Formulario di Statistica Descrittiva I

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 o *osservazioni*.

1.2) Frequenze relative.

Le frequenze relative si calcolano nel seguente modo: $f_r = f_i \Big/ \sum_{i=1}^{n} f_i$ 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: $f_{emp}(x) = \dfrac{\sum f_i \ \text{ minori di x}}{\sum f_i} = \dfrac{f_{ci}}{\sum f_i}$.

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 $Y = \{y_1, y_2, ..., y_n\}$ le 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 $\frac{1}{n}\sum_{i=1}^{n} y_i \wedge y_i = k \quad \Rightarrow \quad \frac{1}{n}\sum_{i=1}^{n} k = \frac{1}{n} nk = k$.
- La media è compresa tra il massimo ed il minimo valore di Y: $y_1 \le \bar{y} \le y_n \wedge y_1 \le y_i \le y_n$.
- La media di una trasformazione lineare dei dati è la stessa trasformazione applicata alla media dei dati: $y = y_i \quad \rightarrow \quad z = a + by_i \quad \rightarrow \quad \bar{z} = a + b\bar{y}$.
- La somma degli scarti dalla media è sempre nulla: $\sum_{i=1}^{n}\left(y_i - \bar{y}\right) = 0$.
- $\sum_{i=1}^{n}(y_i - a)^2 = \sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 + n\left(\bar{y} - a\right)^2$ e $\sum_{i=1}^{n}(y_i - a)^2 > \sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 \wedge a \neq \bar{y}$
- 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: $\frac{\sum_{i=1}^{n}(valore\ cetrale\ dell'intervallo) * f_i}{\sum_{i=1}^{n} f_i} = \frac{\sum_{i=1}^{n}(valore\ cetrale\ dell'intervallo) * f_i}{n}$.

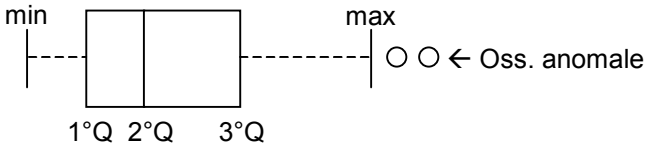
2.2) Mediana e quantili.

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.

- $\sum_{i=1}^{n}\left|y_i - \text{mediana}(Y)\right| \le \sum_{i=1}^{n}\left|y_i - a\right| \wedge a \neq \text{mediana}(Y)$
- 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.

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

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: $y = y_i \quad \rightarrow \quad z = a + by_i \quad \rightarrow \quad \operatorname{var}(Z) = b^2 \operatorname{var}(Y)$.

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: $\operatorname{sqm}(Y) = \sqrt{\operatorname{var}(Y)}$.

3.3) Campo di variazione.

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

3.4) Scarto interquartile.

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

3.5) M.A.D.

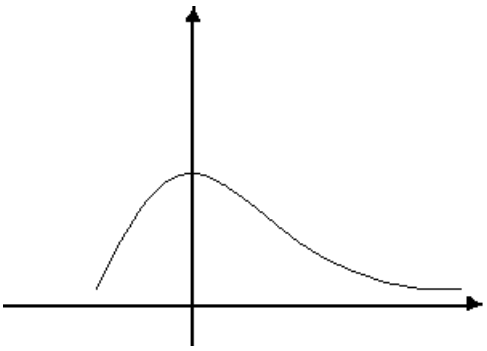
Il MAD o Median Absolute Deviations è pari alla mediana degli scarti in modulo dalla mediana:

$$\operatorname{MAD}(Y) = \operatorname{mediana}\left(\left|y_1 - \operatorname{medana}(Y)\right|; \left|y_2 - \operatorname{medana}(Y)\right|; ...; \left|y_n - \operatorname{medana}(Y)\right|\right)$$

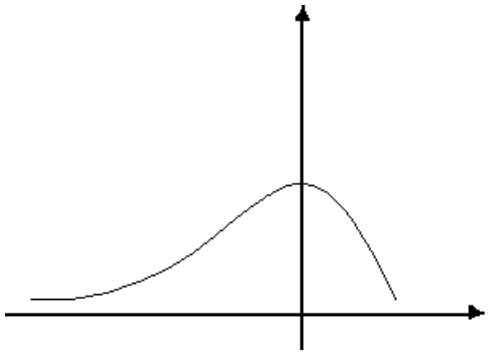
3.6) Coefficiente di variazione.

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

4) Indici di simmetria.

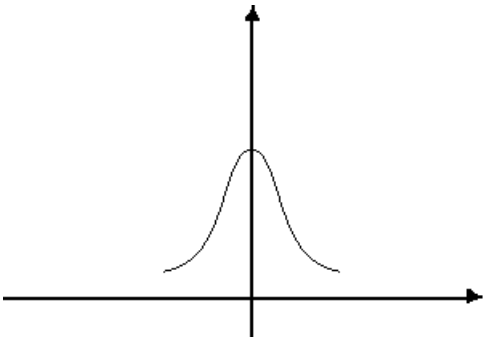


Assimmetria positiva

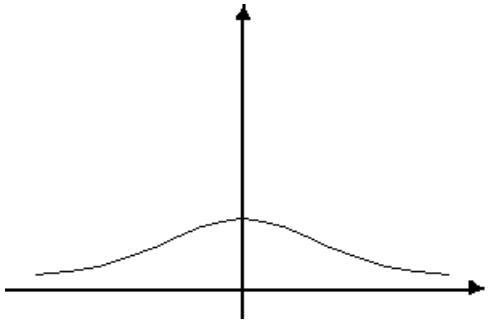


Assimmetria negativa

I seguenti indici misurano la pesantezza della code:



Code meno pesanti



Code più pesanti

4.1) Indice di simmetria standardizzato.

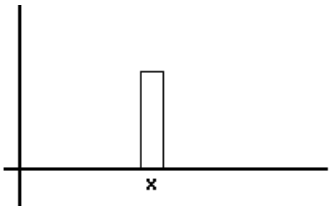
$$\text{i\_simm\_st(Y)} = \frac{1}{n * \text{sqm}^3(Y)} \sum_{i=1}^{n} \left(y_i - \bar{y}\right)^3$$

4.2) Indice di curtosi standardizzato.

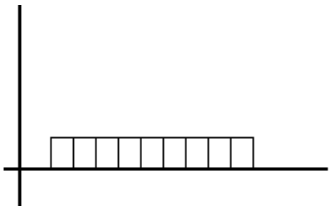
$$\text{i\_simm\_st(Y)} = \frac{1}{n * \text{sqm}^4(Y)} \sum_{i=1}^{n} \left(y_i - \bar{y}\right)^4$$

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 $G = \sum_{i=1}^{n} f_{ri}\left(1 - f_{ri}\right)$ varia tra $0 \le G \le 1 - \frac{1}{n}$.

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

5.2) Entropia di Shannon.

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

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

NB: Entrambi gli indice 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 ε è detto errore.
L'obbiettivo 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 $\sum_{i=1}^{n}(y_i - \alpha - \beta x)^2$, cioè bisogna trovare degli $\hat{\alpha}, \hat{\beta}$ tali che se sostituiti ad α, β rendano minima la sommatoria precedente.

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

6.2) Covarianza.

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

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

6.2.1 Matrice cov/var.

	X1	X2	X3
X1	$n_{1,1}$	$n_{1,2}$	$n_{1,3}$ → cov(Xij, Xij)
X2	$n_{2,1}$	$n_{2,2}$	$n_{2,3}$
X3	$n_{3,1}$	$n_{3,2}$	$n_{3,3}$ → cov(X, X)=var(X); simmetria

6.2.2 Coefficiente di correlazione.

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

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

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

	X1	X2	X3
X1	1	$n_{1,2}$	$n_{1,3}$ → cor(Xij, Xij)
X2	$n_{2,1}$	1	$n_{2,3}$
X3	$n_{3,1}$	$n_{3,2}$	1 → simmetria

6.3) Residui.

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

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

6.4) Coefficiente di determinazione R^2.

$R^2 = 1 - \frac{\text{var}(r_i)}{\text{var}(Y)} = \text{cor}^2(X,Y) \quad \wedge \quad 0 \le R^2 \le 1$, 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_{+j} →
Y2	$f_{2,1}$	$f_{2,2}$	$f_{2,3}$	f_{+j}
Y3	$f_{3,1}$	$f_{3,2}$	$f_{3,3}$	f_{+j}
Tot	f_{i+}	f_{i+}	f_{i+}	f_{++}

Frequenze marginali

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

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: $X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{\left(f_{ij} - \hat{f}_{ij}\right)^2}{\hat{f}_{ij}}$.

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: $Y\left(\text{media}(Y); \text{var}(Y)\right) \Rightarrow Z(0;1)$.

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

8.3) Altri tipi di medie.

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.

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