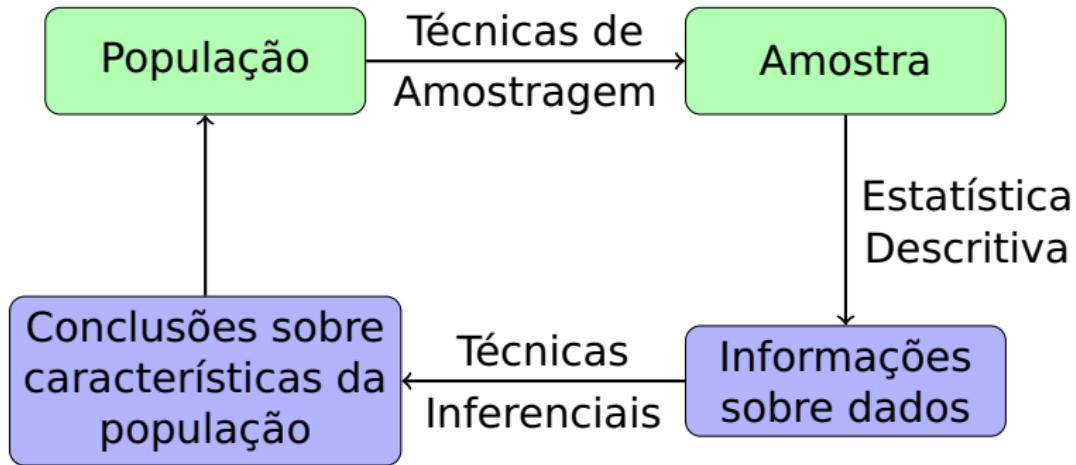


Métodos Estatísticos

Prof. Bruno Santos

Instituto de Matemática e Estatística
Universidade Federal da Bahia

Processo de análise estatística



BASE: PROBABILIDADE

Estatística descritiva:

⇒ Conjunto de técnicas que nos informam sobre propriedades do seu “banco de dados” ou simplesmente sua informação disponível.

- Simetria (Medidas de posição)
- Variabilidade (Medidas de dispersão)
- Gráficos e tabelas

Definições

População: é o conjunto total dos elementos de interesse;

Amostra: é qualquer parcela da população, que é utilizada com o intuito de obter conclusões sobre a população;

Parâmetro: Quantidade desconhecida da população que temos interesse;

Estimativa: O valor numérico com base nas observações amostrais.

Tipos de variável

Variável: característica da população de interesse.

Variável qualitativa

Ordinal

Nominal

Variável quantitativa

Contínua

Discreta

- **Qualitativa**

Nominal: sexo, cor dos olhos, tipos de defeitos, ...;

Ordinal: classe social, grau de instrução, porte de empresa, ...;

- **Quantitativa**

Contínua: peso, altura, vida útil de bateria, ...;

Discreta: número de filhos, número de carros, número de defeitos, ...;

Fontes de informação

Brasil tem diversas fontes de informação, que estão disponíveis para download para análise:

- IBGE
 - ◊ Pesquisa Nacional por Amostragem de Domicílios - PNAD
 - ◊ Pesquisa de Orçamentos Familiares - POF
 - ◊ Censo Demográfico
- IPEADATA
 - ◊ Séries históricas de indicadores

https://www.ibge.gov.br/estatisticas-novoportal/socials/populacao/9127-pesquisa-nacional/

Search

Estatísticas > Sociais > População > Pesquisa Nacional por Amostra de Domicílios

Pesquisa Nacional por Amostra de Domicílios - PNAD

2015 Acesso à Internet e à Televisão e Posse de Telefone Móvel Celular para Uso Pessoal

O que é

Destques
Microdados
Resultados
Publicações
Notas Técnicas
Conceitos e métodos
Downloads
Notícias e releases

O que é

A Pesquisa Nacional por Amostra de Domicílios - PNAD Investiga anualmente, de forma permanente, características gerais da população, de educação, trabalho, rendimento e habitação e outras, com periodicidade variável, de acordo com as necessidades de informação para o País, como as características sobre migração, fecundidade, nupcialidade, saúde, segurança alimentar, entre outros temas.

MAIS INFORMAÇÕES

Estatísticas do site

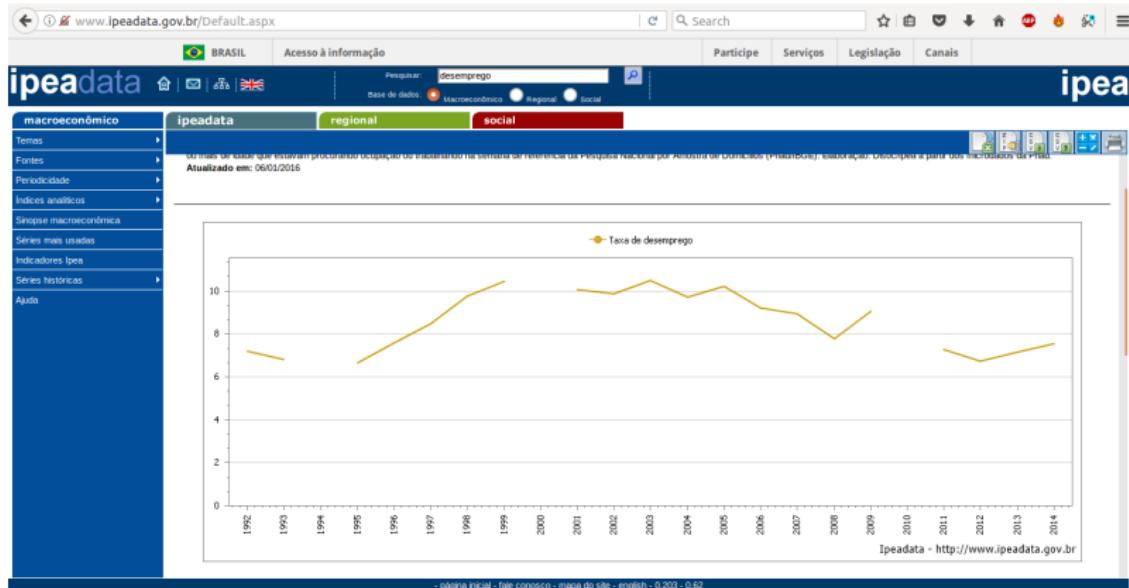
CANAL DE ATENDIMENTO

0800-721-8181

Acesso à Informação

PORTAL DA TRANSPARÉNCIA

IBGE TRANSPARÉNCIA



Exemplo - Tabela

Dados - PNAD 2015

Região	Anos de Estudo	Sexo	Renda (R\$)
Norte	Sem instrução e menos de 1 ano	Masculino	859,15
Norte	Sem instrução e menos de 1 ano	Feminino	626,21
Norte	1 a 3 anos	Masculino	591,75
Norte	1 a 3 anos	Feminino	355,78
Norte	4 a 7 anos	Masculino	635,66
Norte	4 a 7 anos	Feminino	314,72
Norte	8 a 10 anos	Masculino	852,83
Norte	8 a 10 anos	Feminino	361,53
Norte	11 a 14 anos	Masculino	1414,46
Norte	11 a 14 anos	Feminino	757,82
Nordeste	Sem instrução e menos de 1 ano	Masculino	691,24
Nordeste	Sem instrução e menos de 1 ano	Feminino	673,00
Nordeste	1 a 3 anos	Masculino	491,62
Nordeste	1 a 3 anos	Feminino	405,23
Nordeste	4 a 7 anos	Masculino	538,41
Nordeste	4 a 7 anos	Feminino	347,93
Nordeste	8 a 10 anos	Masculino	713,29
Nordeste	8 a 10 anos	Feminino	357,97
Nordeste	11 a 14 anos	Masculino	1334,29
Nordeste	11 a 14 anos	Feminino	735,45
Sudeste	Sem instrução e menos de 1 ano	Masculino	977,74
Sudeste	Sem instrução e menos de 1 ano	Feminino	770,50
Sudeste	1 a 3 anos	Masculino	758,07
Sudeste	1 a 3 anos	Feminino	514,03
Sudeste	4 a 7 anos	Masculino	982,53
Sudeste	4 a 7 anos	Feminino	531,10
.	.	.	.
.	.	.	.
Centro-Oeste	11 a 14 anos	Masculino	2149,36
Centro-Oeste	11 a 14 anos	Feminino	1095,35

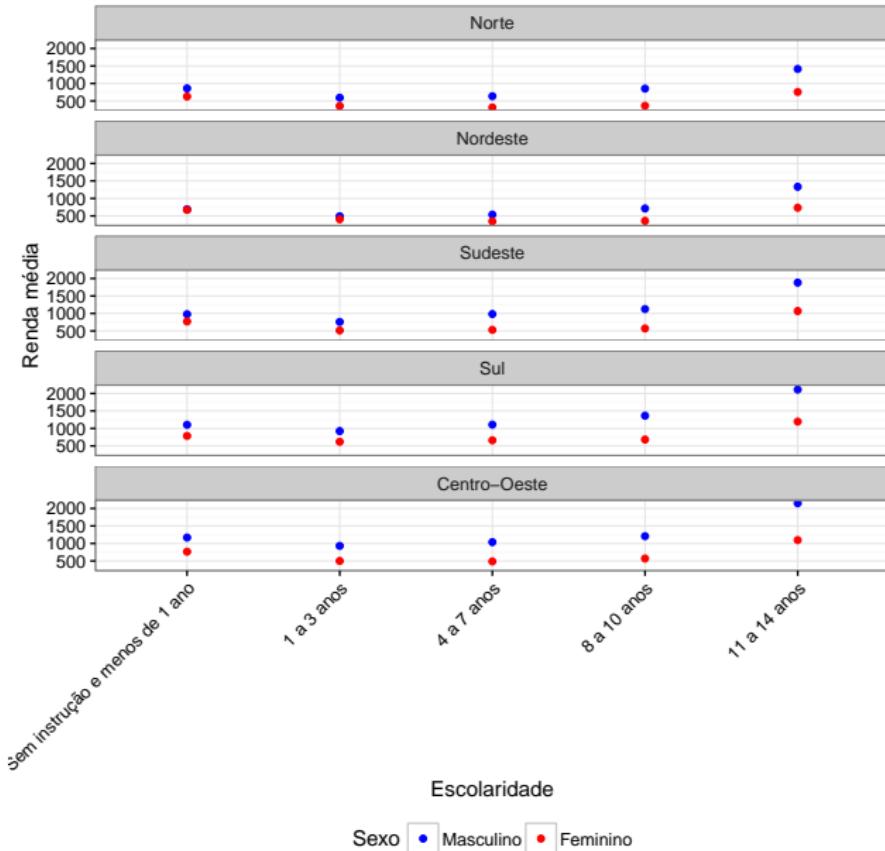
Exemplo - Tabelas Menores

Dados - PNAD 2015

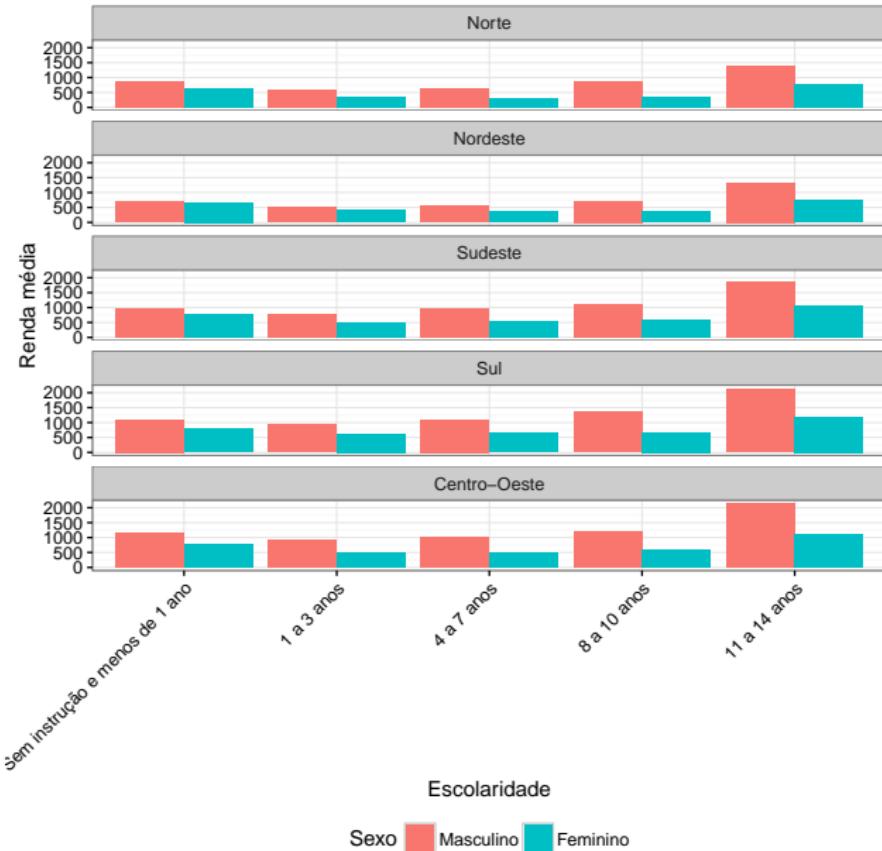
Região	Renda (R\$)
Norte	898,24
Nordeste	833,56
Sudeste	1424,19
Sul	1.539,50
Centro-Oeste	1.560,44

Sexo	Renda (R\$)
Masculino	1.491,58
Feminino	941,00

Exemplo - Gráfico

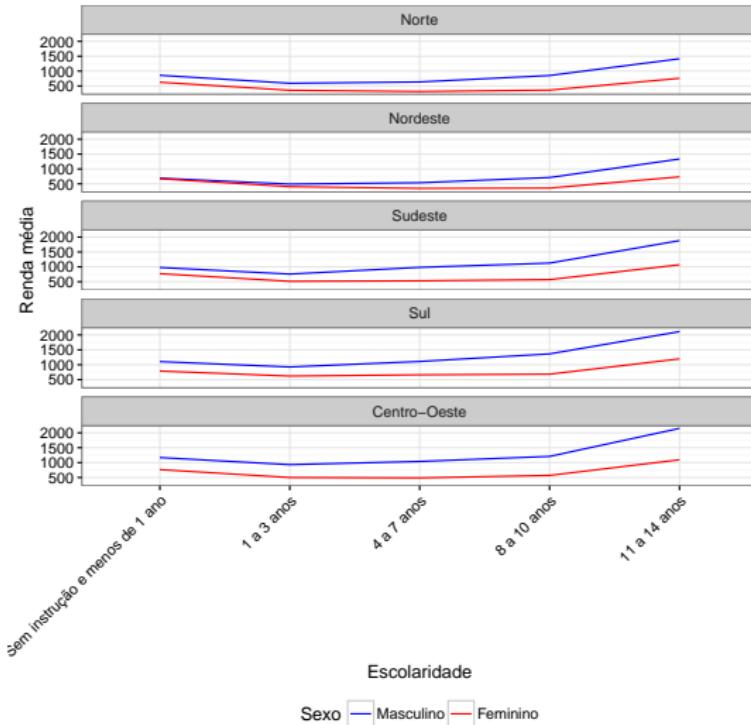


Exemplo - Gráfico

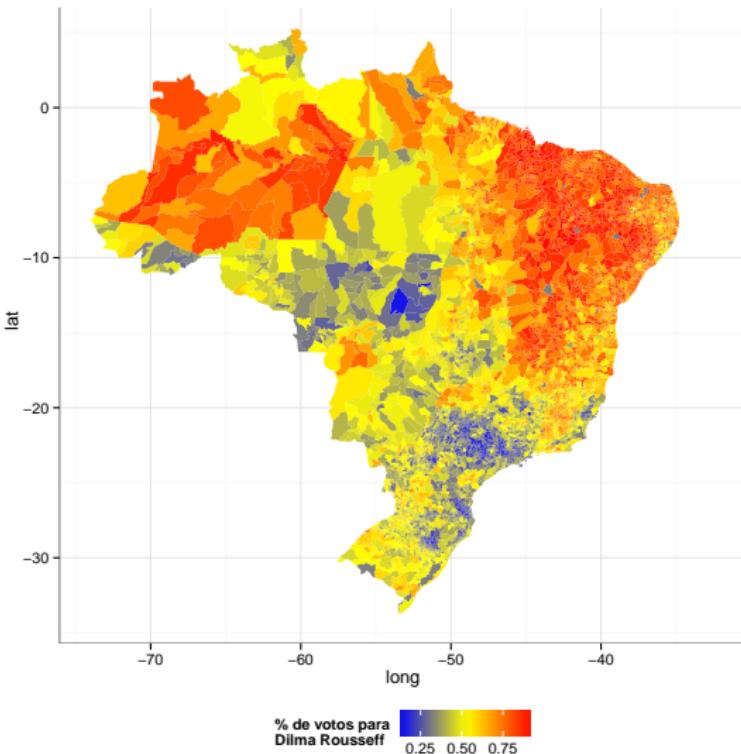


Exemplo - Gráfico

Há algum problema com esse gráfico?



Apresentando informação em mapas

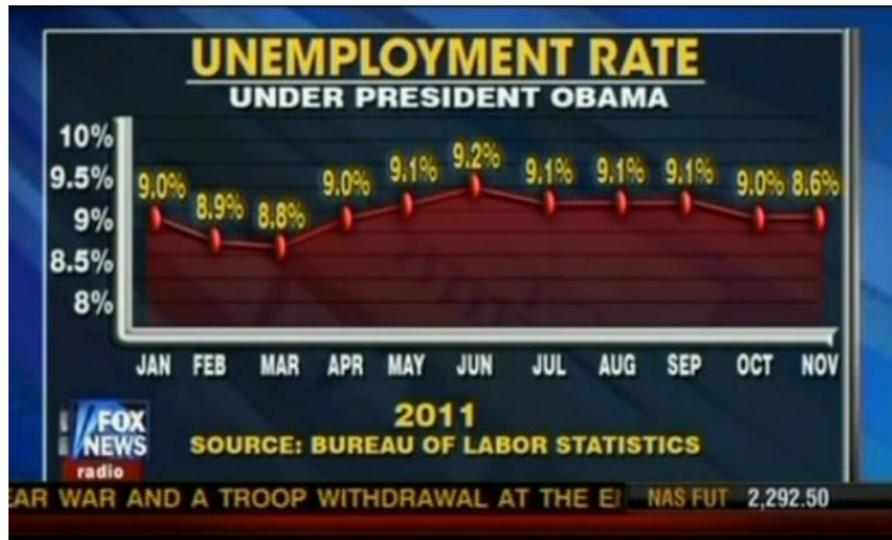


Problemas com gráfico de setores

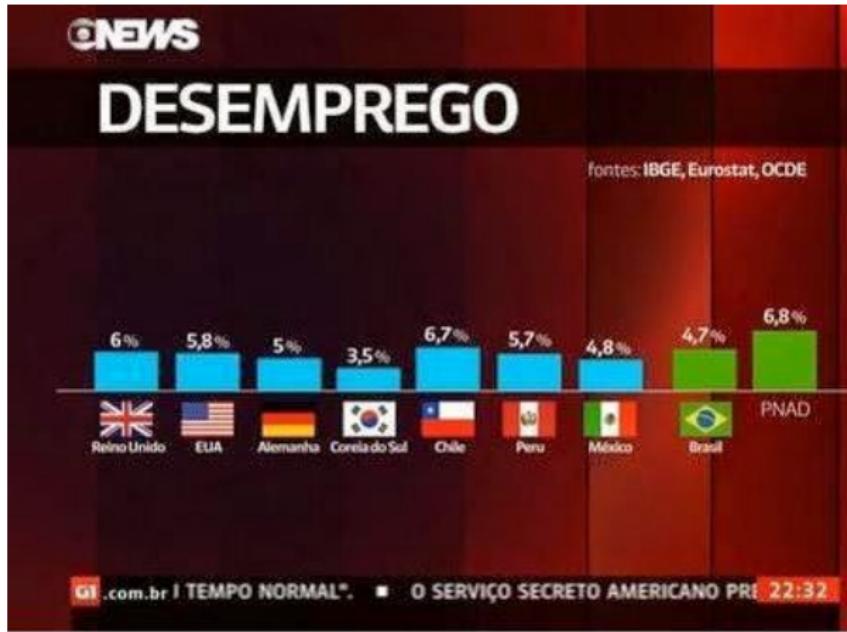
Tabela de participação por região na PNAD

Região	Participação (%)
Norte	16,02
Nordeste	28,76
Sudeste	29,52
Sul	15,27
Centro-Oeste	10,42

Problema no gráfico - Fox News



Problema no gráfico - Globonews



Problema no gráfico - Globonews 2



Conteúdo

Na aula de hoje veremos:

- Distribuição de frequências
- Medidas de tendência central
 - ◊ Média aritmética
 - ◊ Mediana
 - ◊ Moda
- Diferença para dados agrupados ou não agrupados

Dados não agrupados

Notas da turma:

5	6	8	10	5	10	10	8	8	4
5	5	8	6	9	7	9	10	6	9
10	5	8	4	5	6	4	6	10	6
7	8	7	5	9	8	9	4	9	6

Notas da turma ordenadas:

4	4	4	4	5	5	5	5	5	5
5	6	6	6	6	6	6	6	7	7
7	8	8	8	8	8	8	8	9	9
9	9	9	9	10	10	10	10	10	10

Dados agrupados

Dados agrupados sem classe:

Nota	Frequência
4	4
5	7
6	7
7	3
8	7
9	6
10	6

Dados agrupados com classes:

Intervalo da nota	Frequência
4 – 6	11
6 – 8	10
8 – 10	13
10 –	6

Definições

- **Classe:** são os intervalos de variação da variável.
- **Limites de classe:** são os extremos de cada classe.
- **Amplitude do intervalo de classe (h):** é obtida fazendo a diferença entre o limite inferior e o superior de cada classe.
- **Amplitude total:** é a diferença entre os valores mínimo e máximo observados nos dados.
- **Ponto médio da classe:** é o ponto que divide o intervalo de classe em duas partes iguais.

Como construir uma tabela de frequências com classes

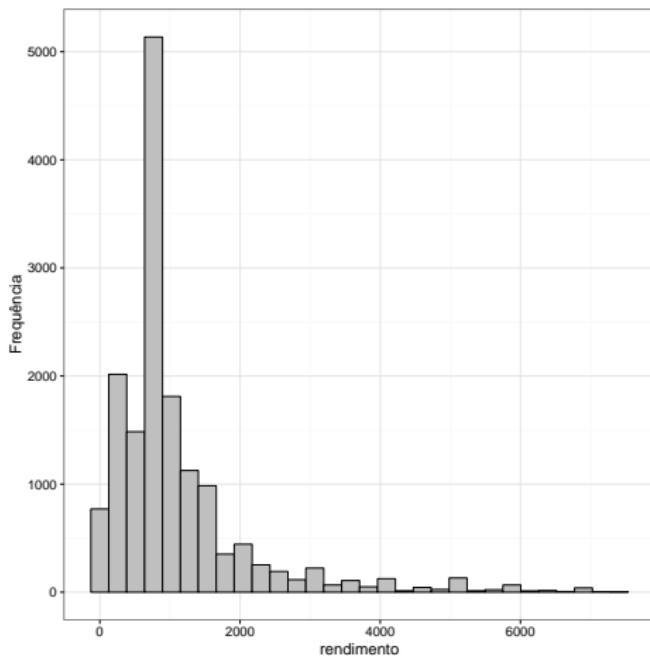
- ➊ Organize os dados de forma ordenada.
- ➋ Calcule a amplitude total (AA).
- ➌ Utilize a “regra de Sturges” ($k \approx 1 + 3,332 \log_{10} n$).

n	Número de classes (k)
6 – 11	4
12 – 22	5
23 – 46	6
47 – 90	7
91 – 181	8
182 – 362	9

- ➍ Calcule a amplitude de cada classe $h > AA/k$.
- ➎ A partir do menor valor, você obtêm a primeira classe e continua assim até todos os dados estarem representados na tabela de frequências.

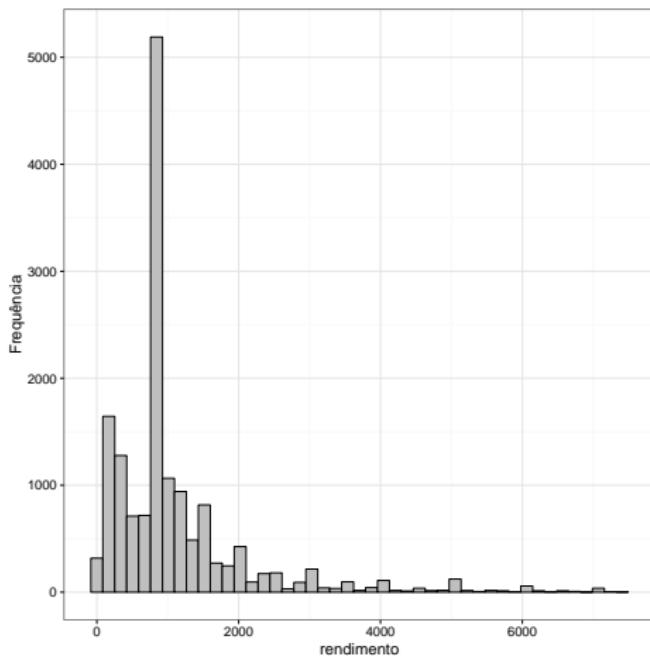
Histograma - Variáveis quantitativas contínuas

Distribuição de renda segundo PNAD 2015 no Estado da Bahia. (30 classes)



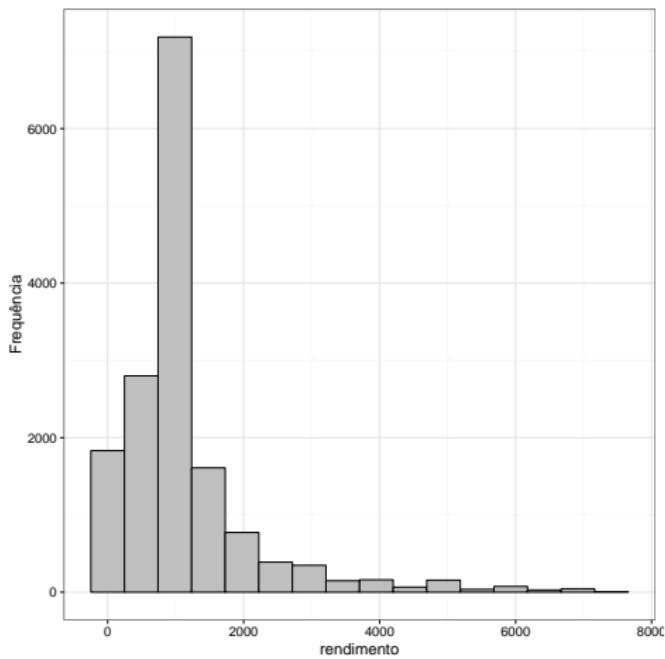
Histograma - Variáveis quantitativas contínuas

Distribuição de renda segundo PNAD 2015 no Estado da Bahia. (45 classes)



Histograma - Variáveis quantitativas contínuas

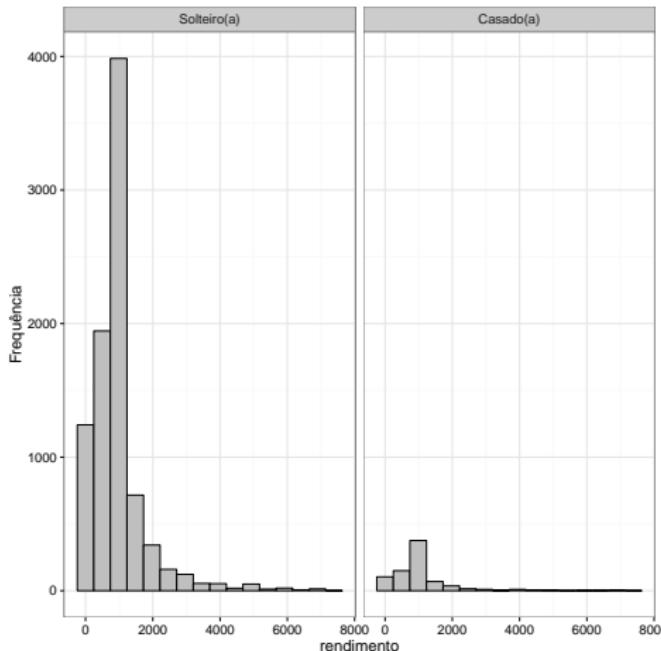
Distribuição de renda segundo PNAD 2015 no Estado da Bahia. (15 classes)



Histograma - Variáveis quantitativas contínuas

Comparação usando variáveis quantitativas discretas

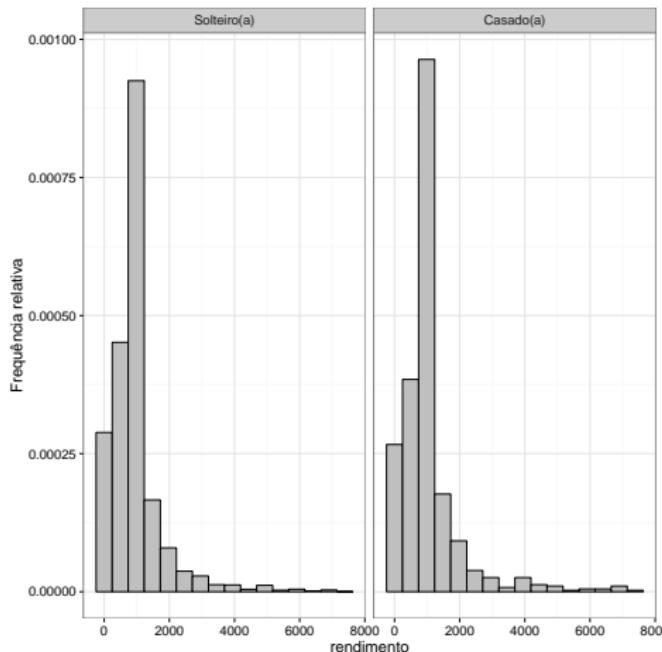
Distribuição de renda segundo PNAD 2015 no Estado da Bahia, segundo Estado Civil.



Histograma - Variáveis quantitativas contínuas

Comparação usando variáveis quantitativas discretas

Distribuição de renda segundo PNAD 2015 no Estado da Bahia, segundo Estado Civil.



Frequências acumuladas

Dados agrupados sem classe:

Nota	f_i	fr_i	F_i	FR_i
4	4	0,10	4	0,10
5	7	0,17	11	0,28
6	7	0,17	18	0,45
7	3	0,07	21	0,53
8	7	0,17	28	0,70
9	6	0,15	34	0,85
10	6	0,15	40	1,00
Total	40			

- f_i : frequência absoluta.
- fr_i : frequência relativa.
- F_i : frequência absoluta acumulada.
- FR_i : frequência relativa acumulada.

Medidas descritivas

Medidas de posição:

- Média
- Mediana
- Moda

Medidas de dispersão:

- Amplitude
- Variância
- Desvio-padrão

Média

Definição: É igual ao quociente entre a soma dos valores do conjunto e o número total dos valores.

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n} = \frac{X_1 + X_2 + \dots + X_n}{n}$$

- Dados não agrupados

Exemplo: Venda de certo produto numa semana:
10, 14, 13, 15, 16, 18 e 12 unidades.

$$\bar{X} = \frac{10 + 14 + 13 + 15 + 16 + 18 + 12}{7} = 14 \text{ unidades.}$$

Propriedades da média

Seja o desvio: $d_i = X_i - \bar{X}$.

- ➊ A soma algébrica dos desvios em relação à média é nula.

No exemplo anterior, $d_1 + d_2 + \dots + d_7 = 0$.

- ➋ Somando-se (ou subtraindo-se) uma constante (c) a todos os valores de uma variável, a média do conjunto fica aumentada (ou diminuída) dessa constante.

No exemplo anterior, caso houvesse um erro na contabilidade das vendas e todos os dias não foram reportadas duas vendas.

$$\bar{X}^* = \bar{X} + 2 = 16.$$

- ➌ Multiplicando-se (ou dividindo-se) todos os valores de uma variável por uma constante (c), a média do conjunto fica multiplicada (ou dividida) por essa constante.

Média para dados agrupados

Sem intervalos de classe:

Nota	f_i
4	4
5	7
6	7
7	3
8	7
9	6
10	6

$$\bar{X} = \frac{\sum_{i=1}^n X_i f_i}{\sum_{i=1}^n f_i} \Rightarrow \bar{X} = 7,1$$

Média para dados agrupados

Intervalo da nota	Frequência	Ponto médio (M_i)
4 – 6	11	5
6 – 8	10	7
8 – 10	13	9
10 –	6	10

$$\bar{X} = \frac{\sum_{i=1}^n M_i f_i}{\sum_{i=1}^n f_i} \Rightarrow \bar{X} = 7,55$$

Moda

É o valor que ocorre com **maior freqüência** em uma série de valores.

Notação: usualmente definimos **Mo** como a moda.

Pode ser definida tanto para variáveis quantitativas quanto para variáveis qualitativas.

Situações de uso:

- Número de peça de roupa mais vendida.
- Salário mais comum numa empresa.

Para dados não agrupados:

- Deve-se buscar o valor mais frequente.
 - ❖ Na série 7 , 8 , 9 , 10 , 10 , 10 , 11 , 12 a moda é igual a 10.

Moda (cont.)

- Há séries nas quais não existe valor modal
 - ◊ Exemplo: 3 , 5 , 8 , 10 , 12 não apresenta moda. A série é **amodal**.
- Em outros casos, pode haver dois ou mais valores de concentração.
 - ◊ Exemplo: 2 , 3 , 4 , 4 , 4 , 5 , 6 , 7 , 7 , 7 , 8 , 9 apresenta duas modas: 4 e 7. A série é **bimodal**.

Mediana

Considerando os dados ordenados, **Mediana** é o valor que separa os dados em duas partes iguais.

Notação: Md .

Dados não agrupados:

Dada uma série de valores como, por exemplo: { 5, 2, 6, 13, 9, 15, 10 }

- ➊ Ordenar dados: { 2, 5, 6, 9, 10, 13, 15 }.
- ➋ Valor que divide a série em duas partes iguais: 9.
- ➌ $Md = 9$.

Método para o cálculo da mediana

Se a série dada tiver número ímpar de termos:

O valor mediano será o termo de ordem dado pela fórmula:

$$\frac{n+1}{2}$$

Exemplo: Calcule a mediana da série { 1, 3, 0, 0, 2, 4, 1, 2, 5 }.

- ① Ordenar a série: { 0, 0, 1, 1, 2, 2, 3, 4, 5 }.
- ② $n = 9$ logo $(n+1)/2$ é dado por $(9+1)/2 = 5$, ou seja, o 5º elemento da série ordenada será a mediana.
- ③ A mediana será o 5º elemento = 2.

Método para o cálculo da mediana (cont.)

Se a série dada tiver número par de termos:

O valor mediano será a media entre os termos de ordem:

$$\left(\frac{n}{2}\right) \text{ e } \left(\frac{n}{2} + 1\right)$$

Exemplo: Calcule a mediana da série { 1, 3, 0, 0, 2, 4, 1, 3, 5, 6 }.

- 1 Ordenar a série: { 0, 0, 1, 1, 2, 3, 3, 4, 5, 6 }.
- 2 $n = 10$ logo $n/2$ e $n/2 + 1$ é dado por 5 e 6, respectivamente.
- 3 O 5º termo é 2 e 6º termo é 3.
- 4 A mediana será

$$Md = \frac{2 + 3}{2}$$

Observações

- Em um conjunto de dados, a mediana, a média e a moda não têm, necessariamente, o mesmo valor.
- A mediana depende da posição e não dos valores dos elementos na série ordenada.
 - ❖ Por esse motivo, essa medida é menos influenciada por valores extremos.
- Exemplo:
 - ❖ Em { 5, 7, 10, 13, 15 } a média = 10 e a mediana = 10.
 - ❖ Em { 5, 7, 10, 13, 65 } a média = 20 e a mediana = 10.
- A mediana permanece a mesma enquanto que a média dobra de valor por influência de uma observação.

Separatrizes

Na aula passada, vimos que a mediana separa o conjunto de dados em duas partes iguais.

Dizemos que é uma medida **separatriz**.

Existem outras medidas separatrizes:

- Quartis
- Decis
- Percentis

Denominamos quartis os valores de uma série que a dividem em quatro partes iguais.

Precisamos portanto de 3 quartis (Q1, Q2 e Q3).

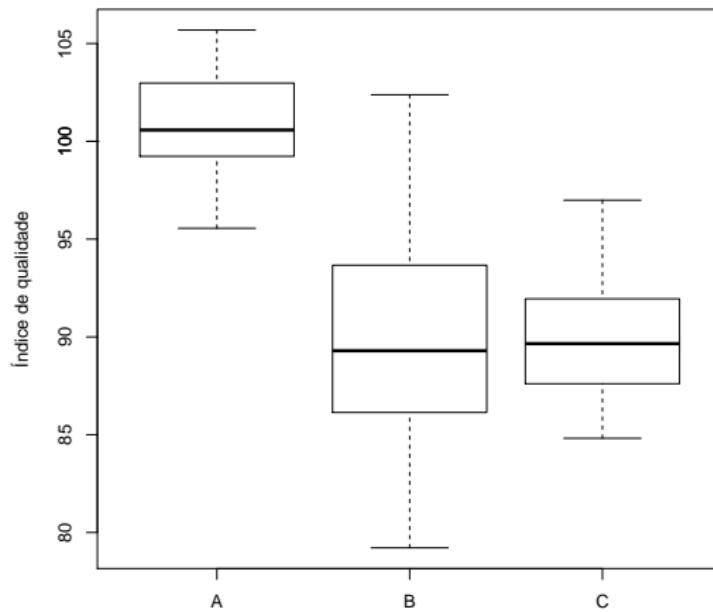
Para dados não agrupados:

- A obtenção dos quartis pode ser feita repetindo o princípio da mediana: na série inteira e nas duas metades obtidas a partir da mediana.

Obs: O 2º quartil (Q2) sempre será igual a mediana da série.

Importância dos quartis

Boxplot:



Exemplo

Calcule os quartis da série: { 5, 2, 6, 13, 9, 15, 10}.

- 1 Ordene os dados: {2, 5, 6, 9, 10, 13, 15}.
- 2 Obtenha a mediana: $Md = 9$, pois esse valor divide a série em duas partes. $\Rightarrow Q2 = 9$.
- 3 Em seguida, obtenha a mediana da série em vermelho: $Q1 = 5$.
- 4 Em seguida, obtenha a mediana da série em azul: $Q3 = 13$.
- 5 Portanto, temos que **$Q1 = 5$, $Q2 = 9$, $Q3 = 13$** .

Decis e percentis

Definição: Denominamos decis os valores de uma série que a dividem em dez partes iguais (D_1, D_2, \dots, D_9).

Definição: Denominamos percentis os valores de uma série que a dividem em cem partes iguais. ($P_1, P_2, \dots, P_{98}, P_{99}$)

Cálculo dos percentis:

A posição do percentil de ordem i no conjunto de dados ordenado será definida como:

$$p_i = i \frac{n}{100}, \quad i = \text{percentil}, \quad i = 1, 2, \dots, 99.$$

- Se $p_i = \text{valor inteiro}$, então o percentil é definido como a média dos valores que ocupam a posição p_i e $p_i + 1$.
- Se $p_i = \text{valor não inteiro}$, então o percentil é definido como o valor que ocupa a posição imediatamente maior que p_i .

Equivalência entre as separatrizes

É possível definir algumas relações de equivalência entre as separatrizes.

- $Md = Q2 = D5 = P50$.
- $D1 = P10; D2 = P20; \dots; D9 = P90$.
- $Q1 = P25; Q3 = P75$.

Exemplo

Calcule os decis do conjunto de notas a seguir:

5.2	5.7	5.8	5.8	6.1	6.5	6.5	7.0
7.3	7.5	7.6	7.7	8.1	8.6	8.6	8.8

Iniciemos pela mediana, que é igual ao 5º decil.

$$r_{50} = 50 \frac{16}{100} = 8 \Rightarrow P50 = D5 = \frac{7.0 + 7.3}{2} = 7.15$$

$$p_{10} = 1.6, p_{20} = 3.2, p_{30} = 4.8, p_{40} = 6.4$$

$$p_{60} = 9.6, p_{70} = 11.2, p_{80} = 12.8, p_{90} = 14.4,$$

Portanto, temos

$$D1 = 5.7, D2 = 5.8, D3 = 6.1, D4 = 6.5$$

$$D6 = 7.5, D7 = 7.7, D8 = 8.1, D9 = 8.6$$

Medidas de dispersão

Por que é necessário obter essas medidas?

Considere o seguinte conjunto de dados:

- $X = \{7, 7, 7, 7, 7\}$;
- $Y = \{6, 6.5, 7, 7.5, 8\}$.
- $Z = \{4, 5, 7, 9, 10\}$.

Para ambos, temos que

$$\bar{X} = \bar{Y} = \bar{Z} = 7.$$

É preciso definir outras medidas para obtermos mais informação dos dados, além de medidas de posição.

Amplitude - exemplo



Neste domingo, em Cuiabá, a temperatura varia entre 17 e 34°C e em Goiânia, a temperatura varia entre 11 e 28°C, com isso, a amplitude térmica chega à 17°C nessas capitais.

Já em Rio Branco, a temperatura varia entre 19 e 34°C e em Porto Velho, a temperatura varia entre 21 e 36°C, e a amplitude térmica chega à 15°C nessas capitais. Assim como em Porto Alegre, onde a temperatura varia entre 14 e 29°C neste domingo.

Com amplitude térmica em 14°C, Curitiba fica com temperatura entre 9 e 23°C e o Rio de Janeiro fica com temperatura entre 14 e 28°C, com isso, a amplitude térmica chega à 17°C nessas capitais.

Amplitude total

Definição: diferença entre o maior e o menor valor observado.

Observações:

- Única medida de dispersão que não tem como referência a média.
- Só leva em consideração duas observações.
 - ◊ $X = \{4, 4, 4, 4, 10\}$;
 - ◊ $Y = \{4, 5, 7, 9, 10\}$.
- Pode ser considerada uma medida de dispersão absoluta.

Exemplos

- $X = \{34, 24, 45, 77, 63\}.$

- ◊ $\min(X) = 24$
 - ◊ $\max(X) = 77$
 - ◊ $AT = 77 - 24 = 53.$

- $X = \{11, 13, 9, 13, 2\}.$

- ◊ $\min(X) = 2$
 - ◊ $\max(X) = 13$
 - ◊ $AT = 13 - 2 = 11.$

Amplitude para dados agrupados em classes

Classe	Freq. abs.	Freq. acumulada
50 — 54	4	4
54 — 58	9	13
58 — 62	11	24
62 — 66	8	32
66 — 70	5	37
70 — 74	3	40
Total	40	

- Não sabemos exatamente os valores das observações.
- Então, a amplitude só pode ser aproximada por $74 - 50 = 24$.

Desvio inter-quartil

Definição: a diferença entre Q_3 e Q_1 .

Observações:

- Também pode ser chamada desvio interquartílico.
- Medida importante na construção do gráfico boxplot.
- Medida que recebe menor influência de pontos extremos.

Exemplos

- $X = \{24, 34, 45, 63, 77\}$.

- ◊ $Q1 = \frac{24 + 34}{2} = 29$

- ◊ $Q3 = \frac{63 + 77}{2} = 70$

- ◊ $IQ = 70 - 29 = 41$

- $X = \{2, 9, 11, 13, 13\}$.

- ◊ $Q1 = \frac{2 + 9}{2} = 5,5$

- ◊ $Q3 = \frac{13 + 13}{2} = 13$

- ◊ $AT = 13 - 5,5 = 7,5$.

Desvio-padrão

- Medida de dispersão mais utilizada.

Definição: raiz quadrada da média aritmética dos desvios ao quadrado.

$$DP(X) = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n}}$$

Exemplo

Dados:

- $X = \{3, 4, 7, 10, 12\}.$

- $\bar{X} = 7, 2$
- $D_1 = (3 - 7, 2)^2 = 17, 64$
- $D_2 = (4 - 7, 2)^2 = 10, 24$
- $D_3 = (7 - 7, 2)^2 = 0, 04$
- $D_4 = (10 - 7, 2)^2 = 7, 84$
- $D_5 = (12 - 7, 2)^2 = 23, 04$

$$DP(X) = \sqrt{\frac{17, 64 + 10, 24 + 0, 04 + 7, 84 + 23, 04}{5}} \\ = 3, 4293.$$

Propriedades

- ➊ Somando-se (ou subtraindo-se) uma constante a todos os valores, o desvio-padrão não se altera.
- ➋ Multiplicando-se (ou dividindo-se) uma constante a todos os valores, o desvio-padrão fica multiplicado (ou dividido) por essa constante.
- ➌ Quando os dados são uma amostra de uma população, usamos a seguinte fórmula

$$\sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}}$$

Desvio-padrão para dados agrupados

Para dados agrupados, consideramos o seguinte:

$$DP(X) = \sqrt{\frac{\sum_{i=1}^n f_i * (X_i - \bar{X})^2}{\sum_{i=1}^n f_i}}$$

Ou ainda,

$$DP(X) = \sqrt{\frac{\sum_{i=1}^n f_i * (X_i - \bar{X})^2}{(\sum_{i=1}^n f_i) - 1}}$$

Exercícios

1 Considere os seguintes conjuntos de números:

- ◊ $A = \{ 10, 20, 30, 40, 50 \}$
- ◊ $B = \{ 100, 200, 300, 400, 500 \}$

Que relação existe entre os desvios padrões dos dois conjuntos de números? **Resposta: $DP(B) = 10 * DP(A)$**

2 Considere os seguintes conjuntos de números:

- ◊ $A = \{ 220, 230, 240, 250, 260 \}$
- ◊ $B = \{ 20, 30, 40, 50, 60 \}$

Que relação existe entre os desvios padrões dos dois conjuntos de números? **Resposta: $DP(B) = DP(A)$**

Exercícios (cont.)

3 Dados os seguintes conjuntos de números:

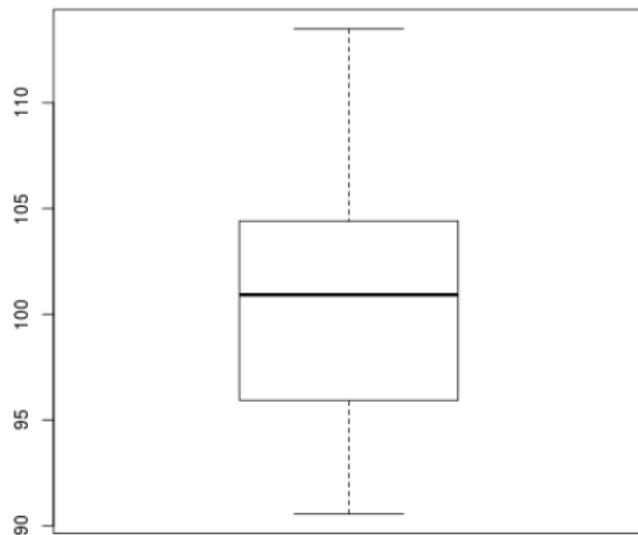
- ◊ $A = \{ -2, -1, 0, 1, 2 \}$
- ◊ $B = \{ 220, 225, 230, 235, 240 \}$

Podemos afirmar, de acordo com as propriedades do desvio padrão, que o desvio padrão de B é igual:

- a) ao desvio padrão de A;
- b) ao desvio padrão de A, multiplicado pela constante 5, e esse resultado somado a 230;
- c) ao desvio padrão de A multiplicado pela constante 5;
- d) ao desvio padrão de A mais a constante 230.

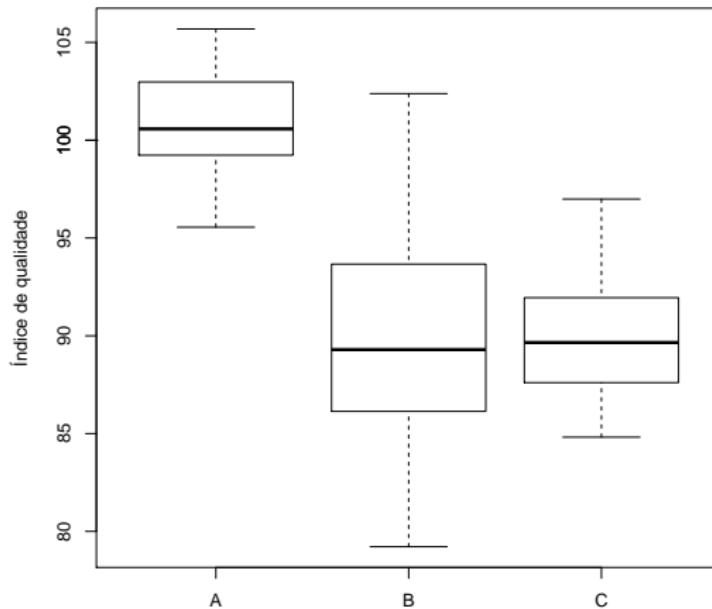
Boxplot

Medidas necessárias: mínimo, máximo, Q1, Mediana (Q2), Q3.



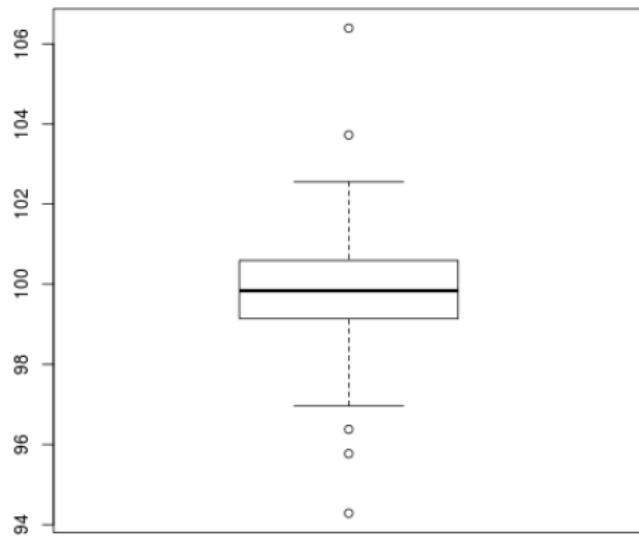
Boxplot

Para comparação de grupos:



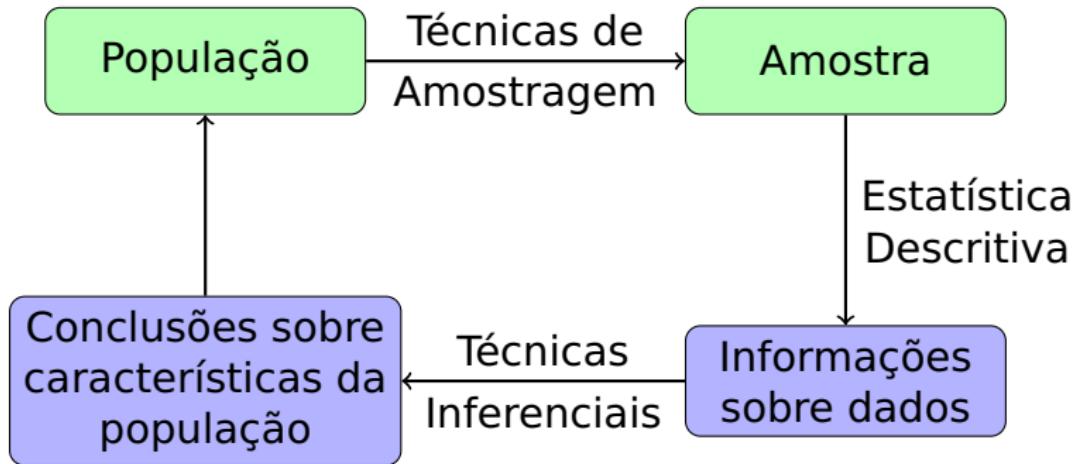
Boxplot

Para apontamento de pontos extremos:



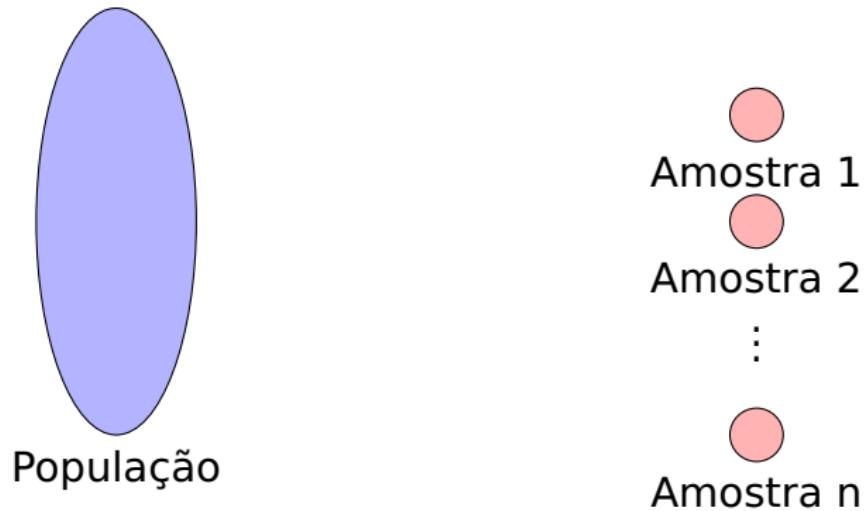
Inferência Estatística

Processo de análise estatística



BASE: PROBABILIDADE

Inferência Estatística



Objetivo: Produzir procedimentos que possam representar os verdadeiros valores da população com bom nível de “confiança”.

Definições

Parâmetro: Quantidade desconhecida da população que temos interesse em estimar.

Estatística: Qualquer função (valor calculado) com base nos valores da amostra.

Estimador: Uma estatística com objetivo de estimar o parâmetro populacional.

Estimativa: O valor numérico com base nas observações amostrais.

Exemplos de estatísticas, estimadores e estimativas

Supondo a amostra X_1, X_2, \dots, X_n ,

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i = \frac{X_1 + \cdots + X_n}{n}$$

A média amostral é um estimador da média populacional.

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

A variância amostral é um estimador da variância populacional.

Distribuição amostral

- Esses estimadores obtidos a partir de uma amostra devem ter uma medida de incerteza associada ao seu valor.
- Estimadores são variáveis aleatórias, com distribuição de probabilidade, valor esperado e variância.
- A distribuição amostral é um dos principais componentes da inferência estatística.

Distribuição amostral de \bar{X}

Seja X_i uma variável aleatória com média μ e variância σ^2 . E seja,

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i = \frac{X_1 + \cdots + X_n}{n}$$

Então,

$$E(\bar{X}) = \mu, \quad \text{Var}(\bar{X}) = \frac{\sigma^2}{n}$$

Se $X \sim N(\mu, \sigma^2)$, então

$$\bar{X} \sim N(\mu, \sigma^2/n)$$

Teorema Central do Limite

Seja $\{X_n\}$, $n > 1$, uma sequência de variáveis aleatórias independentes e identicamente distribuídas, com média μ e variância $\sigma^2 < \infty$. Então, definindo

$$S_n = \sum_{i=1}^n X_i,$$

têm-se o seguinte resultado

$$\frac{S_n - E(S_n)}{\sqrt{\text{Var}(S_n)}} = \frac{S_n - n\mu}{\sigma\sqrt{n}} \sim N(0, 1)$$

Resumo - Distribuição da média amostral

Se X_1, \dots, X_n é uma amostra de uma população com média μ e variância σ^2 , então

$$E(\bar{X}) = \mu, \quad \text{Var}(\bar{X}) = \sigma^2/n.$$

Se $X \sim N(\mu, \sigma^2) \Rightarrow \bar{X} \sim N(\mu, \sigma^2/n).$

Pelo Teorema do Limite Central, para n grande, temos que

$$\bar{X} \sim N(\mu, \sigma^2/n)$$

Distribuição da proporção amostral - \hat{p}

Seja uma determinada característica de interesse na população e suponha que a proporção de pessoas na população com essa característica é p .

Defina a seguinte variável aleatória

$$X = \begin{cases} 1, & \text{se o individuo é portador da característica} \\ 0, & \text{se o individuo não é portador da característica} \end{cases}$$

Então,

$$X \sim \text{Ber}(p) \Rightarrow E(X) = p, \quad \text{Var}(X) = p(1-p)$$

Aproximação da Binomial pela distribuição Normal

Seja $S_n = \sum_{i=1}^n X_i$. Então,

$$S_n \sim \text{Binomial}(n, p)$$

Defina \hat{p} como a proporção amostral, isto é,

$$\hat{p} = \frac{S_n}{n} = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X}.$$

Logo, pelo Teorema Central do Limite, têm-se que, quando n é grande

$$\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$$

Aproximação da distribuição binomial pela normal

Exercícios - Aprox. Binomial para Normal

1. Um procedimento de controle de qualidade foi planejado para garantir um máximo de 10% de itens defeituosos na produção. A cada 60 minutos sorteia-se uma amostra de 50 peças, e, havendo mais de 15% de defeituosos, para-se produção para verificações. Qual a probabilidade de uma parada desnecessária?

$$X \sim \text{Bin}(50, 0.10) \Rightarrow E(X) = 5, \text{Var}(X) = 4.5$$

$$\begin{aligned} P(X > 7,5) &= P\left(\frac{X - 5}{\sqrt{4.5}} > \frac{7.5 - 5}{\sqrt{4.5}}\right) \\ &= P(Z > 1,1785) \\ &= 0,5 - 0,381 = 0,119 \end{aligned}$$

Distribuição amostral de S^2

Considerando uma amostra de n observações de $X \sim N(\mu, \sigma^2)$ e seja S^2 a variância amostral, em que

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2.$$

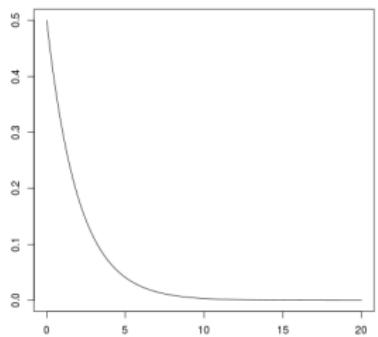
Então, é possível mostrar que

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2,$$

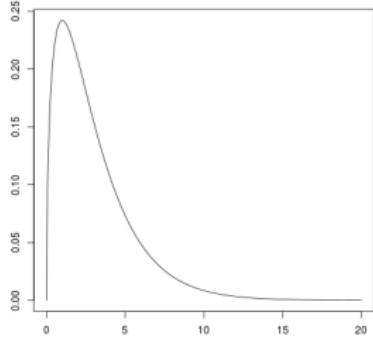
em que χ_{n-1}^2 representa uma distribuição Qui-Quadrado com $n-1$ graus de liberdade.

Distribuição χ^2_{ν}

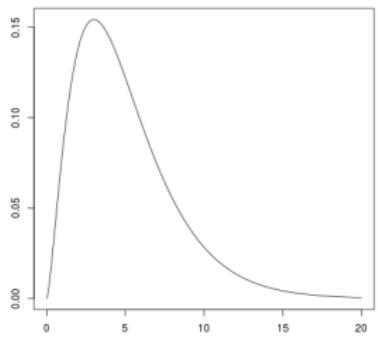
GL = 2



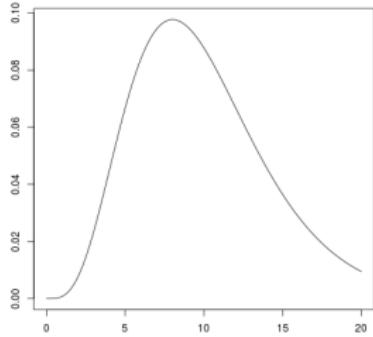
GL = 3



GL = 5



GL = 10



Exercícios - Distribuição amostral de S^2

2. Determine a probabilidade de que uma amostra aleatória de 25 observações, de uma população normal com variância $\sigma^2 = 6$, terá uma variância amostral S^2 .

a) Maior que 9,1.

$$\begin{aligned} P(S^2 > 9,1) &= P\left(\frac{(n-1)S^2}{\sigma^2} > \frac{24 \times 9}{6}\right) \\ &= P(\chi_{24}^2 > 36) \approx 0.05 \end{aligned}$$

Outra distribuição amostral

No caso anterior, σ^2 é suposto conhecido. Essa suposição pode ser considerada **problemática**. Por esse motivo, considere

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}},$$

em que \bar{X} é a média amostral e S^2 é a variância amostral.

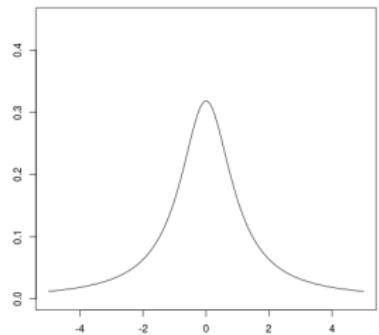
É possível mostrar que

$$T \sim t_{n-1},$$

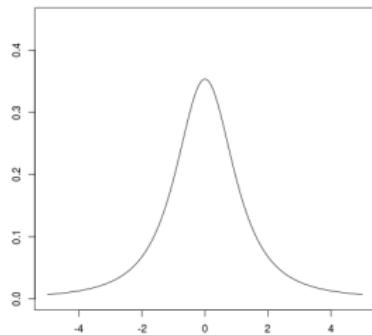
em que t_{n-1} identifica uma distribuição t-Student com $n-1$ graus de liberdade.

Distribuição t_ν

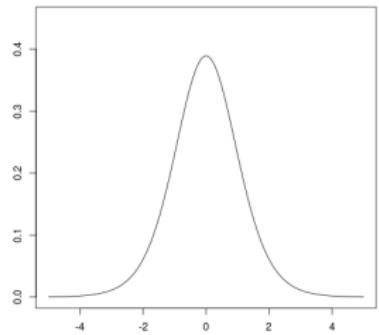
GL = 1



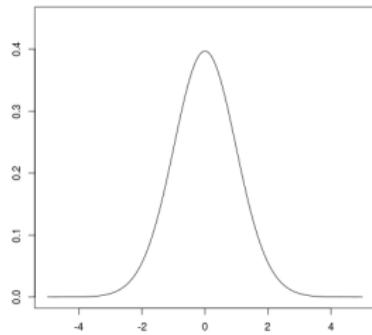
GL = 2



GL = 10



GL = 50



Exercícios - Distribuição amostral (t-Student)

Um engenheiro químico afirma que a média populacional do rendimento de certo lote do processo é 500 gramas por mililitro de matéria-prima. Para verificar essa afirmação, ele amosta 25 lotes a cada mês. Se o valor t calculado ficar entre $-t_{0,05;24}$ e $t_{0,05;24}$, ele fica satisfeito com sua afirmação. A que conclusão ele deveria chegar em relação a uma amostra que tem média $\bar{X} = 518$ gramas por mililitro e desvio padrão 40 gramas? Assuma que a distribuição dos rendimentos é aproximadamente normal.

$$\begin{aligned} T &= \frac{\bar{X} - \mu}{S/\sqrt{n}} \\ &= \frac{518 - 500}{40/\sqrt{25}} \\ &= 2,25 \end{aligned}$$

O valor está fora do intervalo [-1,711; 1,711], portanto ele não deve estar satisfeito com sua afirmação.

Inferência estatística

- Parâmetros em geral são desconhecidos.
- A partir de uma amostra, é possível:
 - ◊ Estimar valores possíveis para esse parâmetro;
 - ◊ Testar se algumas hipóteses são válidas.

X : Renda individual de uma pessoa $\Rightarrow X \sim N(\mu, \sigma^2)$

X : Número de pessoas que têm TV a cabo $\Rightarrow X \sim \text{Bin}(n, p)$

- Estatística paramétrica.

Exemplo

- Interesse: tempo de vida de lâmpadas fabricadas numa certa fábrica.
- Seria importante que o tempo de vida fosse maior que as lâmpadas fabricadas atualmente.
- POPULAÇÃO: todas as lâmpadas feitas nessa fábrica.
- AMOSTRA: algum número de lâmpadas selecionadas.
- Suposição, o tempo de vida $T \sim \text{Exp}(\alpha)$.
- Objetivo: estimar α .
- Estimação pode ser feita de forma pontual ou intervalar.

Alguns estimadores pontuais

Parâmetro	Estimador
Média (μ)	$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$
Variância (σ^2)	$\bar{S} = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$
Proporção (p)	$\hat{p} = \frac{W}{n}$

W é o número de pessoas que têm a característica de interesse

n é o tamanho da amostra.

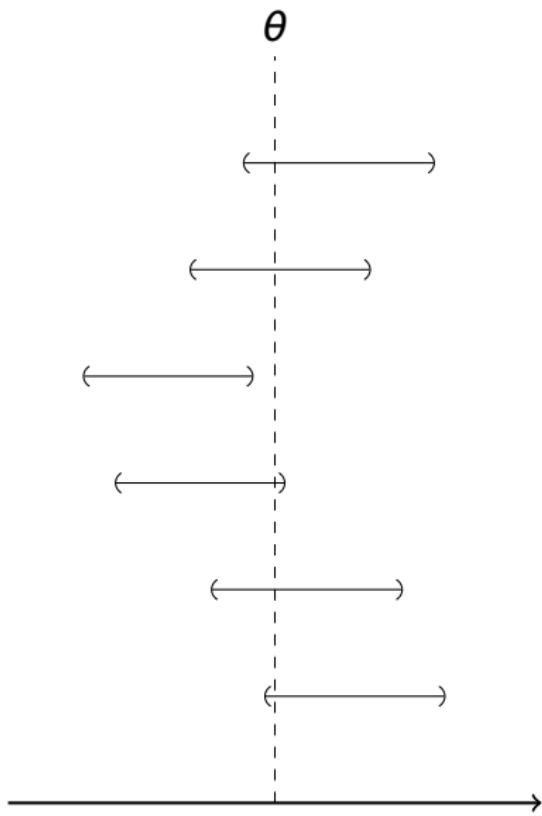
Estimação intervalar

- Procura determinar um intervalo que possa conter o verdadeiro valor do parâmetro.
- Depende do coeficiente de confiança (ou grau de confiança) γ

$$P(L_i < \theta < L_s) = \gamma$$

- Valores mais utilizados:
 - ◊ $\gamma = 0, 95$.
 - ◊ $\gamma = 0, 90$.
 - ◊ $\gamma = 0, 99$.

Explicação do nível de confiança



Espera-se
que $\gamma \times$
100% desses
intervalos
contenham
o verdadeiro
valor de θ

Intervalo de confiança para a média

É preciso distinguir entre as seguintes situações:

- Pequenas amostras ($n < 30$):
 - ◊ Distribuição normal
 - ◊ Distribuição não-normal
- Grandes amostras ($n > 30$):
 - ◊ Distribuição normal
 - ◊ Distribuição não-normal

Pequenas amostras - Distribuição normal

σ^2 conhecido

Considere a seguinte quantidade

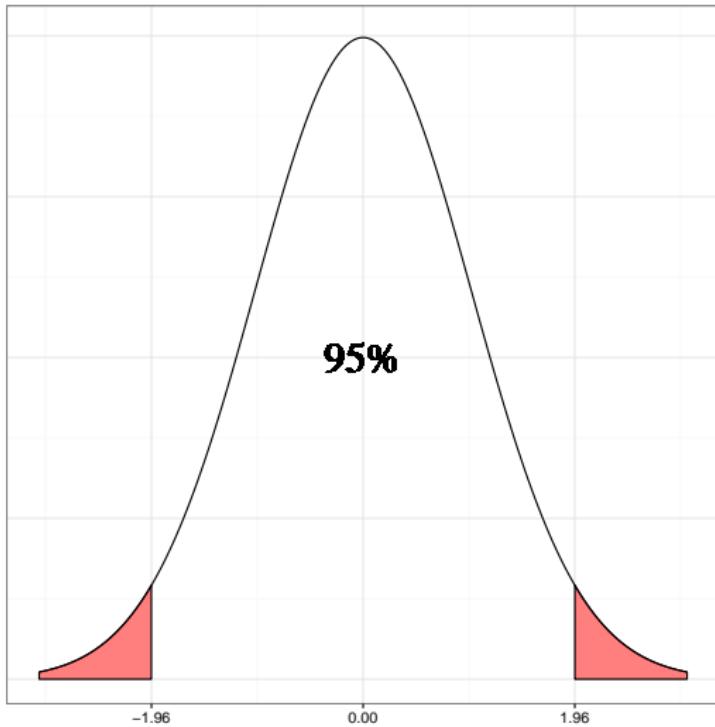
$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

Dessa relação, temos que

$$\begin{aligned}\gamma &= 1 - \alpha = P(-z_{1-\alpha/2} < Z < z_{1-\alpha/2}) \\&= P\left(-z_{1-\alpha/2} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_{1-\alpha/2}\right) \\&= P\left(\bar{X} - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}\right)\end{aligned}$$

$$\Rightarrow \left[\bar{X} - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \right]$$

Ilustração - N(0,1)



Exemplo

Um pesquisador está estudando a resistência média de um determinado material.

$$X \sim N(\mu, 4)$$

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

Intervalo de confiança com coeficiente de confiança 95%:

$$\bar{X} = 6,2, \quad n = 9, \quad \sigma = 2$$

$$\left[6,2 - 1,96 \frac{2}{\sqrt{9}}, 6,2 + 1,96 \frac{2}{\sqrt{9}} \right] = [4,915; 7,529]$$

Pequenas amostras - Distribuição normal σ^2 desconhecido

Considere a seguinte quantidade

$$T = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{n-1}$$

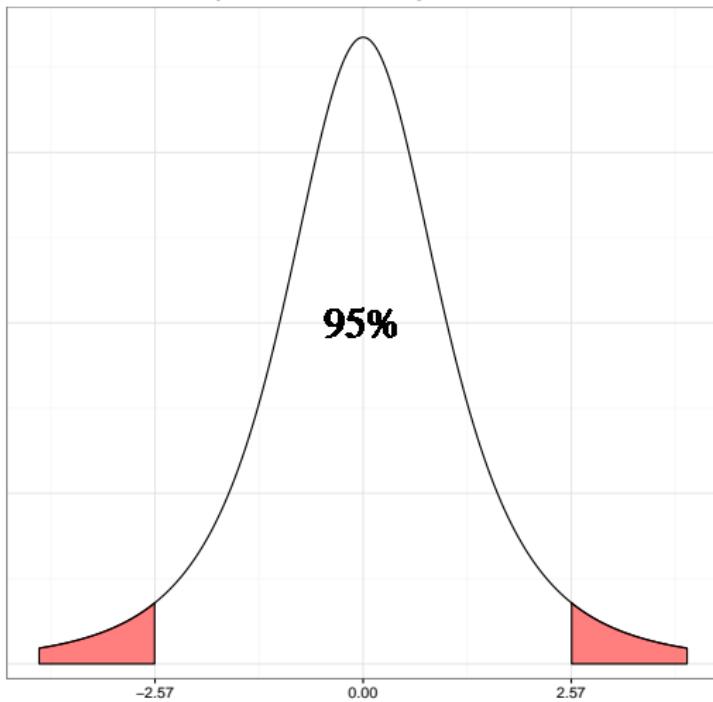
Dessa relação, temos que

$$\gamma = 1 - \alpha = P(-t_{1-\alpha/2;n-1} < T < t_{1-\alpha/2;n-1})$$

$$\Rightarrow \left[\bar{X} - t_{1-\alpha/2;n-1} \frac{s}{\sqrt{n}}, \bar{X} + t_{1-\alpha/2;n-1} \frac{s}{\sqrt{n}} \right]$$

Ilustração - t_5

Distribuição t-Student com 5 graus de liberdade



Exemplo

O consumo diário de alimentos observado em certa amostra da população é, em calorias ($\times 100$)

10	11	11	12	13	13	13	13	13	14	14	14	15	15	16	16
----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----

Objetivo: Construir um intervalo de confiança com 90% de coeficiente de confiança.

Solução: ($\bar{X} = 13,3125$, $s = 1,7404$)

$$\left[\bar{X} - t_{1-\alpha/2;n-1} \frac{s}{\sqrt{n}}, \bar{X} + t_{1-\alpha/2;n-1} \frac{s}{\sqrt{n}} \right]$$

$$\left[13,3125 - 1,753 \frac{1,7404}{4}; 13,3125 + 1,753 \frac{1,7404}{4} \right]$$

$$[12,543; 14,073]$$

Intervalo de confiança para grandes amostras

População normal ou não-normal

- Se n é suficiente grande ($n > 30$);
- Sem conhecimento da distribuição da população;
- Utilizar o desvio-padrão amostral ao invés do desvio-padrão populacional
- O intervalo de confiança para a média fica dado por

$$\Rightarrow \left[\bar{X} - z_{1-\alpha/2} \frac{s}{\sqrt{n}}; \bar{X} + z_{1-\alpha/2} \frac{s}{\sqrt{n}} \right]$$

Exemplo

Resistência à tração de 31 corpos de prova

131	132	134	135	136	135	138
139	140	142	143	144	144	145
146	146	147	147	148	149	150
150	151	151	152	152	153	153
154	160	160				

Objetivo: Obter intervalo de confiança com $\gamma = 0,95$ para a média populacional.

Temos que $\bar{X} = 145,39$ e $s = 7,75$.

$$\left[145,39 - 1,96 \frac{7,75}{\sqrt{31}}, 145,39 + 1,96 \frac{7,75}{\sqrt{31}} \right]$$

$$[142,66; 148,12]$$

Intervalo de confiança para proporção populacional

- Pode haver um interesse em construir um I.C. para p .
- Um estimador pontual para p é dado por

$$\hat{p} = \frac{X}{n}$$

- Para n grande, podemos usar a aproximação da Normal

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0, 1)$$

Intervalo de confiança para a proporção populacional (cont.)

Portanto,

$$\begin{aligned}\gamma = 1 - \alpha &= P(-z_{1-\alpha/2} < Z < z_{1-\alpha/2}) \\ &= P\left(-z_{1-\alpha/2} < \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} < z_{1-\alpha/2}\right)\end{aligned}$$

Como p não é conhecido, faz-se a substituição na variância pelo seu estimador \hat{p} .

Logo, o I.C. para p é dado por

$$\Rightarrow \left[\hat{p} - z_{1-\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}; \hat{p} + z_{1-\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right]$$

Exemplo

Examinam-se 98 animais, encontrando-se 53 infectados com determinado vírus.

Objetivo: Obter I.C. de 95% de confiança para a proporção p de animais infectados.

Solução: ($n = 98$, $\hat{p} = \frac{53}{98} = 0,541$, $(1 - \hat{p}) = 0,459$)

$$\left[0,541 - 1,96 \sqrt{\frac{0,541 \times 0,459}{98}}; 0,541 + 1,96 \sqrt{\frac{0,541 \times 0,459}{98}} \right]$$

$$[0,442; 0,640]$$

Intervalo de confiança para σ^2

- $X \sim N(\mu, \sigma^2)$
- $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$

Dessa maneira é possível considerar o seguinte

$$\gamma = P(q_{\alpha/2; n-1} < \chi^2_{n-1} < q_{1-\alpha/2; n-1})$$

Assim, o I.C. com nível de confiança γ para σ^2 é dado por

$$\left[\frac{(n-1)S^2}{q_{1-\alpha/2; n-1}}; \frac{(n-1)S^2}{q_{\alpha/2; n-1}} \right]$$

Intervalo de confiança para σ

Considerando a seguinte igualdade

$$P\left(q_{\alpha/2;n-1} < \frac{(n-1)S^2}{\sigma^2} < q_{1-\alpha/2;n-1}\right) =$$
$$P\left(\sqrt{q_{\alpha/2;n-1}} < \sqrt{\frac{(n-1)S^2}{\sigma^2}} < \sqrt{q_{1-\alpha/2;n-1}}\right)$$

Assim I.C. com nível de confiança γ para σ é dado por

$$\left[\sqrt{\frac{(n-1)S^2}{q_{1-\alpha/2;n-1}}}; \sqrt{\frac{(n-1)S^2}{q_{\alpha/2;n-1}}} \right]$$

Exemplo

Estudo: Dureza média de folhas-de-flandres

Amostra: 50 observações

Estimativa da variância (S^2): 0,37

$$q_{\alpha/2;n-1} = 32, 36, \quad q_{1-\alpha/2;n-1} = 71, 42.$$

Portanto, o intervalo de confiança para a variância da dureza de folhas-de-flandres é

$$\left(\frac{49 \times 0,37}{71,42}; \frac{49 \times 0,37}{32,36} \right) \\ [0,25; 0,56]$$

Estimação vs Testes de Hipóteses

ESTIMAÇÃO

- Qual é a proporção de desempregados em Salvador?
- Qual é a média da renda salarial no Estado da Bahia?
- Qual é a proporção de alunos que praticam esportes na UFBA?

TESTES DE HIPÓTESES

- A proporção de desempregados em Salvador é maior que 20%?
- A média da renda salarial na Bahia é maior que R\$1.500?
- A proporção de alunos que praticam esportes na UFBA é maior que 50%?

Definições

Hipótese estatística

É uma afirmação sobre os parâmetros de uma ou mais populações.

Exemplo:

- A proporção de peças defeituosas num lote é igual a 20%.

Teste de hipóteses

É o procedimento construído para rejeitar ou não uma hipótese estatística a partir da evidência obtida dos dados.

Exemplo 1

Suponha que o interesse é lançar um novo modelo de lâmpada.

Afirmção:

- O tempo de vida médio da nova lâmpada é igual ao tempo de vida médio da lâmpada produzida atualmente.

Suposição:

- $T_n \sim \text{Exp}(\lambda_n)$, $T_a \sim \text{Exp}(\lambda_a)$

Hipóteses estatística:

- | | | |
|------------------------------------|----|------------------------------------|
| • $H_0 : \lambda_n = \lambda_a$ | ou | • $H_0 : \lambda_n \leq \lambda_a$ |
| • $H_1 : \lambda_n \neq \lambda_a$ | | • $H_1 : \lambda_n > \lambda_a$ |

Exemplo 2

Suponha que o interesse é lançar um novo remédio.

Afirmção:

- A proporção de pessoas curadas pelo novo remédio é igual à proporção de pessoas curadas pelo remédio vendido atualmente.

Suposição:

- $N_n \sim \text{Bin}(n, p_n)$, $N_a \sim \text{Bin}(n, p_a)$

Hipóteses estatística:

- | | | |
|------------------------|----|------------------------|
| • $H_0 : p_n = p_a$ | ou | • $H_0 : p_n \leq p_a$ |
| • $H_1 : p_n \neq p_a$ | | • $H_1 : p_n > p_a$ |

Exemplo 3

Suponha que o interesse é verificar a eficácia de um tratamento.

Afirmção:

- A resposta média das pessoas que recebem ao novo tratamento é igual a das pessoas que não recebem tratamento

Suposição:

- $R_t \sim N(\mu_t, \sigma^2)$, $R_p \sim N(\mu_p, \sigma^2)$

Hipóteses estatística:

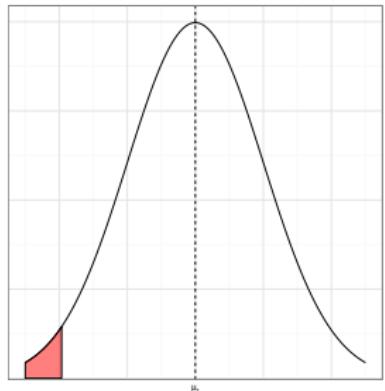
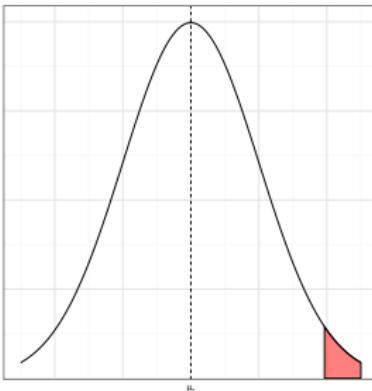
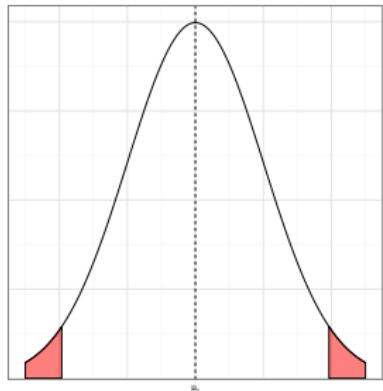
- | | | |
|--|----|---|
| <ul style="list-style-type: none">• $H_0 : \mu_t = \mu_p$• $H_1 : \mu_t \neq \mu_p$ | ou | <ul style="list-style-type: none">• $H_0 : \mu_t \leq \mu_p$• $H_1 : \mu_t > \mu_p$ |
|--|----|---|

Definições

- H_0 é chamada de hipótese nula.
- H_1 é chamada de hipótese alternativa.

Tipos de hipóteses:

- | | | | | |
|----------------------------|----|-------------------------|----|-------------------------|
| ◊ $H_0 : \mu_t = \mu_p$ | ou | ◊ $H_0 : \mu_t = \mu_p$ | ou | ◊ $H_0 : \mu_t = \mu_p$ |
| ◊ $H_1 : \mu_t \neq \mu_p$ | | ◊ $H_1 : \mu_t > \mu_p$ | | ◊ $H_1 : \mu_t < \mu_p$ |



Tipos de erros

O teste de hipótese é construído de forma a tomar uma decisão:

- Rejeitar H_0 ou
- Não rejeitar H_0 .

Podemos cometer dois tipos de erro:

- **Erro tipo I:** Rejeitar H_0 quando H_0 é verdadeira.
- **Erro tipo II:** Não rejeitar H_0 quando H_0 é falsa.

No exemplo 1:

- Erro tipo I: dizer que a nova lâmpada é superior quando na verdade tem o mesmo rendimento que a lâmpada anterior.
- Erro tipo II: dizer que a lâmpada tem rendimento equivalente ao da lampada anterior quando na verdade ela é superior.

Analogias

Considere as seguintes hipóteses num julgamento:

- A pessoa acusada de um crime é **inocente**.
- A pessoa acusada de um crime é **culpada**.

Durante o julgamento dois erros podem ocorrer:

- O juiz pode declarar culpada uma pessoa inocente (**Tipo I**)
- O juiz pode declarar inocente uma pessoa culpada (**Tipo II**)

Pergunta:

- Qual dos erros é mais grave?
- Qual deve ser minimizado ou controlado?

Analogias 2

Considere as seguintes hipóteses durante a confecção de um novo medicamento:

- O novo medicamento é **equivalente** ao medicamento atual.
- O novo medicamento é **superior** ao medicamento atual.

Durante a avaliação da eficácia do novo remédio dois erros podem ocorrer:

- A equipe responsável considerar o novo remédio superior quando na verdade é equivalente (**Tipo I**)
- A equipe responsável considerar o novo remédio equivalente quando na verdade é superior (**Tipo II**)

Pergunta:

- Qual dos erros é mais grave?
- Qual deve ser minimizado ou controlado?

Definição

Nível de significância (α)

É a probabilidade de se cometer um erro do tipo I.

Poder do teste ($1 - \beta$)

É a probabilidade de rejeitar a hipótese nula quando ela é falsa. É a probabilidade complementar do erro do tipo II.

- Não é possível controlar os dois erros ao mesmo tempo.
- Fixa-se um nível de significância de interesse e constrói-se um teste de hipótese a partir disso.
- Valores mais utilizados: $\alpha = \{0, 01; 0, 05; 0, 10\}$

Tipos de erros

		Decisão	
		Rejeitar H_0	Não rejeitar H_0
H_0 é verdadeira	H_0 é verdadeira	Erro do Tipo I	Decisão correta
	H_0 é falsa	Decisão correta	Erro do Tipo II

Nível de significância:

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

Poder do teste:

$$1 - \beta = 1 - P(\text{Erro do tipo II}) = P(\text{Rejeitar } H_0 | H_0 \text{ é falsa})$$

Exemplo

Suposição inicial:

$$X \sim N(\mu, \sigma^2), \quad \sigma^2 \text{ conhecido}$$

Hipóteses:

- $H_0 : \mu = \mu_0$
- $H_1 : \mu \neq \mu_0$

Dada uma amostra, sabemos que

$$\bar{X} \sim N(\mu, \sigma^2/n) \Rightarrow \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

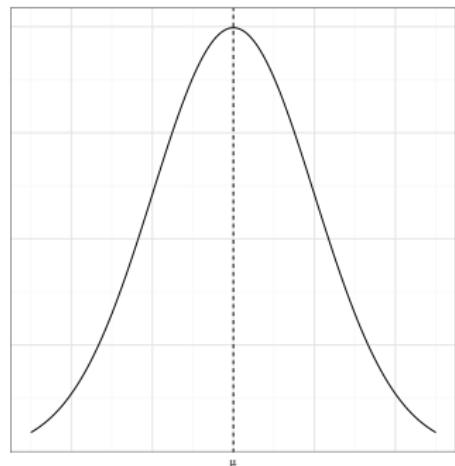
Exemplo (cont.)

Se H_0 é verdadeira (sob H_0):

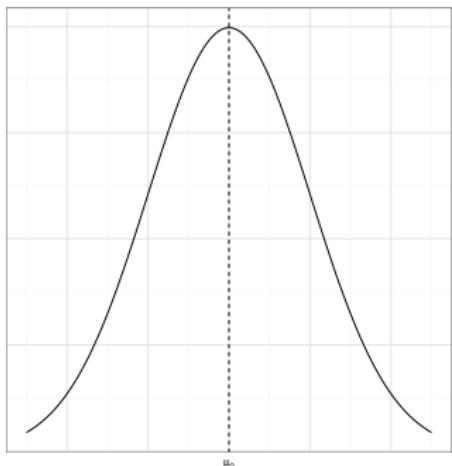
$$\bar{X} \sim N(\mu_0, \sigma^2/n) \Rightarrow \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0, 1)$$

Pergunta:

- Para quais valores de \bar{X} devemos rejeitar H_0 ?



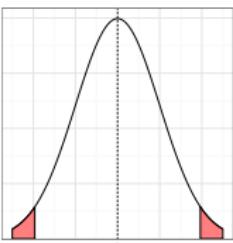
Sob
 $H_0 \Rightarrow$



Exemplo (cont.)

Decisão: Rejeitar H_0 para grandes valores absolutos de

$$\frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$$



Estatística de teste

Estimador com distribuição de probabilidade conhecida que pode ser utilizado para tomada de decisão em um teste de hipótese.

Região crítica

Região de valores do espaço amostral para os quais a hipótese nula é rejeitada

Exemplo (cont.)

Considerando o nível de significância igual a 5% ($\alpha = 0,05$), temos

$$\begin{aligned}\alpha &= P(\text{Erro do tipo I}) \\&= P(\text{Rejeitar } H_0 | H_0 \text{ é verdadeira}) \\&= P\left(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} < -z_c | H_0 \text{ é verdadeira}\right) + \\&\quad P\left(z_c < \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} | H_0 \text{ é verdadeira}\right)\end{aligned}$$

Mas se H_0 é verdadeira, então

$$T = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0, 1)$$

Logo $z_c = 1,96$. Portanto, a RC do teste é dada por $|T| > 1,96$.

Construção de um teste de hipótese

- 1 Definir as hipóteses a serem testadas.
- 2 Usar a teoria estatística para decidir sobre um estimador que pode ser usado para testar uma estatística.
- 3 Fixar a probabilidade α de se cometer um erro e usar esse valor para construir a região crítica do teste.
- 4 Usar as observações para calcular o valor da estatística do teste.
- 5 Tomar a decisão sobre rejeitar ou não a hipótese nula, de acordo com o valor obtido da estatística de teste.

Região crítica

- Para rejeitar uma hipótese:
 - ❖ Necessário construir uma região crítica (RC).
- Se a estatística de teste $\in \text{RC}$, então rejeita-se a hipótese nula.

Por exemplo, para $\alpha = 0,05 = 5\%$, rejeita-se $H_0 : \mu = \mu_0$ sempre que

$$T = \left| \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} \right| > 1,96,$$

em que $X \sim N(\mu, \sigma^2)$, σ^2 conhecido.

Nível descritivo (valor-p, p-valor, p-value)

Ao invés disso, poderíamos ter calculado o seguinte

$$P\left(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} > T\right)$$

⇒ Menor nível de significância para o qual a hipótese poderia ser rejeitada.

- Essa probabilidade é chamada de nível descritivo.

Regra de decisão - Nível de significância e valor-p

- $\text{valor-p} < \alpha \Rightarrow$ rejeitar H_0
- $\text{valor-p} > \alpha \Rightarrow$ não rejeitar H_0

Exercício 41 - Apostila 2, pg 76

- Pesquisador fez vários testes estatísticos:
 - ❖ Um para cada hipótese do trabalho.
- Foi adotado o nível de significância de 5%.
- Responder às seguintes perguntas:
 - ❖ $\text{valor-p} = 0,0001$. Qual deve ser a conclusão? Qual é o risco de tomar uma decisão incorreta?
 - ❖ $\text{valor-p} = 0,25$. Qual deve ser a conclusão? Qual é o risco de tomar uma decisão incorreta?
 - ❖ Em outros dois testes, os p-valores foram de 0,0001 e 0,01, respectivamente. Em qual dos testes o pesquisador deve estar mais convicto na decisão de qual hipótese deve ser escolhida? Por quê?

Teste de hipóteses para a média populacional

- Média populacional é de grande interesse, em geral.
- $E(X) = \mu$.
- Caso particular, $X \sim N(\mu, \sigma^2)$
- Teorema Limite Central mostra que, para grandes amostras

$$\bar{X} \sim N(\mu, \sigma^2/n)$$

- É importante obter testes de hipóteses para esse parâmetro.

Considerações iniciais

Tipos de hipóteses:

- Hipóteses unilaterais
 - ◊ $H_0 : \mu \leq \mu_0$ (ou $\mu = \mu_0$) contra $\mu > \mu_0$
 - ◊ $H_0 : \mu \geq \mu_0$ (ou $\mu = \mu_0$) contra $\mu < \mu_0$
- Hipótese bilateral
 - ◊ $H_0 : \mu = \mu_0$ contra $\mu \neq \mu_0$

Resultado, quando $X \sim N(\mu, \sigma^2)$, mas σ^2 é desconhecido:

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

Regra de decisão 1

É possível mostrar que o teste de hipóteses para a seguinte hipótese,

- $H_0 : \mu \leq \mu_0$ (ou $\mu = \mu_0$)
- $H_1 : \mu > \mu_0$

rejeita a hipótese nula (H_0) quando

$$\frac{\bar{X} - \mu_0}{S/\sqrt{n}} > t_{(1-\alpha;n-1)},$$

em que $t_{1-\alpha;n-1}$ é o quantil de ordem $1 - \alpha$ de

$$W \sim t_{n-1}$$

Regra de decisão 2

É possível mostrar que o teste de hipóteses para a seguinte hipótese,

- $H_0 : \mu \geq \mu_0$ (ou $\mu = \mu_0$)
- $H_1 : \mu < \mu_0$

rejeita a hipótese nula (H_0) quando

$$\frac{\bar{X} - \mu_0}{S/\sqrt{n}} < t_{(\alpha;n-1)},$$

em que $t_{\alpha;n-1}$ é o quantil de ordem α de

$$W \sim t_{n-1}$$

Regra de decisão 3

É possível mostrar que o teste de hipóteses para a seguinte hipótese,

- $H_0 : \mu = \mu_0$
- $H_1 : \mu \neq \mu_0$

rejeita a hipótese nula (H_0) quando

$$\left| \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \right| > t_{(1-\alpha/2;n-1)},$$

em que $t_{1-\alpha/2;n-1}$ é o quantil de ordem $1 - \alpha/2$ de

$$W \sim t_{n-1}$$

Exemplo

- Tempo médio para executar uma tarefa = 100 minutos.
- Introduziu uma modificação para diminuir esse tempo.
- Amostra de 16 operários.

$$\bar{X} = 85, \quad S = 12,$$

- Houve melhora?
- $H_0 : \mu = 100$ contra $H_1 : \mu < 100$.

Exemplo - Solução

Consideremos o nível de significância $\alpha = 0,05$.

Na tabela da distribuição t-Student, temos que $t_{0,05;15} = -1,753$.

A **região crítica** é dada por

$$\text{Rejeitar } H_0 \text{ se } T = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} < t_{0,05;15}$$

O valor da estatística de teste $T = -5$. Portanto, devemos **rejeitar** a hipótese que o processo se manteve igual.

Podemos calcular o valor-p também nesse exemplo

$$\text{valor-p} = P(T < -5 | H_0) \approx 0.0001$$

Teste de hipóteses para grandes amostras

Intervalos de confiança:

- Aproximação normal

Testes de hipóteses:

- Aproximação normal

Caso geral:

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

Grandes amostras:

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim N(0, 1)$$

Regra de decisão 1 - $n \rightarrow \infty$

É possível mostrar que o teste de hipóteses para a seguinte hipótese,

- $H_0 : \mu \leq \mu_0$ (ou $\mu = \mu_0$)
- $H_1 : \mu > \mu_0$

rejeita a hipótese nula (H_0) quando

$$\frac{\bar{X} - \mu_0}{S/\sqrt{n}} > z_{1-\alpha},$$

em que $z_{1-\alpha}$ é o quantil de ordem $1 - \alpha$ de

$$Z \sim N(0, 1)$$

Regra de decisão 2 - $n \rightarrow \infty$

É possível mostrar que o teste de hipóteses para a seguinte hipótese,

- $H_0 : \mu \geq \mu_0$ (ou $\mu = \mu_0$)
- $H_1: \mu < \mu_0$

rejeita a hipótese nula (H_0) quando

$$\frac{\bar{X} - \mu_0}{S/\sqrt{n}} < z_\alpha,$$

em que z_α é o quantil de ordem α de

$$Z \sim N(0, 1)$$

Regra de decisão 3 - $n \rightarrow \infty$

É possível mostrar que o teste de hipóteses para a seguinte hipótese,

- $H_0 : \mu = \mu_0$
- $H_1 : \mu \neq \mu_0$

rejeita a hipótese nula (H_0) quando

$$\left| \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \right| > z_{1-\alpha/2},$$

em que $z_{1-\alpha/2}$ é o quantil de ordem $1 - \alpha/2$ de

$$Z \sim N(0, 1)$$

Exemplo 13.3 - pg 62

Rede de pizzarias:

- Teor de gordura em peças de salame = 15%?

Ao nível de 5% de significância,

$$H_0 : \mu = 15 \text{ contra } H_1 : \mu \neq 15$$

As seguintes estatísticas descritivas foram obtidas:

Teor de Gordura	
Média	14,894
Desvio padrão	6,3871

Exemplo - Solução

- $n = 50$.
- $\alpha = 0,05 \rightarrow z_{1-\alpha/2} = 1,96$.

A decisão do teste bilateral nesse caso define:

$$\text{Rejeitar } H_0 \text{ se } T = \left| \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \right| > 1,96$$

Substituindo os valores, obtemos

$$T = \left| \frac{14,894 - 15}{6,3871/\sqrt{50}} \right| = |-0,1174|$$

Logo, não temos evidências para rejeitar a hipótese nula.

Exemplo 13.3 - pg 63

Dúvida da equipe técnica da indústria siderúrgica:

- Processo de cozimento abaixo do valor nominal de especificação?

Considere o seguinte teste de hipótese:

$$H_0 : \mu \geq 61 \text{ contra } H_1 : \mu < 61$$

As seguintes estatísticas descritivas foram obtidas:

Teor de Gordura	
Média	60,212
Desvio padrão	0,611

Exemplo - Solução

- $n = 50$.
- $\alpha = 0,05 \rightarrow z_{\alpha/2} = -1,96$.

A decisão do teste bilateral nesse caso define:

$$\text{Rejeitar } H_0 \text{ se } T = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} < -1,96$$

Substituindo os valores, obtemos

$$T = \frac{60,212 - 61}{0,611/\sqrt{50}} = -9,12$$

Logo, rejeitamos a hipótese nula.

Teste de hipóteses para a proporção populacional

- Diversas informações são apresentadas como 0's e 1's.
- Por esse motivo, a proporção populacional é de grande interesse.
- Se $X_i \sim \text{Ber}(p) \rightarrow E(\bar{X}) = p$.
- Para grandes amostras, a seguinte aproximação é válida

$$\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$$

- $\hat{p} = \frac{X}{n}$

Exemplos

- Testar se a proporção de itens defeituosos é maior que 10%.
- Testar se a proporção de veículos vendidos por uma certa fabricante é maior que 20%.
- Testar se a proporção de votos de um certo candidato é maior que 50%.
- Testar se a proporção de alunos na UFBA que pratica esportes é maior que 50%.

Considerações iniciais

Tipos de hipóteses:

- Teste de hipóteses unilateral
 - ◊ $H_0 : p \leq p_0$ (ou $p = p_0$) contra $p > p_0$
 - ◊ $H_0 : p \geq p_0$ (ou $p = p_0$) contra $p < p_0$
- Teste de hipóteses bilateral
 - ◊ $H_0 : p = p_0$ contra $p \neq p_0$

Quando $n \rightarrow \infty$, se $X_i \sim \text{Ber}(p)$

$$\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0, 1)$$

Regra de decisão 1 - Proporção populacional

É possível mostrar que o teste de hipóteses para a seguinte hipótese,

- $H_0 : p \leq p_0$ (ou $p = p_0$)
- $H_1: p > p_0$

rejeita a hipótese nula (H_0) quando

$$\frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} > z_{1-\alpha},$$

em que $z_{1-\alpha}$ é o quantil de ordem $1 - \alpha$ de

$$Z \sim N(0, 1)$$

e α é o nível de significância do teste.

Regra de decisão 2 - Proporção populacional

É possível mostrar que o teste de hipóteses para a seguinte hipótese,

- $H_0 : p \geq p_0$ (ou $p = p_0$)
- $H_1: p < p_0$

rejeita a hipótese nula (H_0) quando

$$\frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} < z_\alpha,$$

em que z_α é o quantil de ordem α de

$$Z \sim N(0, 1)$$

e α é o nível de significância do teste.

Regra de decisão 3 - Proporção populacional

É possível mostrar que o teste de hipóteses para a seguinte hipótese,

- $H_0 : p = p_0$
- $H_1 : p \neq p_0$

rejeita a hipótese nula (H_0) quando

$$\left| \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \right| > z_{1-\alpha/2},$$

em que $z_{1-\alpha/2}$ é o quantil de ordem $1 - \alpha/2$ de

$$Z \sim N(0, 1)$$

e α é o nível de significância do teste.

Exemplo 13.5 - pg 64

- Fábrica de automóveis A afirma que detém 60% do mercado.
- Fábrica concorrente gostaria de confirmar tal hipótese.
- Para isso, decide realizar uma pesquisa por amostragem com 300 proprietários de veículos.
- Hipóteses de interesse:
 - ◊ $H_0 : p = 0,60$
 - ◊ $H_1 : p < 0,60$
- p é a proporção de consumidores que possuem carros da fábrica A.

Exemplo - Solução

- Vamos fixar $\alpha = 0,05$.
- Como a amostra é grande podemos utilizar a aproximação normal.
- Suponha que o resultado apontou para 165 proprietários de carros da fábrica A.
- A proporção amostral (\hat{p}) é igual

$$\hat{p} = \frac{165}{300} = 0,55.$$

Exemplo - Solução (cont.)

- A região crítica do teste é dada por

$$\frac{\hat{p} - p_0}{\sqrt{\frac{p(1-p)}{n}}} < z_\alpha$$

Para $\alpha = 0,05 \rightarrow z_\alpha = -1,645$.

- Substituindo os valores, temos

$$\frac{0,55 - 0,60}{\sqrt{\frac{0,60 \times 0,40}{300}}} \approx -1,77 < -1,645$$

- Logo, rejeitamos $H_0 : p = 0,60$ e concluímos que a proporção de veículos da fabricante A deve ser menor que 0,60.

Teste de hipóteses para a variância populacional

- Amostra de tamanho n de uma população $N(\mu, \sigma^2)$.
- O interesse é testar uma hipótese sobre σ^2 .
- Estimador para σ^2

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2,$$

- Sabemos que

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

Hipóteses do tipo bilateral

- Considere que estamos interessados na seguinte hipótese:
 - ◊ $H_0 : \sigma^2 = \sigma_0^2$
 - ◊ $H_1 : \sigma^2 \neq \sigma_0^2$
- A estatística de teste a ser usada é

$$T_{\sigma^2} = \frac{(n-1)S^2}{\sigma_0^2}$$

- Sob H_0 , $T_{\sigma^2} \sim \chi_{n-1}^2$
- E devemos rejeitar H_0 se

$$T_{\sigma^2} < \chi_{\alpha/2;n-1} \text{ ou } T_{\sigma^2} > \chi_{1-\alpha/2;n-1}$$

em que $\chi_{\gamma;n-1}$ é o quantil de ordem γ de χ_{n-1} .

Hipóteses do tipo unilateral - 1

- Considere que estamos interessados na seguinte hipótese:
 - ◊ $H_0 : \sigma^2 = \sigma_0^2$
 - ◊ $H_1 : \sigma^2 > \sigma_0^2$
- A estatística de teste a ser usada é

$$T_{\sigma^2} = \frac{(n-1)S^2}{\sigma_0^2}$$

- Sob H_0 , $T_{\sigma^2} \sim \chi_{n-1}^2$
- E devemos rejeitar H_0 se

$$T_{\sigma^2} > \chi_{1-\alpha;n-1}$$

em que $\chi_{1-\alpha;n-1}$ é o quantil de ordem $1-\alpha$ de χ_{n-1} .

Hipóteses do tipo unilateral - 2

- Considere que estamos interessados na seguinte hipótese:
 - ◊ $H_0 : \sigma^2 = \sigma_0^2$
 - ◊ $H_1 : \sigma^2 < \sigma_0^2$
- A estatística de teste a ser usada é

$$T_{\sigma^2} = \frac{(n-1)S^2}{\sigma_0^2}$$

- Sob H_0 , $T_{\sigma^2} \sim \chi_{n-1}^2$
- E devemos rejeitar H_0 se

$$T_{\sigma^2} < \chi_{\alpha;n-1}$$

em que $\chi_{\alpha;n-1}$ é o quantil de ordem α de χ_{n-1} .

Exemplo 13.8 - pg 66

- Uma linha de montagem produz peças com pesos P_i , em que
 - ◊ $P_i \sim N(\mu, 30g^2)$
- Equipamentos foram modernizados.
- Tomou-se uma amostra de 23 peças
 - ◊ $s^2 = 40g^2$
- Existem evidências para afirmar que a variância mudou, considerando $\alpha = 5\%$?

Exemplo - Solução

As hipóteses a serem testadas são:

- $H_0 : \sigma^2 = 30$ contra $H_1 : \sigma^2 \neq 30$

Temos que

$$T_{\sigma^2} = \frac{22 \times 40}{30} = 29,33.$$

Os valores críticos para esse caso usando $\alpha = 0,05$

- $\chi^2_{\alpha/2;n-1} = \chi^2_{0,025;22} = 10,98.$
- $\chi^2_{1-\alpha/2;n-1} = \chi^2_{0,975;22} = 36,78.$

$$T_{\sigma^2} \notin (0; 10,98] \cup [36,78; \infty)$$

Portanto, não existem evidências de que a variância seja diferente de 30.