DESCRIZIONE DEI DATI - PARTE I

La statistica

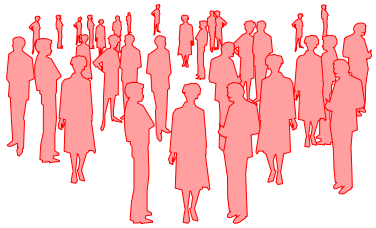
La **statistica** comprende un insieme di metodi per:

- la raccolta
- la descrizione
- l'analisi

di dati relativi a **fenomeni che hanno attitudine a variare**

Statistics is not a bag of tools and math formulas but an evidence-based way of thinking (Frank Harrell)

Se si rilevano in un gruppo di individui



i valori di altezza o il sesso, ad esempio,

1.67, 1.74, 1.94, 1.78 ....

F, F, M, M ....

si può notare che

i valori misurati variano da individuo a individuo

Fonti di variabilità

Strumentale

Biologica

Perché i valori cambiano da individuo a individuo?

Tra le possibili **fonti di variabilità**, quelle più rilevanti sono la:

- ✓ **variabilità strumentale** (legata alla procedura di misurazione, agli strumenti e a chi misura);

  (ad es. la procedura operativa o lo strumento non sono ancora a punto, colui che misura non è sufficientemente esperto)

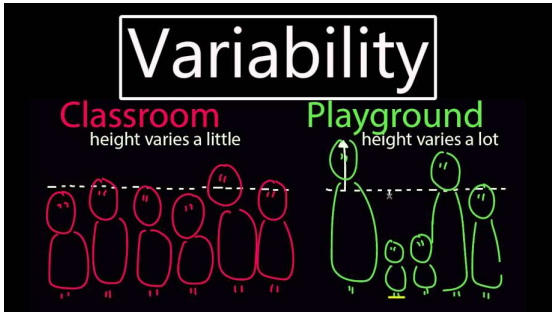
- ✓ **variabilità biologica** (intrinseca).

Perché i valori cambiano da individuo a individuo?

La variabilità strumentale può essere completamente controllata, agendo sulle modalità di misurazione.

(ad es. ottimizzando la procedura operativa, tarando lo strumento, facendo training al personale)

La variabilità biologica può essere solo parzialmente limitata, rendendo più omogeneo l'insieme di soggetti analizzati.



Esempio

Pressione arteriosa del bambino in età scolare

La variabilità strumentale può essere completamente controllata addestrando il personale che effettua le misurazioni

(ad es. scelta del bracciale)

La variabilità biologica può essere parzialmente limitata identificando quei fattori che modificano la pressione arteriosa

(ad es. età, classe ponderale, familiarità)

Terminologia: universo

L'**universo** (o **popolazione**) consiste della **totalità** degli elementi (**unità statistiche**) che hanno certe caratteristiche

Es.:

☞ Studenti che seguono questa lezione di Statistica Medica

Terminologia: campione

Un **campione** è un sottoinsieme di elementi dell' universo che viene utilizzato per trarre conclusioni sulle caratteristiche dell'universo

campione ➡ universo

Il campione non deve essere selezionato ma deve essere scelto in modo casuale

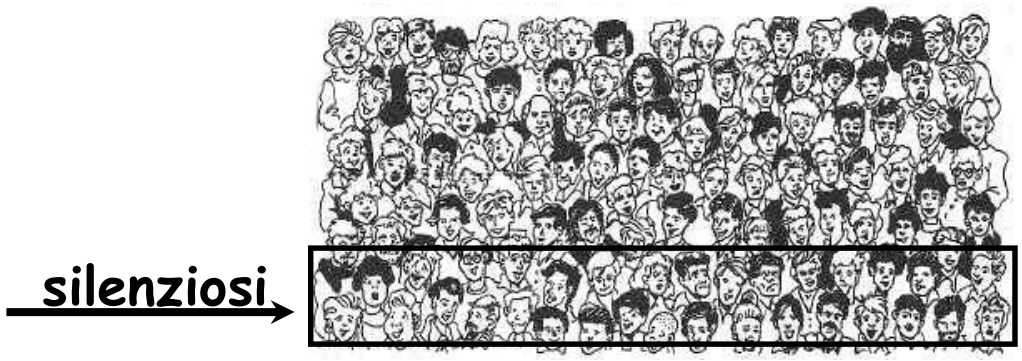
Es: **Universo**: Studenti che seguono questa lezione di Statistica Medica
Campione: 20 studenti presi a caso tra quelli che seguono questa lezione

Terminologia: campione

Un **campione** è un sottoinsieme di elementi dell' universo che viene utilizzato per trarre conclusioni sulle caratteristiche dell'universo

campione ➡ universo

Il campione non deve essere selezionato ma deve essere scelto in modo casuale



Es: **Universo**: Studenti che seguono questa lezione di Statistica Medica
Campione: 20 studenti presi a caso tra quelli che seguono questa lezione

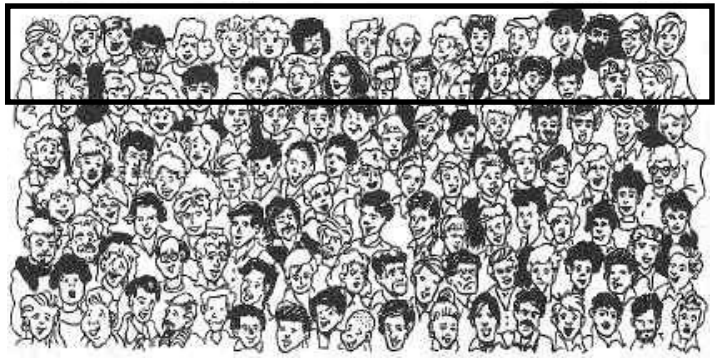
Terminologia: campione

Un **campione** è un sottoinsieme di elementi dell' universo che viene utilizzato per trarre conclusioni sulle caratteristiche dell'universo

campione ➡ universo

Il campione non deve essere selezionato ma deve essere scelto in modo casuale

chiaccheroni



Es: **Universo**: Studenti che seguono questa lezione di Statistica Medica
Campione: 20 studenti presi a caso tra quelli che seguono questa lezione

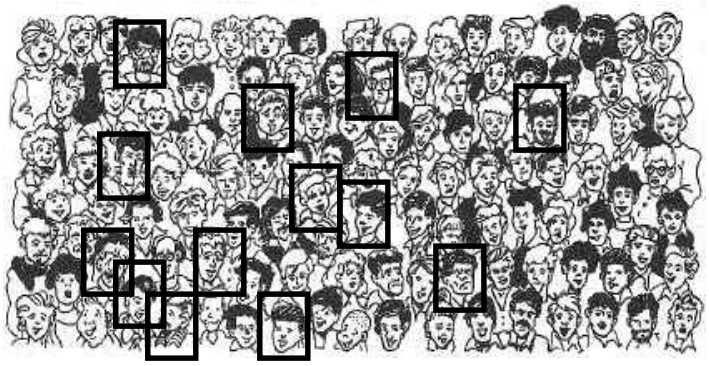
Terminologia: campione

Un **campione** è un sottoinsieme di elementi dell' universo che viene utilizzato per trarre conclusioni sulle caratteristiche dell'universo

campione ➡ universo

Il campione non deve essere selezionato ma deve essere scelto in modo casuale

campione casuale →



Es: **Universo**: Studenti che seguono questa lezione di Statistica Medica
Campione: 20 studenti presi a caso tra quelli che seguono questa lezione

Il campione

Il campione casuale dovrebbe rappresentare una immagine in scala ridotta dell'universo.

campione come **miniatura** dell'universo

… ovvero dovrebbe essere **rappresentativo** dell'universo.

Questa è la condizione (non verificabile) di validità del processo di **generalizzazione dei risultati.**

Il campione

N=100

1) n=5

2) n=25

3) n=75

Quale tra questi tre campioni contiene più informazioni sulla popolazione?

Remdesivir in adults with severe COVID-19: a randomised, double-blind, placebo-controlled, multicentre trial

Yeming Wang, Dingyu Zhang*, Guanhua Du*, Ronghui Du*, Jianping Zhao*, Yang Jin*, Shouzhi Fu*, Ling Gao*, Zhenshun Cheng*, Qiaofa Lu*, Yi Hu*, Guangwei Luo*, Ke Wang, Yang Lu, Huadong Li, Shuzhen Wang, Shunan Ruan, Chengqing Yang, Chunlin Mei, Yi Wang, Dan Ding, Feng Wu, Xin Tang, Xianzhi Ye, Yingchun Ye, Bing Liu, Jie Yang, Wen Yin, Aili Wang, Guohui Fan, Fei Zhou, Zhibo Liu, Xiaoying Gu, Jiuyang Xu, Lianhan Shang, Yi Zhang, Lianjun Cao, Tingting Guo, Yan Wan, Hong Qin, Yushen Jiang, Thomas Jaki, Frederick G Hayden, Peter W Horby, Bin Cao, Chen Wang*

Summary

Background No specific antiviral drug has been proven effective for treatment of patients with severe coronavirus disease 2019 (COVID-19). Remdesivir (GS-5734), a nucleoside analogue prodrug, has inhibitory effects on pathogenic animal and human coronaviruses, including severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) in vitro, and inhibits Middle East respiratory syndrome coronavirus, SARS-CoV-1, and SARS-CoV-2 replication in animal models.

Methods We did a randomised, double-blind, placebo-controlled, multicentre trial at ten hospitals in Hubei, China. Eligible patients were adults (aged ≥18 years) admitted to hospital with laboratory-confirmed SARS-CoV-2 infection, with an interval from symptom onset to enrolment of 12 days or less, oxygen saturation of 94% or less on room air or a ratio of arterial oxygen partial pressure to fractional inspired oxygen of 300 mm Hg or less, and radiologically confirmed pneumonia. Patients were randomly assigned in a 2:1 ratio to intravenous remdesivir (200 mg on day 1 followed by 100 mg on days 2–10 in single daily infusions) or the same volume of placebo infusions for 10 days. Patients were permitted concomitant use of lopinavir–ritonavir, interferons, and corticosteroids. The primary endpoint was time to clinical improvement up to day 28, defined as the time (in days) from randomisation to the point of a decline of two levels on a six-point ordinal scale of clinical status (from 1=discharged to 6=death) or discharged alive from hospital, whichever came first. Primary analysis was done in the intention-to-treat (ITT) population and safety analysis was done in all patients who started their assigned treatment. This trial is registered with ClinicalTrials.gov, NCT04257656.

Findings Between Feb 6, 2020, and March 12, 2020, 237 patients were enrolled and randomly assigned to a treatment group (158 to remdesivir and 79 to placebo); one patient in the placebo group who withdrew after randomisation was not included in the ITT population. Remdesivir use was not associated with a difference in time to clinical improvement (hazard ratio 1·23 [95% CI 0·87–1·75]). Although not statistically significant, patients receiving remdesivir had a numerically faster time to clinical improvement than those receiving placebo among patients with symptom duration of 10 days or less (hazard ratio 1·52 [0·95–2·43]). Adverse events were reported in 102 (66%) of 155 remdesivir recipients versus 50 (64%) of 78 placebo recipients. Remdesivir was stopped early because of adverse events in 18 (12%) patients versus four (5%) patients who stopped placebo early.

Interpretation In this study of adult patients admitted to hospital for severe COVID-19, remdesivir was not associated with statistically significant clinical benefits. However, the numerical reduction in time to clinical improvement in those treated earlier requires confirmation in larger studies.

Lancet 2020; 395: 1569–78

Published Online
April 29, 2020
https://doi.org/10.1016/S0140-6736(20)31022-9

This online publication has been corrected. The corrected version first appeared at thelancet.com on May 28, 2020

See Comment page 1525

*Contributed equally

Department of Pulmonary and Critical Care Medicine, Center of Respiratory Medicine, National Clinical Research Center for Respiratory Diseases (Ye Wang MD, F Zhou MD, Z Liu MD, L Shang MD, Y Zhang MD, Prof B Cao MD, Prof C Wang MD) **and Institute of Clinical Medical Sciences** (G Fan MS, X Gu PhD), **China-Japan Friendship Hospital, Beijing, China; Department of Respiratory Medicine, Capital Medical University, Beijing, China** (Ye Wang, Prof B Cao); **Jin Yin-tan Hospital, Wuhan, Hubei, China** (D Zhang MD, H Li MD, S Wang MS, S Ruan MS); **Institute of Materia Medica, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China** (Prof G Du PhD, Prof K Wang PhD, Prof Y Lu PhD); **Wuhan Lung Hospital, Wuhan, China** (Prof R Du MD, C Yang MD,

Terminologia: variabili

Si dice **variabile** una **caratteristica** delle unità statistiche che può assumere una pluralità di valori al variare dell'unità su cui è rilevata

Es: Altezza, Sesso, Titolo di Studio, peso alla nascita

Le variabili possono essere:

i) quantitative

ii) qualitative

vengono indicate con lettere maiuscole scelte, in genere, tra le ultime lettere dell'alfabeto: Y, X, Z

Es: Y = Altezza X = Sesso

Terminologia: dati

I **dati** sono quei **valori numerici** o quelle **modalità**, assumibili da una variabile.

I dati sono rappresentati da lettere minuscole con un indice che distingue le diverse unità fra loro:

Es: Y=Altezza	y_1= 1.67	y_2= 1.74	y_3= 1.94	y_4= 1.78
X=Sesso	x_1= F	x_2= F	x_3=M	x_4=M

Variabili quantitative discrete

Una variabile quantitativa è **discreta** se può assumere come valore un **qualsiasi numero naturale**

Es.: - Numero automobili per famiglia
- Voto esame di statistica
- Durata dell'allattamento (in mesi)

Le variabili quantitative discrete derivano usualmente da **conteggi**

Variabili quantitative continue

Una variabile quantitativa è **continua**, se può assumere come valore un **qualsiasi numero reale**

Es.: Altezza, Peso, Concentrazione di glucosio nel sangue

I valori assunti da una variabile continua dipendono in realtà dal potere di risoluzione dello strumento di misura

Es.: Una altezza di 1.78324321…. m, potrebbe essere riportata al cm (1.78) o al mm (1.783) a seconda dell'uso

Le variabili quantitative continue derivano usualmente da **misurazioni**

Variabili qualitative nominali

Una variabile qualitativa è **nominale**, quando ogni possibile ordinamento delle modalità è arbitrario

Es: Sesso, Colore degli occhi, tipologia di parto

Etnia pazienti coinvolti in una sperimentazione clinica

caucasico - afroamericano - africano - indiano .. etc.

=

afroamericano - indiano - caucasico - africano .. etc.

Variabili qualitative ordinali

Una variabile qualitativa è **ordinale**, quando è possibile individuare un ordinamento naturale delle modalità.

Es.: Segno zodiacale, Titolo di studio

Misurazione dell'intensità del dolore

nulla < lieve < moderata < forte
forte > moderata > lieve > nulla

moderata - forte - nulla - lieve

NO!!

Variabili qualitative ordinali

Es: Misurazione dell'intensità del dolore

nulla < lieve < moderata < forte

Alle modalità si può associare un **codice numerico**:
(***Es.***: nulla=0, lieve=1, moderata=2, forte=3)
che però non ha significato quantitativo:

- 2 (dolore moderato) **non** è il doppio di 1 (dolore lieve),
  3 (dolore forte) **non** è il triplo di 1
- la differenza tra 2 e 1 **non** è uguale a quella tra 3 e 2

Variabili qualitative a due livelli

Vengono chiamate anche dicotomiche (o binarie), segnalano la presenza (o l'assenza) di una caratteristica.

Es: Presenza di gravi complicazioni dopo un intervento chirurgico. Le uniche modalita' che questa variabile puo' assumere sono 'Sì' , 'No'

Dalla frequenza di 'Sì' si ottiene la frequenza di 'No' calcolando 1 – la frequenza di 'Sì'

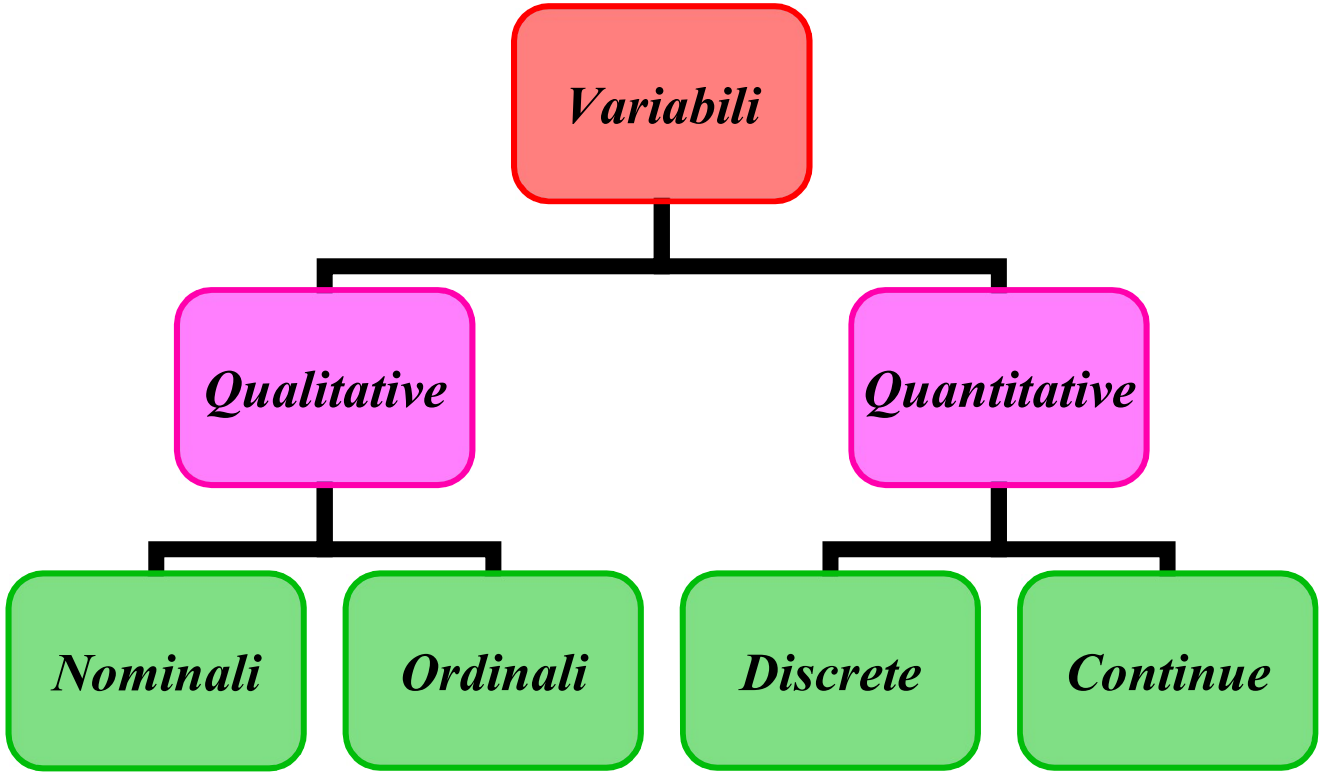
Es: 0.12 (12%) di 'Sì' implica 0.88 (88%) di 'No'

La frequenza di 'Sì' si chiama PROPORZIONE

	Remdesivir group (n=158)	Placebo group (n=78)
Age, years	66·0 (57·0–73·0)	64·0 (53·0–70·0)
Sex		
Men	89 (56%)	51 (65%)
Women	69 (44%)	27 (35%)
Any comorbidities	112 (71%)	55 (71%)
Hypertension	72 (46%)	30 (38%)
Diabetes	40 (25%)	16 (21%)
Coronary heart disease	15 (9%)	2 (3%)
Body temperature, °C	36·8 (36·5–37·2)	36·8 (36·5–37·2)
Fever	56 (35%)	31 (40%)
Respiratory rate >24 breaths per min	36 (23%)	11 (14%)
White blood cell count, × 10^9 per L		
Median	6·2 (4·4–8·3)	6·4 (4·5–8·3)
4–10	108/155 (70%)	58 (74%)
<4	27/155 (17%)	12 (15%)
>10	20/155 (13%)	8 (10%)
Lymphocyte count, × 10^9 per L	0·8 (0·6–1·1)	0·7 (0·6–1·2)
≥1·0	49/155 (32%)	23 (29%)
<1·0	106/155 (68%)	55 (71%)
Platelet count, × 10^9 per L	183·0 (144·0–235·0)	194·5 (141·0–266·0)
≥100	148/155 (95%)	75 (96%)
<100	7/155 (5%)	3 (4%)
Serum creatinine, µmol/L	68·0 (56·0–82·0)	71·3 (56·0–88·7)
≤133	151/154 (98%)	76 (97%)
>133	3/154 (2%)	2 (3%)
Aspartate aminotransferase, U/L	31·0 (22·0–44·0)	33·0 (24·0–48·0)
≤40	109/155 (70%)	49 (63%)
>40	46/155 (30%)	29 (37%)
Alanine aminotransferase, U/L	26·0 (18·0–42·0)	26·0 (20·0–43·0)
≤50	130/155 (84%)	66 (85%)
>50	25/155 (16%)	12 (15%)
Lactate dehydrogenase, U/L	339·0 (247·0–441·5)	329·0 (249·0–411·0)
≤245	36/148 (24%)	17/75 (23%)
>245	112/148 (76%)	58/75 (77%)
Creatine kinase, U/L	75·9 (47·0–131·1)	75·0 (47·0–158·0)
≤185	118/141 (84%)	54/67 (81%)
>185	23/141 (16%)	13/67 (19%)
National Early Warning Score 2 level at day 1	5·0 (3·0–7·0)	4·0 (3·0–6·0)
Six-category scale at day 1		
2—hospital admission, not requiring supplemental oxygen	0	3 (4%)
3—hospital admission, requiring supplemental oxygen	129 (82%)	65 (83%)
4—hospital admission, requiring high-flow nasal cannula or non-invasive mechanical ventilation	28 (18%)	9 (12%)
5—hospital admission, requiring extracorporeal membrane oxygenation or invasive mechanical ventilation	0	1 (1%)

(Table 1 continues on next page)

Per riassumere

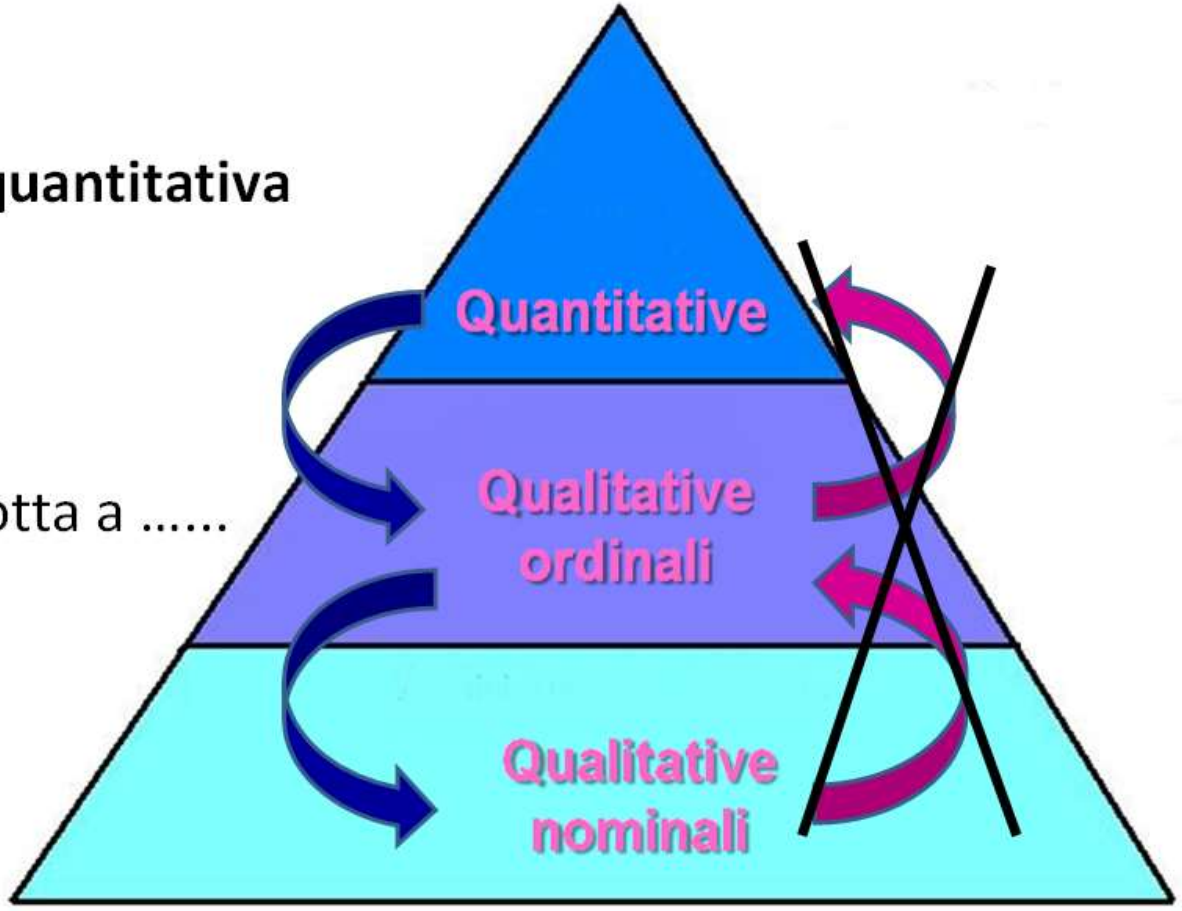


GERARCHIA DELLE VARIABILI

Una **variabile quantitativa**

può essere ridotta a .....

e ancora a .....



Esempi

Numero di carie presenti nell'arcata superiore	Quantitativa Discreta
Stato civile	Qualitativa Nominale
Consumo giornaliero di caffeina (mg)	Quantitativa Continua
Consumo giornaliero di caffè della macchinetta (bicchierini)	Quantitativa Discreta
Albumina sierica (g/l)	Quantitativa Continua
Tipologia Epatite	Qualitativa Nominale
Numero di linfonodi metastatici riscontrati alla TAC	Quantitativa Discreta

Cosa possiamo concludere dai dati relativi al campione di 20 studenti?

Soggetto	Altezza	Sesso	Soggetto	Altezza	Sesso
1	1.76	M	11	1.77	F
2	1.71	F	12	1.69	F
3	1.54	F	13	1.93	M
4	1.82	M	14	1.67	F
5	1.59	F	15	1.72	M
6	1.74	M	16	1.59	F
7	1.95	M	17	1.60	F
8	1.68	M	18	1.81	F
9	1.85	M	19	1.73	F
10	1.74	F	20	1.78	M

Distribuzioni di frequenza

Per riassumere i dati si costruiscono le **distribuzioni di frequenza**

possibili **valori** (modalità) che una variabile può assumere

e

frequenze con cui questi valori si manifestano

Distribuzioni di frequenza

I dati di un'unità per la donazione di sangue mostrano che il numero totale di donatori rispetto ai quattro gruppi sanguigni ammonta a: A 725; B 258; AB 72; e O 1073.

Gruppo sanguigno	f
A	725
B	258
AB	72
0	1073
Totale	n=2128

f = frequenza assoluta

numero di volte in cui una certa modalità si manifesta nel campione

258 dei 2128 donatori hanno gruppo sanguigno B

Distribuzioni di frequenza

p = frequenza relativa ← rapporto tra la frequenza assoluta con cui si manifesta una modalità e la numerosità totale del campione

Gruppo sanguigno	f	f/n	p	p%
A	725	725/2128	0.341	34.1
B	258	258/2128	0.121	12.1
AB	72	72/2128	0.034	3.4
0	1073	1073/2128	0.504	50.4
Totale	n=2128		1.000	100

Il 12% dei donatori ha gruppo sanguigno B

Frequenze assolute e relative

- **frequenze assolute f**
  - ✓ possono assumere valori compresi tra 0 e n (dimensione del campione)
  - ✓ la loro somma è pari a n

- **frequenze relative p**
  - ✓ possono assumere valori compresi tra 0 e 1
  - ✓ la loro somma è pari a 1

- **frequenze relative p%**
  - ✓ possono assumere valori compresi tra 0% e 100%
  - ✓ la loro somma è pari a 100%

Frequenze assolute e relative

Frequenze assolute e relative forniscono le stesse informazioni sulla distribuzione

Tuttavia, le frequenze relative:

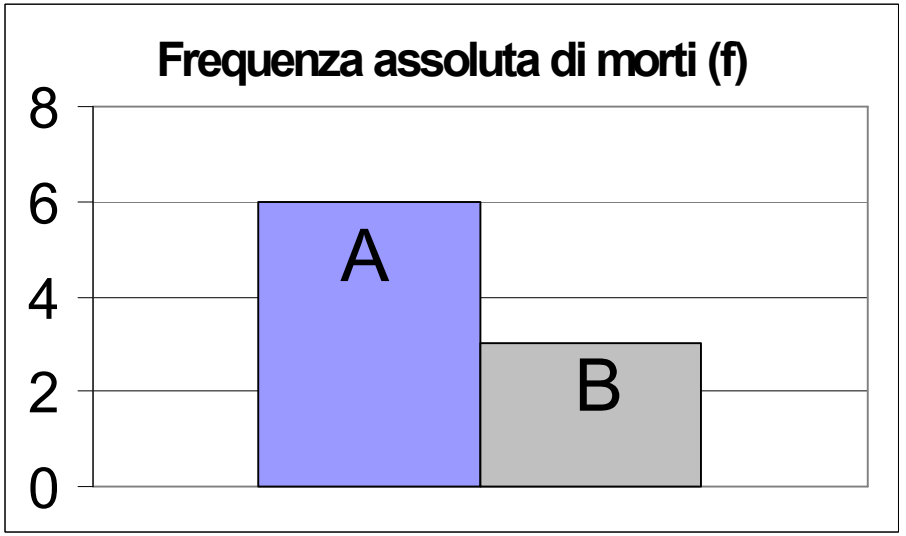
- ✓ facilitano la percezione del peso delle modalità;
- ✓ consentono di confrontare la distribuzione di una variabile in campioni di diversa numerosità.

Andrebbero sempre accompagnate dalla numerosità su cui sono state calcolate!

Esempio

Si vuole valutare l'efficacia di un nuovo farmaco (A) sulla mortalità post-infarto (1 mese). Nello studio vengono coinvolti 150 pazienti: 100 sono randomizzati a ricevere il farmaco sperimentale, 50 il trattamento standard (B).

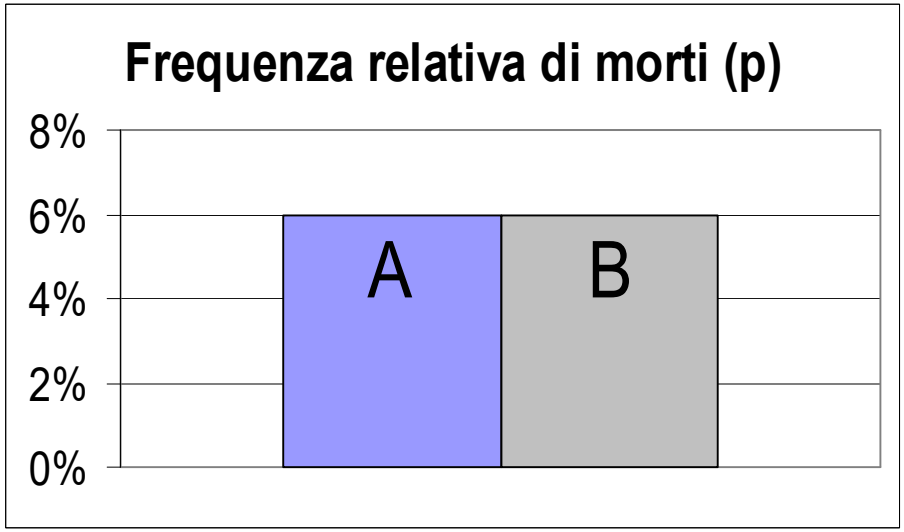
	Trattati con	
	A	B
Morti	6	3
Vivi	94	47
Totale	100	50



Esempio

Si vuole valutare l'efficacia di un nuovo farmaco (A) sulla mortalità post-infarto (1 mese). Nello studio vengono coinvolti 150 pazienti: 100 sono randomizzati a ricevere il farmaco sperimentale, 50 il trattamento standard (B).

	Trattati con	
	A	**B**
Morti	6(**6%**)	3(**6%**)
Vivi	94(**94%**)	47(**94%**)
Totale	100	50



Attenzione alle informazioni fuorvianti!

"The antibiotic phosphomycin is advertised as being 100% effective in chronic urinary tract infections."

L'antibiotico fosfomicin è efficace al 100% nelle infezioni urinarie croniche.

Lo studio su cui si basa questa informazione ha coinvolto 8 pazienti, dopo aver eliminato i pazienti le cui urine contentevano batteri fosfomicina-resistenti.

Grafici per var. qualitative

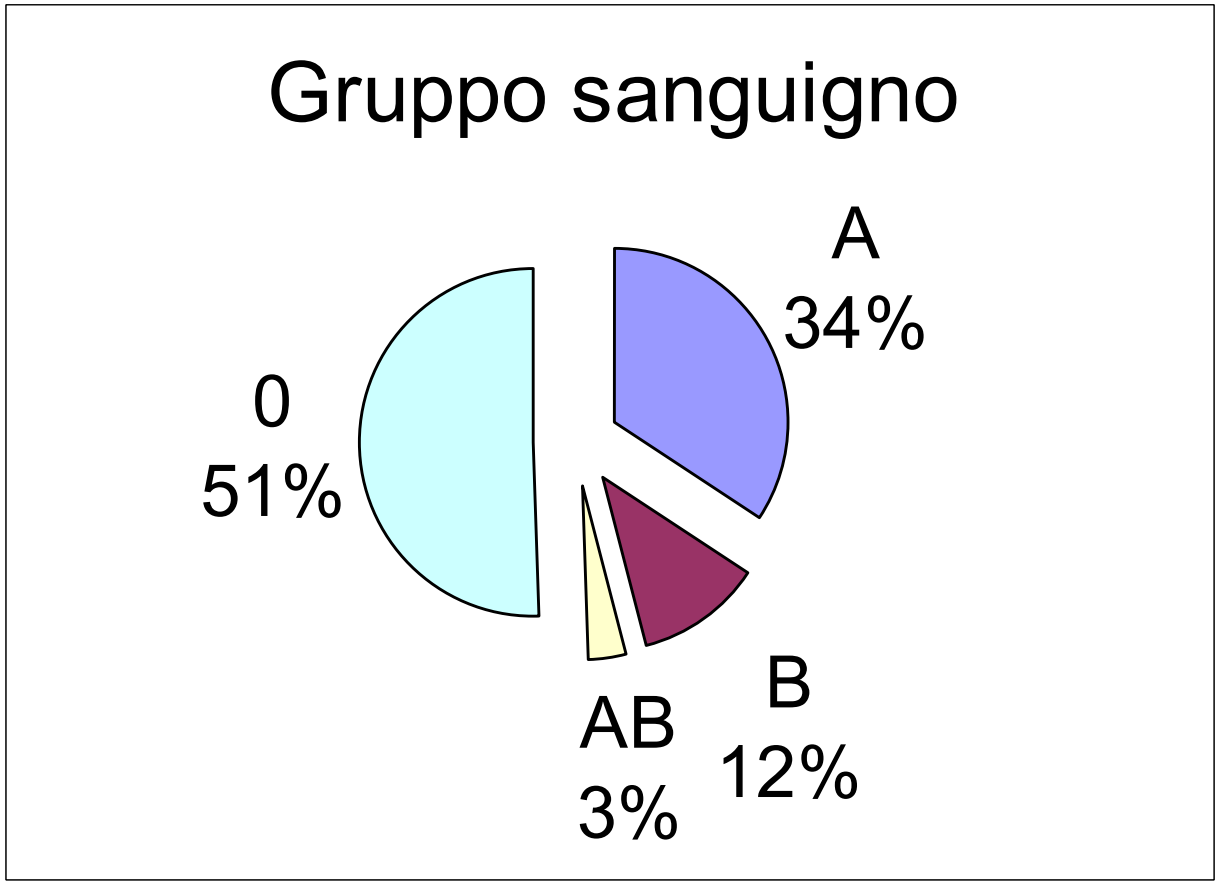


Diagramma areolare (o *a torta*)

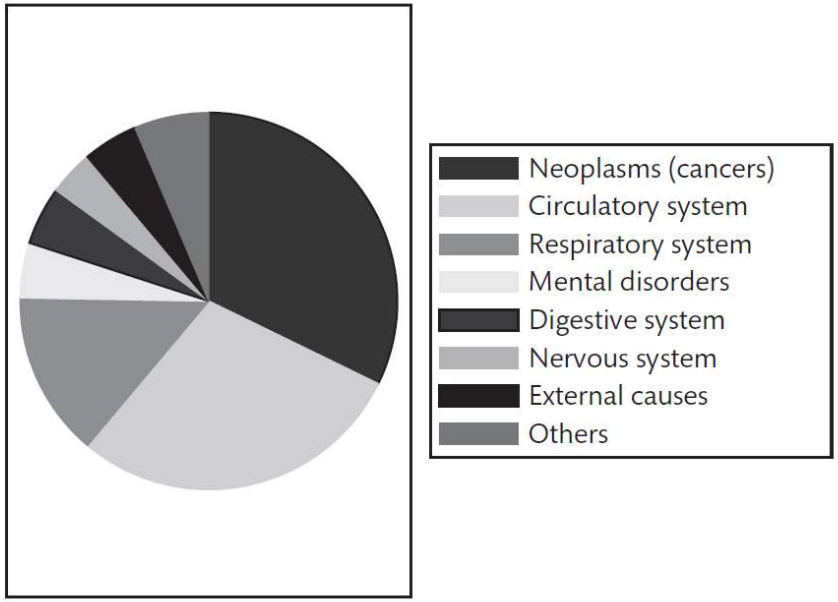


Figure 5.1 Pie chart showing the distribution of cause of death among males, England and Wales, 2012 (data from the Office for National Statistics).

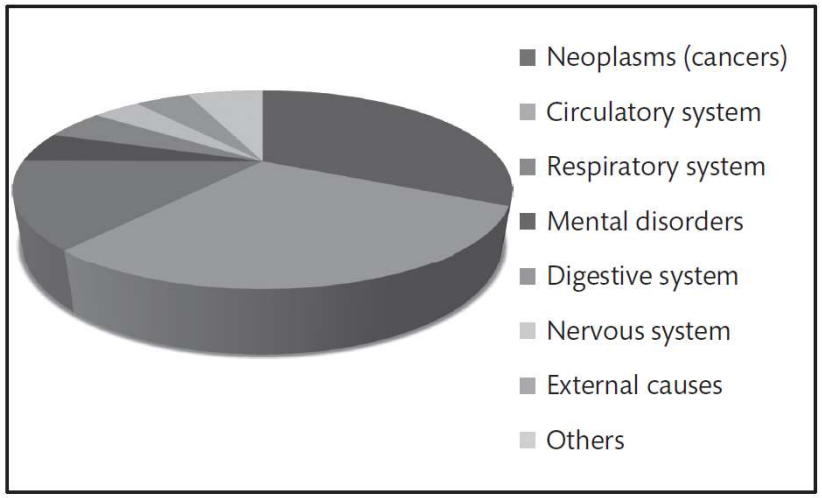


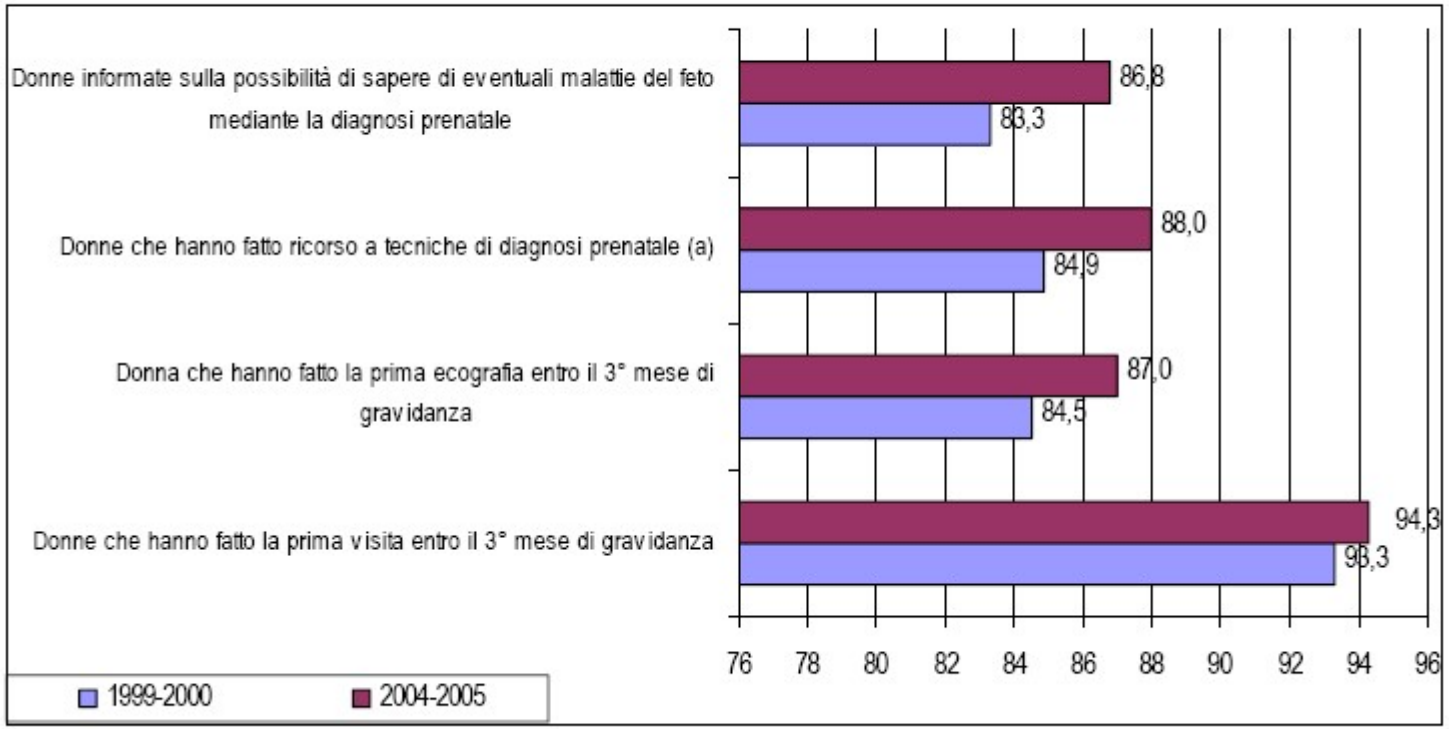
Figure 5.15 Figure 5.1 with three-dimensional effects (data from the Office for National Statistics).

Table 5.4 Calculations for a pie chart of the distribution of cause of death, 2012, men (data from the Office for National Statistics)

Cause of death	Frequency	Relative frequency	Angle (degrees)
Neoplasms (cancers)	76 695	0.319 25	115
Circulatory system	69 516	0.289 36	104
Respiratory system	33 463	0.139 29	50
Mental disorders	11 710	0.048 74	18
Digestive system	11 766	0.048 98	18
Nervous system	9 499	0.039 54	14
External causes	10 993	0.045 76	16
Others	15 400	0.064 10	23
Total	294 227	1.000 00	358

Esempio: Assistenza in gravidanza

Grafico 1 Principali indicatori di assistenza in gravidanza. Confronto 2004-2005 (dati provvisori) con 1999-2000 (per 100 donne con le stesse caratteristiche)



(a) Le tecniche di diagnosi prenatale rilevate sono dosaggio alfa feto proteina, prelievo villi coriali, amniocentesi, ecografia morfologico fetale, tri-test.

Istituto nazionale di statistica

Gravidanza, parto, allattamento al seno

2004 - 2005

Diagramma a barre orizzontali

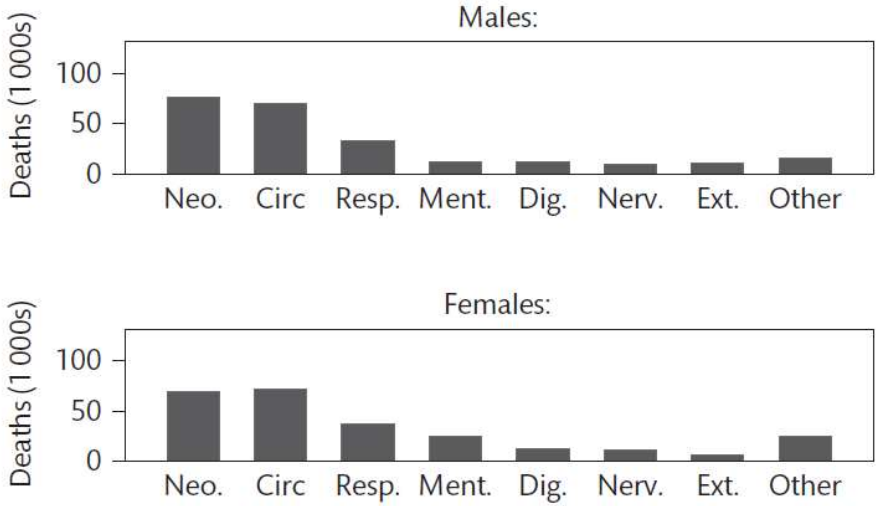


Figure 5.5 Paired bar charts showing data for main causes of death from Table 5.1 (data from the Office for National Statistics).

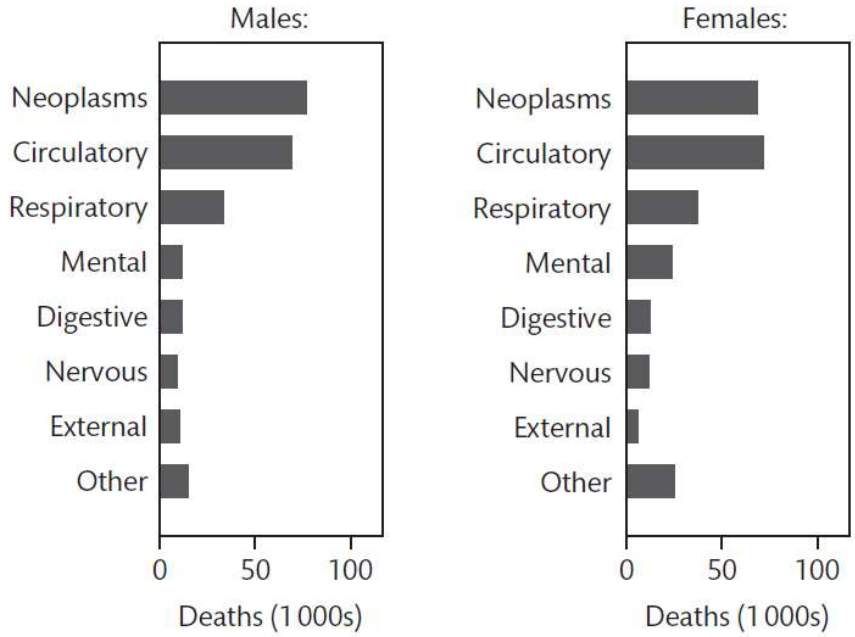


Figure 5.6 Horizontal bar charts showing data for main causes of death from Table 5.1 (data from the Office for National Statistics).

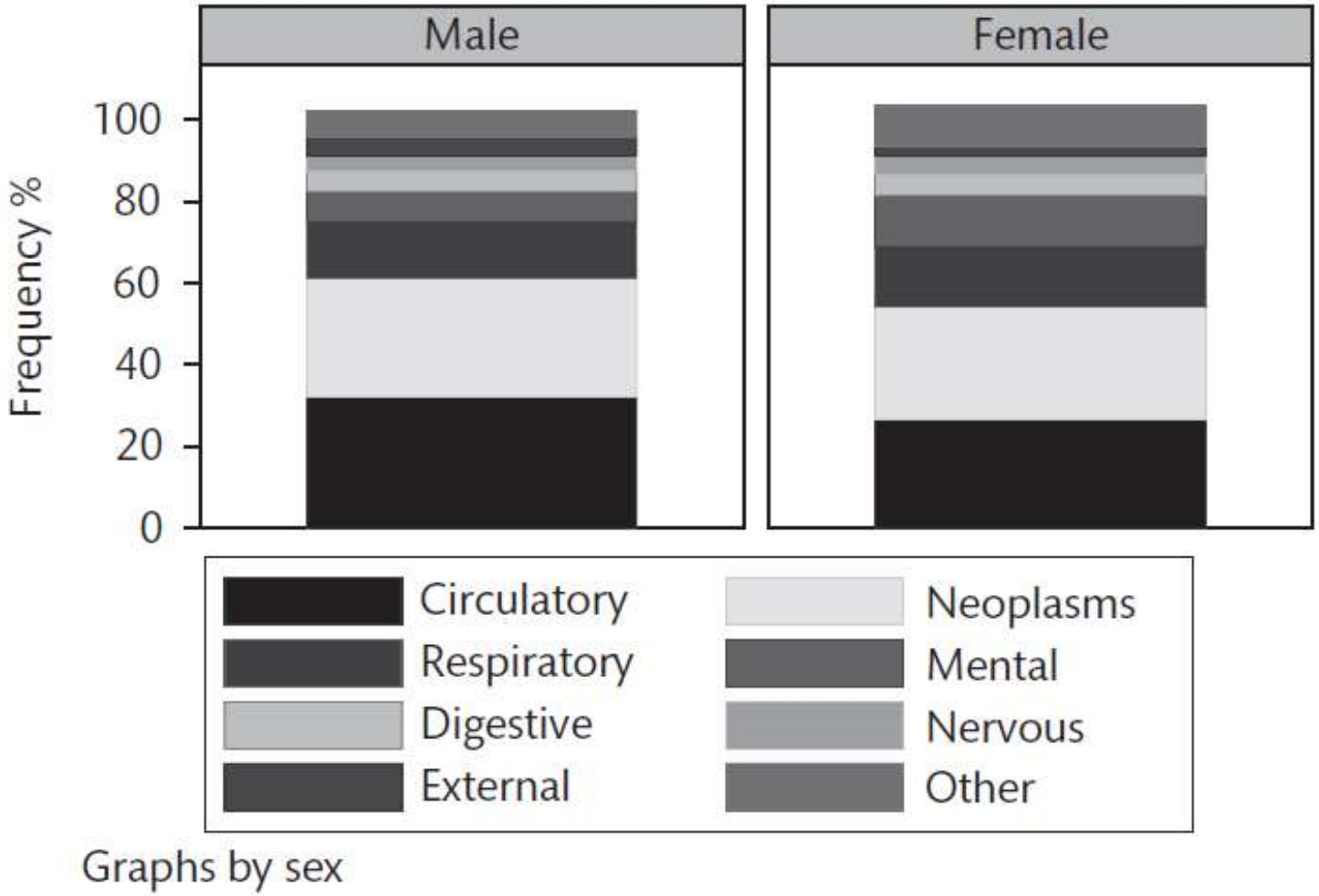


Figure 5.4 Stacked bar chart showing data for main causes of death from Table 5.1 (data from the Office for National Statistics).

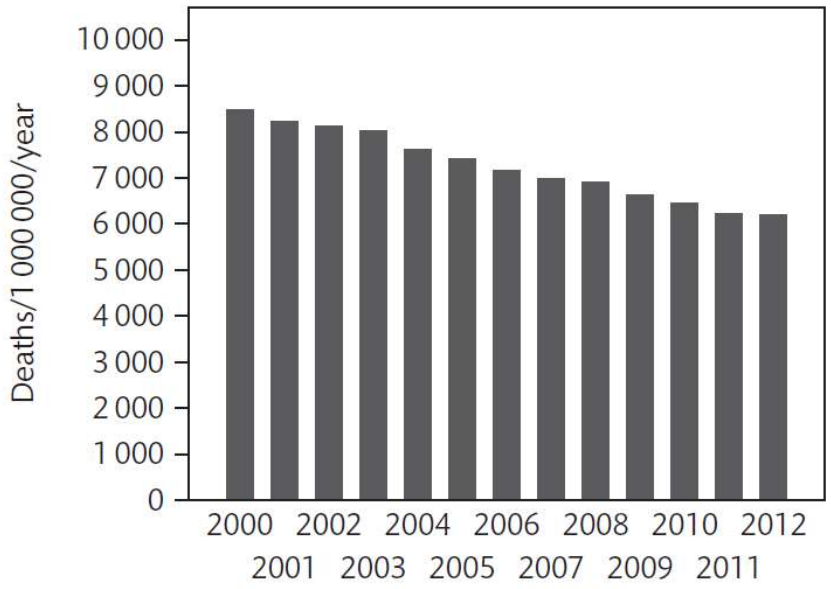


Figure 5.2 Bar chart showing the relationship between mortality caused by cancer and year, males, England and Wales, 2000–2012 (data from the Office for National Statistics).

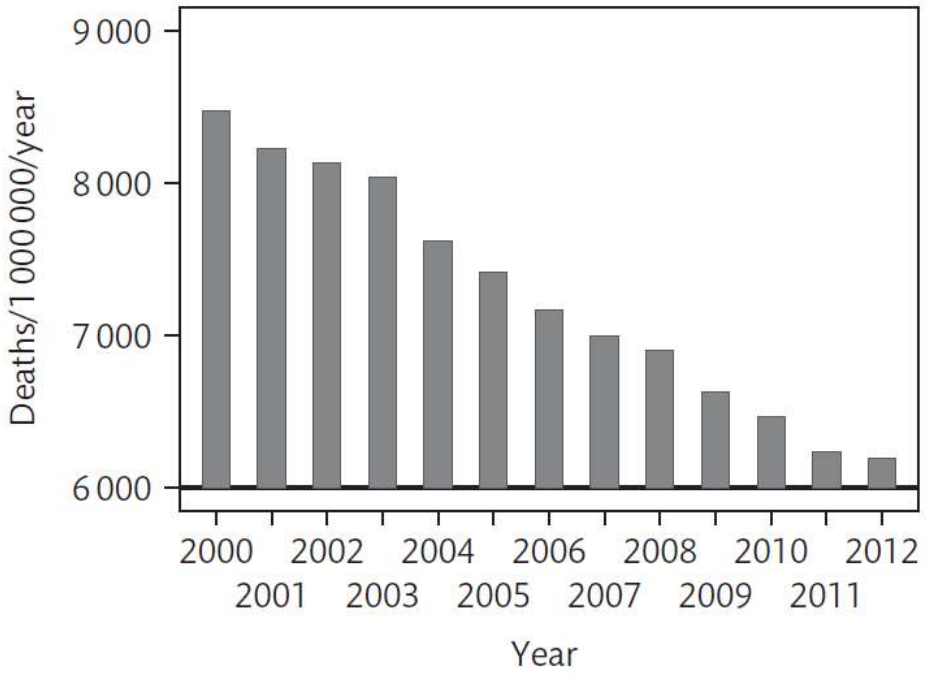


Figure 5.12 Bar chart with zero omitted on the vertical scale.

graphicsprinciples.github.io

graphics principles home three laws cheat sheet tutorial case studies QBV blog resources & references about

Welcome

This is the home page for effective visual communication and good graphical principles for quantitative scientists.

Effective visual communication is a core skill for all quantitative scientists including statisticians, epidemiologists, machine learning experts, bioinformaticians, etc. By using the right graphical principles, we can better understand data, highlight core insights and influence decisions toward appropriate actions. Without it, we can fool ourselves and others and pave the way to wrong conclusions and actions.

The goal of these pages is to help quantitative scientists to get this right. More specifically, you will find:

- The three laws of effective visual communication
- A graphics principles cheat sheet
- A tutorial covering both of the above more comprehensively
- Example case studies from this tutorial with programming code for download
- The concept of Question-Based Visualizations (QBV)
- A blog with various related articles
- Some further background of this initiative, including a related overview paper
- Links to further resources & references

We hope that these page prove beneficial for your work. They will evolve further over time.

https://graphicsprinciples.github.io/threelaws.html

Variabili quantitative discrete

Successione delle **frequenze** che corrispondono ai **valori** assunti da una **variabile quantitativa discreta.**

Numero di morti causate da incidenti stradali rilevate da 14 reparti di emergenza in una regione durante un week-end.

X	frequenze semplici		frequenze cumulate	
	assolute f	relative p	assolute F	relative P
0	7	0.500	7	0.500
1	3	0.214	10	0.714
2	2	0.143	12	0.857
3	1	0.071	13	0.929
4	1	0.071	14	1.000

Frequenze cumulate

X	frequenze semplici		frequenze cumulate	
	assolute f	relative p	assolute F	relative P
0	7	0.500	7	0.500
1	3	0.214	7+3=10	0.714
2	2	0.143	7+3+2=12	0.857
3	1	0.071	7+3+2+1=13	0.929
4	1	0.071	7+3+2+1+1=14	1.000

In 12 dei 14 reparti di emergenza (pari al 86% del totale) sono state riscontrate 2 o meno morti causate da incidenti stradali

0.875=0.5+0.214+0.143=12/14

Frequenze cumulate assolute e relative

- **frequenze cumulate assolute F**
  - ✓ La prima frequenza cumulata è pari alla prima frequenza assoluta.
  - ✓ L'ultima frequenza cumulata è pari alla numerosità campionaria.
- **frequenze cumulate relative P**
  - ✓ La prima frequenza cumulata relativa è pari alla prima frequenza relativa.
  - ✓ L'ultima frequenza cumulata relativa è pari ad uno.

Grafici – Var. quantitative discrete

Morti causate da incidenti stradali

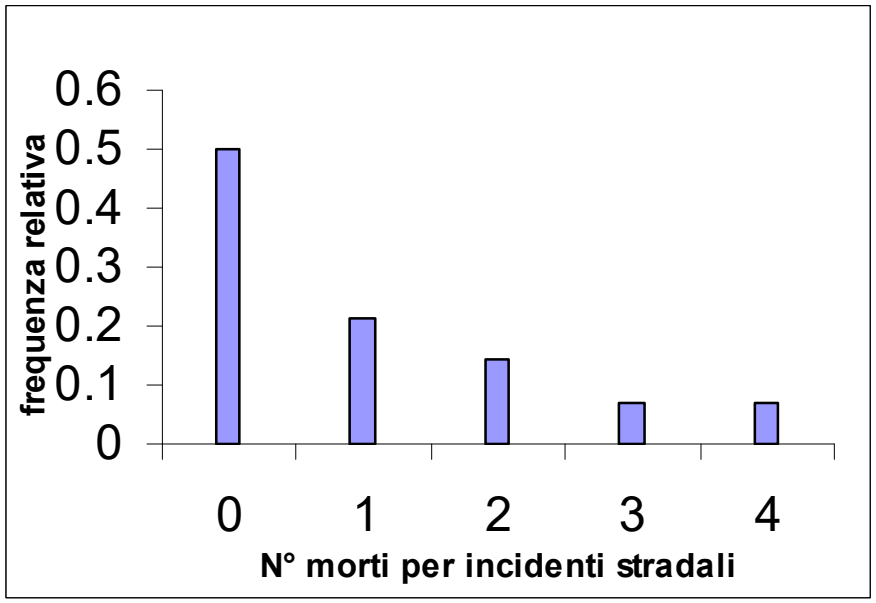


Diagramma ad aghi
(frequenza relative)

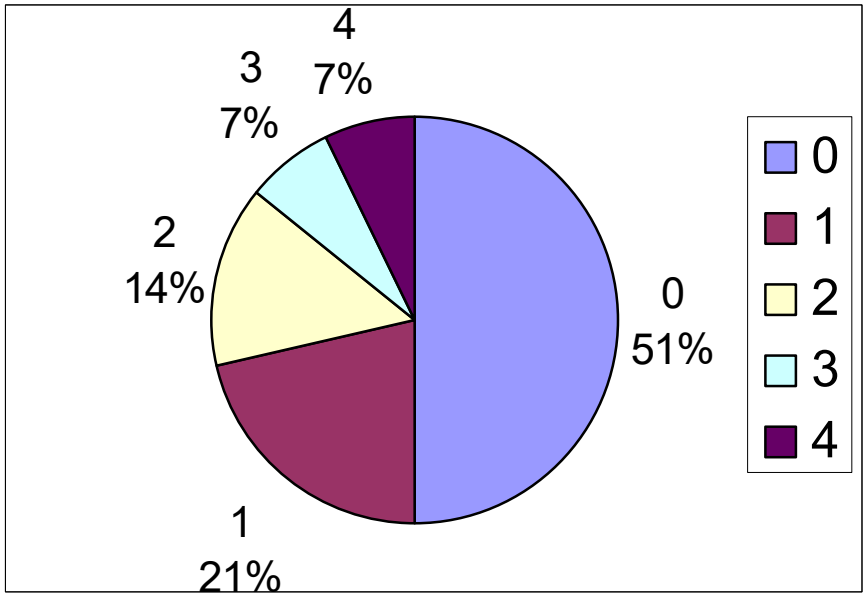


Diagramma a torta
(frequenze relative)

Distribuzioni di frequenza : il caso di variabili continue

In un'indagine condotta da un gruppo di neonatologi si sono rilevati i valori della lunghezza supina (cm) in un campione di 60 neonati. Le misurazioni, eseguite con l'infantometro Harpenden, sono riportate di seguito.

51.0	46.5	48.7	54.5	46.0	51.2	55.0	50.2	44.5	56.3
49.4	47.8	50.0	48.2	52.2	51.1	50.2	53.4	49.2	46.5
49.0	49.7	52.9	48.9	47.0	54.7	50.3	47.4	50.5	51.5
52.5	44.4	50.8	51.2	50.8	52.3	47.7	50.5	49.5	50.9
51.5	49.8	46.2	49.5	50.0	48.2	48.5	51.7	52.9	51.6
51.8	53.0	48.9	54.0	52.5	50.8	53.8	49.5	50.5	52.7

Possiamo migliorare un po' la situazione …

44.4	48.2	49.5	50.5	51.5	52.9
44.5	48.2	49.5	50.5	51.5	52.9
46.0	48.5	49.7	50.8	51.6	53.0
46.2	48.7	49.8	50.8	51.7	53.4
46.5	48.9	50.0	50.8	51.8	53.8
46.5	48.9	50.0	50.9	52.2	54.0
47.0	49.0	50.2	51.0	52.3	54.5
47.4	49.2	50.2	51.1	52.5	54.7
47.7	49.4	50.3	51.2	52.5	55.0
47.8	49.5	50.5	51.2	52.7	56.3

Distribuzioni di frequenza : il caso di variabili continue

La **distribuzione di frequenza** di una **variabile continua** si rappresenta in modo analogo a quella degli altri tipi di variabili, ma….

in questo caso, la frequenza non è riferita ad un singolo valore, ma ad **intervalli (o classi)** di valori.

Distribuzioni di frequenza : il caso di variabili continue

Lunghezza supina (cm) in un campione di 60 neonati.

Estremi di classe	Valore centrale	Freq. semplici		Freq.cumulate	
		f	p%	F	P%
44.25 ⊣ 45.75	**45.0**				
45.75 ⊣ 47.25	**46.5**				
47.25 ⊣ 48.75	**48.0**				
48.75 ⊣ 50.25	**49.5**				
50.25 ⊣ 51.75	**51.0**				
51.75 ⊣ 53.25	**52.5**				
53.25 ⊣ 54.75	**54.0**				
54.75 ⊣ 56.25	**55.5**				
56.25 ⊣ 57.75	**57.0**				

9 classi di uguale ampiezza (1.50cm)

Distribuzioni di frequenza : il caso di variabili continue

Lunghezza supina (cm) in un campione di 60 neonati.

Estremi di classe	Valore centrale	Freq. semplici		Freq.cumulate	
		f	p%	F	P%
44.25 ⊣ 45.75	**45.0**	2	*3.3*	2	*3.3*
45.75 ⊣ 47.25	**46.5**	5	*8.3*	7	*11.7*
47.25 ⊣ 48.75	**48.0**	7	*11.7*	14	*23.3*
48.75 ⊣ 50.25	**49.5**	14	*23.3*	28	*46.7*
50.25 ⊣ 51.75	**51.0**	16	*26.7*	44	*73.3*
51.75 ⊣ 53.25	**52.5**	9	*15.0*	53	*88.3*
53.25 ⊣ 54.75	**54.0**	5	*8.3*	58	*96.7*
54.75 ⊣ 56.25	**55.5**	1	*1.7*	59	*98.3*
56.25 ⊣ 57.75	**57.0**	1	*1.7*	60	*100.0*

5 dei 60 neonati hanno una lunghezza supina compresa fra 45.75 e 47.25

Gli estremi di classe

[44.25-45.75) **o** **44.25 ⊢ 45.75**

classe chiusa a sinistra e aperta a destra

estremo sn incluso

(44.25-45.75] **o** **44.25 ⊣ 45.75**

classe chiusa a destra e aperto a sinistra

estremo dx incluso

[44.25-45.75] **o** **44.25 ⊢⊣ 45.75**

classe chiusa a sinistra e a destra

estremo sn e dx inclusi

(44.25-45.75) **o** **44.25 - 45.75**

classe aperta a sinistra e a destra

estremo sn e dx esclusi

Le classi

- ✓ La scelta del **numero** di classi e degli **estremi** è arbitraria. Entrambi vengono determinati in base a criteri di convenienza.
- ✓ Il **numero** di classi può oscillare e dipende dalla numerosità dei dati.
- ✓ Scegliere **estremi** che siano clinicamente/biologicamente **significativi** o naturali e, preferibilmente, **di uguale ampiezza**.

  NO: 44.137 - 45.541 SI: 44.00 - 45.50
- ✓ Le classi debbono essere mutuamente esclusive (fate attenzione agli estremi!!).

Diagramma a barre con frequenze assolute (erroneamente chiamato istogramma)

f

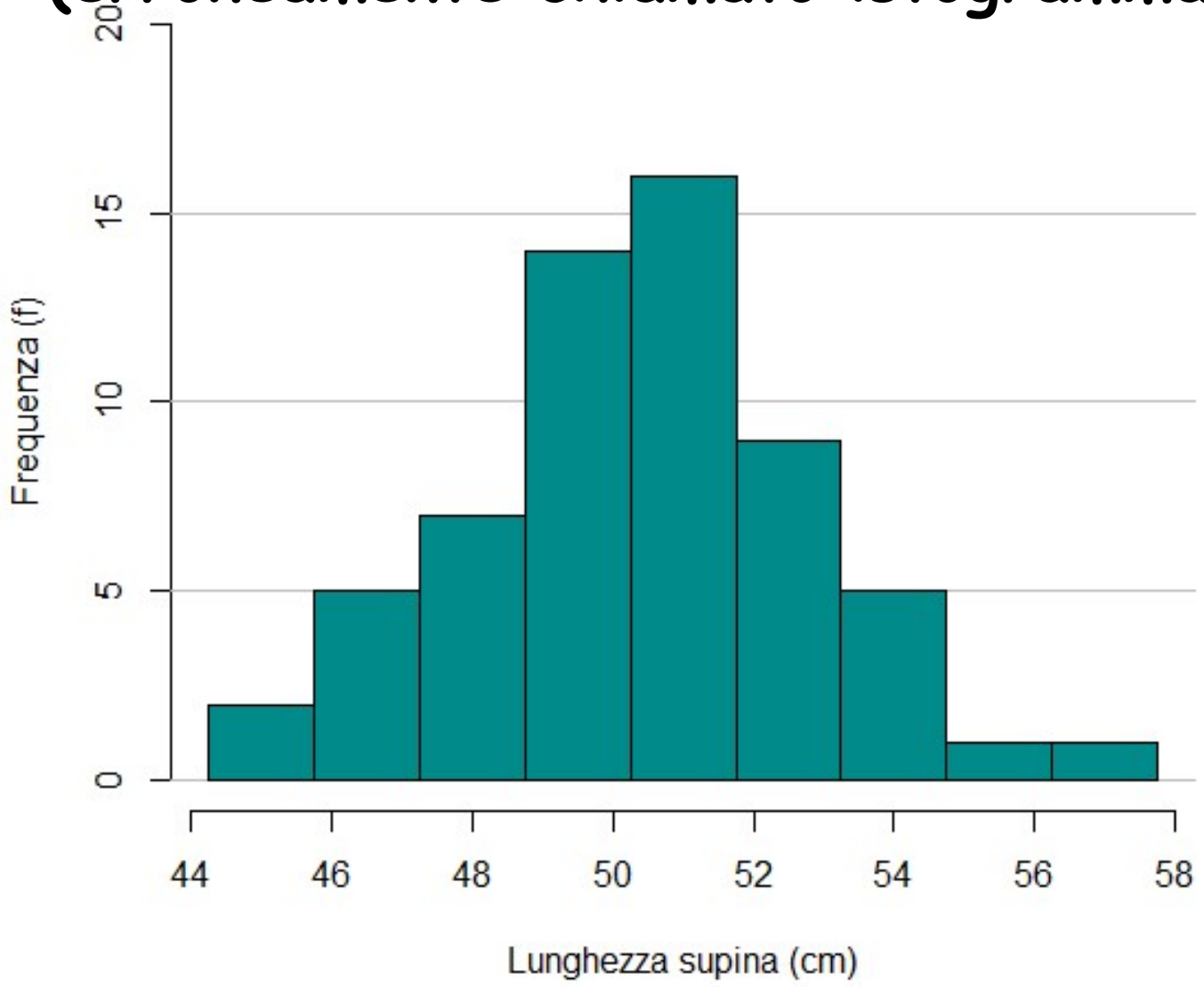
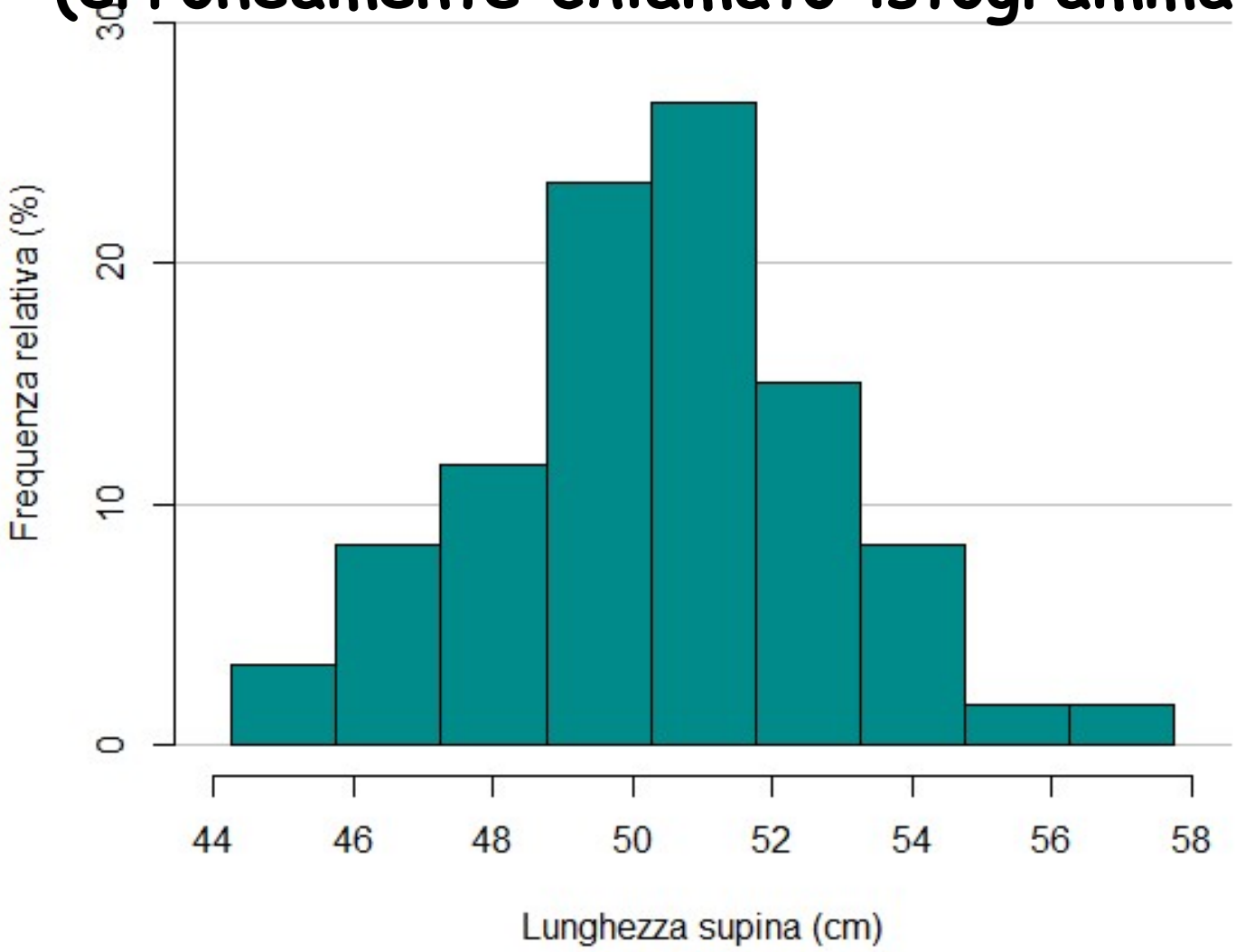


Diagramma a barre con frequenze relative

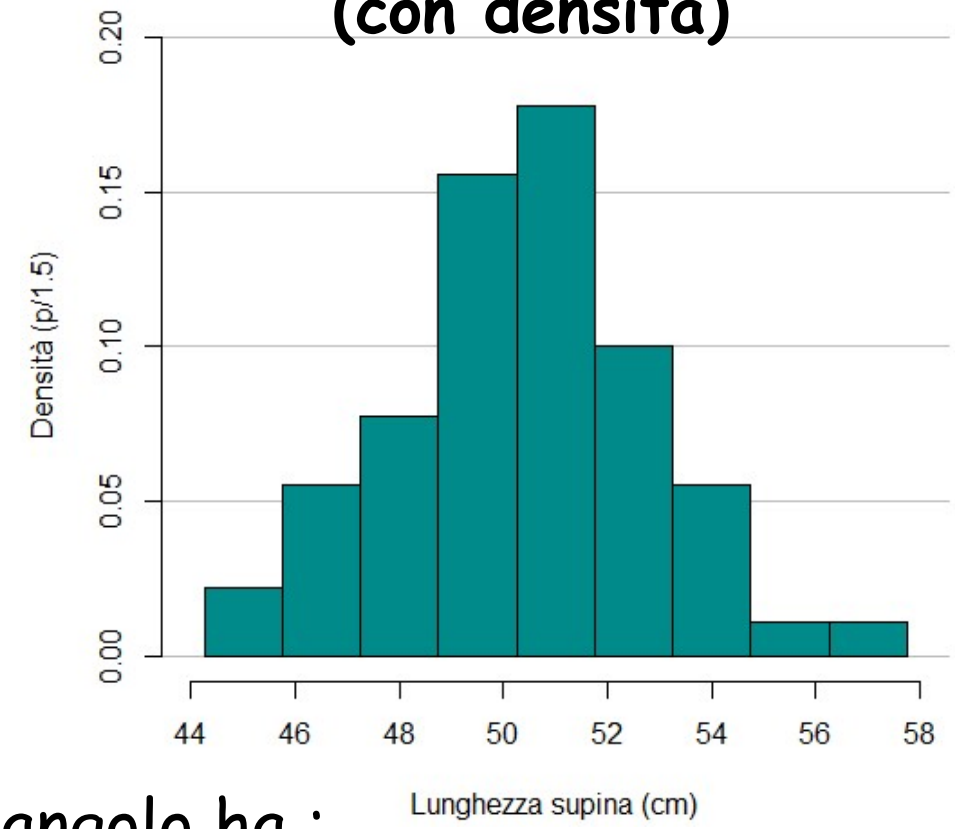
(erroneamente chiamato istogramma)

p%



Istogramma
(con densità)

p/1.5



Ciascun rettangolo ha :

- per base l'ampiezza della classe
- per altezza la frequenza relativa della classe diviso l'ampiezza (densità di frequenza)
- un'area pari alla frequenza relativa

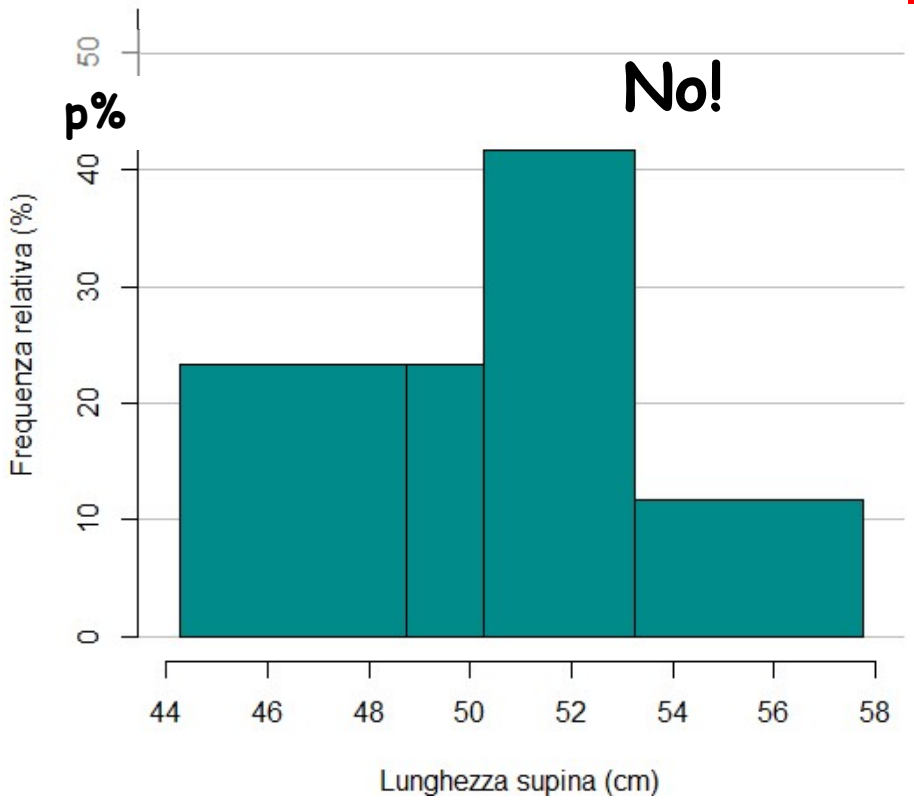
Globalmente i rettangoli ricoprono un'area unitaria

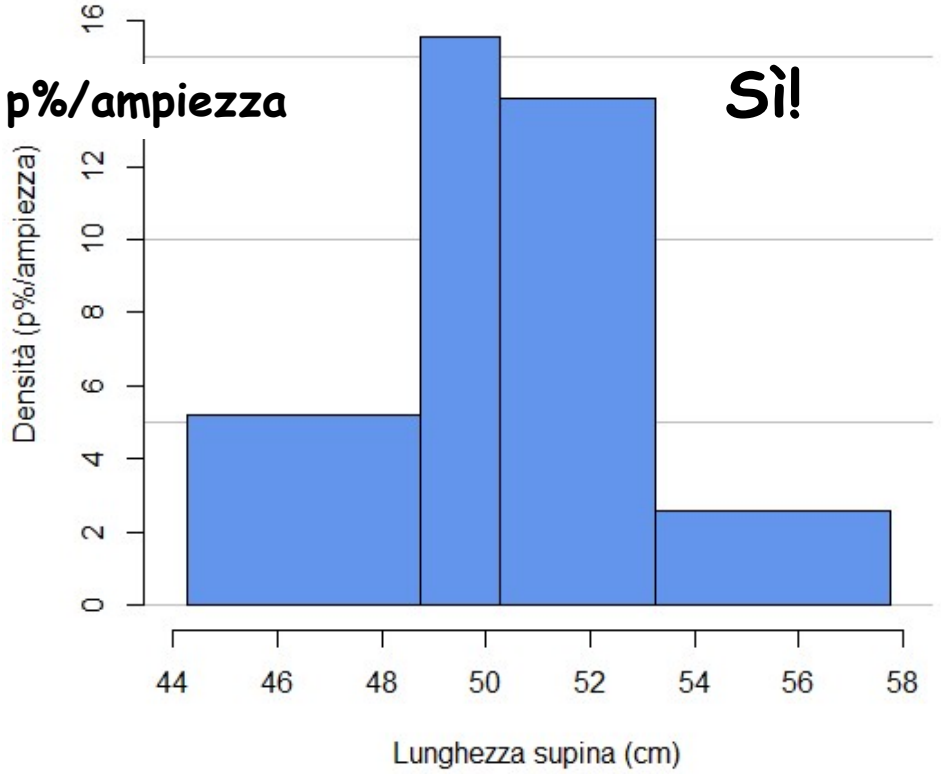
Densità di frequenza

Classe			f	p	d=p/amp =p/1.5
42.8	-|	44.3	0	0.000	0.000
44.3	-|	45.8	2	0.033	0.022
45.8	-|	47.3	5	0.083	0.056
47.3	-|	48.8	7	0.117	0.078
48.8	-|	50.3	14	0.233	0.156
50.3	-|	51.8	16	0.267	0.178
51.8	-|	53.3	9	0.150	0.100
53.3	-|	54.8	5	0.083	0.056
54.8	-|	56.3	1	0.017	0.011
56.3	-|	57.8	1	0.017	0.011
57.8	-|	59.3	0	0.000	0.000

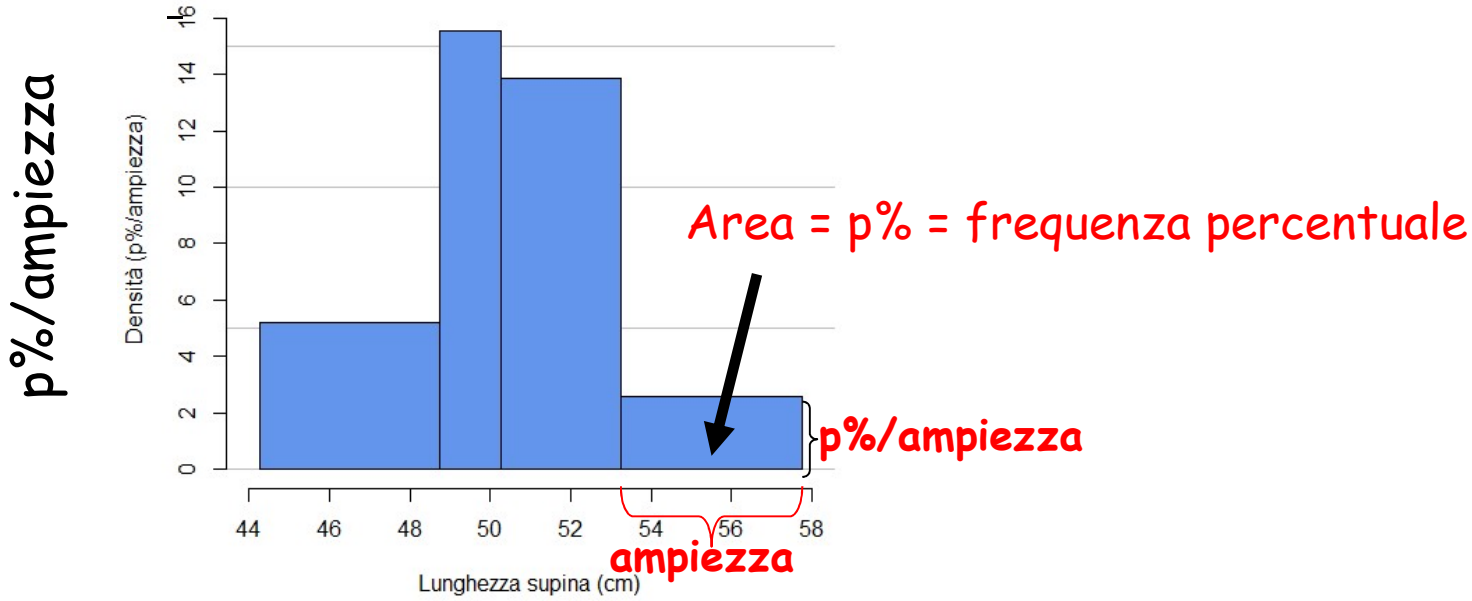
Classi di diversa ampiezza

Estremi di classe	Ampiezza di classe	freq. semplici		Densità freq.	
		f	p%	f/amp	p%/amp
(44.25 , 48.75]	4.5	14	*23.3*	3.1	*5.2*
(48.75 , 50.25]	1.5	14	*23.3*	9.3	*15.5*
(50.25 , 53.25]	3	25	*41.7*	8.3	*13.9*
(53.25 , 57.75]	4.5	7	*11.7*	1.6	*2.6*



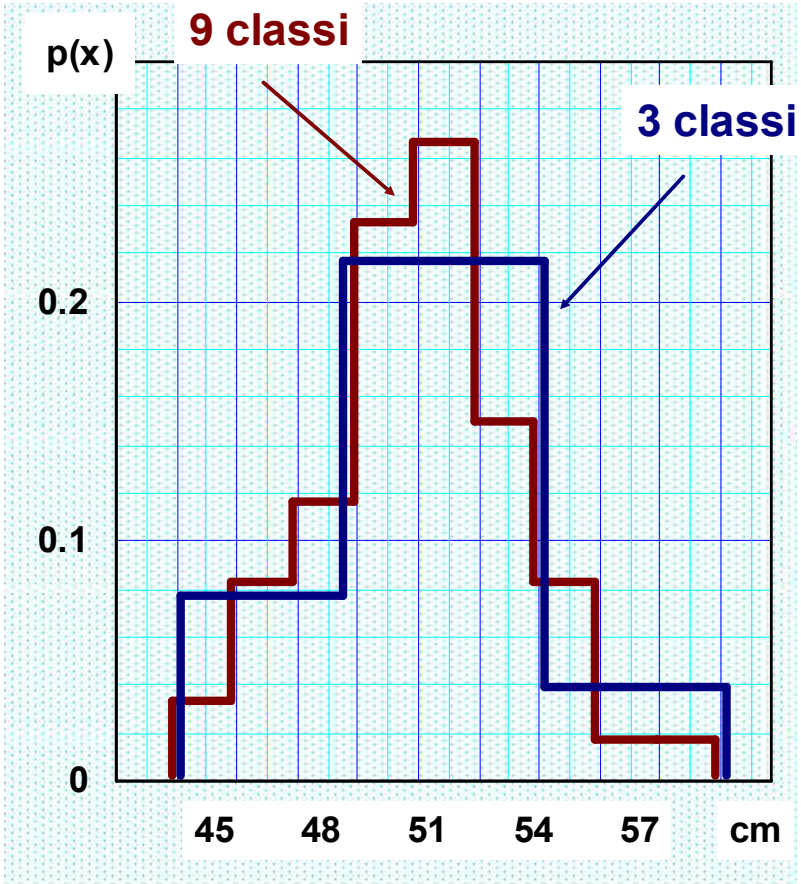


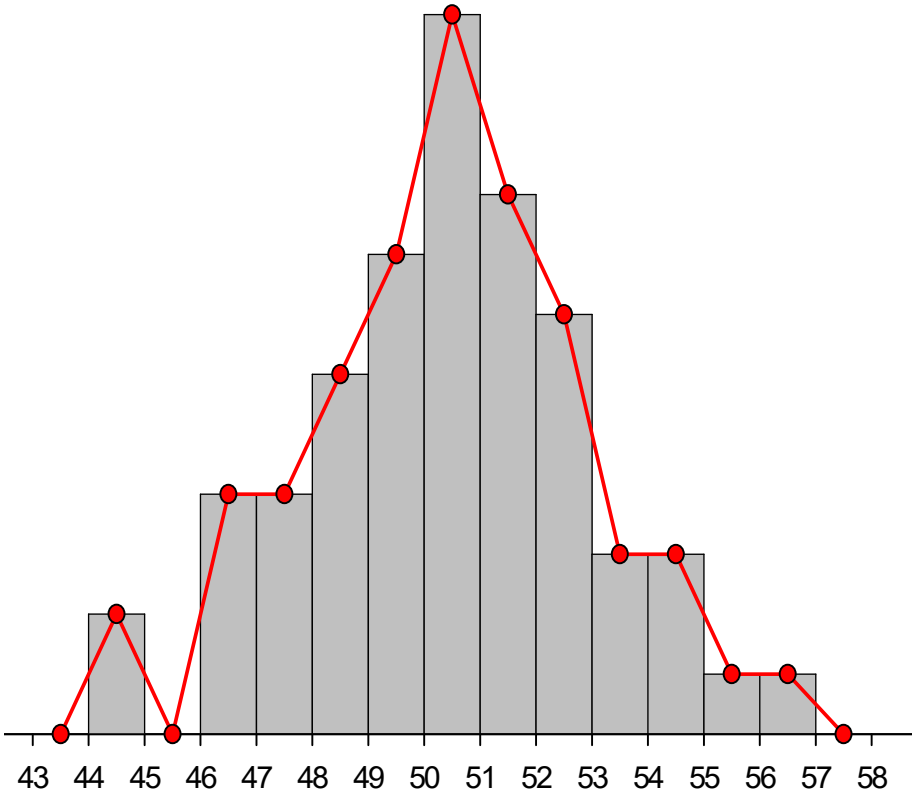
Classi di diversa ampiezza



- ✓ Ogni istogramma (rettangolo) rappresenta una classe:
  - **base** = ampiezza della classe (es. 4.5 cm)
  - **altezza** = frequenza relativa percentuale/ampiezza (es. 2.6)
- ✓ L'area di ogni rettangolo è pari alla frequenza relativa della classe su cui insiste (es. 11.7%)
- ✓ L'area totale è 100

Ampiezza delle classi

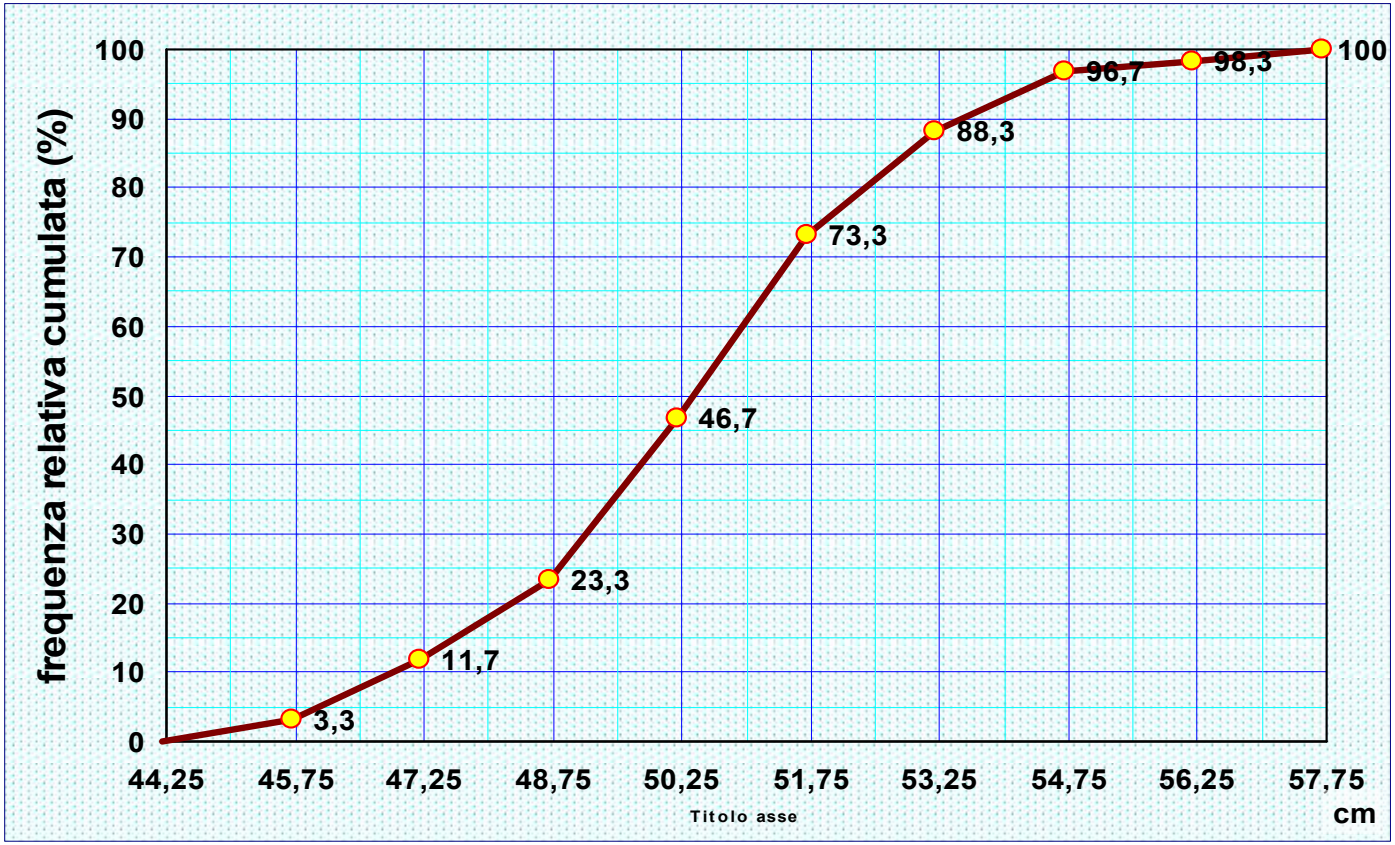




Al diminuire del numero di classi si perdono i dettagli sulla distribuzione.

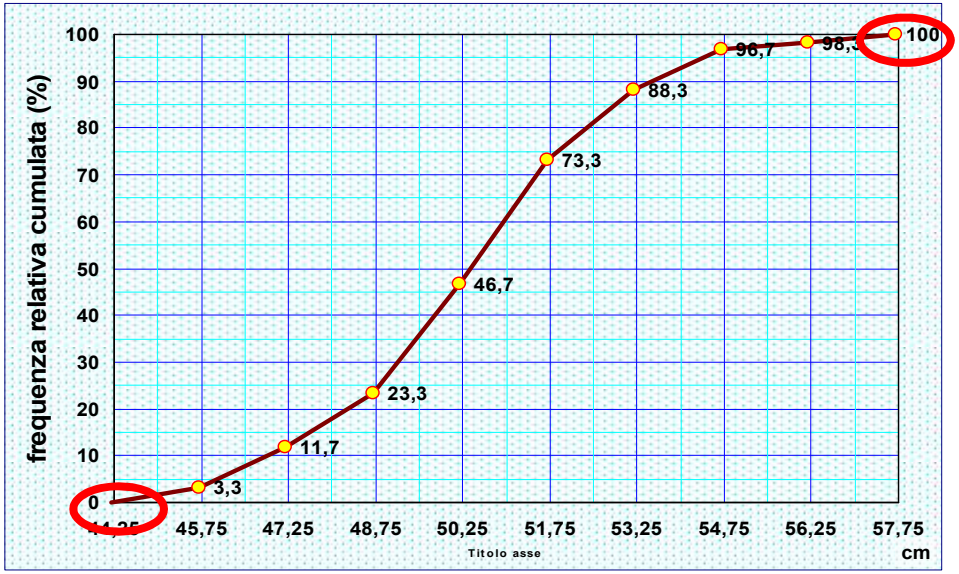
All'aumentare del numero di classi si guadagnano dettagli sulla distribuzione (ma sino ad un certo punto!!)

Grafico delle frequenze cumulate



Ogiva di Galton

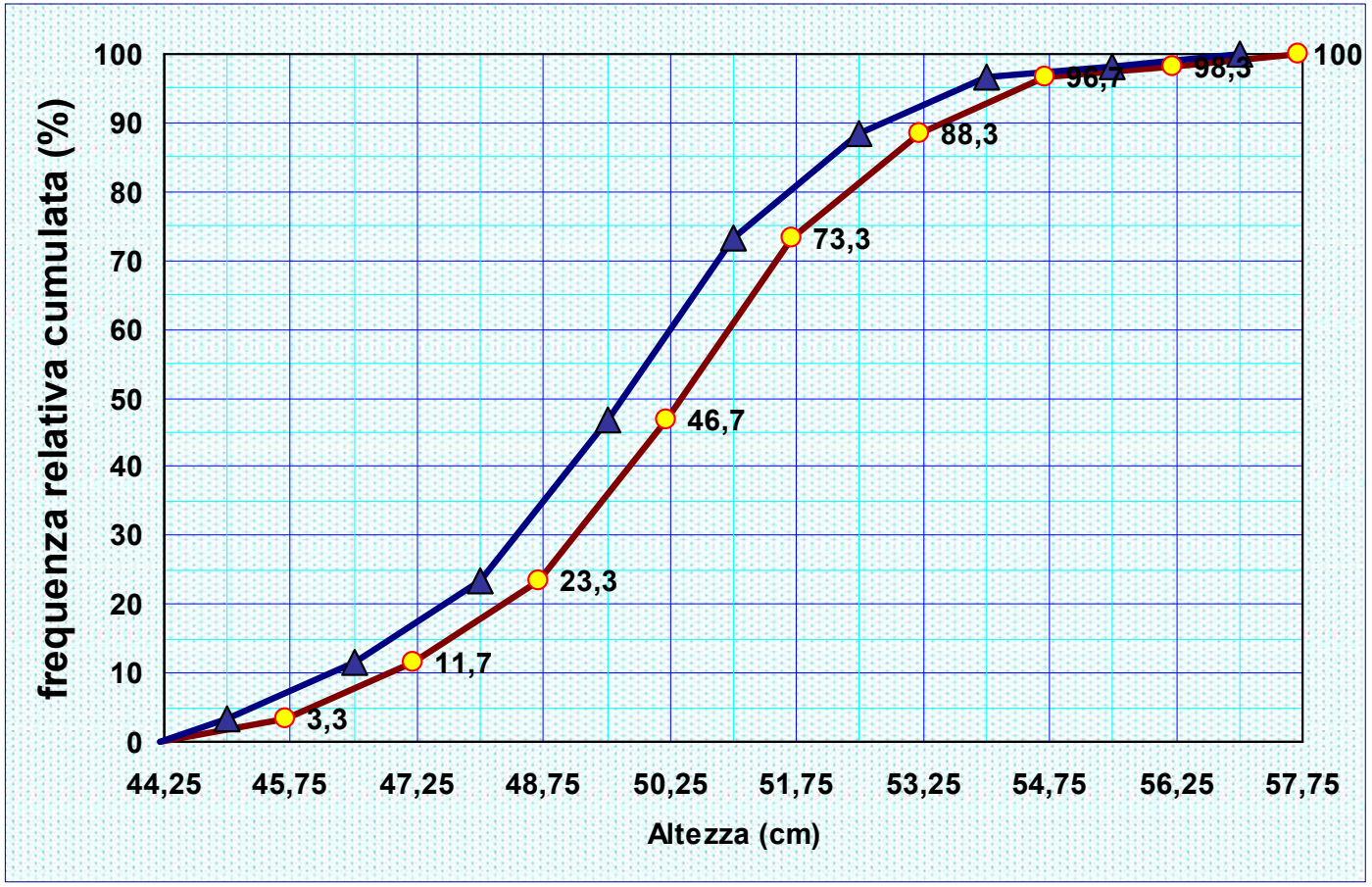
Grafici per var. continue



- ✓ La spezzata parte da 0 e termina a 1 o 100%.
- ✓ La spezzata si ottiene congiungendo con dei segmenti i due punti che hanno per coordinate:

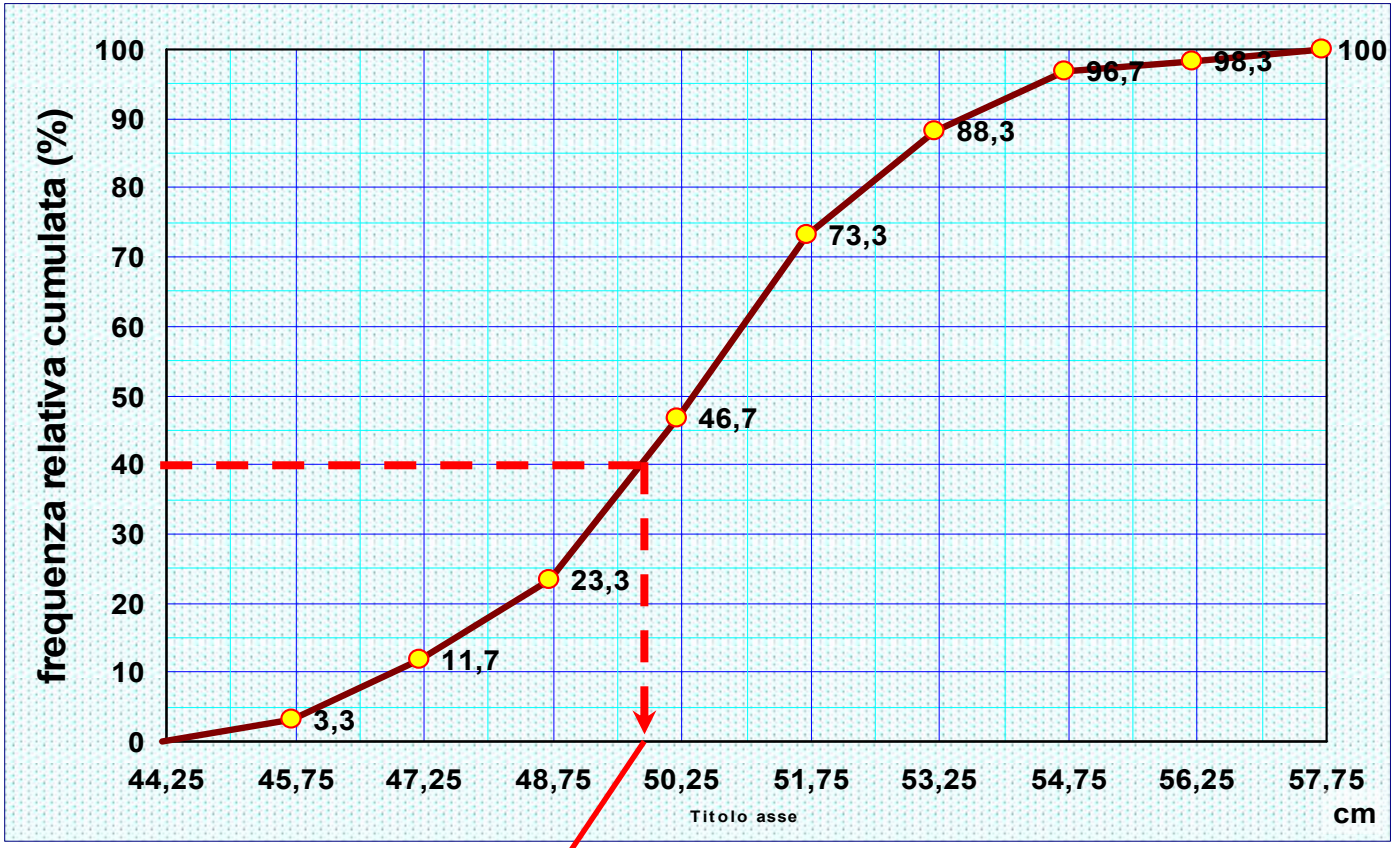
  [estr inf, freq cum prec] — [estr sup, freq cum]
- ✓ Si assume che la distribuzione dei dati nelle classi sia uniforme (interpolazione lineare)

Grafici per var. continue



Se si congiungessero i valori centrali si otterrebbe una rappresentazione scorretta.

Grafici per var. continue



Qual è il valore di altezza sotto il quale trovo il 40% dei neonati?

~ 49.75 cm

Descrizione di una variabile in più popolazioni

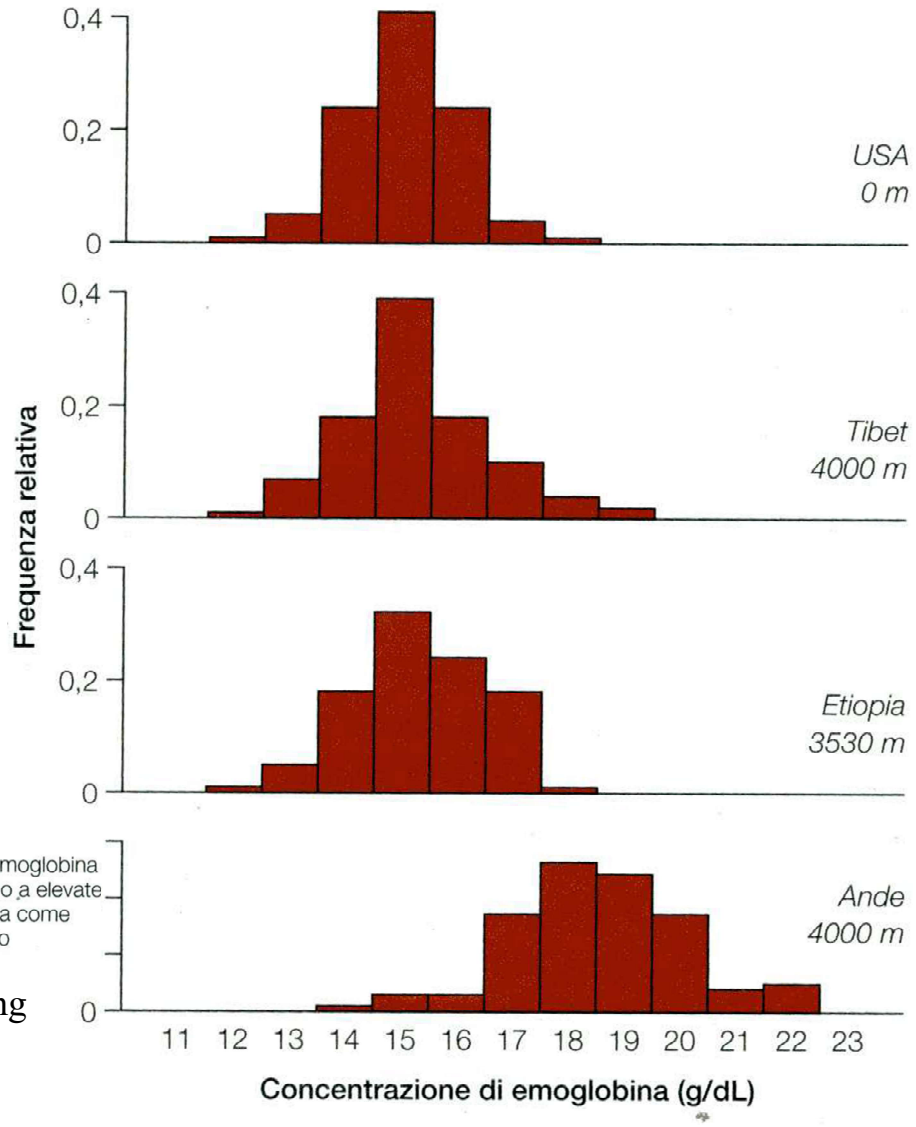


Figura 2.4-1
Istogrammi che mostrano la concentrazione di emoglobina in maschi di popolazioni umane viventi che vivono a elevate altitudini in tre differenti parti del mondo. È inclusa come controllo una quarta popolazione che vive a livello del mare (USA).

da Beal et al. 2002. Proceeding of the National Academy of Science 99:17215-17218.

Descrizione di una variabile in più popolazioni

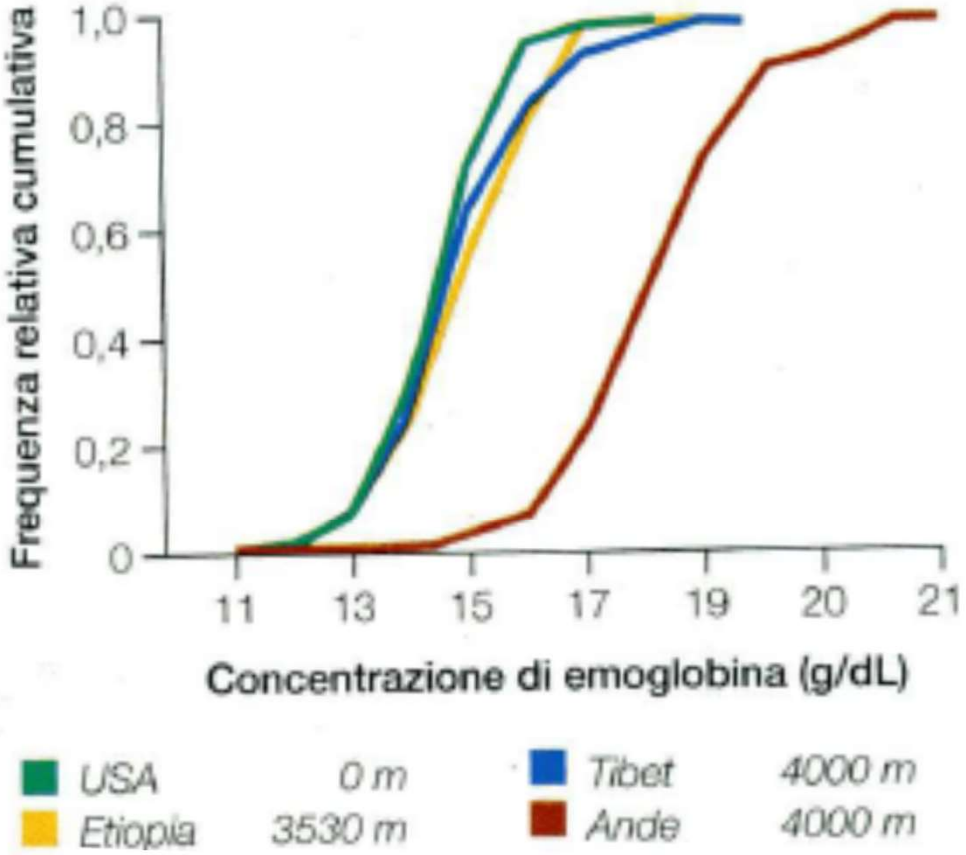


Figura 2.4-2
Distribuzioni di frequenza cumulative della concentrazione di emoglobina in maschi umani che vivono ad altitudini elevate in Etiopia, in Tibet e sulle Ande. È inclusa come controllo una quarta popolazione che vive a livello del mare negli Stati Uniti. (Da Beall et al., 2002; ridisegnato.)

Descrizione di una variabile in più popolazioni

Distribuzione di frequenza a doppia entrata

	Livello ematico di emoglobina (Hb, g/dl)					
	12 (11.5,12.5]	13 (12.5,13.5]	14 (13.5,14.5]	15 (14.5,15.5]	16 (15.5,16.5]	**Totale**
donne	18	65	14	2	1	**100**
uomini	2	40	71	58	29	**200**
Totale	**20**	**105**	**85**	**60**	**30**	**300**

Quale proporzione di soggetti ha livello di Hb > di 14.5 g/dl ?

Quale proporzione di donne ha livello di Hb > di 14.5 g/dl ?

Definizione di Percentile

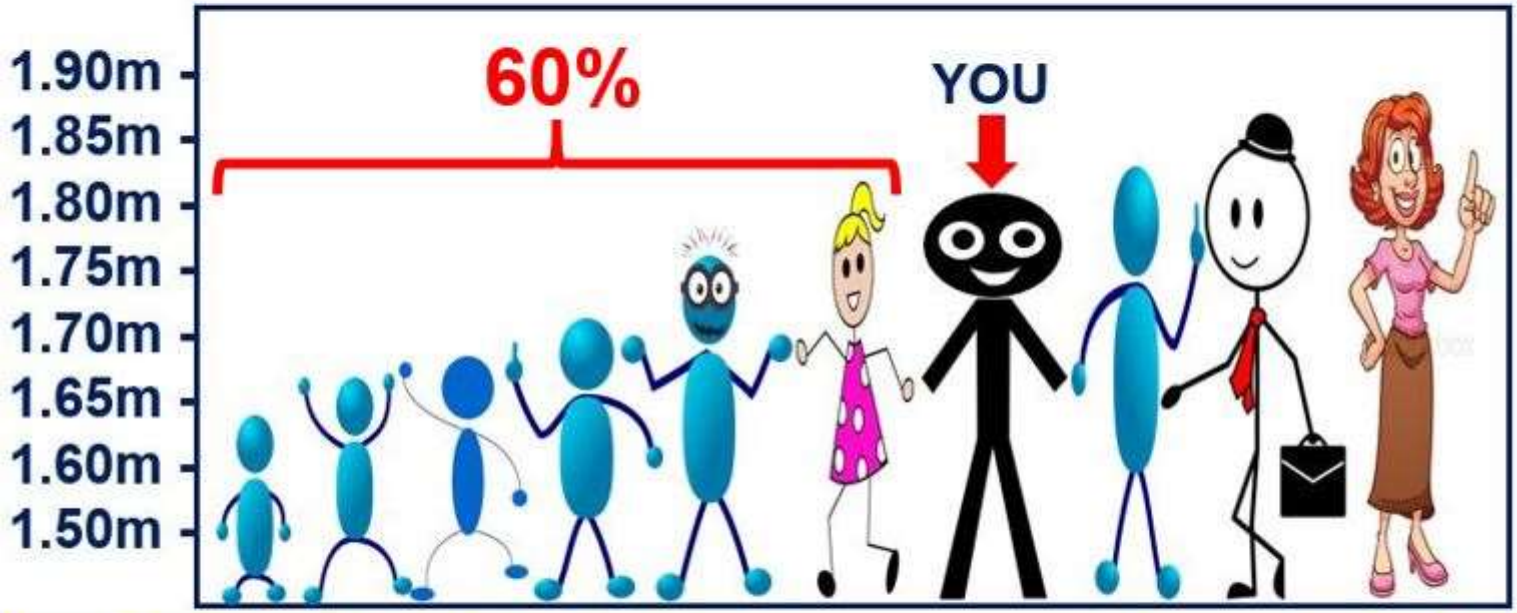
Il **percentile** x_p $(0 \leq p \leq 1)$ della distribuzione di una variabile continua è quel valore della variabile che soddisfa queste condizioni

1. il p% delle osservazioni assume valori < = di x_p,
2. l' (1-p)% delle osservazioni assume valori > di x_p

I percentili sono utili per:

- Descrivere una distribuzione
- Identificare range di normalita'
- Classificare il valore di un soggetto rispetto alla distribuzione del fenomeno

What Percentile are You?



You are the 4th tallest person in the group of 10

60% of people are shorter than you

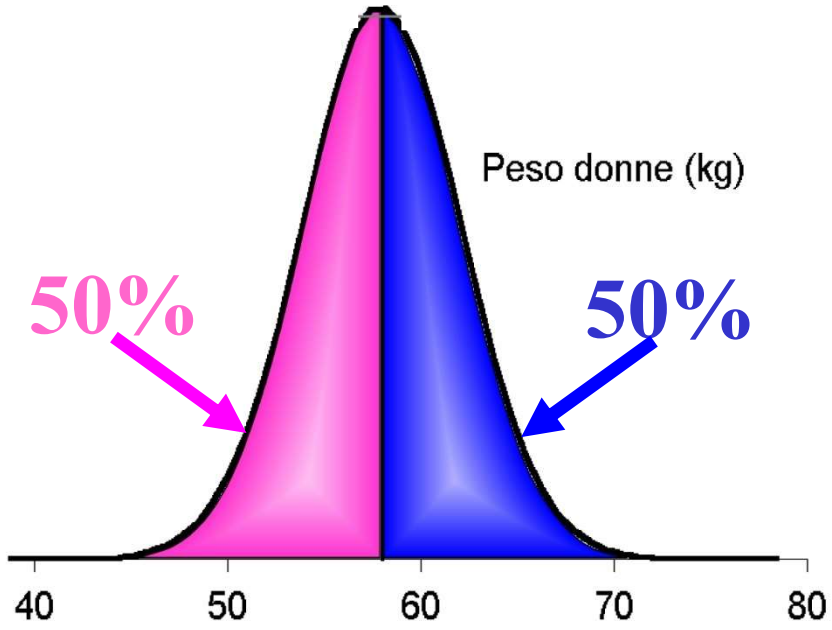
That means that you are at the 60th percentile

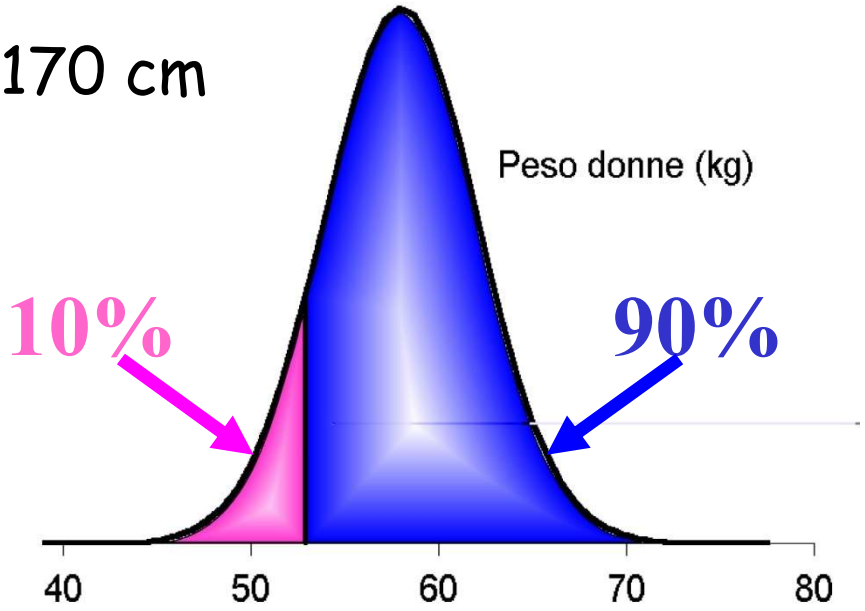
If your height is 1.80m, then 1.80m is the 60th percentile height in that group

Percentili da un istogramma

Peso delle donne di altezza 160-170 cm

p = 0.10 $x_{0.10}$ = **53**



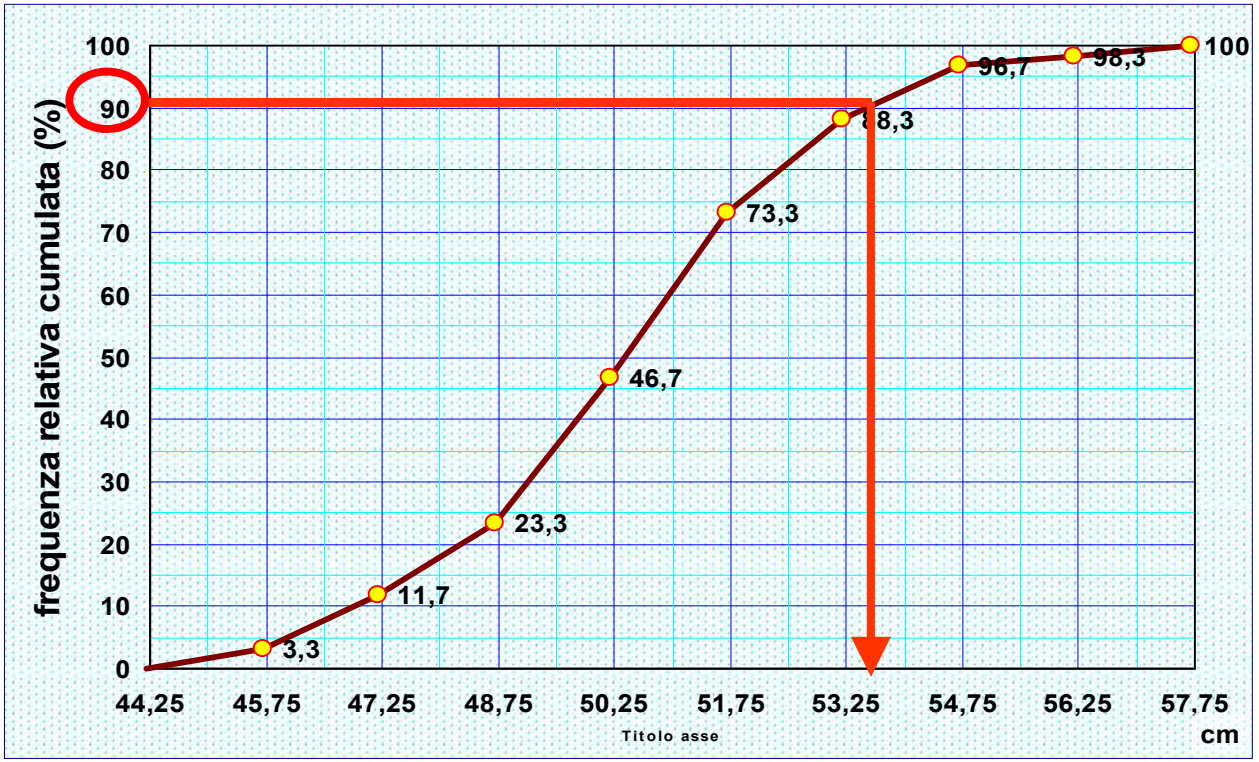


p = 0.50 $x_{0.50}$ =**58**

Percentili a partire dalle frequenze relative cumulate

Lunghezza dei bambini

Es. p=0.90

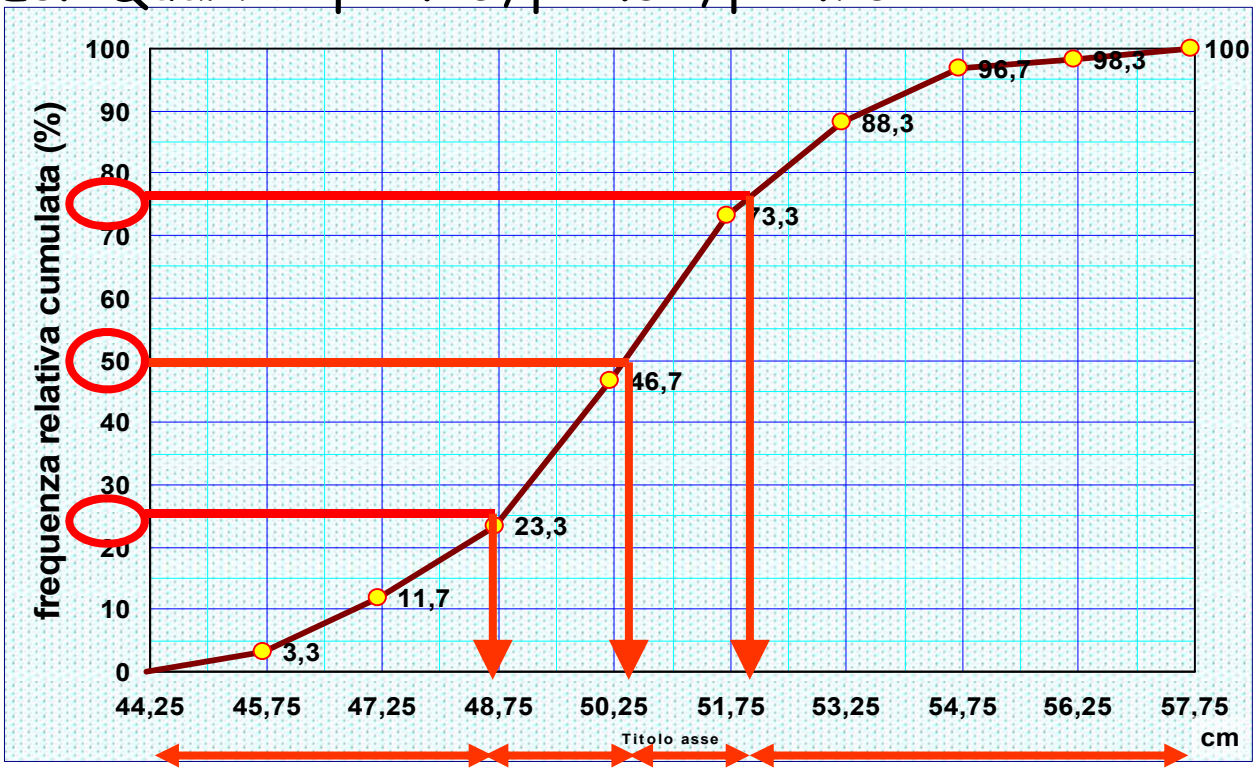


Percentili particolari: quartili

Quartili: suddividono i dati in quattro parti uguali (25%)

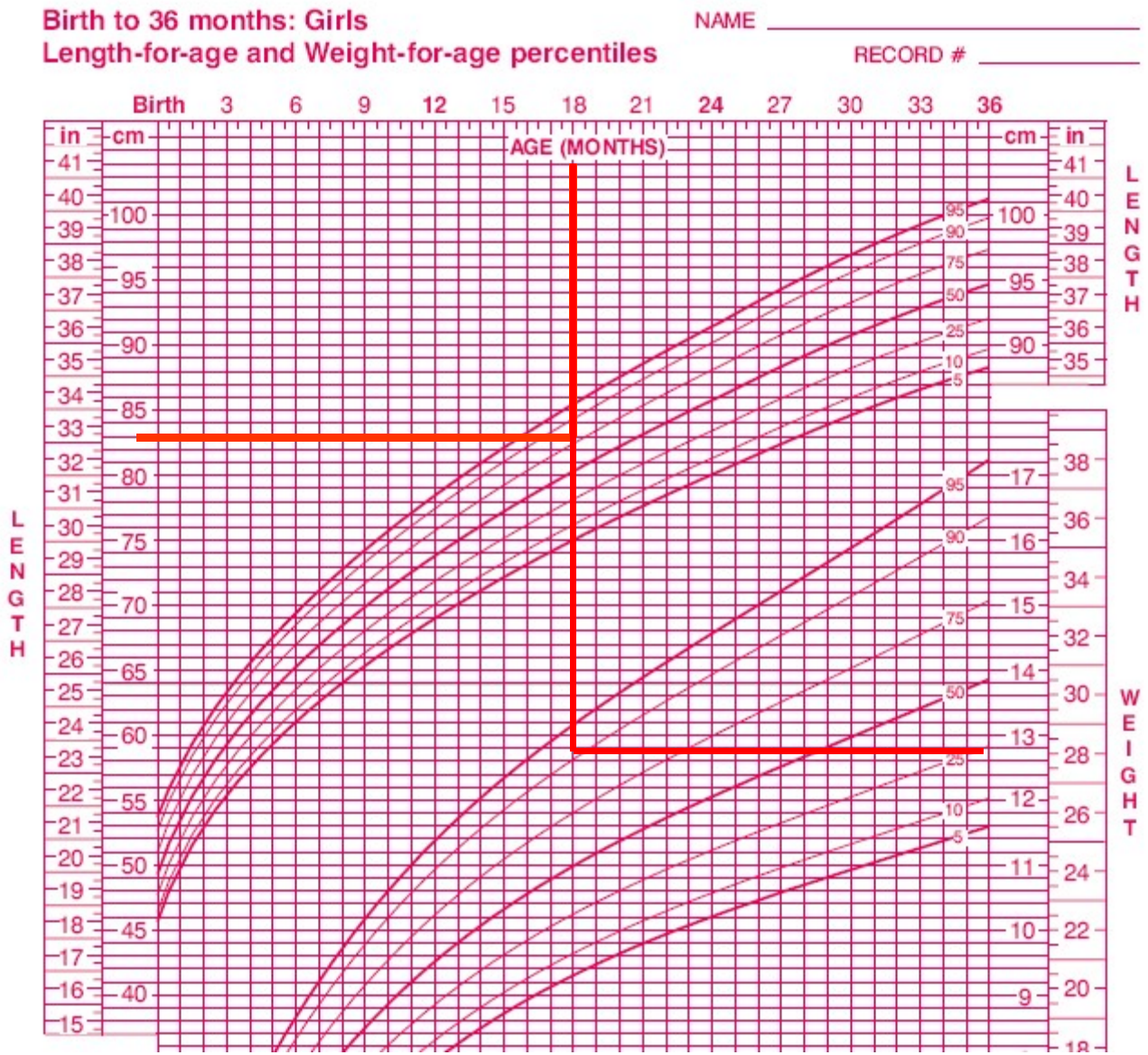
Lunghezza dei bambini

Es. Quartili: p=0.25, p=0.50, p=0.75

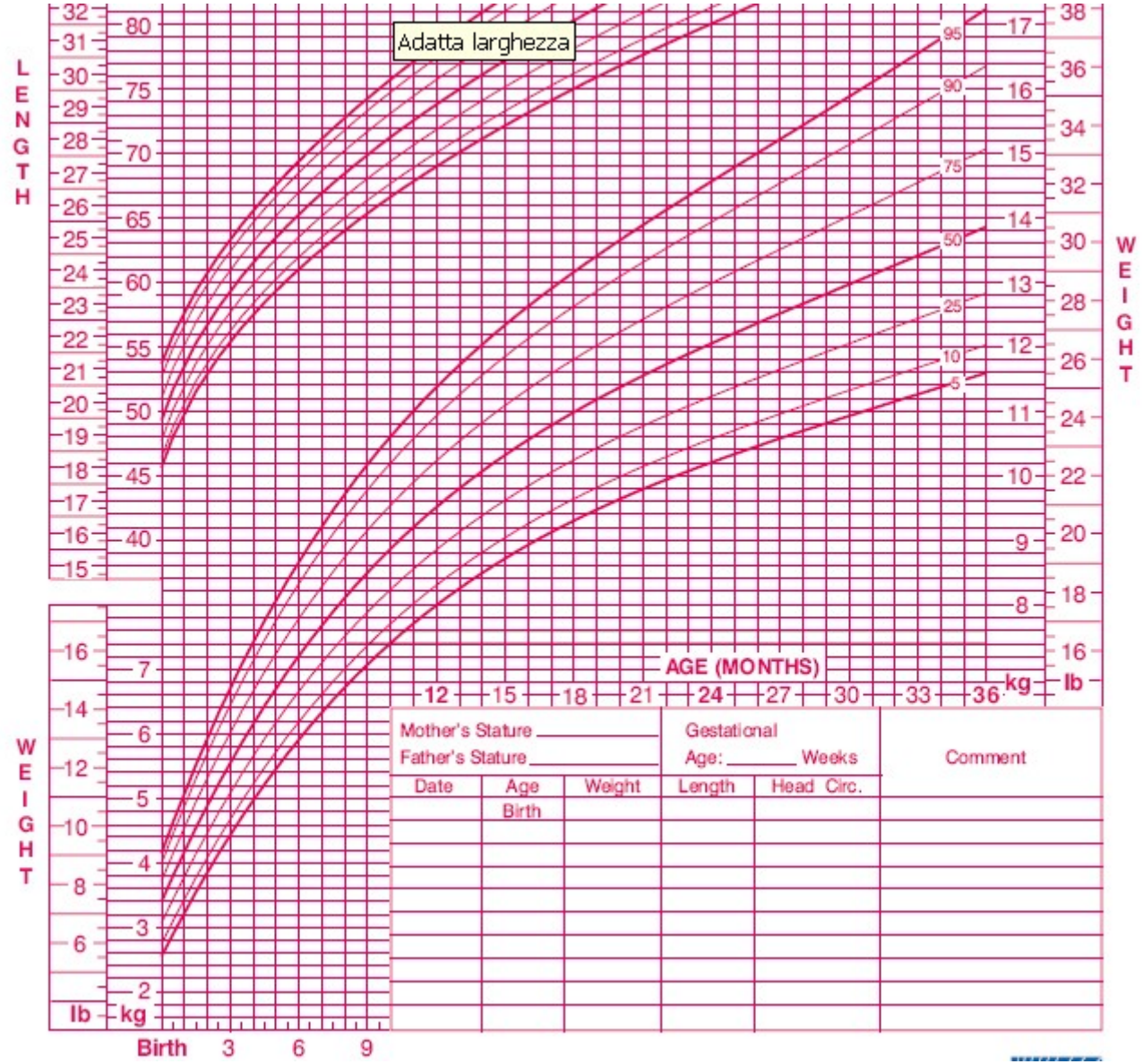


25% 25% 25% 25%

Curve Percentile – lunghezza e peso



Curve Percentile – peso neonate



Esercizio per lo studente

Glicemia (mg/dl) in 500 soggetti anziani
Raggruppamento in 5 classi di uguale ampiezza

Estremi di classe	*valore centrale*	*freq. semplici*		*freq. cumulate*	
		f	p%	F	P%
65-|75	**70**	*75*	*15*	*75*	*15*
75-|85	**80**	100	*20*	175	*35*
85-|95	**90**	150	*30*	225	*65*
95-|105	**100**	125	*25*	450	*90*
105-|115	**110**	50	*10*	500	*100*

- Rappresentare graficamente il fenomeno mediante un istogramma
- Accorpare le ultime due classi e costruire il relativo istogramma
- Rappresentare i dati in un Ogiva di Galton e individuare il valore di glicemia superato solo dal 5% di questi anziani selezionati