Capítulo 14

Análise de variância

'However, perhaps the main point is that you are under no obligation to analyse variance into its parts if it does not come apart easily, and its unwillingness to do so naturally indicates that one's line of approach is not very fruitful.'

– Fisher, 1933

Frequentemente conduzimos experimentos para provar hipóteses científicas. Admitindo que estes sejam delineados de forma adequada e regidos pelos princípios básicos da experimentação (repetição, casualização, controle local), a variação total dos dados pode ser decomposta em partes conhecidas, devidas aos fatores estudados, e em parte desconhecida, o erro experimental. Essa técnica de decomposição é denominada *análise de variância* (ANOVA), e está associada ao teste F para as fontes de variação conhecidas, tais como tratamentos, interação etc.

Veremos neste capítulo, como realizar análise de variância de dados experimentais provenientes do delineamento inteiramente casualizado (DIC) e de blocos casualizados (DBC), ambos envolvendo apenas um fator de tratamento.

14.1 *Nuts and bolts*

Antes de formalizar o conceito de ANOVA, dado o protagonismo desta técnica em análises estatísticas de dados experimentais, faremos uma introdução mais **intuitiva**, por meio de um exemplo

simples, para que o leitor iniciante nesse tipo de procedimento possa entender bem como funciona.

Considere uma amostra com $n = 20$ dados[1] de produção (kg) de grãos de milho. Chamaremos produção de variável resposta y. Cada valor de y foi obtido de uma parcela de 4 m^2 num experimento de campo, isto é, há 20 parcelas no total.

```
> y <- c(25, 26, 20, 23, 21, 31, 25, 28, 27, 24, 22, 26,
    28, 25, 29, 33, 29, 31, 34, 28)
```

Ocorre que estas parcelas foram cultivadas com quatro materiais genéticos (variedades) distintos de milho (A, B, C e D). Cada material, que aqui chamaremos de tratamento, foi cultivado em 5 parcelas, isto é, foi *repetido* cinco vezes.

```
> variedade <- gl(4, 5, labels = LETTERS[1:4])
> variedade
 [1] A A A A A B B B B B C C C C C D D D D D
Levels: A B C D
```

Podemos dizer ainda que, por exemplo, as cinco repetições do tratamento A representam uma amostra 'tirada' da população de valores de A. Estamos interessados em saber se há diferenças entre os tratamentos, em nível de populações. Como não temos os dados das populações, faremos conclusões sobre elas com base nos dados das amostras. Chamamos isso de *inferência estatística*.

Aprendemos em capítulos anteriores a mensurar a variabilidade de uma amostra. Isso pode ser feito, por exemplo, por meio da variância amostral, que basicamente mede o quanto os valores y se desviam da média $\bar{y} = 26.75$ kg. A quantidade

$$SQ_y = \sum (y - \bar{y})^2$$

chamada de soma de quadrados, reflete exatamente essa variação total em y. No exemplo, $SQ_y = 275.75$. A unidade de medida não nos convém. A parte superior da Figura 14.1 mostra a localização de cada valor de y e, abaixo, mostra separadamente os cinco valores de cada tratamento.

Supostamente, as vinte parcelas onde foi instalado o experimento são homogêneas. Apesar disso, observe que ainda assim há variação nos dados de produção entre as parcelas que receberam um mesmo tratamento. Por exemplo, a variação (SQ) dentro no

[1] Extraídos de Vieira & Hoffmann (1999).

tratamento A é de 26; no tratamento B é de 30; e assim por diante. A soma dessas variações internas é de 112. Observe ainda que esses valores de variação interna são semelhantes. E isso é uma das exigências para se realizar a ANOVA. Chamamos isto de *homocedasticidade*.

A variação *entre* um tratamento e outro é de fato esperada, pois estamos *promovendo* isso quando plantamos variedades diferentes. A variação *dentro* de um tratamento também é esperada, embora esperamos que seja baixa, e é causada pelo acaso, ou melhor, por uma soma de fatores que nos são desconhecidos, como variações na fertilidade ou umidade do solo, ou seja, fatores que não foram controlados. Em estatística chamamos essa variação de *resíduo* ou *erro aleatório*.

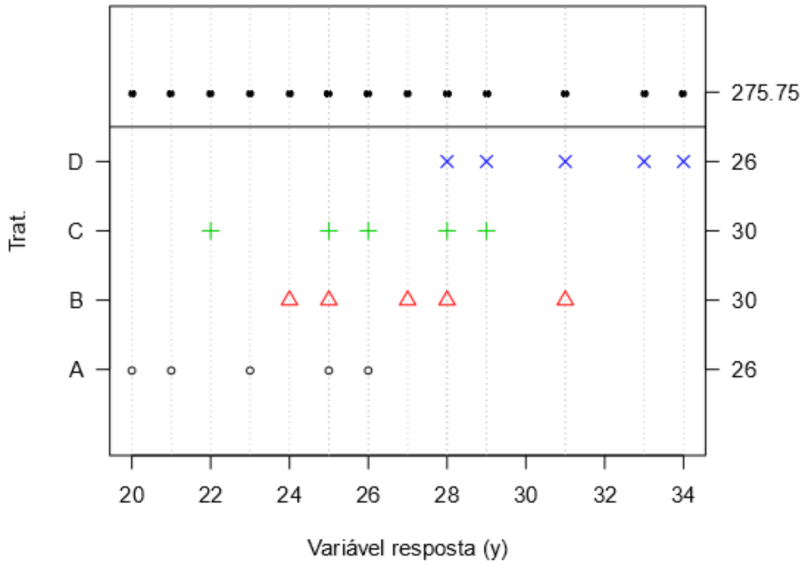

Figura 14.1: Dados da variável resposta y nos tratamentos A a D. No eixo lateral direito, a variabilidade (soma de quadrados) de cada tratamento.

Então faça as contas: se a variação total dos dados é de 275.75 e desta variação uma parte equivalente a 112 é devida ao 'acaso', a variação restante (163.75) só pode ter sido ocasionada pelos tratamentos.

Agora indague-se com o seguinte: se a variação ocasionada pelos tratamentos fosse igual a variação ao acaso, o que poderíamos inferir? A resposta é simples, diríamos que não há diferenças entre os tratamentos. Seria como se as vinte parcelas estivessem sido

cultivadas com uma só variedade de milho.

Mas se, por outro lado, a variação ocasionada pelos tratamentos for relativamente maior que a variação ao acaso, então poderíamos dizer que a produção é realmente diferente entre uma variedade e outra. Pois bem, esse *relativamente maior* quem cuida de nos dizer é a tabela ou distribuição de probabilidade da estatística F, cujos valores são tabelados sob a hipótese de variação equivalente de tratamentos e do acaso. Quando isso ocorre, o valor de F é próximo de 1, pois esta estatística consiste da razão entre as variâncias de tratamento e do acaso.

Perceba que falamos agora em *variância*, cujo símbolo é s^2, que consiste simplesmente de SQ/GL, sendo GL os graus de liberdade. Grau de liberdade, de forma simplista, é o número de *itens* menos 1. Por exemplo, o GL total de y é $20 - 1 = 19$. O GL de tratamentos é $4 - 1 = 3$. Ora, se há 19 para o total e 3 para os tratamentos, restam 16 para o acaso. E assim $s^2_{acaso} = 112/16 = 7$. E $s^2_{trat} = 163.75/3 = 54.583$. Dividindo

$$\frac{s^2_{trat}}{s^2_{acaso}} = \frac{54.583}{7} = 7.79$$

temos exatamente o valor da estatística F da ANOVA. Esse valor pode ser entendido da seguinte forma: a variância causada pelos tratamentos nos dados y é 7.79 vezes maior que a variância devida ao acaso. –*Isso é muito?* Pergunte a tabela ou a distribuição de F. Esta nos diz que, quando não há efeito de tratamentos sobre os dados, o valor esperado em 95% das vezes é de até 3.23. O que nos permite concluir que o nosso valor (7.79) reflete então diferenças significativas.

```
# valor tabelado de F
> qf(0.95, df1 = 3, df2 = 16)
[1] 3.238872
```

De fato, veja na Figura 14.2 o quão *estranho* é o valor obtido em relação aos valores esperados de F nesse exemplo. Note que apenas $100 \times 0.0019 = 0.19\%$ dos valores de F superam o valor calculado. Esse *percentual* que define a *estranheza* do valor calculado de F é chamado de *p-valor*.

```
# p-valor
> 1 - pf(q = 7.79, df1 = 3, df2 = 16)
[1] 0.001984461
```

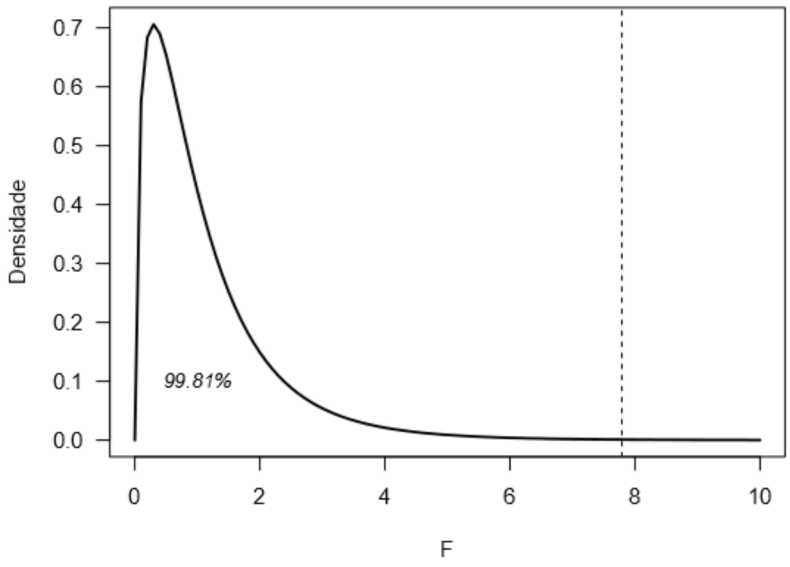

Figura 14.2: Distribuição de probabilidades da variável F com 3 (no numerador) e 16 (no denominador) graus de liberdade. A linha tracejada indica o valor de F calculado.

14.2 *One-way ANOVA*

Nesta seção, uma apresentação um pouco mais formal da ANOVA para **um fator** ou tipo de tratamento. Para tal, suponha estudar um fator com I níveis ($i = 1, 2, ..., I$), cada um repetido r_i vezes ($j = 1, 2, ..., r_i$). O modelo de ANOVA associado é:

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

em que y_{ij} é a observação tomada na j-ésima repetição do i-ésimo nível do fator; μ representa o intercepto ou a média geral de y; τ_i é o efeito do i-ésimo nível do fator; ϵ_{ij} é o erro aleatório associado a observação y_{ij}, suposto ter distribuição normal($\mu = 0$, σ^2). Perceba a semelhança com o modelo de regressão linear $y_i = b_0 + b_1 x_i + \epsilon_i$. A diferença é que aqui a variável preditora X não é quantitativa, mas sim categórica.

O modelo de ANOVA apresentado é conhecido como *modelo de efeitos*, que pode ser reparametrizado para o *modelo de médias*, em que os termos τ_i e μ são somados de modo a representar as médias μ_i do fator: