Análise de Variância

Comparação de duas ou mais médias

Análise de variância com um fator

Exemplo

Um experimento foi realizado para se estudar diabetes gestacional. Desejava-se avaliar o comportamento da hemoglobina (HbA) em gestantes normais (N), com tolerância diminuída (TD) e diabéticas (D).

Foram escolhidas 10 gestantes de cada tipo e mediu-se suas HbA.

$$\text{Tipos de gestante} \begin{cases} \text{normal} \\ \text{tolerância diminuída} \\ \text{diabética} \end{cases}$$

Variável Resposta Y (ou Variável Dependente) :

Hemoglobina glicosilada (HbA)

Tipo de	gestante	
N	TD	D
7,86	6,20	9,67
6,38	7,82	8,08
6,90	8,50	9,25
7,78	6,50	8,20
8,17	8,09	8,64
6,26	6,90	9,67
6,30	7,82	9,23
7,86	7,45	10,43
7,42	7,75	9,97
8,63	7,43	9,59

Fazer análise descritiva

Modelo:

observação Y = sistemática + aleatória

componente sistemática (previsível): incorpora o conhecimento que o pesquisador tem sobre o fenômeno

componente aleatória: representa variações individuais e fatores que não são explicados pela parte sistemática

Subpopulações: P_1, P_2,..., P_k representadas por um nível do fator

No exemplo: fator ⇒ tipo de gestante ⇒ 3 níveis

Suponha que : P_1 ⇒ média de Y : μ_1

P_2 ⇒ média de Y : μ_2

P_k ⇒ média de Y : μ_k

Queremos testar:

H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$

H_A: pelo menos uma das médias é diferente das demais

Análise de Variância : compara a variabilidade entre as médias amostrais dos grupos e a variação dentro dos grupos.

Se H_0 é verdadeira, a variabilidade entre as médias dos grupos deve ser pequena

Em cada nível do fator : uma amostra de observações

$\Downarrow$

k amostras independentes

	Nível 1	**Nível 2**	. . .	**Nível k**
	y_{11}	y_{21}		y_{k1}
	y_{12}	y_{22}		y_{k2}
	$\vdots$	$\vdots$	. . .	$\vdots$
	y_{1n1}	y_{2n2}		y_{knk}
médias:	$\overline{y}_{1.}$	$\overline{y}_{2.}$		$\overline{y}_{k.}$

$$n=n_1+n_2+...+n_k$$

Modelo estatístico (1):

$$Y_{ij} = \mu_i + e_{ij} \qquad i=1,...,k;\ j=1,...,n_i$$

onde: μ_i : média do nível i (efeito do nível i)

e_{ij} : efeito aleatório do j-ésimo indivíduo do nível i

Y_{ij}: variável resposta do j-ésimo indivíduo do nível i

Se a hipótese H_0 for verdadeira, o modelo pode ser reescrito:

Modelo estatístico (0):

$$Y_{ij} = \mu + e^*_{ij} \qquad i=1,...,k;\ j=1,...,n_i$$

Informação que não é explicada pela parte sistemática

Modelo 1:

$$\sum_{i=1}^{k} \sum_{j=1}^{n_i} e_{ij}^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \mu_i)^2$$

Modelo 0:

$$\sum_{i=1}^{k} \sum_{j=1}^{n_i} (e_{ij}^*)^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \mu)^2$$

Modelo 1 (médias diferentes):

$$\hat{\mu}_i = \frac{1}{n_i} \sum_{i=1}^{n_i} y_{ij} = \overline{y}_{i.}$$

Modelo 0 (mesma média):

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij} = \overline{y}_{..}$$

Substituindo temos:

$$SQD = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \overline{y}_{i.})^2$$

$$SQT = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \overline{y}_{..})^2$$

SQE = SQT - SQD

$$\mathbf{SQE} = \sum_{i=1}^{k} n_i(\bar{y}_i - \bar{y}_{..})^2$$

Variabilidade **Total**	=	Variabilidade **entre** grupos	+	Variabilidade **dentro** dos grupos
↓		↓		↓
SQT	=	**SQE**	+	**SQD**

Cada uma das somas de quadrados envolve um certo número de quantidades que estão sendo estimadas. Então definimos os correspondentes quadrados médios:

$$\mathbf{QMT} = \frac{\mathbf{SQT}}{\mathbf{n-1}} \quad \mathbf{QMD} = \frac{\mathbf{SQD}}{\mathbf{n-k}} \quad \mathbf{QME} = \frac{\mathbf{SQE}}{\mathbf{k-1}}$$

Se H_0 não for verdadeira ⇓ modelo 1 é mais adequado do que o modelo 0 (resíduos do modelo 1 são menores)

QME: informação dos dados captada pelo modelo 1

QMD: informação que não é explicada pelo modelo1

$$\mathbf{F} = \frac{\mathbf{QME}}{\mathbf{QMD}}$$

Se QME for grande comparado à QMD, parte sistemática do modelo 1 está captando grande parte da informação. Quanto maior for o valor de F, maiores as evidências contra $H_{0.}$

Para realizarmos o teste, precisamos da distribuição da estatística F

Suposições

1) As amostras são independentes

2) Dentro de cada amostra as observações são independentes.

3) As observações são selecionadas de uma população na qual a variável resposta tem distribuição Normal com variâncias iguais.

$$\mathbf{Y_{ij} \sim N(\mu_i, \sigma^2), \ \text{para todo i e j.}}$$

Ou seja,

Nível 1	**Nível 2**	. . .	**Nível k**
y_{11}	y_{21}		y_{k1}
		. . .	
y_{12}	y_{22}		y_{k2}
$\vdots$	$\vdots$		$\vdots$
y_{1n1}	y_{2n2}		y_{knk}
amostra $N(\mu_1, \sigma^2)$	amostra $N(\mu_2, \sigma^2)$		amostra $N(\mu_k, \sigma^2)$

Se as suposições estiverem satisfeitas, sob a hipótese H_0 temos que:

a estatística do teste F =QME/QMD tem distribuição F-Snedecor com (k-1) e (n-k) graus de liberdade.

$$\mathbf{F = \frac{QME}{QMD} \sim F_{k-1,n-k}}$$

Rejeitamos H_0 para valores grandes de F ou seja

RC = {F ¤ a}

Pelo Teorema de Cochran temos que sob H_0:

$$\frac{\mathbf{SQE}}{\sigma^2} \sim \chi^2_{(k-1)} \qquad \frac{\mathbf{SQD}}{\sigma^2} \sim \chi^2_{(n-k)}$$

$$\mathbf{SQE\ \ e\ \ SQD\ \ independentes}$$

Resumo: Tabela de Análise de Variância - ANOVA

Fonte de variação	Graus de liberdade	Soma de quadrados	Quadrado médio	F
Entre	k-1	SQE	QME	QME/QMD
Dentro	n-k	SDQ	QMD	
Total	n-1	SQT		

com $F \sim F_{(k-1,\ n-k)}$

QMD é um estimador para a variância populacional $\blacklozenge^2$.

Combinação das variâncias amostrais dentro de cada grupo

Só tem sentido se a suposição de igualdade das variâncias populacionais é verdadeira

```
Analysis of Variance
Source  DF        SS        MS        F        P
Factor   2    23,403    11,702    19,36    0,000
Error   27    16,316     0,604
Total   29    39,719
                     Individual 95% CIs For Mean
                          Based on Pooled StDev
Level N  Mean  StDev -----+---------+---------+---------+-
C1   10 7,356  0,8469 (-----*-----)
C2   10 7,446  0,7183  (-----*-----)
C3   10 9,273  0,7614                          (-----*-----)
                      -----+---------+---------+---------+-
Pooled StDev =   0,7774   7,20      8,00      8,80      9,60
```

$$\overline{\mathbf{Y}}_{i.} \sim \mathbf{N}\left(\mu_i, \frac{\sigma^2}{n_i}\right) \qquad \frac{\overline{\mathbf{Y}}_{i.} - \mu_i}{\sqrt{\frac{\mathbf{QMD}}{n_i}}} \sim t_{n-k}$$

Quando rejeitamos a **Hipótese nula**

⇓ Localizar as diferenças através de Técnicas de Comparações Múltiplas.

Alguns Métodos
- Tukey
- Scheffé
- Bonferroni

Comparar os grupos dois a dois por meio de intervalos de confiança para a diferença.

Se o intervalo não contém o zero, podemos obter conclusões sobre a razão da rejeição.

Comparação entre os métodos

a) Tukey deve ser adotado quando tivermos interesse em todas as possíveis comparações de médias duas a duas. Quando o no. for pequeno em relação a k(k-1)/2, Bonferroni é mais preciso que o Tukey.

b) Scheffé deve ser adotado quando temos interesse em comparações envolvendo mais de duas médias

Análise de Resíduos

Verificar se o modelo adotado foi adequado

O resíduo da observação y_{ij} é definido como:

$\boxed{y_{ij} \text{ - média amostral do grupo =}}$ $\mathbf{y_{ij} - \bar{y}_{i.}}$

A média dos resíduos é zero, e a variância é a mesma das observações.

A análise descritiva dos resíduos pode sugerir a validade das suposições de Normalidade , Igualdade de Variâncias e Independência (quando dispusermos da ordem em que as observações foram obtidas)

Teste de Igualdade de Variâncias

Teste de Bartlett (distribuição normal)

Teste de Levene (qualquer distribuição contínua)

Teste de Normalidade

Uma forma de se verificar descritivamente a suposição de normalidade das observações, é construir o gráfico de probabilidade normal dos resíduos

Desvios das Suposições

Se as suposições de **Normalidade ou Igualdade de Variâncias** não estiverem satisfeitas, podem ser feitas transformações nos dados.

No caso de não ser encontrada uma transformação adequada, podem ser adotadas técnicas não paramétricas

Fugas da Normalidade

⇒ ***garantir a validade da distribuição F***

O modelo de ANOVA é robusto

- **Teorema Limite Central (Distr. Amostral da Média)**
- **Em casos extremos** ⇒ Testes Não Paramétricos

Heterocedasticidade

- **Transformação dos dados originais**
- **Testes não paramétricos**
- **Utilização de modelos mais gerais**

Lembrar que "heterocedasticidade" já é uma diferença importante entre os grupos

Transformação de Variáveis

- σ_i^2 **proporcional a** μ_i $\boxed{Y' = \sqrt{Y}}$ **ou** $\boxed{Y' = \sqrt{Y} + \sqrt{Y+1}}$

- σ_i **proporcional a** μ_i $\boxed{Y' = \log Y}$

- σ_i^2 **proporcional a** μ_i^2 $\boxed{Y' = \dfrac{1}{Y}}$

Fugas da Independência

Ajuste de Modelos Específicos

Situações Experimentais

Medidas Repetidas
Dados Longitudinais

A) Comparação das glicemias médias na população de pacientes submetidos à revascularização do miocárdio com a utilização de circulação extracorpórea nos períodos, pré-operatório, primeiro e segundo dias do pós-operatório

B) Comparação de graus médios de melhora em pacientes Esquizofrênicos ou Depressivos submetidos a três tipos de tratamento