Probabilità e Statistica
Esercitazioni

a.a. 2009/2010

C.d.L.: Ingegneria Elettronica e delle Telecomunicazioni, Ingegneria Informatica

Statistica descrittiva II

Ines Campa

Indici di sintesi

Una buona analisi di dati quantitativi richiede anche che le caratteristiche principali delle osservazioni siano *sintetizzate* con opportune **misure** e che tali misure siano *analizzate*.

Distinguiamo tra
statistiche: misure di sintesi calcolate sulla base di un campione;
parametri: misure di sintesi calcolate a partire dall'intera popolazione.

Noi suddividiamo le statistiche in

1) **misure di tendenza centrale**: media campionaria, mediana campionaria e moda campionaria;
2) **misure di variabilità**: varianza campionaria e deviazione standard campionaria.

Noi suddividiamo i parametri in

1) **misure di tendenza centrale**: media, mediana e moda;
2) **misure di variabilità**: varianza e scarto quadratico medio.

Misure di tendenza centrale campionarie

Definizione 1. Supponiamo di avere un campione di n dati $x_1,\ x_2,\ \cdots,\ x_n$. Si dice **media campionaria** e si indica con $\overline{x}_n$

$$\overline{x}_n := \frac{\sum_{i=1}^{n} x_i}{n}$$

Esercizio 1. Si considerino le seguenti altezze in centimetri relative ad un campione di $n = 5$ persone:

$$x_1 = 170,\ x_2 = 160,\ x_3 = 162,\ x_4 = 174,\ x_5 = 161$$

Calcolare la media campionaria.

Risoluzione. La media campionaria risulta $\overline{x}_5 = \frac{170+160+162+174+161}{5} = \frac{827}{5} = 165.4$ centimetri.

Formalizziamo ora il procedimento per il calcolo della media in presenza di una distribuzione di frequenza assoluta o relativa.

–

freq. ass.		
x_i	n_i	$x_i \cdot n_i$
x_1	n_1	$x_1 \cdot n_1$
.	.	.
.	.	.
.	.	.
x_i	n_i	$x_i \cdot n_i$
.	.	.
.	.	.
.	.	.
x_k	n_k	$x_k \cdot n_k$
	$n = \sum_{i=1}^{k} n_i$	$\sum_{i=1}^{k} x_i \cdot n_i$

La media risulta $\overline{x}_n = \frac{1}{n} \sum_{i=1}^{k} x_i \cdot n_i$

–

	freq. rel.	
x_i	f_i	$x_i \cdot f_i$
x_1	f_1	$x_1 \cdot f_1$
.	.	.
.	.	.
.	.	.
x_i	f_i	$x_i \cdot f_i$
.	.	.
.	.	.
.	.	.
x_k	f_k	$x_k \cdot f_k$
	$1 = \sum_{i=1}^{k} f_i$	$\sum_{i=1}^{k} x_i \cdot f_i$

La media risulta $\overline{x}_n = \sum_{i=1}^{k} x_i \cdot f_i$

Definizione 2. **La mediana** è l'osservazione che bipartisce la serie ordinata in senso non decrescente dei dati in due gruppi di ugual numerosità. Nel primo gruppo sono comprese le unità le cui intensità sono al più uguali all'intensità della mediana. Nel secondo gruppo sono comprese invece le unità le cui intensità sono almeno uguali a quelle della mediana.

REGOLA 1

Se l'ampiezza del campione è un numero *dispari*, la mediana coincide con l'osservazione che occupa, nella serie ordinata in senso non decrescente delle osservazioni, la posizione $\frac{n+1}{2}$.

REGOLA 2

Se l'ampiezza del campione è un numero *pari*, la mediana coincide con la media aritmetica delle osservazioni che occupano, nella serie ordinata in senso non decrescente dei dati, la posizione $\frac{n}{2}$ e $\frac{n}{2}+1$.

Esercizio 2. Si considerino le seguenti altezze in centimetri relative ad un campione di $n = 5$ persone:

$$170,\ 160,\ 162,\ 174,\ 161$$

Calcolare la mediana.

Risoluzione. $160 < 161 < 162 < 170 < 174$

n dispari $\Rightarrow \frac{n+1}{2} = 3$. La mediana è $Me = x_3 = 162$

Esercizio 3. Si considerino le seguenti altezze in centimetri relative ad un gruppo di $n = 6$ persone:

$$165,\ 160,\ 163,\ 174,\ 161,\ 165$$

Calcolare la mediana.

Risoluzione. $160 < 161 < 163 < 165 \leq 165 < 174$
n pari $\Rightarrow \frac{n}{2} = 3$ e $\frac{n}{2} + 1 = 4$. La mediana è $Me = \frac{x_3 + x_4}{2} = \frac{163+165}{2} = 164$

Definizione 3. La moda è la modalità che si presenta con maggior frequenza.

La sintesi operata dalla moda è adeguata se la sua frequenza rappresenta almeno il 50% dei casi.

Esercizio 4. Si considerino le seguenti altezze in centimetri relative ad un campione di $n = 5$ persone:

$$165,\ 160,\ 163,\ 174,\ 161,\ 165$$

Calcolare la moda.

Risoluzione. La moda è 165.

Esercizio 5. I dati seguenti rappresentano i tempi di vita in ore di un campione di 25 transistor

104	108	118	128	126
110	104	118	104	112
125	121	114	125	110
108	114	110	110	114
110	121	118	104	114

a) determinare media, mediana e moda campionaria.

b) determinare le frequenze cumulate assoluta e relative.

Risoluzione.

a) Calcoliamo la media aritmetica

$$\overline{x}_{25} = \frac{\sum_{i=1}^{25} x_i}{25} = \frac{104 + 108 + 118 + \cdots + 118 + 104 + 114}{25} = \frac{2850}{25} = 114$$

Compiliamo la seguente tabella

x_i	n_i	$x_i \cdot n_i$	N_i	F_i
104	4	416	4	0.16
108	2	216	6	0.24
110	5	550	11	0.44
112	1	112	12	0.48
114	4	456	16	0.64
118	3	354	19	0.76
121	2	242	21	0.84
125	2	250	23	0.92
126	1	126	24	0.96
128	1	128	25	1
Tot.	25	2850		

Ne segue

$$\overline{x}_{25} = \frac{\sum_{i=1}^{10} x_i \cdot n_i}{25} = \frac{2850}{25} = 114, \; \frac{n+1}{2} = \frac{25+1}{2} = 13 \Rightarrow x_{13} = 114 = Me$$

–

La moda è 110, ma non è rappresentativa perché corrisponde solo al 20% delle osservazioni.

b) Vedere N_i e F_i nella tabella riportata nella pagina precedente.

Esercizio 6. La seguente tabella rappresenta la distribuzione di frequenza della lunghezza di 40 foglie espressa in mm.

classi di lunghezza		frequenza
$(117, 122]$		1
$(122, 127]$		2
$(127, 132]$		2
$(132, 137]$		4
$(137, 142]$		6
$(142, 147]$		8
$(147, 152]$		5
$(152, 157]$		4
$(157, 162]$		2
$(162, 167]$		3
$(167, 172]$		1
$(172, 177]$		2
	totale	40

Determinare la media e la classe modale.

–

Risoluzione.

c_i	n_i	v_{c_i}	$v_{c_i} \cdot n_i$
$(117, 122]$	1	119.5	119.5
$(122, 127]$	2	124.5	249
$(127, 132]$	2	129.5	259
$(132, 137]$	4	134.5	538
$(137, 142]$	6	139.5	837
$(142, 147]$	8	144.5	1156
$(147, 152]$	5	149.5	747.5
$(152, 157]$	4	154.5	618
$(157, 162]$	2	159.5	319
$(162, 167]$	3	164.5	493.5
$(167, 172]$	1	169.5	169.5
$(172, 177]$	2	174.5	349
totale	40		5855

–

Ne segue

$$\overline{x}_{40} = \frac{\sum_{i=1}^{12} v_{c_i} \cdot n_i}{40} = \frac{5855}{40} = 146.375.$$

La classe modale è $(142, 147]$.

Misure di posizione "'non centrate"'

Definizione 4. Sia k un numero intero tale che $0 \leq k \leq 100$. Assegnato un insieme di dati numerici, ordinati in modo non decrescente, ne esiste uno che è contemporaneamente maggiore o uguale di almeno il $k\%$ dei dati, e minore o uguale di almeno il $(100 - k)\%$ dei dati. Se il dato con queste caratteristiche è unico, esso è per definizione il **k-esimo percentile** dell'insieme di dati considerati. Se invece non è unico, allora sono esattamente due, e in questo caso il **k-esimo percentile** è definito come la loro media aritmetica.

I 3 quartili Q_1, Q_2 e Q_3 dividono la serie ordinata in modo non decrescente dei dati in 4 gruppi di ugual numerosità.

I 9 decili $D_1, D_2, \cdots, D_9$ dividono la serie ordinata in modo non decrescente dei dati in 10 gruppi di ugual numerosità.

I 99 centili $C_1, C_2, \cdots, C_{99}$ dividono la serie ordinata in modo non decrescente dei dati in 100 gruppi di ugual numerosità.

–

Esercizio 7. I dati seguenti rappresentano i tempi di vita in ore di un campione di 25 transistor

104	108	118	128	126
110	104	118	104	112
125	121	114	125	110
108	114	110	110	114
110	121	118	104	114

Determinare C_{11}, D_3, D_7 e Q_1.

Risoluzione. Ordiniamo i dati in modo non decrescente

$$
\begin{array}{lllll}
x_1 = 104 & x_2 = 104 & x_3 = 104 & x_4 = 104 & x_5 = 108 \\
x_6 = 108 & x_7 = 110 & x_8 = 110 & x_9 = 110 & x_{10} = 110 \\
x_{11} = 110 & x_{12} = 112 & x_{13} = 114 & x_{14} = 114 & x_{15} = 114 \\
x_{16} = 114 & x_{17} = 118 & x_{18} = 118 & x_{19} = 118 & x_{20} = 121 \\
x_{21} = 121 & x_{22} = 125 & x_{23} = 125 & x_{24} = 126 & x_{25} = 128
\end{array}
$$

–

Risoluzione.

$$\frac{(n+1)\cdot 11}{100} = \frac{26\cdot 11}{100} = 2.86 \Rightarrow C_{11} = \frac{x_2 + x_3}{2} = \frac{104+104}{2} = 104$$

$$\frac{(n+1)\cdot 3}{10} = \frac{26\cdot 3}{10} = 7.8 \Rightarrow D_3 = \frac{x_7 + x_8}{2} = \frac{110+110}{2} = 110$$

$$\frac{(n+1)\cdot 7}{10} = \frac{26\cdot 7}{10} = 18.2 \Rightarrow D_3 = \frac{x_{18} + x_{19}}{2} = \frac{118+118}{2} = 118$$

$$\frac{n+1}{4} = \frac{26}{4} = 6.5 \Rightarrow Q_1 = C_{25} = \frac{x_6 + x_7}{2} = \frac{108+110}{2} = 109$$

Esercizio 8. Il diagramma stem and leaf riporta i giorni di vita di un campione di 19 topi, dopo essere stati sottoposti ad un trattamento radioattivo.

1	59	89	91	98				
2	35	45	50	56	61	65	66	80
3	43	56	83					
4	03	14	28	32				

Determinare C_{15}, D_3, D_7 e Q_1.

Risoluzione.

$$\frac{(n+1)\cdot 15}{100} = \frac{20\cdot 15}{100} = 3 \Rightarrow C_{15} = x_3 = 191$$

$$\frac{(n+1)\cdot 3}{10} = \frac{20\cdot 3}{10} = 6 \Rightarrow D_3 = x_6 = 245$$

$$\frac{(n+1)\cdot 7}{10} = \frac{20\cdot 7}{10} = 14 \Rightarrow D_3 = x_{14} = 356$$

$$\frac{n+1}{4} = \frac{20}{4} = 5 \Rightarrow Q_1 = C_{25} = x_5 = 235$$

Misure di variabilità campionarie

Definizione 5. Supponiamo di avere un campione di n dati $x_1,\ x_2,\ \cdots,\ x_n$. Si dice **varianza campionaria** e si indica con S^2

$$S^2 := \frac{\sum_{i=1}^{n} (x_i - \overline{x}_n)^2}{n-1}$$

Al denominatore si usa $n-1$ per renderla una misura adeguata nell'inferenza statistica.

Definizione 6. Supponiamo di avere un campione di n dati $x_1,\ x_2,\ \cdots,\ x_n$. Si dice **scarto quadratico medio campionario** e si indica con S la radice quadrata della varianza campionaria:

$$S := \sqrt{\frac{\sum_{i=1}^{n} (x_i - \overline{x}_n)^2}{n-1}}$$

Uno dei casi in cui S^2 e S possono essere uguali è quando non c'è variabilità nei dati: $S^2 = S = 0$, l'altro è quando $S^2 = S = 1$.

Esercizio 9. Si considerino le seguenti altezze in centimetri relative ad un campione di $n = 5$ persone:

$$x_1 = 170,\ x_2 = 160,\ x_3 = 162,\ x_4 = 172,\ x_5 = 161$$

Calcolare la varianza campionaria e la deviazione standard campionaria.

Risoluzione. La media campionaria risulta $\overline{x}_5 = \frac{170+160+162+172+161}{5} = \frac{825}{5} = 165$ centimetri. Ne segue

$$S^2 := \frac{\sum_{i=1}^{n} (x_i - \overline{x}_n)^2}{n-1} = \frac{(170-165)^2 + (160-165)^2}{4} +$$

$$+ \frac{(162-165)^2 + (172-165)^2 + (161-165)^2}{4} = 31$$

e $S = \sqrt{S^2} \approx 5.568$

Formalizziamo ora il procedimento per il calcolo della varianza campionaria in presenza di una distribuzione di frequenza assoluta o relativa.

–

	freq. ass.	
x_i	n_i	$(x_i - \overline{x}_n)^2 \cdot n_i$
x_1	n_1	$(x_1 - \overline{x}_n)^2 \cdot n_1$
.	.	.
.	.	.
.	.	.
x_i	n_i	$(x_i - \overline{x}_n)^2 \cdot n_i$
.	.	.
.	.	.
.	.	.
x_k	n_k	$(x_k - \overline{x}_n)^2 \cdot n_k$
	$n = \sum_{i=1}^{k} n_i$	$\sum_{i=1}^{k} (x_i - \overline{x}_n)^2 \cdot n_i$

La varianza campionaria risulta

$$S^2 = \frac{\sum_{i=1}^{k} (x_i - \overline{x}_n)^2 \cdot n_i}{n-1}.$$

–

	freq. rel.	
x_i	f_i	$(x_i - \overline{x}_n)^2 \cdot f_i$
x_1	f_1	$(x_1 - \overline{x}_n)^2 \cdot f_1$
.	.	.
.	.	.
.	.	.
x_i	f_i	$(x_i - \overline{x}_n)^2 \cdot f_i$
.	.	.
.	.	.
.	.	.
x_k	f_k	$(x_k - \overline{x}_n)^2 \cdot f_k$
Tot	$1 = \sum_{i=1}^{k} f_1$	$\sum_{i=1}^{k} (x_i - \overline{x}_n)^2 \cdot f_i$

La varianza campionaria risulta

$$S^2 = \frac{n}{n-1} \sum_{i=1}^{k} (x_i - \overline{x}_n)^2 \cdot f_i$$

Parametri della popolazione

Supponiamo che $x_1, x_2, \cdots, x_n$ sia una raccolta di misurazioni numeriche da una intera *popolazione*. Si definisce

Media della popolazione

$$\mu := \frac{\sum_{i=1}^{n} x_i}{n}$$

Varianza della popolazione

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

Scarto quadratico medio o deviazione standard della popolazione

$$\sigma := \sqrt{\frac{\sum_{i=1}^{n} (x_i - \mu)^2}{n}}$$