Introduzione alla Statistica con Excel

Questo libro è stato aggiornato per Excel 2007 e versioni superiori.

Programma

- **Struttura e visualizzazione dei dati**
- Sintesi e interpretazione
- Relazioni tra variabili
- Inferenza statistica
- Verifica delle ipotesi
- Soluzioni dei test

Raccolta dei dati

L’ utilità delle informazioni che si possono estrarre dall'analisi dei dati dipende da come sono organizzati.

Molte aziende dispongono di un sistema informativo unificato, un data base centralizzato, composto da numerose **tabelle di dati raccolti al livello di dettaglio** nelle più importanti aree di business.

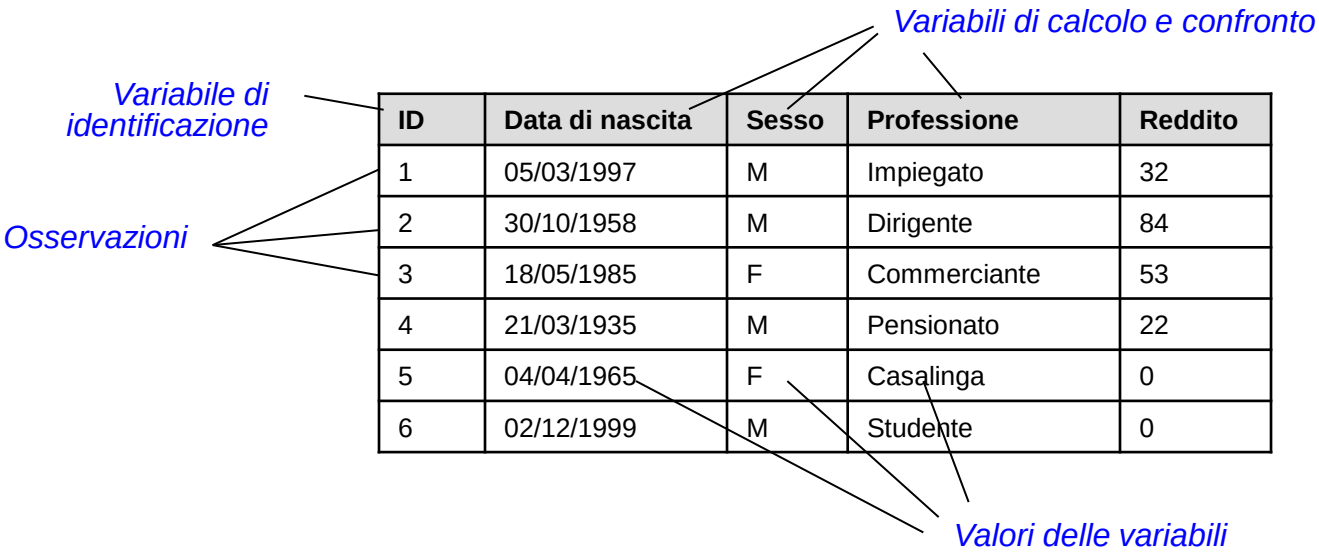
Le informazioni di interesse che si trovano nelle tabelle del data base centralizzato devono essere aggregate e tradotte in una forma che possa essere ***analizzabile in termini statistici***.

Si parla in questo caso di un data base tematico, nel quale ogni tabella è ***strutturata in funzione delle finalità*** che si vogliono perseguire con le analisi (in ambito commerciale, ad esempio, per applicazioni di Marketing).

Le tabelle sono la base essenziale per poter fare analisi statistica dei dati.

La tabella dei dati

Prima di essere analizzati, i dati devono essere raccolti e strutturati in forma di matrice in cui ogni riga rappresenta un'**osservazione** e ogni colonna una **variabile**.



ID	Data di nascita	Sesso	Professione	Reddito
1	05/03/1997	M	Impiegato	32
2	30/10/1958	M	Dirigente	84
3	18/05/1985	F	Commerciante	53
4	21/03/1935	M	Pensionato	22
5	04/04/1965	F	Casalinga	0
6	02/12/1999	M	Studente	0

Le variabili

In statistica si usa il termine ***variabile*** (o carattere) per indicare ogni caratteristica che viene rilevata su ciascuna ***osservazione*** (o unità) di una certa popolazione. Ciascuna variabile si presenta in modo diverso nelle varie osservazioni; per questo motivo, si usa il termine modalità (o ***valore)*** per indicare i modi nei quali una certa variabile si può manifestare.
Si dividono in:

- **Quantitative**

  sono variabili i cui valori sono espressi da **quantità numeriche**[1] che possono essere **discrete** (derivanti di solito da conteggi come, ad esempio, il numero di conti correnti posseduti, il numero dei componenti di una famiglia, ecc.) o **continue** (che derivano di solito da misurazioni come, ad esempio, peso, statura, reddito, ecc.).

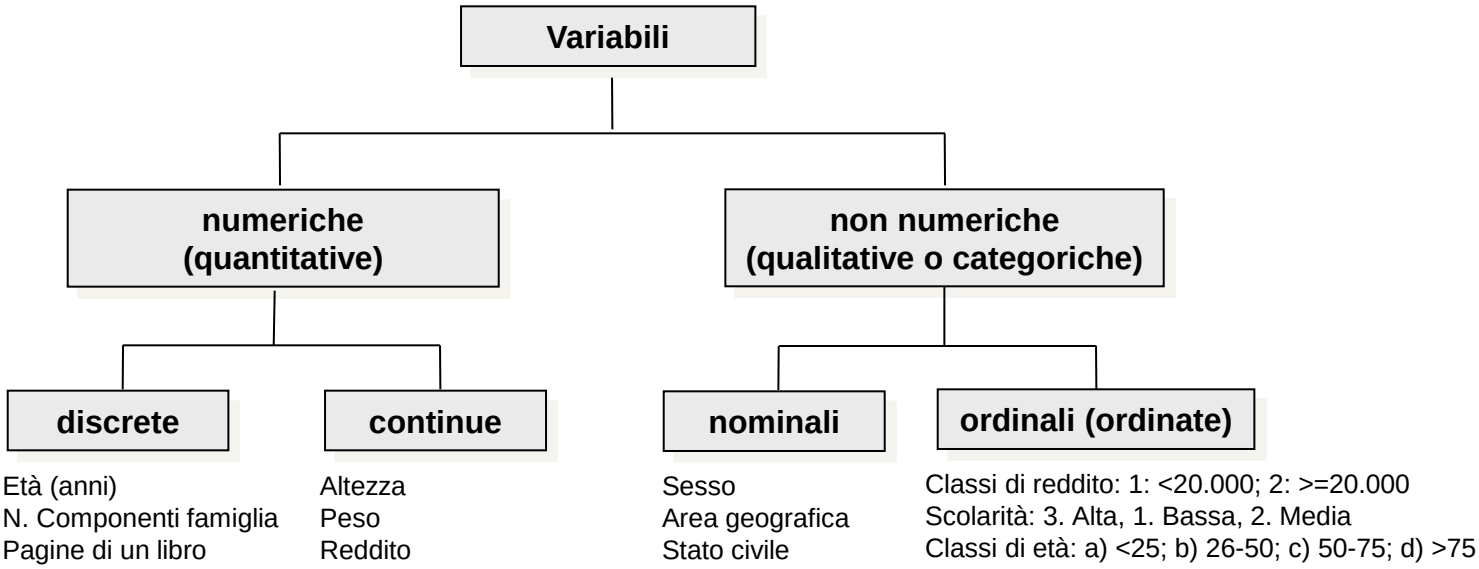
- **Qualitative**[2] **o Categoriche**

  Sono variabili i cui valori sono espressi attraverso **attributi** che possono essere **nominali** (dove ogni possibile ordinamento è arbitrario come, ad esempio, sesso, luogo di residenza, ecc.) oppure **ordinali** (dove è possibile un ordinamento come, ad esempio, titolo di studio, classe di reddito, ecc.).

[1] Tutte le variabili sulle quali si possono eseguire operazioni aritmetiche.
[2] Sono comprese anche le variabili quantitative espresse in classi ordinate di valori.

Le variabili

Classificazione delle diverse tipologie di variabili:



Rappresentazione dei dati: le frequenze assolute e percentuali

La ***frequenza assoluta*** conta il numero complessivo di osservazioni in cui un valore di una variabile si presenta nella totalità delle osservazioni rilevate. La ***frequenza relativa*** (***percentuale***) è data dal rapporto tra la frequenza assoluta e il numero complessivo delle osservazioni. La frequenza ***cumulata*** (assoluta o relativa) è la somma crescente delle frequenze.

Ad esempio, avendo a disposizione un campione di 60 soggetti, si può contare la numerosità per area geografica di residenza ottenendo la seguente tabella di frequenze:

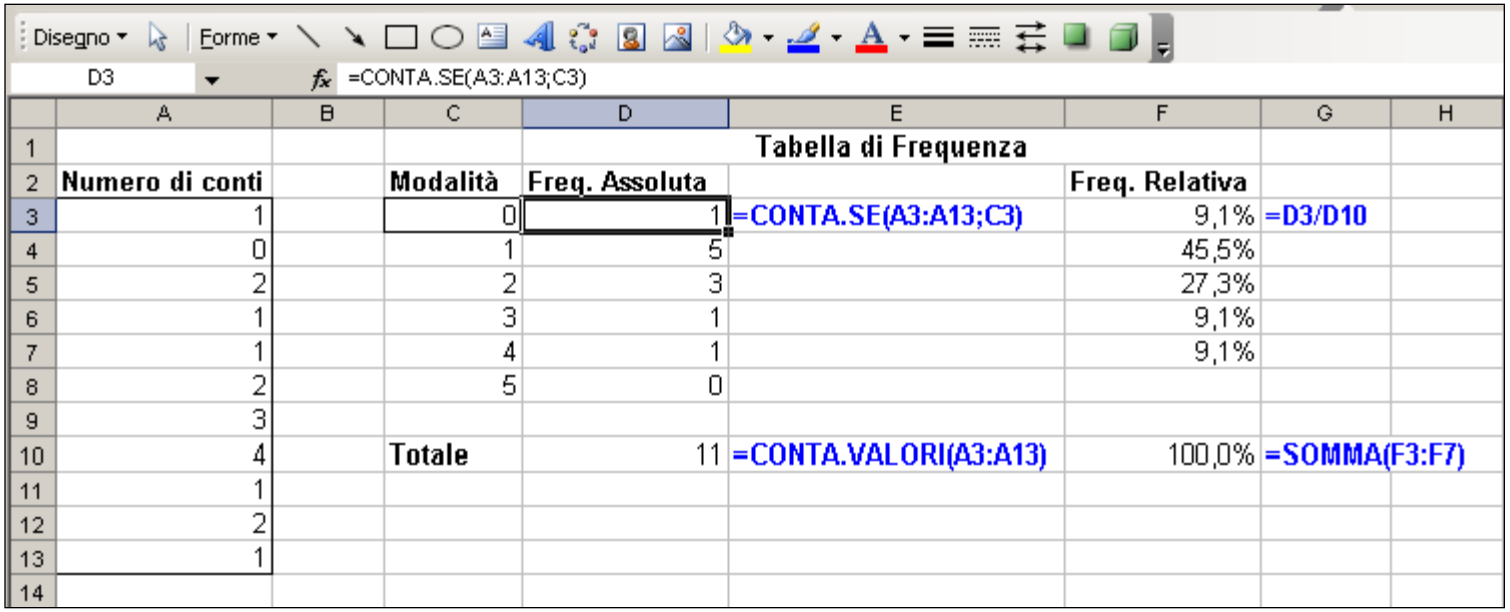
Area
Centro
Nord-Ovest
Nord-Ovest
Nord-Est
Nord-Ovest
Nord-Est
Sud-Isole
Nord-Ovest
Nord-Est

(continua)

Area	Frequenza assoluta	Frequenza relativa	Frequenza cumulata rel
Nord-Est	11	18,3%	18,3%
Nord-Ovest	24	40,0%	58,3%
Centro	17	28,3%	86,7%
Sud-Isole	8	13,3%	100,0%
Totale	60	100,0%	

Tabelle di frequenze – In Excel

In Excel si possono usare le formule statistiche **CONTA.SE()** per calcolare le frequenze assolute e **CONTA.VALORI()** per calcolare il numero complessivo delle osservazioni. Il loro rapporto permette di calcolare le frequenze relative. Ad esempio:



Disegno | Forme | D3 | fx =CONTA.SE(A3:A13;C3)

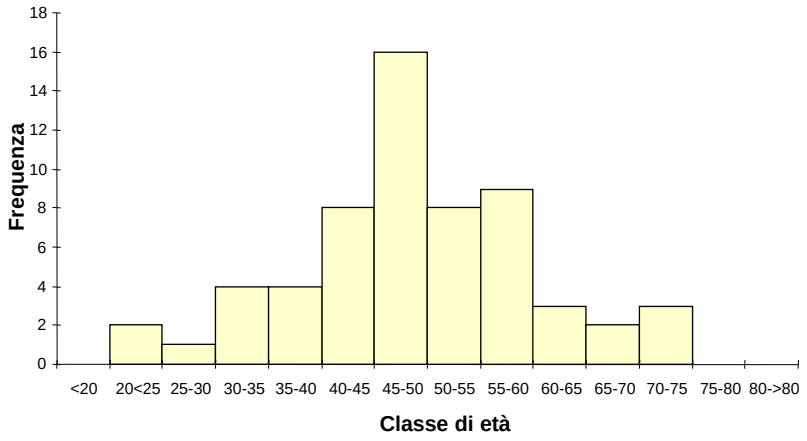
	A	B	C	D	E	F	G	H
1				Tabella di Frequenza				
2	Numero di conti		Modalità	Freq. Assoluta		Freq. Relativa		
3	1		0	1	=CONTA.SE(A3:A13;C3)	9,1%	=D3/D10	
4	0		1	5		45,5%		
5	2		2	3		27,3%		
6	1		3	1		9,1%		
7	1		4	1		9,1%		
8	2		5	0				
9	3							
10	4		Totale	11	=CONTA.VALORI(A3:A13)	100,0%	=SOMMA(F3:F7)	
11	1							
12	2							
13	1							
14								

Variabili qualitative: grafici a istogrammi

La **distribuzione** di frequenza è una prima sintesi di una variabile qualitativa (o di una variabile quantitativa se espressa in classi), ottenuta attraverso la rappresentazione della **frequenza** (assoluta o relativa) con la quale si presentano i suoi valori, che può essere letta in forma tabellare o grafica.

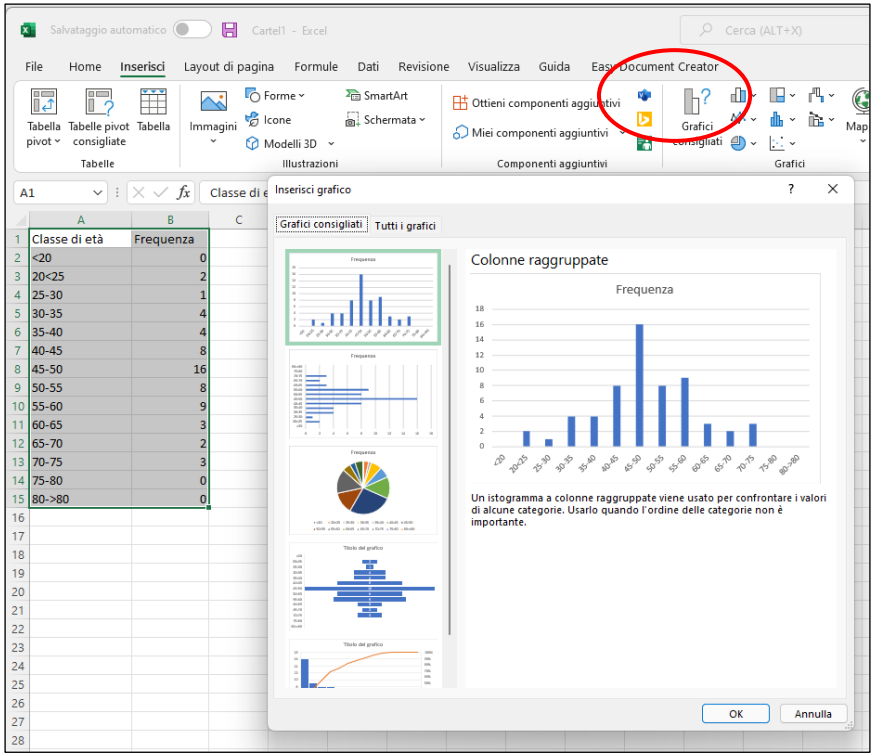
Nella forma grafica la frequenza viene visualizzata per mezzo di **grafici a barre** o **istogrammi**:

Classe di età	*Frequenza*
<20	0
20<25	2
25-30	1
30-35	4
35-40	4
40-45	8
45-50	16
50-55	8
55-60	9
60-65	3
65-70	2
70-75	3
75-80	0
80->80	0



Variabili qualitative: grafici a istogrammi – In Excel

Per creare un grafico in Excel si può utilizzare la sezione ***grafici*** dalla voce di menu **Inserisci** della barra degli strumenti. Bisogna prima selezionare i dati da rappresentare e attivare lo strumento desiderato o sceglierlo da **Grafici consigliati**:



Classificazione delle variabili e distribuzione – Test intermedio

1. Il saldo di un conto corrente è una variabile di tipo:

- ❑ Quantitativa discreta
- ❑ Quantitativa continua

2. La variabile che assume i valori "3. >100.000", "1. 0-49.999", "2. 50.000-100.000" è di tipo:

- ❑ Qualitativa ordinale
- ❑ Qualitativa nominale

3. La frequenza relativa (in percentuale) dei numeri maggiori di 25 nell'insieme 35, 19, 18, 25, 42, 33, 22, 44 è :

- ❑ 62,5%
- ❑ 50%

Programma

- Struttura e visualizzazione dei dati
- **Sintesi e interpretazione**
- Relazioni tra variabili
- Inferenza statistica
- Verifica delle ipotesi
- Soluzioni dei test

Variabili quantitative: indici di posizione

Per rendere più agevole l'interpretazione e i confronti nel tempo e nello spazio di dati quantitativi è opportuno sintetizzarli attraverso semplici elaborazioni matematiche, cioè **riassumere l'informazione che contengono in pochi valori di immediata e semplice lettura**:

- **Indici di posizione**

  sintetizzano in un ***singolo valore*** numerico l'intera distribuzione di frequenza

- **Indici di variabilità**

  misurano la tendenza delle osservazioni ad assumere valori diversi.

  Descrivono quanto l'indice di posizione considerato possa ritenersi ***realmente rappresentativo*** dei valori assunti dalle osservazioni

Variabili quantitative: indici di posizione

Gli indici di posizione sono valori che esprimono una **tendenza centrale**.

- **Media aritmetica**

  E' **la somma dei valori** di un insieme di osservazioni**, diviso** il **loro numero.**

  Ad esempio, avendo i seguenti valori di età di un gruppo 20 clienti:
  [*68,57,65,54,84,48,56,76,73,72,75,76,78,68,69,70,70,71,85,68*]

  la **media** è ***69,15*** (1.383/20).

  La media è l'indice di posizione più utilizzato.

[1] Questo corso tratta due tipi di medie: la media aritmetica e la media geometrica.

La media aritmetica ponderata

La media aritmetica di una variabile X è la somma pesata dei valori delle *N* osservazioni, divisa per *N*.

$$\mu = \frac{\sum_{i=1}^{N} x_i f_i}{\sum_{i=1}^{N} f_i}$$ dove:

μ è la media aritmetica[1]

f_i è il fattore di ponderazione (vale 1 se al valore non vi è associato alcun peso)

$\sum_{i=1}^{N} f_i$ è la somma dei pesi, che equivale al numero delle osservazioni: $\sum_{i=1}^{N} f_i = f_1 + f_2 + \ldots + f_N = N$

x_i è il valore dell'*i*-esima osservazione della variabile X

$\sum_{i=1}^{N} x_i f_i$ è la somma di tutti i valori di x_i per il loro peso: $\sum_{i=1}^{N} x_i f_i = x_1 f_1 + x_2 f_2 + \ldots + x_N f_N$

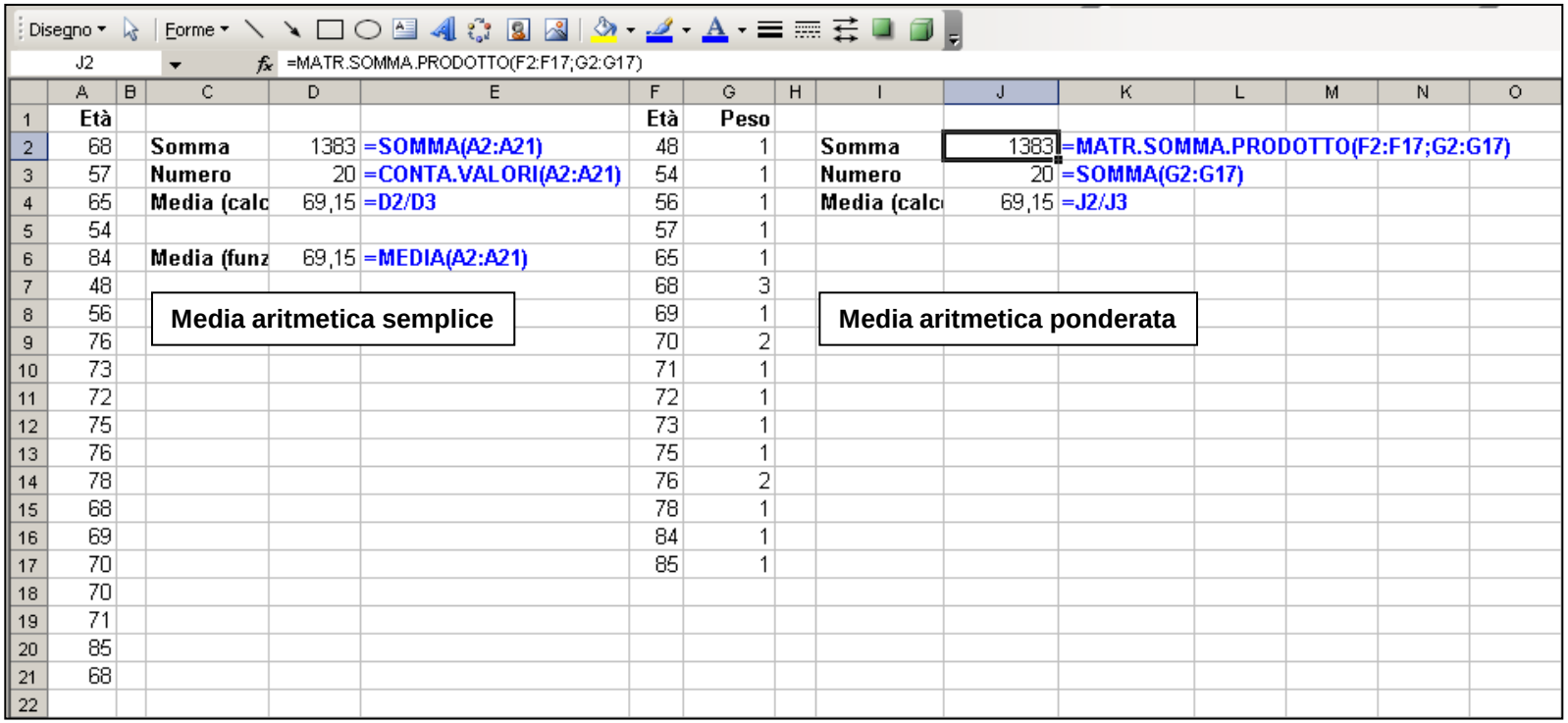
Dall'esempio precedente, avendo i singoli valori di età dei 20 clienti con i relativi pesi:
[48,54,56,57,65,**68,**69,**70,**71,72,73,75,**76,**78,84,85]
[1, 1, 1, 1, 1, **3**, 1, **2,** 1, 1, 1, 1, **2**, 1, 1, 1]

la media è sempre **69,15:**
(48+54+56+57+65+68*3+69+70*2+71+72+73+75+76*2+78+84+85) **/** (1+1+1+1+1+3+1+2+1+1+1+1+2+1+1+1).

[1] In statistica si distingue la media di una popolazione, indicata con il simbolo μ, e la media di un campione, indicata con il simbolo $\bar{x}$.

Calcolo della media aritmetica semplice e ponderata – In Excel

La media aritmetica semplice si calcola in Excel con la funzione **MEDIA().** Per la media aritmetica ponderata non ci sono formule pronte, però si può usare la funzione **MATR.SOMMA.PRODOTTO().** Ad esempio:



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Età					Età	Peso								
2	68		Somma	1383	=SOMMA(A2:A21)	48	1		Somma	1383	=MATR.SOMMA.PRODOTTO(F2:F17;G2:G17)				
3	57		Numero	20	=CONTA.VALORI(A2:A21)	54	1		Numero	20	=SOMMA(G2:G17)				
4	65		Media (calc	69,15	=D2/D3	56	1		Media (calc	69,15	=J2/J3				
5	54					57	1								
6	84		Media (funz	69,15	=MEDIA(A2:A21)	65	1								
7	48					68	3								
8	56					69	1								
9	76					70	2								
10	73					71	1								
11	72					72	1								
12	75					73	1								
13	76					75	1								
14	78					76	2								
15	68					78	1								
16	69					84	1								
17	70					85	1								
18	70														
19	71														
20	85														
21	68														
22															

La media geometrica (opzionale)

La **media geometrica** viene usata quando i valori esprimono **differenze percentuali** tra diversi periodi di tempo (ad esempio il **tasso di crescita** di una popolazione, del reddito, del rendimento di un'azione, ecc.), perché con essa vengono considerati gli **effetti cumulati** nei vari periodi. È definita come la radice *n*-sima del prodotto di *n* numeri:

$$\mathbf{MG} = [(\boldsymbol{x_1} + \mathbf{1})(\boldsymbol{x_2} + \mathbf{1}) \ldots (\boldsymbol{x_n} + \mathbf{1})]^{1/n}$$

Supponiamo che un investimento renda il 10% il primo anno, il 25% il secondo, il -20% il terzo e il 25% il quarto. Il valore di 100€ di investimento alla fine dei 4 anni sarà:
100(1+0,10)*(1+0,25)*(1-0,20)*(1+0,25) = 100*(1,375) = 137,5€*

Applicando la **media geometrica** si ottiene $(1{,}10*1{,}25*0{,}8*1{,}25)^{1/4}$ = **1,0829**,
che è il **rendimento medio** realizzato nel periodo, con un **tasso di crescita** del **8,29%** , infatti:
100(1,0829)*(1,0829)*(1,0829)*(1,0829) = 100*(1,0829)* ^{4} *= 137,5€*.

La **media aritmetica dei rendimenti** è invece $\frac{1{,}10 + 1{,}25 + 0{,}80 + 1{,}25}{4} = 1{,}1$.
Utilizzandola come rendimento medio nei 4 anni, si ha *100*(1,1)*^{4} *= 146,41€*: che è un **valore diverso da quello effettivo** di 137,5€!

La media geometrica – In Excel (opzionale)

Nel caso si conoscessero **solo i valori di inizio e fine periodo**, e si volesse ottenere la media geometrica senza conoscere le percentuali, si può utilizzare il **tasso di crescita annuale composto**, conosciuto come **CAGR** (Compound Annual Growth Rate) derfinito come: $(V_t/V_0)^{1/t}$

dove V_0 e V_t sono rispettivamente i valori di inizio e fine periodo e t è il numero di periodi.

Infatti, riprendendo l'esempio precedente si ha: $(137{,}5/100)^{1/4} = 1{,}375^{0{,}25} =$ **1,0829** (**8,29%**)

La media geometrica con Excel si calcola con la funzione **MEDIA.GEOMETRICA().**

	A	B	C	D	E	F	G	H
1	**Anni**	**0**	**1**	**2**	**3**	**4**		
2	Prezzi	100	110	137,5	110	137,5		
3	Rendimenti		10,0%	25,0%	-20,0%	25,0%		
4	Tassi di incremento		1,10	1,25	0,80	1,25		
5								
6	Val. precedente + media aritm.	100	110,0	121,0	133,1	146,4	=E6*G9	
7	Val. precedente + media geom.	100	108,3	117,3	127,0	137,5	=E7*G11	
8								
9		Media aritmetica (funzione):					**1,10**	=MEDIA(C4:F4)
10								
11		Media geometrica(funzione):					**1,0829**	=MEDIA.GEOMETRICA(C4:F4)
12		Media geometrica(calcolo):					1,0829	=PRODOTTO(C4:F4)^(1/CONTA.VALORI(C4:F4))
13		CAGR (calcolo sui valori di inizio e fine periodo):					1,0829	=(F2/B2)^(1/CONTA.VALORI(C2:F2))
14								

Variabili quantitative: indici di posizione

- **Mediana**

  E' il valore che **divide** la distribuzione in **due parti di uguale numerosità**.

  In una **scala ordinata** di *n* valori è il numero che compare **al centro** se ***n* è dispari**; è invece la **media** aritmetica tra i **due valori centrali** se ***n* è pari**.

  Usando l'esempio precedente, mettendo in ordine crescente i valori delle età:
  [*48,54,56,57,65,68,68,68,69,**70,70**,71,72,73,75,76,76,78,84,85*], la mediana (70+70) / 2 è ***70***.

- **Moda**

  è il valore che si presenta con **maggior frequenza.**

  Dato che la moda dipende soltanto dalla frequenza delle osservazioni, è l'*unica misura di tendenza centrale per **variabili di tipo qualitativo.***

  Usando l'esempio precedente, la moda è ***68*.**

Calcolo della Media, Mediana e Moda – In Excel

La mediana in Excel si calcola con la funzione **MEDIANA()** e la moda con la funzione **MODA().**
Ad esempio:

	A	B	C	D	E	F	G	H
1	Età							
2	48							
3	54							
4	56							
5	57							
6	65							
7	68							
8	68							
9	68							
10	69		Minimo	48,00	=MIN(A2:A21)			
11	70		Moda	68,00	=MODA(A2:A21)			
12	70		Media	69,15	=MEDIA(A2:A21)			
13	71		Mediana	70,00	=MEDIANA(A2:A21)			
14	72		Massimo	85,00	=MAX(A2:A21)			
15	73							
16	75							
17	76							
18	76							
19	78							
20	84							
21	85							
22								

Calcolo della mediana e della moda

Confronto tra media aritmetica e mediana

Molto spesso la media e la mediana presentano valori simili. Ciò accade quando il numero di valori al di sotto del valore centrale e quelli al di sopra più o meno si equivalgono.

Se la mediana ha un valore diverso da quello della media aritmetica, significa che ci sono osservazioni con **valori estremi**[1] (molto alti o molto bassi).

Ad esempio, nell'insieme [**2,5,8,11,48**] la media è **14,8**, mentre la mediana è **8**. Portando il valore estremo da **48** a **96 la mediana non cambia**, mentre la media diventa **24,4**.

La **mediana è meno sensibile ai valori estremi**, per cui è da preferire alla media nel caso di distribuzioni con presenza di valori anomali ed estremi[2].

[1] Sono quei valori che si distinguono dagli altri, come se non appartenessero allo stesso insieme.
[2] Ad esempio, redditi, investimenti, ecc., in tutti i casi in cui i numeri sono "grandi" e molto diversi tra loro.

Variabili quantitative: indici di posizione

• Quantili

Il quantile è il valore di una variabile sotto il quale sta un certo percento delle osservazioni. I quantili sono dunque quei valori che dividono una **scala ordinata** di valori **in gruppi di uguale numerosità**[1], ciascuno dei quali contiene il ***q* per cento** della distribuzione.

*Se si divide la distribuzione ordinata in **cento parti** di uguale numerosità, ciascun valore che separa l'1%, 2%, …, 99%, 100% dal resto si chiamerà **percentile**;*

*se si divide la distribuzione ordinata in **dieci parti** di uguale numerosità, ciascun valore che separa il 10%, 20%, ..., 90%, 100% dal resto si chiamerà **decile**;*

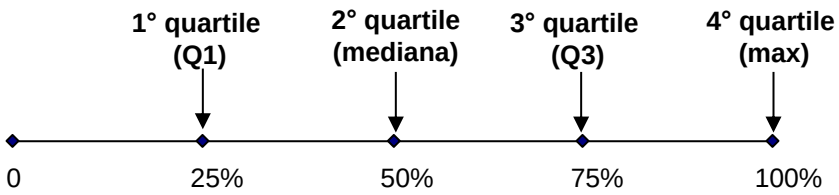
*se di divide la distribuzione ordinata in **cinque parti** di uguale numerosità, ciascun valore che separa il 20%, 40%, …, 80%, 100% dal resto si chiamerà **cinquile**.*

[1] La mediana corrisponde al 50° percentile e al 5° decile in quanto sotto quel valore ci sta il 50% delle osservazioni e, sopra, l'altro 50%.

Variabili quantitative: indici di posizione

Quantili

I quantili usati più frequentemente sono i ***quartili*** che dividono in quattro parti di uguale ampiezza una **scala ordinata** di valori:



Il primo quartile, ***Q1***, è il valore che lascia prima di sé il 25% delle osservazioni che hanno valori inferiori;

il secondo quartile, ***Q2,*** è il valore che divide la distribuzione in due parti uguali, corrisponde alla ***mediana***;

il terzo quartile, ***Q3***, è il valore che lascia prima di sé il 75% delle osservazioni che presentano valori inferiori;

l'ultimo quartile, ***Q4***, corrisponde al ***valore massimo***.

Variabili quantitative: indici di posizione

• Quantili

Ci sono varie metodologie[1] per calcolare i quantili che possono produrre risultati lievemente differenti a causa del metodo selezionato per l'effettuazione del calcolo.

Dai dati dell'esempio precedente, usando il metodo **R-2** (quello usato precedentemente per il calcolo della mediana), Q1 corrisponde a **66,5** *(65+68)/2*, Q3 a **75,5** *(75+76)/2*:

[*48,54,56,57,**65,68**,68,68,69,**70,70**,71,72,73,**75,76**,76,78,84,85*]

1° quartile **2° quartile** **3° quartile**

Usando invece il metodo **R-7**[2] (il più diffuso), Q1 corrisponde a **67,25** e Q3 a **75,25.**

Quindi, con questo metodo, si può affermare che il ***25% dei clienti più giovani ha meno di 67,25 anni, mentre il 25% dei clienti più anziani ha più di 75,25 anni.***

[1] https://en.wikipedia.org/wiki/Quantile
[2] **Excel utilizza il metodo R-7 con la funzione INC.PERCENTILE**

Quantili – In Excel

I quantili in Excel si calcolano con la funzione **INC.PERCENTILE().** Ad esempio:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1		Età	Quartili		Cinquili		Decili		Percentili						
2		48	0	48,0	0	48,0	0	48,0	0	48,0					
3		54	1	67,3	=INC.PERCENTILE($B$2:$B$21;C3*25/100)				1	55,8	=INC.PERCENTILE(R3:R22;I3/100)				
4		56	2	70,0	2	68,6	2	63,4	2	63,4					
5		57	3	75,3	=INC.PERCENTILE($B$2:$B$21;75%)				...	...					
6		65	4	85,0	4	76,0	4	68,6	24	66,7					
7		68			5	85,0	5	70,0	25	67,3					
8		68					6	71,4	=INC.PERCENTILE($B$2:$B$21;G8*10/100)						
9		68					7	73,6	50	70,0					
10		69					8	76,0	...	...					
11		70					9	78,6	74	75,1					
12		70					10	85,0	75	75,3					
13		71							...	...					
14		72							99	84,8					
15		73							100	85,0					
16		75													
17		76													
18		76													
19		78													
20		84													
21		85													
22															

Indici di posizione – Test intermedio

1. **Nell'insieme dei numeri 3,5,2,1,4 la media e la mediana sono rispettivamente:**

   - ❑ 3; 2
   - ❑ 3; 3

2. **Se un insieme di 4 valori con media 30 si aggiunge il valore 80, quale sarà la nuova media?**

   - ❑ 40
   - ❑ 55

3. **Il 20° percentile dell'insieme dei numeri al punto 1 corrisponde al valore (calcolare con Excel):**

   - ❑ 5
   - ❑ 1,8

Variabili quantitative: misure di variabilità

Il poeta romano Carlo Alberto Salustri (1871-1950) noto con lo pseudonimo di Trilussa, è spesso ricordato per l'aforisma del pollo:

> *"da li conti che se fanno*
> *seconno le statistiche d'adesso*
> *risurta che te tocca un pollo all'anno:*
> *e, se nun entra nelle spese tue,*
> *t'entra ne la statistica lo stesso*
> *perch'è c'è un antro che ne magna due."*

In effetti, con solo queste due misurazioni [**0;2**] la media è effettivamente **1**.

Se i prelevamenti di Alberto sono stati [**300,400,300,200**] e quelli di Angela [**300,300,300,300**], per entrambi la media, la mediana e la moda è **300**. E' però evidente che i prelevamenti di Angela sono stati sempre uguali, mentre quelli di Alberto più discontinui.

Come si può esprimere la "bontà" della media quale descrittore di una distribuzione?

La risposta a questa domanda viene data dagli **indici di variabilità**.

Variabili quantitative: misure di variabilità

La ***variabilità*** fornisce informazioni sulla consistenza o sulla stabilità di una misura all'interno di un insieme di dati. Le principali misure di variabilità includono:

- **Campo di variazione (definito anche come Intervallo di variabilità o Range)**

  E' il più semplice da calcolare ed è dato dalla differenza fra il maggiore e il minore dei valori rilevati[1]. Viene usato per verificare che non vi siano valori che cadano oltre limiti prefissati (ad esempio, nel controllo della qualità di un processo produttivo).

- **Deviazione standard**

  E' uno degli indici di variabilità più noti e utilizzati, il simbolo statistico è la lettera greca sigma (σ): dà una **misura dello scostamento dei valori di una variabile dalla media**.

  Una **bassa deviazione standard** indica ch**e la maggior parte dei valori sono vicini al valore medio**, una deviazione standard elevata significa un'ampia variabilità nei valori.

[1] Non è un buon indice di variabilità in quanto risente dell'effetto dei valori anomali.

Varianza e deviazione standard

La **varianza** è la media degli scarti dalla media elevati al quadrato. Viene rappresentata con la lettera greca sigma minuscola al quadrato (σ^2)

$$\sigma^2 = \frac{\sum(x_i - \mu)^2}{N}$$

Di solito viene utilizzata la **deviazione standard** (o scarto quadratico medio) che è la radice quadrata della varianza, rappresentata con il simbolo σ.

$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$$

La radice quadrata consente di mantenere la **stessa unità di misura** della variabile stessa, rendendola più **interpretabile** rispetto alla varianza.

Se i dati non rappresentano l'intera popolazione di riferimento ma un suo **campione *n***, il denominatore dovrà essere $\boldsymbol{n-1}$ invece che $\boldsymbol{N}$, e si indica con la lettera $\boldsymbol{s}$.

$$\boldsymbol{s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}}$$

Indici di variabilità – In Excel

La **deviazione standard** in Excel si calcola con la funzione **DEV.ST.POP()** se riguarda tutta la popolazione, **DEV.ST()** se riguarda un campione. Ad esempio:

	A	B	C	D	E	F
1	Prelevamenti					
2	Angela	Alberto				
3	300	300	N	4	4	
4	300	400	Media	300	300	
5	300	200	Dev. Standard popolazione (calcolo)	0	70,7	=RADQ(((B3-E4)^2+(B4-E4)^2+(B5-E4)^2+(B6-E4)^2)/**E3**)
6	300	300	Dev. Standard popolazione (funzione)	0,0	70,7	=**DEV.ST.PO**P(B3:B6)
7			Dev. Standard campione (calcolo)	0,0	81,6	=RADQ(((B3-E3)^2+(B4-E3)^2+(B6-E3)^2+(B5-E3)^2)/(**E3 - 1**))
8			Dev. Standard campione (funzione)	0,0	81,6	=**DEV.ST**(B3:B6)
9			CV	0	0,272	=E5/E4
10						

Dal valore della deviazione standard si nota che per i prelevamenti **nel primo caso non c'è variabilità** (s = 0), nel secondo c'è variabilità (s = 81,6).

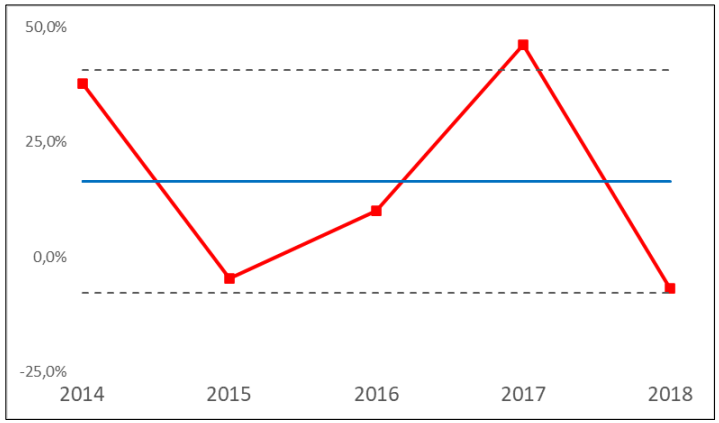
Indici di variabilità – In Excel

Un'applicazione della deviazione standard in ambito finanziario è la misura **della volatilità di una quotazione azionaria** e, quindi, del suo rischio (un'azione molto volatile avrà una deviazione standard molto più grande rispetto a un'azione più stabile).

Ad esempio, se i rendimenti delle azioni della Apple (AAPL) negli anni 2014-2018 sono stati 37,7%, -4,6%, 10%, 46,1%, -6,8%, la media dei rendimenti[1] è stata del 16,5% e la deviazione standard del 24,3%.

La performance dell'azione in quegli anni, quindi, è oscillata tra il ±24,3% intorno alla media del 16,5%.

	Anno	Rendimenti
AAPL	2014	37,7%
	2015	-4,6%
	2016	10,0%
	2017	46,1%
	2018	-6,8%
Media dei rendimenti		16,5%
Deviazione standard (σ)		24,3%



[1] In questo caso si usa la media dei rendimenti e non il rendimento medio perché la **media aritmetica è più adatta per analisi di serie temporali** e stime del rischio; **la media geometrica lo è per le analisi dei rendimenti ex-post**.

Coefficiente di variazione

Detto anche **deviazione standard relativa**, il coefficiente di variazione viene definito come il rapporto tra la deviazione standard e il valore assoluto della media aritmetica.

$$CV_{POPOLAZIONE} = \frac{\sigma}{|\mu|}; \quad CV_{CAMPIONE} = \frac{s}{|\bar{x}|}$$

È un ***indice di variabilità relativa*****,** utilizzato per:

- **Confronti di variabilità** tra variabili quando queste hanno unità di misura o medie diverse tra loro.
- **Misura della "bontà" della media** come indicatore statistico.

Regola empirica:	CV < 0,1	(<10%)	molto buono, la media è rappresentativa
	CV 0,1-0,2	(20%)	buono
	CV 0,2-0,3	(30%)	accettabile
	CV > 0,3	(>30%)	non accettabile

Nel caso dei polli di Trilussa, il CV è $\sqrt{\frac{(0-0,5)^2+(1-0,5)^2}{(2-1)}} / 0,5 = 1$, per cui la media non è un indice accettabile.

Coefficiente di variazione – In Excel

Il coefficiente di variazione in Excel si calcola facendo il **rapporto** tra la funzione DEV.ST.POP() o **DEV.ST()** e la funzione **MEDIA().**

nel caso invece dei prelevamenti

	A	B	C	D	E	F	G	H
1	Prelevamenti							
2	Angela	Alberto						
3	300	300	N	4	4			
4	300	400	Media	300	300			
5	300	200	Dev. Standard campione (funzione)	0,0	81,6	=DEV.ST(B3:B6)		
6	300	300	CV	0	0,272	=E5/E4		
7								

il CV è rispettivamente 0 (ottimo) per Angela e 0,272 (accettabile) per Alberto.

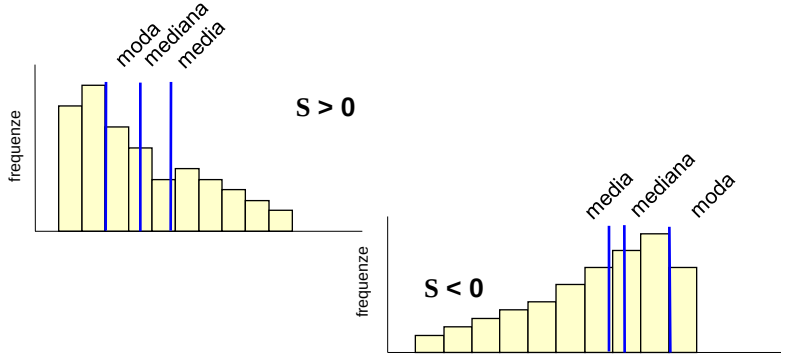
Variabili quantitative: misure di forma

Sono indici utilizzati per evidenziare particolarità nella forma della distribuzione (simmetria e code):

- **Asimmetria**

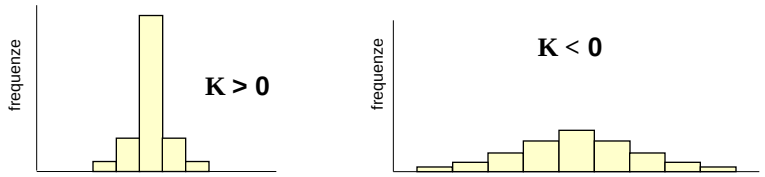
  L'asimmetria è a **destra**[1] quando c'è presenza di valori estremi molto alti.
  moda < mediana < media.

  L'asimmetria è a **sinistra**[1] quando c'è presenza di valori estremi molto piccoli.
  media < mediana < moda.



- **Curtosi**

  La forma di una distribuzione dipende anche dal grado di **addensamento dei valori intorno alla media**, cioè quanto è piatta oppure appuntita.



Quando sia l'indice di asimmetria **S** sia l'indice di curtosi **K** sono uguali a 0, si ha una distribuzione simmetrica, non appiattita né appuntita, detta "***normale***".

[1] La media si trova dalla parte dell'asimmetria, mentre la mediana non lo è, proprio perché non è influenzata dai valori estremi.

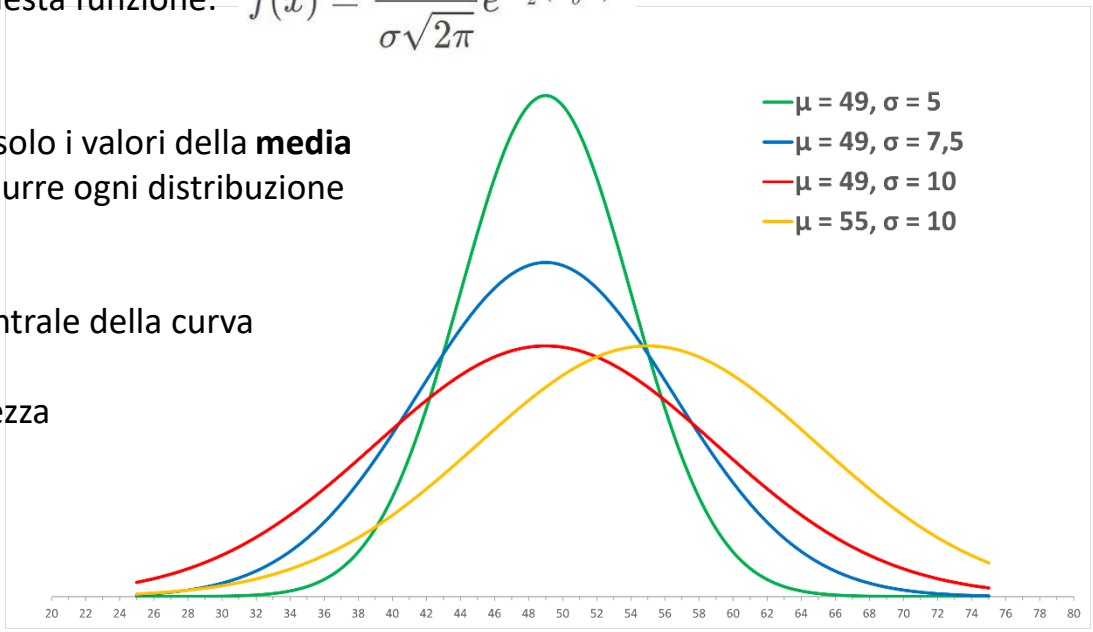
Variabili quantitative: la distribuzione normale

Se i valori possibili di una variabile quantitativa **fossero moltissimi**, addirittura tendenti all'infinito, **il grafico tenderebbe a una curva, detta normale (o *gaussiana*)** che ha l'aspetto di "campana" ed è simmetrica[1] rispetto alla media.

Formalmente, si può esprimere con questa funzione: $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$

Dove si evince che bisogna conoscere solo i valori della **media** e della **deviazione standard** per ricondurre ogni distribuzione alla sua forma teorica.

- La media determina la posizione centrale della curva
- La deviazione standard la sua ampiezza



[1] In questo caso la media, la mediana e la moda coincidono.

Variabili quantitative: la distribuzione normale

Assumendo che il campione oggetto di analisi provenga da una popolazione normale, **si possono ottenere intervalli** delimitati da due limiti (inferiore e superiore) **contenenti una certa percentuale di osservazioni**.

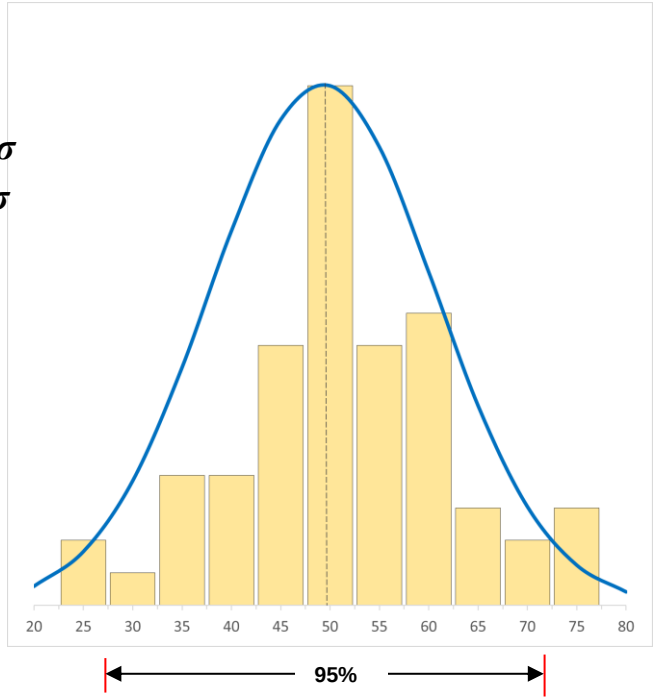
Alcuni esempi di intervalli con le loro percentuali:

- il **68,3**% delle osservazioni cade entro l'intervallo **media ± 1σ**
- il **95,0**%[1] delle osservazioni cade entro l'intervallo **media ± 1,96σ**
- il **99,0**% delle osservazioni cade entro l'intervallo **media ± 2,58σ**

Prendendo come esempio i dati di pag. 10 si ha la distribuzione centrata attorno a una **media di 49 anni** e con una **deviazione standard di 11,37**.

Ipotizzando una distribuzione normale, **il 95% di questi soggetti ha un'età compresa tra i 27 e i 72 anni**.

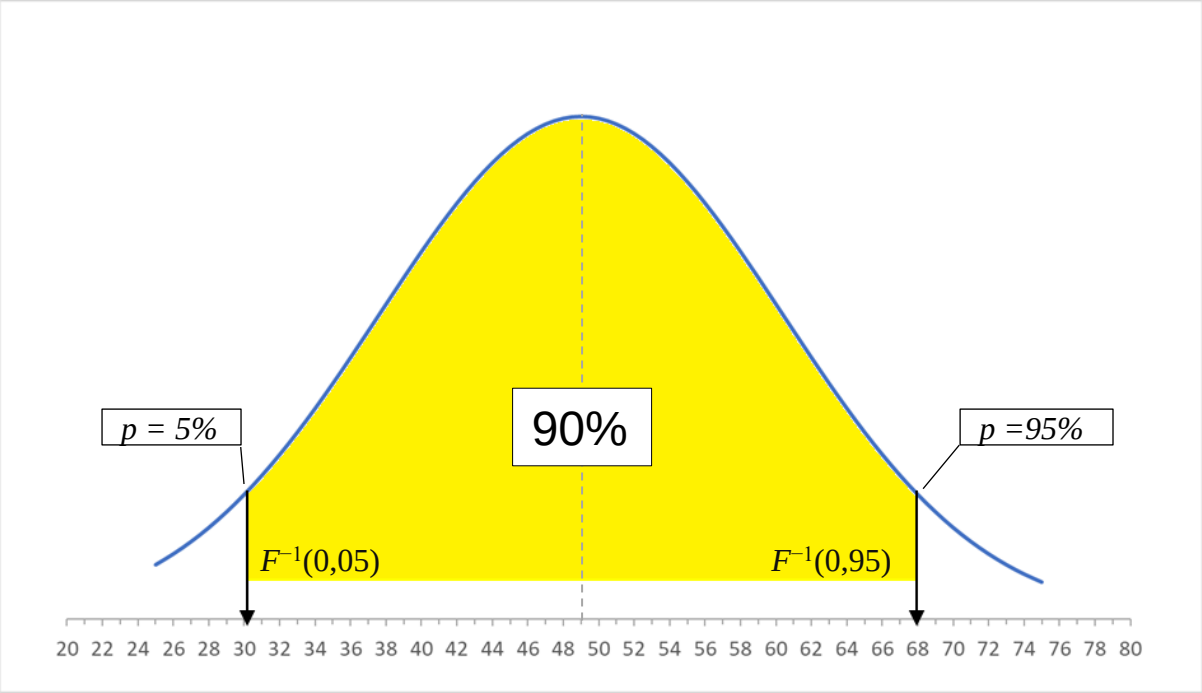
Quelli che cadono fuori dall'intervallo sono detti **valori estremi**.



[1] Di norma si assumono i valori della media $\pm 1{,}96\sigma$ come limiti della normalità; quelli che cadono all'infuori sono valori estremi.

Variabili quantitative: la distribuzione normale

Dato un **valore di probabilità *p***, è possibile calcolare il **valore corrispondente** della variabile in esame attraverso la **funzione inversa di probabilità** (detta anche *funzione di distribuzione cumulata* o *funzione quantile*) o di una distribuzione normale standard $F^{-1}(p)$.



Variabili quantitative: la distribuzione normale in Excel

Dall'esempio precedente, nel caso di una distribuzione normale standard, può essere utilizzata in Excel la funzione inversa di probabilità **INV.NORM**(*prob, media, dev. std*) per trovare i valori entro i quali è compreso il 90% dei dati che corrispondono ai percentili 5 e 95, ovvero i valori di età di 31 e 68 anni:

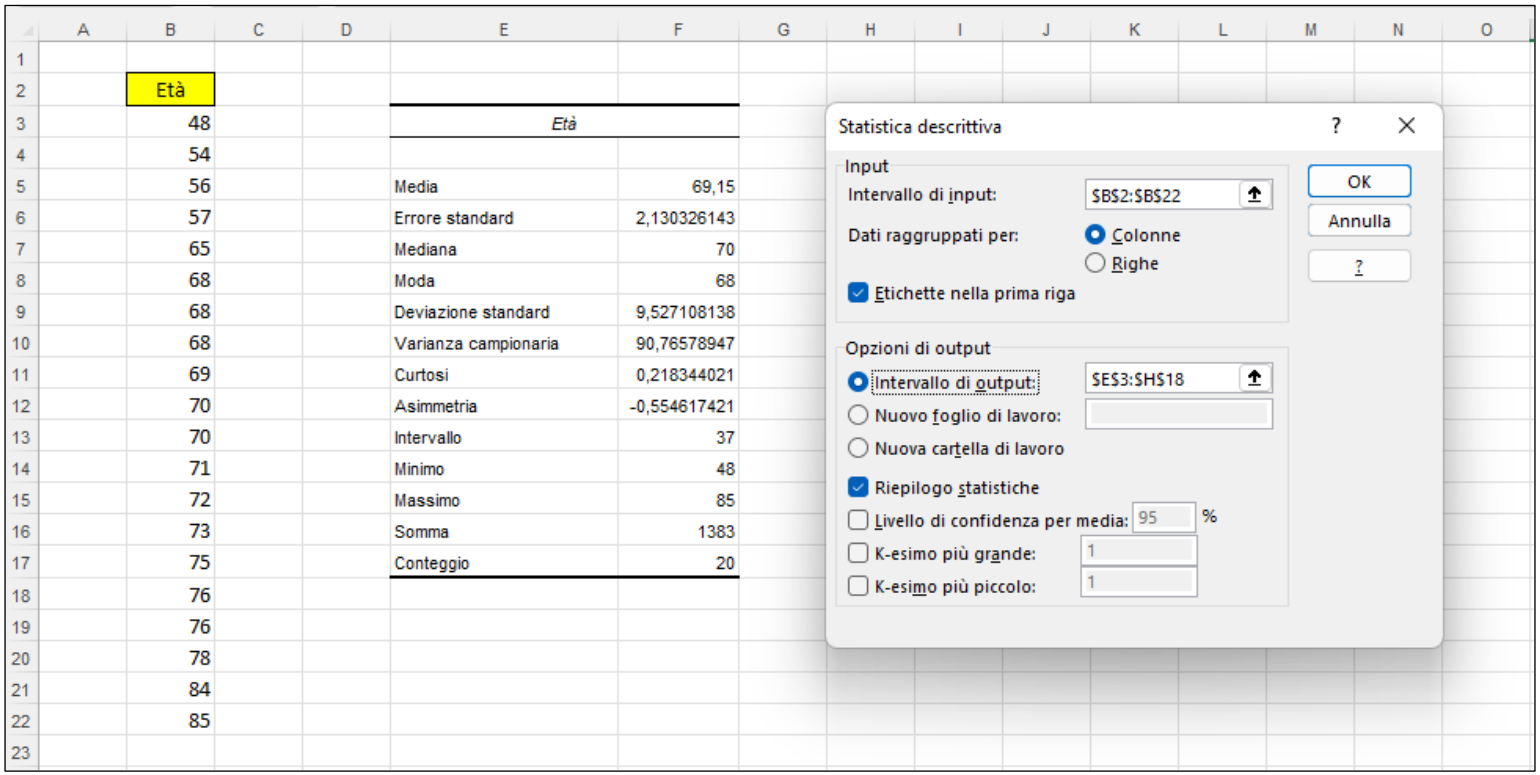
	A	B	C	D	E	F	G	H	I	J
1										
2		Età		media	dev. Standard					
3		23		49,25		11,3774				
4		26								
5		31		Area di interesse:	**90%**					
6		...		Area esterna:	10%		Limiti età per l'intervallo specificato			
7		72		Prob. limite inferiore (0+area esterna/2)	5,0%		**30,54**	=INV.NORM.N(E7;D3;F3)		
8		74		Prob. limite superiore (1 -area esterna/2)	95,0%		**67,96**	=INV.NORM.N(E8;D3;F3)		
9		75								
10										

il 95% dei soggetti, che corrisponde ai percentili 2,5 e 97,5, ha un'età compresa tra 27 e 72 anni:

	A	B	C	D	E	F	G	H	I	J
1										
2		Età		media	dev. Standard					
3		23		49,25		11,3774				
4		26								
5		31		Area di interesse:	**95%**					
6		...		Area esterna:	5%		Limiti età per l'intervallo specificato			
7		72		Prob. limite inferiore (0+area esterna/2)	2,5%		**26,95**	=INV.NORM.N(E7;D3;F3)		
8		74		Prob. limite superiore (1 -area esterna/2)	97,5%		**71,55**	=INV.NORM.N(E7;D3;F3)		
9		75								
10										

Statistiche descrittive – In Excel

Excel permette di calcolare i diversi valori di posizione e di dispersione, in modo più rapido, facendo clic su **Analisi dati**[1] della scheda **Dati** selezionando **Statistica descrittiva**.



[1] Se il comando Analisi dati non è disponibile, è necessario caricare il componente aggiuntivo **Strumenti di analisi**.

Indici di variabilità – Test intermedio

1. **Nell'insieme di numeri 35, 30, 33, 34, 30, 32 qual è la DEVIAZIONE STANDARD CAMPIONARIA?**

   - ❑ 1,88
   - ❑ 2,06

2. **In base al COEFFICIENTE DI VARIAZIONE, la media dei dati precedenti è un buon indicatore? Perché?**

   - ❑ Sì
   - ❑ No

3. **Un secondo insieme ha media 67,3 e deviazione standard campionaria 3,41, qual è in base al COEFFICIENTE DI VARIAZIONE l'insieme che ha minor variabilità?**

   - ❑ Il primo
   - ❑ Il secondo

Programma

- Struttura e visualizzazione dei dati
- Sintesi e interpretazione
- **Relazioni tra variabili**
- Inferenza statistica
- Verifica delle ipotesi
- Soluzioni dei test

Correlazione

Quando si analizzano due variabili quantitative, è possibile calcolare l'indice di **correlazione lineare r**, che misura l'"intensità" di una **relazione tra due variabili numeriche**.

Una **correlazione** tra due variabili **non implica necessariamente un rapporto di causa ed effetto**, poiché misura solo la **tendenza** che hanno le due variabili a **variare congiuntamente** (si prenda, ad esempio, la relazione tra prezzo e domanda di una merce: il prezzo influisce sulla domanda e la domanda influisce sul prezzo).

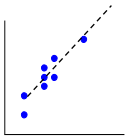
Si misura con il **coefficiente di correlazione *r*** di Pearson:

$$r_{x,y} = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\sum(x - \bar{x})^2 \sum(y - \bar{y})^2}}$$

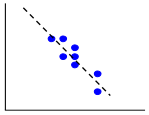
Correlazione

L'**indice *r* di correlazione lineare** tra due variabili X e Y non dipende dalle loro unità di misura, e il suo valore è compreso tra -1 e +1.

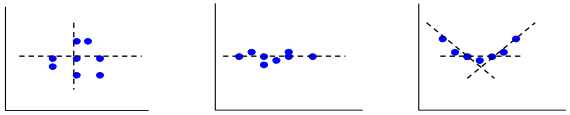
Se ***r*** > 0, la correlazione è **positiva**



se ***r*** < 0, la correlazione è **negativa**



se ***r*** = 0, **non esiste correlazione**



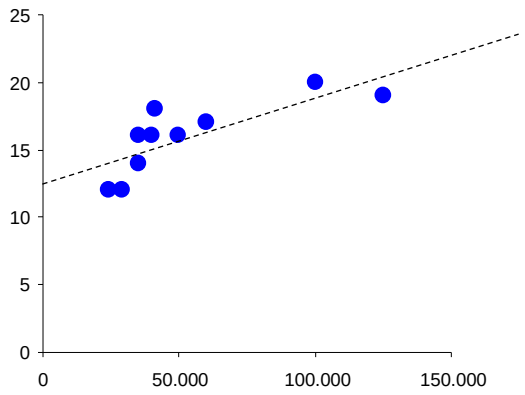
Per la sua valutazione viene spesso usata questa regola empirica:

|r| = 1	perfetta
|r| > 0,9	fortissima
0,7 < |r| < 0,9	forte
0,5 < |r| < 0,7	moderata
0,3 < |r| < 0,5	debole
0 < |r| < 0,3	molto debole
|r| = 0	nessuna

Correlazione – In Excel

L'indice r di correlazione lineare si calcola in Excel con la funzione **CORRELAZIONE().**
Ad esempio, si vuol sapere se esiste una **relazione tra reddito e anni di istruzione**.

	A	B	C	D
1	**Reddito**	**Anni di istruzione**		
2	125.000	19		
3	100.000	20		
4	40.000	16		
5	35.000	16		
6	41.000	18		
7	29.000	12		
8	35.000	14		
9	24.000	12		
10	50.000	16		
11	60.000	17		
12				
13	Indice di			
14	correlazione	0,7887	=CORRELAZIONE(A2:A11;B2:B11)	
15				



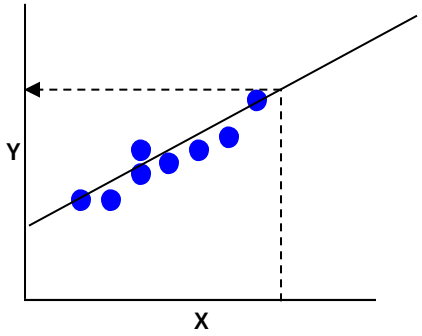
In questo esempio, **c'è correlazione** tra reddito e anni di istruzione ($|r| > 0{,}60$).

Come indicato precedentemente: **correlazione non vuol dire causalità**! Da questo esempio si evince che l'istruzione e il reddito sono positivamente correlati ma **non si sa**, però, **se uno ha causato l'altro** (si può infatti affermare sia che persone con più anni d'istruzione hanno redditi più alti, sia che persone con alti redditi si sono potute permettere un'istruzione avanzata).

Regressione

Per determinare se, e come, una variabile dipende da un'altra, è necessario stabilire una **relazione funzionale**.

Adottando un modello lineare come l'**equazione di una retta**, si può associare a qualsiasi valore della variabile indipendente *(*anche non presente nei dati rilevati) un valore corrispondente della variabile dipendente.



La retta di regressione, essendo una funzione, può essere utilizzata per ***effettuare delle previsioni sul valore assunto da una variabile rispetto a un'altra***.

La misura della bontà del modello è data dal coefficiente di determinazione[1] R^2.

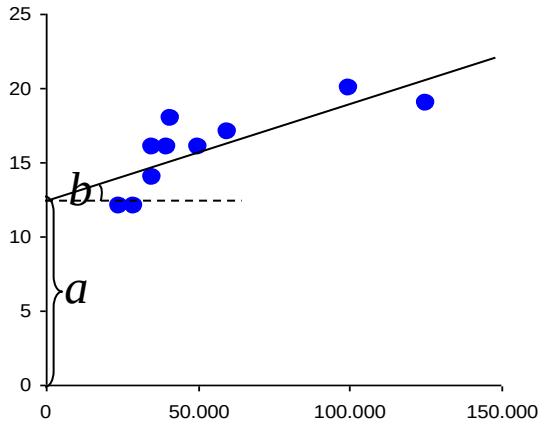
[1] Indica quanta variabilità è spiegata dal modello lineare. Questo numero varia tra 0 e 1; valori superiori a 0,80 indicano un buon adattamento.

Regressione lineare semplice

La regressione lineare tra due variabili *X* e *Y* è la **funzione di una retta**:

$$y = a + bx$$

dove a rappresenta l'**intercetta** (indica il valore assunto da *y* quando *x*=0) e b è il **coefficiente angolare** che rappresenta l'inclinazione della retta (quanto varia in media y al variare di un'unità di x).



Maggiore è la pendenza, maggiore è la variazione di y al variare di x (e quindi è maggiore è l'importanza di x).

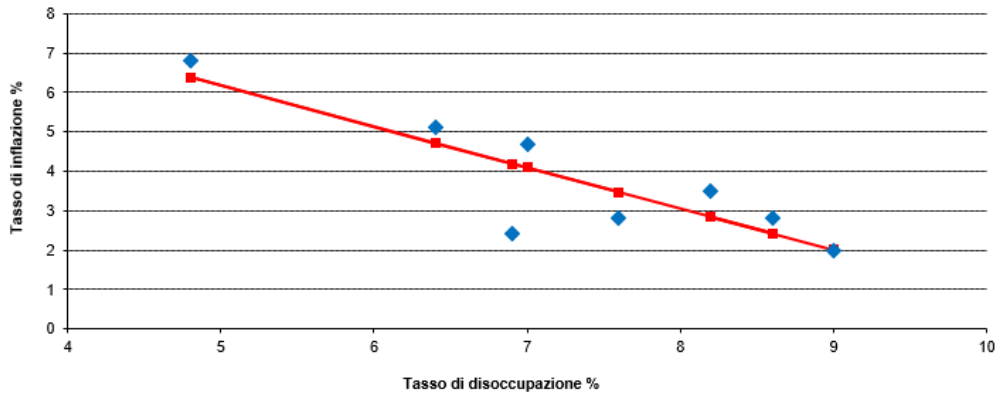
La pendenza e l'intercetta vengono determinati in modo che la somma dei quadrati degli scarti di ogni punto dalla retta siano minimizzate (principio dei minimi quadrati[1]).

[1] Trova **quell'unica retta** che ha la proprietà di ridurre al minimo la somma dei quadrati degli scarti tra i valori osservati e quelli teorici.

Regressione lineare – In Excel

Si vuole stabilire una **relazione** funzionale tra **disoccupazione (variabile indipendente) e inflazione (variabile dipendente)** rilevata in vari anni come riportato nella seguente tabella:

	A	B	C	D	E
1	Tasso di disoccupazione (X)	Tasso di inflazione (Y)	Ŷ = a + bx		
2	9	2	2,0100	=C$13+C$12*A2	
3	8,2	4	2,8420		
4	8,6	3	2,4260		
5	7	5	4,0900		
6	6,4	5	4,7140		
7	4,8	7	6,3780		
8	6,9	2	4,1940		
9	7,6	3	3,4660	=C$13+C$12*A9	
10					
11	Indice di correlazione=		-0,8600	-0,8586	=CORRELAZIONE(B2:B9;A2:A9)
12	b=		-1,0400	-1,0404653	=PENDENZA(B2:B9;A2:A9)
13	a=		11,3700	11,370902	=INTERCETTA(B2:B9;A2:A9)
14	R^2=		0,74	0,7372531	=RQ(B2:B9;A2:A9)



Regressione lineare – In Excel

Dall'esempio precedente si ottiene:

- ✓ La retta di regressione, espressa dall'equazione $\hat{y} = 11,37 - 1,04 * x$
- ✓ Il coefficiente di determinazione $R^2 = 0,74$. Questa misura indica che il modello spiega il 74% della variabilità del fenomeno in esame.

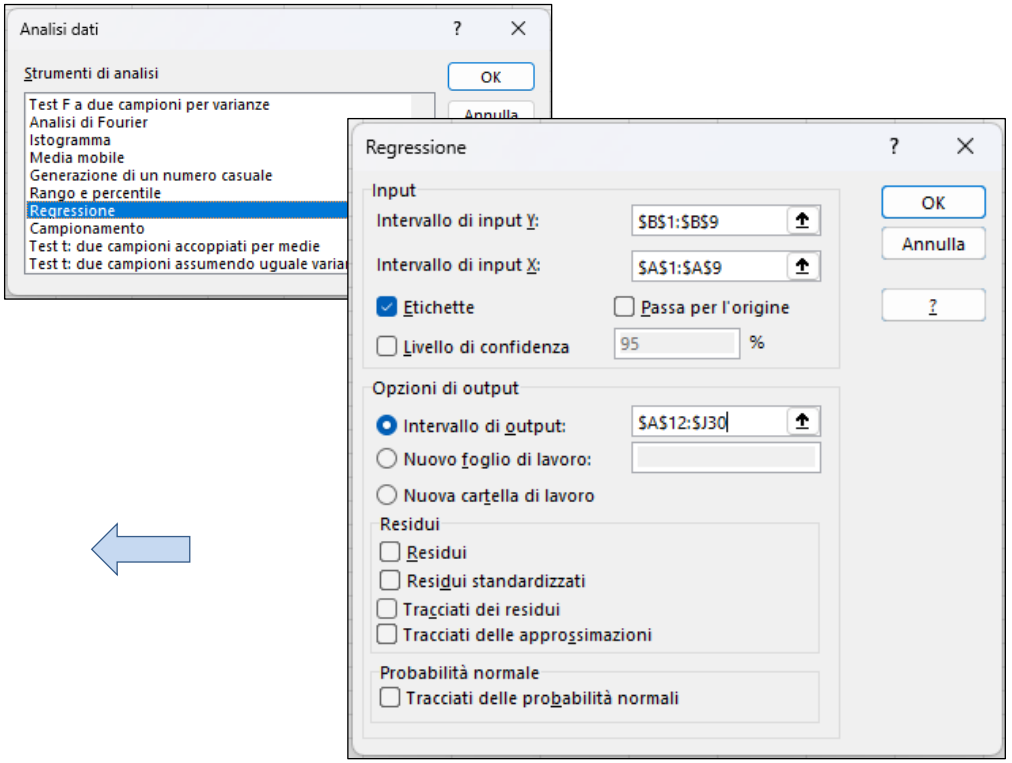
E' allora possibile calcolare a che tasso d'inflazione corrisponde una disoccupazione del 6%: **sostituendo** il valore **6 alla *x*** dell'equazione, **si ottiene 5,13** (11,37 - 1,04 * 6).

Si può quindi affermare che, data la relazione che appare dalle misure di inflazione e di disoccupazione, **un tasso di disoccupazione del 6% porta a valori di inflazione prossimi al 5,13%**.

Regressione lineare – In Excel

Tramite il comando **Analisi dati** presente nel menu **Dati** selezionando **Regressione** è possibile calcolare i parametri della regressione in modo più rapido.

	A	B
1	**Tasso di disoccupazione (X)**	**Tasso di inflazione (Y)**
2	9	2
3	8,2	3,5
4	8,6	2,8
5	7	4,7
6	6,4	5,1
7	4,8	6,8
8	6,9	2,4
9	7,6	2,8
10		



10			
11			
12	OUTPUT RIEPILOGO		
13			
14	*Statistica della regressione*		
15	R multiplo	0,858634412	
16	R al quadrato	**0,737253053**	
17	R al quadrato corretto	0,693461895	
18	Errore standard	0,90683141	
19	Osservazioni	8	
20			
21	ANALISI VARIANZA		
22		*gdl*	*SQ*
23	Regressione	1	13,84469077
24	Residuo	6	4,934059232
25	Totale	7	18,77875
26			
27		*Coefficienti*	*Errore standard*
28	Intercetta	**11,37090216**	1,881806535
29	Tasso di disoccupazione (X)	**-1,040465253**	0,253578545
30			

Correlazione e Regressione – Test intermedio

1. **Il coefficiente di correlazione delle variabili X= 20; 35; 43; 38; 51; 46; 48; 53; e Y= 430; 370; 340; 300; 225; 230; 180; 210; è**

   - ❑ -0.905
   - ❑ 0,905

2. **Se Y fosse la variabile dipendente e X quella indipendente in un modello di regressione lineare, il valore di intercetta e di pendenza sarebbe:**

   - ❑ intercetta: 597,6; pendenza: -7,47
   - ❑ Intercetta: 73; pendenza: -0,11

3. **Dal valore di R^2 si può dire che le stime del modello sono buone?**

   - ❑ Sì
   - ❑ No

4. **Ipotizzando un valore di X pari a 50, quale sarà il valore previsto di Y?**

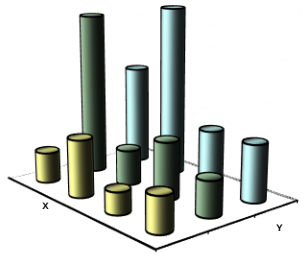
   - ❑ 67,6
   - ❑ 224

Associazione

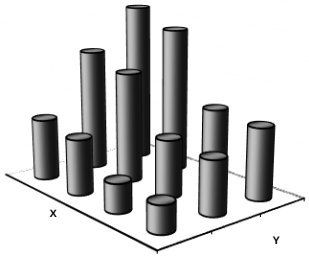
Quando si analizzano due variabili qualitative (o categoriche), vengono utilizzate le tabelle a doppia entrata (***tabelle di contingenza***) in cui sono rappresentate le frequenze incrociate delle modalità delle due variabili.

Attraverso la tabella di contingenza è possibile testare se tra due variabili esiste un'**associazione significativa,** cioè se al manifestarsi di determinate modalità della prima variabile si manifestano determinate modalità della seconda variabile, attraverso l'utilizzo dell'indice chi-quadrato (χ^2)

Se esiste una **differenza significativa tra i dati osservati e i dati teorici**, ovvero se l'indice χ^2 calcolato sui dati osservati è maggiore di quello calcolato sui dati teorici in caso di non associazione (indipendenza), allora si può concludere che esiste un **legame di dipendenza** tra le due variabili.



dati osservati



dati teorici
(in caso di
indipendenza)

Tabelle di contingenza

Lo studio dell'associazione avviene attraverso i seguenti passaggi:

1. Si costruisce la **tabella di contingenza** e si ricavano le **frequenze marginali**

Le tabelle di contingenza sono un particolare tipo di tabelle a doppia entrata dove si riportano le frequenze congiunte (osservate) f_{ij} delle due variabili e i totali di riga e colonna

$f_{i.} = f_{i1} + f_{i2} + f_{i3}$ e $f_{.j} = f_{1j} + f_{2j} + f_{3j}$

che rappresentano le frequenze marginali assolute delle variabili.

	y1	y2	y3	Totale riga
x1	f_{11}	f_{12}	f_{13}	$f_{1.}$
x2	f_{21}	f_{22}	f_{23}	$f_{2.}$
x3	f_{31}	f_{32}	f_{33}	$f_{3.}$
Totale col.	$f_{.1}$	$f_{.2}$	$f_{.3}$	n

2. Si costruisce la **tabella di frequenza teorica**

È la tabella delle **frequenze attese nel caso di indipendenza** (cioè nessuna associazione tra le variabili)

$f^*_{ij} = (f_{i.} * f_{.j}) / n$

	y1	y2	y3	Totale riga
x1	f^*_{11}	f^*_{12}	f^*_{13}	$f_{1.}$
x2	f^*_{21}	f^*_{22}	f^*_{23}	$f_{2.}$
x3	f^*_{31}	f^*_{32}	f^*_{33}	$f_{3.}$
Totale col.	$f_{.1}$	$f_{.2}$	$f_{.3}$	n

L'indice Chi-quadro

3. Si confrontano le due tabelle mediante un indice sintetico

Viene calcolato un apposito indice (detto **chi-quadro**[1] o χ^2) che misura la distanza tra le frequenze osservate e quelle attese:

$$\chi^2 = \sum_i \sum_j \frac{\left(f_{ij} - f_{ij}^*\right)^2}{f_{ij}^*}$$

Questo valore va **confrontato con il valore critico** $\chi^2_{gdl;\alpha}$ che si avrebbe in caso di indipendenza ricavato dalla tabella a lato, fissato un livello di significatività ***α*** (di solito 1%, 5%), in corrispondenza ai suoi gradi di libertà: ***gdl = numero righe-1 * numero colonne-1***.

	Probabilità α	
gdl	***5%***	***1%***
1	3,84	6,63
2	5,99	9,21
3	7,81	11,34
4	9,49	13,28
5	11,07	15,09
6	12,59	16,81

Se $\chi^2 > \chi^2_{gdl;\alpha}$, oppure il ***p-value***[2] ***è inferiore al livello di significatività α,*** le differenze tra le frequenze osservate e quelle attese sono rilevanti, allora **c'è dipendenza.**

[1] Vale se il numero di osservazioni è maggiore di 100

[2] Rappresenta la probabilità che le differenze siano dovute al caso, piuttosto che a un'associazione reale tra le variabili.

Esempio – In Excel

Una banca, attraverso **un'indagine presso tre agenzie** della propria rete, ha rilevato il **principale motivo di insoddisfazione** della clientela.

Motivo principale d'insoddisfazione	Agenzia	Numero di giudizi
Costi elevati dei servizi	A	23
Costi elevati dei servizi	B	7
Costi elevati dei servizi	C	37
Mancato rispetto delle condizioni	A	39
Mancato rispetto delle condizioni	B	13
Mancato rispetto delle condizioni	C	8
Problemi concessione prestiti	A	13
Problemi concessione prestiti	B	5
Problemi concessione prestiti	C	13
Qualità del personale	A	13
Qualità del personale	B	8
Qualità del personale	C	8

Si vuole sapere se i **principali motivi d'insoddisfazione della clientela dipendono dall'agenzia**.

Esempio – In Excel

In base alla procedura descritta alla pag. 51 si sono ottenuti i seguenti risultati:

	A	B	C	D	E	F	G	H
1		**Frequenze osservate**						
2	Motivi di insoddisfazione	Agenzia A	Agenzia B	Agenzia C				
3	Costi elevati dei servizi	23	7	37	67			
4	Mancato rispetto delle condizioni	39	13	8	60			
5	Problemi concessione prestiti	13	5	13	31			
6	Qualità del personale	13	8	8	29			
7		88	33	66	187			
8								
9		**Frequenze attese in caso indipe**					Contributo celle al chi-quadro	
10	Motivi di insoddisfazione	Agenzia A	Agenzia B	Agenzia C			Agenzia A	Agenzia B	Agenzia C
11	Costi elevati dei servizi	31,5	11,8	23,6	67		2,31	1,97	7,54
12	Mancato rispetto delle condizioni	28,2	10,6	21,2	60		4,10	0,55	8,20
13	Problemi concessione prestiti	14,6	5,5	10,9	31		0,17	0,04	0,39
14	Qualità del personale	13,6	5,1	10,2	29		0,03	1,62	0,49
15		88	33	66	187				
16									
17	Indice di associazione χ^2	27,4104	=SOMMA(G11:I14)						
18	Valore critico $\chi^2_{6;0,05}$	12,5916	**=INV.CHI(B19;B20)**						
19	Livello di significatività	5%							
20	gradi di libertà	6	**=(RIGHE(B3:B6)-1)*(COLONNE(B3:D3)-1)**						
21									
22	**p-value**	**0,00012**	**=DISTRIB.CHI(B17;B20)**						
23									
24									

Agenzia A	Agenzia B	Agenzia C	
47,1%	17,6%	35,3%	100,0%
47,1%	17,6%	35,3%	100,0%
47,1%	17,6%	35,3%	100,0%
47,1%	17,6%	35,3%	100,0%

=(B7*E3)/$E$7

=(B3-B11)^2/B11

Agenzia A	**Agenzia B**	**Agenzia C**
35,8%	35,8%	35,8%
32,1%	32,1%	32,1%
16,6%	16,6%	16,6%
15,5%	15,5%	15,5%
100,0%	100,0%	100,0%

Esempio – In Excel

Dall'esempio si ottiene un indice di connessione pari a 27,4. Nel caso di indipendenza, si avrebbe un indice uguale a 12,6. Poiché **27,4>12,6** e il **p-value** (TEST.CHI) è 0,00012 **(<0,05)**, si può sostenere che **il principale motivo d'insoddisfazione della clientela dipende dall'agenzia**.

Dalla tabella delle **differenze standardizzate** (standardized residuals)[1] tra i valori osservati e quelli attesi

$$\left(f_{ij}{}^{oss} - f_{ij}{}^{att}\right) / \sqrt{f_{ij}{}^{att}(1-p_{i.})(1-p_{.j})}$$

si evince che **l'insoddisfazione**[2] dipende:

- **Dall'Agenzia A per il mancato rispetto delle condizioni**
- **Dall'Agenzia C per i costi**

	A	B	C	D	E
1		Frequenze osservate			
2	Motivi di insoddisfazione	Agenzia A	Agenzia B	Agenzia C	
3	Costi elevati dei servizi	23	7	37	67
4	Mancato rispetto delle condizioni	39	13	8	60
5	Problemi concessione prestiti	13	5	13	31
6	Qualità del personale	13	8	8	29
7		88	33	66	187
8					
9		Frequenze attese			
10	Motivi di insoddisfazione	Agenzia A	Agenzia B	Agenzia C	
11	Costi elevati dei servizi	31,5	11,8	23,6	67
12	Mancato rispetto delle condizioni	28,2	10,6	21,2	60
13	Problemi concessione prestiti	=(B3-B11)/RADQ(B11*(1-B7/$E$7)*(1-E3/$E$7))			
14	Qualità del personale	13,6	5,1	10,2	29
15		88	33	66	187
16					
17		**Differenze standardizzate**			
18	Motivi di insoddisfazione	Agenzia A	Agenzia B	Agenzia C	
19	Costi elevati dei servizi	-2,61	-1,93	4,26	
20	Mancato rispetto delle condizioni	3,38	0,99	-4,32	
21	Problemi concessione prestiti	-0,63	-0,24	0,85	
22	Qualità del personale	-0,26	1,53	-0,94	

[1] $p_{i.}$ e $p_{.j}$ sono rispettivamente le percentuali marginali di riga e di colonna.

[2] Se la differenza è positiva, vuol dire che in quella cella ci sono più soggetti di quelle previsti; se negativa, ce ne sono di meno. Questa differenza è distribuita come una normale (media=0, dev.std=1), per cui è significativa se il suo valore assoluto è superiore a 2 (1,96σ) .

Programma

- Struttura e visualizzazione dei dati
- Sintesi e interpretazione
- Relazioni tra variabili
- **Inferenza statistica**
- Verifica delle ipotesi
- Soluzioni dei test

L'errore di campionamento

Un **parametro** (una **media** o una **proporzione**) calcolato su un campione non è necessariamente rappresentativo di quello della popolazione generale di appartenenza.

Ciò che influenza la **differenza** tra il parametro (vero) della popolazione e quello del campione sono due fattori:

- la **numerosità** del campione
- la **variabilità** intrinseca del parametro

Questa discordanza viene definita **errore di campionamento**.

Per trarre allora **indicazioni sulle caratteristiche dell'intera popolazione** si utilizza la **stima** fornita dal campione tenendo conto dell'**errore di campionamento**.

L'errore di campionamento

Conoscendo **la media** o **la proporzione del campione** per la categoria in esame (ad esempio, possessori di un prodotto o clienti soddisfatti), è possibile stimare con una certa sicurezza, chiamata **livello di confidenza**[1], **la media o la proporzione dell'intera popolazione**. Questa stima sarà inclusa **nell'intervallo di confidenza**, che tiene conto dell'**errore di campionamento**.

Ad esempio, se la percentuale di clienti di una banca con mutuo nel campione in esame è del 35%, si può affermare, con un **livello di confidenza del 95%** e **un errore di campionamento del 5%**, che la **vera percentuale** di possessori di mutuo nell'intera base di clientela è compresa nell'intervallo tra il 30% e il 40%.

Le stesse considerazioni valgono anche nel caso in cui si vuole stimare **il valore medio di una popolazione**, data la variabilità del campione.

[1] Rappresenta la probabilità che la stima della media o della proporzione cada nell'intervallo dovuto all'errore; di norma, come livello, si utilizza il 95%.

Margine di errore per una proporzione

Nel caso di un campione in cui il parametro da stimare sia una **proporzione** (nel caso di variabili categoriche), la formula per il calcolo del **margine di errore** sarà:

$$e = z\sqrt{\frac{pq}{n}}$$

e, se la popolazione è conosciuta:

$$e = z\sqrt{\frac{pq}{n}}\sqrt{\frac{N-n}{N-1}}$$

dove

$\boldsymbol{e}$ = margine di errore

$\boldsymbol{N}$ = popolazione

$\boldsymbol{z}$ = coefficiente legato al livello di confidenza[1] (se 95%, è circa **1,96**)

$\boldsymbol{n}$ = dimensione del campione

$\boldsymbol{p}$ = proporzione nel campione della variabile in esame

$\boldsymbol{q = 1-p}$

Dalle formule sopra enunciate si nota che il margine di errore è **direttamente proporzionale** al **livello di confidenza** desiderato e alla **variabilità** del fenomeno studiato, mentre **è inversamente proporzionale** alla **dimensione del campione**.

[1] C'è una probabilità del 95% che la **vera c**aratteristica della popolazione sia compresa fra questi due limiti. .

Margine di errore per una proporzione – In Excel

Un campione di clienti di una banca costituito da 50 unità e 40 di loro dichiarano di possedere una carta di credito. La proporzione di possessori di carta di credito risulta quindi essere del 80% (40/50).

Applicando i calcoli, si ottiene:

MARGINE DI ERRORE			
Livello di confidenza:	95%		
z:	1,96	=INV.NORM.ST(1-(1-C5)/2)	
n:	50		
p:	80,0%		
(1-p):	20,0%		
Margine di errore:	11,1%	=B5*RADQ((B7*B8)/B6)	80 ± 11,1 (68,9 - 91,1)

In conclusione si può affermare, con un livello di confidenza del 95%, che la vera percentuale dei possessori di carta di credito di tutti i clienti della banca è compresa fra il 69% (80-11) e il 91% (80+11).

Margine di errore per una media

Nel caso di un campione in cui il parametro da stimare sia una **media**, come nel caso di variabili numeriche, la formula per il calcolo del **margine di errore** sarà:

$$e = z\,\frac{\sigma}{\sqrt{n}}$$

e, se la popolazione è conosciuta:

$$e = z\,\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

dove

e = margine di errore

N = popolazione

z / t[1] = coefficiente legato al livello di confidenza (per σ conosciuta**,** se 95% è **1,96**)

n = ampiezza del campione

σ = deviazione standard della popolazione (se conosciuto), oppure del campione

[1] Se ***n < 30*** **oppure** ***σ*** **sconosciuto** per il calcolo del coefficiente si deve usare la distribuzione ***t di Student con (n-1) gradi di libertà*****.**

Margine di errore per una media – In Excel

Un campione di clienti di una Banca costituito da 200 unità ha una media dei depositi di 12.500€ con una deviazione standard di 3.750.

Applicando i calcoli in Excel, si ottiene:

	A	B	C	D	E
1	MARGINE DI ERRORE				
2					
3	*N:*				
4	*t:*	**1,97**	**=INV.T.2T((1-E5);C5-1)**	corrispondente ad un livello di confidenza del	**95%**
5	*n:*	**200,0**			
6	σ:	**3750,0**			
7	ẋ	**12500,0**			
8					
9	**Margine di errore:**	**522,89**	**12500 ± 522,894 (11977 ÷ 13023)**		
10					

In conclusione si può affermare, con un livello di confidenza del 95%, che la vera media dei depositi dei clienti di una banca è compresa fra il 11.977 e 13.023€.

[1] Essendo ***σ* sconosciuto** per il calcolo del coefficiente si è usata la ***t di Student con (n-1) gradi di libertà*.**

Dimensione del campione

Una **rilevazione campionaria** richiede **meno tempo** di un censimento, risulta molto **meno onerosa** e portare a una **maggior qualità** della misurazione.

Un **campione troppo piccolo** scelto in base alla necessità di contenere costi e tempi di un'indagine, può però portare un **elevato errore** di campionamento e la stima non essere del tutto attendibile.

La scelta dell'ampiezza del campione dipende dal margine di errore ritenuto accettabile, e dal livello di confidenza desiderato.

Dimensione del campione per una proporzione

Si può calcolare l'ampiezza del campione, in cui il parametro da stimare sia una proporzione, sostituendo alla e della formula a pag. 59 l'errore che si è disposti ad accettare e risolvendo la suddetta equazione rispetto a $\boldsymbol{n}$.

$$n = \frac{z^2 pq}{e^2}$$

e, se la popolazione è finita:

$$n_{corr.} = \frac{nN}{n + (N - 1)}$$

dove

N = popolazione (se conosciuta)

e = margine di errore massimo

z = coefficiente legato al livello di confidenza (se 95% è 1,96)

p = proporzione nel campione della variabile in esame

q = 1-p

Dimensione del campione per una proporzione – In Excel

Una Banca vuol fare un sondaggio per sapere quanti clienti sono soddisfatti dei servizi offerti, con un **margine di errore massimo pari al ±5% e con un livello di confidenza del 95%.** Quale sarebbe la dimensione campionaria necessaria?

Applicando i calcoli[1] in Excel, il campione richiesto dovrebbe essere almeno di **384 intervistati**.

1	NUMEROSITA' DEL CAMPIONE		
2	*N:*		
3	*z:*	1,96	corrispondente ad un livello di confidenza del **95%**
4	*e:*	**5,0%**	
5	*p:*	**50,0%**	
6	*(1-p):*	50,0%	
7	***n*** *:*	**384,1**	

Se il sondaggio riguardasse invece solo un'agenzia, che in **totale ha 1.500 clienti**, quale sarebbe il campione da intervistare?

In questo caso il campione dovrebbe essere di **306 intervistati**.

	A	B	C	D	E
1	NUMEROSITA' DEL CAMPIONE				
2	*N:*	**1.500**			
3	*z:*	1,96	corrispondente ad un livello di confidenza del **95%**		
4	*e:*	**5,0%**			
5	*p:*	**50,0%**			
6	*(1-p):*	50,0%			
7	***n*** *:*	**306,0**			

[1] In questi esempi **p = 50%:** tale valore corrisponde alla situazione di massima variabilità delle stime (caso più sfavorevole).

Dimensione del campione per una media

Si può calcolare l'ampiezza del campione, in cui il parametro da stimare sia una media, sostituendo alla e della formula a pag. 61 l'errore che si è disposti ad accettare e risolvendo la suddetta equazione rispetto a $\boldsymbol{n}$.

$$n = \frac{z^2\sigma^2}{\mathrm{e}^2}$$

e, se la popolazione è finita:

$$n_{corr.} = \frac{nN}{n + (N\ - 1)}$$

dove

N = popolazione (se conosciuta)

e = errore di campionamento

z = coefficiente legato al livello di confidenza (per σ conosciuta, se 95% è 1,96)

$\boldsymbol{\sigma}$ = deviazione standard della popolazione

Dimensione del campione per una media – In Excel

Si vuole stimare, con un intervallo di confidenza del 95%, il reddito medio annuo di un segmento di clientela di una banca con un errore massimo di 1.000€. Si suppone che il range delle entrate non sia superiore a 30.000€ e che, usando la regola empirica, la **deviazione standard della popolazione σ** sia 1/6, ovvero 5.000€. Qual è la dimensione minima del campione richiesta?

Applicando i calcoli in Excel, il campione richiesto dovrebbe essere di 96 intervistati.

	A	B	C	D	E
1	NUMEROSITA' DEL CAMPIONE				
2	*N:*				
3	*z:*	1,96	corrispondente ad un livello di confidenza del		**95%**
4	*e:*	**1.000,00**			
5	σ	**5.000,00**			
7	***n*** *:*	**96,0**			

Se il sondaggio riguardasse una filiale di 250 clienti, quale sarebbe il campione da intervistare?

In questo caso il campione dovrebbe essere di circa 70 intervistati.

	A	B	C	D	E
1	NUMEROSITA' DEL CAMPIONE				
2	*N:*	**250**			
3	*z:*	1,96	corrispondente ad un livello di confidenza del		**95%**
4	*e:*	**1.000,00**			
5	σ	**5.000,00**			
7	***n*** *:*	**69,6**			

Soluzioni dei test

Inferenza statistica – Test intermedio

1. **In un campione di 80 clienti intervistati, il 78% dichiara di essere soddisfatto dei servizi offerti. Qual è l'errore campionario che si avrebbe con un livello di confidenza del 95%?**

2. **Quale dovrà essere la dimensione del campione di clienti da intervistare per il quale l'errore di campionamento non deve superare il 10%, con il livello di confidenza 95%?**

3. **Perché nel secondo esercizio si ottiene, con un errore di campionamento maggiore, un campione più numeroso del primo?**

Programma

- Struttura e visualizzazione dei dati
- Sintesi e interpretazione
- Relazioni tra variabili
- Inferenza statistica
- **Verifica delle ipotesi**
- Soluzioni dei test

I test di verifica di ipotesi

I test di verifica d'ipotesi si utilizzano per verificare la bontà di un'ipotesi, che si presta ad essere confermata o smentita dai dati osservati sperimentalmente.

Per una comprensione completa e approfondita dei concetti con applicazioni a dati e problemi reali, si rimanda alla pubblicazione:

https://www.alfredoroccato.it/corso-test-verifica-ipotesi/

Programma

- **Struttura e visualizzazione dei dati**
- Sintesi e interpretazione
- Relazioni tra variabili
- Inferenza statistica
- Verifica delle ipotesi
- Soluzioni dei test

Soluzioni dei test

Classificazione delle variabili e distribuzione – Test intermedioxxx

1. Il saldo di un conto corrente è una variabile di tipo:

- ❑ Quantitativa discreta
- ☒ Quantitativa continua

2. La variabile che assume i valori "3. >100.000", "1. 0-49.999", "2. 50.000-100.000" è di tipo:

- ☒ Qualitativa ordinale
- ❑ Qualitativa nominale

3. La frequenza relativa (in percentuale) dei numeri maggiori di 25 nell'insieme 35, 19, 18, 25, 42, 33, 22, 44 è :

- ❑ 62,5%
- ☒ 50%

Soluzioni dei test

Indici di posizione – Test intermedio

1. **Nell'insieme dei numeri 3,5,2,1,4 la media e la mediana sono rispettivamente:**

   ❑ 3; 2

   ☒ 3; 3

2. **Se un insieme di 4 valori con media 30 si aggiunge il valore 80, quale sarà la nuova media?**

   ☒ 40

   ❑ 55

3. **Il 20° percentile dell'insieme dei numeri al punto 1 corrisponde al valore (calcolato con Excel):**

   ❑ 5

   ☒ 1,8

Soluzioni dei test

Indici di variabilità – Test intermedio

1. **Nell'insieme di numeri 35, 30, 33, 34, 30, 32 qual è la DEVIAZIONE STANDARD CAMPIONARIA?**

   ❑ 1,88

   ☒ 2,06 (calcolata in Excel con DEV.ST()).

2. **In base al COEFFICIENTE DI VARIAZIONE, la media dei dati precedenti è un buon indicatore? Perché?**

   ☒ Sì, è inferiore al 10% (6,38%)

   ❑ No

3. **Un secondo insieme ha media 67,3 e deviazione standard campionaria 3,41, qual è in base al COEFFICIENTE DI VARIAZIONE l'insieme che ha minor variabilità?**

   ❑ Il primo (6,38%)

   ☒ Il secondo (5,06%)

Soluzioni dei test

Correlazione e Regressione – Test intermedio

1. **Il coefficiente di correlazione delle variabili X= 20; 35; 43; 38; 51; 46; 48; 53; e Y= 430; 370; 340; 300; 225; 230; 180; 210; è**

   ☒ -0.905

   ❑ 0,905

2. **Se Y fosse la variabile dipendente e X quella indipendente in un modello di regressione lineare, il valore di intercetta e di pendenza sarebbe:**

   ☒ intercetta: 597,6; pendenza: -7,47

   ❑ Intercetta: 73; pendenza: -0,11

3. **Dal valore di R^2 si può dire che le stime del modello sono buone?**

   ☒ Sì, è superiore di 0,8

   ❑ No

4. **Ipotizzando un valore di X pari a 50, quale sarà il valore previsto di Y?**

   ❑ 67,6

   ☒ 224

Inferenza statistica – Test intermedio

1. **In un campione di 80 clienti intervistati, il 78% dichiara di essere soddisfatto dei servizi offerti. Qual è l'errore campionario che si avrebbe con un livello di confidenza del 95%?**

   L'errore di campionamento è del 9% per cui si può affermare, con un livello di confidenza del 95%, che la percentuale dei clienti soddisfatti è compresa fra il 69% e l'87%

2. **Quale dovrà essere la dimensione del campione di clienti da intervistare per il quale l'errore di campionamento non deve superare il 10%, con il livello di confidenza 95%?**

   Il campione dovrà essere almeno di 96 clienti.

3. **Perché nel secondo esercizio si ottiene, con un errore di campionamento maggiore, un campione più numeroso del primo?**

   *Perché, nel secondo esercizio, la variabilità 0,5*0,5=0,25 è più elevata del primo 0,78*0,22=0,17.*

L'esperienza è il miglior maestro

Contatti

e-mail: alfredo.roccato(at)fastwebnet.it
www.alfredoroccato.it