

O *minimum minorum* sobre a comparação entre médias

Daniel Corrêa Lobato
< *daniel@lobato.org* >

26 de dezembro de 2008

Resumo

Em algumas situações nós desejamos verificar se dois experimentos produziram o mesmo resultado. Usar 10% ou 20% de algodão em um fibra muda a sua resistência? Variar a memória do sistema de 256MB para 512MB muda o tempo médio de carga de um programa? Aumentar o tempo médio de serviço de um centro de serviço em um modelo de fila de 5 para 12ut modifica o *throughput* do sistema? Na realidade, queremos verificar a igualdade das médias nos dois experimentos (resistência, tempo de resposta, *throughput*...). Este documento apresenta os passos gerais para verificar se duas médias são iguais em modelos homocedásticos ou não. São usados testes estatísticos *t* e *F*. Isso se aplica a várias áreas experimentais.

Sumário

1	Teste de hipóteses	1
2	Ambientação e discussões iniciais	2
3	Verificando homoscedasticidade	3
4	Igualdade de médias para modelos homogêneos	4
5	Igualdade de médias para modelos heterogêneos	6
6	Tudo junto, agora	8
7	E o Excel nisso tudo?	8
8	Conclusões	9

1 Teste de hipóteses

Uma hipótese estatística é uma formulação de parâmetros de uma distribuição. Há, pelo menos, duas hipóteses: (1) H_0 , chamada de hipótese nula, e; (2) H_1 , chamada de hipótese alternativa.

Utilizamos o teste de hipóteses para especificar o conjunto de valores e através desse conjunto rejeitamos ou não a hipótese nula. Tal conjunto de valores é chamado de região crítica o teste.

Quando usamos o teste podem ser gerados dois erros. Se H_0 é rejeitada quando não deveria ser (i.e., ela é verdadeira), temos o erro do tipo I. Se H_0 não é rejeitada quando deveria ser (i.e., ela é falsa), temos o erro do tipo II.

As probabilidades desses dois erros são dadas por:

$$\alpha \rightarrow P(\text{rejeitar } H_0 / H_0 \text{ é verdadeiro}) = \text{erro tipo I}$$

$$\beta \rightarrow P(\text{não rejeitar } H_0 / H_0 \text{ é falso}) = \text{erro tipo II}$$

O poder do teste é dado por $1 - \beta = \alpha = P(\text{rejeitar } H_0 / H_0 \text{ é verdadeiro})$. Poder pode ser encarado como um indicativo de decisão errada, ou seja, um poder de 0,05 indica que, se forem executados vários experimentos, apenas 5% comportar-se-ão de forma diferente de H_0 .

2 Ambientação e discussões iniciais

Vamos considerar os seguintes dados obtidos de dois experimentos diferentes para discutir a teoria:

Obs	1	2	3	4	5	6	7	8	9	10	11	12	13	14	n_i	\bar{X}_i	S_i	S_i^2
AM_1	3	5	7	6	5	8	4	2	1	3	8	9	13	2	14	5,4286	3,3216	11,0330
AM_2	6	3	8	1	8	3	6	9	2	14	7	8	-	-	12	6,2500	3,6213	13,1136

Esses dados podem ser representados como:

$$AM_1 \rightarrow X_{(1,1)}, X_{(1,2)}, \dots, X_{(1,n_1)}$$

$$AM_2 \rightarrow X_{(2,1)}, X_{(2,2)}, \dots, X_{(2,n_2)}$$

Podemos modelar esses dados como:

$$X_{(i,j)} = \mu_i + \varepsilon_{(i,j)}$$

$$\varepsilon_{(i,j)} \sim N(0; \sigma_i^2)$$

$$i = \{1, 2\}$$

$$j = \{1, 2, 3, \dots, n_i\}$$

É importante lembrar que ε indica o resíduo, ou seja, a diferença entre uma observação e a média amostral ($\varepsilon_{(1,j)} = X_{(1,j)} - \bar{X}_1$).

A tabela abaixo mostra os resíduos para cada amostra.

	1	2	3	4	5	6	7
AM_1	-2,4286	-0,4286	1,5714	0,5714	-0,4286	2,5714	-1,4286
AM_2	-0,2500	-3,2500	1,7500	-5,2500	1,7500	-3,2500	-0,2500
	8	9	10	11	12	13	14
AM_1	-3,4286	-4,4286	-2,4286	2,5714	3,5714	7,5714	-3,4286
AM_2	2,7500	-4,2500	7,7500	0,7500	1,7500	-	-

Toda a discussão aqui considera $\varepsilon_{(i,j)} \sim N(0; \sigma_i^2)$, ou seja, os resíduos têm distribuição normal. Para verificar a normalidade do resíduo pode ser aplicado o teste de Kolmogorov-Smirnov. Via de regra, para grandes amostras ($n > 30$), o resíduo é normal. O teste de Kolmogorov-Smirnov faz um teste de hipóteses, onde H_0 é que os dados (no nosso caso, os resíduos) seguem uma distribuição normal. A hipótese H_0 é rejeitada se o *p-valor* for menor que o nível α desejado¹. Como as nossas amostras são pequenas, deveríamos fazer o teste de Kolmogorov-Smirnov antes de continuar, mas vamos assumir que a hipótese de normalidade do resíduo não foi rejeitada e aplicar o que está sendo discutido aqui.

3 Verificando homoscedasticidade

Para que seja possível verificar se duas médias populacionais μ_1 e μ_2 são iguais, a primeira coisa é verificar se as variâncias populacionais σ_1^2 e σ_2^2 são iguais. Se elas forem iguais, iremos fazer um teste; se elas forem diferentes, é necessário outro teste. Na realidade, nós não temos as medidas populacionais, mas as medidas amostrais (\bar{X}_i e S_i^2).

$$F_0 = \frac{S_1^2}{S_2^2} \sim F_{(n_1-1, n_2-1)}$$

É importante lembrar que F_0 deve ser calculado com a maior variância sobre a menor. Na distribuição *Fischer* há dois pontos, F_1 e F_2 , tal que a área entre F_1 e F_2 é igual a $1 - \alpha$, onde α é o nível de significância desejado², e as áreas antes de F_1 e depois de F_2 valem, cada uma, $\frac{\alpha}{2}$. A figura 1 representa a distribuição F , que é mono-caudal.

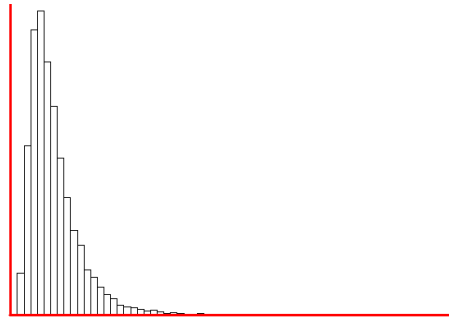


Figura 1: Aparência da distribuição F

Dessa forma, podemos dizer com poder $1 - \alpha$, que se $F_0 < F_1$ ou $F_0 > F_2$, rejeitamos a hipótese de igualdade de variâncias ($S_1^2 \neq S_2^2$). É claro que se $F_1 \leq F_0 \leq F_2$ então não é um absurdo assumir que $S_1^2 = S_2^2$

Aplicando o teste no conjunto acima vemos que:

$$\begin{aligned} S_1^2 &= 11,0330 \\ S_2^2 &= 13,1136 \\ F_0 &= \frac{S_2^2}{S_1^2} = \frac{13,1136}{11,0330} = 1,1886 \end{aligned}$$

¹Em uma leitura simplista, ele é o menor nível de significância α onde se rejeita a hipótese nula H_0

²Nos nosso caso, estamos interessados em coisas com 95% de confiança, ou poder de 0,05; logo, $\alpha = 0,05$

Precisamos identificar F_1 e F_2 . Para isso, encontramos em qualquer tabela estatística o valor de x para áreas acumuladas de 0,025 e 0,975 para 13 graus de liberdade do numerador e 11 graus de liberdade no denominador da distribuição *Fischer*.

$$F_1 = 0,2948$$

$$F_2 = 3,1975$$

Como $0,2948 \leq 1,1886 \leq 3,1975$, então não é um absurdo assumir que as variâncias são iguais.

4 Igualdade de médias para modelos homogêneos

Como as duas variâncias são iguais, podemos fazer um teste t . Para isso, precisamos identificar t_0 , tal que:

$$t_0 = \frac{\bar{X}_1 - \bar{X}_2}{Sp \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{(n_1+n_2-2)}$$

$$Sp = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}$$

Onde Sp é o desvio padrão agrupado. Fazemos um teste de hipóteses de igualdade contra diferença de médias.

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 - \mu_2 \neq 0$$

Se $|t_0| > t_{(\frac{\alpha}{2})}$, então rejeitamos a hipótese de igualdade de médias com poder $1 - \alpha$. Por outro lado, se $|t_0| \leq t_{(\frac{\alpha}{2})}$, não rejeitamos a hipótese de igualdade de médias. A figura 2 mostra a distribuição *t-student* que é bi-caudal.

Colocando números, vamos verificar se os dois conjuntos de dados apresentados no início do texto têm a mesma média. Já verificamos que não é um absurdo assumir que as variâncias são iguais, então é possível aplicar o teste t acima. Precisamos, apenas, calcular o Sp . Assim:

$$Sp = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}$$

$$Sp = \sqrt{\frac{(14 - 1)11,0330 + (12 - 1)13,1136}{14 + 12 - 2}}$$

$$Sp = \sqrt{\frac{143,4286 + 144,2500}{24}}$$

$$Sp = \sqrt{11,9866}$$

$$Sp = 3,4622$$

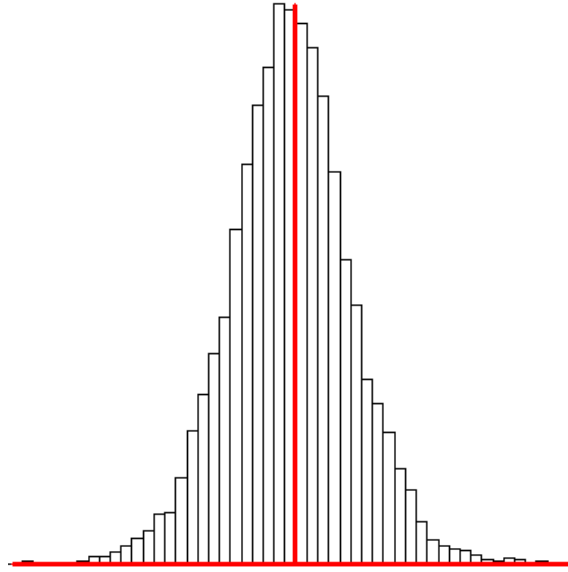


Figura 2: Aparência da distribuição *t-student*

Assim, o nosso t_0 fica:

$$\begin{aligned}
 t_0 &= \frac{\bar{X}_1 - \bar{X}_2}{Sp\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \\
 t_0 &= \frac{5,4286 - 6,2500}{3,4622 \times \sqrt{\frac{1}{14} + \frac{1}{12}}} \\
 t_0 &= \frac{-0,8214}{3,4622 \times \sqrt{0,1548}} \\
 t_0 &= \frac{-0,8214}{3,4622 \times 0,3934} \\
 t_0 &= \frac{-0,8214}{1,3620} \\
 t_0 &= -0,6031
 \end{aligned}$$

Tal como fizemos com o teste F , precisamos encontrar o $t_{\frac{\alpha}{2}}$. Assim, encontramos em qualquer tabela estatística o valor de x para áreas acumuladas de 0,025 e 0,975 de uma distribuição *t-student* com com $(14 + 12 - 2)$ graus de liberdade. Na realidade basta achar um deles: o outro é simétrico, pois a distribuição o é também. Isso nos fornece o valor $t_{\frac{\alpha}{2}} = 2,0639$.

Como $|t_0| = 0,6031$ e $t_{\frac{\alpha}{2}} = 2,0639$, temos que $|t_0| \leq t_{\frac{\alpha}{2}}$ e não rejeitamos a hipótese de igualdade das médias, ou seja, não é um absurdo afirmar que $5,4286 = 6,25$ com poder de 95%.

5 Igualdade de médias para modelos heterogêneos

Quando queremos comparar duas médias μ_1 e μ_2 e desconhecemos as variâncias σ_1^2 e σ_2^2 podemos utilizar Smith-Satterwaithe. Para isso, vamos efetuar, novamente, um teste de hipóteses, onde:

$$\begin{aligned}H_0 & : \mu_1 = \mu_2 \\H_1 & : \mu_1 \neq \mu_2 \\ \overline{X}_1 & \sim N\left(\mu_1; \frac{\sigma_1^2}{n_1}\right) \\ \overline{X}_2 & \sim N\left(\mu_2; \frac{\sigma_2^2}{n_2}\right)\end{aligned}$$

Então,

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

Assim,

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

Utilizando Smith-Satterwaithe temos que:

$$\begin{aligned}t_0 & = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim t_{(\gamma)} \\ \gamma & = \frac{\omega_1^2 + \omega_2^2}{\frac{\omega_1^2}{n_1+1} + \frac{\omega_2^2}{n_2+1}} \\ \omega_i & = \frac{S_i^2}{n_i}\end{aligned}$$

Como a distribuição *t-student* é bi-caudal e simétrica, basta calcular um dos pontos extremos (ou acumulada até $\frac{\alpha}{2}$ ou até $1 - \frac{\alpha}{2}$).

Rejeitamos H_0 quando $|t_0| > t_{(\gamma; 1 - \frac{\alpha}{2})}$ e não rejeitamos H_0 (quer dizer que não é um absurdo considerar as médias iguais) se $|t_0| \leq t_{(\gamma; 1 - \frac{\alpha}{2})}$.

Colocando esses dados no modelo temos que:

$$\begin{aligned}\overline{X}_1 & = 5,429 \quad (S_1^2 = 11,0330) \\ \overline{X}_2 & = 6,250 \quad (S_2^2 = 13,1136)\end{aligned}$$

Agora, queremos testar se as duas médias são iguais. Para isso, precisamos calcular cada um dos ω . Vamos lá:

$$\begin{aligned}\omega_1 &= \frac{S_1^2}{n_1} = \frac{11,0330}{14} = 0,788 \\ \omega_2 &= \frac{S_2^2}{n_2} = \frac{13,1136}{12} = 1,093\end{aligned}$$

Agora, calculamos o γ correspondente aos graus de liberdade da distribuição t :

$$\begin{aligned}\gamma &= \frac{\omega_1^2 + \omega_2^2}{\frac{\omega_1^2}{n_1+1} + \frac{\omega_2^2}{n_2+1}} \\ \gamma &= \frac{0,788^2 + 1,093^2}{\frac{0,788^2}{14+1} + \frac{1,093^2}{12+1}} \\ \gamma &= \frac{0,621 + 1,195}{\frac{0,621}{15} + \frac{1,195}{13}} \\ \gamma &= \frac{1,816}{0,0414 + 0,0919} \\ \gamma &= 13,623\end{aligned}$$

Agora, estimamos t_0 como sendo (sempre lembrando que estamos sob H_0 e, portanto, $\mu_1 = \mu_2$):

$$\begin{aligned}t_0 &= \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \\ t_0 &= \frac{(5,4286 - 6,2500) - (0)}{\sqrt{\frac{11,0330}{14} + \frac{13,1136}{12}}} \\ t_0 &= \frac{-0,8214}{\sqrt{0,7881 + 1,0928}} \\ t_0 &= \frac{-0,8214}{1,3714} \\ t_0 &= -0,5990\end{aligned}$$

Precisamos tirar alguma conclusão com base nesses valores... Precisamos verificar se a hipótese de igualdade das médias deve ou não ser rejeitada (se as médias são iguais ou não). Para isso, precisamos comprar os valor de t_0 com o valor de uma t -student em $1 - \frac{\alpha}{2}$ e γ graus de liberdade. Considerando que queremos 95% de confiança, $\alpha = 0,05$ e, por consequência, $\frac{\alpha}{2} = 0,025$.

Lembrando o critério do teste, rejeitamos H_0 quando $|t_0| > t_{(\gamma; \frac{\alpha}{2})}$ e não rejeitamos H_0 se $|t_0| \leq t_{(\gamma; \frac{\alpha}{2})}$.

Como $|t_0| = 0,5990$ e $t_{(13,623; 0,975)} = 2,1604$, temos que $0,5990 \leq 2,1604$ e, por consequência, não rejeitamos H_0 , ou seja, as médias não podem ser consideradas diferentes.

6 Tudo junto, agora

Conduzir um teste de igualdade de médias se resume aos seguintes passos mínimos:

1. Construir o modelo;
2. Verificar a normalidade do resíduo
3. Verificar a homoscedasticidade através do teste F
4. Fazer o teste t para igualdade de médias adequado ao modelo
 - Homogêneos: calcular o desvio padrão agrupado (Sp) e t_0 ; verificar a validade da hipótese nula de igualdade das médias comparando t_0 com $t_{\frac{\alpha}{2}}$
 - Heterogêneos: calcular γ e ω para, então, calcular t_0 e verificar ou refutar a hipótese nula

7 E o Excel nisso tudo?

Os estatísticos não gostam muito do Excel, pois há vários erros de arredondamento nos algoritmos. Para os nossos casos, é possível usá-la sem muita culpa.

Na versão em inglês³, você usará as seguintes funções

FINV (p ; n ; d) Calcula a inversa da distribuição F com probabilidade acumulada p para n graus de liberdade no numerador e d graus de liberdade no denominador. Perceba que é calculada a probabilidade **acumulada** e, por tanto, se você quiser saber o que sobra a partir de um ponto até o final do eixo x , você precisa fazer **FINV(1- p ; n ; d)**. Por exemplo, para calcular o ponto x tal que sobrem 2,5% no extremo direito da distribuição F com 13 e 11 graus de liberdade use **FINV(1-0,025;13;11)**.

TINV(p ; gl) Calcula a inversa da distribuição t para probabilidade p e gl graus de liberdade. Na hora de calcular $t_{\frac{\alpha}{2}}$, com $\alpha = 0,05$, e 24 graus de liberdade, você deve usar **TINV(0,05;24)**.

Essas funções podem ser usadas se você resolver fazer a análise passo-a-passo, mas os testes F e t já estão implementados no Excel no *add on*⁴ chamado *Analysis toolpak*. Quando você instala esse *add on*, aparece no menu *Tools* uma opção chamada *Data analysis*. Lá, você pode solicitar os cálculos para os testes F e t , fornecendo apenas o conjunto de dados. Os resultados são gerados na própria planilha, e você precisa, apenas, interpretá-los. Abaixo, os resultados da Excel para os testes F e t (homogêneo ou não) com o mesmo conjunto de dados que usamos aqui:

F-Test Two-Sample for Variances

Variable 1	Variable 2
------------	------------

³Por alguma razão que desconheço, os nomes das funções no Excel em português são diferentes. Se a sua for em português, procure a função correta com a ferramenta de ajuda *on line*.

⁴Em português, “suplemento”

Mean	6,2500	5,4286
Variance	13,1136	11,0330
Observations	12,0000	14,0000
df	11,0000	13,0000
F	1,1886	
P(F<=f) one-tail	0,3787	
F Critical one-tail	2,6346	

t-Test: Two-Sample Assuming Equal Variances

	Variable 1	Variable 2
Mean	5,42857	6,25000
Variance	11,03297	13,11364
Observations	14,00000	12,00000
Pooled Variance	11,98661	
Hypothesized Mean Difference	0,00000	
df	24,00000	
t Stat	-0,60310	
P(T<=t) one-tail	0,27605	
t Critical one-tail	1,71088	
P(T<=t) two-tail	0,55210	
t Critical two-tail	2,06390	

t-Test: Two-Sample Assuming Unequal Variances

	Variable 1	Variable 2
Mean	5,4286	6,2500
Variance	11,0330	13,1136
Observations	14,0000	12,0000
Hypothesized Mean Difference	0,0000	
df	23,0000	
t Stat	-0,5989	
P (T<=t) one-tail	0,2775	
t Critical one-tail	1,7139	
P(T<=t) two-tail	0,5551	
t Critical two-tail	2,0687	

8 Conclusões

Este documento serve, apenas, como um roteiro para o assunto. Não substitui, de forma alguma, um estudo apropriado da área.

Há uma infinidade de softwares estatísticos que podem ser usados para fazer essas contas de forma “automágica”, entre eles o MiniTAB⁵ (pago) e o R⁶ (GNU).

Se você resolver agir como o Knuth⁷, fazendo seu próprio software para resolver seus problemas, um livro que deve fazer parte da sua biblioteca é o *Numerical Recipes in C* de Press, Teukoslky, Vetterling e Flannery (Cambridge Press University). A segunda edição (a mais nova) é de 1992, mas continua atual como sempre. Esse livro pode ser encomendado na sua livraria preferida ou, se procurar bem, pode ser baixado de alguns sites de

⁵<http://www.minitab.com/>

⁶<http://cran.r-project.org/>

⁷Donald Knuth escreveu o pacote de software de composição T_EX e criou um pacote de fontes para resolver os problemas de composição do seu livro clássico *The Art of Computer Programming*.

universidades americanas. São mais de 1000 páginas de, literalmente, receitas numéricas nas mais variadas áreas, entre elas, estatística.

Atualizações no documento

Dezembro, 2008 Ajustes nos valores utilizados nos exemplos, e disponibilização pública do texto.

16 de junho 2004 Correção da fórmula de t_0 para modelo homogêneo. Ajuste dos valores e conclusões resultantes do erro na fórmula.

15 de junho de 2004 Acréscimo da seção sobre Excel.

10 de junho de 2004 Versão inicial desse documento.