

Grupo I

10 valores

1. Seja X_1, X_2, \dots, X_n uma amostra aleatória proveniente de uma população $X \sim \text{Bernoulli}(p)$.

(a) Deduza o estimador de máxima verosimilhança do parâmetro p .

(3.0)

$$\mathcal{L}(p|x_1, \dots, x_n) = \prod_{i=1}^n f_X(x_i) \quad (\text{porque } X_i \text{ são va iid})$$

$$= \prod_{i=1}^n p^{x_i} (1-p)^{1-x_i} = p^{\sum_{i=1}^n x_i} (1-p)^{n-\sum_{i=1}^n x_i}, \text{ com } 0 \leq p \leq 1.$$

Para $p \in (0, 1)$,

$$\log(\mathcal{L}(p|x_1, \dots, x_n)) = \sum_{i=1}^n x_i \log p + (n - \sum_{i=1}^n x_i) \log(1-p) \text{ (diferenciável em ordem a } p\text{)}.$$

O procedimento que se segue só deve ser aplicado se $\sum_{i=1}^n x_i \neq 0$ e $\sum_{i=1}^n x_i \neq n$, sendo o resultado obtido, $\hat{p} = \bar{x}$, também válido se $\sum_{i=1}^n x_i \in \{0, n\}$ e $p \in \{0, 1\}$.

Procura-se o valor de p , que se denomina por \hat{p} (a estimativa de máxima verosimilhança de p), que maximiza $\log(\mathcal{L}(p|x_1, \dots, x_n))$. Como

$$\frac{d \log(\mathcal{L}(p|x_1, \dots, x_n))}{dp} = 0 \Leftrightarrow \frac{\sum_{i=1}^n x_i}{p} - \frac{(n - \sum_{i=1}^n x_i)}{1-p} = 0 \Leftrightarrow$$

$$\Leftrightarrow (1-p) \sum_{i=1}^n x_i - p(n - \sum_{i=1}^n x_i) = 0 \Leftrightarrow \sum_{i=1}^n x_i - np = 0 \Leftrightarrow p = \frac{\sum_{i=1}^n x_i}{n} = \bar{x}, \text{ e}$$

$$\left. \frac{d^2 \log(\mathcal{L}(p|x_1, \dots, x_n))}{dp^2} \right|_{p=\bar{x}} = -\frac{\sum_{i=1}^n x_i}{\bar{x}^2} - \frac{(n - \sum_{i=1}^n x_i)}{(1-\bar{x})^2} < 0, \text{ uma vez que } 0 < \sum_{i=1}^n x_i < n, \text{ concluiu-se que } \hat{p} = \bar{x} \text{ é estimativa de máxima verosimilhança de } p \text{ e}$$

$\hat{p}_{MV} = \bar{X}$ é o estimador de máxima verosimilhança de p .

(b) Mostre que $T(X_1, X_2, \dots, X_n) = \sum_{i=1}^n a_i X_i$ é um estimador centrado do parâmetro p , onde a_1, a_2, \dots, a_n são constantes reais tais que $\sum_{i=1}^n a_i = 1$. Que implicação tem este resultado em termos do enviesamento da média da amostra aleatória como estimador do parâmetro p ?

$$E[T] = E\left[\sum_{i=1}^n a_i X_i\right] = \sum_{i=1}^n E[a_i X_i] = \sum_{i=1}^n a_i E[X_i] = \sum_{i=1}^n a_i p = p \sum_{i=1}^n a_i = p,$$

provando-se assim que T é um estimador centrado ou não enviesado de p .

Consequentemente, \bar{X} é um estimador não enviesado de p uma vez que se trata de um caso particular de T com $a_i = 1/n, i = 1, \dots, n$.

2. Para comparar duas técnicas de purificação de ouro foram recolhidas amostras, obtidas ao acaso e de forma independente, de ouro purificado por cada técnica, e analisado o respectivo teor em ouro (X e Y , em percentagem de pureza). A amostra referente à primeira técnica, de dimensão 20, conduziu a: $\sum_{i=1}^{20} x_i = 1880$ e $\sum_{i=1}^{20} x_i^2 = 178000$. Da amostra respeitante à segunda técnica, de dimensão 12, obteve-se: $\sum_{i=1}^{12} y_i = 1140$ e $\sum_{i=1}^{12} y_i^2 = 110000$.

Admitindo que, em ambos os casos, o teor do ouro após purificação tem distribuição normal:

- (a) Obtenha um intervalo de confiança a 99% para a variância do teor do ouro purificado usando a primeira técnica. (3.0)

Sejam $Q = \frac{19S^2}{\sigma^2} \sim \chi^2_{(19)}$, $a = F_{\chi^2_{(19)}}^{-1}(0.005) = 6.844$ e $b = F_{\chi^2_{(19)}}^{-1}(0.995) = 38.58$.

$P(a \leq Q \leq b) = 0.99 \Leftrightarrow P\left(\frac{19S^2}{b} \leq \sigma^2 \leq \frac{19S^2}{a}\right) = 0.99$. Em resumo, define-se

$$IAC_{0.99}(\sigma^2) = \left[\frac{19S^2}{38.58}, \frac{19S^2}{6.844} \right].$$

Para a amostra observada temos $\bar{x} = 1880/20 = 94$ e $19s^2 = \sum_{i=1}^{20} x_i^2 - 20\bar{x}^2 = 1280$ ($s^2 = 67.368$).

$$IC_{0.99}(\sigma^2) = [33.178, 187.025].$$

- (b) Admitindo que a variância do teor do ouro purificado é igual qualquer que seja a técnica usada, teste ao nível de significância de 1% a hipótese de não haver diferença entre as técnicas no que se refere ao teor médio do ouro após purificação. (3.0)

Hipóteses: $H_0 : \mu_X - \mu_Y = 0$ contra $H_1 : \mu_X - \mu_Y \neq 0$.

Dado que $\sigma_X^2 = \sigma_Y^2$, seja $T = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{19S_X^2 + 11S_Y^2}{30} \left(\frac{1}{20} + \frac{1}{12}\right)}} \sim t_{(30)}$.

Sob H_0 , obtemos a estatística do teste: $T_0 = \frac{(\bar{X} - \bar{Y})}{\sqrt{\frac{19S_X^2 + 11S_Y^2}{30} \left(\frac{1}{20} + \frac{1}{12}\right)}} \sim t_{(30)}$.

Para $\alpha = 0.01$ deve rejeitar-se H_0 se $|T_0| > F_{t_{(30)}}^{-1}(0.995) = 2.750$.

Para as amostras observadas temos $\bar{x} = 94$, $s_X^2 = 67.368$, $\bar{y} = 95$, $s_Y^2 = 154.545$ e

$$t_0 = \frac{-1}{3.639} = -0.275.$$

Como t_0 não pertence à região de rejeição então H_0 não é rejeitada para $\alpha = 0.01$.

Alternativa: valor- $p = 2F_{t_{(30)}}(-0.275) = 2 \times 0.393 = 0.786 > 0.01$, então H_0 não é rejeitada para $\alpha = 0.01$.

1. O número de tentativas até se conseguir o acesso a uma página muito popular do *YouTube* foi investigado. Em 100 acessos, escolhidos ao acaso, obtiveram-se os seguintes resultados:

Nº de tentativas	1	2	3	4 ou mais
Nº de acessos	40	30	20	10

- (a) Teste, ao nível de significância de 5%, a hipótese de o número de tentativas até se conseguir o acesso ter distribuição geométrica com valor esperado igual a 2. (3.5)

Seja X a va que representa o “número de tentativas até se conseguir acesso à página”.
Sabe-se que $X \sim Geo(p)$ com $E[X] = 1/p = 2$, logo $X \sim Geo(0.5)$.

Hipóteses: $H_0 : X \sim Geo(0.5)$ contra $H_1 : X \not\sim Geo(0.5)$.

Estatística de teste: $Q_0 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} \underset{H_0}{\sim} \chi_{(k-\beta-1)}^2$, onde β é o número de parâmetros a estimar.

Sejam $p_i^0 = P(X = i | H_0) = 0.5^i$, $i = 1, 2, 3$ e $p_4^0 = P(X \geq 4 | H_0) = 1 - \sum_{i=1}^3 p_i^0 = 0.125$.

	o_i	p_i^0	$E_i = np_i^0$
1	40	0.500	50.0
2	30	0.250	25.0
3	20	0.125	12.5
≥ 4	10	0.125	12.5
	$n = 100$		

Neste caso, não é necessário agrupar classes ($k = 4$) e não há qualquer parâmetro estimado ($\beta = 0$).

Para $\alpha = 0.05$ deve rejeitar-se H_0 se $Q_0 > F_{\chi_{(3)}^2}^{-1}(0.95) = 7.815$.

Como $q_0 = 8$ pertence à região de rejeição então deve rejeitar-se H_0 para $\alpha = 0.05$.

- (b) Calcule, justificando, o valor- p do teste anterior e decida com base no valor obtido. (1.0)

Uma vez que a região crítica é unilateral, rejeitando-se H_0 para valores elevados de Q_0 , tem-se

$$\text{valor-}p = P(Q_0 > 8 | H_0 \text{ verdadeira}) = 1 - F_{\chi_{(3)}^2}(8) = 0.046.$$

Conclui-se que para níveis de significância superiores a 0.046 rejeita-se H_0 , e não se rejeita caso contrário. Assim, aos níveis de significância de 0.05 e 0.10 há evidência para contestar que o número de tentativas até se conseguir acesso à página tem distribuição geométrica com valor esperado 2, o mesmo não acontecendo ao nível de significância de 0.01.

2. Para estudar o efeito do nível de álcool no sangue, x , sobre o tempo de reacção no controlo de uma máquina de corte industrial, Y , foram efectuadas observações em 7 indivíduos escolhidos ao acaso:

x_i : Nível de álcool no sangue (mg/dl)	10	20	30	40	50	60	70
y_i : Tempo de reacção (seg.)	40	50	50	65	70	70	90

$$\sum_{i=1}^7 x_i = 280, \quad \sum_{i=1}^7 y_i = 435, \quad \sum_{i=1}^7 x_i^2 = 14 \times 10^3, \quad \sum_{i=1}^7 y_i^2 = 28725 \quad \text{e} \quad \sum_{i=1}^7 x_i y_i = 19500.$$

Considerando o modelo de regressão linear simples de Y em x , com as hipóteses de trabalho habituais:

- (a) Determine a recta de regressão de mínimos quadrados e obtenha uma estimativa pontual para o incremento no valor esperado do tempo de reacção provocado por um aumento de 5 mg/dl no nível de álcool no sangue. (2.0)

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y}}{\sum_{i=1}^n x_i^2 - n \bar{x}^2} = \frac{19500 - 280 \times 435/7}{14 \times 10^3 - 280^2/7} = \frac{2100}{2800} = 0.75,$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 435/7 - 0.75 \times 280/7 = 32.14,$$

$$\hat{E}[Y|x] = 32.14 + 0.75x.$$

$$\text{Seja } \gamma = E[Y|x+5] - E[Y|x] = \beta_0 + \beta_1(x+5) - (\beta_0 + \beta_1 x) = 5\beta_1.$$

$$\text{Então, } \hat{\gamma} = 5\hat{\beta}_1 = 3.75 \text{ segundos.}$$

- (b) Construa um intervalo de confiança a 99% para o declive da recta de regressão. Poderá concluir que existe uma relação de tipo linear entre o valor esperado do tempo de reacção e o nível de álcool no sangue? Justifique. (3.5)

$$\text{Sejam } T = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\hat{\sigma}^2}{\sum x_i^2 - 7\bar{x}^2}}} \sim t_{(5)} \text{ e } a = F_{t(5)}^{-1}(0.995) = 4.032.$$

$$P(-a \leq T \leq a) = 0.99 \Leftrightarrow P\left(\hat{\beta}_1 - a \sqrt{\frac{\hat{\sigma}^2}{\sum x_i^2 - 7\bar{x}^2}} \leq \beta_1 \leq \hat{\beta}_1 + a \sqrt{\frac{\hat{\sigma}^2}{\sum x_i^2 - 7\bar{x}^2}}\right) = 0.99.$$

$$IAC_{0.99}(\beta_1) = \left[\hat{\beta}_1 - 4.032 \sqrt{\frac{\hat{\sigma}^2}{\sum x_i^2 - 7\bar{x}^2}}, \hat{\beta}_1 + 4.032 \sqrt{\frac{\hat{\sigma}^2}{\sum x_i^2 - 7\bar{x}^2}} \right].$$

$$\hat{\sigma}^2 = \frac{1}{n-2} \left[\left(\sum_{i=1}^n y_i^2 - n\bar{y}^2 \right) - (\hat{\beta}_1)^2 \left(\sum_{i=1}^n x_i^2 - n\bar{x}^2 \right) \right] = \frac{1}{5} [(28725 - 435^2/7) - 0.75^2 \times 2800] = 23.571.$$

$$IC_{0.99}(\beta_1) = [0.38, 1.12].$$

A um nível de significância de 0.01 pode-se concluir que há uma relação de tipo linear entre x e Y ($\beta_1 \neq 0$) uma vez que $0 \notin IC_{0.99}(\beta_1)$.