

Copyright© Alberto C. 2002

# Formulario di Statistica Descrittiva I

<b>1) NOZIONI GENERALI SULLE FREQUENZE.</b>	<b>2</b>
1.1) FREQUENZE ASSOLUTE.	2
1.2) FREQUENZE RELATIVE.	2
1.3) FREQUENZE CUMULATE.	2
1.4) FUNZIONE DI RIPARTIZIONE EMPIRICA.	2
<b>2) INDICI DI POSIZIONE.</b>	<b>3</b>
2.1) MEDIA ARITMETICA.	3
2.1.1 <i>Proprietà della media</i>	3
2.2) MEDIANA E QUANTILI.	3
2.2.1 <i>Boxplot (scatola a baffi)</i>	3
2.2.2 <i>Proprietà della mediana.</i>	3
<b>3) INDICI DI VARIABILITÀ.</b>	<b>4</b>
3.1) VARIANZA.	4
3.1.1 <i>Proprietà della varianza.</i>	4
3.2) SCARTO QUADRATICO MEDIO.	4
3.3) CAMPO DI VARIAZIONE.	4
3.4) SCARTO INTERQUARTILE.	4
3.5) M.A.D.	4
3.6) COEFFICIENTE DI VARIAZIONE.	4
<b>4) INDICI DI SIMMETRIA.</b>	<b>5</b>
4.1) INDICE DI SIMMETRIA STANDARDIZZATO.	5
4.2) INDICE DI CURTOSI STANDARDIZZATO.	5
<b>5) INDICI DI MUTABILITÀ.</b>	<b>6</b>
5.1) INDICE DI GINI.	6
5.2) ENTROPIA DI SHANNON.	6
<b>6) REGRESSIONE LINEARE SEMPLICE.</b>	<b>7</b>
6.1) METODO DEI MINIMI QUADRATI.	7
6.2) COVARIANZA.	7
6.2.1 <i>Matrice cov/var.</i>	7
6.2.2 <i>Coefficiente di correlazione.</i>	7
6.3) RESIDUI.	7
6.4) COEFFICIENTE DI DETERMINAZIONE $R^2$ .	8
6.5) RELAZIONI SPURIE.	8
<b>7) DIPENDENZE.</b>	<b>8</b>
7.1) TABELLA DI CONTINGENZA.	8
7.2) INDICE $\chi^2$ DI PEARSON.	8
<b>8) ALTRE NOZIONI.</b>	<b>8</b>
8.1) STANDARDIZZAZIONE DEI DATI.	8
8.2) SCELTA DEGLI INTERVALLI DEGLI ISTOGRAMMI.	8
8.3) ALTRI TIPI DI MEDIE.	9
8.4) RELAZIONI LINERIZZABILI.	9

---

**1) Nozioni generali sulle frequenze.**

---

**1.1) Frequenze assolute.**

Corrispondono esattamente al numero di osservazioni associate ad una data modalità. La loro somma dà il numero totale di osservazioni.

**1.2) Frequenze relative.**

Le frequenze relative si calcolano nel seguente modo:  $f_r = f_i / \sum_{i=1}^n f_i$  dove  $f_i$  sono le frequenze assolute.

La loro somma dà, come è facile intuire, risultato 1.

NB: Spesso le frequenze relative sono espresse in percentuale (basta moltiplicare la frequenza relativa  $i$ -esima per 100).

**1.3) Frequenze cumulate.**

Ogni frequenza cumulata è pari alla somma di tutte le frequenze delle modalità precedenti più quella attuale:

$$f_{ci} = \sum_{j=1}^i f_j.$$

NB: Le frequenze cumulate possono essere calcolate sia con quelle relative che con le assolute.

**1.4) Funzione di ripartizione empirica.**

La funzione di ripartizione empirica è un altro tipo di rappresentazione grafica dei dati. Ogni valore della

funzione viene calcolato in base alla seguente relazione:  $f_{emp}(x) = \frac{\sum f_i \text{ minori di } x}{\sum f_i} = \frac{f_{ci}}{\sum f_i}.$

NB: La funzione di ripartizione empirica può essere calcolata sia con le frequenze relative che con le assolute.

## 2) Indici di posizione.

### 2.1) Media aritmetica.

Dati  $Y = \{y_1, y_2, \dots, y_n\}$  la media di Y risulta  $Media(Y) = \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$ .

#### 2.1.1 Proprietà della media

- La media di una costante è pari alla costante stessa  $\frac{1}{n} \sum_{i=1}^n y_i \wedge y_i = k \Rightarrow \frac{1}{n} \sum_{i=1}^n k = \frac{1}{n} nk = k$ .
- La media è compresa tra il massimo ed il minimo valore di Y:  $y_1 \leq \bar{y} \leq y_n \wedge y_1 \leq y_i \leq y_n$ .
- La media di una trasformazione lineare dei dati è la stessa trasformazione applicata alla media dei dati:  $y = y_i \rightarrow z = a + by_i \rightarrow \bar{z} = a + b\bar{y}$ .
- La somma degli scarti dalla media è sempre nulla:  $\sum_{i=1}^n (y_i - \bar{y}) = 0$ .
- $\sum_{i=1}^n (y_i - a)^2 = \sum_{i=1}^n (y_i - \bar{y})^2 + n(\bar{y} - a)^2$  e  $\sum_{i=1}^n (y_i - a)^2 > \sum_{i=1}^n (y_i - \bar{y})^2 \wedge a \neq \bar{y}$
- La media è molto suscettibile alla presenza di valori anomali.

NB: Se abbiamo una distribuzione di frequenze per degli intervalli di valori, un'approssimazione della media

è data dalla formula: 
$$\frac{\sum_{i=1}^n (\text{valore centrale dell'intervallo}) * f_i}{\sum_{i=1}^n f_i} = \frac{\sum_{i=1}^n (\text{valore centrale dell'intervallo}) * f_i}{n}$$

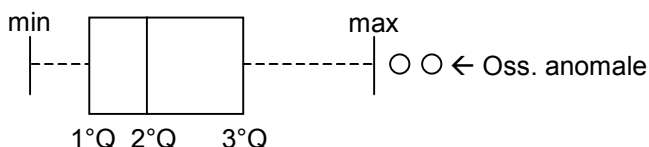
### 2.2) Mediana e quantili.

Un quantile è una particolare osservazione che lascia alla sua sinistra una certa parte della popolazione ed alla sua destra la rispettiva complementare. Esistono quantili particolari:

- 1° Percentile: lascia alla sua sinistra l'1% delle osservazioni (poco usato).
- 1° Quartile: lascia alla sua sinistra l'25% delle osservazioni.
- 2° Quartile o *mediana*: lascia alla sua sinistra l'50% delle osservazioni.
- 3° Quartile: lascia alla sua sinistra l'75% delle osservazioni.

#### 2.2.1 Boxplot (scatola a baffi).

I *BoxPlot* o scatole a baffi sono un modo sintetico di rappresentare i 3 quartili assieme ai massimi e ai minimi della popolazione in osservazione (risulta quindi più semplice studiarne le caratteristiche). Le osservazioni anomale vengono rappresentate come punti esterni ai baffi.



#### 2.2.2 Proprietà della mediana.

- $\sum_{i=1}^n |y_i - \text{mediana}(Y)| \leq \sum_{i=1}^n |y_i - a| \wedge a \neq \text{mediana}(Y)$
- La mediana è poco suscettibile alla presenza di valori anomali.

---

### 3) Indici di variabilità.

---

#### 3.1) Varianza.

È una misura di quanto i dati siano distanti dalla media.

$$\text{var}(Y) = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2 = \frac{1}{2n^2} \sum_{i=1}^n \sum_{j=1}^n (y_i - y_j)^2 = \left( \frac{1}{n} \sum_{i=1}^n y_i^2 \right) - \bar{y}^2$$

##### 3.1.1 Proprietà della varianza.

La varianza di una trasformazione lineare dei dati è pari alla varianza dei dati per il coefficiente angolare al quadrato:  $y = y_i \rightarrow z = a + by_i \rightarrow \text{var}(Z) = b^2 \text{var}(Y)$ .

#### 3.2) Scarto quadratico medio.

Lo scarto quadratico medio o *sqm* è pari alla varianza sotto radice quadrata. Esso ha il vantaggio di mantenere l'unità di misura delle osservazioni rilevate:  $\text{sqm}(Y) = \sqrt{\text{var}(Y)}$ .

#### 3.3) Campo di variazione.

Il campo di variazione è pari alla differenza tra l'osservazione massima e quella minima:  $\text{campo}(Y) = \max(Y) - \min(Y)$ . È estremamente sensibile ai valori anomali.

#### 3.4) Scarto interquartile.

Lo scarto interquartile è pari alla differenza tra il terzo ed il primo quartile:  $\text{siq}(Y) = 3^\circ Q - 1^\circ Q$ . Esso è più resistente della mediana.

#### 3.5) M.A.D.

Il MAD o Median Absolute Deviations è pari alla mediana degli scarti in modulo dalla mediana:

$$\text{MAD}(Y) = \text{mediana}(|y_1 - \text{medana}(Y)|; |y_2 - \text{medana}(Y)|; \dots; |y_n - \text{medana}(Y)|)$$

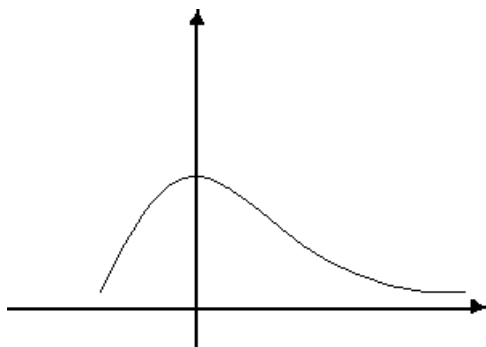
#### 3.6) Coefficiente di variazione.

Il coefficiente di variazione è pari allo *sqm*(Y) fratto la media di Y:  $\text{cvar}(Y) = \frac{\text{sqm}(Y)}{\bar{y}}$ .

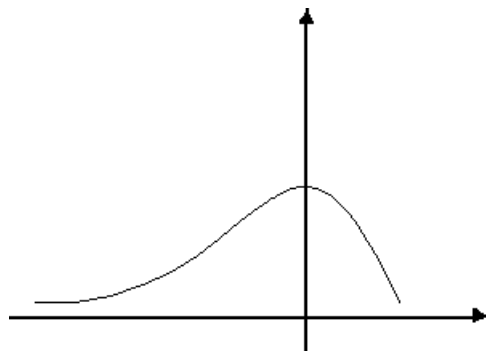
---

#### 4) Indici di simmetria.

---

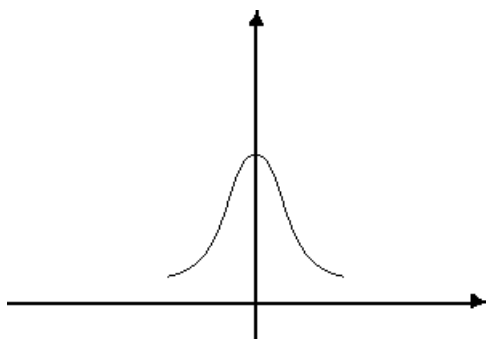


Assimmetria positiva

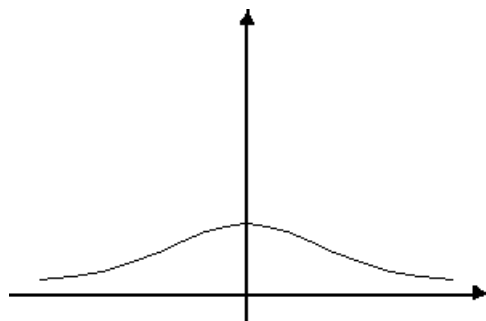


Assimmetria negativa

I seguenti indici misurano la pesantezza della code:



Code meno pesanti



Code più pesanti

##### 4.1) Indice di simmetria standardizzato.

$$i\_simm\_st(Y) = \frac{1}{n * sqm^3(Y)} \sum_{i=1}^n (y_i - \bar{y})^3$$

##### 4.2) Indice di curtosi standardizzato.

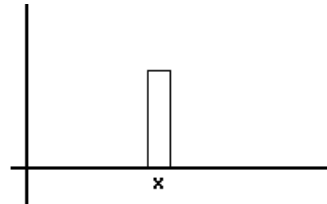
$$i\_simm\_st(Y) = \frac{1}{n * sqm^4(Y)} \sum_{i=1}^n (y_i - \bar{y})^4$$

---

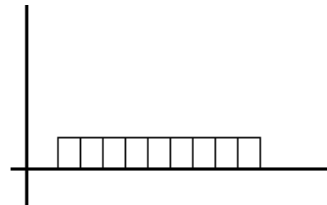
## 5) Indici di mutabilità.

---

Minima mutabilità: si ha quando tutte le osservazioni si concentrano nella stessa modalità



Massima mutabilità: si ha quando tutte le osservazioni si ripartiscono in eguale modo tra le modalità



### 5.1) Indice di Gini.

L'indice normale  $G = \sum_{i=1}^n f_{ri} (1 - f_{ri})$  varia tra  $0 \leq G \leq 1 - \frac{1}{n}$ .

L'indice normalizzato  $G_{norm} = \frac{G}{1 - \frac{1}{n}} = \frac{nG}{n-1}$  varia tra  $0 \leq G \leq 1$ .

### 5.2) Entropia di Shannon.

L'indice normale  $H = - \sum_{i=1}^n f_{ri} \log(f_{ri})$  varia tra  $0 \leq H \leq \log(n)$ , se  $f_{ri} = 0 \rightarrow f_{ri} \log(f_{ri}) = 0$ .

L'indice normalizzato  $H_{norm} = \frac{H}{\log(n)}$  varia tra  $0 \leq H \leq 1$ .

NB: Entrambi gli indici si azzerano in situazioni di minima mutabilità.

## 6) Regressione lineare semplice.

La regressione lineare semplice tenta di spiegare il comportamento delle unità statistiche tramite un modello lineare semplice del tipo:  $y = \alpha + \beta x + \varepsilon$ , dove  $\varepsilon$  è detto errore.

L'obiettivo è quello di minimizzare al massimo l'errore in modo tale che il modello spieghi al meglio il comportamento delle unità statistiche.

### 6.1) Metodo dei minimi quadrati.

Il metodo dei minimi quadrati si pone l'obiettivo di minimizzare gli scarti al quadrato tra i valori osservati ed i valori calcolati dal modello. In pratica bisogna minimizzare  $\sum_{i=1}^n (y_i - \alpha - \beta x_i)^2$ , cioè bisogna trovare degli

$\hat{\alpha}, \hat{\beta}$  tali che se sostituiti ad  $\alpha, \beta$  rendano minima la sommatoria precedente.

Secondo questo metodo i due valori vanno scelti così:  $\hat{\alpha} = \bar{y} - \hat{\beta} \bar{x}$  e  $\hat{\beta} = \frac{\text{cov}(X, Y)}{\text{var}(X)}$ .

### 6.2) Covarianza.

La covarianza misura la dimensione e la forza della relazione tra due variabili.

$$\text{cov}(X, Y) = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x}) = \frac{1}{n} \sum_{i=1}^n (y_i * x_i) - \bar{x} \bar{y}$$

#### 6.2.1 Matrice cov/var.

	X1	X2	X3	
X1	$n_{1,1}$	$n_{1,2}$	$n_{1,3}$	$\rightarrow \text{cov}(X_{ij}, X_{ij})$
X2	$n_{2,1}$	$n_{2,2}$	$n_{2,3}$	
X3	$n_{3,1}$	$n_{3,2}$	$n_{3,3}$	$\rightarrow \text{cov}(X, X) = \text{var}(X); \text{simmetria}$

#### 6.2.2 Coefficiente di correlazione.

La covarianza è sempre compresa tra  $-\text{sqm}(X)\text{sqm}(Y) \leq \text{cov}(X, Y) \leq \text{sqm}(X)\text{sqm}(Y)$ , perciò è possibile introdurre un nuovo indice che corrisponde alla covarianza normalizzata, ossia il coefficiente di correlazione:  $\text{cor}(X, Y) = \frac{\text{cov}(X, Y)}{\text{sqm}(X)\text{sqm}(Y)}$ .

Il nuovo indice varia tra -1 ed 1, se  $|\text{cor}(X, Y)| \rightarrow 1$  il modello spiega tutto altrimenti se  $|\text{cor}(X, Y)| \rightarrow 0$  il modello non spiega nulla.

Analogamente alla covarianza esiste anche per questo indice una matrice detta di correlazione:

	X1	X2	X3	
X1	1	$n_{1,2}$	$n_{1,3}$	$\rightarrow \text{cor}(X_{ij}, X_{ij})$
X2	$n_{2,1}$	1	$n_{2,3}$	
X3	$n_{3,1}$	$n_{3,2}$	1	$\rightarrow \text{simmetria}$

### 6.3) Residui.

La bontà del modello può essere valutata con la varianza dei residui  $r_i = y_i - \hat{\alpha} - \hat{\beta}x_i$ :

$$\text{var}(r_i) = \frac{1}{n} \sum_{i=1}^n (r_i - \bar{r})^2 = \frac{1}{n} \sum_{i=1}^n r_i^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{\alpha} - \hat{\beta}x_i)^2 = \text{var}(Y) - \frac{\text{cov}^2(X, Y)}{\text{var}(X)} \quad \wedge \quad \bar{r} = \frac{1}{n} \sum_{i=1}^n r_i = 0$$



#### 6.4) Coefficiente di determinazione $R^2$ .

$R^2 = 1 - \frac{\text{var}(r_i)}{\text{var}(Y)} = \text{cor}^2(X, Y) \wedge 0 \leq R^2 \leq 1$ , se il coefficiente di determinazione è pari ad 1 il modello spiega viceversa se è pari a 0 non spiega niente.

#### 6.5) Relazioni spurie.

Una relazione spuria del tipo  $X \leftrightarrow Y \leftrightarrow Z$  va trattata nel seguente modo:

1. Trovo i modelli lineari di X,Y ed Z,Y
2. Trovo i residui di entrambi i modelli
3. Calcolo il coefficiente di correlazione tra i residui:  $\text{cor}(r_{xy}, r_{zy})$

---

### 7) Dipendenze.

---

#### 7.1) Tabella di contingenza.

Una tabella di contingenza con Y dipendente da X è solitamente in questa forma:

	X1	X2	X3	Tot
Y1	$f_{1,1}$	$f_{1,2}$	$f_{1,3}$	$f_{1+}$
Y2	$f_{2,1}$	$f_{2,2}$	$f_{2,3}$	$f_{2+}$
Y3	$f_{3,1}$	$f_{3,2}$	$f_{3,3}$	$f_{3+}$
Tot	$f_{+1}$	$f_{+2}$	$f_{+3}$	$f_{++}$

→ Frequenze marginali

Y è dipendente in distribuzione o stocasticamente da X se  $\forall i$  vale:  $\frac{f_{i1}}{f_{+1}} = \frac{f_{i2}}{f_{+2}} = \dots = \frac{f_{ic}}{f_{+c}} = \frac{f_{rel\_i\_c}}{\sum_{j=1}^n f_{rel\_i\_j}}$ ,

da ciò ne consegue che anche le frequenze marginali risultano uguali.

#### 7.2) Indice $X^2$ di Pearson.

Le frequenze attese in caso di indipendenza stocastica sono calcolabili tramite la formula:

$$\hat{f}_{ij} = \frac{f_{i+} * f_{+j}}{f_{++}} = \frac{f_{i+} * f_{+j}}{n}.$$

L'indice di Pearson è una misura della dipendenza in distribuzione:  $X^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(f_{ij} - \hat{f}_{ij})^2}{\hat{f}_{ij}}$ .

NB: Se esiste dipendenza stocastica (o in distribuzione) tra Y e X e le loro medie sono uguali, allora Y è indipendente in media. Le stesse considerazioni valgono per mediana, varianza, ecc...

---

### 8) Altre nozioni.

---

#### 8.1) Standardizzazione dei dati.

La standardizzazione dei dati permette di passare da una distribuzione con una data varianza e media ad una distribuzione equivalente che ha media nulla e varianza unitaria:  $Y(\text{media}(Y); \text{var}(Y)) \Rightarrow Z(0; 1)$ .

In altre parole viene applicata ad Y la seguente trasformazione lineare:  $z_i = \frac{y_i - \text{media}(Y)}{\text{sqm}(Y)}$ .

#### 8.2) Scelta degli intervalli degli istogrammi.

1. Sturges: numero di intervalli =  $1 + \log_2(\text{numero di dati})$
2. Freedman & Diaconis: lunghezza intervalli =  $2(\text{scarto interquartile})(\text{numero di dati})^{-1/3}$

### 8.3) Altri tipi di medie.

$$\text{Media geometrica} = \left( \prod_{i=1}^n x_i^{f_i} \right)^{\left( \sum_{j=1}^n f_j \right)^{-1}} = \left( \prod_{i=1}^n x_i \right)^{\frac{1}{n}}$$

NB: Il pi-greco maiuscolo indica il prodotto.

$$\text{Media armonica} = \frac{\sum_{i=1}^n f_i}{\sum_{i=1}^n \frac{f_i}{x_i}}$$

### 8.4) Relazioni linerizzabili.

Una relazione non lineare del tipo  $y = \alpha x_i^\beta$  corrisponde alla relazione lineare dei logaritmi

$$\log(y) = \log(\alpha) + \beta \log(x_i) \Leftrightarrow z = \hat{\alpha} + \hat{\beta} \quad \wedge \quad z = \log(y) \wedge \hat{\alpha} = \log(\alpha) \wedge \hat{\beta} = \beta.$$