Comparação entre duas médias

Em aplicações práticas é comum que o interesse seja comparar as médias de duas diferentes populações (ambas as médias são desconhecidas).

Na comparação de duas populações, dispomos de duas amostras, em que são possíveis as seguintes situações:

2 amostras
- dependentes
- independentes
  - $\Rightarrow$ variâncias conhecidas
  - $\Rightarrow$ variâncias desconhecidas
    - $\Rightarrow$ iguais
    - $\Rightarrow$ diferentes

Discutiremos apenas os testes conhecidos como *paramétricos*, que assumem que as variáveis se comportam segundo um *modelo Normal*, ou que as *amostras são suficientemente grandes*, de modo que uma boa aproximação pode ser conseguida utilizando o modelo Normal.

Amostras Dependentes (teste t-pareado)

$\rightarrow$ característica das amostras dependentes (pareadas): *para cada unidade amostral realizamos duas medições.*

- As medidas são tomadas em um único indivíduo em dois pontos distintos no tempo.
- Em geral, *observações pareadas* correspondem a medidas tomadas *antes* e *depois* de uma dada intervenção -- cada indivíduo é examinado antes que um certo tratamento seja aplicado e novamente depois que o tratamento foi completado.
- Outro tipo de emparelhamento: o pesquisador “casa” os indivíduos de um grupo com aqueles de um segundo grupo de modo que os membros de um par sejam parecidos (em relação a características tais como a idade e o gênero).

Empregado na tentativa de se controlar fontes de variação que poderiam influenciar os resultados da comparação.

Se as medidas são feitas no mesmo sujeito uma certa variabilidade biológica é eliminada -- não temos que nos preocupar com o fato de um sujeito ser mais velho do que outro ou se um é homem e o outro é mulher.

A intenção do emparelhamento é, portanto, fazer uma comparação mais precisa.

Exemplo 1: Uma companhia farmacêutica está interessada em investigar uma nova droga, que talvez tenha a propriedade de baixar a taxa de colesterol (TC). As determinações foram feitas em mg/100ml antes e depois de o tratamento ser ministrado em cada indivíduo.

Os resultados de um experimento realizado com 6 indivíduos, aleatoriamente escolhidos, foram:

	1	2	3	4	5	6
TC antes	217	252	229	200	209	213
TC depois	209	241	230	208	206	211

População 1 (TC sem tratamento): a taxa média é μ_1(desconhecida); amostra (6 pacientes): taxa média de colesterol antes é $\bar{x}_1 = 219$.

População 2 (TC com tratamento): a taxa média é μ_2(desconhecida); amostra (6 pacientes): taxa média de colesterol depois é $\bar{x}_2 = 217{,}5$.

As medidas tomadas antes e após a intervenção realizada serão representadas pelas variáveis aleatórias X_i e Y_i, repectivamente.

O efeito produzido para o i-ésimo indivíduo pode ser representado pela variável $D_i = X_i - Y_i$ ("antes" – "depois")

Supondo $D_i \sim N(\mu_D, \sigma_D^2)$, para i = 1, ..., *n*,

numa situação geral, queremos testar as hipóteses:

H: $\mu_D = 0$ ⇒ a intervenção não produz efeito

A: $\mu_D \neq 0$ ou
$\mu_D < 0$ ou ⇒ a intervenção produz algum efeito
$\mu_D > 0$

O parâmetro μ_D é estimado pela média amostral das diferenças

$$\overline{D} = \frac{1}{n}\sum_{i=1}^{n} D_i$$

Como não temos informação sobre a variância das diferenças, estimamos seu valor por S_D^2, dado por

$$S_D^2 = \frac{1}{n-1}\sum_{i=1}^{n}(D_i - \overline{D})^2$$

Estatística do teste: $$T = \frac{\overline{D} - \mu_D}{S_D / \sqrt{n}}$$

Sob H_0, T tem distribuição *t*-Student com *n*-1 graus de liberdade.

Observando os dados dos 6 indivíduos temos que para alguns a medida de colesterol depois do tratamento diminui e para outros aumenta.

$\rightarrow$ Gostaríamos de saber se há alguma evidência estatística de que a nova droga realmente reduz a taxa de colesterol.

teste unilateral ao nível de significância de 0,05.

- Hipóteses: $H : \mu_{\mathrm{D}} = 0$

  $A : \mu_{\mathrm{D}} > 0$

$\rightarrow$ Efeito do emparelhamento: eliminar quaisquer distorções que poderiam ser introduzidas ao se comparar pacientes que diferem com relação à idade, peso ou severidade da doença da artéria coronária.

Suponha que os dois grupos de observações possam ser dispostos como a seguir:

Amostra 1	Amostra 2	$d_i = x_i - y_i$
x_1	y_1	$x_1 - y_1$
x_2	y_2	$x_2 - y_2$
...	...	...
x_n	y_n	$x_n - y_n$

variável de interesse: $d_i = x_i - y_i$

Amostra de pares $\Rightarrow$ $d_i = x_i - y_i$: 8, 11, -1, -8, 3, 2

$$\bar{d} = \frac{\sum_{i=1}^{6} d_i}{6} = 2{,}5$$ (média amostral das diferenças)

$$s_D = \sqrt{\frac{\sum_{i=1}^{6} (d_i - \bar{d})^2}{6-1}} = 6{,}71$$ (desvio padrão das diferenças)

• a média da amostra fornece uma estimativa por ponto para a verdadeira diferença das médias das populações $\mu_D = \mu_1 - \mu_2$.

• é razoável supormos que taxa de colesterol tem distribuição normal e, conseqüentemente, podemos considerar que a distribuição das diferenças tem distribuição normal.

Obs.: no caso geral, é necessário uma verificação da suposição de normalidade através de análise gráfica e/ou testes de hipóteses.

As medidas tomadas antes e após a intervenção realizada serão representadas pelas variáveis aleatórias X_i e Y_i, repectivamente.

O efeito produzido para o i-ésimo indivíduo pode ser representado pela variável $D_i = X_i - Y_i$ (“antes” – “depois”)

Supondo $D_i \sim N(\mu_D, \sigma_D^2)$, para i = 1, ..., n,

queremos testar as hipóteses:

$$\begin{aligned} &\boldsymbol{H}: \mu_D = 0 && \Rightarrow \text{a intervenção não produz efeito} \\ &\boldsymbol{A}: \mu_D \neq 0 \text{ ou} \\ &\quad\;\; \mu_D < 0 \text{ ou} && \Rightarrow \text{a intervenção produz algum efeito} \\ &\quad\;\; \mu_D > 0 \end{aligned}$$

Estatística:

$$T = \frac{\overline{D} - \mu_D}{S_D \big/ \sqrt{n}}$$

Sob a hipótese nula *H*, T tem distribuição *t*-Student com 6 -1 = 5 graus de liberdade.

Substituindo valores da amostra resulta:

$$t = \frac{2{,}5 - 0}{6{,}71 \big/ \sqrt{6}} = 0{,}91$$

Nível descritivo:

$P(T \geq 0{,}91) \cong 0{,}20 \quad \Rightarrow$ não rejeitamos *H*.

$\Rightarrow$ não há evidência experimental para concluirmos que a nova droga diminui a taxa de colesterol em indivíduos que a utilizam.

Se a hipótese nula *H* é rejeitada:

Interesse: Encontrar um intervalo de confiança para μ_D

$$IC(\mu_D; (1-\alpha)\%) = \left(\bar{d} - t_{n-1;\,\alpha/2}\frac{s_D}{\sqrt{n}};\ \bar{d} + t_{n-1;\,\alpha/2}\frac{s_D}{\sqrt{n}}\right)$$

$$\begin{aligned} IC(\mu_D; 90\%) &= \left(2{,}5 - 2{,}015 \times \frac{6{,}71}{\sqrt{6}};\ 2{,}5 + 2{,}015 \times \frac{6{,}71}{\sqrt{6}}\right) \\ &= (2{,}5 - 5{,}52;\ 2{,}5 + 5{,}52) \\ &= (-3{,}02;\ 3{,}02) \end{aligned}$$