GRAFICI:

QUANTITATIVA

- CONTINUA→ istogramma
- DISCRETA → diagramma a bastoncini/aste

QUALITATIVA:

- SCONNESSA/NOMINALE→ grafico a torta(se invece c'è un numero enorme di osservazioni si utilizza un grafico a barre)
- ORDINATA→ grafico a barre

MEDIA ARITMETICA	$\mu = \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$
la somma dei quadrati degli scarti dei valori dalla loro media è minima:	$\sum_{i=1}^{N}(x_i - \mu)^2 = min$
RANGE **(valore più grande-valore più piccolo)**	$x_{max} - x_{min}$
MEDIA CON FREQUENZE ASSOLUTE	$\mu = \bar{x} = \frac{1}{N}\sum_{i=1}^{k} x_i \cdot n_i$
MEDIA CON FREQUENZE RELATIVE	$\mu = \bar{x} = \sum_{i=1}^{k} x_i f_i$

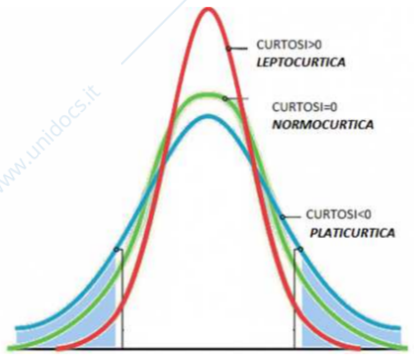
MEDIA CON SOTTOGRUPPI K	$\mu = \bar{x} = \frac{1}{N}\sum_{i=1}^{k}\bar{x}_i n_i$
VARIANZA lascio 3 decimali dopo la virgola	$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$ $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}{x_i}^2 - \mu^2$
VARIANZA SE i dati provengono da un campione	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
Varianza con frequenze assolute	$\sigma^2 = \frac{1}{N}\sum_{i=1}^{k}(x_i - \mu)^2 \cdot n_i$ $\sigma^2 = \frac{1}{N}\sum_{i=1}^{k}{x_i}^2 \cdot n_i - \mu^2$
varianza con frequenze relative	$\sigma^2 = \sum_{i=1}^{k}(x_i - \mu)^2 \cdot f_i$ $\sigma^2 = \sum_{i=1}^{k}{x_i}^2 \cdot f_i - \mu^2$

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

scarto quadratico medio o deviazione standard lascio 3 decimali dopo la virgola	$\sigma = \sqrt{\sigma^2} =$ $\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2}$
proprietà della varianza	$Y = a + bX$ $\bar{y} = a + b\bar{x}$ $\sigma_Y^2 = b^2\sigma_X^2$
VARIANZA CON SOTTOGRUPPI K **VARIANZA WITHIN** **VARIANZA BETWEEN**	$\sigma^2 = \sigma_W^2 + \sigma_B^2$ $\sigma_w^2 = \frac{1}{N}\sum_{i=1}^{k}\sigma_i^2 n_i$ $\sigma_B^2 = \frac{1}{N}\sum_{i=1}^{k}\bar{x}_i{}^2 \cdot n_i - \mu^2$

COEFFICIENTE DI VARIAZIONE % (misura la dispersione nell'insieme di dati relativamente alla media) può variare da meno infinito a più infinito in quanto non è un indice normalizzato	$CV = \sigma^* = \frac{\sigma}{\lvert\bar{x}\rvert}$
Indice di Eterogeneità di Gini (solo con freq. relative) misura la variabilità di una variabile qualitativa e che ci dà dunque l'idea di quanto il campione statistico sia omogeneo o eterogeneo in base a una determinata condizione. nel risultato lasciare 3 decimali dopo la virgola	$G = 1 - \sum_{i=1}^{k} f_i^2$
Il valore massimo che assume l'indice è	$G_{max} = \frac{k-1}{k}$
Per ottenere un indice normalizzato, (ovvero che varia tra 0 ed 1) nel risultato lasciare 4 decimali dopo la virgola(poi trasformalo in %)	$G_{norm} = \frac{G}{G_{max}} = G\,\frac{k}{k-1}$
DIFFERENZA INTERQUARTILE	$DI = Q_3 - Q_1$
presenza di OUTLIERS	se il MASSIMO è maggiore di $Q3 + 1.5\,(DI)$ se il MINIMO è

	minore di $Q1 - 1.5\ (DI)$
presenza di valori "ESTREMI"	se il MASSIMO è maggiore di $Q3 + 3 \cdot (DI)$ se il MINIMO è minore di $Q1 - 3 \cdot (DI)$
STANDARDIZZAZIONE	$z_{score} = z_i = \frac{osservazione\ -\ media}{deviazione\ standard} = \frac{x_i - \mu}{\sigma}$
L'indice di asimmetria $\begin{cases} > 0 & \text{se } X \text{ presenta asimmetria positiva} \\ = 0 & \text{se } X \text{ è simmetrica} \\ < 0 & \text{se } X \text{ presenta asimmetria negativa} \end{cases}$	$\alpha_2 = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^3}{\sigma^3}$
L'indice di curtosi $\begin{cases} > 3 & X \text{ leptocurtica o ipernormale} \\ = 3 & X \text{ è normale} \\ < 3 & X \text{ platicurtica o iponormale} \end{cases}$  	$\beta_2 = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^4}{\sigma^4}$
DISUGUAGLIANZA DI CHEBYSHEV	$\mu - k\sigma$ e $\mu + k\sigma$ $1 - \frac{1}{k^2}$

INDIPENDENZA O CONNESSIONE

STATISTICAMENTE INDIPENDENTI→SONO UGUALI

	Uomo	Donna	
Non Fuma	40%	60%	100%
Fumatore occasionale	40%	60%	100%
Fumatore abituale	40%	60%	100%
	40%	60%	100%

	Uomo	Donna	
Non Fuma	24%	24%	24%
Fumatore occasionale	47%	47%	47%
Fumatore abituale	29%	29%	29%
	100%	100%	100%

CONNESSE→SONO DIVERSE

	Uomo	Donna	
Non Fuma	17%	22%	20%
Fumatore occasionale	57%	26%	40%
Fumatore abituale	26%	52%	40%
	100%	100%	100%

LE DISTRIBUZIONI CONDIZIONATE

$$\frac{n_{11}}{n_{1.}} \cdot 100$$

$$\frac{n_{11}}{n_{.1}} \cdot 100$$

se c'è indipendenza:

$$n_{ij} = n_{ij}^*$$

se c'è connessione,

$$n_{11}^* = \hat{n}_{11} = \frac{n_{1.} \cdot n_{.1}}{N}$$

saranno diverse

CONTINGENZE

(se sono tutte vicine a 0, la connessione è bassa)

A segni positivi delle contingenze corrisponde **ATTRAZIONE** tra le modalità corrispondenti mentre a segni negativi corrisponde **REPULSIONE** tra le modalità corrispondenti

$$c_{ij} = n_{ij} - n_{ij}^*$$

Contingenze	Uomo	Donna	
Non Fuma	-1.2	1.2	0
Fumatore occasionale	7.6	-7.6	0
Fumatore abituale	-6.4	6.4	0
	0	0	0

tra la modalità “2” dell’abitudine al fumo e la modalità “1” del genere c’è **attrazione** (=7.6 è positivo)
tra la modalità “1” dell’abitudine al fumo e la modalità “1” del genere c’è **repulsione**(=-1.2 è negativo)

INDICE DI CONNESSIONE - (Chi quadro di Pearson)

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij} - n_{ij}^*)^2}{n_{ij}^*} = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{{c_{ij}}^2}{n_{ij}^*}$$

formula alternativa:

$$\chi^2 = N\left(\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{{n_{ij}}^2}{n_{i.}n_{.j}} - 1\right)$$

Il valore assoluto dell’indice di Pearson non è interpretabile, il massimo valore che può assumere dipende da *N, r, c* è necessario ricorrere alla sua normalizzazione

$$\chi^2_{norm} = \frac{\chi^2}{N[min(r-1, c-1)]}$$

l’indice normalizzato varia da:
0 = assenza di connessione(ovvero indipendenza statistica)
ad 1 = massima connessione

V di Cramer

(indice per misurare il grado di associazione tra due variabili nominali)
si usa per capire quanta connessione c’è

$$V = \sqrt{\chi^2_{norm}}$$

esempio di indipendenza

X: altezza degli atleti	Y: n° di gare vinte nei primi 2 mesi: 0	1	
160 -| 175	5	5	10
175 -| 185	10	10	20
185 -| 195	5	5	10
	20	20	40

$$\bar{y}|x_1 = \frac{1}{n_{1.}}\sum_{j=1}^{c=2} y_j n_{1j} = \frac{1}{10}(0\cdot 5 + 1\cdot 5) = \frac{5}{10} = 0.5$$

$$\bar{y}|x_2 = \frac{1}{n_{2.}}\sum_{j=1}^{c=2} y_j n_{2j} = \frac{1}{20}(0\cdot 10 + 1\cdot 10) = \frac{10}{20} = 0.5$$

$$\bar{y}|x_3 = \frac{1}{n_{3.}}\sum_{j=1}^{c=2} y_j n_{3j} = \frac{1}{10}(0\cdot 5 + 1\cdot 5) = \frac{5}{10} = 0.5$$

Al cambiare delle altezze non cambia la media del n° di gare vinte

Il numero di gare vinte è indipendente in media dalle altezze degli atleti

Non ha senso un'analisi di indipendenza in media

Si dice che *Y* dipende in media da *X* se la relazione di connessione tra le due variabili si riflette sulle *medie condizionate* di *Y* che risultano diverse tra loro al variare di *X*

$$\bar{y}\,|\,x_1 = \ldots = \bar{y}\,|\,x_i = \ldots = \bar{y}\,|\,x_r = \bar{y}$$

c'è indipendenza in media se tutte le medie condizionate sono tra loro uguali e quindi uguali alla media marginale

formule:

$$\bar{y}\,|\,x_1 = \frac{1}{n_{1.}}\sum_{j=1}^{c} y_j n_{1j}$$

$$\bar{x}\,|\,y_1 = \frac{1}{n_{.1}}\sum_{i=1}^{r} x_i n_{i1}$$

La dipendenza in media

indice di dipendenza eta quadro

$$\sigma_Y^2 = \sigma_B^2 + \sigma_W^2$$

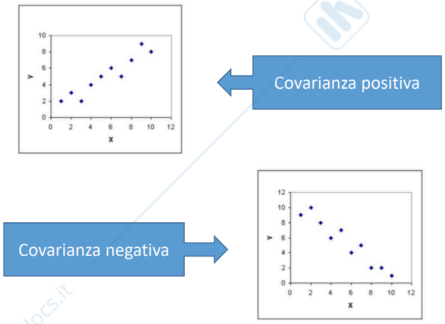
$$\sigma_Y^2 = \frac{1}{N}\sum_{j=1}^{c}(y_j-\bar{y})^2 n_{.j} = \frac{1}{N}\sum_{i=1}^{r}(\bar{y}|x_i-\bar{y})^2 n_{i.} + \frac{1}{N}\sum_{i=1}^{r}\sigma_Y^2|x_i\cdot n_{i.}$$

eta quadro:

$$\eta_{Y|X}^2 = \frac{\sigma_B^2}{\sigma_Y^2} = \frac{\frac{1}{N}\sum_{i=1}^{r}(\bar{y}|x_i-\bar{y})^2 n_{i.}}{\frac{1}{N}\sum_{j=1}^{c}(y_j-\bar{y})^2 n_{.j}}$$

Assume valori compresi tra 0 ed 1

COVARIANZA
la media dei prodotti degli scarti della X e Y dalle loro rispettive medie



$$Cov(X,Y) = \sigma_{XY} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}) =$$

$$= \frac{1}{N}\sum_{i=1}^{N} x_i y_i - \bar{x}\bar{y}$$

La covarianza la puoi calcolare **SOLO SE** le due variabili hanno una relazione lineare:
- **positiva** quando entrambe crescono o decrescono;
- **negativa** quando invece una cresce e l'altra decresce o viceversa infatti la covarianza varia tra -1 e 1

CORRELAZIONE rho (indice di correlazione lineare di Bravais e Pearson)

$$\rho = \frac{Cov(X,Y)}{DS(X)DS(Y)} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{\sigma_{XY}}{\sqrt{\sigma_X^2 \sigma_Y^2}}$$

Attenzione alle correlazioni SPURIE se entrambe le variabili dipendono da una terza variabile

lasciare sempre 4 decimali dopo la virgola!!

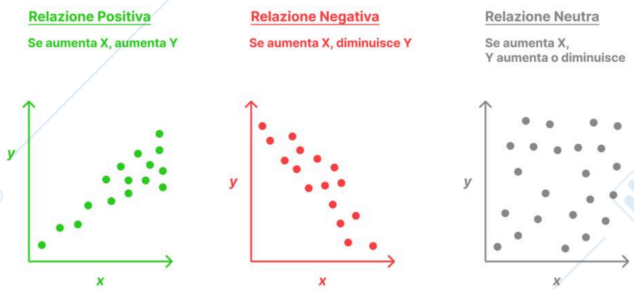
Valore indice di correlazione	Intensità relazione	Direzione relazione
+1	Perfetta correlazione	Positiva
Superiore a +0.7	Forte correlazione	
Tra +0.3 e +0.7	Moderata correlazione	
Fino a +0.3	Debole correlazione	
0	Nessuna correlazione	
Fino a -0.3	Debole correlazione	Negativa
Tra -0.3 e -0.7	Moderata correlazione	
Inferiore a -0.7	Forte correlazione	
-1	Perfetta correlazione	

può assumere valori che vanno da:

–1 (correlazione perfetta negativa)

a **+1** (correlazione perfetta positiva).

legame perfetto=massima connessione=(X^2=1) se i punti sono allineati perfettamente vuol dire che c'è massima connessione



rapporto di composizione(o di parte al tutto)

$$\frac{a_1}{A}\cdot 100 + \frac{a_2}{A}\cdot 100 + \cdots + \frac{a_k}{A}\cdot 100 = 100$$

rapporto di coesistenza

$$\frac{a_1}{a_2} \cdot 100$$

non è mai negativo, ma può essere maggiore di cento

rapporto di derivazione

$$\frac{b_1}{a_1} \cdot 1000 \qquad \frac{b_2}{a_2} \cdot 1000 \qquad \frac{b_k}{a_k} \cdot 1000$$

numeri indici a base fissa

$$_bI_t = \frac{q_t}{q_b} \cdot 100$$

Anno	1997	1998	1999	2000	2001	2002	2003
N° divorzi	33.342	33.510	34.341	35.573	40.051	41.835	43.826
Numeri indice con base fissa b = 1997	(33.342/33.342) 100 100	(33.510 /33.342) 100 100,504	(34.341/33.342) 100 102,996	(35.573/33.342) 100 106,691	(40051/33.342) 100 120,122	(41.835/33.342) 100 125,472	(43.826/33.342) 100 131,444

numeri indici a base mobile

$$_{t-1}I_t = \frac{q_t}{q_{t-1}} \cdot 100$$

Numeri indice con base mobile	----	(33.510 /33.342) 100 100,504	(34.341/33.510) 100 102,480	(35.573/34.341) 100 103,591	(40.051/35.573) 100 112,588	(41.835/40.051) 100 104,454	(43.826/41.835) 100 104,759

Data una serie di numeri indice a base fissa .

è possibile operare un

cambiamento di base,

trasformandola in serie di numeri indice **a base fissa**

$$_cI_t = \frac{_bI_t}{_bI_c} \cdot 100$$

Data una serie di numeri indice a base fissa .
è possibile operare un

cambiamento di base,

trasformandola in serie di numeri indice **a base mobile**

$$_{t-1}I_t = \frac{q_t}{q_{t-1}} = \frac{q_t/q_b}{q_{t-1}/q_b} = \frac{_bI_t}{_bI_{t-1}} \cdot 100$$

Indice dei prezzi: Laspeyres

Se i numeri indici vengono ponderati con i prezzi e le quantità al tempo base b

$$w_j = p_{b,j} \cdot q_{b,j}$$

L'indice di Laspeyres tende ad essere superiore all'indice di Paasche se i prezzi aumentano
e minore se i prezzi diminuiscono

$$_bI_t^L = \frac{\sum_{j=1}^n w_j \frac{p_{t,j}}{p_{b,j}}}{\sum_{j=1}^n w_j} = \frac{\sum_{j=1}^n (p_{b,j} q_{b,j}) \frac{p_{t,j}}{p_{b,j}}}{\sum_{j=1}^n p_{b,j} q_{b,j}} = \frac{\sum_{j=1}^n p_{t,j} q_{b,j}}{\sum_{j=1}^n p_{b,j} q_{b,j}}$$

il risultato lo moltiplichiamo · 100

Indice dei prezzi: Paasche Se come pesi si utilizzano le quantità dei beni del tempo corrente t $w_j = p_{b,j} \cdot q_{t,j}$	${}_bI_t^P = \frac{\sum_{j=1}^n w_j \frac{p_{t,j}}{p_{b,j}}}{\sum_{j=1}^n w_j} = \frac{\sum_{j=1}^n (p_{b,j} q_{t,j}) \frac{p_{t,j}}{p_{b,j}}}{\sum_{j=1}^n p_{b,j} q_{t,j}} = \frac{\sum_{j=1}^n p_{t,j} q_{t,j}}{\sum_{j=1}^n p_{b,j} q_{t,j}}$ il risultato lo moltiplichiamo · 100
Indice dei prezzi: Fisher	${}_bI_t^F = \sqrt{{}_bI_t^L \cdot {}_bI_t^P}$