Formulario per Statistica Corso Base

1. Statistica descrittiva

Notazione.

- N: dimensione del collettivo o della popolazione
- $x_1, x_2, \ldots, x_N$: valori del carattere X osservati sulle N unità del collettivo
- $X_1, X_2, \ldots, X_k$: k possibili modalità che un carattere può assumere.
- $n_1, n_2, \ldots, n_k$: frequenze assolute con cui le modalità vengono osservate nel collettivo
- $f_1, f_2, \ldots, f_k$: frequenze relative con cui le modalità vengono osservate nel collettivo. Ovviamente, per ogni $j = 1, \ldots, k$, $f_j = n_j/N$.
- $N_1, N_2, \ldots, N_k$: frequenze assolute CUMULATE, dove $N_j = \sum_{i=1}^{j} n_i$.
- $F_1, F_2, \ldots, F_k$: frequenze relative CUMULATE, dove $F_j = \sum_{i=1}^{j} f_i$. Ovviamente, per ogni $j = 1, \ldots, k$, $F_j = N_j/N$.
- $x_{(1)}, x_{(2)}, \ldots, x_{(N)}$: valori del carattere X osservati, in ordine non decrescente, sulle N unità del collettivo.

1.1. Indici statistici

- media aritmetica μ

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i \text{ (distribuzioni unitarie)} \qquad \mu = \frac{1}{N}\sum_{j=1}^{k} X_j n_j = \sum_{j=1}^{k} X_j f_j \text{ (distribuzioni di frequenze)}$$

- varianza

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 = \frac{1}{N}\sum_{j=1}^{k} X_j^2 - \mu^2 \text{ (distribuzioni unitarie)}$$

$$\sigma^2 = \frac{1}{N}\sum_{j=1}^{k}(X_j - \mu)^2 n_j = \sum_{j=1}^{k}(X_j - \mu)^2 f_j = \frac{1}{N}\sum_{j=1}^{k} X_j^2 n_j - \mu^2 \text{ (distribuzioni di frequenze)}$$

- se

$$Y = a + bX$$

allora

$$\mu_Y = a + b\mu_X;\ \sigma_Y^2 = b^2\sigma_X^2$$

- In assenza di altre informazioni, se le modalità sono espresse in classi, si sostituiscono i valori delle modalità X_j sopra scritti con i valori centrali delle classi. Se invece si dispone dei totali di classe, i valori delle X_j si sostituiscono con le medie delle classi.

- mediana Me o $X_{0.5}$: $x_{(N+1)/2}$, per N dispari; qualunque valore compreso tra $x_{(N/2)}$ e $x_{(N/2+1)}$, come ad esempio il valore centrale $(x_{(N/2)} + x_{(N/2+1)})/2$, per N pari.

  Nel caso di distribuzioni di frequenza, è la modalità a cui corrisponde la prima frequenza relativa cumulata superiore o uguale a 0.5.

  Nel caso di distribuzioni con classi di modalità, se la classe mediana è la $i-$esima, con estremo superiore $X_{i,sup}$:

$$X_{0.5} = X_{i-1,sup} + \frac{0.5 - F_{i-1}}{F_i - F_{i-1}}(X_{i,sup} - X_{i-1,sup})$$

- coefficiente di variazione: $CV = \sigma/|\mu|$.
- devianza $D = N\sigma^2$
- codevianza:
  $C_{xy} = \sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y) = \sum_{i=1}^{N} x_i y_i - N\mu_x\mu_y$ (distribuzioni unitarie)
  $C_{xy} = \sum_{u=1}^{r}\sum_{v=1}^{c}(X_u - \mu_x)(y_v - \mu_y)n_{uv} = \sum_{u=1}^{r}\sum_{v=1}^{c} X_u y_v n_{uv} - N\mu_x\mu_y$ (distribuzioni di frequenza)
- covarianza $\sigma_{xy} = C_{xy}/N$
- coefficiente di correlazione

$$r = \frac{\sigma_{xy}}{\sigma_x\sigma_y} = \frac{C_{xy}}{\sqrt{D_x D_y}}$$

- tabelle di contingenza: frequenze teoriche di indipendenza: $\hat{n}_{ij} = n_{i0}n_{0j}/N = n_{i.}n_{.j}/n_{..}$
- indice χ^2

$$\chi^2 = \sum_{i=1}^{s}\sum_{j=1}^{t}\left(\frac{n_{ij} - \hat{n}_{ij}}{\hat{n}_{ij}}\right)^2 \hat{n}_{ij} \quad = \quad \sum_{i=1}^{s}\sum_{j=1}^{t}\frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}} = N\left(\sum_{i=1}^{s}\sum_{j=1}^{t}\frac{n_{ij}^2}{n_{i0}n_{0j}} - 1\right); \psi = \sqrt{\frac{\chi^2}{N}}$$

- dipendenza in media

$$\eta^2_{y|x} = \frac{D_S}{D_y} = \frac{D_{TRA}}{D_y} = \frac{\sum_{i=1}^{r}[(\mu_Y|X_i) - \mu_Y]^2 n_{i0}}{\sum_{j=1}^{c}(y_j - \mu_Y)^2 n_{0j}}.$$

- regressione

$$\hat{y} = b_0 + b_1 x; \quad b_1 = \frac{C_{XY}}{D_X}; \quad b_0 = \mu_y - b_1\mu_X; \quad \hat{y}_i = b_0 + b_1 x_i; \quad e_i = y_i - \hat{y}_i$$

$$Devianza\, spiegata = \sum_{i=1}^{N}(\hat{y}_i - \mu_Y)^2 \quad Devianza\, residua = \sum_{i=1}^{N} e_i^2 \quad R^2 = r^2 = \frac{devianza\ spiegata}{devianza\, di\, Y}$$

2. Probabilità

- Unione di eventi: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- Intersezione di eventi: $P(A \cap B) = P(A) \times P(B|A) = P(B) \times P(A|B)$
- indipendenza: $P(A|B) = P(A|\bar{B}) = P(A)$
- coefficiente binomiale: $C_{n,k} = \binom{n}{k} = \frac{n!}{k!(n-k)!}$
- disposizioni semplici: $D_{n,k} = \frac{n!}{(n-k)!}$
- Teorema di Bayes: per ogni $j = 1, \ldots k$,

$$P(A_j|B) = \frac{P(A_j)P(B|A_j)}{\sum_{h=1}^{k} P(A_h)P(B|A_h)}$$

- Distribuzione binomiale, $X \sim \text{Bin}(n, p)$:

$$P(X = j) = \binom{n}{j} p^j (1-p)^{n-j}, \quad \mu_X = E(X) = np, \quad \text{Var}(X) = \sigma_X^2 = np(1-p).$$

- Distribuzione normale, proprietà:

$$\Phi(z) = 1 - \Phi(-z); \quad x_p = \mu + \sigma\, z_p$$

3. Inferenza

- n = dimensione del campione
- $x_1, x_2, \ldots, x_n$ = valori osservati sul campione
- $\bar{X}$ = variabile casuale media campionaria, $\bar{x}$ = suo valore osservato
- Se $E(X_i) = \mu$ e $var(X_i) = \sigma^2$, $E(\bar{X}) = \mu$ e $var(\bar{X}) = \frac{\sigma^2}{n}$

3.1. Intervalli di confidenza

- Intervallo di confidenza di livello $1-\alpha$ per la media (σ^2 noto)

$$\left(\bar{x} - z_{1-\alpha/2}\sqrt{\frac{\sigma^2}{n}}; \quad \bar{x} + z_{1-\alpha/2}\sqrt{\frac{\sigma^2}{n}}\right)$$

- Intervallo di confidenza di livello $1-\alpha$ per la media (σ^2 incognito)

$$\left(\bar{x} - t_{1-\alpha/2}\sqrt{\frac{S^2}{n}}; \quad \bar{x} + t_{1-\alpha/2}\sqrt{\frac{S^2}{n}}\right)$$

  dove S^2 è lo stimatore corretto della varianza e $t_{1-\alpha/2}$ è il quantile di livello $1-\alpha/2$ della t di Student con $n-1$ g.d.l.

- Intervallo di confidenza di livello $1-\alpha$ per la proporzione per grandi campioni (n elevato)

$$\left(\hat{p} - z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}; \quad \hat{p} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

- Intervallo di confidenza di livello $1-\alpha$ per la media (n elevato e σ^2 incognito)

$$\left(\bar{x} - z_{1-\alpha/2}\sqrt{\frac{S^2}{n}}; \quad \bar{x} + z_{1-\alpha/2}\sqrt{\frac{S^2}{n}}\right)$$

3.2. Test di ipotesi

- Test sulla media di una popolazione normale (σ^2 noto)
  Regione di rifiuto per test unilaterale destro e, rispettivamente, sinistro:

$$\bar{x} \geq \mu_0 + z_{1-\alpha}\sqrt{\frac{\sigma^2}{n}}; \quad \bar{x} \leq \mu_0 - z_{1-\alpha}\sqrt{\frac{\sigma^2}{n}}$$

  in alternativa: $\left| z = \frac{\bar{x}-\mu_0}{\sqrt{\sigma^2/n}} \right| \geq z_{1-\alpha}$;
  livello di significatività effettivo: $P(Z \geq z)$; $P(Z \leq z)$;
  potenza del test condizionata a $H_1 : \mu = \mu_1 > \mu_0$ e, rispettivamente, a $H_1 : \mu = \mu_1 < \mu_0$:

$$P\left(\bar{X} \geq \frac{\mu_0 - \mu_1}{\sqrt{\sigma^2/n}} + z_{1-\alpha}\right); \quad P\left(\bar{X} \leq \frac{\mu_0 - \mu_1}{\sqrt{\sigma^2/n}} - z_{1-\alpha}\right)$$

  .
  Regione di rifiuto per test bilaterale:

$$\bar{x} \leq \mu_0 - z_{1-\alpha/2}\sqrt{\frac{\sigma^2}{n}}; \quad \bar{x} \geq \mu_0 + z_{1-\alpha/2}\sqrt{\frac{\sigma^2}{n}};$$

  in alternativa: $\left| z = \frac{\bar{x}-\mu_0}{\sqrt{\sigma^2/n}} \right| \geq z_{1-\alpha/2}$;
  livello di significatività effettivo: $2P(Z \geq |z|)$;
  potenza del test condizionata a $H_1 : \mu = \mu_1 \neq \mu_0$:

$$P\left(\bar{X} \leq \frac{\mu_0 - \mu_1}{\sqrt{\sigma^2/n}} - z_{1-\alpha/2}\right) + P\left(\bar{X} \geq \frac{\mu_0 - \mu_1}{\sqrt{\sigma^2/n}} + z_{1-\alpha/2})\right)$$

  .

- Test sulla media di una popolazione normale (σ^2 incognito)
  Regione di rifiuto per test unilaterale destro e, rispettivamente, sinistro:

$$\bar{x} \geq \mu_0 + t_{n-1;1-\alpha}\sqrt{\frac{S^2}{n}}; \quad \bar{x} \leq \mu_0 - t_{n-1;1-\alpha}\sqrt{\frac{S^2}{n}}$$

  in alternativa: $\left| t = \frac{\bar{x}-\mu_0}{\sqrt{S^2/n}} \right| \geq t_{n-1;1-\alpha}$;
  livello di significatività effettivo: $P(T_{n-1} \geq t)$; $P(T_{n-1} \leq z)$.
  Regione di rifiuto per test bilaterale:

$$\bar{x} \leq \mu_0 - t_{n-1;1-\alpha/2}\sqrt{\frac{\sigma^2}{n}}; \quad \bar{x} \geq \mu_0 + t_{n-1;1-\alpha/2}\sqrt{\frac{\sigma^2}{n}};$$

  in alternativa: $\left| t = \frac{\bar{x}-\mu_0}{\sqrt{S^2/n}} \right| \geq t_{n-1;1-\alpha/2}$;
  livello di significatività effettivo: $2P(T_{n-1} \geq |t|)$.

- Test per grandi campioni (n elevato) sulla proporzione di una popolazione bernoulliana
  Regione di rifiuto per test unilaterale destro e, rispettivamente, sinistro:

$$\hat{p} \geq p_0 + z_{1-\alpha}\sqrt{\frac{p_0(1-p_0)}{n}}; \quad \hat{p} \leq p_0 + z_{1-\alpha}\sqrt{\frac{p_0(1-p_0)}{n}}$$

  in alternativa: $\left| z = \frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)/n}} \right| \geq z_{1-\alpha}$;
  livello di significatività effettivo: $P(Z \geq z)$; $P(Z \leq z)$.
  Regione di rifiuto per test bilaterale:

$$\hat{p} \leq p_0 + z_{1-\alpha/2}\sqrt{\frac{p_0(1-p_0)}{n}}; \quad \hat{p} \geq p_0 + z_{1-\alpha/2}\sqrt{\frac{p_0(1-p_0)}{n}}$$

  in alternativa: $\left| z = \frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)/n}} \right| \geq z_{1-\alpha/2}$;
  livello di significatività effettivo: $2P(Z \geq |z|)$.