Statistica

Maura Mezzetti

maura.mezzetti@uniroma2.it

Statistica Descrittiva

2° parte

Le medie

Obiettivo: sintesi della distribuzione di un carattere mediante una modalità "tipica", altrimenti detto valore medio o caratteristico.

Strumenti:

- **Indici di posizione**: moda (tutte le scale); mediana, quantili (scala almeno ordinale)
- **Medie analitiche**: media aritmetica, media troncata (caratteri quantitativi); media geometrica, media armonica (scala di rapporti).

Le medie analitiche sono una funzione esplicita, analitica, delle modalità del carattere.

La media

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

Media Aritmetica

Apartment Rent Data

445	615	430	590	435	600	460	600	440	615
440	440	440	525	425	445	575	445	450	450
465	450	525	450	450	460	435	460	465	480
450	470	490	472	475	475	500	480	570	465
600	485	580	470	490	500	549	500	500	480
570	515	450	445	525	535	475	550	480	510
510	575	490	435	600	435	445	435	430	440

$$\sum_{i=1}^{70} x_i = 34356$$

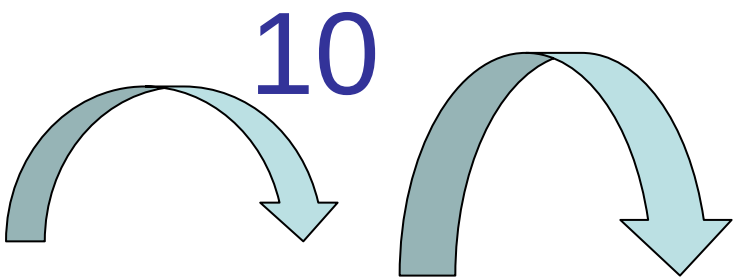
$$\frac{34356}{70} = 490.8$$

Esempio: dati raggruppati (numero di veicoli)

Data la seguente distribuzione del numero di veicoli per famiglia, ottenuta dall'intervista di 100 famiglie. Calcolare la media aritmetica

Numero di veicoli	Numero di famiglie
0	10
1	50
2	10
3	30

Esempio: dati raggruppati (numero di veicoli)



$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{0+\cdots+0+1+\cdots+1+2+\cdots+\ldots}{100}$$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{0\times 10+1\times 50+2\times 10+3\times 30}{100} = \frac{160}{100} = 1.6$$

Esempio: dati raggruppati

$$\bar{x} = \frac{\sum_{j=1}^{m} x_j^* \times n_j}{n} = \sum_{j=1}^{m} x_j^* \times f_j$$

n numero di unità (i=1,..,n)
m numero di modalità (j=1,…,m)

x_j^*	n_j	$n_j \times x_j^*$
0	10	0
1	50	50
2	10	20
3	30	90
	100	**160**

$\bar{x} = 1.6$

Esempio: dati raggruppati

$$\bar{x} = \frac{\sum_{j=1}^{m} x_j^* \times n_j}{n} = \sum_{j=1}^{m} x_j^* \times f_j$$

n numero di unità (i=1,..,n)
m numero di modalità (j=1,…,m)

x_j^*	f_j	$f_j \times x_j^*$
0	0.10	0
1	0.50	0.5
2	0.10	0.2
3	0.30	0.9
		1.6

$\bar{x} = 1.6$

La suddivisione in classi

Arbitrarietà della suddivisione in classi.
Linee guida:

- al fine di facilitare l'interpretrazione della distribuzione, qualora possibile, le classi dovrebbero avere la stessa ampiezza
- evitare di costruire classi caratterizzate da un numero di frequenze molto basso
- equilibrio tra due esigenze in conflitto: sintesi e grado di risoluzione

Attenzione: la suddivisione in classi comporta una perdita di informazioni (le differenze presenti entro la classe).
Tale operazione ha senso soltanto se l'obiettivo finale è produrre una tabella di sintesi o un istogramma.
Per tutti gli altri scopi occorre lavorare con la distribuzione unitaria di partenza.

Esempio: Dati in classe

Considerata la seguente distribuzione del reddito di 100 famiglie

Stipendio	n_j
[0,10)	10
[10,20)	30
[20,30)	20
[30,50]	40

Stipendio	n_j	c_j	$n_j x c_j$
[0,10)	10	5	50
[10,20)	30	15	450
[20,30)	20	25	500
[30,50]	40	40	1600
	100		**2600**

Proprietà della Media Aritmetica

- La m.a. equiripartisce il totale di un carattere tra le unità

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

- Internalità

$$x_{(\min)} \leq \bar{x} \leq x_{(\max)}$$

- La somma degli scarti dalla media aritmetica è nulla:

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

Proprietà media aritmetica

- La m.a. rende minima la somma dei quadrati degli scarti da una costante:

$$\bar{x} = \arg\min_{c} \sum_{i=1}^{n} (x_i - c)^2$$

- Linearità $x_1, x_2, \ldots, x_n$ e $y_1, y_2, \ldots, y_n$ tale che $y_i = a + bx_i$ allora

$$\bar{y} = a + b\bar{x}$$

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{\sum_{i=1}^{n} (a + bx_i)}{n} = \frac{\sum_{i=1}^{n} a + \sum_{i=1}^{n} bx_i}{n}$$

$$\frac{na + b\sum_{i=1}^{n} x_i}{n} = a + b\frac{\sum_{i=1}^{n} x_i}{n} = a + b\bar{x}$$

$$\sum_{i=1}^{n}\left(x_i - c\right)^2$$

$$\frac{d\sum_{i=1}^{n}\left(x_i - c\right)^2}{dc} =$$

$$\sum_{i=1}^{n}\frac{d\left(x_i - c\right)^2}{dc} = \sum_{i=1}^{n} - 2\left(x_i - c\right) = -2\sum_{i=1}^{n} x_i + 2\sum_{i=1}^{n} c$$

$$\sum_{i=1}^{n} \frac{d\left(x_i - c\right)^2}{dc} = -2\sum_{i=1}^{n} x_i + 2\sum_{i=1}^{n} c$$

$$\sum_{i=1}^{n} \frac{d\left(x_i - c\right)^2}{dc} = 0$$

$$\sum_{i=1}^{n} x_i = nc$$

$$c = \bar{x}$$

$$\frac{d^2 \sum_{i=1}^{n} \left(x_i - c\right)^2}{dc} = 2n$$

Proprietà media aritmetica

- La media di un collettivo è la media aritmetica delle medie dei sottogruppi in cui può essere ripartito il medesimo, ponderata per le numerosità relative dei sottogruppi.
- Se $\bar{x}_1$ e $\bar{x}_2$ sono le medie di due campioni di ampiezza rispettivamente n_1 e n_2 la media può essere calcolata come

$$\bar{x} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2}$$

$$\bar{x} = \frac{\sum_{j=1}^{m} n_j \bar{x}_j}{\sum_{j=1}^{m} n_j}$$

Esercizio

In un’azienda gli stipendi annui, in migliaia di euro, sono così distribuiti:

1	direttore	30
3	capi ufficio	20
10	impiegati	16
25	operai	12
30	manovali	10

Calcolare la media aritmetica, la mediana e la moda degli stipendi.

Esempio (stipendio annuo)

- Qual è l’unità statistica?
- Qual è il carattere?
- Cosa rappresenta la parola «impiegati»?

Rappresentazione tabellare

Impiego	**stipendio**	**frequenze**	**x_j*n_j**
Laborer	10	30	300
Worker	12	25	300
Employee	16	10	160
Head office	20	3	60
Director	30	1	30
total		**69**	**850**

$$\bar{x} = \frac{\sum_{i=1}^{69} x_i}{69} = \frac{\sum_{j=1}^{5} x_j^* n_j}{69} = \frac{850}{69} = 12.319$$

Rappresentazione tabellare

Impiego	**stipendio**	**n_j**	**f_j**	**F_j**
Laborer	10	30	0.44	0.44
Worker	12	25	0.36	0.8
Employee	16	10	0.14	0.94
Head office	20	3	0.04	0.98
Director	30	1	0.2	1
total		**69**		

Esercizio

In una stanza ci sono 12 persone con un peso medio pari a 75kg. Se arriva un’altra persona che pesa 60 kg, qual è il peso medio delle 13 persone?

Esercizio

In una stanza ci sono 12 persone con un peso medio pari a 75kg. Se arriva un'altra persona che pesa 60 kg, qual è il peso medio delle 13 persone?

$$\bar{x}_{13} = \frac{\sum_{i=1}^{13} x_i}{13} = \frac{\sum_{i=1}^{12} x_i + x_{13}}{13} = \frac{12\bar{x}_{12} + x_{13}}{13} = \frac{12 \times 75 + 60}{13} = 73.84$$

Esercizio

In una stanza ci sono 6 persone con un peso medio pari a 75kg. Se arriva un'altra persona che pesa 40 kg, qual è il peso medio delle 7 persone?

$$\bar{x}_7 = \frac{\sum_{i=1}^{7} x_i}{7} = \frac{\sum_{i=1}^{6} x_i + x_7}{7} = \frac{6\bar{x}_6 + x_7}{7} = \frac{6 \times 75 + 30}{7} = 70$$

Esercizio

Le temperature della neve in gradi Celsius di una nota località sciistica nel mese di gennaio sono state le seguenti

t_j	-4	-3	-2	-1	0	1
g_j	6	5	8	6	4	2

dove t_j è la temperatura rilevata in gradi Celsius e g_j è il numero di giorni in cui si è registrata la temperatura t_j

Esercizio

- Si calcoli la temperatura media: in gradi Celsius, e in gradi Fahrenheit, dove sapendo che

$$T_{Fahr} = 32 + 1.8\ T_{Cels}$$

t_j	g_j	$t_j \cdot g_j$
-4	6	-24
-3	5	-15
-2	8	-16
-1	6	-6
0	4	0
1	2	2
tot	31	-59

$$\frac{\sum_{j=1}^{6} t_j \cdot g_j}{\sum_{j=1}^{6} g_j}$$

$$\frac{-59}{31} = -1.903$$

Nel periodo di osservazione, la temperatura media della neve nella nota località sciistica è stata pari $-1.903\ C°$. Più precisamente, $-1.903\ C°$ indica la temperatura che si sarebbe dovuta osservare nell'intero mese di gennaio nel caso in cui si fosse avuta la stessa temperatura in ogni giorno.
Si osservi che le relazioni che ci permettono di passare dalle temperature in gradi Celsius a quelle in gradi Fahrenheit e assoluti, sono lineari. In forza della proprietà di linearità della media aritmetica[1] le medie ricercate risultano:

$$\begin{aligned} M_1(T_{Fahrenheit}) &= 32 + 1.8 \cdot M_1(T_{Celsius}) \\ &= 32 + 1.8 \cdot (-1.903) = 28.574 \end{aligned}$$

Esercizio

In una scuola elementare si è misurata l'altezza di 100 bambini di quarta e si è trovato un valore medio di 126 cm. Ci si è accorti però che lo strumento era stato erroneamente posizionato e ciascun bambino è risultato 4 cm più basso, qual è il vero valore dell'altezza media dei bambini?

Esercizio

In una scuola elementare si è misurata l'altezza di 100 bambini di quarta e si è trovato un valore medio di 126 cm. Ci si è accorti però che lo strumento era stato erroneamente posizionato e ciascun bambino è risultato 4 cm più basso, qual è il vero valore dell'altezza media dei bambini?

$$y_i = x_i + 4$$

$$\bar{y} = \bar{x} + 4 = 126 + 4 = 130cm$$

Punti deboli della media aritmetica

- Robustezza. Sensibilità ai valori estremi.
- Rappresentatività nei confronti di distribuzioni asimmetriche. Più avanti argomenteremo che la media aritmetica e un valore di sintesi rappresentativo nei confronti di distribuzioni simmetriche.

Mediana

- La mediana è la modalità pertinente all'unità statistica che occupa la posizione centrale nella distribuzione ordinata delle osservazioni.
- Divide la distribuzione ordinata in due parti ciascuna contenente la meta delle osservazioni.
- Può essere calcolata per i caratteri misurati su scala ordinale e per quelli quantitativi.

Esempio Mediana: numero dispari osservazioni

Dati osservati:	24.1	22.6	21.5	23.7	22.6
Dati ordinati:	21.5	22.6	**22.6**	23.7	24.1
Posizione:	1	2	**3**	4	5

$$\text{Positioning Point} = \frac{n+1}{2} = \frac{5+1}{2} = 3.0$$

$$\text{Mediana} = 22.6$$

Esempio Mediana
Campione pari

Dati grezzi:	10.3	4.9	8.9	11.7	6.3	7.7
Ordinati:	4.9	6.3	**7.7**	**8.9**	10.3	11.7
Posizione:	1	2	**3**	**4**	5	6

$$\text{Positioning Point} = \frac{n+1}{2} = \frac{6+1}{2} = 3.5$$

$$\text{Mediana} = \frac{7.7 + 8.9}{2} = 8.30$$

Mediana

1. Ordinare le osservazioni in ordine crescente
   n = numero di osservazioni.

2a. Se n è dispari, la mediana è l'osservazione che occupa la posizione (n + 1) / 2

2b. Se n è pari, la mediana è la media delle due osservazioni centrali (n/2 e (n/2+1)

Order	Data
1	78
2	91
3	94
4	98
5	99
6	101
7	103
8	105
9	114

← $n = 9$
$(n+1)/2 = 10/2 = 5$
Mediana = 99

$n = 10$
$(n+1)/2 = 5.5$
Mediana = (99+101) /2 = 100

Order	Data
1	78
2	91
3	94
4	98
5	99
6	101
7	103
8	105
9	114
10	121

Median
Apartment Data

Averaging the 35th and 36th data values:

Median = (475 + 475)/2 = 475

425	430	430	435	435	435	435	435	440	440
440	440	440	445	445	445	445	445	450	450
450	450	450	450	450	460	460	460	465	465
465	470	470	472	475	475	475	480	480	480
480	485	490	490	490	500	500	500	500	510
510	515	525	525	525	535	549	550	570	570
575	575	580	590	600	600	600	600	615	615

Note: Data is in ascending order.

Mediana dati raggruppati

Data un collettivo di 100 famiglie, calcola la mediana

Numero di veicoli	Numero di famiglie
0	10
1	50
2	10
3	30

Mediana dati raggruppati

Numero di veicoli	Numero di famiglie	freq	Cumulative freq
0	10	0.1	0.1
1	50	0.5	0.6
2	10	0.1	0.7
3	30	0.3	1

La mediana è la prima modalità la cui frequenza cumulata supera (o è uguale) 0.5

Mediana = 1

Mediana: dati ordinali qualitativi

Rating	Relative Frequency	Percent Frequency
Poor	0.10	10%
Below Average	0.15	15%
Average	0.25	25%
Above Average	0.45	45%
Excellent	0.05	5%
Total	**1.00**	**100%**

Ordinal qualitative data (Maradann Inn example)

Rating	Relative Frequency	Cumulative Frequency
Poor	0.10	0.10
Below Average	0.15	0.25
Average	0.25	0.50
Above Average	0.45	0.95
Excellent	0.05	1
Total	**1.00**	

Median Rating= Average

Caratteri quantitativi suddivisi in classi

Occorre fare riferimento alla distribuzione unitaria di partenza. Altrimenti, non è possibile calcolare la mediana se non in modo approssimativo, sotto l'ipotesi di equidistribuzione del carattere all'interno di ciascuna classe.

Ai fini dell'individuazione della classe entro cui cade la mediana si procede come sopra, facendo riferimento alle frequenze cumulate,

Esempio: Dati in classe

Considerata la seguente distribuzione del reddito di 100 famiglie

Stipendio	n_j
[0,10)	10
[10,20)	30
[20,30)	20
[30,50]	40

Stipendio	n_j	f_j	F_j
[0,10)	10	0.10	0.10
[10,20)	30	0.30	0.40
[20,30)	20	0.20	0.60
[30,50]	40	0.40	1
	100	**1**	

Moda

La moda è la modalità della distribuzione che si presenta con la massima frequenza.
Se la distribuzione di un carattere quantitativo e ripartita in classi di diversa ampiezza e necessario eliminare questa diversità individuando la classe modale.
La classe modale e quella a cui corrisponde la densità di frequenza più elevata.
Una distribuzione può essere unimodale o plurimodale.

Moda

Apartment Data

450 occurred most frequently (7 times)

Mode = 450

425	430	430	435	435	435	435	435	440	440
440	440	440	445	445	445	445	445	450	450
450	450	450	450	450	460	460	460	465	465
465	470	470	472	475	475	475	480	480	480
480	485	490	490	490	500	500	500	500	510
510	515	525	525	525	535	549	550	570	570
575	575	580	590	600	600	600	600	615	615

Note: Data is in ascending order.

Proprià della mediana

La mediana minimizza la somma degli scarti in valore assoluto dei valori di un carattere quantitativo da una costante: definendo

$$S^*(c) = \sum_{i=1}^{n} |x_i - c|,$$

$$\arg\min_c \{S^*(c)\} = M_e.$$

Robustezza. La mediana e resistente, cioe insensibile alla presenza di valori anomali.

Moda dati raggruppati

Numero di veicoli	Numero di famiglie
0	10
1	50
2	10
3	30

Moda dati raggruppati

- La moda del numero di veicoli per famiglia= 1

Numero di veicoli	Numero di famiglie
0	10
1	50
2	10
3	30

Esempio

I seguenti valori si riferiscono ai valori di un titolo rilevati mensilmente:

1.4; 1.7; 2.3; 2.5; 3.2; 3.8

Se il valore 3.8 fosse erroneamente trascritto come 38, quale sarebbe l'effetto sulle misure di posizione calcolate a partire da questi dati?

a) Un incremento della mediana.

b) Un incremento della moda.

c) Un incremento della media aritmetica.

d) Un incremento sia della mediana sia della moda.

e) Un incremento della mediana, della moda e della media aritmetica.

I quantili

I *quantili* sono modalità del carattere che suddividono la distribuzione in q distribuzioni parziali ciascuna contenente $1/q$ della numerosità totale (distribuzioni di frequenza) o della quantità totale (distribuzione di quantità).
Se $q = 10$ si parla di decili; se $q = 5$ di quintili (vedi foto); se $q = 4$ di quartili; se $q = 100$ di percentili.
Ad esempio, i *quartili* ripartiscono la distribuzioni in quattro parti caratterizzate dalla stessa numerosità, pari al 25% della numerosità totale.
La mediana è il 5^o decile, il 2^o quartile e il 50^o percentile.

80th Percentile

"At least 80% of the items take on a value of 542 or less."

56/70 = .8 or 80%

"At least 20% of the items take on a value of 542 or more."

14/70 = .2 or 20%

425	430	430	435	435	435	435	435	440	440
440	440	440	445	445	445	445	445	450	450
450	450	450	450	450	460	460	460	465	465
465	470	470	472	475	475	475	480	480	480
480	485	490	490	490	500	500	500	500	510
510	515	525	525	525	535	549	550	570	570
575	575	580	590	600	600	600	600	615	615

Quartiles

- Quartiles are specific percentiles
- First Quartile = 25th Percentile
- Second Quartile = 50th Percentile = Median
- Third Quartile = 75th Percentile

Third Quartile

Third quartile = 75th percentile

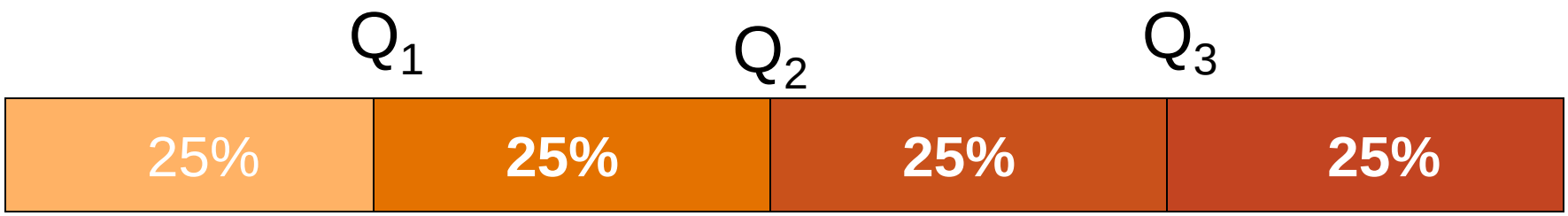
$i = (p/100)n = (75/100)70 = 52.5 = 53$

Third quartile = 525

425	430	430	435	435	435	435	435	440	440
440	440	440	445	445	445	445	445	450	450
450	450	450	450	450	460	460	460	465	465
465	470	470	472	475	475	475	480	480	480
480	485	490	490	490	500	500	500	500	510
510	515	525	525	525	535	549	550	570	570
575	575	580	590	600	600	600	600	615	615

Note: Data is in ascending order.

Quartiles



Q_i element in position $i \times (n+1)/4$

Q_2 median

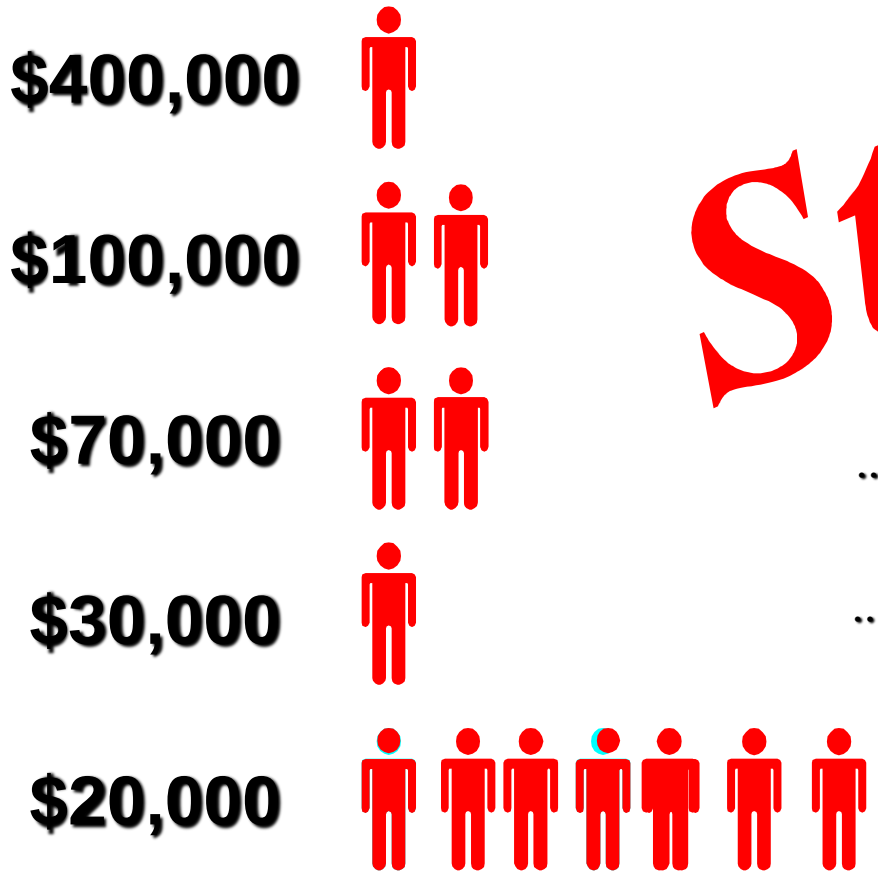
Quartiles grouped data

Number of cars	Number of families	freq	Cumulative freq
0	10	0.1	0.1
1	50	0.5	0.6
2	10	0.1	0.7
3	30	0.3	1

$Q_1=Q_2=1$

$Q_3=3$

Thinking Challenge



Strike!

... employees cite low pay -- most workers earn only $20,000.

... President claims average pay is $70,000!

Introduzione

Obiettivo: ci proponiamo di misurare la variabilità della distribuzione di un carattere.
Questa può essere definita come la tendenza delle unità di un collettivo ad assumere diverse modalità del carattere.
La misura della variabilità dipende dalla natura del carattere.
Tuttavia, un indice di variabilità

- assume il valore minimo se tutte le unità presentano la medesima modalità del carattere (distribuzione degenere)
- risulta crescente al crescere della diversità tra le modalità assunte dalle diverse unità.

1. Misure basate sul confronto dei valori della distribuzione con un valore centrale (medio) - sintesi degli scostamenti o scarti.
2. Misure di mutua variabilità - sintesi di tutti i confronti possibili di ciascun termine della distribuzione con tutti gli altri.
3. Misure basate sugli indici di posizione.

1.1 Varianza

La variabilità delle modalità del carattere viene misurata confrontando ciascuna di esse con un centro della distribuzione, la media aritmetica.
Definizione: per distribuzioni unitarie:

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

Per distribuzioni di frequenze:

$$\sigma^2 = \frac{1}{n}\sum_{j=1}^{K}(x_j^{\star} - \bar{x})^2 n_j = \sum_{j=1}^{K}(x_j^{\star} - \bar{x})^2 f_j$$

La somma dei quadrati degli scarti dalla media viene detta *devianza*.

1.2 Metodo alternativo di calcolo della varianza

La varianza si ottiene sottraendo alla media dei quadrati il quadrato della media.

Distribuzioni unitarie

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

Distribuzioni di frequenza

$$\sigma^2 = \frac{\sum_{j=1}^{k} x_j^{*2} n_j}{n} - \bar{x}^2 = \sum_{j=1}^{k} x_j^{*2} f_j - \bar{x}^2$$

Equivalent Formula

$$\sigma^2 = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n} = \frac{\sum_{i=1}^{n}\left(x_i^2 - 2x_i\bar{x} + \bar{x}^2\right)}{n}$$

$$= \frac{\sum x_i^2 + \sum\left(-2x_i\bar{x}\right) + \sum \bar{x}^2}{n} = \frac{\sum x_i^2 - 2\bar{x}\sum x_i + n\bar{x}^2}{n}$$

$$= \frac{\sum x_i^2 - 2\bar{x}\left(n\bar{x}\right) + n\bar{x}^2}{n} = \frac{\sum x_i^2}{n} - \bar{x}^2$$

1.5 Proprietà della varianza

- $\sigma^2 \geq 0$ (vale zero se la distribuzione è degenere).
- L'unità di misura della varianza è il quadrato della scala originaria. La radice della varianza, σ, è quindi espressa sulla stessa scala ed è nota come *deviazione standard* o *scarto quadratico medio*.
- Trasformazione lineare di un carattere x_i con media $\bar{x}$ e varianza σ^2:

  $$y_i = a + bx_i$$

  La varianza del carattere Y risulta $b^2\sigma^2$

NB La varianza è una media (degli scarti al quadrato). Come tale è sensibile ai valori anomali.

Dimostrazione: per le proprietà della media,

$$\bar{y} = a + b\bar{x}.$$

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 &= \frac{1}{n}\sum_{i=1}^{n}\left[a + bx_i - (a + b\bar{x})\right]^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[b(x_i - \bar{x})\right]^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} b^2(x_i - \bar{x})^2 \\
&= b^2\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \\
&= b^2\sigma^2
\end{aligned}$$

dove $\sigma^2 = n^{-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$.
Pertanto, scriveremo: $\sigma_y^2 = b^2\sigma^2$.
Per lo scarto quadratico medio vale la relazione: $\sigma_y = |b|\sigma$

1.6 Standardizzazione

Introduciamo una particolare trasformazione lineare che da luogo a una operazione importante nota come *standardizzazione*.
Sia x un carattere con media $\bar{x}$ e varianza σ^2:

$$x_i \Rightarrow y_i = \frac{x_i - \bar{x}}{\sigma} = \frac{x_i}{\sigma} - \frac{\bar{x}}{\sigma}$$

Applichiamo la proprietà appena dimostrata per concludere che le misurazioni y_i hanno media 0 e varianza pari a 1 (basta porre $b = 1/\sigma \ a = -\bar{x}/\sigma$).
NB. I valori standardazzati servono a confrontare due distribuzioni caratterizzate da media e varianza diversa.

1.7 Indice di variabilità relativa: il coefficiente di variazione

La varianza è un indice di variabilità assoluto. Se X è il reddito ci si pu chiedere se σ^2 possa costituire una misura della diseguaglianza. Come tale soffre di alcuni problemi.
Si consideri una distribuzione alternativa Y cui si perviene applicando una *poll tax* (testatico) a tutti gli individui del collettivo. In tal caso, $y_i = a + x_i, a < 0$. A quanto ammonta σ_y^2?
La diseguaglianza diminuita, aumentata o rimasta invariata?
Si supponga invece di applicare un'imposta proporzionale sul reddito, per cui si ottiene la distribuzione $z_i = bx_i, 0 < b < 1$. A quanto ammonta σ_z^2?

Deviazione standard

Empirical rule: If distribution is approx. bell-shaped,

- about 68% of data within 1 standard dev. of mean
- about 95% of data within 2 standard dev. of mean
- all or nearly all data within 3 standard dev. of mean

Range

- The range of a data set is the simplest measure of variability: the difference between the largest and smallest data values.
- It is very sensitive to the smallest and largest data values.
- The interquartile range of a data set is the difference between the third quartile and the first quartile.
- It is the range for the middle 50% of the data.
- It overcomes the sensitivity to extreme data values.

Range

Range = largest value - smallest value

Range = 615 - 425 = 190

425	430	430	435	435	435	435	435	440	440
440	440	440	445	445	445	445	445	450	450
450	450	450	450	450	460	460	460	465	465
465	470	470	472	475	475	475	480	480	480
480	485	490	490	490	500	500	500	500	510
510	515	525	525	525	535	549	550	570	570
575	575	580	590	600	600	600	600	615	615

Note: Data is in ascending order.

Interquartile Range

3rd Quartile ($Q3$) = 525

1st Quartile ($Q1$) = 445

Interquartile Range = $Q3$ - $Q1$ = 525 - 445 = 80

425	430	430	435	435	435	435	435	440	440
440	440	440	445	445	445	445	445	450	450
450	450	450	450	450	460	460	460	465	465
465	470	470	472	475	475	475	480	480	480
480	485	490	490	490	500	500	500	500	510
510	515	525	525	525	535	549	550	570	570
575	575	580	590	600	600	600	600	615	615

Note: Data is in ascending order.

Coefficient of Variation

- The coefficient of variation indicates how large the standard deviation is in relation to the mean.
- The coefficient of variation is computed as follows:

$$CV = \left(\frac{\sigma}{\bar{x}} \right) \times 100\%$$

Variance, Standard Deviation, And Coefficient of Variation

- Variance

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} = 2995.36$$

- Standard Deviation

$$\sigma = \sqrt{\sigma^2} = 54.34$$

the standard deviation is about 11% of the mean

- Coefficient Variation

$$\left(\frac{\sigma}{\bar{x}} \times 100\right)\% = \left(\frac{54.34}{490.8} \times 100\right)\% = 11.07\%$$

Equivalent Formula

$$\sigma^2 = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n} = \frac{\sum_{i=1}^{n}\left(x_i^2 - 2x_i\bar{x} + \bar{x}^2\right)}{n}$$

$$= \frac{\sum x_i^2 + \sum\left(-2x_i\bar{x}\right) + \sum \bar{x}^2}{n} = \frac{\sum x_i^2 - 2\bar{x}\sum x_i + n\bar{x}^2}{n}$$

$$= \frac{\sum x_i^2 - 2\bar{x}\left(n\bar{x}\right) + n\bar{x}^2}{n} = \frac{\sum x_i^2}{n} - \bar{x}^2$$

Exercise

You're a financial analyst for Prudential-Bache Securities. You have collected the following closing stock prices of new stock issues: **17, 16, 21, 18, 13, 16, 12, 11.**

What are the **variance** and **standard deviation** of the stock prices?

Variation Solution

Sample Variance

Raw Data: 17 16 21 18 13 16 12 11

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n} \qquad \text{where} \quad \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = 15.5$$

$$\sigma^2 = \frac{(17-15.5)^2 + (16-15.5)^2 + \cdots + (11-15.5)^2}{8}$$

$$= 9.75$$

Esempio: dati raggruppati

$$\bar{x} = \frac{\sum_{j=1}^{m} x_j^* \times n_j}{n} = \sum_{j=1}^{m} x_j^* \times f_j \qquad \sigma_x^2 = \frac{\sum_{j=1}^{m} x_j^{*2} \times n_j}{n} - \bar{x}^2$$

$$\sigma_x^2 = 3.6 - 1.6^2 = 1.04$$

x_j^*	n_j	$x_j^* \times n_j$	x_j^{2*}	$x_j^{2*} \times n_j$
0	10	0	0	0
1	50	50	1	50
2	10	20	4	40
3	30	90	9	270
6	**100**	**160**		**360**
media		**1,6**		**3,6**

Exercize: grouped data (annual income)

The annual salaries of 69 employess in a company are the following (thousands of euros)

1	director	30
3	head office	20
10	employees	16
25	workers	12
30	laborers	10

Compute the variance

income	frequency	x_j*f_j	x_j^2	$x_j^2*f_j$
10	0.435	4.35	100	43.5
12	0.362	4.344	144	52.128
16	0.145	2.32	256	37.12
20	0.043	0.86	400	17.2
30	0.015	0.45	900	13.5
	1	**12.324**		**163.448**

$$\sigma^2 = \frac{\sum_{i=1}^{69} x_i^2}{69} - \bar{x}^2 = 163.448 - 12.324^2 = 11.567$$

$$\sigma = 3.40$$

$$CV = 27.59\%$$

Exercise

Height	Income (thousands euros in a month)
169	2.2
180	1.8
176	5.7
178	9.8
157	1.4
188	2.4
181	14.5

Which of the two characters has more variability?

Exercise

	Height	Income	Height^2	Income^2
	169	2.2	28561	4.84
	180	1.8	32400	3.24
	176	5.7	30976	32.49
	178	9.8	31684	96.04
	157	1.4	24649	1.96
	188	2.4	35344	5.76
	181	14.5	32761	210.25
Sum	**1229**	**37.8**	**216375**	**354.58**
mean	**175.5714**	**5.4**	**30910.71**	**50.65429**

Exercise

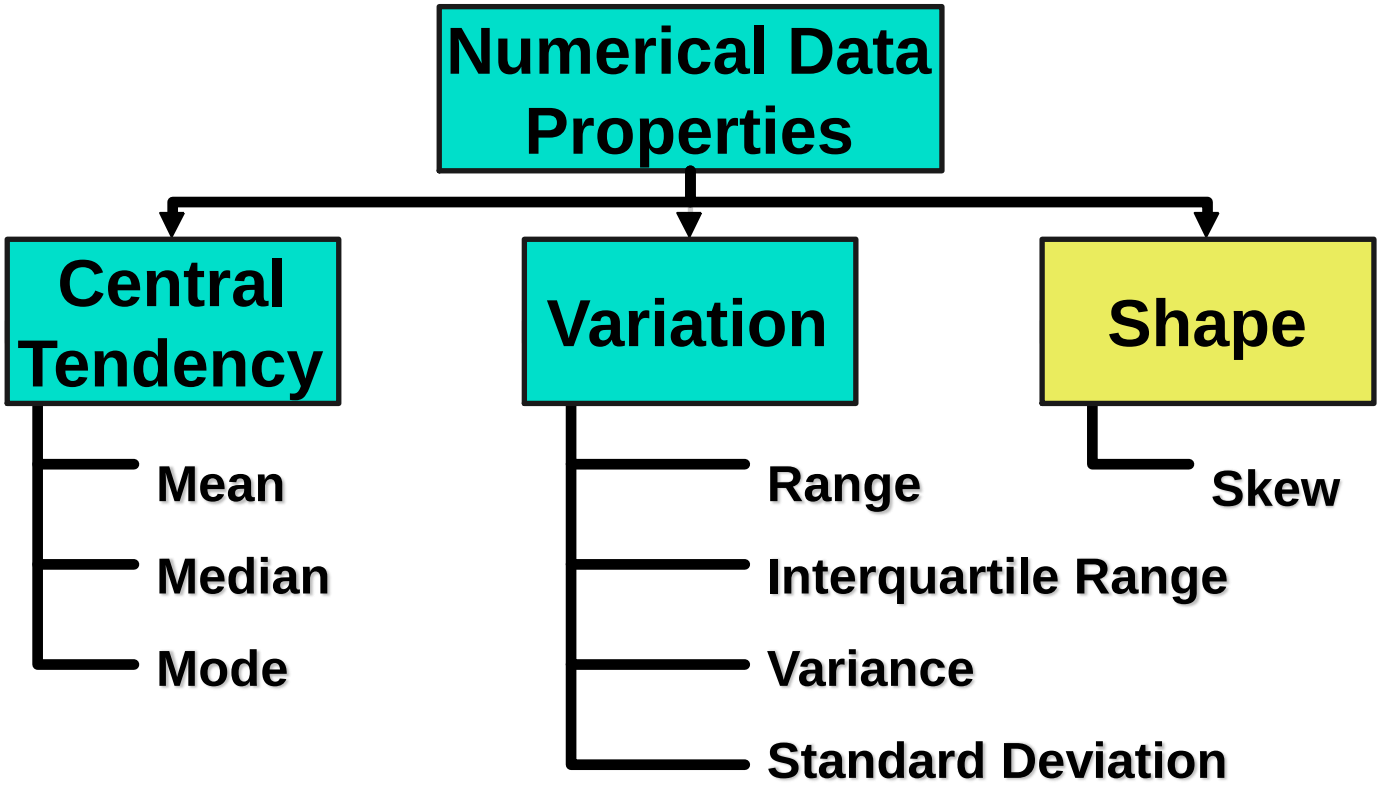
$$\bar{x}_{height} = 175.5714 \qquad \frac{\sum_{i=1}^{n} x_{height}^2}{n} = 30910.71$$

$$\bar{x}_{income} = 5.4 \qquad \frac{\sum_{i=1}^{n} x_{income}^2}{n} = 50.6543$$

$$\sigma_{height}^2 = 30910.71 - 175.5714^2 = 85.3877 \qquad CV_{height} = \frac{\sigma_{height}}{\bar{x}_{height}} = 0.052$$

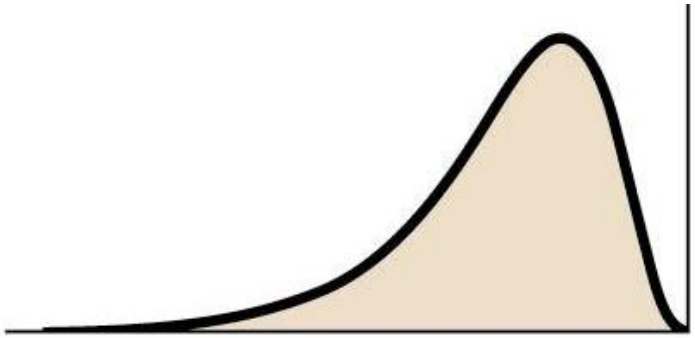
$$\sigma_{income}^2 = 50.6543 - 5.4^2 = 21.4943 \qquad CV_{income} = \frac{\sigma_{income}}{\bar{x}_{income}} = 0.858$$

Numerical Data Properties & Measures



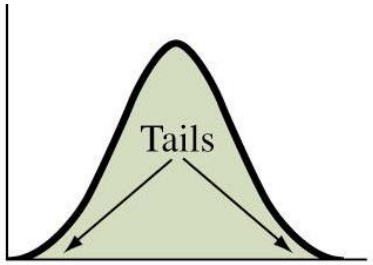
Shape

- Symmetric Distributions: if both left and right sides of the histogram are mirror images of each other

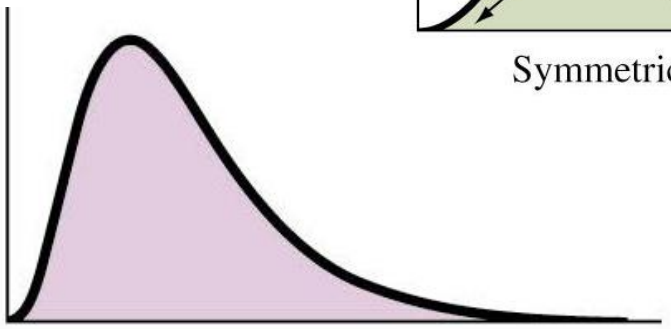


Skewed to the left

A distribution is skewed to the left if the left tail is longer than the right tail



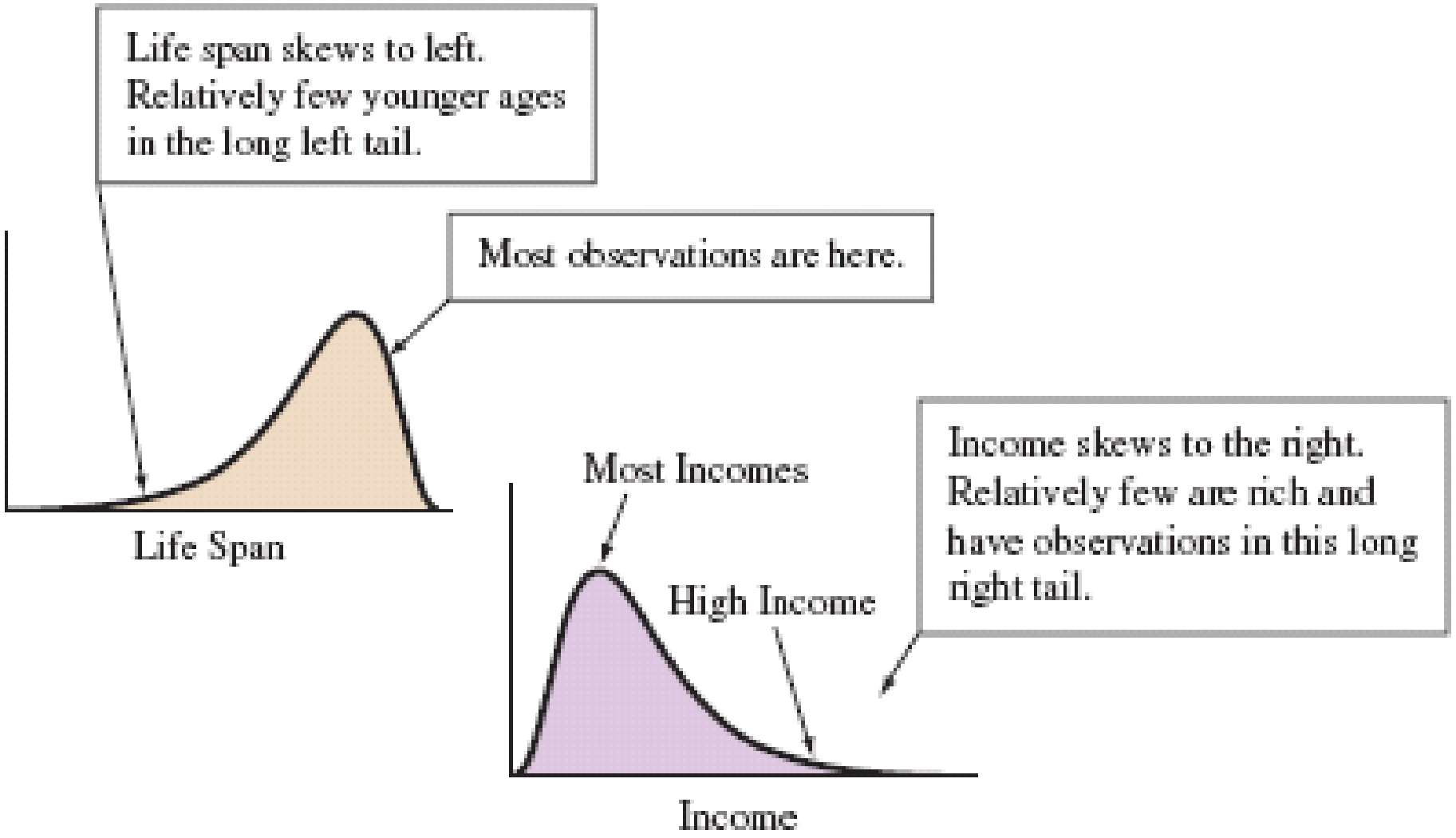
Symmetric



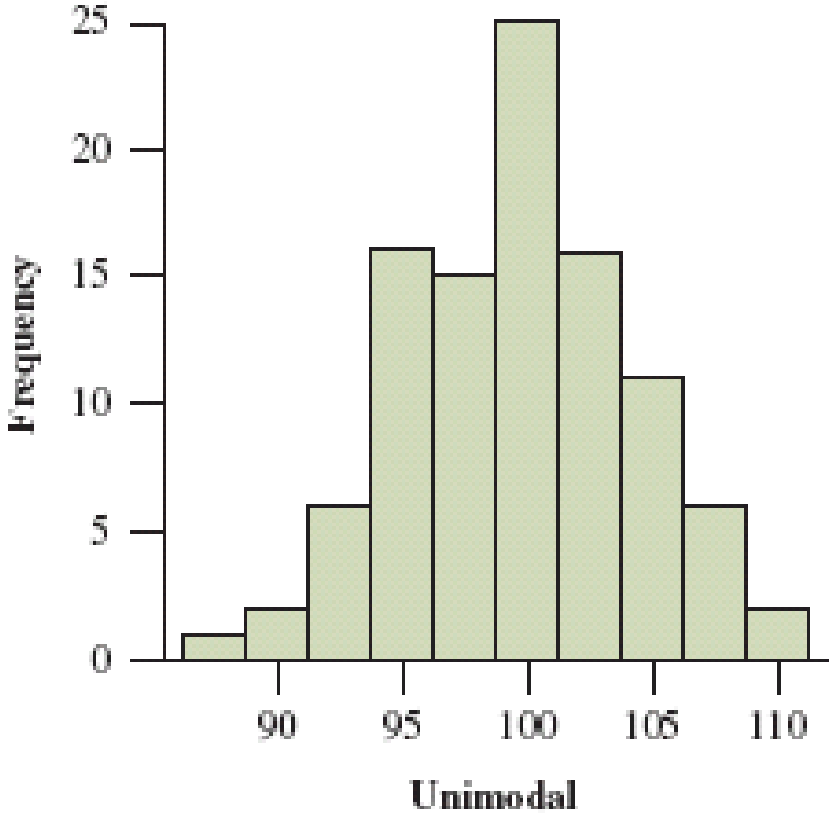
Skewed to the right

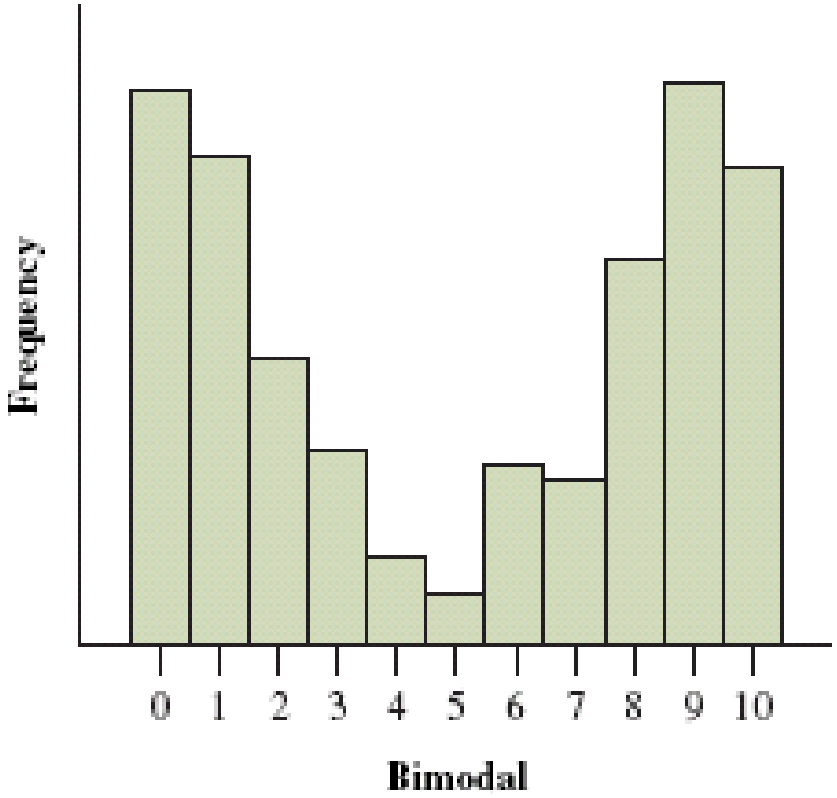
A distribution is skewed to the right if the right tail is longer than the left tail

Examples of Skewness



Shape: Type of Mode





Shape and Skewness

- Consider a data set containing IQ scores for the general public:
- What shape would you expect a histogram of this data set to have?

a. Symmetric

b. Skewed to the left

c. Skewed to the right

d. Bimodal

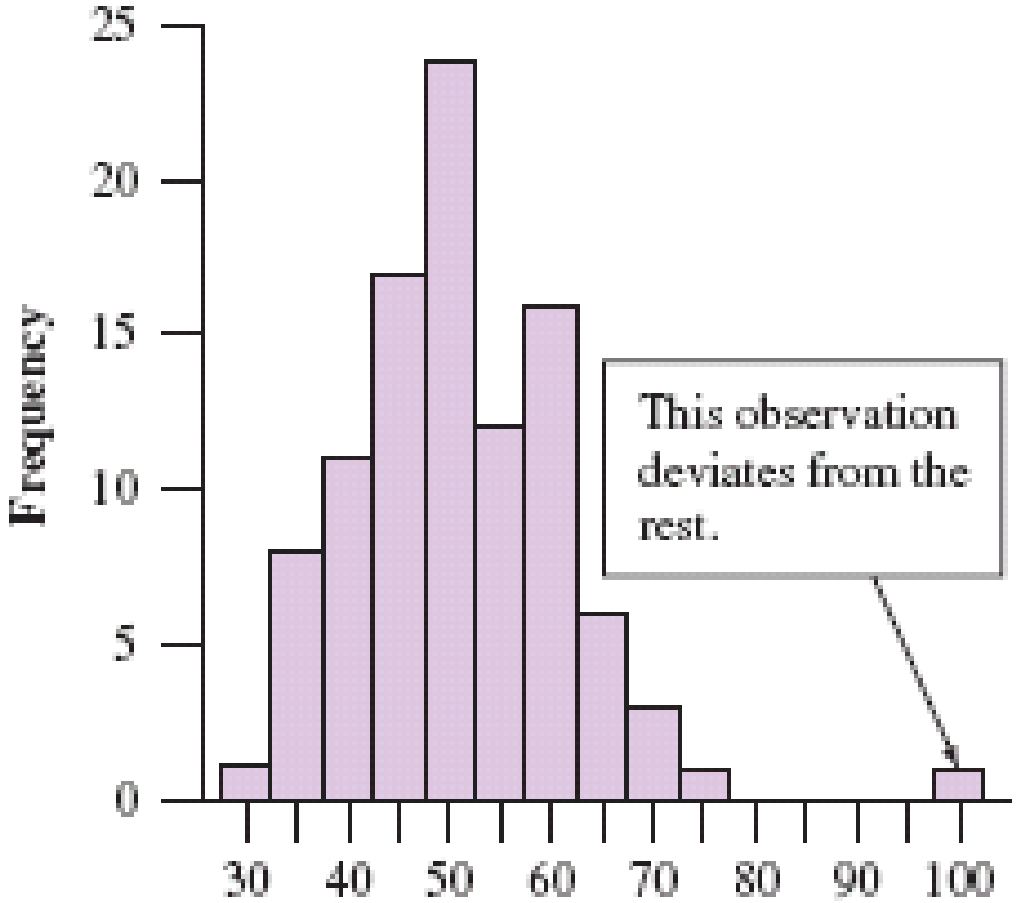
Shape and Skewness

- Consider a data set of the scores of students on a very easy exam in which most score very well but a few score very poorly:
- What shape would you expect a histogram of this data set to have?

a. Symmetric
b. Skewed to the left
c. Skewed to the right
d. Bimodal

Outlier

- An Outlier falls far from the rest of the data



Outlier

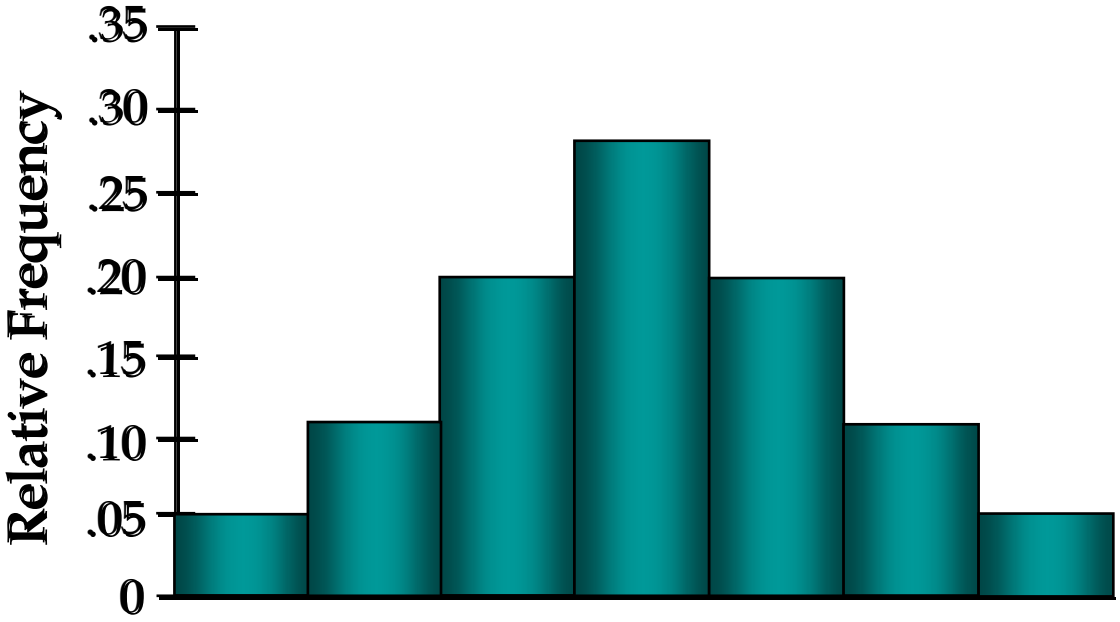
Con valori anomali (in inglese, outlier) ci si riferisce ai valori estremi di una distribuzione che si caratterizzano per essere estremamente elevati o estremamente bassi rispetto al resto della distribuzione e che rappresentano perciò casi isolati rispetto al resto della distribuzione.

In generale, per stabilire se un valore è estremo o anomalo, si fa riferimento alle misure di sintesi della posizione e di dispersione.

Solitamente, vengono considerati come possibili valori anomali quei valori che si discostano dalla media (aritmetica) per più di 3 volte lo scarto quadratico medio.

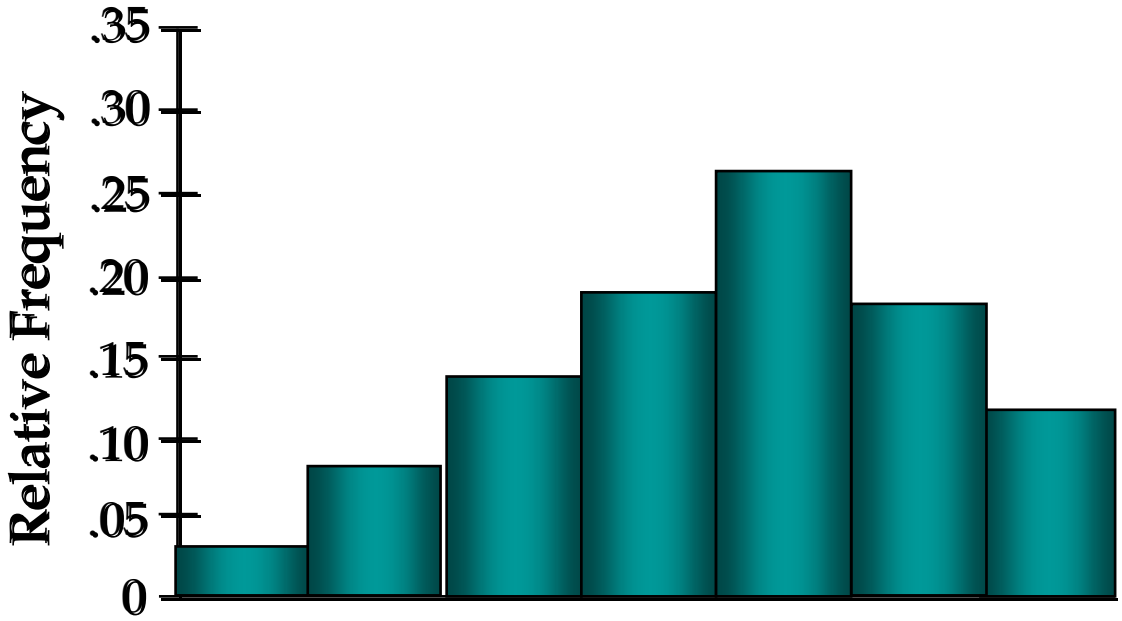
Histogram

- Symmetric
  - Left tail is the mirror image of the right tail
  - Examples: heights and weights of people



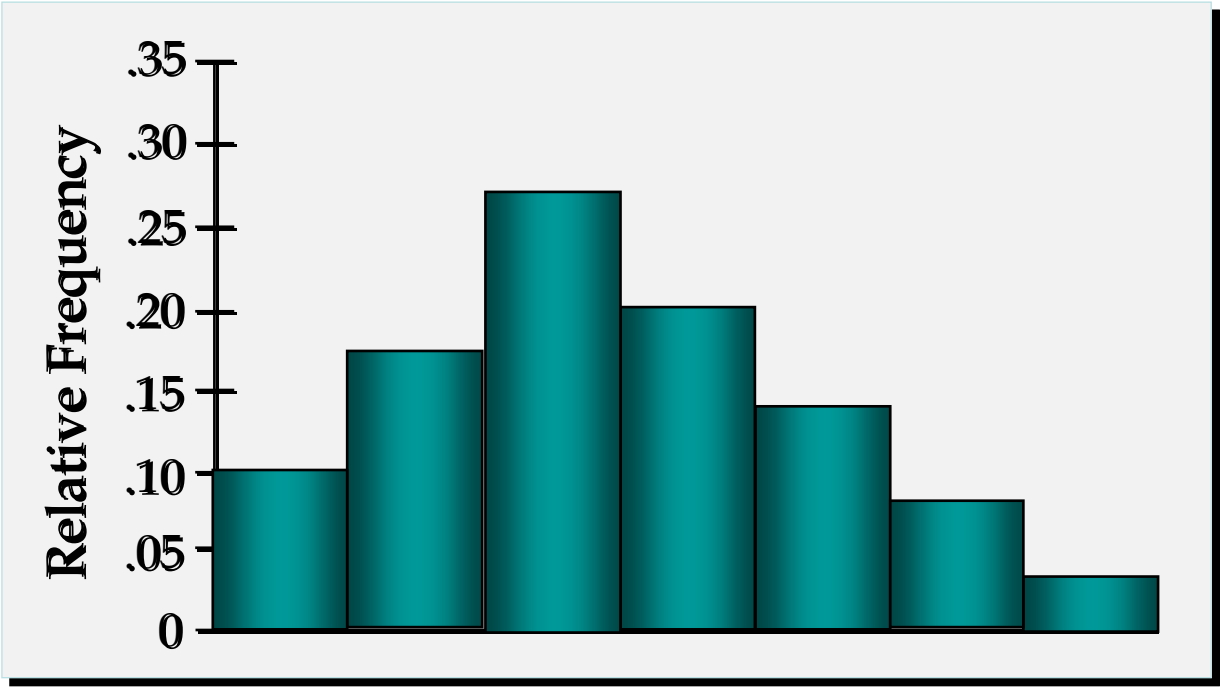
Histogram

- Moderately Skewed Left
  - A longer tail to the left
  - Example: exam scores



Histogram

- Moderately Right Skewed
  - A Longer tail to the right
  - Example: housing values



Box Plot

A **box plot** is a convenient way of graphically represent a distribution through five- number summaries:

the smallest observation, first, second and third quartiles (Q1, Q2 and Q3), and largest observation

A boxplot may also indicate which observations, if any, might be considered outliers

Box Plot

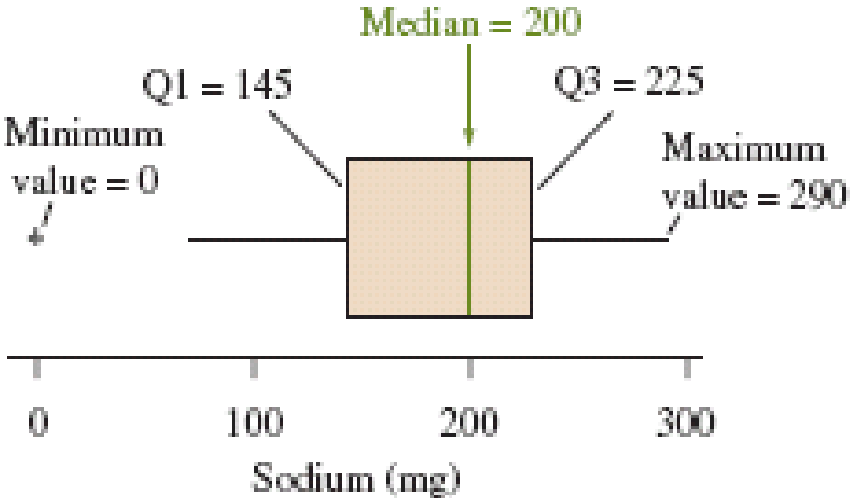
Il grafico a scatola, box plot, è un tipo di grafico proposto dallo statistico americano J. W. Tukey.

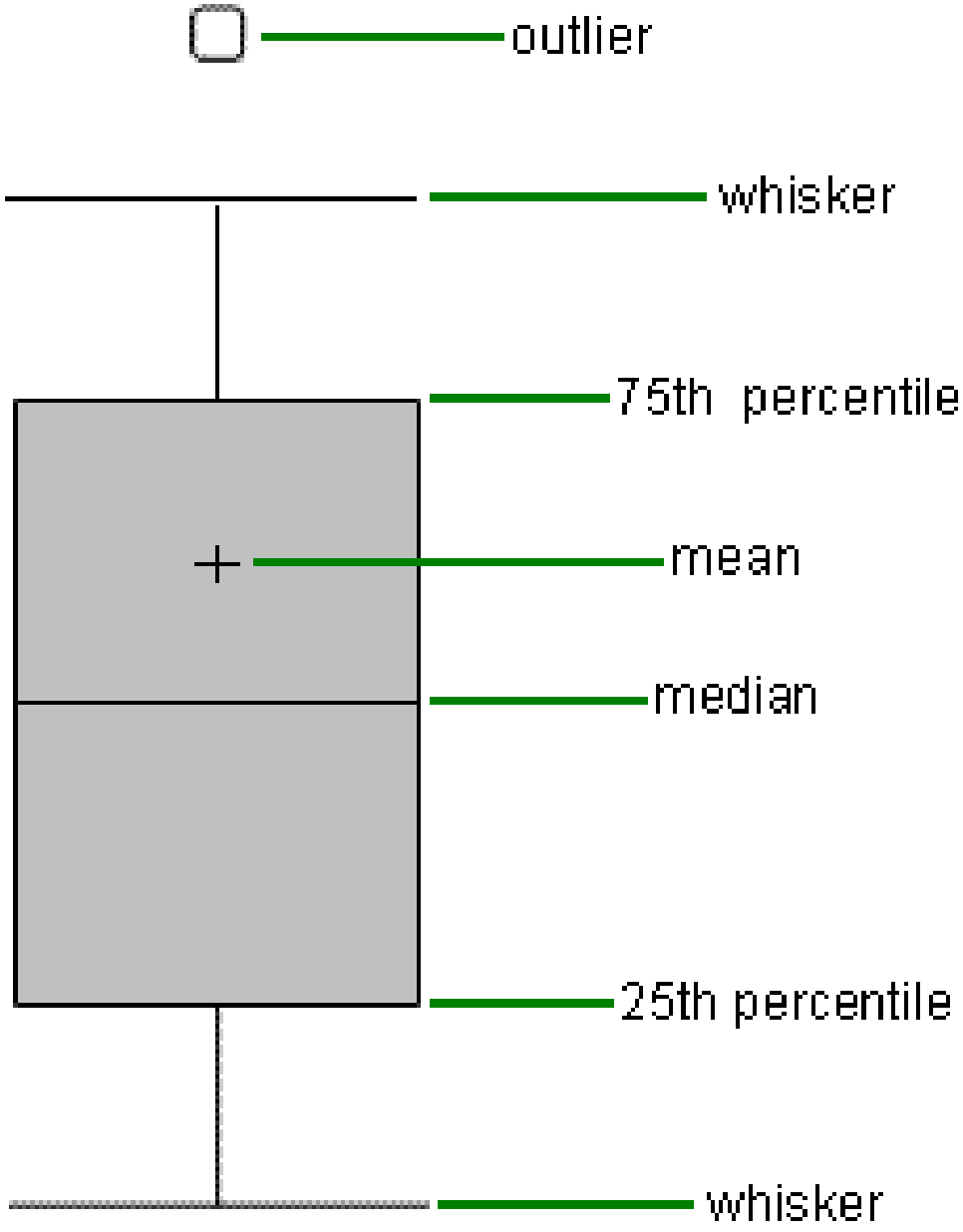
Si ottiene da una serie di indici di una distribuzione, da cui ricava i dati significativi trascurando quelli non importanti.

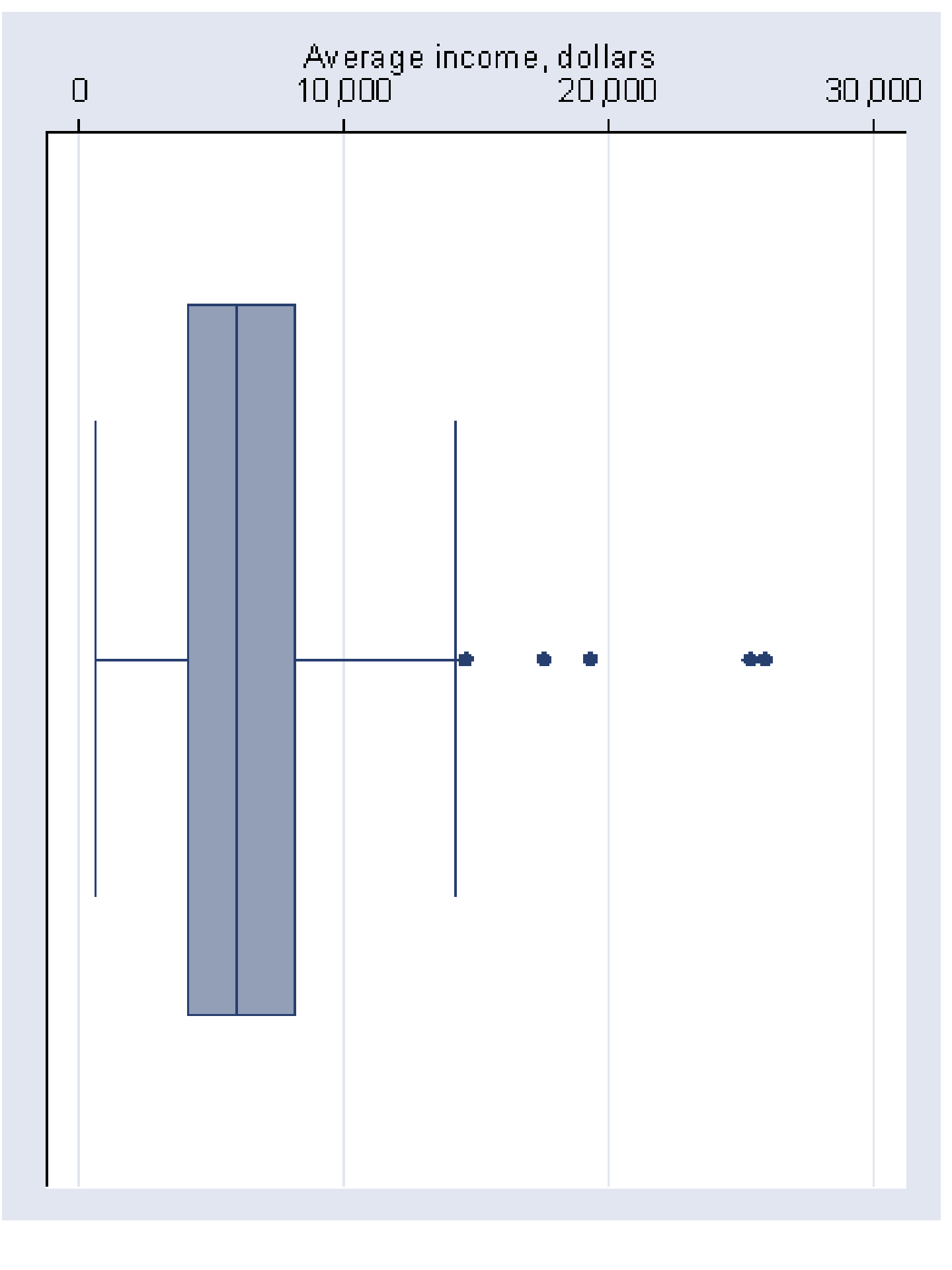
- Una linea che indica la posizione centrale della distribuzione (in generale mediana)
- Un rettangolo la cui altezza indica la variabilità dei valori prossimi alla media (in generale la distanza interquartile)
- Due segmenti che partono dai lati minori del rettangolo e i cui estremi sono determinati in base ai valori estremi della distribuzione (in generale minimo e massimo della distribuzione).

Boxplot

- A box goes from the Q1 to Q3
- A line is drawn inside the box at the median
- A line goes from the lower end of the box to the smallest observation that is not a potential outlier and from the upper end of the box to the largest observation that is not a potential outlier
- The potential outliers are shown separately

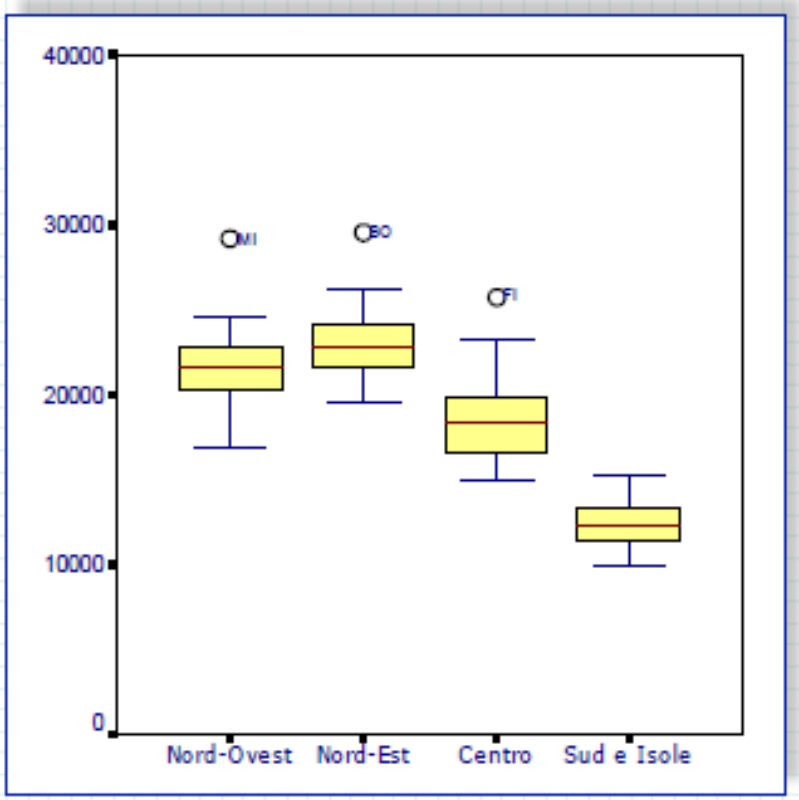



outlier
whisker
75th percentile
mean
median
25th percentile
whisker

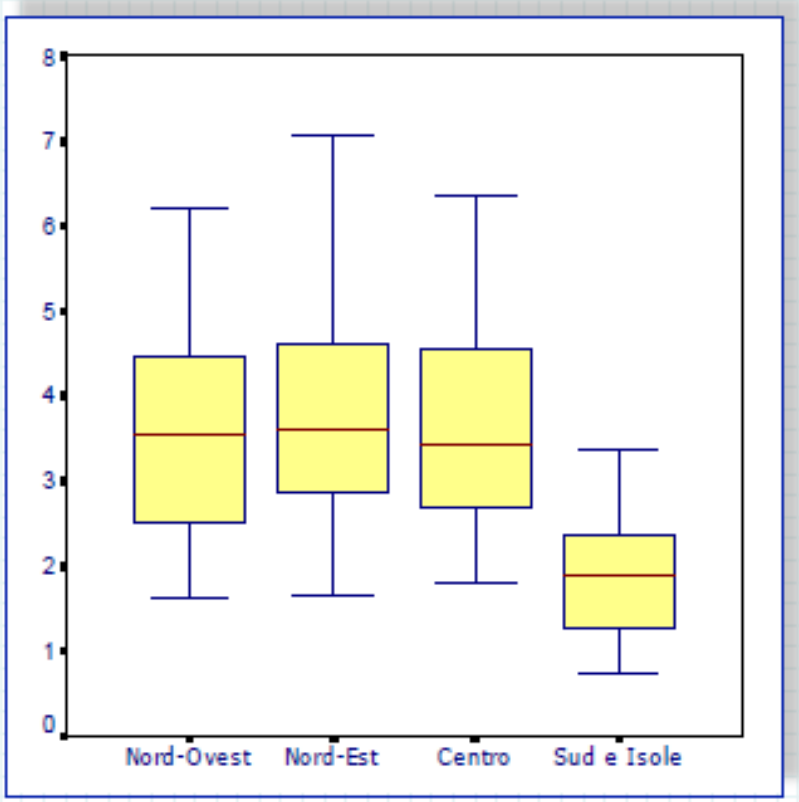

Average income, dollars
0
10,000
20,000
30,000

Il Box Plot agevola il confronto visivo fra due o più distribuzioni

Reddito p.c.
(in €)



Num. sale cinematografiche
(per 100mila ab.)



Example: Income by highest degree