Indicadores para dados univariados

$x_1, x_2, \cdots, x_n$

f.d. empírica $\quad F^*(x) = \dfrac{\text{nº de } x_i \leq x}{n}$

amostra ordenada $\quad x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$

mediana $\quad \tilde{x} = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & n \text{ ímpar} \\ \dfrac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right) & n \text{ par} \end{cases}$

Quantil de ordem θ $(0 < \theta < 1)$

$$Q^*_\theta = \begin{cases} \dfrac{1}{2}\left(x_{(n\theta)} + x_{(n\theta+1)}\right) & \text{se } n\theta \text{ inteiro} \\ x_{([n\theta]+1)} & \text{se } n\theta \text{ não inteiro} \end{cases}$$

$[n\theta]$ designa a parte inteira de $n\theta$

barreira inferior $\quad BI = Q_1 - 1.5(Q_3 - Q_1)$

barreira superior $\quad BS = Q_3 + 1.5(Q_3 - Q_1)$

$Q_1 = Q^*_{0.25}; \quad Q_2 = Q^*_{0.5}; \quad Q_3 = Q^*_{0.75}$

$$s_x^2 = \frac{\sum_{i=1}^n (x_i - \overline{x})^2}{n-1} = \frac{n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2}{n(n-1)}$$

se $x_i' = a + bx_i$ então $\overline{x'} = a + b\overline{x}, \quad s^2_{x'} = b^2 s_x^2$

Dados agrupados em c classes

n_i, f_i, x_i' frequência absoluta, frequência relativa e ponto médio da classe i, respectivamente

$\overline{x} \simeq \frac{\sum_{i=1}^c n_i x_i'}{n} = \sum_{i=1}^c f_i x_i'$

$$s_x^2 \simeq \frac{\sum_{i=1}^c n_i\ (x_i' - \overline{x})^2}{n} = \frac{\sum_{i=1}^c n_i\ x_i'^2}{n} - \overline{x}^2$$

Seja k a primeira classe tal que $F_k \geq \theta$

$$Q^*_\theta \simeq x_k^{\min} + \left(x_k^{\max} - x_k^{\min}\right)\frac{\theta - F_{k-1}}{f_k}$$

Indicadores para dados bivariados

$(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$

$$cov(x,y) = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{n-1} = \frac{n\sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n(n-1)}$$

$r = r_{x,y} = \dfrac{cov(x,y)}{s_x s_y} \qquad$ se $s_x \neq 0, s_y \neq 0$

se $x_i' = a + bx_i$ e $y_i' = c + dy_i \qquad b, d \neq 0$

$cov(x', y') = bd\ cov(x,y)$

$$r_{x',y'} = \begin{cases} r_{x,y} & \text{se } bd > 0 \\ -r_{x,y} & \text{se } bd < 0 \end{cases}$$

Regressão linear simples

$\widehat{y}_i = b_0 + b_1 x_i \qquad \widehat{y}_i - \overline{y} = b_1(x_i - \overline{x})$

$y_i = b_0 + b_1 x_i + e_i$

$$\begin{cases} b_1 = \dfrac{cov(x,y)}{s_x^2} = r\dfrac{s_y}{s_x}, & s_x \neq 0 \\ b_0 = \overline{y} - b_1\overline{x} \end{cases}$$

$\sum_{i=1}^n (y_i - \overline{y})^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - \overline{y})^2$

$\Leftrightarrow SQ_T = SQ_E + SQ_R$

$$R^2 = \frac{SQ_R}{SQ_T} = \frac{cov^2(x,y)}{s_x^2\ s_y^2} = r^2$$

Estimação e Inferência

Erro quadrático médio de $\widehat{\Theta}$, estimador de um parâmetro θ

$EQM[\widehat{\Theta}] = \left(E\left[\widehat{\Theta}\right] - \theta\right)^2 + Var[\widehat{\Theta}]$

$(X_1, X_2, ..., X_n)$ amostra aleatória retirada de uma população $\mathcal{N}(\mu, \sigma)$

$$\overline{X} = \frac{\sum_{i=1}^n X_i}{n} \quad \text{e} \quad \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \frown \mathcal{N}(0,1)$$

$$S^2 = \frac{\sum_{i=1}^n \left(X_i - \overline{X}\right)^2}{n-1} \quad \text{e} \quad \frac{(n-1)S^2}{\sigma^2} \frown \chi^2_{(n-1)}$$

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} \frown t_{(n-1)}$$

$X_1 \frown \mathcal{N}(\mu_1, \sigma_1) \quad$ e $\quad X_2 \frown \mathcal{N}(\mu_2, \sigma_2)$

e tendo duas amostras aleatórias independentes, uma de cada população, com dimensões n_1 e n_2, respectivamente.

$$\overline{X_1} - \overline{X_2} \frown \mathcal{N}\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

$$\frac{\sigma_2^2\ S_1^2}{\sigma_1^2\ S_2^2} \frown F_{(n_1-1, n_2-1)}$$

Testes de hipóteses

H_0 hipótese nula

H_1 hipótese alternativa

P(erro de 1^a espécie)=P(Rejeitar H_0| H_0 verdadeira)= α

P(erro de 2^a espécie)=P(Não rejeitar H_0| H_0 falsa) $=\beta$

regra de decisão:

– rejeitar H_0 se *p-value* $\leq \alpha$

O Teste de Shapiro Wilk

H_0 : X tem distribuição normal

H_1: X não tem distribuição normal

Probabilidade de acontecimentos

$P(A-B)=P(A\cap\overline{B})=P(A)-P(A\cap B)$

$P(\overline{A}|B)=1-P(A|B)$ se $P(B)\neq 0$

$P(A\cup B\cup C)=P(A)+P(B)+P(C)-P(A\cap B)$
$\quad -P(A\cap C)-P(B\cap C)+P(A\cap B\cap C)$

Variáveis aleatórias

X v.a. contínua com função densidade $f_X(x)$ e
$Y=\varphi(X)$, estritamente monótona e derivável, então
$f_Y(y)=f_X(x)\left|\frac{dx}{dy}\right|$ com $x=\varphi^{-1}(y)$

Parâmetros (de funções) de uma v.a. X

$$E[\varphi(X)]=\begin{cases}\sum_i \varphi(x_i)p_i, & \text{se } X \text{ discreta}\\ \int_{-\infty}^{+\infty}\varphi(x)f(x)dx, & \text{se } X \text{ contínua}\end{cases}$$

$\sigma^2=Var[X]=E[(X-\mu)^2]=E[X^2]-\mu^2 \quad (\mu=E[X])$

$Var[a+bX]=b^2Var[X]$

Função geradora de momentos

$M_X(t)=E[e^{tX}]$

$M_{a+bX}(t)=e^{at}M_X(bt)$

Se X e Y v.a.'s independentes $M_{X+Y}(t)=M_X(t)M_Y(t)$

Se $M_X(t)$ está definida numa vizinhança de zero

$\frac{d^{(r)}M_X(t)}{dt^r}_{|t=0}=E[X^r], \quad r=1,2,...$

Par aleatório (X,Y) discreto com função massa de probabilidade conjunta $P[X=x_i,Y=y_j]=p_{ij}$

marginais $p_{i\cdot}=\sum_j p_{ij}$ $\quad p_{\cdot j}=\sum_i p_{ij}$

condicional $P[X=x_i|Y=y_j]=\frac{p_{ij}}{p_{\cdot j}}$

Par aleatório (X,Y) contínuo com função densidade conjunta $f(x,y)$

$f_X(x)=\int_{-\infty}^{+\infty}f(x,y)dy \quad f_Y(y)=\int_{-\infty}^{+\infty}f(x,y)dx$

$f_{X|Y=y}(x)=\frac{f(x,y)}{f_Y(y)}$

Parâmetros de funções de um par aleatório (X,Y).

$$E[g(X,Y)]=\begin{cases}\sum_i\sum_j g(x_i,y_j)p_{ij} & (X,Y) \text{ discreto}\\ \iint_{\mathbb{R}^2} g(x,y)f(x,y)dxdy & (X,Y) \text{ contínuo}\end{cases}$$

$\sigma_{X,Y}=Cov[X,Y]=E[(X-\mu_X)(Y-\mu_Y)]$
$\quad =E[XY]-E[X]E[Y] \quad (\mu_X=E[X],\mu_Y=E[Y])$

$Var[X\pm Y]=Var[X]+Var[Y]\pm 2\,Cov[X,Y]$

$Cov[a+bX,c+dY]=bd\;Cov[X,Y] \quad a,b,c,d\in\mathbb{R}$

$\rho=\rho_{X,Y}=\frac{Cov[X,Y]}{\sigma_X\sigma_Y} \quad \sigma_X,\sigma_Y\neq 0$

$\rho_{a+bX,c+dY}=\rho_{X,Y}$ se $bd>0 \quad a,b,c,d\in\mathbb{R}$

Distribuição binomial

$X\frown\mathcal{B}(n,p)\Leftrightarrow n-X\frown\mathcal{B}(n,q)$ com $q=1-p$

Aproximações das distribuições

$X\frown\mathcal{H}(N,n,k)$ e $\frac{N}{n}>10\Rightarrow X\sim\mathcal{B}(n,p)$ com $p=\frac{k}{N}$

$X\frown\mathcal{B}(n,p)$, $n\geq 20$ e $p\leq 0.05\Rightarrow X\sim\mathcal{P}(\lambda)$ com $\lambda=np$

$X\frown\mathcal{B}(n,p)$, $np>5$ e $nq>5\Rightarrow X\sim\mathcal{N}(\mu,\sigma)$
com $\mu=np,\sigma=\sqrt{npq}$

$X\frown\mathcal{P}(\lambda)$ e $\lambda>12\Rightarrow X\sim\mathcal{N}(\mu,\sigma)$ com $\mu=\lambda,\sigma=\sqrt{\lambda}$

Teorema Limite Central

Sejam $X_1,\ldots,X_n$, v.a.'s independentes e identicamente distribuídas com valor médio μ e variância σ^2 (finita).

$S_n=\sum_{i=1}^n X_i$ e $\overline{X}_n=\frac{1}{n}\sum_{i=1}^n X_i$ tem-se, c/ n grande

$\frac{S_n-n\mu}{\sigma\sqrt{n}}\sim\mathcal{N}(0,1)$ e $\frac{\overline{X}_n-\mu}{\sigma/\sqrt{n}}\sim\mathcal{N}(0,1)$

Construção de modelos (distribuições)

n v.a.'s $Z_i\frown\mathcal{N}(0,1)$ independentes $\Rightarrow\sum_{i=1}^n Z_i^2\frown\chi^2_{(n)}$

$Z\frown\mathcal{N}(0,1)$ e $Y\frown\chi^2_{(n)}$ independentes $\Rightarrow\frac{Z}{\sqrt{Y/n}}\frown t_{(n)}$

$X\frown\chi^2_{(m)}$ e $Y\frown\chi^2_{(n)}$ independentes $\Rightarrow\frac{X/m}{Y/n}\frown F_{(m,n)}$

$X\frown F_{(m,n)})\Rightarrow\frac{1}{X}\frown F_{(n,m)}$

Expressões úteis

Combinações de n elementos k a k, $n,k\in\mathbb{N}_0$

$C_k^n=\binom{n}{k}=\frac{n!}{k!(n-k)!} \quad k\leq n$

$\int_0^{+\infty}x^ne^{-x}dx=n!, \quad n\in\mathbb{N}_0$

$e^x=1+x+\frac{x^2}{2!}+\frac{x^3}{3!}+\cdots=\sum_{n=0}^{\infty}\frac{x^n}{n!}$

$\sum_{k=0}^{n}ar^k=\frac{a(1-r^{n+1})}{1-r} \quad \sum_{k=0}^{\infty}ar^k=\frac{a}{1-r},$ se $|r|<1$

Algumas regras de primitivas

Uma primitiva de xe^{-x} é $-e^{-x}(x+1)$

$P(fg)=Fg-P(Fg') \quad F=Pf$

$P(f'f^\alpha)=\frac{f^{\alpha+1}}{\alpha+1}+C \quad \alpha\in\mathbb{R}\setminus\{-1\}$

$P(f'e^f)=e^f+C \quad P\left(\frac{f'}{f}\right)=log|f|+C$