Universidade Estadual de Santa Cruz
Departamento de Ciências Exatas e Tecnológicas
CET018 - Elementos de Estatística
Curso de Agronomia

Notas de aulas expandidas.

Prof. José Cláudio Faria

Ilhéus – Bahia
Março de 2009

Índice

NOTAS DO AUTOR ..... **IV**

LITERATURA, PROGRAMAS E RECURSOS COMPUTACIONAIS ..... **V**
Programas estatísticos usados na disciplina ..... **v**
Recursos disponíveis na WWW ..... **v**
Laboratórios virtuais selecionados disponíveis na Internet ..... **v**
Site para análises estatísticas on-line ..... **vi**
Exemplos de recursos disponíveis na WWW ..... **vi**

SIMBOLOGIA ADOTADA NO CURSO ..... **VII**

1. CALCULADORAS E APROXIMAÇÕES EM ESTATÍSTICA **1**
1.1. Calculadora adequada **1**
1.2. Comentários sobre os recursos básicos **1**
1.3. Aproximações **1**
1.4. Um teste **2**
1.5. O que não deve ser feito **2**

2. INTRODUÇÃO À ESTATÍSTICA **4**
2.1. Conceitos **4**
2.2. Definições básicas **4**
2.3. A natureza da análise estatística **6**
2.4. Dados **6**
2.5. Análise univariada vs. multivariada **7**
2.6. Objetivos da análise estatística **7**
2.7. Subdivisão e grandes áreas **7**
2.8. Objetos, variáveis e escalas **8**
2.8.1. Objetos 8
2.8.2. Variáveis 8
2.8.3. Escalas 8

3. NOÇÕES DE AMOSTRAGEM **11**
3.1. Introdução **11**
3.2. Amostragem: por que? **11**
3.3. Amostragem: como? **11**
3.4. Métodos probabilísticos **11**
3.4.1. Amostragem aleatória simples 11
3.4.2. Amostragem estratificada 12
3.4.3. Amostragem sistemática 13
3.4.4. Amostragem por áreas 14
3.4.5. Amostragem por conglomerados ou grupos 15
3.5. Métodos não probabilísticos **16**
3.5.1. Amostragem acidental ou de conveniência 16
3.5.2. Amostragem por julgamento 16
3.5.3. Amostragem por quotas 16

4. ESTATÍSTICA DESCRITIVA **18**
4.1. Conceitos **18**
4.2. Método de trabalho **18**
4.3. Apresentações tabulares **19**
4.3.1. Elementos mínimos 19
4.3.2. Séries 22
4.3.3. Erros mais comuns 24
4.4. Apresentações gráficas **26**
4.4.1. Elementos mínimos 26
4.4.2. Gráfico em colunas 26
4.4.3. Gráfico em barras 27
4.4.4. Gráfico em setores (pizza) 27
4.4.5. Gráfico polar 27
4.4.6. Gráfico em curvas 28
4.4.7. Erros mais comuns 28

4.5. Distribuição de freqüências ... 29
4.5.1. Tipos de variável ... 29
4.5.2. Organização dos dados ... 29
4.5.3. Distribuição de freqüências ... 29
4.5.4. Limites das classes ... 30
4.5.5. Número de classes (K) ... 30
4.5.6. Amplitude das classes (h) ... 30
4.5.7. Ponto médio das classes ... 30
4.5.8. Freqüência absoluta acumulada (F_{ac}) ... 31
4.5.9. Freqüência relativa (fi) ... 31
4.5.10. Histograma ... 31
4.5.11. Polígono de freqüências ... 32
4.5.12. Polígono de freqüência acumulada ... 32

5. MEDIDAS ESTATÍSTICAS ... 33
5.1. Introdução ... 33
5.2. Medidas de tendência central ... 33
5.2.1. Média aritmética ... 33
5.2.2. Média geométrica ... 35
5.2.3. Média harmônica ... 35
5.2.4. Mediana ... 35
5.2.5. Moda ... 38
5.3. Comparação entre as medidas de tendência central ... 39
5.3.1. Média ... 39
5.3.2. Mediana ... 39
5.3.3. Moda ... 39
5.4. Medidas de posição ou separatrizes ... 40
5.4.1. Quartis ... 40
5.4.2. Decis ... 40
5.4.3. Percentis ... 41
5.4.4. Situações de uso mais comuns destas medidas ... 42
5.5. Medidas de dispersão ... 43
5.5.1. Amplitude total ... 43
5.5.2. Desvio médio ... 43
5.5.3. Desvio quadrático médio ... 45
5.5.4. Variância ... 46
5.5.5. Desvio padrão ... 52
5.5.6. Desvio padrão relativo e coeficiente de variação ... 52

6. EXEMPLO DE ANÁLISE EXPLORATÓRIA DOS DADOS ... 55
6.1. Dados ... 55
6.2. Análise preliminar ... 55
6.3. Representação tabular dos dados ... 56
6.4. Representações gráficas dos dados ... 57
6.5. Medidas estatísticas ... 58
6.5.1. Tendência central ... 58
6.5.2. Separatrizes ou quantis ... 60
6.5.3. Medidas de dispersão ... 62

7. INTRODUÇÃO AO ESTUDO DE PROBABILIDADE ... 63
7.1. Caracterização de um experimento aleatório ... 63
7.2. Espaço amostral ... 64
7.3. Evento ... 65
7.4. Eventos mutuamente exclusivos ... 67
7.5. Conceito e definição de probabilidade ... 68
7.6. Principais teoremas da probabilidade ... 69
7.7. Probabilidades finitas dos espaços amostrais finitos ... 70
7.8. Espaços amostrais finitos equiprováveis ... 70
7.9. Probabilidade condicional ... 72
7.10. Teorema do produto ... 74
7.11. Independência estatística ... 74

8. VARIÁVEIS ALEATÓRIAS ... 77
8.1. Conceitos ... 77
8.2. Definição ... 77
8.3. Observações ... 78
8.4. Variável aleatória discreta (VAD) e contínua (VAC) ... 78
8.5. Função de probabilidades ... 78
8.6. Função de repartição ou distribuição acumulada ... 80
8.7. Função densidade de probabilidade ... 81
8.8. Esperança matemática (média ou valor esperado) ... 83
8.9. Variância ... 84
8.10. Covariância ... 86

9. CORRELAÇÃO LINEAR SIMPLES ... 88
9.1. Introdução ... 88
9.2. Definição ... 88
9.3. Conceitos e compreensão a partir de um exemplo ... 89
9.4. Pressuposições da correlação: ... 93

10. DISTRIBUIÇÃO NORMAL E NORMAL REDUZIDA ... 96
10.1. Introdução ... 96
10.2. Entendendo a distribuição ... 96
10.3. Simplificando a distribuição para facilitar o uso ... 97
10.4. Entendendo: distribuição normal vs. normal padrão ... 99
10.5. Uso da tabela de distribuição normal padrão ... 99
10.6. Uso da transformação para resolução de probabilidades ... 102

11. DISTRIBUIÇÃO AMOSTRAL DA MÉDIA E TESTE DE HIPÓTESES ... 105
11.1. Teorema do limite central (ou central do limite) ... 105
11.2. Teste de hipóteses ... 108
11.2.1. Hipótese ... 109
11.2.2. Teste de hipóteses ... 109
11.2.3. Tipos de hipóteses ... 109
11.2.4. Tipos de erros ... 109

12. DISTRIBUIÇÃO T DE STUDENT ... 115
12.1. Introdução ... 115
12.2. Aplicação: Intervalo de confiança para a média populacional (μ) ... 117
12.3. Exemplos de Intervalos de confiança para a média populacional ... 122

13. DISTRIBUIÇÃO χ^2 ... 125
13.1. Introdução ... 125
13.2. Entendendo a distribuição χ^2 ... 126
13.3. Exemplos de aplicação da distribuição do χ^2 ... 128
13.4. Teste qui-quadrado ... 129

14. DISTRIBUIÇÃO F DE SNEDECOR ... 131
14.1. Introdução ... 131
14.2. Entendendo a distribuição F ... 133
14.3. Precisão versus exatidão ... 135
14.4. Exemplo de aplicação da distribuição F ... 135

15. EXEMPLOS BÁSICOS DE INFERÊNCIA ESTATÍSTICA ... 140
15.1. Aplicação da distribuição t: teste de hipóteses de uma média com σ desconhecido ... 140
15.1.1. Solução encontrando a média crítica: ... 141
15.1.2. Solução encontrando o valor t crítico: ... 142
15.2. Aplicação da distribuição F: comparação de duas variâncias ... 143

16. TABELAS ESTATÍSTICAS ... I

NOTAS DO AUTOR

10ª edição

Estas anotações contêm, entre outras informações, as transparências utilizadas em sala de aula no curso de CET018 – Elementos de Estatística do curso de Agronomia da Universidade Estadual de Santa Cruz, Ilhéus, Bahia.

Sua reunião, no formato de uma apostila, tem como objetivo fornecer aos estudantes as informações essenciais discutidas em sala de aula, evitando as anotações excessivas, além de servir como referência para as consultas à literatura.

Em hipótese alguma este material deve ser considerado como suficiente para os estudos durante o transcorrer do curso. Adicionalmente, deve ser complementado, de forma pessoal, por anotações decorrentes das discussões.

Este material tem passado por freqüentes atualizações e correções de erros. Assim, é desaconselhado o uso de apostilas de edições anteriores.

O autor agradece quaisquer sugestões que possam contribuir para o aprimoramento do conteúdo.

José Cláudio Faria, 27/03/2009.

joseclaudio.faria@gmail.com

LITERATURA, PROGRAMAS E RECURSOS COMPUTACIONAIS

BUSSAB, W.O. & MORETTIN, P.A. **Estatística básica**. São Paulo, 4 ed. 1987. 321p.

BUCHAFT, G & KELNNER, S.R.O. **Estatística sem mistérios**. Rio de janeiro, Vozes, 1997. 991p.

FONSECA, J.S. & MARTINS, G.A. **Curso de estatística**. São Paulo, Atlas, 6 ed. 1996. 320p.

FREUND, J.E. & SIMON, G.A. **Estatística aplicada**. Porto Alegre, Bookman, 9 ed. 2000. 404p.

TRIOLA, M.F. **Introdução à estatística**. Rio de Janeiro, Livros Técnicos e Científicos Editora, 7 ed. 1998. 410p.

Observações:

- A literatura recomendada está listada por ordem alfabética dos autores.
- Recomendável a realização dos exercícios básicos propostos.
- Todos os livros razoáveis de estatística tratam do assunto.
- Em caso da opção para aquisição de um livro texto de referência para compor a biblioteca pessoal, pela abrangência, atualidade, qualidade de impressão e facilidade de uso, recomenda-se os livros de TRIOLA, M.F., e ou, o de FREUND, J.E. & SIMON, G.A., nesta ordem de preferência.

Programas estatísticos usados na disciplina

- R: http://www.r-project.org/
- BioEstat: http://www.mamiraua.org.br/download/

Recursos disponíveis na WWW

Em função do uso de recursos didáticos avançados, recomenda-se que, na medida do possível, os laboratórios virtuais de estatística disponíveis na internet sejam usados regularmente, uma vez que se constituem de inestimável valia para o aprendizado.

Alguns dos laboratórios disponibilizam programas e links para sites que permitem análises de dados em tempo real, podendo ser úteis no aprendizado, resoluções de exercícios e avaliações.

Laboratórios virtuais selecionados disponíveis na Internet

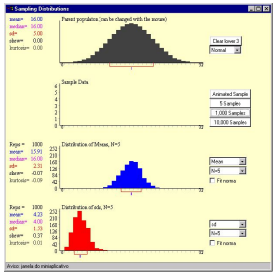
- http://www.ruf.rice.edu/~lane/rvls.html
- http://www.kuleuven.ac.be/ucs/java/
- http://www.stat.vt.edu/~sundar/java/applets/
- http://www.isds.duke.edu/sites/java.html

Site para análises estatísticas on-line

- http://www.webstatsoftware.com/

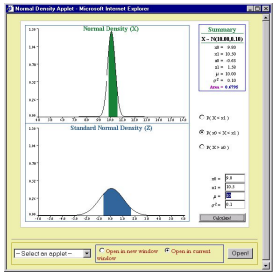
Exemplos de recursos disponíveis na WWW

Distribuições amostrais (excelente para entender o Teorema Central do Limite)



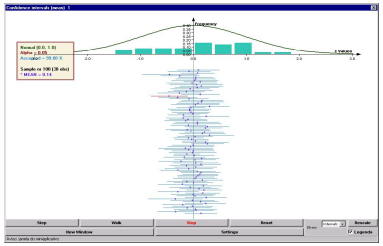
- http://www.ruf.rice.edu/~lane/stat_sim/sampling_dist/index.html

Distribuição normal



- http://www.stat.vt.edu/~sundar/java/applets/

Intervalo de confiança para a média populacional



- http://www.kuleuven.ac.be/ucs/java

SIMBOLOGIA ADOTADA NO CURSO

Medida estatística	**Populacional**	**Amostral (estimativa ou estatística)**
Média	μ, $\overline{Y}$	m, $\overline{y}$
Mediana	Md, $\tilde{Y}$	md, $\tilde{y}$
Moda	Mo	mo
Desvio médio	D_M	d_M
Desvio quadrático médio	DQ_M	dq_M
Amplitude total	AT	at
Variância	σ^2	s^2
Desvio padrão	σ	s
Desvio padrão relativo	DPR	dpr
Coeficiente de variação	CV	cv
Número de elementos	N	n
Correlação	ρ	r
Covariância	COV	cov
Parâmetro genérico	θ	$\hat{\theta}$

1. CALCULADORAS E APROXIMAÇÕES EM ESTATÍSTICA

A experiência no ensino da estatística tem mostrado que uma parte considerável das dificuldades no aprendizado e no rendimento acadêmico relaciona-se ao uso de calculadoras inadequadas, a subutilização dos recursos de calculadoras adequadas e a problemas de aproximações de valores intermediários.

O objetivo destas considerações iniciais é esclarecer previamente o tipo de calculadora científica necessária, o uso adequado dos recursos básicos e as aproximações normalmente usadas em estatística.

1.1. Calculadora adequada

Uma calculadora adequada, não somente para os cursos de estatística, mas para o decorrer das disciplinas dos cursos de graduação deve conter, no mínimo, os seguintes recursos:

- Medidas estatísticas básicas: média, variância, e ou, desvio padrão.
- Somatórios básicos: $\sum x \quad \sum y \quad \sum x^2 \quad \sum y^2 \quad \sum xy$
- Permitir a edição da série de dados armazenada na memória estatística.
- Endereços de memória para armazenar de 5 a 10 resultados parciais.
- Trabalhar com listas de números.

1.2. Comentários sobre os recursos básicos

Medidas estatísticas: são muito usadas e suas determinações, com calculadoras comuns, embora possível, são trabalhosas.

Somatórios básicos: são necessários em várias determinações.

Edição de dados: calculadoras que não possuem este recurso dificultam o trabalho com séries extensas de dados, pois, depois de inseridos na memória estatística, não é possível conferi-los nem corrigi-los. Isso ocasiona incerteza dos resultados e fadiga, devido à necessidade de repetição da digitação.

Endereços de memória: são muito usados para o armazenamento e recuperação de resultados intermediários usados em cálculos sucessivos.

Trabalhar com listas: permite que uma mesma operação seja feita em uma lista de dados, ao invés de elemento por elemento.

Exemplo:

$$\{4\ 3\ 5\ 6\} - 3 = \{1\ 0\ 2\ 3\}^2 = \{1\ 0\ 4\ 9\} \xrightarrow{\sum \text{lista}} = 14$$

1.3. Aproximações

Os cálculos estatísticos, embora simples, são em geral seqüenciais. Isto significa que resultados parciais são usados em novas determinações e assim por diante. Desta forma, os resultados intermediários devem ser sempre armazenados em variáveis de memória com todos os decimais possíveis e usados dessa forma. Apenas no fim dos cálculos é que o resultado final deve ser aproximado para o número de casas decimais suficiente para o problema numérico. Em geral, duas casas ou três decimais são suficientes para a maioria dos problemas acadêmicos.

Se estes cuidados não forem tomados, as aproximações sucessivas levam a distorções consideráveis no resultado final, podendo levar a conclusões equivocadas.

1.4. Um teste

Vamos supor duas séries de dados com 15 elementos cada uma:

$$A = \{12{,}31\ 14{,}52\ 13{,}23\ 14{,}71\ 16{,}82\ 19{,}33\ 14{,}99\ 17{,}98\ 13{,}67\ 14{,}16\ 14{,}85\ 14{,}63\ 13{,}24\ 17{,}65\ 13{,}26\}$$
$$B = \{14{,}13\ 16{,}94\ 11{,}55\ 13{,}36\ 18{,}17\ 13{,}28\ 14{,}19\ 16{,}28\ 12{,}17\ 18{,}46\ 12{,}55\ 11{,}34\ 12{,}13\ 14{,}22\ 18{,}11\}$$

Os seguintes procedimentos são necessários:

a. Calcular a média aritmética simples de cada série:

$$m_A = 15{,}02$$
$$m_B = 14{,}46$$

b. Diminuir cada valor das séries de suas respectivas médias:

$$A = \{(12{,}31 - 15{,}02)\ (14{,}52 - 15{,}02)\ \ldots\ (13{,}26 - 15{,}02)\}$$
$$B = \{(14{,}13 - 14{,}46)\ (16{,}94 - 14{,}46)\ \ldots\ (18{,}11 - 14{,}46)\}$$

c. Para cada série elevar ao quadrado as diferenças e efetuar o somatório:

$$A = \{(-2{,}71)^2 + (-0{,}50)^2 + \ldots + (-1{,}77)^2\}$$
$$B = \{(-0{,}33)^2 + (2{,}48)^2 + \ldots + (3{,}65)^2\}$$

d. Dividir cada resultado da etapa anterior (c) por 14:

$$A = \frac{57{,}40}{14} = 4{,}10$$
$$B = \frac{87{,}91}{14} = 6{,}28$$

e. Dividir o maior pelo menor valor dos encontrados na etapa anterior (d) e expressar o resultado final com duas casas decimais:

$$\frac{6{,}28}{4{,}10} = 1{,}53$$

Este é o resultado trabalhando com todos os resultados intermediários em variáveis de memória. Realizar o teste considerando que afastamentos implicaram na adoção de procedimentos inadequados que necessitam ser revistos e melhorados.

1.5. O que não deve ser feito

a. Não armazenar os valores das médias em variáveis de memória;

b. Subtrair os valores das médias aproximadas (15,02 e 14,46) e não dos valores reais (15,02333... e 14,458666...);

c. Redigitar as diferenças aproximadas para elevar ao quadrado e depois redigitar novamente os valores para efetuar o somatório;

d. Redigitar novamente os resultados anteriores para efetuar a divisão por 14;

e. Redigitar os valores aproximados anteriores para efetuar a divisão final.

É fácil perceber que devido às aproximações de resultados intermediários, pode-se chegar a resultados bem diferentes do real. Adicionalmente, as digitações ocasionam erros (adicionais aos das aproximações) além da fadiga desnecessária.

Alguns estudantes realizam estes cálculos armazenando os valores das médias em variáveis de memória, digitam cada valor da série, que é subtraído da média, elevado e armazenado na memória de soma (M+). Posteriormente a soma final é recuperada e dividida por 14. Embora seja um paliativo, este procedimento encontra-se muito aquém do uso eficiente dos recursos disponíveis. Nas resoluções de exercícios toma muito tempo e, em geral, compromete as avaliações.

Existem varias formas alternativas de realizar os cálculos anteriores utilizando os recursos das calculadoras científicas. A mais simples e usual é informar o valor de cada série na memória estatística e solicitar a medida estatística de dispersão dos dados em torno da média (variância amostral), armazenar cada valor (4,10 e 6,28) em variáveis de memória e, posteriormente, realizar a divisão entre elas.

Outra forma interessante é trabalhar com as séries na forma de listas.

Exemplo:

$$\{12{,}31\ 14{,}52 \ldots 13{,}26\} - 15{,}02 = \{-2{,}71\ -0{,}50 \ldots -1{,}76\}^2 = \{7{,}36\ 0{,}25 \ldots 3{,}11\} \xrightarrow{\sum Lista} \frac{57{,}40}{14} = 4{,}10$$

Deve-se ter em mente que, além da necessidade da calculadora dispor dos recursos necessários, é importante saber usá-los adequadamente. Assim, cada usuário deve estudar o manual de instruções de sua calculadora pessoal a fim de que possa ter clareza e domínio sobre os recursos disponíveis.

2. INTRODUÇÃO À ESTATÍSTICA

"Usa a Estatística do mesmo modo que um bêbado, os postes – mais pelo apoio do que propriamente pela iluminação."

A ênfase da disciplina é na compreensão e no uso adequado dos fundamentos estatísticos, não na memorização de fórmulas e conceitos, comumente medidos como sinônimo de aprendizado.

2.1. Conceitos

A palavra estatística significa, originalmente, uma coleção de informações de interesse para o Estado sobre a população e a economia.

Dessa modesta origem a estatística cresceu e se desenvolveu até tornar-se um método de análise que, hoje, encontra aplicação em todos os ramos da ciência.

a. A estatística é arte e a ciência de coletar, analisar, apresentar e interpretar dados.

b. A estatística é a linguagem universal da ciência. O uso adequado dos métodos estatísticos permite descrever com precisão os objetos da pesquisa científica, tomar decisões e fazer estimativas.

c. O campo da análise estatística é relacionado à coleção, organização e interpretação de dados de acordo com procedimentos bem definidos.

2.2. Definições básicas

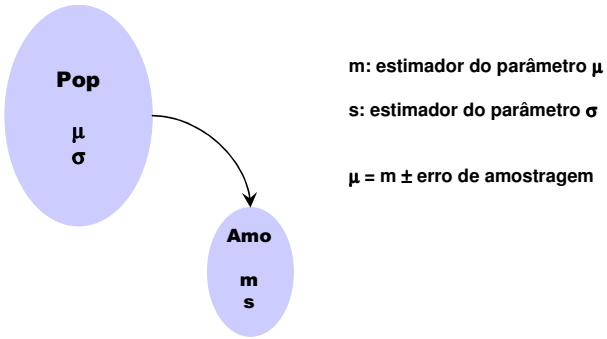


Figura 1.1 – Ilustração de população (parâmetros) e amostra (estimativas).

Um dos grandes objetivos da estatística é a tomada de decisão, em um processo particular qualquer, a respeito de uma população, em geral desconhecida, realizada a partir dos dados amostrais.

População: conjunto, finito ou infinito, de indivíduos, objetos ou medidas que apresentam pelo menos uma característica observável em comum.

Exemplos:

- Os corpos celestes no universo
- Os coqueiros do sul e extremo sul da Bahia
- O rendimento acadêmico em cálculo I dos alunos do curso de Agronomia de determinada universidade ou conjunto de universidades

Amostra: consiste de uma parte (subconjunto) dos indivíduos, objetos ou medidas, selecionados a partir da população.

Parâmetro: qualquer quantidade numérica medindo algum aspecto de uma população.

Exemplos:

- Número de indivíduos ou observações: N
- Média: μ
- Mediana: MD
- Moda: MO
- Variância: σ^2
- Desvio padrão: σ
- Proporção: Π
- Correlação: ρ

Estimador do Parâmetro: qualquer quantidade numérica medindo algum aspecto de uma população obtido, ou estimado, a partir de uma amostra representativa.

Exemplos:

- Número de indivíduos ou observações: n
- Média: m
- Mediana: md
- Moda: mo
- Variância: s^2
- Desvio padrão: s
- Proporção: π
- Correlação: r

Dedução: envolve uma argumentação do geral para o específico – isto é, da população para a amostra.

Indução: envolve uma argumentação do específico para o geral – isto é, da amostra para a população.

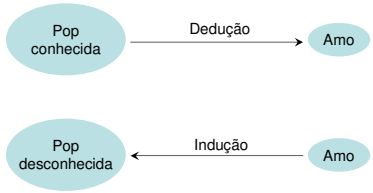


Figura 1.2 – Ilustração dos conceitos dedução e indução.

2.3. A natureza da análise estatística

Os princípios e conceitos básicos da análise estatística são relativamente simples e pouco numerosos, entretanto eles geram uma vasta variedade de técnicas de análise de acordo com a natureza dos dados e das questões.

O maior desafio é a habilidade para reconhecer que tipo de análise é mais adequado em cada situação e como interpretar os resultados.

O advento do computador pessoal tem removido o consumo de tempo e minimizado os aspectos tediosos - que dizem respeito à manipulação e cálculos de medidas estatísticas associadas aos grandes conjuntos de dados - das análises, permitindo a concentração de esforços em sua essência: princípios, lógica associada aos vários métodos, interpretações e aplicações.

2.4. Dados

Dados são os objetos centrais das análises estatísticas e podem ser conceituados como observações feitas sobre o ambiente.

Essas observações são resultantes de mensurações feitas usando instrumentos ou processos de medida (tempo, massa, distância, etc) ou de contagem.

Essas mensurações respondem, em geral, às perguntas: o que, quando, quanto, onde, tipo, intensidade, etc.

Essas observações necessitam ser convertidas em números para evitar a ambigüidade, ou diferentes interpretações das palavras.

Embora palavras como pouco, muito, usualmente, etc, contenham alguma informação, elas não são tão precisas, ou unicamente interpretadas, como através de procedimentos usando números em operações padronizadas.

"Dados podem ser encontrados onde quer que olhemos. Não existe parte de nosso ambiente que não seja uma fonte potencial de dados: nós mesmos, outros indivíduos, unidades familiares, sociedades, culturas, raças, locais, países, planetas, vulcões, moléculas de DNA, partículas atômicas, solos, medicina, plantas e animais, órgãos, células, escolas religiões, etc - em síntese, todos os aspectos de nossa existência." (Kachigan, 1980)

2.5. Análise univariada vs. multivariada

Quando se está interessado em uma característica isolada de um conjunto de objetos, desconsiderando as outras características, o domínio é o da análise univariada.

A análise multivariada, por outro lado, é o ramo da análise estatística que lida simultaneamente com duas ou mais características mensuradas em um conjunto de objetos.

2.6. Objetivos da análise estatística

As observações sobre o ambiente são convertidas em números, estes são manipulados e organizados, segundo procedimentos bem definidos, os resultados podem tornar o ambiente mais compreensivo que antes da análise.

Numa visão mais ampla, a retirada de conclusões e o melhor entendimento da fonte de dados, é o objetivo final da análise estatística.

A manipulação e organização dos dados inevitavelmente atende a um ou mais dos três objetivos básicos de uma análise:

a. Redução dos dados: redução de grandes conjuntos de dados em pequenos conjuntos, que descrevem as observações, sem sacrifício de informações críticas.
b. Inferência: possibilita a tomada de decisão sobre os grandes grupos de observações com base na mensuração de apenas uma parte, ou fração, desse.
c. Identificação de associações ou relacionamentos: o conhecimento sobre um conjunto de variáveis permite descrever ou predizer (inferir) sobre um outro conjunto de variáveis.

2.7. Subdivisão e grandes áreas

A estatística pode ser grosseiramente subdividida em quatro grandes áreas:

a. Amostragem e planejamento de experimentos: tratam dos métodos científicos de amostragem e do planejamento de experimentos.
b. Estatística descritiva ou análise exploratória dos dados: trata dos métodos tabulares, gráficos e numéricos usados para sintetizar dados sem o sacrifício de informações relevantes.
c. Probabilidade: ramo da matemática que trata do estudo da incerteza, ou das medidas numéricas da plausibilidade da ocorrência de eventos. Fornece a base matemática para a inferência estatística, ou seja, a tomada de decisão em situações de incerteza.
d. Estatística inferencial: processo de utilizar dados obtidos a partir de amostras para fazer estimativas ou testar hipóteses sobre as características das populações.

2.8. Objetos, variáveis e escalas

2.8.1. Objetos

Tudo sobre o qual as observações podem ser feitas: indivíduos, entidades, unidades de observação física ou biológica, localização geográfica, período de tempo, eventos, etc.

2.8.2. Variáveis

O fato de mensurar objetos em relação às suas características implica que os mesmos diferem nessas características, ou, as características podem assumir deferentes valores.

Estas características, propriedades ou atributos que podem assumir dois ou mais diferentes valores são denominadas variáveis.

2.8.3. Escalas

Esquema usado para a representação dos possíveis valores de uma variável.

2.8.3.1. Nominal

Os objetos possuem características que se diferenciam apenas pelo tipo.

Exemplo:

Ocupação, tipo sanguíneo, religião, raça, variedade.

Diferenças entre os valores da variável não podem ser interpretadas em termos quantitativos (i.e. quanto se diferenciam), são também denominadas variáveis qualitativas ou categóricas.

2.8.3.2. Ordinal

Os valores numéricos indicam hierarquia dos níveis da variável em questão

(i.e, se A > B > C então A > C).

A limitação é que não podem ser feitas inferências sobre o grau de diferença entre os valores da escala.

Exemplo:

Escala de dureza dos minerais (Moh).

Os números 1 a 10 são atribuídos respectivamente ao talco, gesso, calcita, fluorita, apatita, feldspato, quartzo, topázio, safira e diamante. Podem-se estabelecer desigualdades, mas não quantificá-las, pois diferenças iguais entre valores ordinais não implicam necessariamente em um mesmo significado quantitativo, i.e, a diferença de dureza entre o diamante e a safira (10 - 9 =1) é muito maior que a diferença entre o gesso e o talco (2 – 1 = 1).

Devido à limitação de se interpretar diferenças quantitativas entre os valores ordinais, assim como na escala nominal, elas são chamadas escalas não métricas.

2.8.3.3. Intervalar

É considerada uma escala métrica, pois diferenças iguais entre os valores da escala possuem a mesma magnitude, independente de em que ponto da escala a mesma diferença é considerada.

Exemplo:

$$59 - 56 = 3 \qquad 117 - 114 = 3$$

Variáveis medidas por esta escala são consideradas quantitativas. Entretanto, devido à arbitrariedade do ponto zero, que não representa realmente a quantidade zero, proporções entre valores não tem significado, o que é uma limitação da escala.

Exemplo:

$90^oF - 80^oF$ representa a mesma quantidade de calor que $60^oF - 50^oF$, entretanto não é verdadeiro que 80^oF possa ser interpretado como duas vezes 40^oF:

$$80^oF = 26{,}67^oC$$
$$40^oF = 4{,}44^oC$$
$$\frac{40}{80} = 0{,}5 = 50\% \qquad \frac{4{,}44}{26{,}67} = 0{,}17 = 16{,}67\%$$

2.8.3.4. Proporcional

Armazena mais informações que as anteriores, possuindo as características desejáveis de cada uma delas e não possuindo as limitações de nenhuma delas.

Iguais proporções têm o mesmo significado devido à presença do ponto zero genuíno na escala.

Exemplo:

$$\frac{18}{36} = \frac{50}{100} = \frac{5.000}{10.000} = 0{,}5 = 50\%$$

É considerada uma escala métrica e as mensurações são consideradas quantitativas.

Exemplos:

- Medidas de comprimento: polegada, metro
- Medidas de tempo: segundo
- Desempenho de vendas: reais, dólar
- Medidas de área: hectares, m^2.

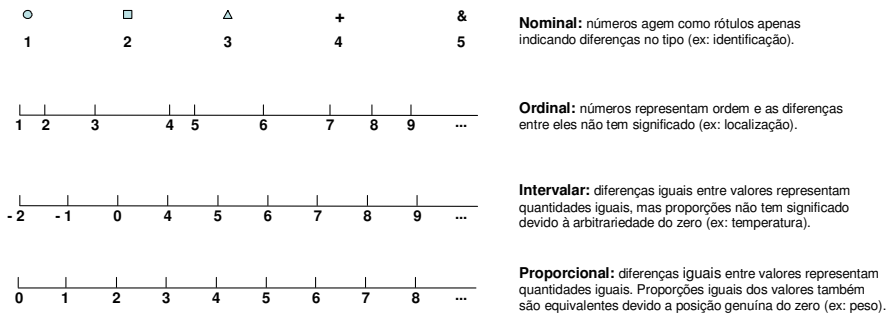


Figura 1.3 – Ilustração comparativa dos tipos de escalas.

Uma escala proporcional que não possui unidade é a contagem de freqüências e o percentual. A contagem responde a questão “quanto” ao invés de “que quantidade”.

2.8.3.5. Escalas binárias ou dicotômicas

Em adição à classificação das variáveis em termos de sua natureza, pode-se classificá-las com base em quantos valores ela pode assumir.

Por definição, o mínimo número de valores que uma variável pode assumir é dois. Essas variáveis são denominadas binárias ou dicotômicas.

Exemplos:

- Sexo: macho ou fêmea
- Aprovação: sim ou não.

3. NOÇÕES DE AMOSTRAGEM

3.1. Introdução

Amostragem consiste na escolha criteriosa dos elementos da população a serem submetidos ao estudo.

3.2. Amostragem: por que?

Por várias razões, ao invés de pesquisar toda uma população, extraí-se uma amostra:

a. Limitações de recursos: orçamentários, humanos, tempo.

b. Escassez de dados: fenômenos raros.

c. Testes destrutivos: eliminação da população.

3.3. Amostragem: como?

Os métodos mais comuns de amostragem são divididos em duas categorias:

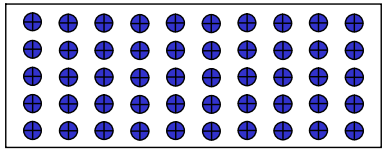
Probabilísticos: a probabilidade de cada elemento da população ser incluído na amostra é, a priori, conhecida.

Não Probabilísticos: não se tem conhecimento da probabilidade de escolha de determinado elemento da população.

3.4. Métodos probabilísticos

3.4.1. Amostragem aleatória simples

É o método de selecionar, sem reposição, n elementos de uma população de tamanho N, conhecido e finito, onde cada elemento tem a mesma chance de ser selecionado:



3.4.1.1. Procedimentos

a. Enumerar os N elementos da população.

b. Sortear, sem reposição, n números compreendidos entre 1 a N.

- Excel: selecionar as células onde será feito o sorteio aleatório e digitar, na barra de fórmulas:=ALEATÓRIOENTRE(num.inferior;num.superior).
- Tabela de números aleatórios (TNA): começar em determinado lugar da TNA e a partir deste ponto, retirar os números de modo que o número de dígitos abranja o maior número desejado.

c. Os elementos correspondentes aos números escolhidos formarão a amostra n elementos.

Exemplo:

Se a população possui 1.000 indivíduos, devemos numerá-los de 1 a 1.000.

Considerar os três primeiros números da TNA (000 = 1.000).

Seguir qualquer direção na TNA: horizontal, vertical ou oblíqua.

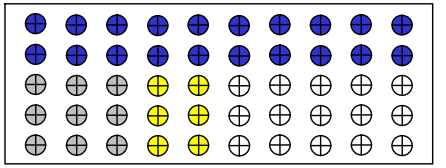
Ao chegar ao final da linha ou coluna muda-se a direção e prossegue-se como antes.

Números repetidos são desprezados.

3.4.2. Amostragem estratificada

A população é dividida em grupos (estratos) que são mutuamente exclusivos de acordo com alguma(s) característica(s) relevante(s). Posteriormente uma amostra aleatória simples é retirada de cada estrato.

O objetivo é melhorar a representatividade da amostra em relação à população, levando a estimativas mais confiáveis dos que as obtidas por outros métodos:



As amostras aleatórias podem ou não ser proporcionais ao tamanho de seus estratos correspondentes, de acordo com os objetivos do estudo.

Dentre as amostras probabilísticas, é a que proporciona estimativas mais seguras acerca da população, especialmente quando se sabe quantos elementos da população fazem parte de cada estrato.

3.4.2.1. Procedimentos

Para uma amostragem estratificada proporcional os seguintes procedimentos são realizados:

a. Dividir a população em L subpopulações chamadas estratos.

b. Realizar a amostragem aleatória simples dentro de cada estrato, observando os seguintes critérios:

Considerando:

N = número de elementos da população

L = número de estratos

N_i = número de elementos do extrato ($N = N_1 + N_2 + ... + N_L$)

n = tamanho da amostra a ser selecionada:

c. Calcular a fração f da amostragem dada por:

$$f = \frac{n}{N}$$

d. Calcular o número de elementos a serem sorteados em cada estrato:

$$n_1 \cong N_1 \cdot f$$
$$n_2 \cong N_2 \cdot f$$
$$\ldots$$
$$n_L \cong N_L \cdot f$$

$$com$$

$$n = n_1 + n_2 + \ldots + n_L$$

Exemplo:

Intenção de votos para governador em 1990 em São Paulo (DataFolha).

Os municípios do estado foram classificados em regiões homogêneas segundo a situação geográfica e o nível socioeconômico.

O DataFolha entrevistou 3.900 eleitores, sorteados em 98 municípios de todo o Estado de São Paulo.

3.4.3. Amostragem sistemática

De posse de uma listagem dos elementos da população, resulta da escolha sistemática, a partir de um número inicial qualquer, onde os demais elementos são selecionados de forma intervalar.

3.4.3.1. Procedimentos

a. Definir a percentagem P% de elementos da população que farão parte da amostra.

b. Obter um valor k inteiro, dado por:

$$k \cong \frac{1}{P} \cdot 100$$

c. Sortear um número r inteiro entre 1 e k.

d. A amostra será composta pelos elementos de ordem:

$$r,\ r+k,\ r+2k,\ r+3k, \ldots$$

Exemplo:

Uma loja deseja conhecer o perfil dos seus clientes e tem condições de entrevistar 20% dos mesmos.

Os compradores que visitaram a loja num certo dia, por ordem de chegada, foram:

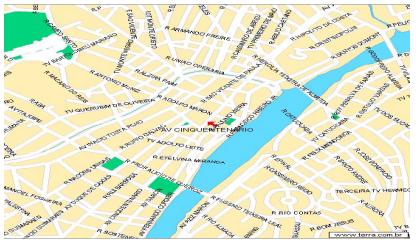
01 JCF	02 KLJ	03 OMI	04 JUI	05 PLW	06 MNH	07 QUR
08 STR	09 JOY	10 LKP	11 NWO	12 GTR	13 LER	14 GFF
15 EQI	16 UPL	17 NMQ	18 JWF	19 DFR	20 PUB	21 NHU
22 PPO	23 QDA	24 NKP	25 HYU	26 DRQ	27 ACD	28 BCV
29 NHU	30 PLK	31 MHZ	32 POP	33 HWR	34 RER	35 BDB

Obtenção do valor k: $k \cong \frac{1}{20} \cdot 100 \cong 5$

Sorteio de r entre 1 e k (entre 1 e 5) = 3

A amostra será composta pelos elementos: 03-OMI, 08-STR, 13-LER, 18-JWF, 23-QDA, 28-BCV e 33-HWR.

3.4.4. Amostragem por áreas



Até esse ponto, a discussão se limitou à amostragem por meio de listas que identificam cada membro da população.

Entretanto, em muitos casos, esta lista não é disponível, mostra-se inadequada ou obsoleta.

A amostragem por áreas baseia-se no sorteio de residências ou pessoas com base no mapa da região a ser pesquisada.

A população regional é transformada numa população de áreas para a qual uma lista, em forma de mapa, existe.

3.4.4.1. Procedimentos

a. Dividir a população (cidade) em áreas menores (quadrículas ou quarteirões) e sortear algumas.

b. Continuar a divisão (quadrículas ou quarteirões em casas) e sortear algumas.

Exemplo:

A cidade do Recife foi dividida em 50 quadrículas.

Utilizando uma TNA 10 dentre as 50 quadrículas foram sorteadas.

Em cada quadrícula sorteou-se 15 pessoas alfabetizadas, com idade entre 18 e 45 anos, para comporem a amostra.

3.4.5. Amostragem por conglomerados ou grupos

O objetivo principal é selecionar amostras quando a população se encontra muito dispersa em termos geográficos.

O princípio da conglomeração se opõe ao da estratificação, pois o que se busca é a heterogeneidade: quanto maior a variabilidade, maior a precisão.

3.4.5.1. Procedimentos

a. Seleciona-se a amostra por meio de vários estágios, indo das unidades mais amplas às menores, até se chegar aos elementos da população que se deseja estudar.

b. Em cada estágio utiliza-se um tipo de seleção probabilística.

A perda da precisão implícita é compensada por sua simplicidade e pela diminuição dos custos operacionais.

Exemplo:

Avaliar a expectativas profissionais dos alunos que cursam Agronomia nos estados do Nordeste.

Elaborar uma lista das faculdades/universidades e realizar o sorteio.

De posse das listas das faculdades/universidades sorteadas, elaborar uma lista de turmas.

Sortear as turmas e, posteriormente, os alunos dentro de cada turma.

3.5. Métodos não probabilísticos

Risco de tendenciosidade.

Problemas relacionados às generalizações dos resultados.

Facilidade, economia e rapidez.

A escolha de um método mais sensível para a coleta de dados pode compensar, em parte, o método de amostragem não muito adequado.

3.5.1. Amostragem acidental ou de conveniência

Os elementos da amostra são escolhidos por serem os mais acessíveis ou fáceis de serem avaliados.

São considerados os casos até que a amostra atinja o tamanho desejado.

Exemplo:

Investigações utilizando como amostra pessoas que passam por determinado lugar.

3.5.2. Amostragem por julgamento

Consiste na escolha dos elementos da amostra por um especialista no assunto, que seleciona os elementos que julga os mais apropriados e representativos para o estudo em questão.

Exemplo:

Em estudos antropológicos podem ser entrevistados os elementos mais proeminentes da cultura que está sendo investigada.

Amostragem intencional ou proposital

Quando o pesquisador está interessado na opinião de determinados elementos da população, considerados como representativos da mesma.

Exemplo:

Realização de pesquisas preferenciais em uma cidade que já se sabe de antemão que seus resultados se aproximam dos resultados gerais da nação.

3.5.3. Amostragem por quotas

É o método não-probabilístico mais empregado, pois acrescenta segurança, incluindo na amostra vários estratos da população.

Difere da amostragem probabilística estratificada pela ausência de escolha aleatória.

Os vários estratos da população não são, necessariamente, amostrados em sua proporção correta.

Deve haver elementos em número suficiente para que seja possível uma estimativa do(s) valor(es) do(s) estrato(s) na população.

3.5.3.1. Procedimentos

a. Seleção das características da população consideradas relevantes para o estudo, escolhidas de modo a se associarem às características que se pretende investigar.

b. Dispor de informações atualizadas sobre sua distribuição na população.

c. Determinação da proporção de cada grupo de características na população com base em dados censitários, cadastros, listagens, etc.

d. Estruturação de células resultantes da divisão do universo nos subuniversos que o compõem.

Exemplo:

Seleção de amostra para avaliação da intenção de votos para governador em 1990 em São Paulo pelo IBOPE:

I. Selecionadas as cidades da região metropolitana e interior.

II. Escolha dos setores a serem pesquisados.

III. Estabelecidos os critérios para a escolha do eleitor: Sexo, faixa etária, nível de instrução, nível sócio-econômico, ocupação profissional.

IV. Atribuição de cotas a cada entrevistador.

De acordo com o gerente de planejamento do IBOPE, foram entrevistados em cada pesquisa 2.000 pessoas, em 88 cidades.

4. ESTATÍSTICA DESCRITIVA

A estatística está interessada nos métodos científicos para a coleta, organização, apresentação e análise de dados, bem como na obtenção de conclusões válidas e na tomada de decisões razoáveis baseadas em tais análises.

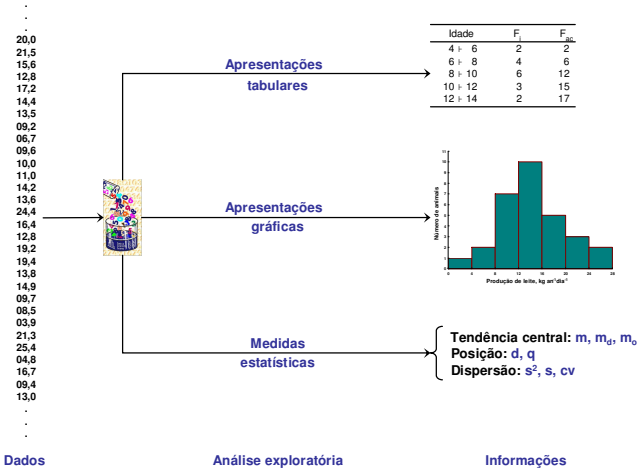


Figura 4.1 – Ilustração da análise exploratória como um conjunto de procedimentos bem definidos que permitem ir dos dados às informações.

4.1. Conceitos

Estatística descritiva (ou análise exploratória dos dados) é a parte da estatística que procura descrever e avaliar um grupo.

O grupo pode ser uma população ou uma amostra.

Em se tratando de amostras, este ramo da estatística não permite retirar quaisquer conclusões ou inferências sobre um grupo maior (população).

4.2. Método de trabalho

a. Definição do problema: identificação das questões a serem investigadas.

b. Planejamento: estabelecimento dos mecanismos de coleta e apresentação dos resultados.

c. Coleta de dados: consiste na apreensão (busca ou compilação) das informações necessárias ao estudo (dados das variáveis).

d. Crítica dos dados: eliminação de erros capazes de provocar futuros enganos de apresentação e análise, através da revisão crítica dos dados e eliminação dos valores estranhos ao levantamento.

e. Apresentação dos dados: organização dos dados de maneira prática e racional, buscando propiciar um melhor entendimento do fenômeno em estudo.

f. Descrição dos dados: feita por meio de medidas que os representem de forma sumária e escolhidos de acordo com os objetivos do pesquisador.

4.3. Apresentações tabulares

São formas não discursivas de apresentação de informações que tem por finalidade a descrição, e ou, o cruzamento de dados numéricos.

A Associação Brasileira de Normas Técnicas (ABNT) é o órgão responsável pela normalização técnica no País, tendo sido fundada em 1940. Regulamenta a construção e apresentação das composições gráficas e tabulares.

NBR6029: informação e documentação – livros e folhetos – apresentação.

NBR6022: apresentação de artigos em publicações periódicas (normas para apresentações tabulares e figuras).

Será dada ênfase nas regulamentações comuns usadas na rotina acadêmica devendo-se consultar a literatura indicada para ampliar o nível de detalhamento.

Apresentações tabulares: tabelas vs. quadros:

a. Tabelas: informações tratadas estatisticamente.

b. Quadros: informações textuais agrupadas em colunas.

A apresentação tabular deve sintetizar os dados de modo a facilitar a leitura e propiciar maior rapidez na interpretação das informações.

Deve-se primar por apresentações simples que possibilitem ao leitor a compreensão do fenômeno em estudo sem muito esforço.

Cada apresentação tabular deve ser vista como uma unidade de informação e, tanto quanto possível, ser auto-explicativa, dispensando consultas ao texto.

4.3.1. Elementos mínimos

a. Número: Usado para identificar a composição.

Exemplo:

Tabela **4.1**

Comentário: Algumas publicações adotam que a palavra TABELA ou QUADRO deve ser preferencialmente escrito com letras maiúsculas. O mais importante, entretanto, é a padronização ao longo do texto.

b. Título: Composto da descrição do conteúdo e o local de referência.

Exemplo:

Tabela 4.1 – **o que e onde**

c. Data de referência: Identifica o período referente aos dados e as informações.

Exemplo:

Tabela 4.1 – o que e onde – **quando**

Observações:

- Deve preferencialmente ser escrito com letras maiúsculas ou seguindo o mesmo padrão definido na escrita do número. Adotou-se na apostila o padrão mais comum encontrado em periódicos de publicação científica, ou seja, apenas as iniciais em maiúsculo.
- Os elementos do título devem ser separados por hífen.
- Quando a descrição do conteúdo utilizar mais de uma linha, a segunda e as demais linhas devem ser alinhadas sob a primeira letra da primeira linha do título.

Exemplo:

Tabela 4.1 – Número de estabelecimentos destinado exclusivamente à comercialização de hortifrutigranjeiros fiscalizados por região administrativa, Bahia – 2003

4.3.1.1. Corpo da composição

a. Cabeçalho: parte superior da composição que especifica o conteúdo das colunas, podendo ser constituído de um ou mais níveis.

Exemplo:

Áreas de ensino		Matrículas	
Ciências exatas	Ciências sociais	2002	2003
	...		

b. Coluna indicadora: especifica o conteúdo das linhas.

Exemplo:

Área de ensino	Matrículas
Ciências biológicas	205
Letras	104
Artes	302

c. Linha do corpo: conjunto de elementos dispostos horizontalmente no corpo da composição onde são registrados os dados numéricos e informações.

Exemplo:

Tratamentos	Repetições					
	1	2	3	4	5	6
A	**58**	**49**	**51**	**56**	**50**	**48**
B	**60**	**55**	**66**	**61**	**54**	**61**
C	**59**	**47**	**44**	**49**	**62**	**60**
D	**45**	**33**	**34**	**48**	**42**	**44**

d. Coluna do corpo: conjunto de elementos dispostos verticalmente no corpo da composição onde são registrados os dados numéricos e informações.

e. Traço: delimitam obrigatoriamente o cabeçalho e a finalização da composição.

f. Fonte: consiste na indicação da(s) entidade(s) responsável(is) pelo fornecimento ou elaboração dos dados e informações contidos na composição.

Observações:

- Deve ser apresentado separado do nome do órgão ou pessoa física responsável pelos dados por dois pontos e um espaço, sem ponto final.
- No caso de varias fontes eles devem vir separadas por vírgula.
- Caso os dados sejam extraídos de publicações, deve-se indicar sua referência completa.
- Quando se tratar de pessoa física, responsável pelos dados levantados e apresentados, comum em trabalhos acadêmicos (monografias, teses e outros), deve-se utilizar como fonte a expressão o autor.
- No caso em que o próprio autor está apresentando dados levantados via pesquisa de campo (utilização de formulários, questionários), pode-se usar tal expressão como fonte.

Exemplos:

Fonte: IBGE

Fonte: SEGRAD, PROPP

Fonte: IPARDES. Indicadores analíticos: Paraná. Curitiba, 1994

Fonte: O autor

Fonte: Pesquisa de campo

g. Nota: apresenta as informações de natureza geral, destinadas a conceituar ou esclarecer o conteúdo, ou indicar a metodologia adotada na coleta ou na elaboração dos dados. E apresentada logo abaixo da fonte.

h. Nota específica: apresenta as informações destinadas a descrever conceitos ou esclarecer dados sobre uma parte ou item específico da composição.

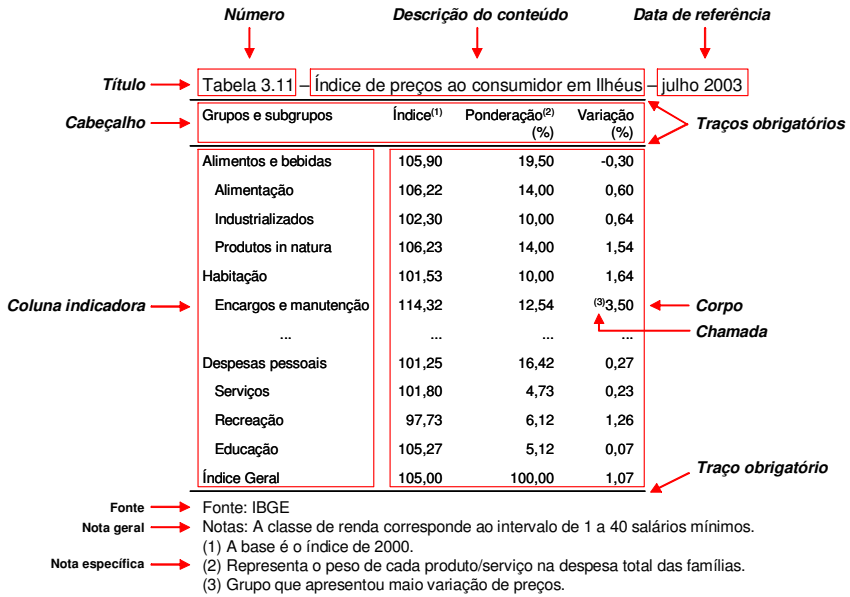


Figura 4.2. Ilustração das partes que compõe uma composição tabular.

4.3.2. Séries

4.3.2.1. Série cronológica, temporal, evolutiva ou histórica

É a série em que os dados são observados segundo a época de ocorrência:

Tabela 4.2 – Vendas da companhia Alfa – (1970 a 1977)

Ano	Vendas (em R$ 1.000,00)
1970	2.181
1971	3.948
1972	5.642
1973	7.550
1974	10.009
1975	11.728
1976	18.873
1977	29.076

Fonte: Departamento de Marketing da Companhia

4.3.2.2. Série geográfica ou de localização

É a série em que os dados são agrupados segundo a localidade de ocorrência:

Tabela 4.3 – Empresas fiscalizadas pelo INAMPS – 1973

Regiões	Empresas fiscalizadas
Norte	7.495
Nordeste	107.783
Sudeste	281.207
Sul	53.661
Centro-Oeste	15.776

Fonte: Mensário Estatístico 259/260

4.3.2.3. Série específica

É a série em que os dados são agrupados segundo a modalidade de ocorrência:

Tabela 4.4 – Matrícula no ensino de terceiro grau, Brasil – 1975

Áreas de ensino	Matrículas
Ciências biológicas	32.109
Ciências exatas e tecnologia	65.949
Ciências agrárias	2.419
Ciências humanas	148.842
Letras	9.883
Artes	7.464

Fonte: Serviço de Estatística da Educação e Cultura
Nota: Ciclo básico.

4.3.2.4. Distribuição de freqüências

É a série em que os dados são agrupados com suas respectivas freqüências absolutas:

Tabela 4.5 – Acidentes por dia na rodovia X – janeiro de 1977

Número de acidentes	Número de dias
0	10
1	7
2	4
3	5
4	3
5	2

Fonte: DNER

Tabela 4.6 – Altura dos alunos da classe – março de 1977

Alturas (m)	Número de alunos
1,50 ⊢ 1,60	5
1,60 ⊢ 1,70	15
1,70 ⊢ 1,80	17
1,80 ⊢ 1,90	3

Fonte: secretaria da escola

4.3.3. Erros mais comuns

A título de ilustração, serão apresentados e discutidos os erros mais comuns e rotineiramente encontrados em trabalhos acadêmicos.

A observação visual desses erros permite a conscientização visual, auxiliando no processo de aprendizagem.

a. Traços não permitidos ou desnecessários:

Tabela 4.7 – Matrícula no ensino de terceiro grau, Brasil – 1975

Áreas de ensino	Matrículas
Ciências biológicas	32.109
Ciências exatas e tecnologia	65.949
Ciências agrárias	2.419
Ciências humanas	148.842
Letras	9.883
Artes	7.464

Fonte: Serviço de Estatística do Ministério da Educação e Cultura

Comentário: os traços verticais somente são admitidos em situações em que, necessariamente, trazem clareza e auxiliam na compreensão do que está sendo representado.

b. Ausência dos traços obrigatórios:

Tabela 4.8 - Matrícula no ensino de terceiro grau, Brasil – 1975

Áreas de ensino	Matrículas
Ciências biológicas	32.109
Ciências exatas e tecnologia	65.949
Ciências agrárias	2.419
Ciências humanas	148.842
Letras	9.883
Artes	7.464

Fonte: Serviço de Estatística do Ministério da Educação e Cultura

c. Ausência dos elementos tornam a apresentação auto-explicativa:

Ciências biológicas	32.109
Ciências exatas e tecnologia	65.949
Ciências agrárias	2.419
Ciências humanas	148.842
Letras	9.883
Artes	7.464

d. Formatações inadequadas:

Tabela 4.9 - Matrícula no ensino de terceiro grau, Brasil – 1975

Áreas de ensino	Matrículas
Ciências biológicas	32.109
Ciências exatas e tecnologia	65.949
Ciências agrárias	2.419
Ciências humanas	148.842
Letras	9.883
Artes	7.464

Fonte: Serviço de Estatística do Ministério da Educação e Cultura

e. Ausência do cabeçalho:

Tabela 4.10 - Matrícula no ensino de terceiro grau, Brasil – 1975

Ciências biológicas	32.109
Ciências exatas e tecnologia	65.949
Ciências agrárias	2.419
Ciências humanas	148.842
Letras	9.883
Artes	7.464

Fonte: Serviço de Estatística do Ministério da Educação e Cultura

Comentário: dificulta o entendimento da composição por não explicitar o que está sendo apresentado nas colunas.

f. Ponto final no fim do título:

Tabela 4.11 - Matrícula no ensino de terceiro grau, Brasil – 1975.

Áreas de ensino	Matrículas
Ciências biológicas	32.109
Ciências exatas e tecnologia	65.949
Ciências agrárias	2.419
Ciências humanas	148.842
Letras	9.883
Artes	7.464

Fonte: Serviço de Estatística do Ministério da Educação e Cultura

Comentário: como o título inicia uma unidade de informação, que é a apresentação tabular vista como um todo, ele não deve ser seguido de um ponto final, que simboliza a finalização de algo.

g. Outros erros:

- Separação do título e dos demais elementos em páginas distintas.
- Fragmentação das composições sem atentar às normas.

4.4. Apresentações gráficas

A apresentação gráfica das séries estatísticas tem por finalidade representar os resultados obtidos.

Facilitam a compreensão de uma série de dados.

Permite chegar-se a conclusões sobre a evolução do fenômeno ou sobre como se relacionam os valores da série.

A escolha do gráfico mais adequado fica a critério do analista.

Os elementos de simplicidade, clareza e veracidade devem ser relevantes e sempre observados.

4.4.1. Elementos mínimos

a. Número: usada para identificar a composição.

b. Título: o que, onde e quando.

c. Identificadores: servem para associar as variáveis e respectivas escalas aos eixos.

d. Legenda: servem para auxiliar o entendimento da composição gráfica.

4.4.2. Gráfico em colunas

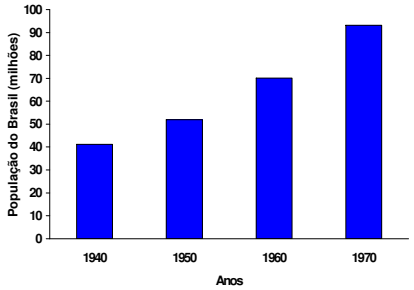


Figura 4.3 – Crescimento da população brasileira (1940-1970).

4.4.3. Gráfico em barras

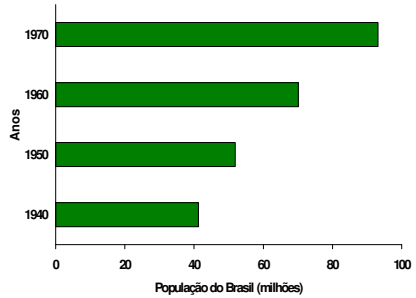


Figura 4.4 – Crescimento da população brasileira (1940-1970).

4.4.4. Gráfico em setores (pizza)

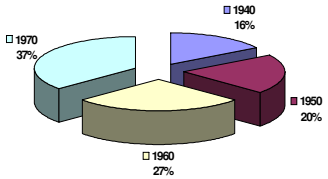


Figura 4.5 – Crescimento da população brasileira (1940-1970).

4.4.5. Gráfico polar

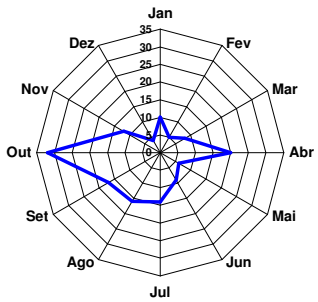


Figura 4.6 – Média mensal de acidentes na rodovia X (1980 a 2000).

4.4.6. Gráfico em curvas

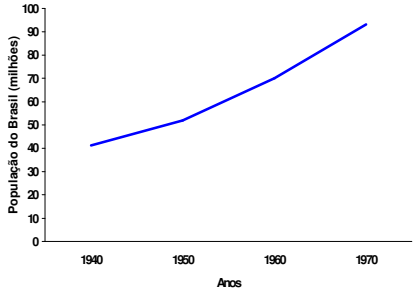


Figura 4.7 – Crescimento da população brasileira (1940-1970).

4.4.7. Erros mais comuns

a. Escalas inadequadas.
b. Ausência dos elementos mínimos.
c. Composição não auto-explicativa obrigando o leitor a buscar esclarecimentos no corpo do texto.

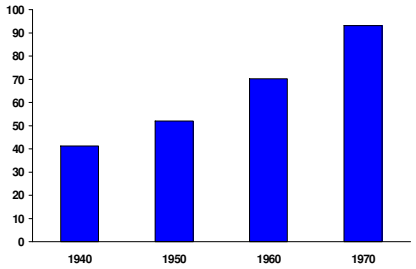
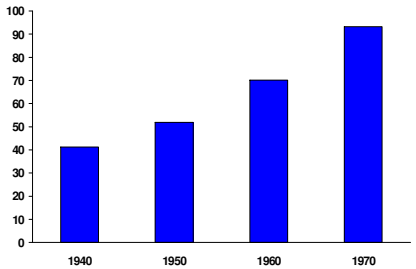


Figura 4.8 – Crescimento da população brasileira (1940-1970).

Ausência da referência e do título:



4.5. Distribuição de freqüências

É o tipo de tabela mais importante para a estatística descritiva.

4.5.1. Tipos de variável

Variável discreta: a variável é discreta quando assume valores em pontos da reta real. Em geral são aquelas que podemos contar utilizando os números inteiros.

Exemplo: número de erros em um livro: 0, 1, 5 ...

Variável contínua: por outro lado, quando a variável pode assumir teoricamente qualquer valor em um certo intervalo da reta real, ela será uma variável contínua. Em geral são aquelas que pesamos ou medimos.

Exemplo: peso de alunos: 50,5 kg; 50,572 kg, 50,574 kg, ...

4.5.2. Organização dos dados

Procedimentos para a apresentação das distribuições de freqüências:

a. Dados brutos: o conjunto dos dados numéricos obtidos após a crítica dos valores coletados constitui-se nos dados brutos:

24 - 23 – 22 – 28 – 35 – 21 – 23 – 33 – 34 – 24 – 21 – 25 – 36 – 26 ...

b. Rol: é o arranjo dos dados brutos em ordem crescente ou decrescente:

21 - 21 – 22 – 23 – 23 – 24 – 24 – 25 – 26 – 28 – 33 – 34 – 35 – 36 ...

c. Amplitude total ou range (R): é a diferença entre o maior e o menor valor observado:

Amplitude total (R) = 36 – 21 = 15

d. Freqüência absoluta (F_i): é o número de vezes que o elemento aparece no conjunto (amostra ou população). Assim:

21 - 21 – 22 – 23 – 23 – 24 – 24 – 25 – 26 – 28 – 33 – 34 – 35 – 36 ...

$F_{(21)} = 2$, $F_{(22)} = 1$, ...

4.5.3. Distribuição de freqüências

É o arranjo dos valores e suas respectivas freqüências. Assim, a distribuição de freqüência para o exemplo será:

21 - 21 – 22 – 23 – 23 – 24 – 24 – 25 – 26 – 28 – 33 – 34 – 35 – 36 ...

Tabela 4.14 – Distribuição de freqüência dos dados

Y_i	F_i
21	2
22	1
23	2
24	2
...	...

Trata-se de uma distribuição de freqüência de uma variável discreta.

Para variáveis contínuas, considerando-se, por exemplo, a altura dos indivíduos, teríamos:

Tabela 4.15 – Distribuição de freqüência dos dados

Classe	F_i
1,50 ⊢ 1,60	5
1,60 ⊢ 1,70	15
1,70 ⊢ 1,80	17
1,80 ⊢ 1,90	3

4.5.4. Limites das classes

Existem diversas maneiras de expressar os limites das classes:

- $Y_1 \vdash Y_2$: Todos os valores (Y_1 e Y_2), incluindo Y_1 e excluindo Y_2.
- $Y_1 \dashv Y_2$: Todos os valores (Y_1 e Y_2), incluindo Y_2 e excluindo Y_1.

4.5.5. Número de classes (K)

Não há uma fórmula exata para o cálculo do número de classes:

- K = 5, para n ≤ 25
- $K \cong \sqrt{n}$, para n > 25
- $K \cong 1 + 3{,}22 \log n$

4.5.6. Amplitude das classes (h)

Razão entre a amplitude total (R) e o número de classes (K):

$$h \cong \frac{R}{K}$$

4.5.7. Ponto médio das classes

É a média aritmética entre o limite superior e o inferior da classe:

$$10 \vdash 20: \quad \frac{10+20}{2} = 15$$

4.5.8. Freqüência absoluta acumulada (F_{ac})

É a soma dos valores inferiores ou iguais ao valor dado.

Exemplo: 0, 0, 0, 1, 1, 1, 1, 1, 2, 2

Tabela 4.16 – Freqüência absoluta e acumulada dos dados

Y_i	F_i	F_{ac}
0	3	3
1	5	8
2	2	10

4.5.9. Freqüência relativa (fi)

A freqüência relativa de um valor é dada por $f_i = \frac{F_i}{n}$, ou seja, é proporção daquele valor no conjunto:

Tabela 4.17 – Freqüência absoluta e relativa dos dados

Y_i	F_i	f_i	f_i, %
0	3	3/10	30
1	5	1/2	50
2	2	1/5	20
Σ	10	1	100

Os dados da F_i da Tabela 4.18 abaixo serão trabalhados para determinar a F_{ac} e posteriormente usados para elaborar as três composições gráficas mais importantes e básicas da análise exploratória dos dados, que serão elaboradas em sala de aula objetivando aprendizado e discussão dos detalhes da construção.

Tabela 4.18 – Idade dos alunos da classe B1 do colégio Alfa (2001)

Idade	F_i	F_{ac}
02 ⊢ 04	3	
04 ⊢ 06	5	
06 ⊢ 08	10	
08 ⊢ 10	6	
10 ⊢ 12	2	

Fonte: Cadastro de educação física dos alunos

4.5.10. Histograma

É a apresentação gráfica de uma distribuição de freqüência por meio de retângulos justapostos (exemplo em sala de aula).

4.5.11. Polígono de freqüências

É a apresentação gráfica de uma distribuição por meio de um polígono (exemplo em sala de aula).

4.5.12. Polígono de freqüência acumulada

É a apresentação gráfica de uma distribuição por meio de um polígono (exemplo em sala de aula).

5. MEDIDAS ESTATÍSTICAS

5.1. Introdução

No tópico anterior foi visto a síntese (ou resumo) de séries de dados sob a forma de apresentações tabulares, apresentações gráficas e as distribuições de freqüências.

Trata-se agora dos cálculos das medidas que possibilitam apresentar e confrontar séries de dados, relativas às observações dos fenômenos, de forma sintética e resumida.

5.2. Medidas de tendência central

Tais medidas orientam quanto aos valores centrais.

Representam os fenômenos pelos seus valores médios, em torno dos quais tendem a se concentrar os dados.

5.2.1. Média aritmética

Medida de tendência central de uso mais comum.

Notação adotada: ($\overline{Y}$ ou μ) para o parâmetro e ($\overline{y}$ ou m) para a estimativa.

5.2.1.1. Dados não agrupados

Sejam y_1, y_2, ..., y_n, portanto (N, n) valores da variável Y. A média aritmética simples de Y representada por ($\overline{Y}$, $\overline{y}$) é definida por:

Parâmetro: $\overline{Y} \; ou \; \mu = \frac{\sum y}{N}$

Estimativa: $\overline{y} \; ou \; m = \frac{\sum y}{n}$

Exemplo: considerando {3, 7, 8, 10, 11} como uma amostra:

$$\overline{y} = \frac{3+7+8+10+11}{5} = 7{,}8$$

5.2.1.2. Dados agrupados

Quando os dados de uma amostra estiverem agrupados numa distribuição de freqüência a média aritmética dos valores de Y (y_1, y_2, ..., y_n), ponderados pelas respectivas freqüências absolutas: F_1, F_2, ..., F_n é calculada como se segue:

$$\overline{y} \; ou \; m = \frac{\sum y \cdot F}{\sum F}$$

Exemplos:

Y	F_i	YF_i
1	1	1
2	3	6
3	5	15
4	1	4
Σ	10	26

$$\overline{y} \; ou \; m = \frac{\sum y \cdot F}{\sum F} = \frac{26}{10} = 2{,}6$$

Idade	F_i	Y	YF_i
02 ⊢ 04	5	3	15
04 ⊢ 06	10	5	50
06 ⊢ 08	14	7	98
08 ⊢ 10	8	9	72
10 ⊢ 12	3	11	33
Σ	40		268

As classes são representadas pelos seus pontos médios:

$$\overline{y} \; ou \; m = \frac{\sum y \cdot F}{\sum F} = \frac{268}{40} = 6{,}7$$

5.2.1.3. Média geral

Sejam $\overline{y}_1$, $\overline{y}_2$, ..., $\overline{y}_k$ as estimativas das médias aritméticas de K séries e n_1, n_2, ..., n_k os números de termos destas séries, respectivamente. A média aritmética da série formada pelos termos da K séries é dada pela fórmula:

$$\overline{y} \; ou \; m = \frac{n_1\overline{y}_1 + n_2\overline{y}_2 + ... + n_k\overline{y}_k}{n_1 + n_2 + ... + n_k}$$

Exemplo: Sejam as séries:

1) {4, 5, 6, 7, 8} em que $n_1 = 5$ e $\overline{y}_1 = 6$

2) {1, 2, 3} em que $n_2 = 3$ e $\overline{y}_2 = 2$

3) {9, 10, 11, 12, 13} em que $n_3 = 5$ e $\overline{y}_3 = 11$

$$\overline{Y} = \frac{n_1\overline{y}_1 + n_2\overline{y}_2 + ... + n_k\overline{y}_k}{n_1 + n_2 + ... + n_k} = \frac{5 \cdot 6 + 3 \cdot 2 + 5 \cdot 11}{5 + 3 + 5} = 7$$

5.2.2. Média geométrica

Usada para médias proporcionais de crescimento quando uma medida subseqüente depende de medidas prévias.

Notação adotada: (MG) para o parâmetro e (mg) para a estimativa.

Sejam y_1, y_2, ..., y_n, valores de Y associados às respectivas freqüências absolutas F_1, F_2, ..., F_n. A média geométrica (MG ou mg) de Y é definida por:

$$MG \text{ ou } mg = \sqrt[n]{y_1^{F_1} \cdot y_2^{F_2} \cdot \ldots \cdot y_n^{F_n}}$$

Exemplo:

Média geométrica de uma amostra {3, 6, 12, 24, 48}

$$mg = \sqrt[5]{3 \cdot 6 \cdot 12 \cdot 24 \cdot 48} = \sqrt[5]{248.832} = 12$$

5.2.3. Média harmônica

Usada para médias de crescimento e proporções de velocidade.

Notação adotada: (MH) para o parâmetro e (mh) para a estimativa.

Sejam y_1, y_2, ..., y_n, valores de Y, associados às respectivas freqüências absolutas F_1, F_2, ..., F_n. A média harmônica (MH ou mh) de Y é definida por:

$$MH \text{ ou } mh = \frac{n}{\frac{F_1}{y_1} + \frac{F_2}{y_2} + \ldots + \frac{F_n}{y_n}} = \frac{n}{\sum_{i=1}^{n} \frac{F_i}{y_i}}$$

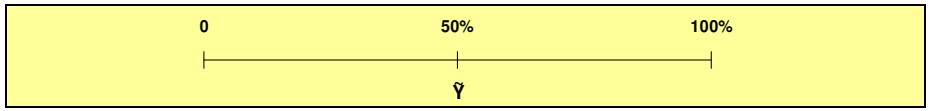
Exemplo:

Média harmônica de uma amostra {2, 5, 8}

$$mh = \frac{3}{\frac{1}{2} + \frac{1}{5} + \frac{1}{8}} = 3{,}64$$

5.2.4. Mediana

Medida de tendência muito usada quando o interesse é a determinação do valor que separa a série de dados em duas partes iguais, 50% situados acima e 50% situados abaixo da medida.



Notação adotada: ($\tilde{Y}$ ou MD) para o parâmetro e ($\tilde{y}$ ou mg) para a estimativa.

Colocados em ordem crescente, a mediana ($\tilde{Y}$, $\tilde{y}$) é o valor que divide a série em duas partes iguais:

5.2.4.1. Cálculo da mediana para variável discreta

Se n for impar, a mediana será o elemento central (de ordem $\frac{n+1}{2}$).

Caso n seja par, a mediana será a média entre os elementos centrais (de ordem $\frac{n}{2}$ e $\frac{n}{2}+1$).

Exemplo 1:

Y_i	F_i	F_{ac}	
1	1	1	
2	3	4	
3	5	9	← contém o 6º elemento
4	2	11	
Σ	11		

n = 11, n é impar, logo $\tilde{y}$ será o elemento de ordem $\frac{n+1}{2}$, ou seja $\frac{11+1}{2} = 6^{o}$

Será, portanto, o sexto elemento.

Por meio das freqüências acumuladas encontra-se o valor y_i correspondente a mediana, que neste exemplo é 3 ($\tilde{y}$ = 3).

Exemplo 2:

Y_i	F_i	F_{ac}	
82	5	5	
85	10	15	
87	15	30	← contém os 21º e 22º elementos
89	8	38	
90	4	42	
Σ	42		

n = 42, n é par, logo $\tilde{y}$ será a média entre os elementos de ordem $\frac{n}{2}$ e $\frac{n}{2}+1$

Ou seja $\frac{42}{2} = 21$ e $\frac{42}{2}+1 = 22$ (21º e 22º elementos)

Identificam-se os elementos de ordem 21º e 22º pela F_{ac}

Assim, o 21º corresponde a 87 e 22º corresponde a 87, logo:

$$\tilde{y} = \frac{87+87}{2} = 87$$

5.2.4.2. Cálculo da mediana para variável contínua

1º passo: calcula-se a ordem $\frac{n}{2}$. Como a variável é contínua, não importa se n é par ou impar.

2º passo: pela F_{ac} identifica-se a classe que contém a mediana (classe md).

3º passo: usa-se fórmula:

$$\tilde{y} = \ell_{md} + \frac{\left(\frac{n}{2} - \sum f\right) \cdot h}{F_{md}}$$

Em que:

ℓ_{md} = limite inferior da classe md

n = tamanho da série

$\sum f$ = soma das freqüências anteriores à classe md

h = amplitude da classe md

F_{md} = freqüência da classe md

Exemplo:

Classe	F_i	F_{ac}	
35 ⊢ 45	5	5	
45 ⊢ 55	12	17	
55 ⊢ 65	18	35	← classe mediana
65 ⊢ 75	14	49	
75 ⊢ 85	6	55	
85 ⊢ 95	3	58	
Σ	58	268	

1º passo: $\frac{58}{2} = 29^o$

2º passo: classe md = 3ª

3º passo: usa-se a fórmula:

ℓ_{md} = 55; n = 58; $\sum f$ = 17; h = 10; F_{md} = 18

$$\tilde{y} = 55 + \frac{\left(\frac{58}{2} - 17\right) \cdot 10}{18} = 61{,}67$$

5.2.5. Moda

Medida de tendência central muito usada quando o interesse é o valor mais freqüente da série.

Notação adotada: (MO) para o parâmetro e (mo) para a estimativa.

A moda pode não existir – o que constitui uma série amodal – ou, mesmo que exista pode não ser única – o que caracteriza uma série multimodal.

Para distribuições simples (sem agrupamento de classes), a identificação da moda é facilitada pela simples observação do elemento que apresenta maior freqüência.

Assim, considerando a distribuição abaixo como uma amostra:

y_i	243	245	248	251	307
F_i	7	17	23	20	8

A moda será 248, e indica-se por mo = 248.

5.2.5.1. Moda para dados agrupados em classes

1º passo: identifica-se a classe modal (maior freqüência).

2º passo: usa-se a fórmula de Czuber

$$mo = \ell + \frac{\Delta_1}{\Delta_1 + \Delta_2} \cdot h$$

Em que:

ℓ = limite inferior da classe mo

Δ_1 = diferença entre a freqüência da classe modal e a imediatamente anterior

Δ_2 = diferença entre a freqüência da classe modal e a imediatamente posterior

h = amplitude da classe

Exemplo:

Classes	0 ⊢ 1	1 ⊢ 2	2 ⊢ 3	3 ⊢ 4	4 ⊢ 5	Σ
F_i	3	10	17	8	5	43

1º passo: indica-se a classe modal: 3ª (2 ⊢ 3)

2º passo: usa-se a fórmula

$$mo = \ell + \frac{\Delta_1}{\Delta_1 + \Delta_2} \cdot h$$

$$mo = 2 + \frac{7}{7+9} \cdot 1 = 2{,}44$$

5.3. Comparação entre as medidas de tendência central

5.3.1. Média

5.3.1.1. Vantagens

- Fácil de compreender e calcular
- Utiliza todos os valores da série
- É um valor único
- É fácil de ser incluída em expressões matemáticas
- Pode ser determinada nas escalas: intervalar e proporcional.

5.3.1.2. Desvantagens

- Muito afetada por valores extremos
- Necessário conhecer todos os valores da série.

5.3.2. Mediana

5.3.2.1. Vantagens

- Fácil de compreender e aplicar
- Não é afetada por valores extremos
- É um valor único
- É fácil de incluir em expressões matemáticas
- Pode ser determinada nas escalas: ordinal, intervalar e proporcional.

5.3.2.2. Desvantagens

- É difícil de ser incluída em expressões matemáticas
- Não usa todos os valores da série.

5.3.3. Moda

5.3.3.1. Vantagens

- Fácil de compreender e calcular
- Não é afetada por valores extremos
- Pode ser aplicada em todas as escalas: nominal, ordinal, intervalar e proporcional.

5.3.3.2. Desvantagens

- Pode estar afastada do centro dos valores
- É difícil de ser incluída em expressões matemáticas
- Não usa todos os valores da série
- A variável pode ter mais de uma moda (bimodal ou multimodal)
- Algumas variáveis não possuem moda.

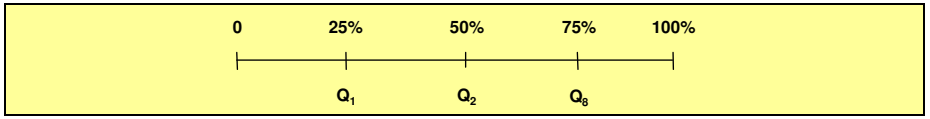
5.4. Medidas de posição ou separatrizes

Genericamente denominadas quantis, orientam quanto à posição na distribuição.

Permitem determinar valores que particionam a série de n observações em partes iguais.

5.4.1. Quartis

Seguindo o mesmo raciocínio da mediana, os três quartis dividem uma série em 4 partes iguais:



Notação adotada: (Q) para o parâmetro e (q) para a estimativa.

$$q_i = \ell_{q_i} + \frac{\left(\frac{i \cdot n}{4} - \sum f\right) \cdot h}{F_{q_i}}$$

Em que:

ℓ_{q_i} = limite inferior da classe q_i (i = 1, ..., 3)

i = 1 para q_1, ..., 3 para q_3

n = tamanho da série

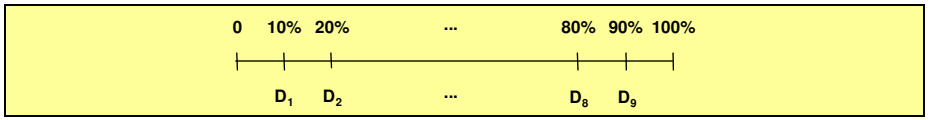
$\sum f$ = soma das freqüências anteriores à classe q_i

h = amplitude da classe q_i

F_{q_i} = freqüência da classe q_i

5.4.2. Decis

Os decis dividem a série em 10 partes iguais.



Notação adotada: (D) para o parâmetro e (d) para a estimativa.

$$d_i = \ell_{d_i} + \frac{\left(\frac{i \cdot n}{10} - \sum f\right) \cdot h}{F_{d_i}}$$

Em que:

ℓ_{d_i} = limite inferior da classe d_i (i = 1, ..., 9)

i = 1 para d_1, ..., 9 para d_9

n = tamanho da série

$\sum f$ = soma das freqüências anteriores à classe d_i

h = amplitude da classe d_i

F_{d_i} = freqüência da classe d_i

5.4.3. Percentis

Os percentis (P para populações e p para amostras) dividem a série em 100 partes iguais.

Notação adotada: (P) para o parâmetro e (p) para a estimativa.

$$p_i = \ell_{p_i} + \frac{\left(\frac{i \cdot n}{100} - \sum f\right) \cdot h}{F_{p_i}}$$

Em que:

ℓ_{p_i} = limite inferior da classe p_i (i = 1, ..., 99)

i = 1 para p_1, ..., 99 para p_{99}

n = tamanho da série

$\sum f$ = soma das freqüências anteriores à classe p_i

h = amplitude da classe p_i

F_{p_i} = freqüência da classe p_i

Os procedimentos para determinar os quartis, decis e percentis são semelhantes aos usados para determinar o valor da mediana.

5.4.4. Situações de uso mais comuns destas medidas

Uma dos usos mais comuns, e importantes, destas medidas na análise exploratória dos dados é o diagrama de caixa (“box plot”), como abaixo:

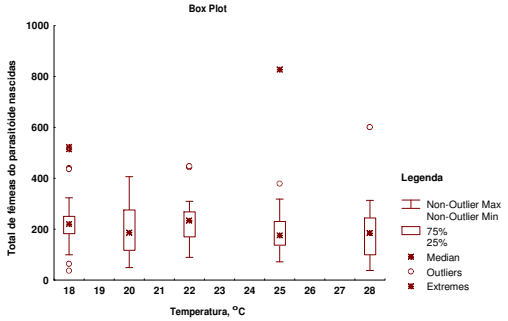


Figura 5.1 – Diagrama de caixa do total de fêmeas do parasitóide nascidas.

5.5. Medidas de dispersão

São medidas estatísticas usadas para avaliar o grau de variabilidade ou dispersão dos valores da série em torno da média.

Juntamente com as medidas tendência central, principalmente a média aritmética, são medidas de extrema importância para o aprendizado e a compreensão da estatística.

5.5.1. Amplitude total

Notação: (AT) para o parâmetro (at) para a estimativa.

5.5.1.1. O que é

Ë uma medida da dispersão dos dados.

É definida como a diferença entre o maior e o menor dos valores da série.

5.5.1.2. O que quantifica

Quantifica a dispersão dos dados.

Permite distinguir séries de dados em relação à homogeneidade:

- Séries homogêneas: menor valor da amplitude total
- Séries heterogêneas: maior valor da amplitude total

5.5.1.3. Como se calcula

Pop: $AT = y_{\max} - y_{\min}$

Amo: $at = y_{\max} - y_{\min}$

Exemplo:

Considerando a série {1, 0, 1, 2, 2, 0, 2, 2, 2, 5, 3, 3, 3, 8} como uma amostra

$$at = 38 - 10 = 28m$$

A amplitude total é uma medida da dispersão muito limitada, pois depende apenas dos valores extremos, não sendo afetada pela dispersão dos valores internos.

5.5.1.4. Unidade de expressão

A unidade de expressão é a mesma da variável aleatória em questão:

5.5.2. Desvio médio

Notação: (DM) para o parâmetro e (dm) para a estimativa.

5.5.2.1. O que é

Ë uma medida da dispersão dos dados em relação à média aritmética.

É definida como a média dos desvios absolutos em relação à média aritmética.

5.5.2.2. O que quantifica

Quantifica a dispersão dos dados.

Permite distinguir séries de dados em relação à homogeneidade:

- Séries homogêneas: menor valor do desvio médio
- Séries heterogêneas: maior valor do desvio médio

5.5.2.3. Como se calcula

Considerando:

Pop: $D_i = (y_i - \mu)$

Amo: $d_i = (y_i - m)$

Considerando que:

$$\sum D_i = \sum d_i = 0$$

Parâmetro: $DM = \frac{\sum D}{n} = \frac{\sum |y - \mu|}{n}$

Estimativa: $dm = \frac{\sum |d|}{n} = \frac{\sum |y - m|}{n}$

Exemplo:

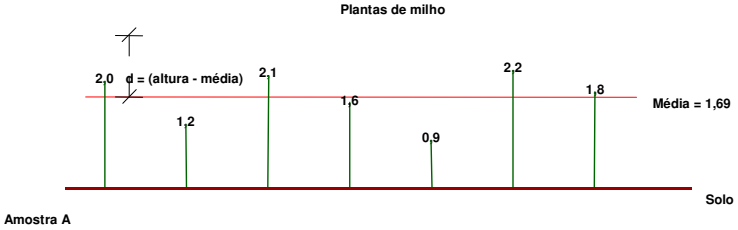


Figura 5.2 – Ilustração de uma amostra de plantas de milho.

$$dm = \frac{\sum |d|}{n} = \frac{\sum |y - \bar{y}|}{n} = \frac{|d_1| + \ldots + |d_7|}{n}$$

$$dm = \frac{|2,0 - 1,69| + ... + |1,8 - 1,69|}{7} = 0,39\,m$$

Trata-se, portanto, de uma medida exata da média dos desvios absolutos em relação à média aritmética.

5.5.2.4. Unidade de expressão

A unidade de expressão é a mesma da variável aleatória em questão:

5.5.3. Desvio quadrático médio

Notação: (DQM) para o parâmetro (dqm) para a estimativa.

5.5.3.1. O que é

Ë uma medida da dispersão dos dados em relação à média aritmética.

É definida como a média do quadrado dos desvios em relação à média aritmética.

5.5.3.2. O que quantifica

Quantifica a dispersão dos dados.

Permite distinguir séries de dados em relação à homogeneidade:

- Séries homogêneas: menor valor do desvio quadrático médio
- Séries heterogêneas: maior valor do desvio quadrático médio

5.5.3.3. Como se calcula

Parâmetro: $DQM = \frac{\sum (D)^2}{N} = \frac{\sum (y - \mu)^2}{N}$

Estimativa: $DQM = \frac{\sum (d)^2}{n} = \frac{\sum (y - m)^2}{n}$

Exemplo:

Considerando os dados da Figura 5.2 relativos a uma amostra de plantas de milho:

$$dqm = \frac{\sum (d)^2}{n} = \frac{\sum (y - \bar{y})^2}{n}$$

$$dqm = \frac{(2,0 - 1,69)^2 + ... + (1,8 - 1,69)^2}{7} = 1,41\,m^2$$

5.5.3.4. Unidade de expressão

A unidade de expressão é a mesma da variável aleatória em questão, porém, elevada ao quadrado. Para o exemplo dado na Figura 5.2, altura de plantas, a unidade é o metro elevado ao quadrado, m^2:

5.5.4. Variância

Notação: (σ^2) para o parâmetro e (s^2) para a estimativa.

5.5.4.1. O que é

Ë uma medida da dispersão dos dados em relação à média aritmética.

É definida como a razão entre a soma de quadrados dos desvios de cada valor em relação à média aritmética, Σd^2, e o número de elementos da série, N para populações ou n-1 para amostras.

5.5.4.2. O que quantifica

Quantifica a dispersão dos dados em relação à média aritmética.

Permite distinguir séries de dados em relação à homogeneidade:

- Séries homogêneas: menor valor da variância
- Séries heterogêneas: maior valor da variância

5.5.4.3. Como se calcula

Populações:

$$\sigma^2 = \frac{\sum D^2}{N} \quad \text{onde} \quad D = y - \mu \quad \text{ou} \quad \sigma^2 = \frac{\sum y^2 - \frac{\left(\sum y\right)^2}{N}}{N}$$

Amostras:

a. μ é conhecido (caso raro):

$$s^2 = \frac{\sum D^2}{n} \quad \text{onde} \quad D = y - \mu \quad \text{ou} \quad s^2 = \frac{\sum y^2 - \frac{\left(\sum y\right)^2}{n}}{n}$$

b. μ é desconhecido (caso comum):

$$s^2 = \frac{\sum d^2}{n-1} \quad \text{onde} \quad d = y - m \quad \text{ou} \quad s^2 = \frac{\sum y^2 - \frac{\left(\sum y\right)^2}{n}}{n-1}$$

5.5.4.4. Unidade de expressão

A unidade de expressão é a mesma da variável aleatória em questão, porém, elevada ao quadrado. Para o exemplo dado na Figura 5.2, altura de plantas, a unidade é o metro elevado ao quadrado, m^2:

$$\sigma^2 = \frac{\sum D^2}{N} \; ou \; s^2 = \frac{\sum d^2}{n-1} = \frac{m^2 + ... + m^2}{número} = m^2$$

É muito comum a dificuldade do estudante compreender o significado das medidas absolutas de dispersão (variância e do desvio padrão). Ou seja, compreender o conceito, o fundamento, antecedendo a qualquer cálculo:

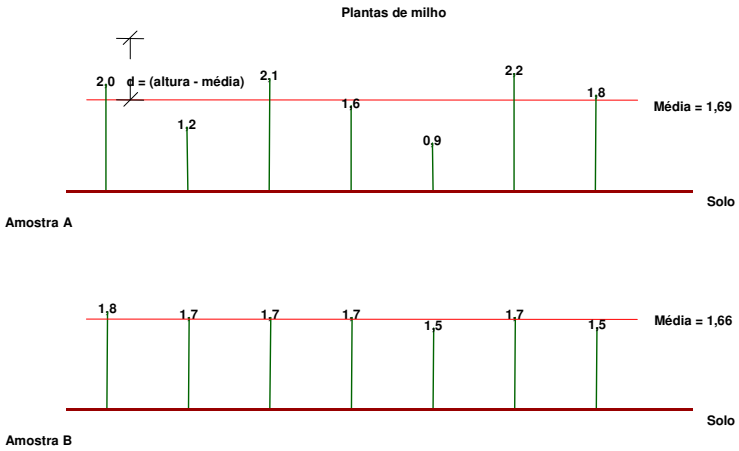


Figura 5.3 – Ilustração do significado da variância, s^2. As barras verdes representam a altura das plantas de milho em relação ao solo e d representa o desvio da altura de uma planta em relação à média da série.

A variância, para uma variável em estudo, nada mais é que uma medida da totalidade dos desvios em relação à média.

Intuitivamente, portanto, a amostra A deve apresentar um maior valor da variância da altura das plantas de milho que a amostra B, pois os dados, em A, encontram-se mais dispersos em relação à média.

Cálculos:

$$s_A^2 = \frac{\sum d^2}{n-1} = \frac{(2{,}0-1{,}69)^2 + (1{,}2-1{,}69)^2 + ... + (1{,}8-1{,}69)^2}{7-1} = 0{,}23\ m^2$$

$$s_B^2 = \frac{\sum d^2}{n-1} = \frac{(1{,}8-1{,}66)^2 + (1{,}7-1{,}66)^2 + ... + (1{,}5-1{,}66)^2}{7-1} = 0{,}01\ m^2$$

5.5.4.5. Formas de cálculo

Amostra A:

$$s_A^2 = \frac{\sum d^2}{n-1} = \frac{d_1^2 + ... + d_7^2}{n-1} = \frac{(2{,}0-1{,}69)^2 + ... + (1{,}8-1{,}69)^2}{7-1} = \frac{(0{,}31)^2 + ... + (0{,}11)^2}{6} = 0{,}23\ m^2$$

ou

$$s_A^2 = \frac{\sum y^2 - \frac{\left(\sum y\right)^2}{n}}{n-1} = \frac{21{,}30 - \frac{(11{,}80)^2}{7}}{6} = 0{,}23\ m^2$$

Amostra B:

$$s_B^2 = \frac{\sum d^2}{n-1} = \frac{d_1^2 + ... + d_7^2}{n-1} = \frac{(1{,}8-1{,}66)^2 + ... + (1{,}5-1{,}66)^2}{7-1} = \frac{(0{,}14)^2 + ... + (-0{,}16)^2}{6} = 0{,}01\ m^2$$

ou

$$s_A^2 = \frac{\sum y^2 - \frac{\left(\sum y\right)^2}{n}}{n-1} = \frac{19{,}30 - \frac{(11{,}60)^2}{7}}{6} = 0{,}01\ m^2$$

5.5.4.6. Demonstração da fórmula para cálculo da variância

$$s_Y^2 = \frac{1}{n-1}\sum d^2$$

$$s_Y^2 = \frac{1}{n-1}\sum (y-m)^2$$

$$s_Y^2 = \frac{1}{n-1}\sum \left(y^2 - 2ym + m^2\right)$$

$$s_Y^2 = \frac{1}{n-1}\sum y^2 - 2m\sum y + \sum m^2 \qquad \therefore \qquad \sum K\cdot y = K\sum y$$

se $m = \frac{\sum y}{n}$ então $\sum y = n\cdot m$

$\sum m^2 = n\cdot m^2$

$$s_Y^2 = \frac{1}{n-1}\sum y^2 - (2m)(n\cdot m) + n\cdot m^2$$

$$s_Y^2 = \frac{1}{n-1}\sum y^2 - 2n\cdot m^2 + n.m^2 \qquad \therefore \qquad 2a - a = a$$

$$s_Y^2 = \frac{1}{n-1}\sum y^2 - n\cdot m^2 \qquad \therefore \qquad m = \frac{\sum y}{n}$$

$$s_Y^2 = \frac{1}{n-1}\sum y^2 - n\cdot\left(\frac{\sum y}{n}\right)^2$$

$$s_Y^2 = \frac{1}{n-1}\sum y^2 - n\cdot\frac{(\sum y)^2}{n^2}$$

$$s_Y^2 = \frac{1}{n-1}\sum y^2 - \not{n}\cdot\frac{(\sum y)^2}{\not{n}^2}$$

$$s_Y^2 = \frac{\sum y^2 - \frac{(\sum y)^2}{n}}{n-1}$$

5.5.4.7. Demonstração da não tendenciosidade da estimativa da variância s^2

$$s^2 = \frac{\sum (y-m)^2}{n \text{ ou } n-1\ ?}$$

$$\sum (y-m)^2 = \sum (y-\mu+\mu-m)^2$$

$$\sum (y-m)^2 = \sum \{(y-\mu)-(m-\mu)\}^2$$

$$\sum (y-m)^2 = \sum \left\{(y-\mu)^2 - 2(y-\mu)\cdot(m-\mu) + (m-\mu)^2\right\}$$

$$\sum (y-m)^2 = \sum (y-\mu)^2 - 2\sum (y-\mu)\cdot(m-\mu) + \sum (m-\mu)^2$$

$$\sum (y-\mu) = \sum y_i - n\cdot\mu \qquad \therefore \qquad m = \frac{\sum y}{n} \qquad \therefore \qquad \sum y = n\cdot m$$

$$\sum (y-\mu) = n\cdot m - n\cdot\mu = n(m-\mu)$$

$\sum (m-\mu)^2 = n(m-\mu)^2$ para uma determinada amostra $(m-\mu) = \text{constante}$

$$\sum (y-m)^2 = \sum (y-\mu)^2 - 2n(m-\mu)\cdot(m-\mu) + n(m-\mu)^2$$

$$\sum (y-m)^2 = \sum (y-\mu)^2 - 2n(m-\mu)^2 + n(m-\mu)^2 \qquad -2a + a = -a$$

$$\sum (y-m)^2 = \sum (y-\mu)^2 - n(m-\mu)^2$$

Considerando $s^2 = \frac{\sum (y-m)^2}{n}$

$$E(s^2) = \frac{1}{n}E\left\{\sum (y-\mu)^2 - n(m-\mu)^2\right\}$$

$$E(s^2) = \frac{1}{n}\left\{\sum E(y-\mu)^2 - n\cdot E(m-\mu)^2\right\}$$

$$E(s^2) = \frac{1}{n}\{n\cdot V(Y) - n\cdot V(m)\} \qquad \therefore \qquad V(m) = \frac{\sigma^2}{n}$$

$$E(s^2) = \frac{1}{n}\left\{n\cdot\sigma^2 - \not{n}\frac{\sigma^2}{\not{n}}\right\}$$

$$E(s^2) = \frac{1}{n}\left\{n\cdot\sigma^2 - \sigma^2\right\} = \frac{1}{n}\left\{\sigma^2(n-1)\right\} = \frac{(n-1)\cdot\sigma^2}{n}$$

Portanto, $s^2 = \frac{\sum (y-m)^2}{n}$, é um estimador tendencioso (subestima) de σ^2.

Considerando $s^2 = \frac{\sum (y-m)^2}{n-1}$

$$E(s^2) = \frac{1}{n-1} E\left\{\sum (y-\mu)^2 - n(m-\mu)^2\right\}$$

$$E(s^2) = \frac{1}{n-1}\left\{\sum E(y-\mu)^2 - n \cdot E(m-\mu)^2\right\}$$

$$E(s^2) = \frac{1}{n-1}\{n \cdot V(Y) - n \cdot V(m)\} \qquad \therefore \qquad V(m) = \frac{\sigma^2}{n}$$

$$E(s^2) = \frac{1}{n-1}\left\{n \cdot \sigma^2 - n\frac{\sigma^2}{n}\right\}$$

$$E(s^2) = \frac{1}{n-1}\{n \cdot \sigma^2 - \sigma^2\} = \frac{1}{n-1}\{\sigma^2(n-1)\} = \frac{(n-1)\cdot\sigma^2}{n-1} = \sigma^2$$

Portanto, $s^2 = \frac{\sum (y-m)^2}{n-1}$, é um estimador não tendencioso de σ^2.

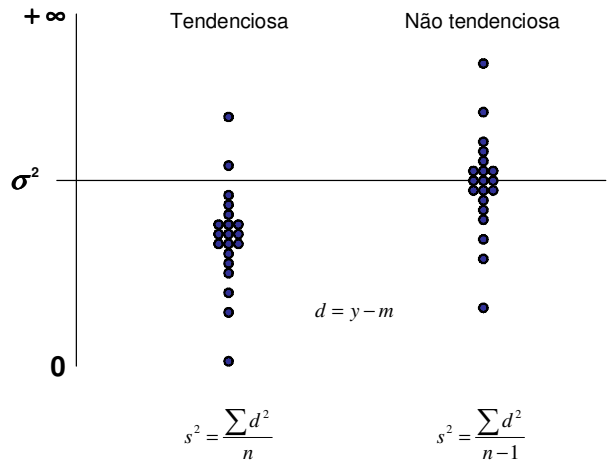


Figura 5.4 – Ilustração da tendenciosidade da estimativa de σ^2 se o somatório dos desvios em relação à média for dividido por n, ao invés de n-1.

5.5.5. Desvio padrão

Notação: (σ) para o parâmetro e (s) para a estimativa.

5.5.5.1. O que é

É uma medida da dispersão dos dados em relação à média aritmética.

É definido como a raiz quadrada da variância.

5.5.5.2. O que quantifica

Quantifica a dispersão dos dados em relação à média aritmética.

5.5.5.3. Como se calcula

Populações: $\sigma = \sqrt{\sigma^2}$

Amostras: $s = \sqrt{s^2} \qquad \therefore \qquad s_A = \sqrt{s_A^2} = \sqrt{0{,}23\ m^2} = 0{,}48m$

5.5.5.4. Unidade de expressão

A unidade de expressão é a mesma da variável aleatória em questão. Para o exemplo dado, a unidade é o metro, m:

$$\sigma \ ou \ s = \sqrt{m^2} = m$$

A variância e o desvio padrão são as medidas mais usadas para quantificar a dispersão dos dados em torno da média.

5.5.6. Desvio padrão relativo e coeficiente de variação

Notação: (DPR e CV) para os parâmetros e (dpr e cv) para as estimativas.

5.5.6.1. O que são

São medidas relativas da dispersão dos dados em relação à média.

São definidas como a razão entre o desvio padrão e a média aritmética.

5.5.6.2. O que quantificam

Quantificam a dispersão relativa dos dados em relação à média aritmética.

5.5.6.3. Como são calculados

Populações: $DPR = \frac{\sigma}{\mu}$ $CV = \frac{\sigma}{\mu} \cdot 100$

Amostras: $dpr = \frac{s}{m}$ $cv = \frac{s}{m} \cdot 100$

5.5.6.4. Justificativas para o uso e unidades de expressão

Freqüentemente em trabalhos de pesquisa são necessárias comparações em situações nas quais as medidas estatísticas das variáveis em estudo foram feitas usando-se unidades distintas. Por exemplo: um pesquisador usou o metro, m, e outro o centímetro, cm.

Como as medidas absolutas de dispersão (variância e desvio padrão) são influenciadas pela unidade de medida das variáveis em estudo, a comparação entre os trabalhos fica dificultada.

Por serem adimensionais, é conveniente determinar uma das medidas relativas de dispersão, sendo a mais usada o coeficiente de variação.

Considerando que a unidade de medida das variáveis estudadas foi o metro, m:

População: $DPR = \frac{\sigma}{\mu} = \frac{\cancel{m}}{\cancel{m}} = \text{admensional}$ $CV = \frac{\sigma}{\mu} . 100 = \frac{\cancel{m}}{\cancel{m}} . 100 = \%$ (admensional)

Amostra: $dpr = \frac{s}{m} = \frac{\cancel{m}}{\cancel{m}} = \text{admensional}$ $cv = \frac{s}{m} . 100 = \frac{\cancel{m}}{\cancel{m}} . 100 = \%$ (admensional)

Desta forma pode-se saber, independentemente da influência das unidades usadas, qual estudo apresentou maior ou menor dispersão.

Exemplo:

Considerando os dados originais da Figura 5.3:

Amostra A em metro (m): $cv = \frac{s}{m} . 100 = \frac{0,48}{1,69} . 100 = 28,74\%$

Amostra B em metro (m): $cv = \frac{s}{m} . 100 = \frac{0,11}{1,66} . 100 = 6,84\%$

Exemplo:

Considerando os dados da Figura 5.3 coletados em outras unidades:

Amostra A em milímetro (mm): $cv = \frac{s}{m} . 100 = \frac{484,52}{1.685,71} . 100 = 28,74\%$

Amostra B em centímetro (cm): $cv = \frac{s}{m} . 100 = \frac{11,34}{165,71} . 100 = 6,84\%$

Observa-se que as unidades de medida das variáveis não exercem influência na magnitude das medidas de dispersão relativas.

6. EXEMPLO DE ANÁLISE EXPLORATÓRIA DOS DADOS

6.1. Dados

{20,0; 21,5; 15,6; 12,8; 17,2; 14,4; 13,5; 9,2; 6,7; 9,6; 112; 10,0; 11,0; 14,2; 13,6; 24,4; 16,4; 12,8; 19,2; 19,4; 13,8; 14,9; 9,7; 8,5; 3,9; 21,3; 25,4; 4,8; 16,7; 9,4; 13,0}.

n = 31

6.2. Análise preliminar

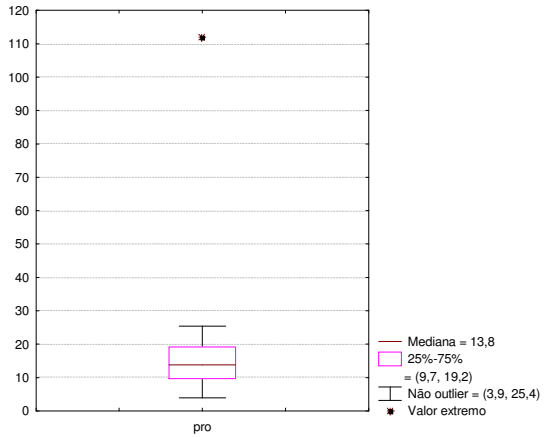


Figura 6.1 – Diagrama de caixa da produção de leite da Fazenda Nova Esperança, município de Itabuna, BA – abril de 2001.

Crítica

Eliminar 112 (1,12 ou 11,2 ou valor estranho 112 ?)

n = 31 – 1 = 30

Dados brutos

{20,0; 21,5; 15,6; 12,8; 17,2; 14,4; 13,5; 9,2; 6,7; 9,6; 10,0; 11,0; 14,2; 13,6; 24,4; 16,4;12,8; 19,2; 19,4; 13,8; 14,9; 9,7; 8,5; 3,9; 21,3; 25,4; 4,8; 16,7; 9,4; 13,0}.

n = 30

Rol

{3,9; 4,8; 6,7; 8,5; 9,2; 9,4; 9,6; 9,7; 10,0; 11,0; 12,8; 12,8; 13,0; 13,5; 13,6; 13,8; 14,2; 14,4; 14,9; 15,6; 16,4; 16,7; 17,2; 19,2; 19,4; 20,0; 21,3; 21,5; 24,4; 25,4}.

6.3. Representação tabular dos dados

Estrutura da tabela:

Amplitude total (at): 25,4 – 3,9 = 21,5 kg an^{-1} dia^{-1}

Número de classes (K): $K \cong \sqrt{n} \cong \sqrt{30} \cong 6{,}0 \rightarrow 7$ (opção)

Amplitude das classes (h): $h \cong \frac{at}{K} \cong \frac{21{,}5}{7} \cong 3{,}1 \rightarrow 4$ (opção)

Tabela 6.1 – Freqüências da produção de leite da Fazenda Nova Esperança, município de Itabuna, BA – abril de 2001

Classes	F_i	f_i	f_i, %	F_{ac}	F_{ac}, %
00 ⊢ 04	1	0,03	3,33	1	3,33
04 ⊢ 08	2	0,07	6,67	3	10,00
08 ⊢ 12	7	0,23	23,33	10	33,33
12 ⊢ 16	10	0,33	33,33	20	66,67
16 ⊢ 20	5	0,17	16,67	25	83,33
20 ⊢ 24	3	0,10	10,00	28	93,33
24 ⊢ 28	2	0,07	6,67	30	100,00

Fonte: dados da ordenha de 22/04/2001
Nota: dados expressos em kg an^{-1} dia^{-1}.

6.4. Representações gráficas dos dados

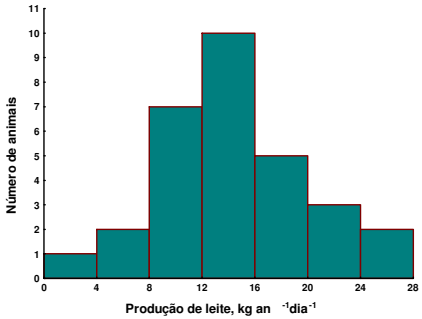


Figura 6.2 – Histograma da produção de leite da Fazenda Nova Esperança, município de Itabuna, BA – abril de 2001.

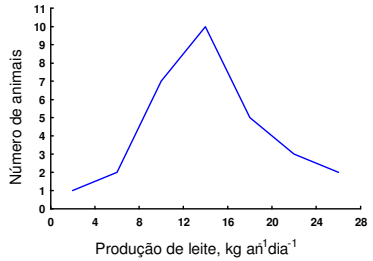


Figura 6.3 – Polígono de freqüência da produção de leite da Fazenda Nova Esperança, município de Itabuna, BA – abril de 2001.

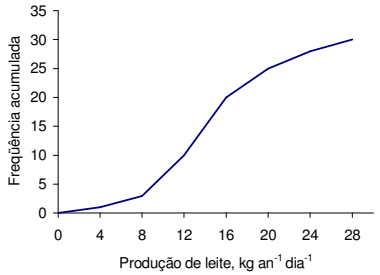


Figura 6.4 – Polígono de freqüência acumulada da produção de leite da Fazenda Nova Esperança, município de Itabuna, BA – abril de 2001.

6.5. Medidas estatísticas

6.5.1. Tendência central

6.5.1.1. Média - dados não agrupados

{3,9; 4,8; 6,7; 8,5; 9,2; 9,4; 9,6; 9,7; 10,0; 11,0; 12,8; 12,8; 13,0; 13,5; 13,6; 13,8; 14,2; 14,4; 14,9; 15,6; 16,4; 16,7; 17,2; 19,2; 19,4; 20,0; 21,3; 21,5; 24,4; 25,4}.

$$m = \frac{\sum y_i}{n} = \frac{(3{,}9 + \ldots + 25{,}4)}{30} = \frac{422{,}90}{30} = 14{,}1\ kg\ an^{-1}\ dia^{-1}$$

6.5.1.2. Média – dados agrupados

Tabela 6.2 – Distribuição de freqüências da produção de leite da Fazenda Nova Esperança, município de Itabuna, BA – abril de 2001

Classes	F_i	y_i	y_iF_i
00 ⊢ 04	1	2	2
04 ⊢ 08	2	6	12
08 ⊢ 12	7	10	70
12 ⊢ 16	10	14	140
16 ⊢ 20	5	18	90
20 ⊢ 24	3	22	66
24 ⊢ 28	2	26	52
	30		432

Fonte: Dados coletados na ordenha de 22/04/2001
Nota: Dados expressos em kg an^{-1} dia^{-1}.

$$m = \frac{\sum y_i F_i}{n} = \frac{432}{30} = 14{,}4\ kg\ an^{-1}\ dia^{-1}$$

Obs: $14{,}1\ kg.an^{-1}.dia^{-1} \approx 14{,}4\ kg\ an^{-1}\ dia^{-1}$

6.5.1.3. Mediana

Tabela 6.3 – Distribuição de freqüências da produção de leite da Fazenda Nova Esperança, município de Itabuna, BA – abril de 2001

Classes	F_i	F_{ac}	
00 ⊢ 04	1	1	
04 ⊢ 08	2	3	
08 ⊢ 12	7	10	
12 ⊢ 16	10	20	Classe mediana
16 ⊢ 20	5	25	
20 ⊢ 24	3	28	
24 ⊢ 28	2	30	

Fonte: Dados coletados na ordenha de 22/04/2001
Nota: Dados expressos em kg an^{-1} dia^{-1}.

$$\frac{30}{2} = 15^{o}$$

Classe md = 4ª

ℓ_{md} = 12; n = 30; $\sum f$ = 10; h = 4; Fmd = 10

$$\tilde{y} = \ell_{md} + \frac{\left(\frac{n}{2} - \sum f\right) \cdot h}{F_{md}} = 12 + \frac{\left(\frac{30}{2} - 10\right) \cdot 4}{10} = 14{,}0 \text{ kg an}^{-1} \text{ dia}^{-1}$$

Obs: o valor exato da mediana é 13,7 kg an^{-1} dia^{-1}

$$13{,}7\ kg\ an^{-1}\ dia^{-1} \approx 14{,}0\ kg\ an^{-1}\ dia^{-1}$$

6.5.1.4. Moda

Tabela 6.4 – Distribuição de freqüências da produção de leite da Fazenda Nova Esperança, município de Itabuna, BA – abril de 2001

Classes	F_i	F_{ac}	
00 ⊢ 04	1	1	
04 ⊢ 08	2	3	
08 ⊢ 12	7	10	
12 ⊢ 16	10	20	Classe modal
16 ⊢ 20	5	25	
20 ⊢ 24	3	28	
24 ⊢ 28	2	30	

Fonte: Dados coletados na ordenha de 22/04/2001
Nota: Dados expressos em kg an^{-1} dia^{-1}.

$$mo = \ell + \frac{\Delta_1}{\Delta_1 + \Delta_2} \cdot h = 12 + \frac{3}{3+5} \cdot 4 = 13{,}5\ kg\ an^{-1}\ dia^{-1}$$

6.5.2. Separatrizes ou quantis

6.5.2.1. Quartis

O método usado será o dos dados tabulados em tabela de freqüências. Existem outros métodos para determinação (a partir do rol) e os resultados nem sempre são coincidentes.

Tabela 6.5 – Distribuição de freqüências da produção de leite da Fazenda Nova Esperança, município de Itabuna, BA – abril de 2001

Classes	F_i	F_{ac}	
00 ⊢ 04	1	1	
04 ⊢ 08	2	3	
08 ⊢ 12	7	10	Classe do q_1
12 ⊢ 16	10	20	Classe do q_2
16 ⊢ 20	5	25	Classe do q_3
20 ⊢ 24	3	28	
24 ⊢ 28	2	30	

Fonte: Dados coletados na ordenha de 22/04/2001
Nota: Dados expressos em kg an^{-1} dia^{-1}.

6.5.2.1.1. Determinação do primeiro quartil (q_1)

Calcula-se $\frac{n}{4} \Rightarrow \frac{30}{4} = 7{,}5 \cong 8^o$

Identifica-se a classe q_1 pela $F_{ac} \Rightarrow$ classe $q_1 = 3^{\underline{a}}$

$\ell_{q_1} = 8$; i = 1; n = 30; $\sum f = 3$; h = 4; $F_{q_i} = 7$

$$q_i = \ell_{q_i} + \frac{\left(\frac{i \cdot n}{4} - \sum f\right) \cdot h}{Fq} = 8 + \frac{\left(\frac{1 \cdot 30}{4} - 3\right) \cdot 4}{7} = 10{,}57 \text{ kg an}^{-1} \text{ dia}^{-1}$$

6.5.2.1.2. Determinação do segundo quartil (q_2)

Calcula-se $\frac{2 \cdot n}{4} \Rightarrow \frac{2 \cdot 30}{4} = 15^o$

Identifica-se a classe q_2 pela $F_{ac} \Rightarrow$ classe $q_2 = 4^{\underline{a}}$

$\ell_{q_1} = 12$; i = 2; n = 30; $\sum f = 10$; h = 4; $Fq_i = 10$

$$q_i = \ell_{q_i} + \frac{\left(\frac{i \cdot n}{4} - \sum f\right) \cdot h}{F_{q_i}} = 12 + \frac{\left(\frac{2 \cdot 30}{4} - 10\right) \cdot 4}{10} = 14{,}00 \text{ kg an}^{-1} \text{ dia}^{-1}$$

Obs: q_2 = md = 14,00 kg an^{-1} dia^{-1}

6.5.2.1.3. Determinação do terceiro quartil (q_3)

Calcula-se $\frac{3 \cdot n}{4} \Rightarrow \frac{30}{4} = 22{,}5 \cong 23^o$

Identifica-se a classe q_3 pela $F_{ac} \Rightarrow$ classe $q_3 = 5^{\underline{a}}$

$\ell_{q_1} = 16$; i = 3; n = 30; $\sum f = 20$; h = 4; $F_{q_i} = 5$

$$q_i = \ell_{q_i} + \frac{\left(\frac{i \cdot n}{4} - \sum f\right) \cdot h}{F_{q_i}} = 16 + \frac{\left(\frac{3 \cdot 30}{4} - 20\right) \cdot 4}{7} = 17{,}43 \text{ kg an}^{-1} \text{ dia}^{-1}$$

6.5.3. Medidas de dispersão

6.5.3.1. Variância

6.5.3.1.1. Via cálculo da média amostral

$$s^2 = \frac{\sum d^2}{n-1} = \frac{(20{,}0 - 14{,}1)^2 + \ldots + (13{,}0 - 14{,}1)^2}{29} = 29{,}2 \left(kg\, an^{-1}\, dia^{-1}\right)^2$$

6.5.3.1.2. Sem utilizar a média amostral

$$s^2 = \frac{\sum y^2 - \frac{\left(\sum y\right)^2}{n}}{n-1} = \frac{6.809{,}5 - \frac{(422{,}9)^2}{30}}{30-1} = 29{,}2 \left(kg\, an^{-1}\, dia^{-1}\right)^2$$

6.5.3.2. Desvio Padrão

$$s = \sqrt{s^2} = \sqrt{29{,}24 \left(kg\, an^{-1}\, dia^{-1}\right)^2} = 5{,}4\ kg\, an^{-1}\, dia^{-1}$$

6.5.3.3. Coeficiente de variação

$$cv = \frac{s}{m} \cdot 100 = \frac{5{,}41\ kg\, an^{-1}\, dia^{-1}}{14{,}1\ kg\, an^{-1}\, dia^{-1}} \cdot 100 = 38{,}46\%$$

7. INTRODUÇÃO AO ESTUDO DE PROBABILIDADE

A teoria da probabilidade é essencial aos procedimentos utilizados na estatística inferencial.

Segundo alguns autores, a teoria da probabilidade originou-se como modelo explicativo para os jogos de azar: dados, moedas, etc.

No estudo dos fenômenos de observação são utilizados modelos:

a. Determinísticos;

b. Probabilísticos ou estocásticos.

Os fenômenos estudados pela estatística são fenômenos que mesmo em condições normais de experimentação variam de uma observação para outra, dificultando a previsão de um resultado futuro.

Para a explicação desses fenômenos adota-se o cálculo matemático probabilístico.

7.1. Caracterização de um experimento aleatório

Experimento:	qualquer processo que gera resultados bem definidos.
Ponto_amostral:	um resultado particular do experimento.

Experimento	Resultado experimental
Jogar uma moeda	cara, coroa
Retirar uma carta de um baralho	copa, ouro, paus, espada
Jogar um dado	1, 2, 3, 4, 5, 6
Selecionar uma peça para inspeção	defeituosa, não defeituosa

A análise desses experimentos revela que:

a. Cada experimento poderá ser repetido indefinidamente sob as mesmas condições.

b. Não se conhece “a priori” um particular resultado do experimento.

c. Quando o experimento for repetido um grande número de vezes surgirá uma regularidade, isto é, haverá uma estabilidade da fração:

$$f_i = \frac{n}{N}$$

onde:

f_i: freqüência relativa

n: número de sucessos de um particular resultado

N: número de repetições

c/k	lan	suc/lan = n/N	fi
c	1	1/1	**1,00**
k	2	1/2	**0,50**
k	3	1/3	**0,33**
k	4	1/4	**0,25**
c	5	2/5	**0,40**
c	6	3/6	**0,50**
k	7	3/7	**0,43**
c	8	4/8	**0,50**
c	9	5/9	**0,56**
k	10	5/10	**0,50**
c	11	6/11	**0,55**
c	12	7/12	**0,58**
k	13	7/13	**0,54**
k	14	7/14	**0,50**
c	15	8/15	**0,53**
k	16	8/16	**0,50**
c	17	9/17	**0,53**
k	18	9/18	**0,50**
c	19	10/19	**0,53**
k	20	10/20	**0,50**

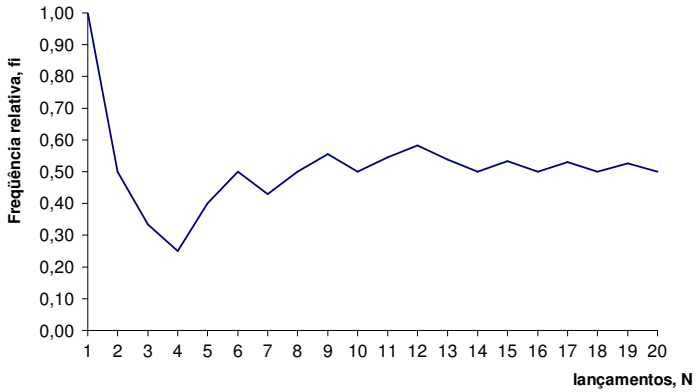


Figura 7.1 – Verificação da estabilização da freqüência relativa do número de caras de uma moeda não viciada em função do aumento do número de lançamentos.

7.2. Espaço amostral

Para cada experimento aleatório, E, define-se espaço amostral, S, o conjunto de todos os possíveis resultados desse experimento.

Exemplos:

E = jogar um dado e observar o número da face de cima: S = {1, 2, 3, 4, 5, 6}

E = jogar duas moedas e observar o resultado: S = {(cc), (ck), (kc), (kk)}

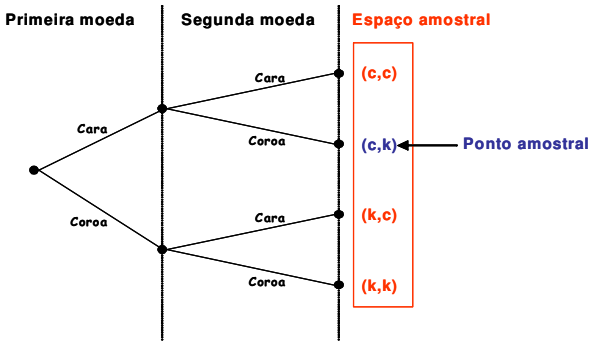


Figura 7.2 – Diagrama de árvore para um experimento de arremesso de duas moedas.

7.3. Evento

É um conjunto particular de resultados do espaço amostral do experimento, em termos de conjuntos, é um subconjunto de S.

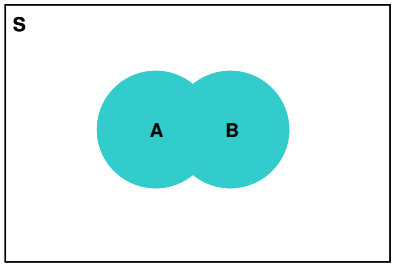
Considerando S e ϕ (conjunto vazio) como eventos:

S: é dito evento certo.

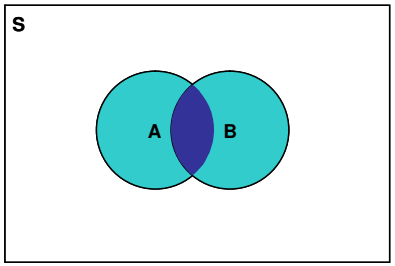
ϕ: é dito evento impossível.

Usando as operações com conjuntos, podem-se formar novos conjuntos, assim:

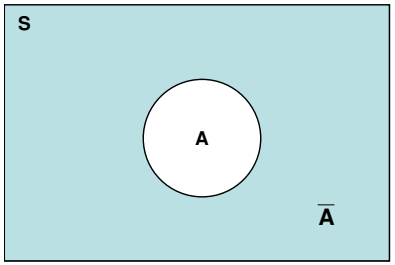
$A \cup B$: é o evento que ocorre se A ocorre, ou B ocorre, ou ambos ocorrem:



$A \cap B$: é o evento que ocorre se A e B ocorrem simultaneamente:



$\overline{A}$: é o evento que ocorre se A não ocorre:



Exemplos:

Seja o experimento E jogar três moedas e observar os resultados:

S = {(ccc), (cck), (ckc), (kcc), (kkk), (kkc), (kck), (ckk)}

Seja o evento A ocorrer pelo menos 2 caras

A = {(ccc), (cck), (ckc), (kcc)}

Seja o experimento E lançar um dado e observar o resultado:

S = {1, 2, 3, 4, 5, 6}

Seja o evento B ocorrer múltiplo de 2

B = {2, 4, 6}

Sendo S o espaço amostral finito, verifica-se que p^n fornece o número total de eventos extraídos de S:

$$S = p^n$$

onde:

p = valores possíveis (moeda = 2; dado = 6)

n = número de elementos do evento

Exemplo:

Seja o experimento E jogar três moedas e observar os resultados:

S = {(ccc), (cck), (ckc), (kcc), (kkk), (kkc), (kck), (ckk)}

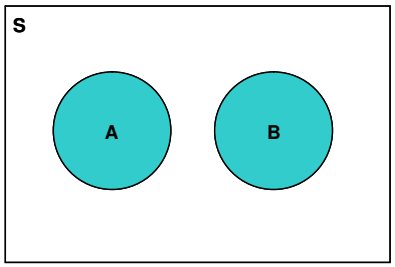
p=2

n=3

$$S = p^n = 2^3 = 8$$

7.4. Eventos mutuamente exclusivos

Dois eventos são denominados mutuamente exclusivos se eles não puderem ocorrer simultaneamente, isto é, $A \cap B = \phi$:



Exemplo:

Seja o experimento E lançar um dado e observar o resultado:

S = {1, 2, 3, 4, 5, 6}

Sejam os eventos:

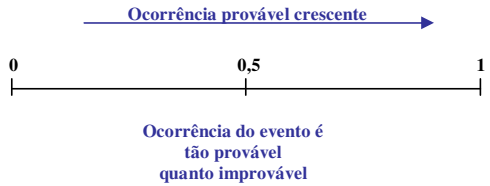
A = ocorrer número par

B = ocorrer número impar

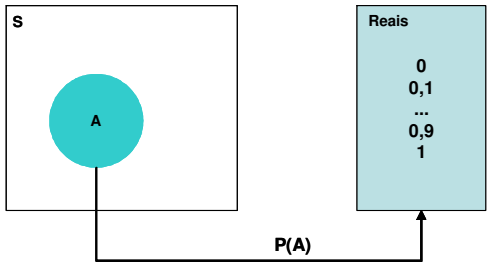
Então, A = {2, 4, 6} e B = {1, 3, 5}: observa-se que $A \cap B = \phi$

7.5. Conceito e definição de probabilidade

Conceito: a probabilidade é uma medida numérica da provável ocorrência de um evento:



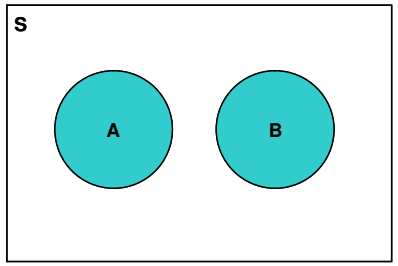
Definição: dado um experimento aleatório E, e S seu espaço amostral, a probabilidade de um evento A, indicada por P(A), é uma função definida em S que associa a cada evento um número real, satisfazendo os seguintes axiomas:



$P(S) = 1$

$0 \le P(A) \le 1$

Se A e B forem eventos mutuamente exclusivos, $A \cap B = \phi$, então $P(A \cup B) = P(A) + P(B)$

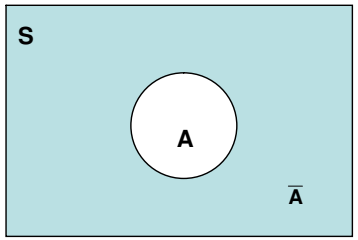


Axioma: proposição geral que não tem demonstração, recebida e aceita por todos como verdadeira e evidente.

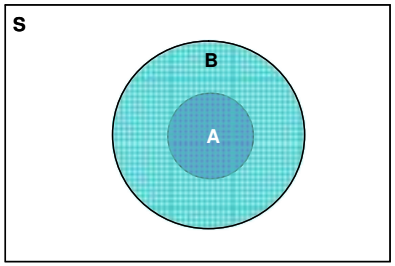
7.6. Principais teoremas da probabilidade

Se ϕ é um conjunto vazio, então: $P(\phi) = 0$

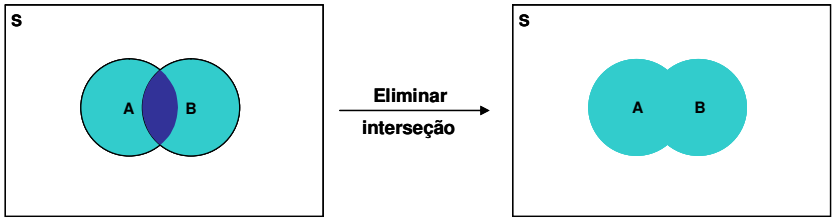
Se $\overline{A}$ é o complemento do evento A, então: $P(\overline{A}) = 1 - P(A)$



Se $(A \subset B)$,então: $P(A) \leq P(B)$



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



7.7. Probabilidades finitas dos espaços amostrais finitos

Seja um espaço amostral finito S= {a_1, a_2, ..., a_n}.

A cada evento simples a_i associa-se um número p_i denominado probabilidade de a_i, $P(a_i)$ ou simplesmente P_i, satisfazendo as seguintes condições:

$$p_i \geq 0 \quad (i = 1, 2, ..., n) \quad \text{e} \quad p_1 + p_2 + ... + p_n = 1$$

A probabilidade P(A) de cada evento composto (mais de um elemento ou ponto amostral) é então definida pela soma das probabilidades dos pontos amostrais de A.

Exemplo:

Três cavalos (A, B e C) estão em uma corrida; A tem duas vezes mais probabilidade de ganhar que B; e B tem duas vezes mais probabilidade de ganhar que C.

Quais são as probabilidades de vitória de cada um, isto é, P(A), P(B) e P(C)?

Fazendo P(C) = p

P(B) = 2p

P(A) = 4p

$$p + 2p + 4p = 1 \quad \therefore \quad p = \frac{1}{7}$$

Logo:

$$P(A) = \frac{4}{7}$$

$$P(B) = \frac{2}{7}$$

$$P(C) = \frac{1}{7}$$

Probabilidade de B ou C ganhar: $P(B \cup C) = \frac{2}{7} + \frac{1}{7} = \frac{3}{7}$

7.8. Espaços amostrais finitos equiprováveis

Quando se associa a cada ponto amostral a mesma probabilidade, o espaço amostral chama-se equiprovável ou uniforme.

Em particular, se S contém N pontos, então, a probabilidade de cada ponto será

$$\frac{1}{N}$$

Por outro lado, se um evento A contém n pontos, então:

$$P(A) = n \cdot \left(\frac{1}{N}\right) = \frac{n}{N}$$

Este método de avaliar P(A) é freqüentemente enunciado da seguinte maneira:

$$P(A) = \frac{\text{número de vezes que o evento } (A) \text{ pode ocorrer}}{\text{número de vezes que o espaço amostral } (S) \text{ ocorre}}$$

ou:

$$P(A) = \frac{NCF \ \ (\text{número de casos favoráveis})}{NTC \ \ (\text{número total de casos})}$$

Exemplo:

Escolher aleatoriamente (a expressão "aleatória" indica que o espaço amostral é equiprovável) uma carta de um baralho com 52 cartas.

A = {a carta é de ouros}

B = {a carta é uma figura}

Calcular P(A) e P(B)

$$P(A) = \frac{\text{número de ouros}}{\text{número de cartas}} = \frac{13}{52} = \frac{1}{4}$$

$$P(B) = \frac{\text{número de figuras}}{\text{número de cartas}} = \frac{12}{52} = \frac{3}{13}$$

Como se observa, o cálculo da probabilidade de um evento se resume a um problema de contagem.

Assim, a análise combinatória (teoria da contagem) tem fundamental importância para se contar o número de casos favoráveis e o total de casos.

A combinação de N elementos tomados (combinados) n a n, sendo n □ N, é calculada por:

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

Exemplo:

Num lote de 12 peças, 4 são defeituosas, duas peças são retiradas aleatoriamente uma após a outra sem reposição. Calcule:

P(A) = a probabilidade de ambas serem defeituosas:

$$A = C_2^4 = \binom{4}{2} = \frac{4!}{2!(4-2)!} = 6$$

$$S = C_2^{12} = \binom{12}{2} = \frac{12!}{2!(12-2)!} = 66$$

$$P(A) = \frac{NCF \ \ (\text{número de casos favoráveis})}{NTC \ \ (\text{número total de casos})} = \frac{6}{66} = \frac{1}{11}$$

P(B) = a probabilidade de ambas não serem defeituosas:

$$B = C_2^8 = \binom{8}{2} = \frac{8!}{2!(8-2)!} = 28$$

$$P(B) = \frac{NCF \ \ (\text{número de casos favoráveis})}{NTC \ \ (\text{número total de casos})} = \frac{28}{66} = \frac{14}{33}$$

A probabilidade de pelo menos uma ser defeituosa (C):

Observar que C é o complemento de B, ou seja C = $\overline{\mathbf{B}}$

$$P(C) = 1 - P(B) = 1 - \frac{14}{33} = \frac{19}{33}$$

7.9. Probabilidade condicional

Seja E lançar um dado, e o evento A = {3}. Então:

$$P(A) = \frac{1}{6}$$

Considere agora o evento B = {impar} = {1, 3, 5}.

É de grande importância para o cálculo das probabilidades calcular a probabilidade condicional.

Ou seja, avaliar a probabilidade do evento A condicionada ao evento B, simbolizada por P(A/B).

Lê-se probabilidade do evento A condicionada à ocorrência do evento B, ou ainda, probabilidade de A dado B:

$$P(A/B) = \frac{P(A \cap B)}{P(B)}; \text{ com } P(B) \neq 0, \text{ pois B já ocorreu}$$

Para aplicações, uma fórmula prática para o cálculo da probabilidade condicional é dada a seguir:

$$P(A/B)=\frac{P(A\cap B)}{P(B)}=\frac{\frac{NCF(A\cap B)}{NTC}}{\frac{NCF(B)}{NTC}}=\frac{NCF(A\cap B)}{NCF(B)}$$

Exemplo:

Dois dados são lançados. Consideremos os eventos:

$$A=\{(x_1,x_2)/x_1+x_2=10\}$$
$$B=\{(x_1,x_2)/x_1>x_2\}$$

Onde x_1 é o resultado do dado 1 e x_2 é o resultado do dado 2.

Avaliar P(A); P(B); P(A/B) e P(B/A)

$$S=\begin{Bmatrix}(1,1)&(1,2)&(1,3)&(1,4)&(1,5)&(1,6)\\(2,1)&(2,2)&(2,3)&(2,4)&(2,5)&(2,6)\\(3,1)&(3,2)&(3,3)&(3,4)&(3,5)&(3,6)\\(4,1)&(4,2)&(4,3)&(4,4)&(4,5)&(4,6)\\(5,1)&(5,2)&(5,3)&(5,4)&(5,5)&(5,6)\\(6,1)&(6,2)&(6,3)&(6,4)&(6,5)&(6,6)\end{Bmatrix}$$

$$A=[(x_1,x_2)/x_1+x_2=10]=[(4,6);(6,4);(5,5)]$$

$$B=[(x_1,x_2)/x_1>x_2]=\begin{Bmatrix}(2,1);\\(3,1);(3,2);\\(4,1);(4,2);(4,3);\\(5,1);(5,2);(5,3);(5,4);\\(6,1);(6,2);(6,3);(6,4);(6,5)\end{Bmatrix}$$

$$P(A)=\frac{NCF(A)}{NTC}=\frac{3}{36}=\frac{1}{12}$$

$$P(B)=\frac{NCF(B)}{NTC}=\frac{15}{36}=\frac{5}{12}$$

$$P(A/B)=\frac{NCF(A\cap B)}{NCF(B)}=\frac{1}{15}$$

Obs: *notar que apenas o par* (6,4) *é favorável ao evento* $(A\cap B)$

$$P(B/A)=\frac{NCF(A\cap B)}{NCF(A)}=\frac{1}{3}$$

7.10. Teorema do produto

A partir da definição de probabilidade condicional pode-se enunciar o teorema do produto:

"A probabilidade da ocorrência simultânea de dois eventos, A e B, do mesmo espaço amostral, é igual ao produto da probabilidade de um deles pela probabilidade condicional do outro, dado o primeiro."

Assim:

$$P(A/B)=\frac{P(A\cap B)}{P(B)}\Rightarrow P(A\cap B)=P(B)\cdot P(A/B)$$

$$P(B/A)=\frac{P(A\cap B)}{P(A)}\Rightarrow P(A\cap B)=P(A)\cdot P(B/A)$$

Exemplo:

Num lote de 12 peças, 4 são defeituosas. 2 peças são retiradas uma após a outra sem reposição. Qual a probabilidade de ambas não serem defeituosas?

A = {a primeira peça é boa}

B = {a segunda peça é boa}

$$P(A\cap B)=P(A)\cdot P(B/A)=\frac{8}{12}\cdot\frac{7}{11}=\frac{56}{132}=\frac{14}{33}$$

7.11. Independência estatística

Um evento A é dito independente de um evento B, se a probabilidade de A ocorrer não é influenciada pelo fato de B ter ocorrido ou não.

Em outras palavras, se a probabilidade de A é igual à probabilidade condicional de A dado B, isto é, se:

$$P(A)=P(A/B)$$

Em conseqüência, se A é independente de B, B é independente de A, assim:

$$P(B) = P(B/A)$$

Considerando o teorema do produto, pode-se afirmar que se A e B são independentes, então:

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

A equação acima é usada como definição formal de independência.

Dados "n" eventos A_1, A_2, ..., A_N, diz-se que eles são independentes se o forem 2 a 2, 3 a 3, n a n.

Isto é, se as igualdades abaixo forem verificadas:

$$P(A_1 \cap A_2) = P(A_1) \cdot P(A_2)$$

$$P(A_{n-1} \cap A_n) = P(A_{n-1}) \cdot P(A_n)$$

$$P(A_1 \cap A_2 \cap A_3) = P(A_1) \cdot P(A_2) \cdot P(A_3)$$

$$P(A_1 \cap A_2 \cap ... \cap A_n) = P(A_1) \cdot P(A_2) \cdots P(A_{n-1}) \cdot P(A_n)$$

Exemplo 1:

Num lote de 10 peças, 4 são defeituosas. 2 peças são retiradas uma após a outra com reposição. Qual a probabilidade de que ambas sejam boas?

A = {a primeira peça é boa}

B = {a segunda peça é boa}

Notar que A e B são independentes, pois $P(B) = P(B/A)$

$$P(A \cap B) = P(A) \cdot P(B) = \frac{6}{10} \cdot \frac{6}{10} = \frac{9}{25}$$

Exemplo 2:

Em um lançamento de um par de moedas não viciadas, então:

S = {(cc), (ck), (kc), (kk)}

A = {cara na primeira moeda} = {(cc), (ck)}

B = {cara na segunda moeda} = {(cc), (kc)}

C = {cara apenas em uma moeda} = {(ck), (kc)}

$$P(A) = P(B) = P(C) = \frac{2}{4} = \frac{1}{2}$$

$$P(A \cap B) = (cc) = \frac{1}{4}$$

$$P(A \cap C) = (ck) = \frac{1}{4}$$

$$P(B \cap C) = (kc) = \frac{1}{4}$$

Os eventos são independentes entre si dois a dois.

Entretanto, os eventos não são todos independentes entre si, pois:

8. VARIÁVEIS ALEATÓRIAS

8.1. Conceitos

Os estudos estatísticos são baseados em amostras vindas de uma população real ou fictícia e nos casos mais simples, uma única medição é feita em cada indivíduo retirado da população.

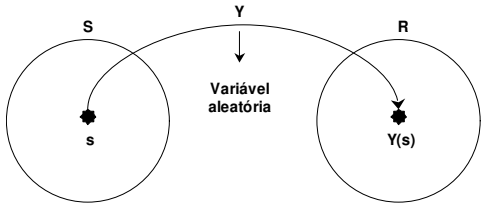
Como não se pode prever com certeza qual o resultado desta medição, ela é uma variável aleatória.

Como toda variável aleatória, a medição acima possui um conjunto de valores que ela pode assumir. Além disto, como em geral nem todo valor possível é igualmente provável, é necessário descrever as diferentes probabilidades associadas a esses valores.

Assim, uma variável aleatória é toda e qualquer variável associada a uma probabilidade, isto é, seus valores estão associados a um experimento aleatório.

Descrição numérica dos resultados de um experimento.

Em geral são identificadas por letras maiúsculas e cada um de seus possíveis valores por letras minúsculas correspondentes.



8.2. Definição

Seja E um experimento e S o espaço amostral associado ao experimento. Uma função Y, que associe a cada elemento (s ∈ S) um número real Y(s) é denominada variável aleatória.

Exemplo:

E: lançamento de duas moedas

Y: número de caras obtidas nas duas moedas

S = {(c,c), (c,k), (k,c), (k,k)}

Y(k,k) = 0	→ com probabilidade 1/4
Y(c,k) = Y(k,c) = 1	→ com probabilidade 2/4
Y(c,c) = 2	→ com probabilidade 1/4

8.3. Observações

Apesar da terminologia "variável aleatória", ela é uma função cujo domínio é S e o contradomínio é R.

O uso das variáveis aleatórias equivale a descrever os resultados de um experimento aleatório por meio de números, ao invés de eventos, o que possibilita o tratamento matemático adequado.

Se S é numérico, então Y(s) = s

8.4. Variável aleatória discreta (VAD) e contínua (VAC)

Uma variável aleatória Y será discreta se o número de valores de Y (seu contradomínio), finito ou infinito, for numerável. Ou seja, entre quaisquer de dois elementos vizinhos não há quantidades intermediárias.

O que implica apenas em números inteiros.

Exemplo: tudo que se conta.

Quadro 8.1 – Exemplos de variáveis aleatórias discretas

Experimento	Variável aleatória (Y)	Possíveis valores para a VAD
Jogar uma moeda	Valor da face virada para cima	Y = 0 para cara Y = 1 para coroa
Inspecionar uma esteira de empacotamento de leite	Número pacotes defeituosos	Y = 0 ... ∞
Vender um lote de animais	Porte do cliente	Y = 0 se grande pecuarista Y = 1 se pequeno criador

Caso seu contradomínio seja um intervalo ou uma coleção de intervalos, ela será contínua. Ou seja, entre quaisquer de dois elementos vizinhos há quantidades intermediárias infinitas, dependentes da sensibilidade do instrumento de medida.

O que pode implicar em valores fracionários.

Exemplo: tudo que se mede (massa, temperatura, tempo, distância, área, etc).

Quadro 8.2 – Exemplos de variáveis aleatórias contínuas

Experimento	Variável aleatória (Y)	Possíveis valores para a VAC
Trabalhar em um projeto	Percentual executado após 30 dias	$0 \leq Y \geq 100\%$
Observar um operador de máquina agrícola	Tempo ocioso em um dia	$0 \leq Y \geq 24$ hs
Pulverizar 10.000 m^2 de uma área agrícola	Volume de água gasto	$0 \leq Y \geq \infty$

8.5. Função de probabilidades

Chama-se função de probabilidade da VAD Y, a função

$$f(Y = y_i) = f(y_i) = p_i$$

que a cada valor y_i associa sua probabilidade de ocorrência.

Tabela 8.1 – Distribuição de probabilidade do número de pacotes de leite defeituosos do laticínios A, em mm/aa, Local

Y	f (Y)
10	0,18
20	0,39
30	0,24
40	0,14
50	0,04
60	0,01
Total	1,00

A função $f(y_i)$ será uma função de probabilidade se satisfizer às seguintes condições:

$$a)\ f(y_i) \geq 0 \ para\ todo\ y_i$$

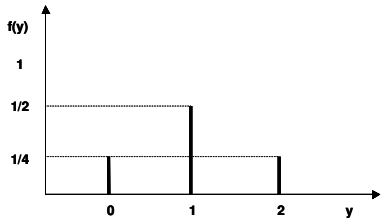
$$b)\sum_{i=1}^{I} f(y_i) = 1$$

À coleção de pares [yi, f(yi)], é denominada distribuição de probabilidade da VAD Y, que pode ser representada por meio de tabela, gráfico ou fórmula:

Tabela

y	f(y)
0	¼
1	½
2	¼

Gráfico



Fórmula

$$f(y) = \frac{1}{4}\binom{2}{y}, \ para\ y = 0,1,2$$

A distribuição de probabilidades para uma variável aleatória descreve como as probabilidades estão distribuídas sobre os valores da variável aleatória.

A principal vantagem de definir uma VA e sua distribuição de probabilidades é que, uma vez que a distribuição de probabilidade seja conhecida, é relativamente fácil determinar a probabilidade de eventos que podem ser do interesse de um tomador de decisões.

8.6. Função de repartição ou distribuição acumulada

Seja Y uma VAD, define-se função de repartição ou função de distribuição acumulada da VAD Y, no ponto y, como sendo a probabilidade de que Y assuma um valor menor ou igual a y, isto é:

$$F(y) = P(Y \leq y)$$

Propriedades:

$$F(y) = \sum_{y_i \leq y} P(y_i) \qquad \text{(cálculo de F(y))}$$

$$F(-\infty) = 0$$

$$F(+\infty) = 1$$

$$P(a < Y \leq b) = F(b) - F(a)$$

$$P(a \leq Y \leq b) = F(b) - F(a) + P(Y = a)$$

$$P(a < Y < b) = F(b) - F(a) - P(Y = b)$$

Exemplo:

Admitamos que a VAD Y tome os valores 0, 1 e 2 com probabilidades 1/3, 1/6 e 1/2 respectivamente.

Então:

$$F(y) = 0 \quad se \quad y < 0$$

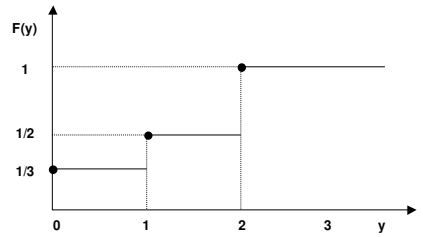
$$F(y) = \frac{1}{3} \quad se \quad 0 \leq y < 1$$

$$F(y) = \frac{1}{2} \quad se \quad 1 \leq y < 2$$

$$F(y) = 1 \quad se \quad y \geq 2$$

y	0	1	2
f(y)	1/3	1/6	1/2
F(y)	1/3	1/3 + 1/6 = 1/2	1/2 + 1/2 = 1

O gráfico de F(y) é:



8.7. Função densidade de probabilidade

Seja Y uma VAC, a função densidade de probabilidade f(y) é uma função que satisfaz as seguintes condições:

$$a)\ f(y) \geq 0 \ \ para\ todo\ y \in [a,b]\ com\ a < b$$
$$b) \int_a^b f(y)dy = 1$$

Além disso, define-se, para qualquer [c < d], contido no intervalo [a,b]

$$P(c < Y < d) = \int_c^d f(y)dy$$

Observações importantes:

A definição acima mostra que a probabilidade de qualquer valor especificado de Y, por exemplo, y_0, tem P(Y = y_0) = 0, pois:

$$P(Y = y_0) = \int_{y_0}^{y_0} f(y)dy = 0$$

Sendo assim, as probabilidades abaixo serão todas iguais, se Y for uma VAC:

$$P(a \leq Y \leq b) = P(a \leq Y < b) = P(a < Y \leq b) = P(a < Y < b)$$

Notar que f(y), densidade de probabilidade, não é probabilidade. Somente quando a função for integrada entre dois limites, ela produzirá uma probabilidade, que será a área sob a curva da função densidade de probabilidade entre y = a e y = b, considerando a < b.

Para VADs, a probabilidade está concentrada em pontos isolados da reta real.

No caso de VACs, a probabilidade está espalhada de modo contínuo em segmentos da reta real.

Quanto à função de repartição, neste caso ela é definida como:

$$F(y) = \int_{-\infty}^{y} f(y)dy$$

A área total sob a curva de probabilidade vale sempre um (1):

$$\int_{-\infty}^{+\infty} f(y)dy = 1$$

Exemplo:

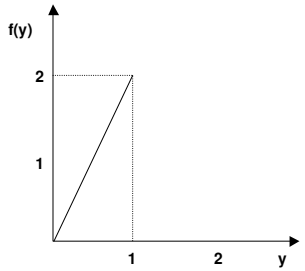
Seja Y uma VAC, com a seguinte função densidade de probabilidade:

$$f(y) = \begin{cases} 2y \ \ para\ \ 0 < y < 1 \\ 0\ para\ quaisquer\ outros\ valores \end{cases}$$

f(y) assim definida, é realmente uma função densidade, pois:

$$f(y) \geq 0 \ \ e\ \ para\ todo\ \ y \in Ry$$
$$\int_{-\infty}^{+\infty} f(y)dy = \int_{-\infty}^{0} dy + \int_0^1 2ydy + \int_1^{+\infty} dy = 2 \times \left(\frac{y^2}{2} \right) \Big|_{y=1} - \left(2 \times \left(\frac{y^2}{2} \right) \Big|_{y=0} \right) = 1 = y^2$$

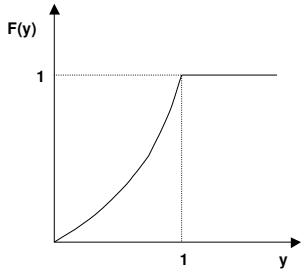
Seu gráfico será:



Quanto a F(y) tem-se:

$$para\ \ y < 0 \qquad F(y) = \int_{-\infty}^{0} 0dy = 0$$
$$para\ \ 0 \leq y < 1 \qquad F(y) = \int_{-\infty}^{0} 0dy + \int_0^1 2ydy = y^2$$
$$para\ \ y \geq 1 \qquad F(y) = \int_{-\infty}^{0} 0dy + \int_0^1 2ydy + \int_1^{+\infty} 0dy = 1$$

Cujo gráfico será:



O gráfico de F(y) no caso de uma VAD é constituído por segmentos de retas horizontais (degraus), e no caso de uma VAC, ele é contínuo para todo y.

8.8. Esperança matemática (média ou valor esperado)

A esperança matemática corresponde ao que se espera que aconteça em média.

Seja Y uma VAD com a seguinte distribuição de probabilidade:

y_i	y_1	...	y_n	Total
$P(y_i)$	$P(y_1)$	…	$P(y_n)$	1

Define-se a esperança matemática de, E(Y), por:

$$E(Y) = \mu_y = \mu = y_i.P(y_i) + ...+ y_n.P(y_n)$$

$$E(Y) = \sum y_i \cdot P(y_i)$$

Exemplo:

E = lançamento de um dado

Y = ponto obtido

Y = 1, 2, 3, 4, 5, 6

P(Y) = $\frac{1}{6}$ para todo Y

$$E(Y) = 1\cdot\frac{1}{6} + 2\cdot\frac{1}{6} + 3\cdot\frac{1}{6} + 4\cdot\frac{1}{6} + 5\cdot\frac{1}{6} + 6\cdot\frac{1}{6} = 3{,}5$$

Interpretação:

Se um dado, não viciado, for lançado um número muito grande de vezes, caracterizando uma população (valores obtidos), a média destes valores será 3,5.

A esperança matemática de uma VAC Y é definida por:

$$E(Y) = \int_{-\infty}^{+\infty} yf(y)dy$$

Exemplo:

$$f(y) = \begin{cases} \frac{1}{2}y, & para\, 0 \le y \le 2 \\ 0, & caso contrário \end{cases}$$

$$E(Y) = \int_{-\infty}^{+\infty} yf(y)dy = \int_0^2 y\frac{1}{2}ydy = \int_0^2 \frac{1}{2}y^2dy = \left[\frac{y^3}{6}\right]_0^2 = \frac{2^3}{6} - \frac{0^3}{6} = \frac{8}{6} - 0 = \frac{8}{6} = \frac{4}{3}$$

Propriedades da esperança matemática

a. E(K) = K

A esperança de uma constante, é a própria constante.

b. E(Y ± K) = E(Y) ± K

Se uma constante é adicionada ou subtraída a cada valor da variável aleatória (Y), a esperança fica adicionada, ou subtraída, desta constante.

c. E(KY) = K.E(Y)

Se uma constante é multiplicada a cada valor da variável aleatória (Y), a esperança fica multiplicada desta constante.

d. E(Y ± Z) = E(Y) ± E(Z)

A esperança da soma ou subtração de duas variáveis aleatórias quaisquer é igual à soma ou subtração de suas esperança.

e. Se Y e Z são independentes: E(YZ) = E(Y). E(Z)

A esperança do produto de duas variáveis aleatórias independentes é o produto das esperanças.

8.9. Variância

Por definição, a variância de uma variável aleatória (VA) Y, de população infinita, é

$$\sigma^2 = V(Y) = E[Y - E(Y)]^2 = E(Y - \mu)^2$$

Uma fórmula utilizada em algumas circunstâncias é dada a seguir

$$V(Y) = E(Y^2) - [E(Y)]^2$$

$$V(Y) = E[Y - E(Y)]^2$$
$$V(Y) = E[Y^2 - 2YE(Y) + [E(Y)]^2]$$
$$V(Y) = E(Y^2) - 2E(Y)E(Y) + [E(Y)]^2$$
$$V(Y) = E(Y^2) - 2[E(Y)]^2 + [E(Y)]^2$$
$$V(Y) = E(Y^2) - [E(Y)]^2$$

Propriedades da variância

a. A variância de uma constante é igual a zero

$$V(K) = 0$$

$$V(K) = E[K - E(K)]^2$$
$$V(K) = E[K - K]^2$$
$$V(K) = 0$$

b. Somando-se ou subtraindo-se uma constante a uma VA, sua variância não se altera

$$V(Y \pm K) = V(Y)$$

$$V(Y \pm K) = E[(Y \pm K) - E(Y \pm K)]^2$$
$$V(Y \pm K) = E[(Y) - E(Y) \pm (K - K)]^2$$
$$V(Y \pm K) = E[Y - E(Y)]^2$$
$$V(Y \pm K) = V(Y)$$

c. Multiplicando-se uma VA por uma constante, sua variância fica multiplicada pelo quadrado da constante

$$V(KY) = K^2 \cdot V(Y)$$

$$V(KY) = E[KY - E(KY)]^2$$
$$V(KY) = E[KY - KE(Y)]^2$$
$$V(KY) = E\{k^2[Y - E(Y)]^2\}$$
$$V(KY) = k^2 \cdot E[Y - E(Y)]^2$$
$$V(KY) = k^2 \cdot V(Y)$$

d. A variância da soma de duas VAs independentes é igual a soma das variâncias das duas variáveis

$$V(Y + Z) = V(Y) + V(Z)$$

$$V(Y + Z) = E[Y + Z]^2 - [E(Y + Z)]^2$$
$$V(Y + Z) = E(Y^2 + 2YZ + Z^2 - [E(Y) + E(Z)]^2$$
$$V(Y + Z) = E(Y^2) + 2E(YZ) + E(Z^2) - \{[E(Y)^2] + 2E(Y)E(Z) + [E(Z)]^2\}$$
$$V(Y + Z) = E(Y^2) + 2E(YZ) + E(Z^2) - [E(Y)^2] - 2E(Y)E(Z) - [E(Z)]^2$$

Se Y e Z são independentes: E(YZ)=E(Y).E(Z)

$$V(Y + Z) = E(Y^2) + 2E(Y)E(Z) + E(Z^2) - [E(Y)^2] - 2E(Y)E(Z) - [E(Z)]^2$$
$$V(Y + Z) = \{E(Y^2) - [E(Y)]^2\} + \{E(Z^2) - [E(Z)]^2\}$$
$$V(Y + Z) = V(Y) + V(Z)$$

Do mesmo modo V(Y- Z) = V(Y) + V(Z)

8.10. Covariância

Dadas duas variáveis aleatórias, Y e Z, quaisquer, a covariância entre Y e Z, denotada por Cov(Y,Z), é por definição:

$$Cov(Y,Z) = E[Y - E(Y)][Z - E(Z)]$$
$$Cov(Y,Z) = E(Y - \mu_Y)(Z - \mu_Z)$$

Será demonstrado que

$$V(Y + Z) = V(Y) + V(Z) + 2Cov(Y,Z)$$
$$V(Y - Z) = V(Y) + V(Z) - 2Cov(Y,Z)$$

$$V(Y + Z) = V(Y) + V(Z) + 2Cov(Y,Z)$$

$$V(Y + Z) = E\{[Y - E(Y)] + [Z - E(Z)]\}^2$$
$$V(Y + Z) = E\{(Y - \mu_Y)^2 + 2(Y - \mu_Y)(Z - \mu_Z) + (Z - \mu_Z)^2\}$$
$$V(Y + Z) = E\{(Y - \mu_Y)^2 + (Z - \mu_Z)^2 + 2(Y - \mu_Y)(Z - \mu_Z)\}$$
$$V(Y + Z) = E(Y - \mu_Y)^2 + E(Z - \mu_Z)^2 + 2E(Y - \mu_Y)(Z - \mu_Z)$$
$$V(Y + Z) = V(Y) + V(Z) + 2Cov(Y,Z)$$

$$V(Y - Z) = V(Y) + V(Z) - 2Cov(Y,Z)$$

$$V(Y - Z) = E\{[Y - E(Y)] - [Z - E(Z)]\}^2$$
$$V(Y - Z) = E\{(Y - \mu_Y)^2 - 2(Y - \mu_Y)(Z - \mu_Z) + (Z - \mu_Z)^2\}$$
$$V(Y - Z) = E\{(Y - \mu_Y)^2 + (Z - \mu_Z)^2 - 2(Y - \mu_Y)(Z - \mu_Z)\}$$
$$V(Y - Z) = E(Y - \mu_Y)^2 + E(Z - \mu_Z)^2 - 2E(Y - \mu_Y)(Z - \mu_Z)$$
$$V(Y - Z) = V(Y) + V(Z) - 2Cov(Y,Z)$$

No estudo de correlação linear simples, será verificado que a covariância fornece o grau de associação linear entre duas variáveis aleatórias.

Ou seja, conhecendo-se uma variável, pode-se saber muito a respeito da outra, se a correlação entre as duas for elevada. Esta medida, correlação, é bastante utilizada na análise quantitativa de experimentos.

9. CORRELAÇÃO LINEAR SIMPLES

9.1. Introdução

A análise de correlação linear simples (Pearson, 1896) , outros tipos de análise de correlação (parcial, múltipla, canônica) e a análise de regressão, são técnicas estatísticas utilizadas no estudo quantitativo de experimentos.

Enquanto a análise de regressão linear simples nos mostra como duas variáveis se relacionam linearmente, a análise de correlação linear simples nos mostra apenas o grau da associação, ou de proporcionalidade, entre estas duas variáveis.

Conquanto a correlação seja uma técnica menos potente que a regressão, as duas se acham tão intimamente ligadas que a correlação freqüentemente é útil na interpretação da regressão.

Muitas técnicas de análise multivariada usam a correlação como medida estatística básica para estudar a associação entre variáveis aleatórias.

9.2. Definição

ρ : Correlação populacional

r : Estimativa da correlação ou correlação amostral

$$\rho = \frac{COV_{Pop}(Y_1, Y_2)}{\sigma(Y_1) \cdot \sigma(Y_2)}$$

$$r = \frac{\mathrm{cov}_{Amo}(Y_1, Y_2)}{s(Y_1) \cdot s(Y_2)}$$

$$COV(Y_1, Y_2) = E\left[\left(Y_1 - E(Y_1)\right) \cdot \left(Y_2 - E(Y_2)\right)\right]$$

$$COV_{Pop}(Y_1, Y_2) = \frac{\Sigma\left[\left(Y_1 - \mu(Y_1)\right) \cdot \left(Y_2 - \mu(Y_2)\right)\right]}{N}$$

$$\mathrm{cov}_{Amo}(Y_1, Y_2) = \frac{\Sigma\left[\left(Y_1 - \mu(Y_1)\right) \cdot \left(Y_2 - \mu(Y_2)\right)\right]}{n}$$

$$\mathrm{cov}_{Amo}(Y_1, Y_2) = \frac{\Sigma\left[\left(Y_1 - m(Y_1)\right) \cdot \left(Y_2 - m(Y_2)\right)\right]}{n-1}$$

9.3. Conceitos e compreensão a partir de um exemplo

Consideremos duas variáveis aleatórias:

M : rendimento acadêmico em matemática

L : rendimento acadêmico em línguas

Quadro 12.1 - Rendimento acadêmico

Obs	01	02	03	04	05	06	07	08
M	36	80	50	58	72	60	56	68
L	35	65	60	39	48	44	48	61

$\sum M = 480$ $\quad$ $\sum L = 400$

$m(M) = 60$ $\quad$ $m(L) = 50$

$s(M) = 13{,}65$ $\quad$ $s(L) = 10{,}93$

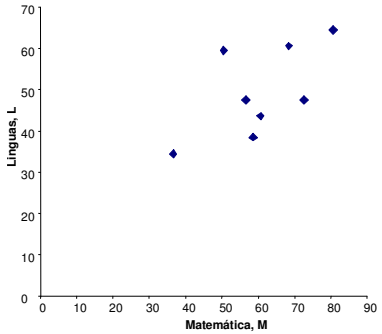
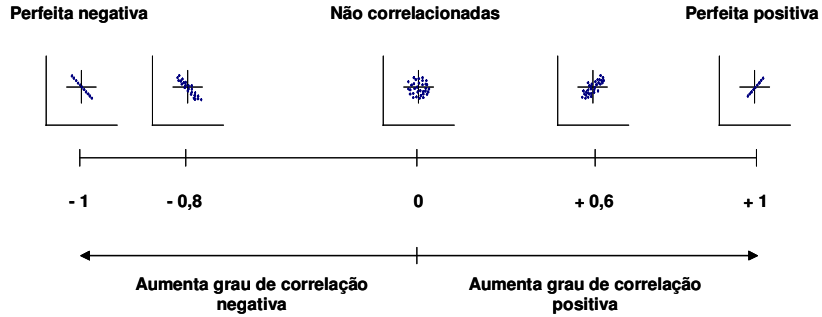


Figura 12.1 - Gráfico da dispersão entre M e L.

Necessita-se de um índice que forneça o grau de associação, ou de proporcionalidade, linear entre as duas variáveis aleatórias (M e L).



Para testar Σml como este índice:

$$m = m_i - m(M)$$

$$l = l_i - m(L)$$

$$\operatorname{cov}(Y_1, Y_2) = \frac{1}{n-1}.\Sigma\left[\left(Y_1 - m(Y_1)\right)\cdot\left(Y_2 - m(Y_2)\right)\right]$$

deve-se sobrepor aos pontos dispersos nos eixos cartesianos, os eixos das médias de matemática e línguas (M e L):

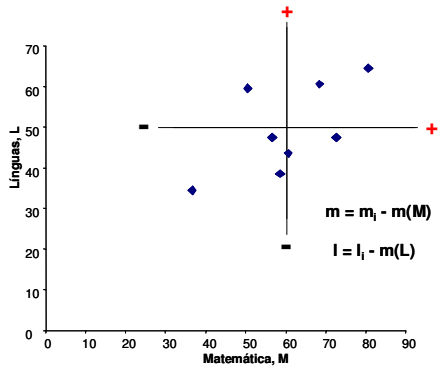
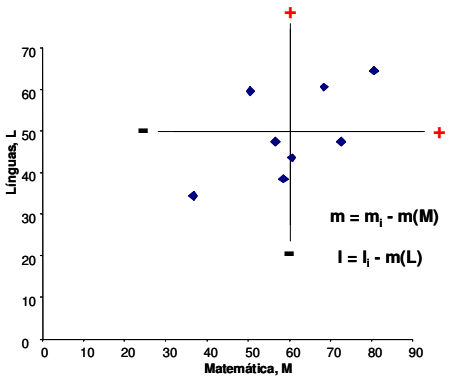


Figura 12.2 - Gráfico da dispersão entre M e L com as médias transladadas.

Quadro 12.2 – Cálculo do índice Σml

Obs	M	L	m = (M_i – m(M))	l = (L_i – m(L))	m.l
1	36	35	- 24	- 15	360
2	80	65	20	15	300
3	50	60	- 10	10	- 100
4	58	39	- 2	- 11	22
5	72	48	12	- 2	- 24
6	60	44	0	- 6	0
7	56	48	- 4	- 2	8
8	68	61	8	11	88
	m(M) = 60	m(L) = 50			Σml = 654
	s(M) = 13,65	s(L) = 10,93			



Se M e L caminharem juntas, isto é, enquanto uma aumenta a outra também aumenta, e enquanto uma diminui a outra também diminui, a maior parte das observações recairão nos 1º e 3º quadrantes.

Conseqüentemente, a maior parte dos produtos (m.l) serão positivos, bem como sua soma (Σml), demonstrando um relacionamento positivo entre M e L.

Mas se M e L estão relacionadas negativamente, isto é, uma aumenta enquanto a outra diminui, a maior parte das observações recairão nos 2º e 4º quadrantes, dando um valor negativo para o índice Σml.

Concluí-se, então, que como índice do grau de associação, ou proporcionalidade, entre as duas variáveis, Σml, pelo menos, tem sinal correto.

Além disso, quando não houver relação entre M e L as observações tenderão a serem distribuídas igualmente pelos quatro quadrantes, os termos positivos e negativos se cancelarão e Σml tenderá para zero.

Há apenas duas maneiras de melhorar Σml como medida do grau de associação, ou proporcionalidade, linear entre duas variáveis aleatórias:

i. Primeiro: Σml é dependente do tamanho da amostra:

Suponha que tivéssemos observado o mesmo diagrama de dispersão para uma amostra com o dobro do tamanho.

Então, Σml também seria o dobro, muito embora a configuração da tendência das variáveis permaneça a mesma.

Para evitar este problema dividimos Σml pelo tamanho da amostra:

$$\frac{\sum ml}{n-1} = \frac{1}{n-1}\left[\sum (M_i - m(M)) \times (L_i - m(L)\right]$$

Ao ser eliminada a influência do tamanho da amostra, nesta medida do grau de associação, ou proporcionalidade, linear entre duas variáveis aleatórias, obtém-se uma medida bastante útil em estatística denominada covariância, neste caso representada por COV(M,L):

$$\text{cov}(M,L) = \frac{\sum ml}{n-1} = \frac{\sum (M_i - m(M)) \times (L_i - m(L))}{n-1}$$

ii. Segundo: pode-se perceber que a covariância tem um ponto fraco: é influenciada pelas unidades de medida das variáveis envolvidas.

Suponha que o teste de matemática tenha valor 50 ao invés de 100.

Os valores relacionados aos desvios de matemática, m, serão apenas a metade, e isto irá influenciar o valor da covariância - muito embora, em essência, o grau da associação, ou proporcionalidade, linear entre matemática e línguas não tenha se modificado.

Em outras palavras, a covariância depende das unidades de medida das variáveis.

Esta dificuldade pode ser contornada se medirmos ambas as variáveis em termos de uma unidade padronizada.

Ou seja, dividindo-se m e l pelos seus respectivos desvios padrões:

$$\frac{1}{n-1}\sum\left(\frac{m}{s(M)}\right)\left(\frac{l}{s(L)}\right) = \frac{1}{n-1}\left[\sum\left(\frac{M_i - m(M)}{s(M)}\right) \times \left(\frac{L_i - m(L)}{s(L)}\right)\right]$$

Ao eliminar a influência do tamanho da amostra (a), obtém-se a covariância; e ao eliminar a influência das unidades de medida das variáveis (b) define-se, finalmente, o que é denominado correlação linear simples entre M e L, r(M,L), por vezes chamada de correlação de Pearson:

$$r(M,L) = \frac{\text{cov}(M,L)}{s(M) \times s(L)}$$

Assim, para calcularmos a correlação entre M e L:

$$\text{cov}(M,L) = \frac{\sum (M_i - m(M)) \times (L_i - m(L))}{n-1} = \frac{654}{7} = 93{,}43$$

$$r(M,L) = \frac{\text{cov}(M,L)}{s(M) \times s(L)} = \frac{93{,}43}{13{,}65 \times 10{,}93} = 0{,}63$$

Observações:

- Limites da correlação: $-1 \leq (\rho \ ou \ r) \leq +1$

9.4. Pressuposições da correlação:

- O relacionamento entre as variáveis tem forma linear.
- As duas variáveis são aleatórias por natureza e medidas em escalas intervalares ou proporcionais, não podendo ser categóricas ou nominais.
- As variáveis apresentam distribuição normal bivariada.

Enquanto medida do grau de associação, ou proporcionalidade, entre duas variáveis aleatórias a covariância possui uma vantagem: não é influenciada pelo tamanho da amostra; e uma desvantagem: é influenciada pela unidade de medida das variáveis.

Ao dividi-la pelos respectivos desvios padrões das variáveis aleatórias obtém-se o coeficiente de correlação linear, r(M,L), que não é influenciado nem pelo tamanho da amostra e nem pelas unidades de medida das variáveis.

O quadrado do coeficiente de correlação indica a proporção da variação em uma variável explicada ou predita pela variação na outra variável:

- r = 0,63 → r^2 = 0,3922
- r^2 = 39,22%
- 39,22% da variação observada em M é explicada pela variação em L, e vice-versa.

Uma fórmula prática para cálculo da correlação linear simples é apresentada abaixo:

$$r(M,L) = \frac{\text{cov}(M,L)}{s(M) \times s(L)} = \frac{\dfrac{\sum (M_i - m(M)) \times (L_i - m(L))}{n-1}}{s(M) \times s(L)}$$

Pode-se calcular a correlação linear na ausência do conhecimento das médias das duas variáveis. A equação acima, retrabalhada, origina:

$$r(M,L) = \frac{n.\sum ML - \sum M \times \sum L}{\sqrt{n\sum M^2 - \left(\sum M\right)^2} \times \sqrt{n\sum L^2 - \left(\sum L\right)^2}}$$

Que é a fórmula mais conhecida e utilizada para o cálculo do coeficiente de correlação linear simples.

Quadro 12.3 – Cálculo do coeficiente de correlação para o exemplo dado

Obs	M	L	ML
1	36	35	1.260
2	80	65	5.200
3	50	60	3.000
4	58	39	2.262
5	72	48	3.456
6	60	44	2.640
7	56	48	2.688
8	68	61	4.148
	ΣM = 480	ΣL = 400	
n=8	ΣM^2= 30.104	ΣL^2= 20.836	ΣML = 24.654
	$(\Sigma M)^2$= 230.400	$(\Sigma L)^2$= 160.000	

$$r(M,L) = \frac{n.\sum ML - \sum M \times \sum L}{\sqrt{n\sum M^2 - \left(\sum M\right)^2} \times \sqrt{n\sum L^2 - \left(\sum L\right)^2}}$$

$$r(M,L) = \frac{8 \times 24.654 - 480 \times 400}{\sqrt{8 \times 30.104 - 230.400} \times \sqrt{8 \times 20.836 - 160.000}} = 0{,}63$$

Considerações finais:

A existência de correlação entre duas variáveis aleatórias não implica em casualidade. Ou seja, não implica que a variação de uma provoca variação na outra. Para esta afirmativa é necessário variar os níveis de uma das variáveis (preditora), mantendo-se fixos todos os outros fatores, e observar o que ocorre com a variável de resposta.

O montante da variação em uma variável é explicada pela variação da outra pode ser medido elevando-se o coeficiente de correlação linear, r, ao quadrado: r^2.

As utilidades básicas da medida são:

- Análise exploratória
- Predição.

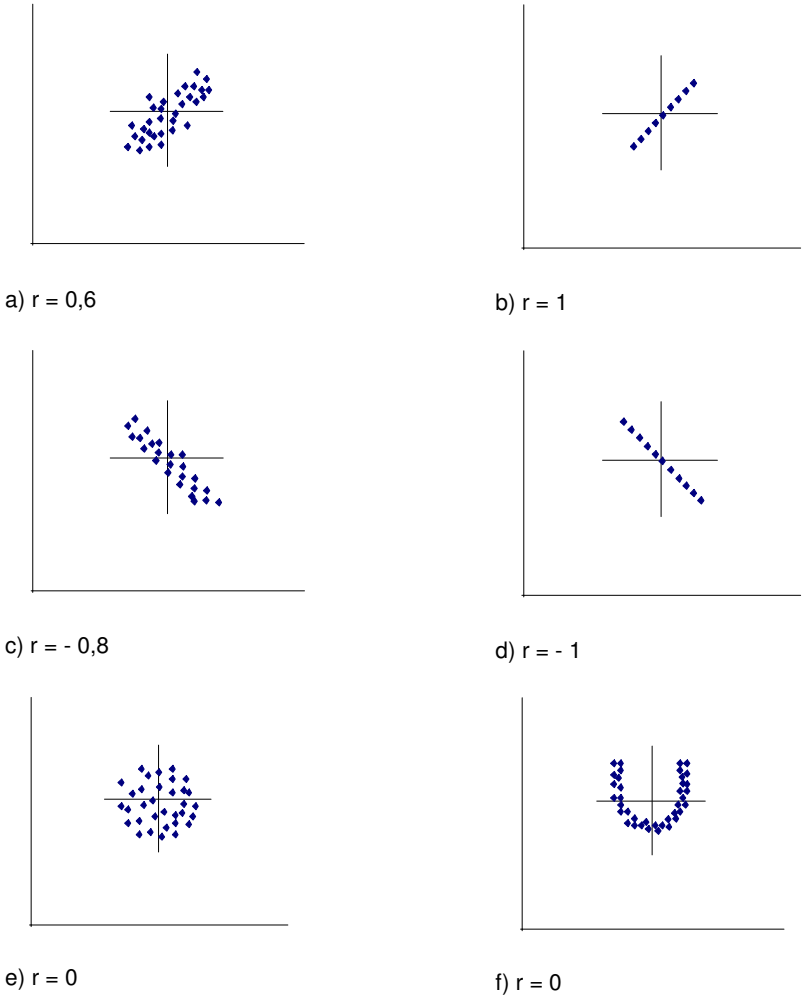


Figura 12.3 - Diagramas ilustrativos dos possíveis valores de r.

Observar que em f, muito embora seja possível identificar um tipo de associação entre as duas variáveis aleatórias, esta associação não é do tipo linear.

10. DISTRIBUIÇÃO NORMAL E NORMAL REDUZIDA

10.1. Introdução

A distribuição normal é a mais importante distribuição de densidade de probabilidade, sendo aplicada em inúmeros fenômenos e utilizada para o desenvolvimento teórico da estatística.

É também conhecida como distribuição de Gauss, Laplace ou Laplace-Gauss.

Seja Y uma variável aleatória contínua. Y terá distribuição normal se:

onde:

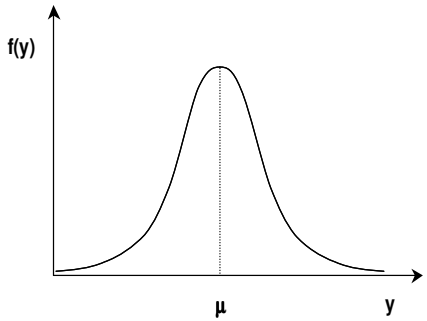
μ = média da população

σ = desvio-padrão da população

π = 3,1416 ...

e = base do logaritmo neperiano (2,718 ...)

$\frac{1}{\sigma\sqrt{2\pi}}$ = fator de escalonamento, faz com que a área sob a curva da função seja sempre igual a 1 (um)



10.2. Entendendo a distribuição

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2}, -\infty < y < \infty$$

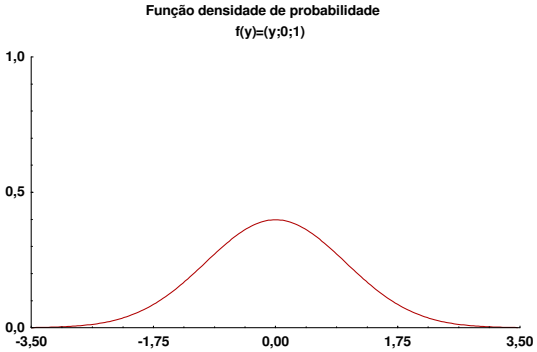
a. Alterações no valor da média:

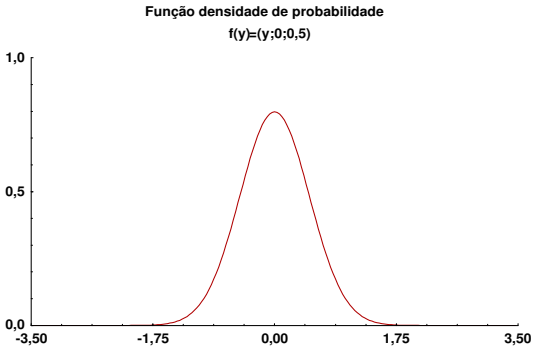
i. implicam no deslocamento do ponto de máximo ao longo do eixo Y, sem alterações na forma básica

b. Alterações no valor do desvio padrão:

i. Aumento: maior dispersão dos dados em torno da média

ii. Redução: menor dispersão dos dados em torno da média





10.3. Simplificando a distribuição para facilitar o uso

Para o cálculo das probabilidades utilizando a função densidade de probabilidades surgem dois problemas:

i. Relativo a integração de f(y), pois é necessário o desenvolvimento em séries, o que é um cálculo relativamente complexo.

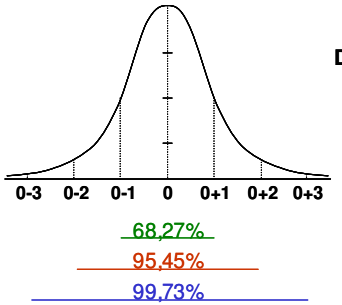
ii. Tabelar todas as probabilidades considerando-se as várias combinações possíveis de **μ** e **σ** acarretaria um grande trabalho, pois, f(y) depende dos parâmetros **μ** e **σ**.

Esses problemas foram solucionados por meio de uma mudança de variável, obtendo-se, assim, a distribuição normal padronizada ou reduzida (**μ** = 0 e **σ** = 1):

$$Z_i = \frac{Y_i - \mu}{\sigma}$$

A equação pode então ser rescrita:

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



Distribuição normal padrão
média = 0
variância = 1

$$\int_{-1}^{1} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz = 0{,}6827$$

$$\int_{-2}^{2} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz = 0{,}9545$$

$$\int_{-3}^{3} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz = 0{,}9973$$

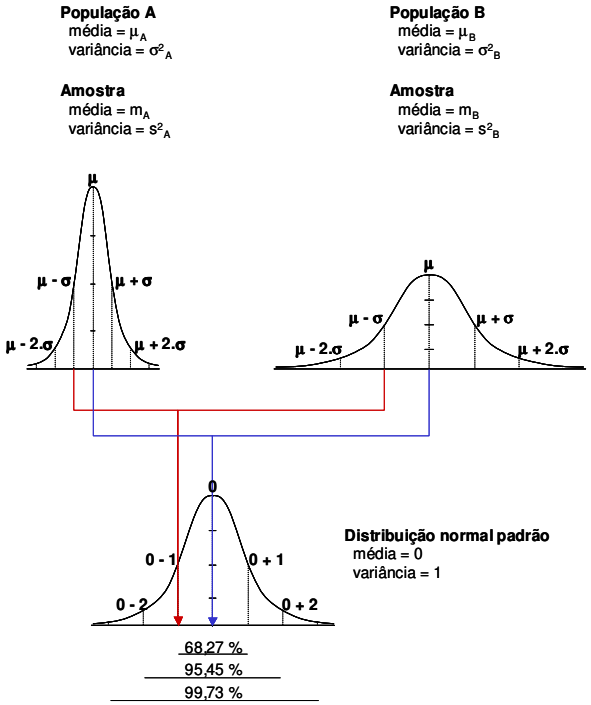
A distribuição apresenta as seguintes características:

- É simétrica em relação à média (μ = 0)
- f(z) possui um máximo para (z = 0), neste caso sua ordenada vale 0,39
- f(z) tende a 0 (zero) quando (z) tende para ± ∞
- A integral de f(z) (- ∞ < z < ∞) é igual a 1 (um)
- Tem dois pontos de inflexão cujas abscissas valem (- σ e + σ)
- 50% da população encontra-se entre - ∞ e 0
- 50% da população encontra-se entre 0 e + ∞
- 68,27% dos indivíduos da população encontram-se entre: - σ e + σ
- 95,45% dos indivíduos da população encontram-se entre: - 2σ e + 2σ
- 99,73% dos indivíduos da população encontram-se entre: - 3σ e + 3σ

10.4. Entendendo: distribuição normal vs. normal padrão

Observa-se que dois parâmetros estatísticos caracterizam uma população cuja variável em estudo possui distribuição normal: a média, μ, e o desvio padrão, σ.

O objetivo fundamental da padronização é facilitar os cálculos de probabilidade, uma vez que foram definidos a média (μ = 0) e o desvio padrão (σ = 1).



Desta forma é possível a utilização de uma tabela, contendo todas as integrais da função (distribuição normal padrão).

Assim, a partir de uma distribuição normal qualquer, pode-se convertê-la para a distribuição normal padrão, obter as informações necessárias sobre as probabilidades, e retornar a variável original.

10.5. Uso da tabela de distribuição normal padrão

Existem basicamente três tipos mais utilizados de tabelas que oferecem as áreas (probabilidades) sob a curva normal padrão:

$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z} e^{-\frac{1}{2}z^2}\,dz$$

$$\frac{1}{\sqrt{2\pi}}\int_{0}^{z} e^{-\frac{1}{2}z^2}\,dz$$

$$1-\left(\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z} e^{-\frac{1}{2}z^2}\,dz\right)$$ - Fornecida na última versão da apostila

Exemplo:

Desejam-se as probabilidades

a. $P(0 \le z \le 1)$

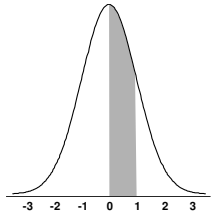


Tabela $\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z} e^{-\frac{1}{2}z^2}\,dz$:

$$P(0 \le z \le 1) = P(z \le 1) - P(z < 0)$$
$$P(0 \le z \le 1) = 0{,}8413 - 0{,}5000$$
$$P(0 \le z \le 1) = 0{,}3413 = 34{,}13\%$$

Tabela $\frac{1}{\sqrt{2\pi}}\int_{0}^{z} e^{-\frac{1}{2}z^2}\,dz$:

$$P(0 \le z \le 1) = 0{,}3413 = 34{,}13\%$$

Tabela $1-\left(\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z} e^{-\frac{1}{2}z^2}\,dz\right)$:

$$P(0 \le z \le 1) = 1 - [P(z > 1) + P(z < 0)]$$
$$P(0 \le z \le 1) = 1 - [P(z > 1) + P(z > 0)]$$
$$P(0 \le z \le 1) = 1 - (0{,}1587 + 0{,}5000)$$
$$P(0 \le z \le 1) = 0{,}3413 = 34{,}13\%$$

b. P(-2,55 < z < 1,2)

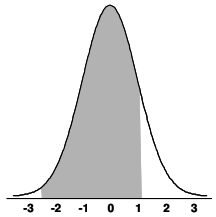


Tabela $\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z} e^{-\frac{1}{2}z^2}\,dz$:

$$P(-2{,}55 < z < 1{,}2) = P(z < 1{,}2) - (P(z \le -2{,}55)$$
$$P(-2{,}55 < z < 1{,}2) = P(z < 1{,}2) - [1 - P(z \le 2{,}55)]$$
$$P(-2{,}55 < z < 1{,}2) = 0{,}8849 - (1 - 0{,}9946)$$
$$P(-2{,}55 < z < 1{,}2) = 0{,}8849 - 0{,}0054$$
$$P(-2{,}55 < z < 1{,}2) = 0{,}8795 = 87{,}95\%$$

Tabela $\frac{1}{\sqrt{2\pi}}\int_{0}^{z} e^{-\frac{1}{2}z^2}\,dz$:

$$P(-2{,}55 < z < 1{,}2) = P(z < 1{,}2) + P(z < 2{,}55)$$
$$P(-2{,}55 < z < 1{,}2) = 0{,}3849 + 0{,}4946$$
$$P(-2{,}55 < z < 1{,}2) = 0{,}8795 = 87{,}95\%$$

Tabela $1 - \left(\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z} e^{-\frac{1}{2}z^2}\,dz\right)$:

$$P(-2{,}55 < z < 1{,}2) = 1 - [P(z \ge 1{,}2) + P(z \le -2{,}55)]$$
$$P(-2{,}55 < z < 1{,}2) = 1 - [P(z \ge 1{,}2) + P(z \ge 2{,}55)]$$
$$P(-2{,}55 < z < 1{,}2) = 1 - (0{,}1151 + 0{,}0054)$$
$$P(-2{,}55 < z < 1{,}2) = 1 - 0{,}1205$$
$$P(-2{,}55 < z < 1{,}2) = 0{,}8795 = 87{,}95\%$$

c. P(z ≥ 1,93)

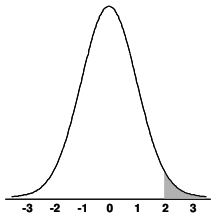


Tabela $\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z} e^{-\frac{1}{2}z^2}\,dz$:

$$P(z \ge 1{,}93) = 1 - P(z < 1{,}93)$$
$$P(z \ge 1{,}93) = 1 - 0{,}9732$$
$$P(z \ge 1{,}93) = 0{,}0268 = 2{,}68\%$$

Tabela $\frac{1}{\sqrt{2\pi}}\int_{0}^{z} e^{-\frac{1}{2}z^2}\,dz$:

$$P(z \ge 1{,}93) = 0{,}5000 - P(z < 1{,}93)$$
$$P(z \ge 1{,}93) = 0{,}5000 - 0{,}4732$$
$$P(z \ge 1{,}93) = 0{,}0268 = 2{,}68\%$$

Tabela $1 - \left(\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z} e^{-\frac{1}{2}z^2}\,dz\right)$:

$$P(z \ge 1{,}93) = 0{,}0268 = 2{,}68\%$$

10.6. Uso da transformação para resolução de probabilidades

Como foi visto, a utilização das tabelas da distribuição normal padronizada polpa tempo para a resolução de problemas envolvendo o cálculo de integrais.

Quando se trabalha com calculadoras científicas potentes, a utilização de tabelas torna-se desnecessária.

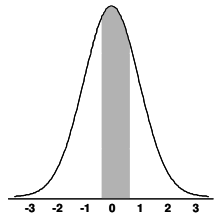
Vejamos um exemplo concreto de utilização da transformação de uma variável aleatória qualquer em uma variável aleatória padronizada para a resolução de problemas:

Exemplo:

As alturas dos alunos de elementos de estatística são normalmente distribuídas com média 1,60 m e desvio padrão 0,30 m. Quais as probabilidades de um aluno medir:

a. Entre 1,50 m e 1,80 m

$$Z_1 = \frac{y - \mu}{\sigma} = \frac{1{,}50 - 1{,}60}{0{,}30} = -0{,}33$$
$$Z_2 = \frac{y - \mu}{\sigma} = \frac{1{,}80 - 1{,}60}{0{,}30} = 0{,}67$$



i. Tabela $\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z} e^{-\frac{1}{2}z^2}\,dz$:

$$P(-0{,}33 \le z \le 0{,}67) = P(z \le 0{,}67) - P(z < -0{,}33)$$
$$P(-0{,}33 \le z \le 0{,}67) = P(z \le 0{,}67) - [1 - P(z < 0{,}33)]$$
$$P(-0{,}33 \le z \le 0{,}67) = 0{,}7486 - [1 - (0{,}6293)]$$
$$P(-0{,}33 \le z \le 0{,}67) = 0{,}7486 - 0{,}3707$$
$$P(-0{,}33 \le z \le 0{,}67) = 0{,}3779 = 37{,}79\%$$

ii. Tabela $\frac{1}{\sqrt{2\pi}}\int_{0}^{z} e^{-\frac{1}{2}z^2}\,dz$:

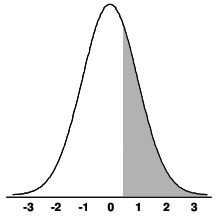
$$P(-0{,}33 \le z \le 0{,}67) = P(z \le 0{,}67) + P(z \le 0{,}33)$$
$$P(-0{,}33 \le z \le 0{,}67) = 0{,}2486 + 0{,}1293$$
$$P(-0{,}33 \le z \le 0{,}67) = 0{,}3779 = 37{,}79\%$$

iii. Tabela $1-\left(\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z} e^{-\frac{1}{2}z^2}dz\right)$:

$$P(-0{,}33 \le z \le 0{,}67) = 1-\left[P(z \ge 0{,}67) + P(z \le -0{,}33)\right]$$
$$P(-0{,}33 \le z \le 0{,}67) = 1-\left[P(z \ge 0{,}67) + P(z \ge 0{,}33)\right]$$
$$P(-0{,}33 \le z \le 0{,}67) = 1-(0{,}2514 + 0{,}3707)$$
$$P(-0{,}33 \le z \le 0{,}67) = 1-0{,}6221$$
$$P(-0{,}33 \le z \le 0{,}67) = 0{,}3779 = 37{,}79\%$$

b. Mais de 1,75 m

$$Z = \frac{y-\mu}{\sigma} = \frac{1{,}75-1{,}60}{0{,}30} = 0{,}50$$



i. Tabela $\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z} e^{-\frac{1}{2}z^2}dz$:

$$P(z > 0{,}50) = 1-P(z \le 0{,}50)$$
$$P(z > 0{,}50) = 1-0{,}6915$$
$$P(z > 0{,}50) = 0{,}3085 = 30{,}85\%$$

ii. Tabela $\frac{1}{\sqrt{2\pi}}\int_{0}^{z} e^{-\frac{1}{2}z^2}dz$:

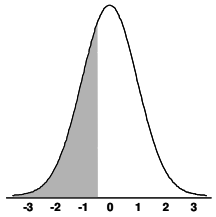
$$P(z > 0{,}50) = 0{,}5000-P(z \le 0{,}50)$$
$$P(z > 0{,}50) = 0{,}5000-0{,}1915$$
$$P(z > 0{,}50) = 0{,}3085 = 30{,}85\%$$

iii. Tabela $1-\left(\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z} e^{-\frac{1}{2}z^2}dz\right)$:

$$P(z > 0{,}50) = 0{,}3085 = 30{,}85\%$$

c. Menos que 1,48 m

$$Z = \frac{y-\mu}{\sigma} = \frac{1{,}48-1{,}60}{0{,}30} = -0{,}40$$



i. Tabela $\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z} e^{-\frac{1}{2}z^2}dz$:

$$P(z < -0{,}40) = 1-P(z \le 0{,}40)$$
$$P(z < -0{,}40) = 1-0{,}6554$$
$$P(z < -0{,}40) = 0{,}3446 = 34{,}46\%$$

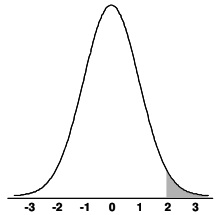
ii. Tabela $\frac{1}{\sqrt{2\pi}}\int_{0}^{z} e^{-\frac{1}{2}z^2}dz$:

$$P(z < -0{,}40) = 0{,}5000-P(z \le 0{,}40)$$
$$P(z < -0{,}40) = 0{,}5000-0{,}1554$$
$$P(z < -0{,}40) = 0{,}3446 = 34{,}46\%$$

iii. Tabela $1-\left(\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z} e^{-\frac{1}{2}z^2}dz\right)$:

$$P(z < -0{,}40) = P(z > 0{,}40)$$
$$P(z < -0{,}40) = 0{,}3446 = 34{,}46\%$$

d. Qual deve ser a medida mínima para escolher-se 10% dos mais altos?



i. Tabela $\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z} e^{-\frac{1}{2}z^2}dz$, valor de z que apresenta probabilidade ≅ 0,90 (= 0,8997).

z = 1,28 $\quad z = \frac{y-\mu}{\sigma} \quad \therefore \quad 1{,}28 = \frac{y-1{,}60}{0{,}30} \quad \therefore \quad y = 1{,}98m$

ii. Tabela $\frac{1}{\sqrt{2\pi}}\int_{0}^{z} e^{-\frac{1}{2}z^2}dz$, valor de z que apresenta probabilidade ≅ 0,40 (= 0,3997)

z = 1,28 $\quad z = \frac{y-\mu}{\sigma} \quad \therefore \quad 1{,}28 = \frac{y-1{,}60}{0{,}30} \quad \therefore \quad y = 1{,}98m$

iii. Tabela $1-\left(\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z} e^{-\frac{1}{2}z^2}dz\right)$, valor de z que apresenta probabilidade ≅ 0,10 (= 0,1003)

z = 1,28 $\quad z = \frac{y-\mu}{\sigma} \quad \therefore \quad 1{,}28 = \frac{y-1{,}60}{0{,}30} \quad \therefore \quad y = 1{,}98m$

11. DISTRIBUIÇÃO AMOSTRAL DA MÉDIA E TESTE DE HIPÓTESES

11.1. Teorema do limite central (ou central do limite)

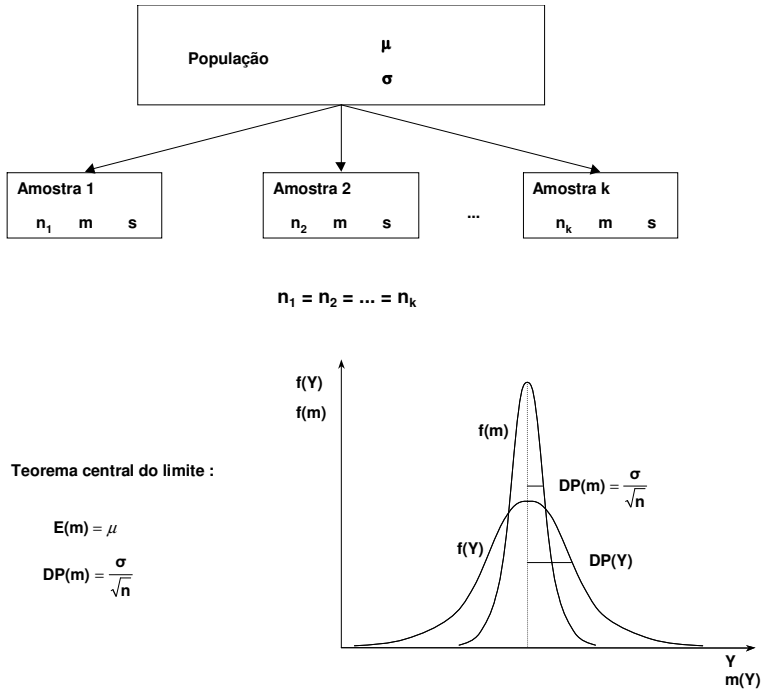


Figura 11.1 Ilustração do teorema central do limite

A estimativa da média, m, de uma variável aleatória é também uma variável aleatória.

A distribuição da estimativa da média, m, tende para a distribuição normal quando o tamanho da amostra, n, aumenta, independentemente do tipo da distribuição básica.

Enunciado do teorema central do limite: a medida em que aumenta o tamanho da amostra, n, a distribuição da estimativa da média, m, de uma amostra aleatória, extraída de praticamente qualquer população, tende para a distribuição normal com média μ e desvio padrão $\frac{\sigma}{\sqrt{n}}$:

O teorema é de grande aplicação prática na inferência pois e especifica completamente a distribuição de m em grandes amostras.

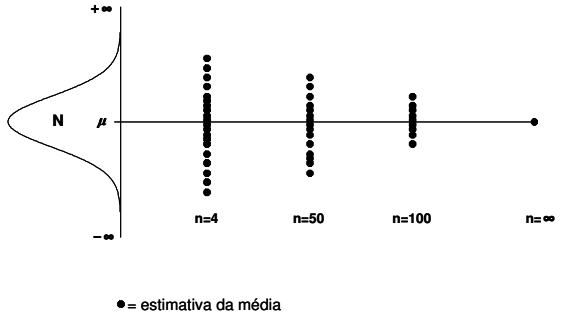


Figura 11.2 - Forma alternativa de compreender o teorema.

Exemplo:

Consideremos um processo de amostragem com n = 2 em uma urna que contém três tipos de fichas (2,4 e 6) na mesma quantidade:

a) Combinações possíveis:

	2	4	6
2	4	6	8
4	6	8	10
6	8	10	12

$\sum y$

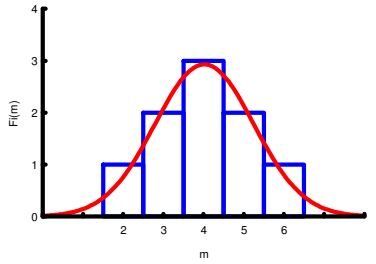
b) Médias possíveis:

	2	4	6
2	2	3	4
4	3	4	5
6	4	5	6

c) Frequência da média:

m	$F_i(m)$
2	1
3	2
4	3
5	2
6	1

d) Histograma:



Observa-se que a amostragem sucessiva em uma distribuição uniforme origina uma distribuição com tendência à normalidade já com n = 2, e mais próximo a normal à medida que n aumenta.

Demonstrações do teorema central do limite

Tendência central da estimativa da média:

$$E(m) = \mu$$

$$E(m) = E\left((y_1 + ... + y_n)\right)$$

$$E(m) = \frac{1}{n}\left[E(y_1 + ... + y_n)\right]$$

$$E(m) = \frac{1}{n}\left[E(y_1) + ... + E(y_n)\right]$$

$$E(m) = \frac{1}{n}\left(\mu + ... + \mu\right)$$

$$E(m) = \frac{1}{n} \cdot n\mu$$

$$E(m) = \mu$$

Dispersão da estimativa da média:

$$V(m) = \frac{\sigma^2}{n}$$

$$m = \frac{(y_1 + ... + y_n)}{n}$$

$$V(m) = V\left[\frac{(y_1 + ... + y_n)}{n}\right]$$

$$V(m) = \frac{1}{n^2}\left[V(y_1 + ... + y_n)\right] \qquad \therefore \qquad \textit{Admitindo independência}$$

$$V(m) = \frac{1}{n^2}\left[V(y_1) + ... + V(y_n)\right]$$

$$V(m) = \frac{1}{n^2}\left(\sigma^2 + ... + \sigma^2\right)$$

$$V(m) = \frac{1}{n^2} \cdot n\sigma^2$$

$$V(m) = \frac{\sigma^2}{n}$$

Exemplo:

Os funcionários da UESC ganham um salário mensal cuja média, μ, é de R$ 800,00, com desvio padrão, σ, de R$ 400,00. Cada estudante da UESC foi encarregado de tomar uma amostra de 40 funcionários e estimar o salário médio mensal, m.

Naturalmente, é de se esperar que cada estudante selecione uma amostra diferente, obtendo assim diferentes estimativas da média, m. Em torno de que valor a estimativa flutuará, E(m), e com que desvio padrão, DP(m)?

Observação: o número de estudantes é suficientemente grande para originar a distribuição de probabilidade da estimativa da média, m.

Solução:

$$E(m) = \mu = 800{,}00$$

$$DP(m) = \frac{\sigma}{\sqrt{n}} = \frac{400{,}00}{\sqrt{40}} = 63{,}25$$

Observações importantes:

A estimativa da média flutua pouco, por causa das compensações: uma amostra típica inclui tanto funcionários de salário alto como de salário baixo, o que contribui para as compensações. Quanto maior a tamanho da amostra, mais isso é evidente.

A flutuação da estimativa da média, V(m) ou DP(m), em relação à dispersão populacional, σ^2 ou σ, é um quociente (ou razão) do tamanho da amostra, n:

$$V(m) = \frac{\sigma^2}{n}$$

$$DP(m) = \frac{\sigma}{\sqrt{n}}$$

11.2. Teste de hipóteses

Basicamente a inferência estatística se dá por dois mecanismos básicos:

a. Intervalos de confiança (μ, σ^2, σ, π)
b. Testar hipósteses

No caso “a” busca-se cercar o parâmetro populacional desconhecido com base nos elementos amostrais.

No caso “b” formulam-se hipóteses quanto ao valor do parâmetro populacional, com base na observação dos elementos amostrais, um teste estatístico permitirá a decisão se a hipótese deve, ou não, ser rejeitada/aceita segundo uma determinada probabilidade de erro.

11.2.1. Hipótese

Trata-se de uma suposição sobre o valor de um parâmetro populacional ou quanto à natureza da distribuição de probabilidade de uma variável.

Exemplos:

A altura média da população brasileira é 1,65 m (μ = 1,65 m).

Peso dos alunos da UESC ~ N (μ, σ).

11.2.2. Teste de hipóteses

É uma regra de decisão para aceitar ou rejeitar uma hipótese estatística, com base nos elementos amostrais.

11.2.3. Tipos de hipóteses

H_0: hipótese da igualdade (ou conservadora)

H_1: hipótese alternativa

Exemplos:

H_0: μ = 1,65 m	H_0: μ = 1,65 m	H_0: μ = 1,65 m
H_1: $\mu \neq$ 1,65 m	H_1: μ > 1,65 m	H_1: μ < 1,65 m

11.2.4. Tipos de erros

São os erros associados às decisões do teste de hipóteses:

		Realidade	
		H_0 verdadeira	H_0 falsa
Decisão	Aceitar H_0	Decisão correta (1 - α)	Erro tipo II (β)
	Rejeitar H_0	Erro tipo I (α)	Decisão correta (1 - β)

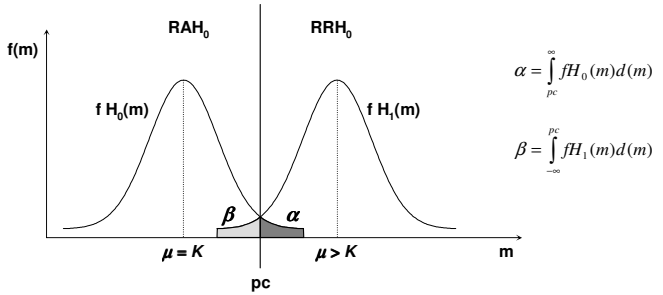
O erro tipo I só poderá ser cometido ao se rejeitar H_0, e o erro tipo II, quando aceitar H_0.

O tomador da decisão (pesquisador) deseja, obviamente, reduzir ao mínimo as probabilidades dos dois tipos de erro.

Infelizmente, esta é uma tarefa difícil, porque, para uma amostra de um determinado tamanho, a probabilidade de se incorrer em um erro tipo II aumenta à medida que diminui a probabilidade do erro tipo I, e vice-versa.

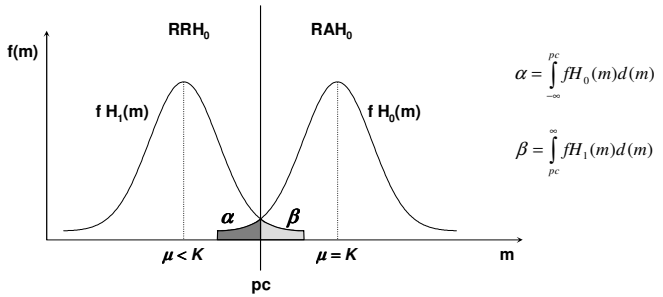
Teste unilateral à direita: H_0: μ = K

H_1: μ > K



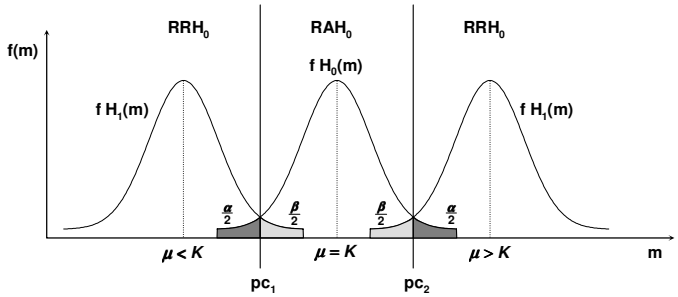
Teste unilateral à esquerda: H_0: μ = K

H_1: μ < K



Teste bilateral: H_0: μ = K

H_1: $\mu \neq$ K



Exemplo:

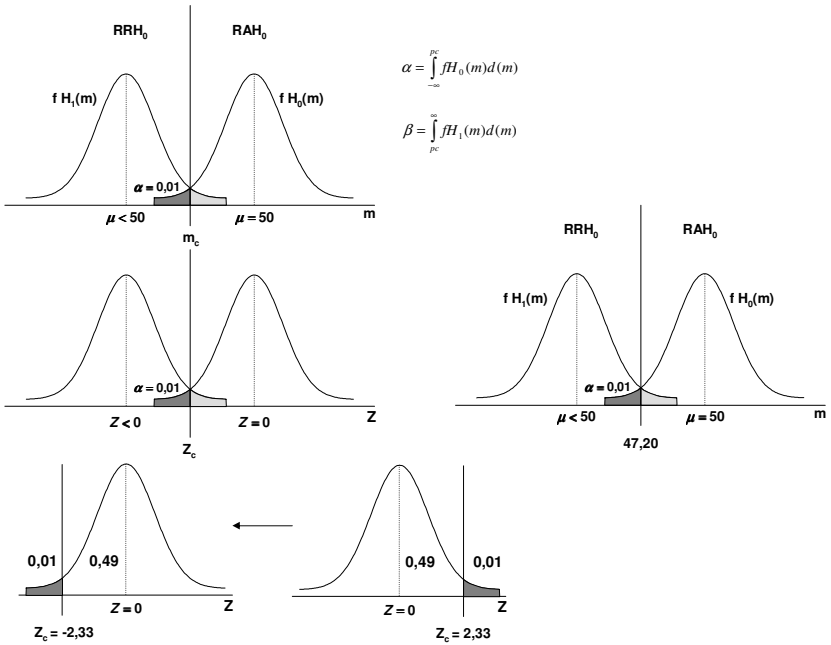
Para compreender o relacionamento dos erros e suas dimensões vamos idealizar um exemplo:

Um tratamento, A, quando aplicado em fêmeas de peixes de uma determinada espécie e peso, provoca ovulação para fecundação artificial em 50 dias, com variância de 36 dias.

Desejando-se reduzir este tempo, um novo tratamento, B, foi desenvolvido e testado em 25 fêmeas e essas apresentaram a desova em média com 48 dias.

Testar, com margem de segurança de erro de 1% se o novo tratamento reduziu o tempo de liberação da ovulação da espécie em questão:

Tratamento A	Tratamento B	Deseja-se testar:
$\mu = 50$ dias	m = 48 dias	H_0: $\mu = 50$ dias
$\sigma^2 = 36$ dias^2	n = 25 (tamanho da amostra)	H_1: $\mu < 50$ dias



$$\alpha = \int_{-\infty}^{pc} fH_0(m)d(m)$$

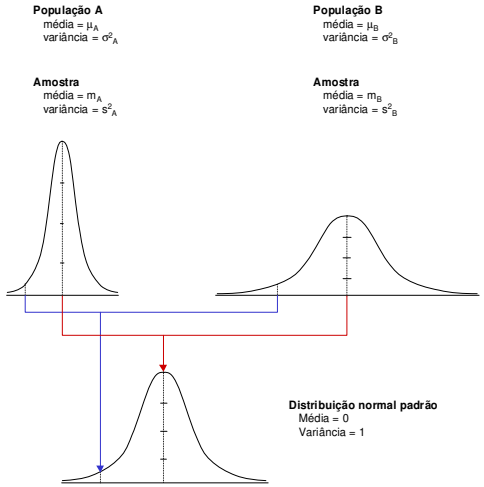
$$\beta = \int_{pc}^{\infty} fH_1(m)d(m)$$

$$Z = \frac{(Y-\mu)}{\sigma} = \frac{(Y-\mu)}{DP(Y)} \therefore Z = \frac{(m-\mu)}{DP(m)} \therefore -2{,}33 = \frac{((m_c-50))}{\frac{6}{\sqrt{25}}} \therefore m_c = 47{,}20$$

A decisão neste caso seria pela aceitação de H_0.

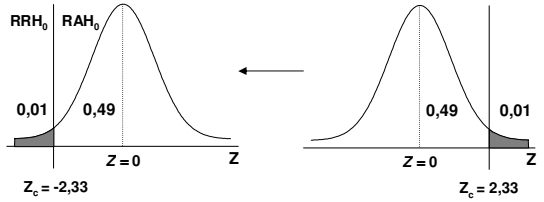
O que implica em afirmar com 99% de certeza, por conseguinte 1% de probabilidade de erro, que o novo tratamento não reduziu o tempo de liberação da ovulação da espécie em questão.

Para haver redução no tempo: m ≤ 47,20.



O objetivo deste mecanismo é uma simplificação dos cálculos utilizando as tabelas de valores associadas às probabilidades de Z.

Solução alternativa:



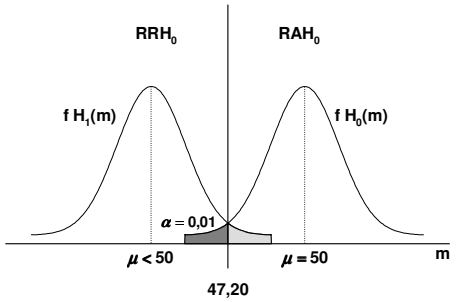
Z_c = -2,33 $\qquad Z_{cal} = \frac{(m-\mu)}{DP(m)} = \frac{(m-\mu)}{\frac{\sigma}{\sqrt{n}}} = \frac{(48-50)}{\frac{6}{\sqrt{25}}} = -1{,}67$

A lógica da decisão

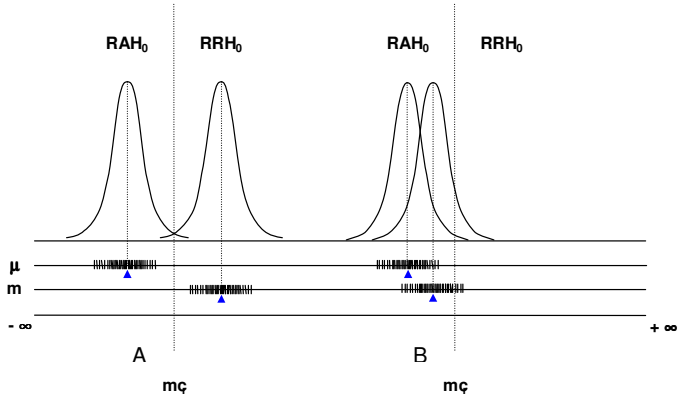
No exemplo dado fixou-se (α = 1%).

Portanto, tem-se 99% de probabilidade de estarmos corretos na decisão:

		Realidade	
		H_0 verdadeira	H_0 falsa
Decisão	Aceitar H_0	Decisão correta (1 - α)	Erro tipo II (β)
	Rejeitar H_0	Erro tipo I (α)	Decisão correta (1 - β)



Uma outra forma de compreender estes testes, com clareza, pode ser visualizada abaixo:



Na primeira situação, A, a média da população e da amostra encontram-se tão distantes que dificilmente poderiam ser consideradas como provenientes de uma mesma população: nestes casos a opção correta é pela rejeição de H_0.

Dada uma população com média μ e considerando uma amostra aleatória de tamanho n, com média m, tal que $m \in \mu$, situações como essas somente seriam possíveis nos casos em que, preponderantemente, os indivíduos da calda superior da população fossem os sorteados para comporem a amostra, o que, embora possível, é pouco provável, principalmente com o aumento de n.

Na segunda situação, B, a situação se inverte, ou seja, as médias da população e da amostra encontram-se tão próximas, que dificilmente poderiam ser concebidas como provenientes de populações distintas: nestes casos a opção correta é pela aceitação de H_0, pois sua rejeição somente seria possível com um erro tipo I muito elevado.

Extrapolar esta figura para os outros testes: unilateral à esquerda e bilateral.

12. DISTRIBUIÇÃO T DE STUDENT

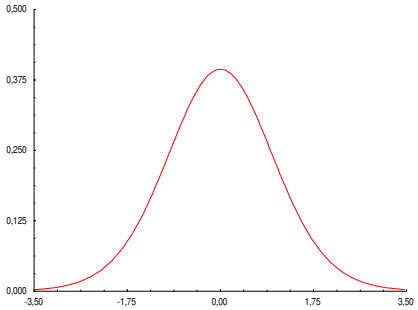
12.1. Introdução

Distribuição t de Student foi elaborada em 1908 por Gosset, sob o pseudônimo de Student, e demonstrada por Fisher em 1926:

$$f(t,\varphi) = c \cdot \left(1 + \frac{t^2}{\varphi}\right)^{-\frac{\varphi+1}{2}}$$

$$c = \frac{\Gamma\left(\frac{\varphi+1}{2}\right)}{\Gamma\left(\frac{\varphi}{2}\right)\sqrt{\pi\varphi}}$$

c é uma constante dependente de φ e determinada pela condição onde a área sob a curva de probabilidade é igual a um.



Trata-se de um modelo de distribuição contínua de densidade de probabilidade que se assemelha à distribuição normal padrão, N(O,1).

É utilizada para inferências estatísticas, particularmente, quando se tem amostras com tamanhos inferiores a 30 elementos.

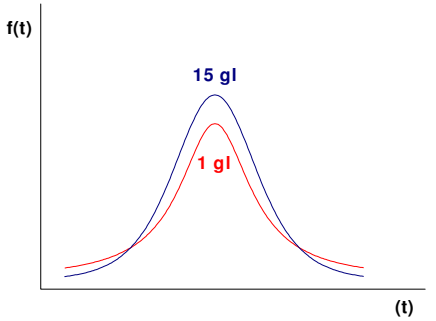
A distribuição possui um parâmetro denominado grau de liberdade φ. A média da distribuição é zero, e sua variância é dada por:

$$Var(t_\varphi) = \sigma^2(t_\varphi) = \frac{\varphi}{\varphi - 2} \quad (\varphi > 2)$$

onde:

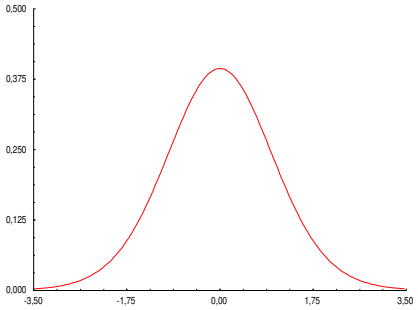
φ = grau de liberdade

Implicando que a variância Var(t_φ)vai se reduzido com o aumento de φ:



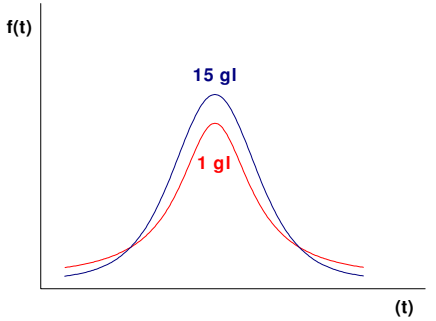
A distribuição é simétrica em relação à sua média.

Abaixo se encontra um exemplo da forma do gráfico da distribuição quando (φ = 20):



Para valores de (φ < 30) a distribuição "t" apresenta maior dispersão do que N(0,1), já que o desvio padrão, nestes casos, é maior do que 1, que é o desvio padrão da distribuição Normal Padrão. Por exemplo:

$$\sigma(t_4) = \sqrt{\frac{4}{4-2}} = 1{,}41 \quad \sigma(t_{35}) = \sqrt{\frac{35}{35-2}} = 1{,}03 \qquad \sigma(t_{60}) = \sqrt{\frac{60}{60-2}} = 1{,}02$$

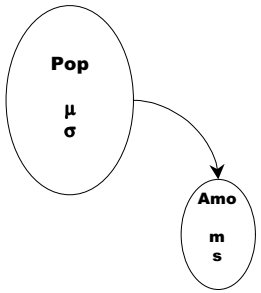


A distribuição "t" encontra-se tabelada para diferentes combinações de probabilidade e graus de liberdade.

Observações:

- Para se fazer inferências estatísticas sobre uma população, geralmente, são utilizadas as distribuições Normal Padrão e "t":
- Quando os valores da média e desvio padrão, μ e σ, são conhecidos, utiliza-se a distribuição normal padrão.
- Quando os valores da média e desvio padrão, μ e σ, não são conhecidos, e fazemos inferências sobre uma população a partir das estimativas da média e do desvio padrão, ou seja, obtidas nas amostras, utiliza-se a distribuição "t".
- Um exemplo clássico de uso desta distribuição é a estimativa do intervalo de confiança para a média populacional a partir de uma amostra representativa.

12.2. Aplicação: Intervalo de confiança para a média populacional (μ)



m: estimador do parâmetro μ

s: estimador do parâmetro σ

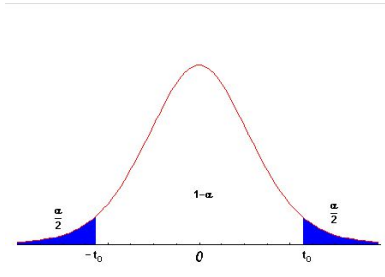
μ = m ± erro de amostragem

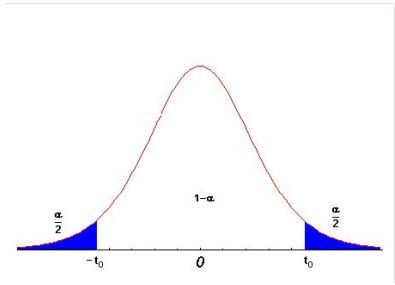
Seja Y uma variável aleatória proveniente de uma população normal, Y~N(μ, σ^2).

Seja uma amostra casual simples extraída desta população.

Sejam m e s, respectivamente, a média e o desvio padrão obtidos a partir da amostra.

A variável (t), $t = \frac{m-\mu}{s_m}$, tem distribuição de Student com n-1 graus de liberdade.





$$t = \frac{m-\mu}{s_m},$$

$$P(-t_0 \le t \le t_0) = 1-\alpha$$

$$P(-t_0 \le \frac{m-\mu}{s_m} \le t_0) = 1-\alpha \ \ (s_m)$$

$$P(-t_0 \cdot s_m \le m-\mu \le t_0 \cdot s_m) = 1-\alpha \ \ (-1)$$

$$P(t_0 \cdot s_m \ge -m+\mu \ge -t_0 \cdot s_m) = 1-\alpha \ \ (somando\ m)$$

$$P(m+t_0 \cdot s_m \ge \mu \ge m-t_0 \cdot s_m) = 1-\alpha$$

$$P(m-t_0 \cdot s_m \le \mu \le m+t_0 \cdot s_m) = 1-\alpha \qquad \therefore \qquad Como\ \ s_m = \frac{s}{\sqrt{n}}$$

$$P\left(m-t_0 \cdot \frac{s}{\sqrt{n}} \le \mu \le m+t_0 \cdot \frac{s}{\sqrt{n}}\right) = 1-\alpha$$

$$\boxed{I.C\,(\mu) = m \pm t_0 \cdot \frac{s}{\sqrt{n}}}$$

$$t_{\alpha\%} = f(gl)$$

Onde:

m = Média amostral

s = Desvio padrão amostral

n = Número de elementos da amostra

$t_{\alpha\%}$ = Valor tabelado em função de gl (graus de liberdade = n – 1)

Exemplo 1:

Um anestésico A foi desenvolvido e possui tempo de ação desconhecido quando aplicado em bovinos de determinado peso e idade. Desejando-se caracterizar o novo produto para que possa ser lançado no mercado, uma amostra de 20 animais, de determinado peso e idade, recebeu uma dose do produto em condições controladas. Os resultados encontrados são mostrados abaixo:

Quadro 12.1 – Tempo de duração do anestésico em minutos, UESC, BA - janeiro de 2001

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
50	55	51	53	58	62	64	54	55	58	59	60	61	61	63	64	57	55	53	52

$$m = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{\sum_{i=1}^{20} y_i}{20} = \frac{(50+55+...+52)}{20} = 57{,}25 \text{ min}$$

$$s^2 = \frac{\sum (Y_i - m)^2}{n-1} \; ou \; s^2 = \frac{\sum Y_i^2 - \frac{\left(\sum Y_i\right)^2}{n}}{n-1} = 19{,}36 \text{ min}^2$$

$$s = \sqrt{s^2} = \sqrt{19{,}36} = 4{,}40 \text{ min}$$

Situação a

Adotando uma probabilidade de erro de 0,01 = 1%

$$1-\alpha$$
$$1-0{,}01 = 0{,}99$$
$$t_{1\%,\,bilateral}\,(19\ gl) = 2{,}861$$

$$I.C\,(\mu) = m \pm t_0 \cdot \frac{s}{\sqrt{n}}$$

$$I.C\,(\mu) = 57{,}25 \pm 2{,}861 \cdot \frac{4{,}40}{\sqrt{20}}$$

$$\boxed{I.C\,(\mu) = 57{,}25 \pm 2{,}81 \text{ min}}$$

A probabilidade do intervalo obtido incluir a média da população é de 99%.

Situação b

Adotando uma probabilidade de erro de 0,05 = 5%

$$1-\alpha$$
$$1-0{,}05 = 0{,}95$$
$$t_{5\%,\,bilateral}\,(19\ gl) = 2{,}093$$

$$I.C\,(\mu) = m \pm t_0 \cdot \frac{s}{\sqrt{n}}$$

$$I.C\,(\mu) = 57{,}25 \pm 2{,}093 \cdot \frac{4{,}40}{\sqrt{20}}$$

$$\boxed{I.C\,(\mu) = 57{,}25 \pm 2{,}06 \text{ min}}$$

A probabilidade do intervalo obtido incluir a média da população é de 95%.

Exemplo 2:

Um anestésico B foi desenvolvido e possui tempo de ação desconhecido quando aplicado em bovinos de determinado peso e idade. Desejando-se caracterizar o novo produto para que possa ser lançado no mercado, uma amostra de 20 animais, de determinado peso e idade, recebeu uma dose do produto em condições controladas. Os resultados encontrados são mostrados abaixo:

Quadro 12.2 – Tempo de duração do anestésico em minutos, UESC, BA - janeiro de 2001

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
45	75	41	73	48	82	44	84	45	78	49	70	41	81	43	84	47	85	43	82

$$m = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{\sum_{i=1}^{20} y_i}{20} = \frac{(45+75+...+82)}{20} = 62{,}00 \text{ min}$$

$$s^2 = \frac{\sum (Y_i - m)^2}{n-1} \; ou \; s^2 = \frac{\sum Y_i^2 - \frac{\left(\sum Y_i\right)^2}{n}}{n-1} = 334{,}95 \text{ min}^2$$

$$s = \sqrt{s^2} = \sqrt{334{,}95} = 18{,}30 \text{ min}$$

Observar que o desvio padrão dos dados do Quadro 12.2 (18,30 min) é substancialmente maior que o do Quadro 12.1 (4,40 min), indicando uma maior dispersão dos dados em torno da média.

Situação c

Adotando uma probabilidade de erro de 0,01 = 1%

$$1-\alpha$$
$$1-0{,}01=0{,}99$$
$$t_{1\%,bilateral}\ (19\ gl)=2{,}861$$

$$I.C\,(\mu)=m\pm t_0\cdot\frac{s}{\sqrt{n}}$$

$$I.C\,(\mu)=62{,}00\pm 2{,}861\cdot\frac{18{,}30}{\sqrt{20}}$$

$$\boxed{I.C\,(\mu)=62{,}00\pm 11{,}71\ \text{min}}$$

A probabilidade do intervalo obtido incluir a média da população é de 99%.

Situação d

Adotando uma probabilidade de erro de 0,05 = 5%

$$1-\alpha$$
$$1-0{,}05=0{,}95$$
$$t_{5\%,bilateral}\ (19\ gl)=2{,}093$$

$$I.C\,(\mu)=m\pm t_0\cdot\frac{s}{\sqrt{n}}$$

$$I.C\,(\mu)=62{,}00\pm 2{,}093\cdot\frac{18{,}30}{\sqrt{20}}$$

$$\boxed{I.C\,(\mu)=62{,}00\pm 8{,}56\ \text{min}}$$

A probabilidade do intervalo obtido incluir a média da população é de 95%.

Tabela 12.3 – Comparativo entre as situações

Situação	Amostra (n)	Probabilidade de acerto	Intervalo de confiança
a	20	99%	57,25 ± 02,81 min
b	20	95%	57,25 ± 02,06 min
c	20	99%	62,00 ± 11,71 min
d	20	95%	62,00 ± 08,56 min

Observa-se com clareza o mecanismo de proteção oferecido pela estatística inferencial à tomada de decisão:

Quando o pesquisador solicita uma maior certeza (passa de 95% para 99%) a estatística amplia o intervalo de confiança:

- Comparar as situações b com a e d com c.

Quando a variável aleatória em questão possui elevada dispersão em torno da média, elevados valores da variância e por conseguinte do desvio padrão, a estatística amplia o intervalo de confiança:

- Comparar as situações (a,b) com (c,d).

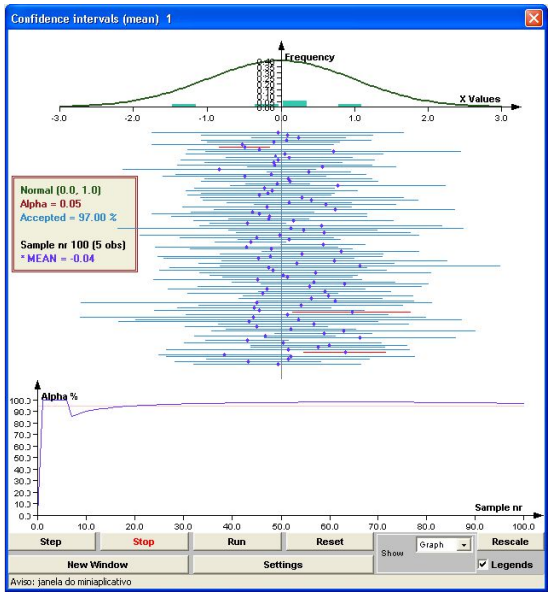
Em síntese, qualquer afirmação estatística sobre uma população, retirada a partir da observação de dados amostrais, envolve sempre alguma incerteza, a quantificação desta incerteza é o grande objetivo da estatística inferencial.

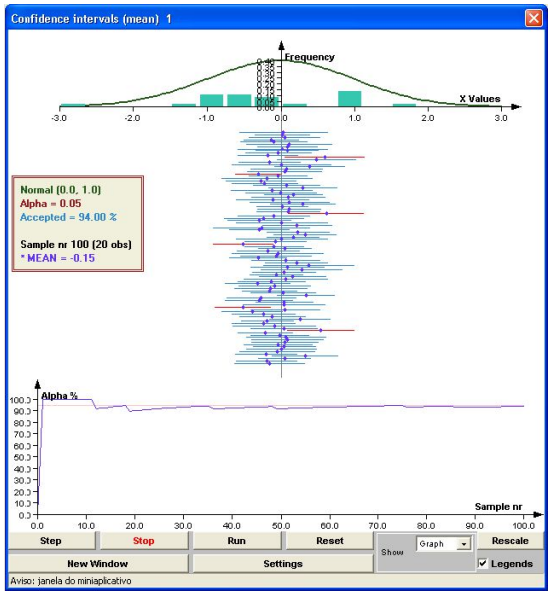
12.3. Exemplos de Intervalos de confiança para a média populacional

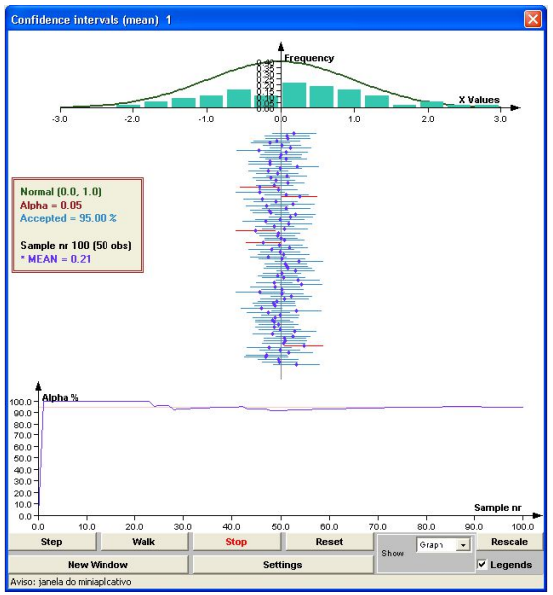
Os intervalos de confiança abaixo foram estimados em um laboratório virtual de estatística (http://www.kuleuven.ac.be/ucs/java/) a partir de uma população Y~N(0, 1)

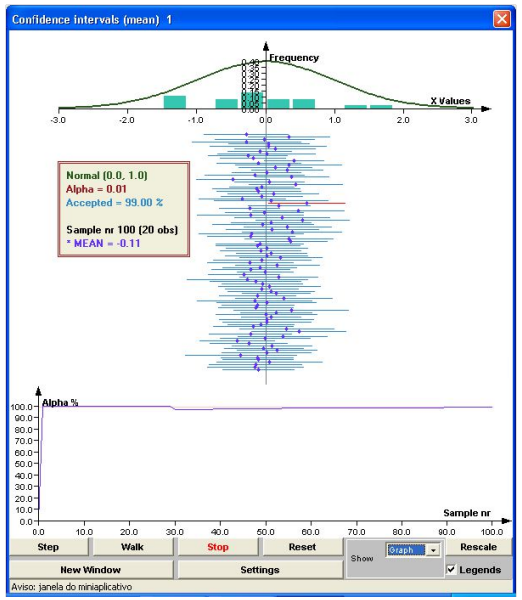
Variou-se a probabilidade de erro, o tamanho da amostra tendo-se solicitado 100 repetições em cada caso.

Recomenda-se que sejam realizadas estas experiências virtuais no laboratório indicado.


Confidence intervals (mean) 1
Frequency
X Values
Normal (0.0, 1.0)
Alpha = 0.05
Accepted = 97.00 %
Sample nr 100 (5 obs)
* MEAN = -0.04
Alpha %
Sample nr
Step
Stop
Run
Reset
Show
Graph
Rescale
New Window
Settings
Legends
Aviso: janela do miniaplicativo


Confidence intervals (mean) 1
Frequency
X Values
Normal (0.0, 1.0)
Alpha = 0.05
Accepted = 94.00 %
Sample nr 100 (20 obs)
* MEAN = -0.15
Alpha %
Sample nr
Step
Stop
Run
Reset
Show
Graph
Rescale
New Window
Settings
Legends
Aviso: janela do miniaplicativo


Confidence intervals (mean) 1
Frequency
X Values
Normal (0.0, 1.0)
Alpha = 0.05
Accepted = 95.00 %
Sample nr 100 (50 obs)
* MEAN = 0.21
Alpha %
Sample nr
Step
Walk
Stop
Reset
Show
Graph
Rescale
New Window
Settings
Legends
Aviso: janela do miniaplicativo


Confidence intervals (mean) 1
Frequency
X Values
Normal (0.0, 1.0)
Alpha = 0.01
Accepted = 99.00 %
Sample nr 100 (20 obs)
* MEAN = -0.11
Alpha %
Sample nr
Step
Walk
Stop
Reset
Show
Graph
Rescale
New Window
Settings
Legends
Aviso: janela do miniaplicativo

13. DISTRIBUIÇÃO χ^2

13.1. Introdução

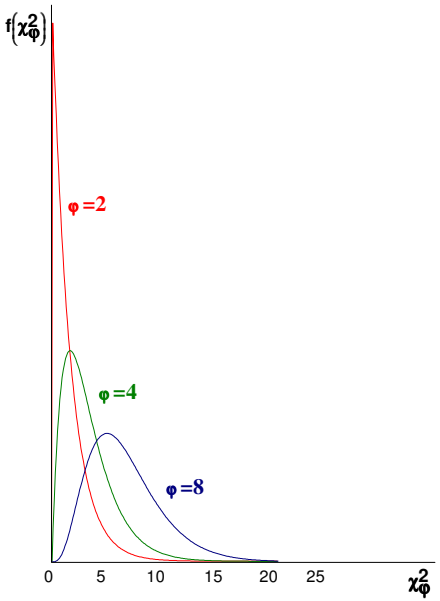
A distribuição qui-quadrado é um modelo de distribuição contínua importante para a teoria da inferência estatística:

$$f(\chi^2, \varphi) = c \cdot \chi^{2^{\left(\frac{n-2}{2}\right)}} \cdot e^{-\frac{\chi^2}{2}}$$

$$c = \frac{1}{\Gamma\left(\frac{\varphi}{2}\right) \cdot 2^{\frac{\pi}{2}}}$$

c é uma constante dependente de φ e determinada pela condição em que a área sob a curva de probabilidade é igual a um.

φ (lê-se fi) é um parâmetro da função densidade denominado grau de liberdade.



Uma das maneiras comumente encontrada na literatura para definir a distribuição χ^2 é fornecida a seguir:

Seja Y_1, Y_2, ..., Y_p variáveis aleatórias independentes, normalmente distribuídas, com média zero e variância 1. Define-se variável aleatória com distribuição qui-quadrado, como:

$$\chi_p^2 = Y_1^2 + Y_2^2 + ... + Y_p^2$$

Pode-se demonstrar que a média de uma distribuição qui-quadrado é igual ao grau de liberdade, e que a variância é igual ao dobro do número de graus de liberdade. Assim:

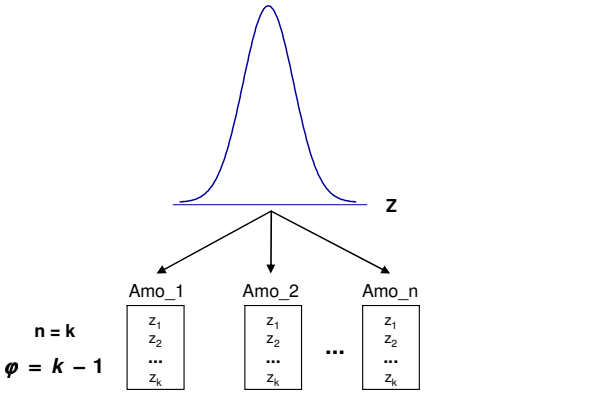
$$E\left(\chi_\varphi^2\right) = \mu\left(\chi_\varphi^2\right) = \varphi$$

$$Var\left(\chi_\varphi^2\right) = \sigma^2\left(\chi_\varphi^2\right) = 2\varphi$$

13.2. Entendendo a distribuição χ^2

A definição apresentada, embora útil sob alguns aspectos, não facilita a compreensão do significado desta distribuição. Assim, vamos conceituá-la de uma forma mais compreensível:

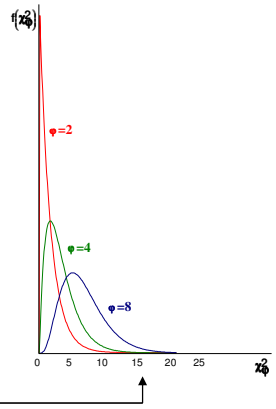
A distribuição χ^2 resulta da seleção aleatória dos desvios reduzidos $z_i = \frac{y_i - \mu}{\sigma}$ da distribuição Z, elevados ao quadrado.



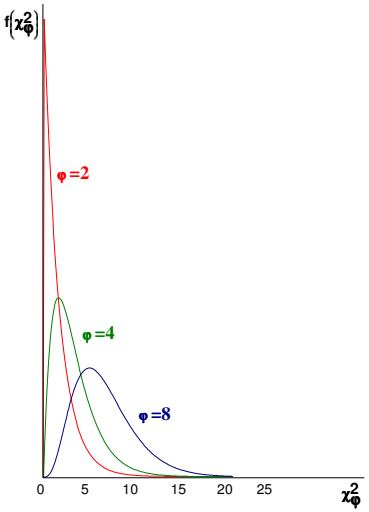
$$\chi^2_\varphi = z_1^2 + z_2^2 + \ldots + z_k^2$$

$$f(\chi^2,\varphi) = c \cdot \chi^{2\left(\frac{n-2}{2}\right)} \cdot e^{-\frac{\chi^2}{2}}$$

$$c = \frac{1}{\Gamma\left(\frac{\varphi}{2}\right) \cdot 2^{\frac{\pi}{2}}}$$



Observa-se que contrariamente às distribuições Normal e t, a χ^2 é assimétrica e sempre positiva (em função dos desvios serem elevados ao quadrado), com seus valores variando de 0 a + ∞.



Utilizando um mecanismo de cálculo, ou uma tabela da distribuição, pode-se observar que para φ = 1 (portanto n = 2):

- 68,27% dos casos estarão entre os valores de $\chi^2 = 0$ e $\chi^2 = 1$
- 95,45% dos casos estarão entre os valores de $\chi^2 = 0$ e $\chi^2 = 4$
- 99,73% dos casos estarão entre os valores de $\chi^2 = 0$ e $\chi^2 = 9$

Observa-se que a medida em que aumenta o número de grau de liberdade, φ, a forma da distribuição se altera, diminuindo a freqüência das observações próximas a 0 e 1, estendendo-se para valores maiores.

A forma da distribuição se altera bastante para o intervalo entre (φ = 1) e

(φ = 30), com intensidade decrescente à medida que φ se aproxima de 30.

A partir deste valor, já com uma conformação mais próxima à simetria e similar a distribuição normal, as alterações da forma são mínimas para pequenos acréscimos em φ.

13.3. Exemplos de aplicação da distribuição do χ^2

Seja ε um experimento aleatório. Sejam E_1, E_2, ... E_k K eventos associados a ε. Admitindo que o experimento é realizado n vezes:

Sejam: Fo_1, Fo_2, ... Fo_k as freqüências observadas dos K eventos.

Sejam: Fe_1, Fe_2, ... Fe_k as freqüências esperadas dos K eventos.

Como $\frac{(Fo_i - Fe_i)}{Fe_i}$ é um desvio padronizado, $\sum_{i=1}^{k}\frac{(Fo_i - Fe_i)^2}{Fe_i}$ apresenta distribuição χ^2.

Pode-se então usar a distribuição χ^2, associada a um teste de hipóteses, para se decidir se as discrepâncias $(Fo_i - Fe_i)$ são devidas ao acaso, ou seja, apresentam a mesma magnitude da variação observada em uma distribuição normal, ou se são maiores que essas, e portanto associadas a outros fatores, que não as flutuações normais da amostra.

13.4. Teste qui-quadrado

Também conhecido como teste de adequação do ajustamento ou aderência.

Procedimentos:

a. Enunciar as hipóteses estatísticas H_0 e H_1:

   H_0: Não existe discrepância entre as freqüências observadas e esperadas.

   H_1: existe discrepância entre as freqüências observadas e esperadas.

b. Fixar α e escolher a variável qui-quadrado com $\varphi = (k-1)$, onde k é o número de eventos.

c. Com o auxílio de uma tabela de χ^2 determinar o valor crítico entre as regiões de aceitação e rejeição de H_0.

d. Calcular o valor da variável $\chi^2_{cal} = \sum_{i=1}^{k}\frac{(Fo_i - Fe_i)^2}{Fe_i}$

   Se $\chi^2_{cal} < \chi^2_{tab}$ → Aceitar H_0

   Se $\chi^2_{cal} \geq \chi^2_{tab}$ → Rejeitar H_0

Exemplo:

Deseja-se testar se o número de acidentes numa rodovia se distribui igualmente pelos dias da semana. Para tanto foram levantados os seguintes dados:

Quadro 13.1 – Acidentes na rodovia X, Local, Estado - janeiro de 2001

Dia da semana	Dom	Seg	Ter	Qua	Qui	Sex	Sab	Total
Número de acidentes	33	26	21	22	17	20	36	175

$$Fe = \frac{1}{7}\cdot 175 = 25$$

Procedimentos

Adotar α = 5% e escolher uma variável qui-quadrado com $\varphi = (k - 1) = 7 - 1 = 6$

$$\chi^2_{tab(5\%,6\,gl)} = 12{,}59$$

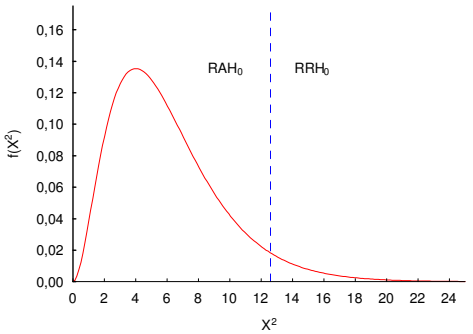
H_0: As freqüências são iguais em todos os dias da semana

H_1: As freqüências não são iguais em todos os dias da semana

Quadro 13.1 – Freqüências observadas e esperadas do número de acidentes na rodovia X, Local, Estado - janeiro de 2001

Dia da semana	Dom	Seg	Ter	Qua	Qui	Sex	Sab
Fo	33	26	21	22	17	20	36
Fe	25	25	25	25	25	25	25

$$\chi^2_{cal} = \sum_{i=1}^{k}\frac{(Fo_i - Fe_i)^2}{Fe_i} = \frac{(33-25)^2}{25} + \ldots + \frac{(36-25)^2}{25} = 12{,}0$$



Conclui-se pela aceitação de H_0, significando que não existe discrepância entre as freqüências observadas ou esperadas, ou ainda, que as freqüências dos acidentes são iguais em todos os dias da semana.

Nestas condições, têm-se 5% de probabilidade de estar errado e 95% de probabilidade de estar certo na decisão.

14. DISTRIBUIÇÃO F DE SNEDECOR

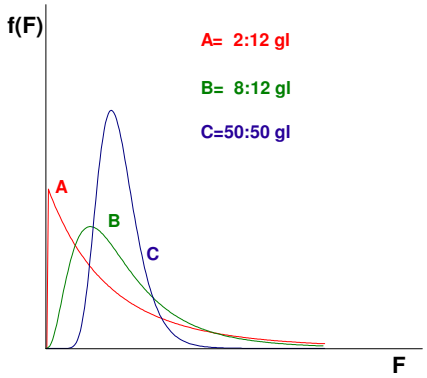
14.1. Introdução

A distribuição F de probabilidade foi reduzida por Snedecor sendo sua denominação uma homenagem a Fisher:

$$f(F,\varphi_1:\varphi_2) = c \cdot \left(\frac{\varphi_1}{\varphi_2}\right)^{\frac{\varphi_1}{2}} \cdot F^{\left(\frac{\varphi_1}{2}-1\right)} \cdot \left(1+\frac{\varphi_1}{\varphi_2}\cdot F\right)^{-\left(\frac{\varphi_1+\varphi_2}{2}\right)}$$

$$c = \frac{\Gamma\left(\frac{\varphi_1+\varphi_2}{2}\right)}{\Gamma\left(\frac{\varphi_1}{2}\right)\cdot\Gamma\left(\frac{\varphi_2}{2}\right)}$$

c é uma constante dependente de φ e determinada pela condição onde a área sob a curva de probabilidade é igual a um.



Entre as distribuições contínuas de probabilidades é uma das mais utilizadas para inferências estatísticas em experimentação.

Na análise de variância de experimentos esta distribuição é intensamente utilizada para a tomada de decisão nos testes de hipóteses (inferências sobre as populações).

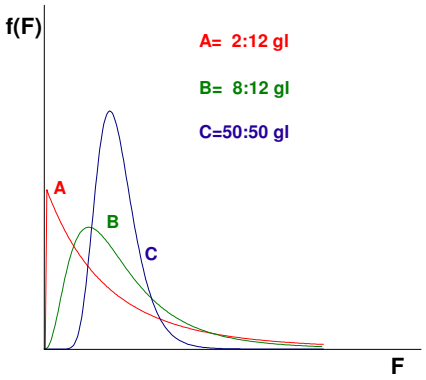
A definição mais comumente encontrada é que: a distribuição F é a razão entre duas variáveis aleatórias independentes com distribuição qui-quadrado.

Assim, uma distribuição F com φ_1 graus de liberdade no numerador, e φ_2 graus de liberdade no denominador é expressa por:

$$F(\varphi_1,\varphi_2) = \frac{\frac{\chi^2_{\varphi_1}}{\varphi_1}}{\frac{\chi^2_{\varphi_2}}{\varphi_2}}$$

Possuindo dois parâmetros: graus de liberdade do numerador e grau de liberdade no denominador, que são denominados, comumente, por φ_1 e φ_2 respectivamente, ela encontra-se tabelada para as probabilidades mais utilizadas nos testes de hipóteses: 1%, 5% e 10%.

Tal como a distribuição χ^2, esta distribuição de probabilidades não apresenta uma forma fixa, mas sim variável de acordo com os graus de liberdade envolvidos:



Em geral, utiliza-se a distribuição F para se tomar decisões sobre as populações a partir das estimativas das variâncias (obtidas das amostras) quando se testa hipóteses (inferências sobre as populações).

As hipóteses são as mais diversas, porém, em geral, esta distribuição é utilizada para se decidir se os dados podem ser considerados como advindos, ou não, de uma mesma população básica.

14.2. Entendendo a distribuição F

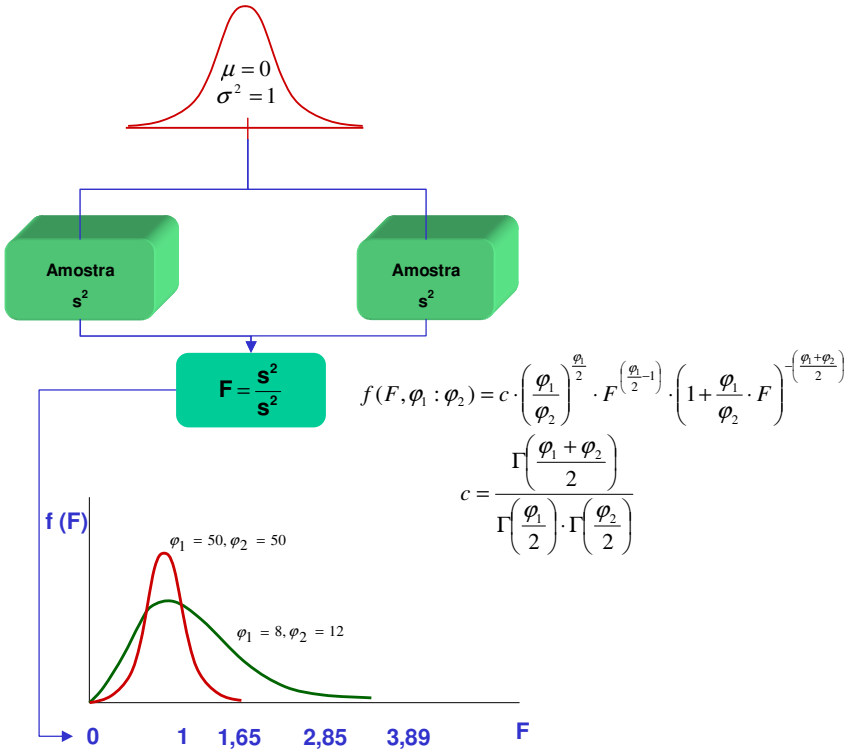


Figura 14.1 – Origem da distribuição F.

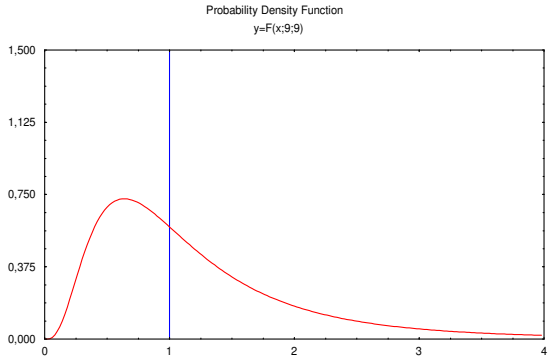
A distribuição F nos informa como se distribui a razão entre duas estimativas da variância de uma variável aleatória que apresenta distribuição normal padronizada.

Se estivermos retirando repetidamente amostras de um determinado tamanho fixo sob uma distribuição normal, calculando as estimativas da variância e calculando a relação:

$$F_{cal} = \frac{s_A^2}{s_B^2}$$

esperaríamos que a probabilidade do valor F_{cal} estar compreendido entre 0 e 1, ou seja, $0 \le F_{cal} \le 1$, seria 0,5 ou 50%.

Vejamos um exemplo concreto, feito via computação, utilizando o tamanho das amostras igual a 10, o que implica em 9 graus de liberdade:



$$\int_0^1 fF\,dF = 0{,}50 = 50\%$$

Da mesma forma:

$$\int_1^\infty fF\,dF = 0{,}50 = 50\%$$

nos fornece a probailidade da relação, ou seja F_{cal} ser maior que 1.

Podemos fixar qualquer valor, F_{Val}, de F nos eixos da abscissas e determinar a probabilidade de F_{cal} assumir valores entre zero e F_{Val}, integrando a função f(F) de zero até o valor desejado (Val).

Portanto, utilizando a distribuição F podemos comparar duas variâncias advindas de amostras de qualquer tamanho, e obter as respectivas distribuições de probabilidades.O que irá permitir a decisão se as variâncias amostrais podem ser, ou não, consideradas como advindas de uma mesma população básica:

Como já citado, sua utilização mais comum na análise de experimentos (análise de variância - ANOVA) é o teste de hipótese, a partir de duas estimativas das variâncias, para se decidir se os dados (variável aleatória) podem ser considerados, ou não, como advindos de uma mesma população básica.

Como vimos, a distribuição F é uma distribuição de probabilidades complexa. A compreensão de seu significado demanda tempo, reflexão e uso para seu completo entendimento. Contudo, seu uso na análise de experimentos é simples.

14.3. Precisão versus exatidão

Exatidão refere-se ao grau de aproximação do real, do objetivo ou do alvo.

Exatidão ⇨ fidelidade ao real ou certo

Precisão refere-se ao grau de repetibilidade na aproximação do real, ou a proximidade de cada observação de sua própria média.

Precisão ⇨ repetibilidade

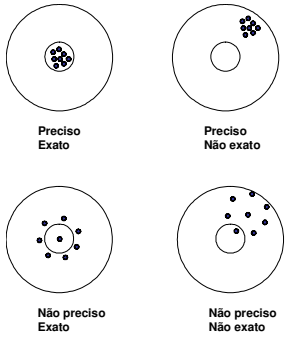


Figura 14.2 – Ilustração do conceito de precisão e exatidão.

Os métodos analíticos "exatos" e precisos são os métodos padrões. Em geral esses métodos são trabalhosos e caros. Assim, em muitas situações eles são substituídos por métodos alternativos, mais rápidos e baratos, cuja principal característica desejável é a elevada repetibilidade ou precisão, uma vez que a inexatidão (distanciamento do real), inerente ao método, pode ser corrigida por um fator de correção obtido entre o método padrão e o alternativo.

14.4. Exemplo de aplicação da distribuição F

Dois métodos de determinação da CTC do solo são usados em uma amostra de controle e fornecem os resultados da Tabela 14.1.

Tabela 14.1 – Resultados da determinação da capacidade de troca catiônica ($cmol_c/kg$) de dois métodos, UESC, BA – março 2009

	r_1	r_2	r_3	r_4	r_5	r_6	r_7	r_8	r_9	r_{10}	n	gl	m	$\mathbf{s^2}$	s
A	10,2	8,7	9,5	12,0	9,0	11,2	12,5	10,9	8,9	10,6	10	9	10,35	**1,76**	1,33
B	9,9	9,2	10,4	10,5	11,0	11,3	9,6	9,4	10,0	10,4	10	9	10,17	**0,46**	0,68

A questão a ser investigada é se é possível, ou não, considerar as precisões dos dois métodos (população de resultados gerados por cada método) estatisticamente iguais, ou seja:

$$H_0 : \sigma_A^2 = \sigma_B^2$$
$$H_1 : \sigma_A^2 > \sigma_B^2$$

Caso de decida que os métodos apresentam igual precisão, $\sigma_A^2 = \sigma_B^2$, as diferenças entre os resultados obtidos serão atribuídas às flutuações estatísticas naturais e, neste caso, os métodos seriam similares e poderiam ser usados indiscriminadamente.

A estatística F pode ser usada para esta decisão.

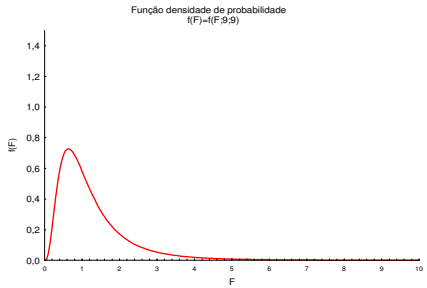
O teste faz uso da razão entre duas estimativas da variância, e como o teste é unilateral à direita, $\sigma_A^2 > \sigma_B^2$, o maior valor ocupa o numerador:

$$F_{cal} = \frac{s_A^2}{s_B^2} \; sendo \; s_A^2 \geq s_B^2$$

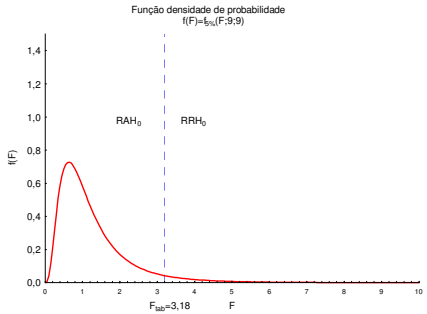
Esta decisão deve ser tomada adotando-se uma probabilidade de erro na decisão. Pode-se estabelecer, por exemplo, um erro máximo aceitável de 5%.

Mecanismo de decisão:

- Escolher a função densidade de probabilidades de F que apresente os graus de liberdade adequados (9:9).



- O valor crítico, F5%(9;9), pode ser obtido na tabela de F a 5% na interseção de 9 gl (numerador) na primeira linha com 9 gl (denominador) na primeira coluna.

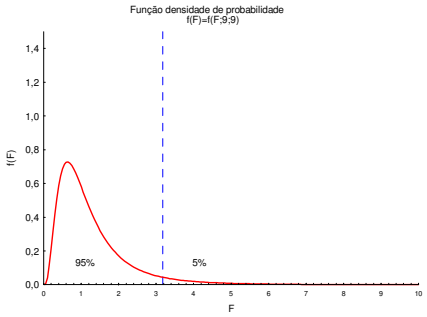


- Considerar os resultados de cada um dos dois métodos como amostras (10 para cada método) aleatoriamente retiradas de uma mesma população normalmente distribuída.

- Calcular o valor de prova (F_{cal}):

$$F_{cal} = \frac{s_A^2}{s_B^2} = 3{,}83$$

- Caso se trate realmente de uma mesma população, o que implica em similaridade dos métodos, em 95% dos casos em que uma amostragem aleatória fosse realizada e o valor F_{cal} determinado ele seria igual ou estaria situado à esquerda da linha pontilhada.

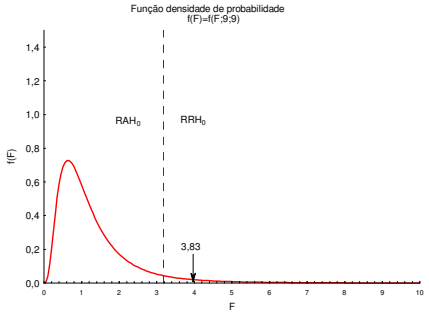


$$\int_0^{3,18} f(F)\,dF = 0{,}95 = 95\%$$

- Nas mesmas condições anteriores (mesma população), em apenas 5% dos casos o valor F_{cal} assumiria valores iguais ou superiores a 3,18:

$$1 - \int_0^{3,18} f(F)\,dF = 1 - 0{,}95 = 0{,}05 = 5\%$$

- Estes casos constituem o possível erro se decidirmos que os dados (resultados analíticos dos dois métodos) não podem ser considerados como provenientes de uma mesma população.



- Portanto, como o valor de prova (F_{cal}=3,83), e admitindo uma probabilidade de 5% de erro, deve-se decidir que os resultados produzidos pelos dois métodos não podem ser considerados como provenientes de uma mesma população.
- A precisão dos métodos não pode ser considerada similar, significando que um método é mais preciso que o outro.
- Implica dizer que o método (A: s^2 = 1,76) é menos preciso que o método (B: s^2 = 0,46), e que, para tomar esta decisão, admitiu-se um erro de 5%.
- O significado do erro tipo I é muito claro:
  - A razão entre duas estimativas da variância advindas de uma mesma população, oriundas de um par de amostras, cada uma com n = 10, pode assumir valores maiores ou iguais a 3,18 em 5% dos casos.
  - Não se tem certeza absoluta se o caso analisado é, ou não, um desses possíveis casos.

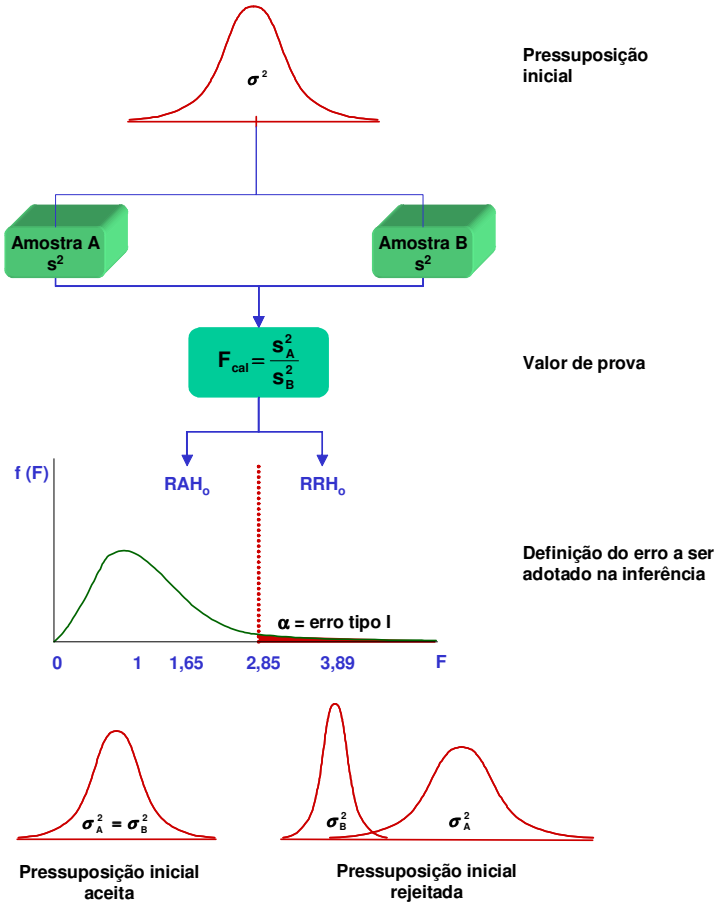
Em síntese:

- Consideraram-se os resultados das determinações dos dois métodos como sendo amostras aleatoriamente retiradas de uma mesma população básica, e admitiu-se que a variável aleatória, ou variável de resposta (determinação da CTC), apresenta distribuição normal.
- A estatística F permitiu decidir, segundo uma determinada probabilidade de erro tipo I (em geral de 1 a 10%, o que implica em 99 a 90% de acerto, respectivamente), se a consideração inicial foi correta ou não, ou seja, se os resultados gerados pelos dois métodos podem ser considerados, ou não, como provenientes de uma mesma população básica:

Hipóteses:

$H_0 : \sigma_A^2 = \sigma_B^2$ (precisão igual = população única)

$H_1 : \sigma_A^2 > \sigma_B^2$ (precisões distintas = populações distintas)



Pressuposição inicial

Valor de prova

Definição do erro a ser adotado na inferência

Pressuposição inicial aceita

Pressuposição inicial rejeitada

Figura 14.3 – Síntese do uso da distribuição F na inferência sobre precisão.

Denominando a linha pontilhada de F_{tab}:

- $F_{cal} < Ftab$: aceita-se a igualdade
- $F_{cal} \geq Ftab$: rejeita-se a igualdade

15. EXEMPLOS BÁSICOS DE INFERÊNCIA ESTATÍSTICA

15.1. Aplicação da distribuição t: teste de hipóteses de uma média com σ desconhecido

Em trabalhos práticos é o teste mais comum.

O desvio padrão σ é estimado à partir da amostra, s.

Utiliza-se a distribuição de Student.

t é uma estatística aproximada enquanto Z é exata.

Amostra		População
n	→	N
s	→	σ
t	→	Z

Quando n ≥ 30, t tende para Z.

Os procedimentos para testar hipóteses são semelhantes aos adotados para a estatística Z, utilizando-se porém a distribuição t.

Exemplo:

Um tratamento A, quando aplicado em fêmeas de peixes de uma determinada espécie e peso, provoca ovulação para fecundação artificial em 50 dias.

Desejando-se reduzir este tempo, um novo tratamento B foi desenvolvido e testado em 25 fêmeas.

Essas apresentaram a desova em média com 48 dias com desvio padrão estimado (s) de 5 dias.

Testar, com margem de segurança de erro de 1% se o novo tratamento reduziu o tempo de liberação da ovulação da espécie em questão:

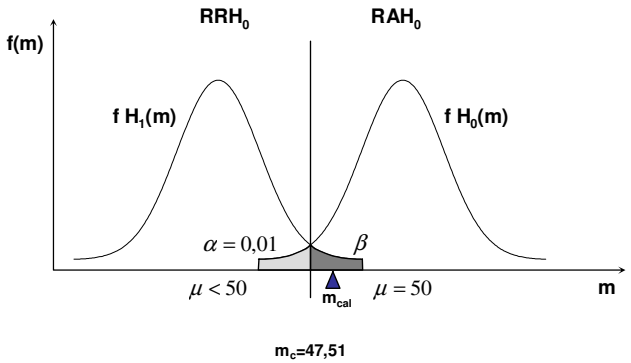
Tratamento A	Tratamento B
μ = 50 dias	m = 48 dias
n = 25 (tamanho da amostra)	
s = 5 dias	

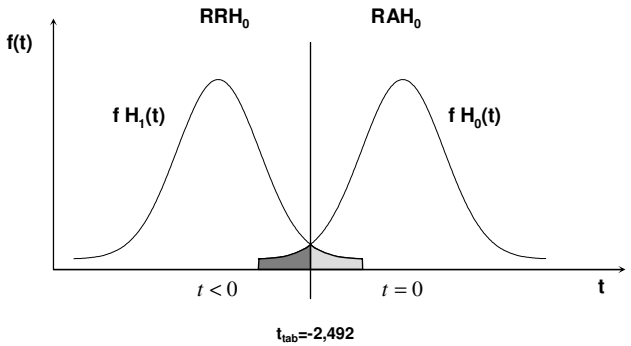
Deseja-se testar:

H_0: μ = 50 dias

H_1: μ < 50 dias

15.1.1. Solução encontrando a média crítica:





$$t = \frac{(\hat{\theta} - \theta)}{s(\hat{\theta})} \therefore\ t = \frac{(m - \mu)}{s(m)} \therefore\ t = \frac{(m - \mu)}{\frac{s}{\sqrt{n}}} \therefore\ -2{,}492 = \frac{(m_c - 50)}{\frac{5}{\sqrt{25}}} \therefore\ m_c = 47{,}51$$

Aceita-se H_0 ao nível de significância de 1%.

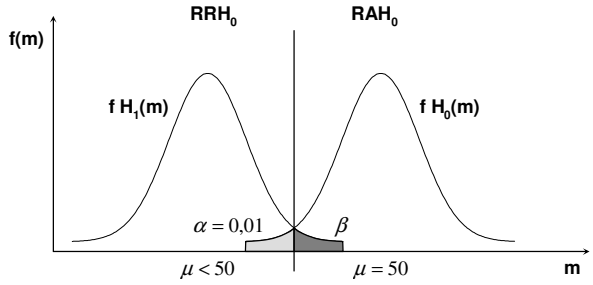
O que implica afirmar que o novo tratamento não reduz significativamente o tempo de ovulação da espécie em questão.

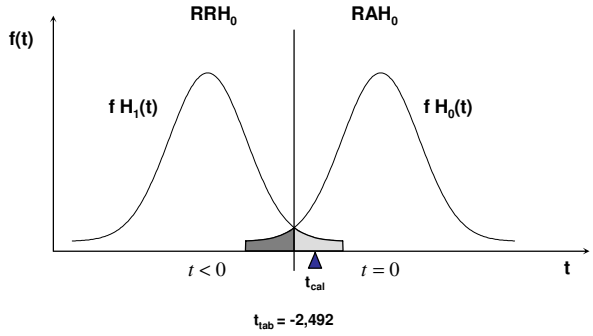
Deseja-se testar:

H_0: μ = 50 dias

H_1: μ < 50 dias

15.1.2. Solução encontrando o valor t crítico:





$$t = \frac{(\hat{\theta} - \theta)}{s(\hat{\theta})} \therefore\ t = \frac{(m - \mu)}{s(m)} \therefore\ t = \frac{(m - \mu)}{\frac{s}{\sqrt{n}}} \therefore\ t_{cal} = \frac{(48 - 50)}{\frac{5}{\sqrt{25}}} \therefore\ t_{cal} = -2{,}0$$

Aceita-se H_0 ao nível de significância de 1%.

O que implica afirmar que o novo tratamento não reduz significativamente o tempo de ovulação da espécie em questão.

15.2. Aplicação da distribuição F: comparação de duas variâncias

Utiliza-se o teste F (distribuição de Snedecor).

Sejam Y_1 e Y_2 duas variáveis aleatórias normalmente distribuídas.

Sejam duas amostras casuais e independentes de tamanho n_1 e n_2 respectivamente.

Sejam as hipóteses:

$$H_0 : \sigma^2_{Y_1} = \sigma^2_{Y_2}$$

$$H_1 : \sigma^2_{Y_1} > \sigma^2_{Y_2}$$

$$H_1 : \sigma^2_{Y_1} < \sigma^2_{Y_2}$$

$$H_1 : \sigma^2_{Y_1} \neq \sigma^2_{Y_2}$$

Para testar H_0 utiliza-se a estatística (F):

$$F = \frac{s^2_{Y_1}}{s^2_{Y_2}}$$

que tem distribuição de Snedecor com $(nY_1 - 1)$ e $(nY_2 - 1)$ gl.

Observação: por convenção o maior valor ocupa a posição do numerador.

Exemplo:

Em uma das turmas, A, da disciplina MEE da UESC, uma amostra de 10 estudantes apresentou, em relação ao rendimento acadêmico, variância de 5 pontos. De uma outra turma, B, foi retirada uma amostra de 6 estudantes, tendo apresentado variância de 2 pontos. Adotando-se α = 5%, pode-se concluir que a variância da turma A é maior que a da turma B?

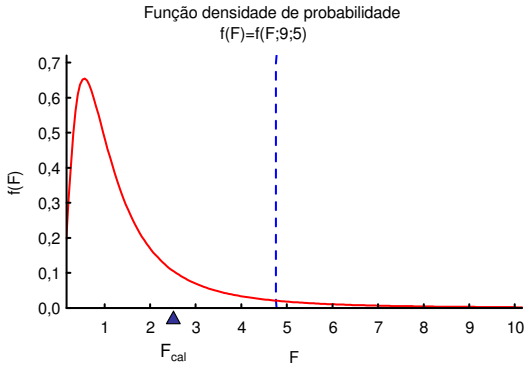
Amostra A	Amostra B
s^2 = 5 pontos	s^2 = 2 pontos
n = 10	n = 6

Deseja-se testar:

$$H_0 : \sigma^2_A = \sigma^2_B$$

$$H_1 : \sigma^2_A > \sigma^2_B$$

$$F = \frac{s^2_A}{s^2_B} = \frac{5}{2} = 2{,}5$$



Portanto, aceita-se H_0.

O que implica em afirmar que a variância da turma A é estatisticamente igual a da turma B ao nível de 5% de significância (probabilidade do erro tipo I).

Tabelas estatísticas

16. TABELAS ESTATÍSTICAS