

Análise de Variância

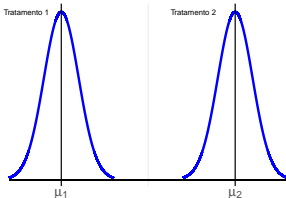
Gilberto Pereira Sassi

Universidade Federal da Bahia
Instituto de Matemática e Estatística
Departamento de Estatística

Considere $N(\mu_1, \sigma^2)$ e $N(\mu_2, \sigma^2)$ conforme ilustrado na Figura 1a.

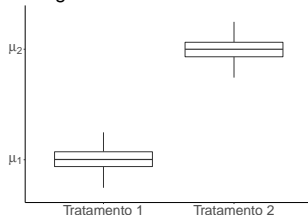
Distribuição normal

Variância: σ^2



(a) Variáveis aleatórias.

Diagrama de caixa



(b) Primeira impressão.

Figura 1: Comparação de médias

Queremos decidir entre as hipóteses:

$$H_0 : \mu_1 = \mu_2 \quad \text{e} \quad H_1 : \mu_1 \neq \mu_2,$$

Figura 1b nos indica que rejeitaremos H_0 .

Distribuição normal
Variância: σ^2



Figura 2: Comparação de médias

$$H_0 : \mu_1 = \mu_2 = \mu_3 \text{ e } H_1 : \mu_i \neq \mu_j, \text{ para menos um par } (i, j), i, j \in \{1, 2, 3\}, i \neq j,$$

Figura 2b nos indica que rejeitaremos H_0 .

- ▶ Suponha que queremos testar $H_0 : \mu_1 = \dots = \mu_a$;
- ▶ Chamamos μ_1, \dots, μ_a de tratamentos;
- ▶ Imagine que temos uma amostra de tamanho n para cada tratamento: $y_{i1}, \dots, y_{in}, i = 1, \dots, a$, conforme ilustrado na Tabela 1;

Tratamento	Observações				Totais	Médias
1	y_{11}	y_{12}	\dots	y_{1n}	$y_{1\cdot}$	$\bar{y}_{1\cdot} = \frac{y_{1\cdot}}{n}$
2	y_{21}	y_{22}	\dots	y_{2n}	$y_{2\cdot}$	$\bar{y}_{2\cdot} = \frac{y_{2\cdot}}{n}$
\vdots	\vdots	\vdots		\vdots	\vdots	\vdots
\vdots	\vdots	\vdots		\vdots	\vdots	\vdots
n	y_{n1}	y_{n2}	\dots	y_{nn}	$y_{n\cdot}$	$\bar{y}_{n\cdot} = \frac{y_{n\cdot}}{n}$
					$y_{\cdot\cdot}$	$\bar{y}_{\cdot\cdot} = \frac{y_{\cdot\cdot}}{n \cdot a}$

Tabela 1: Dados de um estudo completamente aleatórios e balanceados com um fator.

- ▶ Descrevemos os dados da Tabela 1 através de

$$y_{ij} = \mu + \tau_i + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \sigma^2); \quad (1)$$

- ▶ Queremos testar as seguintes hipóteses

$$H_0 : \tau_1 = \tau_2 = \dots = \tau_a = 0,$$

$$H_1 : \tau_i \neq 0 \text{ para pelo menos um } i, i \in \{1, 2, \dots, a\}.$$

Chamamos $\tau_i, i = 1, \dots, a$ de efeitos do tratamentos.

Soma dos quadrados.

Note que

$$\begin{aligned}
 SS_T &= \underbrace{\sum_{i=1}^a \sum_{j=1}^n (y_{ij} - \bar{y}_{..})^2}_{\text{Soma dos quadrados total}} = (N - 1)s^2 \\
 &= \underbrace{\sum_{i=1}^a n(\bar{y}_{i.} - \bar{y}_{..})^2}_{\text{Soma dos Quadrados dos Tratamentos}} + \underbrace{\sum_{i=1}^a \sum_{j=1}^n (y_{ij} - \bar{y}_{i.})^2}_{\text{Soma dos Quadrados do Erro}} \\
 &= \underbrace{\sum_{i=1}^a n(\bar{y}_{i.} - \bar{y}_{..})^2}_{SQ_{Tratamentos}} + \underbrace{\sum_{i=1}^a (n - 1)s_i^2}_{SQ_E} = SQ_{Tratamentos} + SQ_E.
 \end{aligned}$$

em que

- ▶ $s^2 = \frac{1}{N-1} \sum_{i=1}^a \sum_{j=1}^n (y_{ij} - \bar{y}_{..})^2$, em que $N = n \cdot a$ (variância total);
- ▶ $s_i^2 = \frac{1}{n-1} \sum_{j=1}^n (y_{ij} - \bar{y}_{i.})^2$ (variância do tratamento i).

Se $H_0 : \tau_1 = \tau_2 = \dots = \tau_a = 0$ é verdadeira, então $Y_{ij} \sim N(\mu, \sigma^2), \forall i, j$.

Quadrado médio.

Usando a equação (1), podemos provar que

- ▶ $E(SQ_{Tratamentos}) = (a - 1)\sigma^2 + n \cdot \tau_1^2 + \dots + n \cdot \tau_a^2;$
- ▶ $E(SQ_E) = (N - a)\sigma^2;$
- ▶ Quadrado Médio dos Tratamentos: $QM_{Tratamentos} = \frac{SQ_{Tratamentos}}{a - 1};$
- ▶ Quadrado Médio do Erro: $QM_E = \frac{SQ_E}{N - a};$
- ▶ $F_0 = \frac{QM_{Tratamentos}}{QM_E}.$

É possível provar que

- ▶ $E(QM_{Tratamentos}) = \sigma^2 + \frac{n \cdot \tau_1^2 + \dots + n \cdot \tau_a^2}{a - 1};$
- ▶ $E(QM_E) = \sigma^2;$

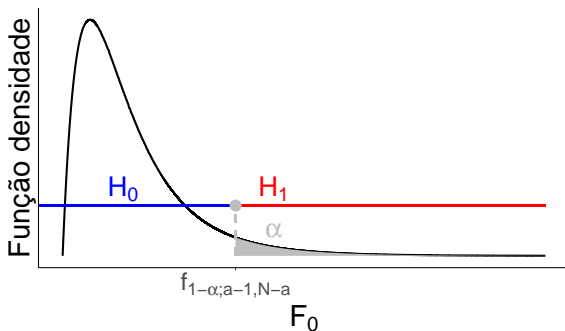
Rejeitamos $H_0 : \tau_1 = \tau_2 = \dots = \tau_a = 0$ se $F_0 = \frac{QM_{Tratamentos}}{QM_E}$ for grande.

Se H_0 for verdadeiro, temos que $F_0 \sim F_{a-1, N-a}$, em que $F_{a-1, N-a}$ é a distribuição F .
 Teste ANOVA é robusto para tratamentos com variâncias ligeiramente diferentes.

ANOVA

Rejeitamos $H_0: \tau_1 = \tau_2 = \dots = \tau_a = 0$ se $f_0 > f_{1-\alpha; a-1, N-a}$

$$RC = \{f_0 \mid f_0 > f_{1-\alpha; a-1, N-a}\}$$



$$f_0 = \frac{QM_{\text{Tratamentos}}}{QM_E} \text{ e } N = n \cdot a$$

Figura 3: Região crítica para ANOVA.

ANOVA: Checando as suposições do modelo.

Análise de resíduo

Depois decidir entre H_0 e H_1 , precisamos checar as suposições do modelo:

- (a) $\epsilon_{ij}, j = 1, \dots, n, i = 1, \dots, a$ são independentes;
- (b) $\epsilon_{ij}, j = 1, \dots, n, i = 1, \dots, a$ tem distribuição normal;
- (c) $\epsilon_{ij}, j = 1, \dots, n, i = 1, \dots, a$ tem variância constante.

$\epsilon_{ij} = y_{ij} - \mu + \tau_i = y_{ij} - \mu_i, j = 1, \dots, n, i = 1, \dots, a$ não são observáveis pois não conhecemos as médias populacionais dentro de cada tratamento, mas podemos aproximar estas médias populacionais μ_i por $\bar{y}_{i\cdot}, i = 1, \dots, a$. Assim, podemos definimos os resíduos por

$$e_{ij} = y_{ij} - \bar{y}_{i\cdot}, j = 1, \dots, n, i = 1, \dots, a. \quad (2)$$

Usando a equação (2), podemos usar gráficos para checar (a), (b) e (c):

- (a) **Independência:** cada par $(i + j, e_{ij}), j = 1, \dots, n, i = 1, \dots, a$, é representado por um ponto no plano cartesiano. Se não existe padrão ou tendência, então assumimos que ϵ_{ij} são independentes;
- (b) **Normalidade:** cada par $\left(q_{(i+j)}; \frac{e_{(i+j)} - \bar{e}_{\cdot\cdot}}{s_e} \right), j = 1, \dots, n, i = 1, \dots, a$ é representado por um ponto no plano cartesiano, em que $\Phi(q_{(i+j)}) = \frac{i+j-0,5}{n}$ e s_e é o desvio padrão dos resíduos – este gráfico é chamado de gráfico de probabilidade normal. Se os pontos estão próximos ou em cima da reta $y = x$, então ϵ_{ij} tem distribuição normal;
- (c) **Variância constante:** cada par $(\bar{y}_{i\cdot}, e_{ij}), j = 1, \dots, n, i = 1, \dots, a$, é representando por um ponto no plano cartesiano. Se não existe padrão ou tendência, então assumimos que ϵ_{ij} tem variância constante.

ANOVA

Exemplo

Em um experimento que deseja analisar se a resistência à tração de um tecido está associada com a porcentagem da fibra de algodão. Cinco níveis de porcentagem da fibra de algodão e cinco observações em cada nível de porcentagem de fibra de algodão. Os dados estão na Tabela 2. Desenhe diagramas de caixa para cada nível de porcentagem de algodão. A porcentagem de algodão afeta a resistência à tração do tecido? Use $\alpha = 5\%$. Calcule o valor-p.

Algodão	Observações					Total	Variância em cada tratamento
15	7	7	15	11	9	49	11,20
20	12	17	12	18	18	77	9,80
25	14	18	18	19	19	88	4,30
30	19	25	22	19	23	108	6,80
35	7	10	11	15	11	54	8,20
							Variância total = 26,54

Tabela 2: Resistência à tração

ANOVA

Solução

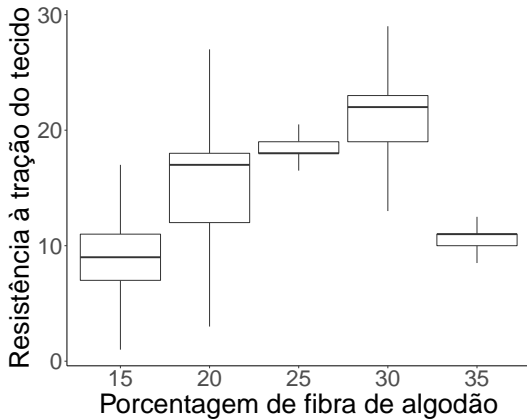


Figura 4: Diagrama de caixa com nível de porcentagem de algodão.

ANOVA

Solução

As médias populacionais de resistência à tração com porcentagem de fibra algodão são: $\mu_{15} = \mu + \tau_{15}$,

$\mu_{20} = \mu + \tau_{20}$, $\mu_{25} = \mu + \tau_{25}$, $\mu_{30} = \mu + \tau_{30}$ e $\mu_{35} = \mu + \tau_{35}$.

Passo 1) Queremos testar as hipóteses: $H_0 : \tau_{15} = \tau_{20} = \tau_{25} = \tau_{30} = \tau_{35} = 0$ e

$H_1 : \tau_{15}^2 + \tau_{20}^2 + \tau_{25}^2 + \tau_{30}^2 + \tau_{35}^2 > 0$;

Passo 2) Nível de significância $\alpha = 5\%$;

Passo 3) Rejeitamos H_0 se $F_0 = \frac{QM_{Tratamentos}}{QM_E}$ for grande. Ou seja, $RC = \{f_0 \mid f_0 > f_{1-\alpha; a-1, N-a}\}$, em que

$N = a \cdot n = 5 \cdot 5 = 25$ e $a = 5$;

Passo 4) Vamos encontrar o valor crítico:

► $P(F_{a-1, N-a} \leq f_{1-\alpha; a-1, N-a}) = P(F_{4, 20} \leq f_{0,95;4,20}) = 1 - \alpha = 0,95$, então $f_{0,95;4,20} = 2,8661$;

Passo 5) Na Tabela 3, mostramos a soma dos quadrados e os quadrados médios:

Fonte de variação	Graus de liberdade	Soma dos Quadrados	Quadrados Médios	F_0
Tratamentos	$a - 1 = 4$	$SQ_{Tratamentos} = 475,76$	$QM_{Tratamentos} = 118,94$	$\frac{QM_{Tratamentos}}{QM_E} = 14,76$
Erro	$N - a = 20$	$SQ_E = 161,20 = (n-1)s_1^2 + (n-1)s_2^2 + (n-1)s_3^2 + (n-1)s_4^2$	$QM_E = 8,06$	
Total	$N - 1 = 24$	$SQ_T = 636,96 = (N-1)s^2$		

Tabela 3: Tabela ANOVA.

Como $F_0 = 14,76 > f_{0,95;4,20} = 2,8661$, rejeitamos H_0 , ou seja, ao nível de significância $\alpha = 5\%$ as resistências médias não são iguais para os diferentes níveis de porcentagem de fibra de algodão.

ANOVA

Solução (valor-p)

O valor-p é dado por

$$p = P(F_0 > f_0 \mid H_0) = 1 - P(F_{a-1, N-a} < f_0).$$

Usando as informações da Tabela 3, temos que $f_0 = 14,76$. Então, o valor-p é dado por

$$\begin{aligned} p &= 1 - P(F_{a-1, N-a} \leq f_0) \\ &= 1 - P(F_{4,20} \leq 14,76) && \text{Calculei usando o R} \\ &= 1 - 1 \\ &= 0 \end{aligned}$$

Como $p = 0 < \alpha = 0,05$, rejeitamos H_0 . Então, as resistências médias não são iguais para os diferentes níveis (15, 20, 25, 30 e 35) de porcentagem de fibra de algodão.

ANOVA

Solução (Análise de resíduos)

Primeiro calculamos os resíduos e médias dentro de cada tratamento:

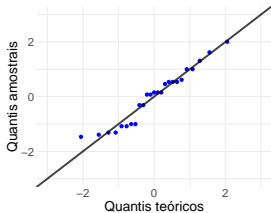
Algodão	Resíduos					Média em cada tratamento	Variância em cada tratamento
15	-2,8	-2,8	5,2	1,2	-0,8	9,80	11,20
20	-3,4	1,6	-3,4	2,6	2,6	15,40	9,80
25	-3,6	0,4	0,4	1,4	1,4	17,60	4,30
30	-2,6	3,4	0,4	-2,6	1,4	21,60	6,80
35	-3,8	-0,8	0,2	4,2	0,2	10,80	8,20
							Variância total = 26,54

Tabela 4: Resistência à tração

- (a) Na Figura 5a, os pontos estão próximos perto da reta $y = x$ e concluímos que ϵ_{ij} tem distribuição normal;
- (b) Na Figura 5b, não existe padrão ou tendência e concluímos que ϵ_{ij} são independentes;
- (c) Na Figura 5c, não existe padrão ou tendência e concluímos que ϵ_{ij} tem variância constante.

ANOVA

Análise de resíduo



(a) Gráfico de probabilidade normal.

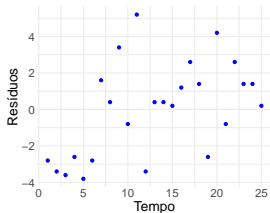
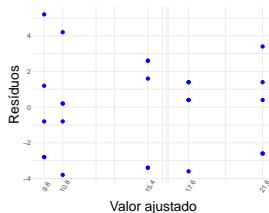
(b) Tempo \times Resíduos.(c) $\bar{y}_{j\cdot} \times$ Resíduos.

Figura 5: Análise de resíduos.

ANOVA

Exemplo

Imagine que um pesquisador realizou um experimento completamente aleatório e balanceado para checar a tratamentos ou níveis de um fator. Algumas informações do experimento estão na Tabela 5. Complete a Tabela 5. As médias dos tratamentos são diferentes? Use $\alpha = 5\%$. Calcule o valor-p.

Fonte de variação	Graus de liberdade	Soma dos Quadrados	Quadrados Médios	F_0
Tratamentos	$a - 1 = 3$	$SQ_{Tratamentos} = ?$	$QM_{Tratamentos} = ?$	$\frac{QM_{Tratamentos}}{QM_E} = ?$
Erro	$N - a = ?$	$SQ_E = (n - 1)s_1^2 + (n - 1)s_2^2 + (n - 1)s_3^2 + (n - 1)s_4^2 = 396,8$	$QM_E = ?$	
Total	$N - 1 = 19$	$SQ_T = (N - 1)s^2 = 514,2$		

Tabela 5: Algumas informações do experimento.

ANOVA

Solução

Passo 1) Note que $a = 4$ e queremos testar as seguintes hipóteses:

$$H_0 : \tau_1 = \tau_2 = \tau_3 = \tau_4 = 0 \text{ e } H_1 : \tau_1^2 + \tau_2^2 + \tau_3^2 + \tau_4^2 > 0;$$

Passo 2) Nível de significância $\alpha = 5\%$;

Passo 3) Rejeitamos H_0 se $F_0 = \frac{QM_{Tratamentos}}{QM_E}$ for grande. Ou seja,

$$RC = \{f_0 \mid f_0 > f_{1-\alpha; a-1, N-a}\}.$$

Passo 4) Note que $N - a = (N - 1) - (a - 1) = 16$. Vamos encontrar o valor crítico:

$$\begin{aligned} \blacktriangleright P(F_{a-1, N-a} \leq f_{1-\alpha; a-1, N-a}) &= P(F_{3, 16} \leq f_{0,95; 3, 16}) = 1 - \alpha = 0,95, \text{ então} \\ f_{0,95; 3, 16} &= 3,2389. \end{aligned}$$

Passo 5) Completamos as informações da Tabela 5 na Tabela 6.

	Graus de liberdade	Soma dos Quadrados	Quadrados Médios	F_0
Tratamentos	$a - 1 = 3$	$SQ_{Tratamentos} = 514,2 - 396,8 = 117,4$	$QM_{Tratamentos} = \frac{117,4}{3} = 39,13$	$\frac{QM_{Tratamentos}}{QM_E} = \frac{39,13}{24,8} = 1,58$
Erro	$N - a = 19 - 3 = 16$	$SQ_E = (n-1)s_1^2 + (n-1)s_2^2 + (n-1)s_3^2 + (n-1)s_4^2 = 396,8$	$QM_E = \frac{396,8}{16} = 24,8$	
Total	$N - 1 = 19$	$SQ_T = (N-1)s^2 = 514,2$		

Tabela 6: Algumas informações do experimento.

Como $f_0 = 1,58 < f_{0,95; 3, 16} = 3,2389$, não rejeitamos H_0 . Ou seja, ao nível de significância $\alpha = 5\%$ não temos evidência as médias dos tratamentos são diferentes.

ANOVA

Solução (valor-p)

O valor-p é dado por

$$p = P(F_0 > f_0 \mid H_0) = 1 - P(F_{a-1, N-a} < f_0).$$

Usando as informações da Tabela 6, temos que $f_0 = 1,58$. Então, o valor-p é dado por

$$\begin{aligned} p &= 1 - P(F_{a-1, N-a} \leq f_0) \\ &= 1 - P(F_{3, 16} \leq 1,58) \quad \text{Calculei usando o R} \\ &= 1 - 0,7668 \\ &= 0,2332 \end{aligned}$$

Como $p = 0,2332 > \alpha = 0,05$, não rejeitamos H_0 . Então, não temos evidência para afirmar que as médias dos tratamentos são diferentes ao nível de significância $\alpha = 5\%$.

Poder do teste (estudos balanceados).

Imagine que

- ▶ Hipóteses: $H_0 : \tau_1 = \dots = \tau_a = 0$ e $H_1 : \tau_1^2 + \dots + \tau_a^2 > 0$;
- ▶ H_1 é verdade, então $\tau_1^2 + \dots + \tau_a^2 > 0$. Chamamos $f = \sqrt{\frac{\tau_1^2 + \dots + \tau_a^2}{a\sigma^2}}$ de *size effect*;
- ▶ $F_0 = \frac{QM_{Tratamentos}}{QME} \sim F_{a-1, N-a}(\Phi^2)$, em que $\Phi^2 = n \cdot f^2$ e n é o número de observações em cada tratamento;
- ▶ Ao nível de significância α , temos $RC = \{f_0 \mid f_0 > f_{1-\alpha; a-1, N-a}\}$.

Poder do teste é dado por

$$1 - \beta = 1 - P(F_0 \leq f_{1-\alpha; a-1, N-a} \mid H_0) = 1 - P(F_{a-1, N-a}(\Phi^2) \leq f_{1-\alpha; a-1, N-a})$$

A **Função Poder**, dado o tamanho da amostra N , o número de tratamentos a e os tamanhos dos tratamentos são n , é $\pi : (0, \infty) \rightarrow [0, 1]$ dada por

$$\pi(f) = 1 - P(F_{a-1, N-a}(n \cdot f^2) \leq f_{1-\alpha; a-1, N-a}), \quad f \in (0, \infty).$$

Alguns livros chamada a Função Poder de **Curva de Característica Operacional**.

Tamanho da amostra (estudos balanceados).

Imagine que

- ▶ Hipóteses: $H_0 : \tau_1 = \dots = \tau_a = 0$ e $H_1 : \tau_1^2 + \dots + \tau_a^2 > 0$;
- ▶ H_1 é verdade, então $\tau_1^2 + \dots + \tau_a^2 > 0$. Chamamos $f = \sqrt{\frac{\tau_1^2 + \dots + \tau_a^2}{a\sigma^2}}$ de *size effect*;
- ▶ $F_0 = \frac{QM_{Tratamentos}}{QME} \sim F_{a-1, N-a}(\Phi^2)$, em que $\Phi^2 = n \cdot f^2$ e n é o número de observações em cada tratamento;
- ▶ Ao nível de significância α , temos $RC = \{f_0 \mid f_0 > f_{1-\alpha; a-1, N-a}\}$.

Dado o número de tratamentos a , o poder do teste $1 - \beta$, o nível de significância α e o *size effect* f . Suponha que todos os tratamentos tem o mesmo número de observações n , então o tamanho mínimo da amostra é $N = n \cdot a$ e n é solução de

$$1 - \beta = 1 - P\left(F_{a-1, n \cdot a - a}(n \cdot f^2) \leq f_{1-\alpha; a-1, n \cdot a - a}\right)$$

Poder do teste (estudos balanceados).

Exemplo

Em um experimento que deseja analisar se a resistência à tração de um tecido está associada com a porcentagem da fibra de algodão. Cinco níveis de porcentagem da fibra de algodão e cinco observações em cada nível de porcentagem de fibra de algodão. Na Tabela 7, mostramos as médias populacionais para cada nível de porcentagem de fibra de algodão, e a média e a variância populacionais sem considerar os níveis de porcentagem de fibra de algodão são, respectivamente, $\mu = 15$ e $\sigma^2 = 27$. Se coletarmos $n = 6$ observações para cada nível de porcentagem de fibra de algodão, qual o poder do teste? Use $\alpha = 5\%$.

Algodão	Médias populacionais
15	10
20	15
25	18
30	22
35	10

Tabela 7: Médias para cada nível de porcentagem de fibra de algodão.

Tamanho da amostra (estudos balanceados).

Solução

Passo 1) Queremos testar as hipóteses: $H_0 : \tau_1 = \dots = \tau_5 = 0$ e $H_1 : \tau_1^2 + \dots + \tau_5^2 > 0$;

Passo 2) Nível de significância $\alpha = 5\%$;

Primeiro vamos calcular os efeitos dos tratamentos:

$$\begin{aligned}\tau_1 &= \mu_1 - \mu = 15 - 15 = 0 & \tau_2 &= \mu_2 - \mu = 20 - 15 = 5 \\ \tau_3 &= \mu_3 - \mu = 25 - 15 = 10 & \tau_4 &= \mu_4 - \mu = 30 - 15 = 15 \\ \tau_5 &= \mu_5 - \mu = 35 - 15 = 20\end{aligned}$$

Primeiro vamos calcular o *size effect*: $f^2 = \frac{\tau_1^2 + \dots + \tau_5^2}{a\sigma^2} = \frac{0^2 + 5^2 + 10^2 + 15^2 + 20^2}{5 \cdot 27} = 5,56$. Pelo enunciado, sabemos que cada tratamento tem $n = 6$ observações, então $N = 5 \cdot 6 = 30$ e $a = 5$. O parâmetro de não-centralidade é dado por $\Phi^2 = n \cdot f^2 = 6 \cdot 5,56 = 33,33$. Vamos encontrar o valor crítico:

$$\blacktriangleright P(F_{a-1, N-a} \leq f_{1-\alpha; a-1, N-a}) = P(F_{a-1, N-a} \leq f_{0,95; 4, 25}) = 1 - \alpha = 0,95, \text{ então } f_{0,95; 4, 25} = 2,7587.$$

O poder do teste é dado por

$$\begin{aligned}1 - \beta &= 1 - P(F_{a-1, N-a}(\Phi^2) \leq f_{0,95; 4, 25}) \\ &= 1 - P(F_{4, 25}(33,33) \leq 2,7587) \\ &= 0,9943.\end{aligned}$$

Tamanho da amostra (estudos balanceados).

Exemplo

Em um experimento que deseja analisar se a resistência à tração de um tecido está associada com a porcentagem da fibra de algodão. Cinco níveis de porcentagem da fibra de algodão e cinco observações em cada nível de porcentagem de fibra de algodão. Na Tabela 8, mostramos as médias populacionais para cada nível de porcentagem de fibra de algodão, e a média e a variância populacionais sem considerar os níveis de porcentagem de fibra de algodão são, respectivamente, $\mu = 15$ e $\sigma^2 = 27$. Para termos um poder de teste de $1 - \beta = 99\%$, quantas observações precisam ser coletadas para cada nível de fibra de porcentagem de algodão? Use $\alpha = 5\%$.

Algodão	Médias populacionais
15	10
20	15
25	18
30	22
35	10

Tabela 8: Médias para cada nível de porcentagem de fibra de algodão.

Tamanho da amostra (estudos balanceados).

Solução

Passo 1) Queremos testar as hipóteses: $H_0 : \tau_1 = \dots = \tau_5 = 0$ e

$$H_1 : \tau_1^2 + \dots + \tau_5^2 > 0;$$

Passo 2) Nível de significância $\alpha = 5\%$;

Primeiro vamos calcular os efeitos dos tratamentos:

$$\tau_1 = \mu_1 - \mu = 15 - 15 = 0 \quad \tau_2 = \mu_2 - \mu = 20 - 15 = 5$$

$$\tau_3 = \mu_3 - \mu = 25 - 15 = 10 \quad \tau_4 = \mu_4 - \mu = 30 - 15 = 15$$

$$\tau_5 = \mu_5 - \mu = 35 - 15 = 20$$

Primeiro vamos calcular o *size effect*: $f^2 = \frac{\tau_1^2 + \dots + \tau_5^2}{a\sigma^2} = \frac{0^2 + 5^2 + 10^2 + 15^2 + 20^2}{5 \cdot 27} = 5,56$. O número de tratamentos é $a = 5$.

Então, o tamanho mínimo de amostra para cada tratamento é solução em n da seguinte equação:

$$\begin{aligned} 1 - \beta = 0,99 &= 1 - P\left(F_{a-1, n \cdot a - a}(n \cdot f^2) \leq f_{1-\alpha; a-1, n \cdot a - a}\right) \\ &= 1 - P\left(F_{4, n \cdot 5 - 5}(n \cdot 5,56) \leq f_{0,95; 4, n \cdot 5 - 5}\right) \end{aligned}$$

Então, precisamos de $n = 5$ observações para cada nível de porcentagem de fibra de tecido.

Solução (R)

```
> samp_size_anova <- function(a, es, alpha = 0.05, power = 0.95){
+
+   f <- function(n){
+     n %>%
+       map_dbl(~ (
+         1 - pf(qf(1 - alpha, df1 = a - 1, df2 = .x * a - a),
+               df1 = a - 1,
+               df2 = .x * a - a,
+               ncp = .x * es^2)-
+         power)^2)
+   }
+
+   f(2:5e+4) %>% which.min()
+ }
> a <- 5
> tau <- c(0, 5, 10, 15, 20)
> s2 = 27
> es <- (sum(tau^2) / (a * s2)) %>% sqrt()
> samp_size_anova(a = 5, es = es, alpha = 0.05, power = 0.99)
[1] 5
```

Figura 6: Tamanho da amostra.

Intervalo de confiança para médias

Considere o modelo dado por

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \sigma^2).$$

Neste contexto, a média para cada tratamento é μ_i e pode ser aproximada por $\bar{y}_{i.} = \frac{y_{i1} + y_{i2} + \dots + y_{in}}{n}$. Pode-se provar que

$$T = \frac{\bar{Y}_{i.} - \mu_i}{\sqrt{\frac{QM_E}{n}}} \sim t_{N-a}.$$

em que $\mu_i = \mu + \tau_i, i = 1, \dots, a$ e $N = n \cdot a$. Lembre que $E[QM_E] = \sigma^2$. Dado o coeficiente de confiança $\gamma = 1 - \alpha$, temos que

$$\gamma = P \left(t_{\frac{\alpha}{2}; N-a} \leq \frac{\bar{Y}_{i.} - \mu_i}{\sqrt{\frac{QM_E}{n}}} \leq t_{1-\frac{\alpha}{2}; N-a} \right),$$

e o intervalo de confiança para a média μ_i do tratamento i é dado

$$IC(\mu_i, \gamma) = \left(t_{\frac{\alpha}{2}; N-a} \sqrt{\frac{QM_E}{n}} + \bar{y}_{i.}; t_{1-\frac{\alpha}{2}; N-a} \sqrt{\frac{QM_E}{n}} + \bar{y}_{i.} \right).$$

Intervalo de confiança para média

Exemplo

Em um experimento que deseja analisar se a resistência à tração de um tecido está associada com a porcentagem da fibra de algodão. Cinco níveis de porcentagem da fibra de algodão e cinco observações em cada nível de porcentagem de fibra de algodão. Os dados estão na Tabela 9. Calcule o intervalo de confiança para média de resistência de tração para cada nível de porcentagem de fibra de algodão. Use $\gamma = 95\%$.

Porcentagem de fibra de algodão (tratamento)	Observações					Total	Variância em cada tratamento	Média em cada tratamento
15	7	7	15	11	9	49	11,20	9,80
20	12	17	12	18	18	77	9,80	15,40
25	14	18	18	19	19	88	4,30	17,60
30	19	25	22	19	23	108	6,80	21,60
35	7	10	11	15	11	54	8,20	10,80

Tabela 9: Resistência à tração

Intervalo de confiança para média

Solução

Primeiro vamos calcular a tabela anova. Mostramos o resultado na Tabela 10.

Fonte de variação	Graus de liberdade	Somas de quadrado	Quadrados médios	F_0	valor-p
Porcentagem de fibra de Algodão (Tratamento)	4	475,76	118,94	14,76	0,00
Erro	20	161,20	8,06		
Total	24	636,96			

Tabela 10: Tabela ANOVA.

Note que $\gamma = 0,95 = 1 - \alpha$, $\alpha = 0,05$, $a = 5$ e $N = a \cdot n = 5 \cdot 5 = 25$. Vamos calcular os quantis da distribuição t -Student:

- ▶ $P\left(t_{N-a} \leq t_{\frac{\alpha}{2}, N-a}\right) = P\left(t_{20} \leq t_{0,025;20}\right) = \frac{\alpha}{2} = 0,025$, então $t_{0,025;20} = -2,086$;
- ▶ $P\left(t_{N-a} \leq t_{\frac{\alpha}{2}, N-a}\right) = P\left(t_{20} \leq t_{0,975;20}\right) = 1 - \frac{\alpha}{2} = 0,975$, então $t_{0,975;20} = 2,086$.

Então a média para cada nível de porcentagem de fibra de algodão é dado:

$$\begin{aligned}
 IC(\mu_1, \gamma) &= \left(\bar{y}_1 + t_{\frac{\alpha}{2}; N-a} \sqrt{\frac{QM_E}{n}}; \bar{y}_1 + t_{1-\frac{\alpha}{2}; N-a} \sqrt{\frac{QM_E}{n}} \right) \\
 &= \left(9,80 - 2,086 \sqrt{\frac{8,06}{5}}; 9,80 + 2,086 \sqrt{\frac{8,06}{5}} \right) = (7,15; 12,45).
 \end{aligned}$$

Então, a resistência média à tração para fibras com 15% de fibra de algodão está entre 7,15 e 12,45.

Intervalo de confiança para média

Solução

Então a média para cada nível de porcentagem de fibra de algodão é dado:

$$IC(\mu_2, \gamma) = \left(\bar{y}_{1\cdot} + t_{\frac{\alpha}{2}; N-a} \sqrt{\frac{QM_E}{n}}; \bar{y}_{1\cdot} + t_{1-\frac{\alpha}{2}; N-a} \sqrt{\frac{QM_E}{n}} \right) = \left(15,40 - 2,086 \sqrt{\frac{118,94}{20}}; 15,40 + 2,086 \sqrt{\frac{118,94}{20}} \right) \\ = (12,75; 18,05)$$

$$IC(\mu_3, \gamma) = \left(\bar{y}_{2\cdot} + t_{\frac{\alpha}{2}; N-a} \sqrt{\frac{QM_E}{n}}; \bar{y}_{2\cdot} + t_{1-\frac{\alpha}{2}; N-a} \sqrt{\frac{QM_E}{n}} \right) = \left(17,60 - 2,086 \sqrt{\frac{118,94}{20}}; 17,60 + 2,086 \sqrt{\frac{118,94}{20}} \right) \\ = (14,95; 20,25)$$

$$IC(\mu_4, \gamma) = \left(\bar{y}_{3\cdot} + t_{\frac{\alpha}{2}; N-a} \sqrt{\frac{QM_E}{n}}; \bar{y}_{3\cdot} + t_{1-\frac{\alpha}{2}; N-a} \sqrt{\frac{QM_E}{n}} \right) = \left(21,60 - 2,086 \sqrt{\frac{118,94}{20}}; 21,60 + 2,086 \sqrt{\frac{118,94}{20}} \right) \\ = (18,95; 24,25)$$

$$IC(\mu_5, \gamma) = \left(\bar{y}_{4\cdot} + t_{\frac{\alpha}{2}; N-a} \sqrt{\frac{QM_E}{n}}; \bar{y}_{4\cdot} + t_{1-\frac{\alpha}{2}; N-a} \sqrt{\frac{QM_E}{n}} \right) = \left(10,80 - 2,086 \sqrt{\frac{118,94}{20}}; 10,80 + 2,086 \sqrt{\frac{118,94}{20}} \right) \\ = (8,15; 13,45).$$

Intervalo de confiança para diferenças das médias

Considere o modelo dado por

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \sigma^2).$$

Neste contexto, a média para cada tratamento é μ_i e pode ser aproximada por $\bar{y}_{i.} = \frac{y_{i1} + y_{i2} + \dots + y_{in}}{n}$. Pode-se provar que

$$T = \frac{(\bar{Y}_{i.} - \bar{Y}_{j.}) - (\mu_i - \mu_j)}{\sqrt{\frac{2 \cdot QM_E}{n}}} \sim t_{N-a}.$$

em que $\mu_i = \mu + \tau_i, i = 1, \dots, a$ e $N = n \cdot a$. Lembre que $E[QM_E] = \sigma^2$. Dado o coeficiente de confiança $\gamma = 1 - \alpha$, temos que

$$\gamma = P \left(t_{\frac{\alpha}{2}; N-a} \leq \frac{(\bar{Y}_{i.} - \bar{Y}_{j.}) - (\mu_i - \mu_j)}{\sqrt{\frac{2 \cdot QM_E}{n}}} \leq t_{1-\frac{\alpha}{2}; N-a} \right),$$

e o intervalo de confiança para a média μ_i do tratamento i é dado

$$IC(\mu_i - \mu_j, \gamma) = \left(t_{\frac{\alpha}{2}; N-a} \sqrt{\frac{2 \cdot QM_E}{n}} + (\bar{y}_{i.} - \bar{y}_{j.}); t_{1-\frac{\alpha}{2}; N-a} \sqrt{\frac{2 \cdot QM_E}{n}} + (\bar{y}_{i.} - \bar{y}_{j.}) \right).$$

Intervalo de confiança para diferenças das médias

Exemplo

Em um experimento que deseja analisar se a resistência à tração de um tecido está associada com a porcentagem da fibra de algodão. Cinco níveis de porcentagem da fibra de algodão e cinco observações em cada nível de porcentagem de fibra de algodão. Os dados estão na Tabela 11. Calcule o intervalo de confiança para a diferença das médias de resistência de tração para cada níveis de porcentagens 20% e 25% e para a diferença das médias de resistência de tração para cada níveis de porcentagens 20% e 30% de fibra de algodão. Use $\gamma = 95\%$.

Porcentagem de fibra de algodão (tratamento)	Observações					Total	Variância em cada tratamento	Média em cada tratamento
15	7	7	15	11	9	49	11,20	9,80
20	12	17	12	18	18	77	9,80	15,40
25	14	18	18	19	19	88	4,30	17,60
30	19	25	22	19	23	108	6,80	21,60
35	7	10	11	15	11	54	8,20	10,80

Tabela 11: Resistência à tração

Intervalo de confiança para diferenças das médias

Solução

Primeiro vamos calcular a tabela anova. Mostramos o resultado na Tabela 12.

Fonte de variação	Graus de liberdade	Somas de quadrado	Quadrados médios	F_0	valor-p
Porcentagem de fibra de Algodão (Tratamento)	4	475,76	118,94	14,76	0,00
Erro	20	161,20	8,06		
Total	24	636,96			

Tabela 12: Tabela ANOVA.

Note que $\gamma = 0,95 = 1 - \alpha$, $\alpha = 0,05$, $a = 5$ e $N = a \cdot n = 5 \cdot 5 = 25$. Vamos calcular os quantis da distribuição t -Student:

- ▶ $P\left(t_{N-a} \leq t_{\frac{\alpha}{2}, N-a}\right) = P\left(t_{20} \leq t_{0,025;20}\right) = \frac{\alpha}{2} = 0,025$, então $t_{0,025;20} = -2,086$;
- ▶ $P\left(t_{N-a} \leq t_{\frac{\alpha}{2}, N-a}\right) = P\left(t_{20} \leq t_{0,975;20}\right) = 1 - \frac{\alpha}{2} = 0,975$, então $t_{0,975;20} = 2,086$.

Intervalo de confiança para diferenças das médias

Solução

Então o intervalo para $\mu_2 - \mu_3$ é dado por

$$IC(\mu_2 - \mu_3; \gamma) = \left(t_{\frac{\alpha}{2}; N-a} \sqrt{\frac{QME}{n}} + (\bar{y}_{2\cdot} - \bar{y}_{3\cdot}); t_{1-\frac{\alpha}{2}; N-a} \sqrt{\frac{QME}{n}} + (\bar{y}_{2\cdot} - \bar{y}_{3\cdot}) \right)$$

$$IC(\mu_2 - \mu_3; 95\%) = \left(-2,086 \sqrt{\frac{8,06}{5}} + (15,40 - 17,60); 2,086 \sqrt{\frac{8,06}{5}} + (15,40 - 17,60) \right) = (-4,85; 0,45).$$

Ou seja, como $0 \in IC(\mu_2 - \mu_3; 95\%)$, as resistências médias à tração para os níveis 20% e 25% de fibra de algodão são iguais com coeficiente de confiança $\gamma = 95\%$.

Então o intervalo para $\mu_2 - \mu_3$ é dado por

$$IC(\mu_2 - \mu_4; \gamma) = \left(t_{\frac{\alpha}{2}; N-a} \sqrt{\frac{QME}{n}} + (\bar{y}_{2\cdot} - \bar{y}_{4\cdot}); t_{1-\frac{\alpha}{2}; N-a} \sqrt{\frac{QME}{n}} + (\bar{y}_{2\cdot} - \bar{y}_{4\cdot}) \right)$$

$$IC(\mu_2 - \mu_4; 95\%) = \left(-2,086 \sqrt{\frac{8,06}{5}} + (15,40 - 21,60); 2,086 \sqrt{\frac{8,06}{5}} + (15,40 - 21,60) \right) = (-8,85; -3,55).$$

Ou seja, como $0 \in IC(\mu_2 - \mu_4; 95\%)$, a resistência média à tração do tecido com 20% de fibra de algodão é menor que a resistência média à tração do tecido com 30% de fibra de algodão.

- ▶ Suponha que queremos testar $H_0 : \mu_1 = \cdots = \mu_a$;
- ▶ Chamamos μ_1, \dots, μ_a de tratamentos;
- ▶ Imagine que temos uma amostra de tamanho n_i para cada tratamento:
 $y_{i1}, \dots, y_{in_i}, i = 1, \dots, a$,
- ▶ Descrevemos os dados através de

$$y_{ij} = \mu + \tau_i + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \sigma^2), \quad j = 1, \dots, n_i, i = 1, \dots, a; \quad (3)$$

- ▶ Queremos testar as seguintes hipóteses

$$H_0 : \tau_1 = \tau_2 = \cdots = \tau_a = 0,$$

$$H_1 : \tau_1^2 + \cdots + \tau_a^2 > 0.$$

Chamamos $\tau_i, i = 1, \dots, a$ de efeitos do tratamentos.

Soma dos quadrados.

Se $H_0 : \tau_1 = \tau_2 = \dots = \tau_a = 0$ é verdadeira, então $Y_{ij} \sim N(\mu, \sigma^2), j = 1, \dots, n_i, i = 1, \dots, a$.

$$\begin{aligned}
 SS_T &= \underbrace{\sum_{i=1}^a \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2}_{\text{Soma dos quadrados total}} = (N - 1)s^2 \\
 &= \underbrace{\sum_{i=1}^a n_i (\bar{y}_{i.} - \bar{y}_{..})^2}_{\text{Soma dos Quadrados dos Tratamentos}} + \underbrace{\sum_{i=1}^a \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2}_{\text{Soma dos Quadrados do Erro}} \\
 &= \underbrace{\sum_{i=1}^a n_i (\bar{y}_{i.} - \bar{y}_{..})^2}_{SQ_{Tratamentos}} + \underbrace{\sum_{i=1}^a (n_i - 1) s_i^2}_{SQ_E} = SQ_{Tratamentos} + SQ_E.
 \end{aligned}$$

em que

- ▶ $s^2 = \frac{1}{N-1} \sum_{i=1}^a \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2$, em que $N = \sum_{i=1}^a n_i$ (variância total);
- ▶ $s_i^2 = \frac{1}{n_i-1} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2$ (variância do tratamento i).

Quadrado médio.

Usando a equação (3), podemos provar que

- ▶ $E(SQ_{Tratamentos}) = (a-1)\sigma^2 + n_1\tau_1^2 + \dots + n_a\tau_a^2;$
- ▶ $E(SQ_E) = (N-a)\sigma^2;$
- ▶ Quadrado Médio dos Tratamentos: $QM_{Tratamentos} = \frac{SQ_{Tratamentos}}{a-1};$
- ▶ Quadrado Médio do Erro: $QM_E = \frac{SQ_E}{N-a};$
- ▶ $F_0 = \frac{QM_{Tratamentos}}{QM_E}.$

em que $N = \sum_{i=1}^a n_i.$

É possível provar que

- ▶ $E(QM_{Tratamentos}) = \sigma^2 + \frac{n_1\tau_1^2 + \dots + n_a\tau_a^2}{a-1};$
- ▶ $E(QM_E) = \sigma^2;$

Rejeitamos $H_0 : \tau_1 = \tau_2 = \dots = \tau_a = 0$ se $F_0 = \frac{QM_{Tratamentos}}{QM_E}$ for grande.

Se H_0 for verdadeiro, temos que $F_0 \sim F_{a-1, N-a}$, em que $F_{a-1, N-a}$ é a distribuição F .
Teste ANOVA é robusto para tratamentos com variâncias ligeiramente diferentes.

Rejeitamos $H_0: \tau_1 = \tau_2 = \dots = \tau_a = 0$ se $f_0 > f_{1-\alpha; a-1, N-a}$

$$f_0 = \frac{QM_{\text{Tratamentos}}}{QM_E} \text{ e } N = n_1 + \dots + n_a$$

Figura 7: Região crítica para ANOVA.

ANOVA: Checando as suposições do modelo.

Análise de resíduo

Depois decidir entre H_0 e H_1 , precisamos checar as suposições do modelo:

- (a) $\epsilon_{ij}, j = 1, \dots, n_i, i = 1, \dots, a$ são independentes;
- (b) $\epsilon_{ij}, j = 1, \dots, n_i, i = 1, \dots, a$ tem distribuição normal;
- (c) $\epsilon_{ij}, j = 1, \dots, n_i, i = 1, \dots, a$ tem variância constante.

$\epsilon_{ij} = y_{ij} - \mu + \tau_i = y_{ij} - \mu + \tau_i, j = 1, \dots, n_i, i = 1, \dots, a$ não são observáveis pois não conhecemos as médias populacionais dentro de cada tratamento, mas podemos aproximar estas médias populacionais μ_i por $\bar{y}_{i.}, i = 1, \dots, a$. Assim, podemos definimos os resíduos por

$$e_{ij} = y_{ij} - \bar{y}_{i.}, j = 1, \dots, n_i, i = 1, \dots, a. \quad (4)$$

Usando a equação (4), podemos usar gráficos para checar (a), (b) e (c):

- (a) **Independência:** cada par $(i + j, e_{ij}), j = 1, \dots, n_i, i = 1, \dots, a$, é representado por um ponto no plano cartesiano. Se não existe padrão ou tendência, então assumimos que ϵ_{ij} são independentes;
- (b) **Normalidade:** cada par $\left(q_{(i+j)}; \frac{e_{(i+j)} - \bar{e}_{..}}{s_e} \right), j = 1, \dots, n_i, i = 1, \dots, a$ é representado por um ponto no plano cartesiano, em que $\Phi(q_{(i+j)}) = \frac{i+j-0,5}{n}$ e s_e é o desvio padrão dos resíduos – este gráfico é chamado de gráfico de probabilidade normal. Se os pontos estão próximos ou em cima da reta com ângulo 45 graus, então ϵ_{ij} tem distribuição normal;
- (c) **Variância constante:** cada par $(\bar{y}_{i.}, e_{ij}), j = 1, \dots, n_i, i = 1, \dots, a$, é representando por um ponto no plano cartesiano. Se não existe padrão ou tendência, então assumimos que ϵ_{ij} tem variância constante.

ANOVA

Exemplo

Um experimento foi realizado para determinar se quatro temperaturas de queima específicas afetam a densidade de um certo tipo de tijolo. Os dados estão na Tabela 13.

Temperatura	Densidade						
40	21,8	21,9	21,7	21,6	21,7	21,5	21,8
50	21,7	21,4	21,5	21,5			
60	21,9	21,8	21,8	21,6	21,5		
70	21,9	21,7	21,8	21,7	21,6	21,8	

Tabela 13: Informações do experimento.

- (a) As temperaturas de queima específicas afetam a densidade de um certo tipo de tijolo? Use $\alpha = 5\%$.
- (b) Calcule o valor-p.
- (c) Faça uma análise de resíduo.

ANOVA

Solução

Passo 1) Queremos testar as hipóteses: $H_0 : \tau_1 = \tau_2 = \tau_3 = \tau_4 = 0$ e

$$H_1 : \tau_1^2 + \tau_2^2 + \tau_3^2 + \tau_4^2 > 0;$$

Passo 2) Nível de significância $\alpha = 5\%$;

Passo 3) Rejeitamos H_0 se $F_0 = \frac{QM_{Tratamentos}}{QM_E}$. Ou seja, $RC = \{f_0 \mid f_0 > f_{1-\alpha; a-1, N-a}\}$;

Passo 4) Note que $n_1 + n_2 + n_3 + n_4 = 22$ e $a = 4$. Vamos encontrar o valor crítico:

$$\begin{aligned} \blacktriangleright P(F_{a-1, N-a} \leq f_{1-\alpha; a-1, N-a}) &= P(F_{3, 18} \leq f_{0,95;3,18}) = 1 - \alpha = 0,95, \text{ então} \\ f_{0,95;3,18} &= 3,1599; \end{aligned}$$

Passo 5) Usando os dados da Tabela 13, obtemos a Tabela ANOVA (mostrada na Tabela 14).

Fatores de variação	Graus de liberdade	Soma de quadrados	Quadrados médios	F_0
Temperatura °C	3	$SQ_{Tratamento} = 0,139$	$QM_{Tratamento} = 0,046$	$\frac{QM_{Tratamentos}}{QM_E} = 2,556$
Erro	18	$SQ_E = 0,319$	$QM_E = 0,018$	
Total	21	$SQ_T = 0,458$		

Tabela 14: Tabela Anova

Lembre que $SQ_E = (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + (n_3 - 1)s_3^2 + (n_4 - 1)s_4^2$, $SQ_T = (N - 1)s^2$ e $N = n_1 + n_2 + n_3 + n_4$.

Como $f_0 = 2,556 \leq 3,1599 = f_{0,95;3,23}$, não rejeitamos H_0 . Ao nível de significância $\alpha = 5\%$, não temos evidência estatística para afirmar que as densidades médias para temperaturas de queima específica são diferentes.

ANOVA

Solução (valor-p)

O valor-p é calculado através de

$$p = P(F_0 | f_0 | H_0) = 1 - P(F_{a-1, N-a} \leq f_0),$$

em que $N = n_1 + \dots + n_a$.

Usando a Tabela 14, temos que $a = 4$, $N = 22$ e $f_0 = 2,556$. Então, temos que

$$\begin{aligned} p &= 1 - P(F_{a-1, N-a} \leq f_0) \\ &= 1 - P(F_{3, 23} \leq 2,556) \\ &= 1 - 0,9199 \\ &= 0,0801. \end{aligned}$$

Como $p = 0,0801 \geq \alpha = 0,05$, não rejeitamos H_0 . Ou seja, não temos evidência estatística para afirmar que as densidades médias são diferentes para as quatro temperaturas de queima específicas.

ANOVA

Solução (Análise de resíduo)

Primeiro calculamos os resíduos e a média dentro de cada tratamento, conforme mostrado na Tabela 15.

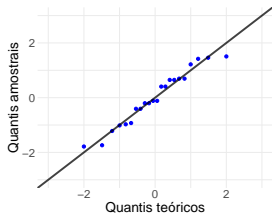
Temperatura	Resíduo							\bar{y}_i	s_i^2
40	0,0857	0,1857	-0,0143	-0,1143	-0,0143	-0,2143	0,0857	21,7143	0,0181
50	0,175	-0,125	-0,025	-0,025				21,5250	0,0158
60	0,18	0,08	0,08	-0,12	-0,22			21,7200	0,0270
70	0,15	-0,05	0,05	-0,05	-0,15	0,05		21,7500	0,0110

Tabela 15: Informações do experimento.

- (a) Na Figura 8a, notamos que os pontos estão próximos da reta e concluímos que ϵ_{ij} tem distribuição normal;
- (b) Na Figura 8b, não notamos qualquer padrão ou tendência e concluímos que as variáveis aleatórias ϵ_{ij} são independentes;
- (c) Na Figura 8c, não notamos qualquer padrão ou tendência e concluímos que a variância é constante para as variáveis aleatórias ϵ_{ij} .

ANOVA

Análise de resíduo



(a) Gráfico de probabilidade normal.

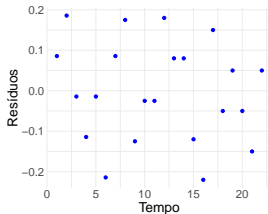
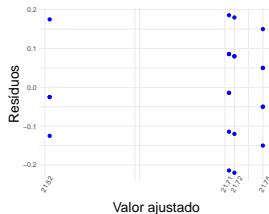
(b) Tempo \times Resíduos.(c) $\bar{y}_{j\cdot} \times$ Resíduos.

Figura 8: Análise de resíduos.

Poder do teste (estudos não-balanceados).

No contexto do modelo da equação (3), imagine que

- ▶ Hipóteses: $H_0 : \tau_1 = \dots = \tau_a = 0$ e $H_1 : \tau_1^2 + \dots + \tau_a^2 > 0$;
- ▶ H_1 é verdade, então $\tau_1^2 + \dots + \tau_a^2 > 0$;
- ▶ $F_0 = \frac{QM_{Tratamentos}}{QME} \sim F_{a-1, N-a}(\Phi^2)$, em que $\Phi^2 = \frac{n_1\tau_1^2 + \dots + n_a\tau_a^2}{a\sigma^2}$, em que n_i é o número de observações do Tratamento i , $i = 1, \dots, a$;
- ▶ Ao nível de significância α , temos $RC = \{f_0 \mid f_0 > f_{1-\alpha; a-1, N-a}\}$.

Poder do teste é dado por

$$1 - \beta = 1 - P(F_0 \leq f_{1-\alpha; a-1, N-a} \mid H_0) = 1 - P(F_{a-1, N-a}(\Phi^2) \leq f_{1-\alpha; a-1, N-a})$$

em que $N = n_1 + \dots + n_a$. A **Função Poder**, dado o tamanho da amostra N , o número de tratamentos a e os tamanhos dos tratamentos são n , é $\pi : (0, \infty) \rightarrow [0, 1]$ dada por

$$\pi(\Phi^2) = 1 - P(F_{a-1, N-a}(\Phi^2) \leq f_{1-\alpha; a-1, N-a}), \quad \Phi^2 \in (0, \infty).$$

Alguns livros chamada a Função Poder de **Curva de Característica Operacional**.

Poder do teste (estudos não-balanceados).

Exemplo

Um experimento foi realizado para determinar se quatro temperaturas de queima específicas afetam a densidade de um certo tipo de tijolo. Algumas informações estão na Tabela 16. Qual o poder do teste? Use $\alpha = 5\%$.

Temperatura	n_i	μ_i	σ_i
40	4	21	2
50	5	14	2
60	6	15	2
70	4	25	2
		$\mu = 21$	$\sigma^2 = 2$

Tabela 16: Informações do experimento.

Note que $\mu_i = \mu + \tau_i$ é a média de cada tratamento, σ_i^2 é a variância de cada tratamento, μ é a média total e σ^2 é a variância total.

Poder do teste (estudos não-balanceados).

Solução

Primeiro vamos calcular os efeitos dentro de cada tratamento:

$$\tau_1 = \mu_j - \mu = 21 - 21 = 0$$

$$\tau_2 = \mu_2 - \mu = 14 - 21 = -7$$

$$\tau_3 = \mu_j - \mu = 15 - 21 = -6$$

$$\tau_4 = \mu_4 - \mu = 25 - 21 = 4$$

Então, o parâmetro de não-centralidade é dada por

$$\Phi^2 = \frac{n_1 \tau_1^2 + \dots + n_a \tau_a^2}{a \sigma^2} = \frac{4 \cdot 0^2 + 5 \cdot (-7)^2 + 6 \cdot (-6)^2 + 4 \cdot 4^2}{4 \cdot 2} = 65,625.$$

Vamos encontrar o quantil da distribuição F :

$$\blacktriangleright P(F_{a-1, N-a} \leq f_{1-\alpha; a-1, N-a}) = P(F_{3, 15} \leq f_{0,95; 3, 15}) = 1 - \alpha = 0,95, \text{ então } f_{0,95; 3, 15} = 3,2874.$$

Então o poder do teste é dado por

$$1 - \beta = 1 - P(F_{a-1, N-a}(\Phi^2) \leq f_{1-\alpha; a-1, N-a}) = 1 - P(F_{3, 15}(65,625) \leq 3,2874) = 1.$$

Intervalo de confiança para médias

Considere o modelo dado por

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \sigma^2), \quad j = 1, \dots, n_i, i = 1, \dots, a.$$

Neste contexto, a média para cada tratamento é μ_i e pode ser aproximada por

$$\bar{y}_{i.} = \frac{y_{i1} + y_{i2} + \dots + y_{in_i}}{n_i}. \text{ Pode-se provar que}$$

$$T = \frac{\bar{Y}_{i.} - \mu_i}{\sqrt{\frac{QM_E}{n_i}}} \sim t_{N-a}.$$

em que $\mu_i = \mu + \tau_i, i = 1, \dots, a$ e $N = n_1 + \dots + n_a$. Lembre que $E[QM_E] = \sigma^2$. Dado o coeficiente de confiança $\gamma = 1 - \alpha$, temos que

$$\gamma = P \left(t_{\frac{\alpha}{2}; N-a} \leq \frac{\bar{Y}_{i.} - \mu_i}{\sqrt{\frac{QM_E}{n_i}}} \leq t_{1-\frac{\alpha}{2}; N-a} \right),$$

e o intervalo de confiança para a média μ_i do tratamento i é dado

$$IC(\mu_i, \gamma) = \left(t_{\frac{\alpha}{2}; N-a} \sqrt{\frac{QM_E}{n_i}} + \bar{y}_{i.}; t_{1-\frac{\alpha}{2}; N-a} \sqrt{\frac{QM_E}{n_i}} + \bar{y}_{i.} \right).$$

Intervalo de confiança para as médias.

Exemplo

Um experimento foi realizado para determinar se quatro temperaturas de queima específicas afetam a densidade de um certo tipo de tijolo. Os dados estão na Tabela 17. Construa um intervalo de confiança para a densidade média do tijolo para temperatura de queima igual a 40. Use $\gamma = 99\%$.

Temperatura	Densidade							$\bar{y}_{i.}$	s_i^2
40	21,8	21,9	21,7	21,6	21,7	21,5	21,8	21,7143	0,0181
50	21,7	21,4	21,5	21,5				21,5250	0,0158
60	21,9	21,8	21,8	21,6	21,5			21,7200	0,0270
70	21,9	21,7	21,8	21,7	21,6	21,8		21,7500	0,0110
								$\bar{y}_{..} = 21,6909$	$s^2 = 0,0218$

Tabela 17: Informações do experimento.

Intervalo de confiança para as médias.

Solução

Primeiro vamos calcular a tabela anova. Mostramos o resultado na Tabela 18.

Fatores de variação	Graus de liberdade	Soma de quadrados	Quadrados médios	F_0
Temperatura °C	3	$SQ_{Tratamento} = 0,139$	$QM_{Tratamento} = 0,046$	$\frac{QM_{Tratamentos}}{QM_E} = 2,556$
Erro	18	$SQ_E = 0,319$	$QM_E = 0,018$	
Total	21	$SQ_T = 0,458$		

Tabela 18: Tabela ANOVA.

Note que $\gamma = 0,99 = 1 - \alpha$, $\alpha = 0,05$, $a = 4$, $N - a = 18$ e $N = 22$. Vamos calcular os quantis da distribuição t -Student:

- ▶ $P(t_{N-a} \leq t_{\frac{\alpha}{2}, N-a}) = P(t_{18} \leq t_{0,005;18}) = \frac{\alpha}{2} = 0,025$, então $t_{0,005;18} = -2,878$;
- ▶ $P(t_{N-a} \leq t_{\frac{\alpha}{2}, N-a}) = P(t_{18} \leq t_{0,995;18}) = 1 - \frac{\alpha}{2} = 0,995$, então $t_{0,995;20} = 2,878$.

Então o intervalo de confiança é dado por

$$IC(\mu_1, \gamma) = \left(\bar{y}_{1\cdot} + t_{\frac{\alpha}{2}, N-a} \sqrt{\frac{QM_E}{n_1}}; \bar{y}_{1\cdot} + t_{1-\frac{\alpha}{2}, N-a} \sqrt{\frac{QM_E}{n_1}} \right) = \left(21,7143 - 2,878 \sqrt{\frac{0,018}{7}}; 21,7143 + 2,878 \sqrt{\frac{0,018}{7}} \right) \\ = (21,57; 21,86).$$

Então, a densidade média população dos tijolos com temperatura de queima igual a 40 está entre 21,57 e 21,86.

Intervalo de confiança para diferenças das médias

Considere o modelo dado por

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \sigma^2), \quad j = 1, \dots, n_i, i = 1, \dots, a.$$

Neste contexto, a média para cada tratamento é μ_i e pode ser aproximada por

$$\bar{y}_{i.} = \frac{y_{i1} + y_{i2} + \dots + y_{in_i}}{n_i}. \text{ Pode-se provar que}$$

$$T = \frac{(\bar{Y}_{i.} - \bar{Y}_{j.}) - (\mu_i - \mu_j)}{\sqrt{QM_E \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}} \sim t_{N-a}.$$

em que $\mu_i = \mu + \tau_i, i = 1, \dots, a$ e $N = n_1 + \dots + n_a$. Lembre que $E[QM_E] = \sigma^2$.
Dado o coeficiente de confiança $\gamma = 1 - \alpha$, temos que

$$\gamma = P \left(t_{\frac{\alpha}{2}; N-a} \leq \frac{(\bar{Y}_{i.} - \bar{Y}_{j.}) - (\mu_i - \mu_j)}{\sqrt{QM_E \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}} \leq t_{1-\frac{\alpha}{2}; N-a} \right),$$

e o intervalo de confiança para a média μ_i do tratamento i é dado

$$IC(\mu_i - \mu_j, \gamma) = \left(t_{\frac{\alpha}{2}; N-a} \sqrt{QM_E \left(\frac{1}{n_i} + \frac{1}{n_j} \right)} + (\bar{y}_{i.} - \bar{y}_{j.}); t_{1-\frac{\alpha}{2}; N-a} \sqrt{QM_E \left(\frac{1}{n_i} + \frac{1}{n_j} \right)} + (\bar{y}_{i.} - \bar{y}_{j.}) \right).$$

Intervalo de confiança para diferenças das médias

Exemplo

Um experimento foi realizado para determinar se quatro temperaturas de queima específicas afetam a densidade de um certo tipo de tijolo. Os dados estão na Tabela 19. Construa um intervalo de confiança para a diferença das densidades médias do tijolo para temperaturas de queima igual a 40 e 60. Use $\gamma = 99\%$.

Temperatura	Densidade							$\bar{y}_{i.}$	s_i^2
40	21,8	21,9	21,7	21,6	21,7	21,5	21,8	21,7143	0,0181
50	21,7	21,4	21,5	21,5				21,5250	0,0158
60	21,9	21,8	21,8	21,6	21,5			21,7200	0,0270
70	21,9	21,7	21,8	21,7	21,6	21,8		21,7500	0,0110
								$\bar{y}_{..} = 21,6909$	$s^2 = 0,0218$

Tabela 19: Informações do experimento.

Intervalo de confiança para diferenças das médias

Solução

Primeiro vamos calcular a tabela anova. Mostramos o resultado na Tabela 20.

Fatores de variação	Graus de liberdade	Soma de quadrados	Quadrados médios	F_0
Temperatura °C	3	$SQ_{Tratamento} = 0,139$	$QM_{Tratamento} = 0,046$	$\frac{QM_{Tratamentos}}{QM_E} = 2,556$
Erro	18	$SQ_E = 0,319$	$QM_E = 0,018$	
Total	21	$SQ_T = 0,458$		

Tabela 20: Tabela ANOVA.

Note que $\gamma = 0,99 = 1 - \alpha$, $\alpha = 0,05$, $a = 4$, $N - a = 18$ e $N = 22$. Vamos calcular os quantis da distribuição t -Student:

$$\blacktriangleright P\left(t_{N-a} \leq t_{\frac{\alpha}{2}, N-a}\right) = P\left(t_{18} \leq t_{0,005;18}\right) = \frac{\alpha}{2} = 0,025, \text{ então } t_{0,005;18} = -2,878;$$

$$\blacktriangleright P\left(t_{N-a} \leq t_{\frac{\alpha}{2}, N-a}\right) = P\left(t_{18} \leq t_{0,995;18}\right) = 1 - \frac{\alpha}{2} = 0,995, \text{ então } t_{0,995;20} = 2,878.$$

Então o intervalo de confiança é dado por:

$$\begin{aligned} IC(\mu_1 - \mu_3, \gamma) &= \left((\bar{y}_1 - \bar{y}_3) + t_{\frac{\alpha}{2}, N-a} \sqrt{QM_E \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}; (\bar{y}_1 - \bar{y}_3) + t_{1-\frac{\alpha}{2}, N-a} \sqrt{QM_E \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} \right) \\ &= \left((21,7143 - 21,72) - 2,878 \sqrt{0,018 \left(\frac{1}{7} + \frac{1}{5} \right)}; (21,7143 - 21,72) + 2,878 \sqrt{0,018 \left(\frac{1}{7} + \frac{1}{5} \right)} \right) = (-0,23; 0,22). \end{aligned}$$

Então, como $0 \in IC(\mu_1 - \mu_3, 99\%)$, as densidades médias dos tijolos para temperaturas de queima iguais a 40 e 60 são iguais.