ESTATÍSTICA

Janeiro de 2004

Reginaldo Rocha Caetano

1 INTRODUÇÃO À ESTATÍSTICA *9*

1.1 APRESENTAÇÃO **9**

1.2 OBJETIVOS **9**
1.2.1 INTERPRETAÇÕES 9
1.2.2 CONSTRUÇÃO 9

1.3 DEFINIÇÕES **10**
1.3.1 ESTATÍSTICA 10
1.3.2 POPULAÇÃO 10
1.3.3 AMOSTRA 10

1.4 RAMOS DA ESTATÍSTICA **11**
1.4.1 ESTATÍSTICA DESCRITIVA 11
1.4.2 TEORIA DA PROBABILIDADE 11
1.4.3 INFERÊNCIA 11

1.5 TIPOS DE DADOS **11**
1.5.1 CONTÍNUOS 11
1.5.2 DISCRETOS 11
1.5.3 NOMINAIS 11
1.5.4 ORDINAIS 12

1.6 ETAPAS DO MÉTODO ESTATÍSTICO **12**
1.6.1 DEFINIÇÃO DO PROBLEMA 12
1.6.2 PLANEJAMENTO 12
1.6.3 COLETA DE DADOS 12
1.6.4 APURAÇÃO DOS DADOS 13
1.6.5 APRESENTAÇÃO DOS DADOS 13
1.6.6 ANÁLISE E INTERPRETAÇÃO DOS DADOS 13

1.7 EMBASAMENTO MATEMÁTICO **13**
1.7.1 RELAÇÕES 13
1.7.2 OPERAÇÕES 14

1.8 EXERCÍCIOS **16**

2 DISTRIBUIÇÃO DE FREQÜÊNCIAS *19*

2.1 INTRODUÇÃO **19**

2.2 PREPARAÇÃO DOS DADOS **19**
2.2.1 DADOS BRUTOS 19
2.2.2 ROL 19

2.3 TABULAÇÃO DOS DADOS **19**
2.3.1 DADOS NÃO AGRUPADOS EM CLASSES 20
2.3.2 DADOS AGRUPADOS EM CLASSES 20

2.4 ELEMENTOS DA DISTRIBUIÇÃO **21**
2.4.1 FREQÜÊNCIA SIMPLES ABSOLUTA 21
2.4.2 AMPLITUDE TOTAL 21
2.4.3 NÚMERO DE CLASSES 21
2.4.4 LIMITES DE CLASSE 22
2.4.5 AMPLITUDE DE CLASSE 22
2.4.6 PONTO MÉDIO DA CLASSE 23

2.5 TIPOS DE FREQÜÊNCIAS **23**
2.5.1 FREQÜÊNCIA SIMPLES RELATIVA 24
2.5.2 FREQÜÊNCIA ABSOLUTA ACUMULADA 24
2.5.3 FREQÜÊNCIA RELATIVA ACUMULADA 24

2.6 EXERCÍCIOS **25**

3 GRÁFICOS *29*

3.1 INTRODUÇÃO **29**

3.2 EXEMPLOS DE DISTRIBUIÇÕES **29**
3.2.1 DADOS NOMINAIS 29
3.2.2 DADOS NUMÉRICOS DISCRETOS 29
3.2.3 DADOS NUMÉRICOS CONTÍNUOS 30

3.3 HISTOGRAMAS **31**
3.3.1 DADOS NOMINAIS 31
3.3.2 DADOS NUMÉRICOS DISCRETOS 31
3.3.3 DADOS NUMÉRICOS CONTÍNUOS 33

3.4 POLÍGONO DE FREQÜÊNCIAS **33**
3.4.1 DADOS NOMINAIS 33
3.4.2 DADOS NUMÉRICOS DISCRETOS 33
3.4.3 DADOS NUMÉRICOS CONTÍNUOS 34

3.5 OGIVA DE GALTON **34**
3.5.1 DADOS NOMINAIS 35
3.5.2 DADOS NUMÉRICOS DISCRETOS 35
3.5.3 DADOS NUMÉRICOS CONTÍNUOS 35

3.6 DIAGRAMA DE SETORES **36**
3.6.1 DADOS NOMINAIS 36
3.6.2 DADOS NUMÉRICOS DISCRETOS 37
3.6.3 DADOS NUMÉRICOS CONTÍNUOS 37

3.7 EXERCÍCIOS **38**

4 MEDIDAS DE POSIÇÃO *41*

4.1 MEDIDAS DE TENDÊNCIA CENTRAL **41**
4.1.1 MÉDIA ARITMÉTICA 41
4.1.2 MODA 43
4.1.3 MEDIANA 43
4.1.4 COMPARAÇÃO ENTRE MÉDIA, MEDIANA E MODA 44
($\bar{x}$) 44

4.2 OUTRAS PROMÉDIAS **44**
4.2.1 MÉDIA GEOMÉTRICA 45
4.2.2 MÉDIA HARMÔNICA 45
4.2.3 MÉDIA QUADRÁTICA 46

4.3 DADOS TABULADOS, NÃO AGRUPADOS **46**
4.3.1 MÉDIA ARITMÉTICA 47
4.3.2 MODA 47
4.3.3 MEDIANA 47

4.4 DADOS TABULADOS, AGRUPADOS **47**
4.4.1 MÉDIA ARITMÉTICA 48
4.4.2 MODA 48
4.4.3 MEDIANA 49

4.5 SEPARATRIZES **50**
4.5.1 QUARTIS 50
4.5.2 DECIS 50
4.5.3 CENTIS OU PERCENTIS 51

4.6 EXERCÍCIOS **52**

5 MEDIDAS DE DISPERSÃO *55*

5.1 MEDIDAS INTERVALARES **55**
5.1.1 AMPLITUDE TOTAL 55
5.1.2 DESVIO QUARTIL 56

5.2 DESVIO MÉDIO **57**
5.2.1 DEFINIÇÃO 57
5.2.2 DADOS BRUTOS 57
5.2.3 DADOS TABULADOS 58

5.3 DESVIO PADRÃO PARA DADOS BRUTOS **59**
5.3.1 VARIÂNCIA 59
5.3.2 DESVIO PADRÃO 60
5.3.3 POPULAÇÃO E AMOSTRA 61
5.3.4 PROPRIEDADES 62

5.4 DESVIO PADRÃO PARA DADOS TABULADOS **62**
5.4.1 CÁLCULO DA VARIÂNCIA E DO DESVIO PADRÃO 62
5.4.2 EXEMPLO DA TABELA 64

5.5 MEDIDAS DE DISPERSÃO RELATIVA **65**
5.5.1 COEFICIENTE DE VARIAÇÃO DE PEARSON 65
5.5.2 COEFICIENTE DE VARIAÇÃO DE THORNDIKE 66
5.5.3 DESVIO QUARTIL REDUZIDO 66
5.5.4 COEFICIENTE QUARTÍLICO DE VARIAÇÃO 66

5.6 EXERCÍCIOS **67**

6 MEDIDAS DE DISTORÇÃO *71*

6.1 ASSIMETRIA OU ENVIESAMENTO **71**
6.1.1 SIMÉTRICAS OU SEM DEFORMAÇÃO 71
6.1.2 ASSIMÉTRICA POSITIVA 71
6.1.3 ASSIMÉTRICA NEGATIVA 71

6.2 CURTOSE **72**
6.2.1 MESOCÚRTICA 72
6.2.2 LEPTOCÚRTICA 72
6.2.3 PLATICÚRTICA 72

6.3 DISTRIBUIÇÕES EXEMPLOS **72**
6.3.1 TABELA A 73
6.3.2 TABELA B 73
6.3.3 TABELA C 73
6.3.4 TABELA D 73
6.3.5 TABELA E 73
6.3.6 GRÁFICOS 74

6.4 MOMENTOS **74**
6.4.1 MOMENTO NATURAL OU ABSOLUTO 74
6.4.2 MOMENTO CENTRADO NA MÉDIA 75

6.5 PRINCIPAIS MEDIDAS DE ASSIMETRIA **76**
6.5.1 COMPARAÇÃO ENTRE PROMÉDIAS 76
6.5.2 COEFICIENTE OU ÍNDICE DE PEARSON 77
6.5.3 COEFICIENTE QUARTIL DE ASSIMETRIA 78
6.5.4 COEFICIENTE PERCENTÍLICO DE ASSIMETRIA 78
6.5.5 COEFICIENTE MOMENTO DE ASSIMETRIA 79

6.6 MEDIDAS DE CURTOSE **80**
6.6.1 COEFICIENTE PERCENTÍLICO DE CURTOSE 80
6.6.2 COEFICIENTE MOMENTO DE CURTOSE 80

6.7 EXERCÍCIOS **82**

7 CORRELAÇÃO E REGRESSÃO *83*

7.1 CORRELAÇÃO LINEAR SIMPLES **83**
7.1.1 COEFICIENTE DE CORRELAÇÃO DE PEARSON 83
7.1.2 CORRELAÇÃO POSITIVA 84
7.1.3 CORRELAÇÃO NEGATIVA 84
7.1.4 CORRELAÇÃO NULA 84
7.1.5 CORRELAÇÃO ESPÚRIA 84
7.1.6 EXEMPLO 84

7.2 CORRELAÇÃO ORDINAL **85**
7.2.1 COEFICIENTE DE SPEARMAN 86
7.2.2 COEFICIENTE GAMA DE GOODMAN E KRUSKAL 87

7.3 REGRESSÃO LINEAR SIMPLES **88**
7.3.1 AJUSTAMENTO DO MODELO 88
7.3.2 PODER EXPLICATIVO DO MODELO 88
7.3.3 EXEMPLO 88

7.4 REGRESSÃO LINEAR POR TRANSFORMAÇÃO **89**
7.4.1 FUNÇÃO POTENCIAL 89
7.4.2 FUNÇÃO EXPONENCIAL 90
7.4.3 FUNÇÃO HIPERBÓLICA 91

7.5 EXERCÍCIOS **93**

8 PROBABILIDADE *97*

8.1 INTRODUÇÃO À PROBABILIDADE **97**
8.1.1 INTRODUÇÃO 97
8.1.2 CÁLCULO DE PROBABILIDADES 97

8.2 DISTRIBUIÇÃO DE POISSON **99**

8.3 DISTRIBUIÇÃO BINOMIAL **99**
8.3.1 INTRODUÇÃO 99
8.3.2 FORMA GERAL 99
8.3.3 TERMO GERAL 100
8.3.4 TABELAS 101

8.4 DISTRIBUIÇÃO NORMAL **101**
8.4.1 DEFINIÇÃO 101
8.4.2 IMPORTÂNCIA 101

8.5 RELAÇÕES ENTRE DISTRIBUIÇÕES **102**
8.5.1 PRINCIPAIS PARÂMETROS 102
8.5.2 APROXIMAÇÃO BINOMIAL POR POISSON 102
8.5.3 APROXIMAÇÃO POISSON POR NORMAL 102
8.5.4 APROXIMAÇÃO BINOMIAL POR NORMAL 103

8.6 CURVA NORMAL E PROBABILIDADE **103**
8.6.1 ÁREA SOB A CURVA NORMAL 103
8.6.2 ACIMA DA MÉDIA 103
8.6.3 EM TORNO DA MÉDIA 104

8.7 CURVA NORMAL PADRONIZADA **104**
8.7.1 USO DA TABELA 105
8.7.2 ESCORE Z 105
8.7.3 EXEMPLO 105

8.8 EXERCÍCIOS **107**

9 AMOSTRAGEM E ESTIMAÇÃO *109*

9.1 AMOSTRAS E POPULAÇÕES **109**

9.2 TIPOS DE AMOSTRAGEM **109**
9.2.1 AMOSTRAGEM NÃO ALEATÓRIA 109
9.2.2 AMOSTRAGEM ALEATÓRIA 110

9.3 DISTRIBUIÇÃO AMOSTRAL DAS MÉDIAS **110**
9.3.1 ERRO AMOSTRAL 110
9.3.2 TEOREMA DO LIMITE CENTRAL 111

9.4 ESTIMATIVA DE MÉDIAS **111**
9.4.1 DISTRIBUIÇÃO NORMAL DAS MÉDIAS 112
9.4.2 ERRO PADRÃO DA MÉDIA 112
9.4.3 INTERVALOS DE CONFIANÇA 113

9.5 ESTIMATIVA DE PROPORÇÕES **115**
9.5.1 DEFINIÇÃO 115
9.5.2 EXEMPLO 115

9.6 EXERCÍCIOS **117**

10 TESTES DE SIGNIFICÂNCIA *119*

10.1 INTRODUÇÃO **119**
10.1.1 TESTES PARAMÉTRICOS 119
10.1.2 TESTES NÃO PARAMÉTRICOS 119

10.2 TESTE DE HIPÓTESES **120**
10.2.1 HIPÓTESE NULA 120
10.2.2 HIPÓTESE ALTERNATIVA 120
10.2.3 NÍVEL DE SIGNIFICÂNCIA 120
10.2.4 TIPOS DE ERROS 121

10.3 DISTRIBUIÇÃO DAS DIFERENÇAS **121**
10.3.1 DEFINIÇÃO 121
10.3.2 ERRO PADRÃO DA DIFERENÇA 122

10.4 ESTATÍSTICA Z **123**
10.4.1 TESTE 123
10.4.2 EXEMPLO 123

10.5 ESTATÍSTICA T **124**
10.5.1 PEQUENAS AMOSTRAS 124
10.5.2 DADOS PAREADOS 125
10.5.3 VARIÂNCIAS IGUAIS 126
10.5.4 VARIÂNCIAS DIFERENTES 127

10.6 ESTATÍSTICA F **127**

10.7 QUI-QUADRADO **128**
10.7.1 O TESTE QUI-QUADRADO 128
10.7.2 LIMITAÇÕES DO TESTE QUI-QUADRADO 130

10.8 EXERCÍCIOS **131**

APÊNDICE TABELAS *135*

A ESTATÍSTICA Z **135**

B ESTATÍSTICA T **137**

C ESTATÍSTICA F **139**

D QUI-QUADRADO $\chi 2$ **143**

1 INTRODUÇÃO À ESTATÍSTICA

1.1 APRESENTAÇÃO

A matemática necessária em estatística gerou uma certa aversão injustificada em relação a esta matéria. Isso pelo fato de que os cálculos estatísticos, via de regra, utilizam muitos dados e as operações são repetitivas, embora bastante simples. Antes do advento das calculadoras e dos computadores, o trabalho cansativo levava a imprecisões que punham por terra imensos esforços, ante qualquer engano. Com a chegada destas tecnologias o estudioso passou a poder ficar restrito a questões mais conceituais e, deste modo, gradualmente, este preconceito deixou de existir.

1.2 OBJETIVOS

1.2.1 INTERPRETAÇÕES

LITERATURA

A compreensão de revistas especializadas exige o domínio da matéria para que se tenha a noção correta da real abrangência das conclusões da publicação. A análise de dados disponíveis, mesmo quando coletados por terceiros, muitas vezes permitem preciosas inferências acerca de determinado assunto.

MODELOS

Os modelos são versões simplificadas de problemas e situações da vida real. Eles permitem a ilustração de aspectos que interessam, sem levar em conta detalhes irrelevantes aos objetivos em questão. São didáticos, funcionam como ideais a serem perseguidos e prestam-se como testes de idéias antes de implementá-las. Tabelas, mapas, gráficos e equações são exemplos de modelos usados em estatística.

1.2.2 CONSTRUÇÃO

REALIZAÇÃO

A introdução do aluno na montagem de estatísticas que o auxiliem em trabalhos de pesquisa é objetivo da disciplina. Se o aprofundamento matemático não é o maior, o desenvolvimento de um espírito criterioso na montagem e análise dos dados coligidos será rigoroso. Isto é necessário para que o estudante possa aprofundar o trato da matéria em estudos subsequentes ou, pelo menos, dialogar com especialistas em um bom nível de profundidade.

TOMADA DE DECISÕES

O número de variáveis que influi em qualquer fenômeno do cotidiano, é muito grande, o que dificulta um tratamento determinístico, quando não o torna impossível. Só esse fato já basta para justificar o estudo de estatística pelos estudantes , sejam eles das áreas das ciências exatas, humanas ou da saúde. Se ele souber utilizar esta ferramenta em sua profissão, suas chances de sucesso ficarão ampliadas, porque suas decisões serão embasadas em um método científico adequado.

1.3 DEFINIÇÕES

1.3.1 ESTATÍSTICA

DUPLO SIGNIFICADO DA PALAVRA

Embora muitos não gostem de qualquer definição, a de Fisher satisfaz: "ramo da matemática aplicada, dedicado à análise de dados de observação". A palavra estatística, além disso, é usada para descrever os próprios dados. Essa dupla significação do termo, ora para falar do ramo da matemática, ora para definir um conjunto de dados numéricos, gera confusão algumas vezes.

DADOS E INFORMAÇÃO

Os dados brutos devem ser processados para que se tornem informações. Antes disso ele parecem não ter sentido, tendendo a confundir ao invés de esclarecer. Com o processamento, a quantidade de detalhes é reduzida, facilitando a constatação das relações. Os dados são condensados e transformados em gráficos, mapas e números simplificados, tudo isso facilitando a compreensão do essencial.

1.3.2 POPULAÇÃO

População ou universo é o conjunto da totalidade dos indivíduos sobre o qual se faz uma inferência. Inferência, em estatística, é o ato de analisar e tirar conclusões. Por indivíduos entende-se os elementos ou objetos sob análise. A população pode ser finita ou infinita - número de medidas de pesos de uma mesma pessoa para determinar a distribuição das medidas, por exemplo. Na maioria dos casos, é antieconômico ou impraticável examinar toda a população, por isso usa-se uma amostra. Em outros casos, a coleta do dado destrói o objeto - testes com cintos de segurança, por exemplo.

1.3.3 AMOSTRA

Amostra é um subconjunto da população. A partir da amostra, faz-se juízo ou inferência sobre as características da população. Estas características da população são denominadas parâmetros. Uma aplicação interessante na área de saúde são os chamados grupos experimentais e de controle. Num são aplicados medicamentos e no outro um falso medicamento, sabidamente inócuo, para que os

indivíduos não saibam a qual grupo pertencem. os dois grupos amostrais são acompanhados a fim de que se obtenha resultados conclusivos.

1.4 RAMOS DA ESTATÍSTICA

1.4.1 ESTATÍSTICA DESCRITIVA

É a parte da estatística que utiliza números para descrever fatos. Compreende a organização, resumo e simplificação de informações que podem ser muito complexas. Lança mão de tabelas e gráficos para atingir esses objetivos.

1.4.2 TEORIA DA PROBABILIDADE

É o ramo da estatística que trabalha com o acaso. Situações que envolvam lançamento de moedas, jogos de cartas, dados, enfim, quaisquer situações onde não exista relação causal, ou mesmo que ela exista, não seja conveniente valer-se dela.

1.4.3 INFERÊNCIA

É a parte da estatística que analisa e interpreta os dados amostrais. A idéia é analisar parte da população que seja típica e a partir daí concluir sobre toda a população. Experimentar um doce e colocar o dedo na água quente são exemplos de amostragens que permitem boas inferências sobre o gosto e temperatura, respectivamente. Uma parte importante da estatística é a que dá a segurança de que a as conclusões tiradas da população a partir da amostra são válidas.

1.5 TIPOS DE DADOS

Os dados selecionados como de interesse da estatística e que admitem certa variabilidade são chamados de variáveis. Estas variáveis originam-se de dados de diversos tipos:

1.5.1 CONTÍNUOS

Podem assumir qualquer valor em um intervalo contínuo. Peso, média de consumo de combustível ,etc...

1.5.2 DISCRETOS

Podem assumir apenas valores determinados, normalmente inteiros. Número de alunos em aula, quantidade de acidentes em uma cidade, etc..

1.5.3 NOMINAIS

Os dados são divididos em categorias e contados os elementos. Como nesse caso os dados não são numéricos, mas qualitativos, as variáveis são denominadas atributos. Sexo (masculino ou feminino) de um grupo de pessoas e conceitos (ótimo, bom, regular, ruim) em provas são

exemplos de dados nominais. Por vezes, os alunos têm alguma dificuldade em distinguí-los dos dados discretos pelo fato de ambos serem inteiros. Lá, a variável discreta pode assumir qualquer valor inteiro; aqui, a variável, que é um atributo, recebe valores inteiros, frutos de contagens.

1.5.4 ORDINAIS

Dados ordinais ou por postos, como são denominados muitas vezes, são valores relativos atribuídos para denotar ordem - classificação em concurso, de preços, etc...

1.6 ETAPAS DO MÉTODO ESTATÍSTICO

1.6.1 DEFINIÇÃO DO PROBLEMA

Certificar-se claramente da finalidade do estudo ou análise e, a partir daí, definir cuidadosamente o problema. É útil examinar outros levantamentos sobre o mesmo assunto ou similares, pois isso pode simplificar ou orientar a pesquisa.

1.6.2 PLANEJAMENTO

Essa fase consiste em determinar o procedimento necessário para resolver o problema. Definir os dados a serem obtidos. Como obtê-los? Escolher as perguntas e o modo de formulá-las para não tornar a pesquisa tendenciosa. O levantamento pode ser censitário, em que todo o universo é contado, ou pode ser amostral, quando devem ser tomados cuidados para que se permita inferir os parâmetros da população.

1.6.3 COLETA DE DADOS

É a parte operacional. Consiste em obter, reunir e registrar os dados de forma sistemática.

TIPOS DE FONTES

- PRIMÁRIAS - quando os dados são recolhidos para o trabalho.
- SECUNDÁRIAS - quando colhidos em outro contexto. Nesse caso, os dados devem ser usados com cautela, pois estão sujeitos a algumas limitações.

MODOS DE COLETA

- DIRETA - quando obtida diretamente da fonte. Pode ser contínua, periódica ou ocasional.
- INDIRETA - quando é inferida a partir de outros dados. Pode ser realizada por analogia, por proporcionalização, por indícios ou por avaliação.

1.6.4 APURAÇÃO DOS DADOS

É a fase de sumarização dos dados, mediante contagem e agrupamento. É um trabalho de condensação e de tabulação. A apuração pode ser manual, mecânica, eletromecânica e eletrônica.

1.6.5 APRESENTAÇÃO DOS DADOS

TABULAR

É a apresentação numérica dos dados através de tabelas. Fornece informação quantitativa e organizada.

GRÁFICA

É a apresentação geométrica dos dados. Fornece uma visão rápida de conjunto.

1.6.6 ANÁLISE E INTERPRETAÇÃO DOS DADOS

É nessa fase que são tiradas as conclusões que auxiliam o pesquisador a resolver o problema. Daí saem os números que resumem as conclusões, as chamadas estatísticas. Se o trabalho for realizado sobre amostras e não sobre toda a população, a generalização para o universo é possível através da Estatística Indutiva ou Inferência Estatística.

1.7 EMBASAMENTO MATEMÁTICO

1.7.1 RELAÇÕES

PROPORÇÃO

Se N de indivíduos são divididos em n categorias, o quociente do número Ni da categoria i, qualquer, pelo número total N é a proporção da categoria i. Então:

$$\text{proporção} = \frac{N_i}{N}$$

E a soma das proporções de todas as categorias deve ser igual à unidade:

$$\frac{N_1}{N} + \frac{N_2}{N} + \ldots + \frac{N_i}{N} + \ldots + \frac{N_n}{N} = 1$$

PERCENTAGEM

É a proporção multiplicada por cem. A soma dos percentuais de todas as categorias é, portanto, cem.

RAZÃO

A razão assemelha-se à proporção, mas o denominador não é o número total e sim o número de outra categoria:

$$\text{razão} = \frac{N_i}{N_j}$$

TAXA

É a relação entre quantidades de natureza diferentes. Por exemplo: consumo de combustível em km/l, percentual de juros ao mês, internos em hospitais por habitante.

1.7.2 OPERAÇÕES

ARREDONDAMENTO

- Se o primeiro algarismo a desprezar for superior a 5 (cinco), arredonda-se o algarismo anterior para cima, isto é, soma-se um.
- Se o primeiro algarismo a desprezar for inferior a 5 (cinco), arredonda-se o algarismo anterior para baixo, isto é, fica como está.
- Se o primeiro algarismo a desprezar for igual a 5 (cinco) e ainda houver qualquer resto, arredonda-se o algarismo anterior para cima, isto é, soma-se um.
- Se o primeiro algarismo a desprezar for igual a 5 (cinco) e não houver qualquer resto, arredonda-se o algarismo anterior de modo a que fique par.

ALGARISMOS SIGNIFICATIVOS

Antes de definir em que casa decimal deve ser feito um arredondamento, é conveniente conhecer o conceito de algarismo significativo. Em uma representação numérica, todos os algarismos a partir do primeiro diferente de zero é significativo. Assim, 0,000008 tem apenas um algarismo significativo enquanto 8.000.000 tem sete.

Ao se efetuar operações aritméticas, deve-se procurar conservar um número mínimo de algarismos significativos em cada operação. Uma pergunta equivocada e comum por parte dos alunos é: quantas casas após a vírgula? Para elucidar esta questão, imagine a operação abaixo em que, com o uso de duas casas após a vírgula, calcula-se segundo dois critérios: fazendo-se antes a multiplicação ou a divisão:

$$457\,\frac{23}{7.634} = \frac{10.511}{7.634} = 1{,}376 \cong 1{,}38$$

$$457\,\frac{23}{7.634} = 457 \times 0{,}003 \cong 457 \times 0 = 0$$

O segundo resultado é completamente absurdo, embora tenha obedecido o critério de em cada operação, terem sido usadas duas casas decimais. Se fosse adotado o critério de conservar três algarismos significativos em cada operação, ambas as operações dariam o mesmo resultado como se pode ver a seguir:

$$457\,\frac{23}{7.634} = \frac{10.511}{7.634} = 1,376 \cong 1,38$$

$$457\,\frac{23}{7.634} = 457\ x\ 0,003012 \cong 457\,x0,00301 = 1,3756 \cong 1,38$$

SOMATÓRIO

O somatório, representado pela letra grega maiúscula sigma, é um operador que designa uma soma. Assim, "o somatório de x_i , i variando de 1 a n" é expresso por:

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + x_3 + ... + x_n$$

1.8 EXERCÍCIOS

1. Na tentativa de correlacionar a capacidade de iniciativa de um grupo de pessoas com o hábito de assistir televisão, foi feita a tabela abaixo, em que aparece o número de pessoas em cada categoria:

	telespectadores assíduos	*telespectadores eventuais*
alta capacidade de iniciativa	46	93
baixa capacidade de iniciativa	127	90
totais	173	183

   a) Calcule a proporção de pessoas com alta capacidade de iniciativa entre os telespectadores.
   b) Calcule a proporção de pessoas com pouca capacidade de iniciativa entre os que não são telespectadores.
   c) Calcule a percentagem de pessoas com pouca capacidade de iniciativa entre os telespectadores.
   d) Calcule a percentagem de pessoas com alta capacidade de iniciativa entre os que não são telespectadores.

2. O quadro seguinte representa, em determinada localidade, a estrutura familiar de crianças negras e brancas, no que diz respeito ao número de progenitores com quem convivem:

	crianças negras	*crianças brancas*
um genitor	53	59
dois genitores	130	167
totais	183	226

   a) Calcular a proporção de crianças negras com dois progenitores.
   b) Calcular a proporção de crianças negras com um único genitor.
   c) Calcular a percentagem de crianças brancas com dois progenitores.
   d) Calcular a percentagem de crianças brancas com um único progenitor.

3. Num canil com 125 machos e 80 fêmeas, qual a razão entre machos e fêmeas e qual a proporção de machos?

4. Numa turma de 35 alunos, 20 foram aprovados. Qual a razão entre aprovados e reprovados e o percentual de aprovação?

5. Arredonde os números abaixo com uma, duas e três casas decimais:

   a) 23,5689
   b) 158,42500
   c) 1,05010
   d) 0,04950
   e) 78,4865

6. Efetue as operações abaixo com quatro algarismos significativos:

   a) 2,368 ÷ 86 ÷ 45,698
   b) 236 895 478 x 5 236 874
   c) 0,201 ÷ 0,0598 x 45 268

d) 569,874 x 14,689 x 41 000
e) 124,56897 + 23,56
f) 32,4567 – 1,459823

7. Monte e calcule usando uma expressão de somatório o número de dias do ano a partir de cada mês.

8. A partir da tabela abaixo calcule os valores solicitados:

j	***Pj***	***Qj***	***Rj***
1	5	12	247
2	45	36	968
3	87	125	21
4	235	20	654
5	12	0	589
6	0	47	23
7	36	65	8
8	2	54	6
9	54	1	4

a) $\sum_{j=1}^{9} P_j$

b) $\sum_{j=1}^{7} Q_j$

c) $\sum_{j=3}^{6} R_j$

d) $\sum_{j=1}^{9} Q_j \div \sum_{j=1}^{9} P_j$

e) $\sum_{j=1}^{9} R_j \div Q_j$

2 DISTRIBUIÇÃO DE FREQÜÊNCIAS

2.1 INTRODUÇÃO

As tabelas condensam uma coleção de dados conforme as freqüências ou repetição de seus valores. Utilizam-se vários tipos de tabelas, mas aqui são apresentadas as mais usuais. Assim, uma vez compreendidas neste formato, facilmente, é possível adaptar-se a outras formas de apresentação. A seguir são definidas as diversas grandezas necessárias para a compreensão e construção das tabelas que refletem as distribuições de freqüência.

2.2 PREPARAÇÃO DOS DADOS

2.2.1 DADOS BRUTOS

São os valores dos dados quando chegam da simples coleta, sem que haja qualquer preocupação com sua ordenação. Como exemplo, imagine a lista abaixo como sendo o número de acidentes em uma cidade a cada mês:

27; 67; 54; 11; 88; 95; 34; 56; 88; 44; 92, 23.

2.2.2 ROL

É a lista dos dados, dispostos em uma certa ordem, crescente ou decrescente. No exemplo acima, se os dados forem organizados numa ordem crescente, chega-se a:

11; 23; 27; 34; 44; 54; 56; 67; 88; 88; 92; 95.

Quando o número de dados for muito grande, esta ordenação exigida pelo rol pode ser muito complicada. Nesses casos, o trabalho pode ser facilitado pelo método dos ramos e folhas que consiste em agrupar por algarismos. A melhor forma de descrever isso é através desse exemplo. O método pode ser extrapolado para um número maior de níveis.

1 ⏐ 1	4 ⏐ 4	7 ⏐
2 ⏐ 3 - 7	5 ⏐ 4 -6	8 ⏐ 8 - 8
3 ⏐ 4	6 ⏐ 7	9 ⏐ 2 - 5

2.3 TABULAÇÃO DOS DADOS

As tabelas de freqüência são representações em que os valores se apresentam correspondendo às suas repetições. Deste modo, valores repetidos não aparecem mais de uma vez como no rol apresentado acima (número 88).

Digamos que ao longo de quatro anos, o número de atendimentos realizados pelo Conselho Tutelar de uma cidade foi:

	jan	fev	mar	abr	mai	jun	jul	ago	set	out	nov	dez
2000	6	2	5	6	0	8	7	6	3	4	5	8
2001	9	9	7	6	3	4	6	4	5	4	0	1
2002	3	6	7	9	3	1	4	6	5	3	5	4
2003	7	2	5	8	6	4	2	5	1	6	5	2

2.3.1 DADOS NÃO AGRUPADOS EM CLASSES

É uma tabela em que aparecem os valores que indicam a repetição de uma determinada variável. Este tipo de tabela é normalmente usada para representar variáveis discretas. No exemplo acima, a variável é o número de atendimentos e a freqüência, a contagem do número de meses que corresponde a esse valor.

j	*Número de atendimentos* ***(x_j)***	*contagem*	*Número de meses* ***(f_j)***
1	0	|_	2
2	1	|_ |	3
3	2	|_ |_	4
4	3	|_ |_ |	5
5	4	|_ |_ |_ |	7
6	5	|_ |_ |_ |_	8
7	6	|_ |_ |_ |_ |	9
8	7	|_ |_	4
9	8	|_ |	3
10	9	|_ |	3
			$\boldsymbol{n} = \sum_{j=1}^{10} f_j = 48$

Usam-se índices ***j*** com dados tabulados e ***i*** para dados brutos.

2.3.2 DADOS AGRUPADOS EM CLASSES

Os dados, em lugar de aparecerem individualmente, são organizados em classes. É preferido esse modo de tabulação para variáveis contínuas ou mesmo discretas quando o número de valores possíveis para ela for muito grande. Assim, evitam-se inconvenientes como tabela muito extensa, valores com freqüência nula e dificuldade para vislumbrar o fenômeno de forma global. Embora o exemplo acima não seja típico da necessidade de agrupamento dos dados em classes, segue a tabulação para caracterizar as diferenças:

Ressalte-se que, embora este tipo de tabela na maioria das vezes facilite a visualização dos dados, parte da informação original é perdida, pois já não discrimina como a freqüência individual se distribui dentro da classe. Aqui, o que é disponível é o valor médio da classe $\boldsymbol{x_j}$.

j	*número de atendimentos* ***classes***	*pontos médios* x_j	*freqüência* f_j
1	0 - 1	0,5	5
2	2 - 3	2,5	9
3	4 - 5	4,5	15
4	6 - 7	6,5	13
5	8 - 9	8,5	6
			$n = \sum_{j=1}^{5} f_j = 48$

2.4 ELEMENTOS DA DISTRIBUIÇÃO

2.4.1 FREQÜÊNCIA SIMPLES ABSOLUTA

Numa distribuição de freqüências, sejam os dados agrupados em classes ou em valores individuais, o número de observações em cada classe ou valor é chamada de freqüência simples absoluta, ou mesmo freqüência, cujo símbolo é f_j. . No exemplo acima é a coluna da direita nas duas tabelas, que por serem diferentes, apresentam valores diferentes para f_j.

2.4.2 AMPLITUDE TOTAL

É a diferença entre o maior e o menor valor da variável em estudo. Cuidado: é a diferença da variável e não da freqüência. No exemplo acima o menor valor é ***0*** e o maior é ***9***. Portanto:

$$At = x_M - x_m \qquad At = 9 - 0 = 9$$

2.4.3 NÚMERO DE CLASSES

Classe é cada um dos grupos de valores em que se subdivide a amplitude total. O número de classes de uma distribuição de freqüências, representado pela letra ***k***, deve ser adequado. Se for muito pequeno, os dados ficam comprimidos e pouco se vê da distribuição e se, por outro lado, for muito grande, aparecerão classes com baixa freqüência, tornando a distribuição irregular.

Aconselha-se um número de classes entre 5 e 20 e uma primeira boa aproximação é que este número deve ser da ordem da raiz quadrada do número total ***n*** de observações. Entretanto, dois métodos são indicados na literatura como mais indicados para a determinação do número de classes. A escolha real deve ficar próxima aos valores indicados por qualquer um dos métodos.

REGRA DE STURGES

O número de classes ***k*** é igual à fórmula abaixo em que ***n*** é o número total de observações.

$$k = 1 + 3{,}3 \log_{10} n$$

TABELA DE KELLEY

Ela dá um relação aproximada e adequada entre n e k:

n	5	10	25	50	100	200	500	1000
k	2	4	6	8	10	12	15	15

2.4.4 LIMITES DE CLASSE

Os limites de classe são seus valores extremos. No exemplo acima, os limites da 3ª classe são 4 e 5, quatro sendo o limite inferior e cinco o superior. A classe seguinte, a quarta, tem por limites 6 e 7. Esse caso é simples porque a variável é discreta, pois representa o número de atendimentos que é sempre um inteiro. Mas se fosse uma variável contínua que pudesse valer 5,7, por exemplo, como ficaria? Por esse motivo, é conveniente definir os limites reais de classe e um modo mais conveniente de estabelecer os limites de classe.

LIMITES REAIS DE CLASSE

Qualquer valor intermediário entre os limites de classes distintas deve se arredondado para definir a que classe pertence. Na situação apresentada, 5,7 deve ser arredondado para 6, indo para a quarta classe, portanto. Assim, o limite real de classe é definido como a média do limite superior de uma classe e o inferior da seguinte. Na determinação da amplitude total, para maior precisão, deve-se adotar os limites reais inferior da primeira classe e superior da última classe.

INTERVALOS DE CLASSE

Para evitar essa dificuldade usa-se o conceito de intervalo de classe aberto ou fechado. Os valores dos limites superior e inferior de classes subseqüentes são iguais, mas é explicitado a qual das duas classes o limite pertence. O símbolo |— significa intervalo fechado à esquerda e aberto à direita, ou seja, o limite inferior da classe está incluído nela , enquanto o superior não. Admitindo que X_E e X_D sejam os limites de uma classe, as alternativas de intervalo são apresentadas abaixo:

símbolo do intervalo	*significado do intervalo*	*situação dos limites em relação à classe*
X_E |— X_D	fechado à esquerda aberto à direita	inclui X_E não inclui X_D
X_E —| X_D	aberto à esquerda fechado à direita	não inclui X_E inclui X_D
X_E — X_D	aberto à esquerda aberto à direita	não inclui X_E não inclui X_D
X_E |—| X_D	fechado à esquerda fechado à direita	inclui X_E inclui X_D

2.4.5 AMPLITUDE DE CLASSE

Para se determinar o intervalo de classe Ac , basta dividir a amplitude total At pelo número de classes k. A partir daí, ajustam-se os limites, mesmo que às vezes seja necessário aumentar ou diminuir uma classe. Procura-se montar a tabela com classes de mesma amplitude. A

amplitude da classe pode ser definida como a diferença entre dois limites superiores (ou inferiores) de classes sucessivas.

$$Ac \cong \frac{At}{k}$$

A amplitude de classe deve ser sempre arredondada para maior, nunca para menor, sob pena de algum dado ficar fora do campo da tabela.

2.4.6 PONTO MÉDIO DA CLASSE

Ponto médio da classe x_j é a média dos limites inferior e superior da classe. Uma vez conhecido o ponto médio de uma classe, pode-se encontrar o da classe seguinte, se elas forem do mesmo tamanho. As fórmulas são as seguintes:

$$x_j = \frac{X_{E_j} + X_{D_j}}{2} \qquad x_{j+1} = x_j + Ac$$

2.5 TIPOS DE FREQÜÊNCIAS

A freqüência simples absoluta, já definida anteriormente, é associada a cada classe e sua soma é igual ao número total de observações. A assim chamada freqüência total é, pois:

$$n = \sum_{j=1}^{k} f_j$$

A tabela a seguir mostra as notas obtidas por 500 alunos em um teste de estatística. Nela, já foram definidas dez classes, a freqüência simples absoluta e os pontos médios de cada classe. Serão definidos outros tipos de freqüências, cujos cálculos serão acrescentados ao final desta seção.

ordem **j**	*Notas* **classes**	*freqüências* f_j	*Nota média* x_j
1	0 |— 10	5	5
2	10 |— 20	15	15
3	20 |— 30	20	25
4	30 |— 40	45	35
5	40 |— 50	100	45
6	50 |— 60	130	55
7	60 |— 70	100	65
8	70 |— 80	60	75
9	80 |— 90	15	85
10	90 |—| 100	10	95
$k = 10$		500 $\sum_{j=1}^{k} f_j = n$, $\sum_{j=1}^{10} f_j = 500$	

2.5.1 FREQÜÊNCIA SIMPLES RELATIVA

Freqüência simples relativa fr_j é a proporção de observações de um valor individual ou de uma classe, em relação número total de observações. Ou seja:

$$fr_j = \frac{f_j}{\sum_{j=1}^{k} f_j} = \frac{f_j}{n}$$

Esta freqüência simples relativa pode ser apresentada na forma de percentagem, em lugar de proporção como acabamos de definir, bastando multiplicar esse valor por cem.

$$fr_j\ (\%) = 100\, fr_j$$

2.5.2 FREQÜÊNCIA ABSOLUTA ACUMULADA

Freqüência absoluta acumulada F_j é a soma da freqüência simples absoluta dessa classe com as freqüências simples absolutas das classes anteriores. Pode-se, mais facilmente somar a freqüência simples absoluta dessa classe com a freqüência absoluta acumulada da classe anterior.

2.5.3 FREQÜÊNCIA RELATIVA ACUMULADA

Freqüência relativa acumulada Fr_j é a soma da freqüência simples relativa dessa classe com as freqüências simples relativas das classes anteriores. Também aqui, como no caso da freqüência simples relativa, a freqüência relativa acumulada pode ser expressa em percentagens, bastando multiplicar por cem.

A tabela de freqüências completa do exemplo anterior, com todos os tipos de freqüências, é apresentada abaixo:

j	*classes*	f_j	x_j	F_j	fr_j	fr_j (%)	Fr_j	Fr_j (%)
1	0 |— 10	5	5	5	0,01	1	0,01	1
2	10 |— 20	15	15	20	0,03	3	0,04	4
3	20 |— 30	20	25	40	0,04	4	0,08	8
4	30 |— 40	45	35	85	0,09	9	0,17	17
5	40 |— 50	100	45	185	0,20	20	0,37	37
6	50 |— 60	130	55	315	0,26	26	0,63	63
7	60 |— 70	100	65	415	0,20	20	0,83	83
8	70 |— 80	60	75	475	0,12	12	0,95	95
9	80 |— 90	15	85	490	0,03	3	0,98	98
10	90 |—| 100	10	95	500	0,02	2	1,00	100
10		500						

$k = 10$ $\quad \sum_{j=1}^{k} f_j = n \ , \ \sum_{j=1}^{10} f_j = 500$

2.6 EXERCÍCIOS

1. Na seguinte distribuição de escores discretos, complete a tabela com as freqüências absolutas acumuladas e freqüências simples relativas e acumuladas - estas duas nas formas proporcional e percentual.

intervalo de classe	40-49	50-59	60-69	70-79	80-89	90-99
freqüência (f_j)	5	8	10	10	9	6

2. Transforme a distribuição abaixo de escores (discretos) numa distribuição de freqüências agrupadas com quatro intervalos de classe e depois determine:

   a) O tamanho dos intervalos de classe.
   b) Os limites superior e inferior de cada classe.
   c) O ponto médio de cada intervalo de classe.
   d) A freqüência para cada intervalo de classe.
   e) A percentagem para cada intervalo de classe.

escores	1	2	3	4	5	6	7	8	9	10	11	12
freqüência (f_j)	2	1	1	2	3	4	5	6	5	4	4	3

3. Na seguinte distribuição de escores discretos, complete a tabela com as freqüências absolutas acumuladas e freqüências simples relativas e acumuladas - estas duas nas formas proporcional e percentual.

intervalo de classe	5-9	10-14	15-19	20-24	25-29	30-34	35-39	40-44
freqüência (f_j)	5	6	8	10	9	8	5	5

4. Numa amostragem salarial de 100 estudantes universitários que trabalham, obteve-se a tabela abaixo:

número de salários mínimos	*número de estudantes*
0 ⊢ 2	40
2 ⊢ 4	30
4 ⊢ 6	10
6 ⊢ 8	15
8 ⊢ 10	5

   a) Faça uma tabela completa, com todos os tipos de freqüências - simples e acumulada, absoluta e relativa, proporcional e percentual.
   b) Quantos alunos ganham menos de dois salários mínimos?
   c) Quantos estudantes ganham menos que 6 salários mínimos?
   d) Qual a percentagem de alunos que recebem seis ou mais salários mínimos, mas menos de 8 salários mínimos?
   e) Qual a percentagem de alunos com salários menores que 8 salários mínimos?
   f) Qual a proporção de alunos com salários iguais ou superiores a 6 salários mínimos?
   g) Quantos alunos recebem 4 ou mais salários mínimos?
   h) Qual a proporção de alunos que ganham menos de 4 salários mínimos?
   i) Qual a percentagem de alunos que ganham dois ou mais salários mínimos?
   j) Qual a proporção de alunos com salários entre 4 e 6 salários mínimos, embora não atinjam este último valor?

5. Mostre os erros na construção da tabela abaixo:

ordem (j)	*classes*	F_j
1	0 |—| 2	80
2	4 —| 6	0
3	6 |— 8	10
4	8 |— 10	10
		$\sum = 100$

6. Três psicólogas de uma cidade do interior fizeram ao longo de um ano 1.800 atendimentos, conforme a tabela abaixo. Determine a distribuição percentual dos atendimentos.

psicóloga	A	B	C
atendimentos	720	480	600

7. Os dados seguintes representam 20 observações relativas ao índice pluviométrico, expresso em milímetros de chuva, em determinado município do Estado:

144	160	154	142	141	152	151	145	146	150
159	157	141	142	143	160	146	150	141	158

   a) Determine o número de classes pela regra de Sturges e tabela de Kelley.
   b) Construa a tabela de freqüências simples absolutas.
   c) Construa a tabela de freqüências absolutas acumuladas.
   d) Construa a tabela de freqüências simples relativas.
   e) Construa a tabela de freqüências acumuladas relativas.

8. A tabela a seguir representa uma distribuição das espessuras de uma amostra de 100 folhas.

   a) Determine o número de classes pela regra de Sturges e tabela de Kelley. Use a média.
   b) Calcule a amplitude total.
   c) Calcule a amplitude de classe.
   d) Construa uma tabela apresentando as ordens de classe, os intervalos de classe, os pontos médios de classe, uma coluna de contagem e as diversas freqüências - simples e acumulada, absoluta e relativa, proporcional e percentual.

2,01	2,08	1,96	3,04	2,01	3,18	1,94	2,19	2,24	2,18
2,59	1,96	2,29	3,18	2,09	1,96	2,06	2,18	2,05	2,04
2,43	1,56	1,94	3,15	2,35	2,08	2,56	2,17	1,96	1,59
2,22	2,34	2,24	1,95	2,01	3,12	3,03	3,12	2,04	1,66
1,87	2,49	3,12	2,24	1,76	3,20	2,38	1,58	1,89	1,98
1,89	1,71	2,42	1,62	1,97	2,18	1,69	3,14	2,18	3,06
2,40	1,96	3,01	2,19	2,25	1,45	1,93	2,06	1,83	1,84
1,91	2,11	1,78	2,36	2,33	3,17	2.03	1,87	3,11	2,17
1,72	1,62	1,99	1,64	1,54	2,26	1,86	2,09	1,74	1,92
2,36	1,82	2,02	2,25	1,75	3,15	3,18	1,99	1,76	2,51

9. A tabela seguinte apresenta as alturas em cm de 40 alunos de uma turma de estatística.

162	163	148	166	169	154	170	166
164	165	159	175	155	163	171	172
170	157	176	157	157	165	158	158
160	158	163	165	164	178	150	168
166	169	152	170	172	165	162	164

a) Calcule a amplitude total.
b) Determine o número de classes pela fórmula de Sturges.
c) Qual a amplitude do intervalo de classes?
d) Construa uma tabela de freqüências completa das alturas dos alunos.

10. Complete o quadro de freqüências abaixo.

*Ordem (**j**)*	*classes*	f_j	***Fj***
1	0 —| 2	?	3
2	2 —| 4	?	?
3	4 —| 6	8	?
4	6 —| 8	10	26
5	8 —| 10	?	28

11. Considere e complemente a tabela abaixo e identifique os seguintes elementos:

a) Freqüência simples absoluta da quinta classe.
b) Freqüência total.
c) Limite inferior da sexta classe.
d) Limite superior da quarta classe.
e) Amplitude do intervalo de classe.
f) Amplitude total.
g) Ponto médio da terceira classe.
h) Número total de classes.
i) Freqüência absoluta acumulada além da sexta classe.
j) Percentagem de valores iguais ou maiores que 3,20.

*ordem da classe (**j**)*	*classe*	f_j
1	2,75 |— 2,80	2
2	2,80 |— 2,85	3
3	2,85 |— 2,90	10
4	2,90 |— 2,95	11
5	2,95 |— 3,00	24
6	3,00 |— 3,05	14
7	3,05 |— 3,10	9
8	3,10 |— 3,15	8
9	3,15 |— 3,20	6
10	3,20 |— 3,25	3
		$\sum_{j=1}^{10} f_j = 90$

12. Considere a seguinte distribuição de freqüências, correspondente aos diferentes preços de um livro de estatística em vinte livrarias pesquisadas.

preços em reais	*número de livrarias*
50	2
51	5
52	6
53	6
54	1
	$\sum = 20$

a) Construa uma tabela de freqüências completa.
b) Quantas livrarias apresentaram um preço de R$ 52,00?
c) Quantas livrarias apresentaram preços iguais ou menores que R$ 52,00?
d) Que percentagem de livrarias apresentaram preços não maiores que R$ 53,00 ?
e) Qual a proporção de livrarias com preços maiores que R$ 51,00 e menores que R$ 54,00?

3 GRÁFICOS

3.1 INTRODUÇÃO

As tabelas, embora resumam adequadamente as informações importantes de uma distribuição estatística, nem sempre fornecem uma boa visão de conjunto. Isso é obtido com maior eficácia através de gráficos, que podem ser apresentados de inúmeras formas, dificultando, de certo modo, um tratamento sistemático da matéria. Contudo, a compreensão de alguns poucos tipos permite, mediante a extensão de conceitos, o domínio da generalidade das apresentações alternativas. Serão apresentados aqui, então, quatro tipos de gráficos: histograma, polígono de freqüência, ogiva de Galton e diagrama de setores.

3.2 EXEMPLOS DE DISTRIBUIÇÕES

Para efeitos de exemplificação, serão utilizadas algumas distribuições que reunam os diferentes tipos de dados. Como foi visto, os dados podem ser nominais, ordinais ou numéricos - contínuos ou discretos. Além disso, quando numéricos, podem ser agrupados ou não em classes.

3.2.1 DADOS NOMINAIS

Como exemplo de dados nominais, será usada a tabela abaixo que resume os conceitos referentes ao comportamento dos alunos de uma escola

Ordem da classe ***j***	*Conceitos quanto ao comportamento*	*Quantidade de alunos* ***fj***	*Proporção de alunos* ***frj***
1	ÓTIMO	118	0,225
2	BOM	134	0,256
3	REGULAR	152	0,290
4	MAU	86	0,164
5	PÉSSIMO	34	0,065
		524	1,000

3.2.2 DADOS NUMÉRICOS DISCRETOS

Será usado o exemplo cuja tabela já foi construída, referente a ao número de atendimentos realizados pelo conselho tutelar de uma cidade ao longo de quatro anos.

NÃO AGRUPADOS EM CLASSES

Primeiramente, a tabela é apresentada sem que os dados estejam agrupados.

j	*Número de atendimentos* (x_j)	*Número de meses* (f_j)	(F_j)	(*frj %*)	(*Frj %*)
1	0	2	2	4,17	4,17
2	1	3	5	6,25	10,42
3	2	4	9	8,33	18,75
4	3	5	14	10,42	29,17
5	4	7	21	14,58	43,75
6	5	8	29	16,67	60,42
7	6	9	38	18,75	79,17
8	7	4	42	8,33	87,50
9	8	3	45	6,25	93,75
10	9	3	48	6,25	100,00
	$\sum$	48		100,00	

AGRUPADOS EM CLASSES

Agora os mesmos dados são agrupados em classes:

j	***classes***	*Pontos médios* x_j	*Freqüência simples absoluta* f_j	*Freqüência simples relativa* fr_j	*Freqüência acumulada absoluta* F_j
1	0 - 1	0,5	5	10,42	5
2	2 - 3	2,5	9	18,75	14
3	4 - 5	4,5	15	31,25	29
4	6 - 7	6,5	13	27,08	42
5	8 - 9	8,5	6	12,50	48
		$\sum$	48	100,00	

3.2.3 DADOS NUMÉRICOS CONTÍNUOS

A tabela completa de freqüências a seguir, já discutida anteriormente, refere-se às notas obtidas por 500 alunos em um teste de estatística..

j	***classes***	f_j	x_j	F_j	fr_j	fr_j *(%)*	Fr_j	Fr_j *(%)*
1	0 |— 10	5	5	5	0,01	1	0,01	1
2	10 |— 20	15	15	20	0,03	3	0,04	4
3	20 |— 30	20	25	40	0,04	4	0,08	8
4	30 |— 40	45	35	85	0,09	9	0,17	17
5	40 |— 50	100	45	185	0,20	20	0,37	37
6	50 |— 60	130	55	315	0,26	26	0,63	63
7	60 |— 70	100	65	415	0,20	20	0,83	83
8	70 |— 80	60	75	475	0,12	12	0,95	95
9	80 |— 90	15	85	490	0,03	3	0,98	98
10	90 |—| 100	10	95	500	0,02	2	1,00	100
10		500						

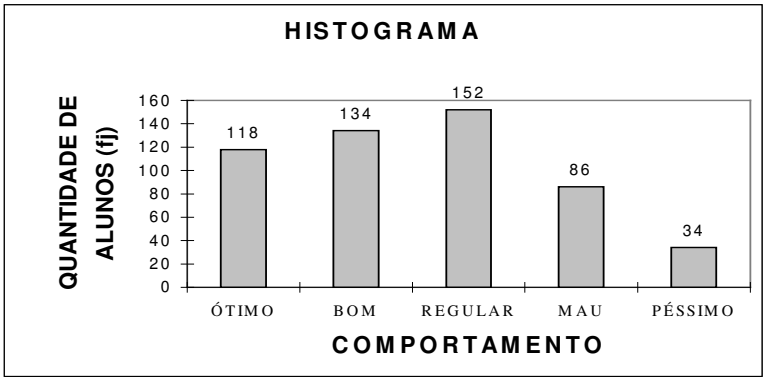
$k = 10$ $\quad \sum_{j=1}^{k} f_j = n \;,\; \sum_{j=1}^{10} f_j = 500$

3.3 HISTOGRAMAS

O histograma, ou diagrama de barras, é utilizado para representar a freqüência simples absoluta (***f***) nas ordenadas e a variável (***x***) nas abcissas. Deve-se ter o cuidado ao definir as escalas dos eixos de forma a suportarem os valores máximos de ***x*** e ***f***. Desenham-se barras de comprimento proporcional aos valores das freqüências.

3.3.1 DADOS NOMINAIS

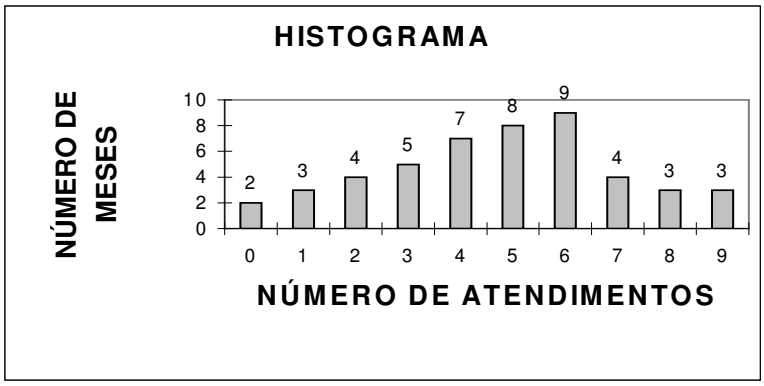
Deve-se ter o cuidado de, nesse caso, evitar que as barras verticais se toquem, para que não seja passada a idéia de continuidade, inadequada quando os dados são nominais, em grande parte das vezes.



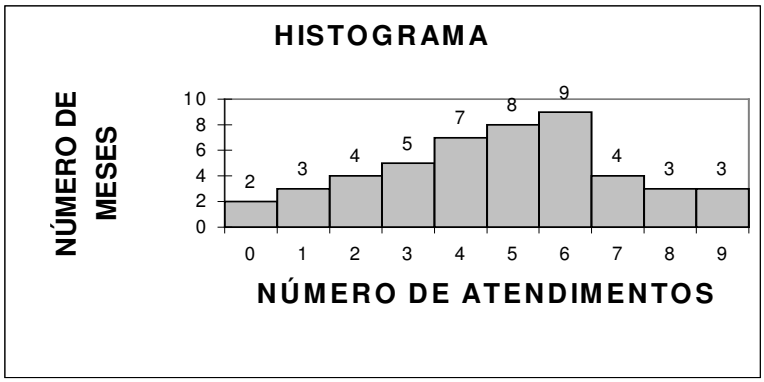
3.3.2 DADOS NUMÉRICOS DISCRETOS

NÃO AGRUPADOS EM CLASSES

Pode-se colocar nos eixos das abcissas e ordenadas o significado da variável e da freqüência sem, necessariamente, escrever ***x*** e ***f***, desde que as medidas lhe correspondam.

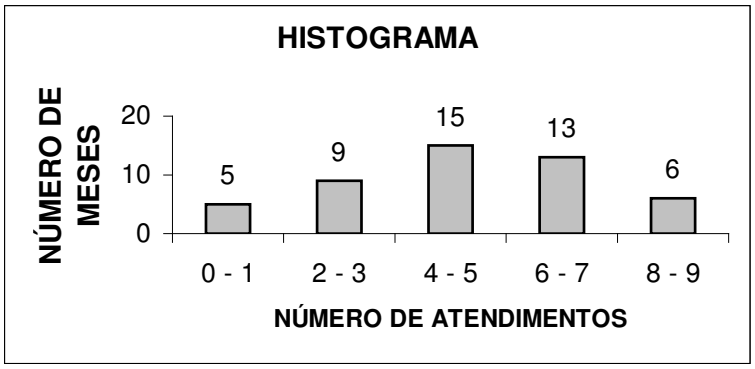


Nesse caso, como os dados são numéricos e discretos, as barras tanto podem estar afastadas como ligadas, como pode ser visto na apresentação alternativa que segue:

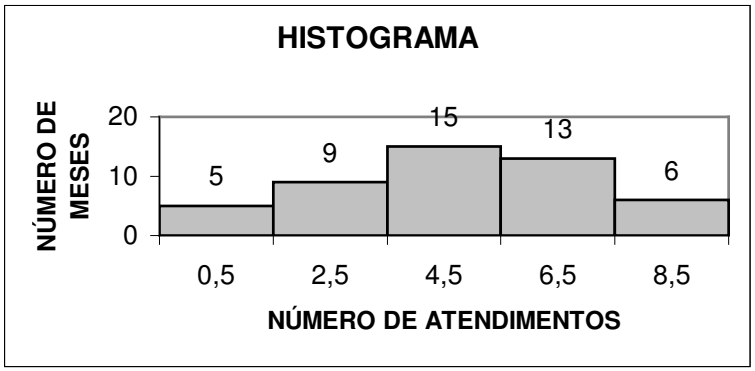


AGRUPADOS EM CLASSES

Aqui, como os dados são discretos, toleram-se que as barras estejam afastadas, embora seja mais recomendado que elas se toquem para dar o sentido de continuidade.

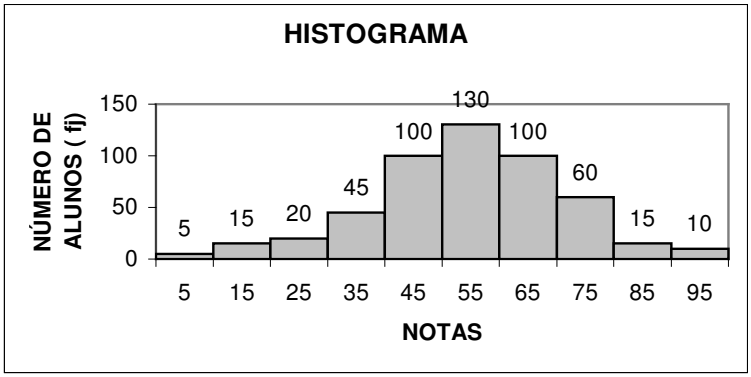


Pode-se usar no eixo das abcissas tanto a amplitude da classe como acima ou o valor médio da classe como é apresentado agora:



3.3.3 DADOS NUMÉRICOS CONTÍNUOS

Para dados contínuos recomenda-se que as barras se toquem para passar a idéia de continuidade.



3.4 POLÍGONO DE FREQÜÊNCIAS

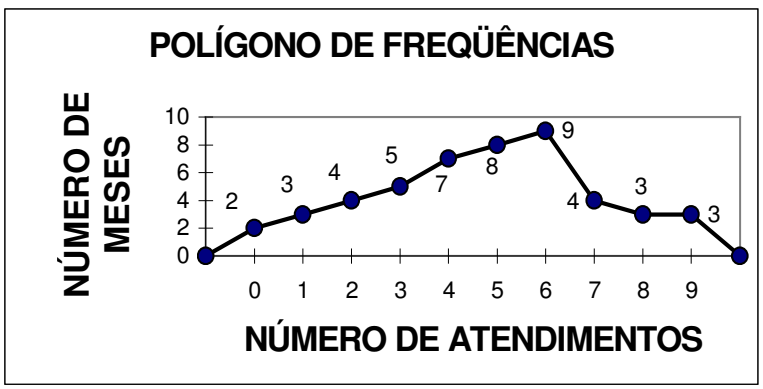
O polígono de freqüências é um gráfico alternativo ao histograma em que, nas ordenadas são marcados os valores da freqüência simples absoluta ***f*** para os diversos valores da variável ***x***. Estes pontos são interligados por uma linha contínua que vem a se constituir o polígono de freqüências. Para evitar que ele fique flutuando no espaço, criam-se duas classes imaginárias, uma antes e outra após a região da variável, ambas com freqüência nula, de modo que o polígono nasça e morra no eixo das abcissas. Esta providência é recomendável dependendo do contexto dos dados representados.

3.4.1 DADOS NOMINAIS

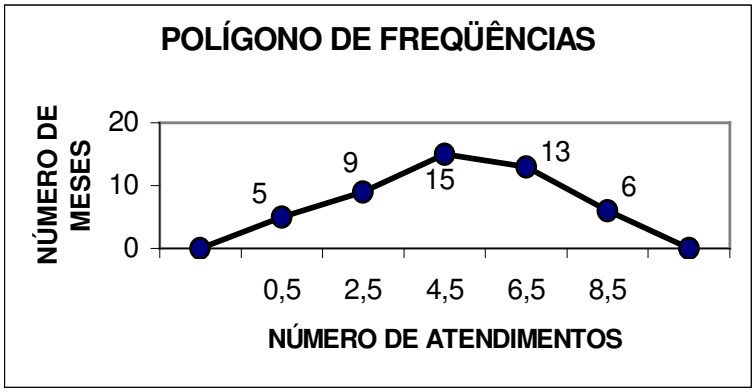
Nem sempre são desenhados polígonos para dados nominais, porque a linha que liga os diversos pontos via de regra não faz sentido.

3.4.2 DADOS NUMÉRICOS DISCRETOS

NÃO AGRUPADOS EM CLASSES



AGRUPADOS EM CLASSES

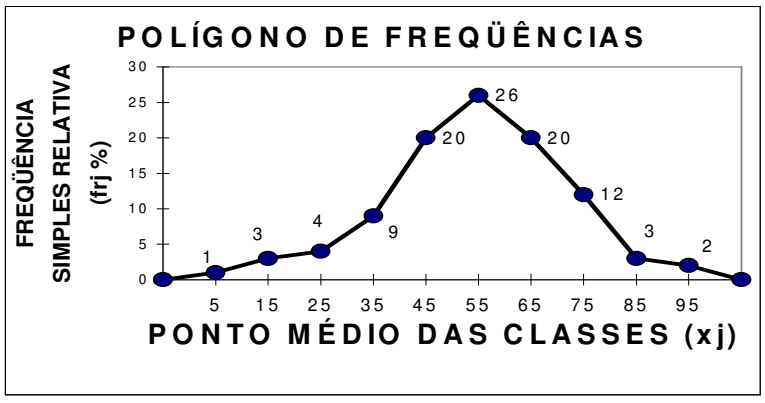


3.4.3 DADOS NUMÉRICOS CONTÍNUOS

O polígono de freqüências para os dados numéricos contínuos toma o ponto médio das classes para caracterizar onde deve ser plotado o ponto que gera o polígono.

É conveniente alertar que tanto o histograma quanto o polígono de freqüências podem ser usados para grafar as freqüências simples relativas proporcional (***fr***) e percentual (***fr%***), da mesma forma que para freqüência simples absoluta (***f***). Em suma, os gráficos são úteis para as freqüências simples.

Abaixo, por exemplo, está desenhada a ***fr %***. Nas abcissas, o título alerta de que se trata da nota média da classe.



3.5 OGIVA DE GALTON

É o tipo de gráfico utilizado para freqüências acumuladas de uma distribuição, sejam elas absolutas ou relativas. O processo de construção é quase idêntico ao do polígono, usado para as freqüências simples. O cuidado a ser tomado aqui é que, enquanto antes a ordenada era plotada no ponto médio da classe, na ogiva o valor acumulado só é atingido ao final da classe.

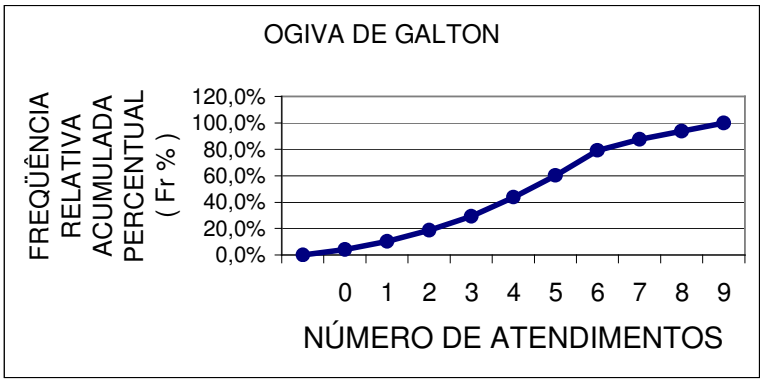
3.5.1 DADOS NOMINAIS

Nem sempre são desenhados polígonos para dados nominais, porque a linha que liga os diversos pontos pode não fazer sentido.

3.5.2 DADOS NUMÉRICOS DISCRETOS

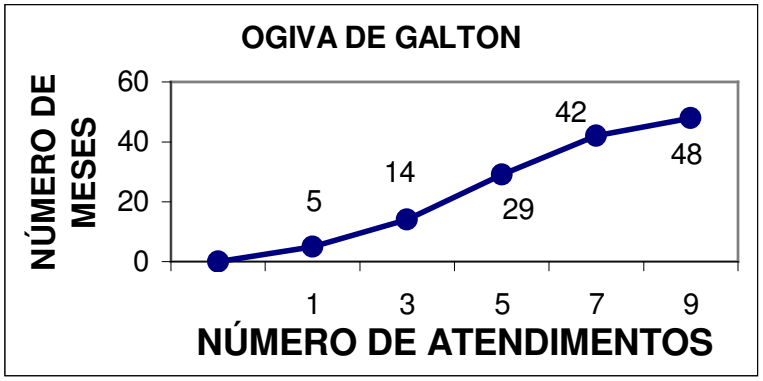
NÃO AGRUPADOS EM CLASSES

Quando os dados são discretos e não estão agrupados, o valor acumulado muda aos saltos, a cada novo valor da variável.



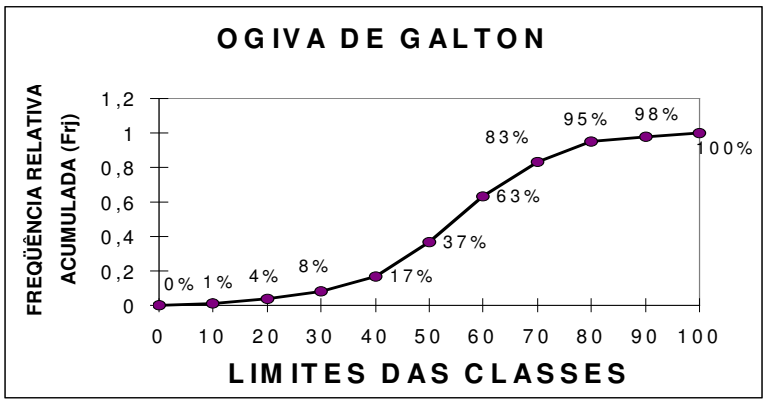
AGRUPADOS EM CLASSES

Quando os dados são discretos, o final da classe é o último valor possível para a variável.



3.5.3 DADOS NUMÉRICOS CONTÍNUOS

Mais uma vez é conveniente alertar que os valores são acumulados ao longo da classe e, por esse motivo, as freqüências acumuladas devem ser plotadas ao final da classe e não em seu ponto médio como no caso das freqüências simples.



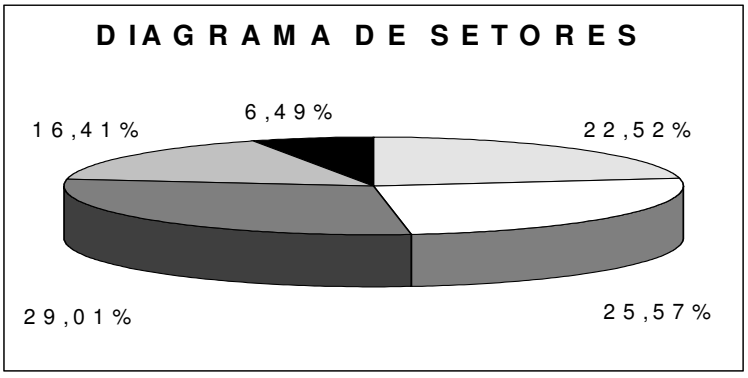
3.6 DIAGRAMA DE SETORES

O diagrama de setores ou setograma ou ainda, como é conhecido vulgarmente, diagrama de pizzas ou tortas, divide um círculo na proporção das partes que compõem o todo. Como a freqüência simples relativa ***fr*** fornece a percentagem ou proporção de cada parte, ela informa o número de graus de cada fatia, bastando para isso multiplicar aquele valor por 3,6 ou 360, num e noutro caso. Isso se deve ao fato que toda a pizza corresponder a 100% e a 360º.

É importante alertar que ***fr*** , a freqüência simples relativa, serve para dividir a torta, mas, normalmente, o que se coloca escrito dentro da fatia é ***f***, a freqüência simples absoluta. As freqüências acumuladas não são usadas nos setogramas.

3.6.1 DADOS NOMINAIS

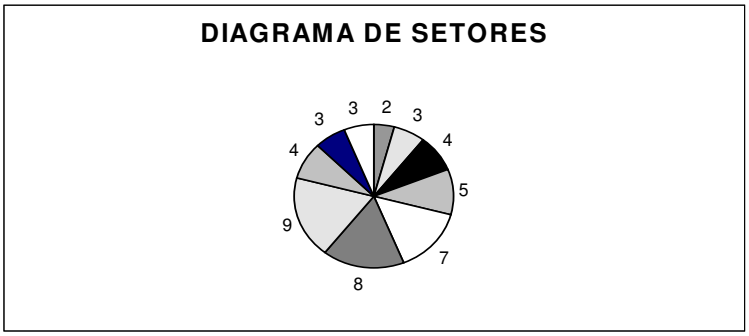
Aqui a torta foi apresentada em perspectiva, mas as percentagens foram obtidas a partir da tabela, multiplicando fr_j por ***100*** e os ângulos foram conseguidos multiplicando fr_j por ***360***.



3.6.2 DADOS NUMÉRICOS DISCRETOS

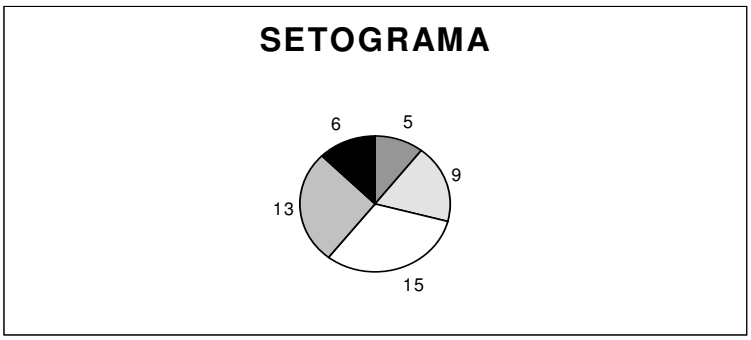
NÃO AGRUPADOS EM CLASSES

Os ângulos foram obtidos multiplicando-se a coluna ***fr %***, freqüência simples relativa, por ***3,6***. Entretanto os valores marcados no gráfico são os de ***f***, freqüência simples absoluta.

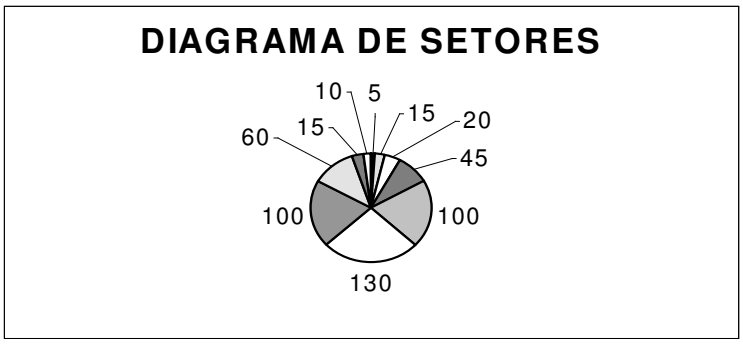


AGRUPADOS EM CLASSES

Quando o número de fatias é muito alto, o setograma não é eficiente como pode ser visto no gráfico anterior, mas quando os dados são agrupados, o número de partes torna-se razoável e fornece uma boa visão contextual.



3.6.3 DADOS NUMÉRICOS CONTÍNUOS



3.7 EXERCÍCIOS

1. Na seguinte distribuição de escores discretos, desenhe:

   k) Histograma.
   l) Polígono de freqüências.
   m) Ogiva de Galton.
   n) Diagrama de setores

intervalo de classe	40-49	50-59	60-69	70-79	80-89	90-99
freqüência (f_i)	5	8	10	10	9	6

2. Transforme a distribuição abaixo de escores (discretos) numa distribuição de freqüências agrupadas com quatro intervalos de classe e depois esboce:

   a) Histograma.
   b) Polígono de freqüências.
   c) Ogiva de Galton.
   d) Diagrama de setores

escores	1	2	3	4	5	6	7	8	9	10	11	12
freqüência (f_i)	2	1	1	2	3	4	5	6	5	4	4	3

3. Na seguinte distribuição de escores discretos, desenhe:

   a) Histograma.
   b) Polígono de freqüências.
   c) Ogiva de Galton.
   d) Diagrama de setores

intervalo de classe	5-9	10-14	15-19	20-24	25-29	30-34	35-39	40-44
freqüência (f_i)	5	6	8	10	9	8	5	5

4. Numa amostragem salarial de 100 estudantes universitários que trabalham, obteve-se a tabela abaixo:

número de salários mínimos	*número de estudantes*
0 ⊢ 2	40
2 ⊢ 4	30
4 ⊢ 6	10
6 ⊢ 8	15
8 ⊢ 10	5

   a) Histograma.
   b) Polígono de freqüências.
   c) Ogiva de Galton.
   d) Diagrama de setores

5. Três psicólogas de uma cidade do interior fizeram ao longo de um ano 1.800 atendimentos, conforme a tabela abaixo. Construa o histograma e o diagrama de setores.

psicóloga	A	B	C
atendimentos	720	480	600

6. Para o quadro de freqüências abaixo, desenhe:

    a) Histograma.
    b) Polígono de freqüências.
    c) Ogiva de Galton.
    d) Diagrama de setores

*Ordem (**j**)*	*classes*	f_j	***Fj***
1	0 —| 2	3	3
2	2 —| 4	5	8
3	4 —| 6	8	16
4	6 —| 8	10	26
5	8 —| 10	2	28

7. Considere e complemente a tabela abaixo e identifique esboce os seguintes gráficos:

    a) Histograma.
    b) Polígono de freqüências.
    c) Ogiva de Galton.
    d) Diagrama de setores

*ordem da classe (**j**)*	*classe*	f_j
1	2,75 |— 2,80	2
2	2,80 |— 2,85	3
3	2,85 |— 2,90	10
4	2,90 |— 2,95	11
5	2,95 |— 3,00	24
6	3,00 |— 3,05	14
7	3,05 |— 3,10	9
8	3,10 |— 3,15	8
9	3,15 |— 3,20	6
10	3,20 |— 3,25	3
		$\sum_{j=1}^{10} f_j = 90$

8. Considere a seguinte distribuição de freqüências, correspondente aos diferentes preços de um livro de estatística em vinte livrarias pesquisadas.

preços em reais	*número de livrarias*
50	2
51	5
52	6
53	6
54	1
	$\sum = 20$

    f) Construa o histograma.
    g) Faça o polígono de freqüências
    h) Desenhe a ogiva de Galton
    i) Esboce o setograma.

4 MEDIDAS DE POSIÇÃO

É difícil trabalhar com uma distribuição completa de freqüências, por isso, utilizam-se outras medidas, umas que fornecem a posição e outras que dão o grau de dispersão. Dentre as medidas de posição, as principais são as medidas de tendência central ou promédias. As promédias mais importantes são a moda, a mediana e a média aritmética. Outras medidas de tendência central menos utilizadas são as médias geométrica, harmônica e quadrática. Além das promédias, outras medidas de posição são as chamadas separatrizes que, como o próprio nome indica, separa a distribuição em partes com um determinado número de elementos. A principal separatriz é a mediana, que também é uma medidas de tendência central. As outras separatrizes que serão estudadas são os quartis, decis e centis.

4.1 MEDIDAS DE TENDÊNCIA CENTRAL

Para efeitos de exemplificação das definições das principais medidas de tendência central, serão usadas duas distribuições, correspondentes aos dados brutos abaixo. Tais dados hipotéticos referem-se às notas obtidas, no "provão" do MEC, pelos alunos formandos em Psicologia de duas universidades distintas.

Universidade **A**: 4, 5, 5, 6, 8, 2, 8, 3, 2, 7, 4, 9, 3, 7, 8, 2, 6, 4, 9, 7, 5, 5, 7, 6.

Universidade **B**: 5, 7, 6, 9, 4, 7, 3, 6, 4, 6, 5, 6, 8, 7, 5, 8, 4, 6, 8.

4.1.1 MÉDIA ARITMÉTICA

MÉDIA ARITMÉTICA SIMPLES

A média aritmética de um conjunto de números é igual ao quociente entre a soma dos valores do conjunto e o número total de valores. Admitindo que existam n números e seus valores sejam genericamente representados por x_i, o valor médio $\overline{x}$ é:

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

Aplicando-se esta fórmula nas distribuições acima, têm-se:

$$\overline{x}_A = \frac{\sum_{i=1}^{24} x_i}{24} = \frac{132}{24} = 5{,}5 \qquad \text{e} \qquad \overline{x}_B = \frac{\sum_{i=1}^{19} x_i}{19} = \frac{114}{19} = 6$$

Note-se que não foi necessária uma prévia organização dos dados. Esta é uma das vantagens da média aritmética.

MÉDIA ARITMÉTICA PONDERADA

Se os dados brutos fossem previamente trabalhados de modo a produzir o rol de uma e outra distribuição, ficaria:

Universidade **A**: 2, 2, 2, 3, 3, 4, 4, 4, 5, 5, 5, 5, 6, 6, 6, 7, 7, 7, 7, 8, 8, 8, 9, 9.

Universidade **B**: 3, 4, 4, 4, 5, 5, 5, 6, 6, 6, 6, 6, 7, 7, 7, 8, 8, 8, 9.

Então poderia ser usada a média aritmética ponderada em substituição à média aritmética simples apresentada. Na média ponderada, conta-se o número se vezes f_j que aparece cada valor da variável x_j, multiplica-se um pelo outro e soma-se os valores obtidos para cada um dos k valores de x_j. O quociente dessa soma pelo número total de valores dá a média ponderada. É mais fácil calcular desse modo quando o número de observações for grande. Genericamente, então, a média ponderada é dada por:

$$\bar{x} = \frac{\sum_{j=1}^{k} x_j f_j}{\sum_{j=1}^{k} f_j} = \frac{\sum x\,f}{\sum f} = \frac{\sum x\,f}{n}$$

Nas distribuições em questão:

$$\bar{x}_A = \frac{\sum_{j=1}^{k} x_j\, f_j}{\sum_{j=1}^{k} f_j} = \frac{\sum_{j=1}^{8} x_j\, f_j}{\sum_{j=1}^{8} f_j} = \frac{\sum x\,f}{\sum f} = \frac{\sum x\,f}{n}$$

$$\bar{x}_A = \frac{\sum x\,f}{n} = \frac{2 \text{ x } 3 + 3 \text{ x } 2 + 4 \text{ x } 3 + 5 \text{ x } 4 + 6 \text{ x } 3 + 7 \text{ x } 4 + 8 \text{ x } 3 + 9 \text{ x } 2}{3 + 2 + 3 + 4 + 3 + 4 + 3 + 2}$$

$$\bar{x}_A = \frac{\sum x\,f}{n} = \frac{132}{24} = 5{,}5$$

$$\bar{x}_B = \frac{\sum_{j=1}^{k} x_j\, f_j}{\sum_{j=1}^{k} f_j} = \frac{\sum_{j=1}^{7} x_j\, f_j}{\sum_{j=1}^{7} f_j} = \frac{\sum x\,f}{\sum f} = \frac{\sum x\,f}{n}$$

$$\bar{x}_B = \frac{\sum x\,f}{n} = \frac{3 \text{ x } 1 + 4 \text{ x } 3 + 5 \text{ x } 3 + 6 \text{ x } 5 + 7 \text{ x } 3 + 8 \text{ x } 3 + 9 \text{ x } 1}{1 + 3 + 3 + 5 + 3 + 3 + 1}$$

$$\bar{x}_B = \frac{\sum x\,f}{n} = \frac{114}{19} = 6$$

Que são os valores encontrados anteriormente.

4.1.2 MODA

A moda (***Mo***)é o valor mais freqüente da distribuição. É uma boa primeira visão de promédia de um conjunto de valores, porque não exige cálculos. Nem sempre, entretanto, é possível definir um valor que caracterize a distribuição. Quando isso ocorre, quando não se encontra a moda, diz-se que o conjunto é amodal. Outras vezes, existe mais de uma moda, sendo então chamada a distribuição de multimodal. Quando a variável é numérica, fica difícil encontrar a moda a partir de dados brutos, mas basta fazer o rol para que a moda salte aos olhos. Quando a distribuição é nominal a moda torna-se importante por ser a única medida possível de tendência central. Nos exemplos apresentados, a partir do rol, vê-se que a moda das notas da universidade **B** é 6 ($Mo_B = 6$), com 6 ocorrências. Já no caso da universidade **A** existem duas modas, 5 e 7 ($Mo_A = 5$ e $Mo_A = 7$), pois ambas as notas apresentam 4 ocorrências. Portanto, nesse caso o conjunto é bimodal.

4.1.3 MEDIANA

Mediana (***Me***) é o valor de uma série ordenada que deixa um número igual de itens maiores (metade) e menores (a outra metade) do que ele. Por ser uma separatriz, a mediana não pode ser encontrada a partir dos dados brutos, mas sim a partir do rol, onde os elementos estão ordenados. O número que indica a ordem ou posição em que se encontra a mediana é chamado de elemento mediano E_{Me}. Calcula-se E_{Me} pela equação abaixo em que n é o número de elementos da distribuição.

$$E_{Me} = \frac{n+1}{2}$$

Isto estabelece duas situações possíveis:

NÚMERO DE OBSERVAÇÕES É ÍMPAR

O resultado de E_{Me} é inteiro e define exatamente o ponto onde se encontra a mediana. No caso da universidade **B**, em que n é impar:

$$E_{Me} = \frac{n+1}{2} = \frac{19+1}{2} = 10$$

Isso define a mediana como o décimo elemento do rol. Portanto:

$$Me = 6$$

NÚMERO DE OBSERVAÇÕES É PAR

O resultado de E_{Me} é um número de ordem não inteiro que define um elemento inexistente. A mediana será a média aritmética dos elementos vizinhos desse elemento imaginário. No caso da universidade **A**, em que n é par, chega-se a:

$$E_{Me} = \frac{n+1}{2} = \frac{24+1}{2} = 12{,}5$$

Os elementos de ordem 12 e 13 são 5 e 6. Portanto:

$$\boldsymbol{Me} = \frac{5+6}{2} = 5{,}5$$

4.1.4 COMPARAÇÃO ENTRE MÉDIA, MEDIANA E MODA

Aqui será feita uma análise comparativa sobre a aplicabilidade das três promédias, segundo três critérios distintos. Antes, porém, os dados obtidos nos exemplos acima serão resumidos.

RESUMO DOS RESULTADOS DOS EXEMPLOS

Pode-se construir uma tabela com os resultados obtidos nos dois exemplos dados:

	Média aritmética ($\bar{x}$)	*Moda* ***(Mo)***	*Mediana* ***(Me)***
Universidade **A**	5,5	5 e 7	5,5
Universidade **B**	6	6	6

Pode-se ver que, enquanto no caso **B** todas as médias deram o mesmo resultado, no caso **A**, a moda está a indicar uma distribuição com dois máximos. Nesse caso, quem melhor informa é a moda, pois com as outras medidas não se pode imaginar a "corcova".

TIPO DE DADOS

Moda	Nominal, ordinais e numéricos contínuos e discretos
Mediana	Ordinais e numéricos contínuos e discretos
Média aritmética	Numéricos contínuos e discretos

TIPO DE DISTRIBUIÇÃO

Moda	Mais próprias para multimodais
Mediana	Mais próprias para assimétricas
Média aritmética	Mais próprias para unimodais simétricas

TIPO DE OBJETIVO

Moda	Medida rápida e simples, mas grosseira
Mediana	Medida confiável, não influenciada por extremos
Média aritmética	Medida exata, útil em estatísticas mais avançadas

4.2 OUTRAS PROMÉDIAS

A título ilustrativo, serão apresentadas outra promédias menos utilizadas. Para exemplificar sua utilização, Imagine as notas de sete alunos de Estatística como sendo: 8, 6, 7, 8, 9, 7, 8. Nesse caso, a média aritmética, a mediana e a moda são respectivamente ***7,571***, ***8*** e ***8***.

4.2.1 MÉDIA GEOMÉTRICA

SIMPLES

$$\bar{x}_g = \sqrt[n]{\prod_{i=1}^{n} x_i}$$ Para os dados do exemplo:

$$\bar{x}_g = \sqrt[7]{\prod_{i=1}^{7} x_i} = \sqrt[7]{8 \text{ x } 6 \text{ x } 7 \text{ x } 8 \text{ x } 9 \text{ x } 7 \text{ x } 8} = \sqrt[7]{1.354.752} = \mathbf{7,516}$$

PONDERADA

$$\bar{x}_g = \sqrt[n]{\prod_{j=1}^{k} x_j^{f_j}}$$ Para os dados do exemplo:

$$\bar{x}_g = \sqrt[7]{\prod_{j=1}^{4} x_j^{f_j}} = \sqrt[7]{6 \text{ x } 7^2 \text{ x } 8^3 \text{ x } 9} = \sqrt[7]{1.354.752} = \mathbf{7,516}$$

4.2.2 MÉDIA HARMÔNICA

SIMPLES

$$\bar{x}_h = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$ Para os dados do exemplo:

$$\bar{x}_h = \frac{7}{\sum_{i=1}^{7} \frac{1}{x_i}} = \frac{7}{\frac{1}{6}+\frac{1}{7}+\frac{1}{7}+\frac{1}{8}+\frac{1}{8}+\frac{1}{8}+\frac{1}{9}} = \mathbf{7,459}$$

PONDERADA

$$\bar{x}_h = \frac{n}{\sum_{j=1}^{k} \frac{f_j}{x_j}}$$ Para os dados do exemplo:

$$\bar{x}_h = \frac{7}{\sum_{j=1}^{7} \frac{f_j}{x_j}} = \frac{7}{\frac{1}{6}+\frac{2}{7}+\frac{3}{8}+\frac{1}{9}} = \mathbf{7,459}$$

4.2.3 MÉDIA QUADRÁTICA

SIMPLES

$$\bar{x}_q = \sqrt{\overline{x^2}} = \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n}}$$ Para os dados do exemplo:

$$\bar{x}_q = \sqrt{\frac{\sum_{i=1}^{7} x_i^2}{7}} = \sqrt{\frac{6^2+7^2+7^2+8^2+8^2+8^2+9^2}{7}} = 7,625$$

PONDERADA

$$\bar{x}_q = \sqrt{\overline{x^2}} = \sqrt{\frac{\sum_{j=1}^{k} x_j^2 f_j}{n}}$$ Para os dados do exemplo:

$$\bar{x}_q = \sqrt{\frac{\sum_{j=1}^{k} x_j^2 f_j}{n}} = \sqrt{\frac{6^2 \text{ x } 1 + 7^2 \text{ x } 2 + 8^2 \text{ x } 3 + 9^2 \text{ x } 1}{7}} = 7,625$$

4.3 DADOS TABULADOS, NÃO AGRUPADOS

Para efeito de exemplificação, considere o exemplo já visto anteriormente que gerou a tabela abaixo. Nela, foram acrescentadas duas colunas. Na primeira, está a freqüência acumulada absoluta. Na segunda, aparece o produto da freqüência simples pelo valor da variável.

j	*Número de atendimentos* (x_j)	*Número de meses* (f_j)	*meses (acumulado)* (F_j)	$x_j f$
1	0	2	2	0
2	1	3	5	3
3	2	4	9	8
4	3	5	14	15
5	4	7	21	28
6	5	8	29	40
7	6	9	38	54
8	7	4	42	28
9	8	3	45	24
10	9	3	48	27
		$n = \sum_{j=1}^{10} f_j = 48$		$\sum_{j=1}^{10} x_j f_j = 227$

4.3.1 MÉDIA ARITMÉTICA

Essa média é mais facilmente obtida dos dados tabulados, partindo da nova coluna introduzida que são as parcelas necessárias aos cálculos da equação:

$$\bar{x} = \frac{\sum_{j=1}^{k} x_j f_j}{\sum_{j=1}^{k} f_j} = \frac{\sum xf}{\sum f} = \frac{\sum xf}{n} = \frac{227}{48} = 4{,}729$$

4.3.2 MODA

A moda é facilmente determinada nesse caso. Basta encontrar o valor da variável mais freqüente que é 6 porque aparece 9 vezes (***Mo = 6***).

4.3.3 MEDIANA

Aplicando a equação que determina o elemento mediano E_{Me} que define o número de ordem da mediana chega-se a:

$$E_{Me} = \frac{n+1}{2} = \frac{48+1}{2} = 24{,}5$$

Que determina a mediana como a média aritmética dos termos de ordem 24 e 25. A coluna das freqüências absolutas acumuladas F_j informa que até a ordem 5 existem 21 incidências e até a 6ª, 29. Portanto, ambos, 24 e 25, encontram-se na ordem 6, cujo valor da variável é 5. Então, a mediana ***Me*** é 5.

4.4 DADOS TABULADOS, AGRUPADOS

Para o caso de dados tabulados, agrupados em classes, será usado outro exemplo já visto, em que foram retiradas algumas colunas desnecessárias neste momento e introduzida uma coluna do produto do valor médio da classe x_j pela freqüência f_j.

j	*classes*	f_j	x_j	F_j	$x_j f_j$
1	0 |---- 10	5	5	5	25
2	10 |---- 20	15	15	20	225
3	20 |---- 30	20	25	40	500
4	30 |---- 40	45	35	85	1575
5	40 |---- 50	100	45	185	4500
6	50 |---- 60	130	55	315	7150
7	60 |---- 70	100	65	415	6500
8	70 |---- 80	60	75	475	4500
9	80 |---- 90	15	85	490	1275
10	90 |----| 100	10	95	500	950
$k = 10$		$n = \sum_{j=1}^{10} f_j = 500$			$\sum_{j=1}^{10} x_j f_j = 27.200$

4.4.1 MÉDIA ARITMÉTICA

A média aritmética é calculada da mesma forma que para os dados tabulados não agrupados em classes. Utiliza-se o ponto médio da classe para definir o valor da variável x_j na fórmula. Isso pode introduzir alguma imprecisão em relação aos dados brutos originais, já que, ao serem agrupados em classes, perde-se parte da informação. Aplicando a fórmula da média aritmética no exemplo tem-se:

$$\bar{x} = \frac{\sum_{j=1}^{k} x_j f_j}{\sum_{j=1}^{k} f_j} = \frac{\sum xf}{\sum f} = \frac{\sum xf}{n} = \frac{27200}{500} = 54{,}4$$

4.4.2 MODA

É extremamente fácil determinar qual a classe modal, bastando para isso verificar qual apresenta a maior incidência. No exemplo, é a 6ª classe, pois tem uma freqüência simples de 130. A questão, agora, é determinar qual o valor modal dentro do intervalo que vai de 50 a 60? Para isso, são usados mais de um critério. Eles serão rapidamente apresentados a seguir. No caso do exemplo dado, o três critérios conduzem a um mesmo valor modal.

MODA BRUTA

Toma-se simplesmente o valor do ponto médio da classe modal. No exemplo, como o valor médio da 6ª classe é 55, este será o valor modal ($Mo_B = 55$).

MÉTODO DE KING

Este método leva em consideração as freqüências simples das classes adjacentes. Procura aproximar a moda do limite da classe vizinha mais populosa. A equação é:

$$Mo_K = Li + Ac \frac{f_P}{f_A + f_P}$$

Onde Li é o limite inferior da classe modal, Ac é a amplitude da classe modal e f_A e f_P são as freqüências simples absolutas das classes anterior e posterior à classe modal.

No caso do exemplo fica:

$$Mo_K = 50 + 10 \frac{100}{100 + 100} = 55$$

MÉTODO DE CZUBER

Este método leva em consideração, além das freqüências das classes adjacentes, a freqüência simples absoluta da classe modal f_{Mo}. É o método mais completo e é calculado por:

$$Mo_C = Li + Ac \frac{f_{Mo} - f_A}{2f_{Mo} - (f_P + f_A)}$$

No caso do exemplo:

$$Mo_C = 50 + 10 \frac{130 - 100}{2 \text{ x } 130 - 100 + 100} = 55$$

4.4.3 MEDIANA

CÁLCULO DA CLASSE MEDIANA

Para encontrar a classe mediana, ou seja, a classe em que se encontra a mediana, acha-se o elemento mediano e procura-se, por este número de ordem, a classe em que ele se encontra. O cálculo do elemento mediano, para os dados agrupados em classes, se forem contínuos, é feito por uma fórmula diferente da vista até aqui. É a seguinte, em que n é o número de elementos:

$$E_{Me} = \frac{n}{2}$$

Para o exemplo:

$$E_{Me} = \frac{500}{2} = 250$$

Pela freqüência absoluta acumulada F_j, vê-se que até a 5ª classe encontram-se 185 observações e, até o final da 6ª, 315. Portanto a 6ª classe é a classe mediana.

CÁLCULO DA MEDIANA

Considerando Li o limite inferior da classe mediana, Ac a amplitude da classe mediana, f_{Me} a freqüência simples absoluta da classe mediana e F_A a freqüência acumulada absoluta da classe anterior à classe mediana, tem-se:

$$Me = Li + Ac \frac{E_{Me} - F_A}{f_{Me}}$$

No exemplo, como a classe mediana é a sexta:

$$Me = 50 + 10 \frac{250 - 185}{130} = 55$$

4.5 SEPARATRIZES

Algumas outras medidas de posição são muito utilizadas. Caracterizam-se por não serem medidas de tendência central, mas nem por isso deixam de ter sua importância. Assemelham-se à mediana, na medida que também dividem a distribuição em partes com o mesmo número de elementos, embora não como aquela que o faz em duas metades, ou seja, a partir do centro. Os próprios nomes esclarecem em quantas partes é dividida a distribuição. Quartis, decis e centis ou percentis repartem o conjunto, respectivamente, em 4, 10 e 100 partes com o mesmo número de elementos. Pela semelhança, as fórmulas são todas muito parecidas com as do cálculo da mediana.

4.5.1 QUARTIS

Calcula-se o elemento quartil da distribuição $E_{Q\,i}$, a partir do número n de elementos, onde i é o número do quartil, através de:

$$E_{Q\,i} = \frac{i.n}{4} \qquad \text{onde} \quad i = 1, 2 \text{ e } 3.$$

Analogamente à mediana, com o elemento quartil, descobre-se a classe em que se encontra o quartil procurado. Então, com o limite inferior Li dessa classe, com a sua amplitude Ac e sua freqüência simples absoluta f_Q , além da freqüência absoluta acumulada da classe anterior F_A, pode-se calcular o quartil desejado.

$$Q_i = Li + Ac\,\frac{E_{Q\,i} - F_A}{f_Q}$$

4.5.2 DECIS

Calcula-se o elemento decil da distribuição $E_{D\,i}$, a partir do número n de elementos, onde i é o número do decil, através de:

$$E_{D\,i} = \frac{i.n}{10} \qquad \text{onde} \quad i = 1, 2, 3,..., 9.$$

Analogamente à mediana, com o elemento decil, descobre-se a classe em que se encontra o decil procurado. Então, com o limite inferior Li dessa classe, com a sua amplitude Ac e sua freqüência simples absoluta f_D, além da freqüência absoluta acumulada da classe anterior F_A, pode-se calcular o decil desejado.

$$D_i = Li + Ac\,\frac{E_{D\,i} - F_A}{f_D}$$

4.5.3 CENTIS OU PERCENTIS

Calcula-se o elemento centil ou percentil da distribuição $E_{C\,i}$, a partir do número n de elementos, onde i é o número do centil, através de:

$$E_{C\,i} = \frac{i.n}{100} \qquad \text{onde} \quad i = 1, 2, 3,..., 99.$$

Analogamente à mediana, com o elemento centil, descobre-se a classe em que se encontra o centil procurado. Então, com o limite inferior Li dessa classe, com a sua amplitude Ac e sua freqüência simples absoluta f_C, além da freqüência absoluta acumulada da classe anterior F_A, pode-se calcular o centil desejado.

$$C_i = Li + Ac\,\frac{E_{C\,i} - F_A}{f_C}$$

4.6 EXERCÍCIOS

1. Sete empregados de uma empresa têm salários de R$ 1530,00, R$ 1360,00, R$ 1530,00, R$ 680,00 R$ 170,00 R$ 1020,00 R$ 510,00. Calcule:

   a) salário modal
   b) O salário mediano
   c) O salário médio

2. Se a empresa acima admitir mais um funcionário, percebendo R$ 170,00, como se modificam os valores anteriormente calculados?

3. Para os escores 205, 6, 5, 5, 5, 2 e 1 calcule a moda a mediana e a média. Qual das medidas de tendência central é inadequada para descrever essa situação?

4. Seis estudantes de psicologia foram testados por um instrumento que produz medidas de natureza intervalar. O objetivo da pesquisa era mensurar suas atitudes em relação a um grupo minoritário. Suas respostas , numa escala de 1 a 10 (quanto maior o escore, maior a tolerância), foram: 5, 2, 6, 3, 1 e 1. Calcule a moda, a mediana e a média.

5. Encontre a moda, a mediana e a média para os escores abaixo:

   a) 10, 12, 14, 8, 6, 7, 10, 10
   b) 3, 3, 4, 3, 1, 6, 5, 6, 6, 4
   c) 8, 8, 7, 9, 10, 5, 6, 8, 8
   d) 5, 4, 6, 6, 1, 3
   e) 8, 6, 10, 12, 1, 3, 4, 4
   f) 12, 12, 1, 12, 5, 6, 7

6. Os escores relativos a atitudes de 31 estudantes frente a um grupo minoritário, foram dispostos na seguinte distribuição de freqüências (quanto maior, mais favorável a atitude). Calcule a média, a moda e a mediana:

Escore	1	2	3	4	5	6	7
freqüências	2	4	5	7	6	4	3

7. Dada a distribuição abaixo, calcule a moda , a média e a mediana:

Escore	1	2	3	4	5	6	7	8	9	10
freqüências	1	1	2	5	7	9	8	6	4	3

8. A tabela abaixo representa a distribuição salarial em número de salários mínimos de 100 estudantes de uma universidade. Encontre o salário médio mediano e modal, este segundo os três critérios (Bruta, King e Czuber).

Nº de salários mínimos	0 ⊢ 2	2 ⊢ 4	4 ⊢ 6	6 ⊢ 8	8 ⊢ 10
Nº de estudantes	40	30	10	15	5

9. A tabela abaixo apresenta os resultados de várias análises de uma substância química, em percentagens. Determine a média e a mediana da distribuição.

%	0 ⊢ 16	16 ⊢ 32	32 ⊢ 48	48 ⊢ 64	64 ⊢ 80	80 ⊢ 96
freqüência	3	3	6	8	4	1

10. Dadas as distribuições de freqüências abaixo, encontrar a moda bruta, a mediana e a média aritmética.

a)

Intervalo de classe	5 -9	10 - 14	15 – 19	20 – 24
freqüência	5	8	4	2

b)

Intervalo de classe	40 - 49	50 - 59	60 - 69	70 - 79	80 - 89	90 - 99
freqüência	3	2	3	15	17	16

c)

Intervalo de classe	5 - 7	8 - 10	11 - 13	14 - 16	17 – 19
freqüência	1	5	6	3	2

11. Seja a distribuição das estaturas dos alunos de um curso:

Estatura (cm)	Número de alunos
140 ⊢ 150	5
150 ⊢ 160	10
160 ⊢ 170	30
170 ⊢ 180	40
180 ⊢ 190	10
190 ⊢ 200	5

Determinar:

a) A estatura média.
b) A estatura modal (Czuber).
c) A estatura mais freqüente (King).
d) A estatura mediana.
e) Os limites onde estão compreendidos 50% das estaturas intermediárias.

12. Diversas pesagens de sacas de determinado produto determinam a distribuição abaixo:

Pesos (kg)	Número de sacas
14,55 ⊢ 15,05	1
15,05 ⊢ 15,55	3
15,55 ⊢ 16,05	8
16,05 ⊢ 16,55	9
16,55 ⊢ 17,05	10
17,05 ⊢ 17,55	6
17,55 ⊢ 18,05	3

a) Qual a média da distribuição?
b) Qual a mediana?
c) Qual a moda segundo os três métodos?
d) Qual o septuagésimo quinto centil?
e) Qual o terceiro decil?
f) A percentagem de sacas entre a mediana e o 75º centil?

13. A tabela abaixo mostra uma avaliação comportamental (quanto maior a nota, mais satisfatório o comportamento).

Notas	50 ⊢ 60	60 ⊢ 70	70 ⊢ 80	80 ⊢ 90	90 ⊢ 100
Nº de alunos	5	10	20	10	5

Calcule:

a) A média.
b) A moda (as três).
c) A mediana.
d) 1º quartil.
e) 3º quartil.
f) 6º decil.
g) 85º centil.

5 MEDIDAS DE DISPERSÃO

As medidas de tendência central dão uma boa idéia das distribuições, mas não fornecem informações a respeito de como os dados se espraiam em torno dos valores médios. Para isso, são adotadas outras medidas, as de desvio ou dispersão. Abaixo são apresentadas três séries numéricas:

Série A: 7, 8, 9, 10, 11, 12. 13.
Série B: 1, 3, 6, 10, 14, 17, 19.
Série C: 1007, 1008, 1009, 1010, 1011, 1012, 1013.

Pode-se ver que dez é tanto a média aritmética como a mediana das duas primeiras séries. Portanto, as promédias não dizem tudo a respeito da natureza de um conjunto de dados, pois é fácil perceber que as duas séries possuem características bem distintas, já que a primeira está muito mais concentrada em torno da média.

Há dois tipos de medidas de dispersão: as absolutas e as relativas. A primeira e a terceira séries acima têm a mesma dispersão absoluta, mas é evidente, pela simples observação, que, relativamente aos valores médios, a série **C** apresenta desvios muito menores. A compreensão e o cálculo das medidas de dispersão absoluta facilita o entendimento e possibilita a determinação quantitativa das medidas de dispersão relativa que serão vistas no final.

Para efeitos de exemplificação, além das séries acima, será usada a tabela abaixo, que já foi apresentada anteriormente:

Ordem (j)	*Variável (x_j)*	*Freqüência (f_j)*
1	0	2
2	1	3
3	2	4
4	3	5
5	4	7
6	5	8
7	6	9
8	7	4
9	8	3
10	9	3
		48

5.1 MEDIDAS INTERVALARES

5.1.1 AMPLITUDE TOTAL

DEFINIÇÃO

É a diferença entre os valores extremos do conjunto:

$$At = x_M - x_m$$

Nas três séries acima:

$$At_1 = 13 - 7 = 6 \quad , \quad At_2 = 19 - 1 = 18 \quad \text{e} \quad At_3 = 1013 - 1007 = 6$$

No caso de dados tabulados não agrupados em classes, como na tabela acima:

$$At = 9 - 0 = 9$$

Quando os dados são agrupados em classes e não dispomos dos dados brutos que originaram a tabela, dois métodos podem ser usados para estimar a amplitude total da distribuição. Num, usam-se os valores médios das classes limites da distribuição. No outro, mais indicado, utilizam-se os valores limites dessas classes limites. Genericamente, as equações ficam num e noutro caso, ficam:

$$At = x_{jM} - x_{jm} \qquad \text{e} \qquad At = x_M - x_m$$

RESTRIÇÕES

A amplitude total dá uma visão rápida e superficial da dispersão em quase todos os casos. Embora em alguns deles ela seja satisfatória na medida da dispersão - como ocorre, por exemplo, com as máximas e mínimas temperaturas ou precipitações pluviométricas -, há restrições à sua utilização, na maioria das vezes. Elas podem ser resumidas em duas:

- Valores anormais afetam a medida de modo acentuado. Um valor atípico ou mesmo uma única medida errada modifica drasticamente o valor da amplitude. Se na série **C**, apresentada acima, houvesse mais um número, treze, por exemplo, a amplitude total passaria de seis para mil.

- Se forem tomadas várias amostras de uma mesma população, os valores da amplitude das amostras vão diferir, o que por si só já demonstra a pouca utilidade quantitativa deste parâmetro.

5.1.2 DESVIO QUARTIL

DEFINIÇÃO

Também chamado de amplitude inter-quartílica, o desvio quartil é a média aritmética das diferenças entre a mediana e o primeiro e terceiro quartil, o que resulta em:

$$d_Q = \frac{Q_3 - Q_1}{2}$$

No intervalo entre estes dois quartis encontram-se exatamente metade dos itens.

CRÍTICA

- Deve ser usado quando a medida de tendência central for a mediana
- O desvio quartil tem a vantagem de não ser afetado por valores extremos. Não apresenta o inconveniente discutido no caso da amplitude total.
- Tem o inconveniente de não ser influenciado pela forma com que os dados se distribuem, internamente, nas três regiões separadas pelos quartis que aparecem na sua fórmula.

5.2 DESVIO MÉDIO

5.2.1 DEFINIÇÃO

É a média aritmética dos valores absolutos dos desvios em relação a uma medida de tendência central que pode ser a média ou a mediana. A mais utilizada das duas é a primeira. Por isso quando se fala sobre desvio médio, simplesmente considera-se que se trata do desvio médio em relação à média aritmética. A principal vantagem do desvio médio em relação às medidas intervalares é que ele leva em conta todos os valores da distribuição. É menos usado que o desvio padrão, que será visto posteriormente, e tem em relação a este a desvantagem adicional de exigir o prévio cálculo da média ou da mediana para que a tabela seja completada.

5.2.2 DADOS BRUTOS

DESVIO MÉDIO EM RELAÇÃO À MÉDIA ARITMÉTICA

Para dados não tabulados, o desvio médio pode ser calculado diretamente a partir deles e de sua média aritmética, mediante:

$$\bar{d} = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n} = \frac{\sum_{i=1}^{n} |d_i|}{n} = \frac{\sum |d|}{n}$$

Para as séries numéricas **A**, **B** e **C** vistas acima, chega-se a:

$$\bar{d}_A = \frac{|7-10| + |8-10| + ... + |13-10|}{7} = \frac{12}{7} = 1{,}714$$

$$\bar{d}_B = \frac{|1-10| + |3-10| + ... + |19-10|}{7} = \frac{40}{7} = 5{,}714$$

$$\bar{d}_C = \frac{|1007-1010| + |1008-1010| + ... + |1013-1010|}{7} = 1{,}714$$

DESVIO MÉDIO EM RELAÇÃO À MEDIANA

Para dados não tabulados, o desvio médio pode ser calculado diretamente a partir deles e de sua mediana, mediante:

$$d_{Me} = \frac{\sum_{i=1}^{n} |x_i - Me|}{n} = \frac{\sum_{i=1}^{n} |d_{Me\,i}|}{n}$$

Como a mediana é igual a média aritmética nas series A, B e C, os valores desses desvios são os mesmos que já foram calculados.

5.2.3 DADOS TABULADOS

DEFINIÇÃO

O cálculo do desvio médio para dados tabulados é feito pelas equações abaixo, em relação à média aritmética ou à mediana, respectivamente:

$$\bar{d} = \frac{\sum_{j=1}^{k} |x_j - \bar{x}| f_j}{\sum_{j=1}^{k} f_j} = \frac{\sum_{j=1}^{k} |d_j| f_j}{n} = \frac{\sum |d| f}{n}$$

$$d_{Me} = \frac{\sum_{j=1}^{k} |x_j - Me| f_j}{\sum_{j=1}^{k} f_j} = \frac{\sum_{j=1}^{k} |d_{Me\,j}| f_j}{n}$$

Se os dados forem agrupados em classes, toma-se o valor do ponto médio da variável para efetuar os cálculos acima.

EXEMPLO DA TABELA

Como apenas o desvio em relação à média aritmética é usualmente utilizado, somente ele será calculado no exemplo abaixo.

Preenche-se os dados da tabela anteriormente apresentada, até a quarta coluna para, com isso, poder calcular a média aritmética necessária para completar a tabela.

A média aritmética é dada por:

$$\bar{x} = \frac{\sum xf}{\sum f} = \frac{\sum xf}{n} = \frac{227}{48} = 4{,}729$$

Após completada a tabela:

$$\bar{d} = \frac{\sum |d|f}{n} = \frac{90{,}626}{48} = 1{,}888$$

j	x_j	f_j	$x_j f_j$	$d_j = x_j - \bar{x}$	$\lvert d_j \rvert$	$\lvert d_j \rvert f_j$
1	0	2	0	0 - 4,729 = - 4,729	4,729	9,458
2	1	3	3	1 - 4,729 = - 3,729	3,729	11,187
3	2	4	8	2 - 4,729 = - 2,729	2,729	10,916
4	3	5	15	3 - 4,729 = - 1,729	1,729	8,645
5	4	7	28	4 - 4,729 = - 0,729	0,729	5,103
6	5	8	40	5 - 4,729 = 0,271	0,271	2,168
7	6	9	54	6 - 4,729 = 1,271	1,271	11,439
8	7	4	28	7 - 4,729 = 2,271	2,271	9,084
9	8	3	24	8 - 4,729 = 3,271	3,271	9,813
10	9	3	27	9 - 4,729 = 4,271	4,271	12,813
		48	227			90,626

5.3 DESVIO PADRÃO PARA DADOS BRUTOS

5.3.1 VARIÂNCIA

A questão das discrepâncias terem de ser tomadas em módulo, para evitar que sejam anuladas no cálculo do desvio médio, sugere um outro tipo de medida que torne desnecessária esta providência. Chama-se variância e usa o quadrado dos desvios individuais. A fórmula da variância é muito importante porque será base para o cálculo do desvio padrão que é a medida de dispersão mais utilizada.

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n} = \frac{\sum_{i=1}^{n} d_i^{\,2}}{n}$$

No caso das séries **A**, **B** e **C**, a variância é:

$$\sigma_A^2 = \frac{(7-10)^2 + (8-10)^2 + \ldots + (13-10)^2}{7} = \frac{28}{7} = 4$$

$$\sigma_B^2 = \frac{(1-10)^2 + (3-10)^2 + \ldots + (19-10)^2}{7} = \frac{292}{7} = 41{,}714$$

$$\sigma_C^2 = \frac{(1007-1010)^2 + (1008-1010)^2 + \ldots + (1013-1010)^2}{7} = 4$$

A fórmula de cálculo acima apresenta o mesmo inconveniente do desvio médio, que é a necessidade de encontrar os desvios individuais. Aqui, todavia, pode-se encontrar a variância mediante a seguinte expressão que é equivalente:

$$\sigma^2 = \frac{\sum_{i=1}^{n} x_i^2}{n} - \left(\frac{\sum_{i=1}^{n} x_i}{n} \right)^2 = \overline{x}_q^2 - \overline{x}^2 = \overline{x^2} - \overline{x}^2$$

Onde $\overline{x}_q$ é chamada de média quadrática e vale a raiz quadrada da média dos quadrados:

$$\overline{x}_q = \sqrt{\overline{x^2}} = \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n}}$$

De onde: $\overline{x}_q^2 = \overline{x^2} = \frac{\sum_{i=1}^{n} x_i^2}{n}$

A variância pode ser calculada como a diferença entre os quadrados da média quadrática e da média aritmética.

Que no caso das séries numéricas apresentadas fica:

$$\sigma_A^2 = \frac{7^2 + 8^2 + 9^2 + 10^2 + 11^2 + 12^2 + 13^2}{7} - 10^2 = \frac{728}{7} - 100 = 4$$

$$\sigma_B^2 = \frac{1^2 + 3^2 + 6^2 + 10^2 + 14^2 + 17^2 + 19^2}{7} - 10^2 = \frac{992}{7} - 100 = 41{,}714$$

$$\sigma_C^2 = \frac{1007^2 + 1008^2 + ... + 1013^2}{7} - 1010^2 = \frac{7140728}{7} - 1020100 = 4$$

Que são os mesmos valores já encontrados.

5.3.2 DESVIO PADRÃO

A variância apresenta uma dificuldade ligada ao fato de não ter a dimensão da unidade da medida. Um conjunto de medidas de comprimento em metros, por exemplo, tem como unidade de variância metros quadrados que é unidade de área e não de comprimento. Essa questão é resolvida pela definição de uma grandeza que é a raiz quadrada da variância, chamada desvio padrão.

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n}} = \sqrt{\frac{\sum_{i=1}^{n} d_i^2}{n}}$$

A fórmula alternativa fica:

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n} - \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^2} = \sqrt{\overline{x^2}_q - \overline{x}^2} = \sqrt{\overline{x^2} - \overline{x}^2}$$

Nas séries **A**, **B** e **C** o desvio padrão é:

$$\sigma_A = \sqrt{\sigma_A^2} = \sqrt{4} = 2$$

$$\sigma_B = \sqrt{\sigma_B^2} = \sqrt{41{,}714} = 6{,}459$$

$$\sigma_C = \sqrt{\sigma_C^2} = \sqrt{4} = 2$$

5.3.3 POPULAÇÃO E AMOSTRA

As fórmulas apresentadas para o cálculo da variância e o desvio padrão são apropriadas para distribuições em que toda a população é considerada. Quando o cálculo é realizado sobre dados retirados de amostragens parciais da população, a conclusão sobre os valores daquelas grandezas precisam sofrer correções. Elas derivam da redução de um dos graus de liberdade, pois a última discrepância não é independente. Assim, o valor de ***n*** deve ser substituído por ***n-1*** nas fórmulas anteriores. Abaixo, são apresentadas as relações que permitem encontrar as variâncias e desvios padrões da população, a partir dos valores correspondentes para as amostras que já foram vistos.

É importante salientar que, em qualquer dos casos, busca-se saber os valores da variância e do desvio padrão da população. Quando se dispõe de todos os dados da população, não há a necessidade das correções e usam-se as fórmulas já vistas.

VARIÂNCIA E DESVIO PADRÃO POPULACIONAL

Os valores da variância e desvio padrão abaixo são obtidos a partir da população:

Variância: $$\sigma^2 = \frac{\sum_{i=1}^{n} x_i^2}{n} - \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^2 = \overline{x^2} - \overline{x}^2$$

Desvio Padrão: $$\sigma = \sqrt{\sigma^2} = \sqrt{\overline{x^2} - \overline{x}^2}$$

VARIÂNCIA E DESVIO PADRÃO AMOSTRAL

Os valores da variância e desvio padrão são obtidos a partir do valores correspondentes encontrados para a amostra extraída da população. Encontram-se σ^2 e σ como se a amostra fosse a própria população e efetuam-se as correções abaixo para a estimação dos valores populacionais:

Variância: $$s^2 = \sigma^2 \frac{n}{n-1}$$

Desvio Padrão: $$s = \sqrt{s^2} = \sigma \sqrt{\frac{n}{n-1}}$$

Quando a população é a própria amostra estas correções não são feitas e pode-se concluir que:

$$\sigma^2 = s^2 \qquad \text{e} \qquad \sigma = s$$

5.3.4 PROPRIEDADES

- Somando (ou subtraindo) todos os números de uma distribuição por uma constante, o desvio padrão não se altera.
- Multiplicando (ou dividindo) todos os números de uma distribuição por uma constante, o desvio padrão fica multiplicado (ou dividido) pela mesma constante
- desvio padrão é sempre maior que o desvio médio

5.4 DESVIO PADRÃO PARA DADOS TABULADOS

5.4.1 CÁLCULO DA VARIÂNCIA E DO DESVIO PADRÃO

Quando os dados são tabelados, para cada valor da variável existe uma freqüência e por isso as fórmulas se alteram para facilitar o cálculo. Não importa se os dados são ou não agrupados em classes, pois aplica-se o valor da variável ou seu valor médio para cada classe. As equações ficam:

VARIÂNCIA POPULACIONAL

$$\sigma^2 = \frac{\sum_{j=1}^{k} (x_j - \bar{x})^2 f_j}{\sum_{j=1}^{k} f_j} = \frac{\sum_{j=1}^{k} d_j^2 f_j}{n} = \frac{\sum d^2 f}{\sum f} = \frac{\sum d^2 f}{n}$$

Ou através da fórmula mais fácil de calcular:

$$\sigma^2 = \frac{\sum_{j=1}^{k} x_j{}^2 f_j}{\sum_{j=1}^{k} f_j} - \left(\frac{\sum_{j=1}^{k} x_j \ f_j}{\sum_{j=1}^{k} f_j} \right)^2 = \frac{\sum_{j=1}^{k} x_j{}^2 f_j}{n} - \left(\frac{\sum_{j=1}^{k} x_j \ f_j}{n} \right)$$

Expressão que apresentada em uma forma simplificada fica:

$$\sigma^2 = \frac{\sum x^2 \ f}{\sum f} - \left(\frac{\sum x \ f}{\sum f} \right)^2 = \frac{\sum x^2 \ f}{n} - \left(\frac{\sum x \ f}{n} \right)^2 = \overline{x^2} - \overline{x}^2$$

Onde a média dos quadrados e a média aritmética podem ser calculadas, respectivamente, através de:

$$\overline{x^2} = \frac{\sum_{j=1}^{k} x_j{}^2 f_j}{\sum_{j=1}^{k} f_j} = \frac{\sum_{j=1}^{k} x_j{}^2 f_j}{n} = \frac{\sum x^2 \ f}{\sum f} = \frac{\sum x^2 \ f}{n}$$

$$\overline{x} = \frac{\sum_{j=1}^{k} x_j f_j}{\sum_{j=1}^{k} f_j} = \frac{\sum_{j=1}^{k} x_j f_j}{n} = \frac{\sum x \ f}{\sum f} = \frac{\sum x \ f}{n}$$

A partir daí, os cálculos e observações referidos para as demais medidas de dispersão são válidos, como pode ser visto a seguir.

DESVIO PADRÃO POPULACIONAL

$$\sigma = \sqrt{\sigma^2} = \sqrt{\overline{x}_q^2 - \overline{x}^2} = \sqrt{\overline{x^2} - \overline{x}^2}$$

VARIÂNCIA AMOSTRAL

$$s^2 = \sigma^2 \frac{n}{n-1}$$

DESVIO PADRÃO AMOSTRAL

$$s = \sqrt{s^2} = \sigma \sqrt{\frac{n}{n-1}}$$

5.4.2 EXEMPLO DA TABELA

MÉTODO DOS DESVIOS OU PELA DEFINIÇÃO

No exemplo da tabela, novas colunas foram introduzidas para que seja possível calcular o desvio padrão.
Para que o preenchimento a partir da quinta coluna possa ser efetivado, há que, anteriormente, calcular a média aritmética, dada por:

$$\bar{x} = \frac{\sum xf}{\sum f} = \frac{\sum xf}{n} = \frac{227}{48} = 4{,}729$$

j	x_j	f_j	$x_j f_j$	$d_j = x_j - \bar{x}$	$d_j^{\,2}$	$d_j^{\,2} f_j$
1	0	2	0	0 - 4,729 = - 4,729	22,363	44,726
2	1	3	3	1 - 4,729 = - 3,729	13,905	41,715
3	2	4	8	2 - 4,729 = - 2,729	7,447	29,788
4	3	5	15	3 - 4,729 = - 1,729	2,989	14,945
5	4	7	28	4 - 4,729 = - 0,729	0,531	3,717
6	5	8	40	5 - 4,729 = 0,271	0,073	0,584
7	6	9	54	6 - 4,729 = 1,271	1,615	14,535
8	7	4	28	7 - 4,729 = 2,271	5,157	20,628
9	8	3	24	8 - 4,729 = 3,271	10,699	32,097
10	9	3	27	9 - 4,729 = 4,271	18,241	54,723
		48	227			257,458

Então, pode-se encontrar a variância populacional:

$$\sigma^2 = \frac{\sum_{j=1}^{k} d_j^2 f_j}{\sum_{j=1}^{k} f_j} = \frac{257{,}458}{48} = 5{,}364$$

MÉTODO DA MÉDIA QUADRÁTICA OU SIMPLIFICADO

Para evitar a necessidade do cálculo dos desvios ou discrepâncias, calcula-se a média quadrática:

j	x_j	f_j	$x_j f_j$			$x_j^{\,2} f_j$
1	0	2	0			0
2	1	3	3			3
3	2	4	8			16
4	3	5	15			45
5	4	7	28			112
6	5	8	40			200
7	6	9	54			324
8	7	4	28			196
9	8	3	24			192
10	9	3	27			243
		48	227			1331

Portanto, com um única coluna de fácil cálculo, tem-se tudo que é preciso para determinar a variância populacional:

$$\sigma^2 = \frac{\sum x^2 f}{\sum f} - \left(\frac{\sum x f}{\sum f}\right)^2 = \overline{x^2} - \bar{x}^2 = \frac{1331}{48} - \left(\frac{227}{48}\right)^2 = 5{,}364$$

CÁLCULO DOS DESVIOS PADRÕES

Com o valor da variância populacional: $\sigma^2 = 5{,}364$

Pode-se achar o desvio padrão populacional:

$$\sigma = \sqrt{\sigma^2} = \sqrt{\bar{x}_q^2 - \bar{x}^2} = \sqrt{\overline{x^2} - \bar{x}^2} = \sqrt{5{,}364} = 2{,}316$$

Também encontra-se a variância amostral, ou seja, pode-se estimar a variância da população de onde foi extraída a amostra:

$$s^2 = \sigma^2 \frac{n}{n-1} = 5{,}364 \frac{48}{48-1} = 5{,}478$$

E o desvio padrão amostral:

$$s = \sqrt{s^2} = \sqrt{5{,}478} = 2{,}341$$

5.5 MEDIDAS DE DISPERSÃO RELATIVA

5.5.1 COEFICIENTE DE VARIAÇÃO DE PEARSON

É o mais utilizado e, por isso, o mais importante. Pode ser apresentado na forma de proporção ou percentual. É definido, num e noutro caso, através de:

$$CV_P = \frac{s}{\bar{x}} \qquad CV_P = 100\ \frac{s}{\bar{x}}$$

Quando for amostrada toda a população, o cálculo do desvio padrão não precisa ser corrigido e as equações acima ficam mais simples:

$$CV_P = \frac{\sigma}{\mu} \qquad CV_P = 100\ \frac{\sigma}{\mu}$$

Para as séries **A**, **B** e **C** e a tabela chega-se aos Coeficientes de Variação de Pearson na tabela abaixo:

	Coeficiente de Variação de Pearson			
	Proporcional	*Percentual*	*Proporcional*	*Percentual*
Série A	0,2000	20,00	0,2160	21,60
Série B	0,6459	64,59	0,6977	69,77
Série C	0,00198	0,198	0,00214	0,214
Tabela	0,4897	48,97	0,4950	49,50

Valores esses consistentes com as séries e tabela apresentadas, pois o pequeno desvio relativo da série **C** aparece nitidamente no Coeficiente de Pearson.

5.5.2 COEFICIENTE DE VARIAÇÃO DE THORNDIKE

Assemelha-se com o de Pearson à exceção de que é calculado em relação à mediana em lugar da média aritmética:

$$CV_T = \frac{\sigma}{Me} \qquad CV_T = 100\ \frac{\sigma}{Me}$$

$$CV_T = \frac{s}{Me} \qquad CV_T = 100\frac{s}{Me}$$

5.5.3 DESVIO QUARTIL REDUZIDO

É o quociente entre o desvio quartil e a mediana. Pode ser apresentado na forma de proporção ou de percentagem:

$$d_{QR} = \frac{d_Q}{Me} = \frac{Q_3 - Q_1}{2Me} \qquad d_{QR} = 100\ \frac{d_Q}{Me} = 50\ \frac{Q_3 - Q_1}{Me}$$

5.5.4 COEFICIENTE QUARTÍLICO DE VARIAÇÃO

Definido pelas relações abaixo, nos dois modos, proporcional ou percentual:

$$CV_Q = \frac{Q_3 - Q_1}{Q_3 + Q_1} \qquad CV_Q = 100\ \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

5.6 EXERCÍCIOS

1. Cinco estudantes obtiveram as seguintes notas em um exame: 7, 5, 3, 2 e 1. Calcule a amplitude total, o desvio médio, a variância, o desvio padrão e o coeficiente de variação de Pearson.

2. Calcule para os seguintes conjuntos de escores, a amplitude total, o desvio médio, a variância e o desvio padrão e o coeficiente de variação de Pearson.

   a) 3, 5, 4, 5, 1
   b) 6, 1, 6, 3, 7, 4
   c) 12, 10, 12, 9, 8

3. Numa escala destinada a mensurar atitudes para com o problema da segregação racial, duas classes de universitários obtiveram os seguintes pontos:

   Classe A → 1,4,1,6,2,1 e Classe B → 3, 1, 2, 4, 3, 2

   Compare a variabilidade de atitudes relativamente à segregação racial entre os membros das classes A e B mediante a determinação da amplitude total, o desvio médio, a variância e o desvio padrão e o coeficiente de variação de Pearson.

4. Calcule a variância e o desvio padrão para as seguintes distribuições freqüencial de escores.

   a)

Escores	1	2	3	4	5
Freqüências	2	2	6	5	3

   b)

Escores	1	2	3	4	5	6	7
Freqüências	1	3	4	7	5	3	2

   c)

Escores	5	6	7	8	9	10
Freqüências	3	4	7	8	5	2

5. Calcule nas tabelas seguintes, para dados agrupados em classes, a amplitude total, a variância, o desvio padrão e o coeficiente de variação de Pearson.

a)

Escores	50 - 59	60 - 69	70 - 79	80 – 89	90 - 99
Freqüências	2	3	4	8	6

b)

Escores	*5 - 7*	*8 - 10*	*11 - 13*	*14 - 16*	*17 - 19*
Freqüências	*1*	*5*	*6*	*3*	*2*

c)

Escores	5 – 9	10 - 14	15 - 19	20 - 24
Freqüências	5	8	4	2

6. A tabela abaixo representa a já apresentada distribuição salarial em número de salários mínimos de 100 estudantes de uma universidade.

Nº de sal. mínimos	0 ⊢ 2	2 ⊢ 4	4 ⊢ 6	6 ⊢ 8	8 ⊢ 10
Nº de estudantes	40	30	10	15	5

a) Encontre as medidas de dispersão absolutas: amplitude total, desvio quartil, desvio médio e desvio padrão.
b) Encontre as medidas de dispersão relativas: coeficientes de variação de Pearson e Thorndike, desvio quartil reduzido e coeficiente quartílico de variação.

7. Discuta a dispersão absoluta e relativa nos dois casos abaixo:

a) Dois conjuntos de números

A = {1.000, 1.001, 1.002, 1.003, 1.004, 1,005} e
B = {0, 1, 2, 3, 4, 5}

b) Resultados de duas provas de estatística em turmas diferentes:

Turma A → $\bar{x} = 5$ e $\mathbf{s} = 2{,}5$
Turma B → $\bar{x} = 4$ e $\mathbf{s} = 2{,}0$

8. A tabela abaixo refere-se ao problema da espessura das folhas, já parcialmente resolvido anteriormente. Encontre a variância, o desvio padrão, o coeficiente de Pearson e o desvio quartil.

Espessura (mm)	Freqüência (f_i)
1,4 ⊢ 1,6	5
1,6 ⊢ 1,8	12
1,8 ⊢ 2,0	23
2,0 ⊢ 2,2	23
2,2 ⊢ 2,4	14
2,4 ⊢ 2,6	7
2,6 ⊢ 2,8	-
2,8 ⊢ 3,0	-
3,0 ⊢ 3,2	16

9. Encontre o desvio padrão e o coeficiente de variação de Pearson abaixo:

Classe	Freqüência ($\mathbf{f_i}$)
0,5 ⊢ 2,5	11
2,5 ⊢ 4,5	47
4,5 ⊢ 6,5	87
6,5 ⊢ 8,5	134
8,5 ⊢ 10,5	200
10,5 ⊢ 12,5	198
12,5 ⊢ 14,5	164
14,5 ⊢ 16,5	102
16,5 ⊢ 18,5	48
18,5 ⊢ 20,5	6
20,5 ⊢ 22,5	3

10. Encontre a variância e o desvio padrão da distribuição abaixo, admitindo que ela seja uma amostra da população ou a própria população.

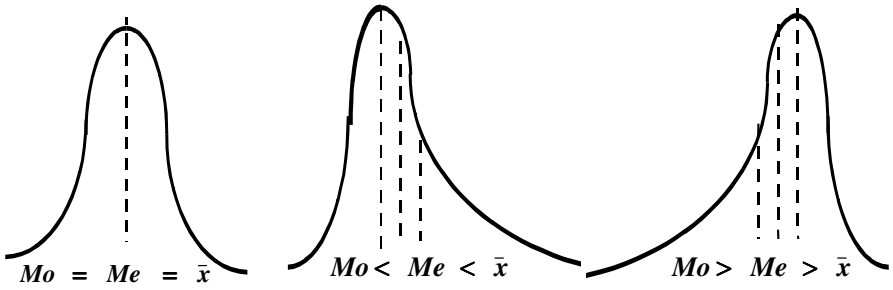
Classe	Freqüência ($\mathbf{f_i}$)	Classe	Freqüência ($\mathbf{f_i}$)
0 ⊢ 10	12	60 ⊢ 70	48
10 ⊢ 20	15	70 ⊢ 80	40
20 ⊢ 30	19	80 ⊢ 90	30
30 ⊢ 40	22	90 ⊢ 10	16
40 ⊢ 50	30	100 ⊢ 100	4
50 ⊢ 60	56	110 ⊢ 120	2

6 MEDIDAS DE DISTORÇÃO

As medidas de posição e de dispersão não bastam para descrever totalmente uma distribuição. Além das diferenças quanto à média e à variabilidade, as distribuições apresentam-se sob formas distintas. O grau de deformação pode ocorrer sob dois aspectos que produzem dois tipos de medidas, as de assimetria e curtose.

6.1 ASSIMETRIA OU ENVIESAMENTO

A figura abaixo apresenta três curvas ou distribuições de freqüências que possuem características distintas.



CURVA SIMÉTRICA SEM DEFORMAÇÃO | CURVA ASSIMÉTRICA À DIREITA | CURVA ASSIMÉTRICA À ESQUERDA

6.1.1 SIMÉTRICAS OU SEM DEFORMAÇÃO

A primeira das curvas acima é simétrica, os valores à esquerda e à direita da moda ficam igualmente distribuídos. Por isso, a moda , a mediana e a média aritmética coincidem.

6.1.2 ASSIMÉTRICA POSITIVA

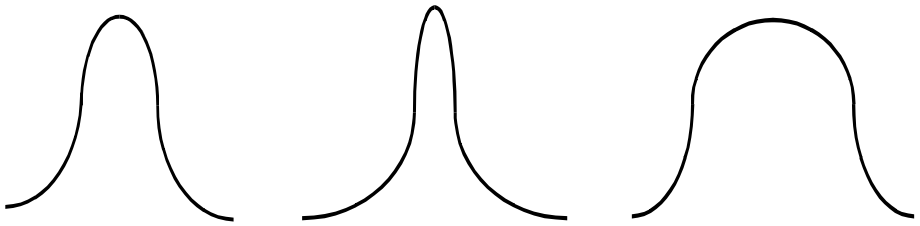
Uma distribuição com deformação positiva ou desviada à direita é inclinada e apresenta a cauda direita mais alongada, como mostra a curva central na figura acima. Neste caso, a média aritmética é maior que a mediana e esta maior que a moda, pois os valores ficam concentrados na extremidade inferior. Este é o tipo de curva mais comum.

6.1.3 ASSIMÉTRICA NEGATIVA

Uma distribuição com deformação negativa ou desviada à esquerda é inclinada e apresenta a cauda esquerda mais alongada, como mostra a curva da direita na figura acima. Neste caso, a média aritmética é menor que a mediana e esta menor que a moda, pois os valores ficam concentrados na extremidade superior. Este tipo de curva é raro.

6.2 CURTOSE

A figura abaixo apresenta três curvas ou distribuições de freqüências que possuem características distintas, embora sejam simétricas - já que têm mesmas medianas , modas e médias. Duas delas apresentam uma distorção caracterizada por um achatamento ou afilamento em relação à primeira que é normal. Esse fenômeno é conhecido por curtose.



MESOCÚRTICA LEPTOCÚRTICA PLATICÚRTICA

6.2.1 MESOCÚRTICA

O grau de achatamento é da curva normal, conforme pode ser visto na curva da esquerda, acima.

6.2.2 LEPTOCÚRTICA

O grau de afilamento é alto, maior que o da curva normal, formando uma espécie de bico, conforme pode ser visto na curva central, acima

6.2.3 PLATICÚRTICA

O grau de achatamento é alto, maior que o da curva normal, formando uma espécie de platô, conforme pode ser visto na curva da direita, acima.

6.3 DISTRIBUIÇÕES EXEMPLOS

Para efeito de cálculos das diversas medidas de assimetria e curtose que serão introduzidas a seguir, serão usadas cinco distribuições diferentes. Elas, de modo imaginário, referem-se às notas obtidas por 5 turmas de 30 alunos cada, de uma disciplina de Estatística.. A quantidade de alunos, em cada classe de notas agrupadas, é dada por:

TURMA A: [0 - 2) 2; [2 - 4) 6; [4 - 6) 14; [6 - 8) 6 e [8 - 10] 2.
TURMA B: [0 - 2) 5; [2 - 4) 10; [4 - 6) 7; [6 - 8) 5 e [8 - 10] 3.
TURMA C: [0 - 2) 1; [2 - 4) 4; [4 - 6) 8; [6 - 8) 11 e [8 - 10] 6.
TURMA D: [0 - 2) 2; [2 - 4) 4; [4 - 6) 18; [6 - 8) 4 e [8 - 10] 2.
TURMA E: [0 - 2) 2; [2 - 4) 8; [4 - 6) 10; [6 - 8) 8 e [8 - 10] 2.

A partir desses dados , abaixo foram construídas as tabelas completas. Para o uso das medidas que não utilizam o conceito de momento, não são necessárias as últimas duas colunas.

6.3.1 TABELA A

j	*Classes*	f_j	x_j	F_j	$x_j f_j$	$x_j^2 f_j$	$x_j^3 f_j$	$x_j^4 f_j$
1	0 ⊢ 2	2	1	2	2	2	2	2
2	2 ⊢ 4	6	3	8	18	54	162	486
3	4 ⊢ 6	14	5	22	70	350	1.750	8.750
4	6 ⊢ 8	6	7	28	42	294	2.058	14.406
5	8 ⊢ 10	2	9	30	18	162	1.458	13.122
		30			150	862	5.430	36.766

6.3.2 TABELA B

j	*Classes*	f_j	x_j	F_j	$x_j f_j$	$x_j^2 f_j$	$x_j^3 f_j$	$x_j^4 f_j$
1	0 ⊢ 2	5	1	5	5	5	5	5
2	2 ⊢ 4	10	3	15	30	90	270	810
3	4 ⊢ 6	7	5	22	35	175	875	4.375
4	6 ⊢ 8	5	7	27	35	245	1.715	12.005
5	8 ⊢ 10	3	9	30	27	243	2.187	19.683
		30			132	758	5.052	36.878

6.3.3 TABELA C

j	*Classes*	f_j	x_j	F_j	$x_j f_j$	$x_j^2 f_j$	$x_j^3 f_j$	$x_j^4 f_j$
1	0 ⊢ 2	1	1	1	1	1	1	1
2	2 ⊢ 4	4	3	5	12	36	108	324
3	4 ⊢ 6	8	5	13	40	200	1.000	5.000
4	6 ⊢ 8	11	7	24	77	539	3.773	26.411
5	8 ⊢ 10	6	9	30	54	486	4.374	39.366
		30			184	1.262	9.256	71.102

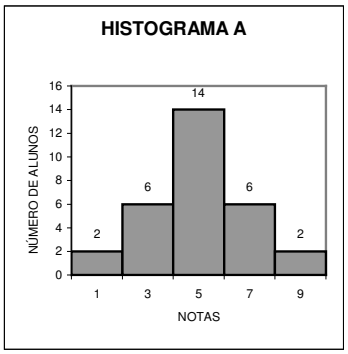
6.3.4 TABELA D

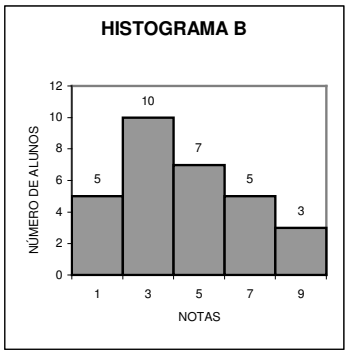
j	*Classes*	f_j	x_j	F_j	$x_j f_j$	$x_j^2 f_j$	$x_j^3 f_j$	$x_j^4 f_j$
1	0 ⊢ 2	2	1	2	2	2	2	2
2	2 ⊢ 4	4	3	6	12	36	108	324
3	4 ⊢ 6	18	5	24	90	450	2.250	11.250
4	6 ⊢ 8	4	7	28	28	196	1.372	9.604
5	8 ⊢ 10	2	9	30	18	162	1.458	13.122
		30			150	846	5.190	34.302

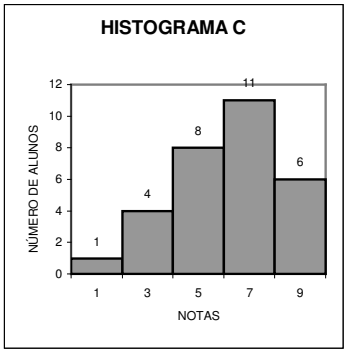
6.3.5 TABELA E

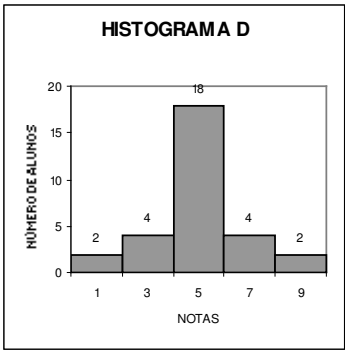
j	*Classes*	f_j	x_j	F_j	$x_j f_j$	$x_j^2 f_j$	$x_j^3 f_j$	$x_j^4 f_j$
1	0 ⊢ 2	2	1	2	2	2	2	2
2	2 ⊢ 4	8	3	10	24	72	216	648
3	4 ⊢ 6	10	5	20	50	250	1.250	6.250
4	6 ⊢ 8	8	7	28	56	392	2.744	19.208
5	8 ⊢ 10	2	9	30	18	162	1.458	13.122
		30			150	878	5.670	39.230

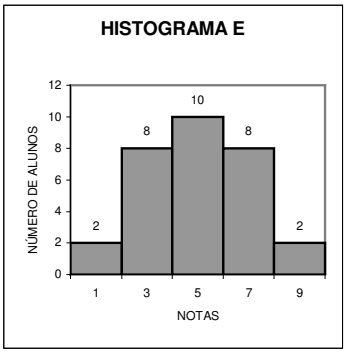
6.3.6 GRÁFICOS











6.4 MOMENTOS

Os momentos são quantidades numéricas obtidas a partir de uma distribuição de freqüências, que permitem resumi-la. Os dois momentos mais importantes já foram estudados: a média aritmética e a variância.

6.4.1 MOMENTO NATURAL OU ABSOLUTO

O momento natural de ordem **r** , cujo símbolo é $\mathbf{m_r'}$ de um conjunto de números é definido da seguinte forma:

PARA DADOS BRUTOS

$$m_r' = \frac{\sum_{i=1}^{n} x_i^r}{n} = \frac{\sum x^r}{n}$$

PARA DADOS TABELADOS

Quando os dados estiverem agrupados em classes, na fórmula abaixo, tomam-se os valores médios das classes para $\mathbf{x_j}$.

$$m_r' = \frac{\sum_{j=1}^{k} x_j^r \, f_j}{\sum_{j=1}^{k} f_j} = \frac{\sum_{j=1}^{k} x_j^r \, f_j}{n} = \frac{\sum x^r f}{n}$$

Pode-se ver, facilmente que, o momento absoluto de primeira ordem, em que **r** = 1, é a média aritmética. Analogamente, o momento absoluto de segunda ordem, em que **r** = 2, é o quadrado da média quadrática ou a média dos quadrados. Assim:

$$\bar{x} = m_1' \qquad \text{e} \qquad \bar{x}_q^2 = \overline{x^2} = m_2'$$

CÁLCULO DOS MOMENTOS ABSOLUTOS PARA OS EXEMPLOS

A partir dos valores tabelados anteriormente para os exemplos A até E, usando a equação acima, pode-se achar os momentos absolutos de ordem 1 até 4. A tabela abaixo apresenta esses resultados:

ordem	*valores de* ***m'*** *para os exemplos*				
r	A	B	C	D	E
1 ($\bar{x}$)	5,00	4,40	6,13	5,00	5,00
2 ($\bar{x}_q^2$)	28,73	25,27	42,07	28,20	29,27
3	181,00	168,40	308,53	173,00	189,00
4	1.225,53	1.229,27	2.370,07	1.143,40	1.307,67

6.4.2 MOMENTO CENTRADO NA MÉDIA

O momento de ordem **r** centrado na média, cujo símbolo é $\mathbf{m_r}$,de um conjunto de números é definido da seguinte forma:

PARA DADOS BRUTOS

$$m_r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^r}{n} = \frac{\sum (x - \bar{x})^r}{n}$$

PARA DADOS TABELADOS

Quando os dados estiverem agrupados em classes, na fórmula abaixo, tomam-se os valores médios das classes para xj.

$$m_r = \frac{\sum_{j=1}^{k} (x_j - \bar{x})^r f_j}{\sum_{j=1}^{k} f_j} = \frac{\sum_{j=1}^{k} (x_j - \bar{x})^r f_j}{n} = \frac{\sum (x - \bar{x})^r f}{n}$$

RELAÇÕES ENTRE MOMENTOS

As fórmulas acima apresentam o inconveniente de exigirem um novo tabulamento com o cálculo das discrepâncias. Felizmente, isso não é necessário, pois os momentos centrados na média **m** podem ser encontrados a partir dos momentos naturais **m'** através das relações abaixo:

$$m_1 = 0$$
$$m_2 = m'_2 - (m'_1)^2 = \sigma^2$$
$$m_3 = m'_3 - 3m'_1 m'_2 + 2(m'_1)^3$$
$$m_4 = m'_4 - 4m'_1 m'_3 + 6(m'_1)^2 m'_2 - 3(m'_1)^4$$

Pode-se ver que o momento centrado na média de segunda ordem, em que **r** = 2, é a variância populacional. Ou seja:

$$\sigma^2 = \bar{x}^2_q - \bar{x}^2 = \overline{x^2} - \bar{x}^2 = m_2 = m'_2 - (m'_1)^2$$

A partir dessas relações, pode-se calcular os valores de **m** para as distribuições exemplos, com os dados tabulados para os valores de **m'**.

ordem	*valores de **m** para os exemplos*				
r	A	B	C	D	E
1	0	0	0	0	0
2 (σ^2)	3,73	5,91	4,45	3,20	4,27
3	0	5,25	-4,05	0,00	0,00
4	40,53	75,97	50,15	38,40	42,67

6.5 PRINCIPAIS MEDIDAS DE ASSIMETRIA

6.5.1 COMPARAÇÃO ENTRE PROMÉDIAS

DEFINIÇÃO

A simples comparação entre duas das três medidas de tendência central permite, de modo rudimentar, saber o tipo de assimetria. Vale a regra seguinte:

$\bar{x} > Mo \Rightarrow$ Assimetria positiva

$\bar{x} = Mo \Rightarrow$ Simetria

$\bar{x} < Mo \Rightarrow$ Assimetria negativa

EXEMPLOS

A moda bruta foi usada na tabela abaixo, por ser mais fácil de determinar que a mediana. Ela é usada na comparação com a média aritmética. Esta já está na tabela dos momentos absolutos (ordem 1). Abaixo foi construída uma tabela com estes valores - média e moda - , cujo sinal do resultado da subtração dá o tipo de assimetria, se houver.

medidas	*valores para os exemplos*				
	A	B	C	D	E
média ($\bar{x}$)	5,00	4,40	6,13	5,00	5,00
*moda (**Mo**)*	5	3	7	5	5
$\bar{x}$ **- Mo**	0	1,40	-0,87	0	0

As distribuições **A**, **D** e **E** são simétricas, a distribuição **B** é assimétrica positiva ou desviada à direita e a distribuição **C** é assimétrica à esquerda, pois apresenta deformação negativa.

6.5.2 COEFICIENTE OU ÍNDICE DE PEARSON

PRIMEIRO COEFICIENTE DE ENVIESAMENTO DE PEARSON

O primeiro coeficiente de enviesamento de Pearson e_1 é definido por:

$$e_1 = \frac{\bar{x} - Mo}{\sigma}$$

O desvio padrão utilizado pode ser o amostral ou populacional, dependendo da situação concreta. A moda pode ser calculada por algum dos métodos alternativos.

SEGUNDO COEFICIENTE DE ENVIESAMENTO DE PEARSON

O segundo coeficiente de assimetria de Pearson e_2 é definido por:

$$e_2 = \frac{3(\bar{x} - Me)}{\sigma}$$

A menos que haja assimetria muito forte este coeficiente deve ser preferido, até porque existe um único modo de calcular a mediana, ao contrário da moda. Em distribuições simétricas ambos coeficientes dão o mesmo valor.

EXEMPLOS

Nos exemplos serão adotados as variâncias já tabeladas acima, presumindo que toda a população foi amostrada, não exigindo correções, portanto. A moda foi calculada pelo método de Czuber.

medidas	*valores para os exemplos*				
	A	B	C	D	E
média ($\bar{x}$)	5,00	4,40	6,13	5,00	5,00
*moda (**Mo**)*	5,00	3,25	6,29	5,00	5,00
*mediana (**Me**)*	5,00	4,00	6,36	5,00	5,00
σ(desvio padrão)	1,93	2,43	2,11	1,79	2,07
e_1 (Pearson)	0,00	0,49	-0,29	0,00	0,00
e_2 (Pearson)	0,00	0,37	-0,33	0,00	0,00

Vê-se novamente, através de ambos os índices de Pearson que as distribuições **A**, **D** e **E** são simétricas, a distribuição **B** é assimétrica positiva e a distribuição **C** é assimétrica à esquerda.

6.5.3 COEFICIENTE QUARTIL DE ASSIMETRIA

DEFINIÇÃO

O coeficiente quartil de assimetria e_Q é definido por:

$$e_Q = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{(Q_3 - Q_2) + (Q_2 - Q_1)} = \frac{Q_3 - 2Me + Q_1}{Q_3 - Q_1}$$

Em que o coeficiente assume valores entre -1 e +1.

EXEMPLOS

Nos exemplos, após encontrar os quartis, sendo que o segundo quartil é a própria mediana, determina-se o coeficiente quartil de assimetria.

medidas	*valores para os exemplos*				
	A	B	C	D	E
primeiro quartil (Q_1)	3,83	2,50	4,63	4,17	3,38
terceiro quartil (Q_3)	6,17	6,20	7,73	5,83	6,63
*mediana (**Me**)*	5,00	4,00	6,36	5,00	5,00
coeficiente quartil (e_Q)	0,00	0,19	-0,12	0,00	0,00

Novamente, as conclusões anteriores são confirmadas pelo coeficiente quartil de assimetria.

6.5.4 COEFICIENTE PERCENTÍLICO DE ASSIMETRIA

DEFINIÇÃO

O coeficiente percentílico de assimetria e_C é definido pelos percentis 10 e 90:

$$e_C = \frac{(C_{90} - C_{50}) - (C_{50} - C_{10})}{(C_{90} - C_{50}) + (C_{50} - C_{10})} = \frac{C_{90} - 2Me + C_{10}}{C_{90} - C_{10}}$$

EXEMPLOS

Nos exemplos, após encontrar os percentis, sendo que o centil 50 é a própria mediana, determina-se o coeficiente percentílico de assimetria.

medidas	*valores para os exemplos*				
	A	B	C	D	E
décimo percentil (C_{10} *)*	2,33	1,20	3,00	2,50	2,25
nonagésimo percentil (C_{90} *)*	7,67	8,00	9,00	7,50	7,75
mediana (Me *)*	5,00	4,00	6,36	5,00	5,00
coeficiente percentil (e_C *)*	0,00	0,18	-0,12	0,00	0,00

Novamente, as conclusões anteriores são confirmadas pelo coeficiente percentílico de assimetria.

6.5.5 COEFICIENTE MOMENTO DE ASSIMETRIA

BASEADO NOS MOMENTOS ATÉ A TERCEIRA ORDEM

O coeficiente momento de assimetria, baseado nos momentos até a terceira ordem e_{M3} , é definido por:

$$e_{M3} = \frac{m_3}{\left(\sqrt{m_2}\right)^3} = \frac{m_3}{\sigma^3} = \frac{m_3}{s^3}$$

BASEADO NOS MOMENTOS ATÉ A QUARTA ORDEM

Antes de definir esse coeficiente convém determinar a relação $\mathbf{b_2}$, que é obtidas a partir dos momentos centrados na média de segunda e quarta ordem. Esta relação é:

$$b_2 = \frac{m_4}{m_2^2} = \frac{m_4}{\sigma^4} = \frac{m_4}{s^4}$$

O coeficiente momento de assimetria, baseado nos momentos até a quarta ordem e_{M4}, é definido, a partir das relações acima definidas, por:

$$e_{M4} = \frac{e_{M3}\left(b_2 + 3\right)}{2\left(5b_2 - 6e_{M3}^2 - 9\right)}$$

EXEMPLOS

Nos exemplos, após encontrar a relação b_2 , a partir de m_2 , m_3 e m_4 , pode-se achar e_{M3} e e_{M4} , que são os coeficientes momentos de assimetria.

medidas	*valores para os exemplos*				
	A	B	C	D	E
m_2	3,73	5,91	4,45	3,20	4,27
m_3	0,00	5,25	-4,05	0,00	0,00
m_4	40,53	75,97	50,15	38,40	42,67
b_2	2,91	2,18	2,53	3,75	2,34
e_{M3}	0,00	0,37	-0,43	0,00	0,00
e_{M4}	0,00	0,86	-0,47	0,00	0,00

6.6 MEDIDAS DE CURTOSE

6.6.1 COEFICIENTE PERCENTÍLICO DE CURTOSE

DEFINIÇÃO

O Coeficiente percentílico de curtose é definido pela relação:

$$k = \frac{d_q}{C_{90} - C_{10}} = \frac{Q_3 - Q_1}{2(C_{90} - C_{10})}$$

A curva mesocúrtica perfeita tem o coeficiente percentílico de curtose **k** igual a 0,263. Quanto mais k superar este valor, mais platicúrtica será a distribuição e quanto mais for inferior, mais leptocúrtica será. Então:

k > 0,263	platicúrtica
k ≅ 0,263	mesocúrtica
k < 0,263	leptocúrtica

EXEMPLOS

Para calcular o coeficiente percentílico de curtose são necessários os quartis 1 e 3 e os percentis 10 e 90. Nos exemplos, todos esse valores já foram tabelados.

medidas	*valores para os exemplos*				
	A	B	C	D	E
décimo percentil (C_{10})	2,33	1,20	3,00	2,50	2,25
nonagésimo percentil (C_{90})	7,67	8,00	9,00	7,50	7,75
primeiro quartil (Q_1)	3,83	2,50	4,63	4,17	3,38
terceiro quartil (Q_3)	6,17	6,20	7,73	5,83	6,63
coeficiente percentílico (k)	0,219	0,272	0,259	0,167	0,295

Segundo o coeficiente percentílico de curtose, as distribuições **A**, **C** e **D** são leptocúrticas e as distribuições **B** e **E** são platicúrticas. Entretanto, apenas a distribuição **D** é fortemente leptocúrtica. Pode-se perceber que as curvas **B** e **C** são quase mesocúrticas.

6.6.2 COEFICIENTE MOMENTO DE CURTOSE

DEFINIÇÃO

O coeficiente momento de curtose utiliza-se dos momentos centrados na média de segunda e quarta ordem, anteriormente definidos. Na verdade, o coeficiente momento de curtose é a própria relação b_2 já apresentada:

$$b_2 = \frac{m_4}{m_2^2} = \frac{m_4}{\sigma^4} = \frac{m_4}{s^4}$$

A distribuição mesocúrtica é aquela que tem o coeficiente igual 3. Valores menores que este indicam curvas platicúrticas e maiores, leptocúrticas. Então vale a regra seguinte:

b_2	< 3	platicúrtica
b_2	$\cong 3$	mesocúrtica
b_2	> 3	leptocúrtica

EXEMPLOS

Repetindo os valores já calculados para os exemplos, fica a tabela.

medidas	*valores para os exemplos*				
	A	B	C	D	E
m_2	3,73	5,91	4,45	3,20	4,27
m_4	40,53	75,97	50,15	38,40	42,67
b_2	2,91	2,18	2,53	3,75	2,34

Segundo o coeficiente momento de curtose apenas a distribuição **D** é leptocúrtica. Pode-se ver que, embora as demais sejam platicúrticas, as distribuições **A** e *C* são quase mesocúrticas.

6.7 EXERCÍCIOS

1. A tabela abaixo apresenta a percentagem de bactérias encontradas por cm3 em 100 amostras de determinado produto. Determine os seguintes parâmetros:

   a) A mediana.
   b) primeiro quartil.
   c) terceiro quartil.
   d) momento natural de primeira ordem.
   e) momento natural de segunda ordem.
   f) desvio quartil.
   g) momento centrado de segunda ordem.
   h) primeiro coeficiente de assimetria de Pearson.
   i) segundo coeficiente de assimetria de Pearson.
   j) coeficiente quartil de assimetria.
   k) coeficiente de percentílico de assimetria.
   l) coeficiente percentílico de curtose;

Classes (%)	Freqüências
0,0 |— 0,1	2
0,1 |— 0,2	5
0,2 |— 0,3	10
0,3 |— 0,4	15
0,4 |— 0,5	18
0,5 |— 0,6	18
0,6 |— 0,7	15
0,7 |— 0,8	10
0,8 |— 0,9	5
0,9 |— 1,0	2

2. Determine os mesmos parâmetros do exercício anterior para as distribuições abaixo:

   a)

Classes	Freqüências
20 |— 25	10
25 |— 30	15
30 |— 35	20
35 |— 40	18
40 |— 45	4

   b)

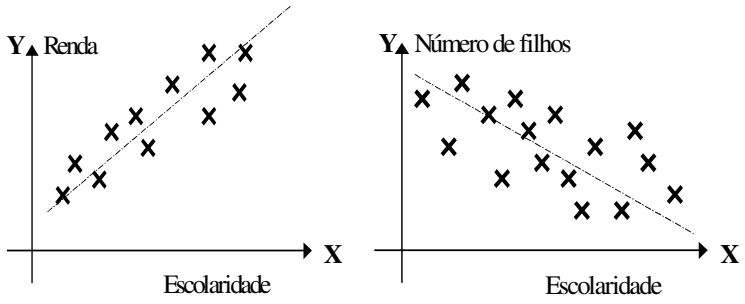
Classes	Freqüências
2 |— 4	2
4 |— 6	8
6 |— 8	10
8 |— 10	8
10 |— 12	2

7 CORRELAÇÃO E REGRESSÃO

A correlação é o estudo da relação entre variáveis, seja da simples verificação da existência ou da determinação do grau dessa relação. A regressão, por sua vez, objetiva descrever a relação quantitativamente, mediante a estimação dos parâmetros da função matemática.

7.1 CORRELAÇÃO LINEAR SIMPLES

Se duas variáveis quaisquer variam de um respondente para outro, elas podem ou não estar correlacionadas. Além disso, essa correlação pode ser maior ou menor e representada por diferentes funções. Se apontarmos, para cada um dos ***N*** pares de valores das duas grandezas, um ponto, em um gráfico que contenha a variável ***X*** nos eixos das abcissas e a variável ***Y*** no eixo das ordenadas, teremos o chamado "diagrama de dispersão". Os dois diagramas de dispersão abaixo mostram, claramente, existir correlação entre as variáveis ***X*** e ***Y***. Mais, os pontos aproximam-se de uma reta , o que determina existir, em ambos os casos, uma correlação linear simples entre as duas variáveis aleatórias. A variável ***X*** será a variável explicativa e a ***Y*** a explicada. Estas variáveis são chamadas, por vezes, independente e dependente, respectivamente, de modo impróprio. Elas estão correlacionadas, mas não apresentam, necessariamente, uma relação causal. A aproximação da correlação por uma reta consiste o objeto da regressão linear.



7.1.1 COEFICIENTE DE CORRELAÇÃO DE PEARSON

A medida da correlação linear é efetuada através do coeficiente de correlação de Pearson ***r***, dado pelas relação abaixo:

$$r = \frac{\sigma_{XY}^2}{\sigma_X \, \sigma_Y}$$

Onde σ_X e σ_Y são os desvios padrões das variáveis ***X*** e ***Y*** e σ_{XY}^2 é a covariância entre elas.

$$\sigma_X^2 = \overline{X^2} - \overline{X}^2 \quad , \quad \sigma_Y^2 = \overline{Y^2} - \overline{Y}^2 \quad \text{e} \quad \sigma_{XY}^2 = \overline{XY} - \overline{X}\,\overline{Y}$$

Em que:

$$\overline{XY} = \frac{\sum XY}{N} \quad , \quad \overline{X^2} = \frac{\sum X^2}{N} \quad , \quad \overline{Y^2} = \frac{\sum Y^2}{N} \, ,$$

$$\overline{X} = \frac{\sum X}{N} \quad \text{e} \quad \overline{Y} = \frac{\sum Y}{N}$$

Onde o intervalo de variação de ***r*** situa-se entre -1 e 1.

7.1.2 CORRELAÇÃO POSITIVA

A correlação será positiva se $r > 0$ e tão mais perfeita quanto mais próxima da unidade. O primeiro diagrama de dispersão acima apresenta uma correlação positiva, pois para ***X*** crescentes ***Y*** também aumenta. Trata-se de uma correlação linear porque os pontos do diagrama aproximam-se de uma reta.

7.1.3 CORRELAÇÃO NEGATIVA

A correlação será negativa se $r < 0$ e tão mais perfeita quanto mais próximo da unidade for seu módulo. O segundo diagrama de dispersão acima apresenta uma correlação negativa, pois para ***X*** crescentes ***Y*** diminui. Trata-se de uma correlação linear porque os pontos do diagrama aproximam-se de uma reta, porém não tão perfeita quanto a primeira, o que leva a crer que o módulo de ***r*** é menor do que naquele caso. Isso acontece porque há menor correlação entre as variáveis.

7.1.4 CORRELAÇÃO NULA

Quando não há relação entre as variáveis ***X*** e ***Y***. Não é possível encontrar uma reta que simbolize a relação entre as variáveis. Quanto mais próximo de zero for ***r***, menor a correlação entre as variáveis.

7.1.5 CORRELAÇÃO ESPÚRIA

O fato de ***r*** ser diferente de zero não significa necessariamente que exista uma correlação. Pode-se tratar de uma correlação espúria ou falsa correlação. Para identificar tais situações existem testes que serão vistos oportunamente.

7.1.6 EXEMPLO

Os dados abaixo mostram os anos de estudos de pares de pais e filhos. Procura-se determinar se existe correlação entre eles. Cada dupla é seguida, entre parênteses, dos anos de escola do pai e do filho, respectivamente:

A(12, 12); B(10 , 8); C(6 , 6); D(16 , 11);
E(8 , 10); F(9 , 8) e G(12 , 11)

A tabela abaixo apresenta os dados necessários para o cálculo do coeficiente de Pearson:

	Anos de Escola				
Dupla	*Pais (X)*	*Filhos (Y)*	X^2	Y^2	**XY**
A	12	12	144	144	144
B	10	8	100	64	80
C	6	6	36	36	36
D	16	11	256	121	176
E	8	10	64	100	80
F	9	8	81	64	72
G	12	11	144	121	132
N = 7	73	66	825	650	720

O coeficiente de correlação linear de Pearson é:

$$\overline{X} = \frac{\sum X}{N} = \frac{73}{7} = 10{,}429 \qquad \overline{Y} = \frac{\sum Y}{N} = \frac{66}{7} = 9{,}429$$

$$\overline{X^2} = \frac{\sum X^2}{N} = \frac{825}{7} = 117{,}857 \qquad \overline{Y^2} = \frac{\sum Y^2}{N} = \frac{650}{7} = 92{,}857$$

$$\overline{XY} = \frac{\sum XY}{N} = \frac{720}{7} = 102{,}857$$

$$\sigma_X^2 = \overline{X^2} - \overline{X}^2 = 117{,}857 - 10{,}429^2 = 9{,}075 \qquad \sigma_X = \sqrt{\sigma_X^2} = 3{,}012$$

$$\sigma_Y^2 = \overline{Y^2} - \overline{Y}^2 = 92{,}857 - 9{,}429^2 = 3{,}951 \qquad \sigma_Y = \sqrt{\sigma_Y^2} = 1{,}988$$

$$\sigma_{XY}^2 = \overline{XY} - \overline{X}\,\overline{Y} = 102{,}857 - 10{,}429 \text{ x } 9{,}429 = 4{,}522$$

$$r = \frac{\sigma_{XY}^2}{\sigma_X \sigma_Y} = \frac{4{,}522}{3{,}012 \text{ x } 1{,}988} = 0{,}755$$

Valor que indica uma forte correlação positiva, isto é, para a amostra em questão, os filhos dos pais que mais estudaram tendem a ser mais estudiosos. Para que se possa concluir a respeito da população, deve-se adotar os testes de significância que serão vistos oportunamente.

7.2 CORRELAÇÃO ORDINAL

Quando os dados não podem ser mensurados para que se calcule o coeficiente de correlação de Pearson, resta ordenar os dados por postos e encontrar os chamados coeficientes de correlação ordinais.

7.2.1 COEFICIENTE DE SPEARMAN

DEFINIÇÃO

O coeficiente de correlação de postos de Spearman r_S é dado por:

$$r_S = 1 - \frac{6 \sum D^2}{N(N^2 - 1)}$$

Onde D é a diferença numérica ordinal entre as duas variáveis para cada um dos N pares. Quando mais de um dado tem a mesma ordem, trabalha-se como se eles fossem diferentes e adota-se a média dos postos empatados ou espelhados.

EXEMPLO

A tabela abaixo mostra o grau de destreza em mecanografia de 12 moças que trabalham em uma empresa. Também são fornecidos seus quocientes de inteligência (Q.I.). Para calcular o coeficiente de correlação de postos de Spearman para essas duas variáveis foi preciso criar algumas colunas.

		Posições			
Nomes	*Q.I.*	*Destreza (X^o)*	*Q.I. (Y^o)*	*Desvio (D)*	D^2
A	100	5	8	-3	9
B	90	2	10,5	-8,5	72,25
C	80	6	12	-6	36
D	130	1	2	-1	1
E	100	8	8	0	0
F	100	9	8	1	1
G	110	7	5,5	1,5	2,25
H	120	3	3,5	-0,5	0,25
I	90	11	10,5	0,5	0,25
J	120	10	3,5	6,5	42,25
K	110	12	5,5	6,5	42,25
L	140	4	1	3	9
12					215,5

Note-se que as variáveis a serem utilizadas são ordinais. No exemplo, foi necessário ordenar os quocientes de inteligência e sobre eles encontrar os desvios.

O coeficiente de correlação de postos de Spearman é:

$$r_S = 1 - \frac{6 \sum D^2}{N(N^2 - 1)} = 1 - \frac{6\,(215{,}5)}{12\,(12^2 - 1)} = -0{,}247$$

Correlação fracamente negativa, indicando que quanto mais inteligentes, menos aptas são as moças àquele trabalho. Ressalte-se que a conclusão exigiria um teste de comprovação de significância para poder ser estendida à população, além do fato da medida indicar uma correlação muito fraca.

7.2.2 COEFICIENTE GAMA DE GOODMAN E KRUSKAL

DEFINIÇÃO

A correlação pode ser vista como o grau de previsão ou adivinhação que se pode ter de uma variável a partir de outra. O coeficiente gama **G** é definido como:

$$G = \frac{f_A - f_I}{f_A + f_I}$$

Onde f_A e f_I são as freqüências de acordos e inversões, respectivamente, que antecedem a cada variável explicada, após os dados serem ordenados pela variável explicativa.

Quanto maior o número de acordos, mais positiva e próxima da unidade será a correlação. Em contrapartida, quanto maior o número de inversões, mais negativa será a correlação. Também aqui o coeficiente varia entre -1 e +1.

EXEMPLO

Várias cidades foram classificadas pelo tamanho proporcional de população negra (quanto maior o posto, menor o percentual de negros) e pelo grau de discriminação na obtenção de emprego (quanto maior a ordem menor a discriminação). Procura-se verificar a existência ou não de uma correlação entre uma coisa e outra.

Os dados são tabulados abaixo nas três primeiras colunas. Nas duas seguintes são calculados o número de acordos e inversões . Para facilitar o trabalho, ordena-se uma das variáveis e conta-se, até cada casilha, quantos acordos ou inversões existem até ali. Quando houver empates eles não são contados nem para um nem para o outro lado.

Cidades	*Tamanho da população negra (X^o)*	*Nível de discriminação no emprego (Y^o)*	*Acordos*	*Inversões*
A	1	2	0	0
B	2	3	1	0
C	3	1	0	2
D	4	6	3	0
E	5	5	3	1
F	6	4	3	2
			10	5

A tabela fornece os dados para que se possa calcular o coeficiente gama.

$$G = \frac{f_A - f_I}{f_A + f_I} = \frac{10 - 5}{10 + 5} = +0{,}33$$

Indicando uma fraca correlação positiva, ou seja, para esta amostra, colhida nos Estados Unidos da América do Norte, quanto maior a população negra da cidade, maior a discriminação no emprego.

7.3 REGRESSÃO LINEAR SIMPLES

A regressão consiste em descrever, mediante um modelo matemático, a relação entre duas variáveis, a partir de inúmeras observações. Uma das variáveis é chamada explicativa (***X***) e a outra, a variável explicada (***Y***).

7.3.1 AJUSTAMENTO DO MODELO

Quando a função $Y = f(X)$ é melhor representada por uma reta, diz-se que se trata de uma regressão linear simples. O processo de determinação dos parâmetros que melhor harmoniza a reta às variáveis é denominado ajustamento.

A reta ajustada é representada por $Y' = a + b\ X$, onde a e b são os parâmetros do modelo. Os erros, diferenças entre Y' (valor da variável explicada a partir da equação de regressão) e Y (valor observado), tem seus quadrados somados e levados ao mínimo na situação de ajustamento - método dos mínimos quadráticos.

A equação da reta pode ser expressa de uma das seguintes formas:

$$Y' = a + bX \qquad \text{ou} \qquad Y' = \overline{Y} + b\left(X - \overline{X}\right)$$

Em que os valores dos parâmetros a e b são:

$$a = \overline{Y} - b\overline{X} \qquad b = \frac{\sigma_{XY}^2}{\sigma_X^2}$$

Onde todos os termos já foram definidos anteriormente.

7.3.2 PODER EXPLICATIVO DO MODELO

O poder explicativo do modelo r^2, também chamado coeficiente de explicação ou determinação, objetiva avaliar a qualidade do ajuste. Nada mais é do que o coeficiente de correlação de Pearson, já visto, ao quadrado. É dado por qualquer uma das expressões que seguem:

$$r^2 = \frac{\sigma_{XY}^4}{\sigma_X^2\,\sigma_Y^2} = b\frac{\sigma_{XY}^2}{\sigma_Y^2} = b^2\frac{\sigma_X^2}{\sigma_Y^2}$$

O valor do poder explicativo é sempre positivo e varia entre 0 e 1, ficando tanto mais próximo da unidade quanto melhor for a qualidade do ajuste, ou seja, mais a variável explicada Y depende da variável explicativa X e menos de outros fatores estranhos à equação de ajustamento.

7.3.3 EXEMPLO

Retomando o exemplo da primeira tabela, em que foi calculado o coeficiente de correlação de Pearson para os anos de estudos de pais e filhos, é possível encontrar a reta de regressão, através da determinação

dos parâmetros ***a*** e ***b***. Os dados necessários para isso podem ser buscados da tabela. Então, com os valores já encontrados anteriormente, pode-se calcular ***a*** e ***b***.

$$b = \frac{\sigma_{XY}^2}{\sigma_X^2} = \frac{4{,}522}{9{,}075} = 0{,}498$$

$$a = \overline{Y} - b\overline{X} = 9{,}429 - 0{,}498 \text{ x } 10{,}429 = 4{,}201$$

E a equação de regressão fica:

$$Y' = a + bX = 4{,}201 + 0{,}498X$$

E o poder explicativo é o próprio coeficiente de correlação de Pearson, já encontrado, elevado ao quadrado:

$$r^2 = 0{,}75^2 = 0{,}5625$$

Que significa que 56,25 % das variações de ***Y*** podem ser explicadas pela variável ***X***.

7.4 REGRESSÃO LINEAR POR TRANSFORMAÇÃO

Na maioria das vezes ***Y*** não varia linearmente com ***X***, casos em que ***Y = f (X)*** é uma função não linear. Algumas dessas funções, entretanto, podem ser tornadas lineares, mediante transformações. Outras, como as polinomiais, não podem ser linearizadas. Destas, as mais importantes são a parábolas de segundo e terceiro graus.

Embora não se pretenda aqui esgotar o assunto, as transformações referidas acima procuram possibilitar que se encontre a reta de regressão para a função transformada. Com isso, pode-se, após resolvidas as situações com a equação encontrada, voltar às variáveis originais através da transformação inversa.

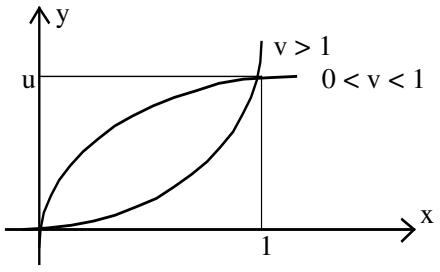
7.4.1 FUNÇÃO POTENCIAL

A função teórica é do tipo:

$y = u\,x^v$ onde:

$u > 0$ e

$v > 0$



Aplicando a transformação abaixo:

$$\log y = \log\left(u\,x^v\right) = \log u + \log x^v = \log u + v \log x$$

Que, comparada à função linear estudada:

$$Y = a + bX$$

Leva às relações abaixo, que permitem a reversão da forma original para a transformada e vice-versa. Assim todo o tratamento dado à função linear pode ser aproveitado por essa função linearizável.

$$Y = \log y \qquad X = \log x$$

$$a = \log u \qquad b = v$$

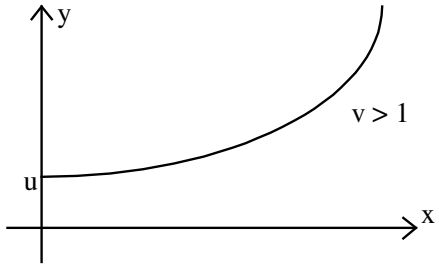
7.4.2 FUNÇÃO EXPONENCIAL

A função teórica é do tipo:

$y = u\,v^{x}$ onde:

$u > 0$ e

$v > 1$



Aplicando a transformação abaixo:

$$\log y = \log\left(u\,v^{x}\right) = \log u + \log v^{x} = \log u + (\log v)\,x$$

Que, comparada à função linear estudada:

$$Y = a + bX$$

Leva às relações abaixo, que permitem a reversão da forma original para a transformada e vice-versa. Assim todo o tratamento dado à função linear pode ser aproveitado por essa função linearizável.

$$Y = \log y \qquad X = x$$

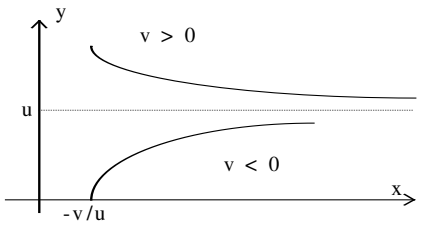
$$a = \log u \qquad b = \log v$$

7.4.3 FUNÇÃO HIPERBÓLICA

PRIMEIRO TIPO

A função teórica é do tipo:

$$y = u + \frac{v}{x}$$



Aplicando a transformação seguinte:

$$y = u + \frac{v}{x} = u + v\ x^{-1}$$

Que, comparada à função linear estudada:

$$Y = a + bX$$

Leva às relações abaixo, que permitem a reversão da forma original para a transformada e vice-versa. Assim todo o tratamento dado à função linear pode ser aproveitado por essa função linearizável.

$$Y = y \qquad X = \frac{1}{x} = x^{-1}$$

$$a = u \qquad b = v$$

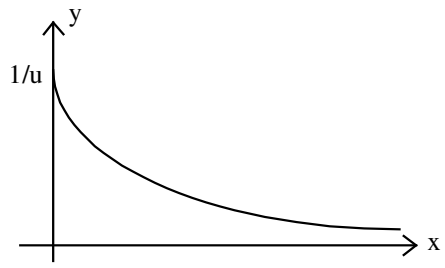
SEGUNDO TIPO

A função teórica é do tipo:

$y = \dfrac{1}{u + v\ x}$ onde

$u > 0$ e

$v > 0$



Aplicando a transformação abaixo:

$$\frac{1}{y} = u + v\ x$$

Que, comparada à função linear estudada:

$$Y = a + bX$$

Leva às relações abaixo, que permitem a reversão da forma original para a transformada e vice-versa. Assim todo o tratamento dado à função linear pode ser aproveitado por essa função linearizável.

$$Y = \frac{1}{y} = y^{-1} \qquad X = x$$

$$a = u \qquad b = v$$

7.5 EXERCÍCIOS

1. A tabela abaixo mostra os resultados de uma pesquisa com dez famílias de uma determinada região. A partir dela calcular o que é solicitado abaixo:

Famílias	Renda (nº de S.M.)	Poupança (nº de S.M.)	Número de filhos	Anos de estudo
A	10	4	8	3
B	15	7	6	4
C	12	5	5	5
D	70	20	1	12
E	80	20	2	16
F	100	30	2	18
G	20	8	3	8
H	30	8	2	8
I	10	3	6	4
J	60	15	1	8

   a) O coeficiente de correlação linear de Pearson, entre Renda Familiar e Poupança.

   b) O coeficiente de correlação linear de Pearson, entre Número de filhos e Poupança.

   c) O coeficiente de correlação linear de Pearson, entre Anos de Estudo e Número de Filhos.

   d) O coeficiente de correlação linear de Pearson, entre Anos de Estudo e Renda Familiar.

2. Dez alunos foram submetidos a uma prova teórica e outra prática. Eles foram classificados de acordo com seus aproveitamentos conforme a tabela abaixo. Encontre o coeficiente de correlação ordinal de Spearman e o coeficiente gama de Goodman e Kruskal.

Alunos	Prova Teórica	Prova Prática
A	3º	3º
B	1º	2º
C	8º	9º
D	6º	6º
E	7º	5º
F	2º	1º
G	9º	8º
H	10º	10º
I	5º	4º
J	4º	7º

3. Em uma escola foi aplicado, no início do ano letivo, um teste para medir o Q.I. (quociente de inteligência) de alguns alunos. No final do ano foi feito um exame para verificação de aproveitamento. Determine o coeficiente de correlação de Spearman entre os resultados dos testes e as médias obtidas, de acordo com a tabela abaixo. Encontre, também, o coeficiente gama de Goodman e Kruskal.

Alunos	Testes de Q.I.	Médias dos exames
A	110	9,0
B	120	8,0
C	125	9,0
D	130	9,5
E	130	8,5
F	140	9,5

4. Determine os seguintes coeficientes de correlação para os dados abaixo: linear de Pearson , ordinal de Spearman e gama de Goodman e Kruskal.

Estudante	A	B	C	D	E
Altura (cm)	183	175	168	178	173
Peso (kg)	77	74	68	82	84

5. Uma empresa classificou seus vinte vendedores segundo dois critérios distintos: conforme a avaliação da chefia e conforme o nível de instrução. Qual o coeficiente de correlação de Spearman e o coeficiente gama de Goodman e Kruskal.

Vendedores	Avaliação da chefia	Nível de instrução
A	1º	1º
B	2º	6º
C	3º	7º
D	4º	9º
E	5º	2º
F	6º	10º
G	7º	3º
H	8º	5º
I	9º	15º
J	10º	8º
L	11º	4º
M	12º	14º
N	13º	17º
O	14º	18º
P	15º	16º
Q	16º	12º
R	17º	13º
S	18º	11º
T	19º	19º
U	20º	20º

6. Cinco sujeitos foram avaliados relativamente a duas variáveis: ***X*** e ***Y***. Em ambas receberam valores inteiros de um a cinco, que podem ser entendidos como postos. Calcule o coeficiente de correlação ordinal de Spearman e o coeficiente gama de Goodman e Kruskal.

Estudante	A	B	C	D	E
X	2	1	3	5	4
Y	3	2	1	5	4

7. Em uma turma de sete alunos procura-se correlacionar a freqüência às aulas e as notas obtidas pelos alunos. Encontre os coeficientes ordinal de Spearman e gama de Goodman e Kruskal.

Alunos	A	B	C	D	E	F	G	H	I	J
Notas	9	5	6	7	8	7	8	8	9	4
Faltas	1	6	7	3	2	5	4	3	2	7

8 PROBABILIDADE

8.1 INTRODUÇÃO À PROBABILIDADE

8.1.1 INTRODUÇÃO

CONCEITO

Probabilidade é o estudo dos fenômenos aleatórios. Por fenômenos aleatórios, ou randômicos, ou ainda casuais, entende-se aqueles que são imprevisíveis, embora uma vez repetidos lhes seja possível prever regras de formação.

A probabilidade de ocorrência de um evento é igual ao número de possibilidades de acontecer tal evento dividido pelo número total de resultados possíveis.

NOÇÃO INTUITIVA

Costuma-se introduzir o aluno no estudo das probabilidades mediante o uso de moedas, dados e cartas, porque nesses jogos os resultados são bastante elucidativos. Quando se joga uma moeda para cima, sabe-se que tanto pode dar cara (doravante simbolizada por K) como coroa (C), mas não se pode prever de que lado cairá. As leis da física permitiriam efetuar esse cálculo a partir das forças e outras grandezas envolvidas, mas seria inviável e absurdamente complicado fazer tal previsão. Por aleatório não se entende, então, a impossibilidade de previsão, necessariamente. Por outro lado, se alguém ficar uma tarde inteira jogando uma moeda concluirá obviamente que mais ou menos 50% das vezes dará cara e outras 50% coroa. Há, portanto, uma regra de formação simples: a probabilidade de C ou K é 0,5. Usando o mesmo raciocínio, pode-se afirmar que a probabilidade de um dado cair no número 4 é de 1/6, bem como de retirar um ás de copas de um baralho é 1/52. Essas conclusões foram possíveis por serem admitidos, dado, moeda e baralho "honestos", o que em estatística significa eqüiprováveis, não viciados, sem nenhuma conotação moral.

8.1.2 CÁLCULO DE PROBABILIDADES

COM REPOSIÇÃO

A idéia de reposição presume que o elemento sorteado é recolocado na amostra antes de novo sorteio.
Se uma moeda for jogada duas vezes, existem quatro possibilidades diferentes de resultado: KK (duas caras), CK (coroa e depois cara}, KC (cara e depois coroa) e CC (duas coroas). Portanto, a probabilidade de dar duas caras é 0,25 (25%), duas coroas também 0,25 (25%) e uma de cada 0,50 (50%). A probabilidades de, jogando um dado duas vezes consecutivas, obter duas vezes o número 4 é (1/6)x(1/6) = 1/36.

Nestes e nos casos seguintes, os eventos são independentes e mutuamente excludentes, o que significa, respectivamente, que um não influi no outro e que se ocorre um não ocorre o outro.

Admitindo que o evento desejado, quando uma moeda é jogada diversas vezes, seja a obtenção de cara e esgotando os resultados possíveis chega-se às tabelas abaixo:

3 VEZES

Evento de interesse (caras)	*Número de resultados favoráveis*	*Probabilidade do evento*	*Resultados que levam ao evento desejado*		
0	1	1/8	CCC		
1	3	3/8	KCC	CKC	CCK
2	3	3/8	KKC	KCK	CKK
3	1	1/8	KKK		
		8/8 = 1			

4 VEZES

Evento de interesse (caras)	*Número de resultados favoráveis*	*Probabilidade do evento*	*Resultados que levam ao evento desejado*					
0	1	1/16	CC CC					
1	4	4/16	CC CK	CC KC	CK CC	KC CC		
2	6	6/16	CC KK	CK CK	CK KC	KC CK	KC KC	KK CC
3	4	4/16	CK KK	KC KK	KK CK	KK KC		
4	1	1/16	KK KK					
		16/16 = 1						

SEM REPOSIÇÃO

Este caso acontece quando, após um sorteio, a amostra fica desfalcada dos elementos retirados, já que eles não são repostos. Assim a probabilidade individual de um evento modifica de um sorteio para o seguinte. Se num baralho, por exemplo, procura-se retirar um ás de copas e após a primeira tentativa fracassada a carta sorteada não for recolocada no baralho, a probabilidade na segunda tentativa aumenta de 1/52 para 1/51.

8.2 DISTRIBUIÇÃO DE POISSON

A distribuição discreta de probabilidade descoberta por Poisson dada pela expressão:

$$P(x) = \frac{\lambda^{x} e^{-\lambda}}{x!} \qquad (x = 0, 1, 2, 3 \ldots)$$

Em que ***P(x)*** é a probabilidade do evento ocorrer ***x*** vezes, λ é um parâmetro e ***e*** é uma constante que vale aproximadamente 2,71828...

8.3 DISTRIBUIÇÃO BINOMIAL

8.3.1 INTRODUÇÃO

Nos exemplos de lançamentos de moedas repetidas, normalmente não interessa a ordem das caras ou coroas, mas a quantidade de vezes que se obtém uma ou outra. Assim, para os casos de 2, 3 e 4 lançamentos pode-se escrever os possíveis resultados na forma potencial:

$n = 2$	$K^2 + 2CK + C^2$
$n = 3$	$K^3 + 3CK^2 + 3C^2K + C^3$
$n = 4$	$K^4 + 4CK^3 + 6C^2K^2 + 4C^3K + C^4$

Que assume a forma do Binômio de Newton $(K + C)^n$, donde a forma de distribuição recebeu a denominação "binomial".

8.3.2 FORMA GERAL

Atribuindo ***p*** à probabilidade de sucesso de determinado evento e ***q*** à probabilidade de fracasso deste mesmo evento, a forma binomial abaixo é válida para ***n*** repetições da experiência. O resultado é obviamente ***1***, ou seja abrange 100% dos casos, pois se os eventos são mutuamente excludentes e complementares - ou ele ocorre ou não ocorre em cada uma das repetições. Quando se joga um dado, por exemplo, ou ele cai no número 4 (***p = 1/6***) ou em qualquer dos outros (***q = 5/6***). A soma destas duas probabilidades dá ***1***, que elevado a qualquer potência, em função do número ***n*** de repetições da experiência, dará ***1***.

A forma geral do Binômio de Newton é então:

$$(p + q)^n$$

Onde, ***n*** é o numero de repetições, ***p*** é a probabilidade de sucesso e ***q*** a probabilidade de fracasso, para cada evento, individualmente.

Para um experimento repetido 5 vezes por exemplo o binômio assume a forma:

$$(p+q)^5 = p^5 + 5qp^4 + 10q^2p^5 + 10q^3p^2 + 5q^4p + q^5$$

Que, quando for o caso de uma moeda. p = 0,5 e q = 0,5. Associando, então, à idéia de sucesso, a obtenção de x caras em cinco lançamentos (n = 5) tem-se:

$$P(x=5) = p^5 = 0{,}5^5 = 0{,}03125 \rightarrow (3{,}125\%)$$
$$P(x=4) = 5\,qp^4 = 5\,(0{,}5)(0{,}5)^4 = 0{,}15625 \rightarrow (15{,}625\%)$$
$$P(x=3) = 10\,q^2 p^3 = 10(0{,}5)^2 (0{,}5)^3 = 0{,}3125 \rightarrow (31{,}25\%)$$
$$P(x=2) = 10\,q^3 p^2 = 10(0{,}5)^3 (0{,}5)^2 = 0{,}3125 \rightarrow (31{,}25\%)$$
$$P(x=1) = 5\,q^4 p = 5(0{,}5)^4 (0{,}5) = 0{,}15625 \rightarrow (15{,}625\%)$$
$$P(x=0) = q^5 = 0{,}5^5 = 0{,}03125 \rightarrow (3{,}125\%)$$

8.3.3 TERMO GERAL

A medida em que n cresce, o desenvolvimento do binômio torna-se complicado. Por esse motivo, a fórmula do termo geral que permite o cálculo de seus coeficientes, necessários ao cálculo da probabilidade de x ocorrências do evento sucesso, fica simplificada:

$$C_n^x = \frac{n!}{x!(n-x)!}$$

E a probabilidade de x sucessos em n tentativas fica:

$$P_n^x = C_n^x\, q^{n-x}\, p^x = \frac{n!}{(n-x)!}\, q^{n-x} p^x$$

No exemplo anterior, para três caras em cinco lançamentos da moeda, tem-se:

$$C_5^3 = \frac{5!}{3!(5-3)!} = \frac{5\text{x}4\text{x}3\text{x}2\text{x}1}{(3\text{x}2\text{x}1)(2\text{x}1)} = \frac{120}{6\text{x}2} = 10$$

$$P_5^3 = C_5^3\, q^{5-3} p^3 = 10(0{,}5)^2 (0{,}5)^3 = 0{,}3125$$

A probabilidade de se conseguir quatro vezes o número quatro em sete lançamentos de um dado é, pois:

$$C_7^4 = \frac{7!}{4!(7-4)!} = \frac{7\text{x}6\text{x}5\text{x}4\text{x}3\text{x}2\text{x}1}{(4\text{x}3\text{x}2\text{x}1)(3\text{x}2\text{x}1)} = \frac{5.040}{24\text{x}6} = 35$$
$$P_7^4 = C_7^4\, q^{7-4} p^4 = 35\left(\frac{5}{6}\right)^3 \left(\frac{1}{6}\right)^4 = 0{,}0156 \rightarrow (1{,}56\%)$$

8.3.4 TABELAS

Existem tabelas que dão diretamente as probabilidades de uma distribuição binomial, a partir do número de eventos ***n***, de sucessos ***x*** e da probabilidade de sucesso ***p*** de cada evento individual. Na tabela ***B (n ; p)***, normalmente uma para cada ***n***, seleciona-se a probabilidade procurada a partir do valor ***x*** da coluna e de ***p*** na linha.

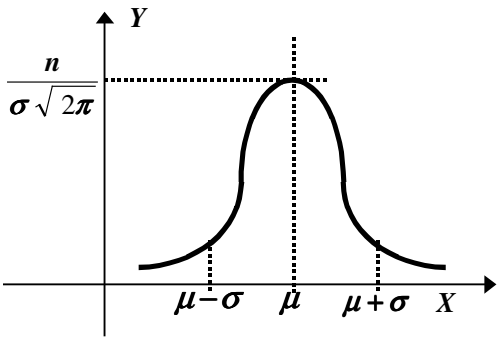
8.4 DISTRIBUIÇÃO NORMAL

8.4.1 DEFINIÇÃO

No estudo das distribuições, viu-se que os histogramas aproximam-se da forma de um sino. A uma distribuição teórica simétrica, que segue uma determinada equação, deu-se o nome de distribuição normal. A curva normal e sua definição matemática encontram- se abaixo:

$$Y = \frac{n}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2} \qquad \text{onde:}$$

$\pi = 3{,}1416$ e $e = 2{,}7183$



A curva normal caracteriza-se por ser unimodal, cujo valor μ é o mesmo da mediana e da média aritmética. Além disso, é assintótica, jamais tocando o eixo ***X***.

8.4.2 IMPORTÂNCIA

A importância da distribuição normal decorre de sua utilização intensiva na pesquisa estatística. O estudo das probabilidades em distribuições reais consagrou a Curva Normal. Ela auxilia, como será visto, a concluir a respeito de populações, partindo de resultados obtidos com amostras.

Muitas distribuições concretas são do tipo normal, como as alturas ou os quocientes de inteligência das pessoas. Noutras situações, as curvas reais não são normalizadas porque são assimétricas - distribuição de renda, por exemplo. A enorme utilidade da distribuição normal, decorre

do fato que as médias de quaisquer distribuições, mesmo as não normais, apresentam uma distribuição normal.

8.5 RELAÇÕES ENTRE DISTRIBUIÇÕES

8.5.1 PRINCIPAIS PARÂMETROS

Os principais parâmetros das três distribuições vistas, média aritmética, variância e desvio padrão, bem como coeficientes de assimetria e curtose são apresentados em uma tabela a seguir:

	NORMAL	*POISSON*	*BINOMIAL*
Média aritmética	μ	$\mu = \lambda$	$\mu = np$
Variância	σ^2	$\sigma^2 = \lambda$	$\sigma^2 = npq$
Desvio padrão	σ	$\sigma = \sqrt{\lambda}$	$\sigma = \sqrt{npq}$
Coeficiente de momento de assimetria	$e_M = 0$	$e_M = \frac{1}{\sqrt{\lambda}}$	$e_M = \frac{q-p}{\sqrt{npq}}$
Coeficiente de momento de curtose	$b_2 = 3$	$b_2 = 3 + \frac{1}{\lambda}$	$b_2 = 3 + \frac{1-6pq}{npq}$
Desvio médio	$\bar{d} = \frac{\sigma}{\sqrt{2/\pi}} = 0{,}7979\sigma$		

8.5.2 APROXIMAÇÃO BINOMIAL POR POISSON

Para eventos raros, em que $p \cong 0$ e $q \cong 1$, pode ser usada a distribuição de Poisson, que é de cálculo mais fácil, em lugar da distribuição binomial.

Para efeitos práticos, o evento pode ser considerado raro quando:

$$n \geq 50 \qquad \text{e} \qquad np < 5$$

E a distribuição binomial pode ser tratada como uma distribuição de Poisson, cujo parâmetro λ vale:

$$\lambda = np$$

8.5.3 APROXIMAÇÃO POISSON POR NORMAL

Quando λ é muito alto a distribuição de Poisson pode ser substituída por uma distribuição normal com média aritmética e variância iguais a λ:

$$\mu = \sigma^2 = \lambda$$

8.5.4 APROXIMAÇÃO BINOMIAL POR NORMAL

Se na distribuição binomial, a seguintes condições forem satisfeitas:

$$np \geq 5 \qquad \text{e} \qquad nq \geq 5$$

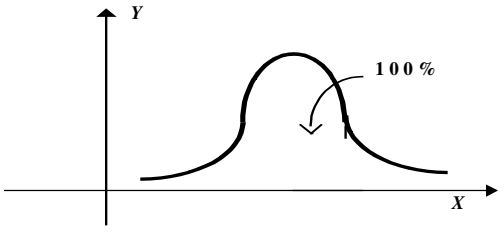
Ela pode ser tratada como uma distribuição normal com as seguintes média aritmética e variância:

$$\mu = np \qquad \text{e} \qquad \sigma^2 = npq$$

8.6 CURVA NORMAL E PROBABILIDADE

8.6.1 ÁREA SOB A CURVA NORMAL

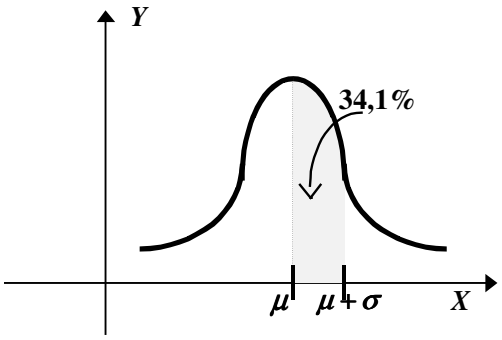
A área situada entre a curva normal e o eixo ***X*** corresponde à totalidade dos dados. Por isso, ela encerra 100% dos casos. A probabilidade de um dado estar situado sob esta curva é, pois, igual a ***1***, já que este evento é certo. A probabilidade é definida como a razão entre o número de vezes que um dado ou evento pode ocorrer e o número total de dados ou eventos.



Vale notar que o ponto onde se encontra o pico divide a distribuição em duas partes com áreas iguais. Até ali encontram-se 50% dos casos e dali para frente os outros 50%. Assim, a probabilidade até o ponto de máximo é 0,5 e dele em diante também 0,5.

8.6.2 ACIMA DA MÉDIA

A área que inclui os dados entre a média e um desvio padrão é de 34,1% do total. A probabilidade, então, de um dado estar entre estes limites é de, usando uma taxa proporcional em lugar da taxa percentual, 0,341. A figura abaixo ilustra essas idéias:

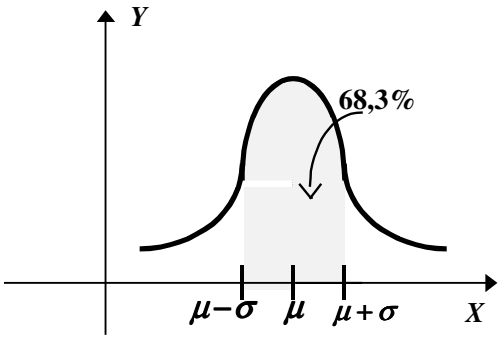


Fazendo as mesmas considerações entre dois e três desvios padrões, tem-se que:

Região entre a média e um número inteiro de desvios padrões:			
Número de desvios padrões	**1**	**2**	**3**
Percentual dos dados dentro da região	**34,1%**	**47,7%**	**49,9%**
Probabilidade do dado estar na região	**0,341**	**0,477**	**0,499**

8.6.3 EM TORNO DA MÉDIA

A área que inclui os dados em torno da média, afastados de no máximo um desvio padrão para cada lado, é de 68,3% do total. A probabilidade, então, de um dado estar entre estes limites é de, usando uma taxa unitária em lugar da taxa percentual, 0,683. A figura abaixo ilustra essas idéias:



Fazendo as mesmas considerações entre dois e três desvios padrões, tem-se que:

Região em torno da média, entre um número de desvios padrões:			
Número de desvios padrões	**1**	**2**	**3**
Percentual dos dados dentro da região	**68,3%**	**95,5%**	**99,7%**
Probabilidade do dado estar na região	**0,683**	**0,955**	**0,997**

8.7 CURVA NORMAL PADRONIZADA

Todas as conclusões até aqui referiram-se a desvios medidos em unidades inteiras de desvios padrões a partir da média. As distribuições reais apresentam suas médias e variâncias em unidades próprias ao fenômeno descrito em cada caso. Isso sugere a utilização de tabelas padronizadas de modo a permitir que qualquer distribuição normal possa ser transformada, achadas as percentagens ou probabilidades na tabela, para, finalmente, mediante uma transformação inversa, interpretar os resultados finais.

8.7.1 USO DA TABELA

A tabela considera uma distribuição normal específica, com os seguintes valores para a média aritmética e o desvio padrão:

$$\mu = 0 \qquad \text{e} \qquad \sigma = 1$$

A variável *z*, que é o desvio padrão dessa distribuição, é a entrada da tabela, cujo valor tabelado dá o percentual de dados entre a média e aquele desvio padrão. A partir daí, somas e diferenças permitem o cálculo das probabilidades em qualquer situação.

8.7.2 ESCORE Z

Para que a tabela possa ser utilizada, é necessário transformar os dados da distribuição em escores padronizados. Para isso, parte-se da média aritmética $\bar{x}$ e do desvio padrão σ da distribuição em questão. Então, define-se o escore *z* como:

$$z = \frac{x - \bar{x}}{\sigma} = \frac{d}{\sigma}$$

Onde:

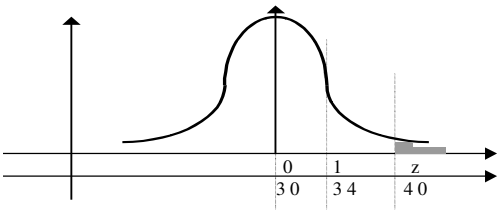
$z \Rightarrow$ escore padronizado,
$x \Rightarrow$ valor da variável a padronizar,
$\bar{x} \Rightarrow$ valor médio d distribuição,
$\sigma \Rightarrow$ desvio padrão da população e
$d \Rightarrow$ desvio, discrepância ou escore diferença

8.7.3 EXEMPLO

Qual a probabilidade que a variável de uma distribuição normal supere 40, se sua média aritmética vale 30 e sua variância 16? Esta distribuição é simbolizada por N(30;16).

A figura esboça o que se pretende, ou seja, achar a probabilidade de que a variável supere 40. Da média para a esquerda estão metade dos dados da distribuição (50%). Da metade direita devem ser extraídos os inferiores a 40. O percentual superior a 40 (à direita da marca) será os 50% menos o valor obtido da tabela, que dá o percentual entre a média e do ponto desejado. Assim:

$$P(x \geq 40) = 0{,}5000 - P(30 \leq x < 40)$$



Achando o escore z para $\boldsymbol{x} = 40$:

$$z = \frac{\boldsymbol{x} - \bar{\boldsymbol{x}}}{\boldsymbol{\sigma}} = \frac{\boldsymbol{x} - \bar{\boldsymbol{x}}}{\sqrt{\boldsymbol{\sigma}^2}} = \frac{40 - 30}{\sqrt{16}} = 2{,}5$$

Indo-se à tabela, encontra-se para z = 2,5, 49,38% que, em termos probabilísticos, dividindo-se por cem, dá 0,4938. Então:

$$\boldsymbol{P}(z \geq 2{,}5) = \boldsymbol{P}(0 \leq z \leq \infty) - \boldsymbol{P}(0 \leq z \leq 2{,}5)$$
$$\boldsymbol{P}(z \geq 2{,}5) = 0{,}5000 - 0{,}4938 = 0{,}0062$$

Ou seja, há 0,62% de possibilidades de que um escore seja maior ou igual a 40.

8.8 EXERCÍCIOS

1. Dado um conjunto de escores brutos com distribuição normal, onde μ = 7,5 e σ = 1,3, exprimir cada um dos seguintes escores brutos em termos de escala *z*:

   a) x = 8,0
   b) x = 6,0
   c) x = 5,3
   d) x = 10,2
   e) x = 7,5
   f) x = 8,5
   g) x = 11,0

2. Seja uma distribuição de ganhos semanais, cuja média é μ = R$ 178,50 e o desvio padrão é σ = R$ 47,60. Supondo-se distribuição normal, como podem ser expressas, em escala *z*, os seguintes salários semanais:

   a) x = R$ 142,80
   b) x = R$ 187,00
   c) x = R$ 224,40
   d) x = R$ 85,00
   e) x = R$ 255,00
   f) x = R$ 195,50
   g) x = R$ 153,00

3. Para a mesma distribuição do problema anterior, determinar:

   a) A percentagem de respondentes à pesquisa que tem ganhos semanais igual ou maior que R$ 255,00.
   b) A probabilidade de sortear um respondente cuja renda semanal seja igual ou superior a R$ 255,00.
   c) A percentagem de respondentes à pesquisa que tem ganhos semanais entre R$ 170,00 e R$ 178,50.
   d) A probabilidade de sortear um respondente cuja renda semanal esteja entre R$ 170,00 e R$ 178,50.
   e) A probabilidade de sortear um respondente cuja renda semanal seja inferior ou igual a R$ 170,00.
   f) A probabilidade de sortear um respondente cuja renda semanal seja inferior a R$ 170,00 ou superior a R$ 187,00.
   g) A probabilidade de sortear dois respondentes e ambos apresentarem renda inferior a R$ 170,00

4. Dado um conjunto de escores brutos, distribuídos normalmente, onde μ = 80 e σ = 7,5, determinar:

   a) A percentagem de escores inferiores ou iguais a 60.
   b) A probabilidade de encontrar um escore inferior ou igual a 60.
   c) A probabilidade de um escore ficar entre 80 e 90.
   d) A percentagem de escores maiores ou iguais a 85.
   e) A probabilidade de encontrar escores menores ou iguais a 70 ou maiores ou iguais a 90.
   f) A probabilidade de sortear três escores, todos maiores ou iguais a 90.

5. Para as distribuições abaixo, do tipo normal, em que são fornecidas a média e a variância – por isso a forma $N(\mu; \sigma^2)$ – encontrar o solicitado.

   a) x é N (20 ; 49) Calcular $P(x < 30)$.
   b) x é N (10 ; 100) Calcular $P(12 \leq x \leq 20)$.
   c) x é N (30 ; 16) Calcular $P(x \leq 19)$.
   d) x é N (20 ; 25) Calcular $P(x \leq 30)$.
   e) x é N (50 ; 81) Calcular $P(40 \leq x \leq 60)$.
   f) x é N (10 ; 16) Calcular $P(5 \leq x)$.
   g) x é N (10 ; 25) Calcular $P(x < 5)$.
   h) x é N (10 ; 25) Calcular $P(5 \leq x \leq 15)$.

6. Uma moeda eqüiprovável vai ser jogada 18 vezes. Calcular, sem recorrer às tabelas, a probabilidade do número de caras, número este representado por x, nos seguintes casos:

   a) $P(x = 12)$.
   b) $P(8 < x < 10)$.
   c) $P(x > 12)$.
   d) $P(3 \leq x < 8)$.

7. Um dado eqüiprovável vai ser jogado 9 vezes. . Calcular, sem recorrer às tabelas, as probabilidades nos seguintes casos, em que x significa o dado cair com a face 4 para cima:

   a) $P(x < 3)$.
   b) $P(2 < x \leq 6)$.
   c) $P(x \geq 8)$.
   d) $P(9 \geq x <. 1)$.

8. Para o desenvolvimento de $(q + p)^{13}$?

   a) Quanto vale o sétimo termo?
   b) Se $q = 0{,}28$, quanto vale a probabilidade expressa pelo sétimo termo?

9. Uma prova tem 40 questões, cada uma com quatro alternativas de resposta, sendo apenas uma correta. Qual o número mais provável de acertos se os alunos chutarem as respostas?

10. Uma moeda honesta foi lançada 100 vezes. Qual a probabilidade de terem saído exatamente 64 caras?

11. Uma moeda honesta foi lançada 100 vezes. Qual a probabilidade de terem saído 80 caras ou mais?

12. Uma moeda honesta foi lançada 60 vezes. Se x for o número de coroas, qual a probabilidade de ($20 < x < 40$)?

13. Sabe-se que a variável x tem distribuição binomial com os parâmetros $n = 40$ e $p = 0{,}6$. Calcular a média aritmética μ e a variância σ^2 dessa distribuição.

9 AMOSTRAGEM E ESTIMAÇÃO

9.1 AMOSTRAS E POPULAÇÕES

O interesse do pesquisador no conhecimento das características de uma população leva a determinar seus parâmetros, normalmente quantificados através de médias, medianas, modas, desvios, variâncias, coeficientes, etc...Ocorre, entretanto, que nem sempre toda a população ou universo está disponível para ser inquirido, ou porque é impraticável sob o ponto de vista econômico e de dispêndio de tempo, ou porque o próprio experimento destrói o elemento - testes de qualidade em palitos de fósforo. A solução consiste em examinar apenas parcelas do universo e procurar tirar conclusões sobre toda a população a partir dessas amostras.

A utilização adequada de técnicas de amostragem e os cuidados a serem tomados ao se realizar inferências é objeto do estudo da estimação e da amostragem. Embora não se pretenda esgotar esse tema, pretende-se introduzi-lo de modo a permitir maior aprofundamento se for desejado.

9.2 TIPOS DE AMOSTRAGEM

O objetivo de qualquer método de amostragem é que a amostra retirada para estudo seja a mais representativa possível. Existem dois grandes grupos: um que os dados são tomados ao acaso, de modo randômico, e outro em que existe alguma forma de interveniência na sua seleção.

9.2.1 AMOSTRAGEM NÃO ALEATÓRIA

ACIDENTAL

É o métodos em que o pesquisador seleciona os dados convenientes, afastando os inconvenientes. Assemelha-se em muito ao procedimentos utilizados no dia a dia.

POR QUOTAS

Neste método, as amostras são tomadas na proporção em que certas características encontram-se na população. A amostra guarda relação proporcional com a população em certos atributos selecionados - sexo, idade ,renda, etc..

POR CONVENIÊNCIA

Consiste em escolher uma amostra que, presumivelmente, é representativa do universo. Um bom exemplo é optar por grupos sociais que, historicamente, melhor representam a média da população.

9.2.2 AMOSTRAGEM ALEATÓRIA

A característica fundamental de uma amostragem aleatória, randômica ou casual é a de que todos os elementos da população têm a mesma probabilidade de serem selecionados para a amostra.

CASUAL SIMPLES

Na amostra casual simples todos os elementos da população são numerados e selecionados por um processo totalmente aleatório. Funciona como uma espécie de sorteio de loteria. Normalmente, os pesquisadores lançam mão de tábuas de números aleatórios para garantir a inexistência de qualquer lei de formação na escolha da amostra.

SISTEMÁTICA

Nesta variante de amostragem casual, em lugar de serem atribuídos números que depois permitem a seleção através de tábuas, cria-se uma lei de formação que, em princípio, não tenha qualquer relação com as variáveis em estudo. Isso facilita porque torna desnecessário vincular números aos elementos da população, bastando retirá-los de quinze em quinze, ou vinte em vinte, por exemplo.

ESTRATIFICADA

A população é dividida por extratos, dentro dos quais são utilizados os métodos aleatórios. Estes extratos reduzem a variância interna, permitindo o uso de amostras menores. Embora assemelhe-se ao método por quotas, difere na essência, porque naquele não são aplicadas as técnicas que garantam a casualidade da amostra.

POR CONGLOMERADOS

De certo modo, é parecido com o método não casual por conveniência visto acima. Só que aqui é um conglomerado, do qual serão extraídos os elementos randomicamente, que será escolhido por julgamento. É de extrema importância na redução dos custos de uma pesquisa.

9.3 DISTRIBUIÇÃO AMOSTRAL DAS MÉDIAS

9.3.1 ERRO AMOSTRAL

Ao se tentar, por exemplo, determinar a altura média da população de uma cidade, escolhe-se uma amostra cuja média não terá exatamente a mesma média da população. Isso ocorre não por falha da metodologia ou de má fé do pesquisador. Isso é inerente ao processo estatístico e é denominado “erro amostral”. Assim, estimar um parâmetro da população a partir de uma amostra, embora seja o objeto da inferência estatística, não garante a confiabilidade. Pode-se, contudo, estabelecer intervalos de confiança, dentro dos quais os resultados merecem credibilidade. A ciência, por vezes, faz do conhecimento do índice de incerteza uma preciosa informação.

Quando se tratar da população, os símbolos utilizados para a média aritmética e para o desvio padrão são, normalmente, μ e σ, respectivamente. Quando esses parâmetros referirem-se a amostras usam-se $\bar{x}$ e s.

9.3.2 TEOREMA DO LIMITE CENTRAL

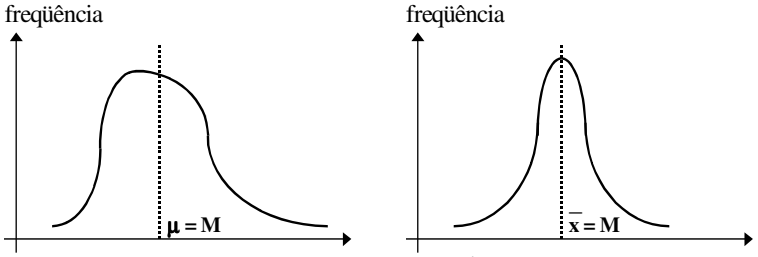
Se no exemplo anterior, em que se pretendia encontrar a média de altura dos habitantes de uma cidade a partir de uma amostra, forem realizadas não uma, mas um grande número de amostras, de tal modo que se possa fazer uma distribuição das médias dessas amostras, chega-se a conclusões bastante interessantes.

A primeira conclusão é de que essa distribuição de médias é uma distribuição do tipo normal. O mais impressionante é que isso é válido mesmo que as distribuições originais, das quais extraíram-se as médias para formar essa distribuição de médias, não sejam normais. Isso constitui o chamado "Teorema do Limite Central". Então, tomada qualquer distribuição, se dela forem extraídas um bom número de amostras, a distribuição amostral de médias é normal.

Além disso, pode-se tomar a média das médias, ou seja, a média dessa distribuição de médias amostrais, como a efetiva média da população.

Finalmente, a variância dessa distribuição de médias amostrais é menor que a variância da população. Esse fato é compreensível, pois, ao serem calculadas as médias de cada amostra, atenuam-se os valores discrepantes.

As duas figuras apresentadas a seguir mostram as distribuições da população e das médias amostrais que dela foram retiradas. A população não segue uma curva normal, a média das duas distribuições é a mesma e o desvio padrão da população é nitidamente superior ao da distribuição amostral das médias.



POPULAÇÃO MÉDIAS AMOSTRAIS

9.4 ESTIMATIVA DE MÉDIAS

Voltando os olhos para a distribuição das médias amostrais que, como foi visto, tem uma curva normal, duas questões podem ser colocadas. A primeira: conhecida a média da população, qual a probabilidade da média de uma amostra se afastar mais ou menos daquele valor? E a segunda: se não for conhecida a média populacional, o quanto pode-se confiar na média obtida a partir de uma amostra? Esta seção pretende trazer um pouco de luz sobre esse assunto.

9.4.1 DISTRIBUIÇÃO NORMAL DAS MÉDIAS

Como a distribuição amostral assume a forma normal, é possível usar o escore z e a tabela para calcular a probabilidade de obter qualquer média amostral. Para isso, é necessário que se conheça a média das médias, que sabe-se ser igual à média da população, e o desvio padrão da distribuição amostral. Encontra-se o escore z da seguinte forma:

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}$$

Onde:

$\bar{x}$ $\Rightarrow$ Média de uma amostra,
μ $\Rightarrow$ Média da população,
$\sigma_{\bar{x}}$ $\Rightarrow$ Desvio padrão da distribuição das médias amostrais.

Com o valor de z vai-se à tabela e acha-se a probabilidade da média da amostra $\bar{x}$ ter esta discrepância da média populacional. Isso sugere, pois, um método interessante para testar médias informadas, mediante o cálculo da média de uma amostra.

9.4.2 ERRO PADRÃO DA MÉDIA

Na prática, a distribuição das médias amostrais não está disponível. Isso eqüivale a dizer que não se conhece a média das médias e, por conseguinte, não se sabe qual a média populacional. Além disso, a variância da distribuição amostral também é desconhecida. Então, a menos que se tome um número muito grande de amostras fica difícil saber a média da população e os cálculos de probabilidades para qualquer média obtida, como foi visto acima.

Normalmente, se toma uma ou outra amostra e a partir daí, quer-se inferir acerca da população. Pode-se, felizmente, estimar o desvio padrão da distribuição amostral a partir de uma única amostra. Essa estimativa chama-se "erro padrão da média" e é calculada do modo a baixo:

$$s_{\bar{x}} = \frac{\sigma}{\sqrt{n-1}} = \frac{s}{\sqrt{n}}$$

Onde:

$s_{\bar{x}}$ $\Rightarrow$ Erro padrão da média (estimativa do desvio padrão da distribuição amostral das médias).

s $\Rightarrow$ Desvio padrão amostral.
σ $\Rightarrow$ Desvio padrão populacional.
n $\Rightarrow$ Tamanho da amostra.

É compreensível que o erro padrão da média, que pode ser usado como estimativa do desvio padrão da distribuição das médias amostrais, seja tão menor que o desvio padrão da amostra, pois os valores discrepantes da média encontrados em uma amostra são amortecidos numa distribuição de médias.

9.4.3 INTERVALOS DE CONFIANÇA

DEFINIÇÃO

Quando se encontra a média aritmética $\bar{x}$ e a variância de uma amostra extraída de uma determinada população, pode-se estimar um intervalo dentro do qual está a verdadeira média populacional μ. Além disso, é possível estabelecer a probabilidade de que este valor esteja dentro do , assim denominado, "intervalo de confiança".

A partir do cálculo da média $\bar{x}$ e do erro padrão da média $s_{\bar{x}}$ da amostra, pode-se garantir que a média da população μ encontra-se entre os limites:

$$\mu = \bar{x} \pm z \, . \, s_{\bar{x}}$$

Em que z é o escore padronizado para o percentual de confiança desejado, obtido na tabela da distribuição normal padronizada. Alguns intervalos de confiança são mais naturais que os de 68,3%, 95,4% e 99,7%, correspondentes a um, dois ou três desvios padrões.

Os intervalos de confiança usuais estão tabelados abaixo, sendo que os mais importantes foram postos em negrito. Adotado um percentual de confiança desejado, retira-se z na tabela abaixo e substitui-se na fórmula acima.

Intervalo de confiança em %	80	85	90	**95**	**99**
Valor de z	1,28	1,44	1,65	**1,96**	**2,58**

Pode-se depreender da tabela que quanto maior for a confiança desejada na informação menor a precisão. Para a confiança de 95% - a mais adotada - a fórmula acima fica:

$$\mu = \bar{x} \pm 1{,}96 . s_{\bar{x}}$$

EXEMPLO

Pretende-se inferir a respeito do quociente de inteligência médio (QI) dos alunos de uma Universidade. Para isso, colhe-se uma amostra aleatória de 10 alunos. Com os dados obtidos abaixo, quais os valores possíveis para os QIs médios dos estudantes da Universidade, com intervalos de confiança de 95% e 99%?

QIs: 80, 120, 100, 110, 130, 90, 120, 110, 90, 110.

A partir dos valores dos QIs, constrói-se a tabela abaixo com as colunas necessárias para o cálculo da média aritmética e variância da amostra.

Ordem	*x (QI)*	x^2
1	80	6.400
2	120	14.400
3	100	10.000
4	110	12.100
5	130	16.900
6	90	8.100
7	120	14.400
8	110	12.100
9	90	8.100
10	110	12.100
$\boldsymbol{n}$ = 10	$\boldsymbol{\Sigma x}$ = 1.060	$\boldsymbol{\Sigma x^2}$ = 114.600

Da tabela chega-se a:

$$\bar{\boldsymbol{x}} = \frac{\sum \boldsymbol{x}}{\boldsymbol{n}} = \frac{1.060}{10} = 106$$

$$\boldsymbol{x_q^2} = \overline{\boldsymbol{x^2}} = \frac{\sum \boldsymbol{x^2}}{\boldsymbol{n}} = \frac{114.600}{10} = 11.460$$

A variância populacional é:

$$\boldsymbol{\sigma}^2 = \overline{\boldsymbol{x^2}} - \bar{\boldsymbol{x}}^2 = 11.460 - 106^2 = 224$$

E o desvio padrão populacional é:

$$\boldsymbol{\sigma} = \sqrt{\boldsymbol{\sigma}^2} = \sqrt{\overline{\boldsymbol{x^2}} - \bar{\boldsymbol{x}}^2} = \sqrt{224} = 14{,}97$$

Que leva ao erro padrão da média:

$$\boldsymbol{s_{\bar{x}}} = \frac{\boldsymbol{s}}{\sqrt{\boldsymbol{n-1}}} = \frac{14{,}97}{\sqrt{10-1}} = 4{,}99$$

Que permite prever uma percentagem de 95% de que a média da população esteja no intervalo de:

$$\boldsymbol{\mu} = \bar{\boldsymbol{x}} \pm \boldsymbol{1{,}96}\ \boldsymbol{s_{\bar{x}}} = 106 \pm 1{,}96 \text{ x } 4{,}99 = 106 \pm 9{,}78$$

Ou seja:

$$96{,}22 < \boldsymbol{\mu} < 115{,}78$$

Para um intervalo de confiança de 99%, a média dos QIs dos universitários apresentaria uma menor precisão:

$$\boldsymbol{\mu} = \bar{\boldsymbol{x}} \pm \boldsymbol{2{,}58}\ \boldsymbol{s_{\bar{x}}} = 106 \pm 2{,}58 \text{ x } 4{,}99 = 106 \pm 12{,}87$$

E o intervalo de:

$$93{,}13 < \boldsymbol{\mu} < 118{,}87$$

9.5 ESTIMATIVA DE PROPORÇÕES

9.5.1 DEFINIÇÃO

Quando o que se pretende é descobrir a proporção ***P*** de determinada ocorrência em uma população, acha-se a proporção ***p*** em que ela ocorre numa amostra de ***n*** elementos. A partir daí, estima-se um intervalo para a proporção populacional com determinada probabilidade de confiança. A proporção da não ocorrência, ou seja ***q*** ,será dada por:

$$q = 1 - p$$

Assim como no caso das médias foi preciso encontrar o erro padrão da média, também aqui é possível estimar o "erro padrão da proporção".

$$s_p = \sqrt{\frac{p \cdot q}{n}}$$

Onde:

s_p $\Rightarrow$ Erro padrão da proporção (estimativa do desvio padrão da distribuição amostral das proporções).

p $\Rightarrow$ Proporção da ocorrência do evento na amostra.

q $\Rightarrow$ Proporção da não ocorrência do evento na amostra.

n $\Rightarrow$ Tamanho da amostra.

A proporção da população ***P*** ficará entre os limites:

$$P = p \pm z \cdot s_p$$

Em que, novamente, *z* é o escore padronizado para o percentual de confiança desejado.

9.5.2 EXEMPLO

Pretende-se inferir, com os dados obtidos na amostra do exemplo anterior, qual a proporção de alunos da Universidade que possuem quocientes de inteligência superiores à média (QI = 100). Adotar um intervalo de confiança de 95%.

A proporção de alunos que, na amostra, suplantam a média pode ser calculada por:

$$p = \frac{m_P}{n} = \frac{6}{10} = 0{,}6 \qquad \text{e} \qquad q = 1 - p = 1 - 0{,}6 = 0{,}4$$

E chega-se ao erro padrão da proporção:

$$s_p = \sqrt{\frac{p \cdot q}{n}} = \sqrt{\frac{0{,}6 \text{ x } 0{,}4}{10}} = \sqrt{0{,}024} = 0{,}155$$

De onde, com confiança de 95%, a proporção de alunos da Universidade, cujos QIs superam a média de 100, é dada por:

$$\boldsymbol{P} = \boldsymbol{p} \pm \boldsymbol{z} \,.\, \boldsymbol{s}_{\boldsymbol{P}} = 0{,}6 \pm 1{,}96 \text{ x } 0{,}155 = 0{,}6 \pm 0{,}3$$

Portanto, o intervalo de confiança de 95% é:

$$0{,}3 < \boldsymbol{P} < 0{,}9$$

9.6 EXERCÍCIOS

1. Para a seguinte amostra composta de 30 escores amostrais:

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

   a) Calcule o erro padrão da média
   b) O intervalo de confiança de 95% para a média amostral
   c) O intervalo de confiança de 99% para a média amostral

2. Para a seguinte amostra composta de 34 escores amostrais:

   10, 4, 10, 5, 5, 6, 7, 3, 5, 4, 4, 5, 6, 6, 7, 5, 8,
   1, 8, 7, 5, 6, 10, 6, 8, 7, 7, 6, 5, 5, 4, 3, 4, 5.

   a) Calcule o erro padrão da média
   b) O intervalo de confiança de 95% para a média amostral
   c) O intervalo de confiança de 99% para a média amostral

3. Para a seguinte amostra composta de 32 escores amostrais:

   4, 2, 5, 6, 1, 1, 7, 8, 7, 8, 8, 2, 6, 5, 6, 4,
   4, 3, 6, 6, 7, 1, 5, 7, 8, 8, 4, 5, 3, 2, 6, 5.

   a) Calcule o erro padrão da média
   b) O intervalo de confiança de 95% para a média amostral
   c) O intervalo de confiança de 99% para a média amostral

4. Em uma amostra aleatória de 50 estudantes de determinado "campus" universitário, 57% deles mostraram-se indiferentes à existência do centro acadêmico. Baseado nessa informação

   a) Calcule o erro padrão da proporção.
   b) Construa um intervalo de confiança de 95%.
   c) Construa um intervalo de confiança de 99%.

5. Dadas uma amostra de tamanho 150 e uma proporção amostral de 0,32:

   a) Calcule o erro padrão da proporção.
   b) Construa um intervalo de confiança de 95%.
   c) Construa um intervalo de confiança de 99%.

6. Dadas uma amostra de tamanho 200 e uma proporção amostral de 0,25:

   a) Calcule o erro padrão da proporção.
   b) Construa um intervalo de confiança de 95%.
   c) Construa um intervalo de confiança de 99%.

7. Um laboratório pretende divulgar a eficácia de um novo medicamento. Aplicando-o em 200 pacientes, 164 ficara curados. Para não incorrer em propaganda enganosa, pretende garantir o percentual mínimo de sucessos. Nos níveis de confiança de 80%, 85%, 90%, 95% e 99%, qual deveria ser o mínimo assegurado?

8. A Secretaria de Educação de determinado município desenvolveu um programa de apoio psicológico para crianças problemáticas de sua rede escolar. Ao final do ano, analisando as notas obtidas por 30 delas, determinou uma média de 7,6 e um desvio padrão de 0,6. Respeitando os níveis de confiança de 80%, 85%, 90%, 95% e 99%, qual deveria ser, em cada caso, o intervalo de médias que poderia ser informado a respeito da totalidade dos alunos?

10TESTES DE SIGNIFICÂNCIA

10.1 INTRODUÇÃO

A estimação de médias ou proporções populacionais a partir de amostras, como já foi estudado, não constitui o principal objeto da inferência estatística. Na maioria das vezes, o que se busca é comparar variáveis de grupos distintos e verificar se as diferenças encontradas são estatisticamente atribuíveis ao erro amostral, simplesmente. Em outras palavras, retiradas amostras de duas ou mais populações diferentes, em relação à variável considerada, estas populações podem ser consideradas indistintas, ou não? A realização de experimentos para responder este tipo de pergunta, é que consiste nos chamados testes de significância ou testes de hipóteses como será visto. Existem duas famílias de testes, genericamente falando: os testes paramétricos e os não paramétricos.

10.1.1 TESTES PARAMÉTRICOS

Testes paramétricos são aqueles cujas variáveis são do tipo intervalar, ou seja, podem-se fazer histogramas contínuos. São dados discretos ou contínuos. A este tipo se contrapõem os dados nominais ou ordinais. Além disso, as distribuições devem ser normais, ou, pelo menos, consideradas como tais em função do tamanho da amostra ser muito grande - Teorema do Limite Central. Resumindo, testes paramétricos são aqueles que requerem normalidade e dados intervalares. Serão vistos os testes paramétricos do escore z, da razão t e da razão F.

10.1.2 TESTES NÃO PARAMÉTRICOS

Quando as condições de normalidade não são satisfeitas - a distribuição é assimétrica, por exemplo - ou os dados não são intervalares, o uso dos testes paramétricos podem levar a resultados errados. Por isso, foram desenvolvidos os chamados testes não paramétricos que não requerem as condições acima e, portanto, não ficam por elas limitados. Apresentam, como inconveniente, um poder mais reduzido da capacidade de identificar amostras de populações distintas. O teste de significância não paramétrico mais utilizado é o teste qui-quadrado que será apresentado aqui. É importante ressaltar que a distribuição Qui-quadrado é paramétrica, embora o teste não o seja. Outros testes não paramétricos que não serão vistos, mas podem ser encontrados na bibliografia são: teste da mediana, prova exata de Fisher, prova binomial e provas de Friedman, de Kruskal-Wallis e de Mann-Whitney.

10.2 TESTE DE HIPÓTESES

10.2.1 HIPÓTESE NULA

A análise estatística inicia pelo teste da chamada hipótese nula ***Ho*** - diz-se "agá-zero". A hipótese nula, também denominada hipótese probanda, afirma que duas ou mais amostras podem ser consideradas como se fossem extraídas de uma mesma população no que concerne à variável considerada. Ela assume que eventuais diferenças entre as amostras são decorrentes do erro amostral, ou seja, são casuais. Considerando duas populações, com médias μ_1 e μ_2, a hipótese nula garante:

$$\mu_1 = \mu_2$$

Uma hipótese nula, por exemplo, afirma que as notas dos alunos não dependem da renda familiar. Analisa-se amostras, então, diferenciadas por essa característica e procura-se confirmar ***Ho***.

10.2.2 HIPÓTESE ALTERNATIVA

Caso a hipótese nula possa ser rejeitada, o que normalmente se procura provar, pois é quando se encontra uma correlação, aceita-se a chamada hipótese alternativa ***Ha*** - diz-se "agá-á". A hipótese alternativa, também chamada experimental, significa que as diferenças das variáveis encontradas nas amostras decorrem de diferenças efetivamente existentes nas populações e não são frutos do erro amostral. ***Ha*** é, portanto, o complemento de ***Ho***, isto é, ou uma ou outra. Assim a hipótese alternativa garante:

$$\mu_1 \neq \mu_2$$

No exemplo citado anteriormente, significaria concluir que as notas dos alunos dependem da renda familiar.

10.2.3 NÍVEL DE SIGNIFICÂNCIA

Quando são encontradas diferenças entre as variáveis para amostras diferentes, surge o questionamento se isto é, ou não, estatisticamente significante. Estabelece-se um nível de significância ou de confiança que é a probabilidade de se rejeitar a hipótese nula com segurança. Considera-se, então, a hipótese nula falsa sempre que a probabilidade da diferença amostral encontrada for menor do que o nível de significância adotado. Assim, define-se zonas ou regiões de aceitação e rejeição. Se superado o valor limite, denominado valor crítico, entra-se na região de rejeição. Em caso contrário, aceita-se a hipótese nula que é a única que é testada.

O nível de confiança adotado, normalmente, é de 5% (ou de 0,05, em termos probabilísticos). Apenas em casos onde é exigido muito rigor usa-se 1%. Encontradas diferenças entre as amostras retiradas de duas populações, testa-se a hipótese nula dentro do limite de 0,05, por exemplo. Se o valor crítico for superado, rejeita-se ***Ho*** e aceita-se ***Ha*** , pois somente em 5% dos casos isso poderia ocorrer ao acaso.

10.2.4 TIPOS DE ERROS

ERRO DO TIPO I

Como foi visto, pode-se rejeitar a hipótese nula mesmo ela sendo verdadeira, já que em 5% dos casos as diferenças amostrais podem ser explicadas pelo acaso e não por diferenças populacionais. Comete-se, nesse caso, um erro do tipo I, ou erro alfa. Este erro deve ser evitado e, para isso, uma solução seria aumentar o nível de confiança - passar de 0,05 para 0,01. Então, a probabilidade de se rejeitar ***Ho*** indevidamente seria reduzida. Infelizmente, isso aumenta a chance de se cometer outro tipo de erro.

ERRO DO TIPO II

O erro do tipo II, ou erro beta, consiste em aceitar a hipótese nula, mesmo ela sendo falsa. Quando se adota um nível de significância muito rigoroso para rejeitar ***Ho*** , acaba-se atribuindo indevidamente ao acaso discrepâncias exageradas entre as amostras. Deste modo, ao se procurar evitar o erro do tipo I, aumenta-se o risco de se cometer um erro do tipo II e vice-versa. Prioriza-se, dentro de certos limites, evitar erros do tipo I, até porque o erro do tipo II pode ser evitado na forma de expressar a conclusão. Diz-se que não se pode afirmar em tal nível de significância que a hipótese nula deva ser rejeitada, ao invés de afirmar que nesse nível de significância a hipótese nula deva ser aceita.

10.3 DISTRIBUIÇÃO DAS DIFERENÇAS

10.3.1 DEFINIÇÃO

Imagine-se que se pretenda determinar se duas populações, com uma característica notoriamente diferente, apresentam resultados significativamente distintos em relação um certo resultado. Por exemplo: as notas de filhos de pais rigorosos são equivalentes aos dos filhos de pais liberais? Toma-se para cada amostra de uma população, uma amostra para a outra, e determinam-se suas médias. Pode-se, então, encontrar a diferença entre estas médias. Se este experimento for repetido para um número muito grande de pares de amostras, a distribuição das diferenças entre as médias de cada par, será, pelo teorema do limite central, uma distribuição normal. A determinação do desvio padrão desta distribuição é muito difícil de calcular, como no caso da distribuição das médias amostrais. Lá, o que se fazia era estimar o erro padrão da média a partir de uma amostra. Aqui, o que se fará, analogamente, é estimar o erro padrão da diferença a partir do erro padrão das médias do par de amostras das duas populações a analisar.

10.3.2 ERRO PADRÃO DA DIFERENÇA

Primeiramente, encontram-se as médias e os desvios padrões das duas amostras extraídas das duas populações. Para a primeira amostra:

$$\bar{x}_1 = \frac{\sum x_1}{n_1} \qquad \overline{x_1^2} = \frac{\sum x_1^2}{n_1}$$

$$\sigma_1^2 = \overline{x_1^2} - \bar{x}_1^2 \qquad \sigma_1 = \sqrt{\sigma_1^2}$$

$$s_1^2 = \frac{n_1 \sigma_1^2}{n_1 - 1} \qquad s_1 = \sqrt{s_1^2}$$

Para a segunda amostra:

$$\bar{x}_2 = \frac{\sum x_2}{n_2} \qquad \overline{x_2^2} = \frac{\sum x_2^2}{n_2}$$

$$\sigma_2^2 = \overline{x_2^2} - \bar{x}_2^2 \qquad \sigma_2 = \sqrt{\sigma_2^2}$$

$$s_2^2 = \frac{n_2 \sigma_2^2}{n_2 - 1} \qquad s_2 = \sqrt{s_2^2}$$

Acha-se o erro padrão da média para cada amostra:

$$s_{\bar{x}1} = \frac{\sigma_1}{\sqrt{n_1 - 1}} \qquad s_{\bar{x}2} = \frac{\sigma_2}{\sqrt{n_2 - 1}}$$

ou

$$s_{\bar{x}1} = \frac{s_1}{\sqrt{n_1}} \qquad s_{\bar{x}2} = \frac{s_2}{\sqrt{n_2}}$$

A partir dos quais estima-se o erro padrão da diferença, mediante:

$$s_\Delta = \sqrt{s_{\bar{x}1}^2 + s_{\bar{x}2}^2}$$

Se as amostras forem de mesmo tamanho n este valor pode ser calculado mais facilmente a partir das variâncias populacionais, sem que sejam necessárias as determinações dos desvios padrões amostrais e erros padrões da média, mediante:

$$s_\Delta = \sqrt{\frac{\sigma_1^2 + \sigma_2^2}{n-1}} = \sqrt{\frac{s_1^2 + s_2^2}{n}}$$

10.4 ESTATÍSTICA Z

10.4.1 TESTE

A distribuição amostral das diferenças entre as médias pode ser considerada normal a partir de duas amostras suficientemente grandes - maiores que 30 elementos, para efeitos práticos. Isso permite que seja utilizada a tabela normal padronizada, mediante a determinação do escore *z*. Assim, a partir de duas amostras, extraídas de duas populações que se pretenda testar, cujos valores médios, desvios padrões, erros padrões médios e erro padrão da diferença foram determinados, é possível encontrar o escore *z*, mediante:

$$z = \frac{\bar{x}_1 - \bar{x}_2}{s_\Delta} \qquad n \geq 30$$

Com o valor encontrado vai-se à tabela e acha-se a percentagem que corresponde ao total de diferenças inferiores a ela. Como existe simetria, deve-se multiplicar essa taxa por dois, achando a percentagem total (%) de casos cujas médias diferem menos do que o valor encontrado.

Na prática, determina-se um ***nível de significância de 0,05***, que corresponde na tabela a um valor de ***1,96*** para *z*. Se fosse adotado o ***nível de confiança de 0,01***, o valor de *z* seria ***2,58***. Encontra-se o *z* pela fórmula acima e se ele superar aquele determinado pelo nível de significância, rejeita-se a hipótese nula e aceita-se a hipótese alternativa. Caso seja inferior aceita-se a hipótese nula.

10.4.2 EXEMPLO

SITUAÇÃO

Duas amostragens imaginárias foram extraídas de dois grupos de 50 alunos cada. Na primeira amostra, todos os alunos são filhos de pais que acompanham seus estudos , verificando seus temas, mantendo diálogo com os professores, etc.. Já os alunos da segunda amostra têm pais que os deixam mais à vontade, sem acompanhá-los no dia a dia. A questão a ser testada é a seguinte: eventuais diferenças de notas entre as duas amostras podem ser consideradas casuais, ou devem ser explicadas pelo modo como os pais tratam uns e outros? Admita que as notas médias e desvios padrões obtidos sejam os seguintes:

Amostra 1	Amostra 2
média das notas $\rightarrow \bar{x}_1 =$ 7,0	média das notas $\rightarrow \bar{x}_2 =$ 6,0
desvio padrão $\rightarrow \sigma_1 =$ 2,0	desvio padrão $\rightarrow \sigma_2 =$ 1,5

SOLUÇÃO

A hipótese nula é de que as médias das populações sejam iguais, isto é, as notas dos alunos não dependam da forma de acompanhamento acadêmico dos pais.

Primeiro, encontram-se os erros padrões da média das duas amostras:

$$s_{\bar{x}1} = \frac{\sigma_1}{\sqrt{n-1}} = \frac{2,0}{\sqrt{50-1}} = 0,286$$

$$s_{\bar{x}2} = \frac{\sigma_2}{\sqrt{n-1}} = \frac{1,5}{\sqrt{50-1}} = 0,214$$

Em seguida, acha-se o erro padrão da diferença:

$$s_{\Delta} = \sqrt{s_{\bar{x}1}^2 + s_{\bar{x}2}^2} = \sqrt{(0,286)^2 + (0,214)^2} = 0,357$$

Este valor poderia ser encontrado, diretamente, através de:

$$s_{\Delta} = \sqrt{\frac{\sigma_1^2 + \sigma_2^2}{n-1}} = \sqrt{\frac{2,0^2 + 1,5^2}{50-1}} = 0,357$$

Pode-se, agora, encontrar o escore *z*:

$$z = \frac{\bar{x}_1 - \bar{x}_2}{s_{\Delta}} = \frac{7,0-6,0}{0,357} = 2,80$$

Que, sendo maior que ***1,96*** garante a rejeição da hipótese nula para um nível de significância de ***0,05***. A hipótese alternativa deve ser aceita, ou seja, as médias dos alunos assistidos pelos pais são melhores do que a daqueles não apoiados. Note-se que para o *z* acima a tabela indica 49,74%, que multiplicado por dois dá 99,48%. Portanto, em menos de meio por cento dos casos uma diferença dessas entre as médias das amostras poderia ser atribuída ao acaso.

10.5 ESTATÍSTICA T

10.5.1 PEQUENAS AMOSTRAS

Se o número de dados obtidos em cada amostras for inferior a 30, não se pode considerar a distribuição das diferenças como normal e a estatística *z* não pode ser aplicada. Para amostras pequenas utiliza-se a estatística ***t***, ou razão ***t***, para realizar os testes, que ainda podem ser chamados ***t de Student***. É necessário ressaltar que no caso da estatística *z*, a distribuição da variável observada não precisava ser normal, porque pelo teorema do limite central, o fato das amostras serem grandes assegurava distribuição normal para as médias e diferenças. Agora, a estatística ***t de Student*** exige que a variável observacional siga uma distribuição normal.

Feitas essas ressalvas, calcula-se t_O do mesmo modo que se calculava z:

$$t_o = \frac{\bar{x}_1 - \bar{x}_2}{s_\Delta} \qquad n < 30$$

E existem outras ***tabela para a estatística t de Student***, para ***níveis de significância 0,05 e 0,01***. Nelas entra-se com o ***número de graus de liberdade*** e acha-se o valor crítico de t_C. Se esse valor t_C for maior que o t_O (valor observado, calculado acima), aceita-se a hipótese nula e em caso contrário ela é rejeitada em detrimento da hipótese alternativa.

Então, resumindo:

t_o	Valor observado, calculado pelo método acima, idêntico ao usado na estatística z.
t_c	Valor crítico, encontrado na tabela, a partir do nível de significância desejado e do número de graus de liberdade gl.
$t_o < t_c$	Aceita-se a hipótese nula e rejeita-se a hipótese alternativa. As diferenças são fruto do erro amostral para este nível de significância.
$t_o \geq t_c$	Rejeita-se hipótese nula e aceita-se a hipótese alternativa. As diferenças não podem ser explicadas pelo erro amostral para este nível de significância.

Os cálculos de s_Δ e de g_l vão depender da natureza do problema em estudo. Se as variâncias das populações de onde foram extraídas as amostras são iguais ou diferentes, o tratamento é distinto. Além disso, se os dados forem pareados, ou seja, quando os dados das duas amostras forem vinculados aos pares, também há um modo diferente de se realizar o teste.

10.5.2 DADOS PAREADOS

Quando os dados forem pareados ou pretende-se fazer testes comparativos em tempos distintos para uma mesma amostra, buscando identificar mudanças em função de certas condições, adota-se este procedimento. Sempre que for possível organizar os dados aos pares, isso deve ser feito, pois reduz diferenças ocasionadas por outras variáveis que não a de estudo. A variável tratada estatisticamente passa a ser a diferença dos pares nas duas situações.

Vale o que foi visto para a estatística t, com o cuidado de se calcular o desvio padrão das diferenças σ e o erro padrão da diferença s_Δ, através das expressões:

$$\sigma^2 = \overline{\Delta^2} - \overline{\Delta}^2 \qquad \text{e} \qquad s_\Delta = \frac{\sigma}{\sqrt{n-1}}$$

Onde Δ é a diferença para cada um dos n pares de medidas da variável.

Como o tamanho da amostra, no cálculo do número de graus de liberdade para efeitos de utilização da tabela da razão t, deve ser n, que é o tamanho da amostra, tanto antes como depois, pois a amostra é a mesma:

$$gl = n - 1$$

10.5.3 VARIÂNCIAS IGUAIS

Quando não é conhecida a variância da população, embora as amostras possam ser consideradas como se fossem retiradas de uma mesma população, diz-se que as variâncias são as mesmas e as amostras são chamadas homocedásticas. Calcula-se o erro padrão da diferença s_Δ, através de:

$$s_\Delta = \sqrt{\hat{s}^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

Onde $\hat{s}$ é a estimativa do desvio padrão da população a partir das duas amostras. Ele pode ser calculado a partir de:

$$\hat{s}^2 = \frac{n_1\sigma_1^2 + n_2\sigma_2^2}{n_1+n_2-2} = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$$

E o número de graus de liberdade, a ser utilizado para que se possa ir à tabela é calculado através de:

$$gl = n_1 + n_2 - 2$$

A regra prática para se admitir que pode ser considerada uma única população e, por conseguinte, adotar-se o método exposto, é que a relação entre as variâncias amostrais seja inferior a quatro.

Se as amostras forem de mesmo tamanho n os valores acima podem ser calculados mais facilmente, mediante:

$$s_\Delta = \sqrt{\frac{\sigma_1^2+\sigma_2^2}{n-1}} = \sqrt{\frac{s_1^2+s_2^2}{n}} \qquad \text{e} \qquad gl = 2(n-1)$$

10.5.4 VARIÂNCIAS DIFERENTES

Quando as amostras provém de populações que apresentam variâncias desiguais, as amostras são heterocedásticas. O critério é que a relação entre as variâncias amostrais seja superior a quatro. Calcula-se o erro padrão da diferença através de:

$$s_{\Delta} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{\sigma_1^2}{n_1 - 1} + \frac{\sigma_2^2}{n_2 - 1}}$$

E o número de graus de liberdade, a ser utilizado para que se possa ir à tabela é calculado através de:

$$gl = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}} = \frac{\left(\frac{\sigma_1^2}{n_1 - 1} + \frac{\sigma_2^2}{n_2 - 1}\right)^2}{\frac{\left(\frac{\sigma_1^2}{n_1 - 1}\right)^2}{n_1 - 1} + \frac{\left(\frac{\sigma_2^2}{n_2 - 1}\right)^2}{n_2 - 1}}$$

10.6 ESTATÍSTICA F

As estatísticas ***z*** e ***t*** são úteis quando se pretende comparar duas médias. Quando se quer comparar três ou mais médias independentes, utiliza-se a razão ***F*** ou estatística ***F***. Aqui só serão consideradas amostras com o mesmo tamanho ***n***.

Na pretensão de comparar ***R*** médias, toma-se ***R*** amostras de ***n*** elementos. Cada amostra terá sua própria variância σ_i^2 . Pode-se imaginar uma única amostra de variância σ^2 , contendo todos os elementos das ***R*** amostras. O valor de ***F*** observado, também simbolizado por ***Fo*** é dado por:

$$F_o = \frac{R(n-1)}{R-1}\left(\frac{R\sigma^2}{\sum \sigma_i^2} - 1\right)$$

Este valor precisa ser comparado com o valor de ***F*** tabelado, o valor crítico ***Fc***. Estas tabelas são dadas normalmente para o ***níveis de significância 0,05 e 0,01***. Como entradas na tabela usam-se os graus de liberdade ***gl*** para o numerador e denominador, definidos por:

$$N = gl(\text{numerador}) = R - 1$$

$$D = gl(\text{denominador}) = R(n-1)$$

10.7 QUI-QUADRADO

10.7.1 O TESTE QUI-QUADRADO

O teste qui-quadrado é o mais popular dos testes não paramétricos. Nele comparam-se freqüências e não médias - como nos testes anteriores. A hipótese nula estabelece que as freqüências das populações não diferem em relação a uma propriedade particular. A hipótese experimental ou alternativa estabelece que as diferenças nas freqüências amostrais refletem diferenças que ocorrem nas populações.

Também aqui, calcula-se um valor que deve ser comparado com o valor crítico encontrado numa tabela. O cálculo do valor observado χ_o^2 (qui-quadrado) parte das freqüências esperadas e observadas e é dado por:

$$\chi_o^2 = \Sigma \frac{(f_O - f_E)^2}{f_E}$$

Imagine-se uma situação em que se pretenda comparar métodos de educação de crianças e orientação política dos pais. Sob o ponto de vista de orientação política os pais serão divididos em conservadores (32), moderados (30) e liberais (27). Quanto aos métodos educacionais existirão três grupos: permissivos, moderados e autoritários. A hipótese nula será de que a freqüência relativa para os diversos métodos educacionais é a mesma, independentemente da orientação política dos pais.

Monta-se uma grade, denominada tabela de contingência, em que todas as combinações apareçam, onde são colocadas as freqüências encontradas na amostragem. São as freqüências observadas f_O da equação acima.

f_O	*Conservadores*	*Moderados*	*Liberais*	*Totais* (T_C)
Permissivo	7	9	14	30
Moderado	10	10	8	28
Autoritário	15	11	5	31
Totais (T_L)	32	30	27	89 (T)

Os valores das freqüências esperadas f_E para cada casela são obtidos multiplicando os valores dos totais marginais da linha T_L e da coluna T_C e dividindo este resultado pela freqüência total T. Ou seja:

$$f_E = \frac{T_L T_C}{T}$$

Assim, os valores esperados idealmente no caso da hipótese nula seriam:

f_E

10,79	10,11	9,10
10,07	9,44	8,49
11,15	10,45	9,40

A partir daí, a fórmula pode ser calculada passo a passo, como na seqüência de tabelas abaixo:

$$(f_O - f_E)^2$$

14,36	1,23	24,01
0	0,31	0,24
14,82	0,30	19,36

Em seguida:

$$\frac{(f_O - f_E)^2}{f_E}$$

1,33	0,12	2,64
0	0,03	0,03
1,33	0,03	2,06

E finalmente:

$$\chi_o^2 = \Sigma \frac{(f_O - f_E)^2}{f_E} =$$

$$\chi_o^2 = 1,33 + 0,12 + 2,64 + 0 + 0,03 + \\ + 0,03 + 1,33 + 0,03 + 2,06 = 7,57$$

Para se comparar o valor calculado χ_O com o valor crítico tabelado χ_C, é necessário saber-se o número de graus de liberdade ***gl*** que é achado pela equação abaixo a partir do numero de linhas ***L*** e de colunas ***C***:

$$gl = (L - 1)(C - 1)$$

Que no caso atual: $gl = (3-1)(3-1) = 4$

E a tabela fornece, para quatro graus de liberdade, um valor crítico do χ_C , para o nível de significância ***0,05***, de 9,488, o que conduz a aceitação da hipótese nula, porque:

$$\chi_o = 7,57 < \chi_c = 9,488$$

10.7.2 LIMITAÇÕES DO TESTE QUI-QUADRADO

O teste Qui-quadrado só pode ser usado em amostras maiores do que 20 elementos e quando elas forem inferiores a 40, todas as freqüências esperadas devem ser superiores a um.

Como a distribuição Qui-quadrado é uma distribuição contínua e está sendo usada em dados discretos devem ser feitas correções de continuidade. A correção de Yates é apresentada abaixo.

$$\chi_o^2 = \Sigma \frac{\left(\left|f_O - f_E\right| - 0{,}5\right)^2}{f_E}$$

Em grandes amostras, a correção é desnecessária e em pequenas amostras torna-se relevante quando as freqüências esperadas são inferiores a dez. Recomenda-se também a correção de Yates quando o número de graus de liberdade é igual a um.

10.8 EXERCÍCIOS

1. Pesquisadores sociais procuraram testar a hipótese de que os jornais lidos pela classe baixa são igualmente "dirigidos" para temas sexuais quanto os lidos pela classe média. Empregando um "índice de sexualidade", eles coligiram dados de uma amostra aleatória composta de 40 artigos publicados em revistas de classe média e outra amostra, também aleatória, com outros 40 artigos publicados em revistas de classe baixa. Enquanto que a amostra da classe média apresentou uma média de sexualidade igual a 3,0, com um desvio padrão de 1,5, a amostra da classe mais baixa apresentou média 4,0 e desvio padrão igual a 2,0 (maior escore, maior grau de sexualidade). Com os dados acima, testar a hipótese de que não há diferença significativa entre os conteúdos das publicações dirigidas a essas duas classes sócio-econômicas. O que é que indicam os seus resultados?

2. Uma amostra de 100 homens indicou um tempo médio de resolução de um problema de 1190 segundos e desvio padrão de 90 segundos. Uma amostra de 75 mulheres acusou uma média de 1230 segundos e um desvio padrão de 120 segundos. Fixado o nível de significância de 0,05, pode-se admitir que a média obtida pelas mulheres é superior a dos homens?

3. Dois grupos fizeram exames finais de Estatística. Um desses grupos recebeu preparação formal para esse exame, enquanto que o outro apenas leu os textos básicos, sem, entretanto comparecer às aulas. O primeiro grupo obteve as notas 2, 2, 3 e 4 enquanto o segundo atingiu 1, 1, 2 e 3. Testar a hipótese nula de que não há diferença significativa entre os dois grupos, no que diz respeito aos escores. O que os resultados indicam?

4. Um psicólogo clínico, trabalhando com outros colegas no mesmo consultório, que verificar a diferença entre duas técnicas terapêuticas de tratamento da fobia de elevador. Para isso, repartiu 20 pacientes fóbicos, de modo aleatório, em dois grupos de dez sujeitos cada um, utilizando duas diferentes técnicas terapêuticas (A e B). Os dados obtidos representam o número de sessões de uma hora que foram necessárias para conseguir que cada paciente usasse o elevador para subir ao 30º andar, sem manifestar sinal de angústia. Os resultados encontram-se abaixo. Usando 5% de significância, pode-se concluir que há diferenças entre os dois tratamentos?

Terapia A	11	8	6	18	10	9	11	7	10	11
Terapia B	3	11	6	8	11	2	7	4	5	12

5. Testar, para as significâncias de 95% e 99%, a diferença entre as médias de cada um dos pares de amostras abaixo:

   a) Amostra 1: 8, 3, 1, 7, 7, 6, 8
      Amostra 2: 1, 5, 8, 3, 2, 1, 2

   b) Amostra 1: 6, 6, 8, 7, 5, 4, 8, 7, 7
      Amostra 2: 6, 5, 7, 7, 3, 3, 5, 6, 3

   c) Amostra 1: 15, 18, 12, 17, 19
      Amostra 2: 10, 11, 12, 10, 10

   d) Amostra 1: 1, 1, 2, 3, 3
      Amostra 2: 2, 2, 4, 2, 2

e) Amostra 1: 5, 7, 7, 3, 6, 5, 4, 6, 7
Amostra 2: 10, 7, 9, 9, 7, 8

f) Amostra 1: 3, 6, 4, 2, 1
Amostra 2: 7, 8, 8, 9, 9, 6, 5

g) Amostra 1: 10, 4, 1, 2, 4, 8, 3, 5
Amostra 2: 10, 10, 8, 7

6. Antes e depois de assistirem a um filme, cujo objetivo era aliviar a discriminação contra grupos minoritários, seis estudantes foram testados. A variável observacional era a "atitude para com judeus", e quanto maiores fossem os escores, mais favoráveis seriam as atitudes. Com os dados abaixo, testar a hipótese de que , com relação às "atitudes", o filme não exerceu a menor influência sobre os estudantes.

Estudantes	*A*	*B*	*C*	*D*	*E*	*F*
Antes	2	2	4	6	7	5
Depois	4	5	3	8	9	8

7. Testar a significância da diferença entre as médias, numa situação do tipo "antes/depois", com os seguintes escores componentes de amostras aleatórias:

a)

Respondente	*A*	*B*	*C*	*D*
Antes	7	6	5	4
Depois	3	4	2	3

b)

Respondente	*A*	*B*	*C*	*D*	*E*
Antes	6	7	10	9	8
Depois	3	4	9	7	5

8. Nas seguintes amostras aleatórias de classe social, testar a hipótese nula de que a "camaradagem entre vizinhos" não varia com a classe social (quanto maior o escore, maior a camaradagem).

Classe Baixa	*Classe Operária*	*Classe Média*	*Classe Alta*
8	7	6	5
4	3	5	2
7	2	5	1
8	8	4	3

9. Testar se as diferenças entre as médias das seguintes amostras são significantes:

a) Amostra 1: 2, 1, 3, 3
Amostra 2: 5, 4, 3, 4
Amostra 3: 8, 9, 7, 8

b) Amostra 1: 12, 6, 8, 7, 6
Amostra 2: 6, 5, 7, 5, 1
Amostra 3: 3, 2, 5, 3, 1

c) Amostra 1: 5, 5, 4, 3, 6
Amostra 2: 4, 3, 2, 2, 1
Amostra 3: 3, 5, 1, 3, 3

d) Amostra 1: 1, 1, 3, 4, 2, 1
Amostra 2: 3, 2, 2, 1, 5, 5
Amostra 3: 4, 4, 2, 2, 3, 3
Amostra 4: 6, 6, 5, 5, 4, 6

10. Amostras aleatórias de homens e de mulheres foram feitas quanto ao vício do cigarro. Descobriu-se que de 29 homens, 15 eram fumantes e que de 30 mulheres, 20 tinham o hábito de fumar. Teste a hipótese nula de que a freqüência relativa de homens e mulheres fumantes é a mesma. O que indicam os resultados?

11. Dois grupos de estudantes fizeram exames finais de Estatística. Somente um grupo recebeu preparação formal para o exame; o outro leu o texto recomendado, mas nunca compareceu às aulas. Enquanto que 22 dos 30 membros do grupo de freqüentadores passaram no exame, apenas 10 dos 28 do outro grupo lograram êxito. Teste a hipótese nula de que os resultados não são significativamente diferentes.

12. Afirma-se que certa droga é eficiente na cura de resfriados. Numa experiência com 164 pessoas a droga foi dada a metade delas e à outra metade foi dado um placebo. As reações dos pacientes ao tratamento estão apresentadas abaixo. Com 5% de significância pode-se concluir que a reação depende do tratamento aplicado?

	Ajudou	Prejudicou	Sem efeito
Droga	50	10	22
Placebo	44	12	26

13. Faça testes qui-quadrado para os problemas abaixo:

a)

	Candidato A	*Candidato B*	*Candidato C*
Região Sul	20	17	5
Região Centro	15	16	16
Região Norte	4	14	18

b)

	Juiz	*Juíza*
Candidata A	25	6
Candidata B	19	10
Candidata C	15	15
Candidata D	8	20

c)

	Produto A	*Produto B*	*Produto C*
jovens	8	10	15
velhos	12	10	9

APÊNDICE TABELAS

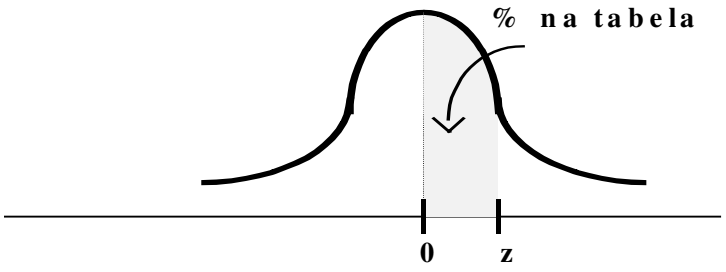
As tabelas a seguir servem de apoio na resolução de problemas. Em livros especializados elas podem ser encontradas em maior número e em maiores detalhes. Estão reunidas as seguintes estatísticas paramétricas: distribuição normal reduzida (estatística *z*), a distribuição ***t*** de Student (estatística ***t***) e a estatística de Fisher-Snedecor (estatística ***f***). A única estatística não paramétrica apresentada é a distribuição qui-quadrado.

A. ESTATÍSTICA Z

A distribuição normal reduzida assume uma média da população igual a zero ($\mu = 0$) e tanto variância como desvio padrão iguais a um ($\sigma^2 = 1$ e $\sigma = 1$). Na forma de representação da curva normal $N(\mu,\sigma^2)$, portanto, fica sendo $N(0,1)$.

A tabela indica a percentagem da área que fica entre a média e o valor de *z*, que é o valor de entrada na tabela. Como a tabela é simétrica, para valores negativos de *z* os percentuais são idênticos aos correspondentes positivos, significando apenas que se localizam à esquerda e não à direita da média.

A procura na tabela faz-se encontrando na coluna até o casa decimal de *z* e na linha a casa centesimal. No ponto de encontro está indicada a percentagem da área entre a média e aquele valor de *z*.



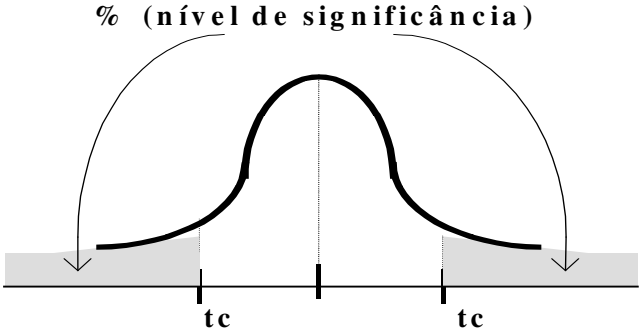
DISTRIBUIÇÃO NORMAL REDUZIDA N(0,1)
ESTATÍSTICA Z

z	X,X0	X,X1	X,X2	X,X3	X,X4	X,X5	X,X6	X,X7	X,X8	X,X9	z
0,0X	0,000	0,399	0,798	1,197	1,595	1,994	2,392	2,790	3,188	3,586	**0,0X**
0,1X	3,983	4,380	4,776	5,172	5,567	5,962	6,356	6,750	7,142	7,535	**0,1X**
0,2X	7,926	8,317	8,706	9,095	9,484	9,871	10,257	10,642	11,026	11,409	**0,2X**
0,3X	11,791	12,172	12,552	12,930	13,307	13,683	14,058	14,431	14,803	15,173	**0,3X**
0,4X	15,542	15,910	16,275	16,640	17,003	17,365	17,724	18,082	18,439	18,793	**0,4X**
0,5X	19,146	19,497	19,847	20,194	20,540	20,884	21,226	21,566	21,904	22,241	**0,5X**
0,6X	22,575	22,907	23,237	23,565	23,891	24,215	24,537	24,857	25,175	25,490	**0,6X**
0,7X	25,804	26,115	26,424	26,731	27,035	27,337	27,637	27,935	28,231	28,524	**0,7X**
0,8X	28,815	29,103	29,389	29,673	29,955	30,234	30,511	30,785	31,057	31,327	**0,8X**
0,9X	31,594	31,859	32,121	32,381	32,639	32,894	33,147	33,398	33,646	33,891	**0,9X**
1,0X	34,135	34,357	34,614	34,850	35,083	35,314	35,543	35,749	35,993	36,214	**1,0X**
1,1X	36,433	36,650	36,864	37,076	37,286	37,493	37,698	37,900	38,100	38,298	**1,1X**
1,2X	38,493	38,686	38,877	39,065	39,251	39,435	39,616	39,796	39,973	40,148	**1,2X**
1,3X	40,319	40,490	40,658	40,824	40,988	41,149	41,309	41,466	41,621	41,774	**1,3X**
1,4X	41,924	42,073	42,220	42,364	42,507	42,647	42,786	42,922	43,056	43,189	**1,4X**
1,5X	43,319	43,478	43,574	43,699	43,822	43,943	44,062	44,179	44,295	44,408	**1,5X**
1,6X	44,520	44,630	44,738	44,845	44,950	45,053	45,154	45,254	45,352	45,449	**1,6X**
1,7X	45,544	45,637	45,728	45,819	45,907	45,994	46,080	46,164	46,246	46,328	**1,7X**
1,8X	46,407	46,485	46,562	46,638	46,712	46,784	46,856	46,926	46,995	47,062	**1,8X**
1,9X	47,128	47,193	47,257	47,320	47,381	47,441	47,500	47,558	47,615	47,671	**1,9X**
2,0X	47,725	47,778	47,831	47,882	47,933	47,982	48,030	48,077	48,124	48,169	**2,0X**
2,1X	48,214	48,257	48,300	48,341	48,382	48,422	48,461	48,500	48,537	48,574	**2,1X**
2,2X	48,610	48,648	48,679	48,713	48,746	48,778	48,809	48,840	48,870	48,899	**2,2X**
2,3X	48,928	48,956	48,983	49,010	49,036	49,061	49,086	49,111	49,134	49,158	**2,3X**
2,4X	49,180	49,202	49,224	49,245	49,266	49,286	49,305	49,324	49,343	49,361	**2,4X**
2,5X	49,379	49,396	49,413	49,430	49,446	49,461	49,477	49,492	49,506	49,520	**2,5X**
2,6X	49,534	49,547	49,560	49,573	49,586	49,598	49,609	49,621	49,632	49,643	**2,6X**
2,7X	49,653	49,664	49,674	49,683	49,693	49,702	49,711	49,720	49,728	49,767	**2,7X**
2,8X	49,745	49,752	49,760	49,767	49,774	49,781	49,788	49,795	49,801	49,807	**2,8X**
2,9X	49,813	48,819	49,825	49,831	49,836	49,841	49,846	49,851	49,856	49,861	**2,9X**
3,0X	49,865	49,869	49,874	49,878	49,882	49,886	49,889	49,893	49,897	49,900	**3,0X**
3,1X	49,903	49,906	49,910	49,913	49,916	49,918	49,921	49,924	49,926	49,929	**3,1X**
3,2X	49,931	49,934	49,936	49,938	49,940	49,942	49,944	49,946	49,948	49,950	**3,2X**
3,3X	49,952	49,953	49,955	49,957	49,958	49,960	49,961	49,962	49,964	49,965	**3,3X**
3,4X	49,966	49,968	49,969	49,970	49,971	49,972	49,973	49,974	49,975	49,976	**3,4X**
3,5X	49,977	49,978	49,978	49,979	49,980	49,981	49,982	49,982	49,983	49,984	**3,5X**
3,6X	49,984	49,985	49,985	49,986	49,986	49,987	49,987	49,988	49,988	49,989	**3,6X**
3,7X	49,989	49,990	49,990	49,990	49,991	49,991	49,992	49,992	49,992	49,993	**3,7X**
3,8X	49,993	49,993	49,993	49,994	49,994	49,994	49,994	49,995	49,995	49,995	**3,8X**
3,9X	49,995	49,995	49,996	49,996	49,996	49,996	49,996	49,996	49,997	49,997	**3,9X**
4,0X	49,997	49,997	49,997	49,997	49,997	49,997	49,998	49,998	49,998	49,998	**4,0X**
z	**X,X0**	**X,X1**	**X,X2**	**X,X3**	**X,X4**	**X,X5**	**X,X6**	**X,X7**	**X,X8**	**X,X9**	**z**

B. ESTATÍSTICA T

A tabela da distribuição ***t*** de Student fornece o valor crítico ***tc*** para vários níveis de significância. Este valor indica o limite a partir do qual existe uma probabilidade, dada pelo nível de significância, de que dois valores não sejam originários de uma mesma população.

A tabela, aqui, é fornecida para os níveis de significância ***0,01***, ***0,05***, ***0,10*** e ***0,20***, indicando que em 1%, 5%, 10% e 20% dos casos, respectivamente, as duas amostras podem não ser oriundas de uma mesma população, a despeito do valor observado para ***to*** ser inferior ao valor crítico ***tc*** encontrado na tabela. Em outras palavras, para duas amostras para as quais foi encontrado um valor de ***to*** menor que o valor crítico ***tc*** tabelado, em cada um dos casos, em 99%, 95%, 90% e 80% das vezes elas provêm de uma mesma população.



A tabela fornece o valor crítico ***tc*** para cada célula ou casela definida pela coluna do nível de significância escolhido e pela linha correspondente ao número de graus de liberdade ***gl***, calculado através de uma das formas abaixo, onde n, n_1 e n_2 são os tamanhos das amostras:

DADOS PAREADOS ANTES / DEPOIS	$gl = n - 1$
VARIÂNCIAS IGUAIS HOMOCEDÁSTICAS	$gl = n_1 + n_2 - 2$
VARIÂNCIAS DIFERENTES HETEROCEDÁSTICAS	$gl = \dfrac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1-1}+\frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2-1}} = \dfrac{\left(\frac{\sigma_1^2}{n_1-1}+\frac{\sigma_2^2}{n_2-1}\right)^2}{\frac{\left(\frac{\sigma_1^2}{n_1-1}\right)^2}{n_1-1}+\frac{\left(\frac{\sigma_2^2}{n_2-1}\right)^2}{n_2-1}}$

DISTRIBUIÇÃO "t" DE STUDENT

	NÍVEL DE SIGNIFICÂNCIA				
gl	**0,20 (20%)**	**0,10 (10%)**	**0,05 (5%)**	**0,01 (1%)**	**gl**
1	3,0777	6,3137	12,7062	63,6559	**1**
2	1,8856	2,9200	4,3027	9,9250	**2**
3	1,6377	2,3534	3,1824	5,8408	**3**
4	1,5332	2,1318	2,7765	4,6041	**4**
5	1,4759	2,0150	2,5706	4,0321	**5**
6	1,4398	1,9432	2,4469	3,7074	**6**
7	1,4149	1,8946	2,3646	3,4995	**7**
8	1,3968	1,8595	2,3060	3,3554	**8**
9	1,3830	1,8331	2,2622	3,2498	**9**
10	1,3722	1,8125	2,2281	3,1693	**10**
11	1,3634	1,7952	2,2010	3,1058	**11**
12	1,3562	1,7823	2,1788	3,0545	**12**
13	1,3502	1,7709	2,1604	3,0123	**13**
14	1,3450	1,7613	2,1448	2,9768	**14**
15	1,3406	1,7531	2,1315	2,9467	**15**
16	1,3368	1,7459	2,1199	2,9208	**16**
17	1,3334	1,7396	2,1098	2,8982	**17**
18	1,3304	1,7341	2,1009	2,8784	**18**
19	1,3277	1,7291	2,0930	2,8609	**19**
20	1,3253	1,7247	2,0860	2,8453	**20**
21	1,3232	1,7207	2,0796	2,8314	**21**
22	1,3212	1,7171	2,0739	2,8188	**22**
23	1,3195	1,7139	2,0687	2,8073	**23**
24	1,3178	1,7109	2,0639	2,7970	**24**
25	1,3163	1,7081	2,0595	2,7874	**25**
26	1,3150	1,7056	2,0555	2,7787	**26**
27	1,3137	1,7033	2,0518	2,7707	**27**
28	1,3125	1,7011	2,0484	2,7633	**28**
29	1,3114	1,6991	2,0452	2,7564	**29**
30	1,3104	1,6973	2,0423	2,7500	**30**
35	1,3062	1,6896	2,0301	2,7238	**35**
40	1,3031	1,6839	2,0211	2,7045	**40**
45	1,3007	1,6794	2,0141	2,6896	**45**
50	1,2987	1,6759	2,0086	2,6778	**50**
60	1,2958	1,6706	2,0003	2,6603	**60**
70	1,2938	1,6669	1,9944	2,6479	**70**
80	1,2922	1,6641	1,9901	2,6387	**80**
90	1,2910	1,6620	1,9867	2,6316	**90**
100	1,2901	1,6602	1,9840	2,6259	**100**
1000	1,2824	1,6464	1,9623	2,5807	**1000**
gl	**0,20 (20%)**	**0,10 (10%)**	**0,05 (5%)**	**0,01 (1%)**	**gl**
	NÍVEL DE SIGNIFICÂNCIA				

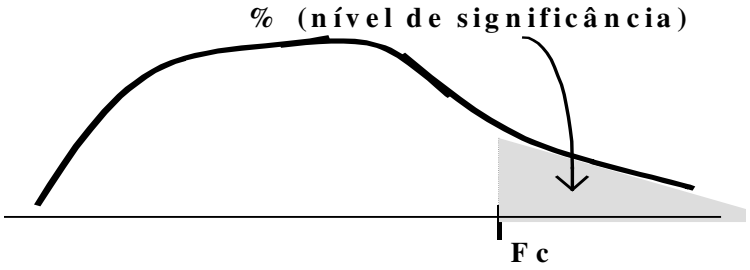
C. ESTATÍSTICA F

Cada tabela da distribuição ***F*** de Fisher-Snedecor fornece os valores críticos ***Fc*** para um nível de significância, em função dos graus de liberdade do numerador (caracteriza a coluna) e do denominador (caracteriza a linha). Se a estatística ***F*** pretende comparar as médias de ***R*** amostras de tamanho ***n***, estes graus de liberdades são dados por:

$$N = gl(\text{numerador}) = R - 1$$

$$D = gl(\text{denominador}) = R(n - 1)$$

Para cada um dos níveis de significância, aqui de ***0,01***, ***0,05***, e ***0,10***, é fornecida uma tabela, indicando que em 1%, 5%, e 10% dos casos, respectivamente, as ***R*** amostras podem não ser oriundas de uma mesma população, a despeito do valor observado para ***Fo*** ser inferior ao valor crítico ***Fc*** encontrado na tabela. Em outras palavras, para ***R*** amostras para as quais foi encontrado um valor de ***Fo*** menor que o valor crítico ***Fc*** tabelado, em cada um dos casos, em 99%, 95%, e 90% das vezes elas provêm de uma mesma população.



DISTRIBUIÇÃO F DE FISHER-SNEDECOR NÍVEL DE SIGNIFICÂNCIA 0,01 (1%)

D\N	1	2	3	4	5	6	8	10	12	15	20	30	50	100	N/D
1	4052,18	4999,34	5403,53	5624,26	5763,96	5858,95	58980,9	6055,95	6106,68	6156,97	6208,66	6560,35	6302,26	6333,92	1
2	98,5019	99,0003	99,1640	99,2513	99,3023	99,3314	99,3750	99,3969	99,4187	99,4332	99,4478	99,4660	99,4769	99,4914	2
3	34,1161	30,8164	29,4567	28,7100	28,2371	27,9106	27,4895	27,2285	27,0520	26,8719	26,6900	26,5045	26,3544	26,2407	3
4	21,1976	17,9998	16,6942	15,9771	15,5219	15,2068	14,7988	14,5460	14,3737	14,1981	14,0194	13,8375	13,6897	13,5769	4
5	16,2581	13,2741	12,0599	11,3919	10,9671	10,6722	10,2893	10,0511	9,8883	9,7223	9,5527	9,3794	9,2377	9,1300	5
6	13,7452	10,9249	9,7796	9,1484	8,7459	8,4660	8,1017	7,8742	7,7183	7,5590	7,3958	7,2286	7,0914	6,9867	6
7	12,2463	9,5465	8,4513	7,8467	7,4604	7,1914	6,8401	6,6201	6,4691	6,3144	6,1555	5,9920	5,8597	5,7546	7
8	11,2586	8,6491	7,5910	7,0061	6,6318	6,3707	6,0288	5,8143	5,6667	5,5152	5,3591	5,1981	5,0654	4,9633	8
9	10,5615	8,0215	6,9920	6,4221	6,0569	5,8018	5,4671	5,2565	5,1115	4,9621	4,8080	4,6486	4,5167	4,4150	9
10	10,0442	7,5595	6,5523	5,9944	5,6364	5,3858	5,0567	4,8491	4,7058	4,5582	4,4054	2,2469	4,1155	4,0137	10
11	9,6461	7,2057	6,2167	5,6683	5,3160	5,0692	4,7445	4,5393	4,3974	4,2509	4,0990	3,9411	3,8097	3,7077	11
12	9,3303	6,9266	5,9525	5,4119	5,0644	4,8205	4,4994	4,2961	4,1553	4,0096	3,5884	3,7008	3,5692	3,4668	12
13	9,0738	6,7009	5,7394	5,2053	4,8616	4,6203	4,3021	4,1003	3,9603	3,8154	3,6646	3,5070	3,3752	3,2723	13
14	8,8617	6,5149	5,5639	5,0354	4,6950	4,4558	4,1400	3,9394	3,8002	3,6557	3,5052	3,3476	3,2153	3,1118	14
15	8,6832	6,3588	5,4170	4,8932	4,5556	4,3183	4,0044	3,8049	3,6662	3,5222	3,3719	3,2141	3,0814	2,9772	15
16	8,5309	6,2263	5,2922	4,7726	4,4374	4,2016	3,8890	3,6909	3,5527	3,4090	3,2587	3,1007	2,9675	2,8627	16
17	8,3998	6,1121	5,1850	4,6689	4,3360	4,1015	3,7909	3,5931	3,4552	3,3117	3,1615	3,0032	2,8694	2,7639	17
18	8,2855	6,0129	5,0919	4,5790	4,2479	4,0146	3,7054	3,5081	3,3706	3,2273	3,0771	2,9185	2,7841	2,6779	18
19	8,1850	5,9259	5,0103	4,5002	4,1708	3,9386	3,6305	3,4338	3,2965	3,1533	3,0031	2,8442	2,7092	2,6023	19
20	8,0960	5,8490	4,9382	4,4307	4,1027	3,8714	3,5644	3,3682	3,2311	3,0880	2,9377	2,7785	2,6430	2,5353	20
21	8,0166	5,7804	4,8740	4,3688	4,0421	3,8117	3,5056	3,3098	3,1729	3,0300	2,8795	2,7200	2,5838	2,4755	21
22	7,9453	5,7190	4,8166	4,3134	3,9880	3,7583	3,4530	3,2576	3,1209	2,9779	2,8474	2,6675	2,5308	2,4218	22
23	7,8811	5,6637	4,7648	4,2635	3,9392	3,7102	3,4057	3,2106	3,0740	2,9311	2,7805	2,6202	2,4829	2,3732	23
24	7,8229	5,6136	4,7181	4,2185	3,8951	3,6667	3,3629	3,1681	3,0316	2,8887	2,7380	2,5773	2,4395	2,3291	24
25	7,7698	5,5680	4,6755	4,1774	3,8550	3,6272	3,3239	3,1294	2,9931	2,8502	2,6993	2,5383	2,3999	2,2888	25
26	7,7213	5,5263	4,6365	4,1400	3,8183	3,5911	3,2884	3,0941	2,9578	2,8150	2,6640	2,5026	2,3637	2,2519	26
27	7,6767	5,4881	4,6009	4,1056	3,7847	3,5580	3,2558	3,0618	2,9526	2,7827	2,6316	2,4699	2,3304	2,2180	27
28	7,6357	5,4529	4,5681	4,0740	3,7539	3,5276	3,2259	3,0320	2,8959	2,7530	2,6018	2,4397	2,2997	2,1867	28
29	7,5977	5,4205	4,5378	4,0449	3,7254	3,4995	3,1982	3,0045	2,8685	2,7256	2,5742	2,4118	2,2713	2,1577	29
30	7,5624	5,3903	4,5097	4,0179	3,6990	3,4735	3,1726	2,9791	2,8431	2,7002	2,5487	2,3860	2,2450	2,1307	30
40	7,3142	5,1785	4,3126	3,8283	3,5138	3,2910	2,9930	2,8005	2,6648	2,5216	2,3689	2,2034	2,0581	1,9383	40
50	7,1706	5,0566	4,1994	3,7195	3,4077	3,1864	2,8900	2,6981	2,5625	2,4190	2,2652	2,0976	1,9490	1,8248	50
100	6,8953	4,8239	3,9837	3,5127	3,2059	2,9877	2,6943	2,1793	2,3676	2,2230	2,0666	1,8933	1,7353	1,5977	100

DISTRIBUIÇÃO F DE FISHER-SNEDECOR NÍVEL DE SIGNIFICÂNCIA 0,05 (5%)

D\N	1	2	3	4	5	6	8	10	12	15	20	30	50	100	N/D
1	161,446	199,500	215,707	224,583	230,160	233,988	238,884	241,882	243,905	245,949	248,016	250,097	251,774	253,043	1
2	18,5128	19,0000	19,1642	19,2467	19,2963	19,3295	19,3709	19,3959	19,4125	19,4291	19,4457	19,4625	19,4757	19,4857	2
3	10,1280	9,5521	9,2766	9,1172	9,0134	8,9407	8,8452	8,7855	8,7447	8,7028	8,6602	8,6166	8,5810	8,5539	3
4	7,7086	6,9443	6,5914	6,3882	6,2561	6,1631	6,0410	5,9644	5,9117	5,8578	5,8025	5,7459	5,6975	5,6640	4
5	6,6079	5,7861	5,4094	5,1922	5,0503	4,9503	4,8183	4,7351	4,6777	4,6188	4,5581	4,4957	4,4444	4,4051	5
6	5,9874	5,1432	4,7571	4,5337	4,3874	4,2839	4,1468	4,0600	3,9999	3,9381	3,8742	3,8082	3,7537	3,7117	6
7	5,5915	4,7374	4,3468	4,1203	3,9715	3,8660	3,7257	3,6365	3,5747	3,5107	3,4445	3,3758	3,3189	3,2749	7
8	5,3176	4,4590	4,0662	3,8379	3,6875	3,5806	3,4381	3,3472	3,2839	3,2184	3,1503	3,0794	3,0204	2,9747	8
9	5,1174	4,2565	3,8625	3,6331	3,4817	3,3738	3,2296	3,1373	3,0729	3,0061	2,9365	2,8637	2,8028	2,7556	9
10	4,9646	4,1028	3,7083	3,4780	3,3258	3,2172	3,0717	2,9782	2,9130	2,8450	2,7740	2,6996	2,6371	2,5884	10
11	4,8443	3,9823	3,5874	3,3567	3,2039	3,0946	2,9480	2,8536	2,7876	2,7186	2,6464	2,5705	2,5066	2,4566	11
12	4,7472	3,8853	3,4903	3,2592	3,1059	2,9961	2,8486	2,7534	2,6866	2,6169	2,5436	2,4663	2,4010	2,3498	12
13	4,6672	3,8056	3,4105	3,1791	3,0254	2,9153	2,7669	2,6710	2,6037	2,5331	2,4589	2,3803	2,3138	2,2614	13
14	4,6001	3,7389	3,3439	3,1122	2,9582	2,8477	2,6987	2,6022	2,5342	2,4630	2,3879	2,3082	2,2405	2,1870	14
15	4,5431	3,6823	3,2874	3,0556	2,9013	2,7905	2,6408	2,5437	2,4753	2,4034	2,3275	2,2468	2,1780	2,1234	15
16	4,4940	3,6337	3,2389	3,0069	2,8524	2,7413	2,5911	2,4935	2,4247	2,3522	2,2756	2,1938	2,1240	2,0685	16
17	4,4513	3,5915	3,1968	2,9647	2,8100	2,6987	2,5480	2,4499	2,3807	2,3077	2,2304	2,1477	2,0769	2,0204	17
18	4,4139	3,5546	3,1599	2,9277	2,7729	2,6613	2,5102	2,4117	2,3421	2,2686	2,1906	2,1071	2,0354	1,9780	18
19	4,3808	3,5219	3,1274	2,8951	2,7401	2,6283	2,4768	2,3779	2,3080	2,2341	2,1555	2,0712	1,9986	1,9403	19
20	4,3513	3,4928	3,0984	2,8661	2,7109	2,5990	2,4471	2,3479	2,2776	2,2033	2,1242	2,0391	1,9656	1,9066	20
21	4,3248	3,4668	3,0725	2,8401	2,6848	2,5727	2,4205	2,3210	2,2504	2,1757	2,0960	2,0102	1,9360	1,8761	21
22	4,3009	3,4434	3,0491	2,8167	2,6613	2,5491	2,3965	2,2967	2,2258	2,1508	2,0707	1,9842	1,9092	1,8486	22
23	4,2793	3,4221	3,0280	2,7955	2,6400	2,5277	2,3748	2,2747	2,2036	2,1282	2,0476	1,9605	1,8848	1,8234	23
24	4,2597	3,4028	3,0088	2,7763	2,6207	2,5082	2,3551	2,2547	2,1834	2,1077	2,0267	1,9390	1,8625	1,8005	24
25	4,2417	3,3852	2,9912	2,7587	2,6030	2,4904	2,3371	2,2365	2,1649	2,0889	2,0075	1,9192	1,8421	1,7794	25
26	4,2252	3,3690	2,9752	2,7426	2,5868	2,4741	2,3205	2,2197	2,1479	2,0716	1,9898	1,9010	1,8233	1,7599	26
27	4,2100	3,3541	2,9603	2,7278	2,5719	2,4591	2,3053	2,2043	2,1323	2,0558	1,9736	1,8842	1,8059	1,7419	27
28	4,196	3,3404	2,9467	2,7141	2,5581	2,4453	2,2913	2,1900	2,1179	2,0411	1,9586	1,8687	1,7898	1,7251	28
29	4,1830	3,3277	2,9340	2,7014	2,5454	2,4324	2,2782	2,1768	2,1045	2,0275	1,9446	1,8543	1,7748	1,7096	29
30	4,1709	3,3158	2,9223	2,6896	2,5336	2,4205	2,2662	2,1646	2,0921	2,0148	1,9317	1,8409	1,7609	1,6950	30
40	4,0847	3,2317	2,8387	2,6060	2,4495	2,3359	2,1802	2,0773	2,0035	1,9245	1,8389	1,7444	1,6600	1,5892	40
50	4,0343	3,1826	2,7900	2,5572	2,4004	2,2864	2,1299	2,0261	1,9515	1,8714	1,7841	1,6872	1,5995	1,5249	50
100	3,9362	3,0873	2,6955	2,4626	2,3053	2,1906	2,0323	1,9267	1,8503	1,7675	1,6764	1,5733	1,4772	1,3917	100

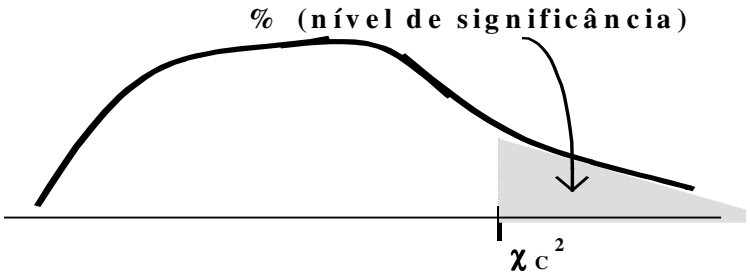
DISTRIBUIÇÃO F DE FISHER-SNEDECOR NÍVEL DE SIGNIFICÂNCIA 0,10 (10%)

D\N	1	2	3	4	5	6	8	10	12	15	20	30	50	100	N/D
1	39,8636	49,5002	53,5933	55,8330	57,2400	58,2045	59,4391	60,1949	60,7051	61,2204	61,7401	62,2649	6,6878	63,0071	**1**
2	8,5263	9,0000	9,1618	9,2434	9,2926	9,3255	9,3668	9,3916	9,4082	9,4247	9,4413	9,4579	9,4713	9,4813	**2**
3	5,5383	5,4624	5,3908	5,3427	5,3091	5,2847	5,2517	5,2304	5,2156	5,2003	5,1845	5,1681	5,1546	5,1443	**3**
4	4,5448	4,3246	4,1909	4,1071	4,0506	4,0097	3,9549	3,9199	3,8955	3,8704	3,8443	3,8174	3,7952	3,7782	**4**
5	4,0604	3,7797	3,6195	3,5202	3,4530	3,4045	3,3393	3,2974	3,2682	3,2380	3,2067	3,1741	3,1471	3,1263	**5**
6	3,7760	3,4633	3,2888	3,1808	3,1075	3,0546	2,9830	2,9369	2,9047	2,8712	2,8363	2,8000	2,7697	2,7463	**6**
7	3,5894	3,2574	3,0741	2,9605	2,8833	2,8274	2,7516	2,7025	2,6681	2,6322	2,5947	2,5555	2,5226	2,4971	**7**
8	3,4579	3,1131	2,9238	2,8064	2,7264	2,6683	2,5893	2,5380	2,5020	2,4642	2,4246	2,3830	2,3481	2,3208	**8**
9	3,3603	3,0064	2,8129	2,6927	2,6106	2,5509	2,4694	2,4163	2,3789	2,3396	2,2983	2,2547	2,2180	2,1892	**9**
10	3,2850	2,9245	2,7277	2,6053	2,5216	2,4606	2,3771	2,3226	2,2841	2,2435	2,2007	2,1554	2,1171	2,0869	**10**
11	3,2252	2,8595	2,6602	2,5362	2,4512	2,3891	2,3040	2,2482	2,2087	2,1671	2,1230	2,0762	2,0364	2,0050	**11**
12	3,1766	2,8068	2,6055	2,4801	2,3940	2,3310	2,2446	2,1878	2,1474	2,1049	2,0597	2,0115	1,9704	1,9379	**12**
13	3,1362	2,7632	2,5603	2,4337	2,3467	2,2830	2,1963	2,1376	2,0966	2,0532	2,0070	1,9576	1,9153	1,8817	**13**
14	3,1022	2,7265	2,5222	2,3947	2,3069	2,2426	2,1539	2,0954	2,0537	2,0095	1,9625	1,9119	1,8686	1,8340	**14**
15	3,0732	2,6952	2,4898	2,3614	2,2730	2,2081	2,1185	2,0593	2,0171	1,9722	1,9243	1,8728	1,8284	1,7929	**15**
16	3,0481	2,6682	2,4618	2,3327	2,2438	2,1783	2,0880	2,0281	1,9854	1,9399	1,8913	1,8388	1,7934	1,7570	**16**
17	3,0262	2,6446	2,4374	2,3077	2,2183	2,1524	2,0613	2,0009	1,9577	1,9117	1,8624	1,8090	1,7628	1,7255	**17**
18	3,0070	2,6239	2,4160	2,2858	2,1958	2,1296	2,0379	1,9770	1,9333	1,8868	1,8368	1,7827	1,7356	1,6976	**18**
19	2,9899	2,6056	2,3970	2,2663	2,1760	2,1094	2,0171	1,9557	1,9117	1,8647	1,8142	1,7592	1,7114	1,6726	**19**
20	2,9747	2,5893	2,3801	2,2489	2,1582	2,0913	1,9985	1,9367	1,8924	1,8449	1,7938	1,7382	1,6896	1,6501	**20**
21	2,9610	2,5746	2,3649	2,2333	2,1423	2,0751	1,9819	1,9197	1,8750	1,8271	1,7756	1,7193	1,6700	1,6298	**21**
22	2,9486	2,5613	2,3512	2,2193	2,1279	2,0605	1,9668	1,9043	1,8593	1,8111	1,7590	1,7021	1,6521	1,6113	**22**
23	2,9374	2,5493	2,3387	2,2065	2,1149	2,0472	1,9531	1,8903	1,8450	1,7964	1,7439	1,6864	1,6358	1,5944	**23**
24	2,9271	2,5383	2,3274	2,1949	2,1030	2,0351	1,9407	1,8775	1,8319	1,7831	1,7302	1,6721	1,6209	1,5788	**24**
25	2,9177	2,5283	2,3170	2,1842	2,0922	2,0241	1,9292	1,8658	1,8200	1,7708	1,7175	1,6589	1,6072	1,5645	**25**
26	2,9091	2,5191	2,3075	2,1745	2,0822	2,0139	1,9188	1,8550	1,8090	1,7596	1,7059	1,6468	1,5945	1,5513	**26**
27	2,9012	2,5106	2,2987	2,1655	2,0730	2,0045	1,9091	1,8451	1,7989	1,7492	1,6951	1,6356	1,5827	1,5390	**27**
28	2,8938	2,5028	2,2906	2,1571	2,0645	1,9959	1,9001	1,8359	1,7895	1,7395	1,6852	1,6252	1,5718	1,5276	**28**
29	2,8870	2,4955	2,2831	2,1494	2,0566	1,9878	1,8918	1,8274	1,7808	1,7306	1,6759	1,6155	1,5617	1,5169	**29**
30	2,8807	2,4887	2,2761	2,1422	2,0492	1,9803	1,8841	1,8195	1,7727	1,7223	1,6673	1,6065	1,5522	1,5069	**30**
40	2,8353	2,4404	2,2261	2,0909	1,9968	1,9269	1,8289	1,7627	1,7146	1,6624	1,6052	1,5411	1,4830	1,4336	**40**
50	2,8087	2,4120	2,1967	2,0608	1,9660	1,8954	1,7963	1,7291	1,6802	1,6269	1,5681	1,5018	1,4409	1,3885	**50**
100	2,7564	2,3564	2,1394	2,0019	1,9057	1,8339	1,6949	1,9267	1,6124	1,5566	1,4943	1,4227	1,3548	1,2934	**100**

D. QUI-QUADRADO $\chi 2$

A tabela da distribuição χ^2 qui-quadrado fornece o valor crítico χ_C^2 para vários níveis de significância. Este valor indica o limite a partir do qual existe uma probabilidade, dada pelo nível de significância, de que as amostras não sejam originários de uma mesma população.

A tabela, aqui, é fornecida para os níveis de significância ***0,01***, ***0,05***, ***0,10*** e ***0,20***, indicando que em 1%, 5%, 10% e 20% dos casos, respectivamente, as amostras podem não ser oriundas de uma mesma população, a despeito do valor observado para χ_O^2 ser inferior ao valor crítico χ_C^2 encontrado na tabela. Em outras palavras, para duas amostras para as quais foi encontrado um valor de χ_O^2 menor que o valor crítico χ_C^2 tabelado, em cada um dos casos, em 99%, 95%, 90% e 80% das vezes elas provêm de uma mesma população.



A tabela fornece o valor crítico χ_C^2 para cada célula ou casela definida pela coluna do nível de significância escolhido e pela linha correspondente ao número de graus de liberdade dado por:

$$gl = (L - 1)(C - 1)$$

Onde ***L*** e ***C*** são o número de linhas e colunas, respectivamente, da grade do teste qui-quadrado.

DISTRIBUIÇÃO QUI-QUADRADO (χ^2)

	NÍVEL DE SIGNIFICÂNCIA				
gl	**0,20 (20%)**	**0,10 (10%)**	**0,05 (5%)**	**0,01 (1%)**	**gl**
1	1,6424	2,7055	3,8415	6,6349	**1**
2	3,2189	4,6052	5,9915	9,2104	**2**
3	4,6416	6,2514	7,8147	11,3449	**3**
4	5,9886	7,7794	9,4877	13,2767	**4**
5	7,2893	9,2363	11,0705	15,0863	**5**
6	8,5581	10,6446	12,5916	16,8119	**6**
7	9,8032	12,0170	14,0671	18,4753	**7**
8	11,0301	13,3616	15,5073	20,0902	**8**
9	12,2421	14,6837	16,9190	21,6660	**9**
10	13,4420	15,9872	18,3070	23,2093	**10**
11	14,6314	17,2750	19,6752	24,7250	**11**
12	15,8120	18,5493	21,0261	26,2170	**12**
13	16,9848	19,8119	22,3620	27,6882	**13**
14	18,1508	21,0641	23,6848	29,1412	**14**
15	19,3107	22,3071	24,9958	30,5780	**15**
16	20,4651	23,5418	26,2962	31,9999	**16**
17	21,6146	24,7690	27,5871	33,4087	**17**
18	22,7595	25,9894	28,8693	34,8052	**18**
19	23,9004	27,2036	30,1435	36,1908	**19**
20	25,0375	28,4120	31,4104	37,5663	**20**
21	26,1711	29,6151	32,6706	38,9322	**21**
22	27,3015	30,8133	33,9245	40,2894	**22**
23	28,4288	32,0069	35,1725	41,6383	**23**
24	29,5533	33,1962	36,4150	42,9798	**24**
25	30,6752	34,3816	37,6525	44,3140	**25**
26	31,7496	35,5632	38,8851	45,6416	**26**
27	32,9117	36,7412	40,1133	46,9628	**27**
28	34,0266	37,9159	41,3372	48,2782	**28**
29	35,1394	39,0875	42,5569	49,5878	**29**
30	36,2502	40,2560	43,7730	50,8922	**30**
35	41,7780	46,0588	49,8018	57,3420	**35**
40	47,2685	51,8050	55,7585	63,6908	**40**
gl	**0,20 (20%)**	**0,10 (10%)**	**0,05 (5%)**	**0,01 (1%)**	**gl**
	NÍVEL DE SIGNIFICÂNCIA				