Articolazione degli argomenti

- 2h, Concetti di base della statistica descrittiva, materiale: slide "introduzione.pdf"
- 4h (1 ora oggi e 3 nel prossimo incontro), Presentazione di un esempio di report e introduzione all'uso di excel come strumento di analisi statistica, materiale: report "esempio report.pdf" e manualino su excel "vademecum.pdf"

Indice

s2 ss1 toc

Utilità della statistica

Viviamo nella società dell'**informazione**

Per qualsiasi fenomeno è sempre più facile ottenere **grandi quantità di dati.**

La società dell'Utilità della statistica

Viviamo nella società dell'informazione... 1/2

Utilità della statistica

Viviamo nella società dell'informazione... 2/2

Utilità della statistica

La statistica è uno strumento

Tuttavia i dati non sempre migliorano la **conoscenza** di un fenomeno o sono di aiuto nel prendere **decisioni**. Esiste un problema di **verifica** e di **interpretazione dei dati**.

La Statistica rappresenta uno **strumento potente** per la soluzione di questi problemi.

Utilità della statistica

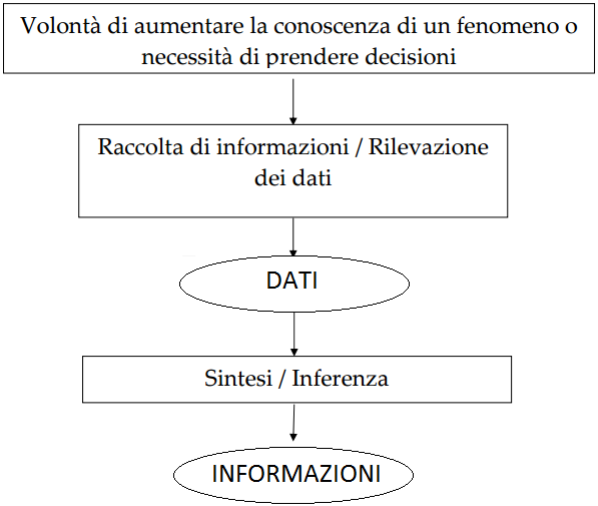
Alcune domande fondamentali

- Come sintetizzare grandi masse di dati? Come estrarre da un insieme apparentemente caotico di dati le informazioni davvero interessanti, quelle utili a prendere decisioni?

Questa domanda corrisponde al grande capitolo della statistica noto come **statistica descrittiva**

Utilità della statistica

Dai dati alle informazioni



s2 ss3 toc

Popolazione o campione?

Indagine statistica

"Oggetto di ogni indagine statistica è la conoscenza di una ***popolazione****, intesa come insieme, come aggregato, di* ***unità elementari*** *in cui il fenomeno allo studio si manifesta.*

. . .

Le informazioni attorno alla popolazione, ossia attorno alle variabili che la caratterizzano, possono essere il frutto di una rilevazione ***totale*** *o* ***censuaria****, oppure di una rilevazione* ***campionaria****."*

(Cicchitelli, Hertzel, Montanari, 1997)

Popolazione o campione?

Spesso, non è possibile studiare tutte le unità statistiche della popolazione (**censimento**) e perciò si procede a rilevare i caratteri oggetto di studio su un suo sottoinsieme (**campione**).

I **vantaggi** delle indagini a campione sono i seguenti:

- sono indispensabili nelle rilevazioni in popolazioni infinite (numero indefinito di elementi, *pezzi che una macchina può produrre, membri di una certa specie, l'insieme dei risultati prodotti dalla sperimentazione di un farmaco)*
- si riducono i costi
- si riducono i tempi di elaborazione dei dati
- si riducono gli errori di rilevazione dei caratteri

Gli **svantaggi** delle indagini a campione sono:

- l'indagine censuaria restituisce il valore vero dei parametri di interesse (proporzioni, percentuali, medie, totali, ...) l'indagine campionaria ne fornisce solo una stima. Mediante i dati rilevati è possibile ottenere un valore approssimato dei parametri, al quale è associabile un grado di fiducia quantificabile se e solo se la formazione del campione risponde a criteri di tipo probabilistico.

s2 ss2 toc

Gli elementi della popolazione sono in genere denominati **unità statistiche**.

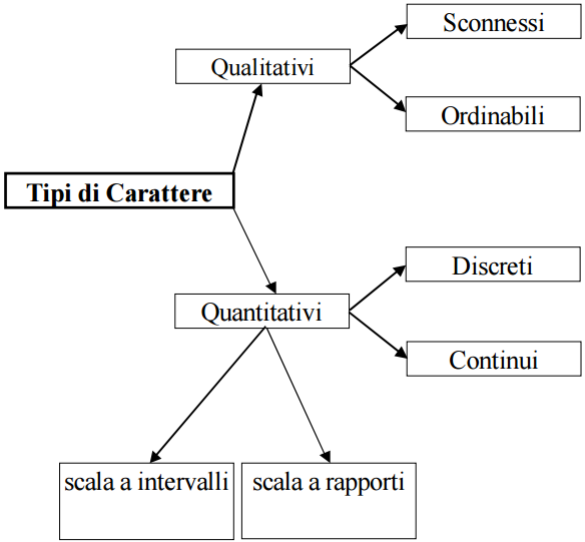
Variabile (carattere) è il fenomeno oggetto di studio rilevato o misurato sulle unità statistiche.

- **Deterministica:** assume valori che sono sotto controllo del ricercatore;
  es. l'ammontare di denaro in un conto corrente, se conosco l'ammontare iniziale e il tasso d'interesse annuo, si può determinare il capitale alla fine del periodo d'interesse.
- **Aleatoria:** è una variabile che può assumere valori diversi in dipendenza da qualche fenomeno aleatorio.
  es. il risultato di un lancio di un dado

I valori distinti assunti da una variabile sono detti **modalità**.

I dati

Le tipologie



I dati

Cross section e Time series

codice	DATA DI NASCITA	SESSO	2015-scuola descrizione	2015-SCUOLA COMUNE	2015_CLASSE	2015_INDIRIZZO	2015_SEZIONE	2015-voto ITALIANO	2015- voto MATEMATICA	2015-voto INGLESE
2273589	12/10/1998	M	I.T.I.G. "O. BELLUZZI - L. DA VINCI"	RIMINI	4	TECNICI TECNOLOGICO MECCANICA E MECCATRONICA	MME	9	.	8
2852835	19/09/1997	F	I.P.S.C.T. "L. EINAUDI" (vecchio codice)	RIMINI	3	PROFESSIONALI SERVIZI PROMOZIONE COMMERCIALE E PUBBLICITARIA - OPZIONE	D	.	.	.
2852919	20/03/1997	M	LICEO "G.CESARE - M.VALGIMIGLI"	RIMINI	4	LICEI SCIENZE UMANE SCIENZE UMANE - OPZ. ECONOMICO SOCIALE	BS	8	8	.
2852950	29/05/1996	M	I.P.S.C.T. "L. EINAUDI" (vecchio codice)	RIMINI	4	PROFESSIONALI SERVIZI PROMOZIONE COMMERCIALE E PUBBLICITARIA - OPZIONE	D	.	.	.
2880331	28/09/1996	M	NOVAFELTRIA POLO I.P.I.A. G. BENELLI	NOVAFELTRIA	4	PROFESSIONALI INDUSTRIA E ARTIGIANATO INDUSTRIA - TRIENNIO	P	7	7	6
3346126	21/07/2001	M	LICEO "A. EINSTEIN"	RIMINI	1	LICEI SCIENTIFICO SCIENTIFICO	C	5.5	5.5	7

I dati

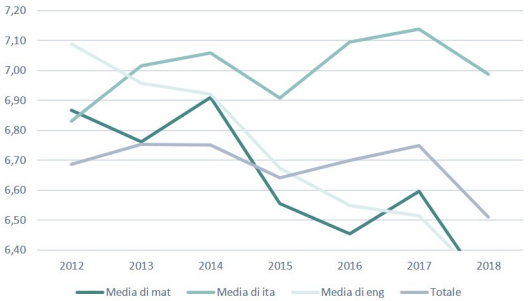
Cross section e Time series

VARIABILI	2012	2013	2014	2015	2016	2017	2018
Media di mat	6.87	6.76	6.91	6.56	6.46	6.60	6.26
Media di ita	6.83	7.02	7.06	6.91	7.10	7.14	6.99
Media di eng	7.09	6.96	6.92	6.67	6.55	6.51	6.28
Totale	**6.69**	**6.75**	**6.75**	**6.64**	**6.70**	**6.75**	**6.51**

I dati

Cross section e Time series

VARIABILI	2012	2013	2014	2015	2016	2017	2018
Media di mat	6.87	6.76	6.91	6.56	6.46	6.60	6.26
Media di ita	6.83	7.02	7.06	6.91	7.10	7.14	6.99
Media di eng	7.09	6.96	6.92	6.67	6.55	6.51	6.28
Totale	**6.69**	**6.75**	**6.75**	**6.64**	**6.70**	**6.75**	**6.51**



L' unità statistica è il **tempo** (anno, trimestre, mese, giorno, ora) e rileviamo il fenomeno su tempi diversi:

$$x_1, x_2, ..., x_{t-1}, x_t, x_{t+1}, ..., x_n \tag{1}$$

L'osservazione nel tempo di un certo fenomeno permette di costruire una **serie storica**. Una serie storica può essere rappresentata su un grafico cartesiano

- in ascissa: dimensione temporale;
- in ordinata: carattere (successione di punti o spezzata)

Se sottraiamo al valore x_t quello precedente x_{t-1} e lo rapportiamo sempre per quello precedente otteniamo un **tasso di variazione relativo** (%).
La serie di tassi che ne risulta non dipende dall'unità di misura né dall'ordine di grandezza del fenomeno: si possono quindi fare confronti!

I dati

Tasso di variazione relativo - Esempio

Tasso di variazione relativo	2013 su 2012	2014 su 2013	2015 su 2014	2016 su 2015	2017 su 2016	2018 su 2017
Media di mat	*-1.5%*	*2.2%*	*-5.1%*	*-1.5%*	*2.2%*	*-5.1%*
Media di ita	*2.7%*	*0.6%*	*-2.1%*	*2.7%*	*0.6%*	*-2.1%*
Media di eng	*-1.9%*	*-0.5%*	*-3.6%*	*-1.9%*	*-0.5%*	*-3.6%*
Totale	***1.0%***	***0.0%***	***-1.6%***	***0.9%***	***0.7%***	***-3.6%***

Mentre i voti medi in matematica ed italiano mostrano un andamento "altalenante" nel corso del tempo, il voto medio in inglese mostra un trend costantemente decrescente nel corso del tempo, seppur con entità diverse.

La comunicazione dei risultati

Il livello di informazione

Obiettivo: Rendere comprensibili al pubblico i risultati ottenuti nella fase di analisi. In tal modo si forniscono strumenti utili alle persone che devono prendere una decisione che produca effetti sulla popolazione osservata. É un ponte tra la scienza e le persone.

La comunicazione è molto importante perchè gli esperti e il pubblico hanno differenti modalità di PERCEZIONE delle informazioni, e molto spesso hanno bisogno di informazioni diverse per decidere.

In teoria come rappresentiamo un certo risultato (ad es. una relazione tra variabili, una criticità in una particolare fetta di popolazione...) dipende dalla conoscenza della situazione, cioè dalle informazioni di cui si è in possesso.

La comunicazione dei risultati

La rappresentazione

Tuttavia, in pratica, la rappresentazione che di solito viene usata, dipende non tanto da quanto si è riusciti a capire dallo studio del fenomeno ma dal messaggio che si vuole trasmettere...

Trova le differenze (prendiamo un esempio molto delicato)

Il vostro dottore, dopo aver esaminato attentamente le vostre ultimi analisi del sangue, afferma...

"Se da oggi inizi ad assumere la *Inegy* le tue probabilità di avere un infarto entro 10 anni si riducono dal 10% all 8%"

"Se da oggi inizi ad assumere il *Lipex* le tue chance di evitare un infarto nei prossimi 10 crescono dal 90% al 92%"

La comunicazione dei risultati

La rappresentazione

Tuttavia, in pratica, la rappresentazione che di solito viene usata, dipende non tanto da quanto si è riusciti a capire dallo studio del fenomeno ma dal messaggio che si vuole trasmettere...

Trova le differenze (prendiamo un esempio molto delicato)

Il vostro dottore, dopo aver esaminato attentamente le vostre ultimi analisi del sangue, afferma...

"Se da oggi inizi ad assumere la *Inegy* le tue probabilità di avere un infarto entro 10 anni si riducono dal 10% all 8%"

"Se da oggi inizi ad assumere il *Lipex* le tue chance di evitare un infarto nei prossimi 10 crescono dal 90% al 92%"

Leggete attentamente le affermazioni, quale dei due medicinali prendereste per avere più probabilità di evitare un infarto nei prossimi 10 anni?

La comunicazione dei risultati

La rappresentazione

Tuttavia, in pratica, la rappresentazione che di solito viene usata, dipende non tanto da quanto si è riusciti a capire dallo studio del fenomeno ma dal messaggio che si vuole trasmettere...

Trova le differenze (prendiamo un esempio molto delicato)

Il vostro dottore, dopo aver esaminato attentamente le vostre ultimi analisi del sangue, afferma...

"Se da oggi inizi ad assumere la *Inegy* le tue probabilità di avere un infarto entro 10 anni si riducono dal 10% all 8%"

"Se da oggi inizi ad assumere il *Lipex* le tue chance di evitare un infarto nei prossimi 10 crescono dal 90% al 92%"

Leggete attentamente le affermazioni, quale dei due medicinali prendereste per avere più probabilità di evitare un infarto nei prossimi 10 anni?

In realtà le due affermazioni hanno lo stesso significato! (Inegy e Lipex sono entrambi nomi commerciali della Simvastatina), la forma in questo caso viene usata per far passare un messaggio diverso.

La comunicazione dei risultati

La rappresentazione (Continua...)

E ancora...

- Ho il 10% di chance di avere un infarto tra 10 anni;
- Il 10% della popolazione con risultati degli esami del sangue come i miei avrà un infarto nei prossimi 10 anni;
- Il 10% degli alternativi mondi futuri includerà me che ho un infarto.

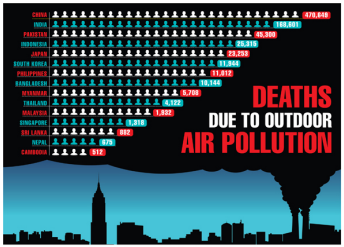
Hanno tutte lo stesso significato ma danno un'impressione diversa del livello di rischio.

La comunicazione dei risultati

La rappresentazione (continua...)

Oltre alla sintassi, il ricercatore deve tener conto della rappresentazione visiva dei contenuti.
Non sempre una percentuale può rendere l'idea o semplicemente può non essere compresa dai destinatari dei risultati.

Si possono produrre grafici come *grafici a barre*, *torte*, *grafici radar* per semplificare la lettura dei risultati.



(SOURCE: WORLD HEALTH ORGANISATION REPORT 2008)

As esempio a seconda del carattere, esiste una appropriata rappresentazione grafica della distribuzione di frequenza:

- Se il carattere è qualitativo sconnesso, il grafico appropriato è il diagramma a torta;
- Se il carattere è qualitativo ordinato o quantitativo discreto, il grafico appropriato è il diagramma a barre verticali;
- Se il carattere è quantitativo continuo, il grafico appropriato è l'istogramma (nel nostro dataset non sono presenti caratteri continui quindi non tratteremo questo argomento).

Vedremo esempi di queste rappresentazioni grafiche nelle slide successive e nella prossima lezione.

s2 ss4 toc

Distribuzioni di frequenza univariate

Variabili qualitative - Frequenze assolute

La distribuzione di frequenza si costruisce raggruppando in classi le n unità statistiche secondo le K modalità del carattere. Una classe raggruppa una o più modalità del carattere.
La k-esima classe conterrà n_k unità statistiche (caratterizzate dalle modalità appartenenti alla classe k).

Distribuzioni di frequenza univariate

Variabili qualitative - Frequenze assolute

La distribuzione di frequenza si costruisce raggruppando in classi le n unità statistiche secondo le K modalità del carattere. Una classe raggruppa una o più modalità del carattere.
La k-esima classe conterrà n_k unità statistiche (caratterizzate dalle modalità appartenenti alla classe k).

Esempio

INDIRIZZO	N
IND_W (solo triennio)	194
IND_V (solo triennio)	74
IND_U (solo biennio)	208
IND_Z	60
IND_Y	321
IND_X	434
Totale	**1291**

Distribuzioni di frequenza univariate

Variabili qualitative - Frequenze percentuali

Possiamo alternativamente esprimere le frequenze come percentuale p_k del numero totale di osservazioni ottenuta dividendo ogni n_k per il numero di unita statistiche osservate n ($= \sum_k n_k$ se non ci sono dati mancanti).

Esempio

INDIRIZZO	N	%
IND_W (solo triennio)	194	*15.0%*
IND_V (solo triennio)	74	*5.7%*
IND_U (solo biennio)	208	*16.1%*
IND_Z	60	*4.6%*
IND_Y	321	*24.9%*
IND_X	434	*33.6%*
Totale	**1291**	***100.0%***

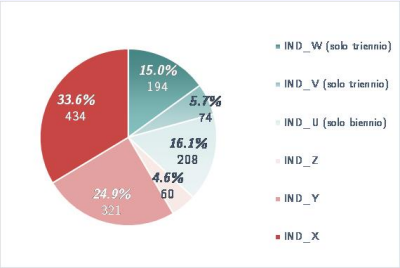
Distribuzioni di frequenza univariate

Variabili qualitative - Frequenze percentuali

Possiamo alternativamente esprimere le frequenze come percentuale p_k del numero totale di osservazioni ottenuta dividendo ogni n_k per il numero di unita statistiche osservate n ($= \sum_k n_k$ se non ci sono dati mancanti).

Esempio

INDIRIZZO	N	%
IND_W (solo triennio)	194	*15.0%*
IND_V (solo triennio)	74	*5.7%*
IND_U (solo biennio)	208	*16.1%*
IND_Z	60	*4.6%*
IND_Y	321	*24.9%*
IND_X	434	*33.6%*
Totale	**1291**	***100.0%***



Distribuzioni di frequenza univariate

Frequenze cumulate

Dato un carattere con K modalità, ORDINABILI in senso crescente, si indica con

$N_k = n_1 + n_2 + \ldots + n_k$ la frequenza assoluta[1] cumulata della modalità k;
$F_k = f_1 + f_2 + \ldots + f_k$ la f. relativa[1] cumulata;
$P_k = p_1 + p_2 + \ldots + p_k$ la f. percentuale[1] cumulata.

Informalmente la distribuzione di frequenza cumulata risponde alla domanda: quante unità del collettivo hanno un valore al massimo uguale a un certo x_0?

[0] n.b. La distribuzione delle frequenze cumulate è definita pertanto solo per caratteri quantitativi e per caratteri qualitativi ordinabili. Si basa infatti sul concetto di ordinamento che non ha senso per i caratteri qualitativi sconnessi.

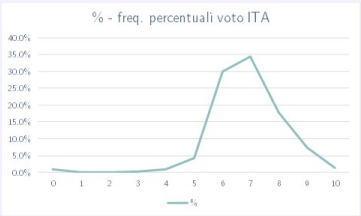
Distribuzioni di frequenza univariate

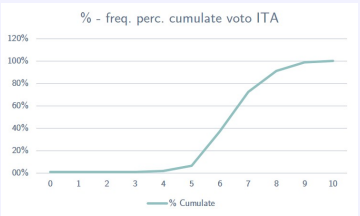
Frequenze cumulate - Esempio

Esempio

Variabile: Voto in italiano, 11 modalità (attenzione allo 0, ci deve essere qualche problema nel dataset, in questi casi le distribuzioni di frequenza possono essere utili anche per individuare errori nei dati).

voto ITA	N	%	% Cumulate
0	11	*0.9%*	0.9%
1	0	*0.0%*	0.9%
2	0	*0.0%*	0.9%
3	2	*0.2%*	1.0%
4	11	*0.9%*	1.9%
5	56	*4.3%*	6.2%
6	387	*30.0%*	36.2%
7	443	*34.3%*	70.5%
8	230	*17.8%*	88.3%
9	96	*7.4%*	95.7%
10	17	*1.3%*	97.1%
	38	*2.9%*	*100.0%*
Totale	**1291**	**100%**	





s2 ss6 toc

Indicatori di sintesi

Intro

La sintesi numerica di una distribuzione statistica è basata sulla costruzione di particolari **indici numerici** che delineano alcuni aspetti essenziali della distribuzione in esame. Questi indici consentono un confronto tra le caratteristiche di distribuzioni diverse.

Possiamo individuare tre famiglie principali di indici:

- indici di tendenza centrale o di posizione
- indici di variabilità o dispersione
- indici di forma (Non li faremo qui)

Alla sintesi numerica si chiede di evidenziare gli aspetti principali di una distribuzione, tenendo conto che tutte le volte che si sintetizzano più dati con un solo valore, **si perdono delle informazioni**.
L'obiettivo dei metodi che vedremo è quella di rendere più possibilmente oggettiva questa sintesi.

I **valori medi** sono strumenti di sintesi che descrivono l'ordine di grandezza del carattere nell'insieme delle unità osservate.

La **media aritmetica** di un insieme di n valori $x_1, x_2, ..., x_n$ di un carattere quantitativo X è pari alla somma dei valori divisa per la loro numerosità ossia risulta dalla **equiripartizione** dell'ammontare complessivo (totale) del carattere fra le unità osservate. Pertanto, la media aritmetica di n osservazioni del carattere X è:

$$\bar{x} = M_1(X) = \frac{\sum_{x=1}^{n} x_i}{n} \tag{2}$$

Quando usare la media aritmetica?

1. Quando le modalità del carattere possono essere pensate come la redistribuzione di un unico ammontare all'interno del collettivo. *(il calcolo di una misura della dimensione media dei capitoli di spesa in un bilancio)*
2. Quando i valori osservati del fenomeno possono essere pensati come approssimazioni di un unico "valore vero". *(Tentativi di misurazione di un fenomeno)*

Quando usare la media aritmetica?

1. Quando le modalità del carattere possono essere pensate come la redistribuzione di un unico ammontare all'interno del collettivo. *(il calcolo di una misura della dimensione media dei capitoli di spesa in un bilancio)*
2. Quando i valori osservati del fenomeno possono essere pensati come approssimazioni di un unico "valore vero". *(Tentativi di misurazione di un fenomeno)*

I difetti della media aritmetica

La media aritmetica può essere molto **sensibile alla presenza di osservazioni anomale**. In queste circostanze la media aritmetica assumerà un valore che è ancora un punto di equilibrio per la somma delle modalità, ma non è più un valore equamente rappresentativo di tutte le osservazioni.

La **moda** è la modalità che nell'insieme delle osservazioni si presenta con la frequenza più alta.
È definita per qualsiasi tipo di carattere (qualitativo o quantitativo), ma può accadere che non identifichi un valore unico (distribuzioni pluri-modali) o che non esista affatto.

Questa definizione non è tuttavia operativa per caratteri quantitativi continui, nei quali accade sovente che ogni modalità ha frequenza $1/n$.
Per tali caratteri ha senso parlare di **classe modale**, ossia la classe cui corrisponde la massima frequenza (ATTENZIONE: Se i dati sono raggruppati in classi di ampiezza disuguale il calcolo è leggermente più elaborato. La moda dovrà essere definita come la classe di modalità con massima **densità di frequenza** $h_k = f_k/a_k = \frac{f_k}{x_k - x_{k-1}}$).

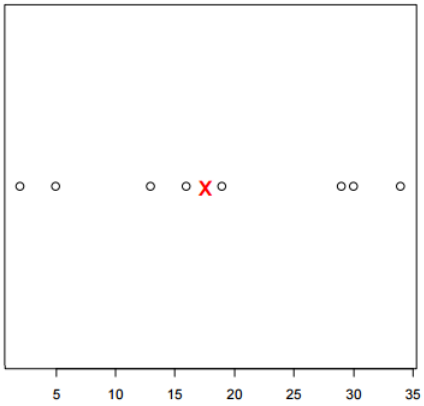
Indicatori di sintesi

Valori medi - la mediana

La **mediana** è la modalità che occupa il posto centrale nella successione ordinata delle n osservazioni.

La mediana bipartisce le osservazioni in modo che le osservazioni maggiori e quelle minori della mediana siano nello stesso numero.

La mediana è calcolabile per qualsiasi carattere ordinabile (ma non necessariamente quantitativo).



Indicatori di sintesi

Valori medi - la mediana

La mediana è poco sensibile (in statistica si dice "robusta") alla presenza di pochi outliers.

Nei casi in cui poche unità hanno valori molto più grandi (o molto più piccoli) della maggioranza delle altre, la mediana è un indicatore di posizione più sensato e "equo" della media aritmetica.

Indicatori di sintesi

Valori medi - la mediana

La mediana è poco sensibile (in statistica si dice "robusta") alla presenza di pochi outliers.
Nei casi in cui poche unità hanno valori molto più grandi (o molto più piccoli) della maggioranza delle altre, la mediana è un indicatore di posizione più sensato e "equo" della media aritmetica.

Allora perchè non usiamo sempre la mediana?

Intanto abbiamo detto che ha un significato diverso dalla media, ma il problema principale sta nella sua determinazione, la media aritmetica si calcola tramite operazioni algebriche sui dati mentre sia moda che mediana sono frutto di un algoritmo più complesso e non sempre analitico.

Indicatori di sintesi

Valori medi - la mediana

La mediana è poco sensibile (in statistica si dice "robusta") alla presenza di pochi outliers.
Nei casi in cui poche unità hanno valori molto più grandi (o molto più piccoli) della maggioranza delle altre, la mediana è un indicatore di posizione più sensato e "equo" della media aritmetica.
Allora perchè non usiamo sempre la mediana?
Intanto abbiamo detto che ha un significato diverso dalla media, ma il problema principale sta nella sua determinazione, la media aritmetica si calcola tramite operazioni algebriche sui dati mentre sia moda che mediana sono frutto di un algoritmo più complesso e non sempre analitico.

Se si vuole calcolare la mediana sulla distribuzione di frequenze di un carattere discreto (oppure sulla distribuzione di frequenze di un carattere qualitativo ordinabile), è utile fare riferimento alle **frequenze relative cumulate**.
Infatti se si individua la modalità x_{i-1} cui corrisponde una frequenza relativa cumulata < 0.5 e tale che per la modalità successiva x_i sia abbia una frequenza cumulata ≥ 0.5, allora la mediana coincide con la modalità x_i Se le modalità sono raggruppate in classi, seguendo lo stesso criterio si individua la **classe mediana**.

Misure di variabilità

Esempio

Qual'è la mediana dei voti ottenuti in Italiano dai nostri studenti?

voto ITA	N	%	% Cumulate
0	11	*0.9%*	0.9%
1	0	*0.0%*	0.9%
2	0	*0.0%*	0.9%
3	2	*0.2%*	1.0%
4	11	*0.9%*	1.9%
5	56	*4.3%*	6.2%
6	387	*30.0%*	36.2%
7	443	*34.3%*	70.5%
8	230	*17.8%*	88.3%
9	96	*7.4%*	95.7%
10	17	*1.3%*	97.1%
	38	*2.9%*	*100.0%*
Totale	**1291**	**100%**	

Misure di variabilità

Esempio

Qual'è la mediana dei voti ottenuti in Italiano dai nostri studenti?

voto ITA	N	%	% Cumulate
0	11	*0.9%*	0.9%
1	0	*0.0%*	0.9%
2	0	*0.0%*	0.9%
3	2	*0.2%*	1.0%
4	11	*0.9%*	1.9%
5	56	*4.3%*	6.2%
6	387	*30.0%*	36.2%
7	443	*34.3%*	70.5%
8	230	*17.8%*	88.3%
9	96	*7.4%*	95.7%
10	17	*1.3%*	97.1%
	38	*2.9%*	*100.0%*
Totale	**1291**	**100%**	

La mediana bipartisce la successione ordinata dei valori individuali in due sottoinsiemi di eguale numerosità. Per una distribuzione di frequenze abbiamo visto che la mediana è dunque quel valore in corrispondenza del quale la frequenza relativa cumulata è pari a 0.5.
L'idea che sta alla base della definizione di valore mediano può essere estesa.

Ad esempio, i **quartili** ripartiscono la graduatoria non decrescente in quattro gruppi, così che in un quarto dei casi l'intensità del carattere non supera il primo quartile (Q1), metà dei valori individuali sono inferiori al secondo quartile (la mediana), ed il terzo quartile (Q3) indica il livello soglia del carattere superato dal 25% delle unità.

Indicatori di sintesi

Valori medi - i quartili

Quali sono il primo e il terzo quartile dei voti ottenuti in Italiano dai nostri studenti?

voto ITA	N	%	% Cumulate
0	11	*0.9%*	0.9%
1	0	*0.0%*	0.9%
2	0	*0.0%*	0.9%
3	2	*0.2%*	1.0%
4	11	*0.9%*	1.9%
5	56	*4.3%*	6.2%
6	387	*30.0%*	36.2%
7	443	*34.3%*	70.5%
8	230	*17.8%*	88.3%
9	96	*7.4%*	95.7%
10	17	*1.3%*	97.1%
	38	*2.9%*	*100.0%*
Totale	**1291**	**100%**	

Indicatori di sintesi

Valori medi - i quartili

Quali sono il primo e il terzo quartile dei voti ottenuti in Italiano dai nostri studenti?

voto ITA	N	%	% Cumulate
0	11	*0.9%*	0.9%
1	0	*0.0%*	0.9%
2	0	*0.0%*	0.9%
3	2	*0.2%*	1.0%
4	11	*0.9%*	1.9%
5	56	*4.3%*	6.2%
6	387	*30.0%*	36.2%
7	443	*34.3%*	70.5%
8	230	*17.8%*	88.3%
9	96	*7.4%*	95.7%
10	17	*1.3%*	97.1%
	38	*2.9%*	*100.0%*
Totale	**1291**	**100%**	

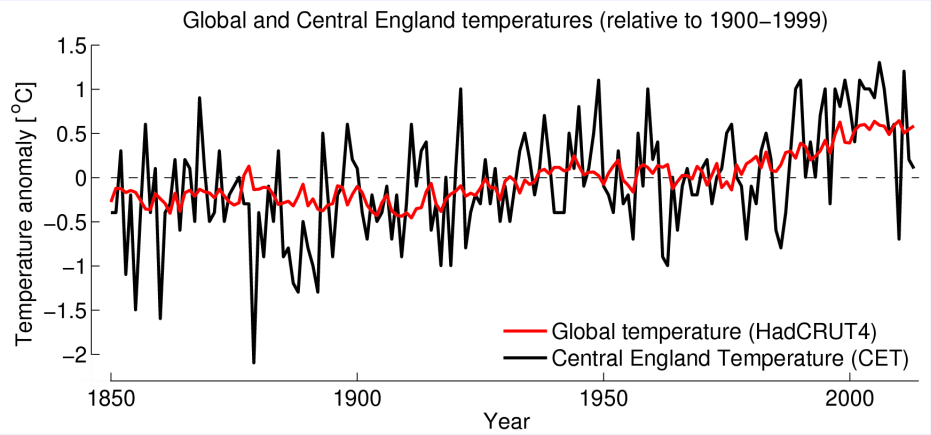
L'attitudine di un carattere quantitativo X ad assumere valori differenti tra le unità componenti un insieme statistico è chiamata **variabilità**.
Essa costituisce una caratteristica dei collettivi statistici e può essere descritta mediante indicatori che godano di particolari proprietà. Più precisamente:

1. una misura di variabilità deve annullarsi quando, e solo quando, tutte le unità del collettivo presentano il medesimo stato di grandezza del carattere;
2. una misura di variabilità deve assumere valori crescenti all'aumentare della variabilità.

ATTENZIONE: Il concetto di variabilità è strettamente legato all'indice di variabilità utilizzato e pertanto non è corretto mettere a confronto misure di variabilità ottenute con indici diversi.

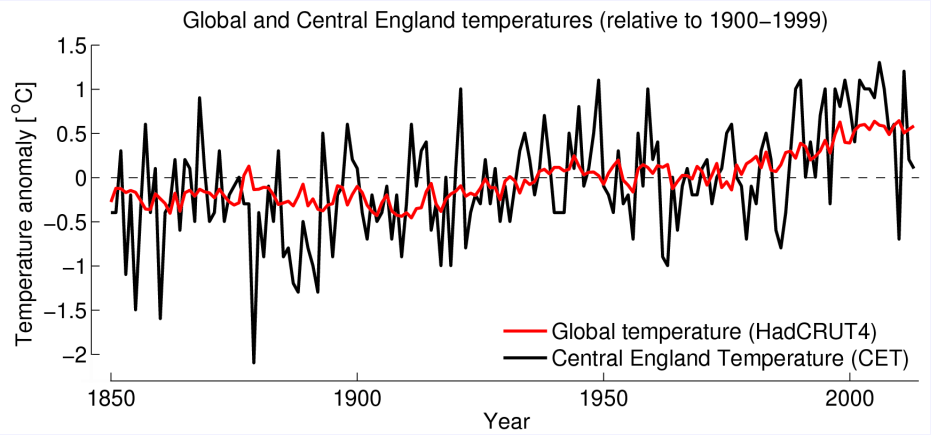
Esempio

Quale delle due serie è più **variabile**?



Esempio

Quale delle due serie è più **variabile**?



Di sicuro la serie dell'indicatore della temperatura dell'inghilterra centrale...

Possiamo distinguere quattro categorie di misure di variabilità:

- indici di variabilità basati sullo scostamento da una media;
- indici che misurano la diversità fra due particolari termini della distribuzione o fra due quantili (intervalli di variabilità);
- indici che misurano la variabilità rispetto alle frequenze relative, adatti per caratteri qualitativi, (indici di eterogeneità);
- indici che misurano la variabilità del carattere mediante una sintesi delle misure della diversità esistente fra le modalità di tutte le possibili coppie di unità (non li faremo qui).

Indicatori di sintesi

Misure di variabilità - Scostamenti dalla media

L'idea è quella di misurare la variabilità di un carattere verificando se le unità statistiche presentano modalità più o meno stabili rispetto a un indice di posizione assunto come rappresentativo della distribuzione.
Se si sceglie come indice di posizione la media aritmetica, una prima idea potrebbe essere quella di sintetizzare gli scarti dalla media definiti come

$$x_1 - \bar{x}; x_2 - \bar{x}; ...; x_n - \bar{x} \tag{3}$$

Indicatori di sintesi

Misure di variabilità - Scostamenti dalla media

L'idea è quella di misurare la variabilità di un carattere verificando se le unità statistiche presentano modalità più o meno stabili rispetto a un indice di posizione assunto come rappresentativo della distribuzione.
Se si sceglie come indice di posizione la media aritmetica, una prima idea potrebbe essere quella di sintetizzare gli scarti dalla media definiti come

$$x_1 - \bar{x}; x_2 - \bar{x}; ...; x_n - \bar{x} \tag{3}$$

Non è una buona idea in quanto la sintesi più semplice e intuitiva, ovvero la media aritmetica degli scarti, sarebbe sempre uguale a zero...
Quindi, come potremmo agire?

Indicatori di sintesi

Misure di variabilità - Scostamenti dalla media

L'idea è quella di misurare la variabilità di un carattere verificando se le unità statistiche presentano modalità più o meno stabili rispetto a un indice di posizione assunto come rappresentativo della distribuzione.
Se si sceglie come indice di posizione la media aritmetica, una prima idea potrebbe essere quella di sintetizzare gli scarti dalla media definiti come

$$x_1 - \bar{x}; x_2 - \bar{x}; ...; x_n - \bar{x} \tag{3}$$

Non è una buona idea in quanto la sintesi più semplice e intuitiva, ovvero la media aritmetica degli scarti, sarebbe sempre uguale a zero...
Quindi, come potremmo agire? Proviamo con gli **scarti al quadrato**, è una buona idea per almeno due motivi:

1. i quadrati degli scarti sono sempre maggiori o uguali a zero, dunque **non si compensano** tra di loro
2. la loro somma costituisce un **minimo** rispetto ad altre proposte (non staremo qui a spiegare perchè).

La misura di variabilità più utilizzata è definita proprio come la media aritmetica degli scarti al quadrato ed è denominata **varianza**.
Su un insieme di valori $x_1, x_2, ..., x_n$, la varianza è definita come:

$$V(X) = \frac{\sum_{x=1}^{n}(x_i - \bar{x})^2}{n} \tag{4}$$

ATTENZIONE: la varianza è espressa nel quadrato dell'unità di misura di X. Si tratta quindi di una *misura dimensionata* il cui valore dipende oltre che dalla variabilità dall'*ordine di grandezza* del fenomeno.

La misura di variabilità più utilizzata è definita proprio come la media aritmetica degli scarti al quadrato ed è denominata **varianza**.
Su un insieme di valori $x_1, x_2, ..., x_n$, la varianza è definita come:

$$V(X) = \frac{\sum_{x=1}^{n}(x_i - \bar{x})^2}{n} \tag{4}$$

ATTENZIONE: la varianza è espressa nel quadrato dell'unità di misura di X. Si tratta quindi di una *misura dimensionata* il cui valore dipende oltre che dalla variabilità dall'*ordine di grandezza* del fenomeno.

Lo **scarto quadratico medio** è la radice della varianza

$$S(X) = \sqrt{V(X)} \tag{5}$$

ed è quindi espresso nella stessa unità di misura di X.

La misura di variabilità più utilizzata è definita proprio come la media aritmetica degli scarti al quadrato ed è denominata **varianza**.
Su un insieme di valori $x_1, x_2, ..., x_n$, la varianza è definita come:

$$V(X) = \frac{\sum_{x=1}^{n}(x_i - \bar{x})^2}{n} \tag{4}$$

ATTENZIONE: la varianza è espressa nel quadrato dell'unità di misura di X. Si tratta quindi di una *misura dimensionata* il cui valore dipende oltre che dalla variabilità dall'*ordine di grandezza* del fenomeno.

Lo **scarto quadratico medio** è la radice della varianza

$$S(X) = \sqrt{V(X)} \tag{5}$$

ed è quindi espresso nella stessa unità di misura di X.

Se vogliamo confrontare la variabilità di due caratteri che hanno caratteristiche diverse tra loro (ad es. media o metrica diversa) dobbiamo ricorrere a degli indicatori *adimensionali* di variabilità, ad esempio il **coefficiente di variazione**

$$CV(X) = S(X)/\bar{x} \tag{6}$$

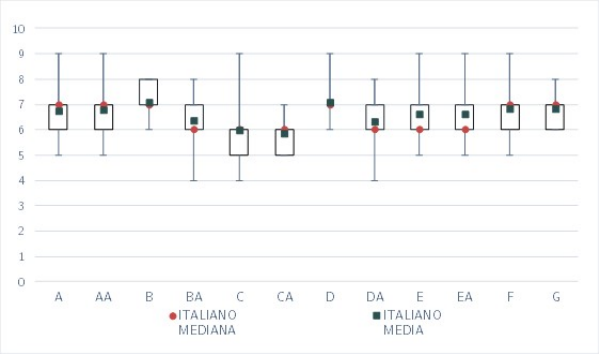
ATTENZIONE: Il CV è normalmente utilizzato solo quando tutti i valori della distribuzione sono positivi. Infatti, per caratteri che assumono valori negativi e positivi, la media aritmetica non rappresenta l'ordine di grandezza effettivo (spesso è vicina allo zero).

Un grafico capace di rappresentare sia valori medi che indici di variabilità è il **box-plot**.
È molto utile nella comparazione di caratteristiche di due o più collettivi.
E' caratterizzato da tre elementi principali:

1. Un indicatore che per la media della distribuzione.
2. Un indicatore che per la mediana della distribuzione.
3. Un rettangolo la cui altezza indica la variabilità dei valori prossimi alla media (di solito il 50% dei valori intorno ad essa quindi delimitato dal primo e dal terzo quartile).
4. Due segmenti che partono dai lati maggiori del rettangolo e i cui estremi sono determinati in base ai valori estremi della distribuzione (min e max).

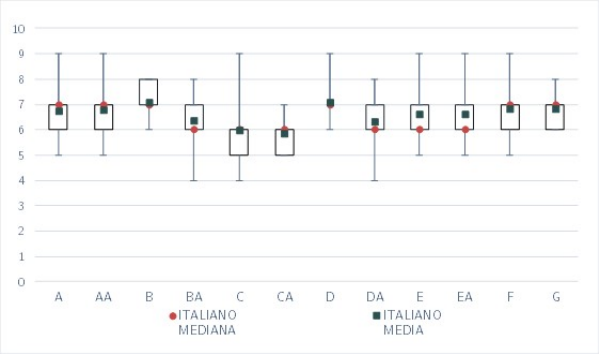
Indicatori di sintesi

Misure di variabilità - Rappresentazione grafica - Esempio



Indicatori di sintesi

Misure di variabilità - Rappresentazione grafica - Esempio



Commenti:

- Osserviamo che, per tutte le sezioni, media e mediana oscillano tra il 6 e il 7: quindi, la valutazione media degli studenti in italiano oscilla tra il sufficiente e il discreto (media) e, in ogni sezione, il 50% degli studenti è almeno sufficiente (mediana);
- Gli studenti con i voti più alti in italiano sono quelli della sezione B, mentre quelli con i voti più bassi sono quelli delle sezioni C e CA;
- Considerando tutte le sezioni, nessuno studente ha più di 9 in italiano e nessuno ha meno di 4.

Analisi congiunta di più variabili

Intro

Finora abbiamo considerato l'analisi di un carattere per volta.
Tuttavia quello che spesso interessa nella ricerca è lo studio congiunto di più caratteri rilevati sullo stesso collettivo.
In questo corso ci limiteremo allo studio congiunto di **due** caratteri.

Due caratteri rilevati (qualitativi o continui opportunamente raccolti in classi) su un collettivo possono essere organizzati in forma di **distribuzione di frequenza bivariata** che è l'analogo in due dimensioni della tabella di distribuzione di frequenze univariate. (sottosezione 5).

Analisi congiunta di più variabili

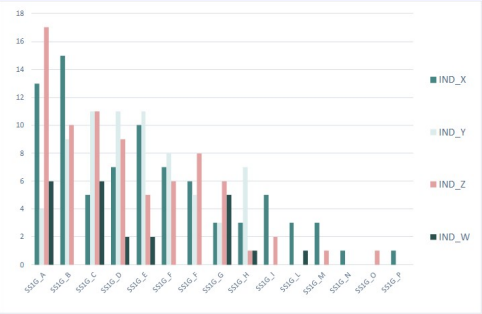
distribuzione di frequenza bivariata

Esempio

Es. di distribuzione di frequenza bivariata, frequenze assolute

SCUOLA DI PROVENIENZA x INDIRIZZO	IND_X	IND_Y	IND_Z	IND_W	vuoto	Totale
SS1G_A	13	4	17	6	0	40
SS1G_B	15	9	10	0	0	34
SS1G_C	5	11	11	6	0	33
SS1G_D	7	11	9	2	0	29
SS1G_E	10	11	5	2	0	28
SS1G_F	7	8	6	0	0	21
SS1G_G	6	5	8	0	0	19
SS1G_H	3	3	6	5	0	17
SS1G_I	3	7	1	1	0	12
SS1G_L	5	0	2	0	0	7
SS1G_M	3	0	0	1	0	4
SS1G_N	3	0	1	0	0	4
SS1G_O	1	0	0	0	0	1
SS1G_P	0	0	1	0	0	1
SS1G_Q	1	0	0	0	0	1
vuoto	0	0	0	0	0	0
Totale	**82**	**69**	**77**	**23**	**0**	**251**

Es. di rappresentazione (grafico a colonne (barre) raggruppate)



Analisi congiunta di più variabili

distribuzione di frequenze bivariata - distribuzioni marginali

La rappresentazione mediante tabella a doppia entrata ci consente anche di vedere: le distribuzioni di frequenze semplici dette **distribuzioni marginali**

Esempio

Nel nostro esempio le distribuzioni marginali sono due e sono:

Distribuzione marginale della scuola di provenienza, frequenze assolute

SCUOLA DI PROVENIENZA	N
SS1G_A	40
SS1G_B	34
SS1G_C	33
SS1G_D	29
SS1G_E	28
SS1G_F	21
SS1G_G	19
SS1G_H	17
SS1G_I	12
SS1G_L	7
SS1G_M	4
SS1G_N	4
SS1G_O	1
SS1G_P	1
SS1G_Q	1
vuoto	0
Totale	**251**

Distribuzione marginale dell'indirizzo di studi, frequenze assolute

INDIRIZZO	N
IND_X	82
IND_Y	69
IND_Z	77
IND_W	23
vuoto	0
Totale	**251**

Analisi congiunta di più variabili

distribuzione di frequenza bivariata - distribuzioni condizionate

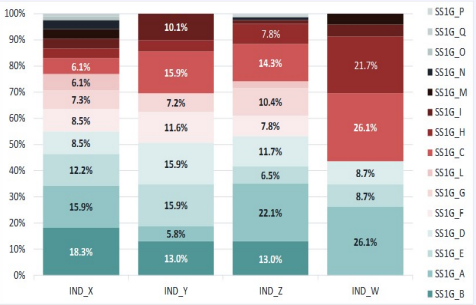
A partire dalle frequenze assolute della distribuzione bivariata si possono costruire due **distribuzioni di frequenze condizionate (in questo caso percentuali)**

Esempio

Distribuzione della scuola di provenienza condizionata all'indirizzo, frequenze percentuali

SCUOLA DI PROVENIENZA x INDIRIZZO	IND_X	IND_Y	IND_Z	IND_W	vuoto	Totale
SS1G_B	*18.3%*	*13.0%*	*13.0%*	*0.0%*		***13.5%***
SS1G_A	*15.9%*	*5.8%*	*22.1%*	*26.1%*		***15.9%***
SS1G_E	*12.2%*	*15.9%*	*6.5%*	*8.7%*		***11.2%***
SS1G_D	*8.5%*	*15.9%*	*11.7%*	*8.7%*		***11.6%***
SS1G_F	*8.5%*	*11.6%*	*7.8%*	*0.0%*		***8.4%***
SS1G_G	*7.3%*	*7.2%*	*10.4%*	*0.0%*		***7.6%***
SS1G_L	*6.1%*	*0.0%*	*2.6%*	*0.0%*		***2.8%***
SS1G_C	*6.1%*	*15.9%*	*14.3%*	*26.1%*		***13.1%***
SS1G_H	*3.7%*	*4.3%*	*7.8%*	*21.7%*		***6.8%***
SS1G_I	*3.7%*	*10.1%*	*1.3%*	*4.3%*		***4.8%***
SS1G_M	*3.7%*	*0.0%*	*0.0%*	*4.3%*		***1.6%***
SS1G_N	*3.7%*	*0.0%*	*1.3%*	*0.0%*		***1.6%***
SS1G_O	*1.2%*	*0.0%*	*0.0%*	*0.0%*		***0.4%***
SS1G_Q	*1.2%*	*0.0%*	*0.0%*	*0.0%*		***0.4%***
SS1G_P	*0.0%*	*0.0%*	*1.3%*	*0.0%*		***0.4%***
vuoto						
Totale	***100.0%***	***100.0%***	***100.0%***	***100.0%***		***100.0%***

Rappresentazione grafica tramite colonne il pila



Analisi congiunta di più variabili

distribuzione di frequenza bivariata - distribuzioni condizionate

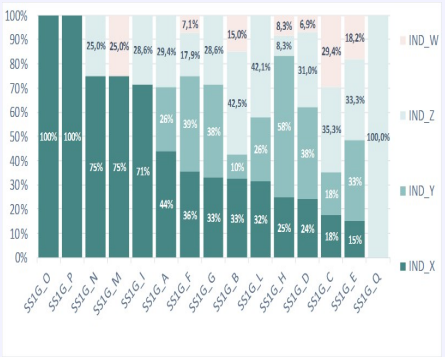
A partire dalle frequenze assolute della distribuzione bivariata si possono costruire due **distribuzioni di frequenze condizionate (in questo caso percentuali)**

Esempio

Distribuzione dell'indirizzo condizionata alla scuola di provenienza, frequenze percentuali

SCUOLA DI PROVENIENZA x INDIRIZZO	IND_X	IND_Y	IND_Z	IND_W	vuoto	Totale
SS1G_O	100,0%	0,0%	0,0%	0,0%		**100,0%**
SS1G_P	100,0%	0,0%	0,0%	0,0%		**100,0%**
SS1G_N	75,0%	0,0%	25,0%	0,0%		**100,0%**
SS1G_M	75,0%	0,0%	0,0%	25,0%		**100,0%**
SS1G_I	71,4%	0,0%	28,6%	0,0%		**100,0%**
SS1G_A	44,1%	26,5%	29,4%	0,0%		**100,0%**
SS1G_F	35,7%	39,3%	17,9%	7,1%		**100,0%**
SS1G_G	33,3%	38,1%	28,6%	0,0%		**100,0%**
SS1G_B	32,5%	10,0%	42,5%	15,0%		**100,0%**
SS1G_L	31,6%	26,3%	42,1%	0,0%		**100,0%**
SS1G_H	25,0%	58,3%	8,3%	8,3%		**100,0%**
SS1G_D	24,1%	37,9%	31,0%	6,9%		**100,0%**
SS1G_C	17,6%	17,6%	35,3%	29,4%		**100,0%**
SS1G_E	15,2%	33,3%	33,3%	18,2%		**100,0%**
SS1G_Q	0,0%	0,0%	100,0%	0,0%		**100,0%**
vuoto						
Totale	**32,7%**	**27,5%**	**30,7%**	**9,2%**		**100,0%**

Rappresentazione grafica tramite colonne il pila



Analisi congiunta di più variabili

Indipendenza lineare

Date due variabili quantitative osservate X e Y

- La **Covarianza** tra X e Y è una misura della loro dipendenza lineare, dipende dall'unità di misura delle variabili.

$$COV(X,Y) = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{n} \tag{7}$$

- L'**coefficiente di Correlazione lineare** è una misura normalizzata della dipendenza lineare. Varia tra –1 (massima associazione negativa) e 1 (massima associazione positiva) e non dipende dall'unità di misura dei due caratteri. E' uguale alla covarianza tra le variabili standardizzate.
  Assume il valore 0 quando tra due caratteri non vi è dipendenza lineare.

$$\rho(X,Y) = \frac{COV(X,Y)}{\sqrt{V(X)V(Y)}} \tag{8}$$

ATTENZIONE: questi due indicatori misurano l'intensità del legame **lineare** e assumono i loro valori soglia solo se i dati sono "allineati".

Esempio

Se consideriamo come **X** la variabile *voto in Italiano* e come **Y** il *voto in matematica* osservate tra i nsotri studenti delle classi prime nell'a.s. 2014/2015, avremo che la loro correlazione $\rho(X, Y) = 0,61$, questo significa che può esistere una relazione positiva abbastanza forte tra le due variabili.

Esempio

Se consideriamo come **X** la variabile *voto in Italiano* e come **Y** il *voto in matematica* osservate tra i nsotri studenti delle classi prime nell'a.s. 2014/2015, avremo che la loro correlazione $\rho(X, Y) = 0,61$, questo significa che può esistere una relazione positiva abbastanza forte tra le due variabili.

Ma cosa succede se tale relazione non è lineare?

Analisi congiunta di più variabili

Dipendenza lineare - Esempio

Prendiamo due variabili continue X e Y, calcoliamone il coeff. di correlazione lineare: $\rho(X, Y) = 0,08$

Analisi congiunta di più variabili

Dipendenza lineare - Esempio

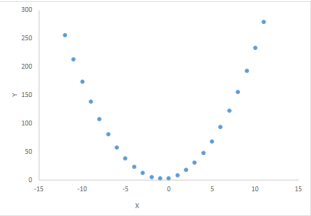
Prendiamo due variabili continue X e Y, calcoliamone il coeff. di correlazione lineare: $\rho(X, Y) = 0,08 --> $associazione praticamente nulla

Analisi congiunta di più variabili

Dipendenza lineare - Esempio

Prendiamo due variabili continue X e Y, calcoliamone il coeff. di correlazione lineare: $\rho(X, Y) = 0,08 \; -- > \text{associazione praticamente nulla}$

Guardiamo il grafico:

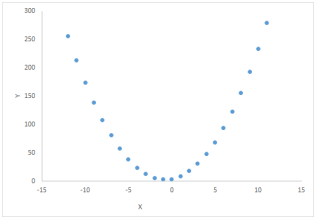


Analisi congiunta di più variabili

Dipendenza lineare - Esempio

Prendiamo due variabili continue X e Y, calcoliamone il coeff. di correlazione lineare: $\rho(X, Y) = 0,08 \; -- > \text{associazione praticamente nulla}$

Guardiamo il grafico:



Si può notare che all'aumentare della X, la Y varia secondo una forma a parabola (fidatevi), ciò significa che esiste una relazione (dipendenza?) tra X e Y ma non è lineare.

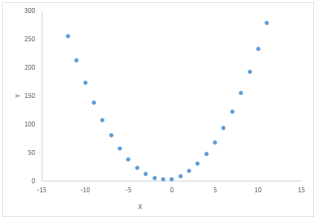
Come facciamo a descrivere questa relazione?

Analisi congiunta di più variabili

Dipendenza lineare - Esempio

Prendiamo due variabili continue X e Y, calcoliamone il coeff. di correlazione lineare: $\rho(X, Y) = 0,08 --> $associazione praticamente nulla

Guardiamo il grafico:



Si può notare che all'aumentare della X, la Y varia secondo una forma a parabola (fidatevi), ciò significa che esiste una relazione (dipendenza?) tra X e Y ma non è lineare.

Come facciamo a descrivere questa relazione?
La conosciamo la funzione della parabola?

$$Y = aX^2 + bX + c \tag{9}$$

con a, b e c parametri incogniti della funzione. La funzione della parabola fa parte di quelle che si chiamano **funzioni lineari nei parametri** e possiamo usare il **modello di regressione lineare** per stimarli...(Fare riferimento alle slide sulla regressione)

Fine

E ora passiamo all'atto pratico...

Ulteriori approfondimenti - Distribuzioni di frequenza univariate

Variabili quantitative

Per i caratteri quantitativi accade spesso (soprattutto per i continui) di osservare n valori distinti *(es. mm di pioggia, superficie di un paese)*.

Non possiamo quindi costruire tabelle di frequenza contando le unità che hanno la stessa modalità.

Come si fa a costruire distribuzioni di frequenza di variabili quantitative?

Ulteriori approfondimenti - Distribuzioni di frequenza univariate

Variabili quantitative

Per i caratteri quantitativi accade spesso (soprattutto per i continui) di osservare n valori distinti *(es. mm di pioggia, superficie di un paese)*.
Non possiamo quindi costruire tabelle di frequenza contando le unità che hanno la stessa modalità.
Come si fa a costruire distribuzioni di frequenza di variabili quantitative?

Possiamo raggruppare le modalità in **classi** e considerare simili le unità all'interno di ogni classe.
La classificazione è basata sul riconoscimento che alcune unità statistiche si rassomigliano secondo una o più caratteristiche, pur rimanendo per altri aspetti diverse.

Ulteriori approfondimenti - Distribuzioni di frequenza univariate

Variabili quantitative

Per i caratteri quantitativi accade spesso (soprattutto per i continui) di osservare n valori distinti *(es. mm di pioggia, superficie di un paese)*.
Non possiamo quindi costruire tabelle di frequenza contando le unità che hanno la stessa modalità.
Come si fa a costruire distribuzioni di frequenza di variabili quantitative?

Possiamo raggruppare le modalità in **classi** e considerare simili le unità all'interno di ogni classe.
La classificazione è basata sul riconoscimento che alcune unità statistiche si rassomigliano secondo una o più caratteristiche, pur rimanendo per altri aspetti diverse.

Esempio

Variabile: mm di pioggia caduti a Rimini nel 2015, dati giornalieri.

Classe	n_k	f_k	p_k
da 0 a 15	345	0,94	94,52%
da 15 a 45	19	0,05	5,20%
da 45 a 90	1	0,01	0,28%
oltre 90	0	0	0%
Totale	365	1	100%

Quesiti:

1. Abbiamo scelto di raggruppare le modalità in 4 classi. Un numero diverso di classi o una diversa ampiezza ci avrebbe permesso di rappresentare meglio la distribuzione del carattere?
2. Le tre classi hanno dimensioni diverse. Non sarebbe stato meglio fare classi tutte di dimensione uguale?
3. Supponiamo che in un giorno sia piovuto esattamente 15mm, lo mettiamo nella prima o nella seconda classe?

Distribuzioni di frequenza univariate

Variabili quantitative - Le classi, regole

Proprietà:

1. Le classi devono essere esaustive (ogni modalità del carattere deve poter essere assegnata ad una classe);
2. Le classi devono essere disgiunte (o esclusive) (ogni modalità del carattere deve poter essere assegnata ad una e una sola classe).

Esempio

non esaustive	non mutuamente esclusive	OK	OK
$0 - 15$	$0 \mid - \mid 15$	$0 - \mid 15$	$0 \mid - 15$
$15 - 45$	$15 \mid - \mid 45$	$15 - \mid 45$	$15 \mid - 45$
$45 - 90$	$45 \mid - \mid 90$	$45 - \mid 90$	$45 \mid - 90$
> 90	> 90	< 90	≥ 90

Esempio

Variabile: mm di pioggia caduti a Rimini nel 2015, dati giornalieri.

Classe	n_k	f_k	p_k
$0 -\vert\ 15$	345	0,94	94,52%
$15 -\vert\ 45$	19	0,05	5,20%
$45 -\vert\ 90$	1	0,01	0,28%
>90	0	0	0%
Totale	365	1	100%

Raggruppare le osservazioni in classi significa perdere informazioni (tutte le unità all'interno della classe sono considerate simili, ma in realtà sono diverse).
Bisogna quindi bilanciare l'esigenza di perdere poca informazione (poche classi: molto "livellamento" all'interno delle classi) con l'esigenza di una visione di sintesi (molte classi: troppi dettagli)

L'intervallo x_{k-1} –|x_k (o x_{k-1}|–x_k) ha ampiezza a_k uguale alla differenza dei suoi estremi:

$$a_k = x_k - x_{k-1}$$

Ad esempio l'intervallo 0 –|15, ha ampiezza 15.
Si preferisce **ampiezza diversa o comunque arbitraria** quando si identifica e qualifica in ogni classe un "tipo" o si ha una certa motivazione teorica che "imponga" una certa classificazione.
Si preferisce **ampiezza costante** quando si studia la forma distributiva del fenomeno.

Esempio ampiezza diversa

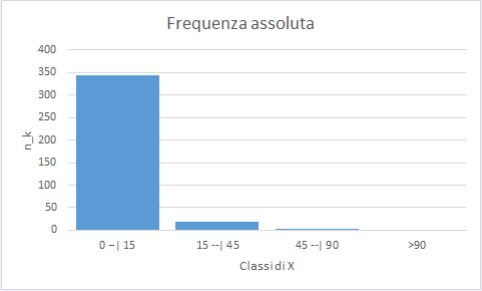
Si prende come riferimento la tabella di scale di piovosità elaborata dall'Arpa del Piemonte ottenute grazie a delle analisi di determinazione delle soglie di rischio pluviometrico:

Classe	n_k	f_k	p_k
$0 \dashv 15$	345	0,94	94,52%
$15 \dashv 45$	19	0,05	5,20%
$45 \dashv 90$	1	0,01	0,28%
>90	0	0	0%
Totale	365	1	100%

Si riporta il **grafico a barre** delle frequenze assolute.

Scale di piovosità

Intensità della pioggia	mm/6h	mm/12h	mm/24h
Debole	0÷5	0÷10	0÷15
Moderata	5÷15	10÷30	15÷45
Forte	15÷30	30÷60	45÷90
Molto Forte	>30	>60	>90

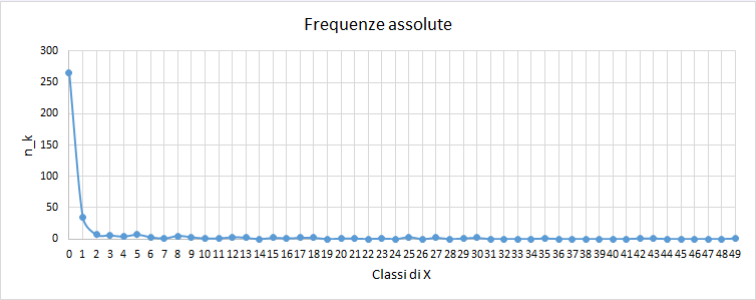


Distribuzioni di frequenza univariate

Variabili quantitative - Le classi, l'ampiezza nel nostro esempio 2/2

Esempio ampiezza costante

Si fissa un'ampiezza costante più o meno piccola e si analizzano le frequenze delle classi risultanti.
Nel nostro caso ho fissato un'ampiezza uguale a 1, ho creato 50 classi[a] (il min e il max di mm di pioggia caduti a Rimini in un giorno dalle 00:00 alle 24:00 sono rispettivamente di 0mm e 48,6mm) e ho rappresentato le frequenze risultanti tramite un grafico[b] a dispersione con linee smussate.



[a] la prima classe contiene solo lo 0

[b] sulle ascisse sono visualizzati gli estremi superiori delle classi

Analisi congiunta di più variabili

Intro

Finora abbiamo considerato l'analisi di un carattere per volta.
Tuttavia quello che spesso interessa nella ricerca è lo studio della **relazione (associazione)** tra più caratteri rilevati sullo stesso collettivo.
In questo corso ci limiteremo allo studio dell'associazione tra due caratteri.
In letteratura sono stati proposti numerosi indici statistici per misurare il grado di **associazione tra caratteri**. Vedremo che esistono indici diversi a seconda della scala di misura dei caratteri (entrambi qualitativi, entrambi quantitativi, uno qualitativo e uno quantitativo, ecc.)

Analisi congiunta di più variabili

Intro

Due caratteri rilevati su un collettivo possono essere organizzati in forma di **distribuzione unitaria semplice** oppure di **tabella a doppia entrata** che è l'analogo in due dimensioni della tabella di distribuzione di frequenze univariate. (sottosezione 5). Prima di passare a questo argomento è bene tenere a mente che:

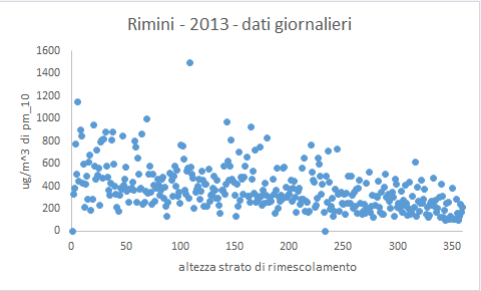
- L'interpretazione dei risultati delle procedure statistiche in termini fenomenici non è mai automatica.
- Nello studio della relazione tra caratteri bisogna tener conto che i sistemi causali del mondo reale sono complessi e le associazioni osservate tra i caratteri possono dipendere da fattori di disturbo (di confondi mento) la cui presenza e influenza va valutata criticamente.
- È sempre utile valutare le relazioni di dipendenza e indipendenza logica, accanto ai risultati delle misurazioni statistiche.

Esempio distribuzione unitaria semplice

Es. di distribuzione unitaria semplice

data	HM_RN	PM10_RN
27/05/2013	900,3333333	9
31/05/2013	843,5	9
26/12/2013	424,4166667	9
13/05/2013	594,5	10
22/11/2013	214,2916667	10
...	...	...

Es. di rappresentazione (grafico a dispersione)



Analisi congiunta di più variabili

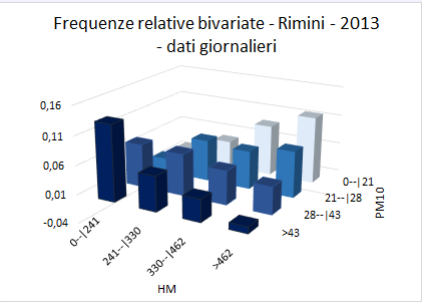
Tabelle a doppia entrata

Esempio tabella a doppia entrata

Es. di tabella a doppia entrata, frequenze relative

f_jk		PM_10				
		0--|21	*21--|28*	*28--|43*	*>43*	*Totale*
Altezza di rimescolamento	*0--|241*	0,02	0,02	0,07	0,13	**0,25**
	241--|330	0,05	0,07	0,07	0,06	**0,25**
	330--|462	0,09	0,07	0,06	0,04	**0,25**
	>462	0,12	0,08	0,05	0,01	**0,25**
	Totale	**0,27**	**0,24**	**0,25**	**0,24**	**1,00**

Es. di rappresentazione (grafico a barre tridimensionale)



Date due variabili osservate X e Y, si possono calcolare diversi indici di associazione:

- misure di associazione basate sull'**indipendenza in distribuzione** (χ^2 , *V di Cramer*)
  - possono essere applicate a coppie di caratteri qualunque (anche entrambe qualitativi)
  - sono simmetriche
  - il loro calcolo si basa sulle contingenze
- misure di associazione basate sull'**indipendenza in media** (non le faremo)
  - non sono misure simmetriche, ossia c'è una variabile dipendente e una indipendente
  - il carattere dipendente deve essere quantitativo
- misure di associazione basate sul concetto di **indipendenza lineare**

Date due variabili quantitative osservate X e Y strutturate su una tabella a doppia entrata, esse sono **indipendenti in distribuzione** se la generica frequenza assoluta n_{jk} è uguale a $\frac{n_{j\cdot} n_{\cdot k}}{n}$

Questa relazione è utilizzata per derivare in una qualsiasi tabella doppia le **frequenze teoriche** sotto l'ipotesi di indipendenza.

$$n'_{jk} = \frac{n_{j\cdot} n_{\cdot k}}{n} \tag{10}$$

Fissato questo concetto è possibile misurare la dipendenza tra X e Y come **allontanamento dalla condizione di indipendenza**. Le misure che consideriamo fanno riferimento alle sole frequenze della tabella a doppia entrata (ma non alle modalità dei caratteri) per cui possono essere calcolate per qualunque sia il tipo dei caratteri.

Analisi congiunta di più variabili

χ^2 d'indipendenza

L'indice di associazione χ^2 è calcolato a partire dalle **contingenze**, c_{jk} ottenute come differenza tra valore osservato n_{jk} e valore teorico n'_{jk}.

$$c_{jk} = n_{jk} - n'_{jk} \tag{11}$$

$$\chi^2 = \sum_{j=1}^{J} \sum_{k=1}^{K} \frac{c_{jk}^2}{n'_{jk}} \tag{12}$$

- χ^2 è sempre non negativo
- Assume valore 0 nel caso di associazione nulla e valori vicini allo 0 nel caso di approssimata indipendenza
- A parità di associazione l'indice aumenta al crescere di n!

Quali sono i valori soglia del χ^2?
Bisogna fare riferimento alle tavole statistiche, operando quello che si chiama un **test** χ^2.

Analisi congiunta di più variabili

χ^2 d'indipendenza - Esempio di test

Esempio

Ritorniamo sempre all'esempio di Rimini, 2013, $J = 4$ e $K = 4$

calcoliamo le n_{jk}:

n_jk		PM_10				
		0--|21	21--|28	28--|43	>43	Totale
Altezza di rimescolamento	0--|241	7	9	27	48	**91**
	241--|330	17	26	26	22	**91**
	330--|462	32	24	21	14	**91**
	>462	42	29	17	4	**92**
	Totale	**98**	**88**	**91**	**88**	365

Analisi congiunta di più variabili

χ^2 d'indipendenza - Esempio di test

Esempio

Ritorniamo sempre all'esempio di Rimini, 2013, $J = 4$ e $K = 4$

calcoliamo le n_{jk}:

n_jk		PM_10				
		0--|21	21--|28	28--|43	>43	Totale
Altezza di rimescolamento	0--|241	7	9	27	48	**91**
	241--|330	17	26	26	22	**91**
	330--|462	32	24	21	14	**91**
	>462	42	29	17	4	**92**
	Totale	**98**	**88**	**91**	**88**	365

Poi le frequenze teoriche n'_{jk}:

n'_jk		PM_10				
		0--|21	21--|28	28--|43	>43	Totale
Altezza di rimescolamento	0--|241	24,43	21,94	22,69	21,94	**91**
	241--|330	24,43	21,94	22,69	21,94	**91**
	330--|462	24,43	21,94	22,69	21,94	**91**
	>462	24,70	22,18	22,94	22,18	**92**
	Totale	**98**	**88**	**91**	**88**	365

Analisi congiunta di più variabili

χ^2 d'indipendenza - Esempio di test

Esempio

Ritorniamo sempre all'esempio di Rimini, 2013, $J = 4$ e $K = 4$

calcoliamo le n_{jk}:

n_jk		PM_10				
		0--|21	21--|28	28--|43	>43	Totale
Altezza di rimescolamento	0--|241	7	9	27	48	91
	241--|330	17	26	26	22	91
	330--|462	32	24	21	14	91
	>462	42	29	17	4	92
	Totale	98	88	91	88	365

Poi le frequenze teoriche n'_{jk}:

n'_jk		PM_10				
		0--|21	21--|28	28--|43	>43	Totale
Altezza di rimescolamento	0--|241	24,43	21,94	22,69	21,94	91
	241--|330	24,43	21,94	22,69	21,94	91
	330--|462	24,43	21,94	22,69	21,94	91
	>462	24,70	22,18	22,94	22,18	92
	Totale	98	88	91	88	365

Le contingenze:

c_jk		PM_10				
		0--|21	21--|28	28--|43	>43	Totale
Altezza di rimescolamento	0--|241	-17,43	-12,94	4,31	26,06	0
	241--|330	-7,43	4,06	3,31	0,06	0
	330--|462	7,57	2,06	-1,69	-7,94	0
	>462	17,30	6,82	-5,94	-18,18	0
	Totale	0	0	0	0	0

Analisi congiunta di più variabili

χ^2 d'indipendenza - Esempio di test

Esempio

Ritorniamo sempre all'esempio di Rimini, 2013, $J = 4$ e $K = 4$

calcoliamo le n_{jk}:

n_jk		PM_10				
		0--|21	21--|28	28--|43	>43	Totale
Altezza di rimescolamento	0--|241	7	9	27	48	91
	241--|330	17	26	26	22	91
	330--|462	32	24	21	14	91
	>462	42	29	17	4	92
	Totale	98	88	91	88	365

Poi le frequenze teoriche n'_{jk}:

n'_jk		PM_10				
		0--|21	21--|28	28--|43	>43	Totale
Altezza di rimescolamento	0--|241	24,43	21,94	22,69	21,94	91
	241--|330	24,43	21,94	22,69	21,94	91
	330--|462	24,43	21,94	22,69	21,94	91
	>462	24,70	22,18	22,94	22,18	92
	Totale	98	88	91	88	365

Le contingenze:

c_jk		PM_10				
		0--|21	21--|28	28--|43	>43	Totale
Altezza di rimescolamento	0--|241	-17,43	-12,94	4,31	26,06	0
	241--|330	-7,43	4,06	3,31	0,00	0
	330--|462	7,57	2,06	-1,69	-7,94	0
	>462	17,30	6,82	-5,94	-18,18	0
	Totale	0	0	0	0	0

Ed infine il χ^2

(c_jk)^2/(n'_jk)		PM_10				
		0--|21	21--|28	28--|43	>43	Totale
Altezza di rimescolamento	0--|241	12,44	7,63	0,82	30,95	51,84
	241--|330	2,26	0,75	0,48	0,00	3,50
	330--|462	2,34	0,19	0,13	2,87	5,54
	>462	12,11	2,10	1,54	14,90	30,65
	Totale	29,16	10,67	2,97	48,73	91,53

Quest'ultimo (91,53) va confrontato con i valori teorici dela distribuzione χ^2 con $(J-1)*(K-1)$ gradi di libertà fissando una probabilità di errore uguale a 0,05, nel nostro caso è uguale a 3,32, nettamente inferiore del valore osservato $-->$ rifiutiamo l'ipotesi nulla di indipendenza quindi possiamo affermare che X e Y sono **dipendenti in distribuzione**.