidas são diferentes que o especificado. Uma amostra de 8 valores foi coletada e indicou média de 0,87 m e desvio padrão de 0,01 m. Sabendo-se que os dados seguem a distribuição normal, teste a hipótese do engenheiro usando um nível de significância $\alpha = 0,05$.

Solução:

(i) H_0: $\mu = 0,87$ e H_1: $\mu \neq 0,87$

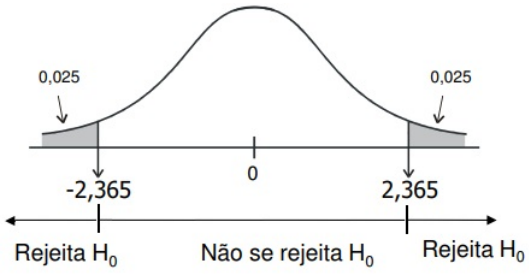
(ii) $\alpha = 0,05$

(iii) A variável em estudo tem distribuição normal.

(iv)

[1] Ideia do grau de liberdade: suponha que a média de 3 números seja igual a 8. Se $x_1 = 7$ e $x_2 = 8$ qual deve ser o valor do terceiro número (x_3)? Se a média é 8, o terceiro valor só pode ser 9, isto é, x_3 não é livre para variar. Neste caso, $n = 3$ e o grau de liberdade é $v = n - 1 = 3 - 1 = 2$ (dois números podem assumir qualquer valor, mas o terceiro não está livre para variar dada uma certa média).

[2] Quando a variância populacional é desconhecida e $v > 30$ podemos usar a distribuição normal substituindo o desvio padrão populacional pelo desvio padrão amostral, pois a distribuição t é próxima da normal.



(v) $t = \dfrac{0,87 - 0,85}{0,01/\sqrt{8}} = 5,66$

(vi) Conclusão: Rejeita-se H_0, ou seja, ao nível de 5% de significância, conclui-se que as bancadas que estão sendo produzidas devem ter altura diferente do especificado, maiores que $0,85$m.

Exemplo 6.2.4. A associação dos proprietários de indústrias metalúrgicas está muito preocupada com o tempo perdido com acidentes de trabalho, cuja média nos últimos tempos, tem sido da ordem de 60 horas/homem por ano. Tentou-se um programa de prevenção de acidentes, após o qual foi tomada uma amostra de 9 indústrias e medido o número de horas/homens perdidas por acidente, que foi de 50 horas e desvio padrão de 20 horas/homem. Você diria, no nível de 5%, que há evidência de melhoria?

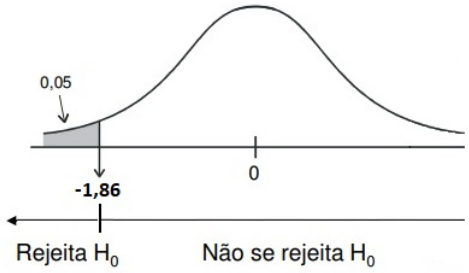
Solução:

(i) H_0: $\mu = 60$ e H_1: $\mu < 60$

(ii) $\alpha = 0,05$

(iii) A variável em estudo tem distribuição normal.

(iv)



(v) $t = \dfrac{50 - 60}{20/\sqrt{9}} = -1,5$

(vi) Conclusão: (não rejeitamos H_0) Não há evidência de melhora, para ter melhoria, teria que cair na região de rejeição.

Exercícios

1. Uma máquina automática para encher pacotes de café enche-os segundo uma distribuição normal, com média μ e desvio padrão igual a 20 g. A máquina foi regulada para $\mu = 500$ g. Desejamos, periodicamente, colher uma amostra de 16 pacotes e verificar se a produção está sob controle, isto é, $\mu = 500$ g ou não. Se uma dessas amostras apresentasse uma média $\overline{x} = 492$ g, você pararia ou não para regular a máquina? Use $\alpha = 1\%$.

2. Determinada firma desejava comprar cabos tendo recebido do fabricante a informação de que a tensão média de ruptura é 8000 Kgf. Para analisar se a afirmação do fabricante é verdadeira, efetuou-se um teste de hipótese unilateral. Se um ensaio com 6 cabos forneceu uma tensão média de ruptura de 7750 Kgf, com desvio padrão de 145 Kgf, a qual conclusão chegar, usando nível de significância de 5%?

3. Uma fábrica anuncia que o índice de nicotina dos cigarros da marca X apresenta-se abaixo de 26 mg por cigarro. Um laboratório realiza 10 análises do índice obtendo: 26, 24, 23, 22, 28, 25, 27, 26, 28 e 24. Sabe-se que o índice de nicotina dos cigarros da marca X se distribui normalmente com variância 5,36 mg^2. Pode-se aceitar a afirmação do fabricante, ao nível de 5%?

4. Uma fábrica de baterias alega que as mesmas têm vida média de 50 meses. Sabe-se que o desvio padrão correspondente é de 4 meses. Se uma amostra de 36 baterias, obtida dessa população, tem vida média de 48 meses, podemos afirmar que a média dessa população é diferente de 50 meses, ao nível de significância de 5%?

5. Um fabricante de lajotas de cerâmica introduz um novo material em sua fabricação e acredita que aumentará a resistência média que é de 206 kg. A resistência das lajotas tem distribuição normal com desvio padrão de 12 kg. Retira-se uma amostra de 30 lajotas, obtendo $\overline{x} = 210$ kg. Ao nível de 10%, pode o fabricante aceitar que a resistência média de suas lajotas tenha aumentado?

6. Uma determinada escola da cidade está treinando seus alunos para participar dos jogos estatuais olímpicos, sendo que a média da impulsão vertical dos alunos é $\mu = 48$ cm. Com o objetivo de melhorar a impulsão vertical, 50 alunos dessa mesma escola foram submetidos a um programa de treino de força explosiva. Os resultados obtidos no final do programa de treino mostraram valores de impulsão vertical de $\overline{x} = 51,5$ cm e $s = 5,92$ cm. Teste a hipótese de haver uma melhoria significativa ($\alpha = 0,05$) nos resultados destes alunos. (Use o teste t).

7. O peso médio de litros de leite de embalagens enchidas em uma linha de produção está sendo estudado. O padrão prevê um conteúdo médio de 1000 ml por embalagem. Sabe-se que o desvio padrão é de 10 ml e que a variável tem distribuição normal. Qual a probabilidade da média ser diferente de 1000 ml ao nível de 5% de significância com 4 unidades amostrais.

8. Uma fábrica de automóveis anuncia que seus carros consomem em média 9,4 km/l, com desvio padrão de 0,79 litros. Uma revista com base em resultados preliminares desconfia da afirmação acima do fabricante e resolve testar tal hipótese analisando 32 automóveis dessa marca, obtendo 9 km/l como consumo médio. O que a revista pode concluir sobre o anúncio da fábrica ao nível de significância de 5%?

9. Um fabricante afirma que seus cigarros contêm não mais que 30 mg de nicotina. Uma amostra de 31 cigarros fornecem média de 31,5mg e desvio padrão de 3mg. Ao nível de significância de 5%, os dados refutam ou não a afirmação do fabricante?

10. Em uma encosta sanitária de certa comunidade, entrevistaram-se 150 pessoas. Um dos detalhes de informação obtida foi o número de receitas médicas que cada pessoa havia pedido durante o ano anterior. O número médio para as 159 pessoas foi de 5,8 com um desvio de 3,1. O investigador deseja saber se esses dados proporcionam evidência suficiente para indicar que a média da população é maior que 5. Considere $\alpha = 5\%$.

11. Em um estudo de hábitos de consumidores, pesquisadores elaboraram um questionário para identificar os compradores compulsivos. Para uma amostra de consumidores que se declararam compradores compulsivos, os resultados dos questionários acusaram média de 0,83 e desvio-padrão de 0,24. Suponha uma amostra de 32 indivíduos selecionados aleatoriamente. No nível de 0,01 de significância, teste a afirmação de que a população dos que se identificam como compradores compulsivos têm média diferente de 0,21.

12. O gerente de um banco presume que a renda média anual de seus clientes é no máximo R$ 3.800,00. Uma amostra aleatória de 60 clientes acusou uma média de R$ 3.950,00 e um desvio padrão de R$ 300,00. Considerando um nível de significância de 2,5%, pode-se dizer que a renda média anual dos clientes desse gerente é superior a R$ 3.800,00?

13. Em fevereiro de 2016, o custo médio para um voo doméstico com passagens de ida e volta com desconto foi de R$ 290,00. Uma amostra aleatória dos preços de 15 passagens de ida e volta com desconto durante o mês de mar¸co forneceu os seguintes valores:

310	260	265	255	300
310	230	250	265	280
290	240	285	250	260

Usando $\alpha = 5\%$ de significância, pode-se dizer que o preço médio da passagem de ida e volta, com desconto, diminui em março, em relação a fevereiro?

14. Numa indústria de autopeças, sabe-se que o nível de dureza de um produto feito a base de cerâmica tem variabilidade $\sigma^2 = 0,49$. Uma amostra de 16 peças foram testadas e o resultado é apresentado abaixo:

18,1	19,0	18,8	18,5	18,1	18,8	18,1	18,0
18,5	19,8	17,8	19,1	18,0	19,2	19,8	19,2

Com um nível de significância de 10%, pede-se:

(a) pode-se afirmar que a média do nível de dureza é superior a 18,4?

(b) testar a hipótese bicaudal de que a média é igual a 18,4.

15. A associação dos proprietários de indústrias metalúrgicas está muito preocupada com o tempo perdido com acidentes de trabalho, cuja média, nos últimos tempos, tem sido da ordem de 60 horas/homem por ano e desvio padrão de 20 horas/homem. Tentou-se um programa de prevenção de acidentes, após o qual foi tomada uma amostra de nove indústrias e medido o número de horas/homem perdidas por acidente, que foi de 50 horas. Você diria no nível de 5%, que há evidências de melhoria?

16. O número de pontos de um exame de inglês tem sido historicamente ao redor de 80. Sorteamos 10 estudantes que fizeram recentemente esse exame e observamos as notas: 65, 74, 78, 86, 59, 84, 75, 72, 81 e 83. Especialistas desconfiam que a média diminuiu e desejam testar essa afirmação através de um teste de hipóteses, com nível de significância de 5%. Fazendo as suposições necessárias qual seria a conclusão do teste? Quais suposições são necessárias para a realização do teste realizado?

17. A resistência de um certo tipo de cabo de aço é uma variável aleatória modelada pela distribuição Normal com desvio padrão igual a 6 kgf. Uma amostra de tamanho 25 desses cabos, escolhida ao acaso, forneceu média igual a 9,8 kgf. Teste as hipóteses $\mu = 13$ versus $\mu = 8$ e tire suas conclusões a um nível de significância de 10%.

18. Um criador tem constatado uma proporção de 10% do rebanho com verminose. O veterinário alterou a dieta dos animais e acredita que a doença diminuiu de itensidade. Um exame em 100 cabeças do rebanho, escolhidas ao acaso, indicou 8 delas com verminose. Ao nível de significância de 8%, há indícios de que a proporção diminuiu?

Capítulo 7

Regressão linear e correlação

Em pesquisas, frequentemente, procura-se verificar se existe relação entre duas (ou mais) variáveis, isto é, saber se as alterações sofridas por uma das variáveis são acompanhadas por alterações nas outras. Por exemplo, peso vs. idade; uso de cigarro vs. incidência do câncer vs. Idade; o consumo de uma família vs. renda; a demanda de um determinado produto vs. preço. A verificação da existência e do grau de relação entre variáveis é o objetivo do estudo da correlação.

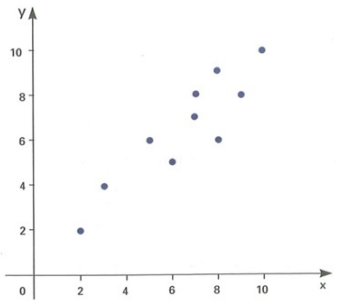
Uma vez caracterizada, procura-se descrever uma relação sob forma matemática, através de uma função. A estimação dos parâmetros dessa função matemática é o objetivo da regressão. Ficaremos restritos às relações entre duas variáveis (correlação/regressão simples).

7.1 Diagrama de dispersão

Considere uma amostra formada por dez dos 98 alunos do curso de Física do IFNMG e pelas notas obtidas por eles em Matemática e Estatística:

Nºs	NOTAS	
	MATEMÁTICA (X)	ESTATÍSTICA (Y)
01	5,0	6,0
08	8,0	9,0
24	7,0	8,0
38	10,0	10,0
44	6,0	5,0
58	7,0	7,0
59	9,0	8,0
72	3,0	4,0
80	8,0	6,0
92	2,0	2,0

Representando em um sistema coordenado cartesiano os pares (X, Y), obtemos uma nuvem de pontos que denominados diagrama de dispersão. Este diagrama nos fornece uma ideia inicial da correlação existente:



Observação 7.1.1. A análise gráfica da relação entre variáveis é importante, mas os olhos nem sempre são um bom juiz para avaliar a intensidade de uma relação. Nossos olhos podem ser enganados por uma mudança de escalas, ou pela quantidade de espaço em branco em torno do aglomerado dos pontos. Deve-se, então, utilizar uma medida numérica para suplementar o gráfico, chamada coeficiente de correlação de Pearson.

7.2 Correlação linear simples

A correlação linear procura medir a relação entre duas variáveis X e Y através da disposição dos pontos (X, Y) em torno de uma reta. O instrumento de medida da correlação linear é dado pelo coeficiente de correlação de Pearson:

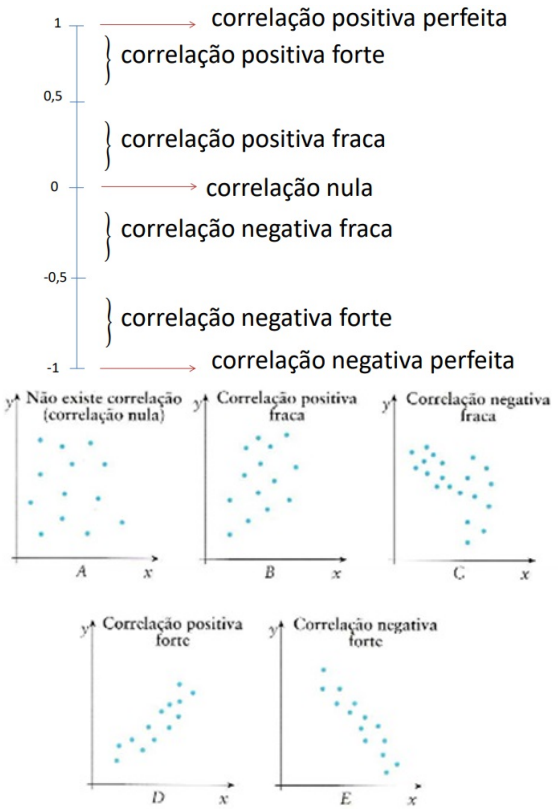
$$r_{XY} = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sqrt{\left[\sum X^2 - \frac{(\sum X)^2}{n}\right]\left[\sum Y^2 - \frac{(\sum Y)^2}{n}\right]}},$$

onde n é o número de observações.

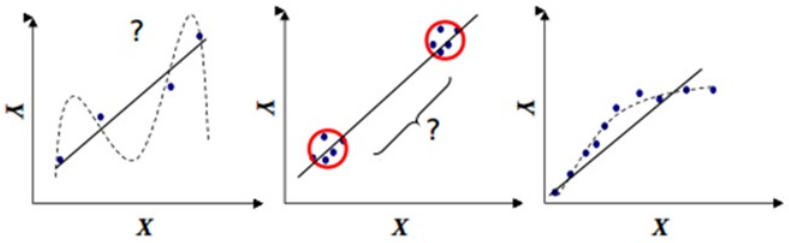
Observação 7.2.1. (i) O valor de r_{XY} estará sempre no intervalo de -1 a 1 e sua interpretação dependerá do valor numérico e do sinal.

(ii) r_{XY} só mede a intensidade ou grau de relacionamentos lineares. Não serve para medir intensidade de relacionamentos não lineares.

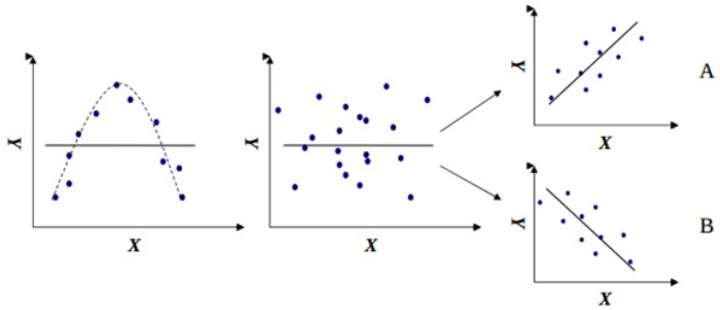
(iii) Conhecido o valor de r_{XY}, avalia-se a intensidade da correlação de acordo com o a seguinte escala:



(iv) Um alto coeficiente de correlação nem sempre indica que a equação de regressão estimada está bem ajustada aos dados.



(v) Um coeficiente de correlação próximo de zero nem sempre indica que X e Y não são relacionadas.



Exemplo 7.2.1. Calcule o coeficiente de correlação linear entre as variáveis X e Y, usando os dados da tabela abaixo.

Y	10	8	6	10	12
X	2	4	6	8	10

Solução: Para auxiliar nos cálculos, construímos a tabela abaixo:

Y	X	X^2	Y^2	XY
10	2	4	100	20
8	4	16	64	32
6	6	36	36	36
10	8	64	100	80
12	10	100	144	120
Σ=46	Σ=30	Σ=220	Σ=444	Σ=288

Assim, $r_{XY} = 0,416$. Este resultado mostra que a correlação linear entre as variáveis X e Y é positiva, porém fraca.

Exemplo 7.2.2. Calcule o coeficiente de correlação linear entre a renda familiar e a poupança das dez famílias segundo a tabela:

FAMÍLIAS	RENDA (R$) (x100)	POUPANÇA (R$) (x1000)	NÚMERO DE FILHOS	MÉDIA DE ANOS DE ESTUDO DA FAMÍLIA
A	10	4	8	3
B	15	7	6	4
C	12	5	5	5
D	70	20	1	12
E	80	20	2	16
F	100	30	2	18
G	20	8	3	8
H	30	8	2	8
I	10	3	6	4
J	60	15	1	8

Exercícios

1. Considerando os dados da tabela do Exemplo 7.2.2, calcule o coeficiente de correlação linear entre:

   (a) Renda e número de filhos.

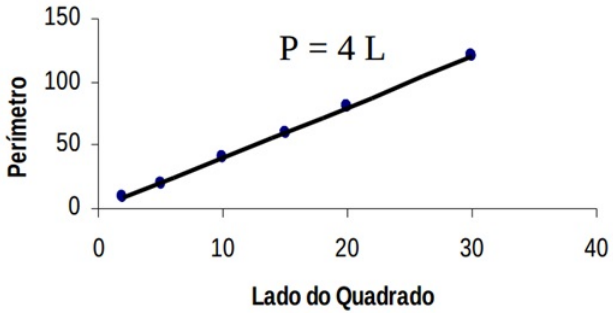
   (b) Poupança e número de filhos.

   (c) Média dos anos de estudo e número de filhos.

   (d) Renda familiar e média de anos de estudo.

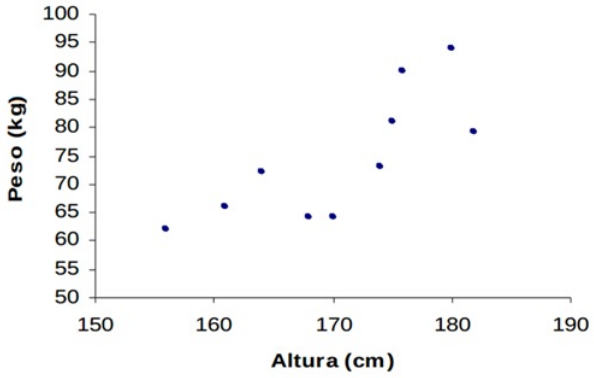
7.3 Regressão linear simples

Algumas variáveis possuem relação expressa por uma fórmula matemática $Y = f(X)$. Por exemplo, o perímetro (P) e o lado de um quadrado (L) estão relacionados segundo a função $P = 4L$. Os pares de valores das duas variáveis (perímetro e lado) podem ser colocados no diagrama de dispersão:



onde todos os pontos caem na curva da relação $P = 4L$.

Em outros casos, não há uma relação perfeita entre as variáveis em estudo e as observações, em geral, não caem exatamente na curva da relação. Por exemplo, considerando a relação entre o peso (P) e a altura (A) de um grupo de pessoas.



Porém, analisando o diagrama de dispersão, percebe-se que, na maioria das observações, quanto maior a altura, maior o peso, indicando a existência de uma relação entre a altura e peso.

A análise de regressão tem por objetivo descrever através de um modelo matemático, a relação existente entre duas variáveis, a partir de ?? observações das mesmas. A variável sobre a qual desejamos fazer uma estimativa recebe o nome de variável dependente e a outra recebe o nome de variável independente.

Sendo assim, considerando X a variável independente e Y a variável dependente, vamos determinar o ajustamento de uma reta entre estas variáveis: $\widehat{Y} = aX + b$, onde a e b são parâmetros calculados por

$$b = \frac{\sum Y}{n} - a\frac{\sum X}{n} \text{ e } a = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sum X^2 - \frac{(\sum X)^2}{n}},$$

onde n é o número de observações.

Exemplo 7.3.1. Os dados a seguir referem-se ao volume de precipitação pluviométrico (mm) e ao volume de produção de leite tipo C (milhões de litros), em determinada região do país.

ANOS	PRODUÇÃO DE LEITE C (1.000.000 L)	ÍNDICE PLUVIOMÉTRICO (mm)
1970	26	23
1971	25	21
1972	31	28
1973	29	27
1974	27	23
1975	31	28
1976	32	27
1977	28	22
1978	30	26
1979	30	25

(a) Ajuste os dados através de um modelo linear.

(b) Admitindo-se, em 1980, um índice pluviométrico de 24 mm, qual deverá ser o volume esperado de produção de leito tipo C?

Exercícios

1. Para cada uma das situações abaixo, diga o que é mais adequado: a análise de regressão ou a análise de correlação. Por quê?

   (a) Uma equipe de pesquisadores deseja determinar se o rendimento na Universidade sugere êxito na profissão escolhida.

   (b) Deseja-se estimar o número de quilômetros que um pneu radial pode rodar antes de ser substituído.

   (c) Deseja-se prever quanto tempo será necessário para executar uma determinada tarefa por uma pessoa, com base no tempo de treinamento.

   (d) Deseja-se verificar se o tempo de treinamento é importante para avaliar o desempenho na execução de uma dada tarefa.

   (e) Um gerente deseja estimar as vendas semanais com base nas vendas das segundas e terças-feiras.

2. Considere um experimento em que se analisa a octanagem da gasolina (Y) em função da adição de um novo aditivo (X). Para isso, foram realizados ensaios com os percentuais de 1, 2, 3, 4, 5 e 6% de aditivo. Os resultados são mostrados na tabela abaixo:

X	1	2	3	4	5	6
Y	80,5	81,6	82,1	83,7	83,9	85,0

(a) Construa o diagrama de dispersão que descreve os dados.

(b) Existe uma relação linear entre a adição de um novo aditivo e a octanagem da gasolina? Qual o grau dessa relação?

(c) Determine a reta de regressão que explica a octanagem da gasolina em função da adição do novo aditivo.

(d) Se adicionarmos 5,5% de aditivo, qual o índice de octanagem esperado?

3. Uma pesquisa foi realizada com sete alunos do IFNMG e foi pesquisado o número total de faltas e a nota final na disciplina de Estatística durante o segundo semestre de 2016. Os dados foram organizados na seguinte tabela:

Faltas	8	2	5	12	15	9	6
Nota final	78	92	90	58	43	74	81

(a) Faça um diagrama de dispersão que represente os dados da tabela.

(b) Existe correlação entre o número de faltas e a nota final deste grupo de alunos? De que forma?

4. A tabela a seguir mostra os resultados de uma pesquisa que apreciou o peso de um veículo (em ton) e o número médio de peças defeituosas que tiveram de ser repostas no primeiro ano de uso do automóvel. Pedem-se:

(a) o coeficiente de correlação linear e a equação de regressão;

(b) o diagrama de dispersão e o gráfico da equação de regressão linear.

P: peso do veículo (ton)	N: número de peças defeituosas
1,00	2
1,25	5
1,50	5
1,75	7
2,00	10
2,25	11
2,50	15

7.4 Coeficiente de determinação

O coeficiente de determinação, denominado R^2, é uma medida descritiva que avalia a "qualidade" do ajuste. Seu valor fornece a proporção da variação total da variável Y explicada pela variável X através da função ajustada e é calculado elevando o coeficiente de correlação de Pearson ao quadrado.

Observação 7.4.1. (i) $0 \leq R^2 \leq 1$.

(ii) Quanto mais próximo de 1 estiver o coeficiente de determinação, melhor será o grau de explicação da variação de Y em termos da variável X. Por exemplo, se $R^2 = 0,98 = 98\%$, isto significa que 98% das variações de Y são explicadas por X através da função escolhida para relacionar as duas variáveis e 2% são atribuídas a causas aleatórias.

Exemplo 7.4.1. Calcule o coeficiente de determinação no Exemplo 7.3.1.

Exercícios

1. Determine se existe correlação entre a altura e o peso de uma amostra de ursos siberianos segundo os dados abaixo:

Altura (pol.)	Peso (Ib.)
53,0	80
67,5	344
72,0	416
72,0	348
73,5	262
68,5	360
73,0	332
37,0	34

2. Suponha que uma população se constitua dos seis pontos seguintes: (1,2), (4,6), (2,4), (2,3), (3,5) e (5,10).

   (a) Grafe os pontos em um diagrama de dispersão.

   (b) Determine a equação de regressão.

3. Quantidade de procaína hidrolisada, em 10 moles/litro, no plasma humano, em função do tempo decorrido após sua administração.

Tempo (min)	Quantidade hidrolisada
2	3,5
3	5,7
5	9,9
8	16,3
10	19,3
12	25,7
14	28,2
15	32,6

   (a) Existe uma relação linear entre a quantidade de procaína e o tempo decorrido após sua administração? Qual o grau dessa relação?

   (b) Determine a reta de regressão que explica a quantidade de procaína em função do tempo. Calcule o coeficiente de determinação do modelo.

   (c) Qual a quantidade de procaína hidrolisada após 6 minutos de sua administração? E após 13 minutos?

4. Um pesquisador deseja verificar se um instrumento para medir a concentração de determinada substância no sangue está bem calibrado. Para isto, ele tomou 15 amostras de concentrações conhecidas (X) e determinou a respectiva concentração através do instrumento (Y), obtendo:

X	2,0	2,0	2,0	4,0	4,0	4,0	6,0	6,0	6,0	8,0	8,0	8,0
Y	2,1	1,8	1,9	4,5	4,2	4,0	6,2	6,0	6,5	8,2	7,8	7,7

   Existe correlação entre os dados? De que forma? Se sim, determine a reta de regressão.

5. A tabela a seguir mostra os resultados de uma pesquisa que apreciou o peso de um veículo (em ton) e o número médio de peças defeituosas que tiveram de ser repostas no primeiro ano de uso do automóvel. Pedem-se:

(a) o coeficiente de correlação linear e a equação de regressão;

(b) o diagrama de dispersão e o gráfico da equação de regressão linear;

(c) o poder explicativo da regressão.

P: peso do veículo (ton)	N: número de peças defeituosas
1,00	2
1,25	5
1,50	5
1,75	7
2,00	10
2,25	11
2,50	15

6. É esperado que a massa muscular de uma pessoa diminua com a idade. Para estudar essa relação, uma nutricionista selecionou 18 mulheres, com idade entre 40 e 79 anos, e observou em cada uma delas a idade (X) e a massa muscular (Y).

Massa muscular (Y)	Idade (X)
82.0	71.0
91.0	64.0
100.0	43.0
68.0	67.0
87.0	56.0
73.0	73.0
78.0	68.0
80.0	56.0
65.0	76.0
84.0	65.0
116.0	45.0
76.0	58.0
97.0	45.0
100.0	53.0
105.0	49.0
77.0	78.0
73.0	73.0
78.0	68.0

(a) Construa o diagrama de dispersão e interprete-o.

(b) Calcule o coeficiente de correlação linear entre X e Y e interprete-o.

(c) Ajuste uma reta de regressão para a relação entre as variáveis Y: massa muscular (dependente) e X: idade (independente).

(d) Considerando a reta estimada dada no item (c), estime a massa muscular média de mulheres com 50 anos.

(e) Calcule e interprete o poder explicativo da regressão.

7. Seja Y uma variável que representa o valor do frete rodoviário de determinada mercadoria e X a variável (em Km) ao destino da mercadoria. Uma amostra de 10 observações das variáveis apresentou os seguintes resultados:

$$n = 10, \sum X = 1200, \sum Y = 6480,5, \sum XY = 842060,$$
$$\sum Y^2 = 4713304,03, \sum X^2 = 186400.$$

(a) Determine a regressão: $\widehat{Y} = a + bX$.

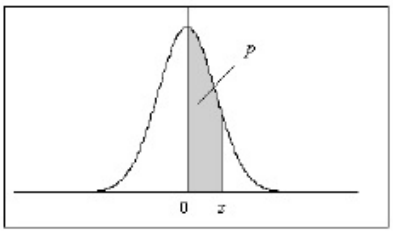
(b) Calcule e interprete o poder explicativo da regressão.

Tabela Z

Tabela
Distribuição normal padrão

Valores de p

$p = \Pr(0 \le Z \le z)$



Casa inteira e 1ª. Decimal	2ª. decimal									
	0	1	2	3	4	5	6	7	8	9
0,0	0,00000	0,00399	0,00798	0,01197	0,01595	0,01994	0,02392	0,02790	0,03188	0,03586
0,1	0,03983	0,04380	0,04776	0,05172	0,05567	0,05962	0,06356	0,06749	0,07142	0,07535
0,2	0,07926	0,08317	0,08706	0,09095	0,09483	0,09871	0,10257	0,10642	0,11026	0,11409
0,3	0,11791	0,12172	0,12552	0,12930	0,13307	0,13683	0,14058	0,14431	0,14803	0,15173
0,4	0,15542	0,15910	0,16276	0,16640	0,17003	0,17364	0,17724	0,18082	0,18439	0,18793
0,5	0,19146	0,19497	0,19847	0,20194	0,20540	0,20884	0,21226	0,21566	0,21904	0,22240
0,6	0,22575	0,22907	0,23237	0,23565	0,23891	0,24215	0,24537	0,24857	0,25175	0,25490
0,7	0,25804	0,26115	0,26424	0,26730	0,27035	0,27337	0,27637	0,27935	0,28230	0,28524
0,8	0,28814	0,29103	0,29389	0,29673	0,29955	0,30234	0,30511	0,30785	0,31057	0,31327
0,9	0,31594	0,31859	0,32121	0,32381	0,32639	0,32894	0,33147	0,33398	0,33646	0,33891
1,0	0,34134	0,34375	0,34614	0,34849	0,35083	0,35314	0,35543	0,35769	0,35993	0,36214
1,1	0,36433	0,36650	0,36864	0,37076	0,37286	0,37493	0,37698	0,37900	0,38100	0,38298
1,2	0,38493	0,38686	0,38877	0,39065	0,39251	0,39435	0,39617	0,39796	0,39973	0,40147
1,3	0,40320	0,40490	0,40658	0,40824	0,40988	0,41149	0,41309	0,41466	0,41621	0,41774
1,4	0,41924	0,42073	0,42220	0,42364	0,42507	0,42647	0,42785	0,42922	0,43056	0,43189
1,5	0,43319	0,43448	0,43574	0,43699	0,43822	0,43943	0,44062	0,44179	0,44295	0,44408
1,6	0,44520	0,44630	0,44738	0,44845	0,44950	0,45053	0,45154	0,45254	0,45352	0,45449
1,7	0,45543	0,45637	0,45728	0,45818	0,45907	0,45994	0,46080	0,46164	0,46246	0,46327
1,8	0,46407	0,46485	0,46562	0,46638	0,46712	0,46784	0,46856	0,46926	0,46995	0,47062
1,9	0,47128	0,47193	0,47257	0,47320	0,47381	0,47441	0,47500	0,47558	0,47615	0,47670
2,0	0,47725	0,47778	0,47831	0,47882	0,47932	0,47982	0,48030	0,48077	0,48124	0,48169
2,1	0,48214	0,48257	0,48300	0,48341	0,48382	0,48422	0,48461	0,48500	0,48537	0,48574
2,2	0,48610	0,48645	0,48679	0,48713	0,48745	0,48778	0,48809	0,48840	0,48870	0,48899
2,3	0,48928	0,48956	0,48983	0,49010	0,49036	0,49061	0,49086	0,49111	0,49134	0,49158
2,4	0,49180	0,49202	0,49224	0,49245	0,49266	0,49286	0,49305	0,49324	0,49343	0,49361
2,5	0,49379	0,49396	0,49413	0,49430	0,49446	0,49461	0,49477	0,49492	0,49506	0,49520
2,6	0,49534	0,49547	0,49560	0,49573	0,49585	0,49598	0,49609	0,49621	0,49632	0,49643
2,7	0,49653	0,49664	0,49674	0,49683	0,49693	0,49702	0,49711	0,49720	0,49728	0,49736
2,8	0,49744	0,49752	0,49760	0,49767	0,49774	0,49781	0,49788	0,49795	0,49801	0,49807
2,9	0,49813	0,49819	0,49825	0,49831	0,49836	0,49841	0,49846	0,49851	0,49856	0,49861
3,0	0,49865	0,49869	0,49874	0,49878	0,49882	0,49886	0,49889	0,49893	0,49896	0,49900
3,1	0,49903	0,49906	0,49910	0,49913	0,49916	0,49918	0,49921	0,49924	0,49926	0,49929
3,2	0,49931	0,49934	0,49936	0,49938	0,49940	0,49942	0,49944	0,49946	0,49948	0,49950
3,3	0,49952	0,49953	0,49955	0,49957	0,49958	0,49960	0,49961	0,49962	0,49964	0,49965
3,4	0,49966	0,49968	0,49969	0,49970	0,49971	0,49972	0,49973	0,49974	0,49975	0,49976
3,5	0,49977	0,49978	0,49978	0,49979	0,49980	0,49981	0,49981	0,49982	0,49983	0,49983
3,6	0,49984	0,49985	0,49985	0,49986	0,49986	0,49987	0,49987	0,49988	0,49988	0,49989
3,7	0,49989	0,49990	0,49990	0,49990	0,49991	0,49991	0,49992	0,49992	0,49992	0,49992
3,8	0,49993	0,49993	0,49993	0,49994	0,49994	0,49994	0,49994	0,49995	0,49995	0,49995
3,9	0,49995	0,49995	0,49996	0,49996	0,49996	0,49996	0,49996	0,49996	0,49997	0,49997
4,0	0,49997	0,49997	0,49997	0,49997	0,49997	0,49997	0,49998	0,49998	0,49998	0,49998

Para abscissas maiores que 4,09, use a probabilidade 0,50000

Tabela T

Unicaudal	75%	80%	85%	90%	95%	97,50%	99%	99,50%
Bicaudal	50%	60%	70%	80%	90%	95%	98%	99%
gl								
1	1,000	1,376	1,963	3,078	6,314	12,710	31,820	63,660
2	0,816	1,061	1,386	1,886	2,920	4,303	6,965	9,925
3	0,765	0,978	1,250	1,638	2,353	3,182	4,541	5,841
4	0,741	0,941	1,190	1,533	2,132	2,776	3,747	4,604
5	0,727	0,920	1,156	1,476	2,015	2,571	3,365	4,032
6	0,718	0,906	1,134	1,440	1,943	2,447	3,143	3,707
7	0,711	0,896	1,119	1,415	1,895	2,365	2,998	3,499
8	0,706	0,889	1,108	1,397	1,860	2,306	2,896	3,355
9	0,703	0,883	1,100	1,383	1,833	2,262	2,821	3,250
10	0,700	0,879	1,093	1,372	1,812	2,228	2,764	3,169
11	0,697	0,876	1,088	1,363	1,796	2,201	2,718	3,106
12	0,695	0,873	1,083	1,356	1,782	2,179	2,681	3,055
13	0,694	0,870	1,079	1,350	1,771	2,160	2,650	3,012
14	0,692	0,868	1,076	1,345	1,761	2,145	2,624	2,977
15	0,691	0,866	1,074	1,341	1,753	2,131	2,602	2,947
16	0,690	0,865	1,071	1,337	1,746	2,120	2,583	2,921
17	0,689	0,863	1,069	1,333	1,740	2,110	2,567	2,898
18	0,688	0,862	1,067	1,330	1,734	2,101	2,552	2,878
19	0,688	0,861	1,066	1,328	1,729	2,093	2,539	2,861
20	0,687	0,860	1,064	1,325	1,725	2,086	2,528	2,845
21	0,686	0,859	1,063	1,323	1,721	2,080	2,518	2,831
22	0,686	0,858	1,061	1,321	1,717	2,074	2,508	2,819
23	0,685	0,858	1,060	1,319	1,714	2,069	2,500	2,807
24	0,685	0,857	1,059	1,318	1,711	2,064	2,492	2,797
25	0,684	0,856	1,058	1,316	1,708	2,060	2,485	2,787
26	0,684	0,856	1,058	1,315	1,706	2,056	2,479	2,779
27	0,684	0,855	1,057	1,314	1,703	2,052	2,473	2,771
28	0,683	0,855	1,056	1,313	1,701	2,048	2,467	2,763
29	0,683	0,854	1,055	1,311	1,699	2,045	2,462	2,756
30	0,683	0,854	1,055	1,310	1,697	2,042	2,457	2,750
40	0,681	0,851	1,050	1,303	1,684	2,021	2,423	2,704
50	0,679	0,849	1,047	1,299	1,676	2,009	2,403	2,678
60	0,679	0,848	1,045	1,296	1,671	2,000	2,390	2,660
80	0,678	0,846	1,043	1,292	1,664	1,990	2,374	2,639
100	0,677	0,845	1,042	1,290	1,660	1,984	2,364	2,626
120	0,677	0,845	1,041	1,289	1,658	1,980	2,358	2,617
∞	0,674	0,842	1,036	1,282	1,645	1,960	2,326	2,576

Referências Bibliográficas

[1] BUSSAB, W. de Q.; MORETTIN, P. A. Estatística Básica. 5. ed. São Paulo: Saraiva, 2004.

[2] FONSECA, J. S. da; MARTINS, G. de A. Curso de Estatística. 6. ed. São Paulo: Atlas, 2010.

[3] MEYER, P. L. Probabilidade: Aplicações à de Estatística. 2. ed. Rio de Janeiro: LTC, 1987.

[4] TRIOLA, M. F. Introdução à Estatística. 9. ed. Tradução Alfredo Alves Farias. Rio de Janeiro: LTC, 2005.