

GA-030 Estatística
Professor Marcio Borges



Laboratório
Nacional de
Computação
Científica

T-Student

Alunos:

Carolina Albuquerque Massena Ribeiro ¹

Eduardo Guerreiro Zilves ¹

Graziele Daiana Sena de Sousa ¹

¹ Programa de Pós-Graduação em Modelagem Computacional

Petrópolis, 01 de Dezembro de 2022



Histórico

- William Sealey Gosset (Student)
- Método científico na fabricação de cerveja.
- Teste t pequenas amostras.

Figura 1 - Retrato de William Sealey Gosset (1876-1938).



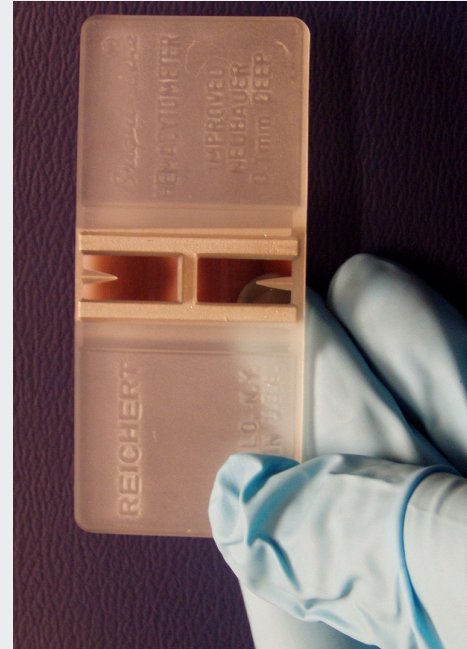
Fonte: Viali, 2016.



Histórico

- 1904 - Inferências de pequenas amostras de malte e lúpulo.
- A contagem das colônias de levedura - hemocitômetro.
- 1905 - Karl Pearson.

Figura 2 - Hemocitômetro



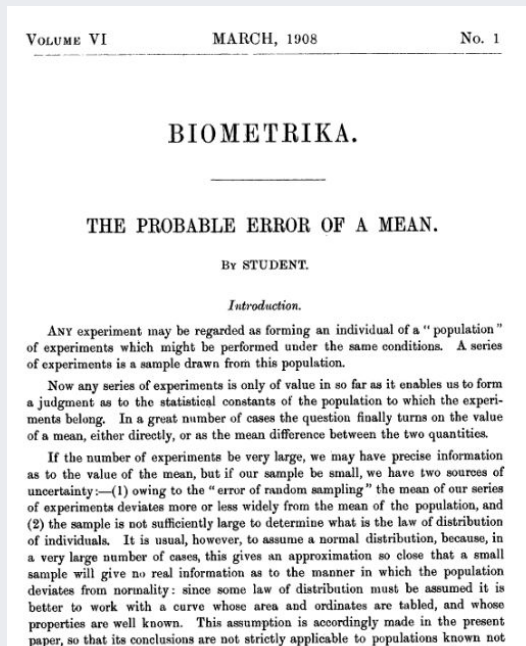
Fonte: wikipedia.org



Histórico

- 1908 - The Probable Error of a Mean (Sobre o Erro Provável de uma Média), na revista Biometrika.
 - S (desvio padrão amostral) era um estimador errático do desvio padrão σ para n pequeno.
- Publicação com pseudônimo Student.

Figura 3 - The probable error of a mean. "Biometrika (1908).



Fonte Viali, 2016.:



Histórico

- População unimodal e simétrica - (Tamanho Amostral)
- Se uma população X de média μ e variabilidade σ for padronizada por meio da transformação:

$$Z = (x - \mu)/\sigma$$

- Z é uma normal padrão.
- Média 0 e variância 1.
- Não altera a forma da variável X .



Histórico

- Comportamento de Z quando o valor de σ fosse desconhecido e estimado por meio de uma amostra.

Grandes Amostras



**Z continuava uma normal
padrão (0,1)**

Pequenas Amostras



?



Histórico

- Os resultados mantinham a simetria em torno de zero;
- A variabilidade dependia do tamanho da amostra utilizada;



Quanto menor a amostra



Maior variabilidade nos resultados

Tamanho da amostra



Teoria

- Modelagem matemática
- Exemplos
 - Hipóteses
 - Formulação



Quando utilizar o T-student

- Comparar grupos com parâmetros desconhecidos
- Experimentos com grupos de controle (grupo experimental e grupo de controle)



Utilizado na comparação

- Amostra e população
- Amostras pareadas (mesmos indivíduos)
- Amostras independentes
 - Populações diferentes ou
 - Tratamentos diferentes

1 Comparação entre população e amostra

- Teste-t para população: Comparar se a amostra pertence à população
- $H_0: \mu_1 = \mu$
- $H_1: \mu_1 \neq \mu$
- Hipóteses:
 - Variáveis X normais, e iid
 - Com $(n-1) S^2/\sigma^2 \sim \chi^2 (n-1)$

1 Comparação entre população e amostra

- Calcular média \bar{X}
- Desvio amostral s_x
- Estatística $t = (\bar{X} - \mu) / (s / \sqrt{n})$
 - $n-1$ Graus de Liberdade
- Requer:
 - população normalmente distribuída

$$T = \frac{\bar{X} - \mu}{s / \sqrt{n}}$$

Exemplo

- Exemplo agronomico
- Adubação de nitrato
- A média amostral é 7,51 mg/L de nitrato
- A média populacional é conhecida como 8,00 mg/L
- Sabendo que o desvio padrão é 1,38 e foram coletadas 27 amostras
- "Qual a probabilidade de se obter uma amostra tão pequena com média = 7,51 mg/L a partir da análise das 27 amostras?"

1

Teste de hipótese:

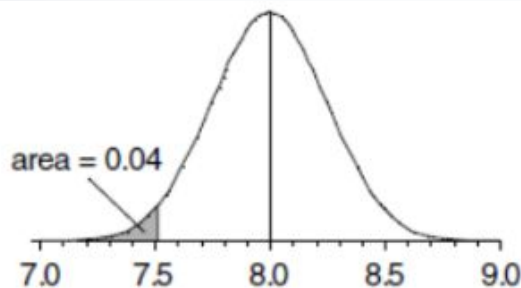
- H_0 : As médias são iguais
- H_1 : As médias são diferentes

$$T = \frac{\bar{X} - \mu}{s / \sqrt{n}}$$

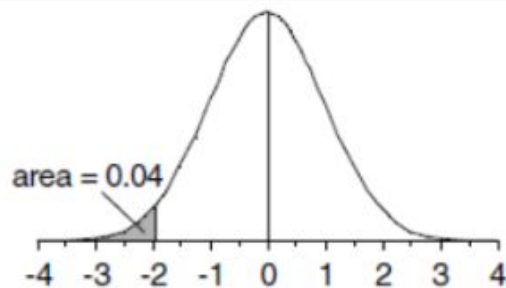
$$t = \frac{7,51 - 8}{1,38 / \sqrt{27}}$$

$$t = \frac{-0,49}{1,38 / 5,19}$$

$$t = \frac{-0,49}{0,2658} = -1,842$$



a) Referência de distribuição de \bar{X}
 $P(\bar{X} \leq 7,51) = 0,04$



b) Referência de distribuição T
 $P(t \leq 1,853) = 0,04$

n	$\alpha = 0.1$	0.05	0.025	0.01	0.005
2	1.886	2.920	4.303	6.965	9.925
4	1.533	2.132	2.776	3.747	4.604
6	1.440	1.943	2.447	3.143	3.707
10	1.372	1.812	2.228	2.764	3.169
20	1.325	1.725	2.086	2.528	2.845
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
40	1.303	1.684	2.021	2.423	2.704
∞	1.282	1.645	1.960	2.326	2.576

Grau de liberdade $v = n - 1$

Como são 27 amostras, temos:
 $v = 27 - 1 = 26$ grau de liberdade

$$t = -1,842$$

Analise do exemplo

- Tenho evidência estatística para rejeitar H_0 .
- Se este resultado é altamente improvável, pode ser que a amostra não representa a população.
- Devido ao Acaso



Nitrato

2 Comparação entre duas amostras pareadas

- Teste-t pareado: Amostras relacionadas:
 - Cada ítem tem dois testes
 - Antes e depois, ou tratamentos no mesmo paciente
- $H_0: \mu_1 - \mu_2 = 0$
- $H_1: \mu_1 - \mu_2 \neq 0$
- Hipóteses:
 - Variáveis X e Y de origem normais, e iid
 - Variáveis devem ser relacionadas entre os dois grupos

2 Comparação entre duas amostras pareadas

- Variáveis aleatórias: X (antes) e Y (depois)
- H0: Diferença $D = Y - X = 0$
 - $d_i = y_i - x_i$
 - Obter média \bar{d}
 - Obter desvio das diferenças s_d e calcular $SE(d) = s_d / \sqrt{n}$
 - Estatística $T = \bar{d} / SE(d)$
 - Graus de liberdade $n-1$
 - Comparar distribuição $t(n-1)$ e achar p-valor
- Requer: diferenças normalmente distribuídas

$$d_i = y_i - x_i$$

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

Exemplo

Exemplo ilustrativo: notas P1 e P2
(Com 20 notas) Média:

- Mean(d) = 2.05
- $s_d = 2.837$
- $SE(d) = s_d / \sqrt{20} = 0.634$
- $t = 2.05 / 0.634 = 3.231$
- 19 Graus de Liberdade
 - $p = 0.004$

	P1	P2	Diferença
1	18	22	+4
2	21	25	+4
3	16	17	+1
4	22	24	+2
5	19	16	-3
6	24	29	+5
7	17	20	+3
8	21	23	+2
9	23	19	-4
10	18	20	+2
...

	P1	P2	Diferença
11	14	15	+1
12	16	15	-1
13	16	18	+2
14	19	26	+7
15	18	18	0
16	20	24	+4
17	12	18	+6
18	22	25	+3
19	15	19	+4
20	17	16	-1
...

2 Comparação entre duas amostras pareadas

- Pros:
 - Amostra pequena
 - Mesma amostra, mesmas capacidades
- Contras:
 - Requer paridade de indivíduos
 - Ordem pode afetar o teste

3 Comparação entre duas amostras não-pareadas

- Teste-t não pareado: duas amostras independentes
 - Comparar média entre dois grupos independentes
- $H_0: \mu_1 = \mu_2$
- $H_1: \mu_1 \neq \mu_2$
- Amostras devem ter mesma variância (realizar teste de variância)
- Hipóteses:
 - Variáveis X normais, iid
 - Variância entre os dados deve ser mesma entre grupos, mesmo desconhecida
 - Variáveis independentes de dois grupos

3 Comparação entre duas amostras não-pareadas

- Amostras 1 e 2
 - Calcular diferença $\bar{x}_1 - \bar{x}_2$
 - Calcular s_p
 - Calcular $SE(\bar{x}_1 - \bar{x}_2)$
 - Estatística $T = (\bar{x}_1 - \bar{x}_2) / SE(\bar{x}_1 - \bar{x}_2)$
 - $n_1 + n_2 - 2$ Graus de Liberdade
 - Comparar com $t_{(n_1+n_2-2)}$ para achar p-valor
- Requer:
 - Amostras normalmente distribuídas
 - Variâncias aproximadamente iguais
 - Deve-se realizar teste de variância antes!

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

$$SE(\bar{x}_1 - \bar{x}_2) = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Exemplo

- Calorias em marcas de presunto defumado ou não-defumado

- $\bar{x}_1 - \bar{x}_2 = 34.38$

- $s_p = 23.98$

- $SE = 7.91$

- $t = 4.346$

- $p < 0.001$

	n	\bar{X}	s
Defumado	20	156.85	22.64
Não defumado	17	122.47	25.48

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

3 Comparação entre duas amostras não-pareadas

- Teste T de Welch
- Se variâncias diferentes, há aproximação:
 - $SE(\bar{x}_1 - \bar{x}_2) = \sqrt{(s_1^2/n_1 + s_2^2/n_2)}$
- Se n_1 e n_2 grandes: estatística $\sim N(0,1)$
- Caso contrário $\sim t_{n'}$ com n' arredondado

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

$$n' = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{(n_1 - 1)} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{(n_2 - 1)}}$$

3 Comparação entre duas amostras não-pareadas

- Pros:
 - Não requer paridade de indivíduos
 - Minimiza efeitos de ordem de testes
- Contras:
 - Requer amostra de tamanho maior
 - Amostras podem ter capacidades diferentes
 - Requer análise de variância

Outros Exemplos


- Pareado:
 - Efeito de medicamentos em mesmo grupo de pessoas
 - Cursos de mesma matéria para grupo de alunos
 - Resultados de exames acadêmicos em diferentes provas para mesmos alunos (P1, P2...)
- Não pareado:
 - Efeito de medicamentos com grupo experimental e grupo de controle
 - Medir nível de glicose para homens e para mulheres
 - Comparar tempo de transporte por rotas diferentes com mesmo destino

Aplicações e exemplos

- Exemplos computacionais
 - Descrição
 - Apresentação dos dados
 - Construção dos testes
 - Análise de variância
 - Hipótese nula e alternativa
- Inferências sobre resultados
- Apresentação do código

Exemplo

- Exemplo agronomico
- Peso de frutos de maracuja
- Azul, amarelo, verde e roxo são oriundos de melhoramento genético
- Vermelho é o controle, cultivar mãe das melhoradas



	Azul	Amarelo	Verde	Roxo	Vermelho
0	222	88	146	294	28
1	235	165	189	303	101
2	188	131	87	207	87
3	239	88	200	282	101
4	165	159	245	199	121
5	200	184	194	127	151
6	138	170	196	121	131
7	215	264	91	150	91
8	135	182	295	273	81
9	122	263	286	203	128

Peso de cada um dos 10 frutos coletados



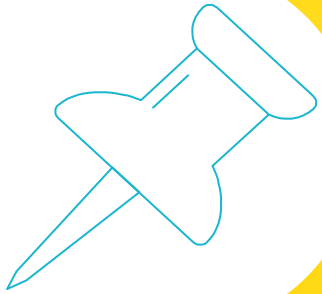
Primeiro faremos uma análise de variância (Anova)

Em seguida teste de hipótese

- H_0 : A média do controle e dos genótipos selecionados é igual
- H_1 : A média do controle e dos genótipos selecionados é diferente



Maracujá-doce



Trabalho Estatística - Colaboratory ([google.com](https://colab.google.com))

Analise do exemplo

- Tenho evidência estatística para rejeitar H_0 em todos os casos analisados

```
[ ] stats.ttest_1samp(a=df['Azul'], popmean=df['Vermelho'].mean())  
  
Ttest_1sampResult(statistic=6.10661563810113, pvalue=0.00017781197141369796)  
  
[ ] stats.ttest_1samp(a=df['Amarelo'], popmean=df['Vermelho'].mean())  
  
Ttest_1sampResult(statistic=3.523246969735404, pvalue=0.0064823031673492525)  
  
[ ] stats.ttest_1samp(a=df['Verde'], popmean=df['Vermelho'].mean())  
  
Ttest_1sampResult(statistic=4.034760979410994, pvalue=0.0029517586924289723)  
  
[ ] stats.ttest_1samp(a=df['Roxo'], popmean=df['Vermelho'].mean())  
  
Ttest_1sampResult(statistic=5.204893957997219, pvalue=0.0005603808784633505)  
  
[ ] stats.ttest_1samp(a=df['Vermelho'], popmean=df['Vermelho'].mean())  
  
Ttest_1sampResult(statistic=0.0, pvalue=1.0)
```



Maracujá-doce

Conclusão

- Teste-t usado para comparar dois grupos
 - Outros testes podem comparar mais de dois grupos
 - Toker, Duncan, Contraste
- Podemos comparar médias mesmo com variâncias desconhecidas
 - Podemos comparar amostras de tamanhos diferentes

Referencias

- Statistics: Paired t-tests. Rosie Shier, 2004. Mathematics Learning Support Centre
- Statistics: Unpaired t-tests. Rosie Shier, 2004. Mathematics Learning Support Centre
- Application of Student's t-test, Analysis of Variance, and Covariance. Mishra et al. 2019.
- Cerveja e Estatística: Vida e Obra de um Mestre Cervejeiro. Viali 2016 UFSM
- William Gosset - Biography. MacTutor, University of St. Andrews.
<https://mathshistory.st-andrews.ac.uk/Biographies/Gosset/>
- Viali, Lorí, and Márcia Elisa Berlikowsky. "Cerveja e estatística: vida e obra de um mestre cervejeiro." VIDYA 36.2 (2016): 507-522.



Obrigado pela atenção!!