Introdução à Probabilidade e à Estatística II

Inferência para duas Populações

Lígia Henriques-Rodrigues

MAE0229 – 1º semestre 2018

1. Introdução

Os dois capítulos anteriores apresentaram intervalos de confiança e testes de hipóteses para o parâmetro de uma única população (a média μ, a variância σ^2 ou uma proporção p).

Neste capítulo iremos comparar duas populações P_1 e P_2, baseados em dados fornecidos por amostras dessas populações, que suporemos seguirem distribuições normais.

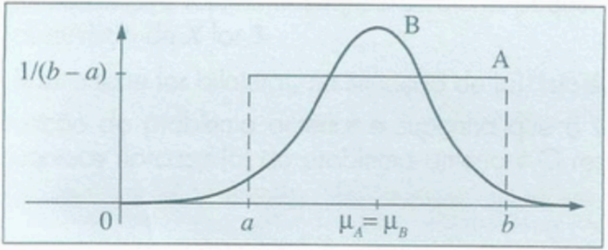
Objetivo: Responder a questões do tipo: O método A é melhor do que o B?

Em termos estatísticos, a questão anterior equivale a comparar dois conjuntos de informações, resultantes das medidas obtidas da aplicação dos dois métodos a dois conjuntos de objetos ou indivíduos.

Uma das dificuldades é a de caracterizar adequadamente a "igualdade" ou "equivalência" entre duas populações.

Exemplo: Suponha que estamos interessados em saber se os alunos de duas regiões, A e B, tiveram desempenhos iguais num mesmo teste nacional.

Figura 13.1 Distribuições das populações A e B, com $\mu_A = \mu_B = 4$, $\sigma_A = \sigma_B = 1{,}16$.



É necessário especificar a forma da distribuição. Especificada a forma, a igualdade dos parâmetros que identificam a curva implica a igualdade das duas populações.

Neste capítulo trataremos de várias situações:

- **Inferências para duas médias: amostras independentes**

  Aqui temos dados na forma de duas amostras, extraídas independentemente de cada população. É muito comum em experimentos do tipo "controle" *versus* "tratamento", nos quais o interesse principal é verificar o efeito desse último.

Exemplo 13.1

(a) Um curso de Estatística é ministrado pela televisão para um grupo de alunos e ao vivo para outro grupo. Queremos testar a hipótese de que o curso ao vivo é mais eficaz que o curso por meio da televisão.

(b) Queremos comparar o efeito de duas rações, A e B, sobre o crescimento de porcos. Dois grupos de porcos em crescimento foram alimentados com as duas rações e após cinco semanas verificam-se quais foram os ganhos de peso dos porcos dos dois grupos.

(c) 20 canteiros foram plantados com uma variedade de milho. Em dez deles um novo tipo de fertilizante é aplicado e nos outros um fertilizante padrão. Examinando-se as produções dos dois canteiros, queremos saber se há diferenças significativas entre as produções.

Modelo:

As v.a.'s $X_1, \ldots, X_n$ representam as respostas do grupo de controle e são v.a.'s independentes com distribuição P_1.

As v.a.'s $Y_1, \ldots, Y_m$ representam as respostas do grupo de tratamento e são v.a.'s independentes com distribuição P_2.

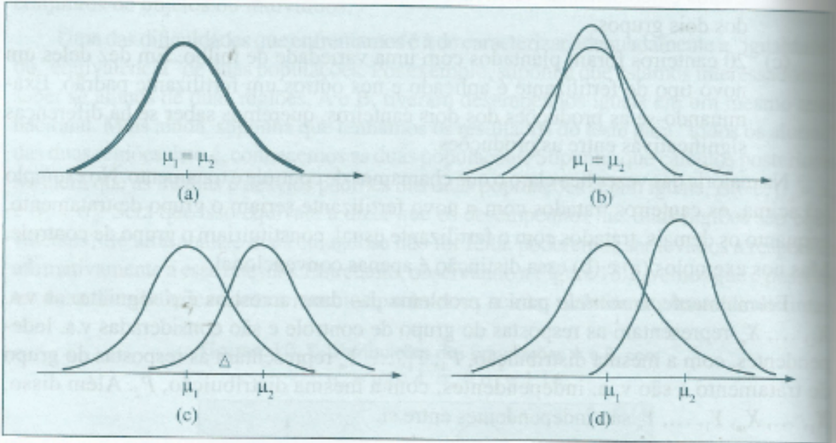
Além disso, $X_1, \ldots, X_n$ e $Y_1, \ldots, Y_m$ são independentes entre si.

A hipótese a ser testada é:

$$H_0 : P_1 = P_2.$$

Suponhamos que $P_1 \sim N(\mu_1, \sigma_1^2)$ e $P_2 \sim N(\mu_2, \sigma_2^2)$

Figura 13.2 (a) $\mu_1 = \mu_2, \sigma_1 = \sigma_2$ (b) $\mu_1 = \mu_2, \sigma_1 \neq \sigma_2$ (c) $\mu_1 \neq \mu_2, \sigma_1 = \sigma_2$ (d) $\mu_1 \neq \mu_2, \sigma_1 \neq \sigma_2$



A estratégia para comparar duas populações, por meio de seus parâmetros envolve suposições sobre a forma das distribuições, para depois testar médias e variâncias.

É comum estarmos interessados em testar apenas que P_1 e P_2 difiram em localização, isto é, a alternativa a H_0 é que P_1 esteja à direita de P_2, ou ao contrário, mas que ambas têm a mesma dispersão. Nesse caso, H_0 será equivalente a

$$H_0 : \mu_1 - \mu_2 = 0.$$

Outro caso interessante é aquele em que queremos testar se as duas médias são iguais, mas as variâncias são diferentes. Neste caso, seria necessário um teste preliminar de igualdade de variâncias.

As hipóteses apresentadas dizem-nos que não há efeito do tratamento. A alternativa usual para H_0 é que o efeito do tratamento aumenta as respostas, isto é, P_2 gera maiores valores do que P_1, com maior frequência. Mas pode ocorrer o contrário: diminuir as respostas.

- **Inferências para duas médias: amostras dependentes**

  Quando se comparam as médias de duas populações, pode ocorrer uma diferença significativa por causa de fatores externos não controlados. Um modo de contornar esse problema é coletar as observaçẽs em pares, de modo que os dois elementos de cada par sejam homogêneos em todos os sentidos, exceto no que diz respeito ao fator que queremos comparar.

- **Inferências para duas variâncias: amostras independentes**

  Como vimos, podemos testar se duas amostras independentes provêm de duas populações com variâncias iguais, mas desconhecidas. Esse teste, sob a suposição de normalidade das duas populações, usa uma estatística que tem uma distribuição especial, chamada F de Snedecor.

2. Comparação das variâncias de duas populações normais

Queremos testar

$$H_0 : \sigma_1^2 = \sigma_2^2 \iff \frac{\sigma_1^2}{\sigma_2^2} = 1 \quad \textit{versus} \quad H_1 : \sigma_1^2 \neq \sigma_2^2 \iff \frac{\sigma_1^2}{\sigma_2^2} \neq 1$$

O estimador de σ_1^2 é $\quad S_1^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2$

A estatística de teste para σ_1^2 é $\quad U = \frac{(n-1)S_1^2}{\sigma_1^2} \sim \chi^2(n-1)$

O estimador de σ_2^2 é $\quad S_2^2 = \frac{1}{m-1} \sum_{i=1}^{m} \left(Y_i - \overline{Y}\right)^2$

A estatística de teste para σ_2^2 é $\quad V = \frac{(m-1)S_2^2}{\sigma_2^2} \sim \chi^2(m-1)$

e portanto

$$W = \frac{S_1^2}{S_2^2} = \frac{\dfrac{\sigma_1^2 U}{n-1}}{\dfrac{\sigma_2^2 V}{m-1}} = \frac{\sigma_1^2}{\sigma_2^2} \frac{\dfrac{U}{n-1}}{\dfrac{V}{m-1}}, \text{ onde } U \sim \chi^2(n-1) \text{ e } V \sim \chi^2(m-1)$$

Sob a validade da hipótese nula:

$$\frac{S_1^2}{S_2^2} = \frac{\dfrac{U}{n-1}}{\dfrac{V}{m-1}} \sim F(n-1, m-1),$$

onde $F(n-1, m-1)$ representa a distribuição F-Snedecor com $n-1$ e $m-1$ graus de liberdades. Assim a estatística de teste é

$$W = \frac{S_1^2}{S_2^2} \underset{H_0}{\sim} F(n-1, m-1)$$

Dado α (nível de significância), a RC é uma reunião de caudas, em que cada cauda tem peso $\alpha/2$, sendo necessário determinar os quantis da distribuição F, f_1 e f_2 que satisfazem a condição:

$$P\{f_1 \leq W \leq f_2\} = 1 - \alpha, \quad P(W < f_1) = \alpha/2 = P(W > f_2).$$

Tendo em conta a tabela da distribuição F (que dá areas à direita), os quantis f_1 e f_2 são dados por

$$f_1 = f_{1-\frac{\alpha}{2}}^{n-1;m-1} \qquad f_2 = f_{\frac{\alpha}{2}}^{n-1;m-1},$$

tendo-se ainda a seguinte relação entre eles

$$f_{1-\frac{\alpha}{2}}^{n-1;m-1} = \frac{1}{f_{\frac{\alpha}{2}}^{m-1;n-1}}.$$

A RC é então,

$$RC_\alpha = (0, f_1) \cup (f_2, \infty)$$

$$P\{W \in RC_\alpha\} = P\{W < f_1 \text{ ou } W > f_2\} = \alpha.$$

Calculado o valor observado da estatística de teste, $w_0 = s_1^2/s_2^2$, a decisão a tomar é a de

- Não rejeitar H_0 se $f_1 \leq w_0 \leq f_2$
- Rejeitar H_0 se $w_0 < f_1$ ou $w_0 > f_2$

NOTA: Na prática escolhe-se s_1 e s_2 de modo a que $w_0 > 1$.

Se rejeitarmos a hipóteses H_0 , de igualdade das variâncias, é conveniente construir um intervalo de confiança para o quociente das duas variâncias. Assim, quando $\sigma_1^2 \neq \sigma_2^2$,

$$W = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{U/(n-1)}{V/(m-1)} \sim F(n-1, m-1),$$

e dado um coeficiente de confiança γ, pretendemos encontrar f_1 e f_2 tais que:

$$P(f_1 < F(n-1, m-1) < f_2) = \gamma \iff P\left(f_1 < \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} < f_2\right) = \gamma$$

Pelo que

$$P\left(f_1\frac{S_2^2}{S_1^2} < \frac{\sigma_2^2}{\sigma_1^2} < f_2\frac{S_2^2}{S_1^2}\right) = \gamma,$$

logo,

$$IC(\sigma_2^2/\sigma_1^2;\gamma) = \left(f_1\frac{s_2^2}{s_1^2}, f_2\frac{s_2^2}{s_1^2}\right)$$

Exemplo (pág. 373 – adaptado): Queremos verificar se duas máquinas produzem peças com a mesma homogeneidade quanto à variabilidade da resistência à tensão. Para isso, sorteamos duas amostras de seis peças de cada máquina, e obtivemos as seguintes resistências:

Máquina A:	145	127	136	142	141	137
Máquina B:	143	128	132	138	142	132

Tome uma decisão, considerando $\alpha = 10\%$.

Nota: $s_A^2 = 40$ e $s_B^2 = 36,97$.

3. Comparação das médias de duas populações normais: amostras independentes

Sejam $P_1 \sim N(\mu_1, \sigma_1^2)$ e $P_2 \sim N(\mu_2, \sigma_2^2)$.

A hipótese a testar é

$$H_0 : \mu_1 = \mu_2,$$

contra a alternativa (Fig. 13.2 (c))

$$H_1 : \mu_1 < \mu_2,$$

ou contra a alternativa

$$H_1 : \mu_1 > \mu_2,$$

ou se estivermos interessados apenas em verificar se existe diferença entre as médias das duas populações, não importando a direção,

$$H_1 : \mu_1 \neq \mu_2.$$

Os estimadores da média e da variância são:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad S_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2;$$

$$\overline{Y} = \frac{1}{m}\sum_{i=1}^{m} Y_i, \qquad S_2^2 = \frac{1}{m-1}\sum_{i=1}^{m}\left(Y_i - \overline{Y}\right)^2.$$

1o Caso: Variâncias conhecidas

Sob a validade de H_0, isto é, quando $\mu_1 = \mu_2$,

$$E(\overline{X} - \overline{Y}) = 0,$$

$$Var(\overline{X} - \overline{Y}) = Var(\overline{X}) + Var(\overline{Y}) = \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m},$$

e portanto,

$$Z = \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n + \sigma_2^2/m}} \sim N(0,1).$$

A estatística de teste, sob a validade de H_0 é então dada por

$$Z = \frac{\overline{X} - \overline{Y}}{\sqrt{\sigma_1^2/n + \sigma_2^2/m}} \underset{H_0}{\sim} N(0,1)$$

2o Caso: Variâncias desconhecidas, mas iguais

Se a hipótese de igualdade das variâncias não foi rejeitada então podemos supor que as variâncias populacionais são iguais, mas não conhecidas.

Como S_1^2 e S_2^2 são dois estimadores não enviesados de σ^2, podemos combiná-los para obter um estimador comum

$$S_p^2 = \frac{(n-1)S_1^2 + (m-1)S_2^2}{n+m-2} = \frac{\sum_{i=1}^n (X_i - \overline{X})^2 + \sum_{i=1}^n (Y_i - \overline{Y})^2}{n+m-2}$$

que também é um estimador não viesado de σ^2. Além disso, temos ainda que

$$\frac{(n+m-2)S_p^2}{\sigma^2} \sim \chi^2(n+m-2)$$

A estatística de teste é então

$$T = \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t(n+m-2)$$

Com base na estatística anterior, podemos construir um intervalo de confiança para $\mu_1 - \mu_2$:

$$P\left\{-t(\alpha/2; n+m-2) \leq \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n} + \frac{1}{m}}} \leq t(\alpha/2; n+m-2)\right\} = 1 - \alpha$$

$$P\left\{\overline{X} - \overline{Y} - t(\alpha/2; n+m-2)S_p\sqrt{\frac{1}{n} + \frac{1}{m}} \leq \mu_1 - \mu_2 \leq \overline{X} - \overline{Y} + t(\alpha/2; n+m-2)S_p\sqrt{\frac{1}{n} + \frac{1}{m}}\right\} = 1 - \alpha$$

$$IC(\mu_1 - \mu_2; 1 - \alpha) = \left\{(\overline{x} - \overline{y}) \pm t(\alpha/2; n+m-2)s_p\sqrt{\frac{1}{n} + \frac{1}{m}}\right\}$$

Como o intervalo de confiança corresponde à região de aceitação do teste bilateral, então:

Se $0 \in IC(\mu_1 - \mu_2; 1 - \alpha)$ então não rejeitamos hipótese $\mu_1 = \mu_2$.
Se $0 \notin IC(\mu_1 - \mu_2; 1 - \alpha)$ então aceitamos a hipótese $\mu_1 \neq \mu_2$.

Exemplo (pág. 376-377): Duas técnicas de vendas são aplicadas por dois grupos de vendedores: a técnica A, por 12 vendedores, e a técnica B, por 15 vendedores. Espera-se que a técnica B produza melhores resultados. No final de um mês, obtiveram-se os resultados da tabela 13.1.

Tabela 13.1 Dados para duas técnicas de vendas.

Dados	Vendas	
	Técnica A	Técnica B
Média	68	76
Variância	50	52
Vendedores	12	15

Teste, para o nível de significância de 5%, se há diferenças significativas entre as vendas resultantes das duas técnicas, supondo que as vendas sejam normalmente distribuídas e que $\sigma_A^2 = \sigma_B^2$.

3o Caso: Variâncias desconhecidas, mas diferentes

Quando a hipótese de igualdade das variâncias for rejeitada, devemos usar a estatística

$$T = \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n} + \frac{S_2^2}{m}}},$$

que sob a validade de $H_0 : \mu_1 = \mu_2$, segue distribuição t-Student com ν graus de liberdade,

$$\nu = \frac{(A+B)^2}{\frac{A^2}{n-1} + \frac{B^2}{m-1}}, \quad A = \frac{s_1^2}{n}, \quad B = \frac{s_2^2}{m},$$

i.e.,

$$T = \frac{(\overline{X} - \overline{Y})}{\sqrt{\frac{S_1^2}{n} + \frac{S_2^2}{m}}} \underset{H_0}{\sim} t(\nu).$$

Como o valor de ν é geralmente um numero fracionário, deve-se arredondar para o inteiro mais próximo para obter o número de graus de liberdade.

Exemplo (pág. 378): Queremos testar as resistências de dois tipos de vigas de aço, A e B. Tomando-se $n = 15$ vigas do tipo A e $m = 20$ vigas do tipo B, obtemos

$$\overline{x}_A = 70,5 \qquad s_A^2 = 81,6$$

$$\overline{x}_B = 84,3 \qquad s_B^2 = 210,8.$$

Admita que as variâncias populacionais são desconhecidas e diferentes. Teste, para o nível de significância de 5%, se há diferenças significativas entre as resistências dos dois tipos de vigas de aço, supondo que as resistências são normalmente distribuídas.

4. Comparação das médias de duas populações normais: amostras dependentes

Consideremos as amostras $X_1, \ldots, X_n$ e $Y_1, \ldots, Y_n$, de igual dimensão, em que as observações são pareadas,

$$(X_1, Y_1), \ldots, (X_n, Y_n).$$

Definindo a v.a. $D = X - Y$, teremos a amostra $D_1, \ldots, D_n$, resultante das diferenças entre os valores de cada par.

População Normal

Vamos então assumir que a v.a. D tem distribuição normal $N(\mu_D, \sigma_D^2)$. Mostra-se que

$$\overline{D} = \frac{1}{n}\sum_{i=1}^{n} D_i = \frac{1}{n}\sum_{i=1}^{n}(X_i - Y_i) = \overline{X} - \overline{Y} \sim N\left(\mu_D, \frac{\sigma_D^2}{n}\right).$$

Considerando o estimador não enviesado da variância σ_D^2

$$S_D^2 = \frac{1}{n-1}\sum_{i=1}^{n}(D_i - \overline{D})^2,$$

mostra-se que a estatística

$$T = \frac{\sqrt{n}(\overline{D} - \mu_D)}{S_D} \sim t(n-1).$$

Nota: $\mu_D = E(D) = E(X - Y) = E(X) - E(Y) = \mu_X - \mu_Y$.

Podemos também construir um intervalo de confiança para μ_D. Assim,

$$P\left\{-t(\alpha/2; n-1) \leq \frac{\sqrt{n}(\overline{D} - \mu_D)}{S_D} \leq t(\alpha/2; n-1)\right\} = 1 - \alpha$$

$$P\left\{\overline{D} - t(\alpha/2; n-1)S_D/\sqrt{n} \leq \mu_D \leq \overline{D} + t(\alpha/2; n-1)S_D/\sqrt{n}\right\} = 1 - \alpha$$

$$IC(\mu_X - \mu_Y = \mu_D; 1 - \alpha) = \left\{\overline{d} \pm t(\alpha/2; n-1)s_D/\sqrt{n}\right\}.$$

Como o intervalo de confiança corresponde à região de aceitação do teste bilateral, então:

Se $0 \in IC(\mu_D; 1-\alpha)$ então não rejeitamos hipótese $\mu_X = \mu_Y$.
Se $0 \notin IC(\mu_D; 1-\alpha)$ então aceitamos a hipótese $\mu_X \neq \mu_Y$.

Exemplo (pág. 379): Cinco operadores de certo tipo de máquinas são treinados em máquinas de duas marcas diferentes, A e B. Mediu-se o tempo que cada um deles gasta na realização de uma mesma tarefa, e os resultados estão na Tabela 13.8

Tabela 13.8 Tempos para realização de tarefa para cinco operadores.

Operador	Marca A	Marca B
1	80	75
2	72	70
3	65	60
4	78	72
5	85	78

Com o nível de significância de 10%, poderíamos afirmar que a tarefa realizada na máquina A demora mais do que a tarefa realizada na máquina B?

5. Comparação de proporções em duas populações

Sejam X_1 e X_2 duas variáveis aleatórias independentes com distribuição Binomial com parâmetros (n_1, p_1) e (n_2, p_2), respectivamente, em que $n_1, n_2 > 30$, com o mesmo atributo.

Sabemos que

$$\hat{p}_1 \approx N\left(p_1, \frac{p_1(1-p_1)}{n_1}\right) \qquad \hat{p}_2 \approx N\left(p_2, \frac{p_2(1-p_2)}{n_2}\right),$$

Pelo que

$$\hat{p}_1 - \hat{p}_2 \approx N\left(p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right).$$

Prova-se que substituindo p_1 e p_2 pelos seus estimadores, a v.a.

$$\frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\dfrac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \dfrac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \approx N(0,1)$$

Admitindo que queremos testar as hipóteses

$$H_0 : p_1 = p_2 \quad \text{contra} \quad H_1 : p_1 \neq p_2$$

e usando um estimador comum de $p_1 = p_2$

$$\hat{p}_c = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2},$$

obtemos, sob a validade de H_0 a seguinte estatística de teste

$$\frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p}_c(1-\hat{p}_c)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \underset{H_0}{\approx} N(0,1) \iff \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\hat{p}_c(1-\hat{p}_c)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \underset{H_0}{\approx} N(0,1)$$

Também podemos construir um intervalo de confiança (otimista) para $p_1 - p_2$

$$P\left(-z_{\alpha/2} < \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} < z_{\alpha/2}\right) = 1 - \alpha.$$

Logo, sendo $\hat{p}_1$ e $\hat{p}_2$ as estimativas de p_1 e p_2,

$$IC(p_1 - p_2; 1 - \alpha) = \left((\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\right).$$

Exemplo (pág. 395): Para lançamento de uma nova embalagem do sabonete X, a divisão de criação estuda duas propostas, A e B. Em cada um de dois supermercados similares, foram colocados sabonetes com cada tipo de embalagem, e a clientes selecionados aleatoriamente foi perguntado se tinham notado o sabonete e que descrevessem o tipo de embalagem. Abaixo estão os resultados:

Proposta	Notaram?		Total
	Sim	Não	
A	168	232	400
B	180	420	600
Total	348	652	1000

Queremos testar a hipótese de que os dois tipos de embalagens são igualmente atraentes.