ANÁLISE DE VARIÂNCIA (ANOVA) CLÁSSICA: TÉCNICA ÚTIL, PORÉM RESTRITIVA!

Questões associadas à verificação de suas suposições:

(aditividade, independência, homocedasticidade e normalidade) → ε_{ijk} i.i.d.~$N(0,\sigma^2)$ → "quadrados mínimos ordinários"

1) Os dados provêm de experimento?

(pesquisa experimental: repetição, casualização) ≠ (pesquisa de observação natural)

SIM **→ reúne as condições prévias p/ a análise, o que não é garantia de sua validade → verificá-las (sempre)!**
NÃO **→ não reúne as condições prévias p/ a análise (mesmo se não houver violação, o uso da análise deve ser justificado)**

2) Os dados (experimentais) são qualitativos?
SIM **→ A análise NÃO se aplica (há uma variação com princípio similar – AMOVA: uso comum em "genética molecular").**

3) Dados quantitativos?
SIM **→ A análise pode ser aplicada, PORÉM, as conclusões estão condicionadas ao atendimento das pressuposições.**

a) *Variáveis discretas ou contínuas truncadas* (ex. contagens; escalas de notas etc.)
→ A princípio, não reúnem condições para garantia de NORMALIDADE → verificá-las (sempre)!
(sobretudo se "n" é pequeno e tratar-se de escala com poucos valores – ex. 1 a 5).

A ANOVA, porém, pode ser válida, se atendidas as pressuposições, ou sob:

- Uso de transformação adequada (mudança da escala dos dados) → ex. Box-Cox (1964) → verificá-las (sempre).
- Obtenção de dados, por exemplo, como médias de contagens (T.L.C. → médias ~ Normal, se "n" for grande).
- Porcentagens (truncadas em 0% e 100%): dados entre 30-70% (ou 20-80%) maior chance de atender ~ Normal.
(0%-20%) e (80%-100%): necessidade de transformação → transf. angular: *arco seno*[*raiz*(Y/100)] → verificá-las!

b) *Variáveis contínuas sem truncamento* (ex. altura, peso, massa, temperatura etc.)
→ A princípio, reúnem as condições; mas não se tem garantia de sua validade → verificá-las sempre!
Caso as pressuposições sejam violadas: buscar alternativas (transformações ou outros métodos)

Atenção p/ possíveis dados discrepantes ("outliers"): podem ser a razão das violações (mas, não os elimine sem razão).

SÍNTESE: inferências a partir de ANOVA só devem ser feitas se as pressuposições forem minimamente atendidas.
(isto, apesar da reportada robustez de testes associados – ex. F-Snedecor; t-Student; Tukey; etc).

Então, faça:

ANOVA → Valores preditos e *resíduos* → Análise de resíduos: verificação das condições de validade.

- Regras práticas: ex. F_{max} < 3ou4 (homog.variâncias); Ampl(i)/méd(i), var(i)/méd(i), s(i)/méd(i) → (heteroc. regular)
- Gráfico de resíduos: configurações de pontos → "outliers", heterocedasticidade e/ou falta de independência.
- Histograma dos resíduos: avaliação de simetria/curtose/"outliers" → desvios de normalidade.
- Q-Q plot (resíduos): "outliers" e normalidade // (correlação de 'resíduos' x 'resíduos padronizados' → teste)
- Testes estatísticos: (maior parte deles aplicada aos resíduos do ajustamento ANOVA)
  - Tukey (1949); Box et al. (1978): aditividade (exclusivo para delineamento em blocos) – (teoria incipiente)
  - Durbin-Watson (1ª, 2ª, ... ordem): independência (vide última página deste documento – item 2.2.3)
  - Barttlet; Cochran; F_{max} (Hartley); Levene: homogeneidade de variâncias
  - Shapiro-Wilk; Kolmogorov-Smirnov; Lilliefors; Cramer-von Mises; Qui-quadrado etc.: normalidade

Busca de transformação:

a) experiência prévia (tradição da área) e regras práticas (ex. dados de contagem e porcentagens; méd(t_i) muito distintas)

- médias α variâncias (distrib. Poisson): raiz quadrada (Y); se houver zeros, raiz (Y+1) ou raiz (Y+0,5)
- médias α desvios padrão (distr. Log-normal): logaritmo - ln(Y) ou $\log_{10}$(Y); se houver zeros, ln(Y+k) ou $\log_{10}$(Y+k); k- c^{te}.
- porcentagens (distrib. Binomial): transformação angular – *arc_sen* [*raiz*(Y/100)]; em que Y é expressa em %.
- verificação da uniformidade de relações como: Ampl./méd → Log(Y); Ampl./raiz(méd.) → Raiz(Y)
- outras transformações específicas: f (tipo de distribuição dos resíduos)

b) critérios técnicos:

- Transformação "ótima" de Box & Cox (1964): $Y' = (Y^{\lambda} - 1)/\lambda$, se $\lambda \neq 0$; ou $Y' = \ln(Y)$, se $\lambda = 0$.
  ("ótima" p/ normalidade) (o parâmetro λ é estimado por máxima verossimilhança[1] ou quadr.mín- aplicativos).
- Transformação "estabilizadora da variância" (Tukey, 1971): $Y' = Y^{\lambda}$, em que $\lambda = 1 - b/2$, sendo 'b' a estimativa do coeficiente β da regressão linear simples: $\ln(\sigma^2_i) = \alpha + \beta \ln(\mu_i)$ → Vide Tabela 1 ao final deste texto.

Apresentação dos resultados → Se uma transformação foi necessária[2]: testes estatísticos só com os dados transformados.
(paralelamente podem ser apresentadas as médias dos dados originais sem qualquer teste[3]).

---

[1] $L = -(\nu/2)\ln(S_T^2) + (\lambda-1)(\nu/n)\Sigma\ln(Y)$; ν é GL_{Erro}; S_T^2 é $QM_{Erro(Y')}$; λ é o expoente de Y'; n é número total de obs.; e *Y* são os dados originais.
[2] Mais de uma transformação pode solucionar o problema de violação das pressuposições encontrado.
[3] Alguns autores recomendam apresentá-las pela reversão da transformação utilizada (outros não).

SE AS TRANSFORMAÇÕES DISPONÍVEIS FALHAREM?

*** Buscar métodos de análise menos restritivos: acomodam, por exemplo, heterocedasticidade e/ou correlação entre observações; admitem outras distribuições de probabilidade ou são livres de distribuições teóricas; adotam modelos lineares generalizados ou modelos não-lineares → análises mais complexas (não triviais)*.**

Alguns exemplos associados ao tipo de violação:

- **Falta de aditividade: abordagem de modelos não-lineares (ex. modelos exponenciais e multiplicativos).**

- **Falta de normalidade: testes não-paramétricos; testes de permutação ou de randomização (ou via *bootstrap*); modelos lineares generalizados.**

- **Heterocedasticidade irregular: eliminação de tratamentos com variâncias muito discrepantes; método dos resíduos específicos (**divisão do erro em componentes aplicáveis aos distintos grupos de tratamentos com variâncias comuns**); quadrados mínimos ponderados ou generalizados.**

- **Falta de independência*: quadrados mínimos generalizados, associado a algum tipo de análise estatística espacial; análise de medidas repetidas (dados longitudinais) e/ou análise de séries temporais.**

*** Há também outras abordagens para análise de dados, como, por exemplo, a Inferência Bayesiana.**

*** Autocorrelação positiva pode resultar de especificação inadequada do modelo (ex. ajuste de regressão linear simples para relações com comportamento quadrático – Hoffmann & Vieira, 1998, p. 252)**

Tabela1. Correspondências entre transformações clássicas e valores de ‘b’ resultante da relação $\ln(Var_i) = a + b \ln(Méd_i)$.

Relação de proporcionalidade (α) entre Variância (σ^2_i) e média (μ_i)	b	$\lambda = 1 - (b/2)$	Transf. $Y' = Y^\lambda$	Transf. clássica
$\sigma^2_i \;\alpha\;$ constante	0	1	$Y' = Y^1 = Y$ (nenhuma)	nenhuma
$\sigma^2_i \;\alpha\; \mu_i^1$ → (Y~ Poisson)	1	½	$Y'=Y^{½} = \sqrt{Y}$	raiz quadrada*
$\sigma^2_i \;\alpha\; \mu_i^2$ → (Y~ Log-Normal)	2	0	Y’=log(Y) ou Y’=ln(Y)	logaritmica*
$\sigma^2_i \;\alpha\; \mu_i^3$	3	-½	$Y'=Y^{-½} = 1/\sqrt{Y}$	raiz recíproca*
$\sigma^2_i \;\alpha\; \mu_i^4$	4	-1	$Y'=Y^{-1} = 1/Y$	recíproca*
Y (<20%; >80%)** → y = 1, 2, ..., n; y ~ Binomial	–	–	Y’ = *arc_sen* ($\sqrt{Y/100}$)	

* - Para estas transformações, sob a presença de zeros, recomenda-se somar uma constante k; ex. $\sqrt{(Y+0,5)}$, log(Y+1) etc. (Nogueira, 1994).

** - Na faixa 20-80% (ou 30-70%, para alguns autores), pouco ou nada se ganha com o uso da transformação angular (Zimmermann, 2004).

Notas:

- Obviamente essa família de transformações potências ($Y' = Y^\lambda$), assim como a de Box & Cox, possibilita a adoção de potências intermediárias àquelas que resultam nas transformações clássicas (Y’→ $Y^{\lambda=0,25}$; $Y^{\lambda=0,75}$; $Y^{\lambda=-0,25}$ etc.).
- O emprego da constante k, usualmente igual a 1,0 ou 0,5, obrigatório no caso da presença de zeros e uso da transformação logarítmica, é útil também para evitar supercorreção, por exemplo, da transformação raiz quadrada, sobretudo quando ocorrem valores (contagens) muito pequenos (Zimmermann, 2004). Nota: Excesso de “zeros” exigem outras abordagens.

Teste de Aditividade do Modelo (Tukey, 1949):

Em DBC: SQ_{NA} = $n.[\sum_{i,j}(Y_{ij}(\overline{Y}_{i}.-\overline{Y}..)(\overline{Y}._j-\overline{Y}..)]^2/(SQ_{Tr}.SQ_{Bl})$ **, com GL_{NA}=1 e novo Resíduo: $QM_{E(na)}$= $(SQe-SQ_{NA})/(GLe-1)$**

Em qualquer delineamento:

⇨ **De posse de:** $\hat{e}_{ij}$ **,** $\hat{Y}_{ij}$ **e SQe =** $\sum_{i,j}\hat{e}_{ij}^2$ **, com GLe graus de liberdade:**

1) **Calcula-se:** $q_{ij} = \hat{Y}_{ij}^2$ **, com i=1,2,...I tratamentos e j=1,2,...J repetições.**
2) **ANOVA:** $q_{ij} = m + t_i + b_j + e'_{ij}$ **# isto p/ D.B.C.(ajustar outro modelo de delineamento, se for o caso)**
3) **Obtenha:** $\hat{e}'_{ij} = q_{ij} - \hat{q}_{ij}$ **; com** $\hat{q}_{ij} = \hat{m} + \hat{t}_i + \hat{b}_j$ **# (ou outra expressão de predição se for o caso).**

   $A = \sum_{i,j}\hat{e}'^2_{ij}$; $B = \sum_{i,j}\hat{e}_{ij}\,q_{ij} = \sum_{i,j}(Y_{ij} - \hat{Y}_{ij})\,\hat{Y}_{ij}^2$

4) $SQ_{N\_Adit} = B^2/A$ **, com 1 grau de liberdade =>** $QM_{N\_Adit} = SQ_{N\_Adit}/1 = SQ_{N\_Adit}$
5) $SQ_{Res2} = SQe - SQ_{N\_Adit}$ **, com GLe-1 graus de liberdade =>** $QM_{Res} = SQ_{Res2}/(GLe-1)$
6) $F_{N\_Adit} = QM_{N\_Adit}/QM_{Res2}$ **=> p-valor na distribuição F-Snedecor com gaus de liberdade (1; GLe-1).**

Testes de Homogeneidade de Variâncias (homocedasticidade):

⇨ **Teste de Hartley ($F_{máx}$)** - presta-se para conjuntos balanceados de dados (j=1,2,...J repetições):

$$F_{máx} = \frac{s_i^2\ (máx)}{s_i^2\ (mín)}$$

A estatística tem distribuição H (ou $F_{máx}$) de Pearson & Hartley, tabulada com entradas "I" (grupos) e "J-1" (graus de liberdade das variâncias em teste) => $H_{\alpha(I;\ J-1)}$.

Para o nível de significância α, se $F_{máx} \geq H_{\alpha(I;\ J-1)}$ rejeita-se H_0 (hipótese de homocedasticidade) e conclui-se que as variâncias são heterogêneas; caso contrário, não se rejeita H_0 e admite-se que as variâncias são homogêneas.

⇨ **Teste de Bartlett (B_a)** - presta-se para conjuntos de dados desbalanceados (j=1,2,...J_i repetições).

De posse das variâncias s_i^2 dentro de tratamentos (i=1,2,...,I) e seus graus de liberdade (ν_i) tem-se:

$$B_a = \frac{2{,}3026[A(\log s_p^2) - C]}{1 + \frac{1}{3(I-1)}\left(D - \frac{1}{A}\right)}\text{, em que: } A = \sum_{i=1}^{I}\nu_i\ ;\ s_p^2 = \sum_{i=1}^{I}\nu_i s_i^2 / \sum_{i=1}^{I}\nu_i\ ;\ C = \sum_{i=1}^{I}\nu_i \log s_i^2 \text{ e } D = \sum_{i=1}^{I}\frac{1}{\nu_i}.$$

Sob ~Normal, B_a tem distribuição $\chi^2_{(I-1)}$; logo, para o nível α de significância, se $B_a \geq \chi^2_{\alpha(I-1)}$ rejeita-se H_0 e conclui-se que não há homocedasticidade; caso contrário, não se rejeita H_0 e admitem-se variâncias homogêneas.

2.2.3. Diagnosticando e estimando a covariância espacial [1/]

A autocorrelação residual foi avaliada por dois tipos de instrumentos: o teste estatístico de Durbin-Watson (d) e algumas representações gráficas dos resíduos. A estatística d, que permite testar a hipótese de ausência de autocorrelação (H_0: $\rho=0$), é definida como (Hoffmann & Vieira, 1998; SAS Institute, 1993a; Gujarati, 1992):

$$d = \frac{\sum_{l=2}^{n}(\hat{e}_l - \hat{e}_{l-1})^2}{\sum_{l=1}^{n}\hat{e}_l^2}$$

sendo: $l=1, 2,..., n$, a ordem de posicionamento da parcela associada ao resíduo $\hat{e}_l$ ($\hat{e}_l$ e $\hat{e}_{l-1}$ indicam resíduos cujas parcelas têm vizinhança de primeira ordem, isto é, são adjacentes) .

A relação entre d e ρ é aproximadamente: $d=2(1-\rho)$. Logo, se não existir autocorrelação o valor esperado de d é $2,0$; valores significativamente inferiores a $2,0$ indicam autocorrelação positiva; e, valores significativamente superiores a $2,0$ indicam autocorrelação negativa. Embora o teste tenha sido delineado, em princípio, para autocorrelação de primeira ordem, o *SAS* permite, através do *PROC AUTOREG* (procedimento para o ajuste de modelos auto-regressivos), obter estatísticas de Durbin-Watson generalizadas até uma dada ordem especificada. As instruções aqui utilizadas para testar autocorrelações residuais de primeira à décima ordem (**dw=10**) foram:

```
proc autoreg data=OBS_PRED;
  model _resid_= / dw=10 dwprob;
run;
```

1/ Fonte:

DUARTE, J. B. *Sobre o emprego e a análise estatística do delineamento em blocos aumentados no melhoramento genético vegetal*. 2000, 293 f. Tese (Doutorado em Agronomia: Genética e Melhoramento de Plantas) - Escola Superior de Agricultura Luiz de Queiroz, Universidade de São Paulo, Piracicaba, 2000.