STATISTICA DESCRITTIVA

STATISTICHE CAMPIONARIE 2

(indici di dispersione e regola empirica)

Prof. Rosario Lo Franco – Lezione 3

Riferimenti: [1] Sheldon M. Ross, *Introduzione alla statistica*, Apogeo Editore; [2] Maria Garetto, *Statistica*, Università di Torino

Indici di dispersione

Gli indici di posizione non tengono conto della variabilità esistente fra i dati; vi sono distribuzioni che, pur avendo la stessa media, sono molto diverse fra loro.

I dati dei seguenti insiemi ad esempio hanno la stessa media ($\bar{x} = 60$)

A = {60 60 60 60 60}
B = {10 20 60 100 110}
C = {50 55 60 65 70}

ma gli insiemi sono molto diversi; il primo è composto da dati tutti uguali, mentre il secondo presenta la maggior differenza tra il valore minimo e il massimo.

Indici significativi per la misura della variabilità di una distribuzione di frequenza sono la **varianza** e lo **scarto quadratico medio**, detto anche **deviazione standard**.

Varianza e Deviazione Standard Campionarie

Si definisce **varianza**, o anche **varianza campionaria**, la quantità

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

dove $\bar{x}$ indica la media dei dati.

Si definisce **scarto quadratico medio** o **deviazione standard**

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

Formule più convenienti per la varianza campionaria (per fini computazionali)

$$s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right] \qquad s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right]$$

Varianza e Deviazione Standard Campionarie

- I valori di s e s^2 , poiché misurano l'effettiva variazione assoluta presente in un insieme di dati, dipendono dall'unità di misura dei dati.

- In particolare la **deviazione standard campionaria *s* misura la dispersione dei dati con la stessa unità di misura della media dei dati**, cosa che non accade per la varianza; questa è la ragione principale per la deviazione standard è più usata della varianza.

- **La media e lo scarto quadratico medio sono i due indici di posizione e di dispersione più usati**; uno dei motivi principali è che la distribuzione normale, che viene largamente utilizzata in molti campi diversi, è definita in termini di questi due parametri.

- Proprietà della varianza e della deviazione standard (vedi appunti)

Esempi con indici di dispersione - 1

$A: 1, 2, 5, 6, 6 \qquad B: -40, 0, 5, 20, 35$

Esempio 3.15 Calcola la varianza campionaria dell'insieme di dati A.

Soluzione Procedendo come segue:

x_i	1	2	5	6	6
$\bar{x}$	4	4	4	4	4
$x_i - \bar{x}$	−3	−2	1	2	2
$(x_i - \bar{x})^2$	9	4	1	4	4

$$s^2 = \frac{9 + 4 + 1 + 4 + 4}{4} = 5.5 \quad \blacksquare$$

Esempio 3.16 Calcola la varianza campionaria dell'insieme B.

Soluzione La media campionaria dell'insieme di dati B è anche in questo caso $\bar{x} = 4$. Quindi, per questo insieme, otteniamo

x_i	−40	0	5	20	35
$x_i - \bar{x}$	−44	−4	1	16	31
$(x_i - \bar{x})^2$	1936	16	1	256	961

Di conseguenza,

$$s^2 = \frac{3170}{4} = 792.5 \quad \blacksquare$$

Esempi con indici di dispersione - 2

I seguenti dati sono i tempi di esecuzione di una certa operazione misurati in minuti

0.6 1.2 0.9 1.0 0.6 0.8

Calcoliamo la varianza e la deviazione standard.

$$\bar{x} = \frac{0.6 + 1.2 + 0.9 + 1.0 + 0.6 + 0.8}{6} = 0.85 \text{ minuti}$$

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
0.6	−0.25	0.0625
1.2	0.35	0.1225
0.9	0.05	0.0025
1.0	0.15	0.0225
0.6	−0.25	0.0625
0.8	−0.05	0.0025
totale		0.2750

x_i	x_i^2
0.6	0.36
1.2	1.44
0.9	0.81
1.0	1
0.6	0.36
0.8	0.64
5.10	4.61

$$s^2 = \frac{0.2750}{5} = 0.055 \text{ minuti}^2$$

$$s = \sqrt{0.055} \cong 0.23 \text{ minuti}$$

$$s^2 = \frac{1}{5}\left(4.61 - \frac{5.10^2}{6}\right) = 0.055 \text{ minuti}^2$$

Esempi con indici di dispersione - 3

Per la partecipazione a una gara di matematica una scuola deve formare una squadra di 6 studenti; con una selezione preliminare, attraverso un test con un punteggio massimo di 100 punti, sulla base della media dei migliori 6 punteggi risultano tre squadre a pari merito.
Con quale criterio può essere scelta la squadra da mandare alla gara?

squadra	*punteggi degli studenti*					
A	73	76	77	85	88	90
B	74	74	78	84	88	91
C	72	77	79	82	84	95

La somma dei punteggi ottenuti da ciascuna squadra è 489; la media aritmetica per le tre squadre vale $\bar{x} = 81.5$ e non è quindi un criterio utilizzabile per la scelta; calcoliamo la varianza e lo scarto quadratico medio

Esempi con indici di dispersione - 3

squadra A		*squadra* B		*squadra* C	
x_i	x_i^2	x_i	x_i^2	x_i	x_i^2
73	5329	74	5476	72	5184
76	5776	74	5476	77	5929
77	5929	78	6084	79	6241
85	7225	84	7056	82	6724
88	7744	88	7744	84	7056
90	8100	91	8281	95	9025
489	40103	489	40117	489	40159

$$s^2 = \frac{1}{5}\left(40117 - \frac{1}{6}489^2\right) = 49.9$$

$$s^2 = \frac{1}{5}\left(40103 - \frac{1}{6}489^2\right) = 52.7$$

$$s^2 = \frac{1}{5}\left(40159 - \frac{1}{6}489^2\right) = 61.1$$

squadra	*varianza*	*scarto quadratico medio*
A	49.9	7.06
B	52.7	7.26
C	61.1	7.82

Utilizzando il criterio dello scarto quadratico medio, la squadra da inviare alla gara è la squadra A, che ha il minor scarto quadratico medio.

Esempi con indici di dispersione - 4

I voti in trentesimi riportati da 25 studenti in un esame sono riportati nella seguente tabella. Individuare quali studenti si discostano dal voto medio per più di una volta oppure due volte lo scarto quadratico medio.

numero studente	1	2	3	4	5	6	7	8	9	10	11	12	13
voto	15	17	27	25	29	14	16	25	27	18	10	15	27

numero studente	14	15	16	17	18	19	20	21	22	23	24	25
voto	28	19	14	30	21	17	24	29	20	13	30	25

$\bar{x} = 21.40$

$s = 6.21$

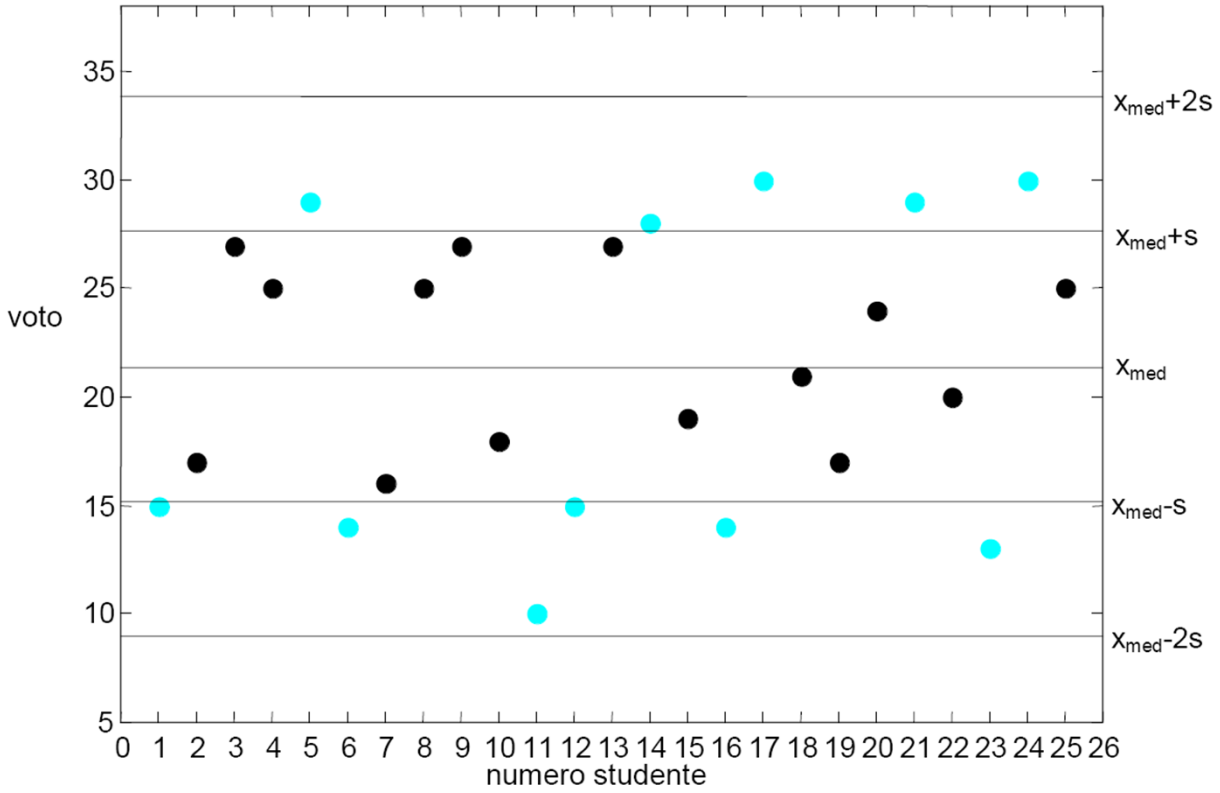
$\bar{x} - s = 15.19 \quad \bar{x} + s = 27.61$

$\bar{x} - 2s = 8.98 \quad \bar{x} + 2s = 33.82$

Tutti i voti appartengono all'intervallo $[\bar{x}-2s\ ,\bar{x}+2s]$, cioè non vi è nessun voto che si discosta dalla media per più di due volte lo scarto quadratico medio; ci sono invece 11 voti che non appartengono all'intervallo $[\bar{x}-s\ ,\bar{x}+s]$, ossia si discostano dalla media per più di una volta lo scarto quadratico medio.

Esempi con indici di dispersione - 4

Per rappresentare la situazione può essere utile un diagramma nel piano cartesiano (figura), con il quale si individuano più facilmente gli studenti che rientrano nella fascia delimitata dai valori $\bar{x}-s$, $\bar{x}+s$.



Media e Varianza per dati raggruppati

- Nel caso in cui i dati siano molto numerosi, non disponendo di un computer il calcolo della media e della varianza viene semplificato se si raggruppano i dati prima di utilizzarli (**quindi, rilevanti solo per calcolo con calcolatrice a mano per un numero grande di dati**). A volte i dati sono forniti già in modo raggruppato.
- Si può calcolare una buona approssimazione di media e varianza, supponendo che i dati di ogni classe siano approssimati dal valore centrale della classe.

Dopo aver raggruppato gli n dati in k classi, indichiamo con m_i il valore centrale della generica classe e con f_i la corrispondente frequenza assoluta della classe.

La **media per dati raggruppati** è definita da

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} m_i f_i$$

La **varianza per dati raggruppati** è definita da

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{k} \left(m_i - \bar{x}\right)^2 f_i$$

Esempio con media e varianza per dati raggruppati

Esempio 2 della Lezione 1

Classe	*Freq. assoluta*	*Freq. relativa*	*Freq. percentuale*
$5.0 \le x \le 8.9$	3	0.0375	3.75%
$9.0 \le x \le 12.9$	10	0.1250	12.5%
$13.0 \le x \le 16.9$	14	0.1750	17.5%
$17.0 \le x \le 20.9$	25	0.3125	31.25%
$21.0 \le x \le 24.9$	17	0.2125	21.25%
$25.0 \le x \le 28.9$	9	0.1125	11.25%
$29.0 \le x \le 32.9$	2	0.0250	2.5%
Totale	80	1	100%

$$\bar{x} = \frac{1}{80}(7 \cdot 3 + 11 \cdot 10 + 15 \cdot 14 + 19 \cdot 25 + 23 \cdot 17 + 27 \cdot 9 + 31 \cdot 2) = 18.90$$

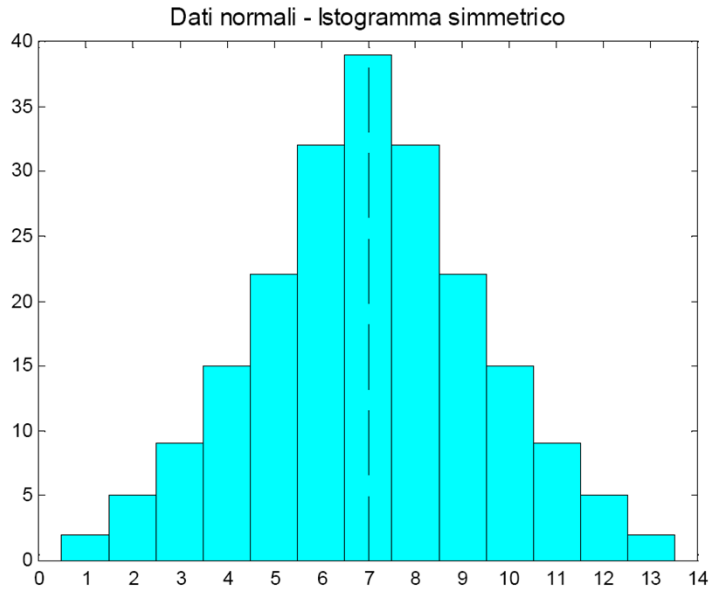
$$s^2 = \frac{1}{79}((7 - 18.9)^2 \cdot 3 + (11 - 18.9)^2 \cdot 10 + \cdots + (31 - 18.9)^2 \cdot 2) = 30.77$$

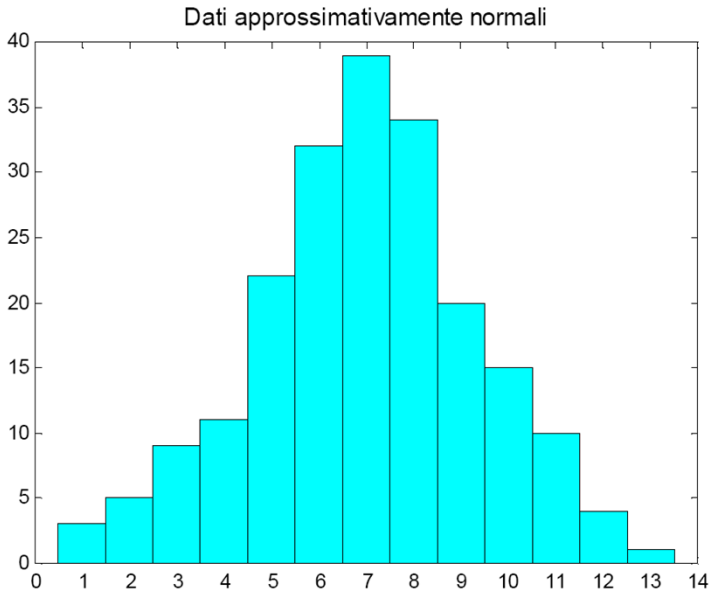
Calcolo esatto sui dati non raggruppati: $\bar{x} = 18.89$

$s^2 = 32.00$

Forma di una distribuzione - Campioni Normali

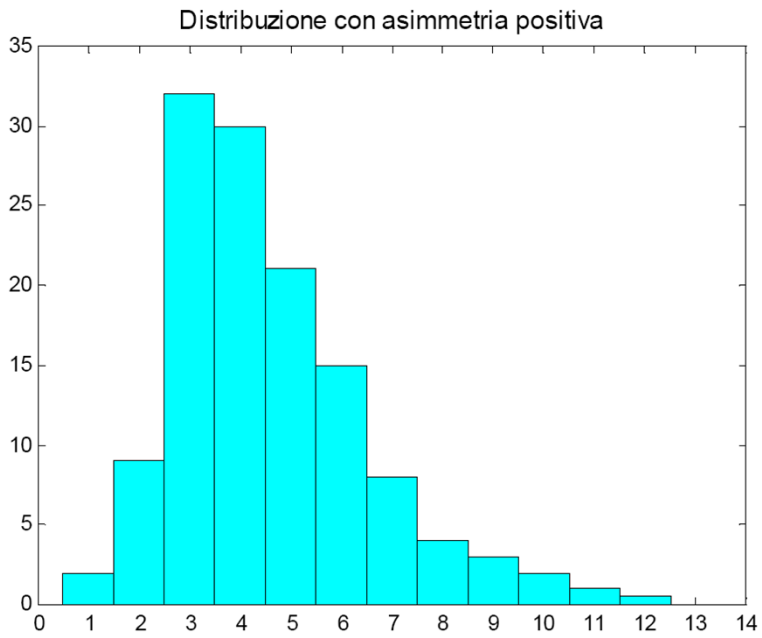
- Le distribuzioni di frequenza possono assumere più forme diverse, e fra queste le più (naturalmente) frequenti e importanti sono quelle che assumono **forma a campana**. In questo caso la distribuzione dei dati è simmetrica rispetto a una linea verticale. **I dati di questo tipo si dicono normali**.

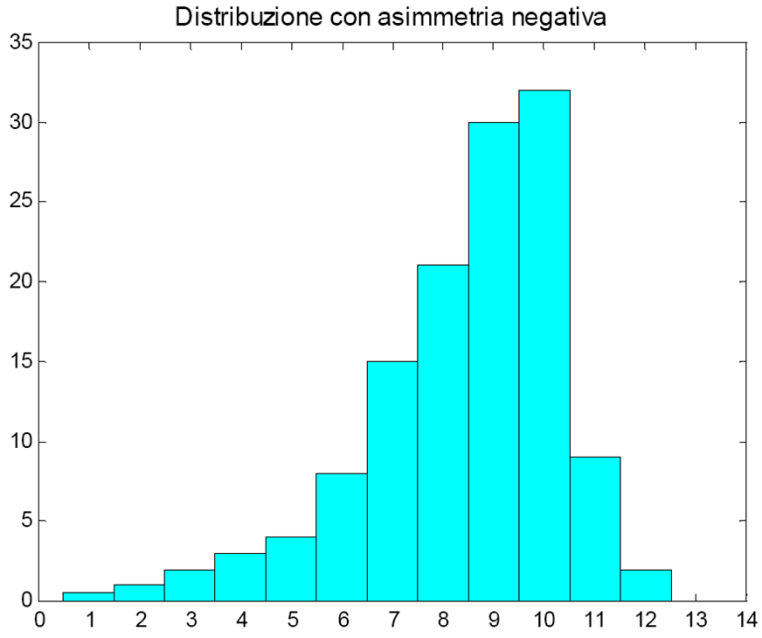




Forma di una distribuzione - Campioni Normali

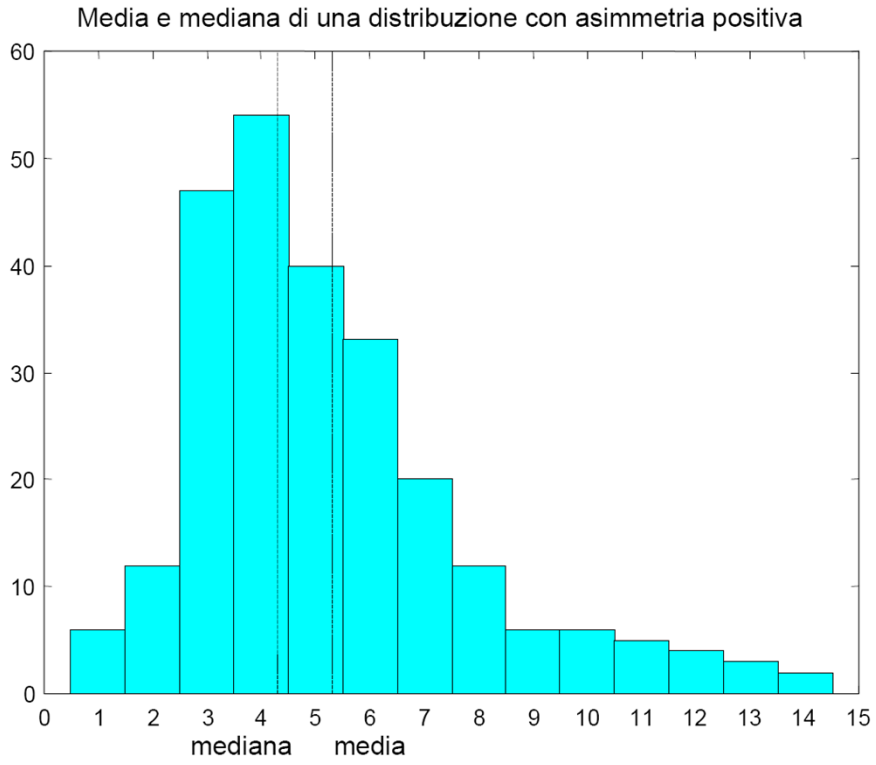
- Una **distribuzione asimmetrica** può avere una **"coda" a destra** e viene detta distribuzione con **asimmetria positiva**.
- Se invece la **coda è a sinistra**, si dice che la distribuzione ha **asimmetria negativa**.





Forma di una distribuzione - Campioni Normali

- Per descrivere la forma della distribuzione è sufficiente **confrontare la media con la mediana**:
  i. se queste due misure sono uguali la distribuzione è simmetrica;
  ii. se la media è maggiore della mediana, la distribuzione ha asimmetria positiva;
  iii. se la media è minore della mediana, la distribuzione ha asimmetria negativa.



Regola Empirica

- Se un insieme di dati è approssimativamente normale, con media $\bar{x}$ e scarto quadratico medio *s*, allora:

$$1 - \text{circa il } 68\% \text{ dei dati è compreso fra } \bar{x} - s \text{ e } \bar{x} + s;$$
$$2 - \text{circa il } 95\% \text{ dei dati è compreso fra } \bar{x} - 2s \text{ e } \bar{x} + 2s;$$
$$3 - \text{circa il } 99.7\% \text{ dei dati è compreso fra } \bar{x} - 3s \text{ e } \bar{x} + 3s;$$

- Questo risultato, noto come **regola empirica** per il fatto che le percentuali indicate sono osservate nella pratica, è in realtà un risultato teorico basato sulle proprietà della distribuzione normale. Gli intervalli 1, 2 e 3 si dicono, rispettivamente, intervalli di una, due e tre deviazioni standard attorno alla media.

Esempio sulla regola empirica

- Per i dati dell'esempio 2 (lezione 1) si trovano le statistiche campionarie:

$$\bar{x} = 18.89 \qquad s^2 = 32.00 \qquad s = 5.66$$

- Abbiamo quindi l'intervallo entro una deviazione standard dalla media:

$$\bar{x} - s = 18.89 - 5.66 = 13.23 \qquad \text{e} \qquad \bar{x} + s = 18.89 + 5.66 = 24.55$$

- Usando la tabella di frequenza, con i dati in ordine crescente, si può facilmente verificare che 14 dati cadono prima di 13.23 e 14 dati cadono dopo 24.55, quindi 80 − 28 = 52 dati cadono nell'intervallo [13.23 , 24.55], ossia il (52/80)x100% = 65% dei dati. La regola empirica ne prevede 68% per dati approssimativamente normali.

Con lo stesso metodo si osserva sulla tabella che il 97.5% dei dati cade fra

$$\bar{x} - 2s = 18.89 - 2 \cdot 5.66 = 7.57 \qquad \text{e} \qquad \bar{x} + 2s = 18.89 + 2 \cdot 5.66 = 30.21$$

e la regola empirica prevede il 95%.

- Si può concludere che i dati sono approssimativamente normali.