Análise de Dados Longitudinais
Aula 01.08.2018

José Luiz Padilha da Silva - UFPR
www.docs.ufpr.br/∼jlpadilha

Sumário

Comparação de duas Médias

Retomar a Comparação dos Colírios A e B

- Pacientes com pressão intra-ocular (PIO) elevada irão participar do estudo.
- A pressão será medida após dois meses de uso do colírio.
- O objetivo é comparar a redução média de PIO dos dois colírios.

Então, queremos o seguinte:

$$\delta = \mu_A - \mu_B.$$

O interesse é então testar a hipótese:

$$H_0 : \delta = 0$$

Comparação de Médias

- Existem duas formas de conduzir o estudo:
  - 50 pacientes são submetidos ao colírio A e ao colírio B (medidas repetidas). Considera-se uma período de descanso de dois meses entre a aplicação dos colírios. É indicado aleatorizar a ordem de aplicação de A e B.
  - 100 pacientes são selecionados e 50 são sorteados para receber o colírio A e os demais recebem o B.
- Ambos estudos são experimentais
  - Pareado: Estudo Cross-over
  - Amostras Independentes: Estudo Clínico Aleatorizado.
- Qual forma você utilizaria?

Amostra Pareada ou Independente?

1. Vantagens de Parear as Amostras
    - Controlar por possíveis fatores de confusão.
    - Menos pacientes/unidades na amostra.
    - Teste mais preciso com menos suposições.
    - Controla pelo efeito de coorte.
2. Vantagens de Amostras Independentes
    - Dados são obtidos de forma mais rápida.

Amostras Pareada ou Independente?

Quando devemos parear?

SEMPRE (que for possível).

- Caso típico: antes e depois.
- Situações impossíveis: fumantes e não-fumantes, etc.

Teste-t pareado

O objetivo é comparar duas medidas pareadas.

$$\delta = \mu_A - \mu_B.$$

Uma estimativa natural para δ é a diferença das médias. Ou seja

$$\widehat{\delta} = \widehat{\mu}_A - \widehat{\mu}_B.$$

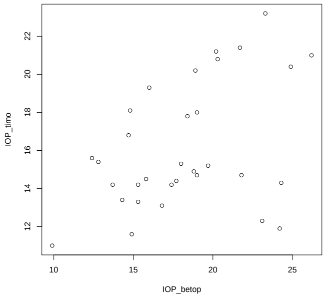
A variância de $\widehat{\delta}$ é

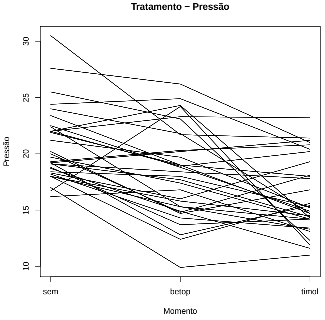
$$Var(\widehat{\delta}) = \frac{1}{N}(\sigma_A^2 + \sigma_B^2 - 2\sigma_{AB})$$

Teste-t pareado

Usualmente dados longitudinais têm correlação positiva. Ou seja $\sigma_{AB} > 0$.

Isto significa que a estatística a ser utilizada tem menor variância do que aquela obtida com dados independentes.
Exemplo: Colírio A: Timoptol (timo) e Colírio B: Betoptic (cor=0,43).





Teste-t pareado

Considere as diferenças:

$$d_i = y_{i1} - y_{i2} \quad i = 1, \ldots, n.$$

A estatística é:

$$t = \frac{\overline{d}}{s_d/\sqrt{n}},$$

que, sob H_0, tem uma distribuição t com $n - 1$ graus de liberdade.

Suposição: d_i vem de uma distribuição normal.

Exemplo Colírio[1]: $t = 2.9934$, df = 31, p-value = 0.005378.

[1] o teste somente é válido se não houver efeito de período e de "carry-over"

Exemplo Colírio

```
> head(colirio)
   PRONT. IDADE SEXO IOP_sem IOP_betop IOP_timo Ordem
1    6649    75    F    19.1      17.7     14.4     0
2    3106    61    M    30.5      21.7     21.4     1
3   15231    57    F    19.1      18.4     17.8     0
4     799    42    F    20.0      14.8     18.1     1
5    9371    59    M    24.0      21.8     14.7     0
6     757    65    M    19.3      20.3     20.8     1

> colMeans(cbind(colirio$IOP_beto,colirio$IOP_timo))
[1] 18.2375 16.1375
> var(cbind(colirio$IOP_beto,colirio$IOP_timo))
         [,1]     [,2]
[1,] 16.20242  5.65629
[2,]  5.65629 10.85919
```

- Calcule e discuta estatística *t* considerando e desconsiderando o pareamento.

Comparação de mais de duas Médias

Comparação dos Colírios A e B e C

- Pacientes com pressão intra-ocular elevada irão participar do estudo.
- A pressão será medida após dois meses de uso do colírio.
- O objetivo é comparar a redução média dos três colírios.

Então, queremos testar a seguinte hipótese:

$$H_0 : \mu_A = \mu_B = \mu_C.$$

ANOVA é válido?

Análise de Dados Longitudinais

1. Características:
   - As respostas de diferentes unidades são independentes;
   - As respostas para a mesma unidade são correlacionadas. De uma forma geral, as respostas próximas no tempo devem ser mais correlacionadas.
2. Medida Temporal
   - Idade;
   - Calendário medido a partir de um certo evento.
3. Objetivos do Estudo:
   - avaliar o comportamento temporal;
   - avaliar o efeito de covariáveis sobre a resposta;
   - predição.
4. Modelos de Regressão
   - Modelos marginais (modelar a média e a estrutura de covariância);
   - Modelo de efeitos aleatórios.
   - Modelo de transição.

Características da Correlação dos Dados

- As correlações usualmente são positivas;
- as correlações usualmente diminuem à medida que aumenta a separação no tempo;
- as correlações entre medidas repetidas raramente aproximam do zero.
- medidas muito próximas tendem a ter correlação um.

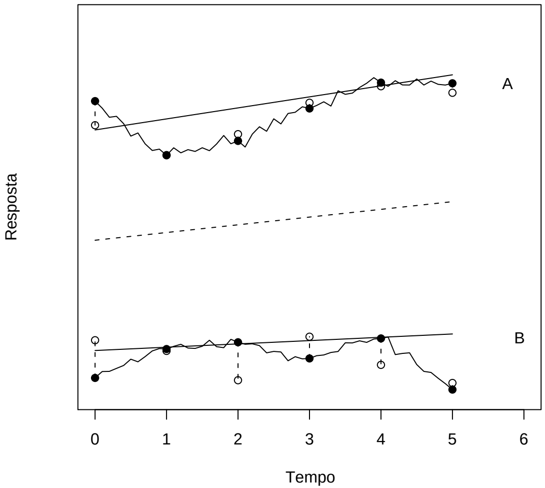
Fontes de Variabilidade em Estudos Longitudinais

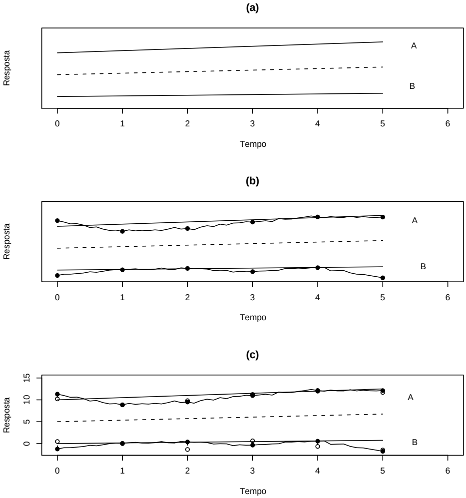
- Variação entre-unidades;
- Variação intra-unidade;
- Erro de medição.

Fontes de Variabilidade em Estudos Longitudinais

Estas três fontes de variação podem ser visualizadas de forma gráfica.

- pontos pretos são respostas livre de erro de medição;
- pontos brancos são as respostas observadas;
- A e B são diferentes indivíduos.


A
B
Resposta
0
1
2
3
4
5
6
Tempo


(a)
Resposta
A
B
0
1
2
3
4
5
6
Tempo
(b)
Resposta
A
B
0
1
2
3
4
5
6
Tempo
(c)
Resposta
15
10
5
0
A
B
0
1
2
3
4
5
6
Tempo

Notação

Seja Y_{ij} a variável resposta para i-ésimo indivíduo ($i = 1, \ldots, N$) na j-ésima ocasião[2] ($j = 1, \ldots, n$).

Dado que temos n medidas repetidas da resposta no mesmo indivíduo, podemos agrupá-las em um vetor $n \times 1$, denotado por

$$Y_i = \begin{pmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{in} \end{pmatrix},$$

ou, por conveniência,

$$Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{in})'.$$

[2] Mais tarde estenderemos a notação para o caso desbalanceado.

Dependência e Correlação

O principal interesse está na média da resposta (em particular, em mudanças da média no tempo e como esta mudança depende de covariáveis).

Denote a média ou esperança de cada resposta Y_{ij} por $\mu_j = E(Y_{ij})$. Adicionalmente, para permitir que a resposta média varie de indivíduo para indivíduo como função de covariáveis medidas em nível de indivíduo, requeremos $\mu_{ij} = E(Y_{ij})$.

Denotando a média condicional de Y_{ij} por μ_{ij}, a variância condicional de Y_{ij} é definida como:

$$\sigma_j^2 = E\left[Y_{ij} - E(Y_{ij})\right]^2 = E\left(Y_{ij} - \mu_{ij}\right)^2.$$

Dependência e Correlação

A *covariância* condicional entre as respostas em duas ocasiões diferentes, digamos Y_{ij} e Y_{ik}, é definida por

$$\sigma_{jk} = E\left[(Y_{ij} - \mu_{ij})(Y_{ik} - \mu_{ik})\right].$$

e fornece uma medida da dependência *linear* entre Y_{ij} e Y_{ik}, dado as covariáveis.

A correlação condicional entre Y_{ij} e Y_{ik} é denotada por

$$\rho_{jk} = \frac{E\left[(Y_{ij} - \mu_{ij})(Y_{ik} - \mu_{ik})\right]}{\sigma_j \sigma_k},$$

que, por definição, assume valores entre -1 e $+1$.

Dependência e Correlação

Definimos a matriz de variância-covariância como segue

$$
Cov\begin{pmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{in} \end{pmatrix} = \begin{pmatrix} Var(Y_{i1}) & Cov(Y_{i1}, Y_{i2}) & \dots & Cov(Y_{i1}, Y_{in}) \\ Cov(Y_{i2}, Y_{i1}) & Var(Y_{i2}) & \dots & Cov(Y_{i2}, Y_{in}) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(Y_{in}, Y_{i1}) & Cov(Y_{in}, Y_{i2}) & \dots & Var(Y_{in}) \end{pmatrix}
$$

$$
= \begin{pmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \dots & \sigma_{nn} \end{pmatrix}.
$$

Assumimos que as variâncias e covariâncias são constantes. Note que há simetria, ou seja, $Cov(Y_{ij}, Y_{ik}) = \sigma_{jk} = \sigma_{kj} = Cov(Y_{ik}, Y_{ij})$.

Dependência e Correlação

Usaremos frequentemente a notação

$$Cov(Y_i) = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \dots & \sigma_n^2 \end{pmatrix}.$$

Definimos a matriz de correlação em termos de

$$Corr(Y_i) = \begin{pmatrix} 1 & \rho_{12} & \dots & \rho_{1n} \\ \rho_{21} & 1 & \dots & \rho_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n1} & \rho_{n2} & \dots & 1 \end{pmatrix},$$

que é simétrica, ou seja, $Corr(Y_{ij}, Y_{ik}) = \rho_{jk} = \rho_{kj} = Corr(Y_{ik}, Y_{ij})$.

Modelo Marginal

O modelo para a resposta média em cada ocasião não incorpora a dependência sobre nenhum efeito aleatório ou sobre respostas anteriores.

Apropriado quando o foco da análise é inferir sobre a população média.

1. $E(Y_{ij}|X_{ij}) = \mu_{ij}$, com $g(\mu_{ij}) = \eta_{ij} = X_{ij}'\beta$.
2. $Var(Y_{ij}|X_{ij}) = \phi v(\mu_{ij})$, em que ϕ é um parâmetro de dispersão e $v(\cdot)$ é uma função conhecida da média.
3. A correlação intra-indivíduos é função de α. Por exemplo:
   - $Corr(Y_{ij}, Y_{ik}) = \alpha^{|k-j|}$ (AR-1 para respostas contínuas);
   - $logOR(Y_{ij}, Y_{ik}) = \alpha_{jk}$ (não-estruturado para respostas categóricas).

Modelos Mistos

Incluem efeitos aleatórios no modelo de efeitos fixos, em nível de indivíduo, modelando a heterogeneidade entre indivíduos e induzindo, assim, uma estrutura de covariância entre as respostas repetidas.

1. $E(Y_{ij}|X_{ij}, b_i) = \mu_{ij}$, com $g(\mu_{ij}) = \eta_{ij} = X_{ij}'\beta + Z_{ij}'b_i$, sendo b_i o efeito aleatório associado com Y_i.
2. $Var(Y_{ij}|X_{ij}, b_i) = \phi v(\mu_{ij})$.
3. Geralmente assume-se $b_i \sim N_q(0, G)$.

Modelos de Transição

A distribuição condicional de Y_{ij} é descrita como uma função explícita das respostas passadas e de um vetor de variáveis preditoras.

1. $E(Y_{ij}|X_{ij}, H_{ij}) = \mu_{ij}$, com $H_{ij} = \{Y_{id}, d = 1, \ldots, j-1\}$. $g(\mu) = X_{ij}'\beta + \sum_{q=1}^{Q} f_q(H_{ij}, \alpha)$, em que $f_q(\cdot)$ são funções conhecidas.
2. $Var(Y_{ij}|X_{ij}, H_{ij}) = \phi v(\mu_{ij})$.
3. A correlação entre $Y_{i1}, \ldots, Y_{in}$ é avaliada através do parâmetro α que aparece na função $f_q(\cdot)$.