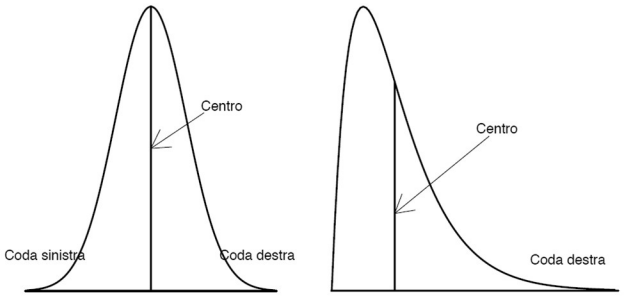
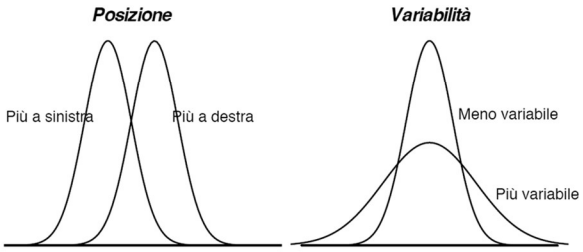
Aspetti notevoli delle distribuzioni

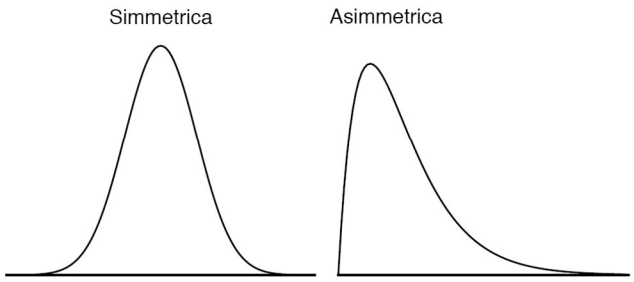
- Posizione
- Variabilità
- Forma



Posizione e variabilità



Forma



- Le misure di tendenza centrale sono particolarmente utili quando vogliamo confrontare due distribuzioni per rispondere a domande del tipo: "*Gli uomini sono più depressi delle donne?*" o "*Agli extra-comunitari arrestati per avere commesso un reato vengono inflitte pene più pesanti che ai cittadini italiani?*"
- Le misure di dispersione sono utili per rispondere a domande del tipo "*C'è una maggiore variabilità nel reddito pro-capite nel nord o nel sud Italia?*"

- Gli indici di posizione consentono una sintesi della distribuzione attraverso un valore rappresentativo.
- **Distribuzioni sconnesse**: moda
- **Distribuzioni ordinate**: mediana
- **Seriazioni**: media aritmetica

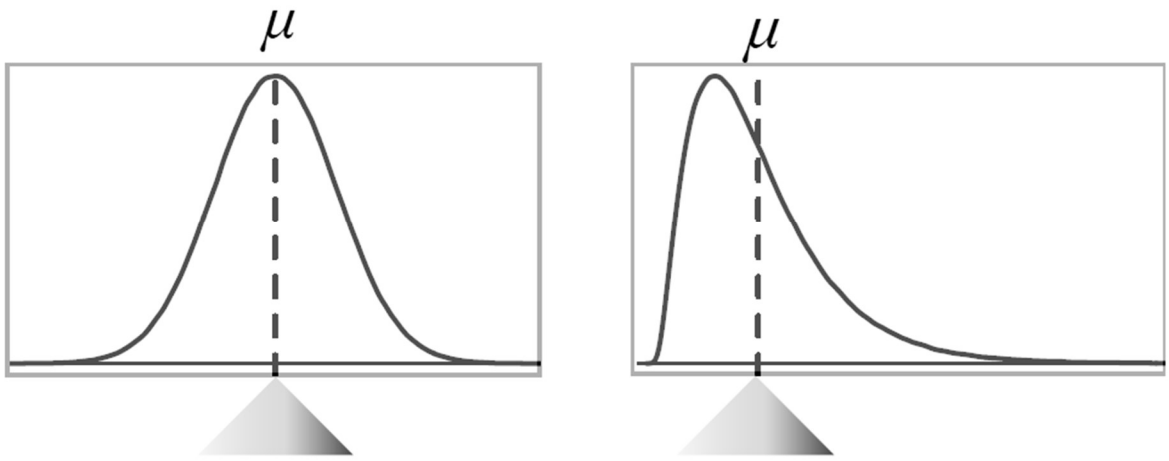
Media aritmetica

- La misura di tendenza centrale più comunemente usata nella statistica descrittiva univariata quantitativa è la media aritmetica.
- **Media del campione**: si indica con $\bar{x}$ la media aritmetica delle modalità che una v.s. quantitativa X ha esibito in un campione di n osservazioni;

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

- Se ciascun dato, x_i, fosse un punto sulla linea dei numeri reali, allora $\bar{x}$ rappresenterebbe il punto di equilibrio.

- **Media della popolazione**: si indica con μ la media aritmetica delle modalità che una v.s. quantitativa X *assume all'interno della popolazione.*



Teorema della somma degli scarti

Data la variabile X che presenta le n modalità $x_1, x_2, \ldots, x_n$, la somma degli scarti di ciascuna modalità dalla propria media aritmetica vale zero.

Dimostrazione.

$$\sum_{i=1}^{n} (x_i - \bar{x}) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \bar{x} = n\bar{x} - n\bar{x} = 0$$

□

Teorema della media di una trasformazione lineare

Data la variabile X che presenta le n modalità $x_1, x_2, \ldots, x_n$, ed avente media $\bar{x}$, se consideriamo la trasformazione lineare $Y = a + bX$ si avrà allora che $\bar{y} = a + b\bar{x}$.

Dimostrazione.

$$\bar{y} = \frac{\sum y_i}{n} = \frac{\sum (a + bx_i)}{n} = \frac{na}{n} + b\frac{\sum x_i}{n} = a + b\bar{x}$$

□

Teorema della devianza

Data la variabile X che presenta le n modalità $x_1, x_2, \ldots, x_n$, la quantità $\sum (x_i - c)^2$ avrà il suo valore *minimo* se e solo se $c = \bar{x}$. Il valore $\sum (x_i - \bar{x})^2$ prende il nome di *devianza*.

Dimostrazione.

Riscriviamo la quantità

$$\sum (x_i - c)^2 ,$$

come

$$\sum \left[(x_i - \bar{x}) + (\bar{x} - c) \right]^2 ;$$

sviluppiamo il quadrato,

$$\sum (x_i - \bar{x})^2 + \sum (\bar{x} - c)^2 + 2 \sum (x_i - \bar{x}) (\bar{x} - c) .$$

□

Dimostrazione.

Essendo la quantità $(\bar{x} - c)$ costante, e per il teorema della somma degli scarti, possiamo scrivere il terzo addendo come

$$2\,(\bar{x} - c) \sum (x_i - \bar{x}) = 0,$$

ottenendo

$$\sum (x_i - \bar{x})^2 + \sum (\bar{x} - c)^2\,,$$

che avrà il valore più basso solo quando $c = \bar{x}$

□

Mediana

- La media aritmetica non è sempre l'indice che meglio rappresenta la tendenza centrale di una distribuzione: se la distribuzione è asimmetrica, la mediana è più adeguata della media quale misura di tendenza centrale.
- La mediana rappresenta la modalità che, una volta ordinate nel senso non decrescente le n unità di P rispetto alle modalità medesime, è posseduta da quella che occupa il posto centrale, ovvero che lascia alla sua destra ed alla sua sinistra un numero uguale di unità.
- Se n è dispari la mediana sarà data dalla modalità a cui corrisponde l'unità statistica di posto $\frac{n+1}{2}$.
- Se n è pari non è detto che la mediana sia univocamente determinata in quanto essa sarà data dalle modalità a cui corrispondono le unità statistiche di posto $\frac{n}{2}$ e $\frac{n}{2} + 1$. **Se queste due modalità sono diverse un procedimento per il calcolo della mediana sarà quello di effettuare la loro media aritmetica**.
- La mediana gode di una importante proprietà che è quella di minimizzare la somma degli scarti assoluti dei valori ossia:

$$\sum_{i=1}^{n} |x_i - M_e| = min$$

- La mediana M_e resta invariata se si sostituiscono i termini $x < M_e$ o $x > M_e$: la mediana non risente di valori anomali.
- Applicabile anche per v.s. ordinali.

Quantili

- I quantili più utilizzati sono i quartili, che dividono la distribuzione in 4 parti uguali e vengono di solito indicati con Q_1, $Q_2 = Me$ e Q_3, i decili, che dividono la distribuzione in 10 parti uguali, ed i percentili, che dividono la distribuzione in 100 parti uguali.
- Chiaramente il 25-esimo percentile è pari a Q_1, il 75-esimo è pari a Q_3, mentre il 50-esimo percentile ed il 5^o decile coincidono entrambi con la mediana $Me = Q_2$.
- Il **primo quartile** Q_1 è quel valore tale che il 25% delle osservazioni ha un valore più piccolo, mentre il restante 75% ha un valore più grande. Cioè Q_1 è l'osservazione di posto $\frac{n+1}{4}$ nell'insieme ordinato dei dati osservati.
- Il **terzo quartile** Q_3 è quel valore tale che il 75% delle osservazioni ha un valore più piccolo, mentre il restante 25% ha un valore più grande. Cioè Q_3 è l'osservazione di posto $\frac{3(n+1)}{4}$ nell'insieme ordinato dei dati osservati.
- Si possono utilizzare procedure diverse per il calcolo dei quantili.

Quantili: 3 situazioni di calcolo

1. Se la posizione $\frac{(n+1)}{4}$ nel caso di Q_1 e $\frac{3(n+1)}{4}$ nel caso di Q_3 è un numero intero, il quartile ha il valore dell'osservazione corrispondente.
2. Se la posizione $\frac{(n+1)}{4}$ nel caso di Q_1 e $\frac{3(n+1)}{4}$ nel caso di Q_3 è a metà tra due numeri interi, si adotta la convenzione di scegliere come quartile la media (delle modalità) delle osservazioni corrispondenti.
3. Se nel calcolo della posizione Se la posizione $\frac{(n+1)}{4}$ nel caso di Q_1 e $\frac{3(n+1)}{4}$ nel caso di Q_3 non cadiamo in uno dei casi precedenti, cioè la posizione non risulta essere né un numero intero né a metà tra due numeri interi, allora si adotta la convenzione di approssimarla per difetto o per eccesso all'intero più vicino e di scegliere come quartile il valore (della modalità) dell'osservazione corrispondente.

Illustrazione

- Supponiamo di misurare una variabile x (almeno su scala di modalità ordinale) su $n = 10$ unità. I dati ordinati di x sono:

$$x_{sort} = \{2.0; 2.9; 3.3; 4.9; 5.2; 6.4; 7.6; 8.1; 9.0; 11.5\}$$

- Dal momento che $\frac{(n+1)}{4} = \frac{(11)}{4} = 2.75$ e $\frac{3(n+1)}{4} = \frac{33}{4} = 8.25$, Q_1 assumerà il valore dell'osservazione di posto 3, mentre Q_3 quello dell'osservazione di posto 8:

$$Q_1 = x_{sort_3} = 3.3$$

e

$$Q_3 = x_{sort_8} = 8.1$$

Interpretazione dei quantili

Leinhardt e Wasserman (1979) riprendono i dati relativi ai tassi di mortalità infantile riportati dal quotidiano The New York Times il 28 settembre 1975. Ciascuna osservazione (riga) è una nazione. Le variabili (colonne) sono:

- il reddito pro-capite in dollari,
- la mortalità infantile per 1000 nati vivi,
- il continente (variabile qualitativa),
- una variabile qualitativa e che riguarda il fatto che quella nazione esporti (modalità $oil = yes$) o meno ($oil = no$) il petrolio.

		income	infant	region	oil
1	Austria	3350	23.7	Europe	no
2	Belgium	3346	17.0	Europe	no
3	Denmark	5029	13.5	Europe	no
4	Finland	3312	10.1	Europe	no
5	France	3403	12.9	Europe	no
6	West.Germany	5040	20.4	Europe	no
7	Ireland	2009	17.8	Europe	no
8	Italy	2298	25.7	Europe	no
9	Japan	3292	11.7	Europe	no
10	Netherlands	4103	11.6	Europe	no
11	Norway	4102	11.3	Europe	no
12	Portugal	956	44.8	Europe	no
13	Sweden	5596	9.6	Europe	no
14	Switzerland	2963	12.8	Europe	no
15	Britain	2503	17.5	Europe	no
16	Greece	1760	27.8	Europe	no
17	Spain	1256	15.1	Europe	no
18	Yugoslavia	406	43.3	Europe	no

```
[1]  9.6 10.1 11.3 11.6 11.7 12.8 12.9 13.5 15.1
[10] 17.0 17.5 17.8 20.4 23.7 25.7 27.8 43.3 44.8
```

- Il primo e il terzo quartile si trovano rispettivamente nella posizione $(18+1)1/4 = 4.75 \approx 5$ e $(18+1)3/4 = 14.25 \approx 14$:

$$Q_1 = x_5 = 11.7$$

$$Q_3 = x_{14 \approx 14} = 23.7$$

- In conclusione, nel 1975 la metà centrale delle nazioni europee (Giappone incluso) rivela una mortalità infantile compresa tra i 12 ed i 23 casi, per ogni 1000 nati vivi.
- Un quarto delle nazioni europee ha una mortalità infantile minore di 12.
- Un quarto delle nazioni europee ha una mortalità infantile maggiore di 23.

Varianza

- Benché il sommario dei cinque numeri fornisca un'utile descrizione numerica di una distribuzione, è più comune usare la media quale misura di tendenza centrale e la varianza quale misura di dispersione.

La varianza ci fornisce una misura della dispersione della distribuzione:

$$s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$$

Deviazione standard

La deviazione standard è la radice quadrata della varianza. È espressa nella stessa unità di misura dei dati originari:

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

- La deviazione standard s misura la dispersione attorno alla media e dovrebbe essere usata soltanto quando la media è adeguata per misurare il centro della distribuzione (ovvero, nel caso di distribuzioni simmetriche).
- Quando tutte le osservazioni sono uguali, $s = 0$, altrimenti $s > 0$.
- Come nel caso della media $\bar{x}$, anche la deviazione standard è fortemente influenzata dai valori anomali.

Proprietà della varianza

- La varianza e la deviazione standard non mutano se i dati vengono traslati sommando (o sottraendo) una costante.
- si considerino di dati $x_1, x_2, \ldots, x_n$ e una costante c. Se

$$y_1 = x_1 + c; y_2 = x_2 + c; \ldots; y_n = x_n + c,$$

allora

$$\begin{aligned} s_Y^2 &= \sum \frac{\left((x_i+c) - \sum (x_i+c)/n\right)^2}{n-1} \\ &= \sum \frac{\left(x_i+c - \sum x_i/n - nc/n\right)^2}{n-1} \\ &= \sum \frac{\left(x_i + \cancel{c} - \bar{x} - \cancel{c}\right)^2}{n-1} = s_X^2 \end{aligned}$$

- Varianza e la deviazione standard sono invece influenzate da un cambiamento della scala di misura.
- Se

$$y_1 = cx_1; y_2 = cx_2; \ldots; y_n = cx_n,$$

allora

$$\begin{aligned} s_Y^2 &= \sum \frac{\left(cx_i - \sum (cx_i)/n\right)^2}{n-1} \\ &= \sum \frac{\left(cx_i - c\sum x_i/n\right)^2}{n-1} \\ &= \sum \frac{c^2\left(x_i - \sum x_i/n\right)^2}{n-1} = c^2 s_X^2 \end{aligned}$$

e $s_Y = |c| s_X$

- Si noti il valore assoluto nell'ultima espressione. Non sono possibili varianze o deviazioni standard negative.

Gradi di libertà

- Nella definizione di varianza, la somma dei quadrati degli scarti dalla media viene divisa per $n-1$, non per n come avverrebbe per una semplice media aritmetica.
- La divisione per $n-1$ trova la sua giustificazione nella teoria degli stimatori ed è legata alla nozione di gradi di libertà.
- Nel caso di n dati, $x_1; x_2; \ldots; x_n$, la conoscenza dei primi $n-1$ valori non ci aiuta a conoscere il valore dell'ultimo dato, x_n.
- Se però i dati vengono espressi come scarti dalla media $(x_1 - \bar{x}; x_2 - \bar{x}; \ldots; x_n - \bar{x})$, allora la loro somma deve essere uguale a zero. $[\text{Teorema della somma degli scarti}]$
- Di conseguenza, l'ultimo scarto dalla media sarà uguale al negativo della somma degli scarti degli altri dati dalla media:

$$(x_n - \bar{x}) = -\left[(x_1 - \bar{x}) + (x_2 - \bar{x}) + \ldots + (x_{n-1} - \bar{x})\right]$$

Per i 18 tassi di mortalità dei paesi europei nel file `Leinhardt`, avremo

	infant	infant-MEDIA
Austria	23,7	4,4
Belgium	17	-2,3
Denmark	13,5	-5,8
Finland	10,1	-9,2
France	12,9	-6,4
West.Germany	20,4	1,1
Ireland	17,8	-1,5
Italy	25,7	6,4
Japan	11,7	-7,6
Netherlands	11,6	-7,7
Norway	11,3	-8,0
Portugal	44,8	25,5
Sweden	9,6	-9,7
Switzerland	12,8	-6,5
Britain	17,5	-1,8
Greece	27,8	8,5
Spain	15,1	-4,2
	SOMMA=	**-24,0**
Yugoslavia	43,3	24,0

Diciamo dunque che i $n = 18$ tassi di mortalità hanno 18 gradi di libertà, ma i 18 scarti dalla media hanno solo $n - 1 = 17$ gradi di libertà.

Possiamo inoltra verificare le due proprietà della varianza:

	infant	infant + c	infant x c	infant x 1/s
Austria	23,7	33,7	237,0	3,3
Belgium	17	27	170,0	2,6
Denmark	13,5	23,5	135,0	2,3
Finland	10,1	20,1	101,0	2,0
France	12,9	22,9	129,0	2,2
West.Germany	20,4	30,4	204,0	3,0
Ireland	17,8	27,8	178,0	2,7
Italy	25,7	35,7	257,0	3,5
Japan	11,7	21,7	117,0	2,1
Netherlands	11,6	21,6	116,0	2,1
Norway	11,3	21,3	113,0	2,1
Portugal	44,8	54,8	448,0	5,4
Sweden	9,6	19,6	96,0	1,9
Switzerland	12,8	22,8	128,0	2,2
Britain	17,5	27,5	175,0	2,7
Greece	27,8	37,8	278,0	3,7
Spain	15,1	25,1	151,0	2,5
Yugoslavia	43,3	53,3	433,0	5,2
VARIANZA (s^2) =	**103,6**	**103,6**	**10361,4**	**1,0**
c=	**10**			

Da notare inoltre che una variabile X moltiplicata per una costante $c = \frac{1}{s_X}$, avrà varianza pari ad 1; l'unità di misura nuova è quindi la variabilità attorno alla media.