STATISTICA DESCRITTIVA

STATISTICHE CAMPIONARIE 1

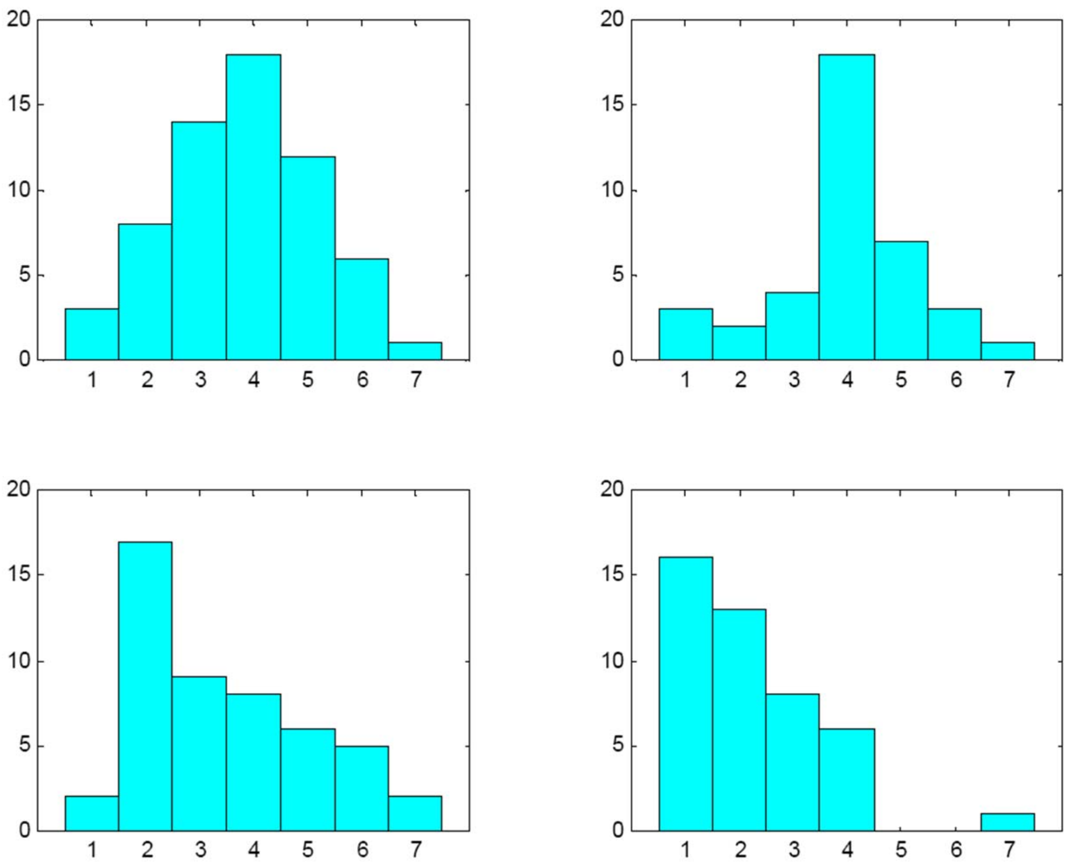
(indici di posizione)

Prof. Rosario Lo Franco – Lezione 2

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

Statistiche campionarie

Si osservino i seguenti istogrammi



Gli indici di posizione centrale e di dispersione (**le statistiche campionarie**) serviranno a misurare quantitativamente alcune caratteristiche qualitative osservate nei grafici

Media campionaria (aritmetica)

Si consideri un insieme (campione) di *n* dati $x_1, x_2, \ldots, x_n$.

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

Per ogni valore x_i della variabile x si definisce lo **scarto dalla media**

$$s_i = x_i - \bar{x}$$

Es.: Dimostra che la somma algebrica degli scarti della media è nulla.

Media dei dati

$$15 \qquad 14 \qquad 2 \qquad 27 \qquad 13$$

$$\bar{x} = \frac{15 + 14 + 2 + 27 + 13}{5} = 14.2$$

Proprietà della media campionaria e altri tipi di medie: pesata, geometrica, armonica

ESERCIZI ED ESEMPI

STATISTICHE CAMPIONARIE 1

(media)

Media Campionaria

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

Esempio 3.3 Il numero di abiti venduti in ciascuno degli ultimi 6 giorni in una boutique per signora è presentato nella seguente tabella delle frequenze:

Valore	Frequenza
3	2
4	1
5	3

Qual è la media campionaria?

Soluzione Visto che l'insieme di dati originale è costituito dai 6 valori

$$3, 3, 4, 5, 5, 5$$

otteniamo che la media campionaria è

$$\bar{x} = \frac{3+3+4+5+5+5}{6} = \frac{3\times 2 + 4\times 1 + 5\times 3}{6} = \frac{25}{6}$$

Media Campionaria

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

Esempio 3.4 In un articolo intitolato "Effetti dell'uso del casco sulla gravità delle contusioni alla testa negli incidenti motociclistici" (pubblicato dal *Journal of the American Statistical Association*, 1992, pp. 48-56), A. Weiss ha analizzato un campione di 770 incidenti motociclistici dalle caratteristiche simili verificatisi nell'area di Los Angeles nel 1976 e nel 1977. Ciascun incidente è stato classificato secondo la gravità del danno alla testa sofferto dal motociclista.

Media Campionaria

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

I danni sono stati classificati come segue:

Categoria	Danno alla testa
0	Nessun danno
1	Minore
2	Moderato
3	Grave, senza pericolo per la vita
4	Grave, pericolo per la vita
5	Critico, sopravvivenza incerta al momento dell'incidente
6	Fatale

In 331 incidenti il motociclista indossava il casco, mentre negli altri 439 il motociclista con indossava il casco. Le tabelle delle frequenze che seguono indicano la gravità degli incidenti che sono capitati quando il motociclista indossava o non indossava il casco.

Categoria	Frequenza con il casco	Frequenza senza il casco
0	248	227
1	58	135
2	11	33
3	3	14
4	2	3
5	8	21
6	1	6
	331	439

Determina la media campionaria delle categorie di danno alla testa per i motociclisti che usano il casco e per quelli che non lo usano.

Media Campionaria

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

Soluzione La media campionaria di chi usa il casco è

$$\bar{x} = \frac{0 \cdot 248 + 1 \cdot 58 + 2 \cdot 11 + 3 \cdot 3 + 4 \cdot 2 + 5 \cdot 8 + 6 \cdot 1}{331} = \frac{143}{331} = 0.432$$

La media campionaria di chi non usa il casco è:

$$\bar{x} = \frac{0 \cdot 227 + 1 \cdot 135 + 2 \cdot 33 + 3 \cdot 14 + 4 \cdot 3 + 5 \cdot 21 + 6 \cdot 6}{439} = \frac{396}{439} = 0.902$$

Di conseguenza, i dati indicano che i motociclisti che usano il casco hanno sofferto, in media, danni alla testa meno gravi di quelli che non usavano il casco. ■

Media Campionaria

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

Esempio 2.3.1. Quelli che seguono sono i punteggi vincenti del torneo di golf U.S. Masters negli anni dal 1982 al 1991:

284 280 277 282 279 285 281 283 278 277

Se ne vuole trovare la media campionaria.

La media campionaria dei dati trasformati si calcola molto facilmente,

$$\bar{y} = \frac{4+0-3+2-1+5+1+3-2-3}{10} = \frac{6}{10}$$

Ne segue che

$$\bar{x} = \bar{y} + 280 = 280.6 \quad \square$$

Media Campionaria

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

4. Considera cinque numeri. Supponiamo che la media dei primi quattro numeri sia 14.
   (a) Se il quinto numero è 24, qual è la media dei cinque numeri?
   (b) Se la media di tutti i cinque numeri è 24, qual è il quinto numero?

15. Metà dei valori di un campione sono uguali a 10, un sesto sono uguali a 20, e un terzo sono uguali a 30. Qual è la media campionaria?

Mediana Campionaria

- La **mediana** *M* di un insieme di *n* dati ordinati in ordine di grandezza crescente è il valore centrale dei dati, se il numero di dati è dispari, o la media aritmetica dei due valori centrali, se il numero dei dati è pari.
- **Lo stesso numero di dati cade sia a sinistra che a destra della mediana stessa**.
- L'uso della mediana come indice per descrivere le caratteristiche dei dati ha lo svantaggio di dover **prima riordinare i dati in ordine crescente**, il che non è richiesto per il calcolo della media.

a – Mediana dei dati

15	14	2	27	13	

Dati ordinati in ordine crescente

2	13	14	15	27	

Mediana

$$M = 14$$

b – Mediana dei dati

11	9	17	19	4	15

Dati ordinati in ordine crescente

4	9	11	15	17	19

Mediana $$M = \frac{11+15}{2} = 13$$

Mediana Campionaria

Regola generale per il calcolo della **mediana** M di un campione di n dati

1. Disporre i dati in ordine di valore crescente
2. Se n è **dispari**, la mediana è il dato nella posizione $(n+1)/2$
3. Se n è **pari**, la mediana è la media aritmetica dei dati nelle posizioni $n/2$ e $n/2+1$

ESERCIZI ED ESEMPI

STATISTICHE CAMPIONARIE 1

(mediana campionaria)

Mediana Campionaria

Esempio 3.6 I seguenti dati rappresentano il numero di settimane impiegate da sette individui per ottenere la patente di guida. Trova la mediana campionaria.

$$2, 110, 5, 7, 6, 7, 3$$

Soluzione Prima di tutto disponiamo i dati in ordine crescente.

$$2, 3, 5, 6, 7, 7, 110$$

Visto che la dimensione del campione è 7, la mediana campionaria è il quarto valore più piccolo. Quindi, la mediana campionaria del numero di settimane necessarie per ottenere la patente è $m = 6$ settimane. ■

Mediana Campionaria

2. (a) Determina la mediana campionaria dell'insieme di dati

$$14, 22, 8, 19, 15, 7, 8, 13, 20, 22, 24, 25, 11, 9, 14$$

(b) Incrementa ciascun valore in (a) di 5, e calcola la nuova mediana campionaria.

(c) Moltiplica ciascun valore in (a) per 3, e calcola la nuova mediana campionaria.

3. Se la mediana dell'insieme di dati x_i, $i = 1, \ldots, n$, è 10, qual è la mediana dell'insieme di dati $2x_i + 3$, $i = 1, \ldots, n$?

10. La mediana campionaria di 10 valori distinti è 5. Che cosa puoi dire sulla nuova mediana campionaria se

(a) Viene aggiunto un dato il cui valore è 7?

(b) Vengono aggiunti due dati, 3 e 42?

Moda Campionaria

- La **moda** $\tilde{x}$ di un insieme di *n* dati è il valore o la classe a cui corrisponde la massima frequenza assoluta.
- La moda è per lo più utilizzata quando si trattano dati di tipo qualitativo, per i quali non è possibile calcolare media e mediana.
- La moda può non esistere o non essere unica; quando è unica, la distribuzione è detta **unimodale**, quando ci sono più mode diverse è detta **bimodale** o **multimodale**

a – Moda dell'insieme di dati

$$3, 3, 5, 4, 7, 7, 7, 9, 2, 1$$

L'insieme ha moda $\tilde{x} = 7$.

b – Moda dell'insieme di dati

$$3, 3, 3, 5, 4, 7, 7, 7, 9, 2, 1$$

L'insieme ha due mode $\tilde{x} = 3$ e $\tilde{x} = 7$.

c – L'insieme di dati

$$3, 5, 4, 7, 8, 6, 9, 2, 1$$

non ha moda, perché ogni dato si presenta una sola volta.

Indici di posizione centrale

- Media, mediana e moda sono detti **indici di posizione o indici di tendenza centrale**, perché descrivono attorno a quale valore è centrato l'insieme di dati.

- La mediana è preferibile alla media quando si vogliono eliminare gli effetti di valori estremi molto diversi dagli altri dati: la ragione è che la mediana non utilizza tutti i dati, ma solo il dato centrale o i due dati centrali.

- La mediana in alcuni casi descrive in modo più adeguato un insieme di dati.

- Tuttavia occorre mettere in evidenza che l'utilizzare solo i dati centrali rende la mediana poco sensibile a tutti gli altri valori dei dati e questo può costituire un limite di questo indice.

Media vs Mediana

Sia dato il seguente insieme di 20 dati, che rappresentano il peso alla nascita (in g) di 20 bambini nati in una settimana in una clinica.

3280	3320	2500	2760
3260	3650	2840	3250
3240	3200	3600	3320
3480	3020	2840	3200
4160	2580	3540	3780

Tabella 38

La media dei dati è

$$\bar{x} = \frac{(3280 + 3320 + \ldots + 3540 + 3780)}{20} = 3241\text{ g}$$

Si può osservare che 9 dati sono minori della media e 11 maggiori.
Come già osservato, uno dei limiti della media come misura della tendenza centrale è che essa è molto sensibile ai valori dei dati che cadono agli estremi dell'intervallo di variabilità; in questo senso può non rappresentare bene la collocazione dei dati. Se ad esempio il primo bambino fosse un nato prematuro del peso di 500 g, la media avrebbe il valore

$$\bar{x} = 3102\text{ g}$$

e in tal caso 7 dati sarebbero minori della media e 13 maggiori.
La mediana in questo caso è

$$M = 3245$$

mentre per l'insieme di dati assegnati inizialmente è

$$M = 3255$$

Media vs Mediana

In una ditta lavorano 4 giovani ingegneri, che guadagnano € 15.000 all'anno ciascuno, e il proprietario, anch'egli ingegnere, che guadagna € 90.000 all'anno.
Stabilire se la ditta è un buon posto di lavoro per un giovane ingegnere.

Media degli stipendi

$$\bar{x} = \frac{4 \cdot 15.000 + 90.000}{5} = \text{€ } 30.000$$

Il valore della media sembra indicare che si tratti di un ottimo posto di lavoro.
Mediana degli stipendi

$$M = \text{€ } 15.000$$

La mediana rappresenta meglio della media quello che guadagna un giovane ingegnere dipendente, quindi il posto di lavoro non è così buono come era stato giudicato con la media.

Media vs Mediana

I dati seguenti rappresentano i valori dei globuli bianchi (in migliaia) rilevati in 10 pazienti ricoverati in una mattina in un ospedale

7 35 5 9 8 3 10 12 8 7

Dati ordinati in modo crescente

3 5 7 7 8 8 9 10 12 35

La media e la mediana di questi dati valgono rispettivamente

$$\bar{x} = 10.4 \qquad\qquad M = 8$$

Se il secondo paziente della tabella avesse un valore di 70.000 globuli bianchi, anziché di 35.000, il valore della mediana resterebbe invariato, mentre la media diventerebbe

$$\bar{x} = 13.9$$

Questi esempi ci ricordano che c'è sempre comunque un rischio a riassumere un insieme di dati con un singolo numero.

Percentili e Quartili Campionari

- Oltre alla mediana, che divide a metà un insieme di dati ordinati, si possono definire altri indici di posizione, detti **quantili** e **percentili**, che dividono l'insieme di dati ordinati in un dato numero di parti uguali.
- Questi indici di posizione non centrale sono **usati soprattutto per ampi insiemi di dati**.
- I **quartili** sono un caso particolare dei quantili, e si ottengono dividendo l'insieme di dati ordinati in quattro parti uguali.

DEFINIZIONI

Il **primo quartile** $\boldsymbol{Q_1}$ è un valore tale che il 25 % dei dati ordinati è minore o uguale a Q_1. Il primo quartile Q_1 è detto anche **25-esimo percentile** e indicato con $\boldsymbol{P_{0.25}}$.
Il **terzo quartile** $\boldsymbol{Q_3}$ è un valore tale che il 75 % dei dati ordinati è minore o uguale a Q_3 ed è detto anche **75-esimo percentile** e indicato con $\boldsymbol{P_{0.75}}$.

Il secondo quartile Q_2 (50-esimo percentile) coincide con la mediana.

Percentili e Quartili Campionari

Regola per il calcolo dei quartili

1 – Si ordinano gli n dati assegnati in ordine crescente;

2 – si calcola il prodotto $k = np$, dove $p = 0.25$ per il primo quartile e $p = 0.75$ per il terzo quartile;

3 – se k è un intero, il quartile si ottiene facendo la media del k-esimo e del $(k+1)$-esimo valore dei dati ordinati;

4 – se k non è intero, si arrotonda k per eccesso al primo intero successivo e si sceglie come quartile il corrispondente valore dei dati ordinati.

La regola può essere generalizzata in modo semplice per trovare un qualsiasi altro percentile.
Ad esempio per trovare il 95-esimo percentile, ossia quel valore tale che il 95 % dei dati ordinati è minore o uguale ad esso, si usa la stessa regola, con $p = 0.95$.

ESERCIZI ED ESEMPI

STATISTICHE CAMPIONARIE 1

(percentili campionari)

Percentili e Quartili Campionari

Esempio 3.9 Quale valore è il 90-esimo percentile campionario se la numerosità del campione è (a) 8, (b) 16, e (c) 100?

Soluzione

(a) Dato che $0.9 \times 8 = 7.2$, che non è un intero, otteniamo che se i dati sono disposti in ordine crescente, allora il 90-esimo percentile campionario è l'ottavo valore dal più piccolo (quindi il valore più grande)

(b) Dato che $0.9 \times 16 = 14.4$, che non è un intero, otteniamo che il 90-esimo percentile è il 15-esimo valore dal più piccolo.

(c) Dato che $0.9 \times 100 = 90$ è un intero, il 90-esimo percentile campionario è la media del 90-esimo valore e del 91-esimo valore in ordine crescente. ■

Esempio 3.10 La Tabella 3.1 elenca le università degli Stati Uniti che ricevono più donazioni. Usando questi dati, calcola

(a) Il 90-esimo percentile campionario.

(b) Il 20-esimo percentile campionario.

Percentili e Quartili Campionari

Esempio 3.10 La Tabella 3.1 elenca le università degli Stati Uniti che ricevono più donazioni. Usando questi dati, calcola

(a) Il 90-esimo percentile campionario.

(b) Il 20-esimo percentile campionario.

Tabella 3.1 20 college e università degli Stati Uniti con le più alte donazioni, 2002*

College/Università	Donazioni totali	College/Università	Donazioni totali
1. Harvard University	$ 17 169 757	11. Washington University	$ 3 517 104
2. Yale University	10 523 600	12. University of Pennsylvania	3 393 297
3. University of Texas System	8 630 679	13. University of Michigan	3 375 689
4. Princeton University	8 319 600	14. University of Chicago	3 255 368
5. Stanford University	7 613 000	15. Northwestern University	3 022 733
6. Massachusetts Institute of Technology	5 359 423	16. Rice University	2 939 804
7. Emory University	4 551 873	17. Duke University	2 927 478
8. Columbia University	4 208 373	18. Cornell University	2 853 742
9. University of California	4 199 067	19. University of Notre Dame	2 554 004
10. The Texas A&M University System and Foundations	3 743 442	20. Dartmouth College	2 186 610

Nota: valore di mercato delle donazioni, escluse le offerte e il capitale operativo.
* Aggiornato al 30 giugno 2002
College and University Business Officers (NACUBO).

Soluzione

(a) Dato che la numerosità del campione è 20 e $20 \times 0.9 = 18$, il 90-esimo percentile campionario è la media del 18-esimo e del 19-esimo valore contati dal più piccolo. In altre parole, è la media del secondo e del terzo valore contati dal più grande. Quindi,

$$\text{90-esimo percentile campionario} = \frac{10\,523\,600 + 8\,630\,679}{2} = 9\,977\,140$$

Percentili e Quartili Campionari

Calcolare il primo e il terzo quartile dell'insieme di dati

32.2 32.0 30.4 31.0 31.2 31.3 30.3 29.6 30.5 30.7

Dati ordinati

29.6 30.3 30.4 30.5 30.7 31.0 31.2 31.3 32.0 32.2

Primo quartile

$$n = 10 \qquad p = 0.25 \qquad k = np = 2.5$$

k non è intero, perciò si arrotonda per eccesso $k = 3$: il primo quartile è il terzo dei dati ordinati

$$Q_1 = 30.4.$$

Terzo quartile

$$n = 10 \qquad p = 0.75 \qquad k = np = 7.5$$

k non è intero, perciò si arrotonda per eccesso $k = 8$: il terzo quartile è l'ottavo dei dati ordinati

$$Q_3 = 31.3.$$

Secondo quartile (mediana)

$$n = 10 \qquad p = 0.5 \qquad k = np = 5$$

k è intero, perciò si fa la media tra il quinto e il sesto dato e si ottiene

$$Q_2 = \frac{30.7 + 31.0}{2} = 30.85$$

Percentili e Quartili Campionari

Calcolare quartili e **95-mo percentile** per i dati seguenti (Esempio 2, Lezione 1)

6.2	7.7	8.3	9.0	9.4	9.8	10.5	10.7	11.0	11.2
11.8	12.3	12.8	13.2	13.3	13.5	13.9	14.4	14.5	14.7
15.2	15.5	15.8	15.9	16.2	16.7	16.9	17.0	17.3	17.5
17.6	17.9	18.0	18.0	18.1	18.1	18.4	18.5	18.7	19.0
19.1	19.2	19.3	19.4	19.4	20.0	20.1	20.1	20.4	20.5
20.8	20.9	21.4	21.6	21.9	22.3	22.5	22.7	22.7	22.9
23.0	23.5	23.7	23.9	24.1	24.3	24.6	24.6	24.8	25.7
25.9	26.1	26.4	26.6	26.8	27.5	28.5	28.6	29.6	31.8

Primo quartile

$$n = 80 \qquad p = 0.25 \qquad k = np = 20$$

k è intero, perciò si fa la media tra il 20-esimo e il 21-esimo dato e si ottiene

95-esimo percentile

$$n = 80 \qquad p = 0.95 \qquad k = np = 76$$

k è intero, perciò si fa la media tra il 76-esimo e il 77-esimo dato e si ottiene

$$P_{0.95} = \frac{27.5 + 28.5}{2} = 28.0$$

Percentili e Quartili Campionari

Calcolare quartili e **95-mo percentile** per i dati seguenti (Esempio 2, Lezione 1)

6.2	7.7	8.3	9.0	9.4	9.8	10.5	10.7	11.0	11.2
11.8	12.3	12.8	13.2	13.3	13.5	13.9	14.4	14.5	14.7
15.2	15.5	15.8	15.9	16.2	16.7	16.9	17.0	17.3	17.5
17.6	17.9	18.0	18.0	18.1	18.1	18.4	18.5	18.7	19.0
19.1	19.2	19.3	19.4	19.4	20.0	20.1	20.1	20.4	20.5
20.8	20.9	21.4	21.6	21.9	22.3	22.5	22.7	22.7	22.9
23.0	23.5	23.7	23.9	24.1	24.3	24.6	24.6	24.8	25.7
25.9	26.1	26.4	26.6	26.8	27.5	28.5	28.6	29.6	31.8

95-esimo percentile

$$n = 80 \qquad p = 0.95 \qquad k = np = 76$$

k è intero, perciò si fa la media tra il 76-esimo e il 77-esimo dato e si ottiene

$$P_{0.95} = \frac{27.5 + 28.5}{2} = 28.0$$

Il 95-esimo percentile fornisce un'importante informazione: soltanto il 5% dei dati sono maggiori di 28.0, ossia, con riferimento al tipo di dati descritti nell'esempio 2, soltanto nel 5% dei giorni l'emissione di gas inquinanti supera la soglia di 28.0 unità.

Percentili Campionari

La tabella riporta la popolazione delle 30 maggiori città americane nel 1990.

Calcoliamo il 10-mo e il 95-esimo percentile campionario

Tabella 2.6 Popolazione delle 30 maggiori città degli Stati Uniti

Posizione	Città	Residenti
1	New York, NY	7 322 564
2	Los Angeles, CA	3 485 557
3	Chicago, IL	2 783 726
4	Houston, TX	1 629 902
5	Philadelphia, PA	1 585 577
6	San Diego, CA	1 110 623
7	Detroit, MI	1 027 974
8	Dallas, TX	1 007 618
9	Phoenix, AZ	983 403
10	San Antonio, TX	935 393
11	San Jose, CA	782 224
12	Indianapolis, IN	741 952
13	Baltimora, MD	736 014
14	San Francisco, CA	723 959
15	Jacksonville, FL	672 971
16	Columbus, OH	632 945
17	Milwaukee, WI	628 088
18	Memphis, TN	610 337
19	Washington, DC	606 900
20	Boston, MA	574 283
21	Seattle, WA	516 259
22	El Paso, TX	515 342
23	Nashville-Davidson, TN	510 784
24	Cleveland, OH	505 616
25	New Orleans, LA	496 938
26	Denver, CO	467 610
27	Austin, TX	465 648
28	Fort Worth, TX	447 619
29	Oklahoma City, OK	444 724
30	Portland, OR	438 802

Fonte: Bureau of the Census, U.S. Dept. of Commerce (100 most populous cities ranked by April 1990 census; revised April 1994).

Percentili e Quartili Campionari

La tabella che segue indica il numero di medici e di dentisti per ogni 100 000 abitanti in 12 stati del Midwest americano nel 2000. I Problemi 4 e 5 sono basati su questa tabella.

4. Calcola il

Stato	Medici	Dentisti
Ohio	188	56
Indiana	146	48
Illinois	206	61
Michigan	177	64
Wisconsin	177	70
Minnesota	207	70
Iowa	141	60
Missouri	186	55
North Dakota	157	55
South Dakota	129	54
Nebraska	162	71
Kansas	166	52

Fonte: American Medical Association, *Physician Characteristics and Distribution in the U.S.*

(a) 40-esimo percentile campionario;

(b) 60-esimo percentile campionario;

(c) 80-esimo percentile campionario;

del numero di medici per ogni 100 000 abitanti.

Percentili e Quartili Campionari

Esempio 2.3.9. Il rumore si misura il decibel, indicati dal simbolo dB. Un decibel è circa la soglia di udibilità in condizioni ideali per una persona con un ottimo udito; 30 dB sono il livello sonoro di un sussurro; un tono di conversazione normale può misurare 70 dB; una radio ad alto volume arriva a 100 dB; la soglia di tollerabilità è intorno ai 120 dB. I valori seguenti sono i livelli di rumore misurati in 36 differenti occasioni in prossimità della stazione centrale di Manhattan.

82	89	94	110	74	122	112	95	100	78	65	60
90	83	87	75	114	85	69	94	124	115	107	88
97	74	72	68	83	91	90	102	77	125	108	65

Determiniamo i quartili campionari