Elementi di statistica

- Definizioni
- Popolazione, campione
- Rappresentazione dei dati
- Analisi di frequenza
- Indici descrittivi
- Box plot

Statistica

- Molti fenomeni presentano un carattere di variabilità e casualità
- Esempi
  - Temperatura massima in un giorno dell'anno
  - Altezza di precipitazione in un dato mese
  - Numero di eruzioni vulcaniche dell'Etna in un secolo
  - Popolazione residente in una certa area
- Generalmente si assume che tale variabilità non possa essere interpretata in maniera deterministica e si ricorre quindi allo strumento statistico per descriverne il carattere sulla base di una serie di osservazioni

Statistica

In termini generali, i metodi statistici possono essere suddivisi in:

- Statistica descrittiva
- Teoria delle probabilità
- Statistica matematica

Statistica

- **Popolazione**: insieme di tutte le possibili modalità con cui un fenomeno casuale può manifestarsi
- **Campione**: insieme limitato di osservazioni estratte da una popolazione
- Esempio: lancio di un dado
  - Popolazione {1,2,3,4,5,6}
  - Campione di dimensione 3: 2,5,4
- Esempio: temperatura media del mese di Gennaio a Catania:
  - Popolazione: -273.16 °C ÷ +∞
  - Campione osservato nel periodo 1986-1996:
    13.3, 14.5, 16.4, 14.7, 13.4, 13.2, 13.7, 13.8, 14.4, 14.0, 15.4 °C
- Esempio: precipitazione annua nella stazione di Caltanissetta:
  - Popolazione: 0 ÷ +∞
  - Campione osservato nel periodo 1990-2000
    475, 687.6, 533.2, 376.2, 453.7, 357.2, 822.2, 618.4, 385, 390.2, 473.6 mm

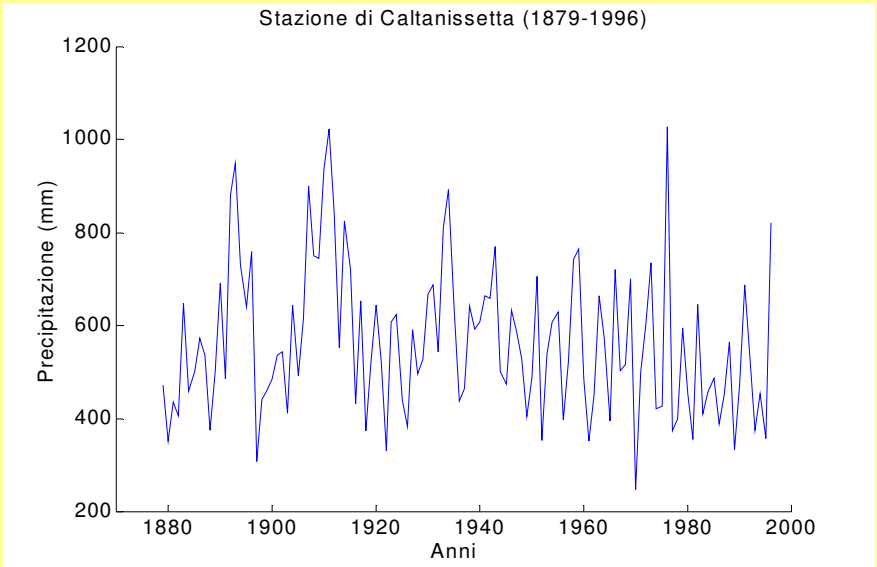
Statistica descrittiva

- La statistica descrittiva si pone l'obiettivo di descrivere in maniera sintetica un campione osservato di dati
- La descrizione viene generalmente effettuata in diversi modi:
  - Rappresentazioni grafiche
  - Analisi di frequenza
  - Calcolo di indici descrittivi
- La statistica descrittiva ha anche la funzione di fornire le informazioni sulla base delle quali effettuare "inferenze" sulla popolazione, attraverso la teoria delle probabilità e la matematica statistica

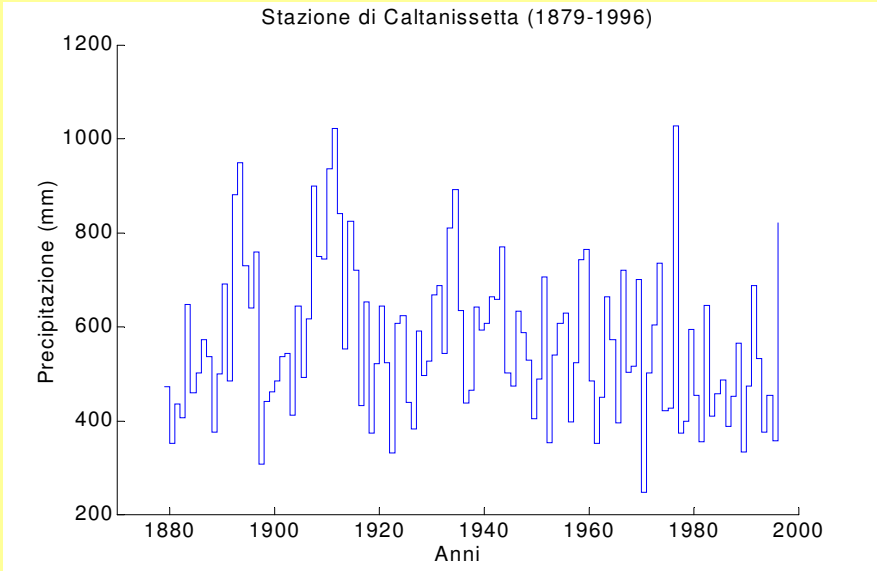
Rappresentazione dei dati

- Un campione osservato di dati si presenta *generalmente* come una successione (cronologica) di valori della variabile in esame
- Esempio:
  - Serie delle temperature massime annue registrate nella stazione di Palermo nel periodo 1921-1990
  - Serie del numero di giorni piovosi in una anno registrati nella stazione di Catania nel periodo 1950-1996
  - Serie delle portate medie (deflussi) mensili del Simeto alla stazione di Giarretta
- Contro esempio:
  - Porosità dello strato superficiale del terreno in una ampia area

Rappresentazione dei dati



Rappresentazione dei dati

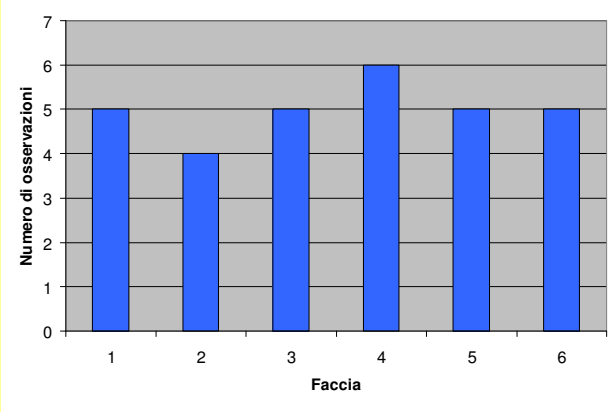


Analisi di frequenza

- L'analisi di frequenza si pone l'obiettivo di valutare la frequenza con cui le osservazioni tendono a ripetersi
- Se la variabile è discreta, le frequenza di ciascuno dei valori della variabile può essere calcolata contando le occorrenze delle osservazioni nel campione
- Tali frequenze possono essere riportate nel cosidetto istogramma di frequenza

Analisi di frequenza

- Esempio: lancio di un dado non truccato 30 volte:
  - No. di 1: 5
  - No. di 2: 4
  - No. di 3: 5
  - No. di 4: 6
  - No. di 5: 5
  - No. di 6: 5



Analisi di frequenza

- Se la variabile è continua (come è il caso in molti fenomeni naturali) non ha senso contare il ripetersi di ciascun valore poiché ciascuna occorrenza è generalmente unica
- E' preferibile quindi suddividere il campo di variazione delle osservazioni in classi di ampiezza finita e contare il numero di osservazioni ricadenti in ciascuna classe

Analisi di frequenza

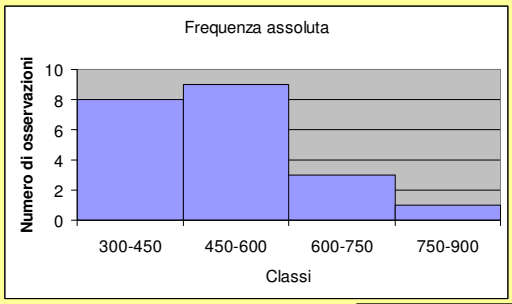
Esempio: serie di precipitazione annue osservate nella stazione di Caltanissetta nel periodo 1980-2000 (21 anni)

Anno	Precipitazione (mm)
1980	454.0
1981	356.2
1982	645.4
1983	409.8
1984	458.6
1985	487.2
1986	387.6
1987	452.6
1988	565.8
1989	332.8
1990	475.0
1991	687.6
1992	533.2
1993	376.2
1994	453.7
1995	357.2
1996	822.2
1997	618.4
1998	385.0
1999	390.2
2000	473.6

Consideriamo 4 classi di ampiezza 150 mm: 300-450, 450-600, 600-750, 750-900

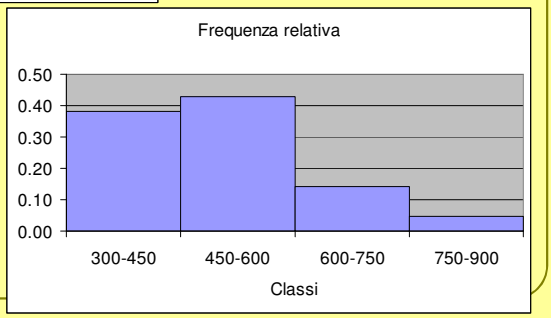
300-450	332.8, 356.2, 357.2, 376.2, 385.0, 387.6, 390.2, 409.8	8
450-600	452.6, 453.7, 454.0, 458.6, 473.6, 475.0, 487.2, 533.2, 565.8	9
600-750	618.4, 645.4, 687.6	3
750-900	822.2	1

Istogramma di frequenza assoluta e relativa



L'istogramma di frequenza assoluta riporta il numero di osservazioni che ricadono in ciascuna classe

L'istogramma di frequenza relativa riporta il numero di osservazioni che ricadono in ciascuna classe in rapporto al numero totale di osservazioni



Calcolo delle frequenze assolute e relative

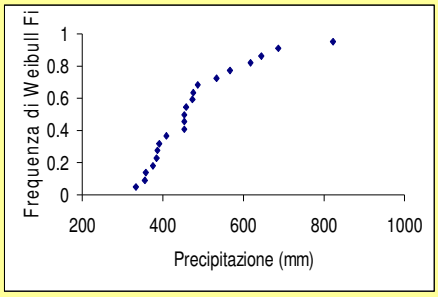
- Ordinare i dati in ordine crescente
- Suddividere il campione in k classi di uguale ampiezza. K può essere calcolato con:
  - Formula di Sturges $K=1+3.3*Log(N)$
  - $K=N^{1/2}$
  - $5 \leq K \leq 25$
- Contare il numero di osservazioni n_i che ricadono nella i-esima classe
- Frequenza assoluta: n_i
- Frequenza relativa: $f_i = n_i/N$

Frequenza cumulata di Weibull

- Consente di calcolare la frequenza di non superamento di un dato osservato
- Ordinati i dati in ordine crescente, la frequenza di non superamento dell'i-esimo valore è data da:

$$Fi = i/(N+1)$$ con N dimensione del campione

Numero d'ordine i	Precipitazione (mm)	Fi
1	332.8	0.0455
2	356.2	0.0909
3	357.2	0.1364
4	376.2	0.1818
5	385.0	0.2273
6	387.6	0.2727
7	390.2	0.3182
8	409.8	0.3636
9	452.6	0.4091
10	453.7	0.4545
11	454.0	0.5000
12	458.6	0.5455
13	473.6	0.5909
14	475.0	0.6364
15	487.2	0.6818
16	533.2	0.7273
17	565.8	0.7727
18	618.4	0.8182
19	645.4	0.8636
20	687.6	0.9091
21	822.2	0.9545



Indici descrittivi

- Gli indici descrittivi consentono di sintetizzare alcuni aspetti essenziali della della distribuzione dei dati in esame
- Essi vengono distinti in:
  - Indici di tendenza centrale
  - Indici di dispersione (variabilità)
  - Indici di forma

Indici di tendenza centrale

Gli indici di tendenza centrale forniscono dei valori attorno ai quali si può ritenere concentrata la variabile statistica in esame, fornendo così una idea (molto) sintetica del fenomeno investigato

Con riferimento ad una serie di osservazioni $x_1, x_2, \ldots, x_n$ possiamo calcolare:

- Media aritmetica: $\bar{x} = \frac{1}{n}\sum_{t=1}^{n} x_t$
- Mediana: valore che non viene superato nel 50% dei casi
  - Ordinati i dati in ordine crescente, valore corrispondente al valore centrale se N è dispari, alla media dei due valori centrali se N è pari
- Moda: valore che si presenta con maggiore frequenza
  - Se la variabile è continua può essere calcolata come la media degli estremi della classe cui corrisponde la massima frequenza assoluta o relativa

Media, mediana e moda generalmente non coincidono!

Indici di tendenza centrale

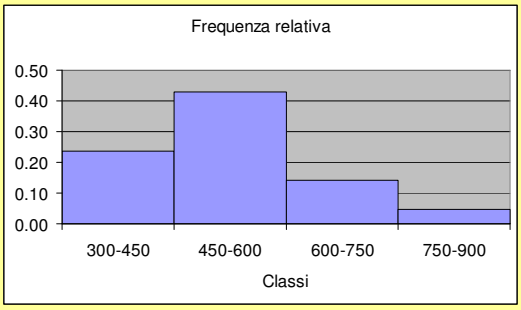
Esempio: serie di precipitazioni annue osservate nella stazione di Caltanissetta nel periodo 1980-2000

Anno	Precipitazione (mm)
1980	454.0
1981	356.2
1982	645.4
1983	409.8
1984	458.6
1985	487.2
1986	387.6
1987	452.6
1988	565.8
1989	332.8
1990	475.0
1991	687.6
1992	533.2
1993	376.2
1994	453.7
1995	357.2
1996	822.2
1997	618.4
1998	385.0
1999	390.2
2000	473.6

Media=482 mm

Mediana=454 mm

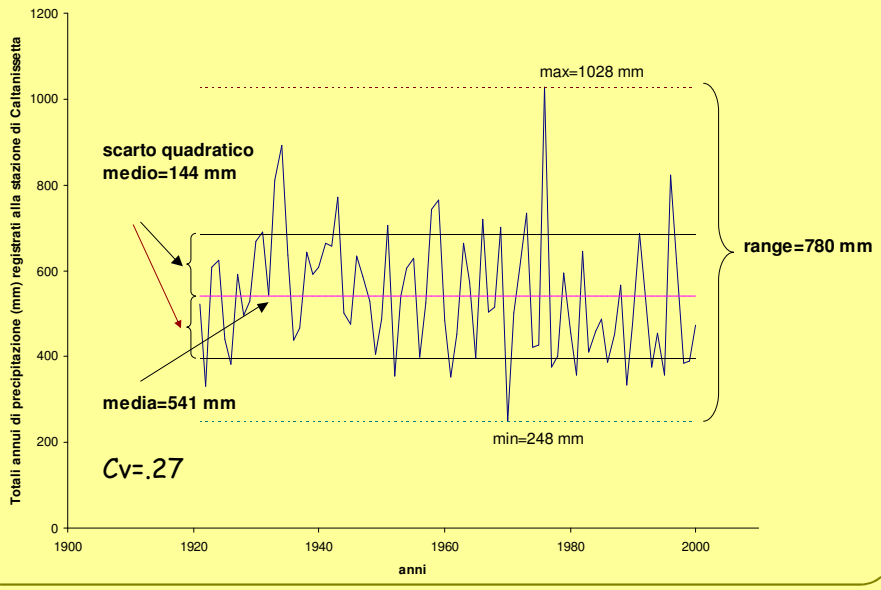
Moda= 525 mm



Indici di dispersione

Misurano la dispersione dei dati attorno ai valori centrali:

- Ampiezza del campione o range: $R=x_{max}-x_{min}$
  - Indica la variabilità totale dei dati, cioè gli estremi dell'intervallo di valori che la variabile assume
- Scarto assoluto medio $D=\frac{1}{n}\sum_{t=1}^{n}|x_t-\bar{x}|$
  - Indica lo scostamento medio (in valore assoluto) dalla media aritmetica
- Varianza campionaria: $S^2=\frac{1}{n-1}\sum_{t=1}^{n}(x_t-\bar{x})^2$
  - Indica la dispersione attorno alla media
- Scarto quadratico medio: $S=\sqrt{\frac{1}{n-1}\sum_{t=1}^{n}(x_t-\bar{x})^2}$
  - E' l'operatore di dispersione per eccellenza e misura la dispersione attorno alla media (stesse dimensioni)
- Coefficiente di variazione: $Cv=\frac{S}{\bar{x}}$
  - Coefficiente adimensionale di dispersione

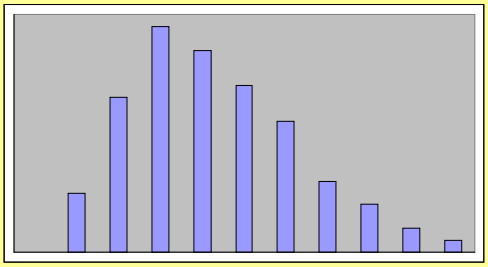


Indici di asimmetria

- Gli indici di forma caratterizzano la forma della distribuzione dei dati intorno alla moda
- Se la distribuzione è simmetrica:
  - Moda=media=mediana
- Se la distribuzione è asimmetrica a **sinistra** (maggiore estensione dell'istogramma per valori **maggiori** della moda):
  - Moda < mediana < media
- Se la distribuzione è asimmetrica a **destra** (maggiore estensione dell'istogramma per valori **minori** della moda):
  - Moda > mediana > media
- L'indice di asimmetria generalmente utilizzato è:

$$g = \frac{\sum_{t=1}^{n} (x_t - \bar{x})^3}{nS^3}$$

Indici di asimmetria

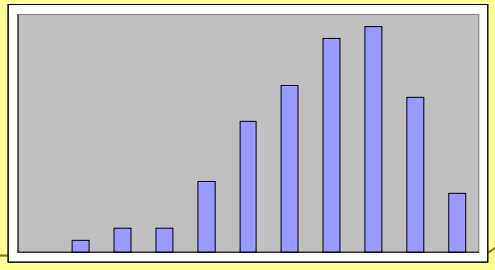


Distribuzione asimmetrica a sinistra

g>0

Distribuzione asimmetrica a destra

g<0



Indici di asimmetria

- Altri indici che possono essere utilizzati per caratterizzare l'asimmetria:
  - Indice di Pearson

$$g_p = \frac{\bar{x} - \text{moda}}{S}$$

$g_p > 0$ distribuzione asimmetrica a sinistra

$g_p < 0$ distribuzione asimmetrica a destra

Quantili

- Si definisce quantile corrispondente ad una frequenza q il valore della variabile osservata che non viene superato o eguagliato nel q% dei casi
- La mediana è il quantile 50%
- Per calcolare il quantile corrispondente ad una data frequenza q occorre:
  - ordinare gli n dati in ordine crescente
  - il quantile x_q è la media tra valori con numero d'ordine $q^*(n+1)-1$ e $q^*(n+1)$ (approssimati per eccesso)

Calcolo dei quantili

Esempio: serie di precipitazioni annue osservate nella stazione di Caltanissetta nel periodo 1980-2000 (21 anni)

Anno	Precipitazione (mm)	Numero d'ordine	Precipitazione (mm)
1980	454.0	1	332.8
1981	356.2	2	356.2
1982	645.4	3	357.2
1983	409.8	4	376.2
1984	458.6	5	385.0
1985	487.2	6	387.6
1986	387.6	7	390.2
1987	452.6	8	409.8
1988	565.8	9	452.6
1989	332.8	10	453.7
1990	475.0	11	454.0
1991	687.6	12	458.6
1992	533.2	13	473.6
1993	376.2	14	475.0
1994	453.7	15	487.2
1995	357.2	16	533.2
1996	822.2	17	565.8
1997	618.4	18	618.4
1998	385.0	19	645.4
1999	390.2	20	687.6
2000	473.6	21	822.2

Quantile x_{25} (detto quartile)

$i_1=q*(n+1)=.25*22=5.5\approx6$

$i_2=i_1-1=5$

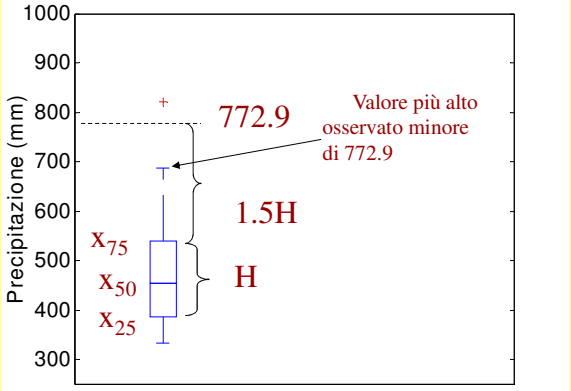
$x_{25}=(385+387.6)/2=386.3$

Box-plot

- Il box and whisker plot (letteralmente grafico a scatola e baffi) è una rappresentazione sintetica ed efficace della distribuzione dei dati
- Presenta il vantaggio di potere facilmente confrontare diverse distribuzioni relative a diverse serie di dati
- La sua costruzione è basata su tre grandezze
  - mediana
  - x_{25}
  - x_{75}

Costruzione del box and whisker plot

Numero d'ordine	Precipitazione (mm)
1	332.8
2	356.2
3	357.2
4	376.2
5	385.0
6	387.6
7	390.2
8	409.8
9	452.6
10	453.7
11	454.0
12	458.6
13	473.6
14	475.0
15	487.2
16	533.2
17	565.8
18	618.4
19	645.4
20	687.6
21	822.2



Confronto tra le precipitazioni annue delle stazioni di Caltanissetta, Catania e Linguaglossa nel periodo 1921-2000