Formulario compatto di Probabilità e Statistica

Di Agostino Manuel

Gennaio 2023

1 Combinatoria

- Permutazioni semplici n elementi: $P_n = n! = n \cdot (n-1) \cdot \ldots \cdot 1$
- Combinazioni semplici classe k: $C_{n,k} = \binom{n}{k} = \frac{n!}{k!(n-k)!}$
- Disposizioni semplici classe k: $D_{n,k} = P_k \cdot C_{n,k}$
- Principio inclusione-esclusione: $|A \cup B| = |A| + |B| - |A \cap B|$
- Probabilità condizionata: $P(A \,|\, B) = \frac{P(A \cap B)}{P(B)}$
- I formula di Bayes: $P(A \,|\, B) = \frac{P(B \,|\, A) \cdot P(A)}{P(B)}$
- II formula di Bayes: $P(A \,|\, B) = \frac{P(B \,|\, A) \cdot P(A)}{P(B \,|\, A) \cdot P(A) + P\left(B \,|\, A^C\right) \cdot P\left(A^C\right)}$
- Eventi indipendenti:
$$P(A \,|\, B) = P(A) \quad \Rightarrow \quad P(A \cap B) = P(A) \cdot P(B)$$
Vale anche $P(B \,|\, A) = P(B)$ e sono indipendenti anche gli eventi $\left(A, B^C\right)$, $\left(A^C, B\right)$ e $\left(A^C, B^C\right)$.

2 Variabili aleatorie

- (CdF) Funzione di ripartizione: $F_X(t) := P(X \leq t)$
- Trasformazioni lineari: $Y := aX + b \quad , a \neq 0$
$$F_Y(t) = F_X\left(\frac{t-b}{a}\right) \quad \Rightarrow \quad f_Y(t) = \frac{1}{a} f_X\left(\frac{t-b}{a}\right)$$

3 Vettori aleatori

- **Leggi marginali**
$$f_X(s) = \int_{\mathbb{R}} f_{X,Y}(s,t)\, dt \quad \wedge \quad f_Y(t) = \int_{\mathbb{R}} f_{X,Y}(s,t)\, ds$$
- **CdF congiunta**
$$F_{X,Y}(s,t) := P(X \leq s, Y \leq t) = \int_{-\infty}^{t} \int_{-\infty}^{s} f_{X,Y}(s,t)\, ds\, dt$$
- **Indipendenza**
  - continuo: $f_{X,Y}(s,t) = f_X(s) \cdot f_Y(t)$
  - discreto: $\varphi_{X,Y}(s,t) = \varphi_X(s) \cdot \varphi_Y(t)$
  - generale: $F_{X,Y}(s,t) = F_X(s) \cdot F_Y(t)$

  Si dimostra che se $\exists\, g, h \mid f_{X,Y}(s,t) = g(s) \cdot h(t)$ allora f, y sono indipendenti e $f_X(s) = g(s)$, $f_Y(t) = h(t)$
- **Leggi condizionali su vettori aleatori**
  - continuo: $f_{X|Y}(s \mid t) = \frac{f_{X,Y}(s,t)}{f_Y(t)}$
  - discreto: $\varphi_{X|Y}(k \mid j) = \frac{\varphi_{X,Y}(k,j)}{\varphi_Y(j)}$
- **Legge di una funz. di un vett. aleatorio**
  - Somma: siano X, Y indipendenti e non negative e $S := X + Y$ allora
$$f_S(t) = \int_{\mathbb{R}} f_X(u, t-u)\, du = \int_{\mathbb{R}} f_Y(t-v, v)\, dv$$
  - Minimo & Massimo: siano $X_1, X_2, \ldots, X_n$ vv.aa. indipendenti e $R := min(X_i)$, $T := max(X_i)$ allora
$$F_T(t) = \prod_i F_{x_i}(t) \quad \wedge \quad 1 - F_R(t) = \prod_i \left(1 - F_{x_i}(t)\right)$$
- **Valore atteso**
  - discreto: $E(X) := \sum_k k \cdot \varphi_X(k)$
  - continuo: $E(X) := \int_{\mathbb{R}} t \cdot f_X(t)\, dt$
  - costante $a \in \mathbb{R}$: $E(a) = a$
  - *Teorema*
    Sia X v.a. con pdf $f_X(t)$ o pmf $\varphi_X(k)$ e una funzione $g : \mathbb{R} \to \mathbb{R}$, $Y = g(X)$ allora
$$E(Y) := \sum_k g(k) \cdot \varphi_X(k) \quad \wedge \quad E(Y) := \int_{\mathbb{R}} g(t) \cdot f_X(t)\, dt$$
    In presenza di un vettore aleatorio X, Y e $h : \mathbb{R}^2 \to \mathbb{R}$, $Z = h(X, Y)$ allora
$$E(Z) := \sum_j \sum_k h(j,k) \cdot \varphi_{X,Y}(j,k)$$
$$E(Z) := \iint_{\mathbb{R}^2} h(u,v) \cdot f_{X,Y}(u,v)\, du\, dv$$
  - vv. aa. positive: $E(X) = \int_0^\infty \left[1 - F_x(t)\right] dt$
- **Covarianza**
  - Definizione
$$\begin{aligned} Cov(X,Y) &:= E(XY) - E(X)E(Y) \\ &= E\left[(X - E(X))(Y - E(Y))\right] \end{aligned}$$
  - La covarianza è simmetrica e bilineare:
$$Cov\left(\sum_i^n a_i x_i, \sum_j^n b_j y_j\right) = \sum_i^n \sum_j^n a_i b_i \cdot Cov\left(x_i, y_i\right)$$
  - se X, Y indipendenti $Cov(X,Y) = 0$
- **Varianza**
  - $Var(X) := E\left[(X - E(X))^2\right] = E(X^2) - E(X)^2$
  - $Cov(X,X) \equiv Var(X)$
  - $Var\left(\sum_i^n X_i\right) = \sum_i^n Var(X_i) \; + \; \sum_{1 \leq i \leq j \leq n}^n Cov(X_i, X_j)$
  - $Var\left(aX + b\right) = a^2 \cdot Var\left(X\right)$
- **Moda**
  È l'insieme delle controimmagini dei valori più probabili di una v. a.
- **Mediana**
  Se esiste è un numero $M_X : P(X \leq M_X) = P(X \geq M_X)$. Si trova ponendo $F_X(M_X) = \frac{1}{2}$

4 Legge debole dei grandi numeri

- **Disuguaglianza di *Markov***
  Se X v. a. non negativa allora $\forall a \in \mathbb{R} \quad P(X \geq a) \leq \frac{E(X)}{a}$
- **Disuguaglianza di *Chebyshev***
  Se X v. a. non negativa allora $\forall a \in \mathbb{R}$
$$P(|X - E(X)| \geq a) \leq \frac{Var(X)}{a^2}$$
  ponendo $\mu = E(X)$, $\sigma = \sqrt{Var(X)}$ si ha
$$P(|X - E(X)| \geq a \cdot \sigma) \leq \frac{1}{a^2}$$
- **Legge debole**
$$\lim_{n \to +\infty} P\left(\left|\overline{X}_n - \mu\right| \geq \varepsilon\right) = 0$$

5 Esempi notevoli

- **Bernoulliana** $bin(1,p)$
  Considera un esperimento con due esiti possibil; generalmente successo con probabilità p e insuccesso con probabilità $1-p$.
$$\varphi_X(k) = \begin{cases} p & k = 0 \\ 1-p & k = 1 \end{cases}$$
$$E(X) = p,\; Var(X) = p(1-p)$$
- **Binomiale** $bin(n,p)$
  Modella la realizzazione di n ripetizioni indipendenti di un esperimento, ciascuna delle quali può concludersi con un successo (prob. p) o insuccesso (prob. $1-p$); X è il numero di successi.
$$\varphi_X(k) = \binom{n}{k} \cdot p^k (1-p)^{n-k}$$
$$E(X) = np\,,\; Var(X) = np(1-p)$$
  - È riproducibile: se le due v.a. indipendenti allora
$$bin(n_1, p) + bin(n_2, p) = bin(n_1 + n_2, p)$$
  - Si approssima con $\mathcal{N}(np, np(1-p))$ (e *correzione di continuità*) se $n \geq 2, 3$
- **Poissoniana** $pois(\nu)$
  Si usa quando n cresce in maniera arbitraria e p è piccolo. μ è il numero medio di successi.
$$\varphi_X(k) = \frac{\nu^k}{k!} e^{-\nu} \quad k = 0, 1, 2, \ldots$$
$$E(X) = Var(X) = \nu$$
  - È riproducibile: $pois(\nu_1) + pois(\nu_2) = pois(\nu_1 + \nu_2)$
  - Si approssima con $\mathcal{N}(n\nu, n\nu)$ se $\nu \geq 20, 30$
- **Uniforme** $unif(a,b)$
  Due parametri a e b, estremi dell'intervallo. $V_x = [a, b]$
$$f_X(t) = \begin{cases} \frac{1}{b-a} & \text{se } a \leq x \leq b \\ 0 & \text{altrimenti} \end{cases}$$
$$F_X(t) = \begin{cases} 0 & t \leq a \\ \frac{t-a}{b-a} & a < t \leq b \\ 1 & t > b \end{cases}$$
$$E(X) = \frac{a+b}{2}\,,\; Var(X) = \frac{(b-a)^2}{12}$$
  - Trasformazioni lineari: $unif \longrightarrow unif$
    Se $U \sim unif(a,b)$ e $V := mU + q \quad , m \neq 0$ allora $V \sim unif(ma + q, mb + q)$
- **Esponenziale** $expo(\lambda)$
  Prevede un parametro λ detto *rate* o tasso di accadimento. $V_x = [0, +\infty)$.
$$f_X(t) = \lambda \cdot e^{-\lambda t} \quad , t \geq 0$$
$$F_X(t) = 1 - e^{-\lambda t}$$
$$E(X) = \frac{1}{\lambda}\,,\; Var(X) = \frac{1}{\lambda^2}$$
  - Omotetie: $expo \longrightarrow expo$
$$\alpha \cdot expo(\lambda) = expo\left(\frac{\lambda}{\alpha}\right) \quad , \alpha > 0$$
  - Assenza di memoria:
$$P(T > t + t_0 \,|\, T > t_0) = P(T > t) \quad \forall t, t_0 > 0$$

- Minimo di esponenziali indipendenti
  Siano $T_1 \sim expo(\lambda_1)$, $T_2 \sim expo(\lambda_2), \dots, T_n \sim expo(\lambda_n)$ indipendenti, allora
  $$S := min\{T_i\} \sim expo(\lambda_1 + \lambda_2 + \ldots + \lambda_n)$$
- Si approssima con $\mathcal{N}(\frac{n}{\lambda}, \frac{n}{\lambda^2})$ se $n \geq 20, 30$

- **Gaussiana o normale** $\mathcal{N}(\mu, \sigma^2)$
$$f_X(t) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(t-\mu)^2}{2\sigma^2}}$$
$$F_X(t) = \Phi\left(\frac{t-\mu}{\sigma}\right)$$
$$E(X) = \mu \, , \, Var(X) = \sigma^2$$
  - Trasformazioni lineari: $\mathcal{N} \longrightarrow \mathcal{N}$
  $$Z \sim \mathcal{N}(\mu, \sigma^2) \, , \, X := aZ + b \quad \Rightarrow \quad X \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$$
  - Valori notevoli:
  $$P(|X-\mu| \leq \sigma) \approx 69.3\%$$
  $$P(|X-\mu| \leq 2\sigma) \approx 95.5\%$$
  $$P(|X-\mu| > 3\sigma) \approx 0.27\%$$
  - Simmetrie :
  $$\Phi(-\alpha) = 1 - \Phi(\alpha) \quad , \quad \Phi^{-1}(\alpha) = -\Phi^{-1}(1-\alpha)$$
  - È riproducibile:
  $$\mathcal{N}(\mu_1, \sigma_1^2) + \mathcal{N}(\mu_2, \sigma_2^2) = \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

- **Teorema del limite centrale (*TLC*)**
  Siano $X_1, X_2, \ldots, X_n$ i.i.d. con $E(X_i) = \mu$, $Var(X_i) = \sigma^2$ allora
$$S_n := \sum_i^n X_i \dot\sim \mathcal{N}(n\mu, n\sigma^2)$$
$$Z_n := \frac{S_n - n\mu}{\sqrt{n}\sigma} \dot\sim \mathcal{N}(0,1)$$
$$\overline{X}_n := \frac{S_n}{n} \dot\sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$$

- **Chi quadro** $\chi^2(n)$
  È somma di quadrati di $\mathcal{N}(0,1)$; ha un parametro n, il numero di Gaussiane da sommare. $V_x = (0, +\infty)$
$$f_{\chi^2_n}(t) = c_n \cdot t^{\frac{n}{2}-1} \cdot e^{-\frac{t}{2}} \quad t > 0$$
$$E(\chi^2_n) = n \, , \, Var(\chi^2_n) = 2n$$
  - È riproducibile:
  $$\chi^2(n_1) + \chi^2(n_2) = \chi^2(n_1 + n_2)$$
  - Collegamento con variabili aleatorie esponenziali
  Siano $T_1, T_2, \ldots, T_n \mid T_i \sim expo(\lambda)$ i.i.d. allora
  $$2\lambda(T_1 + T_2 + \ldots + T_n) \dot\sim \chi^2(2n)$$
  - Per valori grandi di n:
  $$F_{\chi^2(n)}(t) \sim \Phi\left(\frac{t-n}{\sqrt{2n}}\right), \; F^{-1}{}_{\chi^2(n)}(t) \sim n + \sqrt{2n}\Phi^{-1}(t)$$

- t_n **di Student**
  Ha un parametro n che indica i *gradi di libertà*. Date Z normale standard e $C_n \sim \chi^2(n)$ allora
$$f_{t_n}(t) = \frac{Z}{\sqrt{\frac{C_n}{n}}}$$
$$E(t_n) = 0 \quad n \geq 2 \, , \, Var(t_n) = \frac{n}{n-2} \quad n \geq 3$$
  - La funzione di ripartizione ha le stesse proprietà di quella della gaussiana

6 Statistica

- **Media campionaria**
$$\overline{X} := \frac{1}{n}\sum_{i=1}^n X_i$$
- **Deviazione standard campionaria**
$$S_x := \sqrt{\frac{1}{n-1}\sum_{i=1}^n \left(X_i - \overline{X}\right)^2}$$
$$\tilde{S}_x := \sqrt{\frac{1}{n}\sum_{i=1}^n (X_i - \mu)^2} \quad , \mu \text{ nota}$$

Incognita	Funzione ancillare	V.a. di appr.
μ	$\frac{\overline{X}-\mu}{\sigma} \cdot \sqrt{n}$	$\mathcal{N}(0,1)$
	$\frac{\overline{X}-\mu}{S_x} \cdot \sqrt{n}$	$t(n-1)$
σ	$\frac{S_x^2}{\sigma^2} \cdot (n-1)$	$\chi^2(n-1)$
	$\frac{\tilde{S}_x^2}{\sigma^2} \cdot n$	$\chi^2(n)$
[nei test] p	$\frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}}$	$\mathcal{N}(0,1)$
[negli intervalli]	$\frac{\hat{p}-p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$	$\mathcal{N}(0,1)$
λ	$2n \cdot \lambda\overline{X}$	$\chi^2(2n)$
$\mu_x - \mu_y$	$\frac{\overline{X}-\overline{Y}-(\mu_x-\mu_y)}{\sqrt{\frac{\sigma_x^2}{m}+\frac{\sigma_y^2}{n}}}$	$\mathcal{N}(0,1)$
	$\frac{\overline{X}-\overline{Y}-(\mu_x-\mu_y)}{\sqrt{\frac{S_x^2}{m}+\frac{S_y^2}{n}}}$	$\mathcal{N}(0,1)$
$[\sigma_x \approx \sigma_y]$	$\frac{\overline{X}-\overline{Y}-(\mu_x-\mu_y)}{S_p \cdot \sqrt{\frac{1}{m}+\frac{1}{n}}}$	$t(m+n-2)$
σ	$\frac{S_p^2}{\sigma^2} \cdot (m+n-2)$	$\chi^2(m+n-2)$

  - Lo *stimatore pooled* $S_p \approx \sigma$ è dato da
  $$\sqrt{\frac{m-1}{m+n-2} \cdot S_X^2 + \frac{n-1}{m+n-2} \cdot S_Y^2}$$

- **Curva operativa e *potenza* del test**
  La funzione che dà $P(\text{dire } H_0)$ in funzione di un parametro θ si chiama curva OC.
  $$h(\theta) := P_\theta(\text{dire } H_0)$$
  Ricordando che
  $$P(\text{err. I sp.}) = P(\text{dire } H_1 \mid \text{vera } H_0) = \alpha$$
  $$P(\text{err. II sp.}) = P(\text{dire } H_0 \mid \text{vera } H_1) = \beta$$
  allora
  - se vera H_0
  $$\alpha = 1 - h(\theta)$$
  - se vera H_1
  $$\beta = h(\theta)$$
  potenza del test: $1 - h(\theta) = P(\text{dire } H_1)$