Statistica Sociale

08/03/2022
09/03/2022

Casualizzazione

I metodi statistici inferenziali fanno uso delle statistiche campionarie per fare previsioni sui parametri delle popolazioni

L'utilità dell'inferenza dipende da quanto bene il campione rappresenta la popolazione

- È importante ridurre la probabilità di selezionare campioni che per le loro caratteristiche possano distorcere la rappresentatività della popolazione portando ad errate conclusioni inferenziali sui valori dei parametri

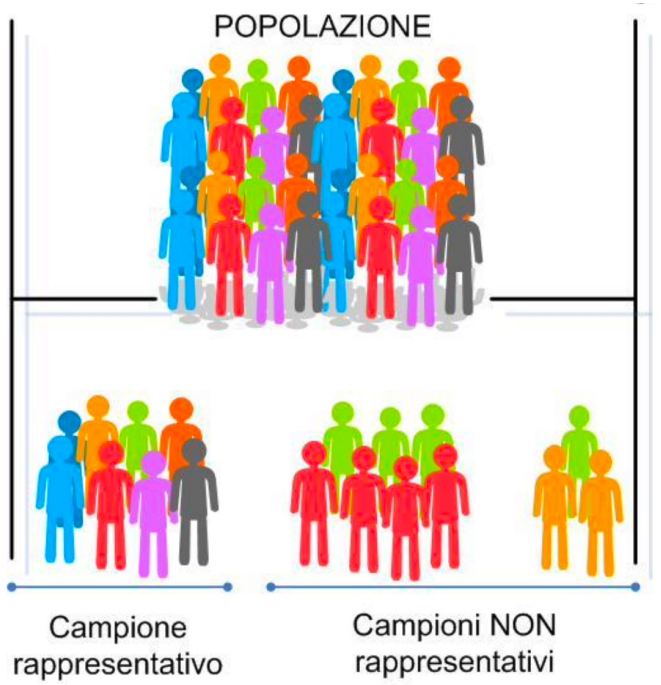
Campione casuale (semplice) di *n* soggetti estratti da una popolazione è tale se a ogni possibile campione di pari numerosità ha uguale probabilità di selezione

- n è la numerosità/dimensione campionaria
- Per selezionare un campione è necessario avere una lista di campionamento

Indagine campionaria: selezione di un campione dalla popolazione di riferimento; le informazioni sono raccolte tramite intervista diretta, telefonica o autocompilata

Esperimento: confrontare le risposte sotto diverse condizioni (trattamenti) su cui si ha controllo sperimentale; il piano sperimentale è il processo attraverso cui il ricercatore assegna i soggetti ai diversi trattamenti (in modo casuale)

Studio osservazionale: si osservano i valori delle variabili di interesse ma non si ha controllo sperimentale; non è possibile determinare i rapporti causa-effetto



Variabilità campionaria

Diversi campioni possono portare a risultati campionari diversi

L'errore campionario di una statistica è l'errore che viene commesso quando viene impiegata una statistica campionaria per prevedere il valore di un parametro della popolazione

L'errore campionario è sempre sconosciuto perchè …

… il valore del parametro di interesse è sempre sconosciuto

Grazie al principio del campionamento casuale possiamo assumere che l'errore campionario oscilli intorno alla 0

Il campionamento casuale ci permette di prevedere l'ampiezza dell'errore campionario

Intuitivamente, che relazione pensate ci sia tra l'errore campionario e la numerosità campionaria?

Ipotizziamo che 3 diversi istituti di sondaggi abbiano selezionato 3 diversi campioni casuali di uguale ampiezza simultaneamente per valutare il sostegno della popolazione al primo ministro in carica

Il primo istituto riporta che il 54% della popolazione sostiene il primo ministro

Il secondo istituto riporta che il 57% della popolazione sostiene il primo ministro

Il terzo istituto riporta che il 51% della popolazione sostiene il primo ministro

Ipotizziamo che il vero valore del parametro sia 55%

- Le domande sono state poste con la stessa formula?
- Ci aspettiamo comunque delle differenze tra i risultati campionari ottenuti da campioni casuali diversi?

Distorsioni

I metodi inferenziali si basano sull'assunzione che i dati provengono da un campione probabilistico

- La probabilità di estrarre un certo campione è nota prima dell'estrazione

Quando non è possibile calcolare la probabilità di estrazione di un campione parliamo di campioni non-probabilistici

I risultati ottenuti da campioni non probabilistici portano a distorsioni campionarie dovute alla possibile bassa rappresentatività dei campioni

- Campionamento non-probabilistico: con il campionamento volontario i soggetti decidono di essere inclusi nel campione

Una sottocopertura nelle liste da cui selezionare un campione casuale può portare comunque ad una distorsione campionaria

Distorsioni dovute alle risposte:

- Domande mal poste o confuse
- Risposte non sincere

Distorsioni dovute a dati mancanti:

- Risposte mancanti
- Soggetti non contattati

Percentuali di non-risposte superiori al 20% vanno valutate con cautela, una possibile soluzione è l'imputazione

L’importanza della numerosità campionaria

Nel 1936, il settimanale Literary Digest inviò per posta oltre 10 milioni di questionari chiedendo di prevedere chi sarebbe stato il prossimo presidente americano tra Landon e Roosvelt.

Solo il 25% degli intervistati rispose. Il risultato atteso dagli intervistati fu la schiacciante vittoria di Landon.

George Gallup ha condotto un’indagine campionaria con lo stesso fine su circa 50.000 persone

Il risultato previsto dagli intevistati da Gallup fu la vittoria di Roosvelt.

- Che tipo di campionamento è quello del Literary Digest? [campionamento volontario]
- Gli intervistati sono stati selezionati da liste di iscritti a riviste, liste di numeri telefonici e liste di appartenenza a club. Sono liste rappresentative della popolazione di riferimento? [no, rappresentano bene la popolazione della classe sociale medio-alta negli USA del 1936]
- Quale altro problema evidenziate nell’indagine del Literary Digest? [l’elevato numero di risposte mancanti]

Altri metodi di campionamento probabilistici

Campione sistematico: dato un passo di estrazione k = N/n, vengono selezionati tutti i soggetti nella lista presenti ogni k soggetti

- Più semplice del campionamento casuale
- Anche se non tutti i campioni hanno la stessa probabilità di essere selezionati, possono essere applicati gli stessi metodi previsti per il campionamento casuale

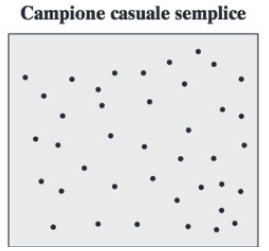
Campione stratificato: la popolazione viene divisa in strati e da ognuno di questi viene estratto un campione casuale semplice

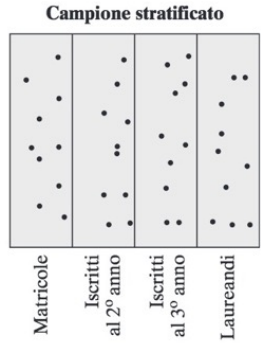
- Proporzionale (se le proporzioni di ogni strato del campione è uguale alla proporzione di popolazione corrispondente al gruppo) o non-proporzionale
- Alcune variabili sono più adatte alla definizione degli strati

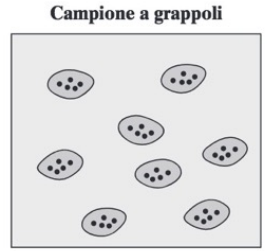
Campione a grappoli: la popolazione viene divisa in molti grappoli ed il campionamento casuale è applicato ai grappoli selezionando tutti i soggetti in essi inclusi

- La maggior parte dei grappoli non viene rappresentata dal campione

Campione a più stadi: è ottenuto attraverso combinazioni dei metodi precedenti. Ad esempio viene applicato prima un campionamento a grappoli e successivamente vengono campionate delle unità in ogni grappolo selezionato.







Statistica descrittiva

Abbiamo già introdotto la distinzione tra metodi statistici *descrittivi e inferenziali*

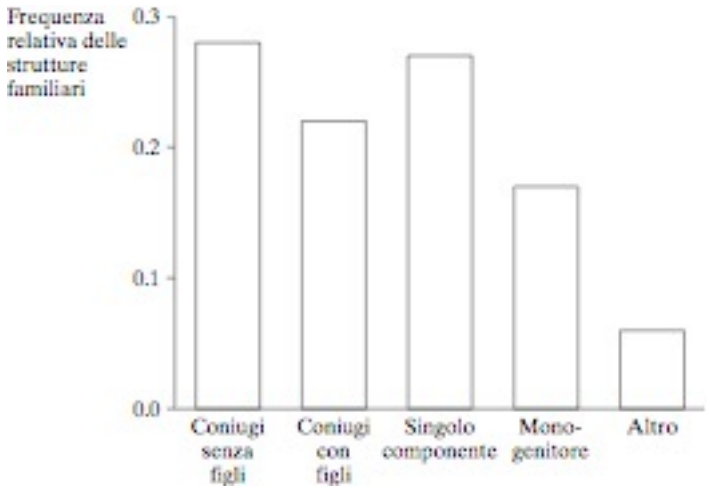
Vedremo adesso le metodologie di base della statistica descrittiva

L'obiettivo delle statistiche descrittive è quello di sintetizzare i dati per rendere fruibili le informazioni in essi contenute

Tabelle e grafici sono utili per sintetizzare tutti i tipi di dati

Famiglia	Numerosità	Proporzione	%
Coppie sposate con figli	24.1	0.22	22
Coppie sposate senza figli	31.1	0.28	28
Monogenitore	19.1	0.17	17
Singolo componente	30.1	0.27	27
Altre tipologie	6.7	0.06	6
Totale	111.1	1.00	100

Fonte: U.S. Census Bureau, *2005 Am.Comm.Survey,* Tav. B11001, C11003.



Distribuzione unitaria di un carattere

Dopo aver acquisito e controllato i dati si passa alla loro sintesi e descrizione

Distribuzione unitaria semplice: elenco, unità per unità, delle modalità di una variabile osservate nel campione

La variabile assume una modalità in corrispondenza di ogni unità statistica

- La variabile età assume valore 21 per l'unità 5
- La distribuzione unitaria del carattere sesso è:
  - F, F, M, F, F, F, M, F

La distribuzione unitaria multipla è l'elenco delle modalità di più variabili osservate nel campione

Unità	Sesso	Età	Statura	Colore occhi
1	F	24	163	Marrone
2	F	21	165	Azzurri
3	M	34	185	Azzurri
4	F	22	164	Marroni
5	F	21	167	Marroni
6	F	22	175	Verdi
7	M	24	178	Verdi
8	F	21	155	Marroni

Distribuzione di frequenze

La frequenza assoluta di una modalità è il numero di volte che questa viene osservata nel campione

- La frequenza assoluta della modalità “Monogenitore” per la variabile Famiglia è 19,1 milioni

La distribuzione di frequenze associa alla distribuzione di una variabile le frequenze osservate

- Si dice semplice se riferita ad un sola variabile, doppia se riferita a due variabili, multipla a più di una variabile.

Famiglia	Numerosità	Proporzione	%
Coppie sposate con figli	24.1	0.22	22
Coppie sposate senza figli	31.1	0.28	28
Monogenitore	19.1	0.17	17
Singolo componente	30.1	0.27	27
Altre tipologie	6.7	0.06	6
Totale	111.1	1.00	100

Fonte: U.S. Census Bureau, *2005 Am.Comm.Survey*, Tav. B11001, C11003.

Costruiamo le distribuzioni di frequenze per le variabili del nostro campione

Unità	Sesso	Età	Statura	Colore occhi
1	F	24	163	Marrone
2	F	21	165	Azzurri
3	M	34	185	Azzurri
4	F	22	164	Marroni
5	F	21	167	Marroni
6	F	22	175	Verdi
7	M	24	178	Verdi
8	F	21	155	Marroni

Sesso	Numerosità
F	6
M	2
	8

Età	Numerosità
21	3
22	2
24	2
34	1
	8

- Qual è la modalità osservata più numerosa della variabile Età? [21]
- Quale variabile è più difficile sintetizzare? [le variabili quantitative o le qualitative con molte modalità]
- Quale delle due rappresentazioni dei dati raccolti offre maggiori informazioni? [la distribuzione unitaria]
- Da quale rappresentazione è più semplice leggere informazioni? [le distribuzioni di frequenze]

Suddivisione in classi

Quando la variabile presenta molte modalità distinte è utile procedere ad una divisione in classi

Non esiste una regola generica per la suddivisione:

- è una scelta soggettiva, dipende dal contesto e per questo deve essere motivata
- si perdono informazioni al prezzo di una maggiore leggibilità dei dati osservati

Se la variabile è qualitativa si possono accorpare le modalità seguendo uno specifico criterio (ad esempio un livello superiore di gerarchia: comuni -> province -> regioni)

Se la variabile è quantitativa la suddivisione in classi ci porta ad un livello ordinale

- Le classi possono avere ampiezza costante o diversa
- Se il numero delle classi è troppo piccolo, rischiamo di sintetizzare troppo e perdere troppa informazione viceversa, se il numero delle classi è troppo alto mateniamo più informazione ma rischia di essere poco leggibile (troppi dettagli)
- Le classi devono essere disgiunte (mutualmente esclusive) e devono includere tutte le possibili modalità della variabile

Suddivisione in classi

L'ampiezza delle classi può essere calcolata come:

- $ampiezza = \frac{valore\ massimo - valore\ minimo}{numero\ delle\ classi}$
  - il minimo e massimo valore osservato non devono coincidere con l'estremo inferiore della prima classe e con l'estremo superiore dell'ultima
  - l'ampiezza ottenuta va approssimata ad un numero intero (es. 9,7 -> 10)

Esempio:

I dati osservati variano tra 11,2 e 98,6 e si vogliono suddividere in 9 classi:

$$ampiezza = \frac{98{,}6 - 11{,}2}{9} \approx 10$$

Scegliamo come valore iniziale per la prima classe 10 (così da non farlo coincidere con 11,2), avremo [10, 20), [20, 30), …, [90,100) oppure 10 -19, 20 - 29, …, 90 - 99

La leggibilità dei dati è la priorità!

Frequenze relative e percentuali

La frequenza relativa è la frequenza assoluta divisa per il numero totale di unità osservate

- È un numero compreso tra 0 e 1
- La somma delle frequenze relative di una variabile è uguale a 1

La frequenza percentuale è la frequenza relativa moltiplicata per 100

Età	Numerosità	Freq Relativa	Percentuale
21	3	3/8 = 0,375	3/8*100 = 37,5%
22	2	2/8 = 0,25	2/8*100 = 25%
24	2	2/8 = 0,25	2/8*100 = 25%
34	1	1/8 = 0,125	1/8*100 = 12,5%
	8	1	100

Le frequenze relative o percentuali sono utili per confrontare frequenze da campioni di diversa numerosità poichè non dipendono dalla numerosità del campione

Esempio (Borra-Di Ciaccio): dalle distribuzioni di frequenze assolute dei due campioni qui sotto sembra che la modalità 1 sia più presente nel secondo gruppo: $(\text{gruppo1}: x_1 = 2; gruppo2: x_2 = 12)$

Gruppo 1	Numerosità
x_1	2
x_2	4
x_3	8
	14

Gruppo 2	Numerosità
x_1	12
x_2	46
x_3	32
	90

Considerando però le frequenze percentuali otteniamo che in realtà la modalità 1 è più presente nel gruppo 1

$$\text{gruppo1}: p_1 = \frac{2}{14} * 100 = 14{,}29\,\%$$

$$gruppo2: p_1 = \frac{12}{90} * 100 = 13{,}33\%$$

Rappresentazioni grafiche

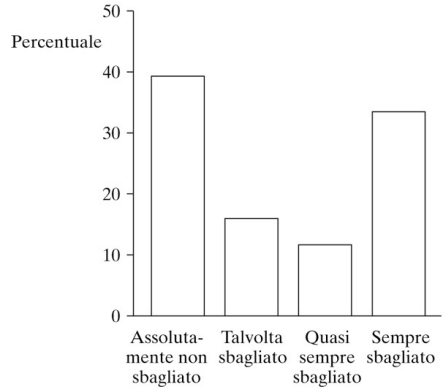
Facilità di lettura delle caratteristiche di una distribuzione osservata di una variabile

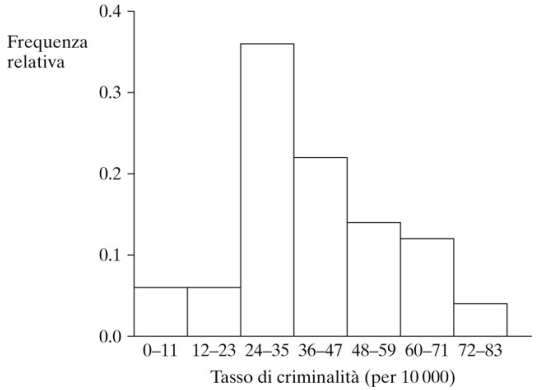
I 5 aspetti da considerare nella valutazione di un grafico:

- Accuratezza: precisione dei dettagli (dimensioni adeguate, legenda)
- Semplicità: limitare elementi grafici superflui, inserire solo ciò che aiuta la lettura del grafico
- Chiarezza: concentrarsi su un messaggio e rappresentare le caratteristiche salienti
- Aspetto: tratti, proporzioni e colori armonici
- Struttura: gli elementi più importanti devono avere più rilievo

Tipi di grafici

Variabile qualitativa		Variabile quantitativa	
Nominale	Ordinale	Discreta	Continua
$\neq$	$\geq o \leq$	$+ \; - \; * \; /$	$+ \; - \; * \; /$
Grafico a torta	[Grafico a torta]		
[Grafico a barre]	Grafico a barre	Grafico a barre	
			Istogramma





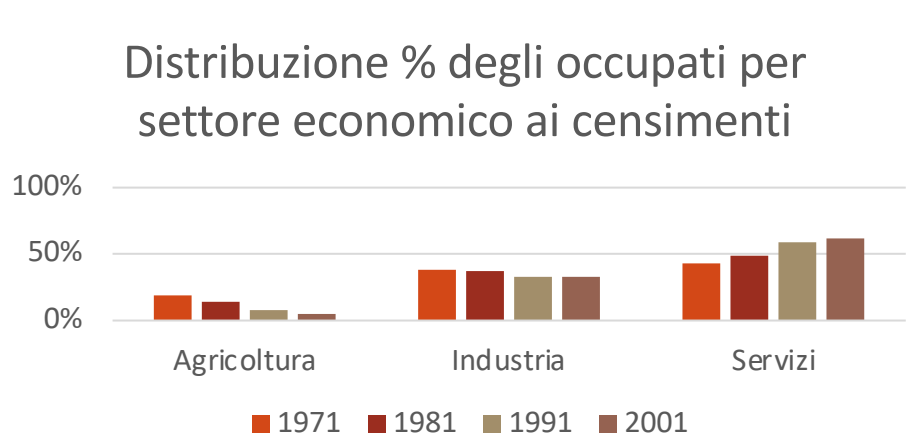
Grafici a barre

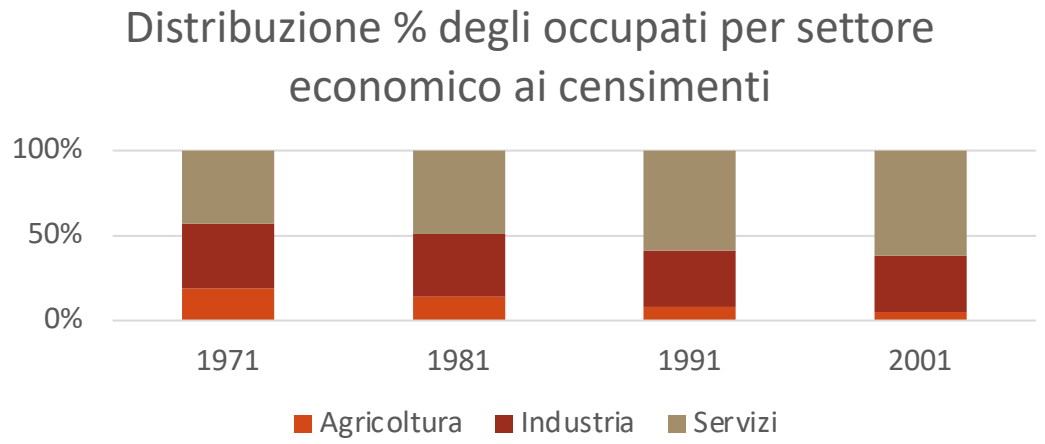
La frequenza (assoluta, relativa o percentuale) di ogni modalità viene rappresentata da una barra/rettangolo

I rettangoli hanno tutti la stessa base mentre le altezze sono proporzionali alle frequenze delle modalità

Mantiene in evidenza l'ordinamento delle modalità del carattere

Grafici a barre multipli permettono il confronto di più distribuzioni





Istogrammi

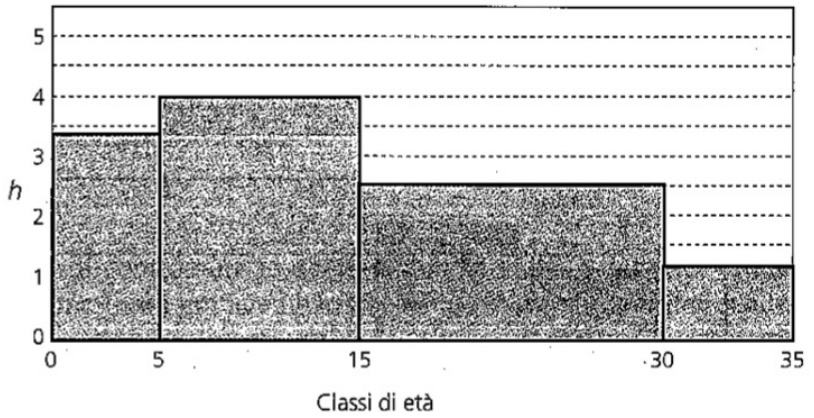
Le barre sono non distanziate

Le basi dei rettangoli possono avere ampiezza diversa, quindi l'area del rettangolo è proporzionale alla corrispondente frequenza

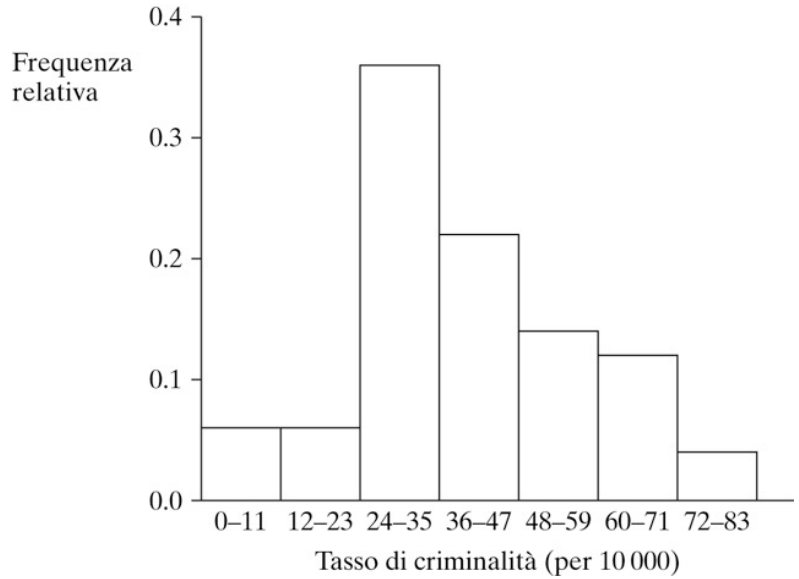
Nel caso di classi di ampiezza diversa, l'altezza del rettangolo è detta densità ed è proporzionale al rapporto tra frequenza e ampiezza

Attenzione all'assunzione che le modalità siano distribuite uniformemente all'interno delle classi

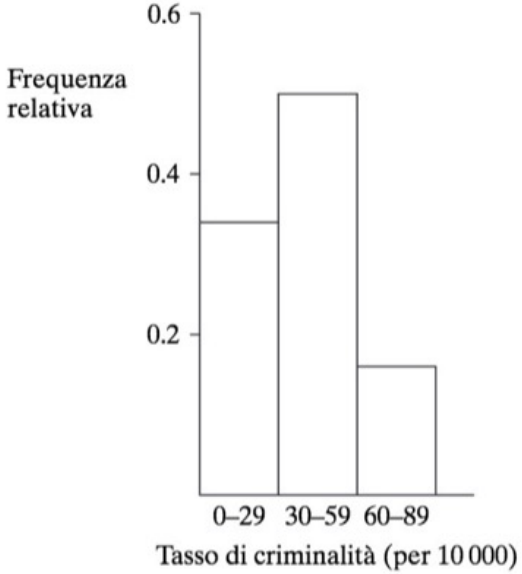
Classi di età	Ampiezza	*Percentuale*	*Densità*
[0-5)	5	17%	3,4 = 17/5
[5-15)	10	40%	4 = 40/10
[15-30)	15	37%	2,5 = 37/15
[30-35)	5	6%	1,2 = 6/5



Istogramma della distribuzione di frequenze relative del tasso di criminalità negli Stati Uniti nel 2005



Istogramma della distribuzione di frequenze relative del tasso di criminalità negli Stati Uniti nel 2005 ottenuto utilizzando intervalli di ampiezza eccessiva

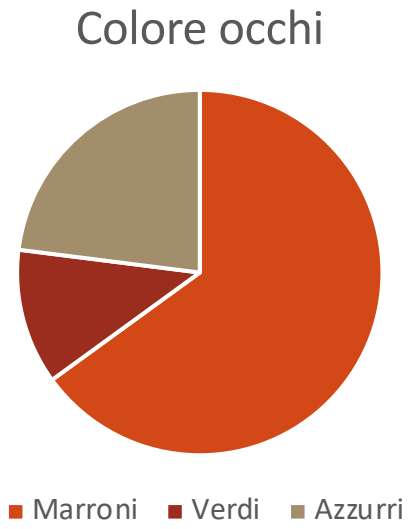


Grafici a torta

Le frequenze sono espresse in termini di angoli

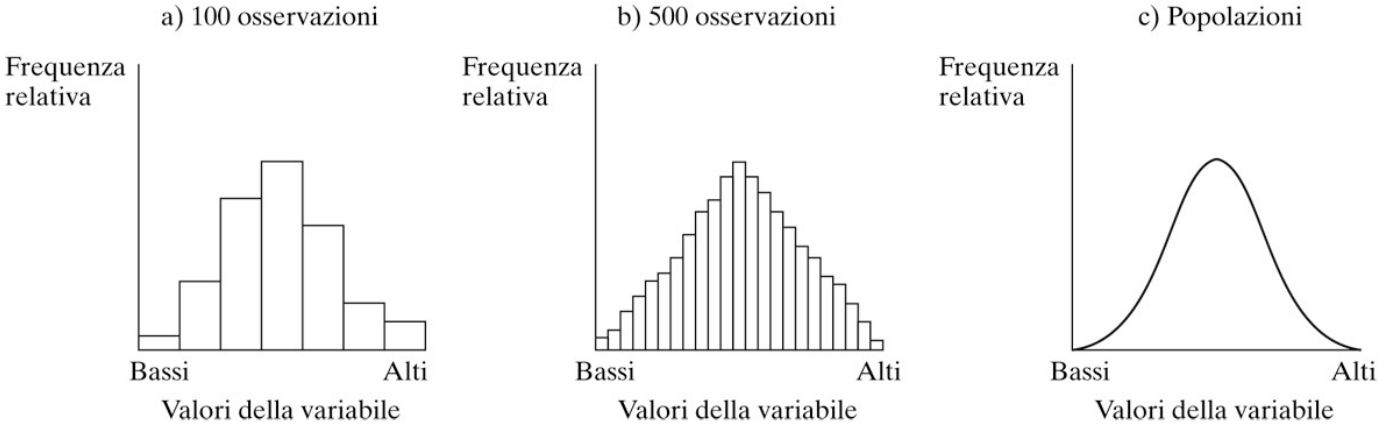
Adatto per distribuzioni con un basso numero di modalità

$$freq\ rel_{marroni} * 360 = 0{,}65 * 360 \approx 234$$
$$freq\ rel_{verdi} * 360 = 0{,}12 * 360 \approx 43$$
$$freq\ rel_{azzurri} * 360 = 0{,}23 * 360 \approx 83$$



Distribuzioni di popolazioni e distribuzioni di dati campionari

- La costruzione di distribuzioni di frequenze o di istogrammi può essere fatta sia per i dati di popolazione che per quelli campionari.
- Nel primo caso facciamo riferimento a distribuzioni di dati di popolazione, nel secondo a distribuzioni di dati campionari.
- Si può dire che una distribuzione di dati campionari è una foto sfuocata della distribuzione dei dati di popolazione.
- A mano a mano che l'ampiezza campionaria aumenta le proporzioni campionarie in ciascun intervallo si approssimano sempre di più alle vere proporzioni della popolazione: la distribuzione dei dati campionari diventa sempre più simile alla distribuzione di popolazione.



Nel grafico c) è sovraimpressa una curva liscia che rappresenta la distribuzione nella popolazione (anche se una variabile è discreta, la sua distribuzione a livello di popolazione può essere adeguatamente approssimata da una curva liscia).

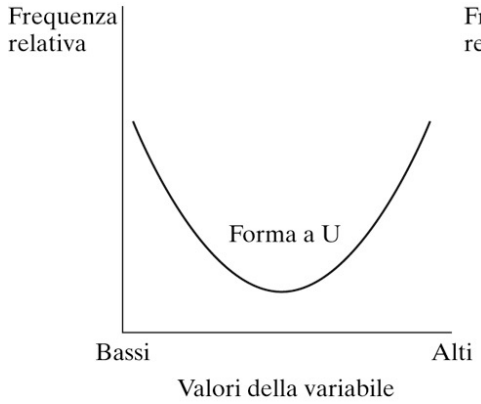
La forma della distribuzione

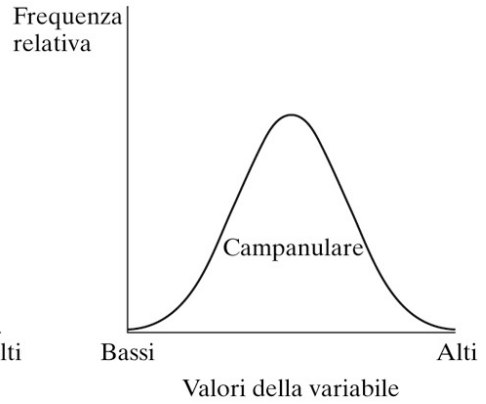
Entrambe le distribuzioni sono simmetriche: il lato della distribuzione a sinistra del valore centrale è l'immagine speculare del lato a destra dello stesso valore.

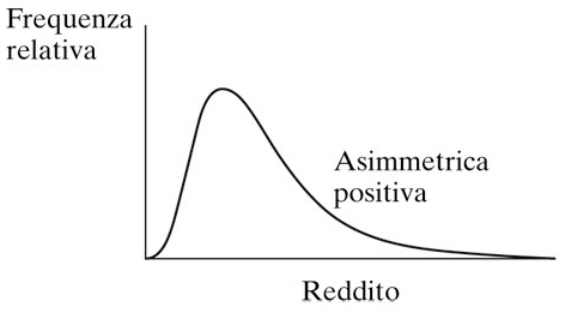
La maggior parte delle distribuzioni che vengono osservate nell'ambito della ricerca sociale non sono simmetriche.

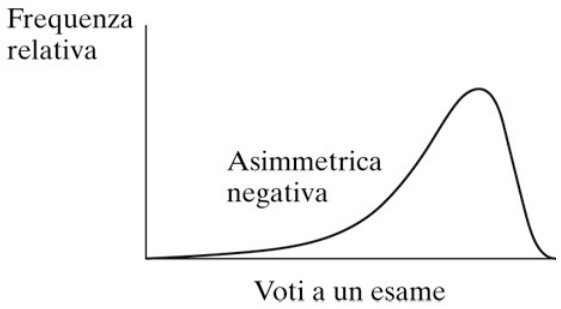
Le parti della figura in corrispondenza dei valori bassi e alti sono chiamate code della distribuzione.

Una distribuzione viene definita asimmetrica positiva o asimmetrica negativa a seconda di quella che è la sua coda più lunga.









Esercizio

Completa la tabella e rappresenta i dati nel modo per te più opportuno

Modalità	Numerosità	Frequenza relativa	Percentuale
A	11	0,22	
B	30		
C			