Universidade Federal Rural do Rio de Janeiro

Profº: Felipe Leite Coelho da Silva
Profª: Josiane Cordeiro Coelho

Apostila de Estatística Aplicada à Economia e Administração

Capítulo 1

Introdução à Inferência

1.0.1 Conceitos Básicos

Definição 1 População: *conjunto de indivíduos (ou objetos, elementos) que tem em comum pelo menos uma variável observável que pode ser representada por uma v.a.* X.

Definição 2 Amostra: *é qualquer subconjunto da população.*

Definição 3 Parâmetro: *é uma medida usada para descrever uma característica da população.*

Definição 4 (Inferência Estatística) *É o uso de técnicas estatísticas para gerar afirmações sobre uma dada característica da população a partir de uma amostra.*
Por exemplo, *a média, a variância, o desvio padrão, a moda, etc.*

Em geral, nosso interesse é estimar o valor de um parâmetro ou estimar o valor de uma função de um parâmetro.

Definição 5 **Amostra Aleatória Simples (aas):** *Uma aas de tamanho* n *de uma v.a.* X *(população) é um conjunto de* n *v.a.'s* $X_1, X_2, \cdots, X_n$ *independentes e identicamente distribuídas (iid). Isto é, quando escolhemos ao acaso um subconjunto de tamanho* n *da população de forma que cada elemento da amostra seja v.a. com a mesma distribuição da população.*

Quando a população tem tamanho finito, então o sorteio é feito com reposição mantendo chance igual de qualquer elemento ser sorteado.

Definição 6 Estatística: *qualquer função da amostra $X_1, X_2, \cdots, X_n$ é dita ser uma estatística, isto é, T é estatística se $T = g(X_1, X_2, \cdots, X_n)$, para g sendo uma função qualquer.*

Assim, T também é uma variável aleatória. As inferências estatísticas são baseadas em estatísticas.

Exemplo 1 *Exemplos de estatísticas: Seja $X_1, \cdots, X_n$ uma aas de uma população X, podemos considerar as estatísticas*

$$\begin{aligned}
\bar{\mathbf{X}} &= \sum_{i=1}^{n} \frac{X_i}{n} && \textit{Média amostral} \\
\hat{\sigma}^2 &= \sum_{i=1}^{n} \frac{\left(X_i - \bar{X}\right)^2}{n} && \\
S^2 &= \sum_{i=1}^{n} \frac{\left(X_i - \bar{X}\right)^2}{n-1} && \textit{Variância amostral} \\
\mathbf{X_{(1)}} &= min\{X_1, \cdots, X_n\} && \textit{Mínimo} \\
\mathbf{X_{(n)}} &= max\{X_1, \cdots, X_n\} && \textit{Máximo}
\end{aligned}$$

Assim, T também é uma variável aleatória. As inferências estatísticas são baseadas em estatísticas.

Definição 7 Estimador: *uma estatística T é dita ser estimador de um parâmetro θ se utilizamos T para estimar θ.*

Notação: $T = \hat{\theta}$

Exemplo 2 *$\bar{X}$ é estimador para média populacional μ;*

Exemplo 3 *$\hat{\sigma}^2$ e S^2 são estimadores para variância populacional σ^2.*

Definição 8 Estimativa: *valor observado de um estimador na amostra.*

Propriedades dos estimadores

Algumas propriedades dos estimadores nos ajudam a escolher qual estimador utilizar para fazer a inferência.

- Estimador não viciado: Um estimador T de θ é dito ser não viciado se $E[T] = \theta$. Exemplos: $E[\bar{X}] = \mu$ e $E(S^2) = \sigma^2$, com $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$.

- Se T_1 e T_2 são dois estimadores não viciados de θ, diz-se que T_1 é mais eficiente do que T_2 se $Var(T_1) < Var(T_2)$.
- Estimador consistente: T_n é um estimador consistente para θ se

  $\lim_{n\to\infty} E[T_n] = \theta$ e $\lim_{n\to\infty} Var(T_n) = 0$,

  com T_n sendo uma estatística baseada na amostra de tamanho n.

1.0.2 Distribuição Amostral

A distribuição amostral de uma estatística T é a distribuição de todos os possíveis valores que ela pode assumir, calculados a partir de todas as possíveis amostras de mesmo tamanho.

Distribuição Amostral da Média

Teorema 1 *Seja $X_1, X_2, \cdots, X_n$ uma aas de tamanho n de uma população representada pela v.a. X com média μ e variância σ^2. Então, $E(\bar{X}) = \mu$ e $Var(\bar{X}) = \frac{\sigma^2}{n}$.*

Importante: Se $X \sim N(\mu, \sigma^2)$, então $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$.

Teorema 2 Teorema Central do Limite

Seja $X_1, X_2, \cdots, X_n$ uma aas de tamanho n de uma população X com média μ e variância σ^2. Então, a distribuição de $\bar{X}$ aproxima-se de uma distribuição Normal com média μ e variância $\frac{\sigma^2}{n}$ quando n tende ao infinito ($n \to \infty$). Assim, $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \approx N(0, 1)$.

Observação: Geralmente amostras de tamanho $n > 30$ fornecem uma aproximação razoável.

Exemplo 4 *A capacidade máxima de um elevador é de $600Kg$. Se a distribuição dos pesos dos usuários é $N(70, 100)$, qual é a probabilidade de que 8 pessoas ultrapassem esse limite?*

1.1 Intervalos de Confiança

Motivação Uma empresa deseja **estimar** a média de vendas, por estabelecimento, durante o último ano de um determinado produto. Sabemos que o desvio padrão populacional é de 200 reais. Suponha que temos disponível uma amostra de 25 estabelecimentos, com média de vendas de 5000 reais.

- Podemos usar $\bar{X}$ para estimar a verdadeira média desconhecida.
- Podemos obter alguma informação sobre a variabilidade desta estimativa (erro de amostragem)?

Queremos obter uma estimação intervalar fornecendo uma margem de erro, além da estimativa pontual que obtemos através dos estimadores dos parâmetros.

1.1.1 Intervalo de confiança para a média de uma população normal com variância conhecida

Seja $X \sim N(\mu, \sigma^2)$ com variância σ^2 conhecida. Se $X_1, \cdots, X_n$ é uma aas dessa população, então o intervalo de confiança de nível de confiança $(1-\alpha)\%$ para a média populacional μ é dado por

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

Importante:

- A margem de erro é dada por: $\epsilon = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$;
- α é o nível de significância, que é um valor no intervalo $(0, 1)$;
- $1-\alpha$ é o nível de confiança, que em geral é um valor alto (por exemplo, 0,90; 0,95; 0,99);
- Mesmo quando a população não é Normal, podemos utilizar este intervalo para a média se $n > 30$.

A ideia é que o intervalo contenha o verdadeiro valor do parâmetro na maioria das vezes (ou na maioria das amostras possíveis), isto é, com probabilidade $(1-\alpha)$. Após a amostra ser observada, ou o intervalo inclui o verdadeiro valor do parâmetro ou não inclui.

Para entendermos a notação $z_{\alpha/2}$, considere o nível de confiança do intervalo $1-\alpha = 0,95$, portanto $\alpha = 0,05$ e $\alpha/2 = 0,025$. Então, $z_{\alpha/2} = z_{0,025}$ é tal que $P(Z \geq z_{0,025}) = 0,025$. A figura 1 apresenta esse exemplo.

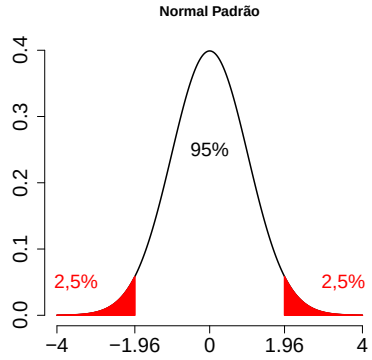


Figura 1.1: Suponha $\alpha = 0,05$, portanto $1 - \alpha = 0,95$ e $z_{\alpha/2} = 1,96$.

No *software* R, podemos utilizar a função *qnorm(p)* (que retorna o valor de z que acumula até ele probabilidade igual a p) para encontrar o valor de $z_{0,025}$ (que tem 0,025 de probabilidade acima dele e, portanto, 0,975 de probabilidade até dele), simplesmente fazendo *qnorm(0.975)*.

Exemplo 5 *Uma empresa deseja* **estimar** *a média de vendas, por estabelecimento, durante o último ano de um determinado produto. Sabemos que o desvio padrão populacional é de 200 reais. Suponha que temos disponível uma amostra de 25 estabelecimentos, com média de vendas de 5000 reais.*

Solução:

- *Estimativa pontual para média de vendas:* $\bar{X} = 5000$*;*
- *Variância da média das vendas:* $\frac{200^2}{25} = 1600 reais^2$*;*
- *Supondo normalidade, para um nível de confiança de 95%, temos que*

$$\epsilon = z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 1,96\frac{200}{\sqrt{25}} = 78,4;$$

- *Logo, o ic de 95% para a média de vendas é dado por:*

$$[5000 - 78,4; 5000 + 78,4] = [4921,6; 5078,4]\,.$$

Ideia básica da construção:

Seja $X_1, \cdots, X_n$ uma aas da distribuição Normal(μ, σ^2), com σ^2 conhecido.

Podemos definir α tal que $P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$ (intervalo simétrico é o de menor comprimento).
Sabemos que $Z = \dfrac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \sim$ Normal$(0, 1)$, logo podemos escrever

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

Portanto, supondo σ^2 conhecido, **o intervalo para μ com coeficiente de confiança** $1 - \alpha$ é dado por

$$[\bar{X} - \epsilon; \bar{X} + \epsilon], \qquad \text{em que a margem de erro } \epsilon = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

Intervalo de confiança para a proporção populacional

Considere uma população em que a proporção de elementos com certa característica é p. Podemos definir uma v.a.:

$$X = \begin{cases} 1, & \text{se o elemento possui a característica;} \\ 0, & \text{se o elemento não possui a característica.} \end{cases}$$

Logo, $X \sim Bernoulli(p)$, com média $E[X] = p$ e variância $Var(X) = p(1 - p)$.
Suponha $X_1, \cdots, X_n$ uma aas de X, logo

$$Y = \sum_{i=1}^{n} X_i \sim Bin(n, p).$$

Podemos estimar a proporção populacional de portadores da caraterística, utilizando a estatística

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

que é um estimador não viciado para p. Pelo TCL, temos que

$$\hat{p} = \bar{X} \dot{\sim} Normal\left(p, \frac{p(1-p)}{n}\right),$$

em que a aproximação é boa se $np \geq 5$ e $n(1 - p) \geq 5$. Daí, podemos proceder usando a mesma ideia do i.c. para média e obter o i.c. para proporção.

O intervalo de confiança para proporção populacional p com $(1-\alpha)\%$ de confiança é dado por

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

Exemplo 6 *Um levantamento com 1018 adultos, concluiu que 255 deles planejaram gastar menos dinheiro em presentes durante a época de férias de 2018 em comparação ao ano anterior.*

a) Qual é a estimativa da proporção de todos os adultos que planejaram gastar menos dinheiro em presentes durante a época de férias de 2018?

b) Usando uma confiança de 95%, qual é a margem de erro associada a essa estimativa?

Intervalo de confiança para média de uma população normal com variância desconhecida

Seja $X_1,\cdots,X_n$ uma aas da distribuição Normal(μ,σ^2), com σ^2 desconhecido. Utilizaremos S^2 para estimar σ^2. Podemos usar a estatística

$$T=\frac{(\bar{X}-\mu)}{S/\sqrt{n}}\sim t_{(n-1)},$$

isto é, a variável T tem distribuição t-Student com $n-1$ graus de liberdade. Logo, o intervalo de confiança para média μ (variância desconhecida) com $(1-\alpha)\%$ de confiança é dado por

$$IC\left(\mu;(1-\alpha)\right)=\left[\bar{X}-t_{(n-1);\alpha/2}\frac{S}{\sqrt{n}};\bar{X}+t_{(n-1);\alpha/2}\frac{S}{\sqrt{n}}\right].$$

Definição 9 *Uma variável aleatória contínua Y é dita seguir a distribuição t-Student com ν graus de liberdade se sua função de densidade de probabilidade é dada por*

$$p(y\mid\nu)=\frac{\Gamma(\frac{\nu+1}{2})\nu^{\nu/2}}{\Gamma(\frac{\nu}{2})(\pi)^{1/2}}\left[\nu+y^2\right]^{-(\nu+1)/2},\qquad \textit{para } -\infty<y<\infty, \tag{1.1}$$

com $\nu>0$.

O gráfico da t-Student também é simétrico, podemos definir α tal que $P(-t_{(n-1,\alpha/2)}<T<t_{(n-1,\alpha/2)})=1-\alpha$.

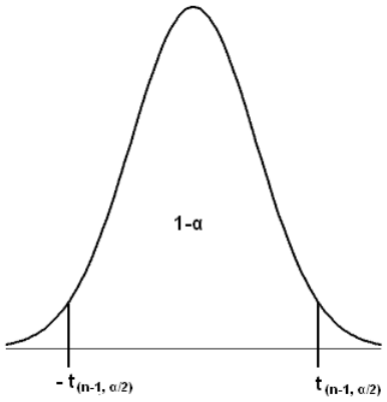


Figura 1.2: Suponha $\alpha = 0,05$ e $n = 36$, logo $t_{(n-1,\alpha/2)} = 2,03$.

Importante:

- A distribuição t-Student aproxima-se da distribuição normal, conforme o valor dos graus de liberdade crescem. Portanto, para graus de liberdade altos (logo, para n alto), podemos aproximar a t pela normal. Podemos perceber que para valores do grau de liberdade maiores do que 30, as duas disribuições são muito parecidas.

- Portanto, com σ^2 desconhecido, para amostras grandes, podemos apenas substituir σ^2 por S^2 no intervalo que depende de $z_{\alpha/2}$.

- O conceito de **graus de liberdade** está relacionado ao número de dados disponíveis (livres) para o calculo da estatística. Por exemplo, ao estimarmos a média populacional com a média amostral perdemos um grau de liberdade. Assim, a estatística t-Student terá $n - 1$ graus de liberdade.

No *software* R, podemos utilizar a função *qt(*p*, df)* (que retorna o valor de t que acumula até ele probabilidade igual a p para a t-Student com df graus de liberdade) para encontrar o valor de $t_{0,025}$ (que tem 0,025 de probabilidade acima dele e, portanto, 0,975 de probabilidade até dele), simplesmente fazendo *qt(0.975, 24)*.

Exemplo 7 *Um empresa deseja estimar o intervalo de confiança de vendas, por estabelecimento, durante o último ano de um determinado produto. Para uma amostra de 25 estabelecimentos, obteve-se uma média de 5000 reais e* $S^2 = 160000\ reais^2$*. Determine o intervalo de confiança de 95%, dado que os valores de venda são considerados normalmente distribuídos com desvio padrão desconhecido.*

Solução:

$$IC\left(\mu; 95\%\right) = \left[5000 - 2,06\frac{400}{\sqrt{25}}; 5000 + 2,06\frac{400}{\sqrt{25}}\right]$$

Intervalo de confiança para variância populacional

Seja uma amostra aleatória simples $X_1, X_2, \cdots, X_n$ retirada de uma população normal com média μ e variância σ^2. O objetivo é construir um intervalo de confiança para σ^2. A hipótese de normalidade da população é fundamental.

Podemos deduzir o ic para σ^2 **(desconhecido)** através da estatística (pivô)

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

que envolve o parâmetro de interesse e o seu estimador.

Assim, o intervalo de confiança para σ^2 de nível de confiança $(1-\alpha)\%$ é dado por

$$\left[\frac{(n-1)S^2}{\chi_2}; \frac{(n-1)S^2}{\chi_1}\right].$$

Embora o gráfico da distribuição qui-quadrado não seja simétrico, podemos ainda considerar o intervalo dado por

$$P\left(\chi_1 < \frac{(n-1)S^2}{\sigma^2} < \chi_2\right) = 1 - \alpha.$$

Note que $\chi_1 \neq \chi_2$.

Exemplo 8 *Retornando a Motivação: Uma empresa deseja* **estimar** *a média de vendas, por estabelecimento, durante o último ano de um determinado produto. Suponha que temos disponível uma amostra de 25 estabelecimentos, com desvio padrão amostral de 200 reais. Podemos obter o ic de 95% para a variância de vendas.*

Solução: *identificando que* $n = 25$ *e* $S^2 = 200^2$*, temos,* $(n-1)S^2 = 960000$*. E, pela tabela da distribuição qui-quadrado* $\chi_1 = 12,401$ *e* $\chi_2 = 39,364$*.*

Portanto,

$$\left[\frac{(n-1)S^2}{\chi_2}; \frac{(n-1)S^2}{\chi_1}\right] = [24387,77; 77413,11].$$

No *software* R, podemos utilizar a função *qchisq(*p*, df)* (que retorna o valor que acumula até ele probabilidade igual a p para a qui-quadrado com df graus de liberdade) para encontrar o valor de χ_1 (que tem 0,025 de probabilidade até ele), simplesmente fazendo *qchisq(0.025, 24)* e de χ_2 (que tem 0,975 de probabilidade até ele), simplesmente fazendo *qchisq(0.975, 24)*.

Intervalo de confiança para diferença de duas médias populacionais: Amostras independentes

Sejam $X_1, \cdots, X_{n_X}$ uma aas da distribuição Normal(μ_X, σ_X^2) e $Y_1, \cdots, Y_{n_Y}$ uma aas da distribuição Normal(μ_Y, σ_Y^2), amostras independentes.

Daí, temos que $E[\bar{X} - \bar{Y}] = \mu_X - \mu_Y$, e $Var(\bar{X} - \bar{Y}) = \dfrac{\sigma_X^2}{n_X} + \dfrac{\sigma_Y^2}{n_Y}$. E, ainda

$$\bar{X} - \bar{Y} \sim Normal\left(\mu_X - \mu_Y, \frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}\right).$$

Logo, um pivô para a diferença entre as médias populacionais $\mu_X - \mu_Y$ é

$$\frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}} \sim \text{ Normal}(0, 1).$$

O intervalo de confiança de nível de confiança $(1-\alpha)$ para diferença entre as duas médias populacionais $(\mu_X - \mu_Y)$, com variâncias conhecidas, é dado por

$$IC\left(\mu_X - \mu_Y; 1-\alpha\right) = \left[(\bar{X} - \bar{Y}) - z_{\alpha/2}\, \sigma_{X-Y}; (\bar{X} - \bar{Y}) + z_{\alpha/2}\, \sigma_{X-Y}\right],$$

tal que

$$\sigma_{X-Y} = \sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}.$$

Supondo as variâncias desconhecidas, com $n_X < 30$ e/ou $n_Y < 30$, o intervalo de confiança de nível de confiança $(1-\alpha)$ para diferença entre as duas médias populacionais $(\mu_x - \mu_y)$ é dado por

$$IC\left(\mu_X - \mu_Y; 1-\alpha\right) = \left[(\bar{X} - \bar{Y}) - \epsilon; (\bar{X} - \bar{Y}) + \epsilon\right],$$

tais que $\epsilon = t_{\nu,\alpha/2} S_p$,

- para variâncias iguais $(\sigma_X^2 = \sigma_Y^2 = \sigma^2)$: $S_p = S_c\sqrt{\dfrac{1}{n_X} + \dfrac{1}{n_Y}}, S_c = \sqrt{\dfrac{(n_X - 1)S_X^2 + (n_Y - 1)S_Y^2}{n_X + n_Y - 2}}$ e $\nu = n_X + n_Y - 2$ graus de liberdade.
- para variâncias diferentes $(\sigma_x^2 \neq \sigma_y^2)$: $S_p = \sqrt{\dfrac{S_X^2}{n_X} + \dfrac{S_Y^2}{n_Y}}$ e $\nu = \dfrac{(A+B)^2}{\dfrac{A^2}{n_X - 1} + \dfrac{B^2}{n_Y - 1}}$, sendo $A = \dfrac{S_X^2}{n_X}$ e $B = \dfrac{S_Y^2}{n_Y}$.

Exemplo 9 *Considere que em uma amostra com 50 funcionários de uma empresa A, o salário médio por hora foi de 15 reais com desvio padrão amostral de 2 reais. Em uma segunda empresa B, o salário médio por hora de 40 funcionários foi de 18 reais com desvio padrão amostral de 2,40 reais. Queremos estimar a diferença dos salários médios por hora das duas empresas.*

Solução:

- *Suponha igualdade das variâncias populacionais* $(\sigma_A^2 = \sigma_B^2)$ *e um nível de 95% de confiança.*
- *Daí,*
  - $(\bar{X} - \bar{Y}) = 18 - 15 = 3$*;*
  - $S_c^2 = \dfrac{(n_X - 1)S_X^2 + (n_Y - 1)S_Y^2}{n_X + n_Y - 2} = \dfrac{39(2,40^2) + 49(2^2)}{50 + 40 - 2} = 4,78;$
  - $\epsilon = t_{\nu,\alpha/2} S_c \sqrt{\dfrac{1}{n_X} + \dfrac{1}{n_Y}} = t_{50+40-2;0,025}\sqrt{4,78}\sqrt{\dfrac{1}{50} + \dfrac{1}{40}} = 2 * 2,186321 * 0,045 = 0,1967689$*, com* $t_{88;0,025} = 2$*;*
  - $\left[(\bar{X} - \bar{Y}) - \epsilon; (\bar{X} - \bar{Y}) + \epsilon\right] = [3-0,1967689; 3+0,1967689] = [2,803231; 3,1967689]$.
- *Como* $n_X > 30$ *e* $n_Y > 30$*, podemos aproximar pelo ic dado quando* σ *é conhecido e substituí-lo por seu estimador* S_p*.*

Exemplo 10

Exemplo 3: Dois processos de conservação de alimentos estão sendo investigados, e a variável de interesse é o tempo de duração do produto. No processo A, o tempo de duração segue a distribuição $N(\mu_A, 100)$*, e no processo B o tempo de duração segue a distribuição* $N(\mu_B, 81)$*. Foram selecionadas duas amostras independentes: 16 latas do processo A com tempo médio de duração de 60 dias; e no processo B foram selecionadas 25 latas com duração média de 50 dias. Para verificar se os dois processos podem ter o mesmo desempenho, decidiu-se construir um IC de 99% para diferença de* $\mu_A - \mu_B$*. Caso o zero pertença ao intervalo, pode-se concluir que existe evidência de igualdade dos processos.*

Exemplo 11 *Dois processos de conservação de alimentos estão sendo investigados, e a variável de interesse é o tempo de duração. No processo A, o tempo de duração segue a distribuição* $N(\mu_A, \sigma^2)$*, e no processo B o tempo de duração segue a distribuição*

$N(\mu_B, \sigma^2)$. *Foram selecionados duas amostras independentes: 16 latas do processo A com tempo médio de duração de 60 dias; e no processo B foram selecionadas 25 latas com duração média de 50 dias. Para verificar se os dois processos podem ter o mesmo desempenho, decidiu-se construir um IC de 95% para diferença de* $\mu_A - \mu_B$*. Caso o zero pertença ao intervalo, pode-se concluir que existe evidência de igualdade dos processos. Suponha que* $S_A = 3$ *e* $S_B = 2$*.*

Quando as populações não seguem a distribuição Normal, então, se ambas as amostras tiverem tamanhos maiores do que 30, podemos utilizar o resultado do TCL e usar os intervalos de confiança derivados anteriormente. Caso as variâncias sejam desconhecidas, podemos usar os estimadores definidos por S_p (variâncias iguais ou diferentes) e, ainda, utilizar o valor crítico $z_{\alpha/2}$ ao invés de $t_{\nu,\alpha/2}$ para simplificar os cálculos. Isso é devido a t-Student aproximar-se da Normal quando seus graus de liberdade aumentam.

Intervalo de confiança para amostras dependentes (emparelhadas)

Sejam $X_1, X_2, \cdots, X_n$ uma aas da população $X \sim N(\mu_X, \sigma_X^2)$ e $Y_1, Y_2, \cdots, Y_n$ uma aas da população $Y \sim N(\mu_Y, \sigma_Y^2)$. Suponha que haja dependência entre as amostras.
Defina a variável aleatória: $D = X - Y$, logo temos uma aas

$$D_1 = X_1 - Y_1, D_2 = X_2 - Y_2, \cdots, D_n = X_n - Y_n$$

e podemos obter:
$\bar{D} = \frac{1}{n}\sum_{i=1}^{n} D_i$ e $S_D^2 = \frac{1}{n-1}\sum_{i=1}^{n}(D_i - \bar{D})^2$,
que são os estimadores da média e da variância populacional dos desvios.

Construção:

Temos que: $D \sim N(\mu_D, \sigma_D^2)$, em que $\mu_D = \mu_X - \mu_Y$ e σ_D^2 é desconhecido porque não conseguimos calcular a $Var(X - Y)$ diretamente (devido a dependência).

Logo, $\bar{D} \sim N(\mu_D, \sigma_D^2/n)$. Além disso, $\frac{(\bar{D} - \mu_D)}{S_D/\sqrt{n}} \sim t_{(n-1)}$.
Daí, simplificamos o problema para o caso de uma amostra de uma normal com média e variância desconhecidos.
Portanto, o intervalo de confiança de $(1-\alpha)\%$ para a média μ_D da população D é dado

por:

$$IC\left(\mu;(1-\alpha)\right)=\left[\bar{D}-t_{(n-1);\alpha/2}\frac{S_D}{\sqrt{n}};\bar{D}+t_{(n-1);\alpha/2}\frac{S_D}{\sqrt{n}}\right].$$

Exemplo 12 *Considere o tempo de treinamento de sete mecânicos utilizando dois métodos* (M_1, M_2) *de produção.*

Mecânico	M_1	M_2
1	6	5,4
2	5	5,2
3	7	4
4	6,2	3
5	10	5
6	8	4
7	10	4

Construa o intervalo de confiança de 95% para essa estimativa de diferença entre as duas médias de tempo de treinamento. Verifique se existe diferença entre os métodos.

Intervalo de confiança para variância - duas amostras

Vamos agora ver o caso que temos duas amostras independentes vindas de populações com mesma variância. Sejam $X_1,\cdots,X_{n_X}$ uma aas da distribuição Normal(μ_X,σ^2) e $Y_1,\cdots,Y_{n_Y}$ uma aas da distribuição Normal(μ_Y,σ^2), amostras independentes. Neste caso, o intervalo de $(1-\alpha)$% de confiança para a σ^2 é baseado no pivô (que envolve o parâmetro σ^2 e seu estimador S_p^2)

$$\frac{(n_X-1)S_X^2+(n_Y-1)S_Y^2}{\sigma^2}\sim\chi^2_{n_X+n_Y-2}.$$

Daí, obtemos que **o intervalo para** σ^2 **com coeficiente de confiança** $1-\alpha$ é dado por

$$\left[\frac{(n_X+n_Y-2)S_p^2}{\chi_2^2};\frac{(n_X+n_Y-2)S_p^2}{\chi_1^2}\right],$$

em que $S_p=\sqrt{\dfrac{S_X^2}{n_X}+\dfrac{S_Y^2}{n_Y}}$.

Para finalizar, vamos ver um intervalo de confiança para o quociente entre duas variâncias de amostras independentes. Sejam $X_1, \cdots, X_{n_X}$ uma aas da v.a. $X \sim$ Normal(μ_X, σ_X^2) e $Y_1, \cdots, Y_{n_Y}$ uma aas da v.a. $Y \sim$ Normal(μ_Y, σ_Y^2), então o intervalo de $(1-\alpha)$% de confiança para a σ_X^2/σ_Y^2 é baseado no pivô

$$\frac{S_X^2 \sigma_Y^2}{S_Y^2 \sigma_X^2} \sim F_{(n_X-1, n_Y-1)},$$

em que $F_{(n_X-1, n_Y-1)}$ denota a distribuição F-Snedecor com $n_X - 1$ e $n_Y - 1$ graus de liberdade.

Obtemos que **o intervalo de $(1-\alpha)$% de confiança para a σ_X^2/σ_Y^2** é

$$\left[\frac{1}{F_2}\frac{S_X^2}{S_Y^2}; \frac{1}{F_1}\frac{S_X^2}{S_Y^2}\right].$$

Tabela 1.1: Probabilidades $p = P[0 \leq Z \leq z]$ da Distribuição Normal padrão com valores de z dados nas margens da tabela

z	0	1	2	3	4	5	6	7	8	9
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990
3.1	0.4990	0.4991	0.4991	0.4991	0.4992	0.4992	0.4992	0.4992	0.4993	0.4993
3.2	0.4993	0.4993	0.4994	0.4994	0.4994	0.4994	0.4994	0.4995	0.4995	0.4995
3.3	0.4995	0.4995	0.4995	0.4996	0.4996	0.4996	0.4996	0.4996	0.4996	0.4997
3.4	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4998
3.5	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998
3.6	0.4998	0.4998	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999
3.7	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999
3.8	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999
3.9	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000

Capítulo 2

Testes de Hipóteses

Na teoria de estimação foi visto que através de estatísticas amostrais adequadas pode-se estimar parâmetros de uma população, dentro de um certo intervalo de confiança. Nos testes de hipóteses, ao contrário de se construir um intervalo de confiança para o parâmetro da população, testa-se a veracidade de uma afirmação sobre um parâmetro da população. Então, a partir do teste de hipótese, procura-se tomar decisões a respeito de uma população, com base em informações obtidas de amostras desta mesma população. Em muitas situações temos interesse em tomar a decisão de aceitar ou rejeitar determinada afirmação baseando-se em um conjunto de evidências.

Motivação

1. Decidir sobre a eficiência ou não de certa vacina utilizada no combate à determinada doença. Os pesquisadores formulam então as hipóteses:

   - a vacina não é eficiente (H_0);
   - a vacina é eficiente (H_1).

   Com base nas evidências, os pesquisadores terão que decidir entre H_0 e H_1.

2. Um exemplo comum é o caso em que um indivíduo está sendo julgado por determinado delito. Com base nas evidências (testemunhas, fatos etc.), o júri terá que decidir pela culpa ou inocência do indivíduo.

   - o indivíduo é inocente (H_0);
   - o indivíduo é culpado(H_1).

Com base nas evidências apresentadas, o júri terá que se decidir por H_0 ou por H_1.

Regra de decisão

Um teste de uma hipótese estatística é baseado em uma **regra de decisão** que leva à ação de considerar a hipótese H_0 como verdadeira ou à ação de considerar a hipótese H_1 como verdadeira.

Decisão tomada	Realidade	
	H_0 é verdadeira	H_0 é falsa
Aceitar H_0 (Rejeitar H_1)	Decisão correta $(1-\alpha)$	Erro tipo II (falso-negativo) (β)
Rejeitar H_0 (Aceitar H_1)	Erro tipo I (falso-positivo) (α)	Decisão correta $(1-\beta)$

Tabela 2.1: Tipos de erros.

- Existe um equilíbrio entre esses dois tipos de erros, no sentido de que ao tentar-se minimizar α, aumenta-se β.
- Logo, não é possível minimizar estas duas probabilidades simultaneamente.
- Fixamos um valor pequeno para α.
- Recomenda-se que a escolha de H_0 seja feita de forma que se rejeitada, quando for verdadeira, leve a um erro do tipo I mais importante de ser evitado.
- No exemplo, uma perda maior seria rejeitar a hipótese da vacina não ser eficiente (aceitando que ela seja eficiente) quando, na verdade, ela não é!
- Se trocássemos as hipóteses, o erro tipo I seria a rejeição da hipótese da vacina ser eficiente (aceitando que ela seja ineficiente) quando, na verdade, ela é eficiente.

Formulação Estatística

- Seja $\mathbf{X} = (X_1, \cdots, X_n)$ uma aas de uma população X que segue uma distribuição que depende de $\theta \in \Theta$.

Definição 10 *Chamamos de hipótese estatística qualquer afirmação sobre a distribuição de probabilidades de uma ou mais variáveis aleatórias.*

- **Hipóteses**:
  - $H_0 : \theta \in \Theta_0$
  - $H_1 : \theta \in \Theta_1$
- Θ_0 e Θ_1 são subconjuntos disjuntos de Θ.
- **Região crítica (RC)**: região com os valores da aas que levam a rejeição de H_0.
- **Região de aceitação ($\bar{RC}$)**: região com os valores da aas que levam a aceitação de H_0.
- Daí:

$$P(\mathbf{X} \in RC \mid \theta \in \Theta_0) = \alpha$$

$$P(\mathbf{X} \in \bar{RC} \mid \theta \in \Theta_1) = \beta$$

Ideia da formulação

- Estabeleça a hipótese a ser testada (H_0) e a alternativa (H_1);
- Determine a estatística do teste $T(\mathbf{X})$;
- Escolha a probabilidade do erro do tipo I (α);
- Construa a região crítica RC (sob H_0);
- Calcule $T(\mathbf{X})$ a partir da aas $\mathbf{X}$;
- Conclusão:
  - Se $T(\mathbf{X}) \in RC$, então rejeite H_0;
  - senão, não rejeite H_0.

Exemplo 13 *Uma companhia de serviços de ônibus intermunicipais planejou uma rota para servir vários locais situados entre duas cidades. Um estudo preliminar afirma que a duração das viagens pode ser considerada uma v.a. com distribuição normal, com média igual a 300 minutos e desvio padrão de 30 minutos. As 10 primeiras viagens realizadas nessa rota apresentaram média igual a 314 minutos. Esse resultado comprova ou não o tempo médio determinado nos estudos preliminares?*

Solução:

- **Hipóteses***:*
  - $H_0 : \mu = 300$
  - $H_1 : \mu \neq 300$

  em que μ *é o tempo médio.*
- $T(\mathbf{X}) = \bar{X}$ *é a estatística de teste, com* $\bar{X} \sim N(\mu, 30^2/10)$*, assumindo a variância conhecida e igual a* $30^2 = 900$*.*
- *Fixe* $\alpha = 0,05$.
- *Sob a hipótese de que* H_0 *é verdadeira, temos que*

$$\bar{X} \sim N(300, 900/10) \Rightarrow Z = (\bar{X} - 300)/\sqrt{90} \sim N(0, 1).$$

- *Pela hipótese alternativa, temos que* H_0 *deve ser rejeitada se* $\bar{X}$ *for muito menor ou muito maior do que 300.*
- *Para* $\alpha = 0,05$*, temos que*

$$P(\bar{X} < x_1) = 0,025 = P(\bar{X} > x_2),$$

  logo

$$P\left(\frac{\bar{X} - 300}{\sqrt{90}} < \frac{x_1 - 300}{\sqrt{90}}\right) = 0,025 = P\left(\frac{\bar{X} - 300}{\sqrt{90}} > \frac{x_2 - 300}{\sqrt{90}}\right),$$

  portanto:
  - $(x_1 - 300)/\sqrt{90} = -1,96 \Rightarrow x_1 \approx 281,41$;
  - $(x_2 - 300)/\sqrt{90} = 1,96 \Rightarrow x_2 \approx 318,59$;
- $RC = \{\bar{x} \in \mathbb{R}; \bar{x} \leq 281,41$ *ou* $\bar{x} \geq 318,59\}$.
- *Conclusão: como* $\bar{X} = 314 \notin RC$*, então aceita-se* H_0*, isto é, o resultado comprova os estudos preliminares.*

Função Poder

- Útil para verificar a adequação de um teste ou para comparar dois ou mais testes.

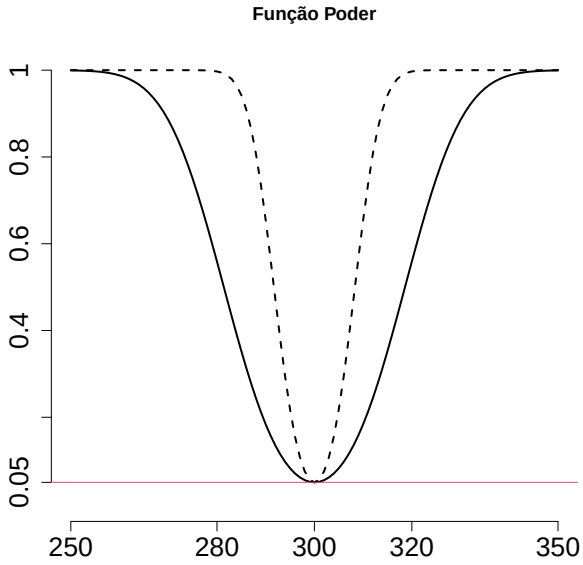


Figura 2.1: Função poder para o exemplo, com $n = 10$ (linha cheia) e $n = 50$ (linha tracejada).

Definição 11 *A função poder $\pi(\theta)$ associa a cada valor possível de θ a probabilidade de rejeitar H_0, isto é,*

$$\pi(\theta) = P(\textit{rejeitar } H_0 \mid \theta) = P(\mathbf{X} \in RC \mid \theta);$$

- Ideal: $\pi(\theta) = 0$, para θ satisfazendo H_0 e $\pi(\theta) = 1$, para θ satisfazendo H_1;
- No exemplo, a função poder sob H_0 (para $\mu = 300$) é

$$\begin{aligned} \pi(\mu) &= P(\text{rejeitar } H_0 \mid \mu = 300) \\ &= P(\bar{x} \leq 281,41 \text{ ou } \bar{x} \geq 318,59 \mid \mu = 300) = \alpha = 0,05. \end{aligned}$$

- Podemos calcular $\pi(\mu)$ para diferentes valores de μ e obter a função poder (isto é, determinando os pontos críticos para cada caso).
- Para valores de θ dentro da região crítica, essa probabilidade corresponde à probabilidade de um acerto.

Probabilidade de significância ou nível descritivo, ou ainda
p-valor (valor p) do teste

Se a hipótese nula for verdadeira, o valor p de um teste de hipótese é a probabilidade de se obter uma estatística amostral com valores tão extremos do que aquela determinada

com os dados da amostra.

Importante: Devemos rejeitar a hipótese nula H_o ao nível de significância α sempre que o valor de p for menor ou igual a α, ou seja, rejeitamos $H_o \Leftrightarrow p \leq \alpha$. Seguem alguns pontos importantes:

- É a probabilidade de significância (ou nível descritivo) do teste.
- Para formular o teste, precisamos fixar α.
- Uma maneira alternativa de se proceder é a partir do cálculo do p-valor.
- Em geral, se $T(\mathbf{X})$ é uma estatística de teste e H_0 é rejeitada, por exemplo, para $T(\mathbf{X}) > c$ então o p-valor é a probabilidade $P(T > t \mid H_0)$ onde $t(\mathbf{x})$ é o valor observado de $T(\mathbf{X})$.
- p-valor é a probabilidade de observar resultados tão extremos quanto aqueles que foram obtidos se a hipótese nula for verdadeira;
- p-valor $>> \Rightarrow$ evidência de que H_0 é verdadeira;
- p-valor $<< \Rightarrow$ evidência nos dados contra H_0;
- Para testes bilaterais, podemos considerar o p-valor bilateral igual a duas vezes o p-valor unilateral, que é razoável quando a distribuição da estatística é simétrica (por exemplo, normal, t).

Exemplo 14 *Uma companhia de serviços de ônibus intermunicipais planejou uma rota para servir vários locais situados entre duas cidades. Um estudo preliminar afirma que a duração das viagens pode ser considerada uma v.a. com distribuição normal, com média igual a 300 minutos e desvio padrão de 30 minutos. As 10 primeiras viagens realizadas nessa rota apresentaram média igual a 314 minutos. Esse resultado comprova ou não o tempo médio determinado nos estudos preliminares?*

Solução:

- **Hipóteses***:*
  - $H_0 : \mu = 300$
  - $H_1 : \mu \neq 300$

– $T(\mathbf{X}) = \bar{X}$ *é a estatística de teste, com* $\bar{X} \sim N(\mu, \sigma^2/10)$*.*

– *Sob a hipótese de que* H_0 *é verdadeira, temos que*

$$\bar{X} \sim N(300, 900/10) \Rightarrow Z = (\bar{X} - 300)/\sqrt{90} \sim N(0, 1),$$

assumindo a variância conhecida e igual a $30^2 = 900$*.*

Como $\bar{X} = 314$*, podemos encontrar a probabilidade de ocorrerem amostras com valores de* $\bar{X}$ *mais extremos do que esse:*

$$\begin{aligned} P(\bar{X} > 314) &= P\left(\frac{(\bar{X} - 300)}{\sqrt{90}} > \frac{(314 - 300)}{\sqrt{90}}\right) \\ &\approx P(Z > 1,48) = 0,5 - P(0 < Z < 1,48) = 0,07. \end{aligned}$$

Logo, p*-valor=*$2 \times 0,07 = 0,14$*, que não é um valor tão pequeno, indicando que não existe evidência contra* H_0*. Portanto, concordando com os estudos preliminares.*

p-valor	Evidência contra H_0
0,1	Sem
0,05	Moderada
0,025	Substancial
0,01	Forte
0,005	Muito forte
0,001	Fortíssima

Tabela 2.2: Escala de significância de Fisher.

Teste de hipótese sobre a Média μ **da** $N(\mu, \sigma^2)$

Seja $X_1, X_2, \cdots, X_n$ uma aaas de uma população $N(\mu, \sigma^2)$, com σ^2 conhecido. O teste de hipótese sobre a média μ é:

- Teste bilateral ($H_0 : \mu = \mu_0; H_1 : \mu \neq \mu_0$)

$$RC(\mu; (1-\alpha)\%) : |z^*| > z_{\alpha/2}, ;$$

- Teste unilateral à esquerda ($H_0 : \mu = \mu_0; H_1 : \mu < \mu_0$)

$$RC(\mu; (1-\alpha)\%) >: z^* < -z_{\alpha}, ;$$

- Teste unilateral à direita ($H_0 : \mu = \mu_0; H_1 : \mu > \mu_0$)

$$RC(\mu; (1-\alpha)\%) : z^* > z_\alpha, ;$$

com $z^* = \dfrac{\left(\bar{X} - \mu_0\right)\sqrt{n}}{\sigma}$. Podemos usar o resultado anterior para σ^2 desconhecido quando $n \geq 30$ (nesse caso, não é necessário ser uma população normal) apenas trocando σ por S, em que $S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$. Quando σ^2 é desconhecido e $n < 30$, o teste bilateral ($H_0 : \mu = \mu_0; H_1 : \mu \neq \mu_0$) é dado por

$$RC(\mu; (1-\alpha)\%) : |t^*| > t_{(n-1),\alpha/2}, \text{ com } t^* = \frac{\left(\bar{X} - \mu_0\right)\sqrt{n}}{S}.$$

No R, podemos usar a função *z.test*.

Teste sobre a Proporção p

Pelo TCL, temos que

$$\hat{p} = \bar{X} \dot{\sim} Normal\left(p, \frac{p(1-p)}{n}\right),$$

em que a aproximação é boa se $np \geq 5$ e $n(1-p) \geq 5$. Portanto,

- Teste bilateral ($H_0 : p = p_0; H_1 : p \neq p_0$)

$$RC(p; (1-\alpha)\%) : |z^*| > z_{\alpha/2};$$

- Teste unilateral à esquerda ($H_0 : p = p_0; H_1 : p < p_0$)

$$RC(p; (1-\alpha)\%) >: z^* < -z_\alpha;$$

- Teste unilateral à direita ($H_0 : p = p_0; H_1 : p > p_0$)

$$RC(p; (1-\alpha)\%) : z^* > z_\alpha;$$

com $z^* = \dfrac{(\hat{p} - p_0)}{\sqrt{\frac{p_0(1-p_0)}{n}}}$.

Teste de hipótese para a variância σ^2 **da** $N(\mu, \sigma^2)$

Utilizaremos a seguinte estatística do teste: $\chi^2 = \dfrac{(n-1)}{\sigma^2} S^2 \sim \chi^2_{(n-1)}$.

- Teste bilateral ($H_0 : \sigma^2 = \sigma_0^2; H_1 : \sigma^2 \neq \sigma_0^2$)

$$RC(\sigma^2; (1-\alpha)\%) : \{\chi_0^2 \in \Re | \chi_0^2 < \chi_{1,\alpha/2} \text{ ou } \chi_0^2 > \chi_{2,\alpha/2}\};$$

- Teste unilateral à esquerda ($H_0 : \sigma^2 = \sigma_0^2; H_1 : \sigma^2 < \sigma_0^2$)

$$RC(\sigma^2; (1-\alpha)\%) >: \{\chi_0^2 \in \Re | \chi_0^2 < \chi_{1,\alpha}\};$$

- Teste unilateral à direita ($H_0 : \sigma^2 = \sigma_0^2; H_1 : \sigma^2 > \sigma_0^2$)

$$RC(\sigma^2; (1-\alpha)\%) >: \{\chi_0^2 \in \Re | \chi_0^2 > \chi_{2,\alpha}\}\};$$

em que $\chi_0^2 = \dfrac{(n-1)}{\sigma_0^2} S^2$.

Teste para comparar as variâncias de duas populações normais

Sejam $X_1, \cdots, X_{n_X}$ uma aas da distribuição Normal(μ_X, σ_X^2) e $Y_1, \cdots, Y_{n_Y}$ uma aas da distribuição Normal(μ_Y, σ_Y^2), amostras independentes.

Utilizaremos a seguinte estatística do teste: $F = \dfrac{S_X^2}{S_Y^2} \sim F_{(n_X-1, n_Y-1)}$.

- Teste bilateral ($H_0 : \sigma_X^2 = \sigma_Y^2; H_1 : \sigma_X^2 \neq \sigma_Y^2$)

$$RC(\sigma_X^2/\sigma_Y^2; (1-\alpha)\%) : \{f \in \Re | f < f_{1,\alpha/2} \text{ ou } f > f_{2,\alpha/2}\};$$

- Teste unilateral à esquerda ($H_0 : \sigma_X^2 = \sigma_Y^2; H_1 : \sigma_X^2 < \sigma_Y^2$)

$$RC(\sigma_X^2/\sigma_Y^2; (1-\alpha)\%) >: \{f \in \Re | f < f_{1,\alpha}\};$$

- Teste unilateral à direita ($H_0 : \sigma_X^2 = \sigma_Y^2; H_1 : \sigma_X^2 > \sigma_Y^2$)

$$RC(\sigma_X^2/\sigma_Y^2; (1-\alpha)\%) >: \{f \in \Re | f > f_{2,\alpha}\}\};$$

em que $f = \dfrac{S_X^2}{S_Y^2}$.

Teste para Diferença entre duas médias populacionais de normais independentes com variâncias conhecidas ou de amostras grandes $n_X \geq 30$ **e** $n_Y > 30$

Sejam $X_1, \cdots, X_{n_X}$ uma aas da distribuição Normal(μ_X, σ_X^2) e $Y_1, \cdots, Y_{n_Y}$ uma aas da distribuição Normal(μ_Y, σ_Y^2), amostras independentes.

Teste bilateral ($H_0 : \mu_X = \mu_Y; H_1 : \mu_X \neq \mu_Y$)

$$RC(\mu_X - \mu_Y; (1-\alpha)\%) : |z^*| > z_{\alpha/2},$$

com

- variâncias iguais $\sigma^2 = \sigma_X^2 = \sigma_Y^2$: $z^* = \frac{\left(\bar{X} - \bar{Y}\right)}{\sigma\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}}$;
- variâncias diferentes $\sigma_X^2 \neq \sigma_Y^2$: $z^* = \frac{\left(\bar{X} - \bar{Y}\right)}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}}$;

e, caso as variâncias sejam desconhecidas e ambas as amostram sejam grandes, trocam-se as variâncias pelos seus estimadores pontuais S_X^2 e S_Y^2.
No R, podemos usar a função *z.test*.

Teste para Diferença entre duas médias populacionais de normais independentes com variâncias iguais e desconhecidas e amostras pequenas

Sejam $X_1, \cdots, X_{n_X}$ uma aas da distribuição Normal(μ_X, σ_X^2) e $Y_1, \cdots, Y_{n_Y}$ uma aas da distribuição Normal(μ_Y, σ_Y^2), amostras independentes.

Teste bilateral ($H_0 : \mu_X = \mu_Y; H_1 : \mu_X \neq \mu_Y$)

$$RC(\mu_X - \mu_Y; (1-\alpha)\%) : |t^*| > t_{\nu,\alpha/2}, \text{ com } t^* = \frac{\left(\bar{X} - \bar{Y}\right)}{S_p\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}};$$

tais que

- $\nu = n_X + n_Y - 2$ graus de liberdade;
- $S_p^2 = \frac{(n_X - 1)S_X^2 + (n_Y - 1)S_Y^2}{n_X + n_Y - 2}$;
- $S_X^2 = \frac{\sum_{i=1}^{n_X}(X_i - \bar{X})^2}{n_X - 1}$;
- $S_Y^2 = \frac{\sum_{i=1}^{n_Y}(Y_i - \bar{Y})^2}{n_Y - 1}$.

Teste para Diferença entre duas médias populacionais de normais independentes com variâncias diferentes e desconhecidas e amostras pequenas

Supondo as variâncias desconhecidas $(\sigma_x^2 \neq \sigma_y^2)$, com $n_X < 30$ e/ou $n_Y < 30$, temos:

Teste bilateral $(H_0 : \mu_X = \mu_Y; H_1 : \mu_X \neq \mu_Y)$

$$RC(\mu_X - \mu_Y; (1-\alpha)\%) : |t^*| > t_{\nu,\alpha/2}, \text{ com } t^* = \frac{\left(\bar{X} - \bar{Y}\right)}{\sqrt{\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}}};$$

tais que

- $\nu = \dfrac{(A+B)^2}{\dfrac{A^2}{n_X - 1} + \dfrac{B^2}{n_Y - 1}}$ graus de liberdade;
- $A = \dfrac{S_X^2}{n_X}$ e $B = \dfrac{S_Y^2}{n_Y}$;
- $S_X^2 = \dfrac{\sum_{i=1}^{n_X}(X_i - \bar{X})^2}{n_X - 1}$;
- $S_Y^2 = \dfrac{\sum_{i=1}^{n_Y}(Y_i - \bar{Y})^2}{n_Y - 1}$.

Comentário: Note que no teste de comparação de duas médias de populações distintas, primeiro verificamos a hipótese de **normalidade**. Se ambas as populações são normalmente distribuídas, então verificamos se as **variâncias são conhecidas ou desconhecidas**. Caso sejam conhecidas, então utilizamos o **teste normal**. As variâncias sendo conhecidas, saberemos se elas são iguais ou distintas e aplicaremos os testes vistos aqui. Caso desconhecidas, então utilizamos o **teste t**. Para utilizar o teste t, temos que antes **testar se as variâncias** podem ser consideradas iguais ou diferentes. Neste teste, precisamos da hipótese de normalidade. Caso o teste aponte que podem ser consideradas iguais, aplicamos o teste t para variâncias iguais. Senão, se são diferentes, utilizamos o teste t para variâncias diferentes. Se a hipótese de normalidade das populações é violada, então precisamos verificar se ambas as **amostras possuem tamanho suficientemente grandes** para utilizar o teste normal. Se sim, então com variâncias conhecidas basta aplicar o teste normal diretamente e, se são desconhecidas, basta substituir pelos seus estimadores pontuais.

Teste de hipótese entre as médias de duas populacões com amostras dependentes (variância desconhecida)

Considere duas amostras dependentes $X_1, \ldots, X_n$ e $Y_1, \ldots, Y_n$. Neste caso considere que as observações são pareadas, isto é, pode-se considerar na realidade uma amostra de pares $(X_1, Y_1), \ldots, (X_n, Y_n)$. Vamos definir $D_i = X_i - Y_i$, para $i = 1, 2, \ldots, n$. Assim remos uma amostra $D_1, \ldots, D_n$, resultante das diferenças entre os valores de cada par. Nesta situação, apesar das amostras serem dependentes, vamos considerar que $D_i \sim N(\mu_D, \sigma_D^2)$.

A estatística de teste é:

$$T = \frac{\bar{D} - \mu_D}{S_D/\sqrt{n}}$$

t-Student com $n - 1$ graus de liberdade.

- Teste bilateral ($H_0 : \mu_D = 0; H_1 : \mu_D \neq 0$)

$$RC(\mu_D; (1-\alpha)\%) : |t^*| > t_{(n-1),\alpha/2}, ;$$

- Teste unilateral à esquerda ($H_0 : \mu_D = 0; H_1 : \mu_D < 0$)

$$RC(\mu_D; (1-\alpha)\%) >: t^* < -t_{(n-1),\alpha}, ;$$

- Teste unilateral à direita ($H_0 : \mu_D = 0; H_1 : \mu_D > 0$)

$$RC(\mu_D; (1-\alpha)\%) : t^* > t_{(n-1),\alpha}, ;$$

em que $t^* = \dfrac{\bar{D}\sqrt{n}}{S}$.

Exemplo 15 *Considere o tempo (minutos) de treinamento de cinco operadores de máquinas de soldagem utilizando dois métodos* (A, B) *.*

Com nível de significância de 10% poderíamos afirmar que a tarefa realizada na máquina A demora mais do que na máquina B?

Fonte: Morettin & Bussab, Estatística Básica 5^a *edição, pag. 376.*

Operadores	A	B
1	80	75
2	72	70
3	65	60
4	78	72
5	85	78

2.0.1 Teste χ^2 de Independência

Este teste é baseado na construção de tabelas de contingência para estudarmos a relação/ dependência/associação entre duas variáveis qualitativas. Os grupos devem ser independentes; os itens de cada grupo são selecionados aleatoriamente; as observações devem ser frequências ou contagens; cada observação pertence a uma e somente uma categoria.

Suponha que p_{ij} é a probabilidade de um elemento ser classificado nas categorias i, para $i = 1, \cdots, I$, e j, para $j = 1, \cdots, J$. Sejam $p_{i\cdot} = \sum_{j=1}^{J} p_{ij}$ e $p_{\cdot j} = \sum_{i=1}^{I} p_{ij}$ as probabilidades marginais de pertencer as categorias i e j, respectivamente. Queremos testar se há independência entre variáveis, isto é,

- H_0 : variáveis são independentes $\Leftrightarrow$ frequências observadas = frequências esperadas $\Leftrightarrow p_{ij} = p_{i\cdot}p_{\cdot j}$, para $i = 1, \cdots, I$ e $j = 1, \cdots, J$,
- H_1 : variáveis são dependentes $\Leftrightarrow$ frequências observadas $\neq$ frequências esperadas $\Leftrightarrow p_{ij} \neq p_{i\cdot}p_{\cdot j}$, para algum par (i, j).

A estatística de teste é:

$$T = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2_{(I-1)(J-1)}.$$

Sob H_0, temos que $E_{ij} = \frac{N_{i\cdot}N_{\cdot j}}{N}$, tais que $N_{i\cdot}$ é o nº de elementos classificados na linha i, $N_{\cdot j}$ é o nº de elementos classificados na coluna j e N é o nº total de elementos.

Quanto mais distintos os valores das frequências observadas e esperadas, maior o valor da estatística de teste T, e mais longe da hipótese de independência. Portanto, H_0 é

rejeitada se $T > \chi^2_{(I-1)(J-1),\alpha}$, fixando-se o nível de significância α.

No R, podemos utilizar a função *chisq.test*.

Exemplo 16 *Considere o seguinte conjunto com observações da variável sexo e uso de hospital.*

	Homens	Mulheres
Usaram hospital	100	150
Não Usaram hospital	900	850

Tabela 2.3: Frequência com que 2.000 segurados utilizaram hospital.

	Homens (E_{ij})	Mulheres (E_{ij})	Total $N_{i.}$
Usaram hospital	100 (125)	150 (125)	250
Não Usaram hospital	900 (875)	850 (875)	1750
Total $N_{.j}$	1000	1000	N=2000

Tabela 2.4: Tabela de Contingência.

Temos que

$$\begin{aligned} T &= \sum_{i=1}^{I} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \\ &= \frac{(100-125)^2}{125} + \frac{(150-125)^2}{125} + \frac{(900-875)^2}{875} + \frac{(850-875)^2}{875} \\ &= 11,43. \end{aligned}$$

Ao nível de significância de 5%, temos uma $\chi^2_{1;0,05}$*, pois* $(2-1)(2-1) = 1$*, e, portanto,* $\chi^2_{1;0,05} = 3,841$*.*

Logo, $T = 11,43 > 3,841 = \chi^2_{1;0,05}$ *levando a rejeição de* H_0*, ou seja, não podemos afirmar que as variáveis são independentes.*

No R:

```
x=as.table(rbind(c(100,150),c(900,850)))
dimnames(x) = list(Uso = c("Uso","não Uso"),Sexo = c("F", "M"))
x  #ou View(x)
x2 = chisq.test(x)
x2
```

```
x2$observed
x2$expected #analisar frequencias esperadas abaixo de 5
x2$residuals #residuos padronizados (não eh o ajustado)
x2$stdres #estes devem ser analisados:residuos padronizados ajustados- estao
    na unidade z
#eh considerado residuo estatisticamente significativos, ao nivel de 5%, se
    estao fora da faixa -196 ateh 1,96

x22 = chisq.test(x,correct = FALSE) #sem correção de continuidade
x22
```

Nas tabelas 2×2, o número de observações esperadas em todas as células deve ser de no mínimo 5 (o que não ocorre usualmente com amostras pequenas). Não satisfeita essa condição, deve-se utilizar o Teste Exato de Fisher (no R, *fisher.test*). **N**as tabelas maiores do que 2×2 o número de observações esperadas em cada célula não deve ser < 1 e não $> 20\%$ das células deve ter < 5 observações. Se for o caso, aumentar a coleta ou diminuir o número de categorias.

Correção de Continuidade de Yates Ao trabalharmos com tabelas de contingência, estamos lidando com quantidades discretas (contagens). No entanto, utilizamos uma distribuição de probabilidades contínua (a Qui-quadrado) para testar a associação entre variáveis. Assim, para melhorar essa aproximação, foi proposta uma correção de continuidade no cálculo da qui-quadrado, que corresponde a subtrair 1/2 de cada diferença entre as frequências antes de elevar ao quadrado. Para tabelas 2×2, com as variáveis A e B como apresentado na tabela abaixo.

	A_1	A_2	Total
B_1	a	b	a+b
B_2	c	d	c+d
Total	a+c	b+d	N

Tabela 2.5: Tabela geral.

Daí, a expressão com a correção será escrita como

$$\chi^2 = \frac{N\left(\mid ad - bc \mid -\frac{N}{2}\right)^2}{(a+b)(c+d)(a+c)(b+d)},$$

que é o valor da estatística de teste T e deve ser comparado com o valor $\chi^2_{1,\alpha}$ (para α fixo), pois temos 1 grau de liberdade.

Em tabelas 2×2, esta correção leva a maior precisão e facilidade nos cálculos.

Apesar de a correção de continuidade não provocar grandes alterações no valor do χ^2, essas alterações podem ser importantes quando se está próximo do limite entre região de rejeição e a região de não rejeição. Desse modo, o uso da correção de continuidade deve ser sempre preferido.

Capítulo 3

Análise de Variância

3.1 Motivação

Podemos querer, com base em dados amostrais:

- determinar se há diferença na média de vendas mensais de uma empresa em cinco regiões diferentes;
- comparar os preços de casas em quatro cidades;
- decidir se há realmente alguma diferença na eficácia de três métodos de ensino de língua estrangeira;
- comparar o rendimento médio por hectare de oito variedades de trigo;
- determinar se há realmente diferença entre quilometragem média obtida com cinco tipos de gasolina;
- determinar se existe realmente alguma diferença na durabilidade de seis marcas de tinta para pintura externa.

3.2 Introdução à Análise de Variância

A análise de variância é uma técnica que pode ser utilizada para determinar se as médias de duas ou mais populações são iguais.

- Nela é utilizado o teste F (Fisher-Snedecor).

- Uma análise de variância só deve ser aplicada a um conjunto de observações sob as seguintes pressuposições:

    1. Independência;
    2. Homocedasticidade (variâncias iguais);
    3. Normalidade.

Na prática, porém, dificilmente essas pressuposições estão, todas, satisfeitas. Então, é importante saber que:

- a não-independência, isto é, a correlação entre as observações é o problema mais grave. Pode comprometer as inferências sobre as médias, isto é, as conclusões obtidas da análise de variância e dos testes de comparação;
- variâncias diferentes (heterocedasticidade) têm, usualmente, efeito apenas moderado nas inferências sobre médias, desde que o número de repetições seja constante;
- a não-normalidade tem pouco efeito nas inferências sobre médias, mas apresenta problemas sérios na inferência sobre variâncias.

Modelo e análise de variância

- Seja Y a v.a. de interesse de uma determinada população (indivíduos, animais, empresas....), e admita-se que os elementos da população podem ser classificados em k níveis de um fator (ex: sexo).
- A questão é saber se o fator exerce alguma influência na variação da característica em estudo.
- Suponha que estamos interessados em comparar as médias dessas k subpopulações. Vamos obter k amostras independentes com m indivíduos em cada uma delas.
- Neste contexto, cada observação Y_{ij} pode ser decomposta no seguinte modelo estatístico:

$$Y_{ij} = \mu_i + e_{ij}, \qquad i = 1, ..., k;\ j = 1, ..., m. \tag{3.1}$$

- Vamos admitir que cada valor observado Y_{ij} é dado pela soma da média (μ_i) da população de onde este valor proveio com um erro aleatório (e_{ij}).

Se μ é a média global de todas as populações e se $\tau_i = \mu_i - \mu$ é o efeito do nível i do fator (tratamento i) ($\sum_{i=1}^{k} \tau_i = 0$), de modo que $\mu_i = \mu + \tau_i$, então a equação (1) toma a forma

$$Y_{ij} = \mu + \tau_i + e_{ij}. \tag{3.2}$$

Na análise de variância, quatro suposições devem ser satisfeitas:

- o modelo deve ser aditivo, ou seja, os efeitos nele contido se somam;
- os erros e_{ij} devem ser independentes;
- os erros e_{ij} devem ter distribuição normal;
- os erros e_{ij} devem ter homogeneidade de variância, isto é, deve haver homocedasticidade.

A violação de uma destas suposições pode levar a uma conclusão errada. Neste contexto, estamos interessados em testar as seguintes hipóteses

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 = \ldots = \mu_k = \mu, \\ H_1 &: \text{pelo menos uma das médias } \mu_i \text{ é diferente das demais.} \end{aligned}$$

Caso a hipótese H_0 seja verdadeira, teremos que todas as médias para as k populações serão iguais a um valor comum μ.

A análise de variância vai estimar a variância por dois métodos diferentes, um sob a validade da hipótese nula e o outro não.

As duas estimativas obtidas são depois comparadas para tomarmos uma decisão: se os grupos tiverem todos a mesma média (isto é, se H_0 é verdadeiro), as duas estimativas devem estar próximas uma da outra, caso contrário (isto é, se H_1 é verdadeiro) devem diferir significativamente.

Sob H_0, o modelo pode então ser escrito como

$$Y_{ij} = \mu + e^*_{ij}, \qquad i = 1, ..., k;\ j = 1, ..., m. \tag{3.3}$$

Temos as seguintes somas de quadrados dos erros:

$$\sum_{i=1}^{k}\sum_{j=1}^{m} e_{ij}^2 = \sum_{i=1}^{k}\sum_{j=1}^{m} (Y_{ij} - \mu_i)^2, \tag{3.4}$$

$$\sum_{i=1}^{k}\sum_{j=1}^{m} (e^*_{ij})^2 = \sum_{i=1}^{k}\sum_{j=1}^{m} (Y_{ij} - \mu)^2. \tag{3.5}$$

Estas expressões envolvem quantidades desconhecidas $\mu_i,\ i = 1, ..., k,$ e μ. Assim, utilizaremos seus estimadores

$$\hat{\mu}_i = \frac{1}{m}\sum_{j=1}^{m} Y_{ij} = \bar{Y}_i, \quad i = 1, ..., k, \tag{3.6}$$

$$\hat{\mu} = \frac{1}{mk}\sum_{i=1}^{k}\sum_{j=1}^{m} Y_{ij} = \bar{Y}. \tag{3.7}$$

Substituindo as estimativas correspondentes nas somas de quadrados, definimos as quantidades

$$\begin{aligned} SQRes &= \sum_{i=1}^{k}\sum_{j=1}^{m}(Y_{ij} - \hat{\mu}_i)^2 = \sum_{i=1}^{k}\sum_{j=1}^{m}(Y_{ij} - \bar{Y}_i)^2 = \sum_{i=1}^{k}\sum_{j=1}^{m} Y_{ij}^2 - m\sum_{i=1}^{k}\bar{Y}_i^2, \\ SQT &= \sum_{i=1}^{k}\sum_{j=1}^{m}(Y_{ij} - \hat{\mu})^2 = \sum_{i=1}^{k}\sum_{j=1}^{m}(Y_{ij} - \bar{Y})^2 = \sum_{i=1}^{k}\sum_{j=1}^{m} Y_{ij}^2 - mk\bar{Y}^2, \end{aligned}$$

sendo

- $SQRes$ a soma de quadrados do resíduo que é a soma de quadrados dentro dos níveis, ou grupos, do fator e mede a variação dentro dos grupos (populações); é por vezes designada por variação não explicada ou residual, pois é atribuída a flutuações dentro da mesma população, portanto não pode ser explicada pelas possíveis diferenças entre os grupos (populações).
- SQT a soma de quadrados total que mede a variação total nos dados.

A diferença entre SQT e $SQRes$ representa a soma de quadrados entre os níveis do fator (tratamentos) (SQTr ou SQDen), isto é,

$$SQTr = SQT - SQRes,$$

que mede a variação entre grupos (populações); é por vezes designada por variação explicada, pois ela é explicada pelo fato de as amostras poderem provir de populações diferentes.

Assim, temos que

$$SQTr = m\sum_{i=1}^{k}(\bar{Y}_i - \bar{Y})^2 = m(\sum_{i=1}^{k}\bar{Y}_i^2 - k\bar{Y}^2).$$

Sabemos que, sob H_0

$$\frac{SQTr}{\sigma^2} \sim \chi^2_{(k-1)},$$

e que a estimativa da variância σ^2 é dada por:

$$QMTr = \frac{SQTr}{k-1},$$

em que $QMTr$ é o quadrado médio dos níveis do fator ou do tratamento. E, também sabemos que, sob H_0

$$\frac{SQRes}{\sigma^2} \sim \chi^2_{k(m-1)},$$

e que a estimativa da variância σ^2 é dada por:

$$QMRes = \frac{SQRes}{km-k} = \frac{SQRes}{k(m-1)},$$

em que $QMRes$ é o quadrado médio do resíduo.

Definimos o quadrado médio total por

$$QMT \quad = \quad \frac{SQT}{km-1}. \tag{3.8}$$

A estimativa do resíduo, $QMRes$, não é afetada pela veracidade ou falsidade de H_0. Ao contrário, a estimativa do tratamento para a variância, $QMTr$, já o é, sendo aproximadamente igual a $QMRes$ quando H_0 é verdadeira e maior do que esta se H_0 é falsa.

Levando este fato em consideração, a estatística de teste é definida como o valor F

$$F = \frac{QMTr}{QMRes} \sim F(k-1, k(m-1)).$$

- Se H_0 é verdadeira, σ^2 pode ser estimada pelos dois processos e como as duas estimativas serão aproximadamente iguais, a razão F será próxima de 1.

- Se H_0 for falsa, as diferenças nas médias populacionais irão provocar maior variabilidade nas médias amostrais e, portanto, $QMTr$ será também grande comparativamente com $QMRes$. A razão F tomará um valor maior que 1.

- O valor de F será utilizado para testar H_0, onde ele será comparado com um valor $F_{tabelado}$ com $(k-1, k(m-1))$ graus de liberdade e com um determinado nível de significância α, e se

  $$F > F_{tabelado},$$

  rejeita-se H_0.

Desta maneira, podemos montar a tabela de análise de variância da seguinte forma:

Tabela 3.1: *ANOVA*

Fontes de Variação	**GL**	**SQ**	**QM**	**F**
Tratamento	$k-1$	$SQTr$	$SQTr/(k-1)$	$QMTr/QMRes$
Resíduo	$k(m-1)$	$SQRes$	$SQRes/k(m-1)$	
Total	$km-1$	SQT	$SQT/(km-1)$	

Observação

As somas dos quadrados também podem ser escritas da forma

$$\begin{aligned} SQT &= \sum_{i=1}^{k}\sum_{j=1}^{m} Y_{ij}^2 - C, \\ SQTr &= \frac{\sum_{i=1}^{k} T_i^2}{m} - C, \\ SQRes &= SQT - SQTr, \end{aligned}$$

tal que $C = \frac{\left(\sum_{i=1}^{k}\sum_{j=1}^{m} Y_{ij}\right)^2}{mk}$ e $T_i = \sum_{j=1}^{m} Y_{ij}$, para $i = 1, 2, \cdots, k$.

Aplicação:

Considere quatro diferentes bancos com agências de mesmo porte em uma determinada avenida de uma cidade. Para testar se essas agências têm movimentos médios equivalentes, foram escolhidos cinco dias de trabalho e os desempenhos nestes dias foram registrados. Os dados obtidos, em milhares de reais, estão apresentados na tabela a seguir:

Fazendo as considerações necessárias, qual é a conclusão ao nível de significância de 5%?

Hipóteses:

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 = \mu_3 = \mu_4, \\ H_a &: \text{ pelo menos uma das médias } \mu_i \text{ é diferente das demais.} \end{aligned}$$

Temos que os graus de liberdade são

Tabela 3.2: *Dados*

Banco 1	**Banco 2**	**Banco 3**	**Banco 4**
25	31	22	33
26	25	26	29
20	28	28	31
23	27	25	34
21	24	29	28
115	135	130	155

- do total: $20 - 1 = 19$;
- do resíduo: $20 - 4 = 16$;
- de tratamentos: $4 - 1 = 3$.

O valor de C é dado por

$$C = \frac{(25 + 26 + \ldots + 28)^2}{5 * 4} = \frac{535^2}{20} = 14.311,25.$$

Obtendo as somas dos quadrados

- $SQT = 25^2 + 26^2 + \ldots + 28^2 - 14.311,25 = 275,75$;
- $SQTr = \frac{115^2+135^2+130^2+155^2}{5} - 14.311,25 = 163,75$;
- $SQRes = 275,75 - 163,75 = 112,00$.

Calculando os quadrados médios

- $QMTr = \frac{SQTr}{k-1} = \frac{163,75}{3} = 54,58$;
- $QMRes = \frac{SQRes}{mk-k} = \frac{112,00}{16} = 7,00$.

Finalmente, o valor de F

$$F = \frac{QMTr}{QMRes} = \frac{54,58}{7,00} = 7,80.$$

Tabela 3.3: *Análise de variância dos dados*

Causas de variação	GL	SQ	QM	F
Trat	3	163,75	54,58	7,80
Res	16	112,00	7,00	
Tot	19	275,75		

Conclusão

Para o nível de significância de 5% com 3 e 16 graus de liberdade, o valor tabelado de F é $3,24$. Temos então

$$F = 7,80 > 3,24 = F_{tabelado},$$

portanto, como visto anteriormente, rejeitamos H_0, isto é, rejeitamos a hipótese de que todas as médias são iguais para os diferentes bancos.

Grupos/Níveis de Tamanhos Diferentes

Anteriormente, supomos que os k grupos têm todos tamanho igual. Consideremos uma situação mais geral, em que isto não acontece.
Denotamos por n_i o número de elementos do nível i do fator. Neste caso, o total de indivíduos nos k grupos será igual a

$$n = n_1 + n_2 + \ldots + n_k.$$

Todos os resultados anteriores, permanecem válidos, mas modificações algébricas são necessárias nas expressões que agora serão escritas da seguinte forma

$$\begin{aligned} SQRes &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}^2 - \sum_{i=1}^{k} n_i \bar{Y}_i^2, \\ SQTr &= \sum_{i=1}^{k} n_i(\bar{Y}_i - \bar{Y})^2 = \sum_{i=1}^{k} n_i \bar{Y}_i^2 - n\bar{Y}^2, \\ SQT &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}^2 - n\bar{Y}^2. \end{aligned}$$

Note que, neste caso, as médias global e dos grupos são dadas por

$$\begin{aligned} \bar{Y} &= \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}, \\ \bar{Y}_i &= \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}, \quad i = 1, ..., k. \end{aligned}$$

A tabela de análise de variância toma a forma

FV	GL	SQ	QM	F
Tratamento	$k-1$	$SQTr$	$QMTr$	$QMTr/QMRes$
Resíduo	$n-k$	$SQRes$	$QMRes$	
Total	$n-1$	SQT		

Teste de Comparação de Médias

Caso rejeitamos H_0, concluímos que existe ao menos um tratamento que difere dos demais.

- **Teste de Tukey:** duas médias são estatisticamente diferentes toda vez que o valor absoluto da diferença entre elas for igual ou maior do que Δ, tal que

  $$\Delta = q\sqrt{QMRes/m},$$

  sendo q um valor tabelado e m o número de repetições de cada um dos tratamentos.

  Exemplo 2: Aplicando o teste de Tukey às médias dos tratamentos do exemplo 1, temos:

  $$\Delta = q\sqrt{QMRes/m} = 4,09\sqrt{7/5} = 4,79$$

  sendo $q = 4,09$ e $\alpha = 0,05$. Importante: o valor de q depende do número de tratamentos e do número de graus liberdade do resíduo. Para a realização do teste, temos que calcular as médias dos tratamentos e ordená-las.

  Podemos concluir pelo teste de Tukey, ao nível 5% de significância, que a média do movimento no banco 4 é diferente das médias dos movimentos nos bancos 1 e 3.

Tabela 3.4: *Dados*

Banco	Média	Diferença entre as médias
Banco 4 B4	31	B4-B2=4, B4-B3=5 e B4-B1=8
Banco 2 B2	27	B2-B3=1 e B2-B1=4
Banco 3 B3	26	B3-B1=3
Banco 1 B1	23	

- **Teste de Duncan:** ordena as médias e estabelece a primeira diferença de média entre a menor e a maior. Se a diferença for significativa, deve-se continuar a análise das diferenças.

Capítulo 4

Anova com dois fatores

4.1 Anova com dois fatores Sem Interação

Motivação: Um estudo foi conduzido para comparar a gasolina entre as marcas diferentes e o consumo de gasolina entre os modelos de automóveis. Foram utilizados quatro modelo de automóveis diferentes.

Tabela 4.1: *Tabela de dupla entrada para as marcas de gasolina e os automóveis.*

Modelo \ Marca	M_1	M_2	M_3	Total
A	32	36	39	107
B	29	29	30	88
C	37	38	39	114
D	34	36	37	107
Total	132	139	145	416

Neste caso, temos dois fatores marca de gasolina e modelo de carro que serão estudados em relação a possíveis efeitos sobre o consumo de gasolina por quilometragem considerada. Abaixo, será apresentada o desenvolvimento geral para problemas desse tipo.

Modelo para dois fatores sem interação:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}, \tag{4.1}$$

em que μ é uma média comum a todos, α_i, $i = 1, ..., a$, efeito do fator A; β_j, $j = 1, ..., b$, efeito do fator B; k é o número de observações para cada combinação entre os fatores e εijk erro aleatório do experimento.

Suposições: Assumimos que os erros ε_{ijk} são independentes e têm distribuição normal com média 0 e variância σ^2 (igualdade de variâncias). Como os erros ε_{ijk} são independentes, obtemos que as observações Y_{ijk} também são independentes. Logo

$$\varepsilon_{ijk} \sim N(0;\ \sigma^2) \quad \text{e} \quad Y_{ijk} \sim N(\mu + \alpha_i + \beta_j;\ \sigma^2).$$

(i) Fator A:

$$\begin{aligned} H_0 &: \alpha_1 = \alpha_2 = \cdots = \alpha_a = 0, \\ H_a &: \text{existe pelo menos um } \alpha_i \neq 0. \end{aligned}$$

(ii) Fator B:

$$\begin{aligned} H_0 &: \beta_1 = \beta_2 = \cdots = \beta_b = 0 \\ H_a &: \text{existe pelo menos um } \beta_j \neq 0. \end{aligned}$$

A tabela 4.1 representa as observações de um experimento de dois fatores com n replicações.

em que

$$Y_{i,,} = \sum_{j=1}^{b}\sum_{k=1}^{n} y_{ijk}\ ;\ \ Y_{,j,} = \sum_{i=1}^{a}\sum_{k=1}^{n} y_{ijk}\ ;\ \ Y_{ij,} = \sum_{k=1}^{n} y_{ijk}\ ;\ \ Y_{,,,} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} y_{ijk}$$

$$\overline{y}_{i,,} = \frac{1}{b\ n}\sum_{j=1}^{b}\sum_{k=1}^{n} y_{ijk}\ ;\ \ \overline{y}_{,j,} = \frac{1}{a\ n}\sum_{i=1}^{a}\sum_{k=1}^{n} y_{ijk}\ ;\ \ \overline{y}_{ij,} = \frac{1}{n}\sum_{k=1}^{n} y_{ijk};$$

$$\overline{y}_{,,,} = \frac{1}{b\ a\ n}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} y_{ijk}$$

Teorema 1:

$$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk} - \bar{y})^2 = bn\sum_{i=1}^{a}(\bar{y}_{i,,} - \bar{y})^2 + an\sum_{i=1}^{b}(\bar{y}_{,j,} - \bar{y})^2 + \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk} - \hat{y}_{ijk})^2, \quad (4.2)$$

onde $\hat{y}_{ijk} = \bar{y}_{ij} = \bar{y}_{i,,} + \bar{y}_{,j,} - \bar{y}$.

Tabela 4.2: *Experimento de dois fatores com n replicações.*

Fator A \ Fator B	1	2		b	Total	Média
1	y_{111} y_{112} $\vdots$ y_{11n}	y_{121} y_{122} $\vdots$ y_{12n}	$\cdots$ $\cdots$ $\cdots$ $\cdots$	y_{1b1} y_{1b2} $\vdots$ y_{1bn}	$Y_{1,,}$	$\bar{y}_{1,,}$
2	y_{211} y_{212} $\vdots$ y_{21n}	y_{221} y_{222} $\vdots$ y_{22n}	$\cdots$ $\cdots$ $\cdots$ $\cdots$	y_{2b1} y_{2b2} $\vdots$ y_{2bn}	$Y_{2,,}$	$\bar{y}_{2,,}$
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$	$\vdots$
a	y_{a11} y_{a12} $\vdots$ y_{a1n}	y_{a21} y_{a22} $\vdots$ y_{a2n}	$\cdots$ $\cdots$ $\cdots$ $\cdots$	y_{ab1} y_{ab2} $\vdots$ y_{abn}	$Y_{a,,}$	$\bar{y}_{a,,}$
Total	$Y_{,1,}$	$Y_{,2,}$	$\cdots$	$Y_{,b,}$	$Y_{,,,}$	
Média	$\bar{y}_{,1,}$	$\bar{y}_{,2,}$	$\cdots$	$\bar{y}_{,b,}$		$\bar{y}_{,,,}$

Podemos escrever a identidade da soma dos quadrados acima por

$$SQT = SQA + SQB + SQRES \tag{4.3}$$

Importante:

Podemos obter as somas dos quadrados da seguinte forma

$$\begin{aligned} SQT &= \sum_i \sum_j \sum_k (y_{ijk})^2 - C, \\ SQA &= \frac{1}{nb} \sum_{i=1}^{a} Y_{i,,}^2 - C, \\ SQB &= \frac{1}{na} \sum_{j=1}^{b} Y_{,j,}^2 - C, \\ SQRes &= SQT - SQA - SQB, \end{aligned}$$

tal que $C = \dfrac{\left(\sum_i \sum_j \sum_k y_{ijk}\right)^2}{abn}$.

Neste caso, a tabela de análise de variância é dada por

Tabela 4.3: *ANOVA com dois fatores (sem interação)*

FV	GL	SQ	QM	F
Fator A	$a-1$	SQA	$SQA/(a-1)$	$QMA/QMRes$
Fator B	$b-1$	SQB	$SQB/(b-1)$	$QMB/QMRes$
Res	$abn-(a+b-1)$	$SQRes$	$SQRes/(abn-(a+b-1))$	
Tot	$abn-1$	SQT	$SQT/(abn-1)$	

Importante: Aplicar, para cada fator, a mesma metodologia do teste de rejeição utilizada para um fator. Isto é, rejeitamos a hipótese de não existência de efeito do Fator A se

$$F = \frac{QMA}{QMRes} > F_{tabelado}(a-1, abn-(a+b-1)),$$

e, rejeitamos a hipótese de não existência de efeito do Fator B se

$$F = \frac{QMB}{QMRes} > F_{tabelado}(b-1, abn-(a+b-1)).$$

Vamos retornar para a solução do problema de motivação dado.

Solução da motivação: Use as equações simplificadas (foram dadas acima) e compare com a tabela gerada pelo R.

Tabela 4.4: *Tabela ANOVA para a Motivação.*

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
factor(Marca)	2	21.17	10.58	5.86	0.0388
factor(Modelo)	3	124.67	41.56	23.02	0.0011
Residuals	6	10.83	1.81		

Neste caso, temos que

$$\frac{QMA}{QMRes} = \frac{10,58}{1,81} = 5,86 > 5,14 = F_{tabelado}(2,6)$$

e

$$\frac{QMB}{QMRes} = \frac{41,56}{1,81} = 23,02 > 4,76 = F_{tabelado}(3,6).$$

Logo, rejeitam-se ambas as hipóteses nulas de inexistência de efeitos dos fatores.
Também podemos observar os valores dos p-valores ($Pr(> F)$) obtidos na tabela 4, levando a rejeitar as duas hipóteses.

4.2 Anova com dois fatores Com Interação

Motivação: Um segundo estudo foi conduzido para comparar o consumo de gasolina entre marcas diferentes de gasolina e modelos de automóveis. Foram utilizados três modelos de automóveis diferentes.

Tabela 4.5: *Tabela de dupla entrada para as marcas de gasolina e os automóveis.*

Modelo (Fator A)	Marca (Fator B)	Replicação 1	Replicação 2	Replicação 3	Total
A (A_1)	M_1	45	39	46	130
A (A_1)	M_2	43	46	41	130
A (A_1)	M_3	55	48	53	156
B (A_2)	M_1	37	32	43	112
B (A_2)	M_2	40	37	46	123
B (A_2)	M_3	56	51	53	160
C (A_3)	M_1	52	52	56	160
C (A_3)	M_2	54	55	48	157
C (A_3)	M_3	46	49	42	137

Neste caso, temos dois fatores marca de gasolina e modelo de carro que serão estudados em relação a possíveis efeitos sobre o consumo de gasolina por quilometragem considerada. Vamos verificar se existe efeito significativo do modelo, da marca ou interação entre modelo e marca. Se houver interação significativa faça o desdobramento da interação; use o teste de Tukey se necessário.
Modelo para dois fatores com interação:

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ijk}, \tag{4.4}$$

com as seguintes restrições $\sum_{i=1}^{a}\alpha_i = 0$, $\sum_{i=1}^{b}\beta_i = 0$, $\sum_{i=1}^{a}(\alpha\beta)_{ij} = 0$, $\sum_{j=1}^{b}(\alpha\beta)_{ij} = 0$. Aqui μ uma média comum a todos, α_i, $i = 1, ..., a$, é efeito do i-ésimo nível do fator A; β_j, $j = 1, ..., b$, efeito do j-ésimo nível do fator B; k é o número de observações para cada combinação entre os fatores; $(\alpha\beta)_{ij}$ é o efeito da interação do i-ésimo nível do fator A com o j-ésimo nível do fator B e e_{ijk} erro aleatório do experimento.
As três hipóteses a serem testadas são as seguintes:

(i) Fator A:

$$\begin{aligned} H_0 &: \alpha_1 = \alpha_2 = \cdots = \alpha_a = 0, \\ H_a &: \text{existe pelo menos um } \alpha_i \neq 0. \end{aligned}$$

(ii) Fator B:

$$\begin{aligned} H_0 &: \beta_1 = \beta_2 = \cdots = \underline{=}0 \\ H_a &: \text{existe pelo menos um } \beta_j \neq 0. \end{aligned}$$

(iii) Interação entre os fatores:

H_0 : Não existe interação entre os fatores, isto é, $(\alpha\beta)_{11} = (\alpha\beta)_{12} = \cdots = (\alpha\beta)_{ab} = 0$,

H_a : Existe interação entre os fatores, isto é, pelo menos um dos $(\alpha\beta)_{ij}$ é diferente de zero.

Importante: É recomendável que o resultado do teste de interação seja realizado primeiro.

Teorema 2: A decomposição da soma dos quadrados é dada por

$$\begin{aligned} \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk} - \bar{y})^2 &= bn\sum_{i=1}^{a}(\bar{y}_{i,,} - \bar{y})^2 + an\sum_{i=1}^{b}(\bar{y}_{,j,} - \bar{y})^2 + n\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{y}_{ij,} - \bar{y}_{i,,} - \bar{y}_{,j,} + \bar{y})^2 \\ &+ \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk} - \bar{y}_{ij,})^2 \end{aligned} \tag{4.5}$$

Podemos escrever a identidade da soma dos quadrados por

$$SQT = SQA + SQB + SQ(AB) + SQRES \tag{4.6}$$

Importante:

Podemos obter as somas dos quadrados da seguinte forma

$$\begin{aligned}
SQT &= \sum_i \sum_j \sum_k (y_{ijk})^2 - C, \\
SQA &= \frac{1}{nb} \sum_{i=1}^{a} Y_{i,,}^2 - C, \\
SQB &= \frac{1}{na} \sum_{j=1}^{b} Y_{,j,}^2 - C, \\
SQTr &= \frac{\sum_{i=1}^{a} \sum_{j=1}^{b} (Y_{ij})^2}{n} - C, \;\; Y_{ij} = y_{ij1} + y_{ij2} + \cdots y_{ijn} \\
SQ(AXB) &= SQTr - (SQA + SQB) \\
SQRes &= SQT - SQA - SQB - SQ(AXB),
\end{aligned}$$

tal que $C = \dfrac{\left(\sum_i \sum_j \sum_k y_{ijk}\right)^2}{abn}$.

Tabela 4.6: *ANOVA com dois fatores*

	GL	**SQ**	**QM**	**F**
Fator A	$a-1$	SQF_A	QMF_A	$QM_A/QMRes$
Fator B	$b-1$	SQF_B	QMF_B	$QM_B/QMRes$
Interação AXB	$(a-1)(b-1)$	$SQ(AXB)$	$QM(AxB)$	$QM(AXB)/QMRes$
Res	$ab(n-1)$	$SQRes$	$QMRes$	
Tot	$abn-1$	SQT		

Solução A ideia é primeiro olhar o problema como apenas um fator e separar os tratamentos como combinações, neste caso, dos modelos e das marcas. Assim, teremos os seguintes tratamentos (combinações dos níveis dos dois fatores): AM_1, AM_2, AM_3, BM_1, BM_2, BM_3, CM_1, CM_2 e CM_3. Daí, procedemos com a análise de variância simples.

Calculando $C = \dfrac{\left(\sum_{i=1}^{k} \sum_{j=1}^{m} Y_{ij}\right)^2}{mk} = \dfrac{1.265^2}{9 \times 3} = 59.267,59$. Podemos obter as somas dos quadrados:

$$\begin{aligned}
SQT &= \sum_{i=1}^{k} \sum_{j=1}^{m} Y_{ij}^2 - C = (45^2 + 39^2 + \cdots + 42^2) - 59.267,59 = 60.389,00 - 59.267,59 = 1.121,41 \\
SQTr &= \frac{\sum_{i=1}^{k} T_i^2}{m} - C = \frac{130^2 + 130^2 + \cdots + 137^2)}{3} - 59.267,59 = 60.142,33 - 59.267,59 = 874,74 \\
SQRes &= SQT - SQTr = 1.121,41 - 874,74 = 246,67.
\end{aligned}$$

Tabela 4.7: *Análise de variância simples dos dados*

Causas de variação	GL	SQ	QM	F
Trat	$k-1$=8	874,74	109,34	7,98
Res	$k(m-1)$=18	246,67	13,70	-
Tot	$km-1$=26	1.121,41	-	-

O valor tabelado da distribuição F com graus de liberdade (8,18), ao nível de 5% de significância, é de $2,51 < 7,98$, portanto, rejeitamos H_0. Isto é, o teste para tratamentos foi significativo, então devemos desdobrar o efeito de tratamentos em marca (A), modelo (B) e interação (AXB). Para isso, precisamos determinar:

$$\begin{aligned} SQA &= \frac{1}{nb}\sum_{i=1}^{a} Y_{i,,}^2 - C = \frac{(416^2+395^2+454^2)}{3\times 3} - 59.267,59 = 59.466,33 - 59.267,59 = 198,74 \\ SQB &= \frac{1}{na}\sum_{j=1}^{b} Y_{,j,}^2 - C = \frac{402^2+410^2+453^2}{3\times 3} - 59.267,59 = 59.434,78 - 59.267,59 = 167,19, \\ SQ(AXB) &= SQTrat - SQA - SQB = 874,74 - 198,74 - 167,19 = 508,81, \end{aligned}$$

tais que $Y_{1,,} = 130+130+156 = 416$,$Y_{2,,} = 112+123+160 = 395$,$Y_{3,,} = 160+157+137 = 454$,$Y_{,1,} = 130+112+160 = 402$,$Y_{,2,} = 130+123+157 = 410$,$Y_{,3,} = 156+160+137 = 453$.

Tabela 4.8: *ANOVA com dois fatores*

	GL	SQ	QM	F
Fator A (Modelo)	$a-1$= 2	198,74	99,370	7,25
Fator B (Marca)	$b-1$= 2	167,19	83,593	6,10
Interação AXB	$(a-1)(b-1)$= 4	508,81	127,204	9,28
Res	$ab(n-1)$= 18	246,6	13,704	-
Tot	$abn-1$= 26	1.121,41	-	-

Os valores tabelados são: $F_{(2,18)} = 3,55 < 7,25$, $F_{(2,18)} = 3,55 < 6,10$ e $F_{(4,18)} = 2,93 < 9,28$, ao nível de 5% de significância. Portanto, rejeitamos as três hipóteses nulas. Como o efeito da interação foi significativo $(F_c > F_{tab})$, então existe interação entre os fatores

A (modelo de carro) e B (marca da gasolina). Logo, o Fator A pode se comportar de maneiras diferentes em cada nível do fator B e vice-versa. Assim, não podemos analisar os efeitos simples do Fator A e do fator B, devemos, então, estudar o fator A dentro de cada nível do fator B e estudar o fator B dentro de cada nível do fator A. Daí, surge o termo desdobramento da interação significativa.

4.2.1 Desdobrando a interação significativa

Desdobrando Modelo (Fator A) dentro de cada nível de Marca (Fator B)

Vamos fazer uma análise de variância obtendo as somas de quadrados do Fator A dentro de cada nível 1,2 e 3 do fator B.

- Desdobrando Modelo (Fator A) dentro do nível M_1 de Marca (Fator B):

$$\frac{\left[\sum_{i=1}^{a}\left(\sum_{k=1}^{n} Y_{i,1,k}\right)^2\right]}{n} - \frac{Y_{,1,}^2}{bn} = \frac{130^2 + 112^2 + 160^2}{3} - \frac{402^2}{9} = 392.$$

- Desdobrando Modelo (Fator A) dentro do nível M_2 de Marca (Fator B):

$$\frac{\left[\sum_{i=1}^{a}\left(\sum_{k=1}^{n} Y_{i,2,k}\right)^2\right]}{n} - \frac{Y_{,2,}^2}{bn} = \frac{130^2 + 123^2 + 157^2}{3} - \frac{410^2}{9} = 214,89.$$

- Desdobrando Modelo (Fator A) dentro do nível M_3 de Marca (Fator B):

$$\frac{\left[\sum_{i=1}^{a}\left(\sum_{k=1}^{n} Y_{i,3,k}\right)^2\right]}{n} - \frac{Y_{,3,}^2}{bn} = \frac{156^2 + 160^2 + 137^2}{3} - \frac{453^2}{9} = 100,67.$$

Daí, construímos uma tabela de análise de variância como mostrada abaixo.

Tabela 4.9: *ANOVA do desdobramento do Fator Modelo dentro de cada nível do Fator Marca*

Fonte de variação	GL	SQ	QM	F
Modelo d.M_1 (Modelo dentro de M_1)	2	392	196	14,30
Modelo d.M_2 (Modelo dentro de M_2)	2	214,89	107,45	7,84
Modelo d.M_3 (Modelo dentro de M_3)	2	100,67	50,33	3,67
Res	18	246,67	13,70	-

O valor da $F_{(2,18)}$ tabelado é dado por 3,55, ao nível de 5%. Portanto, como todos os valores da F calculados foram significativos, então realizaremos um teste de Tukey para comparar os níveis de Modelo A dentro de M_1, outro para comparar os níveis de Modelo dentro de M_2 e, ainda, outro para comparar os níveis de Modelo dentro de M_3.

Teste de Tukey Calculando a diferença mínima:

$$\Delta = 3,61\sqrt{\frac{13,70}{3}} = 7,72,$$

em que $m = 3$ é o número de replicações.
Obtendo as médias do desdobramento de Modelo (Fator A) dentro dos níveis de Marca (Fator B):

- Desdobrando Modelo (Fator A) dentro do nível M_1 de Marca (Fator B): 130/3=43,33 (A dentro de M_1); 112/3=37,33 (B dentro de M_1); 160/3=53,33 (C dentro de M_1).
- Desdobrando Modelo (Fator A) dentro do nível M_2 de Marca (Fator B): 130/3=43,33 (A dentro de M_2); 123/3= 41,00 (B dentro de M_2); 157/3=52,33 (C dentro de M_2).
- Desdobrando Modelo (Fator A) dentro do nível M_3 de Marca (Fator B): 156/3= 52,00 (A dentro de M_3); 160/3=53,33 (B dentro de M_3); 137/3=45,67 (C dentro de M_3).

Obtendo as diferenças entre as médias do desdobramento de Modelo (Fator A) dentro dos níveis de Marca (Fator B):

- Desdobrando Modelo (Fator A) dentro do nível M_1 de Marca (Fator B): 53,33-43,33=10*; 53,33-37,33=16*; 43,33-37,33=6.
- Desdobrando Modelo (Fator A) dentro do nível M_2 de Marca (Fator B): 52,33- 43,33=9*; 52,33-41,00=11,33*; 43,33-41,00=2,33.
- Desdobrando Modelo (Fator A) dentro do nível M_3 de Marca (Fator B): 53,33-52,00=1,33; 53,33-45,67=7,66; 52,00-45,67=6,33.

Os que estão com * representam os casos com diferença significativa, ao nível de 5%.

Desdobrando Marca (Fator B) dentro de cada nível de Modelo (Fator A)

Vamos fazer uma análise de variância obtendo as somas de quadrados do fator Marca dentro de cada nível A,B e C do fator Modelo.

- Desdobrando Marca (Fator B) dentro do nível A de Modelo (Fator A):

$$\frac{\left[\sum_{j=1}^{b}\left(\sum_{k=1}^{n} Y_{1,j,k}\right)^2\right]}{n} - Y_{1,,}^2 an = \frac{130^2 + 130^2 + 156^2}{3} - \frac{416^2}{9} = 150,22.$$

- Desdobrando Marca (Fator B) dentro do nível B de Modelo (Fator A):

$$\frac{\left[\sum_{j=1}^{b}\left(\sum_{k=1}^{n} Y_{2,j,k}\right)^2\right]}{n} - Y_{2,,}^2 an = \frac{112^2 + 123^2 + 160^2}{3} - \frac{395^2}{9} = 421,56.$$

- Desdobrando Marca (Fator B) dentro do nível C de Modelo (Fator A):

$$\frac{\left[\sum_{j=1}^{b}\left(\sum_{k=1}^{n} Y_{3,j,k}\right)^2\right]}{n} - Y_{3,,}^2 an = \frac{160^2 + 157^2 + 137^2}{3} - \frac{454^2}{9} = 104,22.$$

Daí, construindo uma tabela de análise de variância como mostrada abaixo.

Tabela 4.10: *ANOVA do desdobramento do Fator Marca dentro de cada nível do Fator Modelo*

Fonte de variação	**GL**	**SQ**	**QM**	**F**
Marca d.A (Marca dentro de A)	2	150,22	75,11	5,48
Marca d.B (Marca dentro de B)	2	421,56	210,78	15,38
Marca d.C (Marca dentro de C)	2	104,22	52,11	3,80
Res	18	246,67	13,70	-

O valor da $F_{(2,18)}$ tabelado é dado por 3,55, ao nível de 5%. Portanto, como todos os valores da F calculados foram significativos, então realizaremos um teste de Tukey para comparar os níveis de Marca dentro de A, outro para comparar os níveis de Marca dentro de B e, ainda, outro para comparar os níveis de Marca dentro de C.

Teste de Tukey Calculando a diferença mínima:

$$\Delta = 3,61\sqrt{\frac{13,70}{3}} = 7,72,$$

em que $m = 3$ é o número de replicações.
Obtendo as médias do desdobramento de Marca (Fator B) dentro dos níveis de Modelo (Fator A):

- Desdobrando Marca (Fator B) dentro do nível A de Modelo (Fator A): 130/3=43,33 (M_1 dentro de A); 130/3=43,33 (M_2 dentro de A); 156/3= 52,00 (M_3 dentro de A).
- Desdobrando Marca (Fator B) dentro do nível B de Modelo (Fator A): 112/3=37,33 (M_1 dentro de B); 123/3= 41,00 (M_2 dentro de B); 160/3=53,33 (M_3 dentro de B).
- Desdobrando Marca (Fator B) dentro do nível C de Modelo (Fator A): 160/3=53,33 (M_1 dentro de C); 157/3=52,33 (M_2 dentro de C); 137/3=45,67 (M_3 dentro de C).

Obtendo as diferenças entre as médias do desdobramento de Marca (Fator B) dentro dos níveis de Modelo (Fator A):

- Desdobrando Marca (Fator B) dentro do nível A de Modelo (Fator A): 52,00-43,33=8,67*; 52,00-43,33=8,67*; 43,33-43,33=0.
- Desdobrando Marca (Fator B) dentro do nível B de Modelo (Fator A): 53,33- 41=12,33*; 53,33-37,33=16*; 41-37,33=3,67.
- Desdobrando Marca (Fator B) dentro do nível C de Modelo (Fator A): 53,33-52,33=1; 53,33-45,67=7,66; 52,33-45,67=6,66.

Os que estão com * representam os casos com diferença significativa, ao nível de 5%.

Capítulo 5

Regressão Linear Simples

Em muitos problemas envolvendo duas variáveis quantitativas com n pares de observações, (x_i, y_i), observamos uma relação de linearidade entre elas. Nesse caso, o primeiro passo para estudar se há esse tipo de relação, é através do gráfico de dispersão entre elas.

Exemplo 17 *Um gerente de marketing conduziu um estudo para determinar se há uma relação entre o dinheiro gasto com propaganda e as vendas da empresa. Os dados são mostrados na tabela abaixo em milhares de reais.*

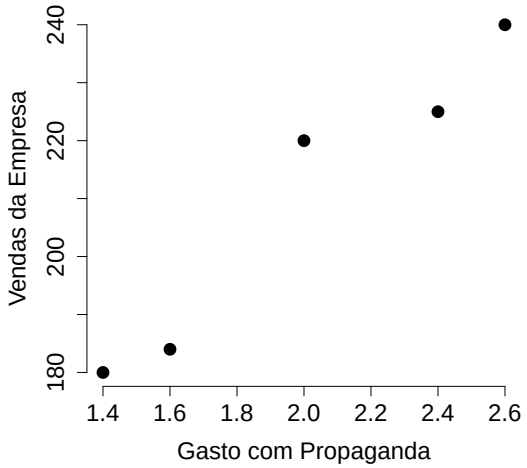
Gasto com propaganda (x_i)	2,4	1,6	2	2,6	1,4
Vendas da empresa (y_i)	225	184	220	240	180

O gráfico de dispersão segue abaixo:

Se o gráfico indicar uma possível relação linear crescente ou descrescente entre as duas variáveis de estudo, podemos calcular o coeficiente de correlação linear de Pearson a fim de verificar essa hipótese. Lembrando que o coeficiente de correlação linear de Pearson é dado por

$$cor(X,Y) = \frac{1}{n}\frac{\sum_{i=1}^{n}[(x_i-\bar{x})(y_i-\bar{y})]}{DP(X)DP(Y)} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right)}}.$$

Exemplo 18 *Continuando o exemplo anterior, note que:* $\sum_{i=1}^{5} x_i y_i = 2150$, $\sum_{i=1}^{5} x_i = 10$, $\sum_{i=1}^{5} y_i = 1049$, $\sum_{i=1}^{5} x_i^2 = 21$, $\sum_{i=1}^{5} y_i^2 = 222881$ *e* $\sum_{i=1}^{5}(x_i-\bar{x})^2 = 1$. *Obs: Alguns cálculos estão*



aproximados (use!). Obtemos que o coeficiente de correlação linear é igual a aproximadamente $0,97$*, indicando alta relação linear positiva/crescente entre as duas variáveis.*

Se a conclusão apontar para relação linear, então podemos estar interessados na função da reta que melhor descreve essa relação. Considere uma das variáveis como resposta (dependente) Y e a outra como uma variável explicativa ou regressora (independente) X, com observações $\{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$. Podemos supor o seguinte modelo de regressão linear simples:

$$y_i = \alpha + \beta x_i + \epsilon_i, \text{ para } i = 1, 2, \cdots, n$$

em que α e β são parâmetros (constantes) desconhecidos e ϵ_i é um erro aleatório ao qual será atribuído uma distribuição simétrica em torno de zero.

Hipóteses do modelo de regressão linear:

- X é controlada pelo experimentador e não sofre variações aleatórias;
- os erros ϵ_i são independentes (não correlacionados);
- os erros ϵ_i possuem média zero;
- homocedasticidade dos erros.

Em geral, assumimos também normalidade. Assim, assumimos que os erros ϵ_i, $\forall \quad i$, possuem distribuição normal com média zero e variância σ^2, isto é,

$$\epsilon_i \sim \text{Normal}(0, \sigma^2).$$

Assim, temos n equações e $n+2$ incógnitas ($\alpha, \beta, \epsilon_1, \cdots, \epsilon_n$) para determinar. O interesse consiste em determinar os valores de α e β da equação da reta de regressão, que é a média do modelo. Precisamos adotar um critério para obtermos essas quantidades.

Seja a soma dos quadrados dos erros

$$SQ(\alpha, \beta) = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2.$$

A ideia é minimizar essa soma e assim obter os valores de (α, β). Esses valores serão as estimativas obtidas para os parâmetros através do método de mínimos quadrados $(\hat{\alpha}, \hat{\beta})$.

Sabemos que precisamos derivar a equação em relação a α e a β e igualar a zero. Daí, temos que

$$n\hat{\alpha} + \hat{\beta} \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i, \qquad \hat{\alpha} \sum_{i=1}^{n} x_i + \hat{\beta} \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i,$$

levando a

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}, \qquad \hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}.$$

Portanto, o estimador para a média $\mu(x_i) = \alpha + \beta x_i$ é dado por

$$\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i = \bar{y} - \hat{\beta}\bar{x} + \hat{\beta} x_i = \bar{y} + \hat{\beta}(x_i - \bar{x}).$$

Exemplo 19 *Retornando ao exemplo,* $\hat{\alpha} = 109,03$ *e* $\hat{\beta} = 50,39$*, o que nos dá o modelo ajustado* $\hat{y}_i = 109,03 + 50,39 x_i$.

A previsão de vendas da empresa para um gasto de 2.700,00 reais com propaganda, por exemplo, é dado por $\hat{y}_i = 109,03 + 50,39 \times 2,7 = 245,083$ *reais. Por outro lado, supondo que a empresa vendeu 250.000,00 reais, qual foi o valor gasto com propaganda? Basta fazer,* $250 = 109,03 + 50,39 x_i$*, logo* $x_i = 2,8$ *que é 2.800,00 reais.*

Interpretação dos coeficientes:

A reta ajustada é dada por: $\hat{y}_i = 109,03 + 50,39 x_i$*, logo a cada aumento de uma unidade em* X*, isto é, a cada aumento de mil reais no gasto em propaganda, temos um aumento de 50,39 unidades de* Y*, ou seja, um aumento de 50,39 mil reais em vendas.*

5.0.1 Estimação da Variância σ^2

Cada resíduo do modelo ajustado é dado por $\hat{e}_i = (y_i - \hat{y}_i)$, e para obtermos uma medida única podemos utilizar a soma dos quadrados dos erros ou residuais que é dada por:

$$SQRes = \sum_{i=1}^{n} \hat{e}_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

É possível mostrar que $E[SQRes] = (n-2)\sigma^2$, logo um estimador não-viciado para σ^2 é dado por

$$S_e^2 = QMRes = \frac{SQRes}{n-2},$$

que é o quadrado médio residual.

Podemos comparar a estimativa da variância S_e^2, sob o modelo de regressão linear, com a estimativa obtida sob o modelo mais simples que seria $y_i = \alpha + \epsilon_i$. Nesse caso, os resíduos são dados por $\hat{e}_i = (y_i - \hat{y})$, cuja soma dos quadrados é a soma dos quadrados total dada por

$$SQT = \sum_{i=1}^{n} \hat{e}_i^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2.$$

Podemos medir o lucro em assumir o modelo de regressão linear ao invés do modelo mais simples através da soma dos quadrados da regressão

$$SQReg = SQT - SQRes = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = \hat{\beta}^2 \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

Daí, quanto maior $\hat{\beta}$, maior a $SQReg$ e maior a redução dos resíduos.

Além disso, podemos medir o lucro relativo em utilizar o modelo de regressão linear através da estatística

$$R^2 = \frac{SQReg}{SQT}.$$

Exemplo 20 *Retornando ao exemplo,*

$$SQRes = \sum_{i=1}^{5} (y_i - \hat{y}_i)^2 = (225 - 229,966)^2 + (184 - 189,654)^2 + \cdots + (180 - 179.576)^2 = 160,6467,$$

logo $S_e^2 = QMRes = \dfrac{160,6467}{5-2} \approx 53,55$. *E,*

$$SQT = \sum_{i=1}^{5} (y_i - \bar{y})^2 = (225 - 209,8)^2 + (184 - 209,8)^2 + \cdots + (180 - 209,8)^2 = 2800,8,$$

logo, obtivemos uma redução dos resíduos utilizando o modelo de regressão linear em relação ao modelo mais simples. Mais especificamente,

$$SQReg = 2800,8 - 160,6467 = 2640,153,$$

representa o ganho em unidades em assumir o modelo de regressão linear. E,

$$R^2 = \frac{2640,153}{2800,8} = 0,9426426,$$

indica que o lucro relativo do modelo de regressão linear é de aproximadamente 94%, isto é, ele explica 94% da variabilidade total.

5.0.2 Tabela ANOVA para o Modelo de Regressão Linear

Podemos resumir algumas informações no formato da tabela ANOVA para um modelo de regressão linear:

Tabela 5.1: *ANOVA*

Fontes de Variação	GL	SQ	QM	F
Regressão	1	$SQReg$	$QMReg = SQReg$	$QMReg/S_e^2$
Resíduo	$n-2$	$SQRes$	$SQRes/(n-2) = S_e^2$	
Total	$n-1$	SQT	$SQT/(n-1) = S^2$	

A estatística F será usada mais adiante.

5.0.3 Distribuições Amostrais dos Estimadores dos Coeficientes

É possível mostrar que

$$\hat{\alpha} \sim Normal\left(\alpha, \frac{\sigma^2 \sum_{i=1}^n x_i^2}{n\sum_{i=1}^n (x_i - \bar{x})^2}\right) \Rightarrow \frac{(\hat{\alpha} - \alpha)}{\sigma}\sqrt{\frac{n\sum_{i=1}^n (x_i - \bar{x})^2}{\sum_{i=1}^n x_i^2}} \sim Normal(0,1).$$

$$\hat{\beta} \sim Normal\left(\beta, \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right) \Rightarrow \frac{(\hat{\beta} - \beta)}{\sigma}\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sim Normal\left(0,1\right).$$

Conhecendo as distribuições amostrais de estatísticas que envolvem o parâmetro de interesse, um estimador para esse parâmetro e constantes conhecidas, podemos construir

intervalos de confiança e testes de hipóteses para os parâmetros. Dessa forma, substituindo σ pelo seu estimador S_e, obtemos que

$$\frac{(\hat{\alpha}-\alpha)}{S_e}\sqrt{\frac{n\sum_{i=1}^{n}(x_i-\bar{x})^2}{\sum_{i=1}^{n}x_i^2}} \sim t_{n-2}.$$

$$\frac{(\hat{\beta}-\beta)}{S_e}\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2} \sim t_{n-2}.$$

5.0.4 Intervalos de Confiança para os Coeficientes do Modelo de Regressão Linear

Para um nível de significância γ, os intervalos de confiança para os coeficientes do modelo de regressão linear são dados por:

$$\begin{aligned} IC(\alpha;\gamma) &= \left[\hat{\alpha} - t_{\gamma,n-2}S_e\sqrt{\frac{\sum_{i=1}^{n}x_i^2}{n\sum_{i=1}^{n}(x_i-\bar{x})^2}}, \hat{\alpha} + t_{\gamma,n-2}S_e\sqrt{\frac{\sum_{i=1}^{n}x_i^2}{n\sum_{i=1}^{n}(x_i-\bar{x})^2}}\right]; \\ IC(\beta;\gamma) &= \left[\hat{\beta} - t_{\gamma,n-2}S_e\sqrt{\frac{1}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}, \hat{\beta} + t_{\gamma,n-2}S_e\sqrt{\frac{1}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}\right]. \end{aligned}$$

Exemplo 21 *Retornando ao exemplo, seja* $\gamma = 5\%$*. Temos que* $t_{\gamma,n-2} = 3,18$*. Pelas contas anteriores:* $\sum_{i=1}^{n}(x_i-\bar{x})^2 = 1$*,* $\sum_{i=1}^{n}x_i^2 = 21$ *e* $S_e = \sqrt{53,55} = 7,32$*. Daí, os i.c.'s de 95% para* α *e* β*, respectivamente, são dados por:*

$$\begin{aligned} IC(\alpha;\gamma) &= \left[109,03 - 3,18\times 7,32\sqrt{\frac{21}{5\times 1}}; 109,03 + 3,18\times 7,32\sqrt{\frac{21}{5\times 1}}\right] \\ &= [109,03 - 47.70488; 109,03 + 47,70488] = [61,32512; 156,7349]; \\ IC(\beta;\gamma) &= \left[50,39 - 3,18\times 7,32\sqrt{\frac{1}{1}}; 50,39 + 3,18\times 7,32\sqrt{\frac{1}{1}}\right] \\ &= [50,39 - 23,2776; 50,39 + 23,2776] = [27,1124; 73,6676]. \end{aligned}$$

5.0.5 Testes de Significância dos Coeficientes do Modelo de Regressão Linear

Suponha que há interesse em testar se o coeficiente linear da função de regressão linear é significativo, isto é, se ele é diferente de zero. Isso ocorre quando, por exemplo R^2 é baixo e surge dúvidas quanto ao ganho em introduzir o modelo de regressão linear. Abaixo é descrito como devemos proceder nesse caso.

Hipóteses do teste:

H_0: $\beta = 0$;

H_a: $\beta \neq 0$

Estatística do teste:

$$\frac{\hat{\beta}}{S_e}\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \sim t_{n-2}.$$

Como é um teste bilateral, a região crítica será delimitada pelos valores tabelados da t-Student com $n-2$ graus de liberdade para um dado nível de significância γ.

Alternativamente, podemos utilizar que

$$\frac{\hat{\beta}^2}{S_e^2}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{QMReg}{S_e^2} \sim F_{(1,n-2)},$$

que é a F que aparece na tabela ANOVA.

Se aceitarmos H_0, então o modelo linear não parece adequado para descrever a relação entre X e Y.

Também podemos estar interessados em testar se o coeficiente linear (ponto que a reta toca o eixo y) da função de regressão linear é significativo, isto é, se ele é diferente de zero. Segue abaixo a descrição do teste.

Hipóteses do teste:

H_0: $\alpha = 0$;

H_a: $\alpha \neq 0$.

Estatística do teste:

$$\frac{\hat{\alpha}}{S_e}\sqrt{\frac{n\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sum_{i=1}^{n} x_i^2}} \sim t_{n-2}.$$

Como é um teste bilateral, a região crítica será delimitada pelos valores tabelados da t-Student com $n-2$ graus de liberdade para um dado nível de significância γ,.

Exemplo 22 *No exemplo, vamos testar ambas as hipóteses.*

Hipóteses:

H_0*:* $\alpha = 0$;

H_a*:* $\alpha \neq 0$.

Calculando: $\frac{\hat{\alpha}}{S_e}\sqrt{\frac{n\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sum_{i=1}^{n} x_i^2}} = \frac{109,03}{7,32}\sqrt{\frac{5*1}{21}} \approx 7,27.$

O valor tabelado para $\gamma = 5\%$ *é* $t_{n-2} = t_3 = 3,18$. *Assim, 7,27 está na região crítica, portanto, rejeitamos* H_0*, indicando que o coeficiente* α *é necessário no modelo.*

Hipóteses do teste:

H_0*:* $\beta = 0$;

H_a*:* $\beta \neq 0$

Calculando: $\frac{\hat{\beta}}{S_e}\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{50,39}{7,32} * 1 \approx 6,88$.

O valor tabelado para $\gamma = 5\%$ *é* $t_{n-2} = t_3 = 3,18$. *Assim, 6,88 está na região crítica, portanto, rejeitamos* H_0*, indicando que o coeficiente* β *é necessário no modelo.*

Utilizando o teste F*,* $\frac{QMReg}{S_e^2} = \frac{2640,153}{53,55} = 49,30 > 10,13 = F_{(1,5-2)}$*, logo pertence a região de rejeição de* H_0.

5.0.6 Intervalo de Confiança para a Média do Modelo de Regressão Linear

Seja $\mu(x_i) = \alpha + \beta x_i$ a média da regressão, então o estimador para a média é dado por $\hat{\mu}(x_i) = \hat{\alpha} + \hat{\beta} x_i$ e temos que

$$\hat{\mu}(x_i) \sim Normal\left(\alpha + \beta x_i, \sigma^2\left[\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right]\right),$$

e, substituindo σ^2 pelo seu estimador S_e^2 obtemos que

$$\frac{\hat{\mu}(x_i) - (\alpha + \beta x_i)}{S_e\sqrt{\left[\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right]}} \sim t_{n-2}.$$

Portanto, o intervalo de $(1-\gamma)\%$ de confiança para a média $\mu(x_i)$ é dado por

$$\left[\hat{\mu}(x_i) - t_{\gamma,n-2}S_e\sqrt{\left[\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right]}, \hat{\mu}(x_i) + t_{\gamma,n-2}S_e\sqrt{\left[\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right]}\right].$$

Exemplo 23 *No exemplo, quanto seria o total de vendas para um gasto com propaganda de 2.400,00 reais? Uma estimação pontual seria calculada por*

$$\hat{\mu}(x_i) = \hat{y}_i = 109,03 + 50,39 \times 2,4 = 229,966,$$

que é o valor de venda médio com 2,4 de gasto. Assim, o intervalo de confiança de 95% é dado por

$$\left[229,966 - 3,18 * 7,32 * \sqrt{\frac{1}{5} + ((2,5-2)^2)}; 229,966 + 3,18 * 7,32 * \sqrt{\frac{1}{5} + ((2,5-2)^2)}\right]$$

$$= [229,966 - 28,02994; 229,966 + 28,02994] = [201,9361; 257,9959].$$

5.0.7 Intervalo de Predição

Podemos estar interessados em prever o comportamento para uma nova observação, dentro do intervalo ou futura, e, ainda, construir um intervalo de confiança para essa previsão. Esse intervalo é chamado de intervalo de predição. Para prever uma nova observação, iremos utilizar o mesmo estimador pontual para a média, entretanto, chamaremos o valor futuro da variável X de x_f e o valor predito de Y por y_f. O modelo nos diz que

$$Y_f = \mu(x_f) + \epsilon_f$$

com estimador $\hat{Y}_f = \hat{y}_f = \hat{\alpha} + \hat{\beta}x_f$.

É possível mostrar que

$$\hat{y}_f \sim Normal\left(\hat{\alpha} + \hat{\beta}x_f, \sigma^2\left[1 + \frac{1}{n} + \frac{(x_f - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right]\right),$$

portanto o intervalo de predição de $(1-\gamma)\%$ para y_f é dado por

$$\left[\hat{y}_f - t_{\gamma,n-2}S_e\sqrt{\left[1 + \frac{1}{n} + \frac{(x_f - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right]}, \hat{y}_f + t_{\gamma,n-2}S_e\sqrt{\left[1 + \frac{1}{n} + \frac{(x_f - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right]}\right].$$

No exemplo, quanto seria o total de vendas para um gasto com propaganda de 2.500,00 reais? Uma estimação pontual seria calculada por

$$\hat{y}_f = 109,03 + 50,39 \times 2,5 = 235,005,$$

que é o valor de venda médio com 2,5 de gasto. Assim, o intervalo de confiança de 95% é dado por

$$\left[235,005 - 3,18 * 7,32 * \sqrt{1 + \frac{1}{5} + ((2,5-2)^2)}; 235,005 + 3,18 * 7,32 * \sqrt{\frac{1}{+}15 + ((2,5-2)^2)}\right]$$

$$= [235,005 - 28,02994; 235,005 + 28,02994] = [206,9751; 263,0349].$$

Capítulo 6

Números Índices

Definição 12 *O índice ou número índice é um quociente que expressa a variação relativa entre os valores de qualquer medida.*

Números índices são úteis para indicar variações relativas em quantidades, preços ou valores de um item (ou itens) ao longo do tempo. Se apenas um item (produto), ou variável, é computado temos um número índice simples. Por outro lado, se vários itens (produtos), ou variáveis, têm suas variações computadas temos um número índice composto.
Neste texto, o foco está nos índices econômicos, que envolvem variações de preços, quantidades e valores ao longo do tempo.

6.1 Relativos

Os relativos (ou índices simples) fazem comparação entre duas épocas - época atual (t) e época base (0) - para um único produto:

Relativo de preço	Relativo de quantidade	Relativo de valor
$p_{0,t} = \dfrac{p_t}{p_0}$	$q_{0,t} = \dfrac{q_t}{q_0}$	$v_{0,t} = \dfrac{v_t}{v_0}$

Na tabela, p_0 e p_t representam os preços nos anos base e atual, respectivamente; q_0 e q_t representam as quantidades nos anos base e atual, respectivamente; v_0 e v_t representam os preços nos anos base e atual, respectivamente, e, lembrando que $v = p \times q$.
Note que: $v_{0,t} = p_{0,t} \times q_{0,t}$.

Na tabela a seguir temos o preço e a quantidade de arroz consumida por uma família no último trimestre de 2001:

	Outubro		Novembro		Dezembro	
	Preço	Quant.	Preço	Quant.	Preço	Quant.
Arroz (kg)	2	5	2	8	3	8
Valor	$2 \times 5 = 10$		$2 \times 8 = 16$		$3 \times 8 = 24$	

O relativo de preço exprime quanto o preço da época atual é maior ou menor que o preço da época base. A partir dele podemos obter a taxa de variação, que mede a variação relativa.

A variação relativa é definida como

$$p\% = \frac{p_t - p_0}{p_0} = \frac{p_t}{p_0} - 1 = p_{0,t} - 1.$$

que, em geral, é apresentada em forma percentual, multiplicando-se o valor por 100. No numerador da taxa de variação, temos a variação absoluta de preços: $p_t - p_0$. Definições análogas valem para quantidade e valor.

Exemplo 24 *Vamos calcular os índices de preço, quantidade e valor tomando como base outubro (apresentar a tabela com as séries dos relativos e mostrar a taxa de variação). Depois, mude o mês base para novembro.*

6.2 Propriedades dos Números Índices

Apresentamos abaixo propriedades que os relativos possuem e que servem como critérios de avaliação para outras formulações de números índices.

1. Identidade: Se a data-base coincidir com a data atual, o índice é sempre 1 (ou 100, no caso de base 100), isto é,

$$I_{t,t} = 1.$$

2. Reversão (ou inversão) no tempo: Invertendo-se os períodos de comparação, os índices são obtidos um como o inverso do outro, isto é,

$$I_{0,t} = \frac{1}{I_{t,0}} \Leftrightarrow I_{0,t} \times I_{t,0} = 1.$$

3. Circular: Se o intervalo de análise é decomposto em vários subintervalos, o índice pode ser obtido como o produto dos índices nos subintervalos, isto é,

$$I_{0,1} \times I_{1,2} \times I_{2,3} \times \ldots \times I_{t-1,t} = I_{0,t}.$$

Esta propriedade é útil quando conhecemos os índices nas épocas intermediárias, o índice de todo o período pode ser calculado sem que haja necessidade de recorrer aos valores que deram origem aos cálculos individuais. Note que, como decorrência desta propriedade, podemos escrever:

$$I_{0,t} = I_{0,t-1} \times I_{t-1,t}.$$

Se o índice satisfizer também o princípio de reversibilidade, então o critério de circularidade é equivalente a

$$I_{0,1} \times I_{1,2} \times I_{2,3} \times \ldots \times I_{t-1,t} \times I_{t,0} = 1.$$

4. Decomposição das causas (ou reversão dos fatores): Denotando por IV , IP e IQ os índices de valor, preço e quantidade respectivamente, o critério da decomposição das causas requer que

$$IV = IP \times IQ$$

5. Homogeneidade: Mudanças de unidade não alteram o valor do índice.

6. Proporcionalidade: Se todas as variáveis envolvidas no índice tiverem a mesma variação, então o índice resultante terá a mesma variação.

Exercício: Verifique que estas propriedades são satisfeitas pelos relativos.

Mudanças de unidade envolvem multiplicação por uma constante (quilo para tonelada, reais para milhões de reais, etc). Tais operações não alteram o valor do relativo, uma vez que numerador e denominador são multiplicados pelo mesmo valor.

6.3 Elos de relativo e relativos em cadeia

Índices envolvendo épocas adjacentes, no caso de relativos, eles são, às vezes, denominados elos relativos. Os elos relativos estabelecem comparações binárias entre épocas adjacentes $\frac{p_t}{p_{t-1}}$, $\frac{q_t}{q_{t-1}}$ e $\frac{v_t}{v_{t-1}}$.

Na tabela a seguir temos dados de preço para 5 anos e calculam-se os elos de relativos e os relativos em cadeia, ano a ano.

Ano	Preço	Elos relativos p_t/p_{t-1}	Relativos em cadeia
1995	200		
1996	250	$250/200 = 1,25$	$1,25 = p_{95,96}$
1997	300	$300\,/\,250 = 1,20$	$1,2 \times 1,25 = 1,5 = p_{95,97}$
1998	390	$390\,/\,300 = 1,30$	$1,2 \times 1,25 \times 1,3 = 1,95 = p_{95,98}$
1999	468	$468\,/\,390 = 1,20$	$1,2 \times 1,25 \times 1,3 \times 1,2 = 2,34 = 9_{95,99}$

o que está em concordância com:

Ano	Relativo de preço Base: 1995 = 100
1995	100
1996	$100 \times 250\,/\,200 = 125 \Rightarrow 25\%$
1997	$100 \times 300\,/\,200 = 150 \Rightarrow 50\%$
1998	$100 \times 390\,/\,200 = 195 \Rightarrow 95\%$
1999	$100 \times 468\,/\,200 = 234 \Rightarrow 134\%$

A propriedade circular também envolve a multiplicação desses índices; para os relativos, tal operação é denominada relativos em cadeia e como a propriedade circular é satisfeita pelos relativos, tal multiplicação resulta no relativo do período. Assim, temos que:

- elos relativos : $p_{1,2}; p_{2,3}; p_{3,4}; ...; p_{t-1,t};$
- relativos em cadeia : $p_{1,2} \times p_{2,3} \times p_{3,4} \times \ \times p_{t-1,t} = p_{1,t}$.

Exemplo 25

Referências

- ANDERSON, D. R.; SWEENEY, D. J.; WILLIAMS. Estatística Aplicada à Economia e Administração. Tradução: Luiz Sérgio de Castro Paiva. 1ed. São Paulo: Pioneira Thomson Learning, 2003.
- BUSSAB. W. O. ; MORETTIN, P. A. Estatística Básica. 5ed. Saraiva, 2003.
- FREUND, JOHN E.; SIMON, GARY A. Estatística Aplicada: economia, administração e contabilidade. Tradução: Alfredo Alves de farias. 9ed. Porto Alegre: Bookman, 2000.

- HAIR, JOSEPH F.; ANDERSON, ROLPH E.; TATHAM, RONALD L.; BLACK, WILLIAN C. Análise Multivariada de Dados. Tradução: AdonaiSchlup SabtáAnna e Anselmo Chaves Neto. Porto Alegre:Bookman,5ed.

- JAMES, Barry R. Probabilidade: um curso em nível intermediário. 2ed., Rio de Janeiro: IMPA, Projeto Euclides, 2002.

- LARSON, Ron; FARBER, Besty. Estatística Aplicada. Tradução: Luciane Ferreira Pauleti Vianna. 4ed. São Paulo: Pearson Prentice Hall, 2010.

- MAGALHÃES, M.N ; LIMA, A.C.P DE. Noções de Probabilidade e Estatística. 5ed., São Paulo: Ed. Edusp, 2005.

- MONTGOMERY, D.C., PECK, E.A.., VINING, G. G.; Introduction to Linear Regression Analysis, John Wiley & Sons, 3rd edition, 2001.

- VIEIRA, Sônia. Análise de Variância:(Anova). São Paulo: Atlas, 2006.

- Walpole, R. E., Myers, R. H., Myers, S. L., Ye, K., 2009. Probabilidade e Estatística para engenharia e ciências. Pearson, 8ed., São Paulo.