Análise de Variância (um fator)

tradução do Jay Davore e do livro de Wilton de Oliveira Bussab

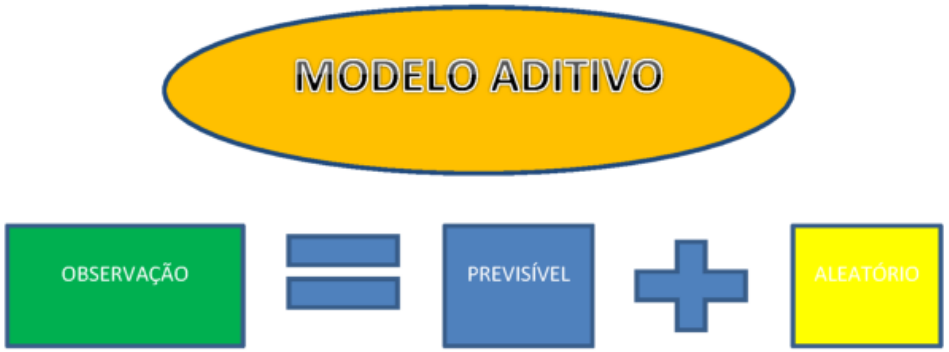
índice

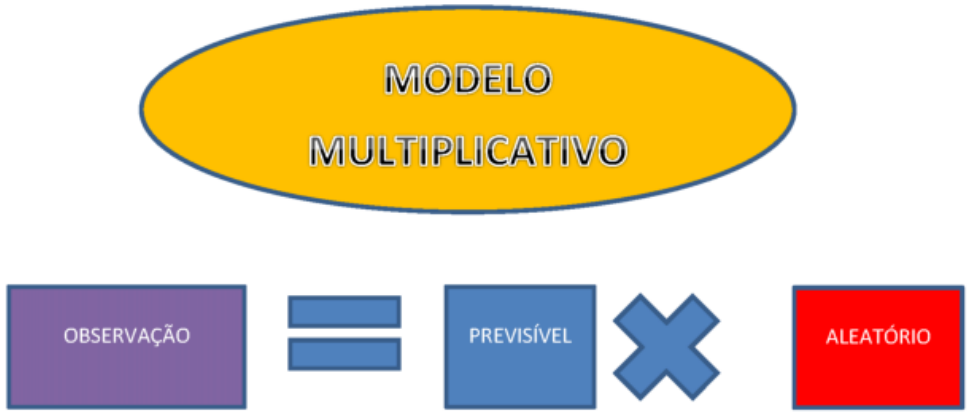
Primeiras idéias

Primeira preocupação ao analisar dados:

Criar Modelos que explicitem estruturas do fenômeno em observação que frequentemente estão misturadas com variações aleatórias

Primeiras Suposições





Primeiras suposições

a parte previsível incorpora o conhecimento que o pesquisador tem sobre o fenômeno, expressada em forma de funções matemáticas
para a parte aleatória impõe-se que os mesmos obedeçam algum modelo de probabilidade

Exemplo

Um psicólogo está investigando a relação entre o tempo que o indivíduo leva para reagir a certo estímulo e algumas de suas características tais como: sexo, idade e acuidade visual (em %). O resultado para 20 indivíduos é apresentado, onde

$i =$ indivíduo

$y_i =$ tempo de reação

$w_i =$ sexo

$x_i =$ idade

$z_i =$ acuidade visual

Exemplo

i	y_i	w_i	x_i	z_i	i	y_i	w_i	x_i	z_i
1	96	H	20	90	11	109	H	30	90
2	92	M	20	100	12	100	H	30	80
3	106	H	20	80	13	112	H	35	90
4	100	M	20	90	14	105	H	35	80
5	98	M	25	100	15	118	H	35	70
6	104	H	25	90	16	108	H	35	90
7	110	H	25	80	17	113	H	40	90
8	101	M	25	90	18	112	H	40	90
9	116	M	30	70	19	127	H	40	60
10	106	H	30	90	20	117	H	40	80

Exemplo

O Modelo e seus estimadores

Admitamos que nenhuma das características tenha influência sistemática sobre o tempo de reação e esses fatores somariam-se a outros não controlados, agindo de maneira aleatória.
Propomos o modelo:
Cada observação (tempo de reação) pode ser decomposto na soma de dois fatores: Um fixo (comum a todas as observações) e outro aleatório (não controlado). Simbólicamente:

$$y_i = \mu + \epsilon_i$$

onde
y_i= tempo de reação da i-ésima observação
μ= efeito fixo
ϵ_i= erro ou efeito residual

Exemplo

ϵ_i pode ser considerado como o efeito resultante de várias características não explicitadas no modelo.

$$\epsilon_i = f(sexo,\ idade,\ acuidade\ visual,\ etc)$$

Para uma melhor interpretação, impomos algumas condições para o modelo:

$$E(\epsilon) = 0 \qquad e \qquad Var(\epsilon) = \sigma^2_\epsilon$$

Exemplo

Conhecendo μ e σ^2_ϵ, teremos idéia do comportamento dos indivíduos, dado que com as condições introduzidas, o tempo médio de reação é μ:

$$E(y_i) = E(\mu) + E(\epsilon_i) = \mu + 0 = \mu$$

$$Var(y_i) = Var(\mu) + Var(\epsilon_i) = 0 + \sigma^2_\epsilon = \sigma^2_\epsilon$$

Exemplo

Estamos propondo um modelo para todos os indivíduos, e não apenas para a amostra (de 20). Precisamos estimar os parâmetros μ e σ^2_ϵ a partir dos dados da amostra.

Usaremos o Princípio de Mínimos Quadrados (MQ)

Temos que:

$$\epsilon_i = y_i - \mu$$

para cada valor de μ, teremos um resíduos diferente. O **melhor** valor de μ, será o que produz os menores resíduos para as 20 observações da amostra. Como os resíduos são positivos e negativos, definos a soma dos quadrados dos resíduos:

$$SQ(\mu) = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \mu)^2$$

Exemplo

O **melhor** valor de μ será o que minimize $SQ(\mu)$. Observe que para uma amostra observada, os valores y_i são constantes, portanto, para encontrar $\hat{\mu}$, basta derivar e igualar a zero a expressão acima.

$$SQ^{`}(\mu) = \sum_{i=1}^{n} 2(y_i - \mu)(-1) = 0$$

de onde

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} y_i = \bar{y}$$

Exemplo

Substituindo em $SQ(\hat{\mu})$, temos:

$$SQ(\hat{\mu}) = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

assim o melhor estimador de σ_ϵ^2, será

$$\hat{\sigma}_\epsilon^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{SQ(\hat{\mu})}{n-1} = s_\epsilon^2$$

Exemplo

Consideramos $SQ(\hat{\mu})$ como a quantidade de informação quadrática perdida pela adoção do modelo.
$\hat{\sigma}_\epsilon^2$ é a quantidade média de informação perdida

Exemplo

Do exemplo: $\sum_{i=1}^{2} 0y_i = 2150$, $\sum_{i=1}^{2} 0y_i^2 = 232498$, n=20, portanto:

$$\hat{\mu} = \bar{y} = \frac{2150}{20} = 107,50$$

$$\begin{aligned} \hat{\sigma}_\epsilon^2 = \quad s_\epsilon^2 = \tfrac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \quad \frac{1}{n-1}\{\sum_{i=1}^{n}(y_i^2) - \frac{(\sum_{i=1}^{n}(y_i))^2}{n}\} = \\ = \quad \tfrac{1}{19}\{232498 - \tfrac{(2150)^2}{20}\} = \quad 72,26 \\ \hat{\sigma}_\epsilon = \quad 8,5 \end{aligned}$$

Suposições Necessárias para Inferência

Pelo método MQ, não foi necessário fazer suposições sobre a distribuição de probabilidade, entretanto, para usar os dados para fazer inferência sobre a população serão necessárias algumas suposições adicionais.

1. o erro $\epsilon_i \sim N(0, \sigma_\epsilon^2)$ para todo i
2. $E(\epsilon_i * \epsilon_j) = 0$ para todo $i \neq j$

das suposições podemos afirmar que

1. $\bar{y} \sim N(\mu, \frac{\sigma_\epsilon^2}{n})$
2. s_ϵ^2 é um estimador não viesado de σ_ϵ^2
3. $(\bar{y} - \mu)\sqrt{n}/s_\epsilon \sim t_{(n-1)}$

Intervalo de Confiança

Construimos um intervalo de confiança para μ com 95% de confiança:

$$IC(\mu, 95\%) = \bar{y} \pm t * \frac{s_\epsilon}{\sqrt{n}} = 107,5 \pm (2,093)\frac{8,5}{\sqrt{20}} =]103,52; 111,48[$$

De um modo Geral,

$$IC(\mu, 1-\alpha) = \bar{y} \pm t_{\frac{\alpha}{2}} * \frac{s_\epsilon}{\sqrt{n}}$$

Estimativas para um novo indivíduo

Se quisermos estimar um novo indivíduo na população, teríamos:

$$y_k = \mu + \epsilon_k$$

estimado por

$$\hat{y}_k = \bar{y} + \epsilon_k$$

, como não podemos estimar ϵ_k o substituimos pela média (que é zero), portanto, temos:

$$\hat{y}_k = \bar{y}$$

com

$$var(\hat{y}_k) = \quad var(\bar{y}) + var(\epsilon_k) = \quad \frac{\sigma_\epsilon^2}{n} + \sigma_\epsilon^2 = \sigma_\epsilon^2(1 + \frac{1}{n})$$

que será estimada por

$$var(\hat{y}_k) = s_\epsilon^2(1 + \frac{1}{n})$$

Intervalo de Confiança para um novo indivíduo

para uma nova observação o IC será

$$IC(y, 1-\alpha) = y \pm t * s_\epsilon \sqrt{1 + \frac{1}{n}}$$

que no exemplo será

$$IC(\hat{y}, 95\%) = 107,5 \pm (2,093) * 8,5 * \sqrt{1 + \frac{1}{20}} =]89,27; 125,73[$$

Análise de residuos

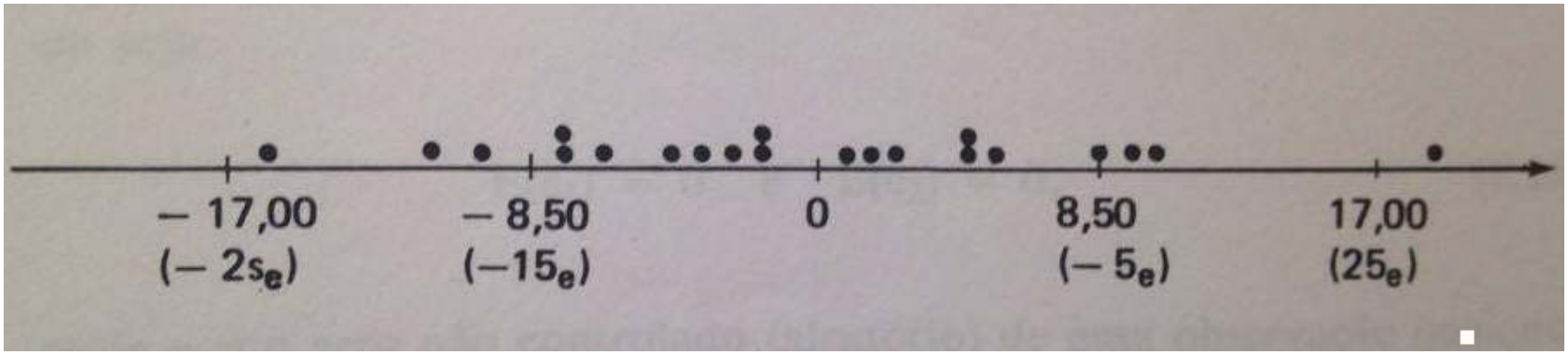


Figura: resíduos do modelo $y_i = \bar{y} + \epsilon_i$

O Modelo

Suponha que desejamos estudar o **Efeito do fator sexo**, isto equivale a retirar esse fator do erro residual. Se esse fator for importante para prever o tempo de reação, o erro residual deverá diminuir drasticamente.

$$y_{ij} = \mu_i + \epsilon_{ij}$$

onde
μ_i= efeito comum a todos os elementos do grupo i ($i = 1$ (homem) e $i = 2$ (mulher))
ϵ_{ij} = efeito aleatório do j-ésimo indivíduo do grupo i
y_{ij} = tempo de reação do j-ésimo indivíduo do grupo i

O Modelo- Suposições

O objetivo consiste em estimar μ_1 e μ_2 e verificar se são diferentes. As primeiras restrições são:

$$E(\epsilon_1) = 0 \qquad e \qquad E(\epsilon_2) = 0$$

onde ϵ_i é o erro de uma observação do grupo i.

$$var(\epsilon_1) = \sigma_1^2 \qquad e \qquad var(\epsilon_2) = \sigma_2^2$$

Os estimadores de μ_1 e μ_2 de MQ, serão aqueles que produzam o mínimo valor para:

$$\begin{aligned} SQ(\mu_1, \mu_2) = & \sum_{i=1}^{2} \sum_{j=1}^{n_i} \epsilon_{ij}^2 = \sum_{i=1}^{2} \sum_{j=1}^{n_i} (y_{ij} - \mu_i)^2 = \\ & = \sum_{j=1}^{n_1} (y_{1j} - \mu_1)^2 + \sum_{j=1}^{n_2} (y_{2j} - \mu_2)^2 \\ & = \sum_{j=1}^{n_1} \epsilon_{1j}^2 + \sum_{j=1}^{n_2} \epsilon_{2j}^2 \end{aligned}$$

Suposições do Modelo

Se as variâncias residuais não forem iguais, a soma será afetada pelo grupo que possui maior variância, e isso deveria influencias a escolha dos estimadores. A sugestão para esse caso é usar os resíduos padronizados:

$$\sum_{j=1}^{n_1}(\frac{\epsilon_{1j}}{\sigma_1})^2 + \sum_{j=1}^{n_2}(\frac{\epsilon_{2j}}{\sigma_2})^2$$

que é chamado de Mínimos quadrados ponderados.
Nesta seção trabalhamos com a seguinte restrição:

$$Var(\epsilon_1) = var(\epsilon_2)$$

$$\sigma_1^2 = \sigma_2^2 = \sigma_\epsilon^2$$

Esta propriedade é conhecida como **Homocedasticidade**

Suposições do Modelo

Se derivarmos a equação acima em relação a μ_1 (depois em relação a μ_2) e igualando a zero, temos:

$$\frac{\partial SQ(\mu_1, \mu_2)}{\partial \mu_1} = 0 \Rightarrow -2 \sum_{j=1}^{n_1} (y_{ij} - \hat{\mu}_1) = 0$$

de onde

$$\hat{\mu}_1 = \frac{1}{n_1} \sum_{j=1}^{n_1} y_{1j} = \bar{y}_1$$

análogamente:

$$\hat{\mu}_2 = \frac{1}{n_2} \sum_{j=1}^{n_2} y_{2j} = \bar{y}_2$$

Suposições do Modelo

A quantidade total de informação perdida será

$$SQ(\hat{\mu}_1, \hat{\mu}_2) = \sum_{i=1}^{2}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2 = \sum_{j=1}^{n_1}(y_{1j} - \bar{y}_1)^2 + \sum_{j=1}^{n_2}(y_{2j} - \bar{y}_2)^2$$

Outra maneira de escrever essa soma de quadrados. Dentro do grupo de homens, a variância σ_1^2 será estimada por

$$s_1^2 = \frac{1}{n_1 - 1}\sum_{j=1}^{n_1}(y_{1j} - \bar{y}_1)^2$$

e para as mulheres

$$s_2^2 = \frac{1}{n_2 - 1}\sum_{j=1}^{n_2}(y_{2j} - \bar{y}_2)^2$$

Suposições do Modelo

Supomos inicialmente que $\sigma_1^2 = \sigma_2^2$, portanto temos acima dois estimadores para o mesmo parâmetro σ_ϵ^2. Assim definimos a variância comum ponderada:

$$s_\epsilon^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{SQ(\mu_1, \mu_2)}{n - 2}$$

Exemplo

Grupo de Homens:

$$\bar{y}_1 = 110,1 \quad \sum_{j=1}^{n_1}(y_{1j} - \bar{y}_1)^2 = 670,9 \quad s_1^2 = 74,54$$

Grupo de Mulheres:

$$\bar{y}_2 = 104,9 \quad \sum_{j=1}^{n_2}(y_{2j} - \bar{y}_2)^2 = 566,9 \quad s_1^2 = 62,99$$

temos então que

$$s_\epsilon^2 = \frac{670,9 + 566,9}{18} = 68,77$$

de onde

$$s_\epsilon = 8,29$$

Suposições para Inferência

1. $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$ para todo $i = 1, 2$ e $j = 1, ..., n_i$
2. $E(\epsilon_{ij} * \epsilon_{ik}) = 0$ para todo $j \neq k$ e $i = 1, 2$ (independência **dentro** de cada população)
3. $E(\epsilon_{1j} * \epsilon_{2k}) = 0$ para todo j e k (independência **entre** observações das duas populações)

Suposições

Com essas suposições, teremos duas amostras aleatórias simples independentes, retiradas das populações $N(\mu_1, \sigma_\epsilon^2)$ e $N(\mu_2, \sigma_\epsilon^2)$ de onde teremos:

$$\bar{y}_1 \sim N(\mu_1, \frac{\sigma_\epsilon^2}{n_1}) \quad e \quad \bar{y}_2 \sim N(\mu_2, \frac{\sigma_\epsilon^2}{n_2})$$

Intervalos de Confiança

Temos:

$$IC(\mu_1, 95\%) = \bar{y}_1 \pm t * \frac{s_\epsilon}{\sqrt{n_1}} = 110,1 \pm 2,101 * \frac{8,29}{\sqrt{10}} =]104,59; 115,61[$$

$$IC(\mu_2, 95\%) = \bar{y}_2 \pm t * \frac{s_\epsilon}{\sqrt{n_2}} = 104,9 \pm 2,101 * \frac{8,29}{\sqrt{10}} =]99,39; 110,41[$$

onde $t_{18,}$

Se querermos produzir uma estimativa para um homem e usando dedução similar à obtida anteriormente, teremos:

$$IC(\hat{y}_{1j}, 95\%) = 110,1 \pm 2,101 * 8,29 * \sqrt{\frac{1}{10} + 1} = 110 \pm 18,27$$

para uma mulher:

$$IC(\hat{y}_{2j}, 95\%) = 104,9 \pm 18,27$$

Intervalos de Confiança

Para a diferença:

$$\bar{y}_1 - \bar{y}_2 \sim N(\mu_1 - \mu_2, \frac{\sigma_\epsilon^2}{n_1} + \frac{\sigma_\epsilon^2}{n_2})$$

ou ainda

$$t = \frac{(\bar{y}_1 - \bar{y}_2) - (\mu_1 - \mu_2)}{s_\epsilon \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

$$IC(\mu_1 - \mu_2, 95\%) = (\bar{y}_1 - \bar{y}_2) \pm t * s_\epsilon * \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} =$$

$$= (110,1 - 104,9) \pm 2,101 * 8,29 * \sqrt{\frac{1}{10} + \frac{1}{10}} =$$

$$=]-2,59; 12,99[$$

Neste caso as duas médias podem ser iguais, pois o zero está contido no intervalo

Análise de resíduos

para os homens:

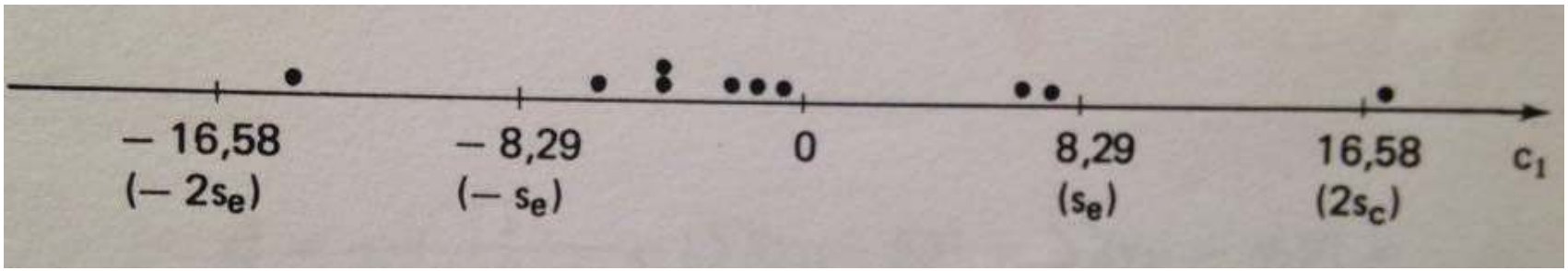


Figura: resíduos do modelo $y_{ij} = \mu_i + \epsilon_{ij}$

para as mulheres:

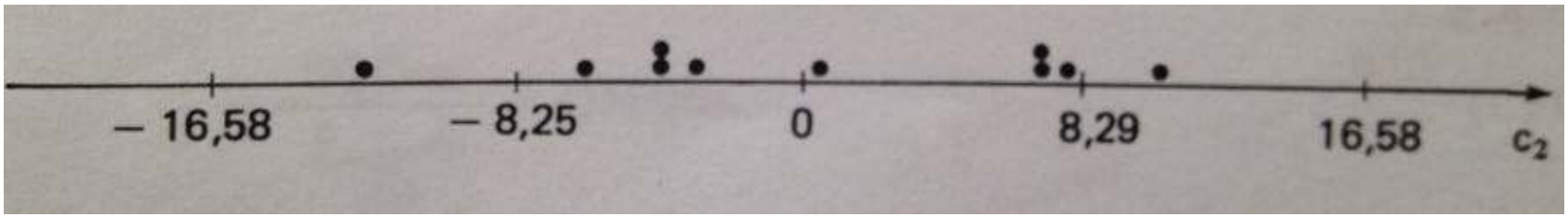


Figura: resíduos do modelo $y_{ij} = \mu_i + \epsilon_{ij}$

Tabela de Análise de Variância

Para o modelo

$$y_i = \mu + \epsilon_i$$

que é equivalente ao modelo

$$y_{ij} = \mu + \epsilon_{ij}$$

a quantidade total de informação perdida é dada por :

$$SQ(\hat{\mu}) = \sum_{i=1}^{n_i} \sum_{j=1}^{n_j} (y_{ij} - \bar{y})^2 = SQT$$

que iremos chamar de ***Soma total de quadrados (SQT)***.

Tabela

Análogamente, para o modelo:

$$y_{ij} = \mu_i + \epsilon_{ij}$$

a quantidade de informação perdida é dada por

$$SQ(\hat{\mu}_1, \hat{\mu}_2) = \sum_{j=1}^{n_1}(y_{1j} - \bar{y}_1)^2 + \sum_{j=1}^{n_2}(y_{2j} - \bar{y}_2)^2 = \sum_{i=1}^{2}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2 = SQDent$$

que iremos chamar de ***Soma de Quadrados dentro dos dos grupos (SQDent)(Ou soma de quadrados residuais (SQRes))***

Tabela

A economia obtida por passar de um modelo para o outro será

$$SQT - SQDent = SQEnt$$

ao que chamaremos de ***Soma de quadrados entre os grupos (SQEnt).*** Podemos verificar que

$$SQEnt = \sum_{i=1}^{2} (\bar{y}_i - \bar{y})^2$$

que mede a distância de cada grupo para a média global. Quanto maior, maior será a economia obtida pelo modelo.

Tabela

Os desvios padrões residuais dos dois modelos eram dados por:

$$s^2 = \frac{1}{n-1}\sum_i\sum_j (y_{ij} - \bar{y})^2 = \frac{SQT}{n-1} = QMT$$

e

$$s^2_\epsilon = \frac{1}{n_1+n_2-2}(\sum_{j=1}^{n_1}(y_{1j} - \bar{y}_1)^2) + \sum_{j=1}^{n_2}(y_{2j} - \bar{y}_2)^2) = \frac{SQDent}{n-2} = QMDent$$

Os quais são referidos como Quadra Médio total (QMT) e Quadrado médio Dentro (QMDent) (ou Residual (QMR))

Tabela ANOVA (ANalysis Of VAriance)

Fonte de Variação (ϕ)	Graus de liberdade (gl)	SQ	QM	F
Entre	1	SQEnt	QMEnt	$\frac{QMEnt}{s^2_\epsilon}$
Dentro	$n-2$	SQDent	QMDent (s^2_ϵ)	
Total	$n-1$	SQT	s^2	

ANOVA para o exemplo

(ϕ)	liberdade (gl)	SQ	QM	F
Entre	1	135,2	135,2	1,97
Dentro	18	1237,8	68,77	
Total	19	1373	72,26	

Para julgarmos a economia do modelo, comparamos 68,77 com 72,26, verificando que não diminuiu muito o erro residual neste novo modelo. Da ANOVA encontramos os desvios padrões residuais:

$$s_\epsilon = \sqrt{68,77} = 8,29 \quad e \quad s = \sqrt{72,26} = 8,5$$

R^2

vemos que ao passar de um modelo para outro, economizamos 135,2 na soma de quadrados, ou seja $\frac{135,2}{1373} = 0,0985 = 9,85\% \approx 10\%$ economizamos aproximadamente 10% na soma dos quadrados dos resíduos. Podemos dizer que essa é a porcentagem de variação total explicada pelo modelo

$$y_{ij} = \mu_i + \epsilon_{ij}$$

Essa medida é chamada de coeficiente de explicação e é definida por

$$R^2 = \frac{SQEnt}{SQT}$$

ANOVA

a conveniência o não da adoção do modelo

$$y_{ij} = \mu_i + \epsilon_{ij}$$

está associada ao teste:

$$H_0 : \mu_1 = \mu_2$$

já que a aceitação desta hipótese implica a adoção do primeiro modelo $(y_{ij} = \mu + \epsilon_{ij})$.
Com as suposições feitas, a estatística do teste é

$$t = \frac{\bar{y}_1 - \bar{y}_2}{s_\epsilon * \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

ANOVA

sabemos que o quadrado da estatística t tem distribuição F de Snedecor com 1 e $n_1 + n_2 - 2$ graus de liberdade.

$$t^2 \sim F_{1,n_1+n_2-2}$$

Contudo,

$$QMEnt = \sum_i (\bar{y}_i - \bar{y})^2 = n_1(\bar{y}_1 - \bar{y})^2 + n_2(\bar{y}_2 - \bar{y})^2$$

e como $\bar{y} = \frac{n_1\bar{y}_1 + n_2\bar{y}_2}{n_1+n_2}$, podemos escrever:

$$\begin{aligned} QMEnt = \quad & n_1\left[\frac{n_2(\bar{y}_1-\bar{y}_2)}{n_1+n_2}\right]^2 + n_2\left[\frac{-n_1(\bar{y}_1-\bar{y}_2)}{n_1+n_2}\right]^2 = \\ & = \frac{n_1 n_2}{n_1+n_2} * (\bar{y}_1 - \bar{y}_2)^2 = \frac{(\bar{y}_1-\bar{y}_2)^2}{\frac{1}{n_1}+\frac{1}{n_2}} \end{aligned}$$

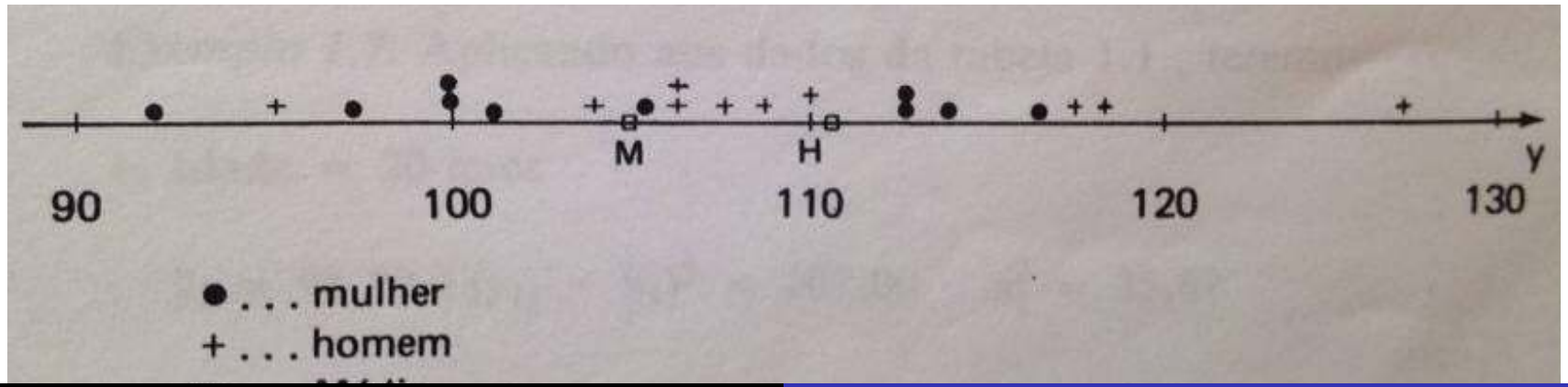
ANOVA

Desse modo teremos:

$$t^2 = \frac{(\bar{y}_1 - \bar{y}_2)^2}{s_\epsilon^2 * (\frac{1}{n_1} + \frac{1}{n_2})} = \frac{QMEnt}{s_\epsilon^2} = F$$

Exemplo

Do exemplo (tabela ANOVA) vemos que $F = 1,97$. Ao consultarmos na tabela F de Snedecor com 1 e 18 graus de liberdade e $\alpha = 5\%$, encontramos o valor crítico para F como sendo 4,41. Portanto, não rejeitamos $H_0 : \mu_1 = \mu_2$. (Não há vantagem em adotar o segundo modelo). O fator sexo não melhora a previsão do tempo de reação do indivíduo.
Na figura não é possível distinguir algum padrão distinto do tempo de reação para homens e mulheres.



Exemplo (efeito da idade)

Propomos o seguinte modelo:

$$y_{ij} = \mu_i + \epsilon_{ij}$$

$i = 1, 2, ..., 5$ (cada grupo de idade)
$j = 1, 2, 3, 4$ (indivíduo)
queremos minimizar:

$$SQ(\mu_1, \mu_2, \mu_3, \mu_4, \mu_5) = \sum_i \sum_j (y_{ij} - \mu_i)^2$$

sujeito à

$$E(\epsilon_i) = 0$$

$$Var(\epsilon_1) = ... = Var(\epsilon_5) = \sigma_\epsilon^2$$

Exemplo

verifica-se que

$$\hat{\mu}_i = \bar{y}_i = \frac{1}{n_i} \sum_j y_{ij}$$

e também que

$$SQDent = SQ(\hat{\mu}_1, ..., \hat{\mu}_5) = \sum_i \sum_j (y_{ij} - \bar{y}_i)^2 = \sum_i (n_i - 1) s_i^2$$

a estimativa ponderada de σ_ϵ^2 será:

$$s_\epsilon^2 = \frac{\sum_i (n_i - 1) * s_i^2}{(n_1 - 1) + ... + (n_5 - 1)} = \frac{SQDent}{n - 5}$$

Exemplo

Para os dados:

1. Idade =20 anos: $\bar{y}_1 = 98,5 \quad \sum_j (y_{1j} - \bar{y}_1)^2 = 107 \quad s_1^2 = 35,67$
2. Idade =25 anos: $\bar{y}_2 = 103,25 \quad \sum_j (y_{2j} - \bar{y}_2)^2 = 78,75 \quad s_2^2 = 26,25$
3. Idade =30 anos: $\bar{y}_3 = 107,75 \quad \sum_j (y_{3j} - \bar{y}_3)^2 = 132,75 \quad s_3^2 = 44,25$
4. Idade =35 anos: $\bar{y}_4 = 110,75 \quad \sum_j (y_{4j} - \bar{y}_4)^2 = 94,75 \quad s_4^2 = 31,58$
5. Idade =40 anos: $\bar{y}_5 = 117,25 \quad \sum_j (y_{5j} - \bar{y}_5)^2 = 140,75 \quad s_5^2 = 46,92$

ANOVA para o fator idade

(ϕ)	(gl)	SQ	QM	F
Entre	4	819	204,75	5,54
Dentro	15	554	36,93	
Total	19	1373	72,26	

É possível ver a redução na soma de quadrados(819), que equivalem a:

$$R^2 = \frac{819}{1373} = 59,65\% \approx 60\%$$

isto é, aproximadamente 60% da variação total é explicada pelo fator idade, reduzindo o erro médio de 8,5 para $\sqrt{36,93} = 6,08$. Parece-nos que a redução foi substancial, sugerindo que a idade influencia na reação.

Exemplo: residuos

Observamos na seguinte figura o tempo de reação segundo a idade:

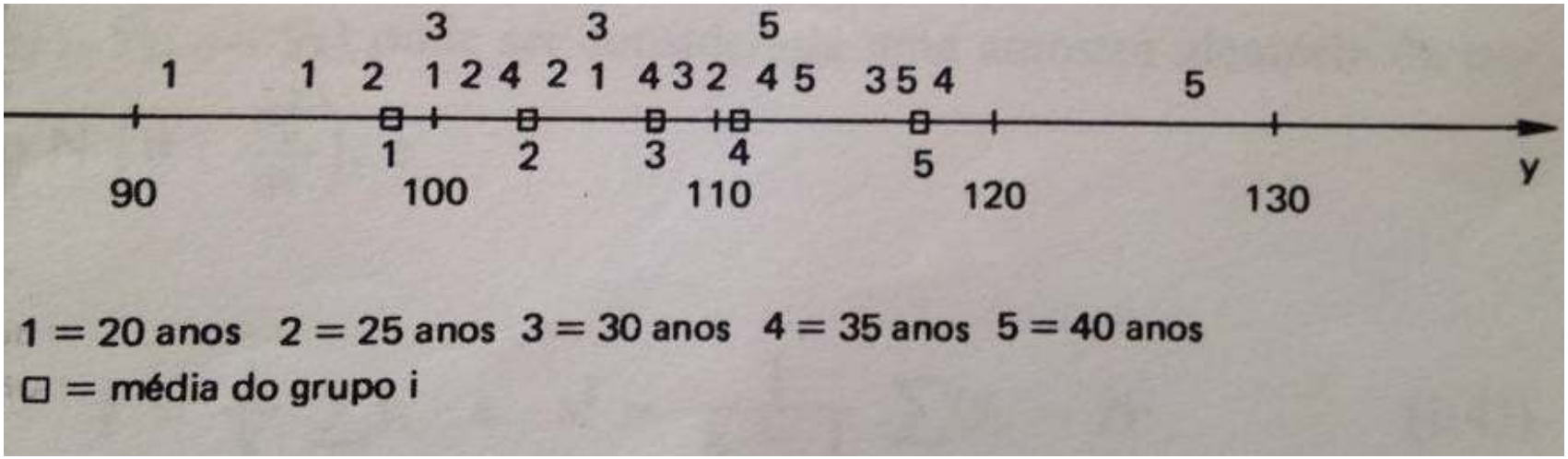


Figura: tempo de reação segundo a idade

ICs para o exemplo

Podemos construir os IC's para os μ_i's ou para observações individuais $\hat{y}_i$. Para o grupo de 25 anos

$$IC(\mu_2, 95\%) = 103,25 \pm 2,131 * \frac{6,08}{\sqrt{4}} = 103,25 \pm 6,48$$

e

$$IC(\hat{y}_2, 95\%) = 103,25 \pm 2,131 * 6.08 * \sqrt{\frac{1}{4} + 1} = 103,25 \pm 14,48$$

Os resíduos para o modelo $y_{ij} = \mu_i + \epsilon_{ij}$ para o fator idade se encontram na seguinte figura

resíduos

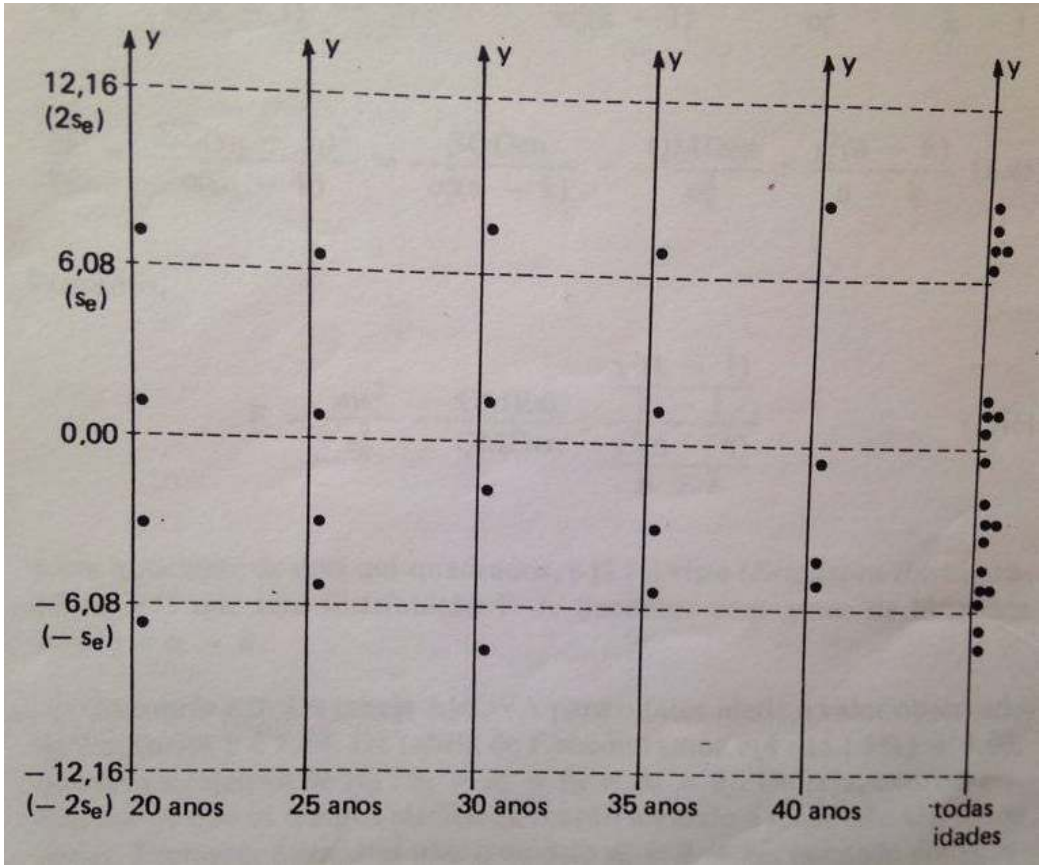


Figura: Resíduos para o fator idade

Teste de igualdade de médias

Testamos agora a hipótese $H_0 : \mu_1 = ... = \mu_k$, contra H_1 : pelo menos uma das igualdades não se verifica.
Trabalhamos inicialmente com a suposição $n_1 = ... = n_k = \frac{n}{k} = m$. A variância comum σ^2_ϵ poderá ser estimada por

$$s^2_\epsilon = \frac{\sum_i (n_i - 1)s^2_i}{n - k} = \frac{SQDent}{n - k} = QMDent$$

Teste de igualdade de médias

sabemos pelo TLC que

$$\bar{y}_i \sim N(\mu_i, \frac{\sigma_\epsilon^2}{m}) \quad i = 1, 2, .., k$$

Quando H_0 é verdadeiro,

$$\bar{y}_i \sim N(\mu, \frac{\sigma_\epsilon^2}{m}) \quad i = 1, 2, .., k$$

de onde $(\bar{y}_1, ..., \bar{y}_k)$ é uma a.a. da população $N(\mu, \frac{\sigma_\epsilon^2}{m})$, e $\bar{y} = \frac{1}{k}\sum_i \bar{y}_i$ e $s_*^2 = \frac{1}{k-1}\sum_i (\bar{y}_i - \bar{y})^2$ são estimadores de μ e de $\frac{\sigma_\epsilon^2}{m}$. Assim, ms_*^2 será outro estimador de σ_ϵ^2

A estatística F

$$F = \frac{ms_*^2}{\sigma_\epsilon^2}$$

tendrá valores proximos de 1. Se H_0 não for verdadeiro, o numerador estará estimando alguma outra coisa e esperam-se resultados distintos da unidade (maiores que 1).
Observando a distribuição de F, vemos que:

$$\frac{ms_*^2}{\sigma_\epsilon^2} = \frac{m}{\sigma_\epsilon^2(k-1)} \sum_i (\bar{y}_i - \bar{y})^2 = \frac{SQEnt}{\sigma_\epsilon^2(k-1)} = \frac{QMEnt}{\sigma_\epsilon^2} \sim \frac{\chi^2_{(k-1)}}{k-1}$$

$$\frac{s_\epsilon^2}{\sigma_\epsilon^2} = \frac{\sum_i \sum_j (y_{ij} - \bar{y}_i)^2}{\sigma_\epsilon^2(n-k)} = \frac{SQDent}{\sigma_\epsilon^2(n-k)} = \frac{QMDent}{\sigma_\epsilon^2} \sim \frac{\chi^2_{(n-k)}}{n-k}$$

A estatística F

Portanto,

$$F = \frac{ms_*^2}{s_\epsilon^2} = \frac{QMEnt}{QMDent} = \frac{\frac{\chi^2_{(k-1)}}{k-1}}{\frac{\chi^2_{(n-k)}}{n-k}} \sim F_{(k-1),(n-k)}$$

Da tabela para o fator idade, o valor observado de F é 5,54. Da tabela F, encontramos que $F(4, 15, 5\%) = 3,06$ o que leva à rejeição de $H_0 : \mu_1 = ... = \mu_5$. Ou seja, existem evidências de que os tempos de reação segundo a idade não são todos iguais.
Portanto, é razoável usar o modelo $y_{ij} = \mu_i + \epsilon_{ij}$, estimado por $\hat{y}_{ij} = \bar{y}_i$

Comparações entre as médias

A Análise de variância verifica apenas a hipótese de igualdade entre as médias, e quando ela é rejeitada, não saberemos entre qual das médias existem as diferenças.
Suponha que $H_0 : \mu_1 = \mu_2 = \mu_3$, as alternativas para esta hipótese são: $\mu_1 = \mu_2 \neq \mu_3$, $\mu_1 \neq \mu_2 = \mu_3$, $\mu_1 = \mu_3 \neq \mu_2$ e $\mu_1 \neq \mu_2 \neq \mu_3$ Uma maneira de verificar as alternativas seria construir intervalos de confiança para as diferenças e verificar em quais intervalor o valor zero não está contido:

$$IC(\mu_1 - \mu_2, 1 - \alpha) = (\bar{y}_1 - \bar{y}_2) \pm t * s_\epsilon * \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \sim t_{n-k}$$

Exemplo

No exemplo, H_0 foi rejeitada, o IC para a diferença de duas médias quaisquer será:

$$IC(\mu_1 - \mu_2, 95\%) = (\bar{y}_1 - \bar{y}_2) \pm 2,131 * 6,08 * \sqrt{\frac{1}{4} + \frac{1}{4}} = (\bar{y}_1 - \bar{y}_2) \pm 9,16$$

Assim, grupos cuja diferença de médias seja superior a 9,16 seriam diferentes.

Grupo	20		25		30		35		40
Média	98,5		103,25		107,75		110,75		117,25
Diferença		4,75		4,50		3		6,5	

Exemplo

No caso anterior, não se pode controlar a probabilidade do erro tipo I: Por exemplo, suponhamos que todas as médias sejam iguais, teriamos então 10 possíveis comparações duas a duas, cada um testada ao 5% e a probabilidade de que pelo menos uma das comparações exceda os 9,16 é bem maior do que 5% (algo de 29%). Esse nível cresce quando cresce o número de comparações. Para isto, utiliza-se a correção de intervalos de Bonferroni, onde o valor de t será agora t_* com $\alpha_* = \frac{\alpha}{\nu}$ onde ν é o número de comparações duas a duas.
Para o exemplo $\alpha_* = \frac{5\%}{10} = 0,5\%$. Da tabela t com 15 g.l. encontramos $t_* = 3,438$, e então:

$$IC(\mu_1 - \mu_2, 95\%) = (\bar{y}_1 - \bar{y}_2) \pm 3,438 * 6,08 * \sqrt{\frac{1}{4} + \frac{1}{4}} = (\bar{y}_1 - \bar{y}_2) \pm 14,78$$

Primeiras Idéias

A Análise de Variância (ANOVA) se refer à coleção de situações experimentais e procedimentos estatísticos para a análise de respostas quantitativas desde unidades experimentais

Terminologia

A característica que diferencia os tratamentos ou uma população de outra é chamada de *fator* sob estudo.
Os diferentes tratamentos ou populações são referidos como os *níveis* do fator.
Uma terminologia alternativa será utilizada nas seguintes apresentações:

$$SQEntre(SQEnt) = SQTr \quad \text{(ou Soma de Quadrados dos Tratamentos)}$$

$$SQDentro(SQDent) = SQE \quad \text{(ou Soma de Quadrados dos Erros)}$$

$$QMEnt = QMTr \quad \text{(Quadrado Médio dos tratamentos)}$$

$$QMDent = QME \quad \text{(Quadrado Médio dos Erros)}$$

ANOVA de um fator

ANOVA de um fator foca na comparação das médias de duas ou mais populações ou tratamentos.
Sejam:

$$
\begin{aligned}
k &= \text{O número de tratamentos (populações) sendo comparadas} \\
\mu_1 &= \text{a média da população 1 ou a resposta promédio verdadeira quando o tratamento 1 é aplicado.} \\
\mu_k &= \text{a média da população k ou a resposta promédio verdadeira quando o tratamento k é aplicado.}
\end{aligned}
$$

ANOVA de um fator

As hipóteses de interesse são:

$$H_0 : \mu_1 = \mu_2 = ... = \mu_k$$

versus

$$H_a : \text{pelo menos dois } \mu_i\text{'s são diferentes}$$

Notação

Y_{ij} = a variável que denota a j-ésima medida tomada da i-ésima população, ou a medida tomada da $j-$ésima unidade experimental que receve o $i-$ésimo tratamento.
y_{ij} = O valor observado de Y_{ij} quando o experimento é realizado.

Suposições

A população k ou as distribuições dos tratamentos são todos normais com a mesma variância σ^2.
Cada Y_{ij} é normalmente distribuida com

$$E(Y_{ij}) = \mu_i$$

$$V(Y_{ij}) = \sigma^2$$

Quadrado Médio para os tratamentos e o Erro

supondo que os tamanhos de cada população (ou fator) sejam iguais, isto é: $n_1 = .. = n_k = m$

Quadrado médio para tratamentos:

$$\begin{aligned} QMTr = & \frac{m}{k-1}\left[(\bar{y}_{1.} - \bar{y})^2 + ... + (\bar{y}_{k.} - \bar{y})^2\right] \\ = & \frac{m}{k-1}\sum_i (\bar{y}_{i.} - \bar{y})^2 \end{aligned}$$

Quadrado médio para o Erro:

$$QME = \frac{s_1^2 + s_2^2 + ... + s_k^2}{k}$$

Teste Estatístico

O teste estatístico para o ANOVA de um fator é

$$F = \frac{QMTr}{QME}$$

quando H_0 foi verdadeiro e as suposições básicas forem satisfeitas, $F \sim F_{(k-1),(n-k)}$.
Quando f representa o valor calculado de F, a região de rejeição $f \geq F_{\alpha,k-1,n-k}$ especifica um teste com nível de significância α

Valor Esperado

Quando H_0 é verdadeiro,

$$E(QMTr) = E(QME) = \sigma^2$$

Quando H_0 é falso,

$$E(QMTr) > E(QME) = \sigma^2$$

A não tendenciosidade de QME é uma consequência de $E(s_i^2) = \sigma^2$, quer H_0 seja ou não verdadeira. Se H_0 for verdadeiro, todo $E(\bar{y}_i) = \mu$ e $var(\bar{y}_i) = \sigma^2/m$ de modo que $\sum_i (\bar{y}_i - \bar{y})^2/(k-1)$ a variância amostral dos $\bar{y}_i$ estima σ^2/m de maneira não tendenciosa.

A distribuição F e o teste

Seja

$$F = \frac{QMTr}{QME}$$

a estatística para o ANOVA de um fator envolvendo k populações ou tratamentos com uma amostra aleatória de m observações de cada uma. Quando H_0 é verdadeira (suposição básica verdadeira),

$$F \sim F(v_1 = k-1, v_2 = n-k)$$

a região de rejeição

$$f \geq F_{\alpha,k-1,n-k}$$

especifica um teste com nível de significância α

Fórmulas para ANOVA

Se considerarmos $n_1 = ... = n_k = m$

Soma de quadrados total

$$SQT = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}^2 - \frac{1}{n}(\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij})^2$$

Soma de quadrados dos tratamentos

$$SQTr = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{y_{i.}} - \bar{y})^2 = \frac{1}{m}\sum_{i=1}^{k} y_{i.}^2 - \frac{1}{n}(\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij})^2$$

Soma de quadrados do Erro

$$SQE = \sum_{i=1}^{k}\sum_{j=1}^{m}(y_{ij} - \bar{y}_{i.})^2$$

$$SQT = SQTr + SQE$$

Quadrados Médios

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

$$QME = \frac{SQE}{n-k}$$

$$F = \frac{QMTr}{QME}$$

Tabela ANOVA

FV	gl	SQ	QM	f
Tratamentos	k-1	SQTr	QMTr	QMTr/QME
Erro	n-k	SQE	QME	
Total	n-1	SQT		

Legenda:
FV=Fonte de Variação
gl=graus de liberdade
SQ= Soma de Quadrados
QM= Quadrado Médio

Distribuição do rank studentizado e pares de diferenças

Com probabilidade $1-\alpha$,

$$\bar{y}_{i.} - \bar{y}_{j.} - Q_{\alpha,k,n-k}\sqrt{\frac{QME}{m}} \leq \mu_i - \mu_j \leq \bar{y}_{i.} - \bar{y}_{j.} + Q_{\alpha,k,n-k}\sqrt{\frac{QME}{m}}$$

para cada i e j com $i < j$

O método T para identificar μ_i's significativamente diferentes

1. Selecione α e extraia $Q_{\alpha,k,n-k}$
2. Calcule $w = Q_{\alpha,k,n-k} \times \sqrt{QME/m}$
3. Liste as médias amostrais em ordem crescente, marque aqueles que diferem por mais do que w. Qualquer par não marcado pela mesma linha corresponde a um par significativamente diferentes

Intervalos de confiança para outras funções paramétricas

Seja $\mu = \sum c_i \mu_i$ os y_{ij}'s são normalmente distribuidos

$$V(\hat{\mu}) = V(\sum_i c_i \bar{y}_i) = \frac{\sigma^2}{m} \sum_i c_i^2$$

Estimando σ^2 pelo QME e formando $\hat{\sigma}_{\hat{\mu}}$ resulta em uma variável t $(\hat{\mu} - \mu)\hat{\sigma}_{\hat{\mu}}$ que leva a :

$$\sum c_i \bar{y}_i \pm t_{\alpha/2, n-k} \sqrt{\frac{QME \sum c_i^2}{m}}$$

Modelo ANOVA

As suposições para um ANOVA de um fator podem ser modeladas por:

$$y_{ij} = \mu_i + \epsilon_{ij}$$

onde ϵ_{ij} representa o desvio aleatório da população ou da média verdadeira do tratamento μ_i

Mais sobre ANOVA de um Fator

$$E(QMTr) = \sigma^2 + \frac{m}{k-1}\sum \alpha_i^2$$

Note que quando H_0 é verdadeiro,

$$\sum \alpha_i^2 = 0$$

β para o teste F

Considere um conjunto de valores paramétricos $\alpha_1, ..., \alpha_n$ para os quais H_0 não é verdadeiro.
A probabilidade de um erro de tipo II, β, é a probabilidade que H_0 não seja rejeitada quando o conjunto é o conjunto de valores verdadeiros.

ANOVA de um fator quando os tamanhos amostrais são diferentes

$$SQT = \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}^2 - \frac{1}{n}y^2 \qquad gl = n-1$$

$$SQTr = \sum_{i=1}^{k}\frac{1}{n_i}y_{i.}^2 - \frac{1}{n}y^2 \qquad gl = k-1$$

$$SQE = SQT - SQTr \qquad gl = n-k$$

ANOVA de um fator quando os tamanhos amostrais são diferentes

Valor do teste estatístico

$$f = \frac{QMTr}{QME}$$

onde

$QMTr = \frac{SQTr}{k-1}$ e $QME = \frac{SQE}{n-k}$

Região de Rejeição

$$f \geq F_{\alpha,k-1,n-k}$$

Comparações múltiplas (amostras de tamanhos diferentes)

Seja

$$w_{ij} = Q_{\alpha,k,n-k}\sqrt{\frac{QME}{2}(\frac{1}{n_i}+\frac{1}{n_j})}$$

Então a probabilidade é aproximadamente $1-\alpha$ que

$$\bar{y}_{i.} - \bar{y}_{j.} - w_{ij} \leq \mu_i - \mu_j \leq \bar{y}_{i.} - \bar{y}_{j.} + w_{ij}$$

para cada i e j com $i \neq j$

Transformação de dados

Se $V(X_{ij}) = g(\mu_i)$,
uma função conhecida de μ_i, então a transformação $h(X_{ij})$ que "estabiliza a variância" tal que $V[h(X_{ij})]$ é aproximadamente o mesmo para cada i é dado por

$$h(x) \propto \int [g(x)]^{-1/2} dx$$

Modelo de efeitos aleatórios

$$X_{ij} = \mu + A_i + \epsilon_{ij}$$

com

$$E(A_i) = E(\epsilon_{ij}) = 0$$

$$V(\epsilon_{ij}) = \sigma^2$$

$$V(A_i) = \sigma_A^2$$

Todos os A_i's e ϵ_{ij}'s são normalmente distribuidos e independentes um dos outros

PROBLEMAS

1. Usando a variável W, acuidade visual, do exemplo 1.1:
   (*a*) Qual o correspondente ao modelo (1.1)? Dê o significado de cada símbolo.
   (*b*) Qual a estimativa de MQ para θ?
   (*c*) E para $\hat{\sigma}_e^2$?
   (*d*) Construa IC(θ : 95%).
   (*e*) Construa IC($\bar{y}$: 90%).
   (*f*) Faça uma análise dos resíduos.

2. No quadro abaixo estão os dados referentes a uma amostra de 21 alunos do primeiro ano de um curso universitário. Os dados referem-se a:
   y : nota obtida na primeira prova do curso
   x : se cursou escola particular(P) ou oficial(O)
   z : o período em que está matriculado: matutino(M), vespertino(V) e noturno(N).

y	56	68	69	70	70	72	75	77	83	84	84
x	P	O	P	P	O	O	O	P	P	P	O
z	N	M	M	M	V	N	M	M	V	N	N

y	85	90	92	95	95	95	100	100	100	100
x	O	P	O	P	P	P	P	P	P	P
z	V	V	M	M	N	V	V	M	M	V

Faça uma análise da variável y análoga à realizada no problema 1.

3. Conduziu-se um estudo-piloto para determinar qual o intervalo de normalidade para o peso de crianças com 10 anos de idade. Usando-se uma amostra de 50 crianças; encontrou-se o peso x de cada uma delas, cujos resultados resumidos são: $\Sigma x = 1\,639{,}5$ kg e $\Sigma x^2 = 56\,950{,}33$ kg^2. Com esses dados, quais seriam os limites de um intervalo, para que crianças com 10 anos de idade fossem consideradas como tendo peso normal? [Pense no IC, e lembre-se que $\Sigma(x - \bar{x})^2 = \Sigma x^2 - (\Sigma x)^2/n$.]

4. Usando a acuidade visual como a variável resposta (exemplo 1.1), e sexo como a variável de classificação:
(*a*) Encontre as estimativas de MQ para θ_1 e θ_2.
(*b*) Encontre a estimativa da variância residual comum σ_e^2.
(*c*) Construa o IC($\theta_1 - \theta_2$: 95%).
(*d*) Construa a tabela ANOVA e analise os resultados obtidos.
(*e*) Calcule o valor de R^2.
(*f*) Você diria que o fator sexo influi na acuidade visual? (Teste a hipótese $H_0 : \theta_1 = \theta_2$.)
(*g*) A análise dos resíduos sugere a violação de algumas das suposições básicas?
(*h*) Qual seria a acuidade visual esperada para uma mulher?

5. Usando os dados do problema 2, você diria que o fato de a pessoa ter cursado a escola particular ou oficial influi no resultado da primeira prova?
(Sugetão: siga todos os passos do problema 4, antes de tomar sua decisão.)

6. Em uma pesquisa sobre rendimentos por hora, entre assalariados segundo o grau de instrução, obtiveram-se os dados do quadro abaixo. Construa a tabela ANOVA e verifique se existe diferença significativa entre os rendimentos das duas categorias.

Escolaridade	n	Σx	Σx^2
1.º grau	50	111,50	259,93
2.º grau	20	71,00	258,89

[Observação: os rendimentos x estão expressos como % do salário mínimo (SM)].

7. Usando os dados do exemplo 1.1, construa um modelo para analisar o efeito da idade sobre a acuidade visual. Siga o roteiro:
   (*a*) Encontre as estimativas para $\hat{\theta}_i$.
   (*b*) Construa a tabela ANOVA.
   (*c*) Teste a hipótese $\theta_1 = \theta_2 = \ldots = \theta_5$.
   (*d*) Caso rejeite, use a técnica descrita no exemplo numérico 1.10, para explorar razão para a rejeição.
   (*e*) Investigue os resíduos, em busca de possíveis transgressões das suposições.
   (*f*) Encontre o valor de R^2.
   (*g*) Qual seria a acuidade média esperada para indivíduos com 40 anos de idade?
   (*h*) Qual a acuidade média esperada para uma pessoa com 40 anos?

8. Usando os dados do problema 2, você diria que o período que o aluno está cursando influencia no seu desempenho da primeira prova?

9. (Continuação do problema 6) Na pesquisa de salários acrescentou-se uma amostra de universitários.
   (*a*) Você diria que o grau de escolaridade influencia os rendimentos?
   (*b*) Qual seria o rendimento médio para uma pessoa com formação universitária?
   (*c*) Existe diferença entre os rendimentos médios daqueles com 1º e 2º graus?

Escolaridade	x	Σx	Σx^2
1.º Grau	50	111,50	259,93
2.º Grau	20	71,00	258,89
3.º Grau	10	84,30	717,94

10. Quer-se verificar a durabilidade de duas marcas de tintas que têm preços de custo bem diferenciados. Para isso foram selecionados 10 casos, 5 pintados com uma das marcas e os 5 restantes com a outra marca. Após um período de seis meses, foi atribuída a cada caso uma nota, resultante de uma série de quesitos. Os resultados foram os seguintes:

1.ª Marca	85	87	92	80	84
2.ª Marca	91	91	92	86	90

Com esses dados, você diria que uma das marcas é melhor do que a outra?

11. Queremos verificar o efeito do tipo de impermeabilização na condutividade de tubos de TV. Com os resultados da tabela abaixo, que tipo de conclusão você poderia obter?

Tipo de Impermeabilização			
I	II	III	IV
56	64	45	42
55	61	46	39
62	50	45	45
59	55	39	43
60	56	43	41

12. Os dados abaixo vêm de um experimento completamente aleatorizado, onde 5 processos de estocagem foram usados em um produto perecível por absorção de água. Vinte e cinco exemplares deste produto foram divididos em 5 grupos de 5 elementos cada, e após uma semana mediu-se a quantidade de água absorvida. Os resultados codificados estão no quadro abaixo. Existem evidências de que os processos de estocagem produzem resultados diferentes?

Estocagem				
A	B	C	D	E
8	4	1	4	10
6	−2	2	6	8
7	0	0	5	7
5	−2	−1	5	4
8	3	−3	4	9

13. A seção de treinamento de uma empresa quer saber qual de três métodos de ensino é mais eficiente. O encarregado de responder a essa pergunta pode dispor de 24 pessoas para verificar a hipótese. Ele as dividiu em 3 grupos de 8 pessoas, de modo aleatório, e submeteu cada grupo a um dos métodos. Após o treinamento os 24 participantse foram submetidos a um mesmo teste, cujos resultados estão no quadro abaixo. Quais seriam as conclusões sobre os métodos de treinamento?

Método 1		*Método 2*		*Método 3*	
3	8	4	7	6	7
5	4	4	4	7	9
2	3	3	2	8	10
4	9	8	5	6	9
Σx	38		37		62
Σx^2	224		199		496

14. Quer-se testar o efeito do tipo de embalagem sobre as vendas do sabonete SEBO. As embalagens são as seguintes:

A: a tradicional embalagem preta
B: cartolina vermelha
C: papel alumínio rosa

Escolheram-se três territórios de venda, com condições de potencial de vendas supostamente idênticas. Cada tratamento foi designado aleatoriamente a cada região, e as vendas observadas durante 4 semanas. Quais seriam suas conclusões e críticas a este experimento?

Réplicas (Semanas)	*Embalagens* A	B	C
1	15	21	9
2	20	23	13
3	9	19	20
4	12	25	18
Total	56	88	

15. Um produtor de gelatina em pó está testando um novo lançamento e quer verificar em que condições de preparo o produto seria melhor aceito. Vinte e quatro donas-de-casa atribuíram notas (0 a 10) para o prato que produziram com aquele produto. Junto com o produto foram fornecidos quatro tipos de receitas: duas para doce (A e D) e duas para salgados (B e C). Feita a análise estatística, que recomendações você faria ao produtor? Discuta a validade das suposições feitas para resolver o problema.

Receita			
A	B	C	D
2	4	3	3
5	7	5	6
1	3	1	2
7	9	9	8
2	4	6	1
6	8	8	4

16. Uma escola avalia o seu curso através de um questionário com 50 perguntas sobre diversos aspectos de interesse. Cada pergunta tem uma resposta numa escala de 1 a 5, onde a maior nota significa melhor desempenho. Para cada aluno é então encontrada a nota média. Na última avaliação usou-se uma amostra de alunos, de cada um dos períodos, e os resultados estão abaixo. Quais as suas conclusões estatísticas sobre a existência ou não de opinião segundo os períodos?

Período		
Manhã	*Tarde*	*Noite*
4,2	2,7	4,6
4,0	2,4	3,9
3,1	2,4	3,8
2,7	2,2	3,7
2,3	1,9	3,6
3,3	1,8	3,5
4,1		3,4
		2.8

17. Em um curso de extensão universitária, entre outras informações, perguntam-se os salários mensais, expressos em uma determinada unidade, e também a área de formação acadêmica. Eliminou-se três salários por serem exageradamente altos, e os restantes produziram as seguintes estatísticas:

Formação	x	$\bar{x}$	s
Humanas	65	28,75	3,54
Exatas	12	35,21	5,46
Biológicas	8	43,90	4,93

Teste a hipótese de que os salários médios nas três áreas de formação acadêmica é o mesmo.

18. Suspeita-se que quatro livros, escritos sob pseudônimo, são de um único autor. Uma pequena investigação inicial selecionou amostras de páginas de cada um dos livros, contando-se o número de preposições. Com os resultados abaixo, quais seriam as suas conclusões?

Livros			
1	2	3	4
28	29	26	39
31	33	24	27
17	35	22	35
25	24	19	34
26	28	23	28
22		25	34
24		29	33
		30	

19. Prove que QMEnt $= \Sigma n_i(\bar{y}_i - \bar{y})^2$.