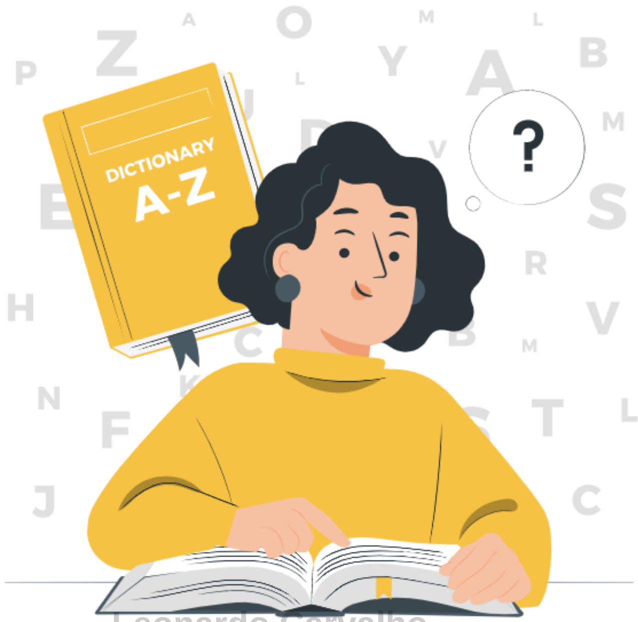
E.B.A - ESTATÍSTICA DO BÁSICO AO AVANÇADO

COM RENATA BIAGGI

TESTES DE HIPÓTESE E INTERVALO DE CONFIANÇA PARA VARIÂNCIA, ANOVA E DEEP DIVE EM CONCEITOS FUNDAMENTAIS

GLOSSÁRIO DE SÍMBOLOS



ITEM DE UMA SEQUÊNCIA

Se temos um conjunto de dados, cada item pode ser representado como sendo x1, x2, x3, etc. Por exemplo, temos o conjunto 10, 21, 32, 43, 58, então:

x1 = 10, x2 = 21, x3 = 32, x4= 43, x5= 58

SOMATÓRIO - ∑

Em matemática, somatório ou somatória é a adição de uma sequência de quaisquer tipos de números. O resultado é sua soma ou total.

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n$$

O índice i significa a partir de qual elemento devemos começar a soma. O símbolo n indica que, nesse caso, vamos somar a sequencia toda - que tem um total de n elementos.

Vamos a alguns exemplos. Vamos supor que queremos somar os números 1, 2 e 3, ou seja, temos um total de 3 elementos para somar:

$$\sum_{i=1}^{3} xi = x1 + x2 + x3 = 1 + 2 + 3 = 6$$

Muitas vezes vemos o somatório em expressões como:

$$\sum_{n=1}^{3} 2n - 1$$

Esse somatório indica que devemos somar a expressão 2*n - 1, substituindo o n por cada um dos valores da série. Logo, para a série 1, 2 e 3 temos

$$\sum_{n=1}^{3} 2n - 1$$

$$= \underbrace{[2(1) - 1]}_{n=1} + \underbrace{[2(2) - 1]}_{n=2} + \underbrace{[2(3) - 1]}_{n=3}$$

$$= 1 + 3 + 5$$

$$= 9$$

PRODUTÓRIO - Π

O produtório é a multiplicação de uma sequência de objetos matemáticos (números, funções, vetores, matrizes, etc.), chamados fatores, que tem como resultado o produto. É uma operação análoga ao somatório, embora seja menos utilizada quanto esse último. É representado pela letra grega pi maiúscula (Π).

$$\prod_{i=1}^{n} x_i = x_1 \times x_2 \times ... \times x_n$$

OUTROS SÍMBOLOS IMPORTANTES

Símbolo	**Parâmetro**
μ	Média populacional
σ	Desvio-padrão populacional
σ^2	Variância populacional
p	Proporção populacional
$\overline{x}$	Média amostral
s	Desvio-padrão amostral
s^2	Variância amostral
$\hat{p}$	Proporção amostral
α	Nível de significância

17. ANOVA - COMPARAÇÃO COM MAIS DE 2 AMOSTRAS

Neste capítulo, continuaremos com dados contínuos e comparando as médias dos grupos. Cobrimos os testes t em profundidade ao longo da primeira metade deste livro, então não vou revisitá-los aqui. No entanto, se você se lembrar, você pode usar testes t para comparar as médias de dois grupos no máximo. O que você faz se tiver três grupos ou mais? Use análise de variância (ANOVA)!

ANOVA COM UM FATOR

Análise de variância com um fator é uma técnica de teste de hipótese usada para comparar as médias de três ou mais populações. A análise de variância geralmente é abreviada como ANOVA. A ANOVA com um fator requer **um fator categórico** para a variável independente **e uma variável contínua para a variável dependente**. Os valores do fator categórico dividem os dados contínuos em grupos. O teste determina se as diferenças médias entre esses grupos são estatisticamente significativas. Por exemplo, se o tipo de fertilizante for sua variável categórica, você pode avaliar se as diferenças entre as médias de crescimento das plantas para pelo menos três fertilizantes são estatisticamente significativas. Para esse teste, usamos a **distribuição F.**

Tecnicamente, você pode usar ANOVA um fator para comparar apenas dois grupos. No entanto, se você tiver dois grupos, normalmente usará um teste t de duas amostras.

As hipóteses padrão para ANOVA 1 fator são as seguintes:

- H0: Todas as médias do grupo são iguais.
- Ha: Nem todas as médias do grupo são iguais.

Se o valor-p for menor que seu nível de significância (geralmente 0,05), rejeite a hipótese nula. Seus dados de amostra suportam a hipótese de que a média de pelo menos uma população é diferente das outras médias populacionais.

Suposições

Para resultados confiáveis de ANOVA unidirecional, seus dados devem atender às seguintes suposições:

- Amostras aleatórias
- Grupos independentes
- A variável dependente é contínua

A variável dependente é o resultado que você está medindo. O procedimento compara as médias do grupo desta variável. Por exemplo, o salário é uma variável contínua e você pode comparar os salários médios por grupos.

A variável independente é categórica

Os níveis da variável categórica definem os grupos que você está comparando. Por exemplo, o curso superior é uma variável categórica. As variáveis categóricas na ANOVA também são conhecidas como fatores.

Seus dados de amostra devem seguir uma distribuição normal ou cada grupo tem mais de 15 ou 20 observações

Os procedimentos de ANOVA pressupõem que seus dados seguem a distribuição normal. No entanto, como você viu para os testes t, você pode dispensar essa suposição se o tamanho da amostra for grande o suficiente (teorema do limite central).

Para ANOVA 1 fator, quando você tem de 2 a 9 grupos e cada grupo é menor que 15, seus dados podem ser distorcidos e os resultados do teste ainda serão confiáveis. Quando você tem 10-12 grupos, você deve ter pelo menos 20 por grupo para dispensar a suposição de normalidade.

Se seus dados não forem normais e seus tamanhos de amostra forem menores do que essas diretrizes, os resultados do teste podem não ser confiáveis.

Os grupos devem ter variações aproximadamente iguais ou usar a ANOVA de Welch

A forma padrão do teste F de ANOVA 1 fator assume que a variância dentro de cada uma das populações é igual. A diretriz padrão é que você pode assumir que as variâncias da população são iguais se nenhum grupo em sua amostra tiver o dobro da variância de outro grupo.

No entanto, se você não tiver certeza de que as variâncias são iguais, use a ANOVA de Welch, que não assume variâncias iguais. Veremos isso em uma outra sessão.

A estatística de teste para um teste ANOVA com um fator é a razão de duas variâncias: a variância entre amostras e a variância dentro das amostras.

$$\text{Estatística de teste} = \frac{\text{variância entre amostras}}{\text{variância dentro das amostras}}$$

A **variância entre amostras** mede as diferenças relacionadas ao tratamento dado a cada amostra. Essa variância, às vezes chamada de quadrado médio entre, é denotada por MSb, MQe ou SE2.

A **variância dentro das amostras** mede as diferenças relacionadas aos valores dentro da mesma amostra e é geralmente devido a erro amostral. Essa variância, às vezes chamada de quadrado médio dentro, é denotada por MSw, MQd, SD2 ou MSe.

As fórmulas necessárias para o cálculo da ANOVA são:

Variação	Soma dos quadrados	Graus de liberdade	Quadrados médios	F
Entre	SS_B	$\text{g.l.}_N = k - 1$	$MS_B = \dfrac{SS_B}{\text{g.l.}_N}$	$\dfrac{MS_B}{MS_W}$
Dentro	SS_W	$\text{g.l.}_D = N - k$	$MS_W = \dfrac{SS_W}{\text{g.l.}_D}$	

Da mesma forma, a notação SSW representa a soma dos quadrados dentro das amostras.

$$SS_W = (n_1 - 1)\, s_i^2 + (n_2 - 1)\, s_2^2 + \cdots + (n_k - 1)\, s_k^2$$
$$= \Sigma\, (n_i - 1)\, s_i^2$$

$$SS_B = n_1\,(\bar{x}_1 - \bar{\bar{x}})^2 + n_2\,(\bar{x}_2 - \bar{\bar{x}})^2 + \cdots + n_k\,(\bar{x}_k - \bar{\bar{x}})^2$$
$$= \Sigma n_i\,(\bar{x}_i - \bar{\bar{x}})^2$$

g.l.N = k − 1 (Graus de liberdade do numerador)

g.l.D. = N − k (Graus de liberdade do denominador)

Em que k é o número de amostras e N é a soma dos tamanhos das amostras

Por fim, calculamos F. Se F estiver dentro da zona de rejeição ou se p-valor menor que alpha, rejeitamos H0

Exemplo

Um pesquisador médico quer determinar se há diferença nas durações médias de tempo que três tipos de analgésicos levam para aliviar a dor de cabeça. Várias pessoas que sofrem com dores de cabeça são selecionadas aleatoriamente e tomam um dos três medicamentos. Cada pessoa registra o tempo (em minutos) que o medicamento levou para começar a fazer efeito. Os resultados estão na tabela abaixo. Para o nível de significância a = 0,01, você pode concluir que pelo menos um tempo médio é diferente dos demais? Suponha que cada população de tempos é normalmente distribuída e que as variâncias populacionais são iguais.

Medicamento 1	Medicamento 2	Medicamento 3
12	16	14
15	14	17
17	21	20
12	15	15
	19	
$n_1 = 4$	$n_2 = 5$	$n_3 = 4$
$\bar{x}_1 = \frac{56}{4} = 14$	$\bar{x}_2 = \frac{85}{5} = 17$	$\bar{x}_3 = \frac{66}{4} = 16{,}5$
$s_1^2 = 6$	$s_2^2 = 8{,}5$	$s_3^2 = 7$

Resposta:

As hipóteses nula e alternativa são:

H0: m1 =m2 =m3.

Ha: Pelo menos uma média é diferente das demais. (Afirmação.)

Como há k = 3 amostras, g.l.N = k − 1 = 3 – 1 = 2. A soma dos tamanhos das amostras é N = n1 + n2 + n3 = 4 + 5 + 4 = 13. Então, g.l.D = N − k = 13 – 3 = 10.

Agora veremos qual é o F para nosso teste (amostras)

$$\bar{\bar{x}} = \frac{\Sigma x}{N} = \frac{56 + 85 + 66}{13} \approx 15{,}92$$

$$MS_B = \frac{SS_B}{\text{g.l.}_\text{N}} = \frac{\Sigma n_i \left(\bar{x}_i - \bar{\bar{x}}\right)^2}{k - 1}$$

$$\approx \frac{(14 - 15{,}92)^2 + 5\ (17 - 15{,}92)^2 + 4\ (16{,}5 - 15{,}92)^2}{3 - 1}$$

$$= \frac{21{,}9232}{2} = 10{,}9616$$

$$MS_W = \frac{SS_W}{\text{g.l.}_\text{D}} = \frac{\Sigma\ (n_i - 1)s_i^2}{N - k}$$

$$= \frac{(4 - 1)(6) + (5 - 1)(8{,}5) + (4 - 1)(7)}{13 - 3}$$

$$= \frac{73}{10} = 7{,}3$$

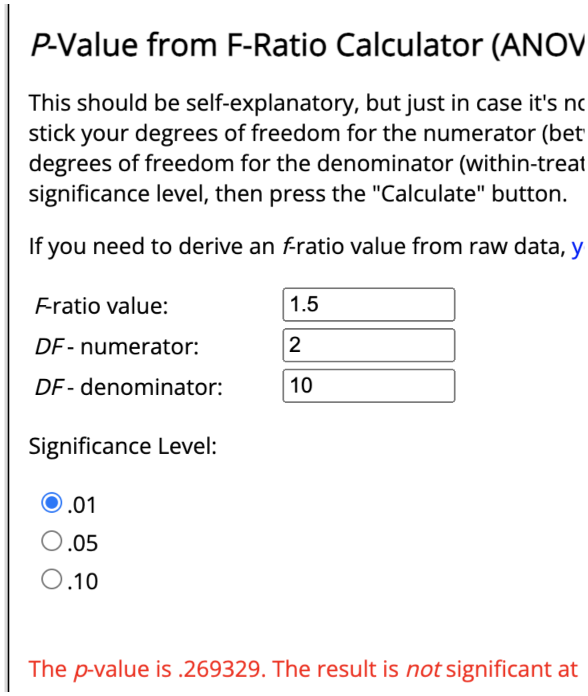
Usando $MS_B \approx 10{,}9616$ e $MS_W = 7{,}3$, a estatística de teste é:

$$F = \frac{MS_B}{MS_W} \approx \frac{10{,}9616}{7{,}3} \approx 1{,}50.$$

Usando nossa calculadora:

https://www.socscistatistics.com/pvalues/fdistribution.aspx

Temos:


P-Value from F-Ratio Calculator (ANOV

This should be self-explanatory, but just in case it's nc
stick your degrees of freedom for the numerator (bet
degrees of freedom for the denominator (within-treat
significance level, then press the "Calculate" button.

If you need to derive an *F*-ratio value from raw data, y

F-ratio value:	1.5
DF - numerator:	2
DF - denominator:	10

Significance Level:

- (●) .01
- ( ) .05
- ( ) .10

The *p*-value is .269329. The result is *not* significant at

Como p-valor é maior do que alpha (0.01) você não rejeita a hipótese nula. Portanto, dizemos que as médias populacionais são estatisticamente iguais.

POST HOC ANOVA

Os testes post hoc são parte integrante da ANOVA. Quando você usa ANOVA para testar a igualdade de pelo menos três médias de grupos, resultados estatisticamente significativos indicam que nem todas as médias de grupos são iguais. No entanto, os resultados da ANOVA não identificam quais diferenças particulares entre pares de médias são significativas. Use testes post hoc para explorar as diferenças entre as médias de vários grupos enquanto controla a taxa de erro do experimento.

Nesta seção, mostrarei o que são análises post hoc, os benefícios críticos que elas fornecem e ajudarei você a escolher a correta para seu estudo.

Começaremos com um exemplo de ANOVA 1 fator e depois o usaremos para ilustrar três testes post hoc.

Imagine que estamos testando quatro materiais que estamos considerando para fazer uma peça de produto. Queremos determinar se as diferenças médias entre as forças desses quatro materiais são estatisticamente significativas. Obtemos os seguintes resultados de ANOVA de uma via.

A	B	C	D
40	26.2	36	38
36.9	24.9	39.4	40.8
33.4	30.3	36.3	45.9
42.3	37.9	29.5	40.4
39.1	32.6	34.9	39.9
34.7	37.5	39.8	41.4

Performando a ANOVA com 1 fator, obtivemos p-valor = 0.004, ou seja, rejeitamos a hipótese nula que afirma que todas as médias são iguais. Porém, quais são os pares de amostras que têm diferenças na média?

Os testes post hoc realizam duas tarefas vitais.

i) Eles informam quais médias de grupo são significativamente diferentes de outras médias de grupo

ii) Eles também controlam a taxa de erro do experimento.

Qual é essa taxa de erro experimental? Para cada teste de hipótese que você executa, há uma taxa de erro do tipo I, que seu nível de significância (alfa) define. Em outras palavras, há uma chance de você rejeitar uma hipótese nula que é realmente verdadeira – um falso positivo. Quando você realiza apenas um teste, a taxa de erro do tipo I é igual ao seu nível de significância, que geralmente é de 5%. No entanto, à medida que você realiza mais e mais testes, sua chance de um falso positivo aumenta. Se você realizar testes suficientes, você aumenta muito sua chance de obter um falso positivo! A taxa de erro para uma família de testes é sempre maior do que para um teste individual.

No contexto ANOVA, você deseja comparar as médias do grupo. Quanto mais grupos você tiver, mais testes de comparação você precisará realizar. Para nosso exemplo de ANOVA com quatro grupos (A B C D), precisaremos fazer as seis comparações a seguir.

• A–B
• A–C
• A–D
• B–C
• B–D
• C–D

Nosso experimento inclui essa família de seis comparações. Infelizmente, a taxa de erro do experimento aumenta com base no número de grupos em seu experimento.

A tabela abaixo mostra como o aumento do número de grupos em seu estudo faz com que o número de comparações aumente, o que, por sua vez, aumenta a taxa de erro experimental. Abaixo temos um resumo de taxa de erro por quantidade de grupo que teríamos quando pensamos em 95% para cada teste:

All Pairwise Comparisons Alpha = 0.05		
Groups	**Comparisons**	**Experimentwise Error Rate**
2	1	0.05
3	3	0.142625
4	6	0.264908109
5	10	0.401263061
6	15	0.53670877
7	21	0.659438374
8	28	0.762173115
9	36	0.842220785
10	45	0.900559743
11	55	0.940461445
12	66	0.966134464
13	78	0.981700416
14	91	0.990606054
15	105	0.995418807

A tabela ilustra sucintamente o problema que os testes post hoc resolvem. Normalmente, ao realizar uma análise estatística, você espera uma taxa de falso positivo de 5%, ou qualquer valor que você definir para o nível de significância. Como mostra a tabela, quando você aumenta o número de grupos de 2 para 3, a taxa de erro quase triplica de 0,05 para 0,143. E, rapidamente piora a partir daí!

Essas taxas de erro são muito altas! Ao ver uma diferença significativa entre os grupos, você teria sérias dúvidas sobre se era um falso positivo em vez de uma diferença real.

Se você usar testes t de 2 amostras para comparar sistematicamente todas as médias dos grupos em seu estudo, encontrará esse problema. Você definiria o nível de significância para cada teste (por exemplo, 0,05) e, em seguida, o número de comparações determinará a taxa de erro do experimento, conforme mostrado na tabela.

Felizmente, os testes post hoc usam uma abordagem diferente. Para esses testes, você define a taxa de erro do experimento que deseja para todo o conjunto de comparações. Em seguida, o teste post hoc calcula o nível de significância para todas as comparações individuais que produzem a taxa de erro experimental especificada.

TESTE DE TUKEY

O teste post hoc que vamos usar é o **método de Tukey**. Há uma variedade de testes post hoc que você pode escolher, mas o método de Tukey é o mais comum para comparar todos os pares de grupos possíveis.

Para usar o teste de Tukey, precisamos:

- As amostras são extraídas independentemente umas das outras.
- As amostras são aleatórias
- Os dados em cada grupo são de uma população normalmente distribuída.
- As populações das quais os dados de cada grupo foram extraídos têm variâncias iguais.
- Os tamanhos de amostra de todos os grupos são iguais. (Se os grupos tiverem tamanhos de amostra diferentes, é realizado um teste de Tukey-Kramer).

O teste é dado por qs, também chamado de qscore:

$$q_s = \frac{Y_A - Y_B}{SE}$$

Em que Ya é a média do grupo A, Yb é a média do grupo B e SE é o erro padrão das duas médias. SE é dado por:

$$SE_{ANOVA} = \sqrt{\frac{MS_E}{n_j}}$$

MSe é a variância dentro das amostras e nj é a quantidade de amostras em cada conjunto.

Depois, podemos comparar o valor de qscore com o valor de qcrítico para entendermos se aceitamos ou não H0.

Q crítico depende da confiança que queremos (em geral, 95%), do g.l, que é dado por N-k, em que N é o total de amostras somadas em cada grupo e k é a quantidade de grupos que temos. Se qcrítico for **menor** que qscore, rejeitamos H0 e falamos que para aquele grupo as diferenças são estatísticas.

Vamos voltar ao nosso exemplo:

A	B	C	D
40	26.2	36	38
36.9	24.9	39.4	40.8
33.4	30.3	36.3	45.9
42.3	37.9	29.5	40.4
39.1	32.6	34.9	39.9
34.7	37.5	39.8	41.4

Vimos que os grupos tem médias diferentes, ou seja, nem todas são iguais. Agora precisamos entender qual delas não é igual. Vamos ao nosso passo a passo:

1. Calcule a média de cada grupo:

A	B	C	D
40	26.2	36	38
36.9	24.9	39.4	40.8
33.4	30.3	36.3	45.9
42.3	37.9	29.5	40.4
39.1	32.6	34.9	39.9
34.7	37.5	39.8	41.4
37.73	31.57	35.98	41.07

Em que a última linha é a média dos valores de cada grupo

2. Separe os pares

Pares
A vs B
A vs C
A vs D
B vs C
B vs D
C vs D

3. Obtenha os valores absolutos das diferenças das médias:

Pares	Abs(dif)
A vs B	6.17
A vs C	1.75
A vs D	3.33
B vs C	4.42
B vs D	9.50
C vs D	5.08

Por exemplo A vs B tem médias:

37.73	31.57

Dif = 37.73 - 31.57 = 6.17

Nesse caso a diferença é positiva, mas caso não seja, precisamos modificar o sinal para deixar todas as diferenças positivas.

4. Calcule o erro padrão: sqrt(mse/ni).

MSE = 15.60

Erro padrão = sqrt(15.60/6)

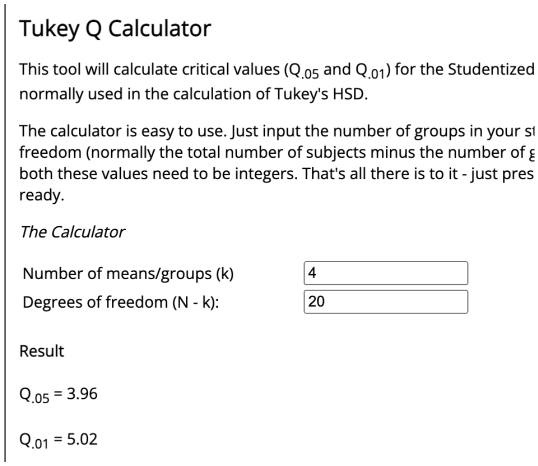
Pares	Abs(dif)	erro padrao
A vs B	6.17	1.612628117
A vs C	1.75	1.612628117
A vs D	3.33	1.612628117
B vs C	4.42	1.612628117
B vs D	9.50	1.612628117
C vs D	5.08	1.612628117

5. calculamos o qscore = abs(dif)/erro padrao

Pares	Abs(dif)	erro padrao	qscore
A vs B	6.17	1.612628117	3.82398558
A vs C	1.75	1.612628117	1.085185097
A vs D	3.33	1.612628117	2.067019233
B vs C	4.42	1.612628117	2.738800483
B vs D	9.50	1.612628117	5.891004813
C vs D	5.08	1.612628117	3.15220433

6. Compare com qcrítico, em que qcrítico é o q para 95% confiança, 4 grupos (k=4) e gl = 6*4 -3 (N-k) e pode ser obtido: https://www.socscistatistics.com/pvalues/qcalculator.aspx

Logo, qcrítico é 3.96:



Tukey Q Calculator

This tool will calculate critical values ($Q_{.05}$ and $Q_{.01}$) for the Studentized
normally used in the calculation of Tukey's HSD.

The calculator is easy to use. Just input the number of groups in your st
freedom (normally the total number of subjects minus the number of g
both these values need to be integers. That's all there is to it - just pres
ready.

The Calculator

Number of means/groups (k) 4
Degrees of freedom (N - k): 20

Result

$Q_{.05}$ = 3.96

$Q_{.01}$ = 5.02

7. qscore pode ser comparado ao qcrítico para acharmos o p-valor. Basta subtrair qcrítico e qscore

Pares	Abs(dif)	erro padrao	qscore	qcrítico	qcrítico-qscore
A vs B	6.17	1.612628117	3.82398558	3.96	0.1360144197
A vs C	1.75	1.612628117	1.085185097	3.96	2.874814903
A vs D	3.33	1.612628117	2.067019233	3.96	1.892980767
B vs C	4.42	1.612628117	2.738800483	3.96	1.221199517
B vs D	9.50	1.612628117	5.891004813	3.96	-1.931004813
C vs D	5.08	1.612628117	3.15220433	3.96	0.8077956703

8. Se qcrítico menor qscore: rejeita H0 e, portanto, há diferença estatística de médias

Pares	Abs(dif)	erro padrao	qscore	qcrítico	qcrítico-qscore
A vs B	6.17	1.612628117	3.82398558	3.96	0.1360144197
A vs C	1.75	1.612628117	1.085185097	3.96	2.874814903
A vs D	3.33	1.612628117	2.067019233	3.96	1.892980767
B vs C	4.42	1.612628117	2.738800483	3.96	1.221199517
B vs D	9.50	1.612628117	5.891004813	3.96	-1.931004813
C vs D	5.08	1.612628117	3.15220433	3.96	0.8077956703

Logo, o único grupo que tem médias estatisticamente diferentes é o B e o D, a 95% de confiança.

Outra forma de verificar isso seria através do **p-valor ajustado**. Caso você o tivesse, se p-valor ajustado do grupo for menor do que alpha, rejeitamos a hipótese nula. Ou seja, aquele grupo tem médias estatisticamente diferentes.

Para consultar como aplicar o teste Tukey-Kramer, acesse o site: https://www.automateexcel.com/stats/tukey-kramer-test/

Existem diversos testes post hoc.

Dunnet

Se o seu estudo tiver um grupo de controle e vários grupos de tratamento, talvez seja necessário comparar os grupos de tratamento apenas com o grupo de controle. Use o método de Dunnett quando o seguinte for verdadeiro:

• Antes do estudo, você sabe qual grupo (controle) deseja comparar com todos os outros grupos (tratamentos).

• Você não precisa comparar os grupos de tratamento entre si.

Leia mais sobre esse teste nos seguintes materiais:

- https://www.statology.org/dunnetts-test/

Hsu's MCB

Se o objetivo do seu estudo for identificar o melhor grupo, talvez não seja necessário comparar todos os grupos possíveis. As comparações múltiplas de Hsu com os melhores (MCB) identificam os grupos que são os melhores, insignificantemente diferentes dos melhores e significativamente diferentes dos melhores.

Use o MCB de Hsu quando você:

• Não sabe de antemão qual grupo você deseja comparar com todos os outros grupos.

• Não precisa comparar grupos que não são os melhores com outros grupos que não são os melhores.

• Pode definir “o melhor” como o grupo com a média mais alta ou com a média mais baixa.

O MCB de Hsu compara cada grupo ao grupo com a melhor média (mais alta ou mais baixa). Usando este procedimento, você pode acabar com vários grupos que não são significativamente diferentes do melhor grupo. Tenha em mente que o grupo que é realmente melhor em toda a população pode não ter a melhor média amostral devido ao erro de amostragem. Os grupos que não são significativamente diferentes do melhor grupo podem ser tão bons ou até melhores do que o grupo com a melhor média amostral.

Por exemplo, o grupo D é o melhor grupo geral porque tem a maior média (41,07). O procedimento compara D a todos os outros grupos. Nesse caso, os melhores grupos seriam D e outro grupo que supere muito o D estatisticamente falando (cuidado, pois não necessariamente é o que tem a maior média além de D devido aos desvios-padrões).

Leia mais sobre esse teste nos seguintes materiais:

- https://support.minitab.com/en-us/minitab/21/help-and-how-to/statistical-modeling/anova/supporting-topics/multiple-comparisons/what-is-hsu-s-mcb/
- https://www.real-statistics.com/one-way-analysis-of-variance-anova/unplanned-comparisons/hsus-mcb/

Games Howell

O teste post-hoc de Games-Howell é outra abordagem **não paramétrica** para comparar combinações de grupos ou tratamentos. Embora bastante semelhante ao teste de Tukey em sua formulação, o teste de Games-Howell **não assume variâncias iguais, tamanhos amostrais iguais e nem amostras normais.** O teste foi projetado com base na correção dos graus de liberdade de Welch:

$$df' = \frac{\left(\frac{s_i^2}{n_i} + \frac{s_j^2}{n_j}\right)^2}{\frac{\left(\frac{s_i^2}{n_i}\right)^2}{n_i - 1} + \frac{\left(\frac{s_j^2}{n_j}\right)^2}{n_j - 1}}$$

Leia mais sobre o teste nesse site: https://aaronschlegel.me/games-howell-post-hoc-multiple-comparisons-test-python.html#:~:text=The%20Games%2DHowell%20test%20is,variances%20or%20equal%20sample%20sizes.

Resumo de todos os testes para ANOVA 1 fator: https://www.ibm.com/docs/en/spss-statistics/saas?topic=anova-one-way-post-hoc-tests

INTRODUÇÃO ANOVA COM DOIS FATORES

Use ANOVA 2 fatores para avaliar as diferenças entre as médias dos grupos que são definidas por dois fatores categóricos. Como todos os testes de hipóteses, a ANOVA 2 fatores usa dados amostrais para inferir as propriedades de toda a população.

Para realizar essa análise, você precisará de duas variáveis categóricas, também chamadas de fatores. Esses fatores são suas variáveis independentes. Cada fator tem um número finito de valores possíveis, que são conhecidos como níveis. Por exemplo, gênero é um fator categórico que tem os dois níveis de masculino e feminino.

Você também precisa de uma variável de resultado contínua, que é a variável dependente. A ANOVA de 2 fatores determina se as diferenças médias entre esses grupos são estatisticamente significativas.

Por exemplo, avaliaremos se os dois fatores categóricos de gênero e curso superior correspondem a diferenças de renda, uma variável contínua.

A ANOVA 2 fatores começa a abordar aspectos dos modelos lineares de mínimos quadrados (que vocês vão ver em regressão). Por isso, não a abordaremos agora, porém voltaremos nesse assunto após a matéria de regressão.

18. PENSAMENTO CRÍTICO: REJEIÇÃO DA HIPÓTESE NULA

Deixar de rejeitar a hipótese nula quando é uma maneira de afirmar que os resultados de seu teste de hipótese não são estatisticamente significativos. O que isso significa exatamente? Há uma excelente razão para essa frase!

Embora a aceitação da hipótese nula pareça mais direta, ela não é estatisticamente precisa!

Antes de prosseguir, vamos recapitular algumas informações necessárias. Quando a hipótese nula tem a **igualdade** de médias ou proporções, temos:

• A hipótese nula afirma que não há efeito ou relação entre as variáveis.

• A hipótese alternativa afirma que o efeito ou relação existe.

Assumimos que a hipótese nula está correta até que tenhamos evidências suficientes para sugerir o contrário.

Depois de realizar um teste de hipótese, há apenas dois resultados possíveis.

• Quando seu valor-p é menor ou igual ao seu nível de significância, você rejeita a hipótese nula. Os dados favorecem a hipótese alternativa. Seus resultados são estatisticamente significativos.

• Quando seu valor-p é maior que seu nível de significância, você deixa de rejeitar a hipótese nula.

Para entender por que não aceitamos o nulo, considere que você não pode provar uma negativa. **A falta de evidência não é prova de que algo não existe. Você só não provou que existe.** Pode existir, mas seu estudo perdeu. Essa é uma diferença enorme, e é a razão para a redação complicada. Vejamos várias analogias.

Em um julgamento, começamos com a suposição de que o réu é inocente até que se prove o contrário. O promotor deve trabalhar duro para exceder um padrão probatório para obter um veredicto de culpado. Se o promotor não cumprir esse ônus, isso não prova que o réu é inocente. Em vez disso, não havia provas suficientes para concluir que ele é culpado.

Talvez o promotor tenha conduzido uma investigação de má qualidade e perdido as pistas? Ou, o réu cobriu com sucesso seus rastros? Consequentemente, o veredicto nesses casos é “inocente”. Esse julgamento não diz que o réu é inocente, apenas que não havia provas suficientes para afastar o júri da suposição padrão de inocência.

Quando você está realizando testes de hipóteses em estudos estatísticos, em muitos casos você pode desejar encontrar um efeito ou relação entre as variáveis. A posição padrão em um teste de hipótese é que a hipótese nula está correta. Como em um processo judicial, a evidência da amostra deve exceder o padrão probatório, que é o nível de significância, para concluir que existe um efeito.

O teste de hipótese avalia a evidência em sua amostra. Se o seu teste não detectar um efeito, isso não é prova de que ele não existe. Significa apenas que sua amostra continha uma quantidade insuficiente de evidências para concluir que ela existe. Como as espécies “extintas” ou o promotor que perdeu as pistas, o efeito pode existir na população geral, mas não em sua amostra específica. Consequentemente, os resultados do teste não rejeitam a hipótese nula, que é análoga a um veredicto de “inocente” em um julgamento. Simplesmente não havia evidência suficiente para mover o teste de hipótese da posição padrão de que o nulo é verdadeiro.

O ponto crítico dessas analogias é que a falta de evidências não prova que algo não existe – apenas que você não o encontrou em sua investigação específica. **Portanto, você nunca aceita a hipótese nula por completo.**

No caso de dizer que as amostras são estatisticamente iguais em H0 (não há efeito), aceitar a hipótese nula indicaria que você provou que um efeito não existe. Como você viu, não é bem assim. Você não pode provar um negativo. Em vez disso, a força de sua evidência fica aquém de ser capaz de rejeitar o nulo. Consequentemente, deixamos de rejeitá-lo.

Deixar de rejeitar o nulo indica que nossa amostra não forneceu evidências suficientes para concluir que o efeito existe. No entanto, ao mesmo tempo, **essa falta de evidência não prova que o efeito não existe.** Quais são as possíveis interpretações de não rejeitar a hipótese nula? Vamos trabalhar com eles.

Primeiro, é possível que o efeito de fato não exista na população, então seu teste de hipótese não o detectou na amostra. Faz sentido, certo? Embora essa seja uma possibilidade, não termina aí.

Outra possibilidade é que o efeito exista na população, mas o teste não o detectou por vários motivos. Esses motivos incluem o seguinte:

• O tamanho da amostra era muito pequeno para detectar o efeito.

• A variabilidade nos dados era muito alta. O efeito existe, mas o ruído em seus dados inundou o sinal (efeito).

• Por acaso, você coletou uma amostra aleatória não representativa da sua população. Ao lidar com amostras aleatórias, o acaso sempre desempenha um papel nos resultados. A sorte do sorteio pode ter feito com que sua amostra não refletisse um efeito que existe na população.

Observe como os estudos que coletam uma pequena quantidade de dados ou dados de baixa qualidade tendem a perder um efeito que existe? Esses estudos tiveram uma capacidade inadequada para detectar o efeito.

Certamente não queremos tomar resultados de estudos de baixa qualidade como prova de que algo não existe!

No entanto, não detectar um efeito (rejeitar H0) não significa necessariamente que um estudo seja de baixa qualidade. O acaso no processo de amostragem pode funcionar contra até mesmo os melhores projetos de pesquisa!

19. PENSAMENTO CRÍTICO: DEEP DIVE NO P-VALOR

Antes de avançarmos e falarmos sobre os outros testes de hipótese e intervalos de confiança para além da média, agora que vocês viram um teste de hipótese na prática precisamos levantar uma discussão.

Muitos dos testes que vocês farão na vida de vocês será comparando 2 ou mais amostras. Vimos que quando é esse o caso, nossa hipótese nula geralmente traz “média 1 = média 2”. Quando fazemos alguma melhoria e testamos essas duas amostras, frequentemente torcemos para que H0 seja rejeitado, uma vez que queremos de fato que essa melhoria seja impactante.

Por exemplo, suponhamos que temos um e-commerce e fizemos um layout diferente para o site na esperança da média de gastos por cliente aumentar, uma vez que supomos que ele vai gastar mais com esse novo site mais intuitivo e bonito. Separamos 2 amostras: uma amostra de pessoas que viram o site antigo e uma amostra de pessoas que viram o site novo. Considerando que essas amostras sejam de fato comparáveis e não temos vieses, medimos a média de gastos de ambas as amostras para entender o impacto do site novo na nossa métrica principal. Quando fazemos esse teste, rezamos com toda força para que as médias sejam diferentes, ou seja, que de fato esse novo layout impacte na média de compras. Portanto, torcemos para que H0 não seja rejeitado, ou seja, para p-valor alto.

MAS AFINAL, O QUE SIGNIFICA UM P-VALOR ALTO?

P-valores altos (aceitar que médias são iguais) não provam que não há efeito. P-valores altos indicam que sua evidência não é forte o suficiente para sugerir que existe um efeito na população. Um efeito pode existir, mas é possível que o tamanho do efeito seja muito pequeno, o tamanho da amostra seja muito pequeno ou haja muita variabilidade para o teste de hipótese detectá-lo.

Embora você possa não gostar de obter resultados que não sejam **estatisticamente significativos**, esses resultados podem impedir que você tire conclusões precipitadas e que tome decisões com base em ruído aleatório em seus dados! Valores altos de p ajudam a evitar erros dispendiosos. Afinal, se você basear suas decisões em erros aleatórios, você não obterá os benefícios esperados. Esta proteção aplica-se a estudos sobre métodos de ensino, eficácia de medicamentos, força do produto e assim por diante.

SIGNIFICÂNCIA PRÁTICA VERSUS ESTATÍSTICA

Anteriormente vimos como um efeito relativamente grande na sua amostra pode realmente ser um erro aleatório (lembrem-se dos boxplots com diversos efeitos mostrados na aula 3) . Vimos como os valores-p altos pode protegê-lo de tirar conclusões precipitadas com base no erro.

Agora imagine que você acabou de realizar um teste de hipótese e seus resultados são estatisticamente significativos. Oba! Esses resultados são importantes, certo? Não tão rápido. A significância estatística não significa necessariamente que os resultados são significativos no mundo real. Você pode ter resultados significativos para um efeito pequeno.

Vamos falar agora sobre as diferenças entre significância prática e significância estatística e como determinar se seus resultados são significativos no mundo real.

Significância estatística

O procedimento de teste de hipótese determina se os resultados da amostra que você obtém são prováveis se você assumir que a hipótese nula é correta para a população. Se os resultados forem improváveis (caem na zona de rejeição), você pode rejeitar a hipótese nula e concluir que existe um efeito estatisticamente significativo.

Consequentemente, pode parecer lógico que os p-valor e a significância estatística estejam relacionados à importância. No entanto, isso é falso porque as condições além de tamanhos de efeito grandes, podem produzir valores p minúsculos, o que nos levaria a rejeitar H0 e assumir que o efeito é estatisticamente significativo.

Isso acontece quando o **tamanho da amostra é muito grande** e/ou os dados têm **baixa variabilidade**. Vamos ver o porquê.

À medida que o tamanho da amostra aumenta, o teste de hipóteses ganha maior **poder estatístico** para detectar pequenos efeitos.

Com um tamanho de amostra grande o suficiente, o teste de hipótese pode detectar um efeito tão minúsculo que não tem sentido prático.

Quando seus dados de amostra têm baixa variabilidade, os testes de hipóteses podem produzir estimativas mais precisas do efeito da população. Essa precisão permite que o teste detecte pequenos efeitos.

A significância estatística indica apenas que você tem evidências suficientes para concluir que existe um efeito. É uma definição matemática que não sabe nada sobre a área temática e o que constitui um efeito importante.

Significância prática

Tamanho importa!

Enquanto a significância estatística se refere à existência de um efeito (rejeição de H0), a significância prática refere-se à sua magnitude. No entanto, nenhum teste estatístico pode dizer se o efeito é grande o suficiente para ser importante em sua área de estudo. Em vez disso, você precisa aplicar seu conhecimento da área de assunto e experiência para determinar se o efeito é grande o suficiente para ser significativo no mundo real.

Como você faz isso? Acho que é útil identificar os menores tamanho do efeito que ainda tem algum significado prático. Mais uma vez, este processo requer que você use seu conhecimento do assunto para fazer essa determinação. Se o tamanho do efeito do seu estudo for maior que este menor efeito significativo, seus resultados são praticamente significativos.

Por exemplo, suponha que você esteja avaliando um programa de treinamento comparando os resultados dos testes dos participantes do programa com aqueles que estudam por conta própria. Além disso, decidimos que a diferença entre esses dois grupos devem ter pelo menos cinco pontos para representar um tamanho de efeito praticamente significativo. Um efeito de 4 pontos ou menos é muito pequeno para ser relevante.

Após a realização do estudo, a análise encontra uma diferença entre os dois grupos. Os participantes do programa de estudos marcaram uma média de 3 pontos a mais em um teste de 100 pontos.

Enquanto estes resultados podem ser estatisticamente significativos (rodamos o teste de hipótese para comprovar), a diferença de 3 pontos é menor do que nosso limite de 5 pontos. Consequentemente, nosso estudo fornece evidências de que esse efeito existe, mas é muito pequeno para ser significativo na realidade mundial. O tempo e o dinheiro que os participantes gastam no treinamento programa não valem uma melhoria média de apenas 3 pontos.

Nem todas as diferenças estatisticamente significativas são interessantes!

Isso é bem direto ao ponto. Infelizmente, há uma pequena complicação. O tamanho do seu efeito é apenas uma estimativa porque vem de uma amostra. Graças ao erro de amostragem, há uma margem de erro em torno dele.

Precisamos de um método para determinar se o efeito estimado ainda é tem significância prática quando você considera essa margem de erro.

Um intervalo de confiança é um intervalo de valores que provavelmente contém o valor da população. Eu escrevi sobre intervalos de confiança anteriormente, então não vou me alongar aqui. A ideia central é que os intervalos de confiança incorporam a margem de erro criando um intervalo em torno do efeito estimado.

É provável que o valor da população caia dentro desse intervalo. Sua tarefa é determinar se todos, alguns ou nenhum desses intervalos representam efeitos com significância prática.

Vamos a um exemplo:

Suponha que conduzimos dois estudos sobre 2 programas de treinamento descritos acima. Ambos os estudos são estatisticamente significativos e produzem uma média de 9 pontos acima do que os estudantes tiravam antes. Esse aumento parece bom porque são maiores do que nosso menor tamanho de efeito significativo de 5.

No entanto, essas estimativas não incorporam a margem de erro. Os intervalos de confiança (ICs) para esse aumento de média em ambos os estudos abaixo fornecem essa informação crucial.

Método A: [3 15]

Método B: [7 11]

O IC do Método A se estende de valores que são muito pequenos para serem significativos (<5) para aqueles que são grandes o suficiente para serem significativos. Mesmo que o método seja estatisticamente significativo e o efeito estimado é 9, o IC cria dúvidas sobre se o efeito populacional real é grande suficiente para ser importante.

Por outro lado, o IC para o Método B contém apenas efeito significativo tamanhos. Podemos estar mais confiantes de que o tamanho do efeito da população é grande o suficiente para nos importarmos!

Intervalos de confiança são ótimos porque você pode usá-los para determinar a significância estatística e importância prática. Os intervalos de confiança focam no tamanho do efeito e da incerteza em torno da estimativa, em vez de apenas se o efeito existe.

CUIDADO PARA NÃO TOMAR CONCLUSÕES PRECIPITADAS

Precisamos lembrar agora dos nossos tipos de erros.

	Rejeita H0	Aceita H0
H0 é verdadeiro (real)	Erro tipo I: Falso positivo (FP)	**Acertamos \o/ Efeito não existe**
H0 é falso (real)	**Acertamos \o/ Efeito existe**	Erro tipo II: Falso negativo (FN)

Sempre que rejeitamos H0 (resultados estatisticamente significativos) corremos o risco de estar cometendo o erro do tipo I, ou seja, os falsos positivos

Neste contexto, um falso positivo ocorre quando você obtém um p-valor baixo e, sem saber, rejeita uma hipótese nula que é realmente verdadeira. Você conclui que existe um efeito (médias não são iguais) na população quando não existe.

Do ponto de vista científico, as altas taxas de falsos positivos são problemáticas por causa dos resultados enganosos. Do ponto de vista prático, se você estiver usando um teste de hipótese para melhorar um produto ou processo, você não obterá os benefícios esperados se os resultados do teste forem falsos positivos. Isso pode te custar muito dinheiro!

Essas dicas ajudarão você a desenvolver uma compreensão dos resultados do seu teste. Vou usar um estudo real de vacina contra a AIDS realizado na Tailândia para trabalhar com essas considerações. O estudo obteve um valor de p de 0,039, o que parece ótimo (mostra aqui que as médias de quem tomou e quem não tomou não são iguais). No entanto, depois de ler o que escrevemos, você pode pensar de forma diferente.

Dica 1: P-valores menores são interessantes

Os analistas geralmente veem os resultados estatísticos como significativos ou não. O foco está em saber se o p-valor é menor que o nível de significância porque os resultados estatisticamente significativos são altamente valorizados. Infelizmente, esse processo de decisão binária é uma simplificação excessiva porque nenhum nível de significância específico determina corretamente quais estudos têm efeitos populacionais reais 100% das vezes. Em vez disso, precisamos nos concentrar em entender a relação entre as taxas de falso-positivos e p-valores.

Existem vários estudos de simulação que mostram que taxas mais baixas de falsos positivos estão associadas com p-valores menores. Por exemplo, um valor de p próximo a 0,05 geralmente tem uma taxa de falso positivo de 25-50%. No entanto, um valor de p de 0,0027 geralmente tem um taxa de falso positivo de cerca de 4,5%. Essa taxa de falso positivo está próxima da taxa que é muitas vezes erroneamente atribuído a um valor p de 0,05.

P-valores mais baixos indicam evidências mais fortes contra a hipótese nula e uma menor probabilidade de um falso positivo.

É importante falar que não há relação diretamente calculável entre os p-valores e a taxa de falsos positivos. No entanto, a simulação estudos e abordagens Bayesiana podem produzir estimativas aproximadas de a taxa de falsos positivos.

Para ajudar a evitar resultados enganosos, você deve considerar o valor exato do p-valor. Usando a abordagem binária de uma determinação sim ou não de significância estatística é muito simplista.

O estudo da vacina contra a AIDS tem um valor de p de 0,039. Com base nas informações acima, devemos ser cautelosos com esse resultado.

Dica 2: A replicação é crucial

Na dica anterior, referi-me aos resultados de um único estudo. Realisticamente, você precisa replicar resultados estatisticamente significativos várias vezes antes de poder ter confiança nas conclusões.

No ambiente de alta pressão no meio corporativo para obter p-valores, um único p-valor é frequentemente considerado conclusivo. No entanto, Ronald Fisher, um grande matemático do século passado, desenvolveu p-valores com a noção de que eles são apenas uma parte do processo científico que inclui experimentação, análise e replicação.

“Um fato científico deve ser considerado como estabelecido experimentalmente somente se um experimento adequadamente projetado raramente deixa de fornecer esse nível de significância." –Ronald Fisher

O ideal é que casos façamos experimentação repetida com resultados consistentemente significativos para ter certeza de que a hipótese alternativa está correta.

Para o estudo da vacina contra a AIDS, o experimento tailandês é o primeiro estudo de vacinação contra a AIDS a produzir resultados estatisticamente significativos. Outros pesquisadores não conseguiram o replicarar, então precisamos ser cautelosos com resultados. Esta vacina não construiu um histórico de resultados significativos.

Dica 3: o tamanho do efeito é importante

A alta pressão para obter p-valores estatisticamente significativos desvia a atenção tanto do tamanho do efeito quanto da precisão da estimativa.

Você pode ter resultados de teste estatisticamente significativos mesmo quando os tamanhos dos efeitos são muito pequenos para serem significativos na prática. Além disso, um p-valor significativo não indica necessariamente que a análise possa estimar o tamanho do efeito com alta precisão.

Para dar mais ênfase ao tamanho e à precisão do efeito, use intervalos de confiança.

Considere se o tamanho do efeito é grande suficiente para ser importante na prática.

Infelizmente, o intervalo de confiança para a eficácia do estudo da vacina contra AIDS se estende de 1% a 52%. A vacina pode funcionar quase nenhuma vez até a metade das vezes. O intervalo de confiança revela que o tamanho do efeito estimado é pequeno e impreciso.

Dica 4: A plausibilidade da hipótese alternativa é importante

À medida que avaliamos os p-valores em testes de hipóteses, há uma tendência achar que p-valores semelhantes em todos os estudos dão suporte comparável para a hipótese alternativa. Por exemplo, um p-valor de 0,04 em um estudo parece fornecer a mesma evidência que um valor de p de 0,04 em outro estudo.

No entanto, estudos de simulação mostram que a plausibilidade da hipótese alternativa do estudo afeta consideravelmente o falso positivo

Por exemplo, com um p-valor de 0,05, uma hipótese alternativa altamente plausível está associada a uma taxa de falsos positivos de pelo menos 12%. Em comparação, uma alternativa implausível tem uma taxa de pelo menos 76%!

Se você está estudando uma hipótese alternativa improvável e você obtém um p-valor significativo, há uma probabilidade maior de que a hipótese alternativa não esteja correta.

Um p-valor significativo não nos absolve de usar nosso sentido ao interpretar os resultados. Se você ouvir falar de um estudo surpreendente que produz resultados maravilhosos, pode ser interessante esperar até que os outros estudos repliquem antes de confiar nele!

Nenhum estudo de outras vacinas contra a AIDS forneceu evidências suficientes para rejeitar a hipótese nula. Este padrão demonstra que é improvável que a hipótese alternativa esteja correta para o estudo tailandês. Nesse cenário, podemos esperar taxas de falso-positivos em torno de 75%!

Dica 5: Use sua experiência

Você deve aplicar seu conhecimento da área de assunto a todas as faces do teste de hipóteses para evitar resultados enganosos. Os pesquisadores e analistas devem usar sua expertise para avaliar a validade do desenho experimental, mecanismos propostos por trás do efeito, significado prático do efeito, a plausibilidade da hipótese alternativa e assim por diante.

Avaliando os resultados do teste de hipóteses para a vacina contra a AIDS

Vimos o seguinte:

• O valor-p de 0,039 não é uma evidência convincente por si só.

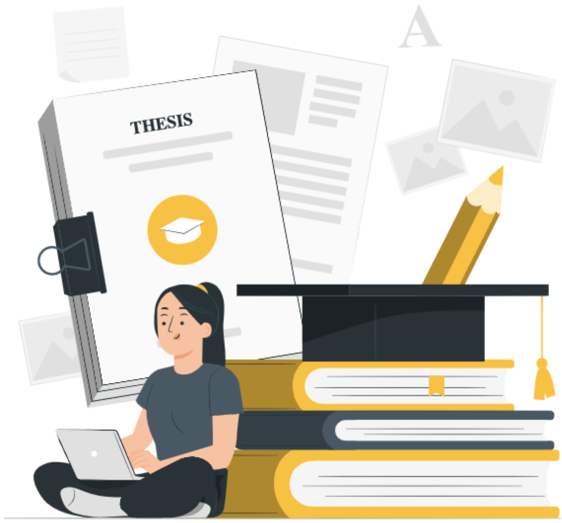
• A vacina não tem um histórico comprovado de resultados.

• O intervalo de confiança indica que o efeito estimado é pequeno e impreciso.

• Estudos de outras vacinas contra a Aids não tiveram resultados significativos, sugerindo que a hipótese alternativa na Tailândia é improvável.

Tomando todos esses pontos juntos, as considerações adicionais devem nos tornar cautelosos sobre resultados potencialmente enganosos. Em outras palavras, não devemos abrir uma garrafa de champanhe e começar a produzir a vacina em massa ainda. Precisamos esperar e ver se outros estudos replicaram esses resultados. Também precisamos ficar de olho no efeito tamanho em estudos futuros para determinar se a eficácia da vacina é significativa na prática.

REFERÊNCIAS BIBLIOGRÁFICAS



Bussan, W., Morettin, P. - Estatística Básica - Editora Saraiva - 2010 - 6th ed

Frost, J. - Hypothesis testing - An intuitive guide for making data driven decisions - Jim Frost - 2020 - 1st ed

Huyen, C. - Designing machine learning systems - Editora O'Reilly - 2022 - 1st ed

Knaflic, C.N - Storytelling com dados: Um guia sobre visualização - Editora Alta Books - 2019

Larson, R., Farber B. - Estatística Aplicada - Editora Pearson - 2016 - 6th ed

Pinheiro, J., Cunha, S., S. Santiago, Gomes, G. - Probabilidade e Estatística: Quantificando a incerteza - Elsevier Editora Ltda - 2012