Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Estatística

Probabilidade e Estatística 2
1^o semestre de 2020
Gustavo L. Gilardoni

Análise de Variância a um fator (one-way)

30 de Setembro de 2020

0 Introdução

No início do semestre temos discutido primeiro como fazer inferências para a média de uma população ou tratamento e como comparar as médias de duas populações ou dois tratamentos. Nesse último caso o experimento foi desenhado de forma que as duas amostras resultassem independentes, seja porque elas eram escolhidas uma de cada população, ou porque os tratamentos eram alocados aleatoriamente às unidades experimentais. Está Unidade vai tratar do problema de como comparar as médias de $I > 2$ populações ou tratamentos.

Parece curioso chamar de "Análise de Variância" uma ferramenta para comparar médias. O motivo pode ser visto a partir do seguinte exemplo.

Exemplo 1. A Figura 1 (a) mostra um histograma para uma amostra $x_1, \ldots, x_{100}$ de 100 gerada da Distribuição Normal Padrão (o código **R** está num apêndice ao final das notas). O desvio padrão amostral foi $s_x = \sqrt{(n-1)^{-1} \sum_{i=1}^{100} (x_i - \bar{x})^2} \doteq 0.943$. A Figura 1 (b) mostra o histograma de uma amostra $y_1, \ldots, y_{100}$ obtida da seguinte forma: Dos primeiros 50 valores de x_i foi subtraído o valor 1, dos últimos 50 foi adicionado 1. Parece claro do segundo histograma que agora temos uma dispersão consideravelmente maior. De fato, o desvio padrão da amostra modificada dessa forma é $s_y \doteq 1.485$ [o histograma 1 (b) pode ser pensado como a superposição dos histogramas das primeiras 50 observações, centradas em -1 e o das últimas 50, centradas em $+1$, mostrados nas Figuras 1 (c) e (d)]. Dessa forma, comparando s_y e s_x podemos ter idéia da diferença entre as médias das duas subamostras. Mais precisamente, denote por $\bar{y}_1$ e $\bar{y}_2$ respectivamente as médias dos primeiros e dos últimos 50 valores de y_i (note que, da forma que os y_i foram obtidos, esperaríamos que $\bar{y}_1 \approx -1$ e $\bar{y}_2 \approx +1$), e por $s_{y,1}$ e $s_{y,2}$ os desvios padrões dos primeiros e dos últimos 50 y_i (de forma semelhante, esperamos que tanto $s_{y,1}$ quanto $s_{y,2}$ sejam próximos de 1, o desvio padrão da distribuição Normal que gerou as observações). Logo

$$
\begin{aligned}
(99)\, s_y^2 = \sum_{i=1}^{100} (y_i - \bar{y})^2 &= \sum_{i=1}^{50} (y_i - \bar{y}_1 + \bar{y}_1 - \bar{y})^2 + \sum_{i=51}^{100} (y_i - \bar{y}_2 + \bar{y}_2 - \bar{y})^2 \\
&= \sum_{i=1}^{50} (y_i - \bar{y}_1)^2 + 2\,(\bar{y}_1 - \bar{y}) \sum_{i=1}^{50} (y_i - \bar{y}_1) + \sum_{i=1}^{50} (\bar{y}_1 - \bar{y})^2 \\
&\quad + \sum_{i=51}^{100} (y_i - \bar{y}_2)^2 + 2\,(\bar{y}_2 - \bar{y}) \sum_{i=51}^{100} (y_i - \bar{y}_2) + \sum_{i=51}^{100} (\bar{y}_2 - \bar{y})^2 \\
= (49)\, s_{y,1}^2 + 2\,(\bar{y}_1 - \bar{y})\,(0) + (50)\,(\bar{y}_1 - \bar{y})^2 &+ (49)\, s_{y,2}^2 + 2\,(\bar{y}_2 - \bar{y})\,(0) + (50)\,(\bar{y}_2 - \bar{y})^2 \\
&= (49)\,(s_{y,1}^2 + s_{y,2}^2) + (50)\,[(\bar{y}_1 - \bar{y})^2 + (\bar{y}_2 - \bar{y})^2]\,. \quad (1)
\end{aligned}
$$

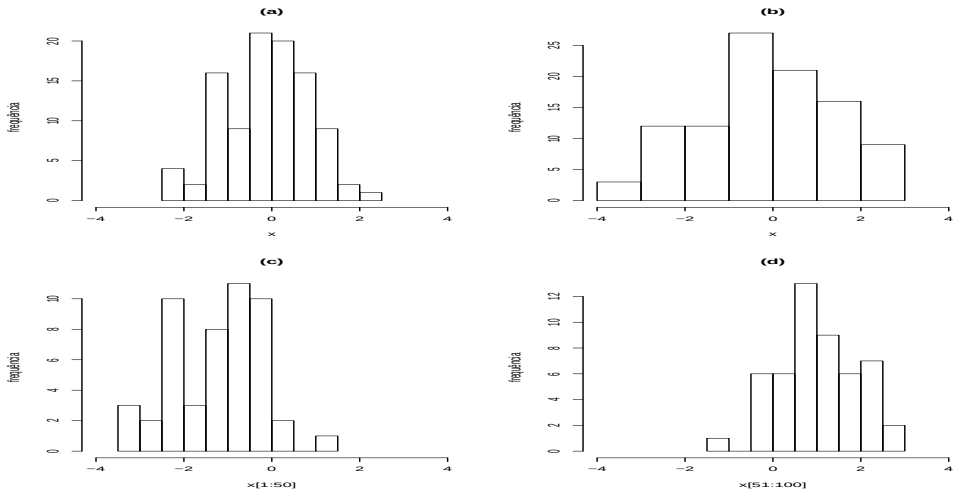

Figura 1: Histogramas para o Exemplo 1 1.

Fertilizante	**Produção (kg)**						**Média**	**DP**	**DP²**
1	10.5	9.7	5.5	11.0	9.3	6.3	8.72	2.28	5.18
2	7.1	10.2	7.2	7.8	10.9	7.3	8.42	1.68	2.84
3	4.7	7.5	6.5	7.7	6.5	6.9	6.63	1.07	1.15
4	8.5	12.4	11.5	7.1	10.6	11.8	10.32	2.08	4.33
						Média:	8.52		3.37

Tabela 1: Colheita em kg para 24 mudas de tomate usando 4 fertilizantes

Nessa decomposição da quantidade (99) s_y^2, o termo (49) $(s_{y,1}^2 + s_{y,2}^2)$ mede uma variabilidade das observações independente da sua média e, de fato, esperamos que seja parecido a (100) s_x^2. O termo restante, (50) $[(\bar{y}_1 - \bar{y})^2 + (\bar{y}_2 - \bar{y})^2]$ mede uma variabilidade entre as médias das duas subamostras.

Para a amostra do exemplo, temos que $\bar{x} = \bar{y} = -0.107$, $\bar{y}_1 \doteq -1.258$, $\bar{y}_2 \doteq 1.043$, $s_1^2 \doteq 0.990$ e $s_2^2 = 0.878$ (veja o cálculo no apêndice). Use esses valores para verificar e equação (1).

O exemplo anterior é artificial. O seguinte é real.

Exemplo 2. A Tabela 1 mostra o resultado de um experimento realizado para estudar o efeito de 4 fertilizantes no cultivo de tomates. Um total de 24 mudas foram alocadas aleatoriamente 6 para cada fertilizante e ao final do experimento o peso da colheita de cada planta foi registrado. O objetivo do estudo é saber se existe algum fertilizante que garante, em média, uma maior colheita.

A Figura 2 mostra o gráfico de pontos (*dotplots*) e de caixa (*boxplots*) para a produção das plantas agrupadas pelos 4 tratamentos ou tipos de fertilizante. Tanto da tabela quanto da figura, perceve-se que as plantas que receberam o fertilizante 4 parecem ter uma produção um pouco maior, enquanto as do fertilizante 3 um pouco menor. A questão

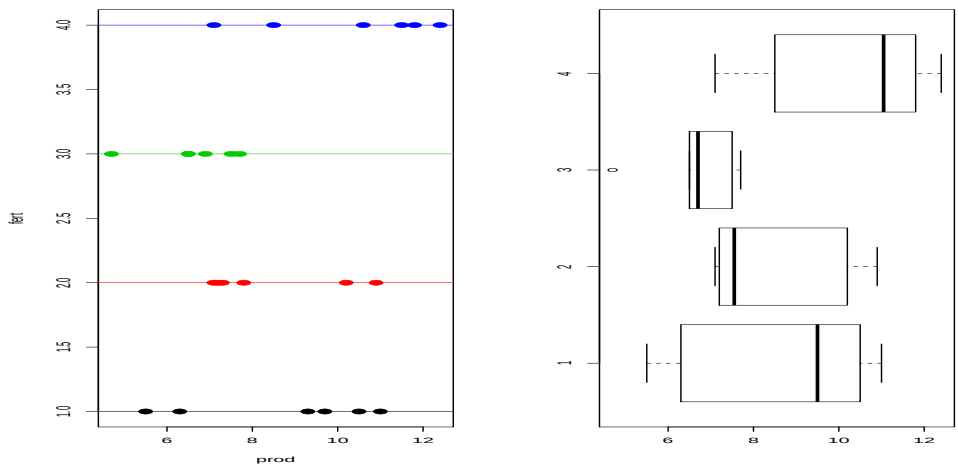

Figura 2: Gráfico de pontos (dotplots) e de caixa (boxplots) para os dados da Tabela 1.

fundamental é se essas diferenças poderiam ser explicadas somente pela aleatoriedade das observações, sem considerar algum tipo de efeito sistemático na produção média devido ao tipo de fertilizante. Para entender melhor essa idéia, a Figura 3 mostra gráficos de pontos e de caixa para 5 amostras geradas aleatoriamente da Distribuição Normal com média $\mu = 8.52$ e desvio padrão $\sigma = 1.83 \doteq \sqrt{3.37}$. Nessas amostras não existe nenhum tipo de efeito sistemático dos fertilizantes e, ainda assim, para algum dos gráficos, por exemplo o boxplot da terceira amostra), pareceria ter algum tratamento com média consideravelmente maior do que a dos outros. □

Usualmente, para analizar dados como os da Tabela 1, testa-se primeiro se pelo menos duas das médias devidas aos tratamentos (fertilizantes) são diferentes. Em outras palavras, considera-se a $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ contra a alternativa $H_a : \mu_i \neq \mu_k$ para pelo menos um par $i \neq k$. Esse problema é abordado na Seção 3. Quando a H_0 desse teste é rejeitada, surge naturalmente a questão de comparar os efeitos dos diferentes tratamentos, para decidir por exemplo se um deles é significativamente melhor que os outros. Esse tipo de problema é conhecido pelo nome de *Comparações múltiplas* e será discutido na Seção 4. A Seção 1 apresenta o modelo e a notação geral. Finalmente, o apêndice contem um *script* na linguagem **R** usado para os cálculos nos exemplos apresentados.

1 Notação e modelo

Consideramos $I > 2$ *tratamentos* ou *populações* com J observações em cada um (esse delineamento é dito *balanceado*; o caso que os números de observações para cada tratamento são diferentes será discutido brevemente mais na frente). A j-ésima observação do i-ésimo tratamento será denotada por $Y_{i,j}$ ou $y_{i,j}$. Na Tabela 1 temos, por exem-

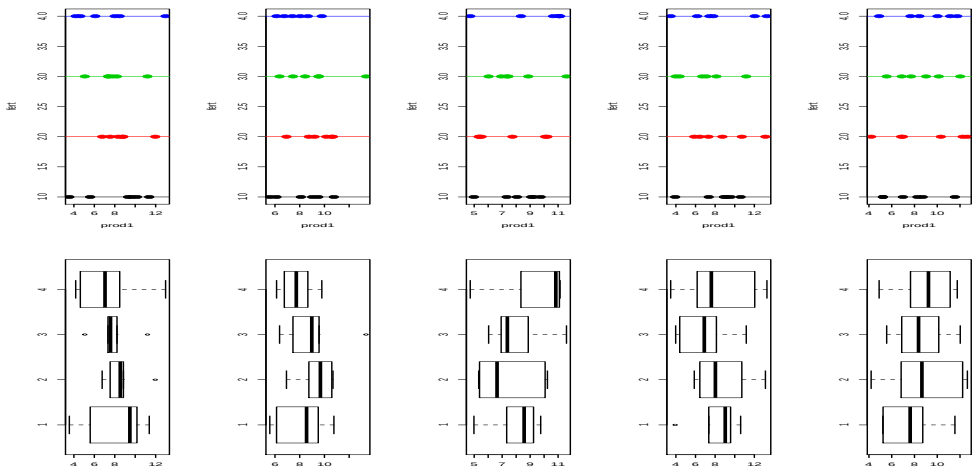

Figura 3: Gráfico de pontos (dotplots) e de caixa (boxplots) para 5 amostras aleatórias de tamanho 24 da distribuição Normal com μ e σ iguais a média dos 24 valores da da Tabela 1.

plo, que $y_{2,3} = 7.2$ e $y_{3,2} = 7.5$. O modelo específica que, para cada tratamento i, $Y_{i,1}, Y_{i,2}, \ldots, Y_{i,J}$ é uma amostra da distribuição Normal com média μ_i e desvio padrão σ, assumido igual para todos os i tratamentos. Dessa forma, as médias μ_i $(i = 1, \ldots, I)$ descrevem as médias dos tratamentos e serão o objeto do nosso estudo. Alternativamente podemos escrever

$$Y_{i,j} = \mu_i + \epsilon_{i,j}\,, \tag{2}$$

onde os *erros aleatórios* $\epsilon_{i,j}$ formam uma amostra de tamanho $(n\,I)$ da distribuição Normal com média $\mu = 0$ e desvio padrão σ. Algumas vezes a equação (2) é escrita da forma alternativa

$$Y_{i,j} = \mu + \alpha_i + \epsilon_{i,j}\,, \tag{3}$$

onde μ é uma média geral e os parâmetros α_i descrevem os efeitos dos tratamentos. Porém, veja que essa versão do modelo permanece idêntica se somarmos uma constante c a média geral μ e subtrairmos o mesmo valor de cada efeito α_i (i.é. $\mu+\alpha_i \equiv (\mu+c)+(\alpha_i-c)$ para todo c). Na literatura diz-se que os parâmetros $\mu, \alpha_1, \ldots, \alpha_I$ não são *identificáveis* e, por esse motivo, é usual acrescentar a restrição

$$\alpha_1 + \alpha_2 + \cdots + \alpha_I = 0 \tag{4}$$

na versão (3). Usualmente queremos testar a hipótese nula que $\mu_1 = \cdots = \mu_I$ contra a alternativa que $\mu_i \neq \mu_k$ para pelo menos um par (i, k).

Com base no modelo (2), um estimador não-viesado para a média μ_i será a correspondente média amostral,

$$\overline{Y}_{i,\cdot} = J^{-1} \sum_{j=1}^{J} Y_{i,j}\,,$$

enquanto a *média geral* para todos os I tratamentos é

$$\overline{Y}_{\cdot,\cdot} = (IJ)^{-1} \sum_{i=1}^{I} \sum_{j=1}^{J} Y_{i,j} = (I\,J)^{-1} \sum_{i=1}^{I} \overline{Y}_{i,\cdot}$$

(a notação anterior é muito conveniente: um "$\cdot$" no lugar de um subíndice significa calcular a média somando sobre ele e dividindo pelo correspondente número de unidades). Cada uma das I amostras ou tratamentos fornece também um estimador não-viesado para σ^2;

$$S_i^2 = (J-1)^{-1} \sum_{j=1}^{J} (Y_{i,j} - \overline{Y}_{i,\cdot})^2 \,.$$

Por exemplo, na Tabela 1 temos $I = 4$ tipos de fertilizante, $J = 6$ mudas foram observadas para cada tratamento, $\overline{Y}_{2,\cdot} \doteq 8.42$, $S_3^2 \doteq 1.15$, $\overline{Y}_{\cdot,\cdot} \doteq 8.52$ e assim por diante.

2 Cuidado! Por que não comparar as médias duas a duas?

Dado que da Unidade anterior sabemos como comparar *pares* de médias, podemos nos perguntar porque não comparar cada par de médias. Por exemplo, no exemplo 2 temos quatro tratamentos (fertilizantes) e poderíamos testar cada par de médias possíveis. Tem no total $\binom{4}{2} = 6$ pares para comparar (μ_1 com μ_2, μ_1 com μ_3, μ_1 com μ_4, μ_2 com μ_3, μ_2 com μ_4 e finalmente μ_3 com μ_4).

Para $i < j$, chame $H_{0;i,j}$ a hipótese nula que $\mu_i = \mu_j$. Suponha que para cada um dos seis pares (i, j) fazemos um teste t para testar $H_{0;i,j}$ ao nível de significância $100\,\alpha\%$, de forma que $P(\text{Rejeitar } H_{0;i,j} | \mu_i = \mu_j) = \alpha$. No final rejeitaríamos a $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ se, e somente se, rejeitamos pelo menos uma das $H_{0;i,j}$. É claro que gostaríamos de saber qual é o nível de significância do teste *ônibus*, isto é, procedendo dessa forma, qual sería a probabilidade de rejeitar H_0 quando H_0 é verdadeira. Veja que

$$\begin{aligned} P(\text{Rejeitar } H_0 | H_0) &= P(\text{Rejeitar pelo menos uma das } H_{0;i,j} | H_0) \\ &= 1 - P(\text{Não rejeitar nenhuma das } H_{0;i,j} | H_0)\,. \end{aligned}$$

Se os seis testes fossem independentes, o resultado anterior seria $1 - (1-\alpha)^6$. Por exemplo, quando $\alpha = 0.10$, o nível de significância do teste ônibus sería $1 - (0.90)^6 \doteq 0.47$, muito maior que o nível 0.10 usado para cada um dos seis testes individuais. O problema é na verdade um pouco mais complicado, pois os seis testes não são independentes. Por exemplo, para testar $H_{0;1,2}$ usamos as amostras dos tratamentos 1 e 2, enquanto para testar $H_{0;1,3}$ usamos também a amostra do tratamento 1, o que vai causar dependência dos estatísticos usados para esses dois testes. O método global que será discutido na próxima Seção controla a probabilidade de erro de forma a garantir que $P(\text{Rejeitar } H_0 | H_0)$ seja efetivamente igual ao nível de significância especificado.

Somente por curiosidade, realizamos uma simulação para ter ideia de qual seria a taxa de rejeição numa situação semelhante ao do Exemplo 2. Para isso, simulamos

$M = 10,000$ vezes dados semelhantes aos da Tabela 1, com $I = 4$ tratamentos e $J = 6$ observações por tratamento. As observações foram todas simuladas de uma distribuição Normal com média $\mu = 8.52$ e desvio padrão $\sigma = (3.37)^{1/2}$, de forma que **todas as $I = 4$ médias dos tratamentos são iguais**. Depois, para cada instância da simulação, comparamos os $\binom{I}{2} = \binom{4}{2} = 6$ pares de médias e rejeitamos H_0 se pelo menos uma das seis diferenças foi significativa ao nível $\alpha = 0.10$ (usamos para isso o teste t com variâncias iguais). No final das $M = 10,000$ simulações verificamos que H_0 foi rejeitada 35% das vezes, consideravelmente mais do que o nível de significância 10% usado para as comparações individuais. O código da simulação na linguagem **R** está no apêndice.

3 A tabela ANOVA

Vamos fazer uma decomposição da variância semelhante à equação (1). Para isso definimos (i) a soma de quadrados total corrigida pela média:

$$SQ_{tot} = \sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{i,j} - \overline{Y}_{.,.})^2\,,$$

(ii) a soma de quadrados explicada pelos tratamentos

$$SQ_{trat} = J\sum_{i=1}^{I}(\overline{Y}_{i,.} - \overline{Y}_{.,.})^2$$

e (iii) a soma de quadrados residual ou dos erros

$$SQ_{erro} = \sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{i,j} - \overline{Y}_{i,.})^2 = (J-1)\sum_{i=1}^{I}S_i^2\,.$$

Essas somas recebem vários outros nomes. Por exemplo, a SQ_{trat} é algumas vezes chamada de soma de quadrados *explicada* ou *entre* os grupos, enquanto a SQ_{erro} é também chamada de soma de quadrados *residual* ou *dentro* dos grupos.

A identidade fundamental é $SQ_{tot} = SQ_{erro} + SQ_{trat}$. Para prová-la, veja que

$$\begin{aligned}
SQ_{tot} &= \sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{i,j} - \overline{Y}_{.,.})^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{i,j} - \overline{Y}_{i,.} + \overline{Y}_{i,.} - \overline{Y}_{.,.})^2 \\
&= \sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{i,j} - \overline{Y}_{i,.})^2 + \sum_{i=1}^{I}\sum_{j=1}^{J}(\overline{Y}_{i,.} - \overline{Y}_{.,.})^2 + 2\sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{i,j} - \overline{Y}_{i,.})\,(\overline{Y}_{i,.} - \overline{Y}_{.,.}) \\
&= SQ_{erro} + J\sum_{i=1}^{I}(\overline{Y}_{i,.} - \overline{Y}_{.,.})^2 + 2\sum_{i=1}^{I}(\overline{Y}_{i,.} - \overline{Y}_{.,.})\sum_{j=1}^{J}(Y_{i,j} - \overline{Y}_{i,.}) = SQ_{erro} + SQ_{trat}\,,
\end{aligned} \tag{5}$$

onde temos usado que $\sum_{j=1}^{J}(\overline{Y}_{i,\cdot}-\overline{Y}_{\cdot,\cdot})^2 = J\,(\overline{Y}_{i,\cdot}-\overline{Y}_{\cdot,\cdot})^2$ e $\sum_{j=1}^{J}(Y_{i,j}-\overline{Y}_{i,\cdot}) = J\,\overline{Y}_{i,\cdot} - J\,\overline{Y}_{i,\cdot} = 0$.

Veja que, independente da H_0 ser verdadeira ou falsa (i.é. das médias μ_i serem iguais ou diferentes),

$$\mathrm{E}(SQ_{erro}) = \mathrm{E}\left[\sum_{i=1}^{I}(J-1)\,S_i^2\right] = (J-1)\sum_{i=1}^{I}\mathrm{E}(S_i^2) = (J-1)\sum_{i=1}^{I}\sigma^2 = I\,(J-1)\,\sigma^2\,. \quad (6)$$

Na terminologia da análise de variância, a razão entre SQ_{erro} e $[I\,(J-1)]$ é chamada de quadrados médio do erro (QM_{erro}), isto é,

$$QM_{erro} = \frac{SQ_{erro}}{I\,(J-1)} = \frac{\sum_{i=1}^{I} S_i^2}{I}\,.$$

A equação (6) diz que QM_{erro} é um estimador não-viesado para σ^2, independente da H_0 ser verdadeira ou falsa.

Para avaliar $\mathrm{E}(SQ_{trat})$, veja do modelo (2) que $\overline{Y}_{i,\cdot} = \mu_i + \overline{\epsilon}_{i,\cdot}$ e $\overline{Y}_{\cdot,\cdot} = \mu_\cdot + \overline{\epsilon}_{\cdot,\cdot}$, onde $\mu_\cdot = I^{-1}\sum_{i=1}^{I}\mu_i \geq 0$, com igualdade se, e somente se, $\mu_1 = \mu_2 = \cdot = \mu_I$. Logo, como $\mathrm{E}(\overline{\epsilon}_{i,\cdot} - \overline{\epsilon}_{\cdot,\cdot}) = 0$, segue que

$$\begin{aligned}
\mathrm{E}[SQ_{trat}] &= \mathrm{E}\left[J\sum_{i=1}^{I}(\overline{\epsilon}_{i,\cdot} - \overline{\epsilon}_{\cdot,\cdot} + \mu_i - \mu_\cdot)^2\right] = J\sum_{i=1}^{I}\mathrm{E}\left[(\overline{\epsilon}_{i,\cdot} - \overline{\epsilon}_{\cdot,\cdot} + \mu_i - \mu_\cdot)^2\right] \\
&= J\left\{\sum_{i=1}^{I}\mathrm{E}(\overline{\epsilon}_{i,\cdot} - \overline{\epsilon}_{\cdot,\cdot})^2 + 2\sum_{i=1}^{I}(\mu_i - \mu_\cdot)\,\mathrm{E}(\overline{\epsilon}_{i,\cdot} - \overline{\epsilon}_{\cdot,\cdot}) + \sum_{i=1}^{I}(\mu_i - \mu_\cdot)^2\right\} \\
&= J\,\mathrm{E}\left[\sum_{i=1}^{I}(\overline{\epsilon}_{i,\cdot} - \overline{\epsilon}_{\cdot,\cdot})^2\right] + J\sum_{i=1}^{I}(\mu_i - \mu_\cdot)^2\,. \quad (7)
\end{aligned}$$

Finalmente, para avaliar o primeiro somando no termo mais à direita, note que $\overline{\epsilon}_{1,\cdot}, \ldots, \overline{\epsilon}_{I,\cdot} \overset{iid}{\sim} \mathrm{Normal}(0,\sigma^2/J)$ e que $\overline{\epsilon}_{\cdot,\cdot} = I^{-1}\sum_{i=1}^{I}\overline{\epsilon}_{i,\cdot}$ é a média "amostral" dos $\overline{\epsilon}_{i,\cdot}$, de forma que agora $(I-1)^{-1}\sum_{i=1}^{I}(\overline{\epsilon}_{i,\cdot} - \overline{\epsilon}_{\cdot,\cdot})^2$ é a variância amostral dos $\overline{\epsilon}_{i,\cdot}$, cujo valor esperado sabemos que é $\mathrm{Var}(\overline{\epsilon}_{i,\cdot}) = \sigma^2/J$. Juntando esse resultado com a equação (7), segue finalmente que

$$\begin{aligned}
\mathrm{E}[SQ_{trat}] &= J\,\mathrm{E}\left[\sum_{i=1}^{I}(\overline{\epsilon}_{i,\cdot} - \overline{\epsilon}_{\cdot,\cdot})^2\right] + J\sum_{i=1}^{I}(\mu_i - \mu_\cdot)^2 = J\,(I-1)\,\frac{\sigma^2}{J} + J\sum_{i=1}^{I}(\mu_i - \mu_\cdot)^2 \\
&= (I-1)\,\sigma^2 + J\sum_{i=1}^{I}(\mu_i - \mu_\cdot)^2 \begin{cases} = (I-1)\,\sigma^2 & \text{quando } H_0 \text{ é verdadeira} \\ > (I-1)\,\sigma^2 & \text{quando } H_0 \text{ é falsa} \end{cases}
\end{aligned}$$

Dessa forma, quando a H_0 é verdadeira, a razão entre SQ_{trat} e $(I-1)$, chamada de quadrados médios devidos aos tratamentos (QM_{trat}), também estima σ^2, mas quando H_0 é falsa, o seu valor esperado é maior do que σ^2.

Fonte de variação	**gl**	**Soma de quadrados**	**Quadrados médios**	F_{obs}
Tratamentos	$(I-1)$	$SQ_{trat} = J\sum_{i=1}^{I}(\overline{Y}_{i,\cdot} - \overline{Y}_{\cdot,\cdot})^2$	$QM_{trat} = \frac{SQ_{trat}}{I-1}$	$\frac{QM_{trat}}{QM_{erro}}$
Erro	$I\,(J-1)$	$SQ_{erro} = \sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{i,j} - \overline{Y}_{i,\cdot})^2$	$QM_{erro} = \frac{SQ_{erro}}{I\,(J-1)}$	
Total	$I\,J-1$	$SQ_{tot} = \sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{i,j} - \overline{Y}_{\cdot,\cdot})^2$		

Tabela 2: Tabela de Análise da Variância ou ANOVA. Note que $\mathrm{E}(QM_{erro}) = \sigma^2$, enquanto $\mathrm{E}(QM_{trat}) = \sigma^2 + (I-1)\sum_{i=1}^{I}(\mu_i - \bar{\mu})^2 \geq \sigma^2$, com igualdade se, e somente se, $\mu_1 = \mu_2 = \cdots = \mu_I$.

O argumento anterior sugere usar como estatístico do teste a quantidade

$$F = QM_{trat}/QM_{erro}\,,$$

no entendido que, sob H_0, deveríamos observar valores de F perto de 1, mas quando H_0 é falsa, esperaríamos valores de F *significativamente* maiores que 1. Para dar um significado preciso à palavra *significativamente*, precisamos calcular a distribuição do estatístico F sob H_0. Em cursos mais avançados de estatística mostra-se que

- SQ_{erro}/σ^2 segue uma distribuição χ^2 com $I\,(J-1)$ gl;
- SQ_{trat}/σ^2 segue, sob H_0, uma distribuição χ^2 com $(I-1)$ gl e
- Sob H_0, SQ_{erro} e SQ_{trat} são independentes.

Dessa forma, sabemos da definição da distribuição F de Snedecor que, sob H_0,

$$F = \frac{QM_{trat}}{QM_{erro}} = \frac{\frac{1}{\sigma^2}\,QM_{trat}/(I-1)}{\frac{1}{\sigma^2}\,QM_{erro}/[I(J-1)]}$$

é um quociente entre duas variáveis aleatórias independentes com distribuição χ^2 divididas pelos respectivos graus de liberdade e segue, portanto, uma distribuição $F_{(I-1);I(J-1)}$ e o nosso teste rejeita H_0 quando $F_{obs} > F_{1-\alpha;(I-1);I(J-1)}$.

Usualmente, a informação necessária para o cálculo do estatístico F é usualmente apresentada na Tabela 2, chamada *Tabela da Análise da Variância* ou ANOVA (por *Analisys of Variance*).

Exemplo 2. (Continuação). No exemplo da Tabela 1 temos $SQ_{erro} \doteq (6-1)\,[2.28^2 + 1.68^2 + 1.07^2 + 2.08^2] \doteq 67.48$, $SQ_{trat} \doteq 6\,[(8.72-8.52)^2 + (8.42-8.52)^2 + (6.63-8.52)^2 +$

$(10.32 - 8.52)^2] \doteq 41.02$ e portanto $SQ_{tot} \doteq 108.50$. Os quadrados médios são $QM_{trat} \doteq 41.02/3 \doteq 13.67$ e $QM_{erro} \doteq 67.48/20 \doteq 3.37$ e portanto $F_{obs} \doteq 13.67/3.37 \doteq 4.05$.

Como o percentil 95% da distribuição F com $4-1 = 3$ gl no numerador e $4(6-1) = 20$ no denominador é $F_{0.05;3,20} \doteq 3.10$, rejeitamos a H_0 que todas as médias são iguais. O p-valor desse teste é $P(F_{3,20} > 4.053) \doteq 0.02$ ou 2%.

É conveniente realizar os cálculos anteriores com alguma rotina computacional. Na linguagem **R**, usamos as funções `aov` (por "*Analisis of Variance*") e `summary` (aplicada ao resultado de `aov`, como em `summary(aov(prod fert))`) (novamente, verifique o código **R** no apêndice). A Tabela ANOVA produzida pelo **R** segue

```
> ajuste<-aov(colheita~fert)    # modelo tomate = (média do trat i) + erro
> summary(ajuste)           # resumo do ajuste (tabela anova)
            Df Sum Sq Mean Sq F value Pr(>F)
fert         3  41.02  13.674   4.053 0.0211 *
Residuals   20  67.48   3.374
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

□

4 Comparações múltiplas

Se o teste da Seção anterior resultou em rejeitar a H_0, possivelmente precisamos decidir quais pares das médias dos tratamentos são diferentes. Como vimos na Seção 2, não é correto usar os métodos para comparar as médias de duas populações discutidos na Unidade anterior, pelo menos sem realizar algum tipo de ajuste no nível de significância.

Existem muitos métodos estatísticos para abordar esse problema, que é usualmente denominado *comparações múltiplas* na literatura. Aqui vamos a apresentar tão somente o método de Tukey ou HSD (por *Honest Significant Differences*).

Suponha que queremos testar **somente** se $H_{0;i,j} : \mu_i = \mu_j$. Nesse caso, e lembrando que os dois tamanhos amostrais são iguais, usaríamos o teste t com variâncias iguais,

$$t_{obs} = \frac{\bar{y}_{j,\cdot} - \bar{y}_{i,\cdot}}{s_{c;i,j}\sqrt{\frac{1}{J} + \frac{1}{J}}} = \frac{\bar{y}_{j,\cdot} - \bar{y}_{i,\cdot}}{s_{c;i,j}\sqrt{2/J}}, \tag{8}$$

onde $s^2_{c;i,j} = (s^2_i + s^2_j)/2$ é o estimador combinado da variância. O fato de querer testar várias comparações ao mesmo tempo introduz duas modificações nesse estatístico. Primeiro, como vimos na Seção anterior, $QM_{erro} = SQ_{erro}/[I\,(J-1)]$ é um estimador para σ^2 que usa as amostras de todos os I tratamentos e ele será melhor do que usar $s^2_{c;i,j}$ baseado somente nas amostras dos tratamentos i e j. Segundo, como discutido na Seção 2, existe uma variabilidade inerente as médias $\bar{y}_{i,\cdot}$, ainda quando H_0 é verdadeira, que deve ser tomada em conta. Por isso, para realizar comparações múltiplas, ao invés da

expressão (8), usa-se

$$T = \frac{\max_i\{\bar{y}_{i,}.\} - \min_i\{\bar{y}_{i,}.\}}{\sqrt{QM_{erro}}\,\sqrt{2/J}}\,. \tag{9}$$

Sob a H_0 que todas as médias são iguais, T segue uma distribuição denominada *de alcance studentizado* (*Studentized Range Distribution* em inglês) com I médias e $I\,(J-1)$ graus de liberdade (referentes a o estimador QM_{erro} de σ^2. Na linguagem **R**, a função de distribuição acumulada dessa distribuição é acessada pela função `ptukey`, enquanto a sua inversa é `qtukey`. Denotaremos por $T_{\alpha;m;gl}$ o valor que deixa $100\,\alpha\%$ de área à sua direita na distribuição de Alcance Studentizado com m médias e gl graus de liberdade. Por exemplo, para obter que $T_{0.05;4;20} \doteq 3.958$, digitamos no **R** `qtukey(0.95,4,20)`

Observação. Existem diferentes tabulações da distribuição de Alcance Studentizado. Algumas usam o termo $\sqrt{2}$ no denominador e outras não. Para o cálculo correto dos intervalos de confiança a seguir, é importante saber qual está sendo usada.

Com base na distribuição de Alcance Studentizado, os intervalos de confiança de Tukey ou HSD para as diferenças $\mu_i - \mu_j$ com confiança simultanea de $100\,(1-\alpha)\%$ são dados por

$$\bar{y}_{j,}. - \bar{y}_{i,}. \pm T_{\alpha;I,I\,(J-1)}\,\sqrt{\frac{QM_{erro}}{J}}\,. \tag{10}$$

Para testar as hipóteses nulas que $\mu_i = \mu_j$ $(1 \le i < j \le I)$, podemos simplesmente calcular os intervalos acima e verificar quais deles contem o ponto $\mu_j - \mu_i = 0$. No **R**, os intervalos de Tukey são calculados pela função `TukeyHSD` (cuidado, **R** faz distinção entre maiúsculas e minúsculas).

Exemplo 2. (Continuação). Como vimos antes que rejeitamos a H_0 que todas as médias são iguais, fazemos agora as comparações múltiplas. Para construir intervalos com 95% de confiança, vimos antes que $T_{0.05;4;20} \doteq 3.958$ e $QME_{erro} \doteq 3.374$. Assim, os intervalos de confiança 10 vão ser da forma $\bar{y}_{j,}. - \bar{y}_{i,}. \pm (3.958)\,\sqrt{3.374/6} \doteq \bar{y}_{j,}. - \bar{y}_{i,}. \pm 2.968$ Por exemplo, para comparar as médias dos fertilizantes 1 e 2, o IC é $8.72 - 8.42 \pm 2.968 \doteq (-2.668; 3.268)$ e, como contém o valor $\mu_1 - \mu_2 = 0$, não é possível concluir que as médias dos fertilizantes 1 e 2 são diferentes. Para fazer o resto das comparações, é possível usar a função `TukeyHSD` no **R**. Abaixo mostramos o resultado dela no exemplo. Veja que a única diferença de médias significativamente diferente de zero são as referentes aos fertilizantes 3 e 4.

```
> ajuste<-aov(colheita~fert)    # modelo colheita = (média do trat i) + erro
> TukeyHSD(ajuste,ordered=T,conf.level=0.95)  # calcule ICs, médias ordenadas
  Tukey multiple comparisons of means
    95% family-wise confidence level
    factor levels have been ordered

Fit: aov(formula = colheita ~ fert)

$fert
```

```
        diff        lwr      upr     p adj
2-3 1.783333 -1.1849103 4.751577 0.3588379
1-3 2.083333 -0.8849103 5.051577 0.2342865
4-3 3.683333  0.7150897 6.651577 0.0118130
1-2 0.300000 -2.6682436 3.268244 0.9918528
4-2 1.900000 -1.0682436 4.868244 0.3062333
4-1 1.600000 -1.3682436 4.568244 0.4510595
```

5 Diagnóstico do modelo

Existem basicamente dois supostos que foram feitos nas Seções anteriores. Primeiro, assumimos que todas as observações seguem uma distribuição Normal. Segundo, assumimos que as variâncias dentro de cada tratamento ou subpopulação são iguais. Nas aplicações, usualmente queremos checar se os dados suportam esses supostos.

Apêndice: Código R

```
#######################################
#         Exemplo 1
#######################################

par(mfrow=c(2,2))   #  faça uma figura com 4 gráficos ordenados 2x2

set.seed(999)
x<-rnorm(100,0,1)   # gere 100 observações da Normal com média 0 e desvio padrão 1


hist(x,xlim=c(-4,4),main='(a)',ylab='frequência')   #  faça um histograma das 100 observações
sd(x)     # calcule o desvio padrão da amostra
mean(x)

y<-rep(NA,100)
y[1:50]<-x[1:50]-1     # subtraia 1 as primeiras 50 observações
y[51:100]<-x[51:100]+1 # some 1 as últimas 50 observações

hist(y,xlim=c(-4,4),main='(b)',ylab='frequência')  # faça um histograma da nova amostra
sd(y)
mean(y)

hist(y[1:50],xlim=c(-4,4),,main='(c)',ylab='frequência')   # faça um histograma das primeiras 5
hist(y[51:100],xlim=c(-4,4),,main='(d)',ylab='frequência')  # faça um histograma das últimas 50
```

```
mean(y[1:50])
mean(y[51:100])
sd(y[1:50])
sd(y[51:100])

99*sd(y)^2
49*(sd(y[1:50])^2+sd(y[51:100])^2)+
  50*((mean(y[1:50])-mean(y))^2+(mean(y[51:100])-mean(y))^2)

#####################################
#         Exemplo 2
#####################################

I<-4      # número de tratamentos
J<-6      # no. de observações por tratamento

fert<-factor(rep(1:I,times=rep(J,I))) # variável qualitativa (factor) para os tratamentos
colheita<-c(10.5,9.7,5.5,11.0,9.3,6.3,     # produção de tomates
              7.1,10.2,7.2,7.8,10.9,7.3,
              4.7,7.5,6.5,7.7,6.5,6.9,
              8.5,12.4,11.5,7.1,10.6,11.8)

for(i in unique(fert)){print(mean(colheita[fert==i]))}   # médias por tratamento
for(i in unique(fert)){print(sd(colheita[fert==i]))}     # dp por tratamento
for(i in unique(fert)){print(sd(colheita[fert==i])^2)}   # var. por tratamento

sqrt((5.177667+2.837667+1.146667+4.333667)/4)  # média das variâncias

ajuste<-aov(colheita~fert)    # modelo tomate = (média do trat i) + erro
summary(ajuste)           # resumo do ajuste (tabela anova)

qf(0.95,3,20)      # percentil 95% da F com 3 e 24 gl
1-pf(4.053,3,20)


qtukey(0.95,nmeans=4,df=20)

ajuste<-aov(colheita~fert)    # modelo colheita = (média do trat i) + erro
TukeyHSD(ajuste,ordered=T,conf.level=0.95)  # calcule ICs, médias ordenadas

par(mfrow=c(1,2))

plot(colheita,fert,pch=19,col=fert,cex=1.5)   # gráficos de pontos (boxplots)
for(i in 1:4){lines(c(0,14),c(i,i),col=i,lwd=0.5)}
```

```
boxplot(colheita~fert,horizontal=T)   # gráficos de caixa (boxplots)

par(mfcol=c(2,5))      # gráficos para dados simulados
for(j in 1:5){
colheita1<-rnorm(I*J,mean(colheita),sd(colheita))
plot(colheita1,fert,pch=19,col=fert,cex=1.5)
for(i in 1:4){lines(c(0,14),c(i,i),col=i,lwd=0.5)}

boxplot(colheita1~fert,horizontal=T)
}

################################
###  Simulação da Seção Cuidado!
################################

M<-10000     # tamanho da simulação
alfa<-0.10   # nivel dos testes individuais
mu<-mean(colheita)   # mesma média que dados de tomates
sigma<-sqrt((5.177667+2.837667+1.146667+4.333667)/4)  # mesma variância também

rejeita.soma<-0    # contar o número de vezes que rejeita a H_0 onibus
for(m in 1:M){     # loop das simulações
  colheita.sim<-rnorm(I*J,mu,sigma)   # gere aleatoriamente um conjunto de dados
  rejeita<-0     # no final rejeita vai medir se rejeitou pelo menos uma comparacao
  for(i1 in 1:(I-1)){
    for(i2 in (i1+1):I){
      x<-colheita.sim[fert==i1]
      y<-colheita.sim[fert==i2]
      s.c<-sqrt((sd(x)^2+sd(y)^2)/2)
      if(abs(mean(y)-mean(x))>qt(1-alfa/2,2*J-2)*sqrt(2/J)*s.c){rejeita<-1}
    }
  }
  rejeita.soma<-rejeita.soma+rejeita
}
rejeita.soma/M
```