

# Formulario di Statistica

Lezioni di Giulio G. Cantone

Simboli	Significato
$x$	indica un valore generico, di solito numero ma non sempre.
$X, Y$	$X$ è il simbolo di una variabile. $Y$ a volte, ma non sempre, indica un fenomeno che accade dopo $X$ .
$x_i, y_i$	Valore specifico assunto dalla variabile $X$ (o dalla $Y$ per $y_i$ ) nel caso $i$ .
$i$	Indica la riga nella tabella (indice dell'osservazione). Può essere ordinato secondo un criterio.
$K$	Di solito l'insieme dei valori possibili (valori $k$ ) di una variabile. In altri casi $k$ è usato come contatore al posto di $i$ .
$\sum_{\text{inizio}}^{\text{fine}} (...)$	SOMMATORIA: Al posto dei puntini ci sarà una operazione da fare indicizzata dal simbolo che si troverà sotto la $\Sigma$ . Per esempio, $i$ o $k$ possono essere questo valore indice. Ogni risultato di queste operazioni va sommato. Bisogna iniziare dal valore minore e finire al valore sopra la $\Sigma$ . Se sopra la $\Sigma$ non c'è niente, bisogna farlo per tutti i valori indicizzati.
$x!$	Il punto esclamativo ! è il simbolo del prodotto fattoriale di un numero intero non negativo, cioè il prodotto di tutti i numeri interi positivi da 1 a $x$
$n, n_g$	Dimensione del campione, ossia numero di righe nella tabella. $n_g$ è la dimensione di sottogruppo $g$ del campione, definito dalla variabile categoriale $G$ .
$n(x)$	Frequenza assoluta del valore $x$ nel campione. Si noti che $n(g) = n_g$ .
$p(x) \simeq \frac{n(x)}{n}$	Frequenza relativa del valore $x$ nel campione rispetto al totale della dimensione campionaria.
$N$	Dimensione di una popolazione finita.
$P(i)$	Posizione relativa della riga $i$ (si suppone, ordinata dal più piccolo di $X$ al più grande)
$ x $	Valore assoluto. Significa che se il numero fosse negativo, diventerebbe positivo.
$ x - y $	Differenza assoluta. Si fa sempre la differenza tra il numero più grande ed il più piccolo.
$\lim_{i \rightarrow y}(x)$	Limite di $x$ mentre $i$ si sta avvicinando a $y$ . Di solito si usa quando $y = \infty$ .
$Pr(x), Prob(x)$	Probabilità dell'evento $x$ .
$X Y$ , o $A B$	Variabili o eventi condizionati. La stanghetta   si legge "quando"; oppure anche "sapendo che".
$\bar{x}$	La sbarretta sopra il valore generico indica la media campionaria.
$\mathbb{E}(X), \mu(X)$	Operatore di valore atteso (expectation). Indica la media teorica della variabile. A fini pratici è intercambiabile con $\mu$ , simbolo del parametro di locazione (media aritmetica) di una popolazione o distribuzione..
$Quant.(X)_{(q,k)}$	Quantile $k$ dopo aver diviso $X$ in $q$ parti (cioè: $q$ quantili) uguali. Il valore del $k$ quantile è il valore di $X$ dove finisce il $k$ quantile, partendo dai valori più piccoli di $X$ .
$Med.(X)$	Mediana di $X$ . Equivalente a $Quant.(X)_{(q=2,k=1)}$ .
$x_i - \bar{x}$	Questa differenza è chiamata scostamento dalla media, o anche "scarto" o "deviation".
$\sigma_x$ o $\sigma(X)$	Deviazione standard nella popolazione (variabili aleatorie, modelli di distribuzioni notevoli).
$s_x$ , o $s(X)$	Deviazione standard campionaria della variabile $X$ .
$Var.(X)$ o $\sigma^2(X)$	Varianza nella popolazione e nelle variabili aleatorie.
$s^2(X)$	Varianza nel campione osservato di $X$
$Z$ e $z_x$	Di solito $Z$ indica la "standardizzazione" di $X$ . $z_x$ è la standardizzazione di $x$
$e(x)$ o $e(x, y)$	Valore teorico secondo una ipotesi. $e(x, y)$ si riferisce ad una tabella di contingenza.

## Statistica Descrittiva

Frequenza assoluta:

$$n(x) = \sum (\text{volte che osservo } x)$$

Frequenza relativa:

$$p(x) = \frac{n(x)}{n}$$

Frequenza cumulata:

$$P(x_i) = \sum_{INIZIO:i=1}^{FINE:i} p(x)$$

HH Index:

$$\sum (p_x)^2$$

Entropia:

$$-\sum (p_x) \cdot \ln(p_x)$$

Gini su tabella dei dati:

$$A_i = \sum X | x \leq x_i$$

$$P(i) = \frac{i}{n}$$

$$Q(i) = \frac{A_i}{\sum X}$$

$$Gini \simeq \frac{2}{n-1} \cdot \sum_i^{n-1} [P(i) - Q(i)]$$

Media campionaria:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

Mediana per  $n$  dispari:

$$Med.(x) = x_i | i = \frac{n+1}{2}$$

Mediana per  $n$  pari:

$$Med.(x) = \frac{(x_i + x_j)}{2} | [P(i), P(j)] \sim \text{i due numeri pi\u00f9 vicini a } P = .5$$

Formula dei quantili:

- Dividere  $i$  per il "quanto" del quantile, es. quartile = 4.
- Moltiplicare il risultato per la posizione del quantile, es: "Primo" significa moltiplicare per uno, "Secondo" per due...
- Cos\u00ec si trova approssimativamente il valore di  $P(i)$  associato a quel quantile.
- Non \u00e8 sbagliato fare la semi-somma dei due valori pi\u00f9 vicini.

Varianza campionaria:

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

Deviazione standard campionaria:

$$s_x = \sqrt{s^2}$$

Covarianza campionaria:

$$\widehat{Cov.}(X, Y) = \frac{\sum_i [(x_i - \bar{x})(y_i - \bar{y})]}{n-1}$$

Correlazione lineare:

$$\rho(X, Y) = \frac{\widehat{Cov.}(X, Y)}{s_x \cdot s_y}$$

Valore teorico secondo ipotesi di uniformit\u00e0 delle contingenze:

- $k$  e  $j$  sono i valori di riga e colonna della tabella di contingenza

$$e(x_k, y_j) = \frac{\sum [\text{riga di } (k,j)] \cdot \sum [\text{colonna di } (k,j)]}{n}$$

Chi Quadrato:

$$\chi^2(X, Y) = \sum \left( \frac{[n(x_k, y_j) - e(x_k, y_j)]^2}{e(x_k, y_j)} \right)$$

V di Cramer per tabelle di contingenza:

$$V_{Cramer} = \sqrt{\frac{\chi^2}{n \cdot [\min(k, j) - 1]}}$$

## Probabilità per l'inferenza

$n$ : numero di casi possibili

$k$ : numero di casi "soglia"

$p$ : probabilità dell'evento ("probabilità di successo")

$1-p$ : probabilità del complemento dell'evento ("probabilità di fallimento")

*CDF*: Probabilità cumulata di molti eventi possibili, da 0 a  $k$ . Quando invece è da  $n$  a  $k$  si chiama "retrocumulata"

### Metodo della probabilità Binomiale

Si adotta solo quando  $n$  è piccolo ( $n < 7$ )

$$Pr.(x) = \frac{n!}{k!(n-k)!} \cdot p^k \cdot (1-p)^{n-k}$$

$$CDF(x \leq k) = \sum_{x=0}^{x=k} Pr.(x)$$

$$CDF(x \geq k) = \sum_{x=k}^{x=n} Pr.(x)$$

### Metodo della approssimazione alla Normale

Si adotta quando  $n$  è grande ( $n \geq 7$ ), ma solo perché se no i calcoli sarebbero troppo faticosi (il computer continua ad usare il metodo della Binomiale)

- Scopo dell'approssimazione Normale: trasformare  $k$  in uno  $z$  di cui la CDF è conosciuta.

$\mu$  e  $\sigma$  sono PARAMETRI DI UNA VARIABILE TEORICA. NON SONO STATISTICHE CAMPIONARIE.

$$\mu = n \cdot p$$

$$\sigma = \sqrt{\mu \cdot (1-p)}$$

$$z_{\{+\}} = z = \frac{k - \mu + 0.5}{\sigma}$$

$$z_{\{-\}} = \frac{k - \mu - 0.5}{\sigma}$$

$$CDF(x \leq k) = \begin{cases} CDF_{Tavola Normale}(|z_{\{+\}}|) & \text{se } z > 0, \\ 1 - CDF_{Tavola Normale}(|z_{\{+\}}|) & \text{se } z < 0 \end{cases}$$

$$CDF(x \geq k) = \begin{cases} 1 - CDF_{Tavola Normale}(|z_{\{-\}}|) & \text{se } z > 0, \\ CDF_{Tavola Normale}(|z_{\{-\}}|) & \text{se } z < 0 \end{cases}$$

### Probabilità di Poisson puntuale

$$Pr.(X = x) = \frac{\lambda^x \cdot e^{-\lambda}}{x!}$$

### Probabilità cumulata di Poisson

$$CDF(x \leq k) = \sum_1^k Pr.(X = x)$$

Applicare dovuti complementi ed accorgimenti per trovare  $CDF(x > k)$  o la retrocumulata.

## Statistiche Test

### t di Student

Per campioni appaiati

$$D = Y - X$$

$$t(D) = \frac{\bar{d}}{\frac{s_d}{\sqrt{n}}}$$

$$g.l. = n - 1$$

**Per gruppi indipendenti chiamati A e B con varianze simili**

Unione delle Varianze, qualora i dati fossero forniti come sommario e non come tabella estesa:

$$s_{POOL}^2 = \frac{(n_A - 1) \cdot s_A^2 + (n_B - 1) \cdot s_B^2}{n_A + n_B - 2}$$

$$t = \frac{(\bar{x}_B - \bar{x}_A)}{\sqrt{s_{POOL}^2 \cdot \left(\frac{1}{n_A} + \frac{1}{n_B}\right)}}$$

$$g.l. = n_A + n_B - 2$$

### Chi-Quadrato

$$e(x, y) = \frac{\text{totale riga} \cdot \text{totale colonna}}{n}$$

$$\chi^2(X, Y) = \sum \frac{[n(x, y) - e(x, y)]^2}{e(x, y)}$$

$$g.l. = (\text{numero di righe} - 1) \cdot (\text{numero di colonne} - 1)$$