Inferência para Várias Populações

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.

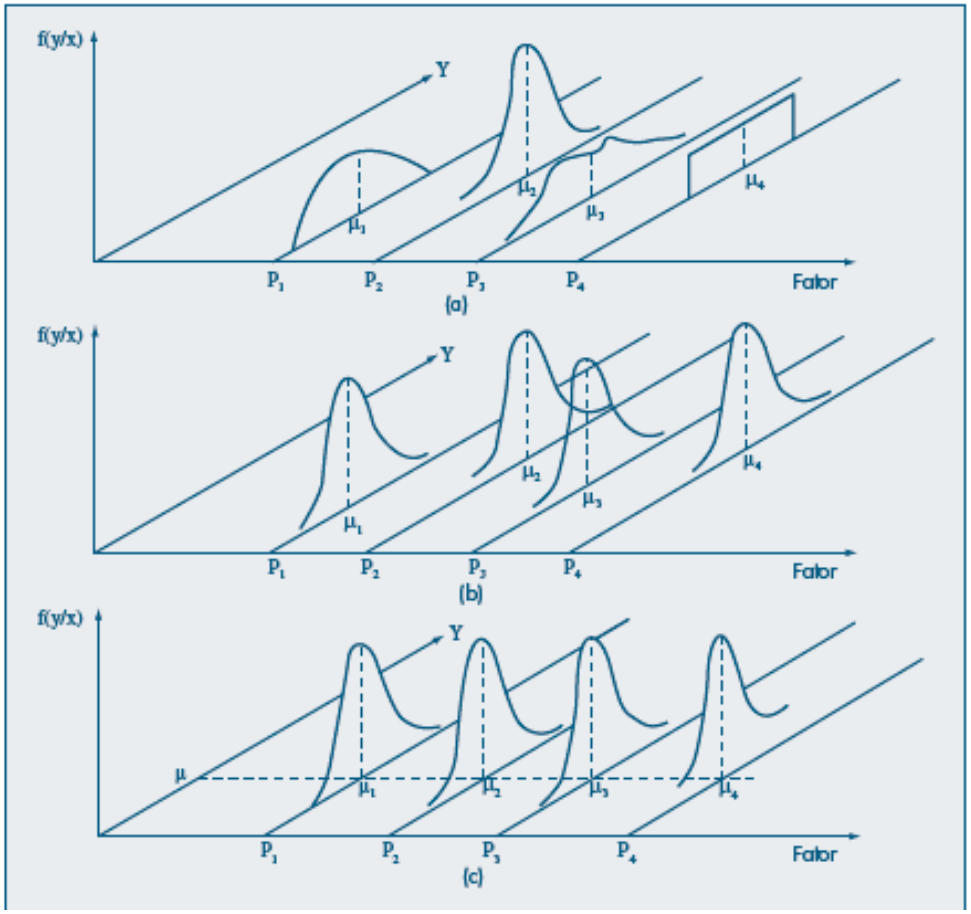
Neste capítulo iremos investigar um modelo simples, chamado de *análise de variância com um fator.*

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, ..., *PI,* 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.

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

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

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

Exemplo 15. 1. 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

Exemplo 15. 1. 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

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 $i = 1, 2, ..., 5$, $j = 1, 2, 3, 4$.

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

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 $n_1 = n_2 = n_3 = 6$.

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, ..., 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

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