“Noções de Estatística”
disciplinas MAE0116 e MAE0110 da USP
Assuno da aula: Comparação entre médias de populações

Ministrante Prof. Dr. Vladimir Belitsky,
IME-USP

14 de junho de 2023

O formalismo geral.

Há duas populações. Suponha que nos interessa o atributo "altura" (denotada por A abaixo) de cada população.
A suposição básica sobre a distribuição populacional:

$$A_1 \sim \mathcal{N}(\mu_1, \sigma_1^2), \quad A_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$$

Temos amostras retiradas das duas populações:

$$a_{1,1}, \ldots, a_{1,n} \text{ e } a_{2,1}, \ldots, a_{2,m}$$

Desejamos (é o que nos interessa) testar a hipótese de igualdade das médias (populacionais)

$$H_0 : \mu_1 = \mu_2$$

contra a desuguladade entre elas, que pode ser da forma

$$H_A : \mu_1 \neq \mu_2$$

ou da forma

$$H_A : \mu_1 < \mu_2 (\text{ ou } \mu_1 > \mu_2)$$

O caso quando as variâncias são conhecidas.

Embora parece impossível na prática, mas vamos conversar sobre o caso quando as variâncias (σ_1^2 e σ_2^2) são conhecidas. (O impossível aqui é imaginar que as médias são desconhecidas–pois se fossem conhecidas, então não teriamos problema em dizer se são iguais ou não,–mas, ao mesmo tempo, as variâncias são conhecidas.)
Ao denotar (recorde, na Estatística, o chapéu $\hat{\cdot}$ significa a estimativa do parámetro feito com base na amostra)

$$\hat{\mu_1} = \frac{1}{n}\left(a_{1,1} + \ldots + a_{1,n}\right), \ \hat{\mu_2} = \frac{1}{m}\left(a_{2,1} + \ldots + a_{2,m}\right)$$

devemos calcular o seguinte valor:

$$z = \frac{\hat{\mu_1} - \hat{\mu_2}}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}}$$

e usar a Distribuição Normal Padrão para calcular o limiar correspondente ao nível de significância desejado do teste e comparar o limiar achado com z.

O caso quando as variâncias são desconhecidas, mas um teste predecessor qualquer indicou que elas são iguais entre si (1/2).

Nesse caso, além das estimativas de médias, precisamos calcular, a fim de construir o valor que determinará a resposta de nosso teste, as estimativas das variâncias:

$$\hat{\sigma_1^2} = \frac{1}{n-1}\sum_{i=1}^{n}(a_{1,i} - \hat{\mu_1})^2\,,\ \hat{\sigma_2^2} = \frac{1}{m-1}\sum_{j=1}^{m}(a_{2,j} - \hat{\mu_2})^2$$

No passo de solução seguinte, usamos as duas estimativas para estimar o valor comum das variâncias; o resultado denota-se por $\hat{\sigma^2}$:

$$\hat{\sigma^2} = \frac{(n-1)\hat{\sigma_1^2} + (m-1)\hat{\sigma_2^2}}{n+m-2}$$

O caso quando as variâncias são desconhecidas, mas um teste predecessor qualquer indicou que elas são iguais entre si (2/2).

Então, o valor

$$t = \frac{\hat{\mu_1} - \hat{\mu_2}}{\sqrt{\hat{\sigma^2}\left(\frac{1}{n} + \frac{1}{m}\right)}}$$

responderá na questão da aceitação/rejeição da hipótes H_0 quando comparado com o limiar obtido da distribuição t-de-Student, cujo número de graus de liberdade é $(n + m - 2)$, de acordo com o nível desejado de significância do teste.

O caso quando as variâncias são desconhecidas e quando um teste predecessor qualquer indicou que elas não são iguais entre si (1/2).

Nesse caso, devemos construir, a fim de posterior uso que rejeitará ou não a hipótese H_0, o seguinte valor

$$\tau = \frac{\hat{\mu_1} - \hat{\mu_2}}{\sqrt{\frac{\hat{\sigma_1^2}}{n} + \frac{\hat{\sigma_2^2}}{m}}}$$

Além de τ, devemos calcular o seguinte valor que auxiliará no cálculo dos graus de liberdade da distribuição t-de-Student:

$$\nu = \frac{(A+B)^2}{\left(\frac{A^2}{n-1} + \frac{B^2}{m-1}\right)}$$

onde

$$A = \frac{\hat{\sigma_1^2}}{n}, \quad B = \frac{\hat{\sigma_2^2}}{m}$$

O caso quando as variâncias são desconhecidas e quando um teste predecessor qualquer indicou que elas não são iguais entre si (2/2).

O valor ν deve ser arredondado para o inteiro mais próximo (a proximidade vai guiar se o arredondamento será para mais ou para menos)

Em seguida, deve ser tomada a distribuição t-de-Student com $\nu - \textit{arredondado}$ número de graus de liberdade. Essa distribuição e o nível de significância desejado do teste determinarão limiar. A comparação do limiar com o valor τ acarreta a rejeição ou não da hipótese H_0.

t Table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002
df										
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485

Métodos não paramétricos (1/2).

Observe que no caso quando as variâncias são iguais, se confirmarmos a hipótese $\mu_1 = \mu_2$, poderemos então concluir que as populações têm a mesma distribuição (do atributo altura).
A pergunta a ser tratada agora é: "Como obter uma conclusão sobre a semelhança de duas distribuições sem saber nada sobre sua forma?"
Nesse caso usam-se métodos paramétricos. Eis um de exemplos (fonte "EStatística Básica de W. de O.Bussab e P. A. Morettin).
Duas populações: crianças ensinadas matemática por método tradicional e criancas ensinadas por um novo método.

Métodos não paramétricos (2/2).

Tomamos, ao acaso, 3 das crianças da população T ("tratamento" – novo método) e 2 crianças da população C ("controle" – metodos de ensino tradicional). Comparamos seu desempenho (aplicando o mesmo teste) e as ordenamos.
Naturalmente, se o resultado de ordenação der

C C T T T

concluiremos que o novo método de ensino é mais eficiente, e se der

T T T C C

concluiremos que o método tradicional é mais eficiente.

O problema de emissão de decisão é quando o resultado é algo assim, por exemplo:

C T T T C

Para a solução existe o teste de Wilcoxon.

Comparação de médias de diversas populações usando a análise de variância (1/4).

Tomei 4 amostras de tamanho 5 cada. Cada amostra adveio da população $\mathcal{N}(0, 1)$.
Eis os valores (cada linha corresponde a uma das 4 amostras

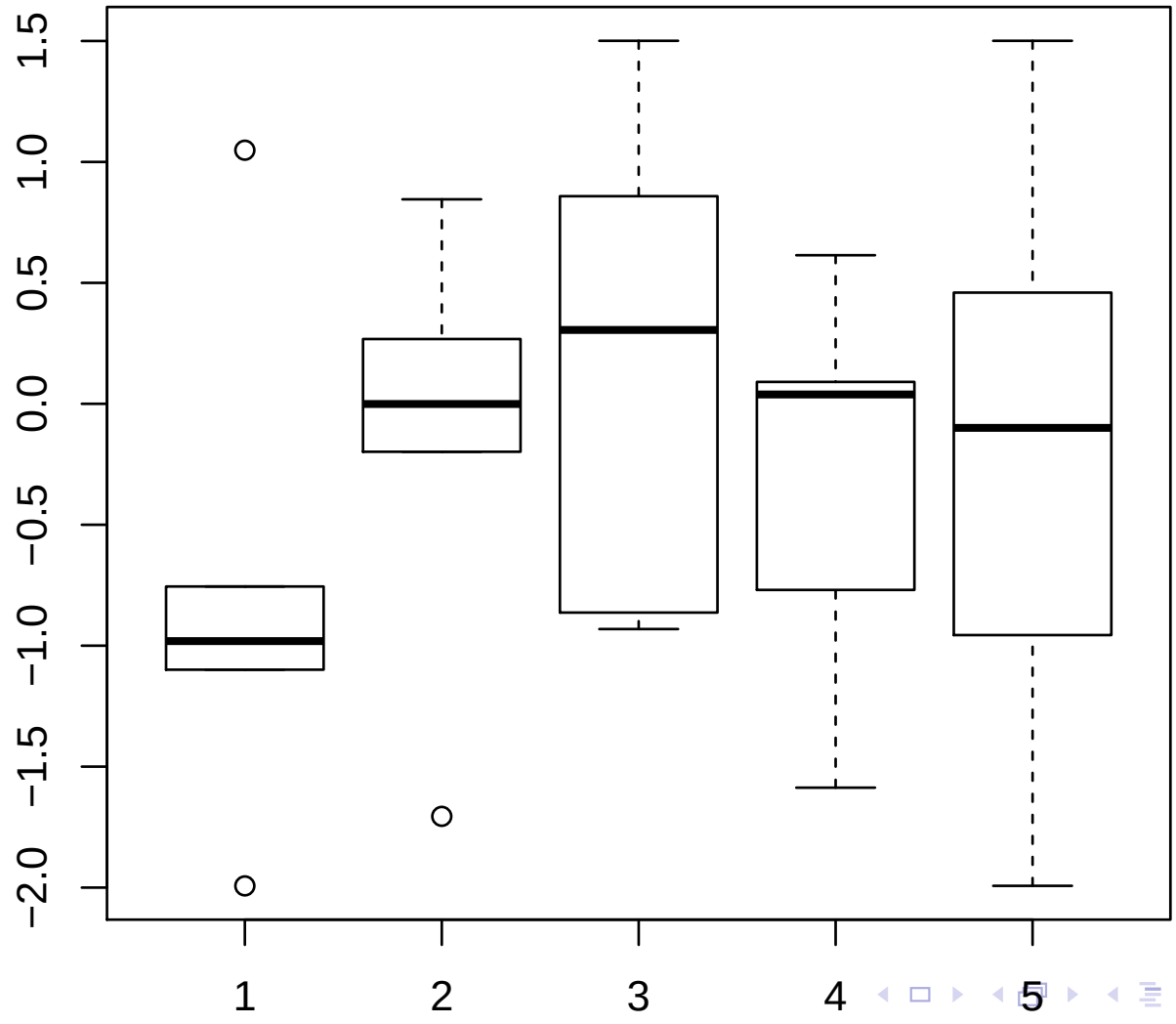
1a amostra:	-1.9928097624	-0.9810024738	1.0485672052	
	-1.0991516544	-0.7552698571		
2a amostra:	-1.7053600268	0.2678797571	0.8454606565	
	-0.0007533413	-0.1981656280		
3a amostra:	0.8583165119	-0.9311885079	-0.8630985081	
	0.3054040174	1.5008013298		
4a amostra:	-1.5869830215	0.0385825038	-0.7693842649	
	0.6142350393	0.0905950287		

As médias **amostrais** são

$$\hat{\mu_1} = -0.7559333 \;\; \hat{\mu_2} = -0.1581877 \;\; \hat{\mu_3} = 0.174047 \;\; \hat{\mu_4} = -0.3225909$$

Isso indica que as médias **populacionais** são iguais entre si? (Hipótese $\mu_1 = \mu_2 = \mu_3 = \mu_4$, contra a hipótese que pelo menos duas médias são diferentes.)

Comparação de médias de diversas populações usando a análise de variância (2/4).



Comparação de médias de diversas populações usando a análise de variância (3/4).

Soma dos quadrados das distâncias entre cada observação e **a média de sua amostra**:

$$SQtotal = 18.32611 = \sum_{j=1}^{4}\sum_{i=1}^{5}(x_{ji} - \hat{\mu}_j)^2$$

Soma dos quadrados das distâncias entre cada observação e a média COMUM ($\hat{\mu}$) definida naturalmente como

$$\hat{\mu} = \frac{\text{soma de todas as 20 observações}}{20}$$

é

$$SQentre = 16.08361 = \sum_{j=1}^{4}\sum_{i=1}^{5}(x_{ji} - \hat{\mu})^2$$

Comparação de médias de diversas populações usando a análise de variância (4/4).

A idéia de teste é assim: se H_0 for verdade, então $SQentre = 16.08361$ será significativamente menor que $SQtotal = 18.32611$.

A quantificação de "menor" faz-se seguindo a seguinte série de cálculos:

$$SQdentro = SQtotal - SQentre = 18.32611 - 16.08361 = 2.242508$$

$$Fobservado = \frac{SQentre/(4-1)}{SQdentro/(20-4)} = 0.7436169$$

A distribuição chamada "F" permite nos dizer se $Fobservado$ é de fato pequeno o suficiente para que podermos não rejeitar H_0 (não rejeitar que $\mu_1 = \mu_2 = \mu_3 = \mu_4$).