ISEC Lisboa

Ciência de Dados

ESCOLA DE GESTÃO, ENGENHARIA E AERONÁUTICA

Formulário

- Axiomática da Probabilidade

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \tag{1}$$

$$P(A \cap B) = 0 \text{ se } A \text{ e } B \text{ são eventos mutuamente exclusivos.} \tag{2}$$

$$P(A \cap B) = P(A)P(B) \text{ se } A \text{ e } B \text{ são eventos independentes.} \tag{3}$$

- Lei da Probabilidade Total

$$P(A) = \sum_i P(A \cap B_i) = \sum_i P(A|B_i)P(B_i) \tag{4}$$

- Teorema de Bayes

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)} \tag{5}$$

Variáveis Discretas

- Função Distribuição de Probabilidade, F

$$F(a) = \sum_{\forall x \leq a} p(x) \tag{6}$$

- Funções de probabilidade marginais de X e Y

$$p_X(x) = \sum_{\forall y} p(x,y) \tag{7}$$

$$p_Y(y) = \sum_{\forall x} p(x,y) \tag{8}$$

onde $p(x,y)$ é a função de probabilidade conjunta de X e Y.

- Valor Esperado & Variância

$$\mathrm{E}[X] = \sum_{\forall i} x_i p(x_i) \tag{9}$$

$$\mathrm{Var}(X) = \sum_{\forall i} (x_i - \mathrm{E}[X])^2 p(x_i) \tag{10}$$

$$\mathrm{E}[g(X)] = \sum_{\forall i} g(x_i)p(x_i) \tag{11}$$

$$\mathrm{Var}(g(X)) = \sum_{\forall i} (g(x_i) - \mathrm{E}[g(X)])^2 p(x_i) \tag{12}$$

onde $g(X)$ é uma função da variável aleatória X.

- Para variáveis aleatórias discretas, X e Y, com função de probabilidade conjunta, $p(x, y)$

$$\mathrm{E}[X] = \sum_{\forall i} \sum_{\forall j} x_i p(x_i, y_j) = \sum_{\forall i} x_i p_X(x_i) \tag{13}$$

$$\mathrm{Var}(X) = \sum_{\forall i} \sum_{\forall j} (x_i - \mathrm{E}[X])^2 p(x_i, y_j) = \sum_{\forall i} (x_i - \mathrm{E}[X])^2 p_X(x_i) \tag{14}$$

$$\mathrm{E}[Y] = \sum_{\forall i} \sum_{\forall j} y_j p(x_i, y_j) = \sum_{\forall j} y_j p_Y(y_j) \tag{15}$$

$$\mathrm{Var}(Y) = \sum_{\forall i} \sum_{\forall j} (y_j - \mathrm{E}[Y])^2 p(x_i, y_j) = \sum_{\forall j} (y_j - \mathrm{E}[Y])^2 p_Y(y_j) \tag{16}$$

$$\mathrm{E}[g(X, Y)] = \sum_{\forall i} \sum_{\forall j} g(X, Y) p(x_i, y_j) \tag{17}$$

$$\mathrm{Var}(g(X, Y)) = \sum_{\forall i} \sum_{\forall j} (g(X, Y) - \mathrm{E}[g(X, Y)])^2 p(x_i, y_j) \tag{18}$$

onde $g(X, Y)$ é uma função das variáveis aleatórias X e Y.

- Covariância entre X e Y com função de probabilidade conjunta, $p(x, y)$

$$\mathrm{Cov}(X, Y) = \sum_{\forall i} \sum_{\forall j} (x_i - \mathrm{E}[X])(y_i - \mathrm{E}[Y]) p(x_i, y_i) \tag{19}$$

Variáveis Contínuas

- Função Distribuição de Probabilidade, F

$$F(a) = \int_{-\infty}^{a} f(x)\ \mathrm{d}x \tag{20}$$

- Funções densidade de probabilidade marginais de X e Y

$$f_X(x) = \int_{-\infty}^{+\infty} f(x, y)\ \mathrm{d}y \tag{21}$$

$$f_Y(y) = \int_{-\infty}^{+\infty} f(x, y)\ \mathrm{d}x \tag{22}$$

onde $f(x, y)$ é a função densidade de probabilidade conjunta de X e Y.

- Valor Esperado & Variância

$$\mathrm{E}[X] = \int_{-\infty}^{+\infty} x f(x)\ \mathrm{d}x \tag{23}$$

$$\mathrm{Var}(X) = \int_{-\infty}^{+\infty} (x - \mathrm{E}[X])^2 f(x)\ \mathrm{d}x \tag{24}$$

$$\mathrm{E}[g(X)] = \int_{-\infty}^{+\infty} g(x) f(x)\ \mathrm{d}x \tag{25}$$

$$\mathrm{Var}(g(X)) = \int_{-\infty}^{+\infty} (g(x) - \mathrm{E}[g(X)])^2 f(x)\ \mathrm{d}x \tag{26}$$

onde $g(X)$ é uma função da variável aleatória X.

– Para variáveis aleatórias contínuas, X e Y, com função densidade de probabilidade conjunta, $f(x,y)$

$$\mathrm{E}[X] = \iint_{-\infty}^{+\infty} x f(x,y)\ \mathrm{d}x\ \mathrm{d}y = \int_{-\infty}^{+\infty} x f_X(x)\ \mathrm{d}x \tag{27}$$

$$\mathrm{Var}(X) = \iint_{-\infty}^{+\infty} (x-\mathrm{E}[X])^2 f(x,y)\ \mathrm{d}x\ \mathrm{d}y = \int_{-\infty}^{+\infty} (x-\mathrm{E}[X])^2 f_X(x)\ \mathrm{d}x \tag{28}$$

$$\mathrm{E}[Y] = \iint_{-\infty}^{+\infty} y f(x,y)\ \mathrm{d}x\ \mathrm{d}y = \int_{-\infty}^{+\infty} y f_Y(y)\ \mathrm{d}y \tag{29}$$

$$\mathrm{Var}(Y) = \iint_{-\infty}^{+\infty} (y-\mathrm{E}[Y])^2 f(x,y)\ \mathrm{d}x\ \mathrm{d}y = \int_{-\infty}^{+\infty} (y-\mathrm{E}[Y])^2 f_Y(y)\ \mathrm{d}y \tag{30}$$

$$\mathrm{E}[g(X,Y)] = \iint_{-\infty}^{+\infty} g(X,Y) f(x,y)\ \mathrm{d}x\ \mathrm{d}y \tag{31}$$

$$\mathrm{Var}(g(X,Y)) = \iint_{-\infty}^{+\infty} (g(X,Y)-\mathrm{E}[g(X,Y)])^2 f(x,y)\ \mathrm{d}x\ \mathrm{d}y \tag{32}$$

onde $g(X,Y)$ é uma função das variáveis aleatórias X e Y.

– Covariância entre X e Y com função densidade de probabilidade conjunta, $f(x,y)$

$$\mathrm{Cov}(X,Y) = \iint_{-\infty}^{+\infty} (x-\mathrm{E}[X])(y-\mathrm{E}[Y]) f(x,y)\ \mathrm{d}x\ \mathrm{d}y \tag{33}$$

Propriedades

- Propriedades do valor esperado

$$\mathrm{E}[X+Y] = \mathrm{E}[X] + \mathrm{E}[Y] \tag{34}$$

$$\mathrm{E}[aX] = a\mathrm{E}[X] \text{ sendo } a \text{ uma constante.} \tag{35}$$

- Propriedades da variância & covariância

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X,Y) \tag{36}$$

$$\mathrm{Var}(aX) = a^2\mathrm{Var}(X) \text{ sendo } a \text{ uma constante.} \tag{37}$$

$$\text{Se } X \text{ e } Y \text{ são variáveis aleatórias independentes}, \mathrm{Cov}(X,Y) = 0 \tag{38}$$

$$\mathrm{Cov}(X,Y) = \mathrm{E}[XY] - \mathrm{E}[X]\mathrm{E}[Y] \tag{39}$$

Distribuições Discretas

- Distribuição Binomial $\rightarrow X$: Número de sucessos numa amostra de n elementos.

$$X \sim \mathcal{B}(n,p) \qquad P(X=i) = \binom{n}{i} p^i (1-p)^{n-i} \tag{40}$$

sendo p a probabilidade de sucesso. A distribuição binomial apresenta $\mathrm{E}[X] = np$ e $\mathrm{Var}(X) = np(1-p)$

- Distribuição Binomial Negativa → X: Número de insucessos até atingir r sucessos.

$$X \sim \mathcal{BN}(r,p) \qquad P(X=i) = \binom{i+r-1}{i}(1-p)^i p^r \tag{41}$$

sendo p a probabilidade de sucesso. A distribuição binomial negativa apresenta $\mathrm{E}[X] = r(1-p)/p$ e $\mathrm{Var}(X) = r(1-p)/p^2$.

- Distribuição Hipergeométrica → X: Número de sucessos numa amostra com dimensão n obtidos de uma população com $N+M$ elementos sem reposição.

$$X \sim \mathcal{H}(N,M,n) \qquad P(X=i) = \frac{\binom{N}{i}\binom{M}{n-i}}{\binom{N+M}{n}} \tag{42}$$

sendo N e M o número de sucessos e insucessos na população, respetivamente. A distribuição hipergeométrica apresenta $\mathrm{E}[X] = nN/(N+M)$ e $\mathrm{Var}(X) = np(1-p)\left(1-\dfrac{n-1}{N+M-1}\right)$.

- Distribuição de Poisson → X: Número de eventos que ocorrem num determinado intervalo de tempo (ou espaço) quantos estes ocorrem a uma taxa constante e independente do tempo (ou espaço) em que o último evento ocorreu.

$$X \sim \mathrm{Poisson}(\lambda) \qquad P(X=i) = \frac{\lambda^i e^{-\lambda}}{i!} \tag{43}$$

sendo λ o número de eventos que ocorrem num determinado intervalo de tempo (ou espaço). A distribuição de Poisson apresenta $\mathrm{E}[X] = \lambda$ e $\mathrm{Var}(X) = \lambda$.

Distribuições Contínuas

- Distribuição Uniforme → Se X é uma variável aleatória contínua que segue uma distribuição uniforme, então a sua função densidade de probabilidade é dada por

$$X \sim \mathcal{U}(a,b) \qquad f(x) = \begin{cases} \dfrac{1}{b-a}, & a \leq x \leq b \\ 0, & \text{caso contrário} \end{cases} \tag{44}$$

apresentando $\mathrm{E}[X] = (a+b)/2$ e $\mathrm{Var}(X) = (b-a)^2/12$.

- Distribuição Exponencial Negativa → Se X é uma variável aleatória contínua que indica a distância (espacial ou temporal) entre eventos de um processo de Poisson, então esta segue uma distribuição exponencial negativa com função densidade de probabilidade dada por

$$X \sim \mathcal{EN}(\lambda) \qquad f(x) = \lambda e^{-\lambda x}, \tag{45}$$

apresentando $\mathrm{E}[X] = 1/\lambda$ e $\mathrm{Var}(X) = 1/\lambda^2$.

- Distribuição Normal → Se X é uma variável aleatória contínua que segue uma distribuição normal, então a sua função densidade de probabilidade é dada por

$$X \sim \mathcal{N}(\mu,\sigma^2) \qquad f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\dfrac{-(x-\mu)^2}{2\sigma^2}}, \tag{46}$$

apresentando $\mathrm{E}[X] = \mu$ e $\mathrm{Var}(X) = \sigma^2$.

Teorema do Limite Central

- Se $X_1, X_2, \ldots, X_n$ é uma sequência de variáveis aleatórias independentes e distribuídas identicamente, cada uma com uma média μ e variância σ^2, então para n grande ($n > 30$), a distribuição de
$$X_1 + X_2 + \cdots + X_n \sim \mathcal{N}(n\mu, n\sigma^2) \tag{47}$$

Estimação Pontual

- Estimador de Máxima Verossimilhança (EMV). O valor de θ que maximiza a função
$$\mathcal{L}(\theta|\mathbf{x}) = \prod_{i=1}^{n} f(x_i|\theta) \text{ ou } \log \mathcal{L} = \sum_{i=1}^{n} \log\left(f(x_i|\theta)\right), \tag{48}$$
é a estimativa de máxima verosimilhança, onde $f(x_i|\theta)$ é a função probabilidade ou densidade de probabilidade das observações $(X_1, X_2, \ldots, X_n)$ parametrizada por θ.

- Qualidade de um estimador (Erro médio quadrático)
$$\text{EMQ}(\hat{\Theta}) = (E(\hat{\Theta}) - \theta)^2 + \text{Var}(\hat{\Theta}) \tag{49}$$

Estatística Descritiva

Considerando uma amostra $(X_1, X_2, \ldots, X_n)$, então

- Média Amostral, $\overline{X}$
$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \tag{50}$$

- Variância Amostral, S^2
$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2 \tag{51}$$

Estimação Intervalar

- Intervalo de confiança (de $1 - \alpha$) para μ com μ desconhecido e variância σ^2 conhecida.
$$\left[\overline{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right] \tag{52}$$

- Intervalo de confiança (de $1 - \alpha$) para μ com μ desconhecido e variância σ^2 desconhecida.
$$\left[\overline{X} - t_{\alpha/2,n-1} \frac{S}{\sqrt{n}}, \overline{X} + t_{\alpha/2,n-1} \frac{S}{\sqrt{n}}\right] \tag{53}$$

- Intervalo de confiança para σ^2 com μ desconhecido e variância σ^2 desconhecida.
$$\left[\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}}, \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}\right] \tag{54}$$

Teste de Hipótese

- Teste à média (com variância σ^2 conhecida)

$$H_0: \mu = \mu_0 \quad H_a: \mu \neq \mu_0 \quad | \quad H_0: \mu \leq \mu_0 \quad H_a: \mu > \mu_0 \quad | \quad H_0: \mu \geq \mu_0 \quad H_a: \mu < \mu_0$$

$$\mathrm{ET} = \sqrt{n}\frac{\overline{X} - \mu_0}{\sigma} \sim \mathcal{N}(0,1) \tag{55}$$

- Teste à média (com variância σ^2 desconhecida)

$$H_0: \mu = \mu_0 \quad H_a: \mu \neq \mu_0 \quad | \quad H_0: \mu \leq \mu_0 \quad H_a: \mu > \mu_0 \quad | \quad H_0: \mu \geq \mu_0 \quad H_a: \mu < \mu_0$$

$$\mathrm{ET} = \sqrt{n}\frac{\overline{X} - \mu_0}{S} \sim t_{n-1} \tag{56}$$

- Teste à variancia

$$H_0: \sigma^2 = \sigma_0^2 \quad H_a: \sigma^2 \neq \sigma_0^2 \quad | \quad H_0: \sigma^2 \leq \sigma_0^2 \quad H_a: \sigma^2 > \sigma_0^2 \quad | \quad H_0: \sigma^2 \geq \sigma_0^2 \quad H_a: \sigma^2 < \sigma_0^2$$

$$\mathrm{ET} = \frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2_{n-1} \tag{57}$$

- Teste de amostras emparelhadas

$$H_0: \mu_w = 0 \quad H_a: \mu_w \neq 0 \quad | \quad H_0: \mu_w \leq 0 \quad H_a: \mu_w > 0 \quad | \quad H_0: \mu_w \geq 0 \quad H_a: \mu_w < 0$$

$$\mathrm{ET} = \sqrt{n}\frac{\overline{W}}{S_W} \sim t_{n-1} \tag{58}$$

- Teste de aderência (χ^2)

$$H_0: X \sim \mathcal{F}_\theta \qquad H_a: X \nsim \mathcal{F}_\theta$$

$$\mathrm{ET} = \sum_{i=1}^{k} \frac{(X_i - n\hat{p}_i)^2}{n\hat{p}_i} \sim \chi^2_{k-1-m} \tag{59}$$

sendo k o número de classes e m o número de parâmetros da distribuição $\mathcal{F}$ estimados pelo método da máxima verossimilhança.

Alerta: De modo a garantir a validade do teste de aderência, verifique que $n\hat{p}_i$ (frequência esperada) é pelo menos 5 para todas as classes. Se tal não for verificado, agrupe classes vizinhas, somando-as, até que a frequência esperada seja superior a 5 para todas as novas classes.