Introduzione alla Statistica

(Metodo delle Scienze Empiriche)

Distribuzioni di Frequenza
Percentili

Sezione di Epidemiologia e Statistica Medica
Università degli Studi di Verona

Distribuzione di Frequenza

Tabella che riporta i livelli assunti da una variabile e le frequenze corrispondenti.

Livelli	Tipo di Variabile
Modalità	**Variabile qualitativa**
Valori	**Variabile quantitativa**

Distribuzione di Frequenza

Rappresentazione tabellare:

- In genere si utilizza per rappresentare variabili di tipo **qualitativo** (nominali od ordinali). A ciascuna modalità assunta dalla variabile corrisponde una determinata frequenza osservata.
- Spesso, per praticità o per carenze degli strumenti di misura, si opera una "**discretizzazione"** di variabili espresse su scala continua, esprimendo i livelli (*valori*) assunti dalla variabile attraverso **categorie intervallari**, a ciascuna delle quali corrisponde una determinata frequenza osservata (relativa o assoluta, semplice o cumulata)

Distribuzione di Frequenza
Variabile Qualitativa

ESEMPIO: V.C. Colore degli Occhi

Modalità	Frequenza		
	Assoluta (n_i)	Relativa (p_i, f_i)	Percentuale (%)
Castani	500	0,714 *(500/700)*	71,4%
Azzurri	100	0,143 *(100/700)*	14,3%
Verdi	100	0,143 *(100/700)*	14,3%
Totale (Σ_i)	*700*	*1*	*100%*

Proprietà:

- Esaustività
- Esclusività *(non ambiguità)*

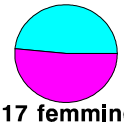
Nella classificazione dei soggetti la distribuzione di frequenza deve essere **esaustiva** (vanno riportati tutti i valori assunti dalla variabile) e **non-ambigua** (ogni soggetto deve appartenere ad una sola classe).

ESEMPIO

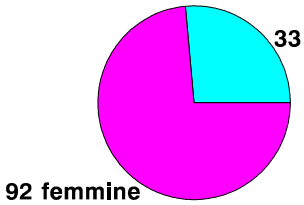
VARIABILE QUALITATIVA = SESSO

Ci sono 16 maschi fra gli specializzandi e 33 fra le matricole di Medicina (FREQUENZE ASSOLUTE, n).
Se consideriamo le frequenze assolute, i maschi tra gli specializzandi sono la metà rispetto ai maschi tra le matricole di Medicina.

SPECIALIZZANDI



MATRICOLE di MEDICINA

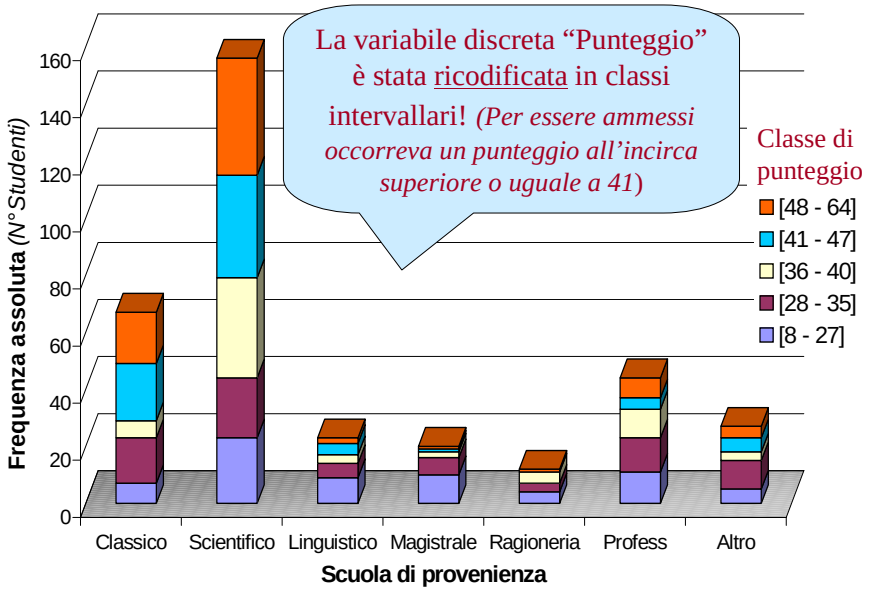


FREQUENZE RELATIVE (f,p)

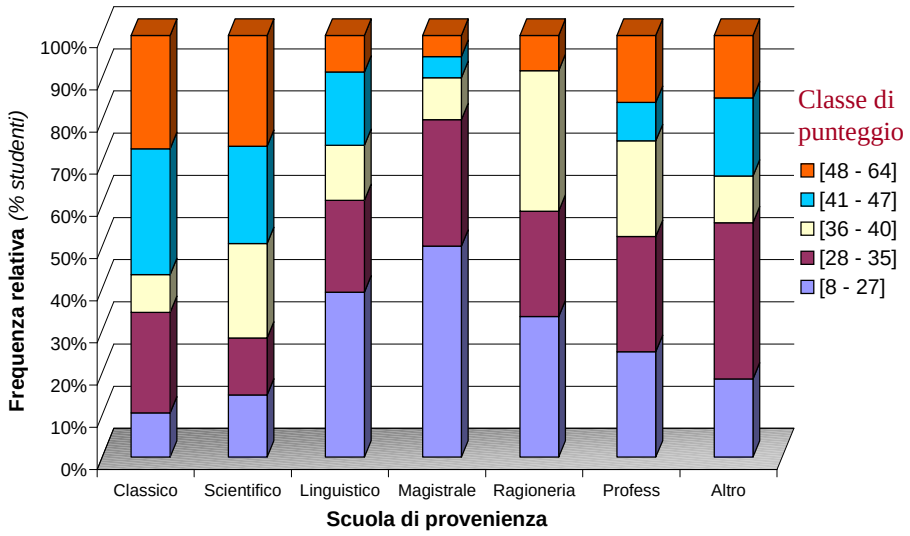
16/33 = 48,5% **33/125 = 26,4%**

In realta', il sesso maschile e' molto piu' frequente tra gli specializzandi che non tra le matricole di medicina.

Distribuzione di frequenza assoluta dei punteggi al test degli studenti di Medicina, per scuola di provenienza *(a.a. 95/96)*



Distribuzione di frequenza relativa dei punteggi al test degli studenti di Medicina, per scuola di provenienza *(a.a. 95/96)*



DISTRIBUZIONE di FREQUENZA di DUE VARIABILI QUALITATIVE

Variabile: Colore degli Occhi

Modalità	Frequenza Assoluta n_i	Percentuale f_i (%)	Cumulativa N_i
Scuri	120	80%	?
Chiari	30	20%	?
Totale (Σ_i)	*150*	*100%*	

Variabile: Colore dei Capelli

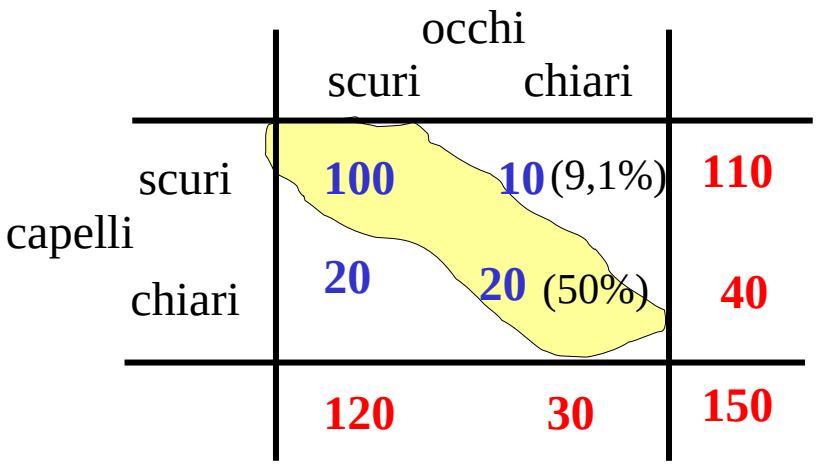
Modalità	Frequenza Assoluta n_i	Percentuale f_i (%)	Cumulativa N_i
Scuri	110	73,3%	?
Chiari	40	26,7%	?
Totale (Σ_i)	*150*	*100%*	

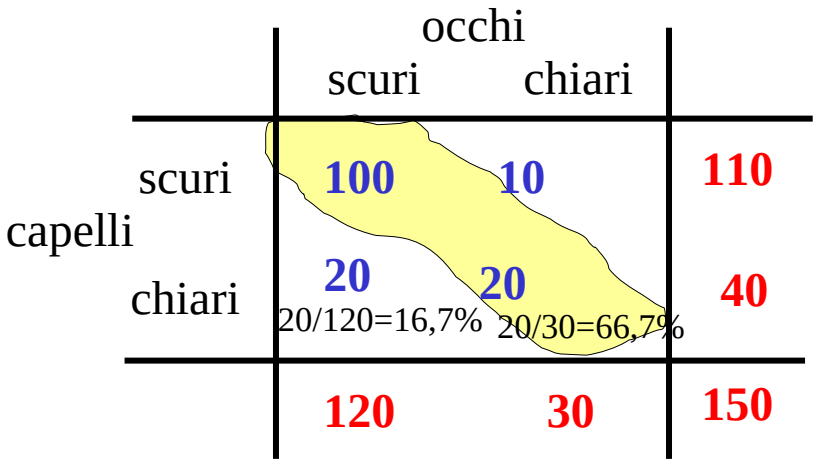
TABELLE di CONTINGENZA 2 x 2
(Fourfold Tables)

		Colore occhi: Scuri	Colore occhi: Chiari	
Colore capelli	Scuri	100 *(90.9%)*	10 *(9.1%)*	*110* *(100%)*
	Chiari	20 *(50%)*	20 *(50)%*	*40* *(100%)*
		120	*30*	*150*

Frequenze Marginali

Le frequenze marginali corrispondono alle frequenze delle distribuzioni di frequenza univariata.





ESERCIZIO: Costruzione di una tabella di contingenza 2*2

DATI: Abbiamo 1000 individui anziani, 100 sono diabetici e 300 sono ipertesi. 70 individui hanno sia il diabete che l'ipertensione.

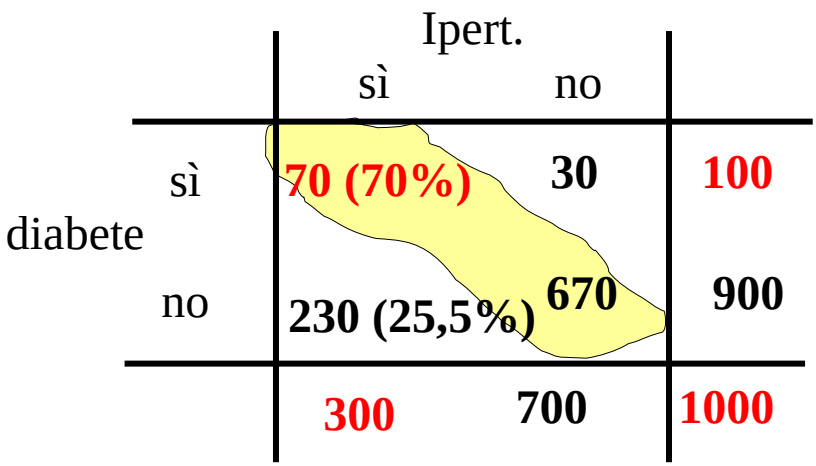
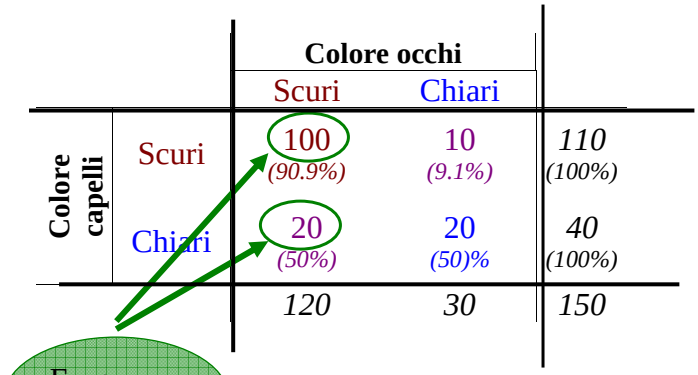


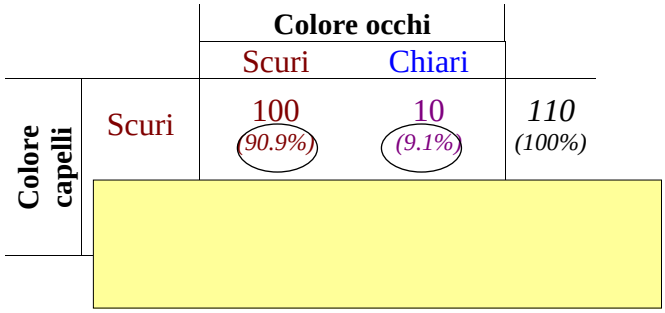
TABELLE di CONTINGENZA 2 x 2
(Fourfold Tables)



Frequenze Congiunte

100 soggetti si trovano all'incrocio tra la prima riga (capelli scuri) e la prima colonna (occhi scuri). Pertanto hanno sia gli occhi che i capelli scuri.

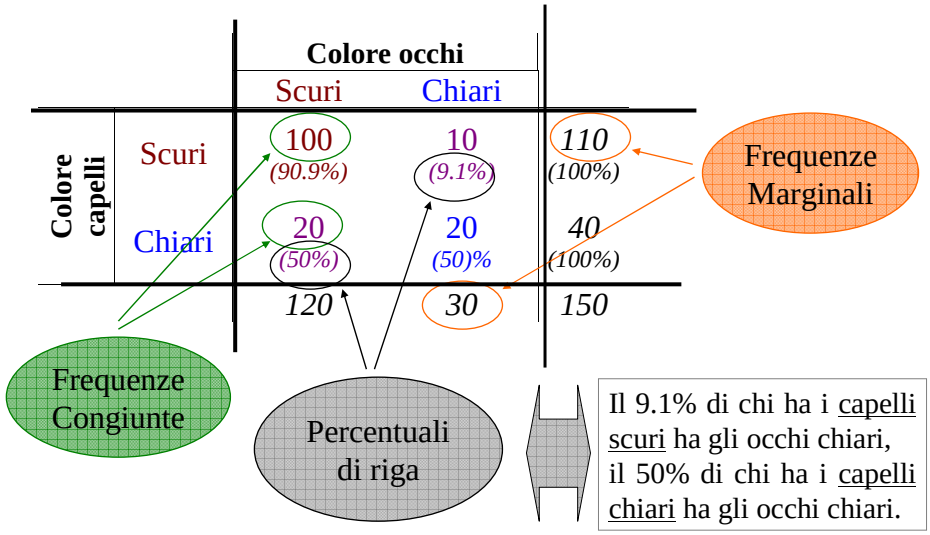
TABELLE di CONTINGENZA 2 x 2
(Fourfold Tables)



Percentuali di riga

Per calcolare una percentuale di riga mi concentro su una sola riga (la prima o la seconda) come se costituisse da sola l'intero campione.

TABELLE di CONTINGENZA 2 x 2
(Fourfold Tables)



ESERCIZIO: Costruzione di una tabella di contingenza 2*2

DATI: Abbiamo 1000 individui anziani, 100 sono diabetici e 300 sono ipertesi. 70 individui hanno sia il diabete che l'ipertensione.

	Iperteso	Normoteso	
Diabetico	70	30	100
Non-diabetico	230	670	900
	300	700	1000

% di ipertesi fra i diabetici =70/100 = 0,70 = 70%
% di ipertesi fra i non- diabetici = 230/900 = 0,255 = 25,5%

CONCLUSIONE: Il diabete e l'ipertensione sono due malattie fortemente collegate.

Esperimento di Mendel:

incrocio di piselli lisci e gialli (caratteri dominanti) e rugosi e verdi (caratteri recessivi),

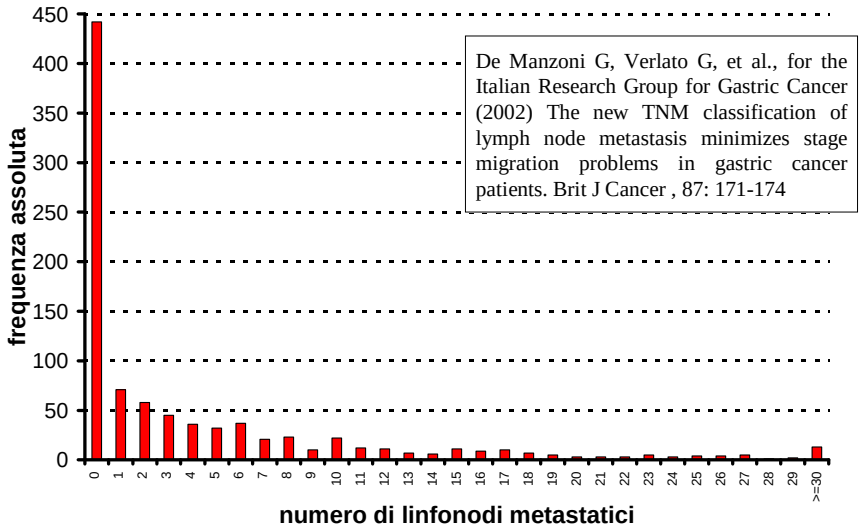
e incrocio degli ibridi di I generazione.

	giallo	verde	
Liscio	315	108	423
Rugoso	101	32	133
	416	140	556

% di piselli verdi fra i piselli lisci = 108/423 = 0,255 = 25,5%
% di piselli verdi fra i piselli rugosi = 32/133 = 0,241 = 24,1%

CONCLUSIONE: I caratteri "caratteristiche della superficie" e "colore" si segregano indipendentemente l'uno dall'altro (**III legge di Mendel**)

Linfonodi metastatici in 921 pazienti con Ca. gastrico
(rappresentazione grafica di una variabile quantitativa discreta - diagramma a barre)



PESO, STATURA e SESSO delle MATRICOLE di MEDICINA dell'UNIVERSITA' di VERONA nell'A.A. 95/96

PESO Kg	STAT. cm	SESSO	PESO Kg	STAT. cm	SESSO	PESO Kg	STAT. cm	SESSO
56	159	F	77	192	M	51	171	F
66	169	F	60	173	F	48	156	F
50	160	F	78	182	M	55	167	F
53	170	F	52	167	F	60	177	M
54	168	F	47.5	164	F	58	170	F
53	161	F	64	166	F	67	167	F
63	172	M	52	160	F	50	172	F
53	170	F	72	184	M	58	169	F
62	161	F	48	169	F	77	179	M
56	163	F	66	170	M	52	162	M
50	160	F	55	172	F	49	160	F
52	170	F	67	177	M	49	165	F
58	173	F	66	170	M	62	178	M
52	167	F	50	160	F	68	174	M
73	178	M	51	167	F	75	181	M
57	166	F	95	193	M	48	167	F
52	165	F	58	160	F	53	160	F
56	171	F	67	178	F	49	167	F
67	175	M	67	175	M	52	165	F
63	182	F	60	160	F	55	155	F
55	169	F	56	165	F	84	188	M
58	165	F	50	165	F	56	170	F
55	175	M	52	170	F	60	171	F
66	176	M	58	172	F	52	176	M
55	164	F	60	170	F	62	180	F
47	160	F	54	166	F			
47	155	F	60	165	F			
63	169	M	74	172	M			
61	177	F	53	173	F			
53	170	F	72	183	M			
55	168	M	52	168	F			
53	162	F	51	164	F			
62	162	F	81	176	M			
45	160	F	50	160	F			
57	167	F	51	171	F			
45	158	F	64	180	F			
53	168	F	82	183	M			
50	160	F	47	156	F			
55	162	F	70	175	M			
70	177	M	58	168	F			
64	178	F	59	173	F			
52	164	F	68	165	F			
75	175	M	63	177	F			
75	178	M	50	159	F			
70	165	F	65	150	F			
58	167	F	60	170	F			
45	160	F	51	167	F			
50	167	F	75	182	M			
56	156	F	62	170	M			
59	165	F	85	174	M			

Distribuzione di frequenza o Seriazione della variabile quantitativa continua Statura

```
fre var=statura.
--------------------------------------------------
STATURA                                 Valid     Cum
      Value  Frequency  Percent  Percent  Percent
        150          1       .8       .8       .8
        155          2      1.6      1.6      2.4
        156          3      2.4      2.4      4.8
        158          1       .8       .8      5.6
        159          2      1.6      1.6      7.2
        160         13     10.4     10.4     17.6
        161          2      1.6      1.6     19.2
        162          4      3.2      3.2     22.4
        163          1       .8       .8     23.2
        164          4      3.2      3.2     26.4
        165         10      8.0      8.0     34.4
        166          3      2.4      2.4     36.8
        167         11      8.8      8.8     45.6
        168          5      4.0      4.0     49.6
        169          5      4.0      4.0     53.6
        170         12      9.6      9.6     63.2
        171          4      3.2      3.2     66.4
        172          5      4.0      4.0     70.4
        173          4      3.2      3.2     73.6
        174          2      1.6      1.6     75.2
        175          5      4.0      4.0     79.2
        176          3      2.4      2.4     81.6
        177          5      4.0      4.0     85.6
        178          5      4.0      4.0     89.6
        179          1       .8       .8     90.4
        180          2      1.6      1.6     92.0
        181          1       .8       .8     92.8
        182          3      2.4      2.4     95.2
        183          2      1.6      1.6     96.8
        184          1       .8       .8     97.6
        188          1       .8       .8     98.4
        192          1       .8       .8     99.2
        193          1       .8       .8    100.0
                 -------  -------  -------
      Total        125    100.0    100.0
```

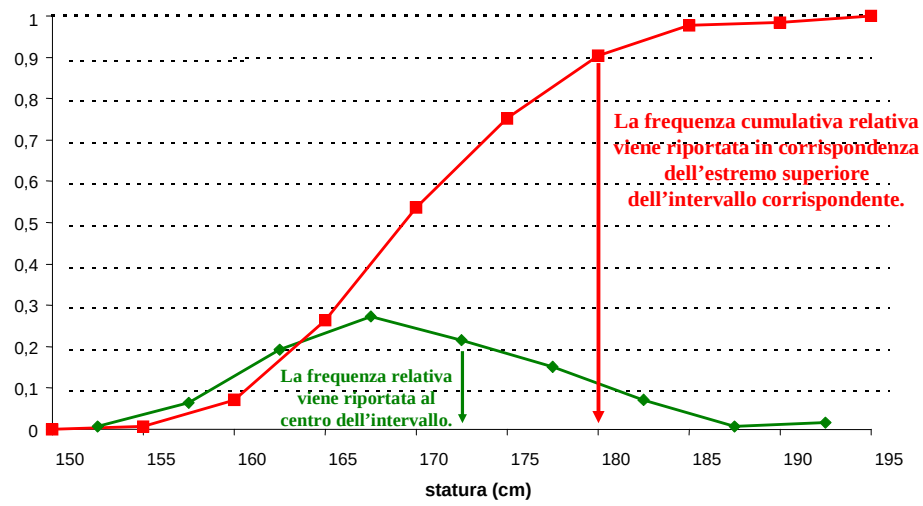
COSTRUZIONE DI UNA DISTRIBUZIONE DI FREQUENZA CON UNA VARIABILE QUANTITATIVA CONTINUA

1) Trovare il valore minimo e il valore massimo	Valore minimo = 150 cm Valore massimo = 193 cm
2) Calcolare il campo di variazione (range) = valore max – valore min	193 cm – 150 cm = 43 cm
3) Stabilire il numero delle classi: tra 5 (pochi individui) e 20 (molti individui)	9 classi
4) Salvo casi particolari, le classi devono avere la stessa ampiezza	
5) Stabilire l'ampiezza dell'intervallo di classe	43 / 9 = 4,78 cm $\approx$ 5 cm
6) Si costruiscono gli intervalli di classe, che devono essere mutuamente esclusivi ed esaustivi	I intervallo: 150,0-154,9 cm II intervallo: 155,0-159,9 cm III intervallo: 160,0-164,9 cm
7) Si conta il numero di individui in ogni classe	I classe: 1 II classe: 8 III classe: 24

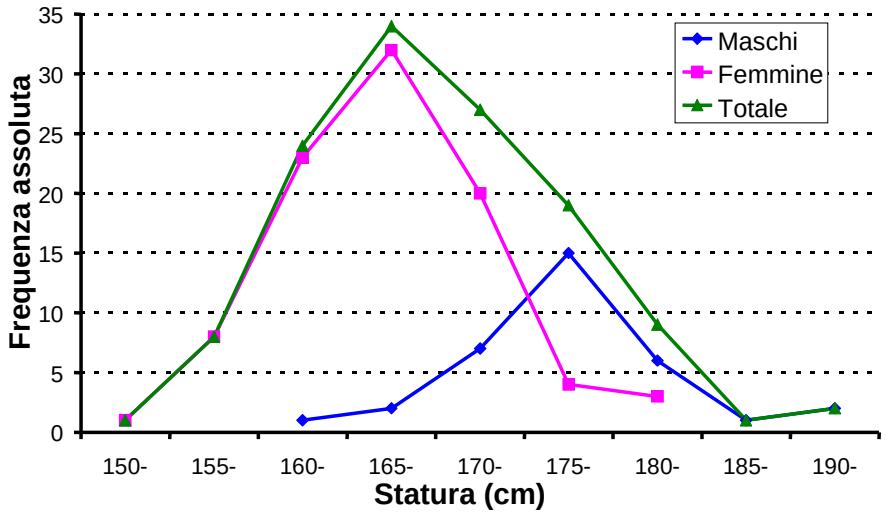
```
compute statCLAS=trunc((statura-145)/5).
fre var=statCLAS.
```

CLASSE	FREQUENZA		FREQUENZA CUMULATIVA	
	ASSOLUTA	RELATIVA %	ASSOLUTA	RELATIVA %
150-154,9	1	1/125= 0,8	1	1/125= 0,8
155-159,9	8	8/125= 6,4	1+8= 9	9/125= 7,2
160-164,9	24	24/125=19,2	1+8+24=33	33/125=26,4
165-169,9	34	34/125=27,2	1+8+24+34=67	67/125=53,6
170-174,9	27	21,6	94	75,2
175-179,9	19	15,2	113	90,4
180-184,9	9	7,2	122	97,6
185-189,9	1	0,8	123	98,4
190-194,9	2	1,6	125	100,0
Totale	125	100,0	125	

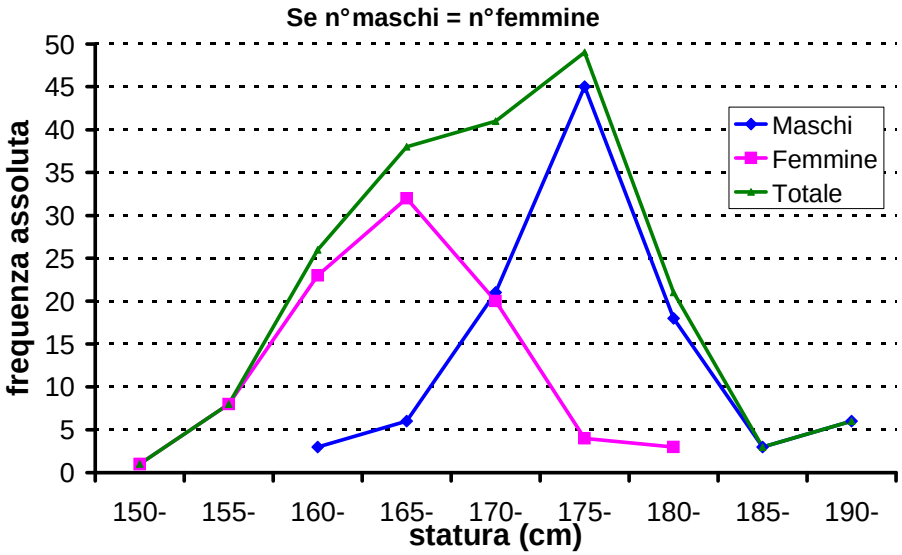
Statura matricole della Facoltà di Medicina (a.a. 95/96) *(rappresentazione grafica - poligoni di frequenza)*



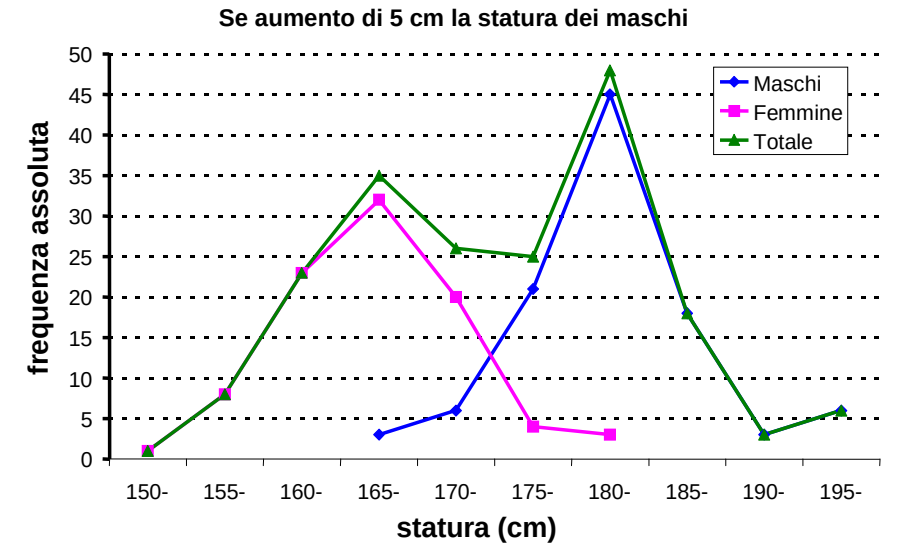
Statura matricole della Facoltà di Medicina (a.a. 95/96), per sesso e per il totale (a)



Statura matricole della Facoltà di Medicina (a.a. 95/96), per sesso e per il totale (b)



Statura matricole della Facoltà di Medicina (a.a. 95/96), per sesso e per il totale (c)



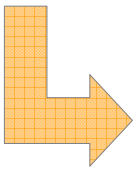
COSTRUZIONE di una DISTRIBUZIONE di FREQUENZA

Costruzione degli intervalli di classe

$$\delta_i = \text{Range} / k$$

δ_i = ampiezza intervallo

k = n° intervalli



range: 160-192 cm

numero degli intervalli di classe = 5

ampiezza degli intervalli = (192-160)/5=32/5=6.4 ÷ 7

Statura (cm)
166
164
170
192
160
174
186
176
160
165
165
173
179
168
168

intervalli di classe

160-166.9 cm

167-173.9 cm

174-180.9 cm

181-187.9 cm

188-194.9 cm

Statura (cm)	n	p	N	P
160-166.9	6	0.40	6	0.40
167-173.9	4	0.26	10	0.67
174-180.9	3	0.20	13	0.86
181-187.9	1	0.07	14	0.93
188-194.9	1	0.07	15	1.00

Costruzione degli intervalli di classe

A) H. Sturges nel 1926, sulla base del numero di osservazioni N, ha indicato il numero ottimale di classi C:

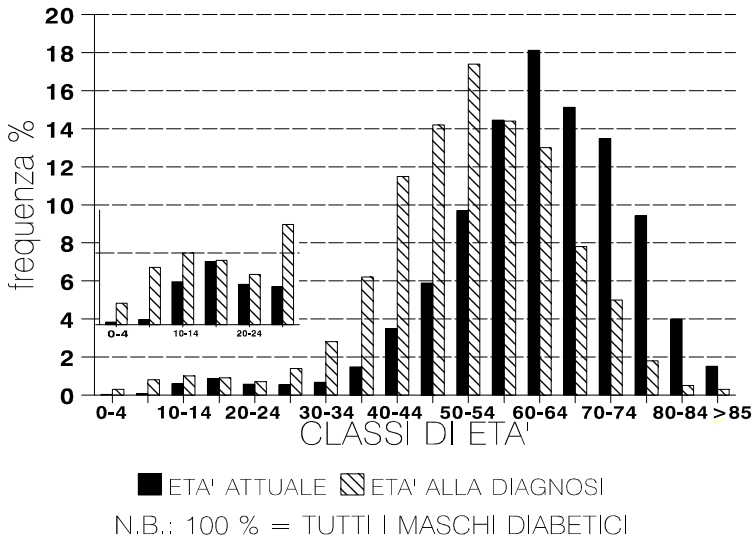
$$C = 1 + \frac{10}{3} \cdot \log_{10}(N)$$

B) D. Scott nel 1979 ha determinato l'ampiezza ottimale h delle classi (dalla quale ovviamente dipende direttamente anche il numero di classi C), mediante la relazione (*dove S = deviazione standard*):

$$h = \frac{3{,}5 \cdot S}{\sqrt{N}}$$

Da: http://www.dsa.unipr.it/soliani/capu1.pdf

MASCHI DIABETICI a VERONA al 31.12.1986



Muggeo M, Verlato G, …, de Marco R (1995) The Verona Diabetes Study: a population-based survey on known diabetes mellitus prevalence and 5-year all-cause mortality. *Diabetologia*, 38: 318-325

Il rango assoluto è la posizione occupata da un'unità statistica in una serie ordinata.

Se due o più individui (unità statistiche) hanno lo stesso valore, si assegna ad esso il rango medio delle posizioni da essi occupati.

RANGO	1	2	3	4	5
NUMERI	3	4	4	5	6
		2,5	2,5		

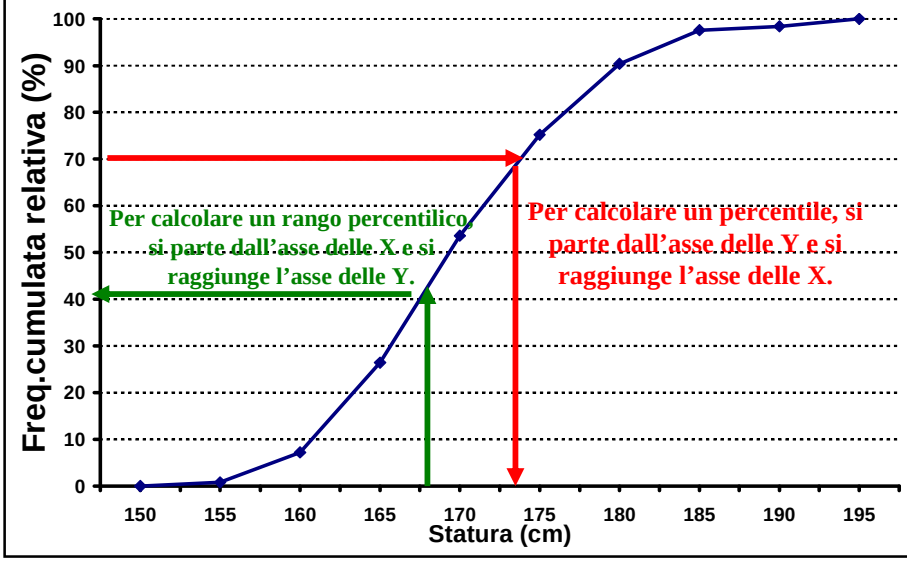
RANGO	1	2	3	4	5
NUMERI	3	4	4	4	5
		3	3	3	

Rango Percentilico

Il **rango percentilico** è la proporzione di punteggi in una distribuzione rispetto alla quale uno specifico valore è superiore o al limite uguale.

Per esempio, se un soggetto ottiene un punteggio pari a 95 in un test […] e tale punteggio è superiore o almeno uguale ai punteggi ottenuti dall'88 % degli studenti che hanno sostenuto il medesimo test, allora il rango percentilico di quel soggetto è 88. Il soggetto rientra dunque nell'88-esimo percentile.

Statura matricole della Facoltà di Medicina (a.a. 95/96)
(rappresentazione grafica - poligoni di frequenza)



Un ragazzo ha la glicemia di 90 mg/dl.

Nella sua scuola ci sono 700 ragazzi.

Se ordiniamo la glicemia in ordine crescente questo ragazzo occupa la posizione 500 (rango assoluto).

Rango assoluto varia in questo caso tra 1 (coma ipoglicemico) e 700 (diabetico di tipo 1 mal controllato).

Qual è il rango percentilico (%)?

500/(700+1) = 500/701 = 0,713 = 71,3 %

RangoAssoluto = (n+1) * *RangoPercentilico*/ 100

RangoPercentilico (%) =100 * RangoAssoluto/(n+1)

Consideriamo un soggetto che ha rango assoluto 50, rispettivamente in un gruppo di 99 soggetti o di 100 soggetti.

	N=99	N=100
Soggetti con rango maggiore	49	49
	50	50
Soggetti con rango minore	49	50
Rango perc =	50/(99+1)=50%	50/(100+1)=49,5%
Calcoli errati =	50/99=50,5%	50/100=50%

RANGO PERCENTILICO = CARATTERISTICA DI UN DETERMINATO INDIVIDUO
PERCENTILE = CARATTERISTICA della POPOLAZIONE

ESEMPIO:
Un individuo pesa 100 Kg. Il suo rango percentilico è 96%. Il 96% degli altri individui pesa meno di lui o come lui e il 4% pesa più di lui.

Nella stessa popolazione qual è il 96esimo percentile? 100 Kg.

L'individuo con rango percentilico 96% si trova esattamente sul 96esimo percentile della popolazione (100 Kg).

Calcolo del *k*-esimo percentile - 1

(Dati individuali disponibili)

- Si individua il rango assoluto corrispondente al ***k*-esimo percentile**

  `Rango Assoluto = (n + 1) * k / 100`

  quindi si riporta il valore dell'osservazione, cui corrisponde quel determinato rango

  Esempio

  la mediana di un campione di 99 individui ha **rango**:

  `(99 + 1) * 50 / 100 = 50`

  il *k*-esimo percentile sarà il valore osservato per la variabile di interesse nell'individuo (più in generale *unità statistica*) con rango 50

Esempio

Qual è il 40esimo percentile della statura nelle matricole di Medicina di Verona nell'anno accademico 1995/96 ?

1) Trovo il rango assoluto corrispondente al ***k*-esimo percentile**

Rango Assoluto = (125 + 1) * 40 / 100 = 126 * 0,4 = 50,4

2) Le osservazioni, con rango assoluto 50 e 51, valgono entrambe 167 cm.

$$X_{40} = 167 \text{ cm}$$

Calcolo del *k*-esimo percentile - 2

(Dati disponibili in classi sotto forma di tabella di frequenza)

- Si individua la classe che contiene il ***k*-esimo percentile**, ovvero la classe in cui la frequenza relativa cumulativa supera o coincide con il *k* per cento
- quindi si procede operando una **interpolazione lineare**

$$x_k = u_{i-1} + \frac{k - F(u_{i-1})}{F(u_i) - F(u_{i-1})} * \delta_i$$

k = rango percentilico
χ_k = k-esimo percentile della distribuzione
u_{i-1} = limite inferiore dello i-esimo intervallo
u_i = limite superiore dello i-esimo intervallo
$F(u_{i-1})$= frequenza cumulativa dell'intervallo precedente
$F(u_i)$ = frequenza cumulativa dell' i-esimo intervallo
δ_i = ampiezza dello i-esimo intervallo

Si assume che all'interno della classe i soggetti siano distribuiti **uniformemente**!

Esempio

Qual è il 40esimo percentile della statura nelle matricole di Medicina di Verona nell'anno accademico 1995/96 ?

Il 40esimo percentile cade nella IV classe (165-169,9 cm)

$$X_{40} = 165 + 5 * \frac{40\% - 26{,}4\%}{53{,}6\% - 26{,}4\%} = 165 + 5 * \frac{13{,}6\%}{27{,}2\%} =$$

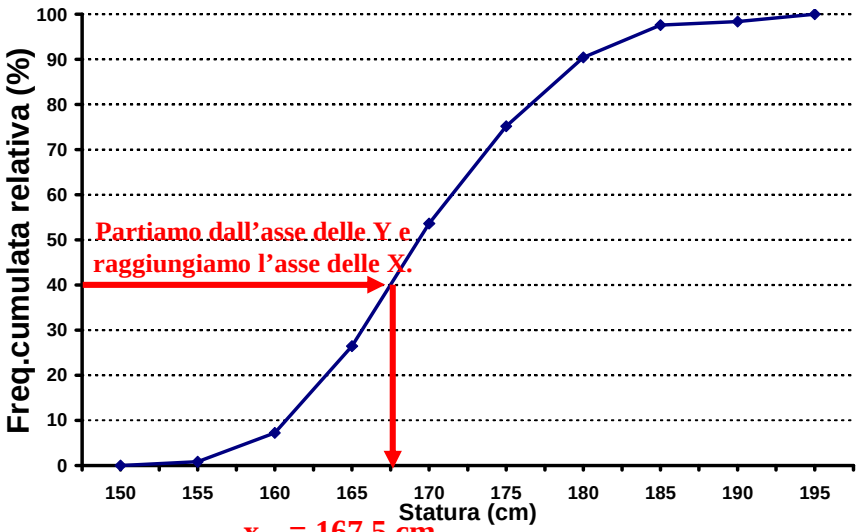
$$= 165 + 5 * 0{,}5 = 165 + 2{,}5 = 167{,}5 \text{ cm}$$

Calcolo del *k*-esimo percentile – 3

(Dati non disponibili, si dispone solamente della rappresentazione grafica della frequenza relativa cumulativa)

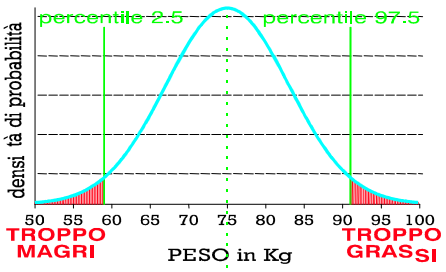
- Sull'asse delle ordinate (*Y*), dove è rappresentata la frequenza relativa cumulata, si individua il punto corrispondente al **rango percentilico (*k*)**
- da qui si traccia una linea orizzontale, che intersechi la linea cosiddetta ***spezzata***, che rappresenta l'andamento della frequenza relativa cumulata
- dal punto d'intersezione così individuato, si traccia una linea verticale fino all'intersezione con l'asse delle ascisse (*X*), che rappresenta i valori della variabile oggetto dello studio
- il valore della variabile in corrispondenza del punto d'intersezione con le *X* rappresenta il ***k*-esimo percentile**

Statura matricole della Facoltà di Medicina (a.a. 95/96)
Troviamo il 40esimo percentile usando il metodo grafico

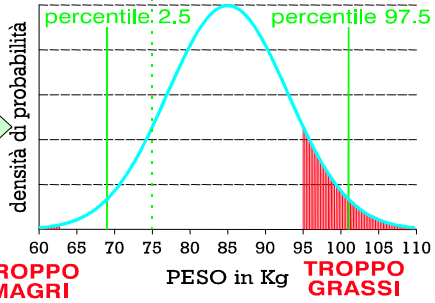


x_{40} = 167,5 cm

NORMALITA' STATISTICA



NORMALITA' CLINICA:
la popolazione è obesa nel suo complesso.
"*The American population is constant in number, but it is ballooning in mass*" (CDC, Atlanta, USA)



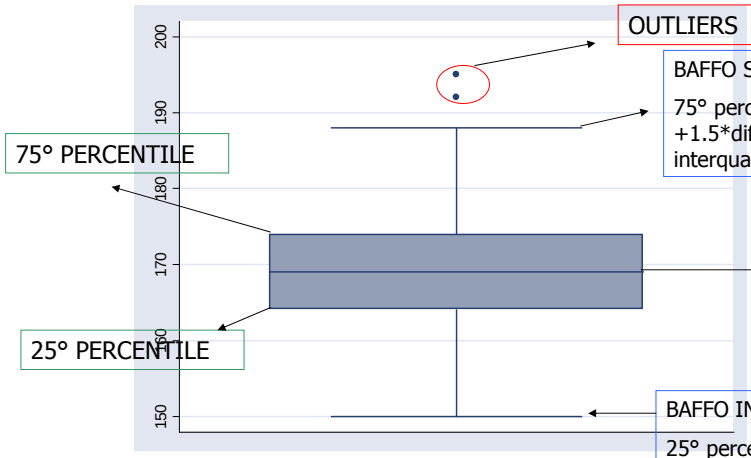
RAPPRESENTAZIONE SINTETICA delle VARIABILI, per via grafica e/o analitica

TIPO VARIABILE	RAPPRESENTAZIONE ANALITICA	RAPPRESENTAZIONE GRAFICA
Categorica	TABELLA di FREQUENZA	Istogramma Torta (*Pie*) Diagramma a Barre
Quantitativa discreta		Diagramma a Barre
Quantitativa continua	Diagramma Stem-and leaf	Istogramma a canne d'organo Poligono di frequenza Box-and-Whiskers plot

```
examine statura/percentiles (2.5 25 50 75 97.5).
------------------------------------------------------------------
          STEM-AND-LEAF DIAGRAM (DIAGRAMMA TRONCO E FOGLIE)
------------------------------------------------------------------
  n  STEM LEAVES                          NUMERI RAPPRESENTATI
------------------------------------------------------------------
  1  15  0                                                  150
  8  15  55666899                    155,155,156,156,156,158,159,159
 24  16  000000000000011222234444
 34  16  55555555566677777777778888899999
 27  17  000000000000111122222333344
 19  17  5555566777777888889
  9  18  001222334
  1  18  8                                                  188
  2  19  23                                               192,193
------------------------------------------------------------------
 Stem width:   10
 Each leaf:     1 case(s)
------------------------------------------------------------------
```

RAPPRESENTAZIONE GRAFICA MEDIANTE BOX-WHISKERS PLOT

(GRAFICO SCATOLA E BAFFI)



OUTLIERS

BAFFO SUPERIORE:

75° percentile +1.5*differenza interquartilica

75° PERCENTILE

MEDIANA

25° PERCENTILE

BAFFO INFERIORE:

25° percentile - 1.5*differenza interquartilica