STATISTICA DESCRITTIVA - SCHEDA N. 2
VARIABILI QUANTITATIVE
Rappresentazioni grafiche e quantili

Una variabile si dice **quantitativa** se è una grandezza misurabile. Per esempio: il peso, l'altezza, il reddito, etc... L'insieme dei valori assunti dalla variabile e le frequenze corrispondenti è detto **distribuzione**: quindi se i differenti valori assunti dalla variabile sono m, indicando con x_k tali valori e con f_k le corrispondenti frequenze, allora la distribuzione è l'insieme delle coppie (x_k , f_k) per k da 1 a m.

Attenzione alle notazioni. Indichiamo:
- con n il numero di osservazioni, con m il numero di *differenti* valori assunti dalla variabile;
- con x_i il valore della i-esima osservazione e con x_k il k-esimo valore dei dati non ripetuti.

ESEMPIO 1.

Nella tabella a sinistra sono riportati alcuni dati osservati su un campione di 18 studenti (ogni riga corrisponde alle diverse rilevazioni su un individuo). Il peso è arrotondato per semplicità ai chilogrammi. La variabile N. SCARPA pur assumendo valori interi, è meglio classificabile come variabile qualitativa ordinale in fatti non corrisponde strettamente a una misura.

Sesso	Scarpa	Peso
M	43	65
M	43	62
F	39	50
F	37	50
F	37	47
F	36	47
F	38	56
F	38	57
M	43	73
M	45	85
M	42	68
M	41	68
M	43	85
F	37	56
M	42	73
M	42	65
M	41	73
M	40	70

Tabella 1. Data set

I dati riguardano 18 studenti (n=18) mentre i *differenti* pesi sono solo 10 (m=10).

La prima e terza riga della tabella qui sotto sono la distribuzione della variabile peso.

Peso (x_k)	47	50	56	57	62	65	68	70	73	85
conteggi (n_k)	2	2	2	1	1	2	2	1	3	2
frequenze (f_k)	0.11	0.11	0.11	0.06	0.06	0.11	0.11	0.06	0.16	0.11

Tabella 2. Distribuzione del peso

Nonostante le rappresentazioni grafiche che introdurremo siano più efficaci se il numero di osservazioni è elevato, negli esempi per semplicità tratteremo poche osservazioni.

1. Diagramma di dispersione

- *Il modo più semplice di rappresentare graficamente la distribuzione di X è quello di costruire il **diagramma di dispersione** (**dotplot**). È simile al diagramma a barre per le variabili qualitative: si ottiene riportando in un grafico un punto per ogni valore assunto dalla variabile. Sull'asse orizzontale sono rappresentati i valori di X: in corrispondenza a ogni valore assunto si disegna un numero di punti proporzionale al numero delle osservazioni.*

ESEMPIO 1 (continua)

Il dot- plot per i dati precedenti è il seguente.

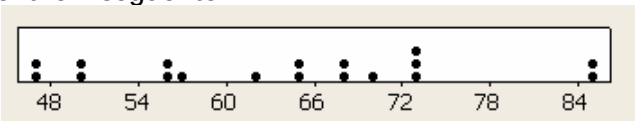


Figura 1. Dot-plot della variabile Peso

Se i valori delle osservazioni differenti sono fitti e numerosi, molti software statistici introducono approssimazioni nella scala dei valori di X e ad ogni punto fanno corrispondere più unità sperimentali.

2. Funzione di distribuzione cumulata.

- *Un'ulteriore rappresentazione di una variabile quantitativa X è la* ***funzione di distribuzione cumulata F*** *(o di ripartizione)****,*** *ovvero F(x) è la* frequenza f di tutte *le osservazioni minori o uguali a x; cioè*

$$F(x) = f(X \le x) = \sum_{k,\ \text{con}\ x_k \le x} f_k$$

Per approndire le notazioni sulle somme vedi Appendice 3 della Scheda 1.

ESEMPIO 1 (continua).

Per costruire la funzione di distribuzione cumulata si aggiungono alla tabella della distribuzione della variabile e le frequenze relative cumulate (ottenute sommando le frequenze relative dei dati inferiori o uguali al valore considerato). Nel caso della variabile PESO.

Peso	47	50	56	57	62	65	68	70	73	85
frequenza	0.11	0.11	0.11	0.06	0.06	0.11	0.11	0.06	0.16	0.11
freq. cum.	0.11	0.22	0.33	0.39	0.45	0.56	0.67	0.73	0.89	1,00

Frequenza delle osservazioni minori o uguali a 56 (0.33=0.11+0.11+0,11)

Tabella 2. Distribuzione e distribuzione cumulata della variabile Peso

Per introdurre un minor numero di errori di approssimazione può essere più opportuno costruire la funzione di distribuzione cumulata a partire dai conteggi cumulati.

Peso	47	50	56	57	62	65	68	70	73	85
conteggi	2	2	2	1	1	2	2	1	3	2
cont. cum.	2	4	6	7	8	10	12	13	16	18
freq. cum.	0.11	0.22	0.33	0.39	0.45	0.56	0.67	0.73	0.89	1,00

Tabella 3. Conteggi, conteggi cumulati e distribuzione cumulata della variabile Peso

La funzione di distribuzione cumulata avrà il seguente grafico

Osserviamo che la funzione di distribuzione cumulata è definita anche in corrispondenza di valori non assunti dai dati.
Ad esempio, anche se nessun soggetto ha peso 60 kg, possiamo comunque dire che la frequenza relativa dei soggetti con peso minore o uguale a 60 è 0.39.

Inoltre la funzione vale 0 per tutti i valori inferiori al più piccolo e 1 per tutti quelli superiori al più grande.

NB: Questo a fianco **è** il grafico di una funzione.

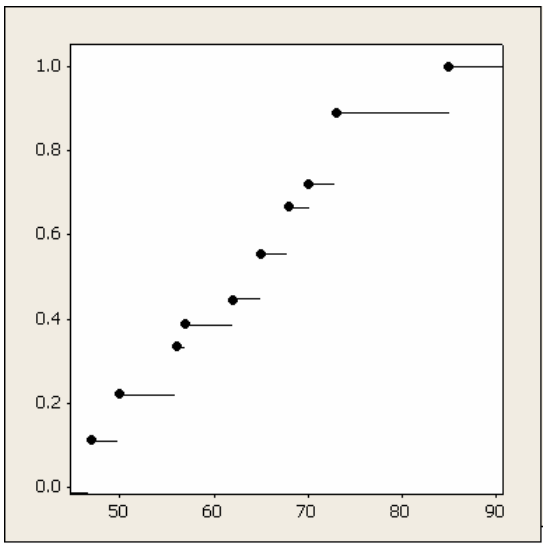
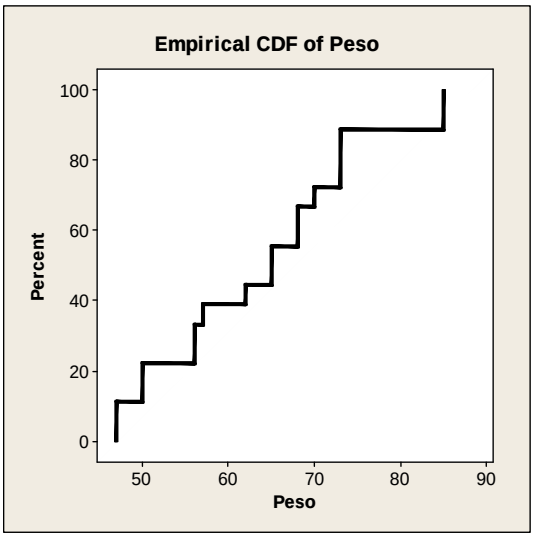


Figura 1. Funzione di distribuzione cumulata della variabile Peso

La funzione di distribuzione cumulata è una funzione a "scalini" e ha le seguenti proprietà

1. F è una funzione crescente o costante;
2. in corrispondenza di ogni punto di salto la funzione assume il valore a sinistra.
3. la funzione vale 0 per ogni valore minore all'osservazione minima e vale 1 per ogni valore maggiore o uguale all'osservazione massima.

Il software Minitab® non costruisce esattamente il grafico della funzione di distribuzione cumulata, ma le seguenti due rappresentazioni grafiche .



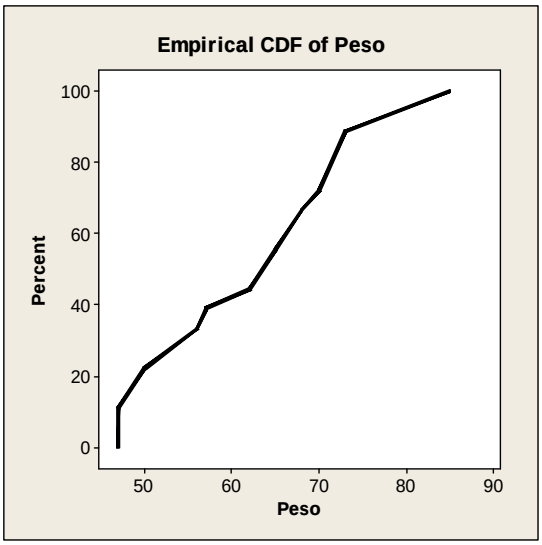


Figura 2. Distribuzione cumulata della variabile Peso con Minitab

Osserviamo che la prima rappresentazione non è il grafico della funzione di distribuzione cumulata ("a una x corrispondono più y"). La seconda è il grafico di una funzione (tranne nel minimo assunto dalla variabile), ottenuta da F "interpolando" i punti. Entrambe le rappresentazioni saranno utilizzate in seguito.

Da quanto visto sopra risulta evidente che la funzione di distribuzione cumulata F e la funzione delle frequenze sono ricavabili una dall'altra.

3. Quartili e quantili.

- *Può essere interessante porsi il problema inverso del precedente., ovvero voler conoscere il valore per cui tra le osservazioni ordinate c'è una frequenza assegnata di valori minori o uguali a tale valore.*

ESEMPIO 1 (continua).

Qual è il valore "centrale" dei pesi? ovvero qual è il valore per cui ci sono metà persone con un peso inferiore e metà persone con un peso superiore?

Oppure qual è il valore del primo quarto dei pesi ordinati? E dell'ultimo quarto?

Le domande poste sono in realtà un po' ambigue.

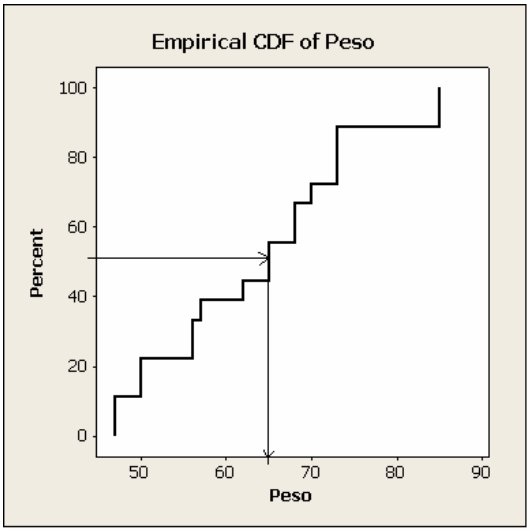


Figura 3. Distribuzione cumulata del Peso

Più precisamente, si definiscono:

- ✓ **Mediana (Q2):** il minimo valore osservato tale che *almeno* il 50% (=1/2)dei dati è *minore o uguali* a questo.
- ✓ **Primo quartile (Q1)** il minimo valore osservato tale che *almeno* il 25% (=1/4) dei dati è *minore o uguali* a questo.
- ✓ **Terzo quartile (Q3)** il minimo valore osservato tale che *almeno* il 75% (=3/4) dei dati è *minore o uguali* a questo.

ESEMPIO 1 (continua).

Per rispondere alle domande precedenti usando le definizioni corrette, utilizziamo il grafico della funzione di distribuzione cumulata.
Consideriamo il primo quartile Q1 (25%).
Sull'asse delle ordinate si individua il punto 0.25 e da questo si traccia una linea orizzontale: in questo caso la linea non interseca il grafico della funzione di distribuzione cumulata. Il minimo valore osservato la cui funzione di distribuzione cumulata **supera** 0.25 è 56; F(50)=0.22 e F(56)=0.33.

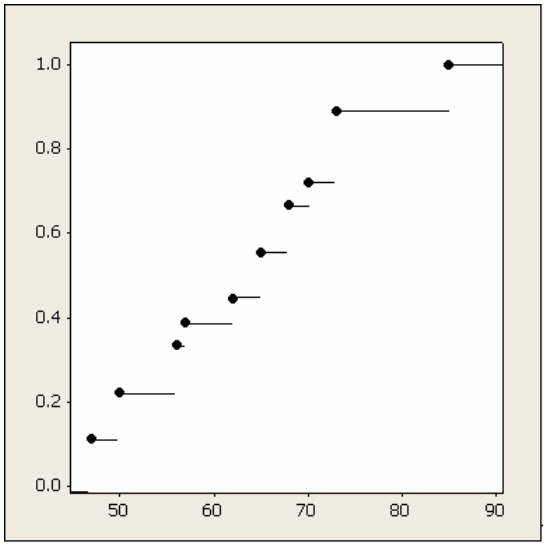


Figura 1. (ripetuta)

Possiamo determinare i quartili anche usando la lista ordinata dei dati ripetuti. Nell'esempio della variabile PESO si ha:

n. ordine	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
valori	47	47	50	50	56	56	57	62	65	65	68	68	70	73	73	73	85	85

Mediana: le due osservazioni centrali (nono e decimo dato) sono 65 quindi la mediana è 65.

Q1: il primo quarto dei dati ordinati, ha un valore minore o uguale a 56. Infatti 0.25x18=4,5 e la quinta osservazione è 56.
Q3: il terzo quarto dei dati ordinati è minore o uguale a 73. Infatti 0.75x18=13,5 e la quattordicesima osservazione è 73.

In generale, dato α, compreso tra 0 e 1, si dice **α-esimo quantile** (ad esempio α=0.20) il minimo valore osservato per cui almeno l'α-esima parte (il 20%) dei dati risultino minori o uguali a questo:

il valore dell'α-esimo quantile è: $\min\{x \text{ osservato tale che } F(x) \geq \alpha\}$

Se α è espresso in forma percentuale, invece che di quantili si parla di percentili.
In pratica, per calcolare il valore dell'α-esimo quantile, è sufficiente scegliere l'i-esimo dato, dove i è l'approssimazione per eccesso del prodotto N$\times\alpha$ (N è il numero totale delle osservazioni). Ad esempio il 0.2-quantile è il quarto dato ($18\times0.2=3,6\rightarrow4$).

Attenzione: i quantili sono dei valori e non delle posizioni.

NB: Molti software (come Minitab® e Excel®) hanno diversi algoritmi per calcolare i quantili. È da osservare che nel caso in cui il numero delle osservazioni sia molto elevato questi tendono a coincidere.

ESEMPIO 1 (continua).

Il software Minitab® calcola i quantili interpolando i valori della funzione: come si vede dal grafico a fianco per il 20-esimo percentile.

I valori forniti da Minitab® per la variabile Peso sono:

Variable	N	Minimum	Q1	Median	Q3	Maximum
Peso	18	47.00	54.50	65.00	73.00	85.00

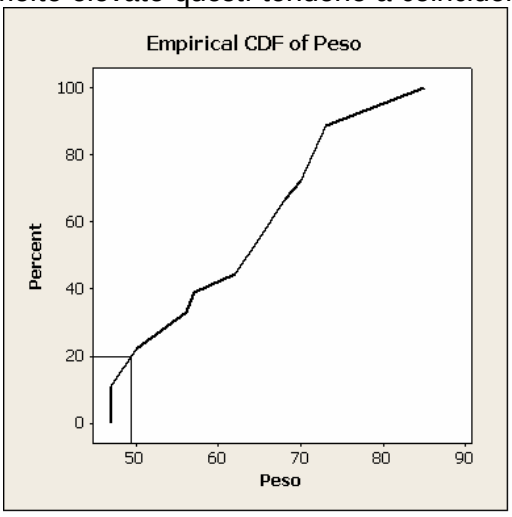


Figura 4. Distribuzione cumulata del Peso

Due indici che forniscono informazioni sulla dispersione dell'insieme dei dati osservati sono:

- il valore Max- Min, detto anche misura dell'intervallo di variazione o range; nell'esempio: 85-47, cioè 38.
- il valore Q3-Q1, detto **distanza interquartile** (**IQR**, dall'inglese Inter Quartile Range) e coincide con l'ampiezza dell'intervallo in cui si trova almeno il 50% dei dati. Nell'esempio precedente, usando i valori di Minitab, IQR=73-54.5, cioè 18.5.

La mediana fornisce informazioni sulla centralità delle osservazioni. Ne conosci altri?

4. Box-plot

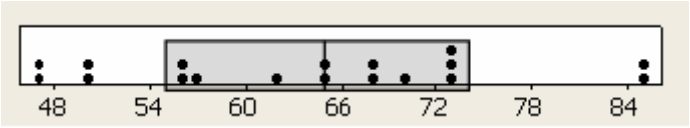
- *Una rappresentazione grafica che si basa sulla definizione dei quantili è il **box-plot**. Pur fornendo minori informazioni rispetto alla funzione di distribuzione cumulata, permette di descrivere la variabile in maniera sintetica ed è molto utile per confrontare sottogruppi di dati.*

L'idea è quella di individuare con una "scatola" le osservazioni centrali e con dei "baffi" o code uscenti dalla scatola le osservazioni più estreme.

Vediamo come si costruisce a partire dal dotplot: si disegna una scatola tra i valori Q1 e Q3. Con una linea verticale si individua la mediana (Q2).

ESEMPIO 1 (continua).

Per il Peso, usando i quartili forniti da Minitab, si ottiene:



Si disegnano poi i baffi che sono lunghi *al più* una volta e mezza la distanza interquartile e terminano in corrispondenza del dato più lontano dalla scatola inferiore a tale valore.
I valori limite L e R per i baffi sono quindi:

L=Q1-1,5XIQR e R=Q3+1,5XIQR.

I valori che rimangono al di fuori dei limiti R e L, si individuano con asterischi.
Nell'esempio: L = 54.5-1.5x 18.5 = 26.75
R = 73+1.5x 18.5 = 100.75
Quindi i baffi si fermano al valore minimo e a quello massimo.

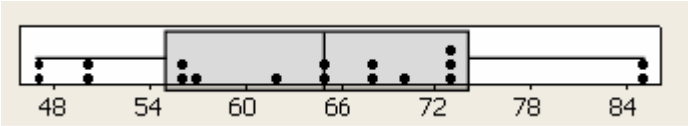


Figura 5. Costruzione del box-plot

Nel boxplot non vengono disegnati i punti rappresentanti i dati, ma solo la scatola, i baffi e gli eventuali dati estremi.

Come già detto i boxplot sono molto utili per confrontare i dati di sottogruppi di soggetti,

ESEMPIO 2.

Sono stati rilevate le pulsazioni cardiache in un minuto di un gruppo di studenti. Alcuni di questi prima della rilevazione hanno effettuato un minuto di corsa, altri no.

Il boxplot a fianco rappresenta i dati dell'intero gruppo di studenti.
Si può osservare che 4 studenti hanno pulsazioni non sono comprese nei baffi.
I dati così rappresentati però non sono omogenei; ovviamente le pulsazioni variano molto tra chi ha corso e chi non ha corso.

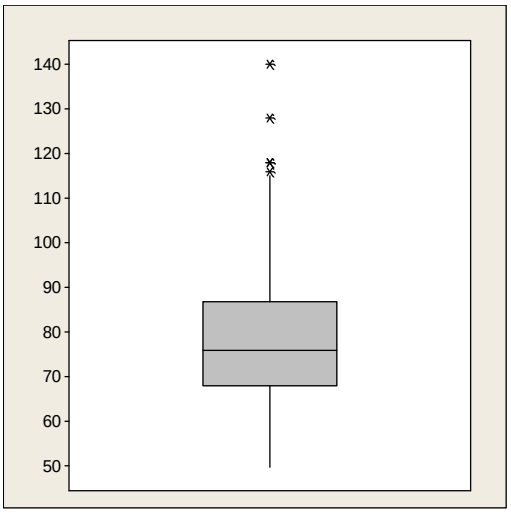
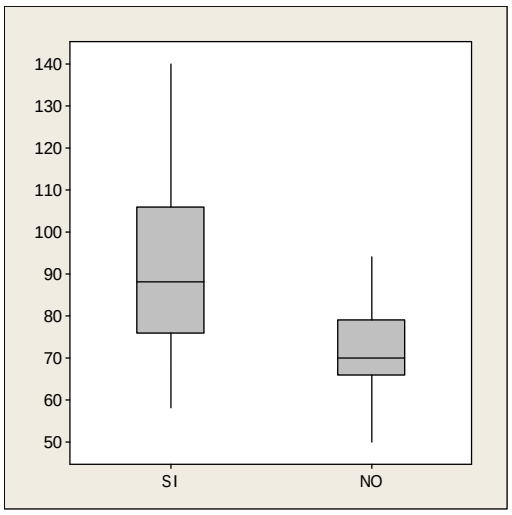


Figura 6. Box-plot delle pulsazioni cardiache

Qui sotto sono rappresentati i boxplot per i due gruppi.



Si può osservare che le due distribuzioni assumono valori su intervalli diversi e che ciascun quartile delle pulsazioni di chi non ha corso è più basso del corrispondente quartile di chi ha corso.
Inoltre le pulsazioni di chi non ha corso sono più concentrate sia nella parte centrale che nelle code.
Anche le simmetrie sono diverse, in particolare i baffi sono più asimmetrici per chi ha corso. Questo e la maggiore dispersione della distribuzione si possono spiegare pensando che la reazione alla corsa varia molto da individuo a individuo: la dispersione aumenta e questo avviene soprattutto per valori alti.

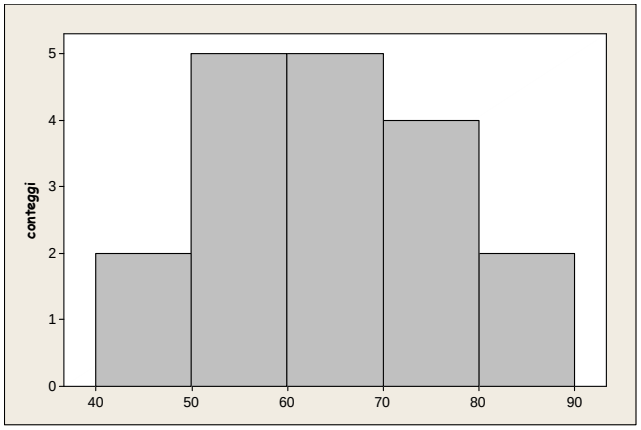
Figura 7. Box-plot delle pulsazioni cardiache per gruppi

4. Istogramma

- *Infine vediamo una rappresentazione grafica non sempre efficace: l'**istogramma.***

Si suddivide l'intervallo in cui variano i dati in classi (preferibilmente di uguale ampiezza) e si assegna ogni osservazione rilevata alla classe corrispondente. La scelta del numero di classi non è indifferente: troppo poche appiattiscono il grafico fino a renderlo insignificante; troppe classi introducono tra le barre oscillazioni eccessive, che potrebbero distruggere l'eventuale "regolarita'" dell'istogramma. L'istogramma si disegna come i diagrammi a barre per le variabili qualitative, ma facendo attenzione che i "rettangoli" verticali devono essere adiacenti ed avere come vertici i punti che separano le classi.

ESEMPIO 1 (continua). Consideriamo la variabile PESO dell'esempio precedente.



peso	Conteggi	frequenze
40-49	2	11,1%
50-59	5	27,5%
60-69	5	27,5%
70-79	4	22,8%
80-89	2	11,1%

Figura 8. Istogramma della variabile Peso

La rappresentazione dei dati tramite istogrammi è da usare con molta cautela perché la suddivisione dei dati in classi è in genere arbitraria, in particolare se le classi sono ampie o se i dati sono pochi. Solo in casi particolari le classi sono stabilite dal contesto che si sta esaminando; ad esempio scaglioni di reddito, ….

ESEMPIO 1 (continua).

Riprendendo l'esempio della variabile Peso. Se si suddivide l'intervallo dei dati osservati in 10 classi si ottiene l'istogramma a fianco e questa rappresentazione sembra indurre conclusioni differenti dalla precedente.

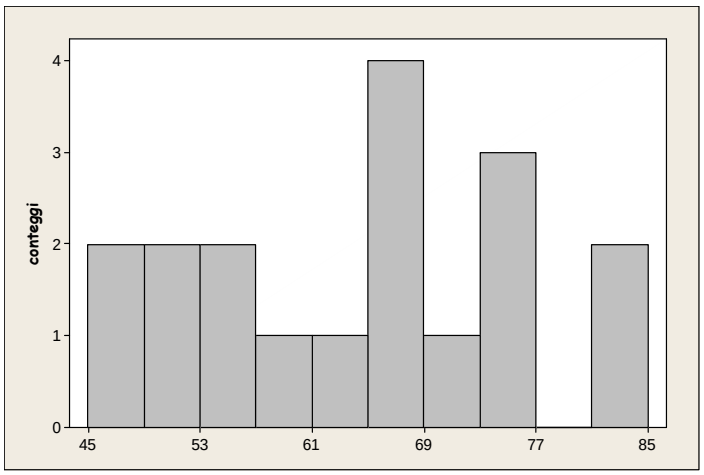


Figura 9. Istogramma della variabile Peso con classi di uguale ampiezza

Nel caso in cui si scelgano classi con ampiezza differente si devono costruire rettangoli la cui AREA sia proporzionale alla frequenza.

Nel caso in cui si scelgano classi di uguale ampiezza, il fatto che l'area sia proporzionale all'ampiezza segue dal fatto che le altezze delle barre lo sono.

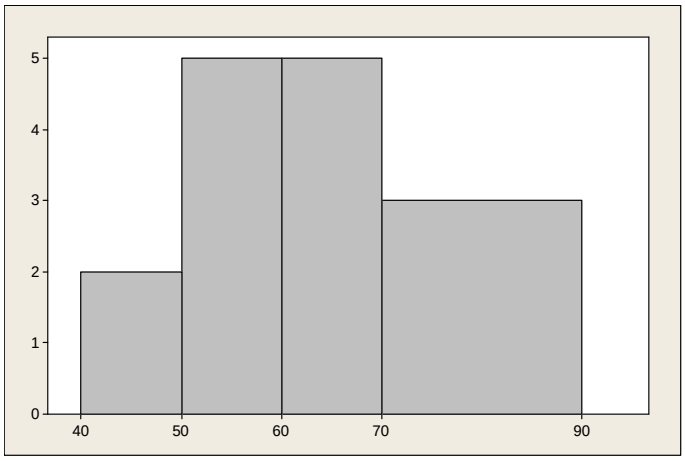


Figura 10. Istogramma della variabile Peso con classi di diversa ampiezza

ESEMPIO 3.

Nella seguente tabella sono riportati i dati relativi all'epoca di costruzione delle abitazioni del Comune di Genova (Censimento generale della popolazione 1991). Il numero di abitazioni è in migliaia.

epoca	pre 1919	19 ├ 46	46 ├ 61	61 ├ 71	71 ├ 82	82 ├ 87	0ltre 86
n. abitaz.	136	70	104	130	35	11	5

dove con ├ si è indicato il fatto che la classe è chiusa a sinistra, cio è, ad esempio 19 ├ 46 corrisponde all'intervallo [19, 46). Consideriamo, per semplicità della rappresentazione, che la prima classe inizi al 1850.

Come si vede, le classi hanno ampiezza diversa; quindi per avere una rappresentazione grafica significativa bisogna fare in modo, ad esempio, che le 136 mila case costruite dal 1850 al 1918 (69 anni) abbiano minor "peso" rispetto a quello delle 70 mila abitazioni costruite dal 1919 al 1945 (27 anni). Per questo l'istogramma viene costruito con le aree di ciascuna classe proporzionali alle frequenze: le altezze devono essere quindi proporzionali rapporto fra l'ampiezza dell'intervallo e le frequenze.

Calcoliamo quindi, per ciascuna classe, l'ampiezza dell'intervallo (base) e la corrispondente altezza.

epoca	850├ 919	19 ├ 46	46 ├ 61	61 ├ 71	71 ├ 82	82 ├ 87	87├ ├
amp. classe	69	27	15	10	11	5	5
n. abitaz.	136	70	104	130	35	11	5
alt. istogr	1.97`	2.59	6.93	13.0	3.18	2.20	1

Quindi il corrispondente istogramma è:

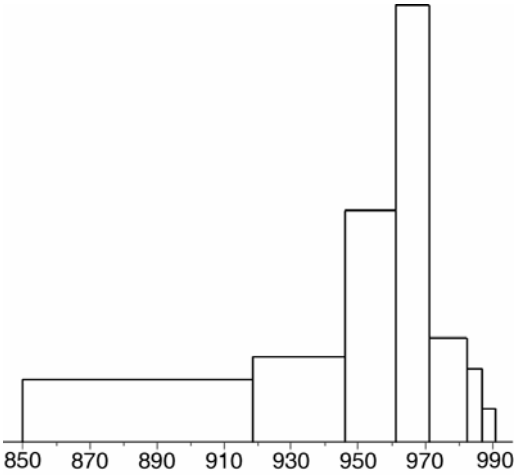
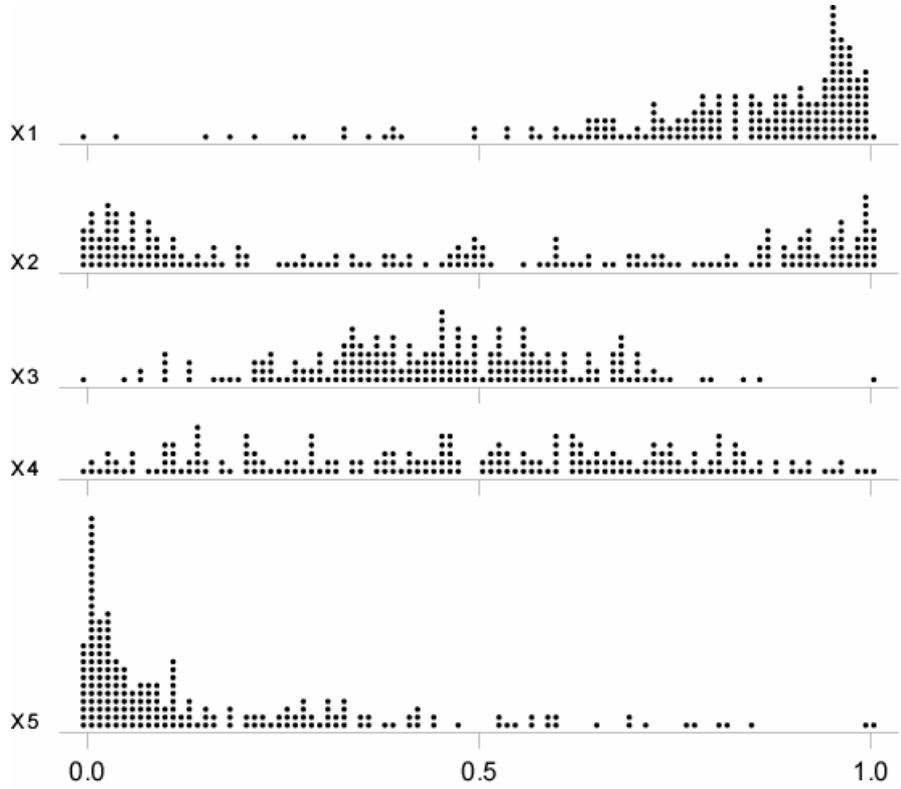


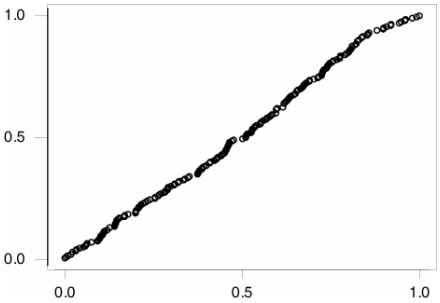
Figura 11. Istogramma della variabile Età delle case con classi di diversa ampiezza

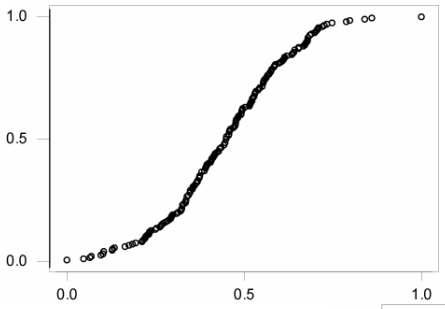
ESERCIZI

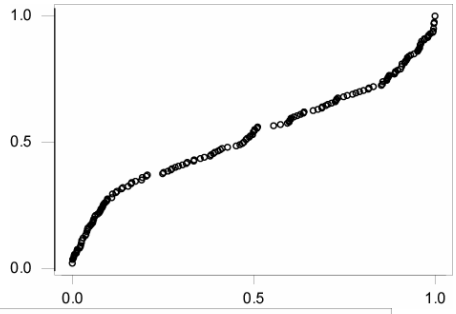
1) Qui sotto sono riportati i diagrammi di dispersione di 5 rilevazioni statistiche, indicate con X1, X2, X3, X4 e X5.



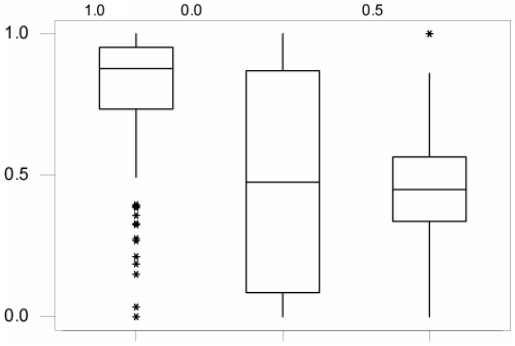
a) Qui sotto sono riportati i grafici delle funzioni di distribuzione cumulata di 3 di queste rilevazioni. Assegnare a ciascun grafico la rilevazione corrispondente.







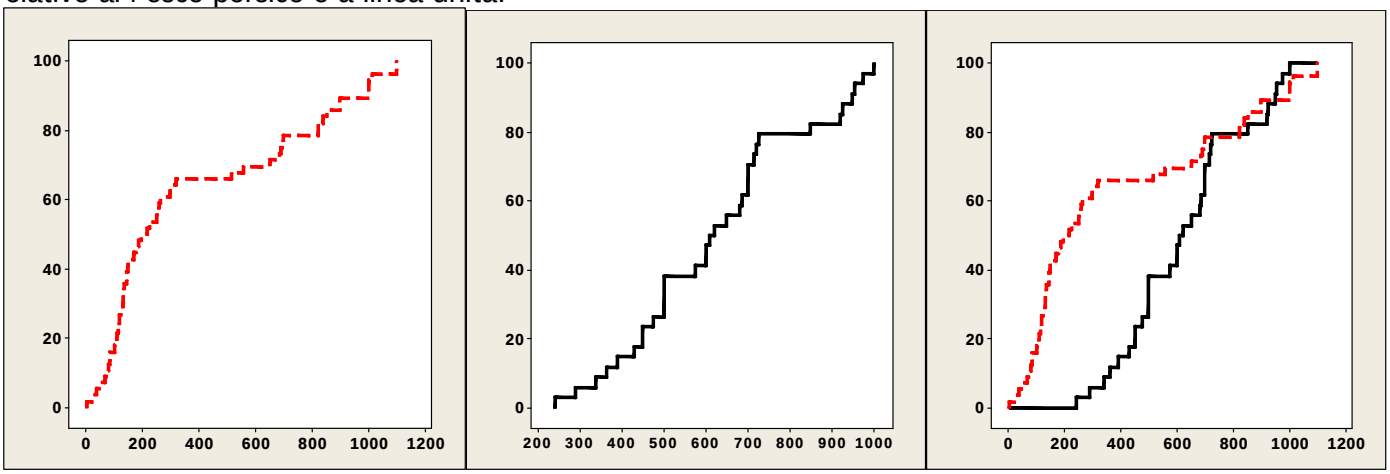
b) Qui a fianco sono riportati i box-plot di 3 rilevazioni (non necessariamente le stesse del punto precedente). Assegnare a ciascun box-plot la rilevazione corrispondente.



2) I box-plot di due rilevazioni statistiche sono perfettamente uguali. Si può concludere che i due insiemi di dati:

- hanno la stessa mediana?
- hanno la stessa funzione di distribuzione cumulata?
- sono ugualmente simmetrici (o ugualmente asimmetrici) rispetto al valore medio?

3) Qui sotto sono riportati i grafici delle funzioni di distribuzione cumulata del peso in grammi di due gruppi di pesci di specie diversa, prima su piani cartesiani diversi (attenzione: le scale delle ascisse non sono uguali) e poi nello stesso piano. Il grafico relativo all'Abramide è a linea tratteggiata e quello relativo al Pesce persico è a linea unita.

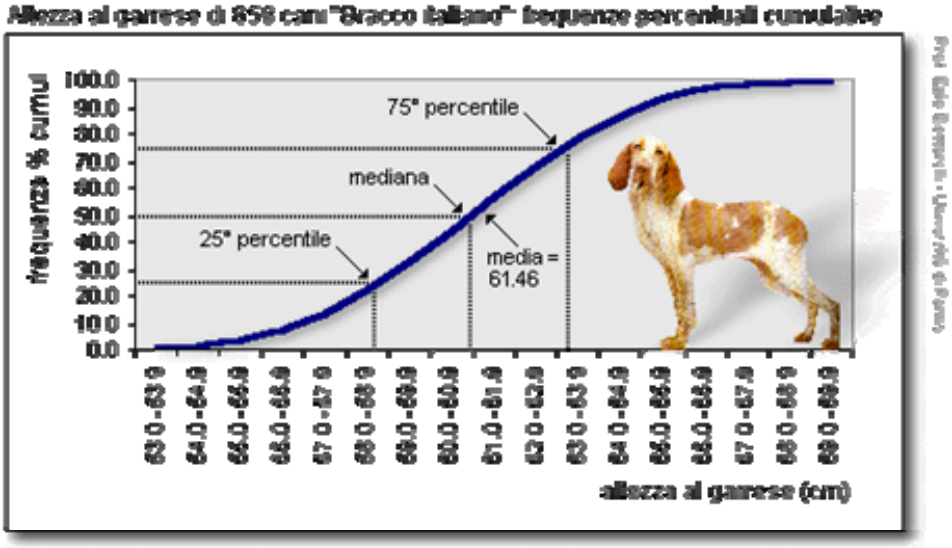


a) Indicare quali delle seguenti affermazioni relative alla "forma" della distribuzione del peso delle due specie sono vere e quali false.

- Il peso del Pesce persico ha una distribuzione simmetrica rispetto alla mediana
- Il peso dell'Abramide ha una distribuzione simmetrica rispetto alla mediana
- I pesi dei due tipi di pesci assumono valori su uno stesso intervallo
- La maggior parte dei Pesci persisco è più leggera di di tutti i pesci Abramide
- La maggior parte degli Abramide è più pesante di tutti i Pesci persico
- Nell'intervallo corrispondente all'ultimo 25% dei dati le distribuzioni dei pesi delle due specie sono molto diverse

b) Calcolare (approssimativamente) la distanza interquartile dei due pesi.

7) Il seguente grafico mostra la funzione di distribuzione cumulata dell'altezza al garrese di 659 cani di razza Bracco italiano.
(tratto dal sito http://www2.unipr.it/~bottarel/epi/varbio/freq_cu.htm)



Disegnare il box-plot corrispondente.

4) Qui a fianco sono riportati i 45 risultati di una rilevazione quantitativa, ordinati in modo crescente.

a) Indicare il primo quartile, la mediana, il secondo quartile e la distanza interquartile.

b) Costruire un box-plot per i dati.

c) Commentare la forma della distribuzione sulla base delle informazioni che si possono trarre dal box-plot.

	X		X
1	0.3957	24	10.0013
2	0.6894	25	11.7987
3	0.7948	26	13.6622
4	1.3256	27	14.5463
5	1.7152	28	14.7880
6	2.0866	29	15.5402
7	2.1393	30	17.9997
8	2.2957	31	20.0579
9	2.3387	32	20.2377
10	2.5338	33	21.0529
11	2.7304	34	21.5234
12	2.9806	35	24.0151
13	3.4939	36	28.3879
14	3.7586	37	28.8789
15	4.6255	38	29.6174
16	5.9514	39	34.2406
17	6.5671	40	39.6083
18	6.5890	41	41.5054
19	7.0588	42	45.1018
20	8.0917	43	49.7889
21	8.4679	44	57.1067
22	9.1634	45	90.0761
23	9.9958		