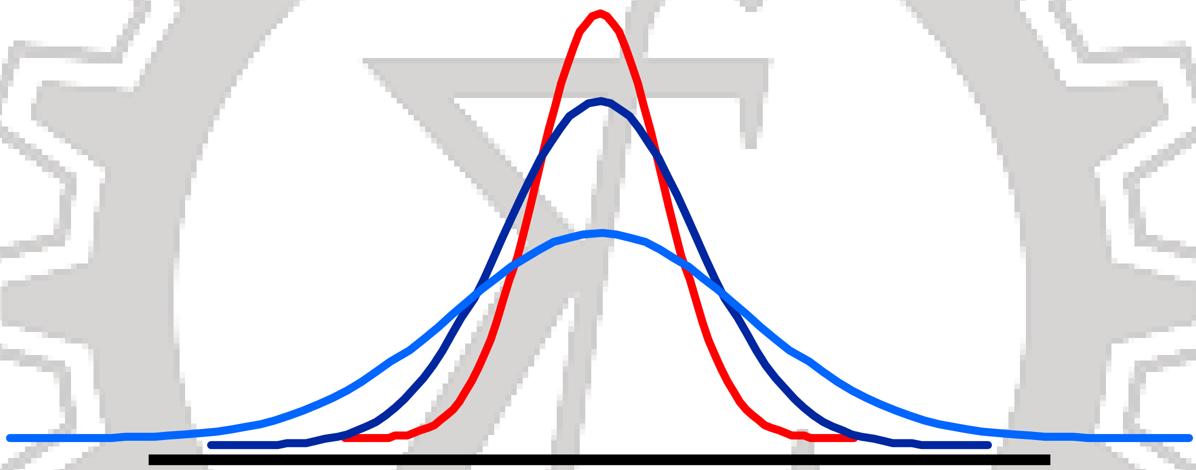


Estatística Indutiva

Gilbert Queiroz dos Santos

Rio de Janeiro - RJ
2016

SUMÁRIO

1. Estatística Indutiva

1.1 Introdução

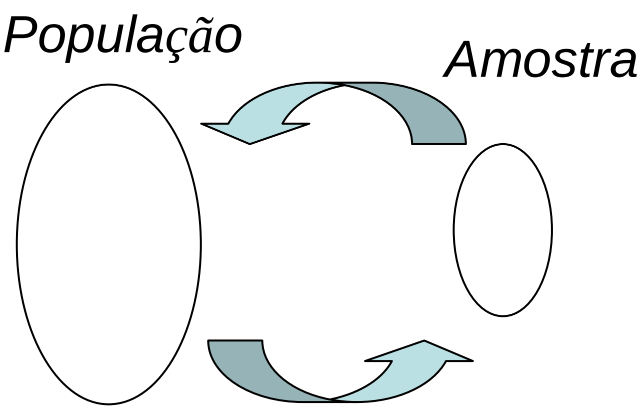
Estatística Indutiva (ou Estatística Inferencial ou Inferência Estatística, ou ainda Indução Estatística), cuida da análise e interpretação dos dados experimentais.

Dois conceitos fundamentais devem ser apresentados:

1. *população* (ou universo): conjunto de elementos com pelo menos uma característica comum;
2. *amostra*: subconjunto da população, necessariamente finito, pois todos seu elementos serão examinados para efeito da realização do estudo estatístico desejado.

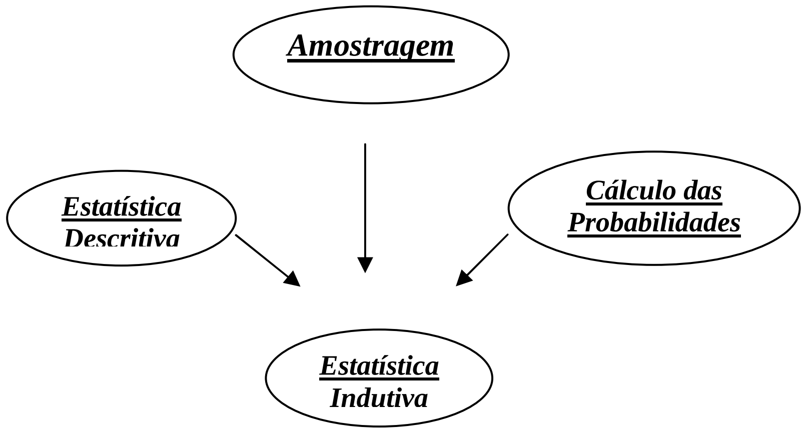
O objetivo da Estatística Indutiva é o de tirar conclusões sobre as populações com base nos resultados observados em amostras extraídas dessas populações.

Este termo "indutiva" decorre da existência de um processo de indução, isto é, um processo de raciocínio em que, partindo-se do conhecimento de uma parte, procura-se tirar conclusões sobre a realidade, no todo.
Este processo de indução não pode ser exato, pois ao induzir, estamos sempre sujeitos a erro. A Estatística Indutiva irá nos dizer até que ponto poderemos estar errado em nossas induções, e com que probabilidade.



Antes de iniciar qualquer análise dos dados através dos métodos da Estatística Indutiva, é preciso organizar os dados da amostra, o que é feito com técnicas de Estatística Descritiva.
Uma outra ferramenta utilizada em Estatística Indutiva, e que surge paralelamente, é a amostragem, onde certos cuidados básicos devem ser tomados no processo de obtenção das amostras.

Em resumo, um estudo estatístico completo que recorra às técnicas da Estatística Indutiva irá envolver também, direta e indiretamente, tópicos de:
- Estatística Descritiva;
- Cálculo das Probabilidade;
- Amostragem.



1.2 Amostragem

É o processo pelo qual obtêm-se amostras, que contenham informações a respeito de valores populacionais desconhecidos.

A amostra ou amostras selecionadas devem ser representativas da população. Isto significa que, a menos de certas pequenas discrepâncias inerentes à aleatoriedade sempre presente, em maior ou menor grau, no processo de amostragem, a amostra deve possuir as mesmas características básicas da população, no que diz respeito à(s) variável(is) que desejamos pesquisar.

Tipos de amostragem

Existem dois tipos de amostragem: a probabilística e a não-probabilística.
Amostragem probabilística
Neste tipo, todos os elementos da população possuem probabilidade conhecida e não nula de pertencer a amostra.
É a melhor recomendação que se deve fazer no sentido de se garantir a representatividade da mostra, pois o acaso será o único responsável por eventuais discrepâncias entre população e amostra, o que é levado em consideração pelos métodos de análise Estatística Indutiva.
As principais técnicas de amostragem probabilística são:
-amostragem casual simples;
-amostragem estratificada;
-amostragem por conglomerados

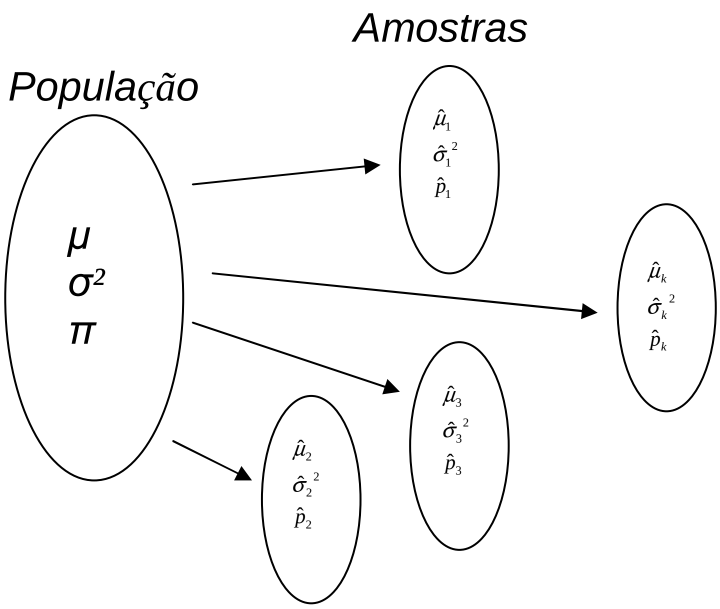
Amostragem não-probabilística
É um processo de amostragem subjetivo e seu rendimento depende do conhecimento que possui o pesquisador a respeito da estrutura das populações e a mostra é uma parcela proporcional desta estrutura.
Ela é empregada, muitas vezes, por simplicidade ou pela impossibilidade de se obter uma amostragem probabilística.

As principais técnicas de amostragem não-probabilística são:

- amostragem a esmo;
- amostragens intencionais.

1.3 Distribuição amostral das Estatísticas

Seja uma população de tamanho N com média μ, variância σ² e proporção π. Ao retirarmos várias amostras desta população, teremos:



1.3.1 Distribuição amostral da média

Seja uma população de tamanho N e X uma variável aleatória dessa população com E[X] = μ e Var[X] = σ², logo X~N(μ,σ²).
Seja uma amostra aleatória (X1, X2,...,Xn) retirada desta população, onde se tem:

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Podemos calcular $E[\overline{X}]$ e $Var[\overline{X}]$ da seguinte forma:

$$E[\overline{X}] = E\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) = \frac{1}{n} E\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{n} n\mu = \mu$$

$$Var[\overline{X}] = Var\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) = \frac{1}{n^2} Var\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var[X_i] = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

Então, $E[\overline{X}] = \mu$ e $Var[\overline{X}] = \dfrac{\sigma^2}{n}$

Com isto, podemos dizer que a média amostral é um estimador justo e consistente da média populacional

1.3.2 Distribuição amostral da variância

Seja uma população de tamanho N e X uma variável aleatória dessa população com E[X] = μ e Var[X] = σ^2, logo X~N(μ,σ^2).
Seja uma amostra aleatória (X1, X2,...,Xn) retirada desta população, onde se tem:

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

Podemos calcular $E\left[S^2\right]$ e $Var\left[S^2\right]$, mas primeiro devemos saber que:

$$\chi_n^2 = \sum z^2 \text{, com } \mathrm{E}\left(\chi_n^2\right) = n \text{ e } \mathrm{Var}\left(\chi_n^2\right) = 2n$$

É a Distribuição Qui-Quadrado

Fazendo

$$\frac{(n-1)S^2}{\sigma^2} = \frac{\sum\left(X_i - \overline{X}\right)^2}{\sigma^2}$$

$$\frac{(n-1)S^2}{\sigma^2} = \chi_{n-1}^2\text{, onde } \mathrm{E}\left(\chi_{n-1}^2\right) = n-1 \text{ e } \mathrm{Var}\left(\chi_{n-1}^2\right) = 2(n-1)$$

Então: $S^2 = \dfrac{\sigma^2}{n-1}\chi_{n-1}^2$

Logo:

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

$$Var(S^2) = Var\left(\frac{\sigma^2}{n-1}\chi_{n-1}^2\right) = \frac{\sigma^4}{(n-1)^2}Var(\chi_{n-1}^2) = \frac{\sigma^4}{(n-1)^2}\left[2(n-1)\right] = \frac{2\sigma^4}{n-1}$$

Com isto, podemos dizer que a variância amostral é um estimador justo e consistente da variância populacional

1.3.3 Distribuição amostral da proporção

Seja uma população de tamanho N e X uma variável aleatória dessa população com E[X] = π e Var[X] = π(1-π), logo X~Be(π, π(1-π),).
Seja uma amostra aleatória (X1, X2,...,Xn) retirada desta população, onde se tem:

$$p = \frac{n_S}{n}$$

Onde n_S é o número de sucessos nos n elementos

Podemos calcular $E[p]$ e $Var[p]$, da seguinte forma:

$$E[p] = E\left[\frac{n_S}{n}\right] = \frac{1}{n}E[n_S] = \frac{1}{n}E\left[\sum_{i=1}^{n}X_i\right] = \frac{1}{n}\sum_{i=1}^{n}E[X_i] = \frac{1}{n}n\pi = \pi$$

$$Var[p]=Var\left[\frac{n_S}{n}\right]=\frac{1}{n^2}Var[n_S]=\frac{1}{n^2}Var\left[\sum_{i=1}^{n}X_i\right]=\frac{1}{n^2}\sum_{i=1}^{n}Var[X_i]=\frac{1}{n^2}n[\pi(1-\pi)]=\frac{\pi(1-\pi)}{n}$$

Com isto, podemos dizer que a proporção amostral é um estimador justo e consistente da proporção populacional

Resumindo:

Estimador($\hat{\theta}$)	E($\hat{\theta}$)	Var($\hat{\theta}$)
$\overline{X}=\frac{\sum_{i=1}^{n}X_i}{n}$	μ	$\frac{\sigma^2}{n}$
$S^2=\frac{\sum_{i=1}^{n}(X_i-\overline{X})^2}{n-1}$	σ^2	$\frac{2\sigma^4}{n-1}$
$p=\frac{n_S}{n}$	π	$\frac{\pi(1-\pi)}{n}$

1.3.4 Erro-padrão

O desvio-padrão da distribuição amostral das estatísticas é freqüentemente denominado de erro-padrão da estatística.

Estimador($\hat{\theta}$)	Erro-padrão - EP($\hat{\theta}$)
$\overline{X}=\frac{\sum_{i=1}^{n}X_i}{n}$	$\frac{\sigma}{\sqrt{n}}$
$S^2=\frac{\sum_{i=1}^{n}(X_i-\overline{X})^2}{n-1}$	$\sigma^2\sqrt{\frac{2}{n-1}}$
$p=\frac{n_S}{n}$	$\sqrt{\frac{\pi(1-\pi)}{n}}$

A variância do estimador depende sempre dos parâmetros populacionais, que são, em geral, desconhecidos. Neste caso, pode-se substituí-lo pelo erro-padrão estimado, usando, neste caso, os valores obtidos pela amostra. Assim,

Estimador($\hat{\theta}$)	Erro-padrão - EP($\hat{\theta}$)	EP Estimado($\hat{\theta}$)
$\overline{X} = \dfrac{\sum_{i=1}^{n} X_i}{n}$	$\dfrac{\sigma}{\sqrt{n}}$	$\dfrac{S}{\sqrt{n}}$
$S^2 = \dfrac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n-1}$	$\sigma^2\sqrt{\dfrac{2}{n-1}}$	$S^2\sqrt{\dfrac{2}{n-1}}$
$p = \dfrac{n_S}{n}$	$\sqrt{\dfrac{\pi(1-\pi)}{n}}$	$\sqrt{\dfrac{p(1-p)}{n}}$

2 Intervalo de confiança

É o intervalo que, com probabilidade conhecida, deverá conter o valor real do parâmetro populacional.
A probabilidade, que é obtida por 1 – α, é chamada de nível de confiança do respectivo intervalo.
O valor α será a probabilidade de erro de estimação, isto é, a probabilidade de errarmos ao afirmar que o valor do parâmetro populacional está contido no intervalo de confiança, normalmente chamada de nível de significância.
A estrutura de um intervalo de confiança é dada por:

$$P(\hat{\theta} - \varepsilon \leq \theta \leq \hat{\theta} + \varepsilon) = 1 - \alpha$$

Onde: $\hat{\theta}$ – é o estimador; θ é o parâmetro a ser estimado; 1 – α é a probabilidade de o valor estar no intervalo; α é a probabilidade de erro; ε é o erro de estimação

Este intervalo pode ser reescrito desta forma:

$$P(\hat{\theta} - \varepsilon \leq \theta \leq \hat{\theta} + \varepsilon) = 1 - \alpha$$

$$P\left(\left|\hat{\theta} - \theta\right| \leq \varepsilon\right) = 1 - \alpha$$

Mas como determinar o valor ε?

Podemos dizer que num primeiro momento:

$$\varepsilon = \hat{\theta} - \theta$$

Ou seja, o erro ε é a diferença entre o estimador $\hat{\theta}$ e o parâmetro verdadeiro θ.

2.1 Intervalo de confiança para média

Caso I – o desvio-padrão populacional é conhecido

Sabendo que $\overline{X} \approx N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$, então $Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$. Aqui o estimador é $\overline{X}$ e o parâmetro verdadeiro é μ, logo $\varepsilon = \overline{X} - \mu$.

Disto, podemos reescrever Z da seguinte forma:

$$Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \rightarrow \varepsilon = Z\frac{\sigma}{\sqrt{n}}.$$

Substituindo em $P(\hat{\theta} - \varepsilon \leq \theta \leq \hat{\theta} + \varepsilon) = 1 - \alpha$, temos:

$$IC\left(\overline{X} - Z\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + Z\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

Aqui, Z é tabelado em função de α/2.

Exemplo: Seja uma amostra de 33 notas de estudantes de um determinado colégio, dada abaixo. Vamos calcular o intervalo de confiança para média das notas, supondo que o desvio-padrão da população de onde foi retirada a amostra seja igual a 7.

84, 93, 100, 86, 82, 86, 88, 94, 89, 94, 93, 83, 95, 86, 94, 87, 91, 96, 89, 79, 99, 98, 81, 80, 88, 100, 90, 100, 81, 98, 87, 95, 94

Temos então que calcular a média. Para isto, vamos montar a tabela de distribuição de freqüências, da seguinte forma;

	Class limits	PM	f	rf	rf(%)	cf	cf(%)
1	[75,80)	77.5	1	0.03030303	3.030303	1	3.030303
2	[80,85)	82.5	6	0.18181818	18.181818	7	21.212121
3	[85,90)	87.5	9	0.27272727	27.272727	16	48.484848
4	[90,95)	92.5	8	0.24242424	24.242424	24	72.727273
5	[95,100)	97.5	6	0.18181818	18.181818	30	90.909091
6	[100,105)	102.5	3	0.09090909	9.090909	33	100.000000

Sabendo que $\overline{X} = \frac{\sum_{i=1}^{n} X_i f_i}{n} = \frac{2992,50}{33} = 90,68$, o intervalo dado por:

$$IC\left(\overline{X} - Z\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + Z\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

Fica assim:

$$IC\left(90,68 - 1,96\frac{7}{\sqrt{33}} \leq \mu \leq 90,68 + 1,96\frac{7}{\sqrt{33}}\right) = 1 - 5\%$$

Resolvendo:

$$IC(88,29 \leq \mu \leq 93,07) = 95\%$$

Usando o R:

```
x.bar<-weighted.mean(PM, f) # calcula a média
x.bar # verifica o resultado
[1] 90.68182

n<-length(data) # tamanho da amostra
n # verifica o valor
[1] 33

dp.p<-7 # valor do desvio-padrão populacional

se = dp.p/sqrt(n) # calcula o valor do erro-padrão (se)
se # verifica o resultado
[1] 1.218544

IC.média<-x.bar + c(-qnorm(.975)*se, qnorm(.975)*se) # calcula o intervalo de
                                                        confiança
IC.média # verifica os resultados
[1] 88.29352     93.07012
```

Caso II – o desvio-padrão populacional é desconhecido

Inicialmente, devemos levar em consideração que $S^2 = \frac{\sigma^2}{n-1}\chi^2_{n-1}$ e que $t = \frac{Z}{\sqrt{\frac{\chi^2_n}{n}}}$ tem distribuição t-student com n graus de liberdade.

Mas, sabemos que $Z = \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$. Fazendo algumas operações, temos:

$$Z = \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\overline{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}} = \frac{\overline{X}-\mu}{\sqrt{\frac{S^2(n-1)}{n\chi^2_{n-1}}}} \rightarrow \frac{Z}{\sqrt{\frac{\chi^2_{n-1}}{n-1}}} = \frac{\overline{X}-\mu}{\sqrt{\frac{S^2}{n}}} \rightarrow t_{n-1} = \frac{\overline{X}-\mu}{\frac{S}{\sqrt{n}}} \rightarrow \varepsilon = t_{n-1}\frac{S}{\sqrt{n}}$$

Substituindo em $P(\hat{\theta}-\varepsilon \leq \theta \leq \hat{\theta}+\varepsilon) = 1-\alpha$, temos:

$$IC\left(\overline{X} - t_{n-1}\frac{S}{\sqrt{n}} \leq \mu \leq \overline{X} + t_{n-1}\frac{S}{\sqrt{n}}\right) = 1-\alpha$$

Aqui, t_{n-1} é tabelado em função de $\alpha/2$ e de n – 1 (graus de liberdade).

Exemplo: Seja uma amostra de 15 sacos de leite, produzidos por uma determinada usina de beneficiamento de leite, dada abaixo. Vamos calcular o intervalo de confiança para média da quantidade de leite, em ml, não sabendo o valor do desvio-padrão da população original.

341, 345, 338, 339, 340, 343, 341, 343, 341, 328, 343, 347, 337, 348 e 339

Inicialmente, temos que calcular a média e o desvio-padrão da amostra. Neste caso, não precisamos montar a tabela de distribuição de freqüências porque a quantidade de dados é pequena (n<25).

Sabendo que a média é dada por $\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$ e o desvio-padrão dado por $S = \sqrt{\frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n-1}}$, temos:

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{5113}{15} = 340{,}87$$

$$S = \sqrt{\frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n-1}} = \sqrt{\frac{\sum (X_i - 340{,}87)^2}{15-1}} = \sqrt{\frac{315{,}73}{14}} = 4{,}75$$

O intervalo dado por:

$$IC\left(\overline{X} - t_{n-1}\frac{S}{\sqrt{n}} \le \mu \le \overline{X} + t_{n-1}\frac{S}{\sqrt{n}}\right) = 1-\alpha$$

Fica assim:

$$IC\left(340{,}87 - 2{,}145\frac{4{,}75}{\sqrt{15}} \le \mu \le 340{,}87 + 2{,}145\frac{4{,}75}{\sqrt{15}}\right) = 1-5\%$$

Resolvendo:

$$IC\left(338{,}24 \le \mu \le 343{,}50\right) = 95\%$$

Usando o R:

```
x<-c(341, 345, 338, 339, 340, 343, 341, 343, 341, 328, 343, 347, 337, 348, 339)
n<-length(x) # tamanho da amostra
n # verifica os resultados
[1] 15
x.bar<-mean(x) # calcula a média
x.bar # verifica o resultado
[1] 340.8667

dp<-sd(x) # calcula o desvio-padrão
dp # verifica o resultado
[1] 4.748935

se = dp/sqrt(n) # calcula o erro-padrão
se # verifica o resultado
[1] 1.22617

IC.média<-x.bar + c(-qt(.975, n-1)*se,qt(.975, n-1)*se)  # calcula o intervalo de
                                                           confiança
IC.média # verifica o resultado
```

[1] 338.2368 343.4965

OBSERVAÇÂO: neste caso, mesmo não se conhecendo o desvio-padrão populacional, mas com uma amostra grande, podemos calcular o intervalo de confiança para média usando:

$$IC\left(\overline{X} - Z\frac{S}{\sqrt{n}} \leq \mu \leq \overline{X} + Z\frac{S}{\sqrt{n}}\right) = 1-\alpha$$

2.2 Intervalo de confiança para a variância

Neste caso, vamos trabalhar com a mesma relação:

$$S^2 = \frac{\sigma^2}{n-1}\chi^2_{n-1}$$

Fazendo uma pequena adaptação:

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

O intervalo fica da seguinte forma:

$$IC\left(\chi^2_1 \leq \chi^2 \leq \chi^2_2\right) = 1-\alpha$$

Com algumas adaptações:

$$IC\left(\frac{S^2(n-1)}{\chi^2_{Sup}} \leq \sigma^2 \leq \frac{S^2(n-1)}{\chi^2_{\inf}}\right) = 1-\alpha$$

Onde: $\chi^2_{Inf} = \chi^2_{\left(\alpha/2\right)}$ e $\chi^2_{Sup} = \chi^2_{\left(1-\alpha/2\right)}$

Ambos com n – 1 graus de liberdade

Exemplo: utilizando os dados do saco de leite, temos os seguintes valores:

341, 345, 338, 339, 340, 343, 341, 343, 341, 328, 343, 347, 337, 348 e 339

Calculando a variância usando: $S^2 = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n-1}$, temos:

$$S^2 = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n-1} = \frac{\sum(X_i - 340{,}87)^2}{15-1} = 22{,}55$$

O intervalo dado por:

$$IC\left(\frac{S^2(n-1)}{\chi^2_{Sup}} \leq \sigma^2 \leq \frac{S^2(n-1)}{\chi^2_{\inf}}\right) = 1-\alpha$$

Fica assim:

$$IC\left(\frac{22{,}55(15-1)}{26{,}119} \le \sigma^2 \le \frac{22{,}55(15-1)}{5{,}629}\right) = 95\%$$

Resolvendo:

$$IC(12{,}008 \le \sigma^2 \le 56{,}093) = 95\%$$

Usando o R, temos:

```
var.x<-var(x) # calcula a variância
var.x # verifica o resultado
[1] 22.55238

qui.i<-qchisq(0.025, n-1) # calcula o Qui-quadrado inferior
qui.i # verifica o resultado
[1] 5.628726

qui.s<-qchisq(0.975, n-1) # calcula o Qui-quadrado superior
qui.s # verifica o resultado

IC.Var<-c(var.x*(n-1)/qui.s, var.x*(n-1)/qui.i) # calcula o intervalo de confiança
IC.Var # verifica os resultados
[1] 12.1 56.1
```

2.3 Intervalo de confiança para o desvio-padrão

Com base no intervalo de confiança da variância, temos:

$$IC\left(S\sqrt{\frac{(n-1)}{\chi^2_{Sup}}} \le \sigma \le S\sqrt{\frac{(n-1)}{\chi^2_{Inf}}}\right) = 1-\alpha$$

Onde: $\chi^2_{Inf} = \chi^2_{\left(1-\alpha/2\right)}$ e $\chi^2_{Sup} = \chi^2_{\alpha/2}$

Ambos com n – 1 graus de liberdade

Exemplo: utilizando os dados do IC para variância, temos:

$$IC\left(\sqrt{\frac{22{,}55(15-1)}{26{,}119}} \le \sigma \le \sqrt{\frac{22{,}55(15-1)}{5{,}629}}\right) = 95\%$$

Resolvendo:

$$IC(3{,}48 \le \sigma \le 7{,}49) = 95\%$$

Usando o R, temos:

```
IC.dp<-c(sqrt(var.x*(n-1)/qui.s), sqrt(var.x*(n-1)/qui.i)) # calcula o intervalo de
                                                              confiança
IC.dp # verifica os resultados
[1] 3.48 7.49
```

2.4 Intervalo de confiança para a proporção

Neste caso, sabendo que $\hat{p} \approx N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$, temos:

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

Lembrando que $\varepsilon = \hat{p} - p$, então:

$$\varepsilon = Z\sqrt{\frac{p(1-p)}{n}}$$

Substituindo em $P(\hat{\theta} - \varepsilon \leq \theta \leq \hat{\theta} + \varepsilon) = 1 - \alpha$, temos:

$$IC\left(\hat{p} - Z\sqrt{\frac{p(1-p)}{n}} \leq p \leq \hat{p} + Z\sqrt{\frac{p(1-p)}{n}}\right) = 1 - \alpha$$

Aqui, Z é tabelado em função de $\alpha/2$.

Exemplo: utilizando os dados dos sacos de leite, vamos considerar os casos com menos de 340 ml, que são 5 unidades; então a proporção de sacos de leite com menos de 340 ml na amostra é:

$$p = \frac{NCF}{NTC} = \frac{5}{15} = 0{,}333$$

O intervalo dado por:

$$IC\left(\hat{p} - Z\sqrt{\frac{p(1-p)}{n}} \leq p \leq \hat{p} + Z\sqrt{\frac{p(1-p)}{n}}\right) = 1 - \alpha$$

Fica assim:

$$IC\left(0{,}333 - 1{,}96\sqrt{\frac{0{,}333(1-0{,}333)}{15}} \leq p \leq 0{,}333 + 1{,}96\sqrt{\frac{0{,}333(1-0{,}333)}{15}}\right) = 1 - 5\%$$

Resolvendo:

$$IC(0{,}095 \leq p \leq 0{,}572) = 95\%$$

Usando o R, temos:

```
x<-c(341, 345, 338, 339, 340, 343, 341, 343, 341, 328, 343, 347, 337, 348, 339)
                                                    # os valores dos sacos de leite

n<-length(x) # tamanho da amostra
n # verifica o resutlado
[1] 15
```

##o comando "for" abaixo junto com o comando "if" ajudam a selecionar os valores abaixo de 340##

```
for(i in 1:n){
 if(x[i]<340){x[i]<-1}else{x[i]<-0}
 }

x # verifica o resultado
[1] 0 0 1 1 0 0 0 0 0 1 0 0 1 0 1 # cada valor "1" indica um valor menor do que
                                    340

p<-mean(x) # calcula a proporção
p # verifica o resultado
[1] 0.333

IC.p<-c(p-z*sqrt((p)*(1-p)/n), p+z*sqrt((p)*(1-p)/n)) # calcula o intervalo
IC.p # verifica o resultado
[1] 0.0948  0.5719
```

3 Testes de Hipótese

Seja H_0 a hipótese existente a ser testada e H_1 a hipótese alternativa.
O teste irá levar a rejeição ou a não rejeição da hipótese H_0, o que corresponde, respectivamente, à negação ou afirmação de H_1.
Em um teste de hipótese, podem ocorrer dois tipos de erros:
Erro tipo I: rejeitar H_0, sendo H_0 verdadeira.
Erro tipo II: aceitar H_0, sendo H_0 falsa.
As probabilidades destes dois tipos de erros serão designadas, respectivamente, por α e β.
A probabilidade α do erro tipo I é denominada nível de significância do teste.
Deve-se notar que as probabilidades α e β são probabilidades condicionadas à realidade.

As faixas de valores da variável de teste que leva à rejeição de H_0 é denominada região crítica (RC). A faixa restante constitui a região de aceitação (RA), ou não rejeição.
Um resultado experimental obtido pode ser ou não significante, dependendo do α fixado. Um resultado significativo a um determinado nível α nos levará à rejeição da hipótese H_0, pois admitiremos que, a menos de um risco pré-fixado α, ele seja incompatível com a hipótese H_0.
Por outro lado, se o valor experimental da variável de teste cair na região de aceitação, não terá havido, no nível α considerado, evidência significativa suficiente para a rejeição da hipótese H_0, a qual deverá ser aceita. Note-se que neste caso, estaríamos sujeitos a cometer o erro tipo II, cuja a probabilidade é β.

Se providências não tiverem sido tomadas no sentido de controlar a probabilidade β do erro tipo II, então a aceitação da hipótese H_0 será acompanhada de uma avaliação probabilística da possibilidade do erro, conforme sempre ocorre no caso de chegar-se à rejeição de H_0 (pois o nível de significância α será sempre pré-fixado). A aceitação de H_0 corresponde à insuficiência da evidência experimental, no nível de significância desejado, para chegar à sua rejeição. Essa aceitação, como o próprio termo sugere, não deve ser entendida como uma afirmação de H_0.

3.1 Poder do teste

É a capacidade do teste em rejeitar H_0, sendo H_0 falsa, logo o valor de p será dado por $1 - \beta$.

Os estatísticos aplicados dão cada vez mais preferência ao poder do teste p, em lugar aos testes clássicos, porque um teste clássico envolve a fixação arbitrária de α (usualmente em 5%). Ao invés de introduzir tal elemento arbitrário, muitas vezes é preferível indicar o poder do teste p, deixando-se a tarefa de formular o julgamento sobre H_0. (Formalmente, determinado o nível de α que se julgue adequado aos seus propósitos, pode-se chegar a uma decisão individual).
O poder está relacionado com a natureza do teste escolhido e , de modo geral, o poder aumenta com o tamanho n da amostra.

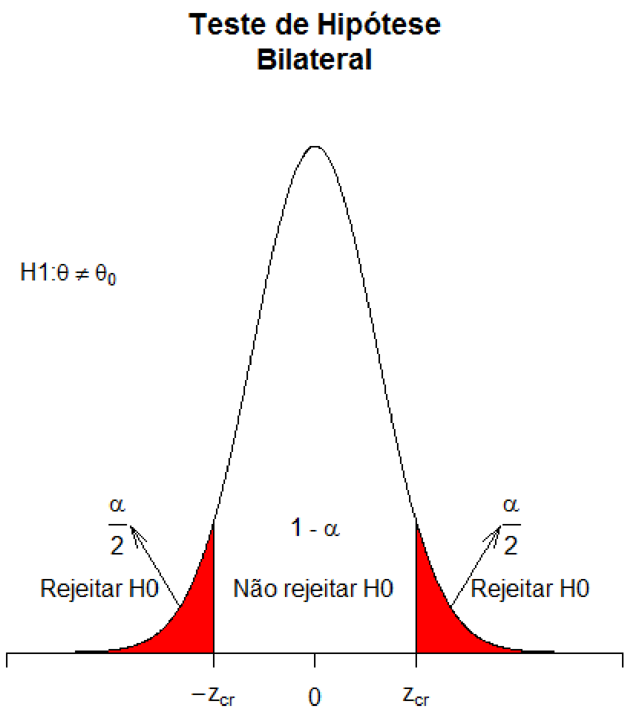
3.2 Procedimentos

Basicamente, os procedimentos para o teste de hipótese são os seguintes:
1. Enunciar as hipóteses, sendo:
H_0: $\theta = \theta_0$
H_1: $\theta < \theta_0$ ou $\theta \neq \theta_0$ ou $\theta > \theta_0$
2. Estabelecer o nível de significância α;
3. Calcular a variável de teste, de acordo com a distribuição amostra da estatística do teste;
4. Decidir sobre a rejeição ou não de H_0, comparando o valor da variável de teste com o valor tabelado da distribuição teórica correspondente.

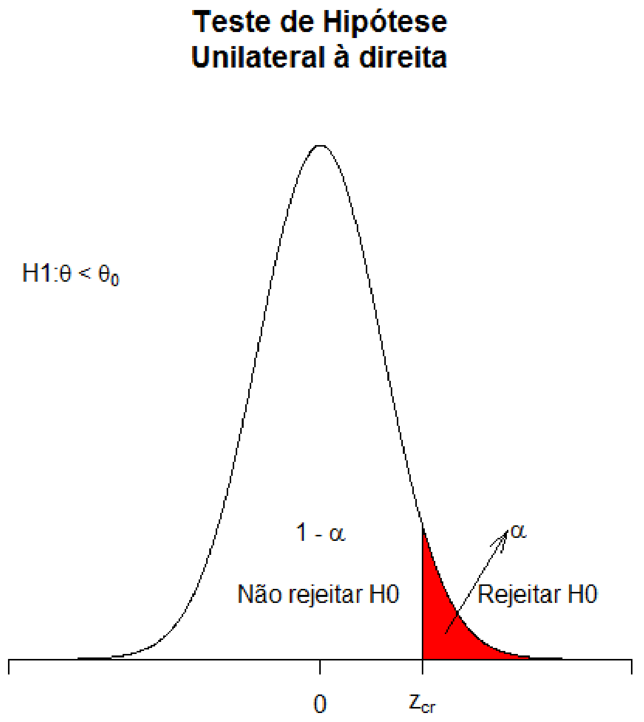
OBSERVAÇÃO: pode-se usar na decisão de rejeitar ou não H0 o p-valor da variável de teste; a hipótese H0 será rejeitada se p-valor < α.

Graficamente, temos:

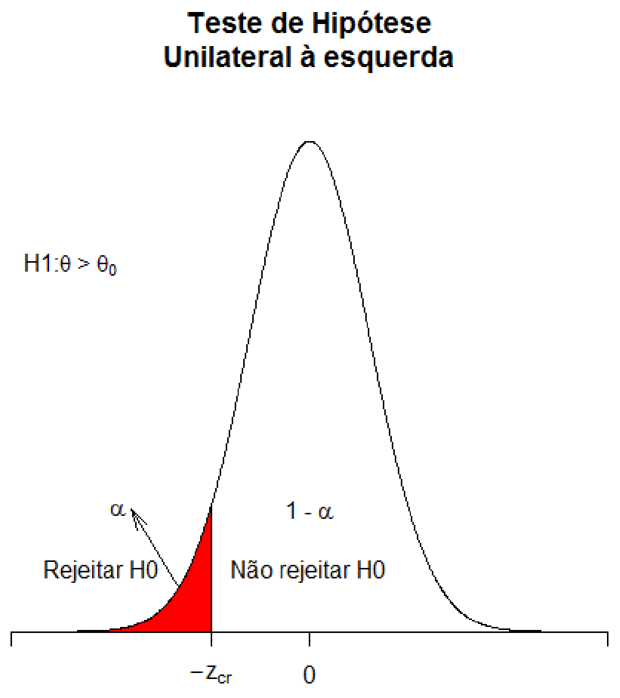
Teste de Hipótese Bilateral



Teste de Hipótese Unilateral à direita



Teste de Hipótese Unilateral à esquerda



3.3 Testes de hipóteses para uma amostra

3.3.1Teste de uma média populacional

1. σ conhecido

a) Estabelecer as hipóteses

$H_o:\ \mu = \mu_0$

$$H_1: \begin{cases} \mu < \mu_0 \\ \mu \neq \mu_0 \\ \mu > \mu_0 \end{cases}$$

B) estabelecer o nível de significância α;

C) calcular a variável de teste, dada por $z = \dfrac{\bar{x} - \mu_0}{\dfrac{\sigma}{\sqrt{n}}}$, pois $\bar{x} \approx N\left(\mu; \dfrac{\sigma}{\sqrt{n}}\right)$

D) Rejeita-se H_0 se $\begin{cases} z < -z_\alpha, \text{ p/ } \mu < \mu_0 \\ z < -z_{\frac{\alpha}{2}} \text{ ou } z > z_{\frac{\alpha}{2}}, \text{ p/ } \mu \neq \mu_0 \\ z > z_\alpha, \text{ p/ } \mu > \mu_0 \end{cases}$

Exemplo: uma fábrica anuncia que o índice de nicotina dos cigarros da marca X apresenta-se abaixo de 26 mg por cigarro. Um laboratório realizou 10 análise do índice obtendo: 26, 24, 23, 22, 28, 25, 27, 26, 28, 24. Sabe-se que o índice de nicotina dos cigarros da marca X se distribui normalmente com variância 5,36 mg²; pode-se aceitar a afirmação do fabricante, ao nível de 5%?

Então, temos:
H0: μ = 26
H1: μ < 26
α = 5%

$$\bar{X} = \frac{\sum x}{n} = \frac{253}{10} = 25{,}3$$

σ^2 = 5,36 mg² → σ = $\sqrt{5{,}36} = 2{,}32$

$$z = \frac{\bar{x} - \mu_0}{\dfrac{\sigma}{\sqrt{n}}} = \frac{25{,}3 - 26}{\dfrac{2{,}32}{\sqrt{10}}} = -0{,}959$$

Considerando o valor do nível de significância(α) igual a 5%, temos o valor de "z" tabelado igual a 1,64. Como o teste é unilateral à esquerda, o valor fica -1,64. Como z > -1,64, não rejeitamos H0.

2. σ desconhecido

a) Estabelecer as hipóteses

$$H_o:\ \mu=\mu_0$$

$$H_1:\begin{cases}\mu<\mu_0\\ \mu\neq\mu_0\\ \mu>\mu_0\end{cases}$$

B) estabelecer o nível de significância α;

C) calcular a variável de teste, dada por $t=\dfrac{\bar{x}-\mu_0}{\dfrac{s}{\sqrt{n}}}$, pois $\bar{x}\approx t\left(\mu;\dfrac{s}{\sqrt{n}}\right)$

D) Rejeita-se H_0 se $\begin{cases}t<-t_{n-1;\alpha\,\text{tab}},\ \text{p/ }\mu<\mu_0\\ t<-t_{n-1;\frac{\alpha}{2}}\ \text{ou t > t}_{n-1;\frac{\alpha}{2}},\ \text{p/ }\mu\neq\mu_0\\ t>t_{n-1;\alpha},\ \text{p/ }\mu>\mu_0\end{cases}$

Exemplo: um empresa tem constatado um volume médio de vendas de seus produtos comercializados no varejo na ordem de 200 mil reais mensais; contudo, um pesquisador selecionou um amostra de 16 estabelecimentos onde são comercializados os produtos e constatou um volume médio de vendas de 198 mil reais mensais com desvio-padrão de 4 mil reais; o pesquisador suspeita que o volume médio de vendas se alterou e não está mais em torno de 200 mil reais mensais; verifique se a suspeita do pesquisador está correta, a um nível de significância de 1%.

Então, temos:
H0: μ = 200.000
H1: μ ≠ 200.000
α = 1%

$\bar{X}=198.000$
σ = 4.000

$$t=\frac{\bar{x}-\mu_0}{\dfrac{s}{\sqrt{n}}}=\frac{198000-200000}{\dfrac{4000}{\sqrt{16}}}=-2,00$$

Considerando o valor do nível de significância(α) igual a 1%, temos o valor de "t" tabelado igual a 2,95. Como o teste é bilateral, temos -2,95 e 2,95. Como "t" está entre estes valores, não rejeitamos H0.

3.3.2 Teste de uma variância populacional

a) Estabelecer as hipóteses

$$H_0:\sigma^2=\sigma^2{}_0$$

$$H_1: \begin{cases} \sigma^2 < \sigma^2{}_o \\ \sigma^2 \neq \sigma^2{}_o \\ \sigma^2 > \sigma^2{}_o \end{cases}$$

B) estabelecer o nível de significância α;

C) calcular a variável de teste, dada por

$\chi^2 = \dfrac{(n-1)s^2}{\sigma_0^2}$, que tem distribuição χ^2_{n-1}

D) Rejeita-se H_0 se $\begin{cases} \chi^2 < \chi^2_{n-1;(1-\alpha)},\ \text{p/ } \sigma^2 < \sigma^2{}_o \\ \chi^2 < \chi^2_{n-1;\left(1-\alpha/2\right)}\ ou\ \chi^2 > \chi^2_{n-1;\frac{\alpha}{2}},\ \text{p/}\sigma^2 \neq \sigma^2{}_o \\ \chi^2 > \chi^2_{n-1;\alpha},\ \text{p/}\sigma^2 > \sigma^2{}_o \end{cases}$

Exemplo: uma amostra de 10 elementos de uma população forneceu uma variância $s^2 = 24{,}8$; pergunta-se: esse resultado é suficiente para se concluir, ao nível de significância $\alpha = 5\%$, que a variância dessa população é inferior a 50?

Então, temos:
H0: $\sigma^2 = 50$
H1: $\sigma^2 < 50$
$\alpha = 5\%$

$s^2 = 24{,}8$

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{(10-1)24{,}8}{50} = 4{,}464$$

Considerando o valor do nível de significância(α) igual a 5%, temos o valor de "χ^2" tabelado igual a 3,33. Como o teste é unilateral à esquerda, "χ^2" é maior que este valor e, não rejeitamos H0.

3.3.3 Teste de uma proporção populacional

a) Estabelecer as hipóteses

$$H_0: P = P_0$$

$$H_1: \begin{cases} P < P_0 \\ P \neq P_0 \\ P > P_0 \end{cases}$$

B) estabelecer o nível de significância α;

C) calcular a variável de teste, dada por $z = \dfrac{p - P_0}{\sqrt{\dfrac{P_0(1-P_0)}{n}}}$, pois $p \approx N\left(P, \dfrac{P(1-P)}{n}\right)$

D) Rejeita-se H_0 se $\begin{cases} z < -z_\alpha, \text{ p/ } P < P_0 \\ z < -z_{\frac{\alpha}{2}} \text{ ou } z > z_{\frac{\alpha}{2}}, \text{ p/ } P \neq P_0 \\ z > z_\alpha, \text{ p/ } P > P_0 \end{cases}$

Exemplo: um estatístico selecionou uma amostra aleatória de 2000 eleitores, constatando uma intenção de voto de 43% para um candidato á presidência na época das eleições; o político desconfia que sua intenção de voto se alterou, não estando mais em torno de 52%; pede, então ao estatístico que verifique se sua suspeita está correta ao nível de confiança de 99%.

Então, temos:
H0: $\pi = 52\%$
H1: $\pi \neq 52\%$
α = ? , mas foi dito que $1 - \alpha = 99\%$, logo $\alpha = 1\%$

p = 43%
n = 2000

$$z = \frac{p - P_0}{\sqrt{\frac{P_0(1-P_0)}{n}}} = \frac{0{,}43 - 0{,}52}{\sqrt{\frac{0{,}52(1-0{,}52)}{2000}}} = -8{,}18$$

Considerando o valor do nível de significância(α) igual a 1%, temos o valor de "z" tabelado igual a 2,58. Como o teste é bilateral, temos -2,58 e 2,58. Como "z" é menor do que -2,58, rejeitamos H0. Ou seja, a suspeita do político pode ter sentido.

3.4 Testes de Hipóteses para 2 amostras

3.4.1 Teste para comparação de duas médias

I. Dados emparelhados (experimento tipo antes - depois)

$H_o : \mu_d = \Delta$

a) $H_1 : \begin{cases} \mu_d < \Delta \\ \mu_d \neq \Delta \\ \mu_d > \Delta \end{cases}$, onde $\mu_d = \mu_{antes} - \mu_{depois}$

b) estabelecer o nível de significância α;

c) calcular a variável de teste, dada por

$$t = \frac{\bar{d} - \Delta}{\frac{s_d}{\sqrt{n}}}, \text{ pois } \bar{d} \approx t\left(\mu_d ; \frac{s_d}{\sqrt{n}}\right), \text{ onde } \bar{d} = \frac{\sum d_i}{n} \text{ e } d_i = X_{i\,\text{antes}} - X_{i\,\text{depois}}$$

$$s_d = \sqrt{\frac{\sum d_i^{\,2} - \frac{\left(\sum d_1\right)^2}{n}}{n-1}}$$

d) Rejeita-se H_0 se: $\begin{cases} t < t_{n-1;\alpha \text{ tab}}, \text{ p/ } \mu_d < \Delta \\ t < -tn-1;_{\frac{\alpha}{2}} \text{ ou t > t}_{n-1;\frac{\alpha}{2}}, \text{ p/ } \mu_d \neq \Delta \\ t > t_{n-1;\alpha}, \text{ p/ } \mu_d > \Delta \end{cases}$

Exemplo: uma amostra de 9 carros foi submetida a melhoramentos técnicos e mecânicos; foi verificado o consumo dos carros antes e depois dos melhoramentos; deseja-se verificar se existe diferença significativa no consumo dos carros antes e depois deste melhoramentos:

	Consumo(Km/L)		
Carro	Antes	Depois	Diferença
1	5	5,8	-0,8
2	10,3	11,5	-1,2
3	12	12,3	-0,3
4	9,7	10,6	-0,9
5	14	13,6	0,4
6	13,5	13,2	0,3
7	9	9,4	-0,4
8	7	9,6	-2,6
9	8	8,8	-0,8
Total	88,5	94,8	-6,3

Então temos:

$H_o : \mu_d = \Delta$

$H_1 : \mu_d \neq \Delta$

$\mu_d = \mu_{antes} - \mu_{depois}$

nível de significância(α) : 5%

Estatística de teste: $t = \dfrac{\bar{d} - \Delta}{\dfrac{s_d}{\sqrt{n}}}$

n = 9

$$\bar{d} = \frac{\sum d_i}{n} = \frac{-6,3}{9} = -0,7$$

$$s_d = \sqrt{\frac{\sum d_i^{\,2} - \frac{\left(\sum d_1\right)^2}{n}}{n-1}} = \sqrt{\frac{10,79 - \frac{(-6,3)^2}{9}}{9-1}} = 0,893$$

$$t = \frac{\bar{d} - \Delta}{\frac{s_d}{\sqrt{n}}} = \frac{-0,7-0}{\frac{0,893}{\sqrt{9}}} = -2,35$$

OBSERVAÇÂO: se n ≥ 30, podemos utilizar a seguinte estatística para o teste:

$$Z = \frac{\bar{d} - \Delta}{\frac{s_d}{\sqrt{n}}}$$

Considerando o valor do nível de significância(α) igual a 5%, temos o valor de "t" tabelado igual a 2,306. Como o teste é bilateral, temos -2,306 e 2,306. Como "t" calculado é menor do que -2,306, rejeitamos H0. Ou seja, existe diferença significativa do consumo após o melhoramento.

II. Dados não emparelhados

II.1 Com σ_1 e σ_2 conhecidos e diferentes

a)
$$H_o : \mu_d = \Delta$$
$$H_1 : \begin{cases} \mu_d < \Delta \\ \mu_d \neq \Delta \\ \mu_d > \Delta \end{cases}, \text{ onde } \mu_d = \mu_1 - \mu_2$$

b) estabelecer o nível de significância α;

c) calcular a variável de teste dada por

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta}{\sqrt{\frac{\sigma^2{}_1}{n_1} + \frac{\sigma^2{}_2}{n_2}}}, \text{ pois } (\bar{x}_1 - \bar{x}_2) = N\left(\mu_d ; \frac{\sigma^2{}_1}{n_1} + \frac{\sigma^2{}_2}{n_2}\right)$$

d) Rejeita-se H_0 se: $\begin{cases} z < z_{\alpha\,\text{tab}}, \text{ p/ } \mu_d < \Delta \\ z < -z_{\frac{\alpha}{2}} \text{ ou } z > z_{\frac{\alpha}{2}}, \text{ p/ } \mu_d \neq \Delta \\ z > z_{\alpha}, \text{ p/ } \mu_d > \Delta \end{cases}$

Exemplo: uma grande empresa quer comprar peças de 2 fornecedores diferentes; o fornecedor "A" alega que a durabilidade média de suas peças é de 1000 horas com desvio-padrão de 120 horas; enquanto que o fornecedor "B" diz que a durabilidade média de suas peças é de 1050 horas com desvio-padrão de 140 horas; duas amostra foram obtidas de cada fornecedor com um tamanho de 64, ou seja, $n_A = n_B = 64$; a duração média da amostra de "A" foi de 995 horas e a de "B" foi de 1025; qual a conclusão a 5% de significância, sabendo que os desvios-padrões são conhecidos e diferentes?

Então temos:

$$H_o : \mu_d = \Delta$$
$$H_1 : \mu_d \neq \Delta$$

$$\mu_d = \mu_1 - \mu_2$$

o nível de significância α = 5%

Da amostra de "A", sabemos que: $n_A = 64$ e $\overline{X}_A = 995$

Da amostra de "B", sabemos que: $n_B = 64$ e $\overline{X}_B = 1025$

Ainda são conhecidos de "A" o desvio-padrão $\sigma_A = 120$

e de "B" o desvio-padrão $\sigma_B = 140$

A estatística de teste é:

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta}{\sqrt{\frac{\sigma^2{}_1}{n_1} + \frac{\sigma^2{}_2}{n_2}}} = \frac{(995 - 1025) - 0}{\sqrt{\frac{120^2}{64} + \frac{140^2}{64}}} = -1{,}30$$

Considerando o valor do nível de significância(α) igual a 5%, temos o valor de "z" tabelado igual a 1,96. Como o teste é bilateral, temos -1,96 e 1,96. Como "z" calculado está entre o menor e o maior valor tabelado, não rejeitamos H0. Ou seja, não existe diferença significativa entre os fornecedores.

II.2 σ_1 e σ_2 conhecidos e iguais

$$H_o : \mu_d = \Delta$$

a) $H_1 : \begin{cases} \mu_d < \Delta \\ \mu_d \neq \Delta \\ \mu_d > \Delta \end{cases}$, onde $\mu_d = \mu_1 - \mu_2$

b) estabelecer o nível de significância α;

c) calcular a variável de teste dada por

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \text{ pois } (\bar{x}_1 - \bar{x}_2) = N\left(\mu_d ; \sigma^2\left[\frac{1}{n_1} + \frac{1}{n_2}\right]\right)$$

d) Rejeita-se H_0 se: $\begin{cases} z < z_{\alpha\,\text{tab}}, \text{ p/ } \mu_d < \Delta \\ z < -z_{\frac{\alpha}{2}} \text{ ou } z > z_{\frac{\alpha}{2}}, \text{ p/ } \mu_d \neq \Delta \\ z > z_{\alpha}, \text{ p/ } \mu_d > \Delta \end{cases}$

II.3 Com σ_1 e σ_2 desconhecidos e diferentes

$$H_o:\ \mu_d = \Delta$$

a) $H_1: \begin{cases} \mu_d < \Delta \\ \mu_d \neq \Delta \\ \mu_d > \Delta \end{cases}$, onde $\mu_d = \mu_1 - \mu_2$

b) estabelecer o nível de significância α;

c) calcular a variável de teste dada por

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta}{\sqrt{\frac{s^2{}_1}{n_1} + \frac{s^2{}_2}{n_2}}}\text{, pois } (\bar{x}_1 - \bar{x}_2) = t\left(\mu_d; \left[\frac{s^2{}_1}{n_1} + \frac{s^2{}_2}{n_2}\right]\right)$$

d) Rejeita-se H_0 se: $\begin{cases} t < t_{\upsilon;\varepsilon}\text{, p/ } \mu_d < \Delta \\ t < -t_{\upsilon;\frac{\alpha}{2}} \text{ ou t} > t_{\upsilon;\frac{\alpha}{2}}\text{, p/ } \mu_d \neq \Delta \\ t > t_{\upsilon;\alpha}\text{, p/ } \mu_d > \Delta \end{cases}$

Onde , o grau de liberdade é calculado pelo método de Aspin-Welch dado por $\upsilon = \frac{(w_1 + w_2)^2}{\frac{w_1^2}{(n_1 - 1)} + \frac{w_2^2}{(n_2 - 1)}}$, sendo $w_1 = \frac{s_1^2}{n_1}$ e $w_2 = \frac{s_2^2}{n_2}$

Exemplo: um empresa fabrica transistores do tipo "A" e do tipo "B"; a marca "A", mais cara, é, pelo menos 60 horas mais durável do que a marca "B"; um usuário quer saber se vale a pena pagar mais pela marca "A", e resolve testar, de fato, se ela é mais durável; testa 20 itens de "A", encontrando uma vida média de 1.000 horas com desvio-padrão de 60 horas, enquanto que 20 itens da marca "B" apresentam uma vida média de 910 horas e um desvio-padrão de 40 horas; qual a conclusão ao nível de 5% de significância, sabendo que os desvios-padrões são desconhecidos e diferentes?

Então, temos:

$H_o:\ \mu_d = \Delta$
$H_1: \mu_d > \Delta$

Onde Δ = 60 e $\mu_d = \mu_1 - \mu_2$

o nível de significância α = 5%

n = m = 20

Da amostra "A" sabemos: $\overline{X}_A = 1000$ horas e $\sigma_A = 60$ horas

Da amostra "B" sabemos: $\overline{X}_B = 910$ horas e $\sigma_B = 40$ horas

A estatística de teste é:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta}{\sqrt{\frac{s^2{}_1}{n_1} + \frac{s^2{}_2}{n_2}}} = \frac{(1000 - 910) - 60}{\sqrt{\frac{60^2}{20} + \frac{40^2}{20}}} = 1{,}861$$

Considerando o valor do nível de significância(α) igual a 5%, temos o valor de "t" tabelado, só que neste caso, o grau de liberdade será dados por:

$$\upsilon = \frac{(w_1 + w_2)^2}{\frac{w_1{}^2}{(n_1 - 1)} + \frac{w_2{}^2}{(n_2 - 1)}}, \text{ sendo } w_1 = \frac{s^2{}_1}{n_1} \text{ e } w_2 = \frac{s^2{}_2}{n_2}$$

Então:

$$w_1 = \frac{s^2{}_1}{n_1} = \frac{60^2}{20} = 180 \quad \text{e} \quad w_2 = \frac{s^2{}_2}{n_2} = \frac{40^2}{20} = 80, \quad \text{assim:}$$

$$\upsilon = \frac{(w_1 + w_2)^2}{\frac{w_1{}^2}{(n_1 - 1)} + \frac{w_2{}^2}{(n_2 - 1)}} = \frac{(180 + 80)^2}{\frac{180^2}{(20-1)} + \frac{80^2}{(20-1)}} = \frac{67600}{\frac{32400}{19} + \frac{6400}{19}} = 33{,}103 \cong 33$$

Então "t" será igual a 1,692. Como o teste é unilateral à esquerda, temos 1,861 > 1,692. Então rejeitamos H0. Ou seja, a vida média da marca "A" é pelo menos maior do que a da marca "B".

II. 4 Com σ_1 e σ_2 desconhecidos e iguais

a) $H_o : \mu_d = \Delta$

$$H_1 : \begin{cases} \mu_d < \Delta \\ \mu_d \neq \Delta \\ \mu_d > \Delta \end{cases}, \text{ onde } \mu_d = \mu_1 - \mu_2$$

b) estabelecer o nível de significância α;

c) calcular a variável de teste dada por

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \text{ pois } (\bar{x}_1 - \bar{x}_2) = t\left(\mu_d ; s_p{}^2 \left[\frac{1}{n_1} + \frac{1}{n_2}\right]\right)$$

$$s_p = \sqrt{\frac{(n_1 - 1)s^2{}_1 + (n_2 - 1)s^2{}_2}{n_1 + n_2 - 2}}$$

d) Rejeita-se H_0 se: $\begin{cases} t < t_{n_1+n_2-2;\alpha\,\text{tab}}, \text{ p/ } \mu_d < \Delta \\ t < -t_{n_1+n-2;\frac{\alpha}{2}} \text{ ou } t > t_{n_1+n-2;\frac{\alpha}{2}}, \text{ p/ } \mu_d \neq \Delta \\ t > t_{n_1+n-2;\alpha}, \text{ p/ } \mu_d > \Delta \end{cases}$

OBSERVAÇÂO: Se $n_1 \geq 30$ e $n_2 \geq 30$, podemos utilizar a seguinte estatística para teste:

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Exemplo: um relatório de defesa do consumidor mostrou que um teste com 8 pneus da marca "A" apresentaram uma vida média de 37.500 km com desvio-padrão de 3.500 km e que 12 pneus de uma marca concorrente "B", testados nas mesmas condições, tiveram uma durabilidade média de 41.400 km com um desvio-padrão de 4.200 km; supondo que os desvios-padrões populacionais sejam os mesmos e admitindo um nível de significância de 5%, verifique se é possível afirmar que as duas marcas diferem quanto a durabilidade média.

Então, temos:

$$H_o: \mu_d = \Delta$$
$$H_1: \mu_d \neq \Delta$$

sendo $\Delta = 0$ e $\mu_d = \mu_1 - \mu_2$

o nível de significância $\alpha = 5\%$

Sabemos da marca "A" que $\bar{X}_A = 37.500\,\text{Km}$, n = 8, $\sigma_A = 3.500\,\text{Km}$

Sabemos da marca "B" que $\bar{X}_B = 41.400\,\text{Km}$, n = 12, $\sigma_B = 4.200\,\text{Km}$

A estatística do teste é:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \text{ onde } s_p = \sqrt{\frac{(n_1 - 1)s^2{}_1 + (n_2 - 1)s^2{}_2}{n_1 + n_2 - 2}}$$

Resolvendo, temos:

$$s_p = \sqrt{\frac{(n_1 - 1)s^2{}_1 + (n_2 - 1)s^2{}_2}{n_1 + n_2 - 2}} = \sqrt{\frac{(8-1)3.500^3 + (12-1)4.200^2}{8+12-2}} = 4012{,}9651$$

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{(37.500 - 41.400)}{4012{,}9651\sqrt{\frac{1}{8} + \frac{1}{12}}} = -2{,}129$$

Considerando o valor do nível de significância(α) igual a 5%, temos o valor de "t" tabelado, só que neste caso, o grau de liberdade será dados por $n_1 + n_2 - 2$, que pelos dados corresponde a

18, então "t" tabelado é igual a 2, 101. Como o teste é bilateral, temos -2,101 e 2,101. Como "t" calculado é menor do que -2,101, rejeitamos H0. Ou seja, existe diferença significativa entre as marcas.

3.4.2 Teste para comparação de duas variâncias

a) enunciar as hipóteses

$$H_0 : \sigma^2{}_1 = \sigma^2{}_2$$

$$H_1 : \begin{cases} \sigma^2{}_1 < \sigma^2{}_2 \\ \sigma^2{}_1 \neq \sigma^2{}_2 \\ \sigma^2{}_1 > \sigma^2{}_2 \end{cases}$$

b) estabelecer o nível de significância α;

c) calcular a variável de teste dada por

$$F = \frac{s^2{}_2}{s^2{}_1}, \text{ se } H_1 : \sigma^2{}_1 < \sigma^2{}_2$$

$$F = \frac{s^2{}_1}{s^2{}_2}, \text{ se } H_1 : \sigma^2{}_1 > \sigma^2{}_2$$

$$F = \frac{\max(s^2{}_1; s^2{}_2)}{\min(s^2{}_1; s^2{}_2)}$$

d) decidir pela rejeição ou não de H_0 comparando F com $F_{n_1-1;n_2-1;\alpha}$

Exemplo: o desvio-padrão de uma dimensão particular de um componente de metal é satisfatório para a montagem deste componente; um novo fornecedor está sendo considerado e será preferível se o desvio-padrão de seu componente de metal for menor do que o do atual fornecedor; uma amostra de 100 itens de cada fornecedor foi obtida, onde foi verificada $S_A^2 = 0{,}0058$ e $S_B^2 = 0{,}0041$; com base nestes dados, a empresa deve trocar de fornecedor, a um nível de significância de 5%?

Então, temos:

$$H_0 : \sigma^2{}_1 = \sigma^2{}_2$$

$$H_1 : \sigma^2{}_1 > \sigma^2{}_2$$

o nível de significância α = 5%

n = m = 100

a variável de teste dada por: $F = \frac{\max(s^2{}_1; s^2{}_2)}{\min(s^2{}_1; s^2{}_2)} = \frac{\max(0{,}0058;0{,}0041)}{\min(0{,}0058;0{,}0041)} = \frac{0{,}0058}{0{,}0041} = 1{,}415$

Considerando o valor do nível de significância(α) igual a 5%, temos o valor de "F" tabelado, só que neste caso, os graus de liberdade serão dados por n - 1 e m - 1, que pelos dados corresponde a 1,394, então "F" tabelado é igual a 1,394. Como o teste é unilateral à esquerda, temos "F" calculado é maior do que 1,394, logo rejeitamos H0. Ou seja, a variância do fornecedor "A", o atual, é maior do que a do novo fornecedor, fornecedor "B".

3.4.3 Teste para comparação de duas proporções

a) enunciar as hipóteses

$$H_0 : P_1 - P_2 = \Pi$$

$$H_1 : \begin{cases} P_1 - P_2 < \Pi \\ P_1 - P_2 \neq \Pi \\ P_1 - P_2 > \Pi \end{cases}$$

b) estabelecer o nível de significância α;

c) calcular a variável de teste, dada por

I - Se P_1 e P_2 forem conhecidos:

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - \Pi}{\sqrt{\dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}}}\text{, onde } \hat{p}_1 = \frac{f_1}{n_1} \text{ e } \hat{p}_2 = \frac{f_2}{n_2}$$

II - Se P_1 e P_2 não forem conhecidos:

$$z = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{p'(1-p')\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}\text{, sendo } p' = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{f_1 + f_2}{n_1 + n_2}$$

a) Rejeita-se H_0 se: $\begin{cases} z < z_{\alpha\,\text{tab}}\text{, p/ } P_1 - P_2 < \Pi \\ z < -z_{\frac{\alpha}{2}} \text{ ou } z > z_{\frac{\alpha}{2}}\text{, p/ } P_1 - P_2 \neq \Pi \\ z > z_{\alpha}\text{, p/ } P_1 - P_2 > \Pi \end{cases}$

Exemplo: a reitoria de uma grande universidade entrevistou 600 alunos, dos quais 350 eram mulheres e 250 eram homens, para colher a opinião sobre a troca do sistema de avaliação da universidade; desta amostra, 140 mulheres e 115 homens estavam a favor da referida troca; verifique se existe diferença significativa de opinião entre homens e mulheres, ao nível de 5% de significância.

Então temos:

$$H_0 : P_1 - P_2 = \Pi$$

$H_1 : P_1 - P_2 \neq \Pi$

o nível de significância α = 5%

A estatística de teste é $z = \dfrac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{p'(1-p')\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$, pois $\Pi = 0$.

Sabe-se que $p' = \dfrac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \dfrac{f_1 + f_2}{n_1 + n_2} = \dfrac{140 + 115}{350 + 250} = \dfrac{255}{600} = 0{,}425$, $p_1 = \dfrac{f_1}{n_1} = \dfrac{140}{350} = 0{,}40$ e

$p_2 = \dfrac{f_2}{n_2} = \dfrac{115}{250} = 0{,}46$. Então,

$$z = \frac{(p_1 - p_2)}{\sqrt{p'(1-p')\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}} = \frac{0{,}40 - 0{,}46}{\sqrt{0{,}425(1-0{,}425)\left(\dfrac{1}{350} + \dfrac{1}{250}\right)}} = \frac{-0{,}06}{\sqrt{0{,}244(0{,}007)}} = \frac{-0{,}06}{0{,}045} = -1{,}33$$

Considerando o valor do nível de significância(α) igual a 5%, temos o valor de "z" tabelado igual a 1,96. Como o teste é bilateral, temos -1,96 e 1,96. Como "z" calculado está entre o menor e o maior valor tabelado, não rejeitamos H0. Ou seja, não existe diferença significativa entre as opiniões.

3.5 Teste para várias amostras

Existem situações em que desejamos trabalhar com várias amostras e tirar algumas conclusões sobre elas. Neste sentido, temos dois testes a utilizar:

I. Teste de homocedasticidade

II. ANOVA - Análise da variância

3.5.1 Testes de Homocedasticidade

Homocedasticidades significa que a variância da variável em estudo é a mesma em todos os níveis. Este é o pressuposto básico para a aplicação da ANOVA - Análise de Variância e para a Regressão Linear. A hipótese nula aqui é:

$$\text{H0: } \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$$

Temos, então:

I. para amostras do mesmo tamanho – Teste de Cochran

a) enunciar as hipóteses

$H_0 : \sigma^2{}_1 = \sigma^2{}_2 = \ldots = \sigma^2{}_k$

H_1: pelo menos uma variância difira das demais.

b) estabelecer o nível de significância α;

c) calcular a variável de teste, dada por

$$g = \frac{\max s^2{}_i}{\sum s^2{}_i}, \text{ com i = 1, 2,..., k}$$

d) decidir pela rejeição ou não de H_0 comparando g com o valor de g tabelado em função de n e K, onde n = número de grupos e k = número de réplicas(repetições). Se $g > g_{n,k,\alpha}$, rejeita-se H_0

Exemplo:

Um laboratório de metrologia contratou um novo metrologista que passou por diversos treinamentos para integrar a equipe. Antes de liberarmos o metrologista para realizar o procedimento de calibração, realizamos um teste para comparar a variabilidade das medições do metrologista novato com os demais metrologistas do laboratório. Em um experimento completamente aleatorizado, um bloco padrão de 50mm foi medido 5 vezes por cada metrologista. As medições estão na tabela a seguir:

	Metrologistas			
Medidas	**João**	**Novato**	**Moacir**	**Roberto**
Medida 1	50,0071	50,007	50,0072	50,0073
Medida 2	50,0072	50,0076	50,0074	50,0074
Medida 3	50,0072	50,0075	50,0073	50,0073
Medida 4	50,0071	50,0071	50,0072	50,0072
Medida 5	50,0072	50,0078	50,0072	50,0072

	Sumário Estatístico			
Estatísticas	**João**	**Novato**	**Moacir**	**Roberto**
Média	50,00716	50,0074	50,00726	50,00728
Desvio Padrão	0,000055	0,00034	0,000089	0,000084
Variância	0,000000003	0,000000115	0,000000008	0,000000007

Calculando g, temos:

$$g = \frac{\max s^2{}_i}{\sum s^2{}_i} = \frac{0{,}000000115}{0{,}000000003 + 0{,}000000115 + 0{,}000000008 + 0{,}000000007} = 0{,}864$$

Agora, temos que comparar com os dados da tabela abaixo, levando em consideração n - número de grupos, k - número de réplicas e α - nível de significância. No nosso exemplo, vamos adotar α = 5%. Então temos:

Número de Grupos	**Tamanho do grupo (réplicas)**				
	2	**3**	**4**	**5**	**6**
2	-	0,975	0,939	0,906	0,877
3	0,967	0,871	0,798	0,746	0,707
4	0,906	0,768	0,684	0,629	0,69
5	0,841	0,684	0,598	0,544	0,506
6	0,781	0,616	0,532	0,48	0,445
7	0,727	0,561	0,48	0,431	0,397
8	0,68	0,516	0,438	0,391	0,3

No caso, $g_{4,5,5\%} = 0{,}629$. Como g calculado é maior do que este valor tabelado, rejeitamos H0, ou seja a variância do metrologista Novato não é homogênea em relação a dos demais metrologistas.

II. para amostras de tamanhos diferentes – Teste de Bartlett

a) enunciar as hipóteses

$H_0 : \sigma^2{}_1 = \sigma^2{}_2 = ... = \sigma^2{}_k$

H_1: pelo menos uma variância difira das demais.

b) estabelecer o nível de significância α;

c) calcular a variável de teste, dada por

$B = \frac{1}{C}\left[(n-k)\ln S_p^2 - \sum_{i=1}^{k} \upsilon_i \ln s_i^2\right]$, que tem distribuição χ^2_{k-1}

onde:

$$n = \sum_{i=1}^{k} n_i$$

$$\upsilon_i = n_i - 1$$

$$S_p^2 = \frac{\sum_{i=1}^{k} \upsilon_i s^2{}_i}{n-k}$$

$$s_i^2 = \frac{\sum (x_i - \bar{x}_i)^2}{n_i - 1}$$

$$C = 1 + \frac{1}{3(k-1)}\left(\sum_{i=1}^{k} \frac{1}{\upsilon_i} - \frac{1}{n-k}\right)$$

d) Rejeita-se H_0 se $B > \chi^2_{k-1;\alpha}$

Exemplo: vamos comparar diversas resistências de fibras para diversas porcentagens de algodão, da seguinte forma:

Resistência da Fibra				
Fator: 15%	Fator: 16%	Fator: 17%	Fator: 18%	Fator: 19%
7	12	14	19	7
7	17	18	25	10
15	12	18	22	11
11	18	19	19	15
9	18	19	23	11

Sumário Estatístico					
Fator: 15%	Fator: 16%	Fator: 17%	Fator: 18%	Fator: 19%	Fator: 15%
Média	9,8	15,4	17,6	21,6	10,8
Desvio Padrão	3,346640106	3,130495168	2,073644135	2,607680962	2,863564213
Variância	11,2	9,8	4,3	6,8	8,2
n	5	5	5	5	5

$$S_p^2 = \frac{\sum_{i=1}^{k} \upsilon_i s^2{}_i}{n-k} = \frac{(5-1)*11,2+(5-1)*9,8+(5-1)*4,3+(5-1)*6,8+(5-1)*8,2}{25-5} = 8,06$$

$$B = \frac{1}{C}\left[(n-k)\ln S_p^2 - \sum_{i=1}^{k} \upsilon_i \ln s_i^2\right] =$$

$$= \frac{1}{C}\left[(25-5)\ln(80,6) - \left((5-1)*\ln(11,2)+(5-1)*\ln(9,8)+(5-1)*\ln(4,3)+(5-1)*\ln(6,8)+(5-1)*\ln(8,2)\right)\right] =$$

$$= \frac{1,026}{C}$$

$$C = 1+\frac{1}{3(k-1)}\left(\sum_{i=1}^{k}\frac{1}{\upsilon_i} - \frac{1}{n-k}\right) = 1+\frac{1}{3(5-1)}\left[\left(\frac{1}{5-1}+\frac{1}{5-1}+\frac{1}{5-1}+\frac{1}{5-1}+\frac{1}{5-1}\right)-\frac{1}{25-5}\right] = 1,10$$

Então,

$$B = \frac{1,026}{C} = \frac{1,026}{1,10} = 0,933$$

Devemos compara este valor com o valor tabelado de Qui-quadrado. Se $B > \chi^2_{k-1;\alpha}$, rejeitamos H0. O valor tabelado $\chi^2_{k-1;\alpha} = \chi^2_{5-1;5\%} = 9,49$, logo não rejeitamos H0, ou seja a de que todos as variâncias são iguais.

3.5.2 Análise de variância

- Comparação das técnicas de pesquisa

a) o teste de χ^2 para associação compara freqüências observadas com freqüências esperadas, pois isso apresenta uma fragilidade devido a falta de sensibilidade em comparação ao teste de associação angular

b) o teste de associação angular da conta de uma associação que o χ^2 não acusa em certas situações;

c) o teste t aplica-se apenas à duas populações.

Definições

- teste para a comparação de várias médias;
- desenvolvida por Sir R. A. Fisher, estatístico britânico;

- é um método suficientemente poderoso para poder identificar diferenças entre as médias populacionais devidas a várias causas atuando simultaneamente sobre os elementos da população;
- visa analisar a variação de uma resposta e associar partes dessa variação a cada variável de um conjunto de variáveis independentes;
- o objetivo é localizar as variáveis independentes importantes em um estudo e determinar como elas interajam e afetam a resposta;
- a variabilidade (dispersão) de um conjunto de n medidas é proporcional à soma dos quadrados dos desvios $SQ_x = \sum_{i=1}^{k}\sum_{j=1}^{n}\left(x_{ij} - \overline{X}\right)^2$;
- a análise da variância subdivide $SQ_x = \sum_{i=1}^{k}\sum_{j=1}^{n}\left(x_{ij} - \overline{X}\right)^2$, chamada Soma Total dos Quadrados dos Desvios, em duas parcelas, cada uma das quais é atribuída a uma variável independente do experimento, mais uma outra variável que responda pelos erros aleatórios;
- cada uma das parcelas da Soma Total dos Quadrados dos Desvios divididas por uma constate apropriada indica um estimador independente e imparcial de σ^2;
- as hipóteses são $H_0 : \mu_1 = \mu_2 = ... = \mu_k$ e H_1: pelo menos uma das médias difere das demais;
- para cada caso deve-se estabelecer o nível de significância;
- as K populações têm a mesma variância σ^2 (condição de homoscedasticidade)
- a variável de interesse é normalmente distribuída em todas as populações.

Hipóteses básicas:

- no modelo matemático adotado, os diversos efeitos (de tratamento ou de bloco) são aditivos;
- normalidade dos valores observados em cada grupo
- homogeneidade da variância dentro de cada grupo
- os erros ou desvios e_{ij} são independentes, de onde resulta que não são correlacionados;
- os erros ou desvios e_{ij} têm todos a mesma variância σ^2;
- os erros ou desvios e_{ij} têm distribuição normal;

1) Uma classificação – Um Fator – Em grupos simples

a) é o caso mais simples;
b) deseja-se averiguar a influência do tratamento T_i , i = 1, 2,...,K, dado ao elemento x_{ij} da amostra j, j = 1, 2,..., n.
c) o modelo matemático linear é $x_{ij} = \mu + \alpha_i + e_{ij}$, onde μ é a média global, α_i é o efeito do tratamento i e e_{ij} é o erro aleatório.

d) apresentação dos dados

Tratamentos i	amostras	Média i
Tratamento 1	$x_{11}\ x_{12}\ x_{13}$ ....x_{1n}	$\bar{x}_1$
Tratamento 2	$x_{21}\ x_{22}\ x_{23}$ ....x_{2n}	$\bar{x}_2$
Tratamento k	$x_{k1}\ x_{k2}\ x_{k3}$ ....x_{kn}	$\bar{x}_k$

Seja $SQ = SQ_E + SQ_D$, onde:

$SQ = \sum_{i=1}^{k}\sum_{j=1}^{n}\left(x_{ij} - \bar{X}\right)^2$, é a soma dos quadrados totais, onde $\bar{X} = \dfrac{\sum_{i=1}^{k}\sum_{j=1}^{n} x_{ij}}{kn}$

$SQ_E = n\sum_{i=1}^{k}\left(\bar{x}_i - \bar{X}\right)^2$, onde $\bar{x}_i = \dfrac{\sum_{j=1}^{n} x_{ij}}{n}$, soma dos quadrados entre as amostras

$SQ_D = \sum_{i=1}^{k}\sum_{j=1}^{n}\left(x_{ij} - \bar{x}_i\right)^2$, dentro das amostras.

Prova-se pelo Teorema de Fisher que $SQ \approx \chi^2_{nk-1}$, $SQ_E \approx \chi^2_{k-1}$ e $SQ_D \approx \chi^2_{k(n-1)}$

A estimativa da variância total é $\hat{\sigma}^2{}_T = S^2{}_T = \dfrac{SQ}{nk-1}$;

A estimativa da variância entre é $\hat{\sigma}^2{}_E = S^2{}_E = \dfrac{SQ_E}{k-1}$;

A estimativa da variância dentro é $\hat{\sigma}^2{}_D = S^2{}_D = \dfrac{SQ_D}{k(n-1)}$

A variável de teste é $F = \dfrac{S^2{}_E}{S^2{}_D}$ e é comparada com o valor tabelado de $F_{k-1;k(n-1);\alpha}$

Quadro de análise de variância

Causa das variações	Soma dos quadrados	Graus de liberdade	Média da soma dos quadrados	F
Entre amostras	$SQ_E = n\sum_{i=1}^{k}\left(\bar{x}_i - \bar{X}\right)^2$	k-1	$\hat{\sigma}^2{}_E = S^2{}_E = \dfrac{SQ_E}{k-1}$	$F = \dfrac{S^2{}_E}{S^2{}_D}$
Dentro da	$SQ_D = \sum_{i=1}^{k}\sum_{j=1}^{n}\left(x_{ij} - \bar{x}_i\right)^2$	k(n-1)	$\hat{\sigma}^2{}_D = S^2{}_D = \dfrac{SQ_D}{k(n-1)}$	

amostra				
Total	$SQ=\sum_{i=1}^{k}\sum_{j=1}^{n}\left(x_{ij}-\overline{X}\right)^2$	nk-1		

Para amostras de tamanhos diferentes, temos

$$SQ_E=\sum_{i=1}^{k}n_i\left(\overline{x}_i-\overline{X}\right)^2$$

E as estimativas das variâncias passam a ser

A estimativa da variância total é $\hat{\sigma}^2{}_T=S^2{}_T=\dfrac{SQ}{n-1}$;

A estimativa da variância entre é $\hat{\sigma}^2{}_E=S^2{}_E=\dfrac{SQ_E}{k-1}$;

A estimativa da variância dentro é $\hat{\sigma}^2{}_D=S^2{}_D=\dfrac{SQ_D}{n-k}$

Quadro de análise de variância

Causa das variações	Soma dos quadrados	Graus de liberdade	Média da soma dos quadrados	F
Entre amostras	$SQ_E=\sum_{i=1}^{k}n_i\left(\overline{x}_i-\overline{X}\right)^2$	k-1	$\hat{\sigma}^2{}_E=S^2{}_E=\dfrac{SQ_E}{k-1}$	$F=\dfrac{S^2{}_E}{S^2{}_D}$
Dentro da amostra	$SQ_D=\sum_{i=1}^{k}\sum_{j=1}^{n}\left(x_{ij}-\overline{x}_i\right)^2$	n-k	$\hat{\sigma}^2{}_D=S^2{}_D=\dfrac{SQ_D}{n-k}$	
Total	$SQ=\sum_{i=1}^{k}\sum_{j=1}^{n}\left(x_{ij}-\overline{X}\right)^2$	n-1		

Exemplo:

O resultado das vendas efetuadas por 3 vendedores de uma indústria durante certo período é dado abaixo. Deseja-se saber, ao nível de significância de 5%, se há diferenças entre os vendedores:

RESUMO ESTATÍSTICO

Grupo	*Contagem*	*Soma*	*Média*	*Variância*
Vendedor A	6	178	29,66667	3,066667
Vendedor B	4	112	28	2
Vendedor C	5	127	25,4	106,3

ANOVA

Fonte da variação	*SQ*	*gl*	*MQ*	*F*	*valor-P*	*F crítico*
Entre grupos	49,8666667	2	24,93333	0,670051	0,529825	3,885294
Dentro dos	446,533333	12	37,21111			

grupos		
Total	496,4	14

Comparando o valor da coluna valor-P acima com o nível de significância 5%, não rejeitamos H0, pois valor-P > 0,05, ou seja não há diferenças significativas entre os vendedores.

2) Duas classificações – 2 fatores – Em blocos ao acaso

a) os elementos observados x_{ij} são classificados segundo dois critérios;
b) o primeiro critério corresponde às linhas i (i = 1, 2, ...,k) e o segundo às colunas j (j = 1, 2, ..., n) da matriz k x n;
c) as hipóteses são $H_{01}: \mu_{1\cdot} = \mu_{2\cdot} = ... = \mu_{k\cdot}$ e $H_{01}: \mu_{\cdot 1} = \mu_{\cdot 2} = ... = \mu_{\cdot k}$
d) a aceitação de H_{01} significa a não-comprovação de diferença significativa entre as médias devida à classificação segundo o critério das linhas;
e) a aceitação de H_{02} significa a não-comprovação de diferença significativa entre médias devida à classificação segundo o critério das colunas;
e) o modelo matemático linear é $x_{ij} = \mu + \alpha_i + \beta_j + e_{ij}$, onde onde μ é a média global, α_i é o efeito linha do tratamento i, β_j é o efeito do tratamento coluna j e e_{ij} é o erro aleatório.
f) apresentação dos dados

	Segundo critério j	Médias
Primeiro critério i	X_{11} x_{12} ....x_{1j}.................................x_{1n}	$\bar{x}_{1\cdot}$
	....	
	X_{k1} x_{k2} ....x_{kj}.................................x_{kn}	$\bar{x}_{k\cdot}$
Médias	$\bar{x}_{\cdot 1}$ $\bar{x}_{\cdot 2}$ .... $\bar{x}_{\cdot n}$	$\bar{X} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n} x_{ij}}{kn}$

Seja $SQ = SQ_C + SQ_L + SQ_{Erro}$, onde

$SQ = \sum_{i=1}^{k}\sum_{j=1}^{n}\left(x_{ij} - \bar{X}\right)^2$, é a soma dos quadrados totais, onde $\bar{X} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n} x_{ij}}{kn}$

$SQ_C = k\sum_{j=1}^{n}\left(\bar{x}_{\cdot j} - \bar{X}\right)^2$, é a soma dos quadrados nas colunas, e $\bar{x}_{\cdot j} = \frac{\sum_{i=1}^{k} x_{ij}}{k}$

$SQ_L = n\sum_{i=1}^{k}\left(\bar{x}_{i.} - \bar{X}\right)^2$, é a soma dos quadrados das linhas, e $\bar{x}_i. = \dfrac{\sum_{j=1}^{n} x_{ij}}{n}$

$$SQ_{erro} = \sum_{j=1}^{k}\sum_{i=1}^{n}\left(x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{X}\right)^2$$

Prova-se pelo Teorema de Fisher que $SQ \approx \chi^2_{nk-1}$, $SQ_C \approx \chi^2_{k-1}$, $SQ_L \approx \chi^2_{n-1}$ e $SQ_{Erro} \approx \chi^2_{(k-1)(n-1)}$

A estimativa da variância total é $\hat{\sigma}^2{}_T = S^2{}_T = \dfrac{SQ}{nk-1}$;

A estimativa da variância nas colunas é $\hat{\sigma}^2{}_C = S^2{}_C = \dfrac{SQ_C}{k-1}$;

A estimativa da variância nas linhas é $\hat{\sigma}^2{}_L = S^2{}_L = \dfrac{SQ_L}{n-1}$

A estimativa da variância residual é $\hat{\sigma}^2{}_{Erro} = S^2{}_{Erro} = \dfrac{SQ_{Erro}}{(k-1)(n-1)}$

As variáveis de teste são $F_L = \dfrac{S^2{}_L}{S^2{}_{Erro}}$ e $F_C = \dfrac{S^2{}_C}{S^2{}_{Erro}}$ e são comparadas, respectivamente, com o valores tabelados de $F_{k-1;(k-1)(n-1);\alpha}$ e $F_{n-1;(k-1)(n-1);\alpha}$

Quadro de análise de variância

Causa das variações	Soma dos quadrados	Graus de liberdade	Média da soma dos quadrados	F
Entre linhas	$SQ_L = n\sum_{i=1}^{k}\left(\bar{x}_{i.} - \bar{X}\right)^2$	n-1	$\hat{\sigma}^2{}_L = S^2{}_L = \dfrac{SQ_L}{n-1}$	$F_L = \dfrac{S^2{}_L}{S^2{}_{Erro}}$
Entre colunas	$SQ_C = k\sum_{j=1}^{n}\left(\bar{x}_{.j} - \bar{X}\right)^2$	k-1	$\hat{\sigma}^2{}_C = S^2{}_C = \dfrac{SQ_C}{k-1}$	$F_C = \dfrac{S^2{}_C}{S^2{}_{Erro}}$
Aleatório	$SQ_{erro} = \sum_{j=1}^{k}\sum_{i=1}^{n}\left(x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{X}\right)^2$	(k-1)(n-1)	$\hat{\sigma}^2{}_{Erro} = S^2{}_{Erro} = \dfrac{SQ_{Erro}}{(k-1)(n-1)}$	
Total	$SQ = \sum_{i=1}^{k}\sum_{j=1}^{n}\left(x_{ij} - \bar{X}\right)^2$	nk-1		

Exemplo:

Em uma experiência agrícola, foram usados diferentes fertilizantes em duas variedades de trigo. A produção está indicada abaixo. Verifique ao nível de 5% se:

a) há diferenças na produção devido ao fertilizantes

b) há diferença na safra devido à variedade do trigo

RESUMO	Contagem	Soma	Média	Variância
Variedade 1	5	232	46,4	36,8
Variedade 2	5	247	49,4	36,3
A	2	111	55,5	4,5
B	2	80	40	8
C	2	91	45,5	0,5
D	2	103	51,5	4,5
E	2	94	47	18

ANOVA

Fonte da variação	SQ	gl	MQ	F	valor-P	F crítico
Linhas	22,5	1	22,5	6,923077	0,058115	7,708647
Colunas	279,4	4	69,85	21,49231	0,005754	6,388233
Erro	13	4	3,25			
Total	314,9	9				

Comparando os valores da coluna valor-P acima com o nível de significância 5% podemos rejeitar a hipótese nula H0 para as colunas, ou seja, fator fertilizante, já que valor-P < 0,05, o que significa que o tipo de fertilizante tem influência na produção de trigo. Já com relação às linhas, ou seja, o fator variedade do trigo, não podemos rejeitar H0, ou seja, a variedade de trigo não altera a produção.

4 Comparações múltiplas

a) se H_0 for rejeitada, estaremos admitindo que pelo menos uma das médias é diferente das demais;

b) mas quais médias devem ser consideradas deferentes de quais outras ?

4.1 Teste de Tukey

Se o modelo é o de classificação única, temos

$\left|\bar{x}_i - \bar{x}_m\right| > q_{k,\upsilon,\alpha}\sqrt{\dfrac{S^2{}_D}{n}}$, onde q é tabelado em função de k, $\upsilon = k(n-1)$ e α e

Se for o de classificação dupla, temos

Para as colunas

$$\left|\bar{x}._j - \bar{x}._m\right| > q_{n,\upsilon,\alpha}\sqrt{\frac{S^2{}_{erro}}{n}}$$

Para as linhas

$$\left|\bar{x}_i. - \bar{x}_m.\right| > q_{k,\upsilon,\alpha}\sqrt{\frac{S^2{}_{erro}}{k}} \text{ ,}$$

onde q é tabelado em função de k, $\upsilon = (k-1)(n-1)$ e α .

4.2 Teste de Scheffe

a) utiliza os mesmos valores do quadro de Análise da variância;

b) pode ser usado no caso de amostras de tamanhos diferentes;

c) no caso de classificação única

$$|\bar{x}_i - \bar{x}_m| > \sqrt{S^2{}_D \left[\frac{2(k-1)}{n}\right] F_{k-1;k(n-1);\alpha}}$$

d) no caso de classificação única com amostras de tamanhos diferentes

$$|\bar{x}_i - \bar{x}_m| > \sqrt{S^2{}_D (k-1)\left[\frac{1}{n_i} + \frac{1}{n_m}\right] F_{k-1;\sum n_i - k;\alpha}}$$

e) para o caso de classificação dupla

Para as colunas

$$|\bar{x}._j - \bar{x}._m| > \sqrt{S^2{}_D \left[\frac{2(k-1)}{k}\right] F_{n-1;(k-1)(n-1);\alpha}}$$

Para as linhas

$$|\bar{x}_i. - \bar{x}_m.| > \sqrt{S^2{}_D \left[\frac{2(k-1)}{n}\right] F_{k-1;(k-1)(n-1);\alpha}}$$

4.3 Teste para múltiplas proporções

Para comparação de proporções de mais de duas populações utiliza-se o Teste de Qui-quadrado, considerando como hipóteses:

H0: $\pi_1 = \pi_2 = ... = \pi_n$
H1: nem todas as π_i são iguais, com i = 1, 2, ..., n

Ou seja, a hipótese nula (H0) é a hipótese de que todas as proporções são iguais e a hipótese alternativa (H1) é a de que pelo menos uma proporção seja diferente das demais.

A estatística do teste é dada por:

$$\chi_c^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(fo_{ij} - fe_{ij})^2}{fe_{ij}}$$

Onde:
fo_{ij} = é a freqüência observada na linha i e na coluna j
fe_{ij} = é a freqüência esperada na linha i e na coluna j

Esta estatística tem uma distribuição de Qui-quadrado com n – 1 graus de liberdade(gl) e com um nível de significância α.
Após o cálculo de χ_c^2, o valor obtido é comparado com o valor de Qui-quadrado tabelado segundo gl e α.

Rejeita-se H0 se $\chi_c^2 > \chi_{\alpha;gl}^2$, ou seja, se χ_c^2 for maior do que o valor tabelado.

Atualmente, com o constante uso de programas estatísticos, podemos rejeitar H0 com base no p-value. Neste caso, comparamos este valor com o nível de significância α. Se p-value < α, rejetamos H0.

Uma vez rejeitada H0, resta-nos identificar qual diferença entre os pares de π_i é significante. Para isto, calculamos as diferenças entre as proporções, usando:

$$|\pi_i - \pi_j|$$

Uma vez calculadas as diferenças entre as proporções, devemos calcular o valor crítico pelo `processo de Marascuillo, descrito em http://www.itl.nist.gov/div898/handbook/prc/section4/prc474.htm,` dado por:

$$C_{ij} = \sqrt{\chi_{\alpha;gl}^2}\sqrt{\frac{p_i(1-p_i)}{n_i} + \frac{p_j(1-p_j)}{n_j}}$$

Compara-se o valor das diferenças entre as proporções com o valor obtido de C_{ij}. As diferenças são significantes quando:

$$|\pi_i - \pi_j| > C_{ij}$$

5 Análise de Regressão

Conjunto de métodos e técnicas para o estabelecimento de fórmulas empíricas que interpretem a relação funcional entre variáveis com boa aproximação.
Deseja-se encontrar alguma forma de medir a relação entre as variáveis de cada conjunto, de tal modo que essa medida pudesse mostrar:

a) se há relação entre as variáveis e, caso afirmativo, se é fraca ou forte;
b) que, se essa relação existir, estabeleceremos um modelo que interprete a relação funcional existente entre as variáveis;
c) que construindo o modelo, usá-lo-emos para fins de predição.

Suponhamos que Y seja uma variável que nos interessa estudar e prever o seu comportamento. É de se esperar que os valores da variável Y (dependente) sofram influências dos valores de um número infinito de variáveis $X_1, X_2, ..., X_N$ (independentes) e que exista uma função g que expresse tal dependência, ou seja

$$Y = g(X_1, X_2, ..., X_N)$$

É impraticável a utilização das N variáveis ou por desconhecimento dos valores de algumas ou pela dificuldade de mensuração e tratamento de outras, logo se usa um número menor de variáveis (k) e o modelo fica

$$Y = f(X_1, X_2, ..., X_k) + h(X_{k+1}, X_{k+2}, ..., X_N)$$

Todas as influências das variáveis $X_{k+1}, X_{k+2}, ..., X_N$, sobre as quais não exercemos controle, serão consideradas como casuais, e associaremos uma variável aleatória U, obtendo o seguinte modelo:

$$Y = f(X_1, X_2, ..., X_k) + U$$

onde $f(X_1, X_2, ..., X_k)$ é a componente funcional do modelo e U a parte aleatória.

Problemas na análise de regressão:

a) o problema da especificação do modelo
   Consiste em determinar qual o tipo de função f que melhor explique a relação entre Y e $X_1, X_2, ..., X_k$
b) o problema da estimação dos parâmetros
   Consiste em estimar o valor dos diversos parâmetros que aparecem na especificação adotada.
c) o problema da adaptação e significância do modelo adotado
   Consiste em verificar se a especificação adotada na primeira etapa se adapta a convenientemente aos dados observados.

5.1 Modelo de regressão linear simples

Quando a função f que relaciona X e Y é da seguinte forma:

$$Y_i = \alpha + \beta X_i + U_i$$

onde: - $\alpha + \beta X_i$ é a componente funcional, que representa a influência da variável independente X sobre o valor de Y e define o eixo da nuvem de pontos, que nesse caso será uma reta;

- U_i é a componente aleatória, que representa a influência de outros fatores.

Sobre U_i temos:

a) tem distribuição Normal;
b) é uma variável aleatória com média igual a 0 e variância igual a σ^2, ou seja
   $E(U_i) = 0$ e $Var(U_i) = \sigma^2$, logo $U_i \approx N(0;\sigma^2)$
c) a $Cov(U_i;U_j) = \sigma^2$ para i = j e $Cov(U_i;U_j) = 0$ para i ≠ j

5.1.1 O modelo matemático

Quando desejamos fazer inferências sobre a população da qual foi extraída uma amostra, devemos considerar o modelo matemático que vai nos permitir construir intervalos de confiança e testar hipóteses.

- Hipóteses simplificadoras
São as hipóteses básicas sobre a regularidade da população:
1ª as distribuições de probabilidade $P(Y_i \mid X_i)$ possuem a mesma variância σ^2 para todo X_i;
2ª as médias $E(Y_i) = \mu_i = \alpha + \beta X_i$ se dispõem sobre uma linha reta, conhecida como a verdadeira reta de regressão (da população); os parâmetros α e β que especificam esta reta devem ser estimados a partir da informação da amostra;
3ª as variáveis aleatórias Y_i são estatisticamente independentes, com $E(Y_i) = \mu_i = \alpha + \beta X_i$ e $Var(Y_i) = \sigma^2$

5.1.2 Estimação de parâmetros

Seja $\hat{Y}_i = a + bX_i$ uma estimativa de $Y_i = \alpha + \beta X_i + U_i$, onde a e b são os estimadores de α e β e seja $e_i = (Y_i - \hat{Y}_i)$ o erro de estimação ou desvio.

Deseja-se minimizar a soma dos desvios ao quadrado, ou seja minimizar $\sum e_i^2 = \sum (Y_i - \hat{Y}_i)^2$.

Usando o Método dos Mínimos ao Quadrado, encontramos

$b = \frac{S_{XY}}{S_{XX}}$, onde $S_{XY} = \sum XY - \frac{\sum X \sum Y}{n}$ e $S_{XX} = \sum X^2 - \frac{\left(\sum X\right)^2}{n}$

$a = \overline{Y} - b\overline{X}$

5.1.3 Teorema de Gauss-Markov

A justificativa principal para utilizarmos o Método dos Mínimos Quadrados para estimar os parâmetros de $Y_i = \alpha + \beta X_i + U_i$ é a seguinte:

"Na classe dos estimadores lineares não-tendenciosos, o estimador *b* de mínimos quadrados tem variância mínima (é o mais eficiente). Analogamente, o estimador *a* também tem variância mínima".

Aplica-se somente a estimadores simultaneamente lineares e não-tendenciosos.

5.1.4 Significância das estimativas

Prova-se que:

$$a \approx N\left(\alpha; \sigma^2\left[\frac{1}{n} + \frac{\overline{X}^2}{S_{XX}}\right]\right),\ b \approx N\left(\beta; \frac{\sigma^2}{S_{XX}}\right) \text{ e } \hat{Y} \approx N\left(\alpha + \beta X; \sigma^2\left[\frac{1}{n} + \frac{\left(X - \overline{X}\right)^2}{S_{XX}}\right]\right)$$

onde σ^2 é a variância homoscedástica e desconhecida

Um estimador não-viesado de σ^2 é $\hat{\sigma}^2 = S^2 = \frac{S_{YY} - b^2 S_{XX}}{n-2} = \frac{S_{YY} - bS_{XY}}{n-2}$, onde $S_{XY} = \sum XY - \frac{\sum X \sum Y}{n}$, $S_{XX} = \sum X^2 - \frac{\left(\sum X\right)^2}{n}$ e $S_{YY} = \sum Y^2 - \frac{\left(\sum Y\right)^2}{n}$

5.1.5 Teste de hipóteses

As hipóteses são $H_0 : \beta = 0$ e $H_1: \beta > 0$ ou $\beta < 0$ ou $\beta \neq 0$

A variável de teste é $t = \frac{b - \beta}{\frac{S}{\sqrt{S_{XX}}}}$ que tem distribuição t-Student com n - 2 graus de liberdade

Para o modelo como um todo, se usa a variável de teste $F = \frac{SQM_E}{SQM_R}$, que tem distribuição F de Snedecor com α fixado e 1 grau de liberdade no numerador e n-2 graus de liberdade no denominador. Caso F > $F_{tabelado}$ rejeita-se H_0 (hipótese de que não existe regressão entre os dados observados).

Quadro de Análise de variância

Fonte de variação	Soma dos quadrados	Graus de liberdade	Quadrados médios	F
Explicada (devido a regressão)	$VE = b^2 S_{XX}$	1	$SQM_E = \frac{VE}{1}$	$F = \frac{SQM_E}{SQM_R}$
Residual	$VR = S_{YY} - b^2 S_{XX}$	n – 2	$SQM_R = \frac{VR}{n-2}$	

Total	$VT = S_{YY}$	n - 1		

Uma vez que

$\sum\left(Y_i-\overline{Y}\right)^2 = \sum\left(Y_i-\hat{Y}\right)^2 + \sum\left(\hat{Y}-\overline{Y}\right)^2$ - VT = VR + VE

$$VT = \sum\left(Y_i-\overline{Y}\right)^2 = \sum\left(Y_i^2-2Y_i\overline{Y}+\overline{Y}^2\right) = \sum Y_i^2 - 2\overline{Y}\sum Y_i + \sum \overline{Y}^2 = \sum Y_i^2 - 2\frac{\sum Y_i}{n}\sum Y_i + n\left(\frac{\sum Y_i}{n}\right)^2 =$$

$$\sum Y_i^2 - 2\frac{\left(\sum Y_i\right)^2}{n} + \frac{\left(\sum Y_i\right)^2}{n} = \sum Y_i^2 - \frac{\left(\sum Y_i\right)^2}{n} = S_{YY}$$

$$VR = \sum\left(Y_i-\hat{Y}\right)^2 = \sum\left(Y_i-\left[a+bX_i\right]\right)^2 = \sum\left(Y_i-\left[\left(\overline{Y}-b\overline{X}\right)+bX_i\right]\right)^2 = \sum\left(Y_i-\left[\overline{Y}-b\overline{X}+bX_i\right]\right)^2 =$$

$$\sum\left(Y_i-\left[\overline{Y}+b\left(X_i-\overline{X}\right)\right]\right)^2 = \sum\left(Y_i-\overline{Y}-b\left(X_i-\overline{X}\right)\right)^2 = \sum\left(\left(Y_i-\overline{Y}\right)-b\left(X_i-\overline{X}\right)\right)^2 =$$

$$\sum\left(\left(Y_i-\overline{Y}\right)^2-2b\left(X_i-\overline{X}\right)\left(Y_i-\overline{Y}\right)+b^2\left(X_i-\overline{X}\right)^2\right) = \sum\left(Y_i-\overline{Y}\right)^2 - 2b\sum\left(X_i-\overline{X}\right)\left(Y_i-\overline{Y}\right) + b^2\left(X_i-\overline{X}\right)^2 =$$

$$S_{YY} - 2bS_{XY} + b^2S_{XX} = S_{YY} - 2b\left(bS_{XX}\right) + b^2S_{XX} = S_{YY} - b^2S_{XX}$$

$$VE = \sum\left(\hat{Y}_i-\overline{Y}\right)^2 = \sum\left(\left[a+bX_i\right]-\left[a+b\overline{X}\right]\right)^2 = \sum\left(b\left(X_i-\overline{X}\right)\right)^2 = b^2\sum\left(X_i-\overline{X}\right)^2 = b^2S_{XX}$$

5.1.6 Coeficiente de Explicação ou determinação

Explica a relação entre a variação explicada VE e a variação total VT e é dado por $R^2 = \frac{VE}{VT} = \frac{b^2S_{XX}}{S_{YY}}$, onde $0 \le R^2 \le 1$ e se $R^2 = 0$ o modelo adotado não explica nada da realidade e se $R^2 = 1$ o modelo adotado explica a realidade com perfeição.

O R² indica quantos por cento a variação explicada pela regressão representa da variação total do modelo.

O valor da raiz quadrada de R^2 representa o coeficiente de correlação linear

O R^2 ajustado é dado por

$R^2_{ajustado} = 1 - \left[\left(1-R^2\right)\frac{n-1}{n-k-1}\right]$, onde k é número de variáveis independentes

5.1.7 Previsão

Uma vez encontrado os valores de *a* e *b* podemos fazer a previsão usando $\hat{Y} = a + bX$, e prova-se que

1) a previsão média tem distribuição

$$E\left(\hat{Y}_i \mid X\right) = \alpha + \beta X$$

$$Var\left(\hat{Y}_i \mid X\right) = Var\left(a+bX\right) = Var(\overline{Y}-b\overline{X}+bX) = Var\left(\overline{Y}+b\left[X-\overline{X}\right]\right) = Var\left(\overline{Y}\right) + \left(X-\overline{X}\right)^2 Var(b) =$$

$$= \frac{1}{n^2}\sum Var(Y) + \left(X-\overline{X}\right)^2\frac{\sigma^2}{S_{XX}} = \frac{1}{n^2}n\sigma^2 + \left(X-\overline{X}\right)^2\frac{\sigma^2}{S_{XX}} = \sigma^2\left[\frac{1}{n} + \frac{\left(X-\overline{X}^2\right)}{S_{XX}}\right]$$

$$P\left(\hat{Y}_i - t_{\frac{\alpha}{2}}\sqrt{\hat{\sigma}^2\left[\frac{1}{n}+\frac{\left(X-\overline{X}\right)^2}{S_{XX}}\right]} \le Y_i \le \hat{Y}_i + t_{\frac{\alpha}{2}}\sqrt{\hat{\sigma}^2\left[\frac{1}{n}+\frac{\left(X-\overline{X}\right)^2}{S_{XX}}\right]}\right) = 1-\alpha$$

2) a previsão individual tem distribuição

$$E\left(\hat{Y}_0 \mid X\right) = \alpha + \beta X$$

$$Var\left(\hat{Y}_0 \mid X\right) = \sigma^2\left[1 + \frac{1}{n} + \frac{\left(X - \overline{X}\right)^2}{S_{XX}}\right]$$

$$P\left(\hat{Y}_i - t_{\frac{\alpha}{2}}\sqrt{\hat{\sigma}^2\left[1 + \frac{1}{n} + \frac{\left(X - \overline{X}\right)^2}{S_{XX}}\right]} \le Y_i \le \hat{Y}_i + t_{\frac{\alpha}{2}}\sqrt{\hat{\sigma}^2\left[1 + \frac{1}{n} + \frac{\left(X - \overline{X}\right)^2}{S_{XX}}\right]}\right) = 1 - \alpha$$

Exemplo:

Suponha que exista uma relação linear entre as variáveis X = despesas com propaganda e Y = vendas de certo produto. Considerando os dados abaixo, determine a reta de mínimos quadrados, os testes e o coeficiente de explicação:

X (milhões de reais)	Y (milhares de unidades)
1,5	120
5,5	190
10,0	240
3,0	140
7,5	180
5,0	150
13,0	280
4,0	110
9,0	210
12,5	220
15,0	310

Primeiramente, devemos fazer o seguinte:

X (milhões de reais)	Y (milhares de unidades)	XY	X²	Y²
1,5	120	180	2,25	14400
5,5	190	1045	30,25	36100
10,0	240	2400	100	57600
3,0	140	420	9	19600
7,5	180	1350	56,25	32400
5,0	150	750	25	22500
13,0	280	3640	169	78400
4,0	110	440	16	12100
9,0	210	1890	81	44100
12,5	220	2750	156,25	48400
15,0	310	4650	225	96100
86	2150	19515	870	461700

Usando as fórmulas dadas, temos:

$$\overline{Y} = \frac{\sum Y}{n} = \frac{2150}{11} = 195{,}45 \qquad \overline{X} = \frac{\sum X}{n} = \frac{86}{11} = 7{,}82$$

$$S_{XY} = \sum XY - \frac{\sum X \sum Y}{n} = 19515 - \frac{86(2150)}{11} = 2705{,}91$$

$$S_{XX} = \sum X^2 - \frac{\left(\sum X\right)^2}{n} = 870 - \frac{(86)^2}{11} = 197{,}64$$

$$S_{YY} = \sum Y^2 - \frac{\left(\sum Y\right)^2}{n} = 461700 - \frac{(2150)^2}{11} = 41472{,}73$$

$$b = \frac{S_{XY}}{S_{XX}} = \frac{2705{,}91}{197{,}64} = 13{,}69$$

$$a = \overline{Y} - b\overline{X} = 195{,}45 - 13{,}69(7{,}82) = 88{,}39$$

Então, o modelo $\hat{Y}_i = a + bX_i$, fica $\hat{Y}_i = 88{,}39 + 13{,}69X_i$

Teste dos coeficientes do modelo

i) hipótese
  H_0: α e $\beta = 0$
  H_1 α e $\beta \neq 0$

ii) para α = 5%, temos t, com n – 2 g. l. igual a 2,2622

iii) cálculo da variável de teste

$$t = \frac{b - \beta}{\frac{S}{\sqrt{S_{XX}}}} = \frac{13{,}69}{22{,}18 / \sqrt{197{,}64}} = 8{,}71$$

Onde: $S^2 = \frac{S_{YY} - b^2 S_{XX}}{n-2} = \frac{S_{YY} - bS_{XY}}{n-2} \rightarrow S^2 = \frac{41472{,}73 - 13{,}69(2705{,}91)}{9} = 492{,}06$

$$S = \sqrt{S^2} = \sqrt{492{,}06} = 22{,}18$$

Como o valor da variável de teste é maior que valor de t tabulado, rejeitamos H_0.

Teste F para a regressão

i) hipótese
  H_0: não existe regressão
  H_1 : existe regressão

ii) para α = 5%, temos F, com 1 e n – 2 g. l. igual a 5,12.

iii) cálculo da variável de teste

$$F = \frac{SQM_E}{SQM_R} = \frac{VE}{S^2} = \frac{bS_{XY}}{S^2} = \frac{13{,}69(2705{,}91)}{492{,}06} = 75{,}28$$

Como o valor da variável de teste é maior que valor de F tabulado, rejeitamos H_0.

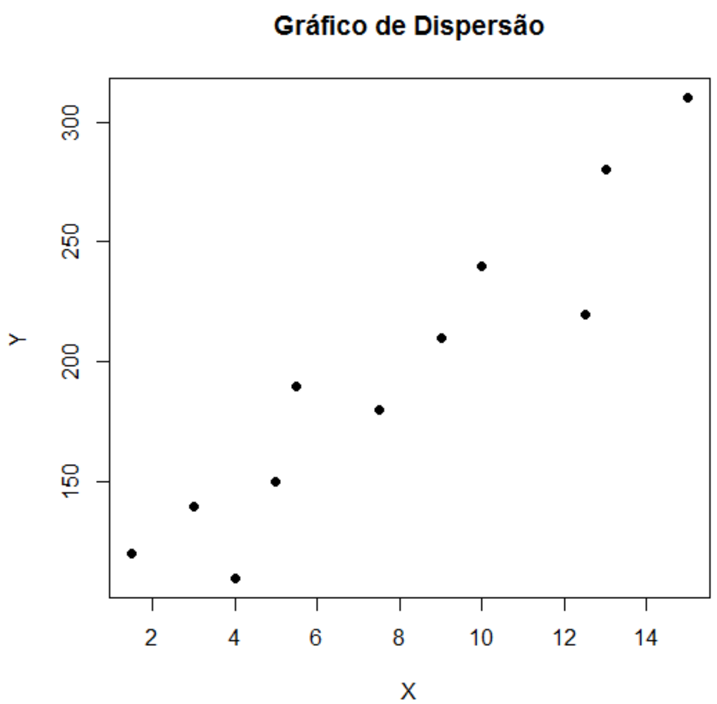
O Coeficiente de explicação é dado por: $R^2 = \frac{VE}{VT} = \frac{bS_{XY}}{S_{YY}} = \frac{(13,69)(2705,91)}{41472,73} = 0,89$ ou 89%.

Este resultado indica que o modelo explica 89% da variação total de Y

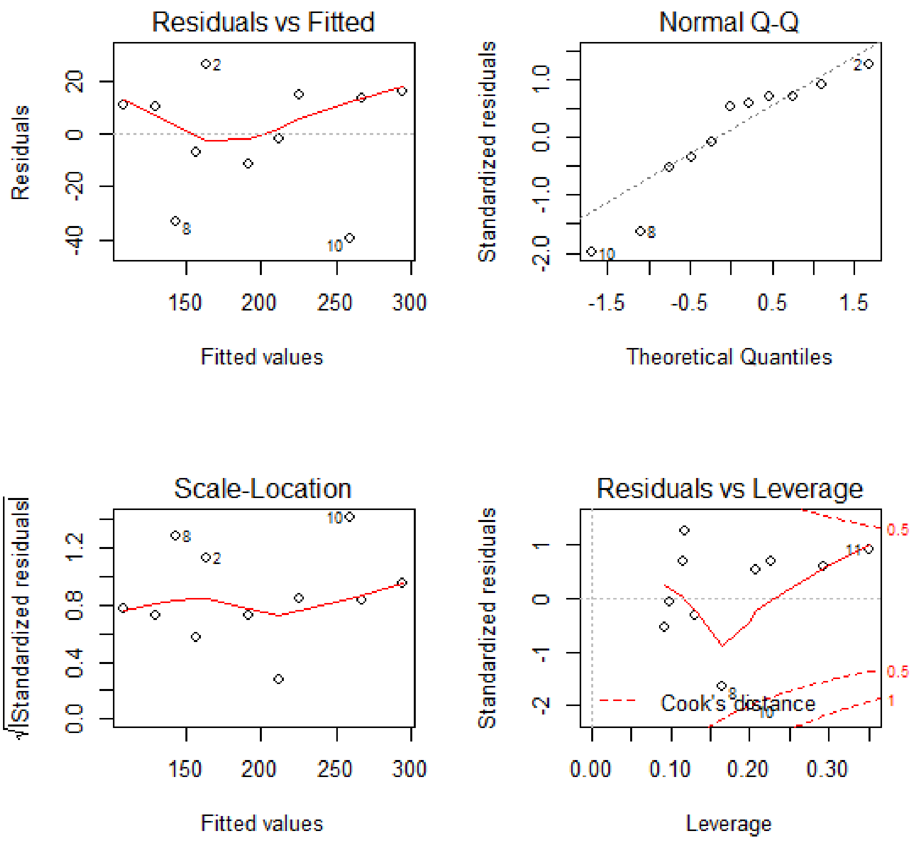
Saída de um Pacote Estatístico - R

```
Call:
lm(formula = dados$Y ~ dados$X)

Residuals:
    Min      1Q  Median      3Q     Max
-39.555  -8.984  10.513  14.136  26.284

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    88.41      14.03    6.30  0.00014 ***
X              13.69       1.58    8.68 0.000011 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 22.17 on 9 degrees of freedom
Multiple R-squared: 0.8933,    Adjusted R-squared: 0.8814
F-statistic: 75.35 on 1 and 9 DF,  p-value: 1.147e-05
```

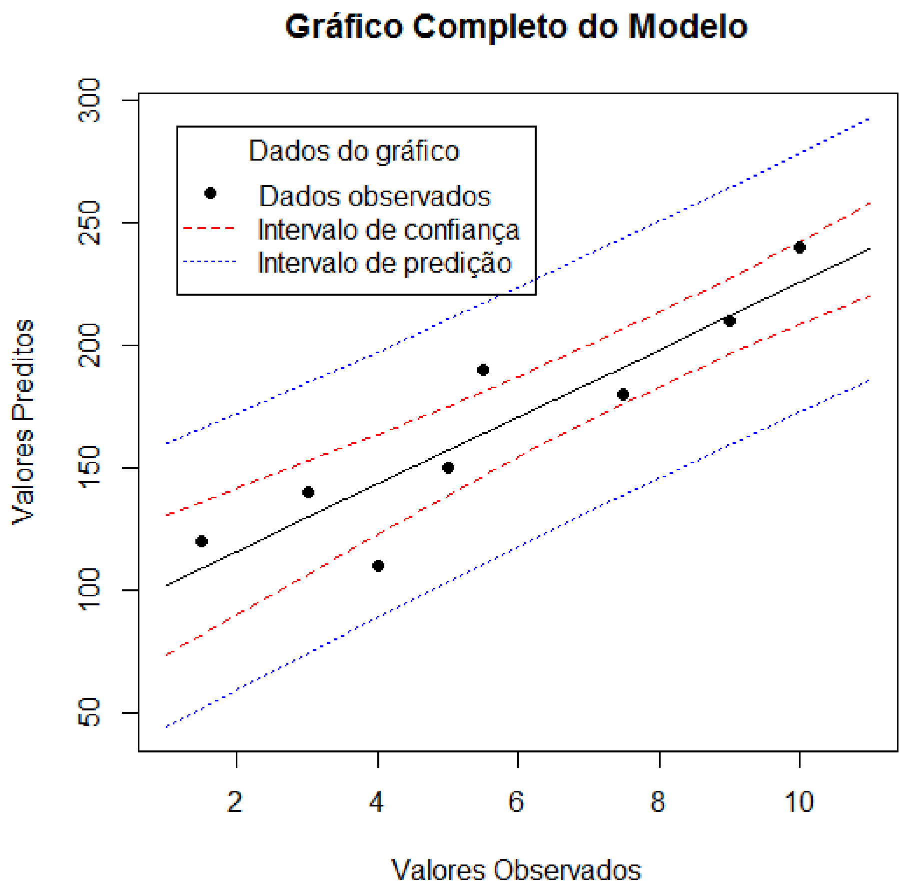
Graficamente, temos:



Com o modelo, temos:


Residuals vs Fitted
Residuals
Fitted values
Normal Q-Q
Standardized residuals
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage

E ainda:



5.2 Modelo de regressão linear múltipla

5.2.1 Introdução

O modelo de regressão da população de k +1 variáveis envolvendo a variável dependente Y e k variáveis independentes ou explicativas $X_1, X_2, \ldots, X_k$, pode ser escrita com

$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \ldots + \beta_k X_{ki} + U_i$, onde

β_0 é o intercepto

β_1 a β_k são os coeficientes parciais de inclinação

U_i é o termo de perturbação estocástica

Esta expressão é uma abreviação do seguinte conjunto de n equações

$$Y_1 = \beta_0 + \beta_1 X_{11} + \beta_2 X_{21} + \beta_3 X_{31} + \ldots + \beta_k X_{k1} + U_1$$

$$Y_2 = \beta_0 + \beta_1 X_{12} + \beta_2 X_{22} + \beta_3 X_{32} + \ldots + \beta_k X_{k2} + U_2$$

....

$$Y_n = \beta_0 + \beta_1 X_{1n} + \beta_2 X_{2n} + \beta_3 X_{3n} + ... + \beta_k X_{kn} + U_n$$

Que pode ser escrito em forma de matriz

$$\begin{bmatrix} Y_1 \\ Y_2 \\ . \\ . \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{21} & ... & X_{k1} \\ 1 & X_{12} & X_{22} & ... & X_{k2} \\ . & . & . & ... & . \\ . & . & . & ... & . \\ 1 & X_{1n} & X_{2n} & ... & X_{kn} \end{bmatrix} \times \begin{bmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ \beta_k \end{bmatrix} + \begin{bmatrix} U_1 \\ U_2 \\ . \\ . \\ U_n \end{bmatrix}$$

ou $Y_{n\times 1} = X_{n\times k} \times \beta_{k\times 1} + U_{n\times 1}$

que é conhecida como representação matricial do modelo geral (k +1 variáveis) de regressão linear e pode ser escrito ainda $Y = X\beta + U$

5.2.2 Hipóteses básicas do modelo de regressão linear

a) aleatoriedade dos U_i: a variável U_i é real e aleatória;
b) a variável U_i tem média 0, ou seja E(U_i) = 0, para todo i;
c) Homoscedasticidade de U_i , ou seja tem variância constante $E\left(U_i^{\,2}\right) = \sigma^2$
d) Ausência de autocorrelação ou independência serial dos resíduos U_i , ou seja $E\left(U_i U_j\right) = 0$, para i ≠ j
e) Independência entre U_i e X_{ij} , ou seja $E(X_{ij}, U_i) = 0$
f) Ausência de multicolineariedade perfeita: as variáveis explicativas não apresentam correlação linear perfeita.

5.2.3 Estimação dos parâmetros

Para obter as estimativas de β usaremos o MQO a partir de uma função de regressão da amostra – FRA , do tipo

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + ... + \hat{\beta}_k X_{ki}$$

que em forma matricial fica

$$\begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ . \\ . \\ \hat{Y}_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{21} & \dots & X_{k1} \\ 1 & X_{12} & X_{22} & \dots & X_{k2} \\ . & . & . & \dots & . \\ . & . & . & \dots & . \\ 1 & X_{1n} & X_{2n} & \dots & X_{kn} \end{bmatrix} \times \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ . \\ . \\ \hat{\beta}_k \end{bmatrix}$$ que corresponde a $\hat{Y} = X\hat{\beta}$

Fazendo $U_i = Y_i - \hat{Y}_i$ e posteriormente $\sum U^2{}_i = \sum \left(Y_i - \hat{Y}_i\right)^2$, que é a soma dos quadrados dos resíduos, poderemos achar os valores $\hat{\beta}$ que minimizam esta soma, ou sejam as estimativas de MQO.

Em termos de matrizes, temos que

$$\sum U^2{}_i = U'U = \begin{bmatrix} U_1 & U_2 & U_3 & \dots & U_n \end{bmatrix} \times \begin{bmatrix} U_1 \\ U_2 \\ U_3 \\ \dots \\ U_n \end{bmatrix} = U_1{}^2 + U_2{}^2 + \dots + U_n{}^2$$

Então

$$\sum U_i{}^2 = U'U = \left(Y - \hat{Y}\right)'\left(Y - \hat{Y}\right) = Y'Y - Y'\hat{Y} - \hat{Y}'Y + \hat{Y}'\hat{Y} = Y'Y - Y'\left(X\hat{\beta}\right) - \left(X\hat{\beta}\right)'Y + \left(X\hat{\beta}\right)'\left(X\hat{\beta}\right) =$$ Sendo

$$= Y'Y - Y'X\hat{\beta} - X'\hat{\beta}'Y + X'\hat{\beta}'X\hat{\beta} = Y'Y - 2\hat{\beta}'X'Y + \hat{\beta}'X'X\hat{\beta}$$

$$Y'X\hat{\beta} = \hat{\beta}'X'Y$$

Usando as regras de diferenciação matricial

$\dfrac{\partial(U'U)}{\partial\hat{\beta}} = -2X'Y + 2X'X\hat{\beta}$, sabendo que $\hat{\beta}'\hat{\beta} = \hat{\beta}^2$

Igualando a zero temos $-2X'Y + 2X'X\hat{\beta} = 0 \therefore \hat{\beta} = X'Y\left(X'X\right)^{-1}$

5.2.4 Distribuição amostral de $\hat{\beta}$

Fazendo $\hat{\beta} = X'Y\left(X'X\right)^{-1} = X'\left(X\beta + U\right)\left(X'X\right)^{-1} = \beta + X'\left(X'X\right)^{-1}U$, logo

$E(\hat{\beta}) = E(\beta) + X'\left(X'X\right)^{-1}E(U) = E(\beta) = \beta$

$Var(\hat{\beta}) = \sigma^2\left(X'X\right)^{-1}$, onde σ^2 é a variância homoscedástica e desconhecida

Um estimador não-viesado de σ^2 é

$\hat{\sigma}^2 = \dfrac{U'U}{n-k-1}$, logo $Var(\hat{\beta}) = \hat{\sigma}^2\left(X'X\right)^{-1}$, ou seja a diagonal principal de $\hat{\sigma}^2\left(X'X\right)^{-1}$

5.2.5 Propriedade dos estimadores $\hat{\beta}$

São estimadores lineares não-viesados, ou seja de variância mínima (Teorema de Gauss-Markov)

5.2.6 Teste de Hipóteses e Intervalo de confiança de $\hat{\beta}$

Etapas dos Teste de hipótese

a) enunciar as hipóteses

$H_0 : \beta = 0$, ou seja, se todos os valores de β são iguais a zero e a regressão não existe

$H_1 : \beta \neq 0,$ ou $\beta > 0,$ ou $\beta < 0$

b) estabelecer o nível de confiança α

c) usar a variável de teste

$$t = \frac{\hat{\beta}_i - \beta}{\hat{\sigma}^2 (X'X)}$$

que tem distribuição t-Student com n-k-1 graus de liberdade

d) decidir sobre a rejeição ou não de H_0 comparando t com o valor de t tabelado.

5.2.7 Intervalo de confiança

$$P\left(\hat{\beta}_i - t_{\frac{\alpha}{2}} \hat{\sigma} \sqrt{(X'X)} \leq \beta \leq \hat{\beta}_i + t_{\frac{\alpha}{2}} \hat{\sigma} \sqrt{(X'X)} \right) = 1 - \alpha$$

5.2.8 Análise de variância – teste da significância global da regressão

Quadro de Análise de variância

Fonte de variação	Soma dos quadrados	Graus de liberdade	Quadrados médios	F
Explicada (devido a regressão)	$VE = b^2 S_{XX}$	k	$SQM_E = \frac{VE}{k}$	$F = \frac{SQM_E}{SQM_R}$
Residual	$VR = S_{YY} - b^2 S_{XX}$	n – k – 1	$SQM_R = \frac{VR}{n-k-1}$	
Total	$VT = Y'Y - n\overline{Y}^2$	n – 1		

Uma vez que

$$\sum\left(Y_i - \overline{Y}\right)^2 = \sum\left(Y_i - \hat{Y}\right)^2 + \sum\left(\hat{Y} - \overline{Y}\right)^2 \text{ - VT = VR + VE}$$

$$VT = \sum\left(Y_i - \overline{Y}\right)^2 = \sum\left(Y_i^2 - 2Y_i\overline{Y} + \overline{Y}^2\right) = \sum Y_i^2 - 2\overline{Y}\sum Y_i + \sum \overline{Y}^2 = \sum Y_i^2 - 2\frac{\sum Y_i}{n}\sum Y_i + n\left(\frac{\sum Y_i}{n}\right)^2 =$$

$$\sum Y_i^2 - 2\frac{\left(\sum Y_i\right)^2}{n} + \frac{\left(\sum Y_i\right)^2}{n} = \sum Y_i^2 - \frac{\left(\sum Y_i\right)^2}{n} = Y'Y - n\overline{Y}^2$$

$$VR = \sum\left(Y_i - \hat{Y}\right)^2 = \sum U_i^2 = U'U = \left(Y - \hat{Y}\right)'\left(Y - \hat{Y}\right) = Y'Y - Y'\hat{Y} - \hat{Y}'Y + \hat{Y}'\hat{Y} =$$

$$= Y'Y - Y'\left(X\hat{\beta}\right) - \left(X\hat{\beta}\right)'Y + \left(X\hat{\beta}\right)'\left(X\hat{\beta}\right) = Y'Y - Y'X\hat{\beta} - X'\hat{\beta}'Y + X'\hat{\beta}'X\hat{\beta} =$$

$$= Y'Y - \hat{\beta}'X'Y - X'\hat{\beta}'Y + X'X\hat{\beta}'\left(X'Y(X'X)^{1}\right) = Y'Y - 2\hat{\beta}'X'Y + \hat{\beta}'X'Y = Y'Y - \hat{\beta}'X'Y$$

$$VE = \sum\left(\hat{Y}_i - \overline{Y}\right)^2 = \hat{\beta}'X'Y - n\overline{Y}^2$$

5.2.9 Coeficiente de Explicação ou determinação

Explica a relação entre a variação explicada VE e a variação total VT e é dado por $R^2 = \frac{VE}{VT} = \frac{\hat{\beta}'X'Y - n\overline{Y}^2}{Y'Y - n\overline{Y}^2}$, onde $0 \le R^2 \le 1$ e se $R^2 = 0$ o modelo adotado não explica nada da realidade e se $R^2 = 1$ o modelo adotado explica a realidade com perfeição.

O R^2 ajustado é dado por

$R^2_{ajustado} = 1 - \left[\left(1 - R^2\right)\frac{n-1}{n-k-1}\right]$, onde k é número de variáveis independentes

5.2.10 Previsão

Uma vez encontrado os valores $\hat{\beta}$ podemos fazer a previsão usando $\hat{Y} = X_0{}'\hat{\beta}$, onde $X_0{}' = [1 \quad X_{01} \quad X_{02} \quad \ldots \quad X_{0k}]$ e prova-se que

1) a previsão média tem distribuição

$$E\left(\hat{Y}_i \mid X_0{}'\right) = X_0{}'\hat{\beta}$$

$$Var\left(\hat{Y}_i \mid X_0{}'\right) = \hat{\sigma}^2\left(X_0{}'X_0\right)\left(X'X\right)^{-1}$$

$$P\left(\hat{Y}_i - t_{\frac{\alpha}{2}}\sqrt{\hat{\sigma}^2\left(X_0{}' X_0\right)(X'X)^{-1}} \le X_0'\hat{\beta} \le \hat{Y}_i + t_{\frac{\alpha}{2}}\sqrt{\hat{\sigma}^2\left(X_0{}' X_0\right)(X'X)^{-1}}\right) = 1-\alpha$$

2) a previsão individual tem distribuição

$$E\left(\hat{Y}_0 \mid X_0{}'\right) = X_0{}'\hat{\beta}$$

$$Var\left(\hat{Y}_0 \mid X_0{}'\right) = \hat{\sigma}^2\left[1+\left(X_0{}' X_0\right)(X'X)^{-1}\right]$$

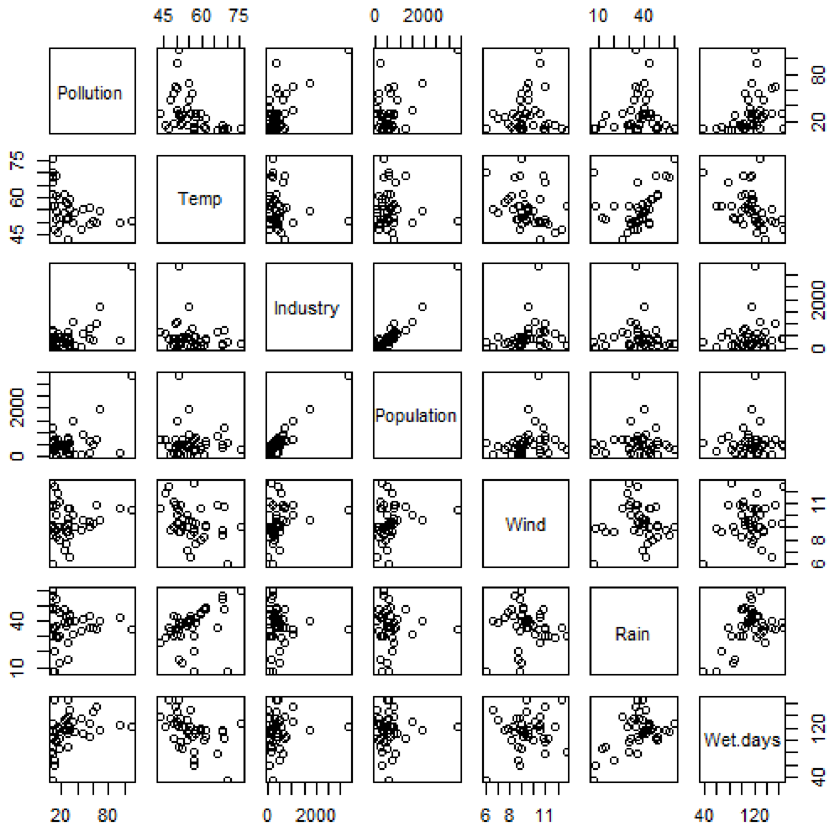
$$P\left(\hat{Y}_i - t_{\frac{\alpha}{2}}\sqrt{\hat{\sigma}^2\left[1+\left(X_0{}' X_0\right)(X'X)^{-1}\right]} \le X_0'\hat{\beta} \le \hat{Y}_i - t_{\frac{\alpha}{2}}\sqrt{\hat{\sigma}^2\left[1+\left(X_0{}' X_0\right)(X'X)^{-1}\right]}\right) = 1-\alpha$$

Exemplo:
Vamos usar o conjunto de dados *Pollute*, que tem dados do nível de poluição em algumas cidades e atributos destas cidades que podem servir como variáveis preditoras. O conjunto de dados pode ser obtido no seguinte endereço: http://www.bio.ic.ac.uk/research/mjcraw/therbook/data/Pollute.txt

Os dados são:

```
  Pollution Temp Industry Population Wind Rain Wet.days
1        24 61.5      368        497  9.1 48.3      115
2        30 55.6      291        593  8.3 43.1      123
3        56 55.9      775        622  9.5 35.9      105
4        28 51.0      137        176  8.7 15.2       89
5        14 68.4      136        529  8.8 54.5      116
6        46 47.6       44        116  8.8 33.4      135
:         :    :        :          :    :    :        :
:         :    :        :          :    :    :        :
```

Vamos verificar o comportamento das variáveis:



Vamos criar um modelo com as variáveis:

Pollution ~ Temp + Industry + Population + Wind + Rain + Wet.days

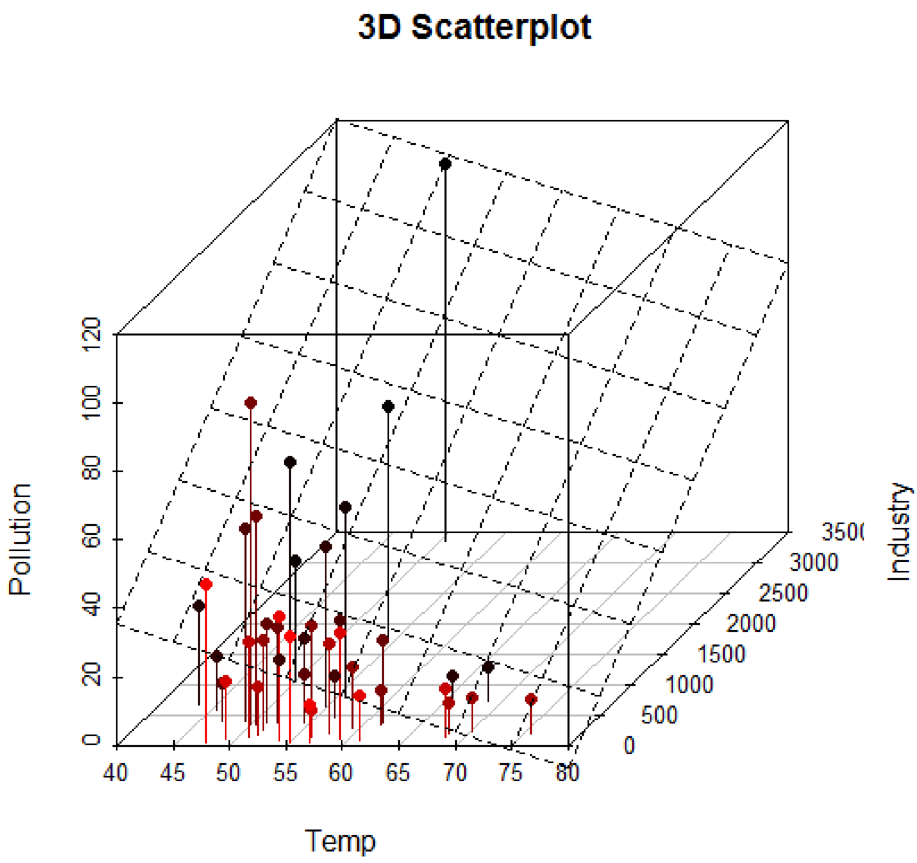
No R fica assim:

mod1<- lm(Pollution ~ Temp + Industry + Population + Wind + Rain + Wet.days, data=poluicao)
summary(mod1)

```
Call:
lm(formula = Pollution ~ Temp + Industry + Population + Wind +
    Rain + Wet.days, data = poluicao)

Residuals:
   Min      1Q Median      3Q     Max
-23.16  -8.52  -1.15   5.82   48.59

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 111.5936    47.1113    2.37   0.0237 *
Temp         -1.2651     0.6180   -2.05   0.0484 *
Industry      0.0654     0.0157    4.18   0.0002 ***
Population   -0.0397     0.0151   -2.64   0.0124 *
Wind         -3.1796     1.8140   -1.75   0.0886 .
Rain          0.5051     0.3612    1.40   0.1711
Wet.days     -0.0491     0.1612   -0.30   0.7626
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.6 on 34 degrees of freedom
Multiple R-squared:  0.672,     Adjusted R-squared:  0.615
F-statistic: 11.6 on 6 and 34 DF,  p-value: 0.000000472
```

Gráfico 3d usando as variáveis Pollution, Temp e Industry:



Bibliografia:

Bussab, Wilton de O., Morettin, Pedro A. Estatística Básica. 8. Ed. São Paulo: Saraiva, 2013.

Morettin, Luiz Gonzaga. Estatística Básica: Probabilidade e Inferência. Volume único. São Paulo: Ed. Pearson, 2011.

Belfiore, Patrícia, Estatística Aplicada a Administração, Contabilidade e Economia com Excel e SPSS. 1. Ed. Rio de Janeiro: Elsevier, 2015.

Pinheiro, João Ismael D. et al. Estatística Básica: a arte de trabalhar com dados. 2. Ed. Rio de Janeiro: Elsevier, 2015

Martins, Gilberto de Andrade. Estatística Geral e Aplicada. 3. Ed. São Paulo: Atlas, 2008.