

Esercitazione 1

Corso di Statistica

Info

Daniele Pasquini

Phd Student in Data Science

- **Email:** daniele.pasquini@uniroma2.eu
- **Ricevimento:** Online su Teams
- **Ricevimento:** Ing. Informatica, Corpo D, Piano 2, Stanza 20

Statistica

- La **statistica** è lo studio di come raccogliere, organizzare, analizzare e interpretare informazioni numeriche e dati
- Aiuta a prendere **decisioni**
- Scienza dell'**incertezza**, le decisioni non sono perfette, ma migliori di un'ipotesi casuale

Pianificazione di uno studio

1. Dichiarare l'ipotesi (o la domanda di ricerca).
- 2. Identificare gli individui di interesse.**
- 3. Specificare le variabili da misurare.**
4. Determinare se utilizzare l'intera popolazione (censimento) o un campione, scegliendo il metodo di campionamento appropriato.
5. Considerare le preoccupazioni etiche (es. consenso informato).
6. Raccogliere i dati (tramite dati esistenti o raccolta diretta).
7. Utilizzare la statistica (descrittiva o inferenziale) per rispondere all'ipotesi.
8. Redigere i risultati, includendo raccomandazioni e preoccupazioni sull'analisi o raccolta dati.

Concetti base

- **Individui** (Individuals): Sono le **persone** o gli **oggetti** inclusi in uno studio
- **Variabili** (Variables): Sono i fattori o le caratteristiche che vengono **misurate** sugli individui e che possono variare. Ogni individuo può avere molte variabili misurate su di sé

Variabili

- **Quantitative** (o Continue): Misure numeriche da cui è possibile calcolare una **media**. Si suddividono ulteriormente in:
 - **Intervallo** (Interval)
 - **Rapporto** (Ratio)

Variabili

- **Qualitative** (o Categorie): Caratteristiche o categorie. Si suddividono ulteriormente in:
 - **Nominale** (Nominal)
 - **Ordinale** (Ordinal)

Variabili

Categoria	Definizione generale	Caratteristica distintiva	Esempio e Spiegazione
Intervallo	I valori sono numerici e le differenze tra i valori dei dati sono significative .	Non esiste uno zero vero (True Zero) . Un valore zero non implica l'assenza della quantità misurata, ma solo un punto arbitrario sulla scala.	Anno di Ammissione / Anno di Diagnosi . Non esiste un "Anno Zero" (Year Zero) assoluto o un "Tempo Zero".
Rapporto	I valori sono numerici e le differenze tra i valori dei dati sono significative .	Esiste uno zero vero (True Zero) . La presenza di un valore zero indica l'assenza della caratteristica o della quantità misurata.	Pressione Sanguigna Sistolica / Conteggio delle Piastrine Nella sanità, la maggior parte dei dati rientra nel livello di misurazione "Rapporto".

Variabili

Categoria	Definizione generale	Caratteristica distintiva	Esempio e Spiegazione
Nominale (Nominal)	Si applica a categorie, etichette o nomi .	Non possono essere ordinate dal più piccolo al più grande. Le categorie devono essere considerate uguali in termini di ordine.	Tipo di Assicurazione Sanitaria (Medicare, Medicaid, assicurazioni private). Paese di Origine .
Ordinale (Ordinal)	Si applica a dati che possono essere disposti in un ordine o gerarchia.	Esiste un ordine naturale , ma la differenza tra i valori dei dati è indeterminabile o non significativa . Sebbene vi sia un ordine, la distanza tra i livelli non è uguale.	Stadio del Cancro (Stadio 1, 2, 3, 4). Livello Triage (Livello 1, 2, 3, 4). La differenza tra Stadio 3 e 4 è materialmente diversa da quella tra Stadio 1 e 2.

Variabili

- il livello di misurazione determina quali **operazioni** e metodi statistici possono essere eseguiti sui dati

Esercizio

1. Dichiarare l'ipotesi (o la domanda di ricerca) in questo contesto
2. Identificare gli individui di interesse.
3. Specificare le variabili da misurare.

Pianificazione di uno studio

1. Dichiarare l'ipotesi (o la domanda di ricerca).
2. Identificare gli individui di interesse.
3. Specificare le variabili da misurare.
4. **Determinare se utilizzare l'intera popolazione (censimento) o un campione, scegliendo il metodo di campionamento appropriato.**
5. Considerare le preoccupazioni etiche (es. consenso informato).
6. Raccogliere i dati (tramite dati esistenti o raccolta diretta).
7. Utilizzare la statistica (descrittiva o inferenziale) per rispondere all'ipotesi.
8. Redigere i risultati, includendo raccomandazioni e preoccupazioni sull'analisi o raccolta dati.

Concetti base

- **Popolazione** (Population): Un gruppo di persone o oggetti con un tema comune, considerato nella sua totalità (ogni membro è incluso)
- **Campione** (Sample): Un sottoinsieme della popolazione.

Concetti base

- **Parametro (Parameter)**: Una misura che descrive l'intera popolazione (inizia con 'P'). Ad esempio
- **Statistica (Statistic)**: Una misura che descrive solo un campione di una popolazione (inizia con 'S')

Metodi di campionamento

- **Sampling Frame** (Lista di Campionamento): La lista da cui viene selezionato il campione.
- Non deve esserci **sottocopertura** (undercoverage), ovvero l'omissione di membri della popolazione dalla lista

Metodi di campionamento

- **Campionamento Casuale Semplice (Simple Random Sampling)**: Tecnica di base
- Ogni individuo della popolazione ha la stessa probabilità di essere selezionato

Metodi di campionamento

- **Campionamento Sistemático (Systematic Sampling)**: Consiste nello scegliere un punto di partenza casuale e poi selezionare ogni k -esimo individuo
- È facile da eseguire con o senza una lista
- Utile in contesti clinici quando c'è un flusso continuo di pazienti (ad esempio, in un pronto soccorso o in una clinica)

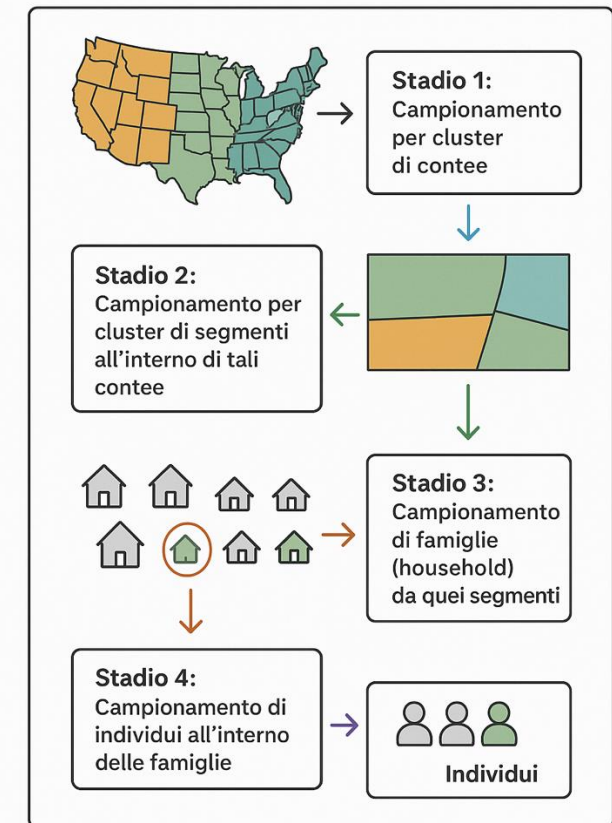
Metodi di campionamento

- **Campionamento Stratificato (Stratified Sampling)**: Scelto quando si vuole assicurare la rappresentazione di sottogruppi specifici
- Si divide la popolazione in gruppi omogenei (strati) e si preleva un campione casuale semplice da ciascuno strato

Metodi di campionamento

- **Campionamento a Più Stadi (Multi-stage Sampling):**
Combinazione di diversi approcci (cluster, casuale, stratificato) in fasi sequenziali. Utilizzato tipicamente in grandi studi governativi

Campionamento a Più Stadi (Multi-stage Sampling)



Metodi di campionamento

- Parametro di Popolazione (P) o una Statistica di Campione (S)?

1. Il tempo medio di attesa in una clinica, calcolato intervistando 200 volontari che hanno partecipato a uno studio.

Metodi di campionamento

- Parametro di Popolazione (P) o una Statistica di Campione (S)?

2. La percentuale di donne in Africa che ha contratto l'HIV, secondo i dati raccolti dalle Nazioni Unite.

Metodi di campionamento

- Parametro di Popolazione (P) o una Statistica di Campione (S)?

3. La media aritmetica degli stipendi di 16 professori universitari scelti a caso da una grande università privata

Metodi di campionamento

- Parametro di Popolazione (P) o una Statistica di Campione (S)?

Soluzioni:

1. **S (Statistica di Campione)**. Lo studio ha reclutato volontari, che rappresentano solo un campione.
2. **P (Parametro di Popolazione)**. Ci si riferisce alla percentuale dell'intera Africa (popolazione) e i dati sono probabilmente raccolti da governi o organizzazioni internazionali, tipici dei parametri.
3. **S (Statistica di Campione)**. Sebbene si riferisca a "professori" (individui), il calcolo è basato su un "campione casuale di 16 professori".

Metodi di campionamento

- Studiare la presenza di metalli pesanti nei chicchi di caffè prodotti da una grande Azienda
- Come si applica il Campionamento Sistemático per selezionare 50 lotti da un totale di 2000 lotti prodotti in un mese?

$$k = \text{popolazione} / \text{campione} = 2000/50 = 40$$
$$\text{lotto iniziale casuale} = 1 \leq i \leq k$$

Esercizio

1. Applicare il campionamento casuale semplice

Misure di Frequenza

- primo passo della **statistica descrittiva** è organizzare e visualizzare i dati, utilizzando tabelle di frequenza
- La frequenza si applica a dati che assumono un numero relativamente **ridotto** di valori distinti o a **dati raggruppati in intervalli** (classi)

Frequenza assoluta

- La **frequenza assoluta** (f) di un particolare valore o classe è il **numero di volte in cui tale valore si verifica** all'interno del set di dati.
 - **Scopo:** Indica quanto è comune un determinato valore (o classe) nel campione.

Frequenza relativa

- La **frequenza relativa** è la **proporzione** del set di dati totale che è associata a un particolare valore o classe.
 - **Calcolo:** Se f è la frequenza di un valore e n è la dimensione totale del set di dati (o la somma totale delle frequenze), la frequenza relativa si calcola come il **rapporto f/n** .
 - **Significato:** Rappresenta la proporzione (o *fraction*) dei dati che possiedono quel valore.

Frequenza relativa

Età (Classe d'età) (x_i)	N. auto (Frequenza assoluta f_i)
30-40	4
40-50	6
50-60	7
Over 60	3
Totale (N)	20

- Calcolare la frequenza relativa
- Calcolare la frequenza relativa percentuale

Frequenza cumulata

- La frequenza **cumulata** è una misura progressiva che tiene conto delle frequenze accumulate fino a un determinato punto della distribuzione.
 - **Concetto:** La parola "accumulare" significa raccogliere continuamente; la frequenza cumulata raccoglie e **somma tutte le frequenze** delle classi precedenti e della classe attuale.
 - **Calcolo:** Per una data classe, è la somma della sua frequenza assoluta e delle frequenze assolute di tutte le classi che la precedono. La frequenza cumulata della prima classe è sempre uguale alla sua frequenza assoluta.
 - **Proprietà:** Ogni frequenza cumulata è uguale o superiore alla precedente. L'ultima frequenza cumulata deve essere uguale al totale delle osservazioni (N)

Frequenza relativa cumulata

- La frequenza relativa cumulata tiene traccia della **proporzione accumulata** dei dati.
 - **Calcolo:** Si calcola sommando le frequenze relative delle classi precedenti e della classe attuale. Oppure, si ottiene dividendo la frequenza cumulata assoluta per la dimensione totale del campione (N).
 - **Interpretazione:** Indica la proporzione totale del campione che cade fino o sotto un certo valore o classe. Ad esempio, se la frequenza relativa cumulata fino a 900 è 0.40, ciò significa che circa il 40% dei valori dei dati è inferiore o uguale a 900.

Frequenza relativa

Età (Classe d'età) (x_i)	N. auto (Frequenza assoluta f_i)
30-40	4
40-50	6
50-60	7
Over 60	3
Totale (N)	20

- Calcolare la frequenza relativa
- Calcolare la frequenza relativa percentuale

Soluzione

Età (Classe d'età)	N. auto (f_i)	Frequenza Relativa (r_i)	Frequenza Relativa Percentuale ($\%r_i$)	Frequenza Cumulata (F_i)	Frequenza Cumulata Percentuale ($\%R_i$)
30-40	4	0.20	20%	4	20%
40-50	6	0.30	30%	10	50%
50-60	7	0.35	35%	17	85%
Over 60	3	0.15	15%	20	100%
Totale	20	1.00	100%		

la frequenza cumulata percentuale del 50% per la classe d'età 40-50 significa che **il 50% degli acquirenti ha un'età pari o inferiore a 50 anni**

Misure di sintesi dei dati

- Le statistiche descrittive si occupano della descrizione e **della sintesi dei dati**
 - Misure di tendenza centrale
 - Misure di Dispersione
 - Misure di Posizione

Misure di sintesi dei dati

- Le statistiche descrittive si occupano della descrizione e **della sintesi dei dati**
 - **Misure di tendenza centrale**
 - Queste statistiche indicano il "centro" del set di dati

Media

- La media campionaria è la media aritmetica di un set di dati
- somma di tutti i valori del campione divisa per il numero di osservazioni (n)
- Se $x_1, x_2, \dots, x_n$ sono i valori del campione, la media campionaria è:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

Media

- la **media ponderata** è una media in cui **non tutti i valori contano allo stesso modo**, ma ognuno ha un **peso** (o "importanza") diverso.

$$\bar{x} = \frac{\sum w_i x_i}{\sum w_i}$$

- Quando i dati sono **raggruppati in frequenze**, cioè quando per ogni valore x_i sappiamo **quante volte si ripete**, allora le frequenze f_i **giocano il ruolo dei pesi**.
- ogni valore contribuisce **in proporzione alla sua frequenza** → cioè ai **"pezzi di realtà"** che rappresenta.

Media

- dati **raggruppati in classi di intervallo**, non conosciamo i singoli valori osservati, ma solo **quante osservazioni** f_i cadono in ciascun intervallo.
- si usa il **punto medio** (o **classe centrale**) di ciascun intervallo come rappresentante di tutti i valori in quella classe.

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i}$$

Media

Classe di età (anni)	Punto medio x_i	Numero di persone f_i	$f_i x_i$
20 - 25	22,5	80	1 800
25 - 30	27,5	100	2 750
30 - 35	32,5	90	2 925
35 - 40	37,5	30	1 125
Totale		N = 300	8 600

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i} = \frac{8600}{300} = 28,67$$

Media

- **Input, durata in ore di 10 transistor**

[112, 121, 126, 108, 141, 104, 136, 134, 118,
129]

$$M = \frac{112 + 121 + \dots + 129}{10} = 122.9$$

Media

N. stanze (x_i)	N. abitazioni (frequenza f_i)
2	70
3	200
4	150
5	50
6	30
Totale	N = 500

- non serve calcolare il **punto medio** (perché ogni x_i è già un valore esatto)
- basta moltiplicare ciascun valore per la sua frequenza e dividere per il totale

Media

Superficie (in mq) (Classe)	N. di locali commerciali (Frequenza f_i)
400 - 1000	351
1000 - 2000	393
2000 - 3000	77
3000 - 5000	39
Totale (N)	860

- serve calcolare il **punto medio**
- basta moltiplicare ciascun valore per la sua frequenza e dividere per il totale

Mediana

- Valore centrale di un set di dati (ordinati)
- valore che occupa la posizione centrale, se n è **dispari**
- Media dei due valori centrali, se n è **pari**
- Nei dati **discreti** , la mediana si trova **guardando la frequenza cumulata.**

Mediana

- Nei dati raggruppati, si calcola per interpolazione

$$M_e = L + \left(\frac{\frac{N}{2} - F_c}{f_m} \right) \cdot h$$

dove:

L = limite inferiore della **classe mediana**

N = totale delle frequenze

F_c = somma delle frequenze **precedenti** la classe mediana

f_m = frequenza della classe mediana

h = ampiezza della classe

Interpolazione

- L'interpolazione serve per **ricavare un valore "interno"** a un intervallo, quando non lo conosciamo esattamente ma sappiamo come varia la grandezza ai suoi estremi
- Quando i dati sono **raggruppati in classi non conosciamo i valori singoli**, ma solo:
 - L'ampiezza della classe
 - La frequenza di ogni classe
- Stimare (interpolare) il punto preciso dentro la classe che contiene il valore cercato

Mediana

$$M_e = L + \left(\frac{\frac{N}{2} - F_c}{f_m} \right) \cdot h$$

dove:

L = limite inferiore della **classe mediana**

N = totale delle frequenze

F_c = somma delle frequenze **precedenti** la classe mediana

f_m = frequenza della classe mediana

h = ampiezza della classe

Classe di età (anni)	Numero di persone f_i	Cumulata
20 - 25	80	80
25 - 30	100	180
30 - 35	90	270
35 - 40	30	300
Totale	N = 300	

- Totale $N = 300$
- La **mediana** è il valore che divide il campione a metà, la posizione mediana è: $N/2 = 150$
- La 150^a osservazione cade nella **classe 25-30**, perché: fino alla prima classe (20-25) abbiamo 80 osservazioni, fino alla seconda (25-30) arriviamo a 180.
- la 150^a è dentro la classe 25-30.

Mediana

$$M_e = L + \left(\frac{\frac{N}{2} - F_c}{f_m} \right) \cdot h$$

dove:

L = limite inferiore della **classe mediana**

N = totale delle frequenze

F_c = somma delle frequenze **precedenti** la classe mediana

f_m = frequenza della classe mediana

h = ampiezza della classe

Classe di età (anni)	Numero di persone f_i	Cumulata
20 - 25	80	80
25 - 30	100	180
30 - 35	90	270
35 - 40	30	300
Totale	N = 300	

stimare dove si trova esattamente **dentro** quella classe con interpolazione lineare:

L = limite inferiore della **classe mediana** → 25

N = totale delle frequenze → 300

F_c = somma delle frequenze **precedenti** la classe mediana → 80

f_m = frequenza della classe mediana → 100

h = ampiezza della classe → 30-25 = 5 anni

Mediana

$$M_e = L + \left(\frac{\frac{N}{2} - F_c}{f_m} \right) \cdot h$$

dove:

L = limite inferiore della **classe mediana**

N = totale delle frequenze

F_c = somma delle frequenze **precedenti** la classe mediana

f_m = frequenza della classe mediana

h = ampiezza della classe

Classe di età (anni)	Numero di persone f_i	Cumulata
20 - 25	80	80
25 - 30	100	180
30 - 35	90	270
35 - 40	30	300
Totale	N = 300	

Mediana = 28.5

L'interpolazione serve a "entrare dentro" la classe 25-30 e **stimare** il valore esatto (28,5) in cui si trova la 150^a persona, **assumendo che i dati siano distribuiti uniformemente** dentro la classe.

Mediana

- **Input, durata in ore di 10 transistor**

[112, 121, 126, 108, 141, 104, 136, 134, 118,
129]

Ordiniamo i dati:

[104, 108, 112, 118, 121, 126, 129, 134, 136, 141]

Mediana

N. stanze (x_i)	N. abitazioni (frequenza f_i)	Frequenza cumulata
2	70	70
3	200	270
4	150	420
5	50	470
6	30	500
Totale	N = 500	

Mediana

Superficie (in mq) (Classe)	N. di locali commerciali (Frequenza f_i)
400 - 1000	351
1000 - 2000	393
2000 - 3000	77
3000 - 5000	39
Totale (N)	860