Comparação entre duas médias (cont.)

Amostras Independentes

Exemplo: Medidas de nível sérico de ferro para 2 amostras de crianças: um grupo saudável e outro que sofre de fibrose cística(doença congênita das glândulas das mucosas)

As duas populações, de onde as amostras são provenientes, são *independentes* e *normalmente distribuídas*;

- a população das *crianças doentes* tem nível sérico médio de ferro $\mu_1 \Rightarrow X_1 \sim N(\mu_1, \sigma_1^2)$.
- a população das *crianças saudáveis* tem nível sérico médio de ferro $\mu_2 \Rightarrow X_2 \sim N(\mu_2, \sigma_2^2)$.

$\rightarrow$ Interesse: *Comparar as médias das duas populações*.

Hipóteses estatísticas:

$H: \mu_1 = \mu_2$

$A: \mu_1 \neq \mu_2$
ou $\mu_1 > \mu_2$
ou $\mu_1 < \mu_2$

ou, equivalentemente, usando diferenças $\Rightarrow$

$H: \mu_1 - \mu_2 = 0$

$A: \mu_1 - \mu_2 \neq 0$
ou $\mu_1 - \mu_2 > 0$
ou $\mu_1 - \mu_2 < 0$

$\rightarrow$ da pop. normal com média μ_1 e desvio padrão $\sigma_1 \Rightarrow$ extrai-se uma a.a. de tamanho $n_1 \Rightarrow$ $\bar{x}_1$: média da amostra 1

s_1 : desvio padrão da amostra 1

$\rightarrow$ da pop. normal com média μ_2 e desvio padrão $\sigma_2 \Rightarrow$ extrai-se uma a.a. de tamanho $n_2 \Rightarrow$ $\bar{x}_2$: média da amostra 2

s_2 : desvio padrão da amostra 2

Obs.: note que os números de observações nas 2 amostras, n_1 e n_2 não precisam ser iguais.

		grupo 1	grupo 2
população	média	μ_1	μ_2
	desvio padrão	σ_1	σ_2
amostra	média	$\bar{x}_1$	$\bar{x}_2$
	desvio padrão	s_1	s_2
	tamanho	n_1	n_2

Situações possíveis com respeito às variâncias σ_1^2 e σ_2^2:

1. conhecidas: *teste Z*
2. desconhecidas:
   - iguais: *teste-t de duas amostras.*
   - diferentes: *teste-t modificado.*

Obs.: O teste de comparação de variâncias pode ser utilizado como um procedimento preliminar em teste de comparação de médias, auxiliando a escolha da técnica adequada.

- Hipóteses: $H : \mu_1 - \mu_2 = 0$

  $A : \mu_1 - \mu_2 \neq 0$

- Estimador de $\mu_1 - \mu_2$: $\overline{X}_1 - \overline{X}_2$

- Distribuição do estimador:

  Como X_1 e X_2 são independentes com distribuição normal com médias μ_1 e μ_2 e desvio padrão σ_1^2 e σ_2^2, respectivamente, então

$$\overline{X}_1 - \overline{X}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right),$$

  resultando

$$Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

- Situação 1:
  as variâncias das populações são iguais: $\sigma_1^2 = \sigma_2^2 = \sigma^2$

$$\Rightarrow \quad \overline{X}_1 - \overline{X}_2 \sim N(\mu_1 - \mu_2, \sigma^2(\frac{1}{n_1} + \frac{1}{n_2})) ,$$

resultando,

$$Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma^2(\frac{1}{n_1} + \frac{1}{n_2})}} \sim N(0,1)$$

É comum que o valor verdadeiro de σ^2 não seja conhecido. Nesse caso, precisamos estimá-lo.

- estatística do teste:

$$T = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{(n_1+n_2-2)} \quad ,$$

sendo $s_p^2 = \dfrac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$.

- A estimativa s_p^2 combina informação de ambas amostras para se produzir uma estimativa mais confiável de σ^2;
- Na verdade, s_p^2 é média ponderada das duas variâncias amostrais s_1^2 e s_2^2, onde cada variância é ponderada pelos seus graus de liberdade associados;
- Se n_1 é igual a n_2, s_p^2 é a média aritmética simples; caso contrário, maior peso é dado à variância da maior amostra.

- Distribuição de *T*:

  Sob a hipótese nula *H*: μ_1-μ_2 = 0, a estatística de teste *T* tem uma distribuição *t-student* com (n_1+n_2-2) g.l.

  $\Rightarrow$ A partir dessa distribuição podemos encontrar *P* (nível descritivo), para verificar a significância da diferença amostral $\bar{x}_1 - \bar{x}_2$, quando μ_1=μ_2.

  Como concluir ?

  Se ***P*** $\leq \alpha$, rejeitamos a hipótese nula *H*.

  Se ***P*** > α, não rejeitamos *H*.

Voltando ao exemplo, para as hipóteses

$$H : \mu_1 - \mu_2 = 0$$
$$A : \mu_1 - \mu_2 \neq 0$$

Uma a. a. é selecionada de cada população:

$\rightarrow n_1$=9 crianças sadias $\Rightarrow \bar{x}_1$= 18,9 μmol/l e s_1 = 5,9 μmol/l

$\rightarrow n_2$=13 crianças com fibrose cística $\Rightarrow \bar{x}_2$= 11,9 μmol/l e s_1 = 6,3 μmol/l.

Pergunta: É provável que a diferença observada nas médias das amostras -- 18,9 *versus* 11,9 μmol/l -- seja o resultado de uma variação ao acaso, ou devemos concluir que a discrepância seja devida a uma verdadeira diferença das médias das populações?

Comentário:

Algumas vezes inicia-se uma análise construindo um IC separado para a média de cada população; por ex., IC de 95% para os níveis séricos médios de ferro de crianças com e sem fibrose cística;

$\Rightarrow$ em geral, se os dois intervalos não se sobrepõem, isso sugere que as médias das populações são diferentes; no entanto, essa técnica não é um teste de hipótese formal.

Em nosso exemplo, há uma pequena quantidade de sobreposição entre os intervalos; conseqüentemente, não é possível extrair qualquer tipo de conclusão significativa.

$\rightarrow$ No exemplo, notar que as 2 amostras de crianças foram aleatoriamente selecionadas de populações normais distintas; assumiu-se, ainda, que as variâncias das populações são iguais e desconhecidas $\Rightarrow$ o *teste-t de duas amostras* é a técnica apropriada.

teste bilateral e nível de significância $\alpha = 0{,}05$.

$$\bar{x}_1 - \bar{x}_2 = 18{,}9 - 11{,}9 = 7{,}0$$

$$s_p^2 = \frac{(9-1)5{,}9^2 + (13-1)6{,}3^2}{9+13-2} = 37{,}74$$

$$\Rightarrow \quad t = \frac{7-0}{\sqrt{37{,}74\left(\frac{1}{9}+\frac{1}{13}\right)}} = 2{,}63$$

nível descritivo (usando *t-Student* com 9+13-2= 20 g.l.):
$P = 2 \times P(T \geq 2{,}63) = 2 \times 0{,}008 = 0{,}016 < 0{,}05 \Rightarrow$ rejeitamos *H.*

$\Rightarrow$ A diferença entre o nível sérico médio de ferro de crianças saudáveis e o nível médio de crianças com fibrose cística é estatisticamente significante.
Baseado nessas amostras, parece que as crianças com fibrose cística sofrem de uma deficiência de ferro.

A quantidade $\bar{x}_1 - \bar{x}_2$ fornece uma estimativa por ponto para $\mu_1 - \mu_2$;

- Intervalo de 95% confiança para $\mu_1 - \mu_2$:

$$((18{,}9 - 11{,}97) \mp 2{,}086\sqrt{37{,}74[\frac{1}{9} + \frac{1}{13}]}) =$$

$$(1{,}4\,;12{,}6)$$

$\Rightarrow$ Diferentemente dos intervalos separados para cada média, esse intervalo de confiança para a diferença das médias é equivalente ao teste de hipótese de duas amostras conduzido ao nível de 0,05.
(Note que, como esperado, o intervalo não contém o valor *zero*.

- Situação 2:

  as variâncias das populações são desiguais: $\sigma_1^2 \neq \sigma_2^2$

Nesse caso, aplica-se uma modificação do teste-*t* de duas amostras. Ao invés de usar s_p^2 como uma estimativa da variância comum σ^2, estimamos σ_1^2 por s_1^2 e σ_2^2 por s_2^2.

$\rightarrow$ A estatística do teste apropriada é:

$$T = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$\Rightarrow T$ tem distribuição *t-Student* com graus de libertade ν, sendo

$$\nu = \frac{[(s_1^2/n_1) + (s_2^2/n_2)]^2}{[(s_1^2/n_1)^2/(n_1 - 1) + (s_2^2/n_2)^2/(n_2 - 1)]},$$

e o valor de ν é, então, arredondado para baixo para o inteiro mais próximo.