Capítulo 15

Inferência para Várias Populações

15.1 Introdução

Como vimos no Capítulo 1, uma das preocupações de um estatístico ao analisar um conjunto de dados é criar modelos que explicitem estruturas do fenómeno sob observação, as quais frequentemente estão misturadas com variações acidentais ou aleatórias. A identificação dessas estruturas permite conhecer melhor o fenómeno, bem como fazer afirmações sobre possíveis comportamentos.

Portanto, uma estratégia conveniente de análise é supor que cada observação seja formada por duas partes, como vimos em (1.1) do Capítulo 1:

$$\text{observação} = \text{previsível} + \text{aleatório}. \qquad (15.1)$$

Aqui, a primeira componente incorpora o conhecimento que o pesquisador tem sobre o fenômeno e é usualmente expressa por uma função matemática, com parâmetros desconhecidos. A segunda parte, a aleatória (ou não previsível), representa aquilo que o pesquisador não pode controlar e para a qual são impostas algumas suposições, como, por exemplo, que ela obedeça a algum modelo probabilístico específico, que, por sua vez, também contém parâmetros desconhecidos.

Dentro desse cenário, o trabalho do estatístico passa a ser o de estimar os parâmetros desconhecidos das duas partes do modelo, baseado em amostras observadas.

Neste capítulo iremos investigar um modelo simples, chamado de *análise de variância com um fator*. No capítulo seguinte iremos estudar o modelo de regressão linear simples. As técnicas de análise de variância foram desenvolvidas principalmente pelo estatístico inglês Ronald A. Fisher, a partir de 1918. O leitor interessado pode consultar os trabalhos pioneiros de Fisher (1935, 1954) ou Peres e Saldiva (1982) para mais informações sobre esse assunto.

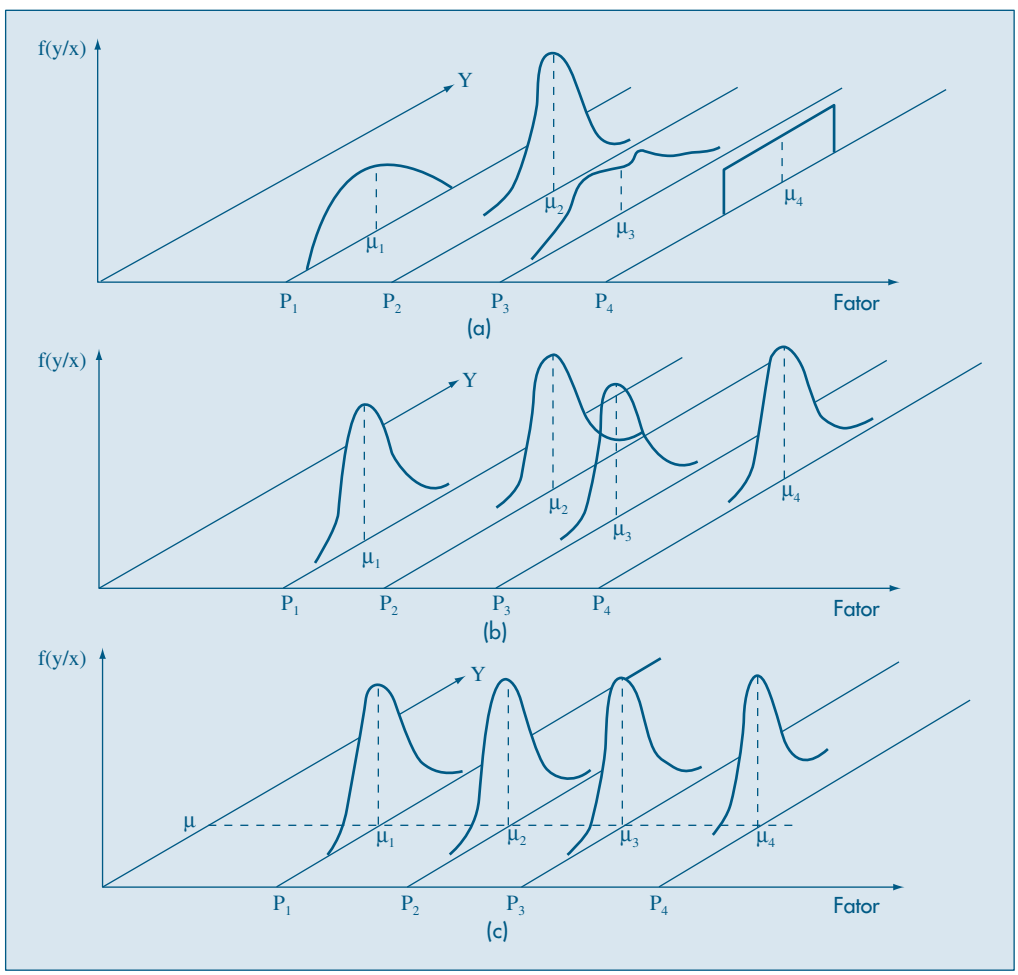
A situação geral pode ser descrita como segue. Temos uma população P de unidades experimentais (indivíduos, animais, empresas etc.), para a qual temos uma v.a. Y de interesse.

Suponha, agora, que possamos classificar as unidades dessa população segundo *níveis* de um *fator*. Por exemplo, o fator pode ser o sexo, com dois níveis, arbitrariamente denotados por l: sexo masculino e 2: sexo feminino. A v.a. Y pode ser a altura de cada indivíduo.

Genericamente podemos ter I níveis para esse fator. A população fica, então, dividida em I subpopulações (ou estratos), P_1, ..., P_I, cada uma representada por um nível i do fator, $i = 1, 2, ..., I$. No exemplo citado teríamos duas subpopulações: a dos indivíduos do sexo masculino e a dos indivíduos do sexo feminino.

Na Figura 15.1 mostramos graficamente as suposições adotadas para o comportamento da população neste modelo. A Figura 15.1 (a) mostra um comportamento mais amplo, com distribuições distintas para cada subpopulação. Na Figura 15.1 (b), aparece a suposição mais comum, em que a parte aleatória segue uma distribuição normal, com a mesma variância σ^2 para todas as subpopulações P_i, $i = 1, 2, ..., I$.

Figura 15.1: Formas da distribuição de *y* para os diversos níveis do fator.



Para cada nível i, observamos a v.a. Y em n_i unidades experimentais selecionadas ao acaso da subpopulação correspondente, ou seja, teremos uma amostra $(y_{i_1}, ..., y_{in_i})$ dessa subpopulação. No exemplo citado acima, temos $i = 1, 2$, ou seja, dois níveis para o fator sexo. Extraímos uma amostra de tamanho n_1 de P_1: pessoas do sexo masculino, $(y_{11}, ..., y_{1n_1})$, e uma amostra de tamanho n_2 de P_2: pessoas do sexo feminino, $(y_{21}, ..., y_{2n_2})$. Essas amostras são independentes.

Suponha que $E(Y) = \mu$ para a população toda, ou seja, a *média global* da v.a. Y para P. Suponha, também, que $E(Y|P_i) = \mu_i$, $i = 1, ..., I$, ou seja, as médias da v.a. Y para as subpopulações sejam $\mu_1, ..., \mu_I$. No nosso exemplo, μ é a média das alturas da população de todos os indivíduos, μ_1 é a média das alturas dos homens, e μ_2 é a média das alturas das mulheres.

O objetivo é estimar μ_i, $i = 1, ..., I$ e testar hipóteses sobre essas médias. Uma hipótese de interesse é

$$H_0: \mu_1 = \mu_2 = ... = \mu_I = \mu, \tag{15.2}$$

contra a alternativa

$$H_1: \mu_i \neq \mu_j, \text{ para algum par } (i, j). \tag{15.3}$$

O teste acima corresponde a verificar se as duas populações estão dispostas como na Figura 15.1 (c), ou seja, os centros das distribuições têm a mesma ordenada e estão sobre uma reta paralela ao eixo do fator. Isso significa que o fator não tem influência sobre a média da variável sob observação.

A análise da variância pode ser pensada como um método para testar a hipótese H_0 acima, por meio da análise das variâncias das diversas amostras. Esse método estende aquele visto no Capítulo 13, onde comparávamos apenas duas médias. A teoria desenvolvida naquele capítulo envolvia situações mais amplas do que as que serão vistas aqui. Sob as mesmas suposições os dois métodos são equivalentes. Porém, não podemos usar os métodos do Capítulo 13 para comparar mais do que duas populações. Poderia ser aventada a possibilidade de testar as hipóteses duas a duas, mas isso traz problemas relacionados no nível de significância do teste global, já que efetuaremos $\binom{I}{2}$ testes parciais. Voltaremos a esse assunto na seção 15.4,

Um modelo conveniente para descrever essa situação é

$$y_{ij} = \mu_i + e_{ij} \qquad i = 1, ..., I, \qquad j = 1, ... n_i, \tag{15.4}$$

para o qual supomos que e_{ij} são v.a. independentes, de média zero e variância σ_e^2, desconhecida, por exemplo. Podemos adicionar a hipótese de que esses "erros" sejam normais, ou seja,

$$e_{ij} \sim N(0, \sigma_e^2), \tag{15.5}$$

para $i = 1, 2, ..., I$, $j = 1, 2, ..., n_i$.

Logo, além de estimar μ_1, ..., μ_I, temos que estimar também σ_e^2. Se (15.4) e (15.5) valerem, teremos I subpopulações normais $N(\mu_i, \sigma_e^2)$, $i = 1, 2, ..., I$, que têm médias diferentes e mesma variância. A Figura 15.1 (b) ilustra essa situação, com $I = 4$.

O modelo (15.4) é chamado *modelo com efeitos fixos*, no sentido de que as subpopulações determinadas pelos níveis do fator são aquelas de interesse do pesquisador. Se o experimento fosse repetido, amostras aleatórias das *mesmas* subpopulações seriam extraídas e analisadas. Pode-se considerar, também, modelos com efeitos aleatórios, mas esse caso não será tratado neste livro.

Exemplo 15. l. Um psicólogo está investigando a relação entre o tempo que um indivíduo leva para reagir a um estímulo visual (Y) e alguns fatores, como sexo (W), idade (X) e acuidade visual (Z, medida em porcentagem). Na Tabela 15.1 temos os tempos para $n = 20$ indivíduos (valores da v.a. Y). O fator sexo tem dois níveis: $i = 1$: sexo masculino (H) e $i = 2$: sexo feminino (M), com $n_1 = n_2 = 10$. O fator idade tem cinco níveis: $i = 1$: indivíduos com 20 anos de idade, $i = 2$: indivíduos com 25 anos etc., $i = 5$: indivíduos com 40 anos. Aqui, $n_1 = ... = n_5 = 4$. A acuidade visual, como porcentagem

Tabela 15.1: Tempos de reação a um estímulo (Y) e acuidade visual (Z) de 20 indivíduos, segundo o sexo (W) e a idade (X).

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

da visão completa, também gera cinco níveis: $i = 1$: indivíduos com 100% de visão, $i = 2$: indivíduos com 90% de visão, e assim por diante. Não foi possível controlar essa variável *a priori* como as outras duas, já que ela exige exames oftalmológicos para sua mensuração. Daí o desbalanceamento dos tamanhos observados: $n_1 = 2$, n_2, $= 10$, $n_3 = 5$, $n_4 = 2$ e $n_5 = 1$. Fatores desse tipo são chamados de *co-fatores*.

Assim, para o fator sexo, teremos o modelo (15.4) com $i = 1, 2$, $j = 1, 2, 3, ..., 10$, e para o fator idade, o mesmo modelo com $i = 1, 2, ..., 5$, $j = 1, 2, 3, 4$.

Exemplo 15.2. Uma escola analisa seu curso por meio de um questionário com 50 questões sobre diversos aspectos de interesse. Cada pergunta tem uma resposta, numa escala de 1 a 5 (v.a. Y), onde a maior nota significa melhor desempenho. Na última avaliação usou-se uma amostra de alunos de cada período, e os resultados estão na Tabela 15.2. Aqui, o fator é período, com três níveis: $i = 1$: manhã, $i = 2$: tarde e $i = 3$: noite; temos $n_1 = 7$, $n_2 = 6$ e $n_3 = 8$.

Tabela 15.2: Avaliação de um curso segundo o período.

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

Exemplo 15.3. Num experimento sobre a eficácia de regimes para emagrecer, homens, todos pesando cerca de 100 kg e de biotipos semelhantes, são submetidos a três regimes. Após um mês, verifica-se a perda de peso de cada indivíduo, obtendo-se os valores da Tabela 15.3.

Tabela 15.3: Perdas de peso de indivíduos submetidos a três regimes.

	Regime	
1	2	3
11,8	7,4	10,5
10,5	9,7	11,2
12,5	8,2	11,8
12,3	7,2	13,1
15,5	8,6	14,0
11,4	7,1	9,8

Aqui, o fator é regime, com $I = 3$ níveis e cada regime é indexado por; $i = 1, 2, 3$. A v.a. Y é a perda de peso depois de um mês. $E(Y) = \mu$ é a perda de peso global dos 18 homens, μ_i é a perda média de peso para o regime i. As amostras têm todas o mesmo tamanho $n_1 = n_2 = n_3 = 6$.

Problemas

1. O modelo (15.4) pode ser escrito na forma

$$y_{ij} = \mu + \alpha_i + e_{ij},$$

com $i = 1, \ldots I$ e $j = 1, \ldots n_i$. Dizemos que α_i, é o efeito diferenciado da subpopulação P_i ou do nível i do fator. Mostre que os estimadores de mínimos quadrados para μ e α_i são dados por

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{I} \sum_{j=1}^{n_i} y_{ij} = \bar{y},$$

$$\hat{\alpha}_i = \bar{y}_i - \bar{y}, \quad \text{com} \quad \bar{y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij},$$

se impusermos a condição $\sum_{i=1}^{I} n_i \alpha_i = 0$.

2. Obtenha $\hat{\mu}$, $\hat{\alpha}_i$, para os Exemplos 15.2 e 15.3.

15.2 Modelo para Duas Subpopulações

Inicialmente, consideremos o caso em que temos um fator com dois níveis, como no Exemplo 15.1, com o fator sexo. Ou seja, queremos avaliar o efeito do sexo do indivíduo sobre o seu tempo de reação ao estímulo. Temos, então, o modelo

$$y_{ij} = \mu_i + e_{ij}, \tag{15.6}$$

onde

μ_i = efeito comum a todos os elementos do nível $i = 1, 2$;

e_{ij} = efeito aleatório, não-controlado, do j-ésimo indivíduo do nível i,

y_{ij} = tempo de reação ao estímulo do j-ésimo indivíduo do nível i.

15.2.1 Suposições

É necessário introduzir suposições sobre os erros e_{ij} a fim de fazer inferências sobre μ_1 e μ_2. Iremos admitir que:

(i) $e_{ij} \sim N(0, \sigma_e^2)$, para todos $i = 1, 2$ e $j = 1, 2, \ldots, n_i$.

(ii) $E(e_{ij}\, e_{ik}) = 0$, para $j \neq k$ e $i = 1, 2$, indicando independência entre observações dentro de cada subpopulação.

(iii) $E(e_{1j}\ e_{2k}) = 0$, para todo j e k, indicando independência entre observações das duas subpopulações.

Com essas suposições, temos duas amostras aleatórias simples, independentes entre si, retiradas das duas subpopulações $N(\mu_1, \sigma_e^2)$ e $N(\mu_2, \sigma_e^2)$.

Queremos testar a hipótese

$$H_0: \mu_1 = \mu_2$$

contra a alternativa

$$H_1: \mu_1 \neq \mu_2.$$

Como já salientamos acima, esse teste pode ser conduzido com os métodos do Capítulo 13, mas o objetivo aqui é introduzir a metodologia da análise de variância, com um caso simples. A extensão para mais de dois níveis será estudada na seção 15.3.

Note que estamos supondo que as variâncias residuais dos níveis 1 e 2 são iguais, ou seja,

$$\text{Var}(e_{1j}) = \text{Var}(e_{2j}) = \sigma_e^2, \text{ para todo } j = 1, ..., n_i. \tag{15.7}$$

Essa é a propriedade conhecida como *homoscedasticidade*, isto é, estamos admitindo que a variabilidade residual é a mesma para os dois níveis (ou que P_1 e P_2 têm a mesma variabilidade segundo a v.a. Y). Note também que

$$E(y_{ij}) = \mu_i, \qquad \text{Var}(y_{ij}) = \text{Var}(e_{ij}) = \sigma^2 e. \tag{15.8}$$

15.2.2 Estimação do Modelo

Nosso objetivo é estimar μ_1, μ_2 e σ_e^2 no modelo (15.6), para podermos testar H_0. Usaremos estimadores de mínimos quadrados. Poderíamos usar também estimadores de máxima verossimilhança, pois sabemos que nossas observações têm distribuição normal. Temos que, de (15.6), os *resíduos* são dados por

$$e_{ij} = y_{ij} - \mu_i, \tag{15.9}$$

e a soma dos quadrados dos resíduos é dada por

$$SQ(\mu_1, \mu_2) = \sum_{i=1}^{2} \sum_{j=1}^{n_i} e_{ij}^2 = \sum_{i=1}^{2} \sum_{j=1}^{n_i} (y_{ij} - \mu_i)^2$$

$$= \sum_{j=1}^{n_1} (y_{1j} - \mu_1)^2 + \sum_{j=1}^{n_2} (y_{2j} - \mu_2)^2,$$

ou seja,

$$SQ(\mu_1, \mu_2) = \sum_{j=1}^{n_1} e_{1j}^2 + \sum_{j=1}^{n_2} e_{2j}^2 . \tag{15.10}$$

Observe que essa soma de quadrados é uma função de μ_1 e, μ_2 Se as variâncias residuais das duas subpopulações não fossem iguais, essa soma seria mais afetada por aquele nível que tivesse maior variância, e isso deveria influenciar a escolha dos estimadores. Nesse caso, uma sugestão seria então minimizarmos a expressão (15.10) com e_{ij}^2 substituída por $(e_{ij}/\sigma_i)^2$, com $\text{Var}(e_{ij}) = \sigma_i^2$, o que conduz a *estimadores de mínimos quadrados ponderados*.

Derivando (15.10) em relação a μ_1 e μ_2 obtemos:

$$\frac{\partial SQ(\mu_1, \mu_2)}{\partial \mu_i} = -2\sum_{j=1}^{n_i}(y_{ij} - \mu_i) = 0, \quad i = 1, 2,$$

do que segue que os estimadores são dados por

$$\hat{\mu}_1 = \frac{1}{n_1}\sum_{j=1}^{n_1} y_{1j} = \overline{y}_{1,}, \tag{15.11}$$

$$\hat{\mu}_2 = \frac{1}{n_2}\sum_{j=1}^{n2} y_{2j} = \overline{y}_{2,}, \tag{15.12}$$

que são as médias das observações dos níveis 1 e 2, respectivamente. Logo,

$$SQ(\hat{\mu}_1, \hat{\mu}_2) = \sum_{j=1}^{n_1}(y_{1j} - \overline{y}_1)^2 + \sum_{j=1}^{n_2}(y_{2j} - \overline{y}_2)^2. \tag{15.13}$$

Podemos pensar em (15.13) como a *quantidade total de informação quadrática perdida* pela adoção do modelo (15.6). Essa soma é também denominada *soma dos quadrados dos resíduos*.

Vejamos outra maneira de escrever essa soma. Dentro do grupo dos homens, a variância da subpopulação P_1 pode ser estimada por

$$S_1^2 = \frac{1}{n_1 - 1}\sum_{j=1}^{n_1}(y_{1j} - \overline{y}_1)^2, \tag{15.14}$$

e a variância da subpopulação P_2 das mulheres é estimada por

$$S_2^2 = \frac{1}{n_2 - 1}\sum_{j=1}^{n_2}(y_{2j} - \overline{y}_2)^2. \tag{15.15}$$

Segue-se que

$$SQ(\hat{\mu}_1, \hat{\mu}_2) = (n_1 - 1)S_1^2 + (n_2 - 1)S_2^2. \tag{15.16}$$

Temos, acima, dois estimadores não-viesados do mesmo parâmetro σ_e^2 e, portanto, podemos definir uma variância amostral ponderada

$$S_e^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}, \tag{15.17}$$

e, usando (15.16), podemos escrever

$$S_e^2 = \frac{\text{SQ}(\hat{\mu}_1, \hat{\mu}_2)}{n-2}, \tag{15.18}$$

se $n = n_1 + n_2$. Vemos que S_e^2 é a *quantidade média* de informação quadrática perdida e é um estimador não-viesado de σ_e^2. Observe que esse é o mesmo estimador definido em (13.10).

Temos, portanto, um primeiro enfoque para estimar a variância desconhecida, σ_e^2, por meio da *variância devida ao erro* ou *variância dentro de amostras,* dada por S_e^2, que é baseada nas *variâncias amostrais,* dadas por (15.14) e (15.15). A soma de quadrados (15.16) é também chamada de *soma de quadradros dentro dos grupos*.

Um outro enfoque será visto mais adiante, e que consiste em estimar σ_e^2, através de uma *variância entre amostras,* baseada na variabilidade *entre as médias amostrais,* também chamada *variação devida ao fator.*

Exemplo 15.1. (continuação) Para os dados da Tabela 15.1, temos:

Grupo dos Homens (nível l): $\bar{y}_1 = 110{,}1$, $\sum_{j=1}^{10}(y_{1j} - \bar{y}_1)^2 = 670{,}9$, $S_1^2 = 74{,}54$;

Grupo das Mulheres (nível 2): $\bar{y}_2 = 104{,}9$, $\sum_{j=1}^{10}(y_{2j} - \bar{y}_2)^2 = 566{,}9$, $S_2^2 = 62{,}99$.

Segue-se que

$$S_e^2 = \frac{670{,}9 + 566{,}9}{18} = \frac{1.237{,}8}{18} = 68{,}77, \quad S_e = 8{,}29.$$

Note que a soma dos quadrados dos resíduos é

$$SQ(\hat{\mu}_1, \hat{\mu}_2) = \text{SQ}(\bar{y}_1, \bar{y}_2) = 1.237{,}8.$$

Observe, também, que $\bar{y}_1$ e $\bar{y}_2$, denotam os tempos médios estimados de reação ao estímulo dos homens e mulheres, respectivamente.

Uma questão de interesse é a seguinte: será que o conhecimento do sexo de um indivíduo ajuda a melhorar a previsão do tempo de reação dele ao estímulo? Para responder a essa questão, devemos ter algum modelo alternativo para poder comparar os ganhos. O modelo usualmente adotado é o mais simples de todos, ou seja, aquele

que considera os dados vindos de uma única população. Suponha que os valores da v.a. Y para todos os $n = 20$ indivíduos sigam o modelo

$$y_i = \mu + e_i, \quad i = 1, 2, ..., 20. \tag{15.19}$$

Podemos considerar esse modelo como sendo para *uma população,* ou seja, aquela de todos os indivíduos para a qual queremos investigar o tempo de reação ao estímulo, independentemente do sexo, idade e outros fatores.

Para o modelo (15.19) a soma dos quadrados dos resíduos é

$$SQ(\mu) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \mu)^2, \tag{15.20}$$

e o estimador de mínimos quadrados de μ, *é* obtido derivando-se (15.20) com relação a μ e igualando a zero, chegando-se a

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} y_j = \bar{y}, \tag{15.21}$$

ou seja, a média de todas as observações. Como aqui $y_i \sim N(\mu, \sigma_e^2)$, um estimador da variância residual σ_e^2 é

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_1 - \bar{y})^2 = \frac{\mathrm{SQ}(\mu)}{n-1} \tag{15.22}$$

ou seja, a nossa conhecida variância amostral.

Para os dados da Tabela 15.1, encontramos

$$\bar{y} = \frac{2.150}{20} = 107{,}50,$$

$$S^2 = \frac{1.373}{19} = 72{,}26, \quad S = 8{,}5.$$

Assim, sem informação adicional, podemos prever o tempo de reação de um indivíduo como sendo 107,50, com um desvio padrão de 8,5. Os resíduos desse modelo e do modelo (15.6) estão na Tabela 15.4, colunas *e*(l) e *e(2),* respectivamente. Comparando esses resíduos, vemos que os segundos melhoram um pouco as previsões, isto é, fazem cair o erro quadrático médio de 8,5 para 8,29. Mas essa queda nos parece pequena para justificar a inclusão do fator *sexo* no modelo, e talvez fosse preferível adotar o modelo mais simples (15.19).

Tabela 15.4: Resíduos para vários modelos ajustados aos dados do Exemplo 15.1.

Variáveis				Resíduos dos Modelos		
				e(1)	e(2)	e(3)
Indivíduo	Tempo de Reação	Sexo	Idade	$y_i - \overline{y}$	$y_{ij} - \overline{y}_i$	$y_{ij} - \overline{y}_i$
1	96	H	20	–11,50	–14,1	–2,50
2	92	M	20	–15,50	–12,9	–6,50
3	106	H	20	–1,50	–4,1	7,50
4	100	M	20	–7,50	–4,9	1,50
5	98	M	25	–9,50	–6,9	–5,25
6	104	H	25	–3,50	–6,1	0,75
7	110	H	25	2,50	–0,1	6,75
8	101	M	25	–6,50	–3,9	–2,25
9	116	M	30	8,50	11,1	8,25
10	106	H	30	–1,50	–4,1	–1,75
11	109	H	30	1,50	–1,1	1,25
12	100	M	30	–7,50	–4,9	–7,75
13	112	M	35	–4,50	7,1	1,25
14	105	M	35	–2.50	0,1	–5,75
15	118	H	35	10,50	7,9	7,25
16	108	H	35	0,50	–2,1	–2,75
17	113	M	40	5,50	8,1	–4,25
18	112	M	40	4,50	7,1	–5,25
19	127	H	40	19,50	16,9	9,75
20	117	H	40	9,50	6,9	–0,25
d.p.				8,50	8,29	6,08
2d.p.				17,00	16,58	12,16

Nota: Nesta tabela estão expressos os resíduos de diversos modelos ajustados aos dados e colocados juntos para comparar os "lucros" na adoção de cada modelo. No texto aparece o significado de cada coluna dos resíduos.

15.2.3 Intervalos de Confiança

Com as suposições feitas sobre os erros, podemos escrever

$$\overline{y}_1 \sim N(\mu_1, \sigma_e^2 / n_1),\ \overline{y}_2 \sim N(\mu_2, \sigma_e^2 / n_2), \tag{15.23}$$

o que permite construir intervalos de confiança separados para os dois parâmetros μ_1 e μ_2, como já vimos anteriormente. Esses têm a forma

$$\overline{y}_i \pm t_\gamma \frac{S_e}{\sqrt{n_i}}, \quad i = 1, 2, \tag{15.24}$$

onde t_γ é o valor crítico da distribuição t de Student com $v = n - 2$ graus de liberdade, tal que $P(-t_\gamma < t(n-2) < t_\gamma) = \gamma$, $0 < \gamma < 1$. Observe que o número de graus de liberdade é $(n - 2)$ e não $n_i - 1$, porque

$$Z_i = \frac{(\bar{y}_i - \mu_i)\sqrt{n_i}}{\sigma_e} \sim N(0,1),$$

$$W = \frac{(n-2)S_e^2}{\sigma_e^2} \sim \chi^2(n-2)$$

e, portanto, $\dfrac{Z_i}{\sqrt{W/(n-2)}} = \dfrac{\sqrt{n_i}(\bar{y}_i - \mu_i)}{S_e}$ tem distribuição $t(n - 2)$ pelo Teorema 7.1. Daqui, obtemos (15.24).

Exemplo 15.1. (continuação) Para o Exemplo 15.1, temos:

$$\text{IC}(\mu_1;\ 0{,}95) = 110{,}10 \pm (2{,}101)8{,}29/\sqrt{10} =]104{,}59;\ 115{,}61[,$$

$$\text{IC}(\mu_2;\ 0{,}95) = 104{,}90 \pm (2{,}101)8{,}29/\sqrt{10} =]99{,}39;\ 110{,}41[,$$

com $t_{0{,}95} = 2{,}101$ encontrado na Tabela V, com $v = 18$ graus de liberdade.

Ainda, com as suposições feitas, podemos concluir que

$$\bar{y}_1 - \bar{y}_2 \sim N(\mu_1 - \mu_2,\ \sigma_e^2/n_1 + \sigma_e^2/n_2), \tag{15.25}$$

de modo que a estatística

$$T = \frac{(\bar{y}_1 - \bar{y}_2) - (\mu_1 - \mu_2)}{S_e\sqrt{1/n_1 + 1/n_2)}} \tag{15.26}$$

tem distribuição t de Student com $v = n_1 + n_2 - 2 = n - 2$ graus de liberdade, e um intervalo de confiança para a diferença $\mu_1 - \mu_2$ pode ser construído.

Exemplo 15.1. (continuação) Para o exemplo,

$$\text{IC}(\mu_1 - \mu_2;\ 0{,}95) = (\bar{y}_1 - \bar{y}_2) \pm t_y S_e\sqrt{1/n_1 + 1/n_2}$$

$$= (110{,}1 - 104{,}9) \pm (2{,}101)(8{,}29)\sqrt{1/10 + 1/10} =]-2{,}59;\ 12{,}99[.$$

Este resultado implica que a hipótese

$$H_0 : \mu_1 = \mu_2 \tag{15.27}$$

não pode ser rejeitada no nível $\alpha = 0{,}05$, já que o zero pertence ao intervalo. Isso está de acordo com o resultado já apontado de que o conhecimento do sexo de um indivíduo não irá ajudar a prever o tempo de reação ao estímulo.

O teste da hipótese para (15.27), com as suposições adotadas, é feito usando a estatística (15.26), com $n_1 + n_2 - 2$ g.l., obtendo-se o valor observado $t_0 = 1{,}40$, que, comparado com o valor crítico de 2,101($\alpha = 5\%$ e l8 g.l.), leva à não-rejeição da hipótese, como foi visto acima.

15.2.4 Tabela de Análise de Variância

As operações processadas anteriormente podem ser resumidas num quadro, para facilitar a análise. Se (15.27) for válida, o modelo adotado será

$$y_{ij} = \mu + e_{ij},$$

e a quantidade de informação perdida (devida aos resíduos) será dada por

$$SQ(\hat{\mu}) = \sum_{i=1}^{2}\sum_{j=1}^{n_i}(y_{ij} - \overline{y})^2, \tag{15.28}$$

que iremos chamar de *soma de quadrados total,* abreviadamente, SQTot.

Analogamente, adotado o modelo (15.4), a quantidade de informação perdida é dada por (15.13) ou (15.16), e que chamamos de *soma de quadrados dos resíduos,* abreviadamente, SQRes, ou *soma de quadrados dentro dos dois grupos*, abreviadamente, SQDen.

A *economia* obtida ao passarmos de um modelo para outro será

$$\text{SQTot} - \text{SQDen} = \text{SQEnt}, \tag{15.29}$$

que chamaremos de *soma de quadrados entre grupos*. Não é difícil provar que (veja o problema 18)

$$\text{SQEnt} = \sum_{i=1}^{2} n_i(\overline{y}_i - \overline{y})^2. \tag{15.30}$$

Observando essa expressão, vemos que ela representa a variabilidade *entre as médias amostrais,* ou seja, uma "distância" entre a média de cada grupo e a média global. Donde o nome "soma de quadrados entre grupos". Quanto mais diferentes forem as médias $\overline{y}_i$, $i = 1, 2$, maior será SQEnt e, conseqüentemente, menor será SQDen.

As quantidades

$$\text{QMTot} = \frac{\text{SQTot}}{n-1} \tag{15.31}$$

e

$$\text{QMDen} = \frac{\text{SQDen}}{n-2} \tag{15.32}$$

são chamadas *quadrado médio total* e *quadrado médio dentro* (ou residual), respectivamente.

Todas essas informações são agrupadas numa única tabela, conhecida pelo nome de ANOVA (abreviação de ANalysis Of VAriance), descrita na Tabela 15.5.

Tabela 15.5: Tabela de Análise de Variância (ANOVA).

F.V.	g.l.	SQ	QM	F
Entre	1	SQEnt	QMEnt	QMEnt/S,^{2}
Dentro	$n-2$	SQDen	QMDen (ou S_e^2)	
Total	$n-1$	SQTot	QMTot (ou S^2)	

Na primeira coluna temos as descrições das diferentes somas de quadrados, tecnicamente indicadas por fontes de variação (F.V.). Os graus de liberdade (g.l.) da segunda coluna estão associados às respectivas somas de quadrados, sendo que o número de g.l. da SQE é obtido por subtração. Falaremos abaixo sobre QMEnt e a razão F = QMEnt/QMDen.

Exemplo 15.1. (continuação) Com os dados obtidos anteriormente para o Exemplo 15.1, podemos construir a tabela ANOVA para o modelo (15.4). O resultado está na Tabela 15.6.

Tabela 15.6: Tabela ANOVA para o Exemplo 15.1.

F.V.	g.l.	SQ	QM	F
Entre	1	135,20	135,20	1,97
Dentro	18	1.237,80	68,77	
Total	19	1.373,00	72,26	

Da ANOVA encontramos os desvios padrões residuais $S_e = \sqrt{68{,}77} = 8{,}29$ do "modelo completo" (15.4) e $S = \sqrt{72{,}26} = 8{,}50$, do "modelo reduzido" (15.19). A economia propiciada ao passar de um modelo para outro, em termos de soma de quadrados, é 135,20, e em termos de quadrados médios, comparando 72,26 e 68,77. Proporcionalmente, economizamos

$$\frac{135{,}20}{1.373{,}00} = 0{,}0985 \approx 9{,}85\%,$$

ou seja, aproximadamente 10% na SQ de resíduos. Podemos dizer que essa é *a proporção da variação explicada pelo modelo* (15.9). Essa medida é chamada *coeficiente de explicação* do modelo, denotada por

$$R^2 = \frac{\text{SQEnt}}{\text{SQTot}}. \tag{15.33}$$

Essa medida já foi usada na seção 4.6. Veja o problema 27.

A conveniência ou não do modelo (15.4) está associada ao teste (15.27), já que aceitar essa hipótese implica a adoção do modelo (15.19). Com as suposições feitas, a estatística para o teste é (15.26), que, sob H_0 fica

$$T = \frac{\bar{y}_1 - \bar{y}_2}{S_e\sqrt{1/n_1 + 1/n_2}}, \tag{15.34}$$

que tem distribuição $t(n_1 + n_2 - 2)$. Também sabemos que o quadrado de T tem distribuição $F(1, n_1 + n_2 - 2)$ (ver seção 13.3). Contudo,

$$\text{QMEnt} = \text{SQEnt} = n_1(\bar{y}_1 - \bar{y})^2 + n_2(\bar{y}_2 - \bar{y})^2,$$

e como

$$\bar{y} = \frac{n_1\bar{y}_1 + n_2\bar{y}_2}{n_1 + n_2},$$

podemos escrever

$$\text{QMEnt} = \frac{n_1 n_2}{n_1 + n_2}(\bar{y}_1 - \bar{y}_2)^2 = \frac{(\bar{y}_1 - \bar{y}_2)^2}{1/n_1 + 1/n_2}. \tag{15.35}$$

Logo, concluímos que

$$T^2 = \frac{(\bar{y}_1 - \bar{y}_2)^2}{S_e^2(1/n_1 + 1/n_2)} = \frac{\text{QMEnt}}{S_e^2} = F. \tag{15.36}$$

Essa é a estatística que aparece na última coluna da tabela ANOVA. Portanto, podemos usar F, com $(1, n - 2)$ graus de liberdade para testar a hipótese (15.27). Rejeitaremos H_0 se $F > c$, c determinado pelo nível de significância do teste.

Exemplo 15.4. Da ANOVA da Tabela 15.6, vemos que o valor da estatística F é 1,97. Consultando a Tabela VI, com (1,18) g.l. e α = 0,05, encontramos o valor crítico 4,41. Logo, não rejeitamos H_0: $\mu_1 = \mu_2$. Isso significa que não há vantagem em usar o modelo (15.4) no lugar de (15.19).

Problemas

3. Na tabela abaixo estão os dados referentes a uma amostra de 21 alunos do primeiro ano de um curso universitário. As variáveis são:

   Y: nota obtida na primeira prova do curso;

   X: se cursou escola particular (P) ou oficial (O);

   Z: o período em que está matriculado: manhã (M), tarde (T), noite (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	T	N	M	M	T	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	T	T	M	M	N	T	T	M	M	T	

   Considere o modelo $y_i = \mu + e_i$, $i = 1, 2, ..., 21$, $e_i \sim N(0, \sigma^2)$. Obtenha os erros quadráticos médios de $\hat{\mu}$ e $\hat{\sigma}^2$. Construa intervalos de confiança para μ e σ^2, com coeficiente de confiança 95%. Analise os resíduos do modelo.

4. Usando os dados do problema 3, você diria que o fato de a pessoa ter cursado a escola particular ou oficial influi no resultado da primeira prova? Siga todos os passos do Exemplo 15.1 para responder a essa pergunta.

5. Usando os dados do Exemplo 15.2, você diria que o fato de estudar durante o dia ou à noite afeta o desempenho dos alunos?

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

Escolaridade	n	Σx_i	Σx_i^2
Fundamental	50	111,50	259,93
Médio	20	71,00	258,89

   [Observação: rendimentos (x) expressos como porcentagem do salário mínimo.]

7. Obtenha a tabela ANOVA para o Exemplo 15.3, usando o fator regime com os níveis I e 2.

15.3 Modelo para Mais de Duas Subpopulações

Para ilustrar essa situação, vamos considerar o fator idade para o Exemplo 15.1. Consideremos o modelo

$$y_{ij} = \mu_i + e_{ij}, \tag{15.37}$$