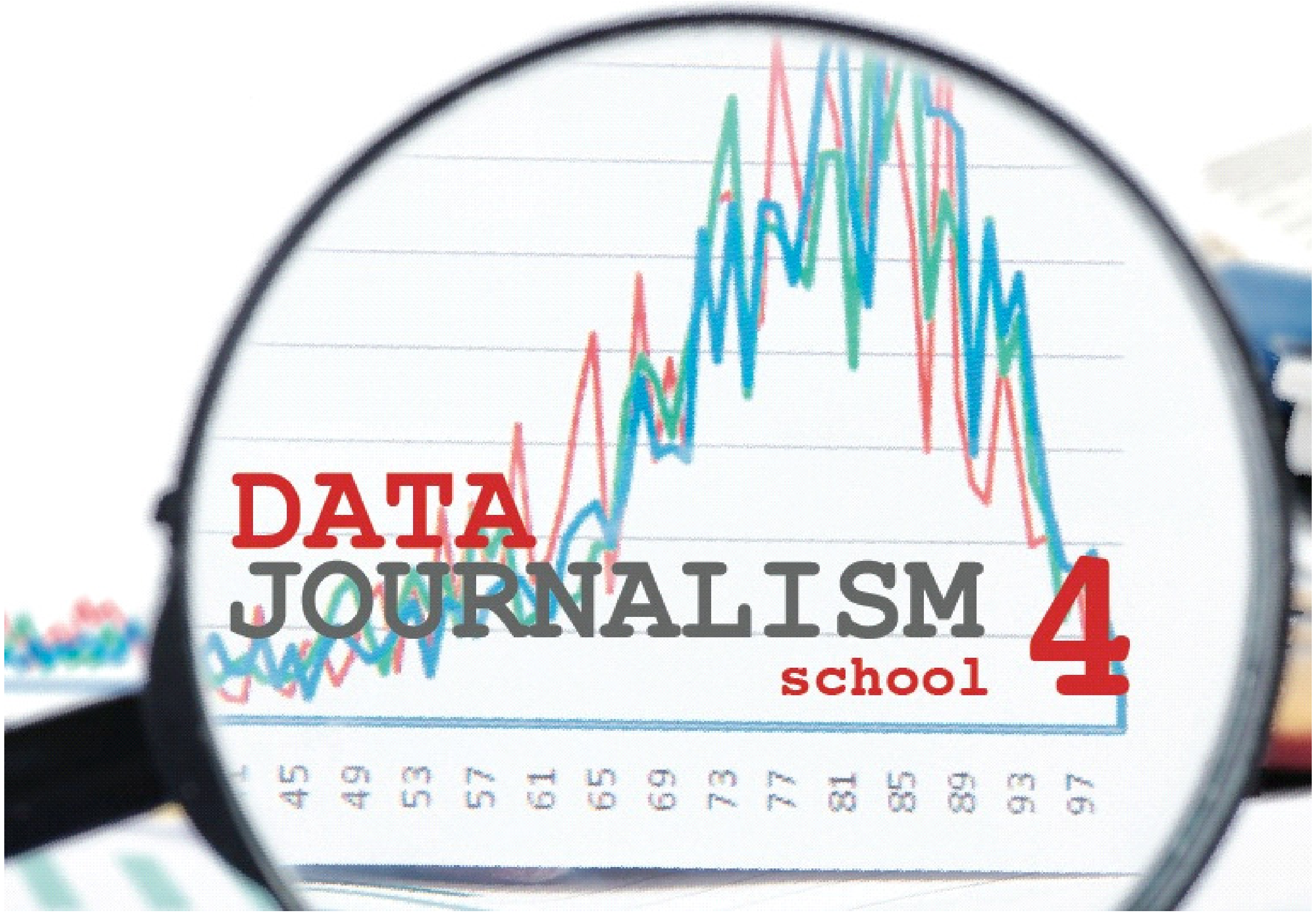


Strumenti statistici di base

Giancarlo Bruno

Roma, 18 dicembre 2013

fondazione ahref living links

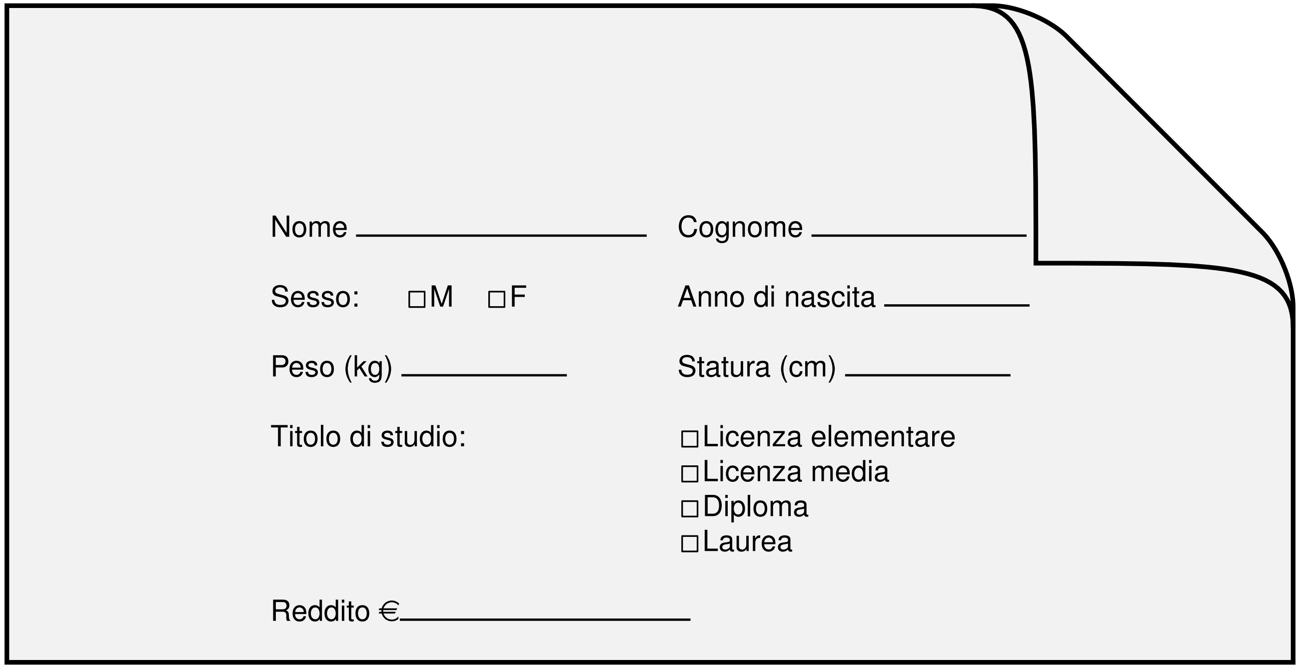
Istat Istituto Nazionale di Statistica

Indice

- ✗ Possiamo definire la statistica come la scienza dei fenomeni collettivi.
- ✗ I metodi statistici, infatti, trovano applicazione ovunque occorra investigare fenomeni di massa che possano essere in qualche modo quantificati e che siano caratterizzati da variabilità.
- ✗ La statistica definisce le procedure per misurare e rappresentare i fenomeni collettivi e studia i metodi per stimare i valori caratteristici di una popolazione a partire da un campione.

Questionario

Tratteremo alcuni concetti statistici con un esempio. Il nostro punto di partenza è costituito da una rilevazione, condotta mediante un questionario in cui sono richieste alcune informazioni personali:


Nome ____________ Cognome ____________

Sesso: ☐M ☐F Anno di nascita ________

Peso (kg) ________ Statura (cm) ________

Titolo di studio:
☐Licenza elementare
☐Licenza media
☐Diploma
☐Laurea

Reddito €____________

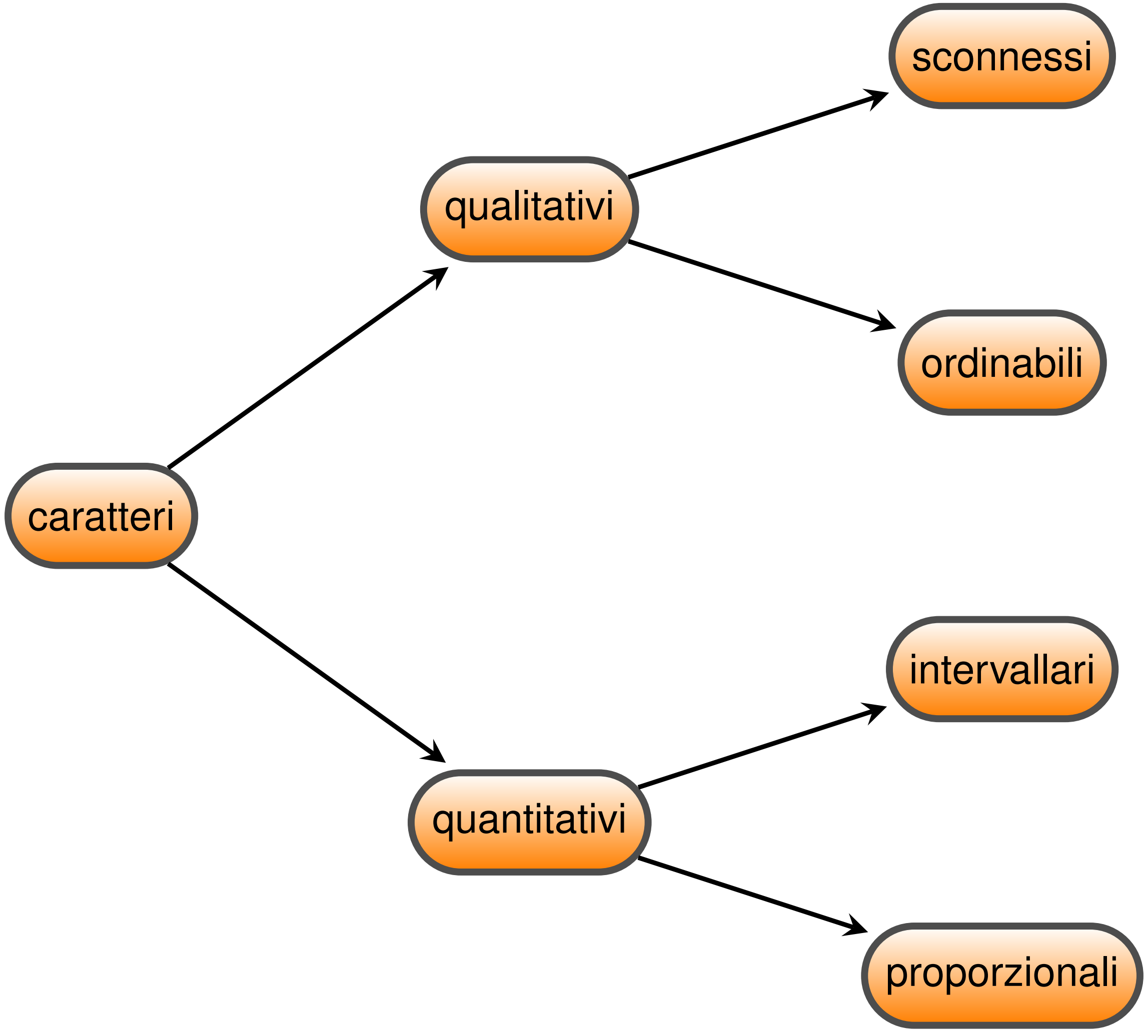
La matrice dei dati

N.id.	Sesso	Titolo di studio	Anno	Statura	Peso	Reddito
1	M	Licenza elementare	1971	182	79,8	14.300
2	M	Licenza elementare	1957	159	69,9	19.800
3	F	Diploma	1938	170	52,0	27.100
4	F	Licenza media	1939	171	56,4	23.200
5	F	Licenza media	1968	165	59,3	23.000
6	F	Licenza media	1973	167	59,6	22.400
7	F	Diploma	1940	173	63,9	23.600
8	F	Licenza media	1944	161	53,1	15.500
9	M	Licenza media	1965	172	80,4	17.600
10	F	Licenza elementare	1952	169	53,1	15.600
11	M	Licenza elementare	1979	180	81,7	18.600
12	M	Laurea	1972	187	68,4	49.700
13	F	Diploma	1985	161	56,5	21.000
14	M	Licenza media	1969	178	77,5	22.600
15	F	Diploma	1946	168	56,4	21.900
16	F	Laurea	1986	156	46,1	25.500
17	M	Diploma	1983	173	69,2	32.400
18	M	Diploma	1953	176	79,8	20.100
19	F	Licenza media	1963	175	59,6	19.800
20	M	Laurea	1982	170	65,0	34.200

Caratteri

- ✗ La matrice dei dati così costruita contiene delle osservazioni su venti unità statistiche (i rispondenti al questionario) rispetto alle quali sono stati rilevati i valori di diversi caratteri: sesso, reddito, ecc.
- ✗ Nell'analisi si può privilegiare la dimensione delle unità statistiche o dei caratteri.
- ✗ E' immediato osservare che vi sono diverse tipologie di caratteri, e ciò comporta delle conseguenze, in particolare, sul modo in cui vengono misurati.

Caratteri


caratteri
qualitativi
quantitativi
sconnessi
ordinabili
intervallari
proporzionali

Caratteri qualitativi

- ✗ I caratteri qualitativi sono quelli che assumono modalità non numeriche: nel nostro esempio, sesso e titolo di studio.
- ✗ Il carattere “sesso” è di tipo *sconnesso*: tra i valori che può assumere si può stabilire solo la relazione di uguaglianza/disuguaglianza.
- ✗ Il carattere “titolo di studio”, invece, è suscettibile di *ordinamento*.

Caratteri quantitativi

- ✗ I caratteri quantitativi assumono valori numerici, che possono essere continui (es. statura, peso) o discreti (anno di nascita).
- ✗ Quelli che non hanno uno zero assoluto possono essere misurati su una scala di intervalli. Ad esempio, si può affermare che l'intervallo tra l'anno di nascita 1990 e il 2000 ha la stessa ampiezza dell'intervallo tra il 2000 e il 2010, ma non ha senso dire che l'anno 2000 è del 33% maggiore del 1500!
- ✗ I caratteri quantitativi che hanno uno zero assoluto possono essere misurati su una scala proporzionale: nel nostro caso peso, statura, reddito. È possibile affermare che un reddito di 2.000 € è il doppio di 1.000 €.

Distribuzioni di frequenza

- ✗ La semplice elencazione dei valori dei caratteri delle unità statistiche non consente di sintetizzare l'informazione presente, di individuare dei sottogruppi, di confrontare agevolmente questi dati con altre situazioni.
- ✗ Per questo motivo si ricorre alle distribuzioni statistiche. In particolare ci soffermeremo sulle distribuzioni di frequenza di un certo carattere.
- ✗ La distribuzione di frequenza rappresenta un modo per sintetizzare le informazioni sulle variabili del collettivo.

Frequenze assolute e relative

Ad esempio, consideriamo il carattere titolo di studio. La distribuzione di frequenze riporta, per ogni i-esima modalità del carattere, il numero di occorrenze n_i:

titolo di studio	**frequenza (n_i)**	**frequenza relativa (f_i)**
licenza elementare	4	0,20
licenza media	7	0,35
diploma	6	0,30
laurea	3	0,15
totale	20 (N)	1,00

Per eliminare l'effetto della numerosità del collettivo e rendere possibili confronti con popolazioni di ampiezza diversa, è opportuno ricorrere alle frequenze relative: $f_i = n_i/N$.

L'utilizzo delle distribuzioni permette di confrontare agevolmente collettivi o sottopopolazioni diverse.

titolo di studio	**M**	**F**	**M**	**F**
licenza elementare	3	1	0,33	0,09
licenza media	2	5	0,22	0,45
diploma	2	4	0,22	0,36
laurea	2	1	0,22	0,09
totale	9	11	1,00	1,00

Diagramma a nastri

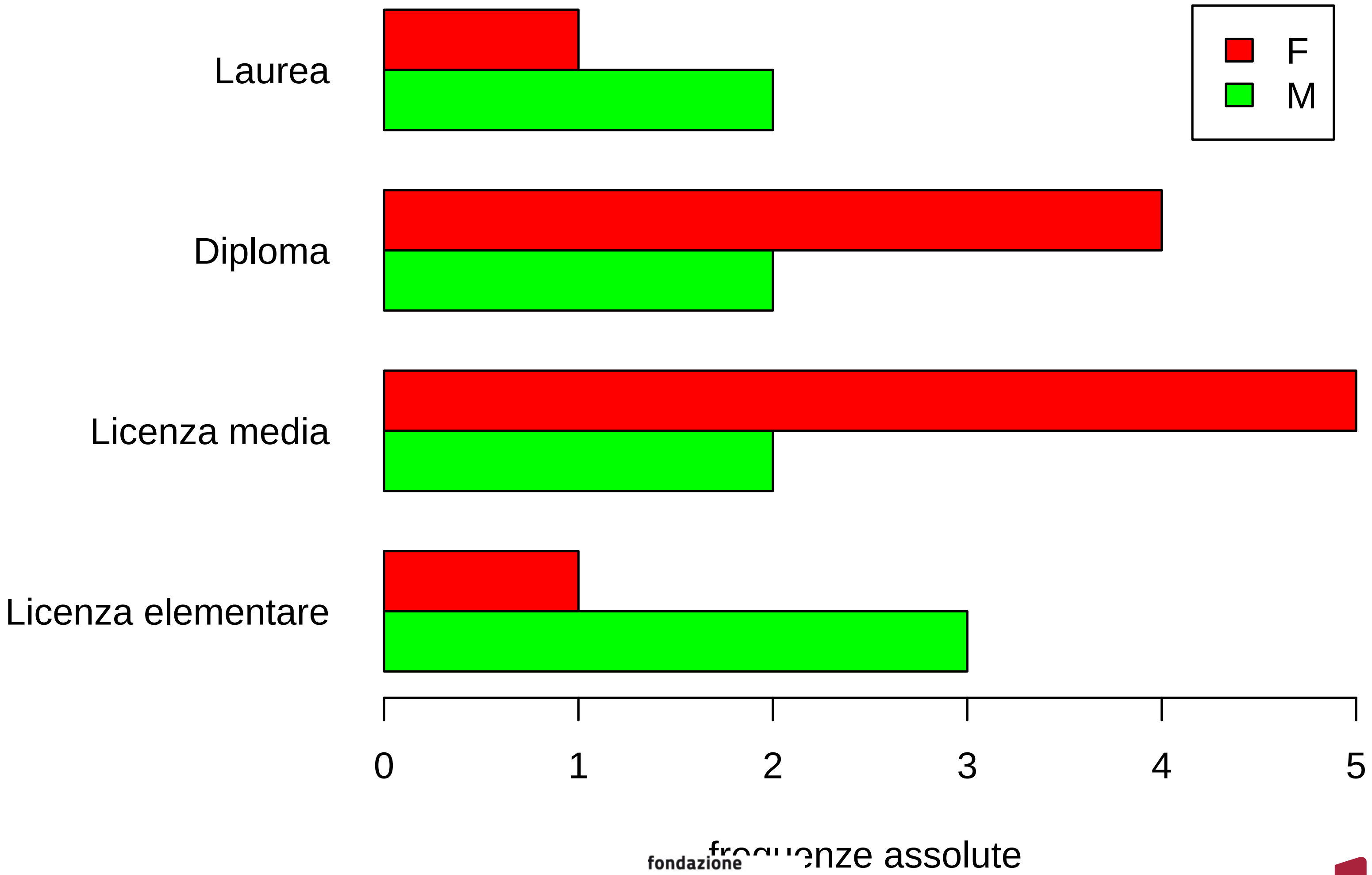
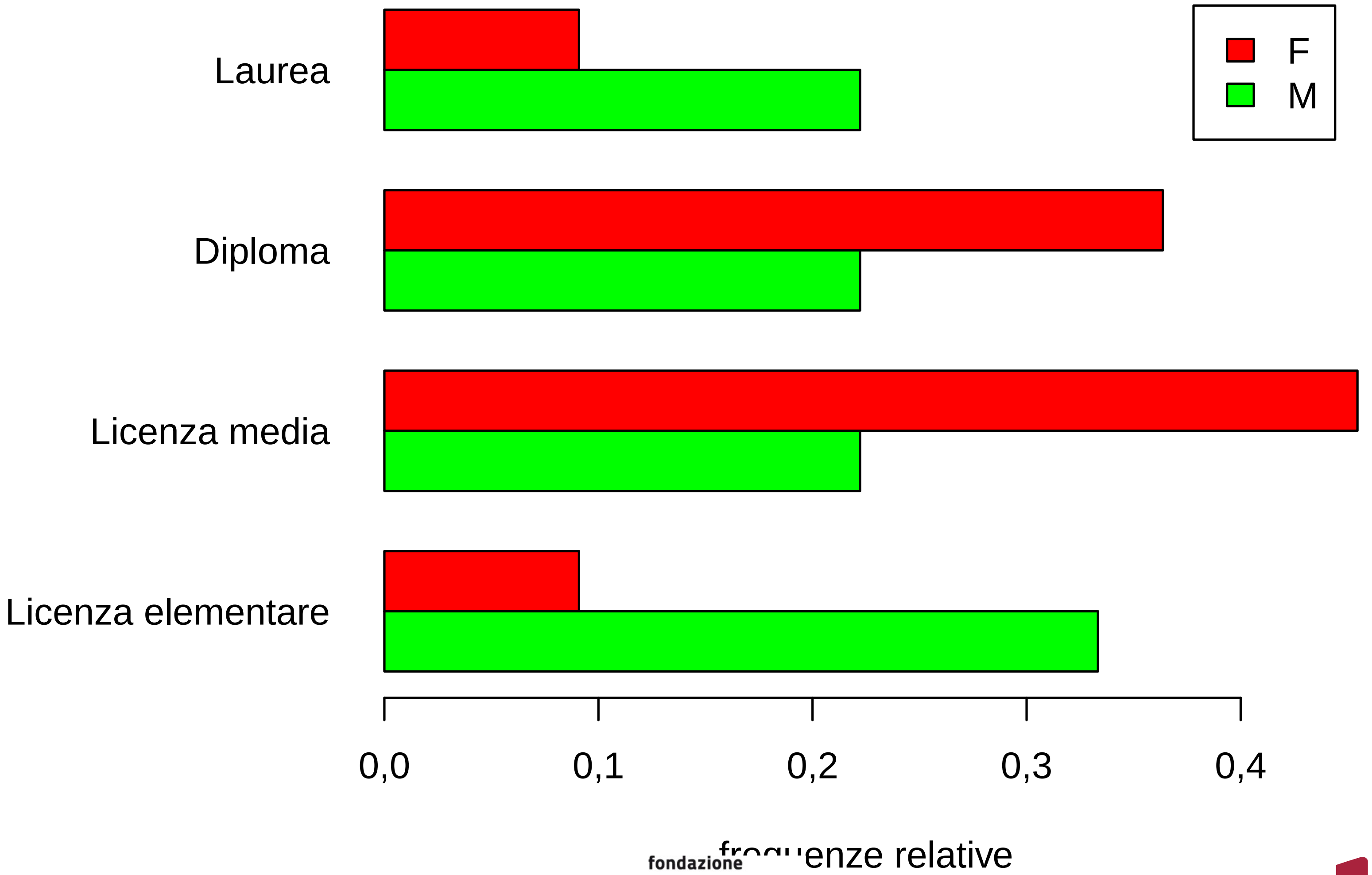


Diagramma a nastri



Due distribuzioni di frequenze relative, A e B, possono essere confrontate mediante un indice di dissomiglianza, variabile tra 0 (distribuzioni uguali) e 1 (massima differenziazione):

$$D = \frac{1}{2} \sum_{i=1}^{N} |f_{A,i} - f_{B,i}|$$

Indici di dissomiglianza

titolo di studio	$f_{M,i}$ M	$f_{F,i}$ F	$f_{M,i} - f_{F,i}$	$\lvert f_{M,i} - f_{F,i} \rvert$
licenza elementare	0,33	0,09	0,24	0,24
licenza media	0,22	0,45	-0,23	0,23
diploma	0,22	0,36	-0,14	0,14
laurea	0,22	0,09	0,13	0,13
totale	1,00	1,00	0	0,74

Nel caso della distribuzione di frequenza per titolo di studio la dissomiglianza tra maschi e femmine è pari a $0{,}74/2 = 0{,}37$.

Distribuzione di un carattere quantitativo

La distribuzione di frequenza può essere ugualmente calcolata per un carattere quantitativo. Ad esempio per il carattere statura abbiamo:

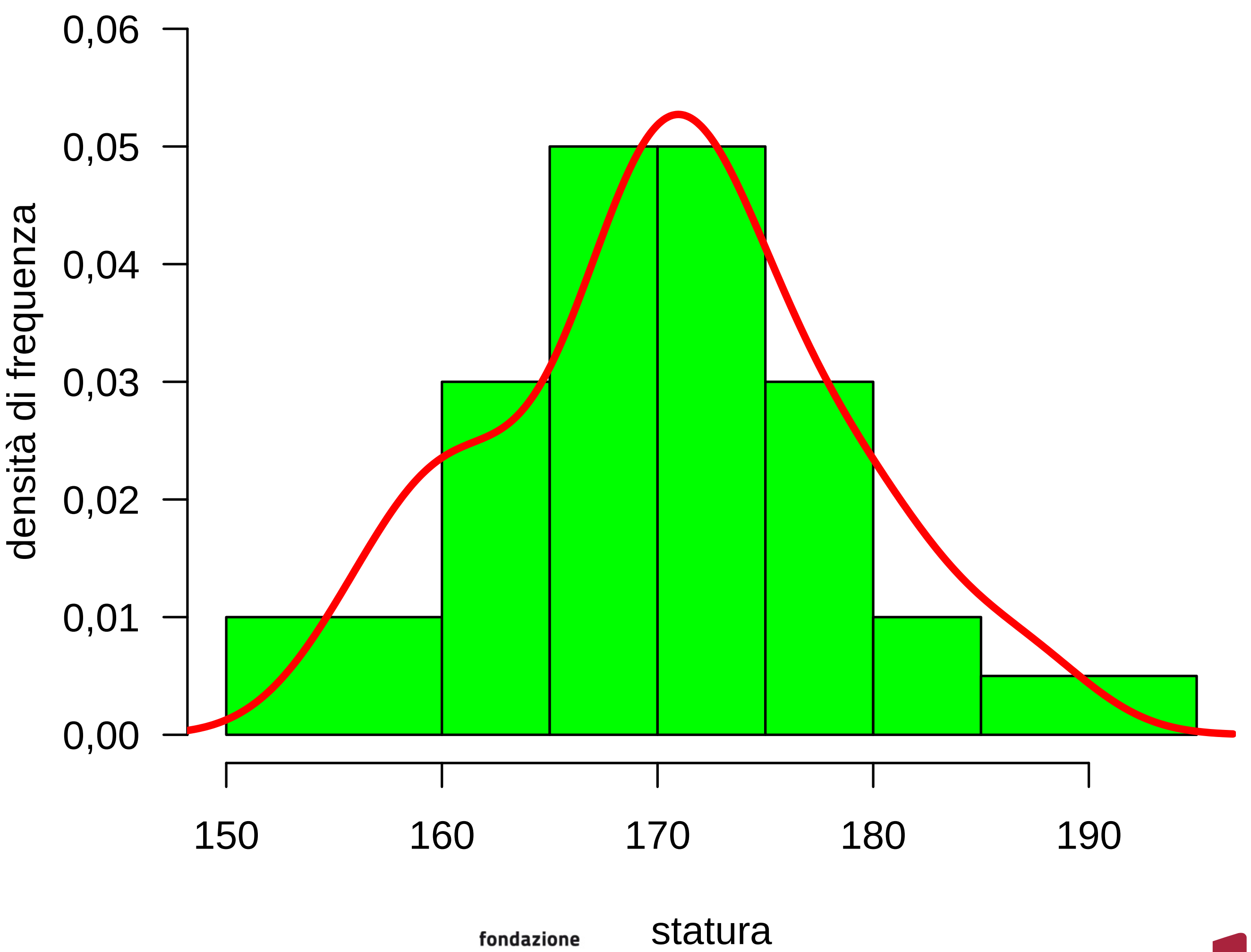
statura (cm)	n_i
156	1
159	1
161	2
...	...
169	1
170	2
...	...
187	1
totale	20

Distribuzione di un carattere quantitativo

Nel caso di un carattere quantitativo è talvolta necessario e spesso opportuno raggruppare i dati in classi. Ad esempio, nel caso precedente, un possibile raggruppamento è rappresentato nella seguente tabella:

statura (cm)	n_i
fino a 160	2
$160 \dashv 165$	3
$165 \dashv 170$	5
$170 \dashv 175$	5
$175 \dashv 180$	3
$180 \dashv 185$	1
oltre 185	1
totale	20

Istogramma


0,06
0,05
0,04
0,03
0,02
0,01
0,00
densità di frequenza
150
160
170
180
190
statura

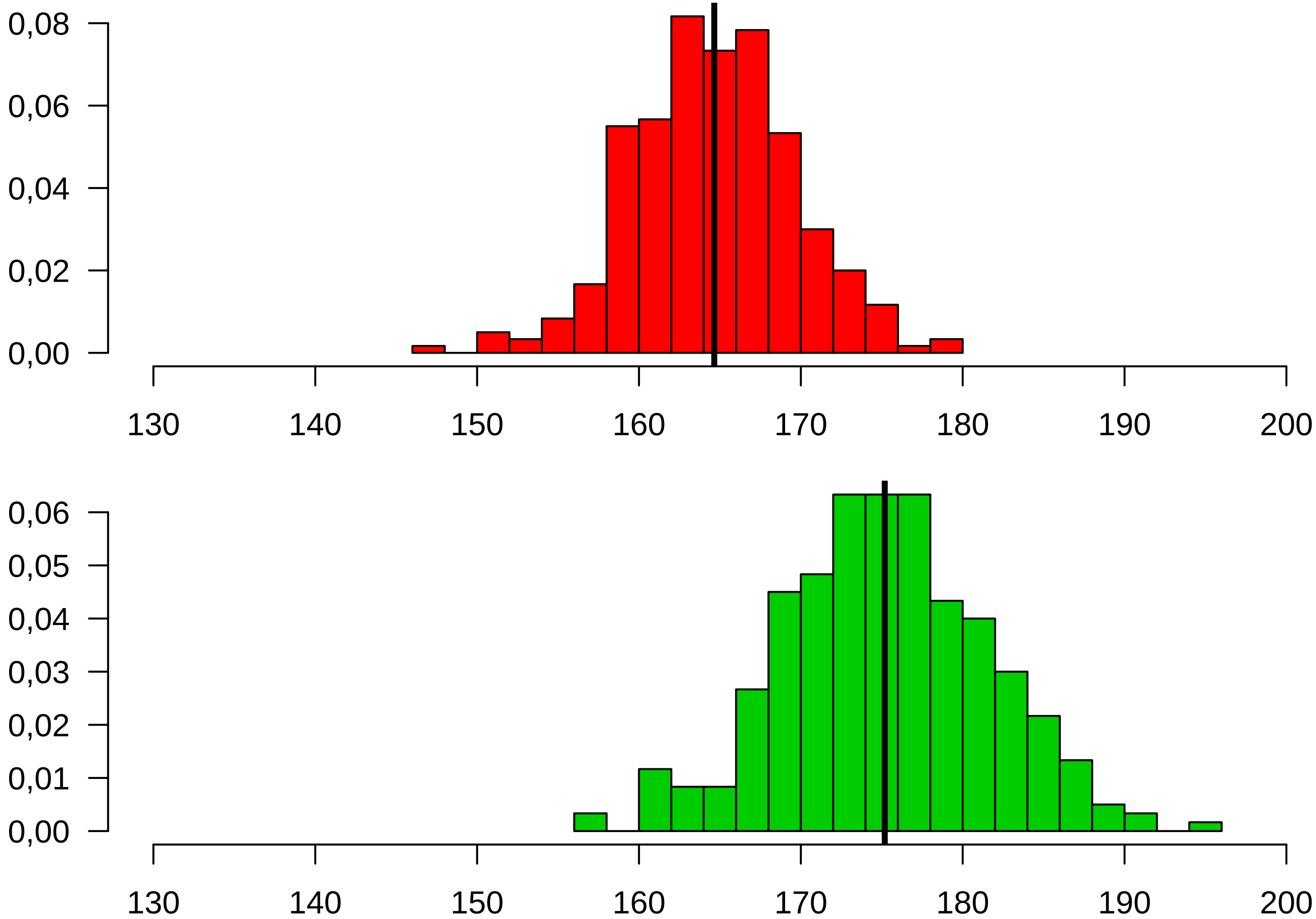
Indicatori sintetici

Per confrontare distribuzioni diverse si ricorre sovente a indicatori sintetici.

Solitamente si distinguono indicatori di:

- ✗ posizione
- ✗ variabilità
- ✗ forma

Indicatori di posizione centrale



Media aritmetica

L'indicatore di posizione centrale più conosciuto è sicuramente la media aritmetica:

$$\mu = \frac{x_1 n_1 + x_2 n_2 + \cdots + x_N n_N}{N} = \frac{\sum_{i=1}^{N} x_i n_i}{N}$$

La media aritmetica è quel valore che, sostituito al valore del carattere di ogni unità statistica, non muta l'ammontare totale del carattere.

Mediana

- ✗ Tuttavia, la media aritmetica può essere utilizzata solamente per caratteri quantitativi.
- ✗ Se vogliamo calcolare un indicatore di posizione per un carattere qualitativo ordinabile (ad esempio per la variabile “titolo di studio”), si può utilizzare la mediana, ossia il valore che bipartisce la distribuzione ordinata della variabile statistica.
- ✗ Nel caso di un numero dispari di osservazioni l’elemento è unico, altrimenti abbiamo due valori mediani.

Mediana

Per il calcolo è necessario ordinare la variabile e calcolare le frequenze cumulate.

titolo di studio	**frequenza (n_i)**	**frequenza cumulata ($\sum_i n_i$)**
licenza elementare	4	4
licenza media	7	11
diploma	6	17
laurea	3	20
totale	20	

Con 20 osservazioni abbiamo due valori mediani, corrispondenti alla decima e all'undicesima osservazione.

Dalla colonna delle frequenze cumulate è facile constatare come in entrambi i casi ciò corrisponda al valore "Licenza media"

Moda

- ✗ Nel caso di un carattere qualitativo sconnesso, neanche la mediana è utilizzabile.
- ✗ In tal caso si ricorre alla moda, ossia il valore che presenta la frequenza relativa più elevata.
- ✗ Ad esempio, nel caso della distribuzione per sesso, il valore modale è "femmina".

sesso	n_i	f_i
maschi	9	0,45
femmine	11	0,55
totale	20	1,00

Uso delle medie

tipo di variabile	moda	mediana	media aritmetica
qualitativa sconnessa	✓		
qualitativa ordinale	✓	✓	
quantitativa	✓	✓	✓

Utilizzo delle medie

- ✗ L’utilizzo di una media, operando una sintesi, implica sempre una perdita di informazione.
- ✗ Normalmente è preferibile adottare una media più adeguata alla modalità di misurazione utilizzabile per una certa variabile.
- ✗ Quindi, per variabili qualitative ordinabili è preferibile utilizzare la mediana anziché la moda, mentre per variabili quantitative è opportuno utilizzare la media aritmetica.
- ✗ A volte si ricorre alla mediana anche nel caso di dati quantitativi a causa della sua possibile maggiore “robustezza” rispetto a valori eccezionali o anomali.

Esempio

✗ Consideriamo la seguente variabile statistica:

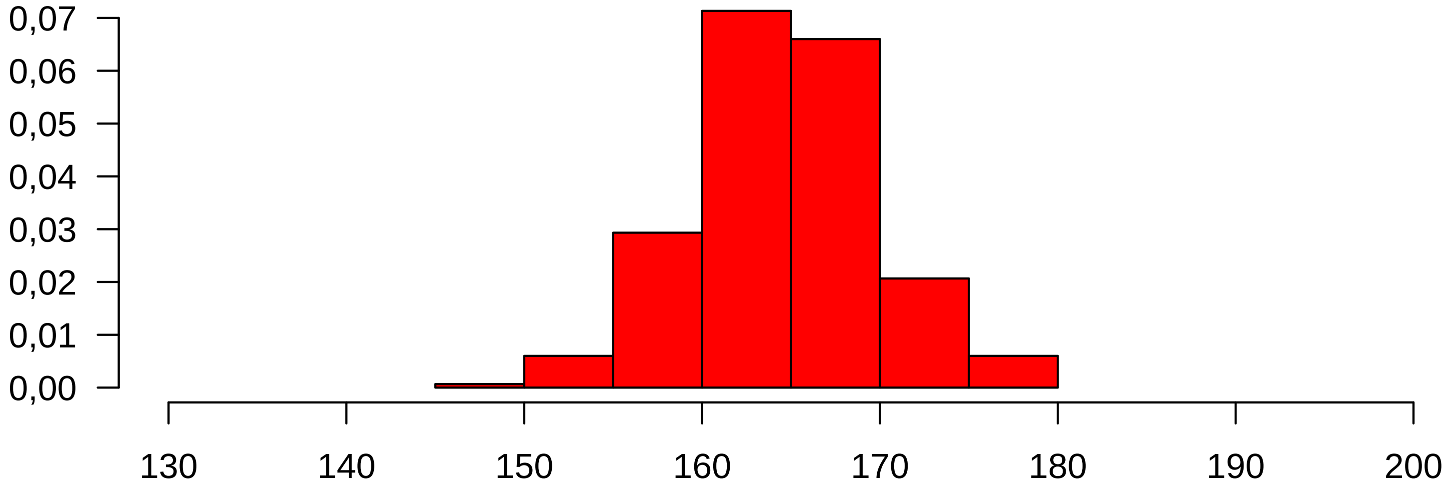
$$180, 173, 192, 160, 165$$

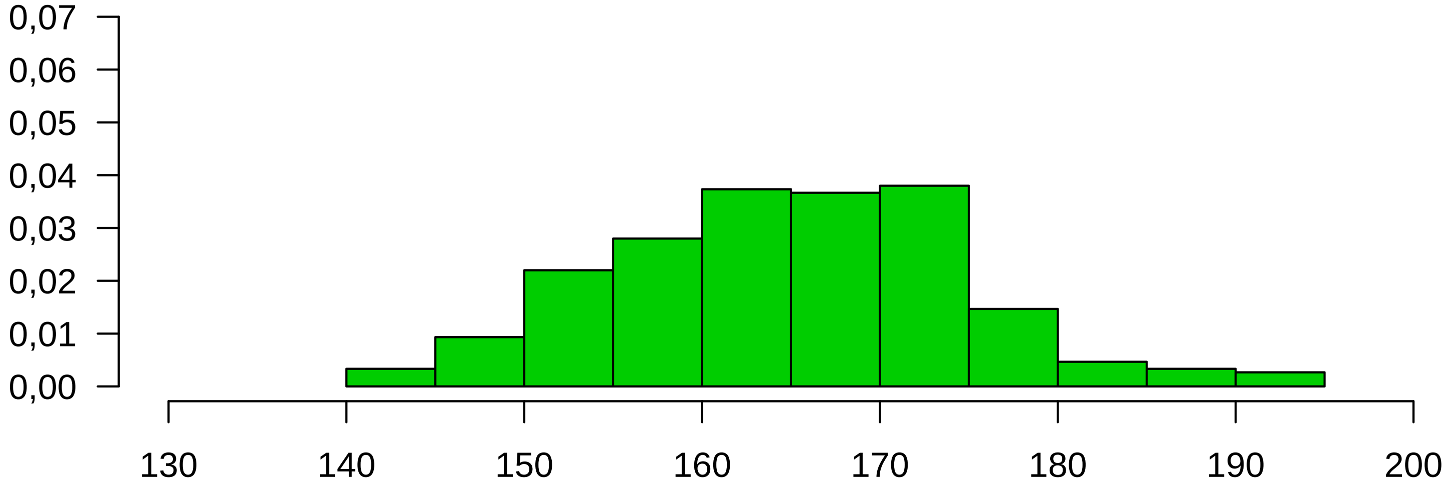
La media aritmetica è pari a 174; la mediana ha un valore assai prossimo, pari a 173.

✗ Supponiamo che, per un errore, il valore 160 sia stato registrato come 16. Mentre la mediana non muta, la media aritmetica diventa 145,2 !

- ✗ L'uso della media per sintetizzare una distribuzione è particolarmente utile quando si vogliono confrontare due o più distribuzioni che abbiano approssimativamente la stessa forma.
- ✗ Tuttavia, nei casi concreti, possiamo avere il caso di popolazioni che, pur presentando la stessa media, hanno un grado di addensamento diverso intorno a quest'ultima, ossia una diversa variabilità.
- ✗ In questi casi è bene presentare sia un indicatore sintetico della posizione centrale, sia un indicatore di variabilità.

Variabilità





✗ L'indicatore di variabilità più semplice è il campo di variazione, che è pari alla differenza tra il valore massimo e minimo della variabile:

$$\omega = x_{max} - x_{min}.$$

La misura precedente è molto grossolana e risente significativamente della presenza di valori anomali.

Variabilità

- ✗ Un modo per misurare la variabilità per dati quantitativi è quello di calcolare una misura sintetica di una funzione delle distanze dei valori della variabile dalla propria media.
- ✗ Una misura assai utilizzata è la varianza, ossia la media dei quadrati delle distanze tra i valori della variabile e la sua media.

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

- ✗ Per avere un indicatore espresso nella stessa unità di misura del fenomeno di interesse si utilizza lo scarto quadratico medio, ossia la radice positiva della varianza:

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

Variabilità

- ✗ Le misure di variabilità illustrate sono misure assolute, ovvero dipendono dall'unità di misura del fenomeno considerato.
- ✗ Per ottenere dei cosiddetti numeri puri, al fine di confrontare la variabilità di distribuzioni di tipo molto diverso, è opportuno ricorrere a indici relativi di variabilità.
- ✗ Nel caso di intensità positive, si può ricorrere al coefficiente di variazione, rappresentato dal rapporto tra lo scarto quadratico medio e la media aritmetica:

$$CV = \frac{\sigma}{\mu}$$

Questo indicatore consente, ad esempio, di confrontare la variabilità di distribuzioni di stature espresse in pollici e centimetri, o distribuzioni di redditi espressi in monete diverse, senza necessità di convertire le misurazioni in una misura comune.

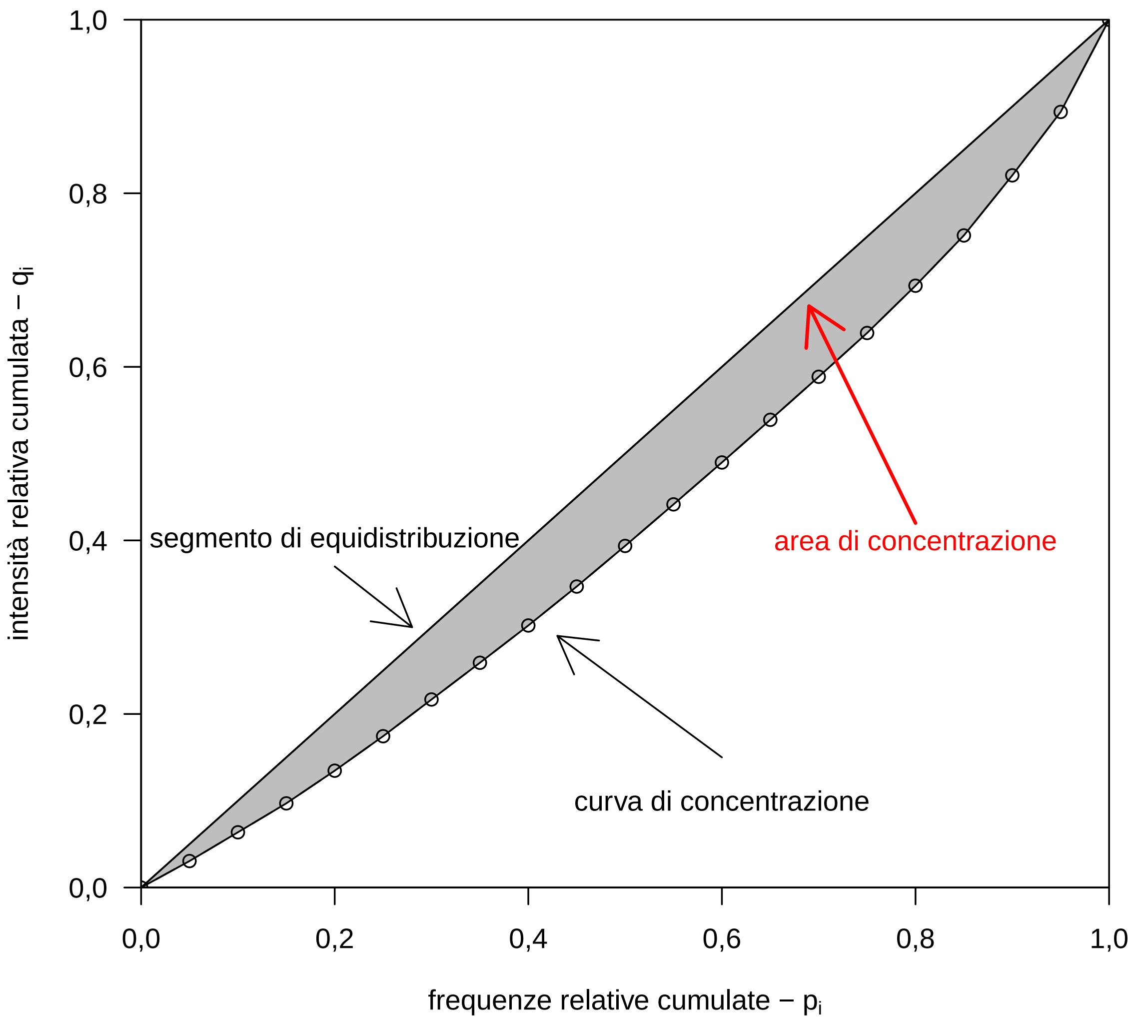
Concentrazione

Un particolare tipo di misure di variabilità relative riguarda la cosiddetta concentrazione. Si parla di concentrazione in particolare per caratteri quantitativi trasferibili, come il reddito.

A questo proposito è necessario ordinare in senso crescente la variabile di interesse, e costruire la sua distribuzione di frequenze e di intensità cumulata:

intensità	**frequenze**	**intensità cumulata**	**frequenze cumulate**	**intensità relativa cumulata**	**frequenze relative cumulate**
x_i	n_i	$\sum x_i$	$\sum n_i$	q_i	p_i
14.300	1	14.300	1	0,03	0,05
15.500	1	29.800	2	0,06	0,10
15.600	1	45.400	3	0,10	0,15
		...			
34.200	1	418.200	19	0,89	0,95
49.700	1	467.900	20	1,00	1,00

Concentrazione



✗ La concentrazione è tanto più elevata quanto maggiori sono le differenze $p_i - q_i$. Infatti, nel caso di equidistribuzione, per ogni i, $p_i = q_i$. Al contrario, nel caso di massima concentrazione, le q_i sono tutte nulle, salvo $q_N = 1$.

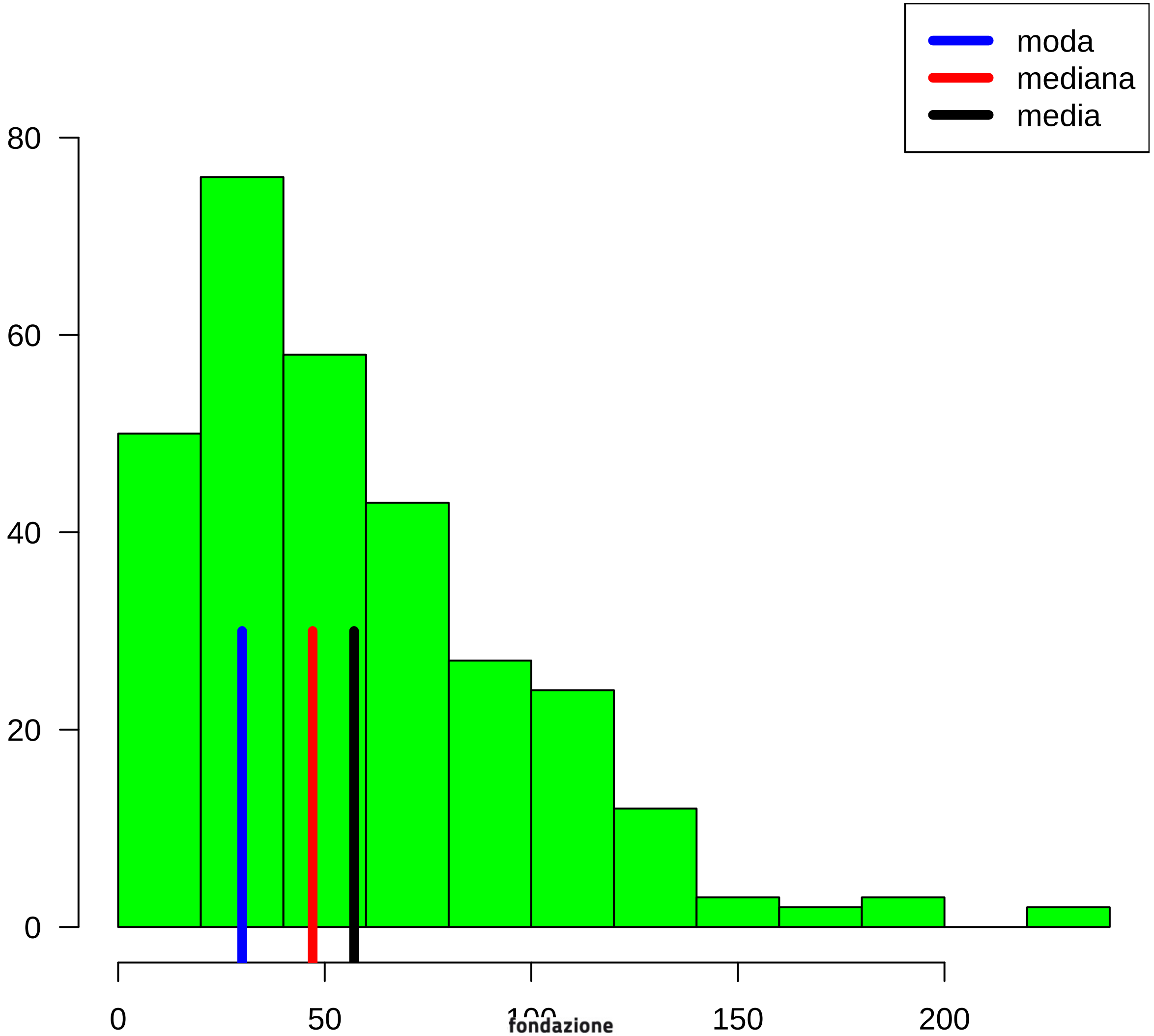
✗ Un indicatore di concentrazione che varia tra 0 (equidistribuzione) e 1 (massima concentrazione) è il rapporto di concentrazione di Gini:

$$R = \frac{\sum_{i=1}^{N-1}(p_i - q_i)}{\sum_{i=1}^{N-1} p_i}$$

Asimmetria

- ✗ Un ultimo gruppo di parametri di forma riguarda la possibile asimmetria della distribuzione.
- ✗ In particolare, nel caso di una distribuzione simmetrica unimodale, moda, media e mediana coincidono. Ciò non avviene per una distribuzione asimmetrica.
- ✗ Nel caso di una distribuzione unimodale asimmetrica a destra, abbiamo sempre che $Mo > Me > \mu$. Il contrario accade per una distribuzione unimodale asimmetrica a sinistra.

Asimmetria



Distribuzioni statistiche multiple

- ✗ Finora ci siamo concentrati prevalentemente su distribuzioni semplici di una variabile statistica.
- ✗ D'altra parte, spesso vi è interesse ad analizzare congiuntamente la distribuzione secondo più caratteri. In tal caso si considerano le distribuzioni statistiche multiple.
- ✗ Nel seguito, per semplicità di esposizione, faremo riferimento esclusivamente alle distribuzioni doppie.

Distribuzioni statistiche doppie

In questo caso è possibile continuare a utilizzare un'estensione della rappresentazione vista in precedenza. La seguente tabella contiene la distribuzione per peso e altezza del nostro collettivo, con dati raggruppati in classi:

Statura (cm)	Peso (kg)			
	fino a 60	60 $\dashv$ 75	oltre 75	totale
fino a 165	4	1	0	5
165 $\dashv$ 175	6	3	1	10
oltre 175	0	1	4	5
totale	10	5	5	20

Analogamente al caso delle distribuzioni semplici possiamo calcolare le frequenze relative $f_{ij} = n_{ij}/N$:

Statura (cm)	Peso (kg)			
	fino a 60	60 ⊣ 75	oltre 75	totale
fino a 165	0,20	0,05	0	0,25
da 165 ⊣ 175	0,30	0,15	0,05	0,50
oltre 175	0	0,05	0,20	0,25
totale	0,50	0,25	0,25	1,00

Nella tabella sono presenti la distribuzione delle frequenze relative congiunte di peso e statura e le distribuzioni marginali delle due variabili:

Statura (cm)	Peso (kg)			
	fino a 60	60 $\dashv$ 75	oltre 75	totale
fino a 165	0,20	0,05	0	0,25
165 $\dashv$ 175	0,30	0,15	0,05	0,50
oltre 175	0	0,05	0,20	0,25
totale	0,50	0,25	0,25	1,00

Distribuzioni condizionate

Inoltre, possiamo calcolare le distribuzioni condizionate. Ad esempio, le distribuzioni condizionate della statura dato un certo valore del peso:

Statura (cm)	Peso (kg)			
	fino a 60	60 ⊣ 75	oltre 75	totale
fino a 165	0,40	0,20	0	0,25
165 ⊣ 175	0,60	0,60	0,20	0,50
oltre 175	0	0,20	0,80	0,25
totale	1,00	1,00	1,00	1,00

- La distribuzioni statistiche multiple vengono naturalmente utilizzate per analizzare le relazioni tra le variabili. In particolare si è sovente interessati all'indipendenza di due o più variabili.

X \ Y	y_1	y_2	y_3	y_4	totale
x_1	n_{11}	n_{12}	n_{13}	n_{14}	$n_{1.}$
x_2	n_{21}	n_{22}	n_{23}	n_{24}	$n_{2.}$
x_1	n_{31}	n_{32}	n_{33}	n_{34}	$n_{3.}$
totale	$n_{.1}$	$n_{.2}$	$n_{.3}$	$n_{.4}$	$n_{..}$

- Si ha indipendenza tra X e Y se ogni frequenza della distribuzione congiunta è data da:

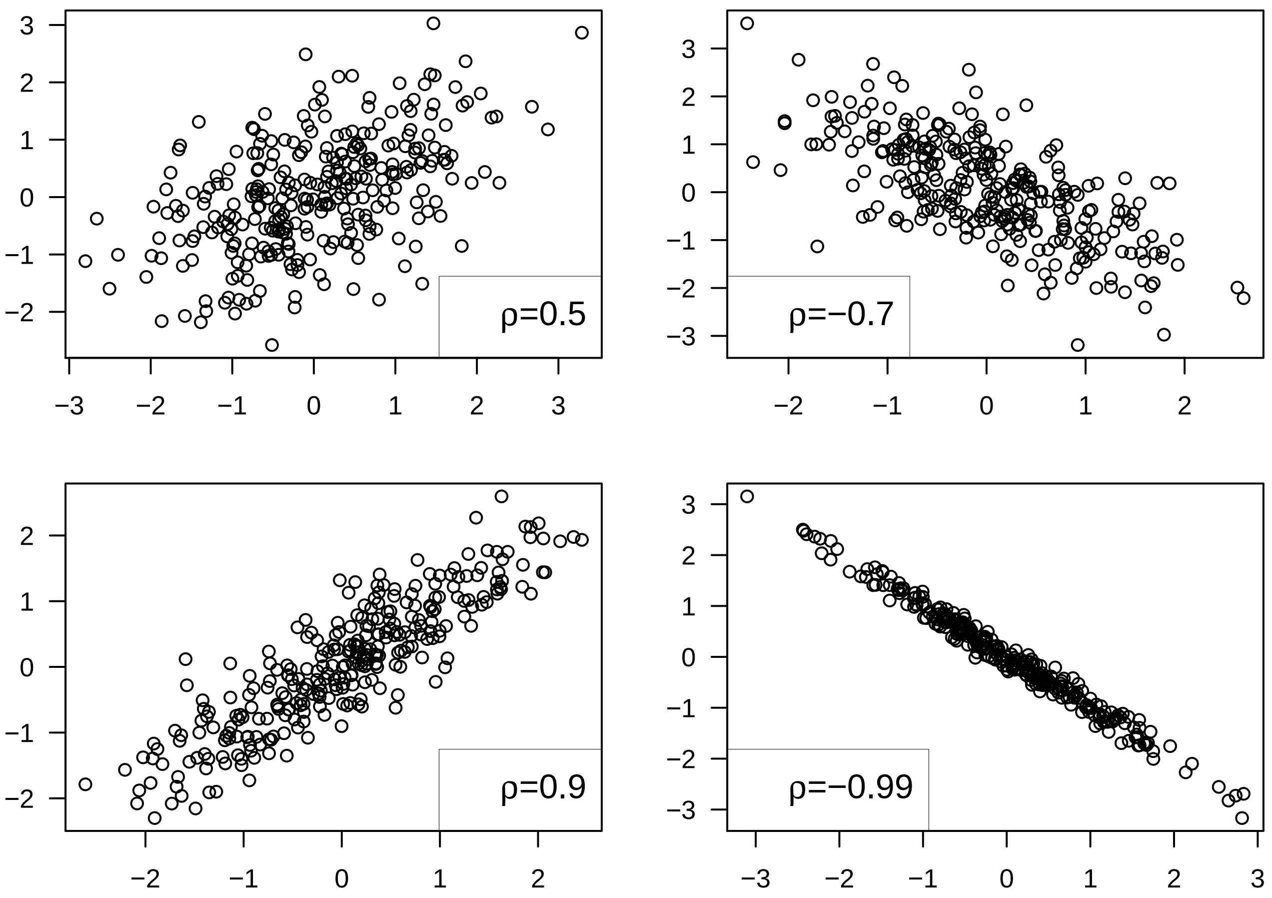
$$n_{ij} = \frac{n_{i.}n_{.j}}{n_{..}}$$

- Quanto più la distribuzione osservata si allontana dalla situazione suddetta, tanto più si è lontani dalla situazione di indipendenza.

Correlazione lineare

- ✗ Un concetto molto importante quando si parla di relazioni tra variabili quantitative è quello di correlazione lineare, ossia del legame lineare tra due variabili.
- ✗ Questo può essere misurato dal coefficiente di correlazione, che assume valori compresi tra -1 e 1.
- ✗ In particolare, il legame è tanto più forte quanto più il valore del coefficiente è prossimo a 1 o -1. Il segno indica se il legame è diretto (al crescere di una variabile cresce anche l'altra) o inverso.
- ✗ Attenzione: la correlazione non ha implicazioni in termini di causalità!

Correlazione lineare



Inferenza

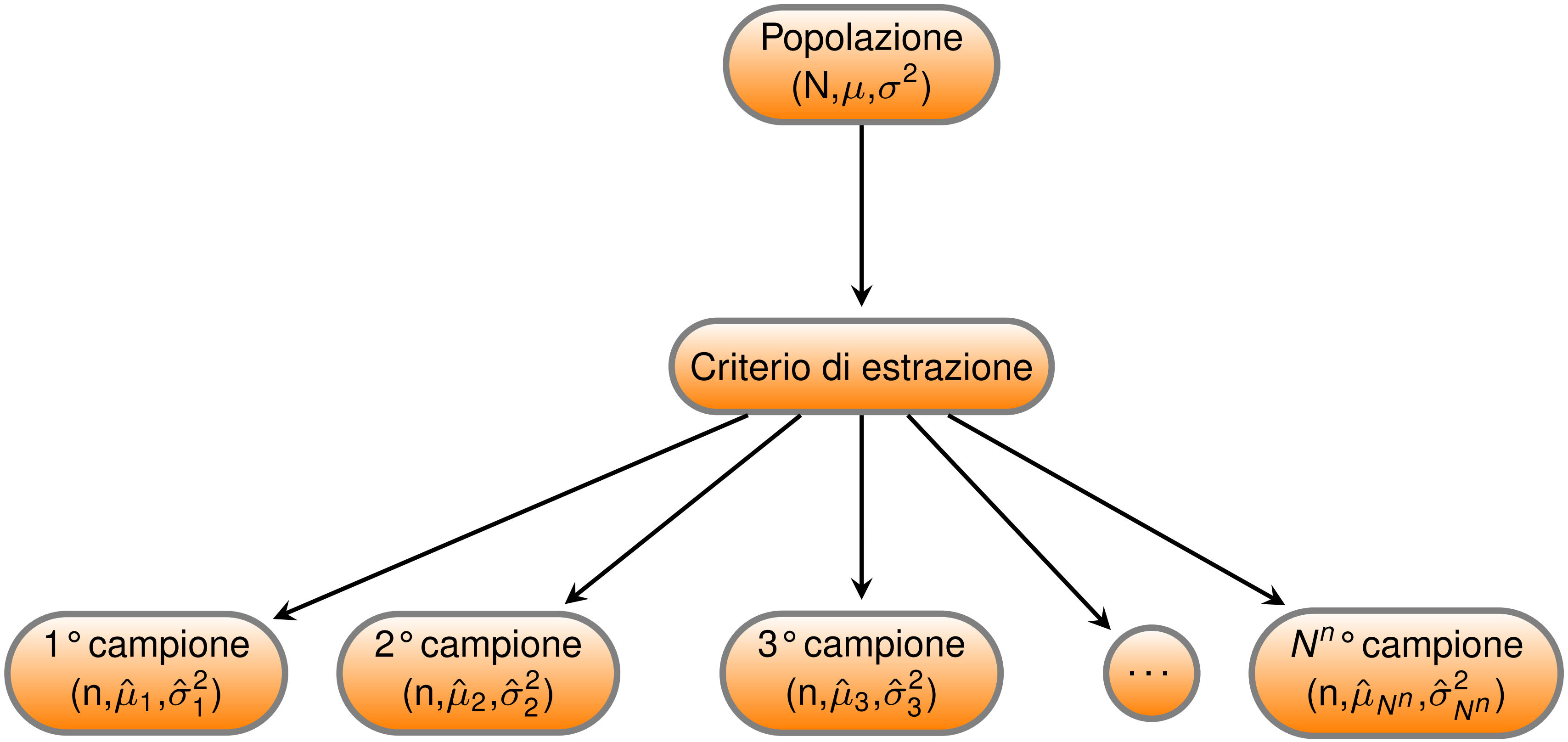
- ✗ Finora abbiamo calcolato delle statistiche descrittive su un certo collettivo. Il problema di calcolare delle statistiche su una popolazione partendo solamente dall'osservazione di una parte di essa (campione) è definito inferenza statistica.
- ✗ L'inferenza studia le condizioni sotto le quali le statistiche calcolate su un campione sono generalizzabili all'intera popolazione di interesse.
- ✗ Ovviamente le statistiche calcolate su un campione non saranno (quasi) mai uguali a quelle calcolabili sulla popolazione; è infatti assai improbabile che il campione riproduca, in piccolo, tutte ed esattamente le caratteristiche della popolazione di partenza.
- ✗ Quello che possiamo sperare è che il procedimento di estrazione del campione non introduca elementi di distorsione sistematica rispetto alla misurazione del carattere di interesse presente nella popolazione.

- ✗ Uno schema di sorteggio è quello dell'estrazione casuale semplice (con o senza ripetizione). Data una popolazione, di numerosità N, di media μ e scarto quadratico medio σ, procediamo all'estrazione casuale di un campione di ampiezza $n < N$.
- ✗ Questo procedimento è assimilabile concettualmente all'estrazione di palline da un'urna, dove ogni pallina rappresenta un elemento della popolazione.
- ✗ Il campionamento può essere con ripetizione, ossia la pallina, una volta estratta viene reimmessa nell'urna; quindi, nel campione la stessa unità può osservarsi più volte. Oppure, al contrario, senza ripetizione.
- ✗ Il campionamento casuale semplice è quello che assegna a ogni unità la stessa probabilità di essere estratta.

Esempio

- ✗ Immaginiamo di avere una popolazione, di numerosità N, di media μ e scarto quadratico medio σ. In particolare consideriamo una popolazione di 6 individui nella quale siamo interessati al carattere reddito, che assume i valori (10,20,30,40,50,60).
- ✗ Determiniamo una numerosità campionaria n ed estraiamo un campione casuale di tale ampiezza.
- ✗ Dal campione potremo calcolare delle stime di μ e di σ^2, che definiremo $\hat{\mu}$ e $\hat{\sigma}^2$.
- ✗ Le domande immediate sono: quanto sono attendibili tali stime? Come possiamo avere un'idea della loro precisione?
- ✗ Per rispondere in maniera intuitiva a queste domande faremo un esperimento concettuale, supponendo di estrarre tutti i possibili campioni.

Campionamento

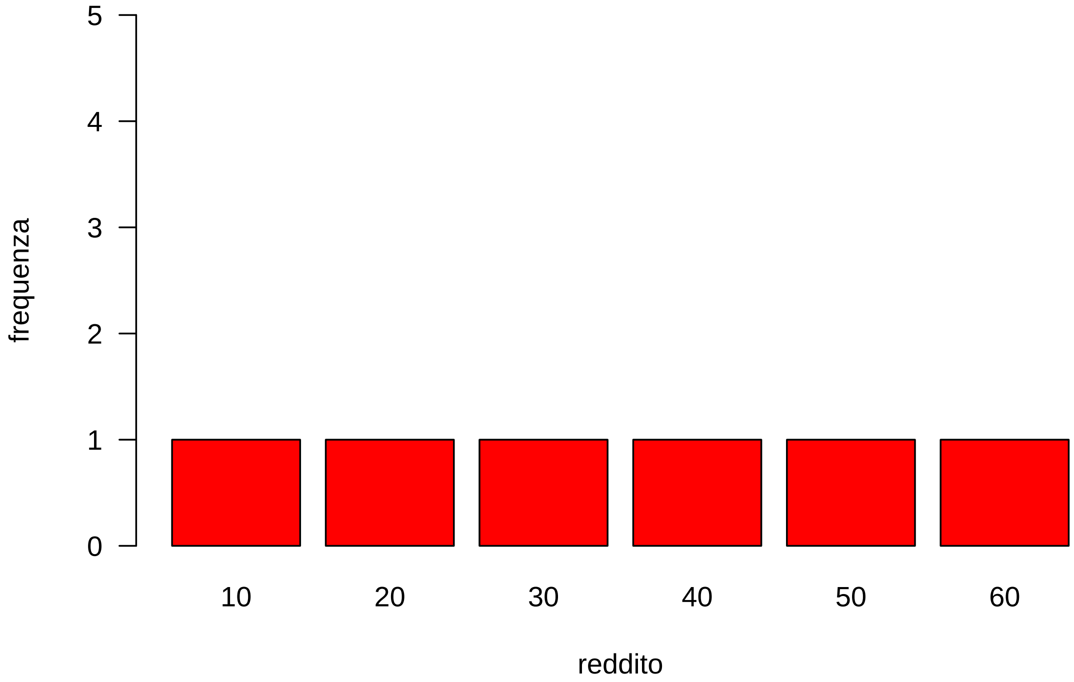


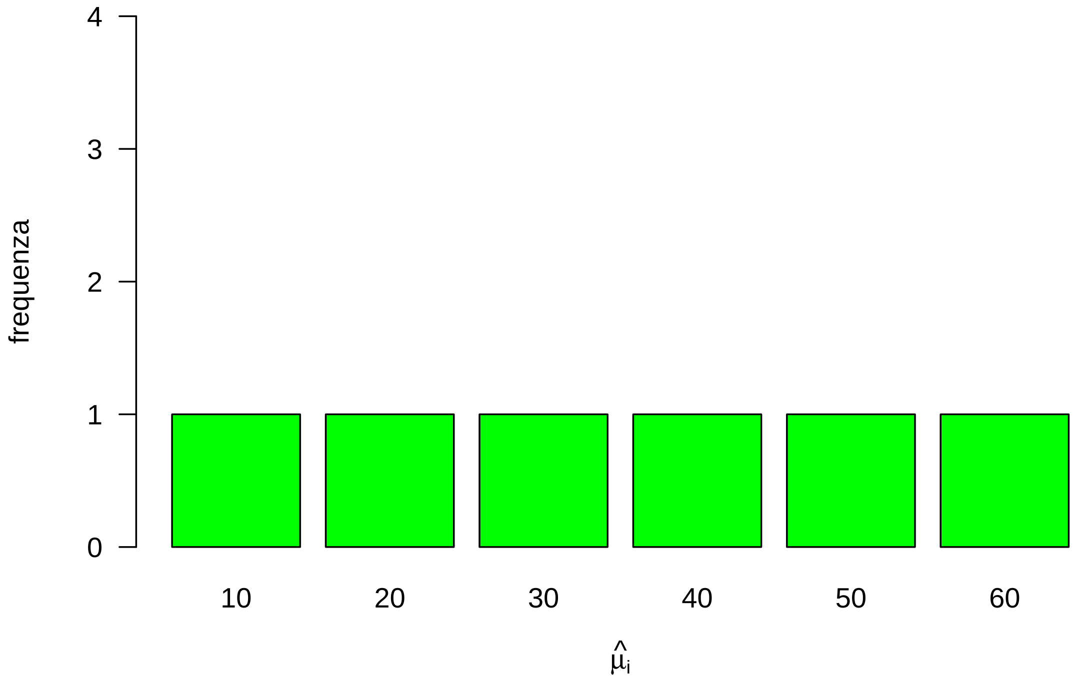
Distribuzione della popolazione

reddito (x_i)	frequenza (n_i)
10	1
20	1
30	1
40	1
50	1
60	1
totale	6

La media aritmetica della popolazione è pari a 35 e lo scarto quadratico medio è 17,08.

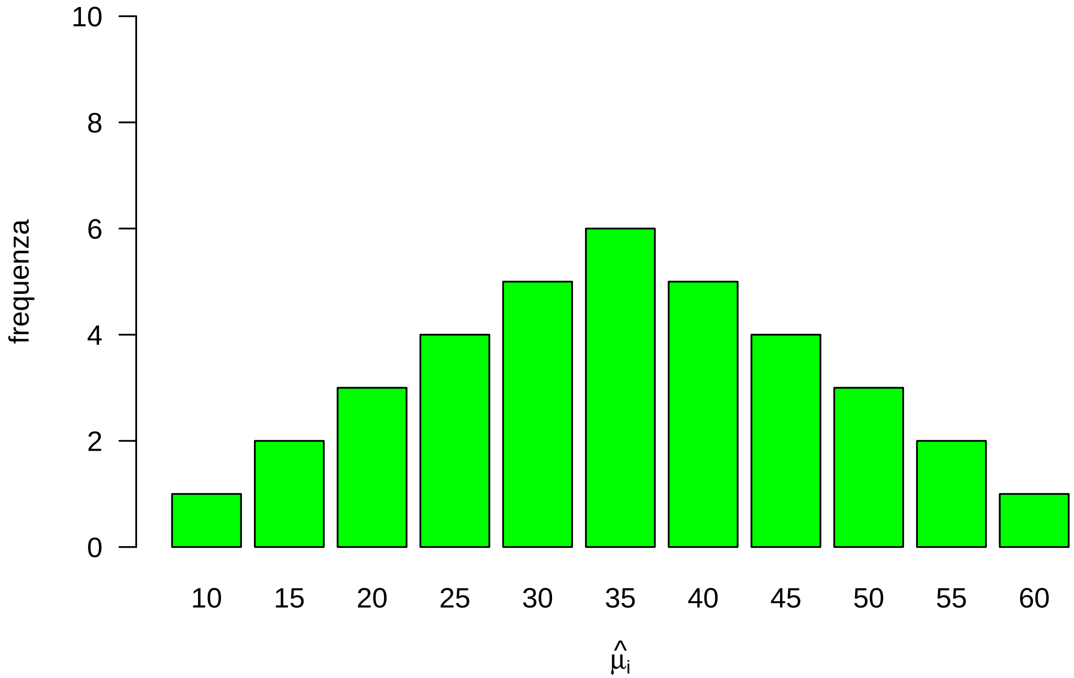
Distribuzione della popolazione





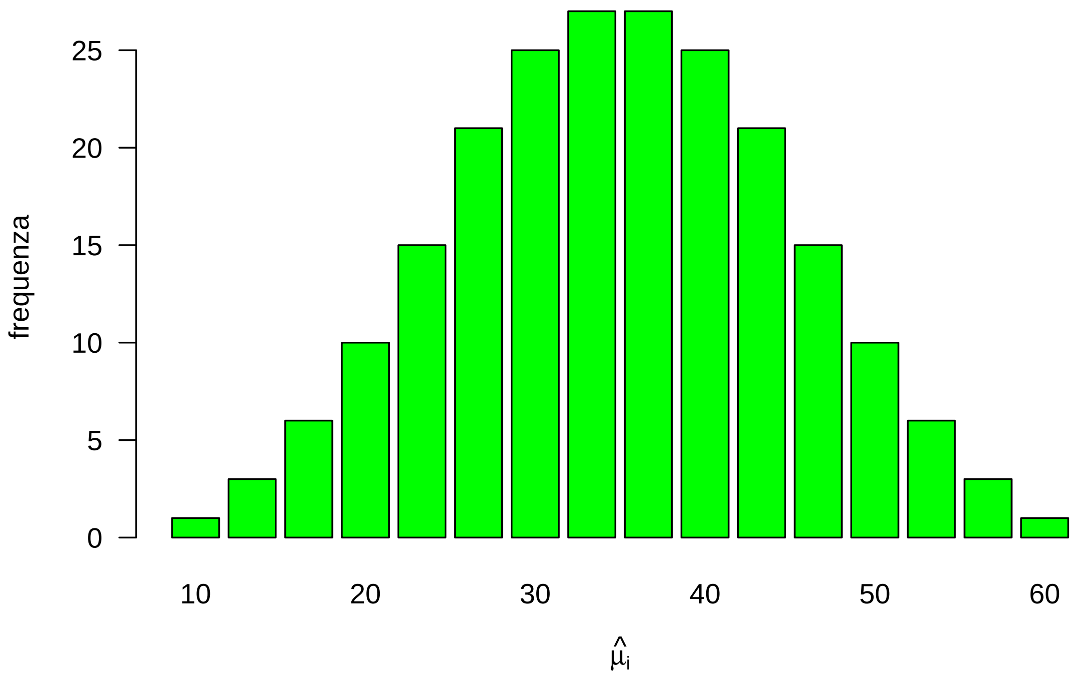
Considerando campioni di numerosità 1, è possibile estrarne $6^1 = 6$. Di essi 2 (pari al 33%) presentano una media compresa tra 30 e 40.

Distribuzione della media campionaria



Considerando campioni di numerosità 2, è possibile estrarne $6^2 = 36$. Di essi 16 (pari al 44%) presentano una media compresa tra 30 e 40.

Distribuzione della media campionaria



Considerando campioni di numerosità 3, è possibile estrarne $6^3 = 216$. Di essi 104 (pari al 48%) presentano una media compresa tra 30 e 40.

Osservazioni

- ✗ Nonostante la popolazione non sia unimodale, la distribuzione della media campionaria lo è, e la sua moda tende a coincidere con il valore della media della popolazione.
- ✗ All'aumentare della numerosità del campione questo comportamento è più evidente; inoltre, la distribuzione delle medie campionarie è più concentrata intorno al valore della media della popolazione.

Intervallo di fiducia

- ✗ Nella pratica, naturalmente, viene estratto un solo campione; tuttavia, con opportune ipotesi, possiamo avere un'idea della precisione della stima della media campionaria ottenuta.
- ✗ Ciò avviene grazie all'utilizzazione dell'intervallo di fiducia, spesso fornito insieme alla stima puntuale del parametro (media).
- ✗ Esempio: la stima della media della statura di una popolazione derivante da un campione è pari a 175 cm, e l'intervallo di fiducia al 95% è pari a 168 – 182 cm.

Intervallo di fiducia

- ✗ Così come abbiamo ipotizzato di calcolare tutte le possibili medie campionarie, possiamo ipotizzare di calcolare tutti i possibili intervalli di fiducia. L'affermazione dell'esempio precedente significa che il 95% degli intervalli calcolabili con la stessa procedura conterebbe il valore vero, ma incognito, della media della popolazione.
- ✗ Cosa fare se vogliamo diminuire l'ampiezza dell'intervallo, oppure se desideriamo un livello di fiducia maggiore (ad esempio 99%)?
- ✗ Tra i due aspetti c'è un *trade-off*: aumentare il livello di fiducia implica aumentare l'ampiezza dell'intervallo e viceversa. L'unico modo per raggiungere entrambi gli obiettivi consiste nell'aumentare la numerosità del campione.

Conclusione

L'incontro di oggi ci ha permesso di:

- ✗ definire cos'è una variabile statistica semplice e la sua distribuzione;
- ✗ illustrare alcuni indicatori sintetici di posizione e variabilità di una distribuzione;
- ✗ introdurre le distribuzioni multiple e la correlazione lineare;
- ✗ accennare al processo di inferenza statistica.

Grazie dell'attenzione