Delineamento e Análise de Experimentos

Professora Juliana Betini Fachini Gomes
e-mail: jfachini@unb.br

Brasília - 2023

Revisando

Kutner et al. (2013) enfatizam que o desenho adequado de um estudo científico é muito mais importante do que as técnicas específicas usadas na análise.

Unidade Experimental ou Parcela

- Segundo Barbin (2003), unidade experimental ou parcela são os índivíduos(plantas, animais, aparelho eletrônico, vaso, etc) aos quais será aplicado um tratamento. Elas devem ser tão homogêneas quanto possível, para que , quando submetidas a diferentes tratamentos, sejam os efeitos facilmente detectados.

Princípios Básicos em Estudos Experimentais

1. Casualização
2. Repetição
3. Controle Local (Bloco)

ANOVA

- Para o modelo de efeitos (ANOVA):

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij}, \quad i = 1, 2, ..., a; \quad j = 1, 2, ..., n \qquad (1)$$

em que μ é a média geral, τ_i é a média ou efeito dos tratamentos e ε_{ij} componente de erro aleatório.

- As hipótese de interesse são definidas por:

$$\begin{cases} H_0 : \tau_1 = \tau_2 = ... = \tau_a = 0, & \text{(O efeito de tratamento é nulo)} \\ H_1 : \exists \tau_i \neq 0 \end{cases}$$

- Ou de forma equivalente:

$$\begin{cases} H_0 : \mu_1 = \mu_2 = ... = \mu_a \\ H_1 : \exists \mu_i \neq \mu_j, \ i \neq j, \end{cases}$$

em que $\mu_i = \mu + \tau_i$.

Anova

- Só é possível realizar a análise de variância se certas condições, ou seja, certas exigências do modelo matemático forem satisfeitas:
    - Os erros devem ter distribuição normal;
    - Os erros devem ser independentes;
    - Os erros deve ter a mesma variância, ou seja, deve existir homocedasticidade.
- Pode-se representar essas condições por:

$$\varepsilon_{ij} \sim N(0, \sigma^2). \tag{2}$$

ANOVA

- O nome análise de variância é derivado de uma partição da variabilidade total em suas partes componentes. A soma total dos quadrados:

$$SS_T = \sum_{i=1}^{a} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{..})^2 \qquad (3)$$

é usada como uma medida da variabilidade geral dos dados;

- Intuitivamente, isso é razoável porque se dividirmos SS_T pelo número apropriado de graus de liberdade (neste caso, $an - 1 = N - 1$), teríamos a variância amostral dos $y^{'}$s;

- A variância da amostra é, obviamente, uma medida padrão de variabilidade.

ANOVA

- Observe que a SS_T pode ser escrita como:

$$\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{..})^2 = n\sum_{i=1}^{a}(\bar{y}_{i.}-\bar{y}_{..})^2 + \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i.})^2 \quad (4)$$

- A equação (4) é conhecida com a identidade fundamental da ANOVA;
- Ela afirma que a variabilidade total dos dados, medida pela soma total de quadrados, pode ser dividida em uma soma de quadrados das diferenças entre as médias de tratamento e a média geral mais uma soma de quadrados das diferenças das observações com a média do respectivo tratamento;

ANOVA

- Sendo assim, a equação (4) também pode ser escrita como

$$SS_T = SS_{Trat} + SS_{Res}. \tag{5}$$

- SS_{Trat} - soma de quadrados devido aos tratamentos - representa a diferença entre as médias de cada tratamento e a média geral, ou seja, é uma medida de diferenças entre as médias de tratamento,

- SS_{Res} - soma dos quadrados devido ao erro (ou seja, dentro dos tratamentos) - representa que as diferenças das observações em relação a média do respectivo tratamento pode ser devido à um erro aleatório.

ANOVA

- Há um total de $an = N$ observações. Então, a SS_T tem $an - 1 = N - 1$ graus de liberdade;
- Existem a níveis de um fator (a médias de tratamentos). Então, a SS_{Trat} tem $a - 1$ graus de liberdade;
- E existem n repetições dentro de qualquer tratamento fornecendo $n - 1$ graus de liberdade para estimar o erro experimental. Como existem a tratamentos, tem-se $a(n-1) = an - a = N - a$ graus de liberdade para o erro.

ANOVA

- É importante examinar o lado direito da identidade fundamental da ANOVA;
- Considere a soma de quadrados do erro:

$$SS_{Res} = \sum_{i=1}^{a} \sum_{j=1}^{n} (y_{ij} - \bar{y_{i.}})^2 = \sum_{i=1}^{a} \left[\sum_{j=1}^{n} (y_{ij} - \bar{y_{i.}})^2 \right],$$

- É fácil observar que o termo entre colchetes, se dividido por $n - 1$, é a variância da amostra do i-ésimo tratamento, ou

$$S_i^2 = \frac{\sum_{j=1}^{n} (y_{ij} - \bar{y_{i.}})^2}{n - 1}, \tag{6}$$

para $i = 1, 2, \ldots, a$.

ANOVA

- Ao considerar os a tratamentos, as variâncias da amostra podem ser combinadas para fornecer uma única estimativa da variância da população:

$$\begin{aligned}
\frac{(n-1)S_1^2 + (n-1)S_2^2 + \ldots + (n-1)S_a^2}{(n-1) + (n-1) + \ldots + (n-1)} &= \frac{\sum_{i=1}^{a}\left[\sum_{j=1}^{n}(y_{ij} - \bar{y_{i.}})^2\right]}{\sum_{i=1}^{a}(n-1)} \\
&= \frac{SS_{Res}}{(N-a)},
\end{aligned}$$

que é uma estimativa combinada da variância dentro de cada um dos a tratamentos.

ANOVA

- Se não houvesse diferenças entre as médias de tratamento, poderia-se usar a variação das médias de tratamento em relação a média geral para estimar σ^2. Ou seja:

$$\frac{SQ_{Trat}}{a-1} = \frac{n\sum_{i=1}^{a}(\bar{y}_{i.} - \bar{y}_{..})^2}{a-1}$$

  é uma estimativa de σ^2 se as médias de tratamento forem iguais.
- A razão para isso pode ser intuitivamente vista como segue:
  - A quantidade $\sum_{i=1}^{a}(\bar{y}_{i.} - \bar{y}_{..})^2/(a-1)$ estima σ^2/n, que é a variância das médias de tratamento;
  - Então, $n\sum_{i=1}^{a}(\bar{y}_{i.} - \bar{y}_{..})^2/(a-1)$ deve estimar σ^2 se não houver diferenças significativas entre os tratamentos.

ANOVA

- Logo, é possível observar que a ANOVA fornece duas estimativas para σ^2;
- Uma baseada na variabilidade inerente dentro dos tratamentos e a outra baseada na variabilidade entre tratamentos;
- Se não houver diferenças nas médias de tratamento, essas duas estimativas devem ser muito semelhante. E se não forem, suspeita-se que a diferença deve ser causada por diferenças nas médias de tratamento;
- Embora tenha-se usado um argumento intuitivo para desenvolver este resultado, uma abordagem um pouco mais formal pode ser utilizada.

ANOVA

- As quantidades:

$$QM_{Trat} = \frac{SQ_{Trat}}{a-1} \tag{7}$$

e

$$QM_{Res} = \frac{SQ_{Res}}{N-a} \tag{8}$$

são chamadas de quadrados médios.

- Agora será examinado o valor esperado dos quadrados médios:

ANOVA

$$E(QM_{Res}) = E\left(\frac{SQ_{Res}}{N-a}\right) = \frac{1}{N-a}E\left[\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y_{i.}})^2\right] = \sigma^2$$

e

$$E(QM_{Trat}) = \sigma^2 + \frac{n\sum_{i=1}^{a}\tau_i^2}{a-1}$$

- Ou seja, $QM_{Res} = SQ_{Res}/(N-a)$ estima σ^2;
- E $QM_{Trat} = SQ_{Trat}/(a-1)$ também estima σ^2, caso não há diferença entre as médias dos tratamentos (o que implica em $\tau_i = 0$).

ANOVA

- Ao considerar as hipóteses de interesse:

$$\begin{cases} H_0 : \tau_1 = \tau_2 = ... = \tau_a = 0, & \text{(O efeito de tratamento é nulo)} \\ H_1 : \exists \tau_i \neq 0 \end{cases}$$

- Ou de forma equivalente:

$$\begin{cases} H_0 : \mu_1 = \mu_2 = ... = \mu_a \\ H_1 : \exists \mu_i \neq \mu_j, \ i \neq j, \end{cases}$$

em que $\mu_i = \mu + \tau_i$.

- Qual a estatística do teste?

ANOVA

- Ao supor $\varepsilon_{ij} \sim N(0, \sigma^2)$;
- As observações, y_{ij}, também seguem distribuição $N(\mu + \tau_i, \sigma^2)$;
- Logo, SQ_T também seguem distribuição Normal e SQ_T/σ^2 segue distribuição χ^2 com $(N-1)$ graus de liberdade;
- Além disso, SQ_{Res}/σ^2 segue distribuição χ^2 com $(N-a)$ graus de liberdade e SQ_{Trat}/σ^2 segue distribuição χ^2 com $(a-1)$ graus de liberdade se a hipótese nula for verdadeira, $H_0 : \tau_i = 0$;

ANOVA

- No entanto, as três somas de quadrados não são necessariamente independente porque ao somar SQ_{Trat} e SQ_{Res} tem-se a SQ_T;

- Ao usar o Teorema de Cochran definido por:
  - Seja $Z_i \sim N(0,1)$ para $i = 1, 2, \ldots, \nu$ e

$$\sum_{i=1}^{\nu} Z_i^2 = Q_1 + Q_2 + \ldots + Q_s$$

em que $s \leq \nu$, e Q_i tem ν graus de liberdade $(i = 1, 2, \ldots, s)$. Então, $Q_1, Q_2, \ldots, Q_s$ são variáveis aleatórias independentes com distribuição qui-quadrado com $\nu_1, \nu_2, \ldots, \nu_s$ graus de liberdade, respectivamente, se e somente se

$$\nu = \nu_1 + \nu_2 + \ldots + \nu_s$$

ANOVA

- Como os graus de liberdade de SQ_{Trat} e SQ_{Res} somam $N-1$, que é o número de graus de liberdade da SQ_T, o Teorema de Cochran estabelece que SQ_{Trat}/σ^2 e SQ_{Res}/σ^2 são variáveis aleatórias independentes.

- Portanto, se a hipótese nula de nenhuma diferença nas médias de tratamento for verdadeira, a razão

$$F_0 = \frac{SQ_{Trat}/(a-1)}{SQ_{Res}/(N-a)} = \frac{QM_{Trat}}{QM_{Res}}, \tag{9}$$

  segue distribuição F com $a-1$ e $N-a$ graus de liberdade.

- A equação (9) é a estatística do teste para a hipótese de não haver diferenças nas médias de tratamento.

ANOVA

- A partir da esperança dos quadrados médios, verifica-se que, em geral, QM_{Res} é um estimador imparcial de σ^2;
- Além disso, sob a hipótese nula ser verdadeira, QM_{Trat} também é um estimador imparcial de σ^2;
- No entanto, se a hipótese nula for falsa, o valor esperado de QM_{Trat} é maior que σ^2;
- Portanto, sob a hipótese alternativa, o valor esperado do numerador da estatística de teste (9) é maior que o valor esperado do denominador, e deve-se rejeitar H_0 para valores da estatística do teste que são muito grandes;
- Portanto, deve-se rejeitar H_0 e concluir que existem diferenças nas médias de tratamento se $F_0 > F_{\alpha,(a-1),(N-a)}$.

ANOVA

Table: Tabela de Análise de Variância

Fonte de Variação	SQ	g.l.	QM	F
Tratamentos	SQ_{Trat}	a - 1	QM_{Trat}	QM_{Trat}/QM_{Res}
Resíduo	SQ_{Res}	N - a	QM_{Res}	
Total	SQ_T	N - 1		

em que $N = an$.

ANOVA

- Uma fórmula alternativa para calcular SQ_T e SQ_{Trat} é dada por:

$$SQ_T = \sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij}^2 - \frac{\bar{y}_{..}^2}{N}, \quad (10)$$

$$SQ_{Trat} = \frac{1}{n}\sum_{i=1}^{a} y_{i.}^2 + \frac{\bar{y}_{..}^2}{N} \quad (11)$$

e

$$SQ_{Res} = SQ_T - SQ_{Trat}. \quad (12)$$

Diagnóstico

- É fundamental verificar a adequabilidade do modelo antes de tirar conclusões sobre o processo inferencial.

- Verificar as características do modelo.

- Para avaliar a adequabilidade do modelo serão usados:
    - Métodos gráficos
    - Testes estatísticos

Resíduo

- Violações das suposições básicas e adequação do modelo podem ser facilmente investigadas através da análise dos resíduos.
- O resíduo para a j-ésima observação no i-ésimo tratamento é definido por:

$$e_{ij} = y_{ij} - \hat{y}_{ij}, \tag{13}$$

em que $\hat{y}_{ij}$ é o valor estimado obtido por:

$$\begin{aligned} \hat{y}_{ij} &= \hat{\mu} + \hat{\tau}_i \\ &= \bar{y}_{..} + (\bar{y}_{i.} - \bar{y}_{..}) \\ &= \bar{y}_{i.} \end{aligned} \tag{14}$$

Resíduo

- Para o modelo (1), assume-se que o erros, ε_{ij} são variáveis aleatórias independentes, identicamente distribuídos com distribuição Normal com média 0 e variância σ^2.

- Se o modelo (1) for apropriado para os dados em estudo, os resíduos observados, e_{ij}, devem refletir as propriedades assumidas para o erro do modelo, ε_{ij}.

Suposição de Normalidade

- Pode ser inicialmente verificada ao construir um histograma dos resíduos. Se a suposição sobre os erros for satisfeita, este gráfico deve se parecer com uma amostra de uma distribuição normal centrada em zero;
- Outro recurso gráfico importante é o **Gráfico de Probabilidade Normal dos resíduos** - cada resíduo é colocado em um gráfico versus seu valor esperado sob normalidade.
- Se o resultado gráfico for aproximadamente linear, há evidências da suposição de normalidade.
- Se o resultado gráfico afasta substancialmente da linearidade, não há evidências de que a distribuição do erro é normal.

Suposição de Normalidade

- As vezes é útil padronizar os resíduos para a análise de resíduos.
- Como o desvio padrão dos erros, ε_{ij}, é σ, que pode ser estimado por $\hat{\sigma} = \sqrt{\text{MSRes}}$. É natural considerar a seguinte padronização:

$$d_{ij} = \frac{e_{ij}}{\sqrt{\text{MSRes}}}. \tag{15}$$

- Se ε_{ij} sin $N(0, \sigma^2)$, os resíduos padronizados (15) deve ser aproximadamente normal com média zero e variância constante;
- Teste de hipótese também pode ser utilizado, como o teste de Shapiro-Wilk.

Suposição de independência

- O objetivo é verificar se existe qualquer correlação entre os erros que são próximos um dos outros em uma sequência;
- Quando os erros são independentes espera-se que os resíduos versus a ordem das observações flutuem em um padrão mais ou menos aleatório em torno da linha de referência Zero.

Os erros não são independentes

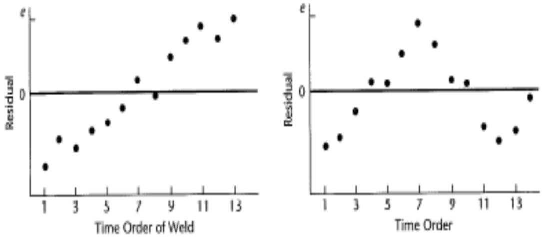

Figure: Exemplo

Variância constante

- Gráficos dos resíduos versus valores ajustados podem ser usados para verificar se a variância dos erros é constante. Pois, os sinais dos resíduos não são importantes para examinar a constância da variância do erro;
- Testes de hipóteses.

Heterocedasticida e homogeneodade de Variância

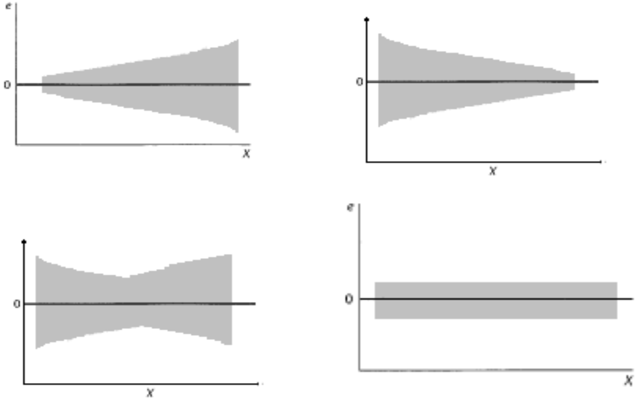

Figure: As três primeiras figuras representam exemplos de situação com heterocedasticida de variância e a quarta figura representa o caso de homogeneodade de variância, em que x representa os valores ajustados

Variância constante

Para verificar as hipóteses:

$$\begin{cases} H_0 : \sigma_1 = \sigma_2 = ... = \sigma_a = 0 \\ H_1 : \exists \sigma_i \neq 0 \end{cases}$$

pode-se utilizar os seguintes testes de hipóteses:

- F;
- Bartlett, e
- Levene

Variância constante

- O teste de hipótese F possui estatística do teste dada por:

$$F = \frac{S_1^2}{S_2^2} \sim F_{(n_1-1),(n_2-1)}.$$

- Como nos delineamentos inteiramente casualizados existem a tratamentos, uma alternativa é fazer o teste F dado por:

$$F_{max} = \frac{S_{max}^2}{S_{min}^2} \sim F_{(n_{max}-1),(n_{min}-1)},$$

em que (n_{max}) é o número de observações do tratamento com maior variância e (n_{min}) é o número de observações do tratamento com menor variância.

Variância constante

- O teste de hipótese de Bartlett envolve o cálculo de uma estatística cuja distribuição amostral é aproximada pela distribuição qui-quadrado com $a-1$ graus de liberdade quando as a amostras aleatórias são de populações normais independentes. A estatística do teste é definida por:

$$\chi_0^2 = 2,3026\frac{q}{c}$$

em que

$$q = (N-a)\log_{10} S_p^2 - \sum_{i=1}^{a}(n_i-1)\log_{10} S_i^2,$$

$$c = 1 + \frac{1}{3(a-1)}\left(\sum_{i=1}^{a}(n_i-1)^{-1} - (N-a)^{-1}\right),$$

Variância constante

$$S_p^2 = \frac{\sum_{i=1}^{a}(n_i - 1)S_i^2}{N - a}$$

e S_i^2 é a variância amostral da população i.

- A quantidade q é grande quando as variâncias da amostra S_i^2 diferem muito e é igual a zero quando todos os S_i^2 são iguais. Portanto, deve-se rejeitar H_0 em valores de χ_0^2 que são muito grandes. Ou seja, rejeitamos H_0 somente quando

$$\chi_0^2 > \chi^2_{\alpha,(a-1)}.$$

- O teste de Bartlett é muito sensível à suposição de normalidade. Consequentemente, quando o validade desta suposição é duvidosa, o teste de Bartlett não deve ser usado.

Variância constante

- Como o teste de Bartlett é sensível à suposição de normalidade, é importante ter um procedimento alternativo;
- Um teste alternativo muito utilizado é o teste de Levene modificado;
- Esse teste usa o desvio absoluto das observações y_{ij} em relação a mediana do respectivo tratamento, $\tilde{y}_i$. Denote esses desvios por

$$d_{ij} = |y_{ij} - \tilde{y}_i|$$

em que $i = 1, 2, \ldots, a$ e $j = 1, 2, \ldots, n_i$.

Variância constante

- O teste de Levene modificado então avalia se as médias desses desvios são ou não iguais para todos os tratamentos;
- Acontece que se os desvios médios são iguais, as variâncias das observações em todos os tratamentos serão as mesmas;
- Logo, a estatística de teste para o teste de Levene é simplesmente a estatística ANOVA para testar a igualdade de médias aplicada aos desvios absolutos.