

Introdução

A estatística reúne os métodos para coleta, resumo, apresentação e análise de dados bem como na obtenção de conclusões válidas e na tomada de decisões razoáveis baseadas em tais análises. A estatística é importante em quase todas as áreas da ciência, mas é especialmente importante na Hidrologia, nas ciências da terra como Geografia e Geologia e nas Engenharias.

No passado, tratar uma grande quantidade de dados numéricos era uma tarefa tediosa, cansativa e sujeita a erros de cálculo diversos. O desenvolvimento dos computadores e de diversos programas permite que grande quantidade de dados possa ser analisada rapidamente. Por outro lado, o uso de computadores e programas sofisticados não exclui a possibilidade de ocorrerem erros na interpretação dos resultados, especialmente no caso de pessoas sem o mínimo conhecimento teórico.

A Estatística pode ser dividida em três áreas:

- Estatística Descritiva
- Probabilidade
- Inferência Estatística

A *Estatística Descritiva* é utilizada nas etapas iniciais de análise, quando os dados são coletados e verificados pela primeira vez. Para uma primeira análise dos dados são usados métodos para apresentação dos dados e métodos para resumo dos dados. A Estatística Descritiva pode ser definida como um conjunto de técnicas destinadas a descrever e resumir os dados, a fim de que possamos tirar conclusões a respeito de características de interesse.

A *Probabilidade* pode ser pensada como a teoria matemática utilizada para se estudar a incerteza oriunda de fenômenos de caráter aleatório. Esta área não será aprofundada neste curso, mas apenas abordada de forma superficial em casos de problemas simples.

A *Inferência Estatística* é o conjunto de técnicas que possibilitam tirar conclusões sobre fenômenos que atingem um grande conjunto de dados a partir da análise de um pequeno conjunto de dados, como no caso das pesquisas eleitorais, que buscam prever o resultado de uma eleição em que votam vários milhões de eleitores a partir da entrevista de apenas alguns milhares de eleitores. Assim, o objetivo da Inferência Estatística é obter respostas corretas de questões específicas, atendendo a um determinado grau de acerto.

Para alguns estudiosos a estatística é uma arte; para outros a estatística é a simples aplicação do bom senso. Em qualquer caso, a estatística ajuda a tomar decisões com informações incompletas, tendo presente que o sucesso da decisão dependerá da habilidade do analista para compreender os resultados das informações contidas nos dados. A primeira parte do processo decisório é a estatística descritiva e a outra é a inferência estatística.

Este texto está organizado em 10 capítulos, com as referências bibliográficas e informações adicionais no décimo primeiro.

Este primeiro capítulo apresenta apenas uma introdução ao assunto. O segundo capítulo descreve a coleta de dados que podem ser utilizados para análises estatísticas. O terceiro capítulo descreve algumas formas de apresentar dados. O quarto capítulo apresenta ferramentas práticas para simplificar a análise através da transformação de uma grande massa de dados em alguns valores numéricos que representam a essência das características dos dados coletados. O quinto capítulo apresenta o histograma, que é uma forma especialmente útil de apresentar dados. O sexto capítulo descreve a curva de permanência de vazões, que é especialmente útil na hidrologia. O sétimo capítulo apresenta o Box-Plot, que é uma forma gráfica de apresentar e resumir dados. O oitavo capítulo apresenta uma breve introdução à exploração estatística da relação entre variáveis. No nono capítulo é introduzido o conceito de distribuição de probabilidades, com ênfase à distribuição normal, ou gaussiana. O décimo capítulo apresenta algumas técnicas para a estimativa de vazões máximas e mínimas em rios. Finalmente o décimo capítulo apresenta uma introdução à distribuição amostral, e à estimativa de erros e testes de diferenças de médias.

Aplicações da estatística aos problemas de medição de variáveis hidrológicas e à análise de problemas em hidrologia são apresentadas em quase todos os capítulos.

Parte dos exercícios sugeridos pode ser realizada utilizando uma calculadora científica comum, mas alguns exercícios podem ser resolvidos mais facilmente utilizando o software Excel, da Microsoft. Em diversos capítulos deste texto são apresentadas as ferramentas de estatística mais comuns do Excel.

Alunos interessados em aprofundar seus conhecimentos de estatística podem buscar complementação em diversos livros texto, alguns que foram consultados na elaboração desta apostila são apresentados no último capítulo.

Coleta de dados

A Estatística lida com dados, números dentro de um contexto. Entretanto, a utilização de estatística é mais do que trabalhar com números, pois embora a organização dos números e a construção de gráficos possa ser mecanizada com softwares e modelos, as idéias e bons julgamentos, por enquanto, não podem ser automatizados. O analista deve ter o hábito de perguntar, por exemplo, o que mostram os resultados dentro de um determinado contexto? Quais as respostas que os dados podem dar a perguntas específicas?

Tenha em mente que durante a apresentação da disciplina Estatística é realizada uma análise explanatória de dados conhecidos, não havendo, em geral, nenhuma pergunta em mente. Entretanto, na prática diária da estatística são procuradas respostas a perguntas específicas, por exemplo, quais indivíduos (medições, pessoas, animais, taxas de juros e outras coisas) devem ser estudados? Que variáveis devem ser medidas? Nesses casos, em geral, os dados devem ser gerados. Este é o caso na hidrologia, em que devem ser medidos os dados de chuva, vazão, nível dos rios, granulometria dos sedimentos etc.

Os dados requeridos pela análise são obtidos pesquisando dados disponíveis, ou gerando novos dados. Em hidrologia e hidrometria podem ser realizadas análises estatísticas sobre dados já existentes, como as séries de dados de chuva em um determinado posto pluviométrico, ou sobre dados novos, como no caso de uma série de testes de esvaziamento de um tanque, ou de medição de capacidade de infiltração do solo.

Classificação dos dados

Como o procedimento estatístico a ser aplicado dependerá da natureza dos dados ou das observações de cada variável, deve-se desenvolver a habilidade de distinguir os tipos de dados possíveis e suas unidades de medida. Quanto a sua natureza, as observações ou dados se classificam em quantitativas discretas e contínuas, qualitativas nominais e ordinais, de corte transversal e séries temporais.

Dados quantitativos. Refere-se a quantidades medidas numa escala numérica, em geral, acompanhadas de alguma unidade de medida e podem ser de dois tipos discretos ou contínuos.

Dados discretos. Referem-se aos valores numéricos que assumem somente números inteiros positivos 0,1,2,3 Os dados discretos resultam, em geral, de contagens: a quantidade de vendas diárias de uma empresa, o número de filhos das famílias de uma região do país, o número de movimentos da conta corrente dos clientes de um banco comercial, a quantidade de peças defeituosas em um lote de produção, o número de transações financeiras com erro de lançamentos, o número de acidentes nas estradas durante as férias anuais de verão etc.

Dados contínuos. Referem-se aos valores numéricos que assumem qualquer valor do conjunto dos números reais. Os dados contínuos resultam, em geral, de medições que podem ter grande precisão: a estatura dos alunos de uma turma, o consumo mensal de energia elétrica de uma casa, o tempo de espera na fila do banco, o tempo de espera na parada de ônibus, o coeficiente de escoamento de um vertedor.

Dados qualitativos. Refere-se às observações não numéricas e são classificadas em nominais e ordinais.

Dados nominais. Estes dados não tem ordenamento nem hierarquia. Por exemplo, o sexo dos funcionários registrados no cadastro de uma empresa, o estado civil, o nome das empresas que tem ações negociadas na bolsa de valores, etc.

Dados ordinais. Estes dados são semelhantes aos nominais, mas incluem uma ordem, uma hierarquia. Por exemplo, o cargo dos funcionários numa empresa: presidente, diretor, gerente, etc. Ou a escala em um questionário de avaliação de satisfação de um cliente: ótimo, bom, regular, ruim, péssimo. O grau de nebulosidade de um dia: claro, nublado, encoberto.

Tipos de variáveis

As variáveis podem ser obtidas de duas formas, dependendo de como os valores são obtidos ao longo do tempo, ou se o tempo desempenha algum papel na análise: séries temporais ou cortes transversais numa data ou período.

Séries temporais. As observações são dados de uma mesma variável em diferentes períodos de tempo: o valor do PIB anual de um país, a taxa mensal de desemprego numa região, as cotações diárias de uma ação, a rentabilidade mensal de uma empresa, a demanda de energia elétrica diária na região Sudeste medida às dezoito horas etc.

Corte transversal numa data ou período. Se na coleta dos dados não for considerada a sequência temporal; por exemplo, amostras da quantidade produzida e do preço médio dos produtos, ou das vendas e do

investimento em propaganda, a média de apartamentos vendidos durante o último mês pelas primeiras dez imobiliárias da cidade, o número de operações fechadas por cinco ações numa determinada data etc.

População e amostra

População é o conjunto total unidades elementares de pessoas, objetos ou coisas sobre as quais se querem obter informações. Um subconjunto de unidades elementares selecionadas de uma população é denominado amostra.

Uma população pode ser fornada por todos os habitantes de um país, ou de um estado, ou de um município etc. Um exemplo de pesquisa de uma população completa é o censo demográfico do Brasil realizado pelo IBGE. A análise das vendas de um segmento da economia, por exemplo, o de montadoras de carros, durante o mesmo ano é outro exemplo de população. Entretanto, nem sempre é conveniente obter informações de todas as pessoas, objetos ou coisas de uma população. Os resultados de uma pesquisa de intenção de voto de todos os eleitores do país numa eleição presidencial não conseguiriam captar do que os

Uma amostra é dita representativa quando ela tem as mesmas características da população de onde foi tirada.

partidos políticos necessitam, pois o tempo necessário para coletar todas as opiniões comprometeria os resultados, além de ser muito cara para a finalidade que se propõe. Em alguns casos, a restrição de consultar toda a população é econômica, como é o caso da determinação da vida útil das lâmpadas que obrigaria a testar todas

as lâmpadas produzidas, não restando nenhuma para a venda! Dessa maneira, o procedimento recomendado é escolher uma amostra representativa de um lote de lâmpadas produzidas.

Uma amostra aleatória de tamanho n , retirada de uma população é uma das muitas possíveis e igualmente prováveis combinações de n unidades elementares que podem ser retiradas de uma população. Portanto, qualquer amostra de tamanho n tem a mesma probabilidade de ser selecionada.

Na hidrologia é comum termos variáveis na forma de séries temporais, como por exemplo, as chuvas totais anuais em um determinado local. Neste caso a população seriam todos os anos (a idade da Terra), enquanto uma amostra seriam, por exemplo, 30 anos de dados.

Mas na hidrologia também são comuns os dados na forma de cortes transversais, onde o tempo não importa. As características do solo, por exemplo, variam no espaço. Para caracterizar perfeitamente a porosidade do solo esta característica deveria ser medida em uma infinidade de pontos (população) cobrindo toda a região de interesse. É claro que isso é impossível de realizar, assim é necessário fazer apenas umas poucas medições, talvez dez ou vinte, para estimar as características. Este conjunto também seria chamado amostra.

Apresentação de dados

A apresentação dos dados de forma organizada é necessária para obter informações a partir de uma amostra, especialmente se a amostra for relativamente grande, o que sempre é desejável. A organização dos dados é normalmente feita na forma de gráficos e tabelas.

Tabelas de dados discretos

A forma mais simples de apresentar dados estatísticos é na forma de tabelas. Qualquer análise estatística normalmente parte da elaboração de uma tabela com os dados. Uma das tabelas mais utilizadas é a tabela de frequências. A frequência do valor de uma variável é o número de repetições deste valor dentro da amostra.

Considere a sequência de números a seguir, que correspondem ao número de dias de chuva no mês de janeiro em um determinado local nos últimos 26 anos.

14	12	13	11	12	13	16	14	14	15	17	14	11
13	14	15	13	12	14	13	14	13	15	16	12	12

Podemos obter algumas informações apenas olhando para estes números, como o máximo (17) e o mínimo (11), mas temos dificuldades em aprofundar a análise. Uma tabela de frequências absolutas mostra quantas vezes ocorreu cada um dos valores da variável *número de dias de chuva em janeiro*.

Olhando a tabela de frequências absolutas observamos facilmente que o valor mais frequente é 14, e que o valor 17 ocorreu apenas uma vez.

Muitas vezes as tabelas de frequência absoluta são transformadas em tabelas de frequência relativa, dividindo as frequências absolutas pelo tamanho da amostra.

Número de dias de chuva em janeiro	Frequência absoluta
11	2
12	5
13	6
14	7
15	3
16	2
17	1
	Total=26

No exemplo anterior, dividindo a coluna da direita por 26, que é o tamanho da amostra, temos:

Número de dias de chuva em janeiro	Frequência absoluta
11	7,69%
12	19,23%
13	23,08%
14	26,92%
15	11,54%
16	7,69%
17	3,85%
	Total=100%

Esta tabela permite observações ainda mais aprofundadas, por exemplo, podemos verificar que em pouco mais de um quarto (26,92%) dos anos da amostra o mês de janeiro apresentou 14 dias de chuva.

Em alguns casos o interesse da análise dos dados reside em conhecer os valores da variável que são maiores que um determinado limite, por exemplo, o número de anos em que janeiro teve mais de 15 dias de chuva. Neste caso é útil elaborar a tabela de frequências acumuladas. A frequência acumulada do valor de uma variável é a soma das frequências absolutas ou relativas desde o valor inicial da variável.

No exemplo anterior relativo ao número de dias de chuva em janeiro podemos elaborar a tabela com frequências acumuladas (tanto absolutas como relativas):

Número de dias de chuva em janeiro	Frequência absoluta	Frequência relativa
11	2	7,69%
12	7	26,92%
13	13	50,00%

14	20	76,92%
15	23	88,46%
16	25	96,15%
17	26	100%

Tabelas de dados contínuos

No caso de dados contínuos recomenda-se trabalhar com intervalos de valores para a contagem de frequência. Isto ocorre porque é praticamente impossível contar a frequência de dados contínuos, e a tabela teria um número excessivo de linhas e frequências baixas para cada um dos valores, inviabilizando qualquer análise.

Considere os dados de chuva anual em um determinado local (medidos em mm), dados na tabela a seguir.

1421	1234	1326	1187	1281	1311	1600	1489	1492	1522	1709	1490	1101
1393	1414	1505	1333	1201	1444	1380	1477	1329	1540	1603	1267	1299

Para criar uma tabela de frequências como a anterior, teríamos que ter 609 linhas, para cada um dos valores entre 1101 e 1709. Isto não é prático e não faz sentido. Assim, na contagem de frequência somamos o número de anos em que o valor da variável está em cada intervalo. Podemos definir, por exemplo, intervalos de 100 mm: de 1000 a 1100; de 1100 a 1200; de 1200 a 1300; de 1300 a 1400; de 1400 a 1500; de 1500 a 1600; de 1600 a 1700; e de 1700 a 1800.

A tabela de frequências absolutas e relativas resultante é dada abaixo:

Chuva anual (mm)	Frequência absoluta	Frequencia relativa
1000 a 1100	0	0%
1100 a 1200	2	7,69%
1200 a 1300	5	19,23%
1300 a 1400	6	23,08%
1400 a 1500	7	26,92%
1500 a 1600	3	11,54%
1600 a 1700	2	7,69%
1700 a 1800	1	3,85%
	Total=26	Total=100%

O número (k) de intervalos em que deve ser dividida uma amostra é arbitrário, mas um critério que pode ser utilizado é dado por:

$$k = \sqrt{n}$$

onde n é o tamanho da amostra e k deve ser arredondado para o valor mais próximo.

Histograma

O histograma é um gráfico que permite visualizar as tabelas de frequência de forma rápida, utilizando barras verticais para representar as frequências. O histograma pode ser feito representando frequências relativas ou absolutas (a figura será igual, somente os valores serão diferentes).

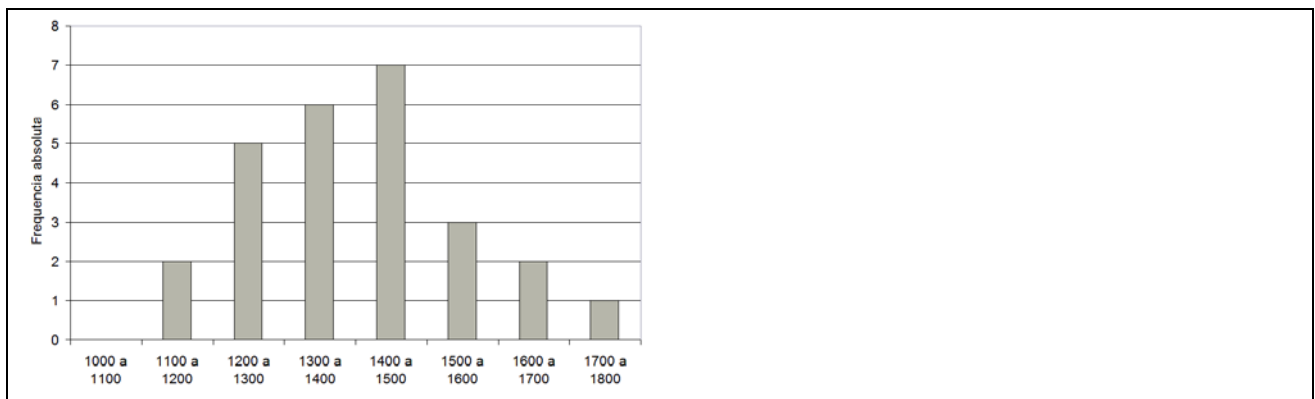


Figura 1: Histograma de frequências de chuva anual (mm) em um determinado local a partir de dados de uma amostra de 26 anos.

A partir de um histograma podemos ver facilmente qual é a faixa normal de precipitações anuais neste local e quais são os valores que mais ocorrem. Com base nesta figura seria fácil dizer que um ano com 2500 mm de chuva foi um ano extremamente chuvoso, e que um ano com apenas 600 mm de chuva foi um ano extremamente seco. Por outro lado, um ano com 1250 mm de chuva seria um ano razoavelmente normal, pelo que nos mostra o histograma.

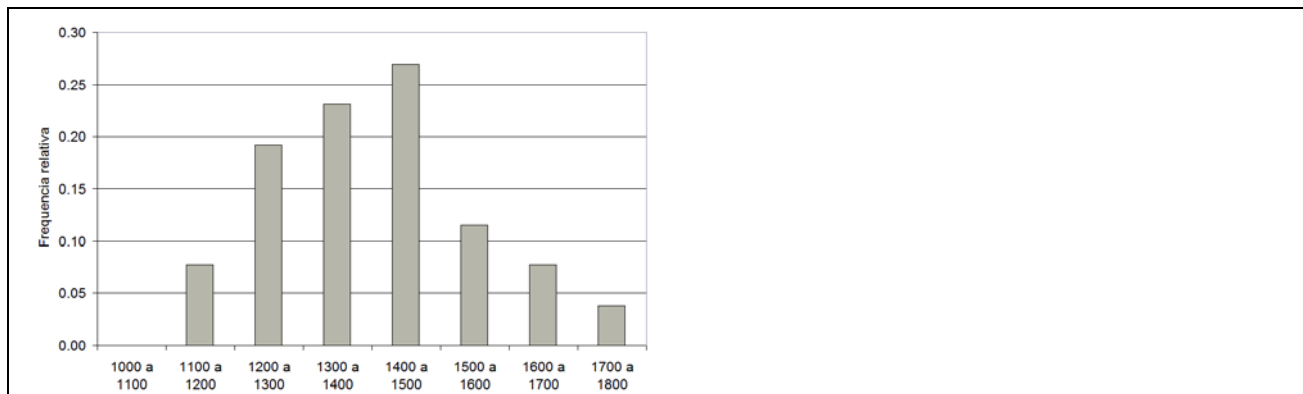


Figura 2: Histograma de frequências relativas de chuva anual (mm) em um determinado local a partir de dados de uma amostra de 26 anos.

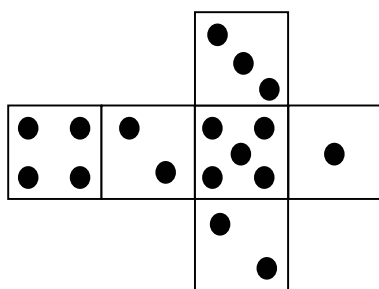
O histograma pode ser elaborado de forma manual ou de forma automática em softwares como o Excel. As ferramentas de elaboração de gráficos do Excel podem ser usadas para elaborar o histograma e o histograma também pode ser obtido diretamente com uma ferramenta específica, incluída entre as Ferramentas de Análise do Excel.

Exercícios

- Um grupo de pedagogos estuda a influência da troca de escolas no desempenho de alunos. Como parte do levantamento realizado, foi anotado o número de escolas cursadas pelos alunos participantes do estudo, conforme apresentado na tabela abaixo. Com base nestes dados responda qual é a porcentagem dos alunos que cursaram mais de uma escola. Classifique os alunos em dois grupos de rotatividade: alta – quando o aluno cursou mais de duas escolas e baixa – para os outros alunos.

Escolas Cursadas	Número de alunos
1	46
2	57
3	21
4	15
5	4

- Considere o dado com as seis faces apresentadas abaixo. Construa o histograma que é esperado a partir de uma série de 1000 lançamentos deste dado.



Resumo de dados estatísticos

Para tentar conhecer as características de uma população, extraímos uma amostra desta população e analisamos esta amostra. A análise pode começar organizando os dados como descrito no capítulo anterior. A partir daí procuramos extrair mais informação utilizando formas de resumir os dados da amostra em alguns poucos valores que representam de forma razoavelmente fiel a variabilidade dos dados da amostra.

Entre os valores que são usados para resumir os dados de uma amostra estão a **média**, o **desvio padrão**, a **moda**, o **coeficiente de variação** e o **coeficiente de assimetria**.

Outra ferramenta útil para resumir os dados de uma amostra é a análise baseada no ordenamento, baseada no simples ato de colocar em ordem crescente ou decrescente todos os valores da amostra. A partir do ordenamento da amostra é possível obter informações que resumem a amostra como a **mediana**, os **percentis** ou **quantis**, e, especialmente, os **quartis**, que são bastante utilizados.

A média

A média é o valor obtido pela soma de todos os valores dos dados da amostra dividida pelo tamanho da amostra, como apresentado na equação abaixo:

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

onde x_i são os valores; n é o tamanho da amostra (número de valores que devem ser somados); e \bar{x} é a média.

Calculadoras permitem obter o valor da média de uma sequência de valores de forma muito rápida.

A mediana

A mediana é o valor que é superado por 50% dos pontos da amostra. A média e a mediana podem ter valores relativamente próximos, porém não iguais.

A mediana pode ser obtida organizando os n valores x_i da amostra em ordem crescente.

Sendo x_i com $i = 1$ a n , os valores de x organizados em ordem decrescente, a mediana é obtida por:

$$\text{Mediana} = x_p \text{ com } p = \frac{n-1}{2} + 1 \text{ se } n \text{ for ímpar;}$$

$$\text{e } \text{Mediana} = \frac{x_p + x_{p+1}}{2} \text{ se } n \text{ for par.}$$

Ao contrário da média, a mediana não é uma função encontrada em qualquer calculadora. Para calcular a mediana de um conjunto grande de dados o ideal é utilizar uma planilha de cálculo no computador, como o programa Excel, por exemplo.

O desvio padrão

O desvio padrão é uma medida de dispersão dos valores de uma amostra em torno da média. O desvio padrão é dado por:

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$$

o quadrado do desvio padrão s^2 é chamada variância da amostra.

A moda

A moda é o valor que mais freqüente da amostra. Assim como a média e a mediana, a moda indica uma tendência dos valores aleatórios.

A moda pode ser utilizada para valores numéricos ou para dados não numéricos, como categorias, cores, nomes de pessoas ou modelos de carros. Também é possível definir a moda para intervalos de valores numéricos.

Exemplo: Um conjunto de dois dados foi lançado 20 vezes, revelando os resultados da soma dos valores apresentados na tabela abaixo. Qual é o valor da moda?

Valor da soma dos dois dados	Número de ocorrências
3	1
4	3
5	1
6	4
7	5
8	2
9	3
10	0
11	1

O valor da moda é 7, porque este é o valor da soma mais freqüente na amostra. De certa forma isto poderia ser esperado, uma vez que a soma 7 pode ocorrer pela combinação de 1+6; 2+5; 3+4; 4+3; 5+2 e 6+1; enquanto os outros valores, como o 2, o 6 e o 12, tem um número de combinações possíveis menores.

A moda também pode ser calculada para variáveis contínuas, desde que se definam intervalos para a contagem de freqüência.

Exemplo: Os pesos de 15 alunos de uma turma foram medidos, com os resultados apresentados na tabela abaixo. Qual é o valor da moda?

Aluno	Peso
1	81,1
2	73,2
3	61,5
4	64,3
5	55,7
6	62,9
7	73,4
8	70,0

9	71,5
10	78,0
11	78,1
12	73,9
13	68,5
14	66,9
15	59,0

Definindo intervalos de 10 quilos, entre 40 e 50; 50 e 60; 60 e 70; 70 e 80 e 80 e 90 é possível contar a frequência (número de alunos) em cada intervalo. Assim, a tabela anterior pode ser resumida pela tabela abaixo:

Peso (kg)	Frequência <i>a</i>
40 a 50	0
50 a 60	2
60 a 70	5
70 a 80	7
80 a 90	1

Assim, a moda é o intervalo de peso entre 70 e 80 Kg.

No exemplo anterior, para ser mais rigoroso, os intervalos deveriam ter sido definidos como de 40 Kg ao valor imediatamente menor a 50 Kg (por exemplo 49,999), e assim por diante.

Exemplo: Muitas coisas na natureza tem uma tendência interessante de apresentarem um padrão em que o algarismo 1 é mais comum do que outros no início de um número.. Este padrão é conhecido como Lei de Benford. A população de um país, ou de uma cidade, por exemplo, tem a tendência de ser um número iniciado por um algarismo de valor mais baixo, como 1, 2 ou 3. Cidades com populações expressas por números iniciados com os algarismos 7, 8 e 9 são mais raras. A tabela abaixo apresenta a população de um conjunto aleatório de cidades da França. Identifique a moda do algarismo inicial dos números que expressam esta população e verifique se estes dados seguem a Lei de Benford.

Cidade	Numero de habitantes
Bleis	49300
La Rochelle	120000
Rouen	108750
Paris	2100000
Millau	22500
Grasse	44790
Marseille	807071
Toulouse	398423

<i>Fécamp</i>	21500
<i>Le Havre</i>	193250
<i>Honfleur</i>	8350
<i>Trouville</i>	5600
<i>Caen</i>	117000
<i>Bayeux</i>	15400
<i>Valence</i>	63400
<i>Amboise</i>	11000
<i>Uzerche</i>	3500
<i>Aubusson</i>	5000
<i>St Etienne</i>	200000
<i>Autun</i>	18000
<i>Belle Île</i>	5200
<i>Monaco</i>	30000
<i>Nice</i>	342738
<i>Les Baux de Provence</i>	468
<i>Grenoble</i>	156000
<i>Briançon</i>	11300
<i>Besançon</i>	125000
<i>Clermont-Ferrand</i>	137000
<i>Vichy</i>	27000
<i>Orcival</i>	290
<i>Murat</i>	2400
<i>Thiers</i>	14800
<i>Limoges</i>	200000
<i>Périgueux</i>	33294

Solução: Uma tabela com a frequência de cada um dos algarismos iniciais de 1 a 9 (tabela a seguir) mostra que o 1 é o algarismo inicial mais frequente (a moda do algarismo inicial), com 12 ocorrências.

<i>Algarismo inicial</i>	<i>Frequência</i>
1	12
2	8
3	5
4	3
5	3
6	1
7	0
8	3
9	0

Procure na Internet a página do IBGE (Instituto Brasileiro de Geografia e Estatística) e selecione variáveis como a população dos municípios brasileiros. Você poderá comprovar que a Lei de Benford (regra dos algarismos iniciais) também se aplica no Brasil. O mesmo poderá ser testado para a área de cada município (em Km²).

O coeficiente de variação

O coeficiente de variação é uma relação entre o desvio padrão e a média. O coeficiente de variação é uma medida da variabilidade dos valores em torno da média, relativamente a própria média.

$$cv = \frac{s}{\bar{x}}$$

Exemplo: O seguinte conjunto de valores apresenta a chuva anual ocorrida em uma cidade ao longo de 30 anos. Calcule a média, o desvio padrão e o coeficiente de variação destes dados.

1671
1485
1766
1565
2082
1370
1926
2042
1691
1491
2024
1305
1644
1908
1913
1485
1693
1313
1567
1493
1357
2023
1390
1641
1585
1526
1962
1672
1404
1352

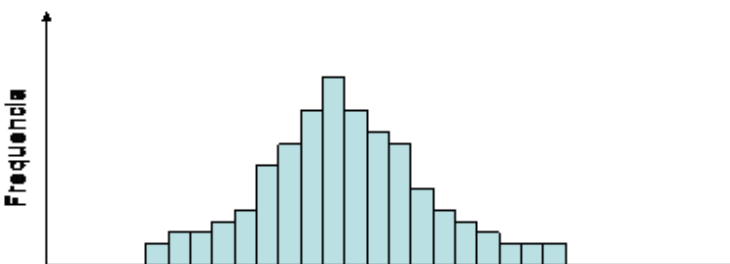
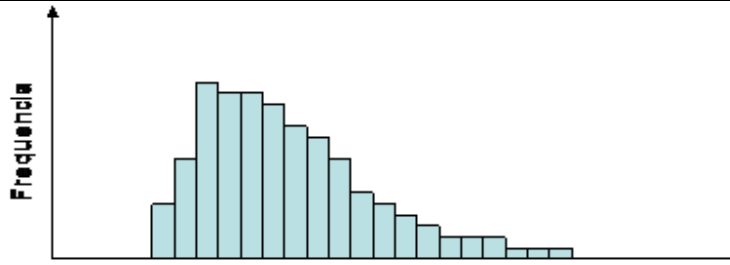
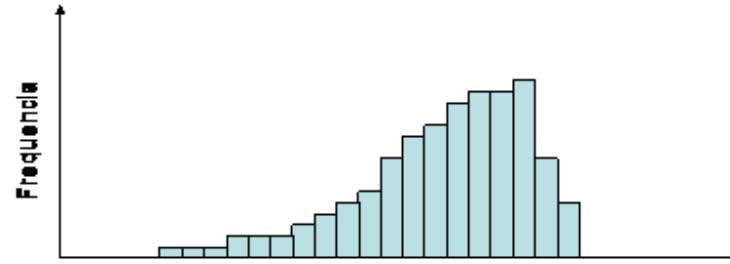
A média é de 1645,1 mm por ano, o desvio padrão é de 241,9 mm por ano e o coeficiente de variação é de 0,15.

O coeficiente de assimetria

O coeficiente de assimetria é um valor que caracteriza o quanto uma amostra de dados é assimétrica com relação à média. Uma amostra é simétrica com relação à média se o histograma dos dados revela o mesmo comportamento de ambos os lados da média.

$$G = \frac{\sum_{i=1}^n (x_i - \bar{x})^3}{n \cdot S^3}$$

A assimetria é chamada positiva quando o valor de G é positivo e a assimetria é negativa quando o valor de G é negativo. Algumas variáveis importantes na hidrologia, como as vazões máximas anuais em rios, apresentam uma assimetria positiva.

Assimetria	Valor de G	Exemplo de histograma
Nula	0 ou próximo de zero	
Positiva	$G > 0$	
Negativa	$G < 0$	

O cálculo da assimetria de uma amostra é um pouco mais complexo do que o da média e do desvio padrão. A maior parte das calculadoras simples não permite calcular diretamente o coeficiente de assimetria.

Quartis e quantis

Quartis separam a amostra de forma semelhante à mediana, porém em intervalos diferentes. Enquanto a mediana separa a amostra em dois grupos, com 50% dos dados com valores inferiores e 50% dos dados com valores superiores à mediana, os quartis e os quantis dividem a amostra em grupos de tamanhos diferentes. O primeiro Quartil é o valor que separa a amostra em dois grupos em que 25% dos pontos tem valor inferior ao quartil e 75% tem valor superior ao quartil. O terceiro Quartil é o valor que separa a amostra em dois grupos em que 75% dos pontos tem valor inferior ao quartil e 25% tem valor superior ao quartil. Já o segundo quartil é a própria mediana.

Além dos três quartis, que separam a amostra em quatro, podem ser definidos quantis arbitrários, que dividem a amostra arbitrariamente em frações diferentes. Por exemplo, o quantil 90 % divide a amostra em dois grupos. O primeiro (90% dos dados) tem valores inferiores ao quantil 90% e o segundo (10% dos dados) tem valores superiores ao quantil 90%.

Aplicações em hidrologia

As variáveis hidrológicas como chuva e vazão têm como característica básica uma grande variabilidade no tempo. Para analisar a vazão de rio e a sua variabilidade temporal é necessário utilizar alguns valores estatísticos que resumem, em grande parte, o comportamento hidrológico do rio ou da bacia. Entre as estatísticas mais importantes estão: a vazão média, a vazão média mensal, a vazão média específica, as vazões mínimas e as vazões máximas de cada ano.

A vazão média é a média de toda a série de vazões diárias registradas, e é muito importante na avaliação da disponibilidade hídrica total de uma bacia. A vazão mediana é a vazão que é superada em 50% dos dias da série. Normalmente, a vazão média e a vazão mediana têm valores próximos, porém não iguais. A vazão média específica é a vazão média dividida pela área de drenagem da bacia.

As vazões médias mensais representam o valor médio da vazão para cada mês do ano, e são importantes para analisar a sazonalidade de um rio. A Figura 3 apresenta um gráfico das vazões médias mensais do rio Cuiabá na seção da cidade de Cuiabá, com base nos dados de 1967 a 1999.

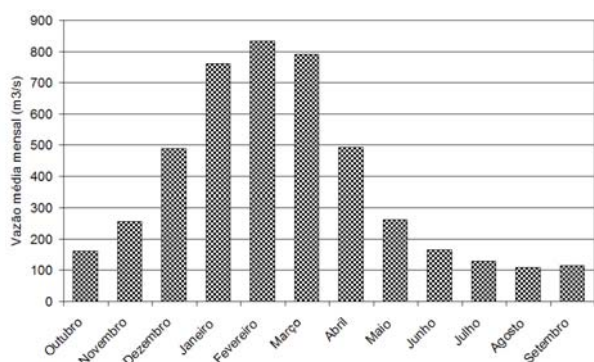


Figura 3: Vazões medias mensais do rio Cuiabá em Cuiabá (dados de 1967 a 1999).

Observa-se nesta figura que há uma sazonalidade marcada, com estiagem no inverno e vazões altas no verão. As maiores vazões mensais médias ocorrem em Fevereiro e as menores em Agosto, o que é consequência direta da sazonalidade das chuvas, que ocorrem de forma concentrada no período de verão.

Exercícios

1. Exercício: Os seguintes automóveis estavam estacionados no IPH numa tarde de inverno. Qual é a moda dos modelos presentes nesta amostra?

Carr o	Modelo
1	Fiat Palio
2	Ford Fiesta
3	Fiat Uno
4	Fiat Uno
5	GM Corsa
6	GM Celta
7	Fiat Uno
8	VW Gol
9	GM Celta
10	Fiat Uno
11	GM Celta
12	Honda Civic
13	Citroen C3
14	Fiat Palio
15	Fiat Palio
16	GM Celta
17	Fiat Palio
18	GM Corsa
19	Peugeot 206
20	Peugeot 206
21	Fiat Palio
22	Honda Fit
23	Peugeot 206
24	Ford Ecosport

25	Peugeot 206
26	VW Gol
27	VW Santana
28	VW Gol
29	Fiat Uno
30	VW Gol

2. Utilizando os dados de vazão média mensal da tabela abaixo, calcule as médias das vazões em cada mês.

Ano	Jan	Fev	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov	Dez
1981	132	154	122	98	67	34	23	18	15	19	39	88
1982	126	191	122	75	62	43	27	19	13	22	45	101
1983	119	155	180	94	45	37	24	19	13	29	53	99
1984	108	125	170	91	83	56	29	15	15	21	39	88
1985	172	167	199	108	70	23	18	12	10	14	28	68
1986	145	187	100	93	60	30	20	18	20	23	49	108

A curva de permanência de vazões

A elaboração da curva de permanência é uma das análises estatísticas mais simples e mais úteis na hidrologia. A curva de permanência auxilia na análise dos dados de vazão com relação a perguntas como as destacadas a seguir.

- O rio tem uma vazão aproximadamente constante ou extremamente variável entre os extremos máximo e mínimo?
- Qual é a porcentagem do tempo em que o rio apresenta vazões em determinada faixa de valores?
- Qual é a porcentagem do tempo em que um rio tem vazão suficiente para atender determinada demanda?

A curva de permanência expressa a relação entre a vazão e a frequência com que esta vazão é superada ou igualada. A curva de permanência pode ser elaborada a partir de dados diários ou dados mensais de vazão.

A figura a seguir apresenta o hidrograma de vazões diárias do rio Taquari, em Muçum (RS), e a curva de permanência que corresponde aos mesmos dados apresentados no hidrograma. Observa-se que a vazão de $1000 \text{ m}^3.\text{s}^{-1}$ é igualada ou superada em menos de 10% do tempo. Apesar de apresentar picos de cheias com $7000 \text{ m}^3.\text{s}^{-1}$ ou mais, na maior parte do tempo as vazões do rio Taquari neste local são bastante inferiores a $500 \text{ m}^3.\text{s}^{-1}$.

Para destacar mais a faixa de vazões mais baixas a curva de permanência é apresentada com eixo vertical logarítmico, como mostra a Figura 5.

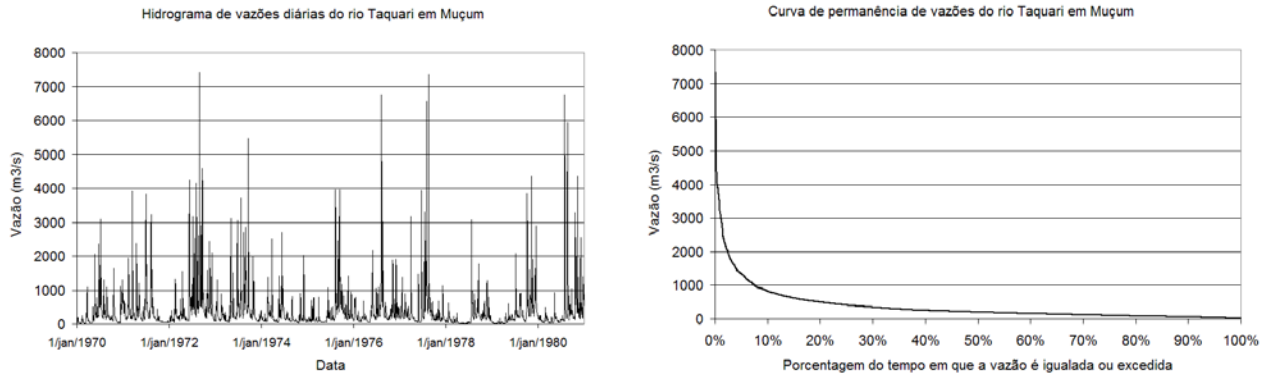


Figura 4: Hidrograma de vazões diárias do rio Taquari em Muçum (RS) e a curva de permanência correspondente.

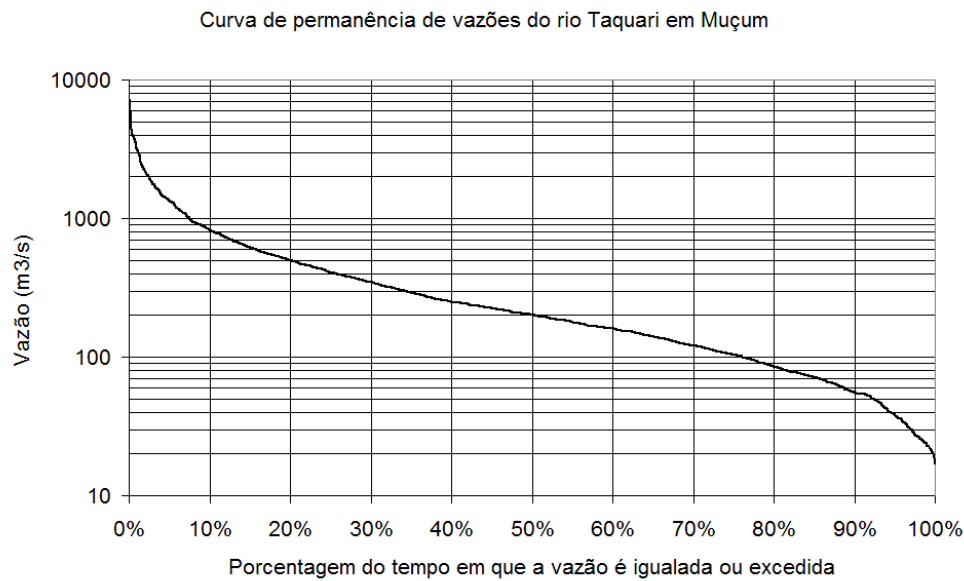


Figura 5: Curva de permanência do rio Taquari em Muçum com eixo das vazões logarítmico para dar destaque à faixa de vazões mais baixas.

Alguns pontos da curva de permanência recebem atenção especial:

- A vazão que é superada em 50% do tempo (mediana das vazões) é a chamada Q_{50} .

- A vazão que é superada em 90% do tempo é chamada de Q_{90} e é utilizada como referência para legislação na área de Meio Ambiente e de Recursos Hídricos em muitos Estados do Brasil.
- A vazão que é superada em 95% do tempo é chamada de Q_{95} e é utilizada para definir a Energia Assegurada de uma usina hidrelétrica.

EXEMPLO

- 1) Os dados de vazão do rio Descoberto em Santo Antônio do Descoberto (GO) foram organizados na forma de uma curva de permanência, como mostra a Figura 6. Um empreendedor solicita outorga de $2,5 \text{ m}^3 \cdot \text{s}^{-1}$ num ponto próximo no mesmo rio. Considerando que a legislação permite outorgar apenas 20% da Q_{90} a cada solicitante, responda: é possível atender a solicitação?

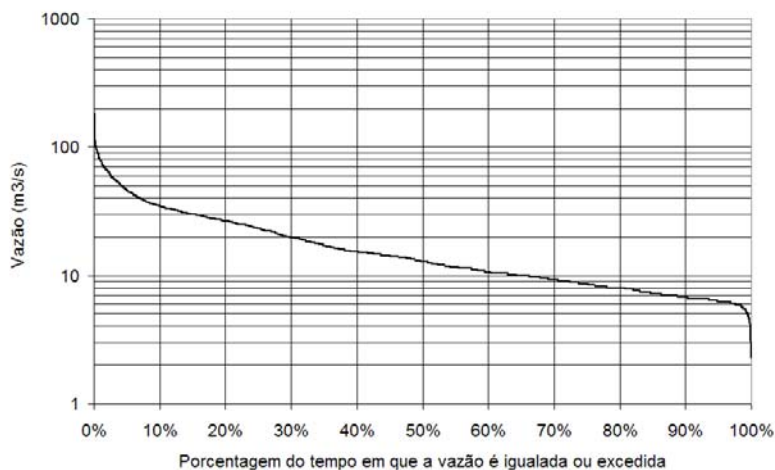


Figura 6: Curva de permanência do rio Descoberto, em Santo Antônio do Descoberto (GO), para o exemplo 1.

Observa-se na curva de permanência que a vazão Q_{90} é de $7 \text{ m}^3 \cdot \text{s}^{-1}$ aproximadamente. Portanto a máxima vazão que pode ser outorgada para um usuário individual neste ponto corresponde a:

$$Q_{\max} = 0,2 \cdot 7 = 1,4 \text{ m}^3 \cdot \text{s}^{-1}$$

Como o empreendedor solicitou $2,5 \text{ m}^3 \cdot \text{s}^{-1}$ não é possível atender sua solicitação.

A curva de permanência também é útil para diferenciar o comportamento de rios e para avaliar o efeito de modificações como desmatamento, reflorestamento, construção de reservatórios e extração de água para uso consuntivo.

A Figura 7 apresenta as curvas de permanência dos rios Cuiabá, em Cuiabá (MT), e Taquari, em Coxim (MS), baseadas nos dados de vazão diária de 1980 a 1984. As duas bacias tem áreas de drenagem de tamanho semelhante. A bacia do rio Cuiabá tem, aproximadamente, 22.000 km², e a do rio Taquari cerca de 27.000 km². O relevo e a precipitação média anual são semelhantes. A vazão média do rio Cuiabá é de 438 m³.s⁻¹ neste período, enquanto a vazão média do rio Taquari é de 436 m³.s⁻¹, ou seja, são praticamente idênticas. Entretanto, observa-se que as vazões mínimas são mais altas no rio Taquari do que no rio Cuiabá e as vazões máximas são maiores no rio Cuiabá.

O rio Cuiabá apresenta maior variabilidade das vazões, que se alternam rapidamente entre situações de baixa e de alta vazão, enquanto o rio Taquari permanece mais tempo com vazões próximas da média. Esta diferença ocorre basicamente porque a geologia da bacia do rio Taquari favorece mais a infiltração da água no solo, e esta água chega ao rio apenas após um longo período em que fica armazenada no subsolo. A vazão do rio Taquari é naturalmente regularizada pelos aquíferos existentes na bacia, enquanto que na bacia do rio Cuiabá este efeito não é tão importante.

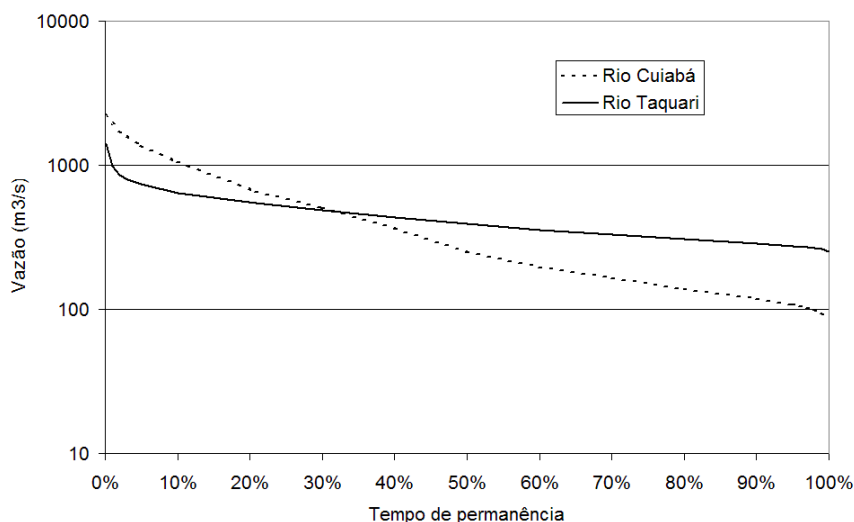


Figura 7: Comparação entre as curvas de permanência dos rios Taquari (MS) e Cuiabá (MT).

A Figura 8 apresenta as curvas de permanência de vazão afluente (entrada) e efluente (saída) do reservatório de Três Marias, no rio São Francisco (MG). Este reservatório tem um grande volume e uma grande capacidade de regularização, permitindo reter grande parte das vazões altas que ocorrem durante o período do verão, aumentando a disponibilidade de água no período de estiagem. Como resultado observa-se que a vazão Q_{90} é alterada de 148 m³.s⁻¹ para 379 m³.s⁻¹ pelo efeito de regularização do reservatório, enquanto a vazão Q_{95} é alterada de 120 m³.s⁻¹ para 335 m³.s⁻¹.

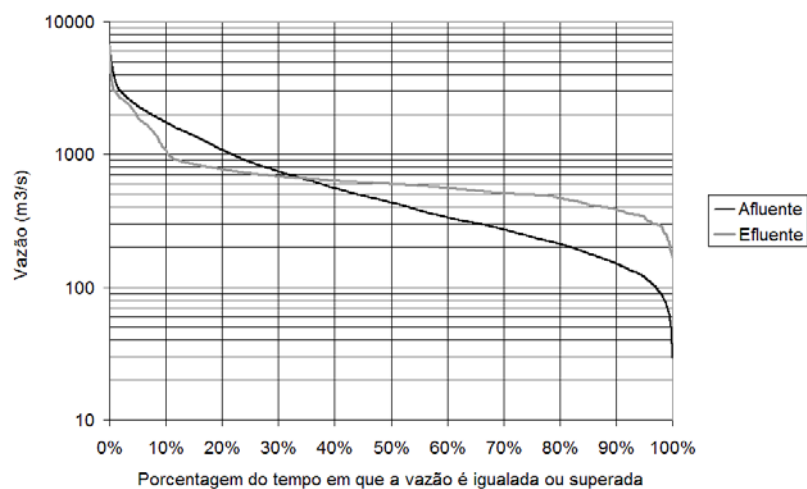


Figura 8: Curvas de permanência de vazão afluente e efluente do reservatório de Três Marias, no rio São Francisco (MG).

Portanto o efeito da regularização da vazão sobre a curva de permanência é torná-la mais horizontal, com valores mais próximos da mediana durante a maior parte do tempo.

O Box-Plot

O Box plot, também conhecido como gráfico de caixa, é uma forma simples de representar graficamente a faixa de variação de uma variável, bem como algumas características de seu histograma. O Box-plot é uma representação gráfica envolvendo os quartis, a mediana, os valores máximo e mínimo.

Para elaborar o Box-plot define-se uma caixa em que os limites superior e inferior são dados pelo terceiro e pelo primeiro quartil, respectivamente. A mediana é representada por um traço no interior da caixa e segmentos de reta são colocados da caixa até os valores máximo e mínimo.

O primeiro Quartil é o valor que separa a amostra em dois grupos, em que 25% dos pontos tem valor inferior ao primeiro quartil e 75% tem valor superior ao quartil.

O segundo quartil é a mediana e o terceiro quartil é o valor que separa a amostra em dois grupos, em que 75% dos pontos tem valor inferior ao terceiro quartil e 25% tem valor superior ao terceiro quartil. A Figura 9 apresenta um exemplo de um Box-plot.

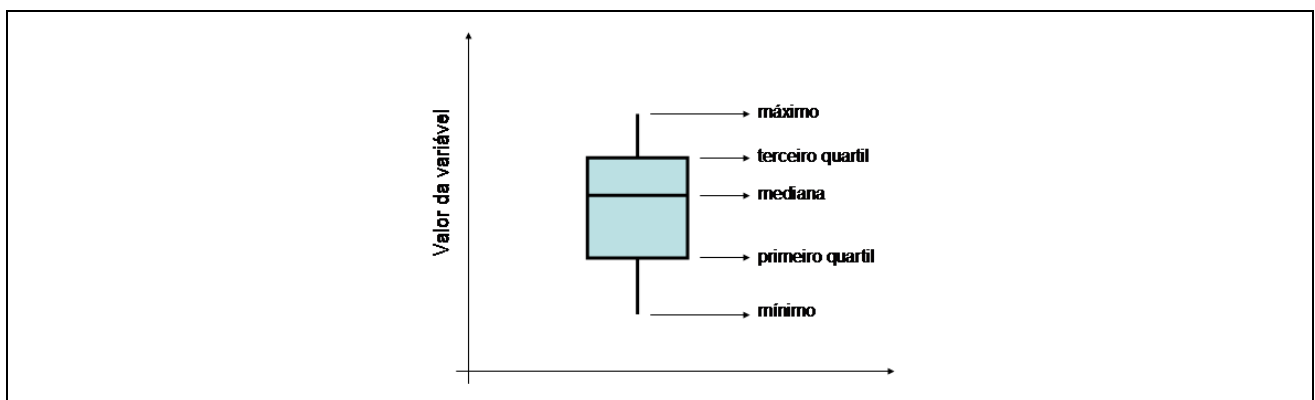


Figura 9: Exemplo de elaboração de um Box-plot de uma amostra de dados.

Exemplo: Elabore um Box plot para representar os dados referentes ao seguinte problema: Um técnico em hidrologia realizou 20 ensaios de infiltração em uma área de solo utilizado para agricultura, obtendo os seguintes resultados em mm/hora: 48; 35; 37; 52; 43; 29; 61; 33; 44; 55; 69; 43; 22; 35; 38; 57; 53; 67; 62; 48.

Solução: Estes valores podem ser organizados em ordem crescente ou decrescente permitindo encontrar os seguintes valores: a mediana é 46; o primeiro quartil (25%) é 36,5; o terceiro quartil é 55,5; o valor máximo é 69 e o mínimo é 22.

minimo	22
q1	36.5
mediana	46
q3	55.5
máximo	69

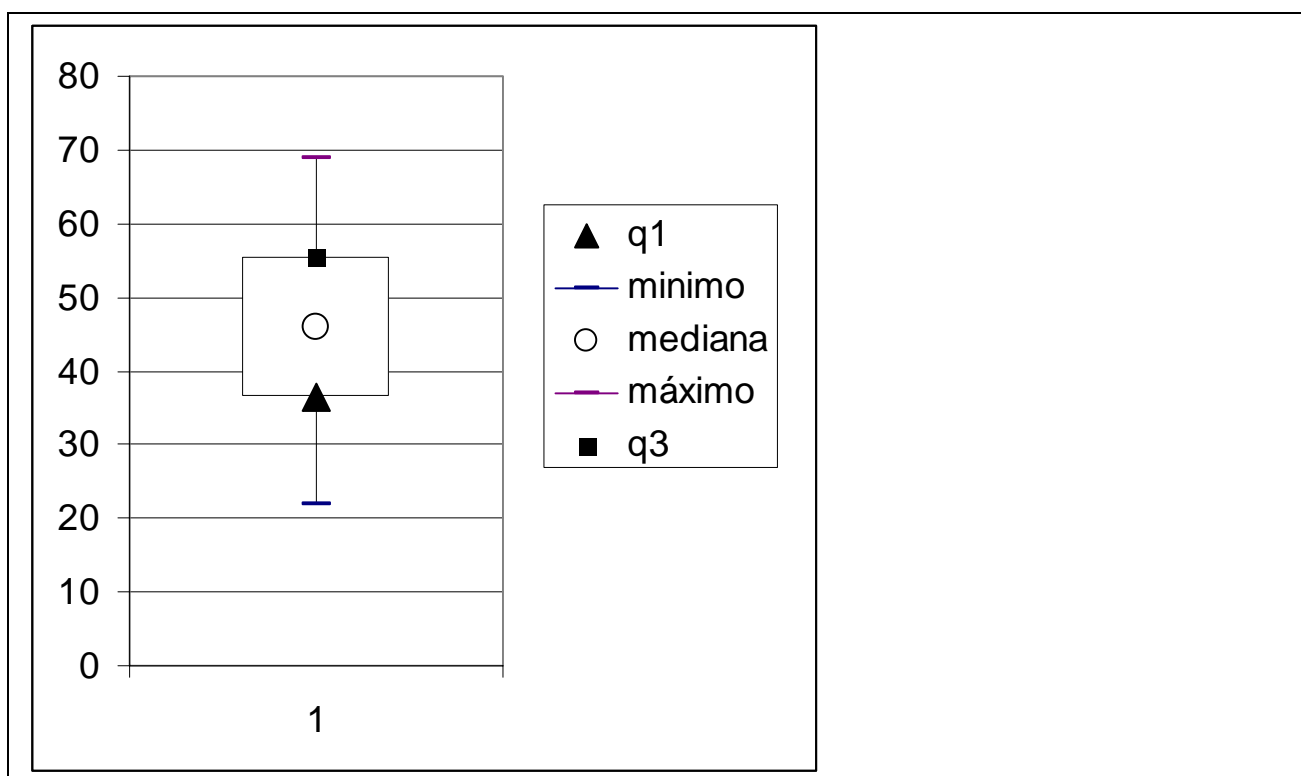
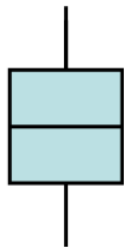
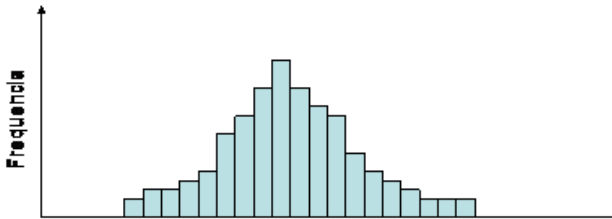
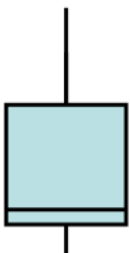
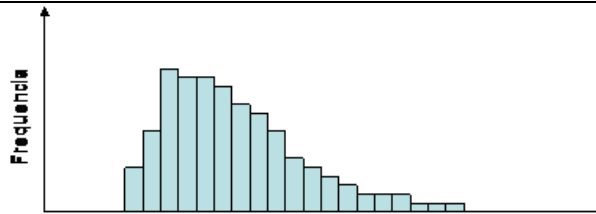
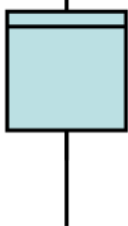
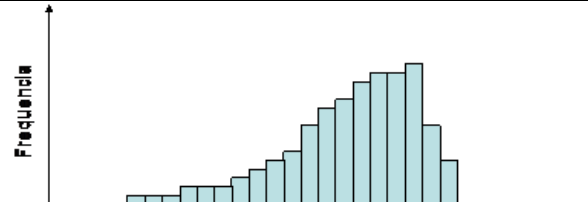


Figura 10: Box plot relativo aos dados do exemplo.

Uma importante aplicação do Box-plot é a avaliação rápida de algumas características da distribuição estatística dos dados da amostra. Neste sentido o Box-plot é quase tão útil como o histograma. Por exemplo, é relativamente simples avaliar se a distribuição dos dados é simétrica ou assimétrica observando a forma do Box-plot. Se o traço representando a mediana está no centro da caixa definida pelo primeiro e pelo terceiro quartil, a distribuição é simétrica. Se o traço representando a mediana está deslocado para cima, então a

distribuição tem assimetria negativa. Se o traço da mediana está mais próximo do lado inferior da caixa, então o Box-plot mostra que os dados tem distribuição com assimetria positiva.

Os traços verticais que representam os valores máximo e mínimo também ajudam a avaliar a assimetria da distribuição pelo Box-plot. Se o valor do máximo está muito distante da caixa, então é provável que a distribuição tenha assimetria positiva. Se o valor do mínimo é que se afasta mais da caixa, então a distribuição tem assimetria negativa. Caso a distância do máximo e do mínimo até a caixa seja a mesma, então a distribuição é simétrica.

Assimetria	Valor de G	Box-plot	Exemplo de histograma
Nula	0 ou próximo de zero		
Positiva	$G > 0$		
Negativa	$G < 0$		

Outra aplicação muito importante do Box-plot é a comparação rápida entre duas amostras, ou de uma amostra com um ou mais valores individuais.

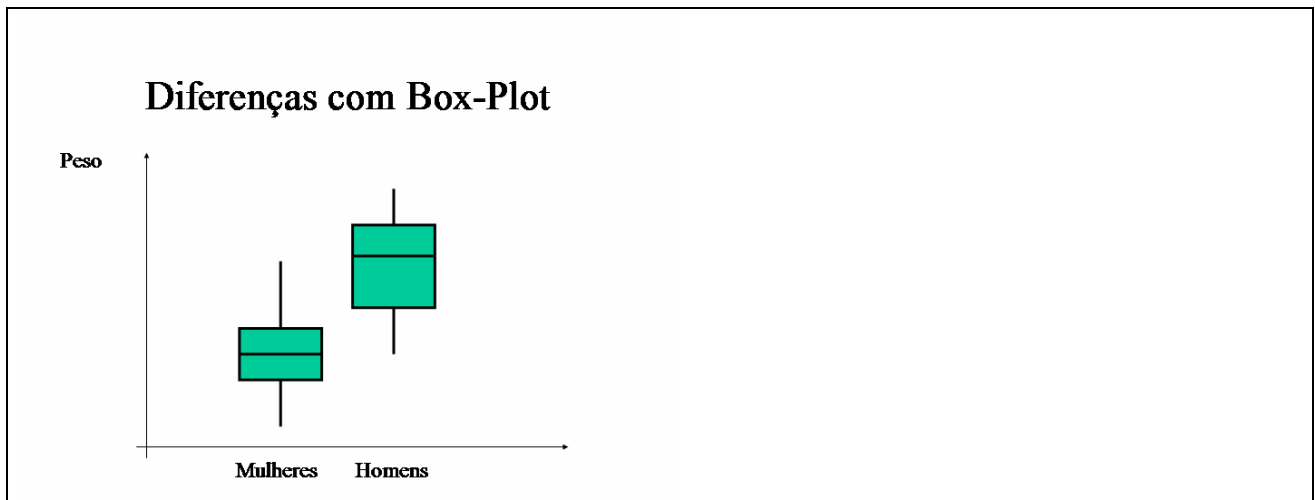


Figura 11: Avaliação de diferenças entre amostras de peso de homens e mulheres utilizando o Box plot..

Relações entre variáveis

Nos capítulos anteriores foram apresentados métodos para analisar dados de uma variável pertencente a uma população. Outro tipo de análise importante na hidrologia é como uma variável se relaciona com as outras, da mesma população. Existem formas de medir o grau de associação entre variáveis e existem métodos para prever o valor de uma variável A, desde que se conheça o valor da variável B que mantém uma relação com a variável A.

Existem muitos exemplos de relações entre variáveis: a velocidade da água de um rio tem relação com a concentração de sedimentos; o nível da água de um rio tem relação com a vazão que está passando por ele; a altura das ondas em um lago tem relação com a velocidade do vento; a temperatura média do ar em Porto Alegre tem relação com o dia do ano; as notas dos alunos do CTH em Estatística tem relação com o número de horas por semana que eles dedicam ao estudo; e assim por diante.

Análise gráfica de relações

Em alguns casos é útil elaborar o gráfico relacionando duas variáveis de um experimento para identificar possíveis relações entre estas variáveis.

Sabe-se que existe uma relação entre a temperatura do ar e a altitude de um determinado local. Podemos testar esta relação com os dados de várias estações meteorológicas do Rio Grande do Sul, coletados ao longo de 10 anos ou mais entre 1957 e 1977. A tabela abaixo apresenta os dados de altitude e de temperatura de 22 estações meteorológicas no RS. São apresentados os dados de temperatura média anual e de temperatura média das máximas no mês de Dezembro.

Podemos explorar se existe uma relação entre os dados de altitude e temperatura elaborando um gráfico relacionando estas duas variáveis, com a altitude no eixo horizontal e a temperatura no eixo vertical. Com base neste gráfico observamos que existe uma tendência de que as temperaturas sejam mais baixas em locais mais altos.

Tabela 1: Dados de altitude e temperatura de 22 estações meteorológicas do RS (fonte IPAGRO)

Estação	Altitude (m)	Temperatura média anual (°C)	Temperatura média das máximas de Dezembro (°C)
Bajé	214	18.7	28.4
Encruzilhada do Sul	420	17.6	26.1
Erexim	760	18.7	27.3
Farroupilha	702	17.3	26.5
Guaíba	46	19.6	28.9
Ijuí	448	20.5	30
Jaguarão	11	18.3	28.2
Júlio de Castilhos	514	18.6	28
Osório	32	19.8	27.8
Passo Fundo	709	18.4	28
Quaraí	100	19.5	30
Rio Grande	16	18.8	26.9
Santa Maria	153	19.7	29.3
Santana do Livramento	210	18.7	28.5
Santo Augusto	380	20.1	29.7
São Borja	99	21	30.4
São Gabriel	109	19.3	27
Taquari	76	20.2	29.4
Tramandaí	3	19.6	25.8
Uruguaiana	74	20.2	30.7
Vacaria	955	16.4	25.4
Veranópolis	705	17.5	26.1

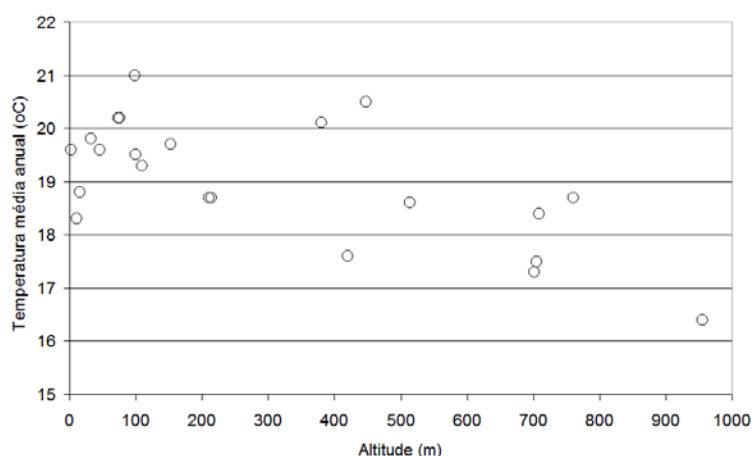


Figura 12: Relação entre altitude e temperatura média anual em 22 estações meteorológicas do Rio Grande do Sul.

Análise de relações em tabelas

Também podemos explorar relações entre variáveis qualitativas e quantitativas através de uma tabela. Considere um conjunto de medições de capacidade de infiltração de solos com o método dos anéis concêntricos. Foram medidos dados em diferentes tipos de solos, que foram classificados como Arenosos ou Argilosos. Uma tabela apresenta a frequência de dados com capacidade de infiltração em faixas: menor que 10; entre 10 e 20; e maior do que 20 mm/hora. A tabela mostra que foram medidos 82 locais, dos quais 39 em locais de solo arenoso e 43 em locais de solo argiloso. Em 36 locais a capacidade de infiltração medida foi inferior a 10 mm/hora, em 21 a capacidade ficou entre 10 e 20 e em 25 pontos a capacidade de infiltração foi superior a 20 mm/hora.

Dos 36 pontos de capacidade de infiltração baixa, 29 ocorreram em solos argilosos, indicando que existe uma relação entre a classe de solo e a capacidade de infiltração. Da mesma forma, dos 25 pontos com capacidade de infiltração alta (>20 mm/hora), apenas 2 ocorreram em solos argilosos e 23 em solos arenosos.

	Aren	Argil	Tot
<10	7/36	29/36	36
10-20	9/21	12/21	21
>20	23/25	2/25	25
tot	39/82	43/82	82

A mesma tabela pode ser elaborada com valores percentuais, deixando ainda mais clara a relação entre capacidade de infiltração e tipos de solos.

	Aren	Argil	Tot
<10	19,4%	81,6%	100%
10-20	42,9%	57,1%	100%
>20	92%	8%	100%
tot	47,6%	52,4%	100%

O coeficiente de correlação

Podemos medir o grau ou a intensidade da relação entre duas variáveis utilizando a estatística. O coeficiente de correlação de Pearson (r), definido nas equações abaixo, permite avaliar o quanto duas variáveis estão relacionadas. O valor de r pode estar entre -1 e 1, e a interpretação dos valores de r é dada na tabela que segue.

$$r = \frac{\left(\sum_{i=1}^n x_i \cdot y_i \right) - n \cdot \bar{x} \cdot \bar{y}}{\sqrt{\left[\sum_{j=1}^n x_j^2 - n \cdot \bar{x}^2 \right] \cdot \left[\sum_{j=1}^n y_j^2 - n \cdot \bar{y}^2 \right]}}$$

$$r = \frac{n \cdot \left(\sum_{i=1}^n x_i \cdot y_i \right) - \left(\sum_{i=1}^n x_i \right) \cdot \left(\sum_{i=1}^n y_i \right)}{\sqrt{n \cdot \left(\sum_{i=1}^n x_i^2 \right) - \left(\sum_{i=1}^n x_i \right)^2} \cdot \sqrt{n \cdot \left(\sum_{i=1}^n y_i^2 \right) - \left(\sum_{i=1}^n y_i \right)^2}}$$

Tabela 2: Interpretação dos valores do coeficiente de correlação de Pearson.

Valor do coeficiente de correlação de Pearson (r)	Interpretação
+1	Existe uma relação linear perfeita e positiva entre as variáveis. À medida que x aumenta y também aumenta.
+0,8 até +1	Existe forte correlação positiva entre as variáveis. À medida que x aumenta y também aumenta.
+0,4 até +0,8	Existe uma correlação positiva moderada entre as variáveis. À medida que x aumenta y também aumenta.
-0,4 a +0,4	Existe pouca correlação linear entre os dados. Os dados podem não ter relação nenhuma ou pode ser necessário avaliar relações não lineares.
-0,4 até -0,8	Existe uma correlação negativa moderada entre as variáveis. À medida que x aumenta y diminui.
-0,8 até -1	Existe forte correlação negativa entre as variáveis. À medida que x aumenta y diminui.
-1	Existe uma relação linear perfeita e negativa entre as variáveis. À medida que x aumenta y diminui.

A Figura 13 apresenta a relação entre pontos ganhos e número de vitórias dos times da primeira divisão de futebol do Brasil na 33ª rodada do campeonato de 2006. O coeficiente de correlação é de 0,98, o que indica uma forte correlação entre pontos e número de vitórias.

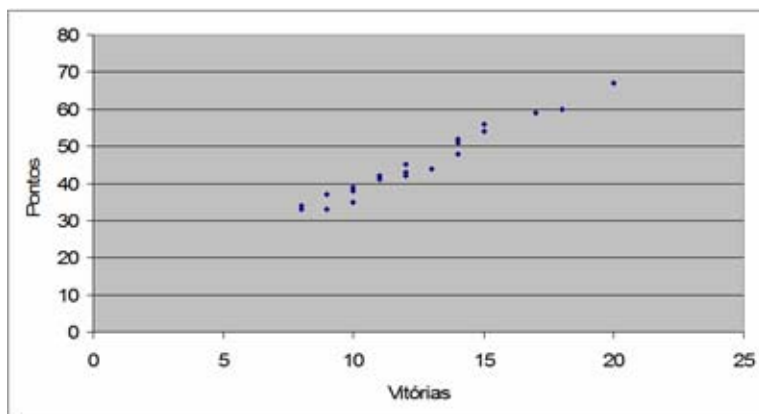


Figura 13: Gráfico de dispersão relacionando vitórias e número de pontos no campeonato brasileiro de futebol.

A Figura 14 apresenta a relação entre pontos ganhos e número de gols marcados na mesma situação. O coeficiente de correlação é de 0,74, o que indica uma forte correlação entre pontos e número de vitórias, mas a relação é mais fraca do que no caso anterior.

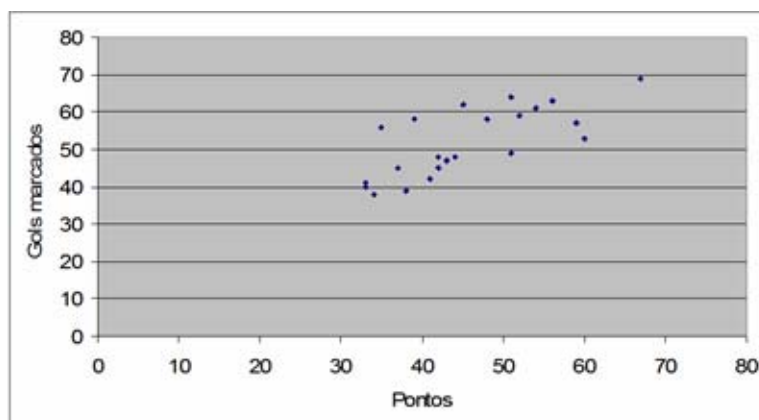


Figura 14: Gráfico de dispersão relacionando pontos e gols marcados no campeonato brasileiro de futebol.

Regressão linear simples

Constatada uma correlação relativamente alta entre duas variáveis, pode ser interessante encontrar uma equação que representa adequadamente a relação entre estas variáveis.

A equação mais simples que pode ser explorada é a equação de uma linha reta. Tomando como exemplo a relação entre altitude e temperatura média analisada antes, podemos tentar traçar manualmente uma linha reta relacionando as duas variáveis diretamente sobre a figura. O problema é que duas pessoas diferentes vão obter retas diferentes, assim é necessário definir matematicamente a equação da reta que melhor representa os dados.

O formato básico de uma equação de reta, ou equação linear é:

$$y = a \cdot x + b$$

ou

$$y = b \cdot x + a$$

ou

$$y = m \cdot x + n$$

Procurando a resposta para a pergunta - Qual é a linha reta que melhor representa os pontos? - chegou-se a conclusão que o ideal é escolher uma linha que minimiza os erros. Pode-se mostrar que é melhor trabalhar com erros médios quadrados, ao invés de erros simples ou dos módulos dos erros. Assim, a equação escolhida é a que minimiza o somatório dos erros quadrados.

Considerando que a equação da reta é $y = a + b \cdot x$; então o erro cometido ao utilizar esta equação para representar um ponto qualquer x_i, y_i , é dado por

$$\text{erro}_i = (a + b \cdot x_i) - y_i$$

e o somatório dos erros quadrados é:

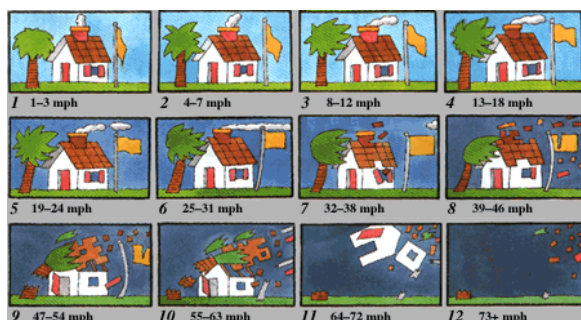
$$SQ = \sum_{i=1}^n ((a + b \cdot x_i) - y_i)^2$$

Derivando esta equação com relação a a e depois b obtemos duas novas equações cujo valor deve ser zero (para entender por que, estude Cálculo I e Cálculo II). Estas duas novas equações representam um sistema de duas equações e duas incógnitas a e b . Resolvendo este sistema chegamos aos valores:

$$y = a + b \cdot x$$

$$b = \frac{n \cdot \left(\sum_{i=1}^n x_i \cdot y_i \right) - \left(\sum_{i=1}^n x_i \right) \cdot \left(\sum_{i=1}^n y_i \right)}{n \cdot \left(\sum_{i=1}^n x_i^2 \right) - \left(\sum_{i=1}^n x_i \right)^2}$$

$$a = \bar{y} - b \cdot \bar{x}$$



A escala de Beaufort classifica o vento em diferentes categorias e relaciona as velocidades do vento com os efeitos que causa no oceano e na terra. Foi criada pelo meteorologista britânico Francis Beaufort no início do século XIX.

Um dos aspectos interessantes da escala de Beaufort é a relação entre velocidade do vento e altura de ondas no mar aberto (Tabela 3: Velocidade do vento e altura das ondas na escala de Beaufort.). Tente desenvolver uma equação que relacione estas duas variáveis.

Tabela 3: Velocidade do vento e altura das ondas na escala de Beaufort.

Vento (nós)	Altura de onda (m)
0,5	0
2,0	0,1
5,0	0,5
8,5	1
13,5	2
19,0	3
24,5	4
30,5	5
37,0	6
44,0	7
51,5	9

Podemos calcular todos os termos que precisamos para estimar o coeficiente b da reta. Temos 11 pontos na tabela o que significa que $n=11$. O somatório dos x é igual a 236, e assim por diante, conforme pode se observar na tabela que segue.

Vento (nós)	Altura de onda (m)		x	y	x.y	x ²
0.5	0		0.5	0	0	1
2	0.1		2	0.1	0.2	4
5	0.5		5	0.5	2.5	10
8.5	1		8.5	1	8.5	17

13.5	2		13.5	2	27	27
19	3		19	3	57	38
24.5	4		24.5	4	98	49
30.5	5		30.5	5	152.5	61
37	6		37	6	222	74
44	7		44	7	308	88
51.5	9		51.5	9	463.5	103
		Soma	236	37.6	1339.2	8132.5

$$\sum_{i=1}^n x_i = 236$$

$$\sum_{i=1}^n y_i = 37.6$$

$$\sum_{i=1}^n x_i \cdot y_i = 1339.2$$

$$\sum_{i=1}^n x_i^2 = 472$$

$$b = \frac{n \cdot \sum xy - \sum x \cdot \sum y}{n \cdot \sum x^2 - (\sum x)^2} = \frac{11 \cdot 1339.2 - 236 \cdot 37.6}{11 \cdot 8132.5 - 236^2} = 0.1735$$

$$\bar{x} = 21.45$$

$$\bar{y} = 3.42$$

$$a = \bar{y} - b \cdot \bar{x} = 3.42 - 0.1735 \cdot 21.45 = -0.3042$$

Portanto, a equação da reta é

$$y = 0.1735 \cdot x - 0.3042$$

A figura apresenta os dados e a reta que foi ajustada a estes dados. O valor R² que aparece no gráfico é o coeficiente de determinação, que é igual ao r, coeficiente de correlação elevado ao quadrado. O coeficiente R² mede o quão bem a reta aproximou os dados. Com R² igual a 1, a aproximação é perfeita. Com R² próximo de zero a reta não representa bem os dados.

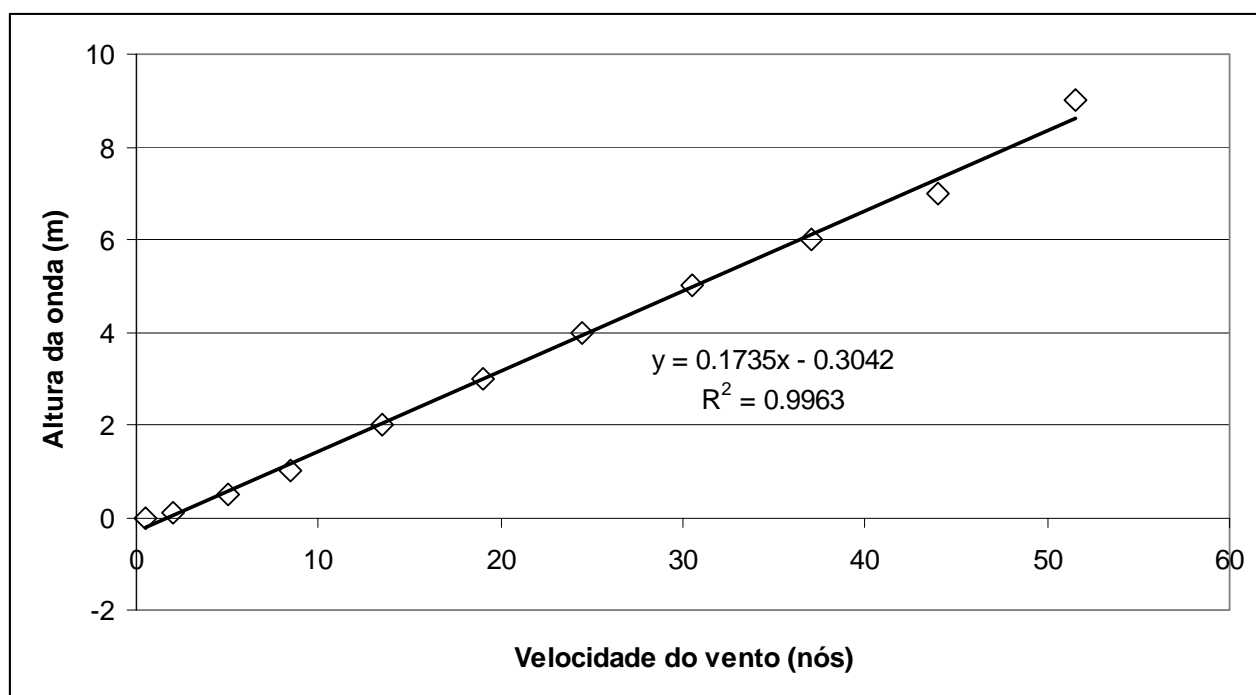


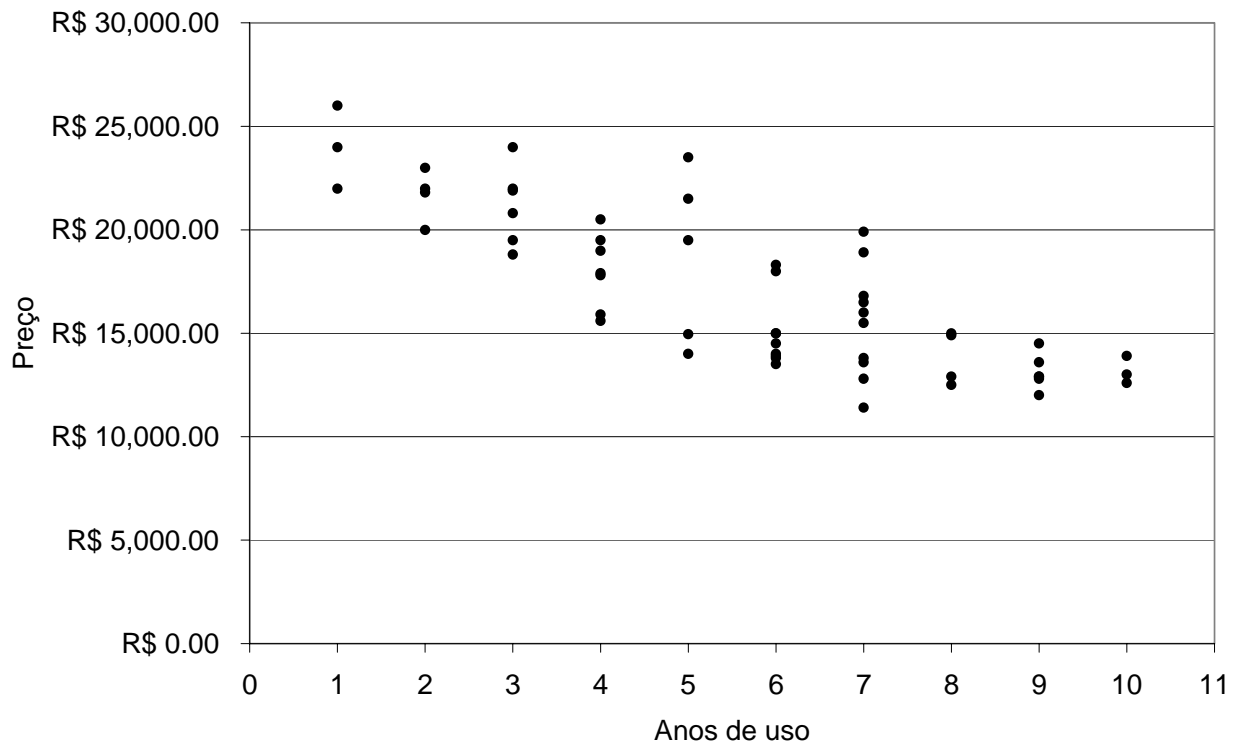
Figura 15: Equação ajustada aos dados de velocidade do vento e altura de ondas da escala de Beaufort.

Regressão não lineares

Algumas variáveis podem apresentar um alto grau de relação, facilmente visível num gráfico, porém eventualmente esta relação pode ser não linear. A tabela a seguir apresenta preços de automóveis usados do modelo Fiat Palio. Considerando que estes dados correspondem ao ano de 2007, podemos contar os anos de uso desde a data de fabricação até 2007 e elaborar um gráfico.

Ano	Preço
2005	R\$ 19990.00
2003	R\$ 18990.00
2006	R\$ 23990.00
2005	R\$ 22990.00
2005	R\$ 21990.00
2004	R\$ 23990.00
2006	R\$ 26000.00
2004	R\$ 20800.00
2006	R\$ 21990.00
2002	R\$ 23500.00
1998	R\$ 14500.00
2003	R\$ 17800.00

1998	R\$ 12800.00
1999	R\$ 15000.00
2002	R\$ 21500.00
1999	R\$ 12900.00
2004	R\$ 21990.00
2001	R\$ 18000.00
2001	R\$ 15000.00
2001	R\$ 15000.00
2000	R\$ 16800.00
2001	R\$ 14000.00
2001	R\$ 18300.00
2000	R\$ 13600.00
2000	R\$ 18900.00
2005	R\$ 21800.00
2000	R\$ 16000.00
2000	R\$ 13800.00
2000	R\$ 11400.00
1998	R\$ 13600.00
2004	R\$ 21900.00
2001	R\$ 14500.00
2001	R\$ 13900.00
2004	R\$ 18800.00
2000	R\$ 15500.00
1999	R\$ 14900.00
1997	R\$ 12600.00
2002	R\$ 19500.00
2000	R\$ 16500.00
2001	R\$ 13800.00
1998	R\$ 12900.00
2003	R\$ 15600.00
1998	R\$ 12000.00
1997	R\$ 13900.00
2001	R\$ 13500.00
2002	R\$ 14950.00
1999	R\$ 12500.00
2003	R\$ 20500.00
2000	R\$ 12800.00
1997	R\$ 13000.00
2003	R\$ 17900.00
2003	R\$ 15900.00
2000	R\$ 19900.00
2003	R\$ 19500.00
1998	R\$ 12900.00
2004	R\$ 19500.00
2002	R\$ 14000.00



Observa-se que a relação é negativa, isto é, quanto maior o tempo de uso, menor o preço do carro. Entretanto, observa-se também que carros muito velhos aproximam-se de um patamar entre 12 e 15 mil reais a partir do oitavo ano. Neste caso, um outro tipo de equação deveria ser avaliado para estimar a relação entre tempo de uso e preço.

O Excel pode ser utilizado para testar várias equações de regressão a um conjunto de dados. Isto pode ser feito usando a ferramenta Solver ou Adicionar Linha de Tendência sobre um gráfico. Alguns critérios devem ser lembrados, entretanto:

- É melhor evitar polinômios de grau alto, mesmo que o R^2 seja mais alto.
- Nunca usar polinômio com grau igual ou superior ao número de pontos.
- R^2 alto nem sempre significa bom ajuste.
- É desejável fazer um gráfico de resíduos da regressão.

Probabilidade e Distribuições de probabilidade

Na tomada de decisões na vida profissional, e até mesmo na vida pessoal, estamos constantemente sendo obrigados a fazer escolhas sem ter certeza sobre diversos aspectos e variáveis que gostaríamos de saber. Muitas vezes estamos considerando implicitamente a probabilidade de termos sucesso ou insucesso. O exemplo mais evidente e mais antigo são os jogos de azar, como o jogo de dados ou o lançamento de uma moeda.

A idéia da probabilidade

O comportamento do acaso é imprevisível a curto prazo mas tem um padrão regular e relativamente previsível a longo prazo.

Quando você lança uma moeda, há apenas dois resultados possíveis: cara ou coroa. A Figura 16 mostra os resultados de lançar uma moeda em duas seqüências de 400 vezes cada uma. Para cada número de lançamentos o gráfico apresenta a proporção dos lançamentos que resultaram em uma cara. A seqüência 1 inicia com “coroa”, e portanto a proporção de caras inicia em zero. No segundo lançamento da seqüência 1 ocorre a face cara, e a proporção de cara passa a ser 0,5. Nos próximos 3 lançamentos também ocorre cara, e assim a proporção cresce até 0,8. A partir daí volta a ocorrer uma coroa, e a proporção volta a cair.

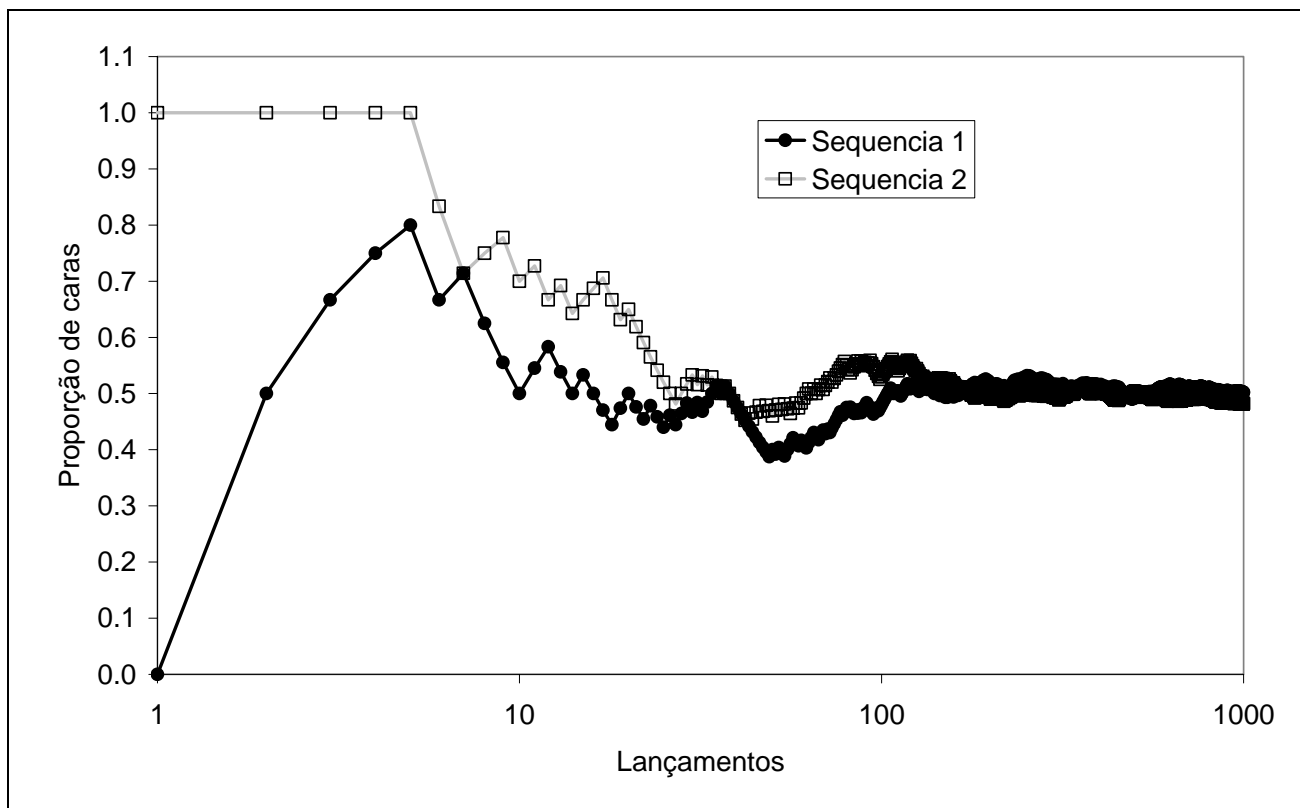


Figura 16: A proporção de lançamentos de uma moeda em que aparece a face “cara” voltada para cima muda à medida em que fazemos mais lançamentos. Finalmente, a proporção atinge 0,5, que é a probabilidade de ocorrer “cara”. Esta figura mostra os resultados de dois ensaios denominados sequência 1 e sequência 2, cada um com 400 lançamentos.

A sequência 2 inicia com “cara” nos cinco primeiros lançamentos, e por isso a proporção de “caras” é de 1 no início do gráfico. Depois ocorrem algumas coroas, e a proporção cai.

Algumas pessoas não ficaram satisfeitas em analisar o problema das probabilidades no lançamento de moedas apenas do ponto de vista teórico, e puseram-se a realizar experimentos. No século XVIII o naturalista francês Conde de Buffon lançou uma moeda 4040 vezes, anotando todos os resultados. Ele observou 2048 caras, o que é uma proporção de 0,5069. Por volta de 1900 o estatístico inglês Karl Pearson heroicamente lançou uma moeda 24.000 vezes, observando 12012 caras, ou seja, uma proporção de 0,5005. Com bastante tempo disponível enquanto esteve preso na segunda guerra mundial, o matemático sul africano John Kerrich lançou uma moeda 10.000 vezes e obteve 5067 caras, uma proporção de 0,5067.

Nas duas sequências a proporção varia entre 0,4 e 0,6 por algum tempo e lentamente vai se aproximando de 0,5, à medida que aumenta o número de lançamentos. Isto significa que a proporção lançamentos que resultam em caras é muito variável no começo. À medida que são feitos mais e mais lançamentos, contudo, a proporção de caras aproxima-se de 0,5 e lá permanece. Se uma terceira sequência

for testada, a proporção de caras, a longo prazo, novamente se estabilizará em 0,5. Dizemos que 0,5 é a probabilidade de ocorrer uma cara.

Uma piada diz que quando viajam de avião os matemáticos sempre levam com eles uma bomba. A justificativa é que a probabilidade de existirem duas bombas no mesmo avião é muito pequena.



Aleatório em Estatística não é sinônimo de “descontrolado”, mas uma descrição de um tipo de ordem que emerge apenas em longo prazo.

Poderíamos suspeitar que uma moeda tem probabilidade 0,5 de aparecer cara apenas porque a moeda tem dois lados, o que significaria que há metade de probabilidade de ocorrência de cada um dos lados. Estas suspeitas não são sempre corretas. No lançamento de um percevejo, por exemplo, há dois resultados possíveis: cair com a ponta para cima, ou cair deitado. Entretanto a probabilidade não é igual para os dois resultados. O percevejo cai deitado com uma probabilidade maior do que com a ponta para cima.

É importante destacar que os lançamentos de um dado ou de uma moeda são eventos independentes, isto é, o resultado de um lançamento não interfere nas probabilidades do próximo. Algumas pessoas que gostam de jogos de azar tendem a esquecer este detalhe, acreditando em “ondas de sorte” ou “marés de azar”.

O estudo da estatística nos coloca frente a frente com situações de incerteza. O seguinte diálogo apresenta um exemplo típico do desafio que representa a compreensão da incerteza:

Estudante: Professor, eu tenho um problema. Decidi fazer um experimento de estatística lançando uma moeda para checar tudo isto que você vem ensinando.

Professor: Muito bem, estou feliz pelo seu interesse. E o que você fez?

Estudante: Eu lancei uma moeda 1000 vezes. Você disse que a probabilidade de obter coroa é um meio. Isto significa que se eu lanço a moeda 1000 vezes eu deveria obter 500 coroas. Mas não funcionou, eu tive 513 coroas. O que está errado?

Professor: Você esqueceu a margem de erro. Quando você lança a moeda um certo número de vezes, a margem de erro é mais ou menos igual a raiz quadrada do número de lançamentos. Para 1000 lançamentos a margem de erro é próxima de 30. Assim, com 513 coroas você está dentro da margem de erro.

Estudante: Ah, agora eu entendi! Toda a vez que eu lançar a moeda 1000 vezes eu vou obter alguma coisa entre 470 e 530 coroas.

Professor: Não, na verdade você provavelmente terá um número de coroas entre 470 e 530, mas não pode ter certeza disso.

Estudante: Você está dizendo que eu poderia ter 200 coroas? Ou talvez 850? Ou mesmo 1000 coroas?

Professor: Provavelmente não.

Estudante: Talvez eu devesse lançar a moeda mais vezes. Se eu tentar um milhão de lançamentos vai funcionar melhor.

Professor: Provavelmente.

Estudante: Por favor, professor, me diga algo em que eu possa acreditar e confiar! Me diga os números!

Professor: Bem, tudo o que eu posso dizer é que eu ficaria surpreso se em mil lançamentos o número de coroas não ficasse entre 470 e 530.

A distribuição normal

Muitos fenômenos aleatórios na natureza seguem a distribuição de probabilidade conhecida como distribuição normal, ou gaussiana. A distribuição normal é descrita em qualquer livro introdutório de estatística e se aplica a muitos tipos de informações da natureza. Um gráfico da função densidade de probabilidade da distribuição normal tem uma forma de sino e é simétrico com relação à média, que é o valor central. A forma em sino indica que existe uma probabilidade maior de ocorrerem valores próximos à média do que nos extremos mínimo e máximo.

A função densidade de probabilidade (PDF) da distribuição normal é uma expressão que depende de dois parâmetros: a média e o desvio padrão da população, conforme a equação seguinte:

$$f_x(x) = \frac{1}{\sqrt{2 \cdot \pi \cdot \sigma_x}} \cdot \exp \left[-\frac{1}{2} \cdot \left(\frac{x - \mu_x}{\sigma_x} \right)^2 \right]$$

onde μ_x é a média da população e σ_x é o desvio padrão da população. Para o caso mais simples, em que a média da população é zero e o desvio padrão igual a 1, a expressão acima fica simplificada:

$$f_z(z) = \frac{1}{\sqrt{2 \cdot \pi}} \cdot \exp \left[-\frac{z^2}{2} \right]$$

onde z é uma variável aleatória com média zero e desvio padrão igual a 1.

O gráfico desta última é apresentado na Figura 17. A área total sob a curva é igual a 1. A área hachurada representa a probabilidade de ocorrência de um valor maior do que z (figura de cima) ou menor do que z (figura de baixo).

A área sob a curva pode ser calculada por integração analítica, mas resulta numa série infinita. Por este motivo, as aplicações práticas são mais comuns na forma de tabelas que relacionam o valor de z com a probabilidade de ocorrer um valor maior do que z ou menor do que z . Existem, também, tabelas que fornecem valores da área entre 0 e z , ou de $-z$ a z .

No final do capítulo é apresentada uma tabela de probabilidades da distribuição normal. No programa Excel é possível obter os valores das probabilidades utilizando a função `DIST.NORMP(z)`, que dá a probabilidade de ocorrer um valor inferior a z .

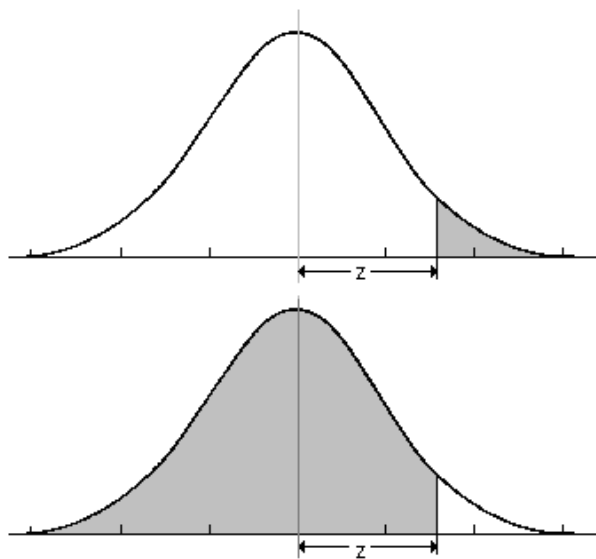


Figura 17: Gráfico da distribuição normal (na figura superior é indicada a área hachurada que representa a probabilidade de ocorrer um valor maior do que z ; e na figura inferior é indicada a área hachurada que representa a probabilidade de ocorrer um valor menor do que z).

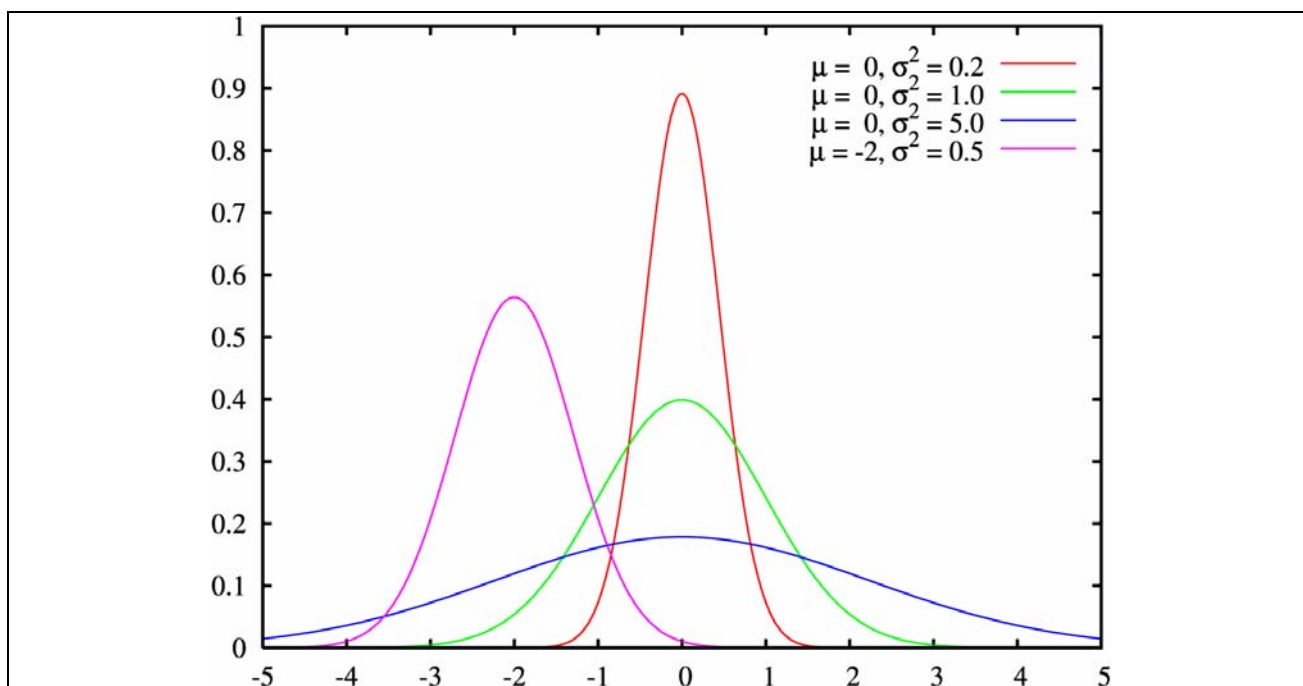


Figura 18: Gráficos da distribuição normal com diferentes valores de média e desvio padrão.

A regra 68-95-99,7

Algumas propriedades da distribuição normal são muitas vezes utilizadas na análise de dados estatísticos. É importante conhecer, por exemplo, a probabilidade de uma observação ocorrer em determinado intervalo de valores. Assim, pode se dizer que, para uma distribuição normal com média μ e com desvio padrão σ :

- 68% das observações estão a menos de σ da média μ ;
- 95% das observações estão a menos de 2σ da média μ ;
- 99,7% das observações estão a menos de 3σ da média μ .

Lembrando destes valores é possível visualizar distribuições normais sem ter de constantemente fazer cálculos ou consultar tabelas.

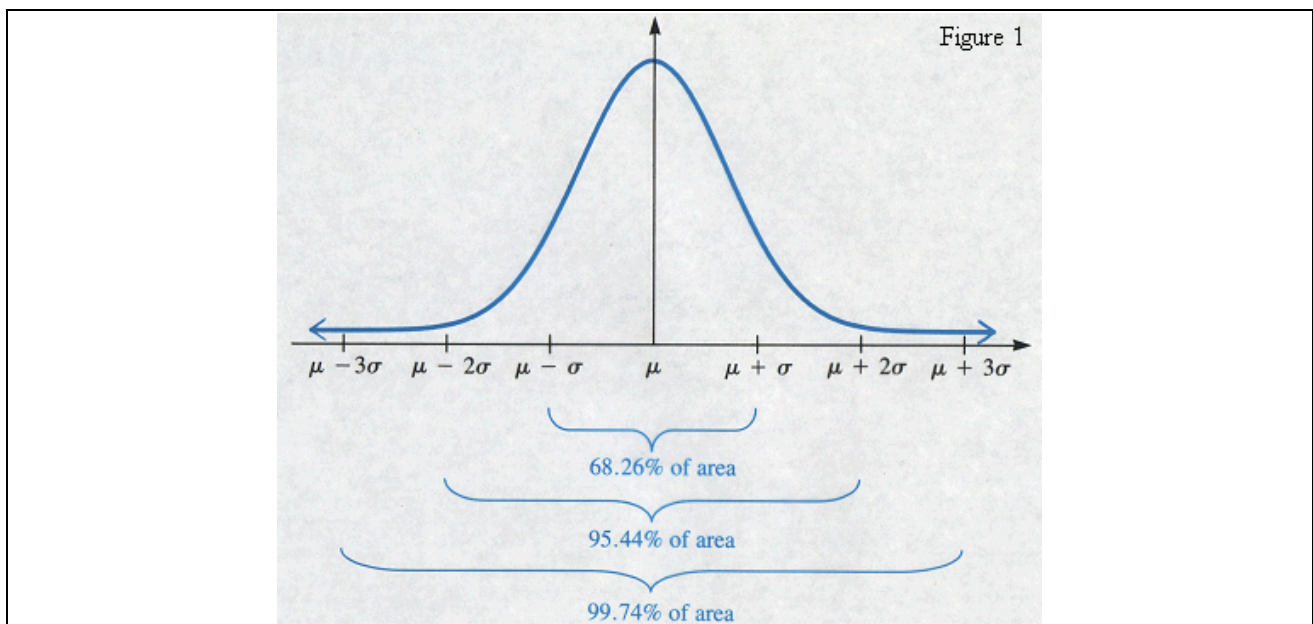


Figura 19: Percentual das observações em diferentes intervalos da distribuição normal.

Por exemplo, se em um vestibular os escores dos alunos se distribuem de forma normal (seguem a distribuição normal) com média 600 e desvio padrão igual a 100, podemos inferir que aproximadamente 68% dos alunos tem escores entre 500 e 700, e que cerca de 99,7% dos alunos tem escores entre 300 e 900.

A distribuição normal padrão

Como sugere a regra 68-95-99,7, todas as distribuições normais compartilham algumas propriedades. Na verdade, todas as distribuições normais são idênticas, se medirmos em unidades de tamanho σ em torno da média μ como centro. A mudança para estas unidades é chamada padronização. Para padronizar o valor de uma observação, subtrai-se a média e divide-se o resultado pelo desvio padrão, como expresso na equação que segue:

Carl F. Gauss foi um dos maiores matemáticos de todos os tempos, tendo se dedicado ainda à topografia e ao estudo da eletricidade e do magnetismo. Viveu entre os anos 1777 e 1855. Quando tinha apenas 3 anos de idade, Gauss surpreendeu seus pais ao corrigir um erro de cálculo que seu pai havia feito. Sem o auxílio de ninguém Gauss havia aprendido a calcular quando mal havia deixado de usar fraldas! Com 15 anos, aproximadamente, Gauss freqüentava diariamente uma escola em sua cidade para a qual se dirigia a pé. Incansavelmente interessado por números, Gauss passava o tempo da caminhada entre a casa e a escola contando os passos que dava. Ao final de algumas semanas percebeu que encontrava números diferentes de passos a cada ida ou volta da escola. Por mais que se esforçasse para dar passos sempre iguais, e para percorrer sempre o mesmo caminho, os números de passos teimavam em variar de forma mais ou menos aleatória. Gauss encontrou o mesmo problema medindo o comprimento de objetos, a área de polígonos e a abertura de ângulos. Ele continuou a explorar este tema e começou a organizar os dados na forma de tabelas e gráficos de frequência. Em quase todos os casos Gauss percebeu que os gráficos mostravam uma curva em forma de sino, que é hoje conhecida como a curva de distribuição normal, curva do sino, ou curva de Gauss.

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

onde x é o valor da observação oriunda de uma distribuição com média μ e desvio padrão σ ; e z é o valor padronizado da observação.

Um valor padronizado z nos diz de quantos desvios padrão a observação original está afastada da média, e em que direção. As observações maiores do que a média tem z positivo enquanto as observações menores do que a média tem z negativo.

A padronização de uma variável que tem uma distribuição normal qualquer gera uma nova variável que tem uma *distribuição normal padrão*. A distribuição normal padrão tem média igual a zero e desvio padrão igual a 1.

Cálculos com a distribuição normal

Uma área sob uma curva de densidade de probabilidade é uma proporção das observações em um determinado intervalo da distribuição. Podemos responder qualquer pergunta acerca de qual proporção de observações está em uma determinada faixa de valores, determinando a área sob a curva. Como todas as distribuições normais são iguais quando padronizadas, podemos determinar áreas sob qualquer curva normal utilizando uma única tabela que forneça as áreas sob a curva para a distribuição normal padrão. No final do capítulo está apresentada uma tabela deste tipo.

Exemplo: As alturas das mulheres adultas seguem uma distribuição normal com média 164,2 e desvio padrão 6,8 cm. Qual é a proporção de todas as mulheres que tem altura superior a 180 cm?

$$Z = (180 - 164,2) / 6,8 = 2,32$$

Podemos procurar na tabela ao final do capítulo qual é a probabilidade de ocorrerem valores maiores do que $z=2,3$, que é suficientemente próximo de 2,32. Neste caso o valor é 0,0107, o que significa que 1,07% das mulheres tem altura superior a 180 cm.

Utilizando as tabelas da distribuição normal também podemos encontrar qual é o valor da variável que cumpra um certo requisito de probabilidade. Por exemplo, no exemplo anterior poderia ser encontrado a altura para a qual 10 % das mulheres são mais altas.

Tabelas da distribuição normal

Tabela A: Probabilidade de ocorrer um valor maior do que Z , considerando uma distribuição normal com média zero e desvio padrão igual a 1.

Z	Probabilidade
0.0	0.5000
0.1	0.4602
0.2	0.4207
0.3	0.3821
0.4	0.3446
0.5	0.3085
0.6	0.2743
0.7	0.2420
0.8	0.2119
0.9	0.1841
1.0	0.1587
1.1	0.1357
1.2	0.1151
1.3	0.0968
1.4	0.0808
1.5	0.0668
1.6	0.0548
1.7	0.0446
1.8	0.0359
1.9	0.0287
2.0	0.0228
2.1	0.0179
2.2	0.0139
2.3	0.0107
2.4	0.0082
2.5	0.0062
2.6	0.0047
2.7	0.0035
2.8	0.0026
2.9	0.0019
3.0	0.0013

Exercícios

- 1) O nível de colesterol no sangue é importante, pois níveis elevados de colesterol podem aumentar o risco de doença do coração. A distribuição de níveis de colesterol no sangue em uma grande população de pessoas da mesma idade e do mesmo sexo é aproximadamente Normal. Para meninos de 14 anos de idade, a média é de 170 miligramas de colesterol por decilitro de sangue (mg/dl) e o desvio padrão é de 30 mg/dl. Os níveis de colesterol acima de 240 mg/dl podem exigir cuidados médicos. Que percentual de meninos com 14 anos tem mais que 240 mg/dl de colesterol?
- 2) Considerando que a chuva média anual em Caxias é 1600 mm por ano, e que o desvio padrão é de 250 mm, qual é a probabilidade de que no ano que vem a chuva anual seja superior a 2500 mm?
- 3) Considere que os salários dos trabalhadores do Unguistão tem uma distribuição normal com média 1500 dólares e desvio padrão de 300 dólares, e que no país vizinho, o Danistão, os salários também tem distribuição normal, mas com média 2000 dólares e desvio padrão de 350 dólares. O sr. Al Amunamed mora no Unguistão e recebe 1800 dólares e o sr. Mustafah Umarah mora no Danistão e recebe 2100 dólares. Comparando com os seus vizinhos do mesmo país, qual dos dois é o que recebe mais? Utilize a padronização dos valores dos salários.

Vazões mínimas e máximas

A vazão de um rio é uma variável que se modifica de forma contínua no tempo, e pode ser representada em um hidrograma, que é o gráfico que relaciona os valores de vazão com o tempo, como na figura 7.7.

Diversas análises estatísticas de dados hidrológicos são realizadas de forma mais conveniente sobre valores discretos no tempo, ao contrário das seqüências contínuas. A partir de uma seqüência contínua de vazões é possível identificar séries temporais de valores discretos, como, por exemplo, as vazões médias anuais, as vazões máximas anuais e as vazões mínimas anuais, conforme representado na figura 7.8 e na tabela 7.1.

As séries discretas que são obtidas a partir da observação de alguns anos de dados de vazão são tratadas como amostras do comportamento de um rio ou de uma bacia. A população, neste caso, seriam todos os anos de existência de um rio. A vazão é considerada uma variável aleatória porque depende de fenômenos climáticos complexos e de difícil previsibilidade a partir de um certo horizonte.

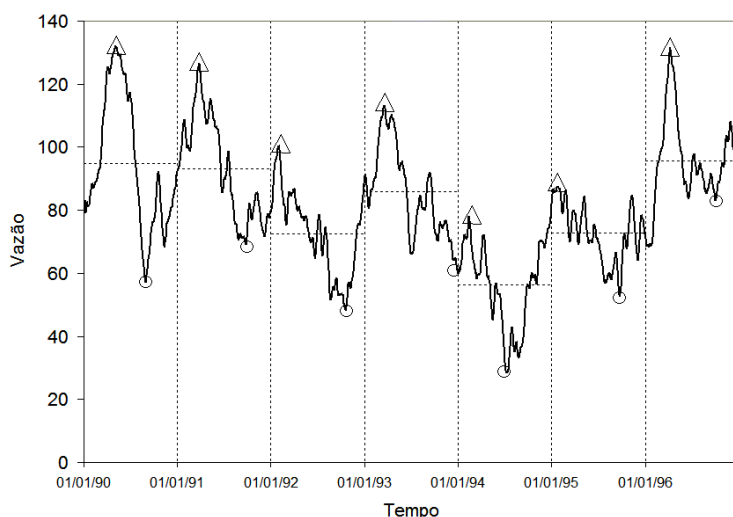


Figura 20: As vazões variam continuamente no tempo (linha) mas a partir dos dados de vazão é possível gerar séries temporais discretas, como as médias, máximas (triângulos) e mínimas (círculos) anuais (adaptado de Dingman, 2002).

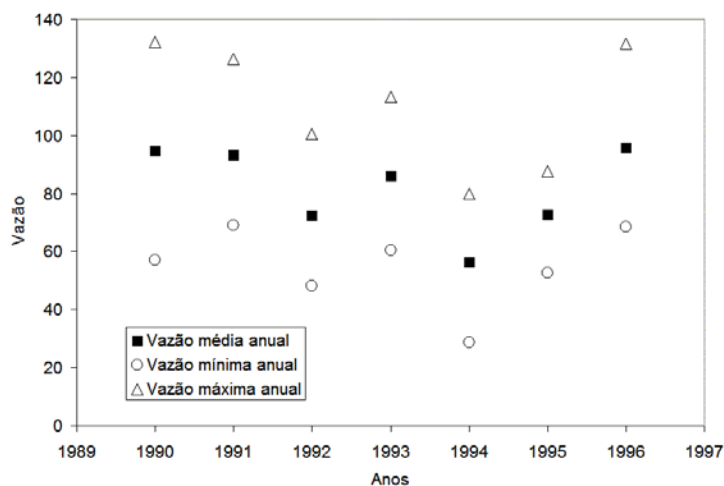


Figura 21: Gráfico das séries discretas de médias, mínimas e máximas anuais.

Tabela 7. 1: Valores das séries temporais discretas de vazões médias, mínimas e máximas anuais relativos à figura anterior.

Ano	Vazão média anual	Vazão mínima anual	Vazão máxima anual
1990	95	57	132
1991	93	69	126
1992	72	48	100
1993	86	60	113
1994	56	29	80
1995	73	53	88
1996	96	68	132

Risco, probabilidade e tempo de retorno

Séries temporais discretas são convenientes para avaliar riscos em hidrologia. Risco é muitas vezes entendido como um sinônimo de probabilidade, mas em hidrologia é mais adequado considerar o risco como a probabilidade de ocorrência de um evento multiplicada pelos prejuízos que se espera da ocorrência deste evento.

Projetos de estruturas hidráulicas sempre são elaborados admitindo probabilidades de falha. Por exemplo, as pontes de uma estrada são projetadas com uma altura tal que a probabilidade de ocorrência de uma cheia que atinja a ponte seja de apenas 1% num ano qualquer. Isto ocorre porque é muito caro dimensionar as pontes para a maior vazão possível, por isso admite-se uma probabilidade, ou risco, de que a estrutura falhe. Isto significa que podem ocorrer vazões maiores do que a vazão adotada no dimensionamento.

A probabilidade admitida pode ser maior ou menor, dependendo do tipo de estrutura. A probabilidade admitida para a falha de uma estrutura hidráulica é menor se a falha desta estrutura provocar grandes prejuízos econômicos ou mortes de pessoas. Assim, a probabilidade de falha admitida para um dique de proteção de uma cidade é a probabilidade de que ocorra uma cheia em que o nível da água supere o nível de proteção do dique. Diques que protegem grandes cidades deveriam ser construídos admitindo uma probabilidade menor de falha do que diques de proteção de pequenas áreas agrícolas. A tabela 7.2 apresenta o tempo de retorno em anos adotado, normalmente, para diferentes tipos de estrutura.

Tabela 4: Tempo de retorno adotado para diferentes estruturas, de acordo com o risco associado.

Estrutura	TR (anos)
Bueiros de estradas pouco movimentadas	5 a 10
Bueiros de estradas muito movimentadas	50 a 100
Pontes	50 a 100
Diques de proteção de cidades	50 a 200
Drenagem pluvial	2 a 10
Grandes barragens (vertedor)	10.000
Pequenas barragens	100

O risco também pode estar relacionado a situações de vazões mínimas. Por exemplo, considere uma cidade que utilize a água de um rio para abastecimento da população. Dependendo do tamanho da população e das características do rio, existe um sério risco de que, num ano qualquer, ocorram alguns dias em que a vazão do rio é inferior à vazão necessária para abastecer a população.

No caso da análise de vazões máximas, são úteis os conceitos de *probabilidade de excedência* e de *tempo de retorno* de uma dada vazão. A probabilidade anual de excedência de uma determinada vazão é a probabilidade que esta vazão venha a ser igualada ou superada num ano qualquer. O tempo de retorno desta vazão é o intervalo médio de tempo, em anos, que decorre entre duas ocorrências subseqüentes de uma vazão maior ou igual. O tempo de retorno é o inverso da probabilidade de excedência como expresso na seguinte equação:

$$TR = \frac{1}{P} \quad (7.1)$$

onde TR é o tempo de retorno em anos e P é a probabilidade de ocorrer um evento igual ou superior em um ano qualquer. No caso de vazões mínimas, P refere-se à probabilidade de ocorrer um evento com vazão igual ou inferior.

A equação acima indica que a probabilidade de ocorrência de uma cheia de 10 anos de tempo de retorno, ou mais, num ano qualquer é de 0,1 (ou 10%).

A vazão máxima de 10 anos de tempo de retorno ($TR = 10$ anos) é excedida em média 1 vez a cada dez anos. Isto não significa que 2 cheias de $TR = 10$ anos não possam ocorrer em 2 anos seguidos. Também não significa que não possam ocorrer 20 anos seguidos sem vazões iguais ou maiores do que a cheia de $TR=10$ anos.

Existem duas formas de atribuir probabilidades e tempos de retorno às vazões máximas e mínimas: métodos empíricos e métodos analíticos.

Probabilidades empíricas podem ser estimadas a partir da observação das variáveis aleatórias. Por exemplo, a probabilidade de que uma moeda caia com a face “cara” virada para cima é de 50%. Esta probabilidade pode ser estimada empiricamente lançando a moeda 100 vezes e contando quantas vezes cada uma das faces fica voltada para cima.

O problema das probabilidades empíricas é que quando o tamanho da amostra é pequeno, a estimativa tende a ser muito incerta. Suponha, por exemplo, que apenas 6 lançamentos sejam feitos para estimar a probabilidade de que uma moeda caia com a face “cara” voltada para cima. É possível que seja estimada uma probabilidade muito diferente de 50%.

Para contornar este problema é comum supor que os dados hidrológicos sejam aleatórios e que sigam uma determinada distribuição de probabilidade analítica, como a distribuição normal, por exemplo. Esta metodologia analítica permite explorar melhor as amostras relativamente pequenas de dados hidrológicos, como se descreve na seqüência deste capítulo.

Chuvas anuais e a distribuição normal

O total de chuva que cai ao longo de um ano pode ser considerado uma variável aleatória com distribuição aproximadamente normal. Esta suposição permite explorar melhor amostras relativamente pequenas, com apenas 20 anos, por exemplo.

A distribuição normal é descrita em qualquer livro introdutório de estatística e se aplica a muitos tipos de informações da natureza. Um gráfico da função densidade de probabilidade da distribuição normal tem uma forma de sino e é simétrica com relação à média, que é o valor central. A forma em sino indica que existe uma probabilidade maior de ocorrerem valores próximos à média do que nos extremos mínimo e máximo.

A função densidade de probabilidade (PDF) da distribuição normal é uma expressão que depende de dois parâmetros: a média e o desvio padrão da população, conforme a equação seguinte:

$$f_x(x) = \frac{1}{\sqrt{2 \cdot \pi} \cdot \sigma_x} \cdot \exp \left[-\frac{1}{2} \cdot \left(\frac{x - \mu_x}{\sigma_x} \right)^2 \right] \quad (7.2)$$

onde μ_x é a média da população e σ_x é o desvio padrão da população. Para o caso mais simples, em que a média da população é zero e o desvio padrão igual a 1, a expressão acima fica simplificada:

$$f_z(z) = \frac{1}{\sqrt{2 \cdot \pi}} \cdot \exp\left[-\frac{z^2}{2}\right] \quad (7.3)$$

onde z é uma variável aleatória com média zero e desvio padrão igual a 1.

O gráfico desta última é apresentado na figura 7.9. A área total sob a curva é igual a 1. A área hachurada representa a probabilidade de ocorrência de um valor maior do que z (figura de cima) ou menor do que z (figura de baixo).

A área sob a curva pode ser calculada por integração analítica, mas resulta numa série infinita. Por este motivo, as aplicações práticas são mais comuns na forma de tabelas que relacionam o valor de z com a probabilidade de ocorrer um valor maior do que z ou menor do que z . Existem, também, tabelas que fornecem valores da área entre 0 e z , ou de $-z$ a z .

No final do capítulo é apresentada uma tabela de probabilidades da distribuição normal. No programa Excel é possível obter os valores das probabilidades utilizando a função `DIST.NORMP(z)`, que dá a probabilidade de ocorrer um valor inferior a z .

Lembrando a relação entre probabilidades e tempos de retorno, é interessante saber os valores de z que correspondem a alguns valores específicos de probabilidade, como 0,1 0,01 e 0,001. Estes valores correspondem aos tempos de retorno de 10, 100 e 1000 anos. No final do capítulo é apresentada uma tabela de probabilidades da distribuição normal, indicando os valores de z correspondentes aos tempos de retorno de 2 a 10000 anos.

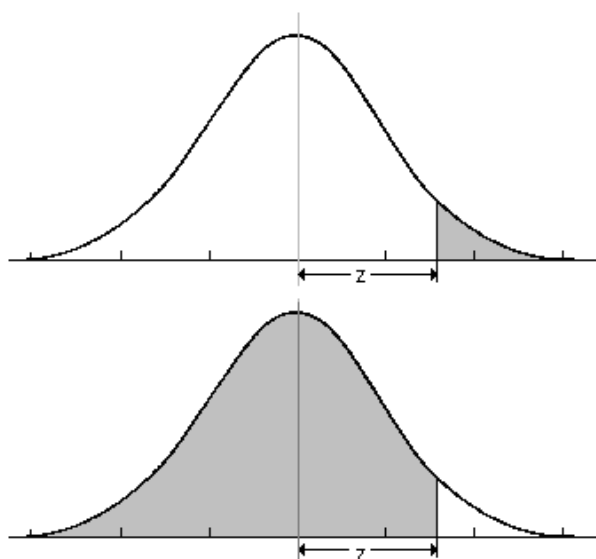


Figura 22: Gráfico da distribuição normal (na figura superior é indicada a área hachurada que representa a probabilidade de ocorrer um valor maior do que z ; e na figura inferior é indicada a área hachurada que representa a probabilidade de ocorrer um valor menor do que z).

Uma variável aleatória x com média μ_x e desvio padrão σ_x pode ser transformada em uma variável aleatória z , com média zero e desvio padrão igual a 1 pela transformação abaixo:

$$z = \frac{x - \mu_x}{\sigma_x} \quad (7.4)$$

Esta transformação pode ser utilizada para estimar a probabilidade associada a um determinado evento hidrológico em que a variável segue uma distribuição normal.

Considere, por exemplo, a chuva anual em um determinado local. Anos com chuva próxima da média são relativamente freqüentes, enquanto anos muito chuvosos ou muito secos são menos freqüentes. Em muitos locais as chuvas anuais seguem, aproximadamente uma distribuição normal, como mostra a figura 7.10.

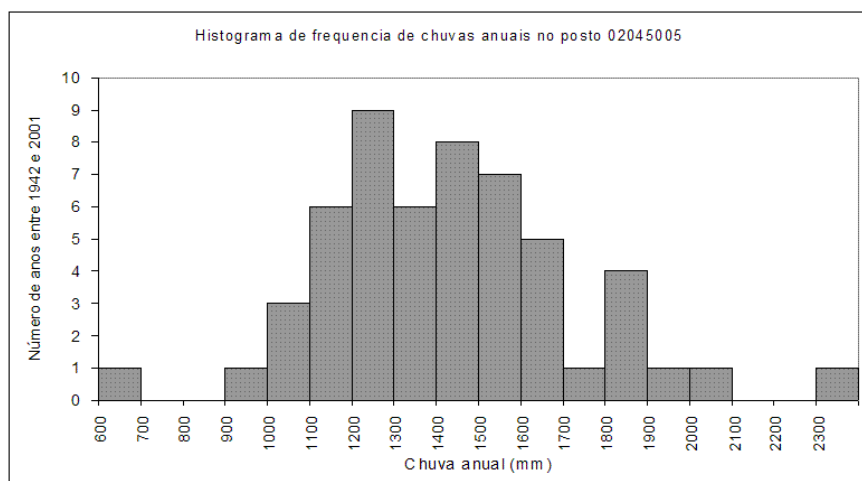


Figura 23: Histograma de freqüências de chuvas anuais no posto pluviométrico localizado em Lamounier, MG (código 02045005 - ver capítulo 3).

A probabilidade de ocorrência de chuvas anuais superiores a 2000 mm, por exemplo, pode ser estimada a partir da análise dos dados de n anos, e da suposição de que os dados seguem uma distribuição normal.

EXEMPLOS

- 2) As chuvas anuais no posto pluviométrico localizado em Lamounier, em Minas Gerais (Código 02045005) seguem, aproximadamente, uma distribuição normal, com média igual a 1433 mm e desvio padrão igual a 299 mm. Qual é a probabilidade de ocorrer um ano com chuva total superior a 2000 mm?

Considerando que a média e o desvio padrão da amostra disponível sejam boas aproximações da média e do desvio padrão da população, pode-se estimar o valor da variável reduzida z para o valor de 2000 mm:

$$z = \frac{x - \mu_x}{\sigma_x} \cong \frac{x - \bar{x}}{s} = \frac{2000 - 1433}{299} = 1,896$$

de acordo com a Tabela A, no final do capítulo, a probabilidade de ocorrência de um valor maior do que $z=1,896$ é de aproximadamente 0,0287 (valor correspondente a $z=1,9$). Portanto, a probabilidade de ocorrer um ano com chuva total superior a 2000 mm é de, aproximadamente, 2,87%. O tempo de retorno correspondente é de pouco menos de 35 anos. Isto significa que, em média, um ano a cada 35 apresenta chuva total superior a 2000 mm neste local.

- 3) As chuvas anuais no posto pluviométrico localizado em Lamounier, em Minas Gerais (Código 02045005) seguem, aproximadamente, uma distribuição normal, com média igual a 1433 mm e desvio padrão igual a 299 mm. Qual é a probabilidade de ocorrer um ano com chuva total inferior a 550 mm?

A distribuição normal é simétrica. A probabilidade de ocorrer um valor superior a z é igual à probabilidade de ocorrer um valor inferior a $-z$. Assim,

$$z = \frac{x - \mu_x}{\sigma_x} \cong \frac{x - \bar{x}}{s} = \frac{550 - 1433}{299} = -2,95$$

de acordo com a Tabela A, no final do capítulo, a probabilidade de ocorrência de um valor maior do que $z=2,95$ está entre 0,0012 e 0,0019. Portanto, a probabilidade de ocorrer um ano com chuva total superior a 2000 mm é de, aproximadamente, 0,15%. O tempo de retorno correspondente é de pouco menos de 666 anos. Isto significa que, em média, um ano a cada 666 apresenta chuva total inferior a 550 mm neste local.

Vazões máximas

Selecionando apenas as vazões máximas de cada ano em um determinado local, é obtida a série de vazões máximas deste local e é possível realizar análises estatísticas relacionando vazão com probabilidade. As séries de vazões disponíveis na maior parte dos locais (postos fluviométricos) são relativamente curtas, não superando algumas dezenas de anos.

Analisando as vazões do rio Cuiabá no período de 1984 a 1992, por exemplo, podemos selecionar de cada ano apenas o valor da maior vazão, e analisar apenas as vazões máximas (tabela 7.3). Reorganizando as vazões máximas para uma ordem decrescente, podemos atribuir uma probabilidade de excedência empírica a cada uma das vazões máximas da série, utilizando a fórmula de Weibull:

$$P = \frac{m}{N + 1} \quad (7.5)$$

onde N é o tamanho da amostra (número de anos); e m é a ordem da vazão (para a maior vazão $m=1$ e para a menor vazão $m=N$). O resultado é apresentado na tabela 7.4.

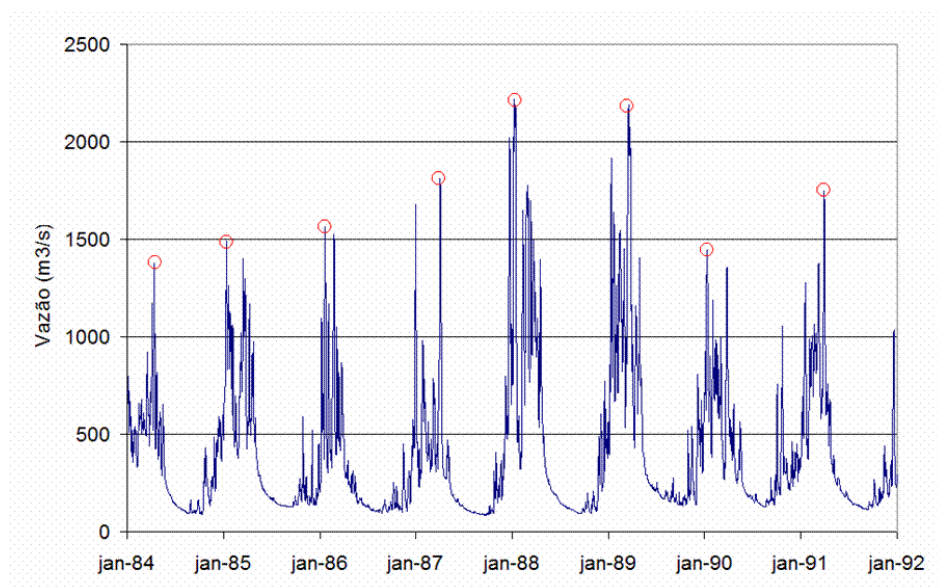


Figura 24: Série de vazões do rio Cuiabá em Cuiabá, de 1984 ao final de 1991, evidenciando a vazão máxima de cada ano.

Tabela 5: Vazões máximas anuais entre 1984 e 1991.

Ano	Q máx
1984	1796.8
1985	1492.0
1986	1565.0
1987	1812.0
1988	2218.0
1989	2190.0
1990	1445.0
1991	1747.0

Tabela 6: Vazões máximas reorganizadas em ordem decrescente, com ordem e probabilidade empírica associada.

Ano	Vazão (m³/s)	Ordem	Probabilidade	TR (anos)
1988	2218.0	1	0.11	9.0
1989	2190.0	2	0.22	4.5
1987	1812.0	3	0.33	3.0
1984	1796.8	4	0.44	2.3
1991	1747.0	5	0.56	1.8
1986	1565.0	6	0.67	1.5
1985	1492.0	7	0.78	1.3

1990 1445.0 8 0.89 1.1

O problema da estimativa empírica de probabilidades é que não é possível extrapolar a estimativa para tempos de retorno maiores. Por exemplo, se é necessário estimar a vazão máxima de 100 anos de tempo de retorno, mas existem apenas 18 anos de dados observados, as probabilidades empíricas permitem estimar vazões máximas de TR próximo de 18 anos.

Para extrapolar as estimativas de vazão máxima é necessário supor que as vazões máximas anuais seguem uma distribuição de probabilidades conhecida, como no caso das chuvas anuais. Infelizmente, porém, as vazões máximas não seguem a distribuição normal. Histogramas de vazões máximas anuais tendem a apresentar uma forte assimetria positiva (longa cauda na direção dos maiores valores), o que invalida o uso da distribuição normal (figura 7.12).

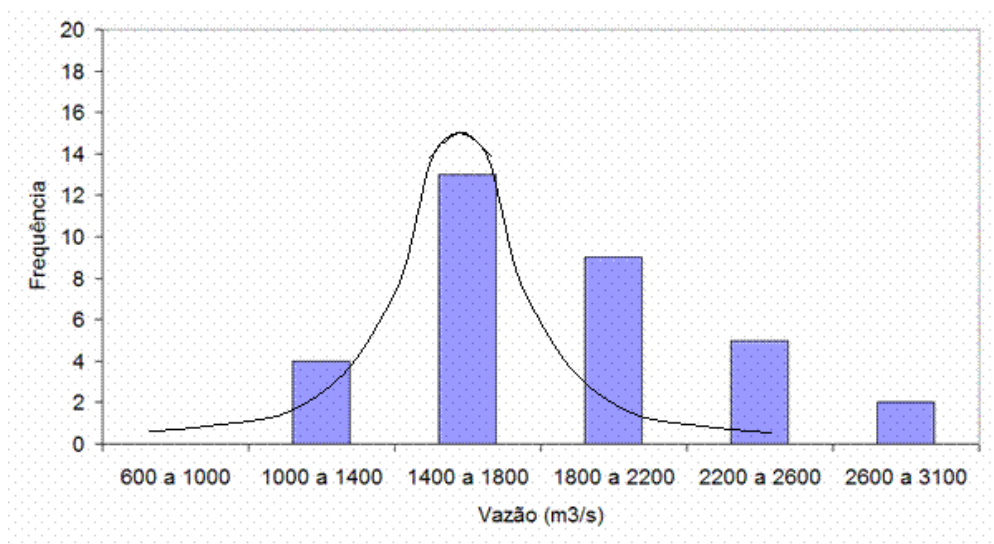


Figura 25: Comparação entre um histograma de vazões máximas observadas do rio Cuiabá em Cuiabá entre 1967 e 1999 e a distribuição normal.

Para superar este problema existem outras distribuições de probabilidade que são, normalmente, utilizadas para a análise de vazões máximas. A mais simples destas distribuições é a denominada log-normal. Nesta distribuição a suposição é que os logaritmos das vazões seguem uma distribuição normal.

Se o objetivo da análise é determinar a vazão de 100 anos de tempo de retorno em um determinado local, por exemplo, a seqüência de etapas para a estimativa supondo que os dados correspondem a uma distribuição log-normal é a seguinte:

- Obter vazões máximas de N anos
- Calcular os logaritmos das vazões máximas
- Calcular a média e o desvio padrão dos logaritmos das vazões máximas

- Obter o valor de z para a probabilidade correspondente ao tempo de retorno de 100 anos
- Obter o valor do logaritmo da vazão de tempo de retorno de 100 anos a partir da equação 7.4
- Obter o valor da vazão através da função inversa do logaritmo.

Esta seqüência de etapas fica mais clara na aplicação em um exemplo.

EXEMPLOS

- 4) As vazões máximas anuais do rio Guaporé no posto fluviométrico Linha Colombo são apresentadas na tabela abaixo. Utilize a distribuição log-normal para estimar a vazão máxima com 100 anos de tempo de retorno.

ANO	MAXIMA	ANO	MAXIMA	ANO	MAXIMA	ANO	MAXIMA	ANO	MAXIMA	ANO	MAXIMA
1940	953	1950	1192	1960	falha	1970	365	1980	653	1990	falha
1941	1171	1951	356	1961	718	1971	671	1981	537	1991	falha
1942	723	1952	246	1962	503	1972	1785	1982	945	1992	falha
1943	267	1953	1093	1963	falha	1973	726	1983	1650	1993	1115
1944	646	1954	840	1964	457	1974	397	1984	1165	1995	639
1945	365	1955	622	1965	915	1975	480	1985	888		
1946	1359	1956	falha	1966	742	1976	falha	1986	728		
1947	411	1957	598	1967	840	1977	673	1987	809		
1948	480	1958	646	1968	331	1978	760	1988	945		
1949	365	1959	953	1969	320	1979	780	1989	1380		

Este exemplo apresenta uma situação muito comum na análise de dados hidrológicos: as falhas. As falhas são períodos em que não houve observação. As falhas são desconsideradas na análise, assim o tamanho da amostra é $N=48$. Utilizando logaritmos de base decimal, a média dos logaritmos das vazões máximas é 2,831 e o desvio padrão é 0,206. Para o tempo de retorno de 100 anos a probabilidade de excedência é igual a 0,01. Na tabela B, ao final do capítulo, pode-se obter o valor de z correspondente ($z=2,326$). A vazão máxima de $TR=100$ anos é obtida por:

$$z \cong \frac{x - \bar{x}}{s}$$

$$2,326 \cong \frac{x - 2,831}{0,206}$$

$$x = 2,326 \cdot 0,206 + 2,831 = 3,31$$

$$Q = 10^{3,31} = 2041$$

Portanto, a vazão máxima de 100 anos de tempo de retorno é 2041 m³/s.

Este procedimento pode ser repetido para outros valores de TR, e o resultado pode ser apresentado na forma de um gráfico, relacionando vazão com tempo de retorno, como na figura 7.13. Nesta figura fica claro, também, que a suposição de uma distribuição log-normal é muito mais adequada do que a suposição de uma distribuição normal.

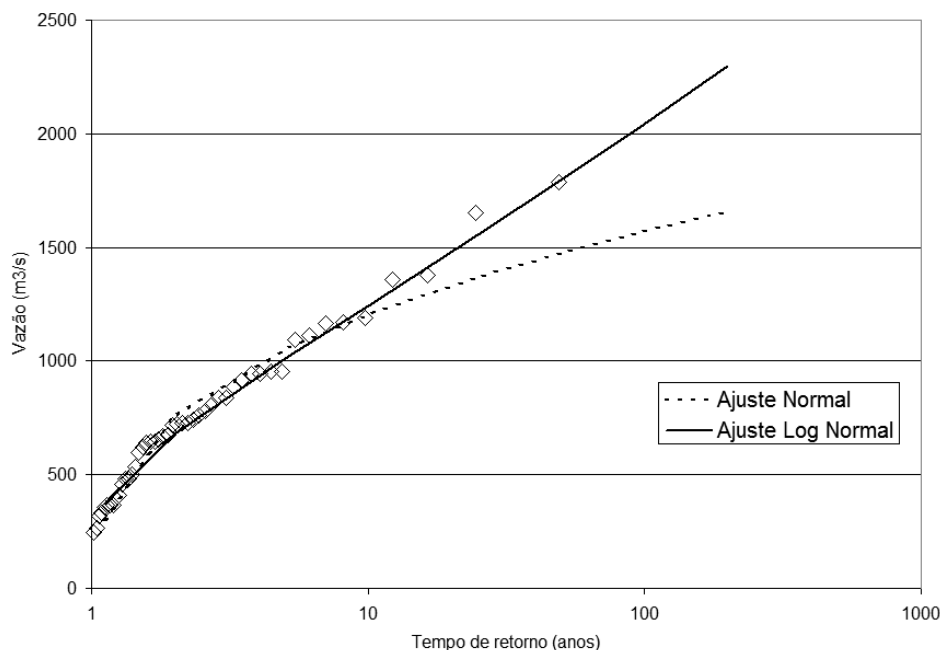


Figura 26: Vazões máximas do rio Guaporé em Linha Colombo. Comparação entre o ajuste e as probabilidades empíricas (pontos), supondo distribuição normal (linha pontilhada) e distribuição log-normal (linha contínua).

Os métodos de estimativa de vazões máximas apresentados neste texto são relativamente simples e a forma de apresentação é resumida. Para realizar análises de vazões máximas mais rigorosas normalmente é necessário testar três ou mais distribuições de probabilidade teóricas, e avaliar qual é a distribuição que melhor se adequa aos dados. Metodologias mais aprofundadas podem ser encontradas em Tucci (1993), Maidment (1993) e Wurbs e James (2001).

Vazões mínimas

A análise de vazões mínimas é semelhante à análise de vazões máximas, exceto pelo fato que no caso das vazões mínimas o interesse é pela probabilidade de ocorrência de vazões iguais ou menores do que um determinado limite.

No caso da análise utilizando probabilidades empíricas, esta diferença implica em que os valores de vazão devem ser organizados em ordem crescente, ao contrário da ordem decrescente utilizada no caso das vazões máximas.

A aplicação da análise estatística para vazões mínimas é analisada através de um exemplo.

EXEMPLOS

- 5) A tabela abaixo apresenta as vazões mínimas anuais observadas no rio Piquiri, no município de Iporã (PR). Considerando que os dados seguem uma distribuição normal, determine a vazão mínima de 5 anos de tempo de retorno. A distribuição normal se ajusta bem aos dados observados?

ano	Vazão mínima
1980	202
1981	128.6
1982	111.4
1983	269
1984	158.2
1985	77.5
1986	77.5
1987	166
1988	70
1989	219.6
1990	221.8
1991	111.4
1992	204.2
1993	196
1994	172
1995	130.4
1996	121.6
1997	198
1998	320.6
1999	101.2
2000	118.2
2001	213

Os valores de vazão mínima são reorganizados em ordem crescente e a probabilidade empírica para cada valor é calculada. A seguir é calculada a média e o desvio padrão do conjunto de dados.

ano	ordem	probabilidade	TR empírico	Vazão mínima
1988	1	0.04	23.0	70
1985	2	0.09	11.5	77.5
1986	3	0.13	7.7	77.5
1999	4	0.17	5.8	101.2
1982	5	0.22	4.6	111.4
1991	6	0.26	3.8	111.4
2000	7	0.30	3.3	118.2
1996	8	0.35	2.9	121.6

1981	9	0.39	2.6	128.6
1995	10	0.43	2.3	130.4
1984	11	0.48	2.1	158.2
1987	12	0.52	1.9	166
1994	13	0.57	1.8	172
1993	14	0.61	1.6	196
1997	15	0.65	1.5	198
1980	16	0.70	1.4	202
1992	17	0.74	1.4	204.2
2001	18	0.78	1.3	213
1989	19	0.83	1.2	219.6
1990	20	0.87	1.2	221.8
1983	21	0.91	1.1	269
1998	22	0.96	1.0	320.6

Média = 163

Desvio padrão = 65.2

Os valores da vazão para diferentes tempos de retorno são calculados por:

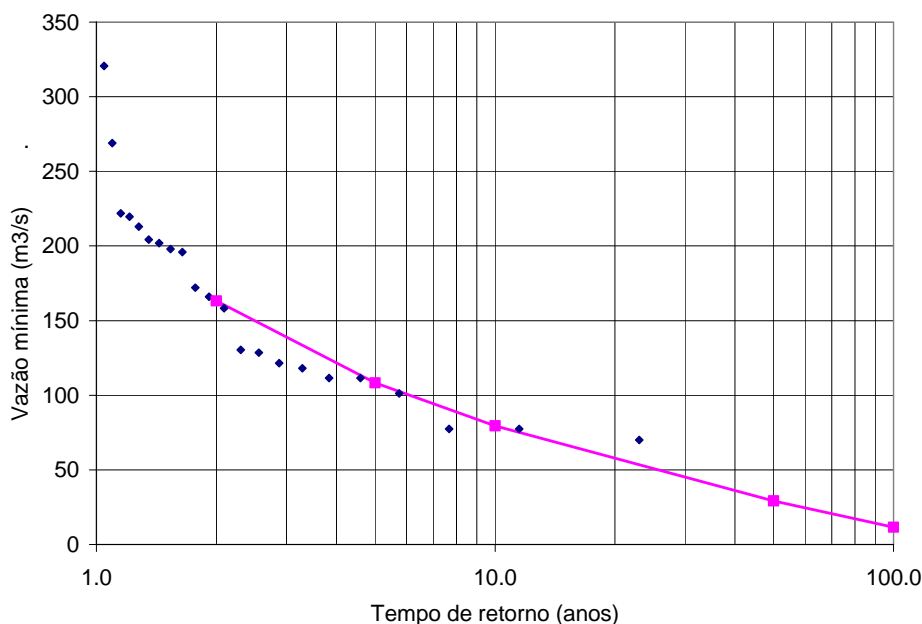
$$Q = \bar{Q} - S_Q \cdot K$$

Onde K é o valor da tabela da distribuição normal para as probabilidades (veja tabela B ao final do capítulo).

Tempo
de
retorno

<i>n</i> o	<i>K</i>	<i>Q</i>
2	0	163.1
5	0.842	108.2
10	1.282	79.5
50	2.054	29.2
100	2.326	11.5

Na figura abaixo vê-se que o ajuste da distribuição normal não é muito bom para estes dados. A vazão mínima com tempo de retorno de 5 anos é estimada em 108 m³/s.



Normalmente, as análises estatísticas de vazões mínimas são realizadas sobre as vazões mínimas de 7 dias, 15 dias ou 30 dias de duração. Neste caso, para cada ano do registro histórico encontra-se a vazão mínima média de 7 dias (médias móveis de 7 dias). O restante do procedimento de análise é semelhante ao apresentado aqui.

A distribuição binomial

A distribuição de probabilidades binomial é adequada para avaliar o número (x) de ocorrências de um dado evento em N tentativas.

As seguintes condições devem existir para que seja válida a distribuição binomial: 1) são realizadas N tentativas; 2) em cada tentativa o evento pode ocorrer ou não, sendo que a probabilidade de que o evento ocorra é dada por P enquanto a probabilidade de que o evento não ocorra é dada por $1-P$; 3) a probabilidade de ocorrência do evento numa tentativa qualquer é constante e as tentativas são independentes, isto é, a ocorrência ou não do evento na tentativa anterior não altera a probabilidade de ocorrência atual.

Estas propriedades ficam mais claras considerando o exemplo de um dado de seis faces. A probabilidade de obter um “seis” num lançamento qualquer é de $1/6$. A probabilidade de não obter um “seis” num lançamento qualquer é de $5/6$. Se um dado é lançado uma vez, resultando em um “seis”, isto não altera a probabilidade de obter um “seis” no lançamento seguinte.

De acordo com a probabilidade binomial, a probabilidade de que um evento ocorra x vezes em N tentativas, é dada pela equação 7.7.

$$P_x(X = x) = \frac{N!}{x!(N-x)!} \cdot P^x \cdot (1-P)^{N-x} \quad (7.7)$$

Nesta equação $P_x(X=x)$ é a probabilidade de que o evento ocorra x vezes em N tentativas. P é a probabilidade que o evento ocorra numa tentativa qualquer e (1-P) é a probabilidade que o evento não ocorra numa tentativa qualquer.

EXEMPLOS

- 6) Calcule a probabilidade de obter exatamente 5 “coroas” em 10 lançamentos de uma moeda.

Neste caso $x=5$ e $N=10$. A probabilidade de obter “coroa” num lançamento qualquer é de 50%, ou $1/2$. A probabilidade de obter exatamente 5 “coroas” pode ser calculada pela equação 7.7.

$$P_x(X = 5) = \frac{10!}{5!(10-5)!} \cdot \left(\frac{1}{2}\right)^5 \cdot \left(1 - \frac{1}{2}\right)^{10-5} = 0,246$$

Portanto, a probabilidade de obter exatamente 5 “coroas” em 10 lançamentos é de 24,6%.

- 7) A probabilidade da vazão de 10 anos de tempo de retorno seja igualada ou excedida num ano qualquer é de 10%. Qual é a probabilidade que ocorram duas cheias iguais ou superiores à cheia de TR = 10 anos em dois anos seguidos?

Neste caso $x=2$ e $N=2$. A probabilidade de ocorrer a cheia num ano qualquer é de 10%, ou $1/10$. A probabilidade de ocorrer exatamente 2 cheias em 2 anos pode ser calculada pela equação 7.7.

$$P_x(X = 2) = \frac{2!}{2!(2-2)!} \cdot \left(\frac{1}{10}\right)^2 \cdot \left(1 - \frac{1}{10}\right)^{2-2} = \left(\frac{1}{10}\right)^2 = 0,01$$

Portanto, a probabilidade de ocorrerem exatamente 2 cheias em 2 anos é 1%.

Tabelas da distribuição normal

Tabela A: Probabilidade de ocorrer um valor maior do que Z, considerando uma distribuição normal com média zero e desvio padrão igual a 1.

Z	Probabilidade
0.0	0.5000
0.1	0.4602
0.2	0.4207
0.3	0.3821
0.4	0.3446
0.5	0.3085
0.6	0.2743
0.7	0.2420
0.8	0.2119
0.9	0.1841
1.0	0.1587
1.1	0.1357
1.2	0.1151
1.3	0.0968
1.4	0.0808
1.5	0.0668
1.6	0.0548
1.7	0.0446
1.8	0.0359
1.9	0.0287
2.0	0.0228
2.1	0.0179
2.2	0.0139
2.3	0.0107
2.4	0.0082
2.5	0.0062
2.6	0.0047
2.7	0.0035
2.8	0.0026
2.9	0.0019
3.0	0.0013

Tabela B: Probabilidade de ocorrer um valor maior do que Z, considerando uma distribuição normal com média zero e desvio padrão igual a 1.

z	Probabilidade	TR
0.000	0.5	2
0.842	0.2	5
1.282	0.1	10
1.751	0.04	25
2.054	0.02	50
2.326	0.01	100
2.878	0.002	500
3.090	0.001	1000
3.719	0.0001	10000

Exercícios

- 3) O que é a curva de permanência?
- 4) Qual é a porcentagem do tempo em que é superada ou igualada a vazão Q_{90} ?
- 5) Se um rio intermitente passa mais da metade do tempo completamente seco, qual é a sua Q_{80} ?
- 6) Qual é o efeito de um reservatório sobre a curva de permanência de vazões de um rio?
- 7) Estime a vazão máxima de projeto para um galeria de drenagem sob uma rua numa área comercial de Porto Alegre, densamente construída, cuja bacia tem área de 35 hectares, comprimento de talvegue de 2 km e diferença de altitude ao longo do talvegue de 17 m.
- 8) Na cidade de Porto Amnésia um apresentador de televisão defende a remoção do dique que protege a cidade das cheias do rio Goiaba. Ele argumenta afirmando que o dique foi dimensionado para a cheia de 50 anos, e que há 65 anos não ocorre na cidade nenhuma cheia que justificaria a construção de qualquer dique. Analise as idéias do apresentador. Calcule qual é a probabilidade de que não ocorra nenhuma cheia de tempo de retorno igual ou superior a 50 anos ao longo de um período de 65 anos.
- 9) Na mesma cidade um arquiteto propõe a substituição de 2000 metros do dique por uma estrutura composta por peças móveis removíveis de 10 m de comprimento. Quando estas peças são expostas à pressão da água equivalente a que ocorreria durante uma cheia, a probabilidade de falha (para cada uma) é de 0,01 %. Qual é a probabilidade de que, durante uma cheia, pelo menos uma das peças venha a falhar?

Distribuição amostral e erros

Após extrair uma amostra de uma população sabemos as respostas dos indivíduos da amostra. Mas muitas vezes não basta ter as informações sobre a amostra. Queremos inferir a partir dos dados amostrais alguma conclusão sobre a população mais ampla que a amostra representa. A *inferência estatística* fornece métodos para extrair conclusões sobre uma população a partir dos dados amostrais.

Quando obtemos dados de uma amostra não podemos ter certeza de que nossas conclusões são corretas. Uma amostra diferente poderia conduzir a conclusões diferentes. A inferência estatística usa a linguagem de probabilidade para expressar o grau de confiança das conclusões.

Considere, por exemplo, os dados da seguinte tabela, que expressam a perda de cálcio (%) de 47 mães durante três meses de amamentação. Considere que as 47 mães são a população. Com a média de uma amostra de 20 mães, extraída desta população, será estimada a média desta população. No entanto, desta mesma população pode-se tomar muitas amostras diferentes de mesmo tamanho ($n=20$). Estas amostras provavelmente terão médias um pouco diferentes entre si. Qual delas é a melhor estimativa da média da população?

-4,7	-2,5	-4,9	-2,7	-0,8	-5,3	-8,3	-2,1	-6,8	-4,3
2,2	-7,8	-3,1	-1,0	-6,5	-1,8	-5,2	-5,7	-7,0	-2,2
-6,5	-1,0	-3,0	-3,6	-5,2	-2,0	-2,1	-5,6	-4,4	-3,3
-4,0	-4,9	-4,7	-3,8	-5,9	-2,5	-0,3	-6,2	-6,8	1,7
0,3	-2,3	0,4	-5,3	0,2	-2,2	-5,1			

Usando os dados das duas primeiras linhas (n=20) a média de perda de cálcio é de menos 4,025 %, já usando os dados das linhas 3 e 4, a média é de -3,705. Nenhuma das duas estimativas é perfeita, já que a média da população, neste caso, é de -3,587.

É claro que, à medida que aumenta o tamanho da amostra, a média da amostra se aproxima cada vez mais da média da população. A lei dos grandes números diz que a média da amostra \bar{x} será próxima da média da população μ se tomarmos uma amostra suficientemente grande. Entretanto, \bar{x} raramente será igual a μ logo, nossa estimativa terá algum erro.

Intervalos de confiança para as médias

O intervalo de confiança indica o quanto eu posso confiar no valor da média obtida da amostra. Por exemplo: A partir dos dados da amostra, eu tenho 95% de certeza que a média está entre 10,54 e 11,21 mg/l.

O intervalo de confiança para a estimativa da média pode ser calculado por:

$$\mu = \bar{x} \pm Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{N}}$$

Isto significa que dada uma amostra com média \bar{x} tirada de uma população com desvio padrão σ , espera-se que a média da população da qual foi extraída a amostra esteja dentro de um intervalo dado por:

$$\bar{x} - Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{N}} < \mu < \bar{x} + Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{N}}$$

O valor de $Z_{\alpha/2}$ depende do grau de confiança que se deseja para a estimativa. Para uma confiança de 95% o valor de $Z_{\alpha/2}$ é de 1,960. Outros valores são dados na tabela abaixo. Estes valores de $Z_{\alpha/2}$ são obtidos da distribuição normal.

Nível de confiança	Valor crítico de $Z_{\alpha/2}$
90%	1,645
95%	1,960
99%	2,576

Intervalos de confiança para as médias com amostras pequenas

Com amostras pequenas ($n < 30$), ou nos casos em que não se conhece o desvio padrão da população, mas apenas o da amostra, o intervalo de confiança para as médias pode ser obtido

$$\mu = \bar{x} \pm t_{\alpha/2} \cdot \frac{s}{\sqrt{N}}$$

Isto significa que dada uma amostra com média \bar{x} e desvio padrão s , espera-se que a média da população da qual foi extraída a amostra esteja dentro de um intervalo dado por:

$$\bar{x} - t_{\alpha/2} \cdot \frac{s}{\sqrt{N}} < \mu < \bar{x} + t_{\alpha/2} \cdot \frac{s}{\sqrt{N}}$$

O valor de $t_{\alpha/2}$ depende do grau de confiança que se deseja para a estimativa e do tamanho da amostra. Com amostras pequenas a distribuição normal tem menor validade. Neste caso é mais adequado utilizar a distribuição de Student (t). Quando o tamanho da amostra supera 30 ($N > 30$) a distribuição de Student e a distribuição normal ficam muito próximas.

Os valores de $t_{\alpha/2}$ da distribuição de Student são dados na tabela abaixo.

Exemplo: A amostra a seguir foi obtida de uma população com distribuição normal. Estime a média da população com intervalo de confiança 90%. (7; 4; 2; 5; 7).

Solução: A média da amostra é 5. O desvio padrão é 2,12. O tamanho da amostra é 5. O número de graus de liberdade é $N-1 = 4$.

Então o valor de $t_{\alpha/2}$ é de aproximadamente 2,132 (este valor pode ser obtido da tabela para a confiança de 90% com 4 graus de liberdade).

$$\bar{x} - t_{\alpha/2} \cdot \frac{s}{\sqrt{N}} < \mu < \bar{x} + t_{\alpha/2} \cdot \frac{s}{\sqrt{N}}$$

$$5 - 2,132 \cdot \frac{2,12}{\sqrt{5}} < \mu < 5 + 2,132 \cdot \frac{2,12}{\sqrt{5}}$$

Assim, a média da população deve estar entre $5-2,02$ e $5+2,02$. Podemos dizer isso com 90% de confiança.

Testes de diferenças de médias

Ver Apresentação Power point

Exercícios

- Foram realizadas 12 medições para calcular o coeficiente C de um orifício a partir de dados de vazão observados em um laboratório utilizando a equação $Q = C \cdot A \cdot \sqrt{2 \cdot g \cdot h}$. Os valores obtidos em cada uma das medições são apresentados na tabela abaixo. Determine o intervalo de confiança para o valor médio de C . Considere um grau de confiança de 95%.

Medição	Coefficiente
1	0,613
2	0,598
3	0,620
4	0,610
5	0,615
6	0,623
7	0,604
8	0,609
9	0,611
10	0,608
11	0,603
12	0,610

- Foram obtidas 15 amostras de sedimentos do leito de um rio. O diâmetro mediano característico d_{50} foi calculado para cada uma das amostras (veja na tabela abaixo). Determine o intervalo de confiança para o valor médio de d_{50} do leito deste rio com base nestes dados. Considere um grau de confiança de 90%.

Amostra	Diâmetro d_{50} (mm)
1	0,739
2	1,098
3	0,820
4	0,910
5	1,015
6	0,623
7	1,504
8	1,609

9	0,811
10	0,758
11	0,923
12	0,918
13	1,001
14	1,200
15	0,998

- 3) Um estagiário do curso técnico de hidrologia fez uma série de medições de velocidade com um molinete chegando ao valor médio de 0,74 m/s. O seu chefe ficou desconfiado do resultado e enviou outro estagiário para repetir as medições. Este obteve os resultados da tabela ao lado. O valor encontrado pelo primeiro estagiário é suspeito? Resolva o problema usando o intervalo de confiança de 90% para a média.

Medição	Velocidade (m/s)
1	0,739
2	0,739
3	0,720
4	0,710
5	0,739
6	0,723
7	0,739
8	0,739
9	0,711
10	0,758
11	0,723
12	0,721
13	0,721
14	0,734
15	0,748



Bibliografia

- Chapra, S.; Canale, R. P. 2006 Numerical methods for engineers. McGraw Hill. New York.
- Gonick, L.; Smith, W. 1993 The cartoon guide to statistics. Harper Perennial. New York.
- Huff, D. 1954 How to lie with Statistics. Editora Norton. New York.
- Lapponi, J. C. 2005 Estatística usando Excel. Editora Campus.
- Magalhães, M. N.; Lima, A. C. P. 2002 Noções de probabilidade e estatística. Editora da USP (EDUSP).
- Moore, D. S. 2005 A estatística básica e sua prática. Editora LTC.
- Smailes, J.; McGrane, A. 2002 Estatística aplicada à administração com Excel. Editora Atlas.
- Tent, M. B. W. 2006 The prince of mathematics Carl Friedrich Gauss. Editora A. K. Peters.