Università di Firenze
Corso di laurea in Statistica
A.A. 2010/2011

STATISTICA
parte I, A

Carla Rampichini
rampichini@ds.unifi.it

Leonardo Grilli
grilli@ds.unifi.it

http://www.ds.unifi.it/rampichini/statistica2010_11.htm

Programma del corso

- PARTE 1 – Statistica descrittiva univariata
  - Nozioni di base
  - Distribuzioni, rappresentazioni grafiche
  - Medie
  - Indici di variabilità, indici di forma
- PARTE 2 – Statistica descrittiva bivariata
  - Distribuzioni doppie
  - Connessione, dipendenza in media
  - Correlazione
  - Regressione

Programma del corso

- PARTE 3 – Calcolo delle probabilità
  - Introduzione alla probabilità
  - Variabili casuali discrete
  - Variabili casuali continue
- PARTE 4 – Inferenza statistica
  - Distribuzioni campionarie
  - Stima puntuale
  - Stima per intervallo
  - Verifica delle ipotesi

Libri

Libro di testo:

- Cicchitelli G. (2008) ***Statistica. Principi e Metodi***. Pearson.

Libri di utile consultazione:

- Borra S., Di Ciaccio A. (2008) ***Statistica. Metodologie per le scienze economiche e sociali***, Secoda edizione, McGraw-Hill.
- Moore D.S. (2005) ***Statistica di base***. Apogeo.
- Newbold P., Carlson W.L., Thorne B. (2007) **Statistica**. Pearson / Prentice Hall.

Quali prospettive di lavoro?

- Alta probabilità di ottenere un lavoro di qualità
  - Vedi Società Italiana di Statistica www.sis-statistica.it/ alla voce "Didattica della Statistica"
  - Negli USA CareerCast ha valutato la professione di Statistico come la terza migliore www.careercast.com/jobs/content/JobsRated_10BestJobs

I keep saying the sexy job in the next ten years will be statisticians. The ability to take data—to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it—that's going to be a hugely important skill in the next decades, not only at the professional level but even at the educational level for elementary school kids, for high school kids, for college kids. Because now we really do have essentially free and ubiquitous data. So the complimentary scarce factor is the ability to understand that data and extract value from it.

Hal Varian
Professor of information sciences, business, and economics at the University of California at Berkeley and Google's chief economist
McKinseyQuarterly, January 2009

Nozioni di base

Origini della Statistica

- Il termine "statistica" deriva da "stato": all'inizio la statistica riguardava la raccolta di dati relativi allo stato (numerosità della popolazione, numero di cannoni, quantità di raccolto di grano …)
- La formalizzazione matematica della statistica è recente
  - XVIII e XIX secolo: calcolo delle probabilità
  - prima metà del XX secolo: inferenza statistica, disegno degli esperimenti, campionamento statistico
  - anni 40 - anni 70: sviluppi teorici
  - dagli anni 70: sviluppi legati alle capacità di calcolo dei computer

`http://it.wikipedia.org/wiki/Storia_della_statistica`

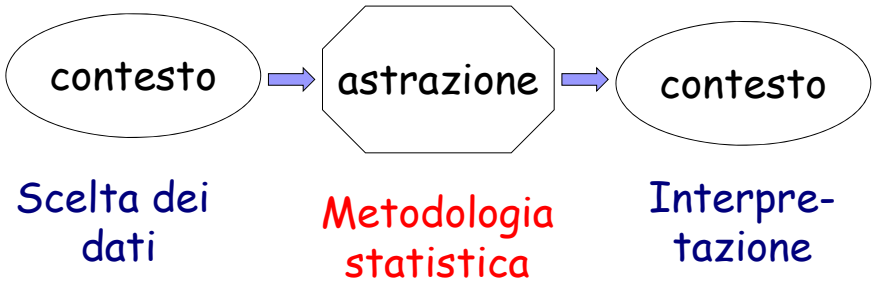
Cos'è la Statistica?

- *Statistica: l'arte e la scienza di imparare dai dati*
- Esistono molte definizioni formali, gli ingredienti essenziali sono i **dati** e l'uso di **strumenti di analisi di tipo quantitativo**

"E' facile mentire con la statistica, ma è difficile dire la verità senza di essa" (Andrejs Dunkels).

Cfr. D. Huff (1954) *How to lie with statistics*, recentemente tradotto in italiano (*Come mentire con la statistica*)

Cos'è la Statistica?



Apprendimento e valutazione del metodo prescindono dal tipo di applicazione

Statistica e matematica

- La statistica è una scienza quantitativa, ma il modo di pensare 'statistico' è diverso da quello 'matematico' per almeno 2 aspetti
  - la statistica non può prescindere dal **contesto** (dati)
  - la logica dell'inferenza statistica non è basata sulla deduzione (come la matematica) ma sull'**induzione**: dal particolare (ciò che si è osservato) al generale
- La matematica ha un ruolo strumentale, cioè consente di costruire gli strumenti che permettono l'analisi statistica (la matematica sta alla statistica come il martello sta al fabbro)

Statistica descrittiva vs inferenziale

Statistica Descrittiva

Metodi di

- raccolta
- presentazione (grafici)
- caratterizzazione (statistiche)

di un insieme di dati allo scopo di descriverne le caratteristiche

Statistica Inferenziale

Metodi di

- stima

di una particolare caratteristica relativa alla popolazione di interesse, sulla base dell'osservazione di un campione, allo scopo di generalizzare il risultato all'intera collettività

Esempio di inferenza statistica

Qual è la proporzione di persone che scrivono con la mano sinistra?

N= numero di persone; M= n. di "mancini"

Quanto vale p=M/N ?

Campione di 100 persone, di cui 5 sono mancini

$\hat{p}$ = 5/100 = 0.05

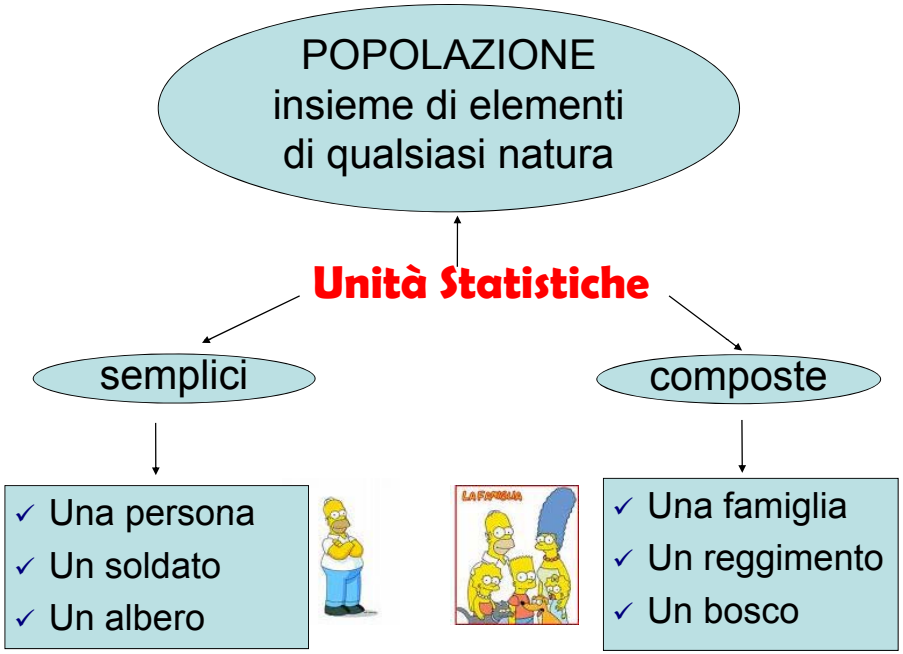
$p \neq \hat{p}$ per errore di campionamento

Inferenza statistica → quantificazione dell'errore

Es. si arriva ad affermare che, con elevata probabilità, $p \in [0.02; 0.08]$

Alcuni termini 'statistici'...

- **Popolazione** o **Collettivo statistico**: insieme che interessa studiare
- **Unità statistica**: elemento della popolazione
- **Dati**: risultato della rilevazione-misurazione di caratteristiche delle unità statistiche
- **Variabile** o **Carattere**: caratteristica rilevata-misurata sulle unità statistiche
- **Modalità**: valori distinti assunti da una variabile
- **Campione**: sottoinsieme della popolazione oggetto della rilevazione



Statistica, dati, variabilità

- La Statistica è una scienza che mira ad estrarre informazioni dai dati
- La ragione della Statistica risiede nella variabilità dei dati: ogni carattere assume valori diversi nelle unità statistiche
  - es. con riferimento al carattere "Esito dell'esame", alcuni presentano la modalità "Promosso", altri la modalità "Respinto"
- Se il mondo fosse perfettamente prevedibile e non ci fosse variabilità, non ci sarebbe bisogno della Statistica

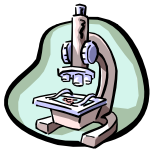
Fonti di variabilità

La variabilità nei dati si riscontra:

- *in due misurazioni dello stesso oggetto* (errore di misura: es. due misurazioni in contemporanea del battito cardiaco)
- *misurazione di due oggetti diversi* (es. battito cardiaco di due persone, oppure battito cardiaco della stessa persona in due momenti)
- *nei processi casuali* (es. due estrazioni con reintroduzione da un'urna contenente palline numerate da 1 a 20)

Genesi dei dati

- Indagini statistiche
  - Popolazione finita
  - Censuarie vs campionarie (lista, inferenza)
- Esperimenti



- Studi osservazionali

Esperimento vs studio osservazionale

Esperimento

Fenomeno:
- replicabile;
- controllabile.

Dati rilevati secondo protocollo sperim.

Studio osservazionale

Fenomeno:
- esistente in natura;
- non controllabile.

Dati rilevati come si presentano.

La strategia di acquisizione dei dati determina la distinzione tra osservazione e sperimentazione

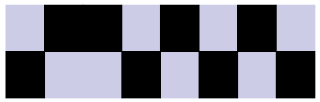
Esperimento vs studio osservazionale

ESPERIMENTO:
efficacia di un fertilizzante

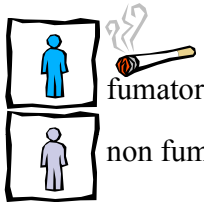
non fertilizzato

fertilizzato

assegnazione casuale



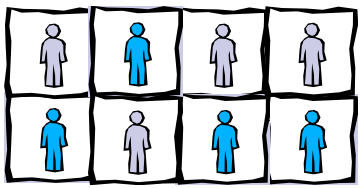
STUDIO OSSERVAZIONALE:
danni del fumo all'apparato respiratorio



fumatore

non fumatore

assegnazione non casuale
(individui scelgono)



Esperimento vs studio osservazionale

Esperimento: efficacia fertilizzante

Tratt.: fertilizzante
Risp.: quantità prodotto

Assegnazione casuale dei lotti al trattamento

Fonti di variabilità sotto controllo

Differenze sistematiche nelle risposte dovute al trattamento

Studio osservazionale: danni del fumo

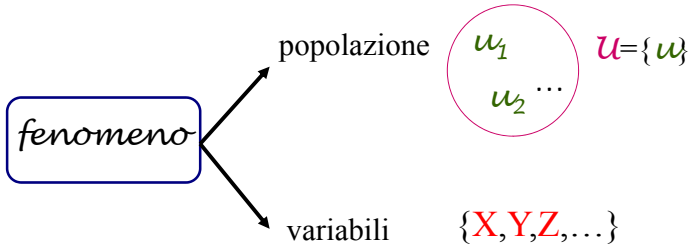
Tratt.: fumo
Risp.: sviluppo malattie repiratorie

Auto-selezione degli individui al trattamento

Differenze sistematiche nelle risposte dovute al fumo e/o altri fattori non controllati (fattori ereditari, età, sesso)

Definizione e rilevazione dei dati

- Individuazione del fenomeno
- Individuazione della popolazione di riferimento e delle unità che la compongono
- Individuazione delle variabili rilevanti e loro definizione operativa



Matrice dei dati

- I dati sono di solito raccolti in forma RETTANGOLARE: matrice righe × colonne
- ogni riga della matrice corrisponde ad una **unità di osservazione**
- ogni colonna della matrice corrisponde ad una **variabile**
- Es. si intervistano 39 persone con un questionario di 6 domande
  → **39** osservazioni × **6** variabili

Esempio: matrice dei dati

VARIABILI

Etichette di identificazione

UNITÀ STATISTICHE

unità u	id	SESSO	ETA' (a.c.)	LIVISTR	DIST(KM)
Alpio	1	M	28	2	5
Caio	2	M	17	4	7.5
Prima	3	F	20	4	12
Velio	4	M	32	2	3.2
Rufa	5	F	16	1	-
Sesto	6	M	34	2	12.3
Beowulf	7	M	18	1	25
Sebaste	8	F	25	2	7.7

Modalità delle variabili

Solitamente i nomi vengono eliminati (privacy)

DIST(KM): distanza casa-lavoro in Km
LIVISTR: livello di istruzione (1=Lic. Elem., 2=Lic. Media, 3=Diploma, 4=Laurea)

Esempio: matrice dei dati

Attenzione alla qualità dei dati!!

unità u	id	SESSO	ETA' (a.c.)	LIVISTR	DIST(KM)
Alpio	1	M	28	2	5
Caio	2	M	17	4	7.5
Prima	3	F	20	4	12
Velio	4	M	32	2	3.2
Rufa	5	F	16	1	-
Sesto	6	M	34	2	12.3
Beowulf	7	M	18	1	25
Sebaste	8	F	25	2	7.7

Controllare coerenza dei dati!

Dato mancante (missing)

Prima di iniziare l'analisi occorre controllare i dati e correggere gli errori riscontrati (data cleaning) ▶

Classificazione delle variabili

Variabili e modalità

Variabile
caratteristica delle unità statistiche che al variare delle unità può assumere almeno due valori

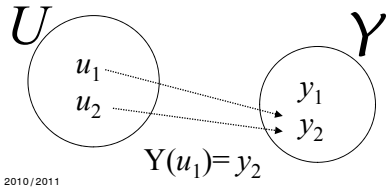
Notazione
X, Y, Z …
$X_1, X_2, \ldots, X_p$

Modalità
Valori assumibili da una variabile (generalmente note a priori).
L'insieme di tali valori è detto INSIEME delle MODALITÀ della variabile

$x, y, z, \ldots$ modalità

X, Y, Z insieme delle modalità

$$Y(u): U \rightarrow Y$$



Tipi di variabili

- Le variabili QUANTITATIVE misurano caratteristiche numeriche: es. il numero di figli e l'altezza di una persona
- Le variabili QUALITATIVE misurano delle qualità: es. il colore degli occhi
  - In particolare le variabili dicotomiche sono variabili qualitative con due sole modalità: es. la variabile sesso assume le modalità maschio e femmina

Codifica numerica delle modalità

- Spesso nella matrice dei dati le modalità delle variabili qualitative sono espresse tramite **numeri** (es. 1 per maschio, 2 per femmina)
- Questi numeri NON sono quantità ma sono dei CODICI che facilitano la registrazione dei dati
- Attenzione: poiché la codifica è arbitraria è importante associare alla matrice dei dati un documento con la codifica (**tracciato record**)

Codifica disgiuntiva di un carattere

X carattere qualitativo con *K* modalità

X_k=1 se X=k,
X_k=0 altrimenti
k=1,2, …, K

ID	SCUOLA	X1	X2	X3
1	3	0	0	1
2	2	0	1	0
3	2	0	1	0
4	3	0	0	1
5	2	0	1	0
6	1	1	0	0
7	2	0	1	0
8	3	0	0	1
9	2	0	1	0
10	3	0	0	1
11	2	0	1	0
12	2	0	1	0
13	3	0	0	1
14	3	0	0	1
15	2	0	1	0
16	3	0	0	1
17	2	0	1	0
18	3	0	0	1
totale		1	9	8

Classificazione delle variabili

Variabile qualitativa
Modalità espresse da nomi, aggettivi, attributi

Variabile quantitativa
Modalità espresse da numeri

QUALITATIVE

- SCONNESSE o NOMINALI — Modalità non ordinabili — SESSO, RESIDENZA, PROFESSIONE
- ORDINALI — Modalità ordinabili — LIV. ISTRUZIONE, VOTAZIONE LIC. MEDIA, LIV. SODDISFAZIONE

QUANTITATIVE

- DISCRETE — Insieme modalità finito o numerabile — N. FIGLI, N. DI STANZE, N. AUTO
- CONTINUE — Insieme modalità infinito non numerabile — ETÀ, ALTEZZA, PESO

Trasformazioni e ricodifiche

Quantitativa continua ⟵ ✕ ⟵ Quantitativa discreta

Y=(100.2, 102.7, …) ⟶ X=(101, 102, …)

X=101 se Y in (100,101]

Attenzione: perdita di informazione!!!

scoring

Qualitativa ordinale ⟵ ✕ ⟵ Qualitativa nominale

Z=(basso,medio,alto) ⟶ W=(normale, estremo)

Z=basso se Y in (100,150]
Z=medio se Y in (150,190]
Z=alto se Y in (190,220]

W=normale se Z=medio
W=estremo se Z=(alto,basso)

Scale di misurazione dei caratteri

- *SCALA NOMINALE* $x_i = x_j, \quad x_i \neq x_j$
- *SCALA ORDINALE* $x_i = x_j, \quad x_i \prec x_j, \quad x_i \succ x_j$
- *SCALA DI INTERVALLI*
  Fissare unità di misura e **origine** del sistema

$$(x_i - x_j) = (x_k - x_h), \quad (x_i - x_j) > (x_k - x_h), \quad (x_i - x_j) < (x_k - x_h)$$

- *SCALA DI RAPPORTI*
  Fissare unità di misura, 0=assenza fenomeno

$$(x_i / x_j) = (x_k / x_h), \quad (x_i / x_j) > (x_k / x_h), \quad (x_i / x_j) < (x_k / x_h)$$

Scala di intervalli vs scala di rapporti

- Scala di rapporti (lo 0 significa assenza del carattere): es. il peso
  - Se A pesa 50kg e B pesa 100kg, allora B pesa il doppio di A
- Scala di intervalli (lo 0 è arbitrario): es. la temperatura in gradi Celsius o Fahrenheit
  - Se A ha una temperatura di 10°C e B di 20°C, non si può dire che B ha una temperatura doppia di A
  - Infatti in gradi Fahrenheit A ha una temperatura di 50°F e B di 68°F

$$F = \frac{9}{5}C + 32$$

Ricordate …

- La distinzione tra variabili qualitative e quantitative è importante per scegliere il metodo di analisi da utilizzare
- Talvolta la classificazione di una variabile dipende da come viene misurata
- Una variabile che assume valori numerici corrispondenti a codici (es. CAP) è qualitativa
- La variabile continua è un concetto astratto: qualunque sia la precisione dello strumento il numero di modalità ottenibili è discreto
  - Es. una bilancia che misura alla precisione dell'hg fornisce valori come 66.0 kg, 66.1 kg, 66.2 kg … → i valori osservabili sono un insieme discreto ma ***il carattere peso è continuo***!

Errore frequente: affermare che un carattere continuo (peso, tempo …) è discreto in quanto si osserva un insieme discreto di valori

Distribuzioni statistiche

Cicchitelli Cap. 3

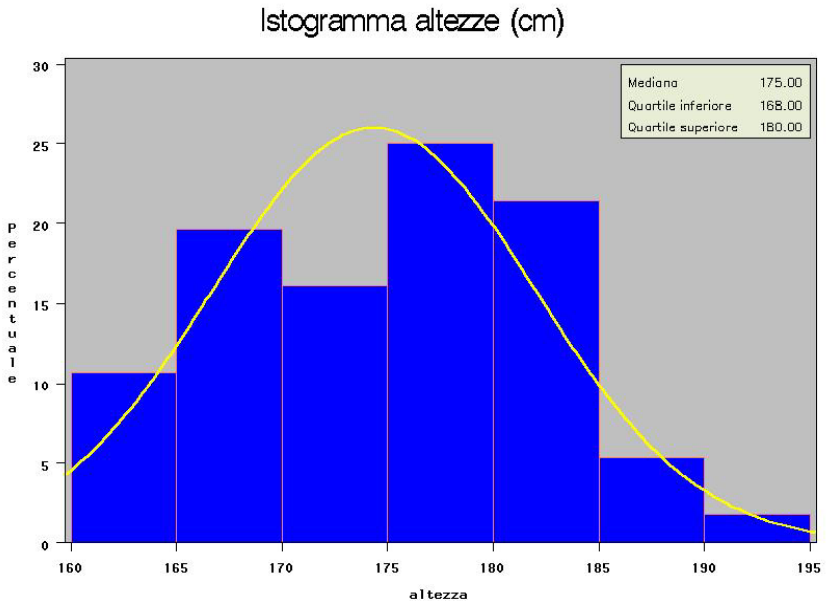
Distribuzione e sintesi dei dati

- I dati sono un lungo elenco di valori ed è difficile trovare una regolarità
- Come fa uno studente a confrontare la sua altezza con quella dei suoi compagni di classe?
- Meglio usare una sintesi dei valori. Ad esempio
  - la metà delle altezze è superiore a 175 cm e l'altra metà è inferiore a questo valore
  - il 50% centrale dei valori è compreso tra 168 e 180 cm

Distribuzione e sintesi dei dati

- Questa sintesi fornisce due informazioni: il valore centrale è 175 cm e le altezze si distribuiscono intorno a questo valore, variando tra 168 e 180 cm nella parte centrale (50% delle altezze) della distribuzione
- Ad es. se uno studente è alto 178 cm, in base a questa sintesi sa subito che la sua altezza si trova nella parte centrale della distribuzione, poco sopra il valore centrale

- Spesso questo tipo di sintesi fornisce tutte le informazioni necessarie per capire l'andamento del fenomeno, soprattutto quando la forma della distribuzione è una di quelle tipiche



Come si esplorano i dati?

Ogni analisi esplorativa dovrebbe seguire questi passi

grafico → forma → centro → dispersione

1. Tracciare il **grafico** più appropriato
2. Descrivere la **forma** della distribuzione in base al grafico e indici di forma
3. Calcolare una misura del **centro** della distribuzione, appropriata in base alla forma della distribuzione
4. Calcolare un indice di **dispersione** appropriato in base alla forma della distribuzione e coerente con la misura di centro utilizzata

UTILIZZARE GRAFICI E INDICI APPROPRIATI IN BASE AL TIPO DI VARIABILE (qualitativa sconnessa, …)

Le analisi statistiche

Univariate: ogni variabile separatamente

l'**età** media è 23.75 anni

unità u	etichetta	SESSO	ETA' (a.c.)	LIVISTR	DIST(KM)
Alpio	1	M	28	2	5
Caio	2	M	17	4	7.5
Prima	3	F	20	4	12
Velio	4	M	32	2	3.2
Rufa	5	F	16	1	-
Sesto	6	M	34	2	12.3
Beowulf	7	M	18	1	25
Sebaste	8	F	25	2	7.7

Bivariate: le variabili a coppie

l'**età** media dei maschi è più elevata di quella delle femmine (variabili considerate: età, sesso)

Multivariate

Al crescere dell'**età** il **livello di istruzione** aumenta per le femmine e diminuisce per i maschi (variabili considerate: età, livello di istruzione, sesso)

Distribuzioni di frequenza

Un esempio di matrice dei dati

Unità u	Etichetta	SESSO	ETA'	LIVIST	DIST
Alpio	1	M	28	2	5
Caio	2	M	17	4	7.5
Prima	3	F	20	4	12
Velio	4	M	32	2	3.2
Rufa	5	F	16	1	-
Sesto	6	M	34	2	12.3
Beowful	7	M	18	1	25
Sebaste	8	F	25	2	7.7

Distribuzione statistica disaggregata

M, M, F, M, F, M, M, F

$x_1, x_2, \ldots, x_N$

Distribuzione di frequenza

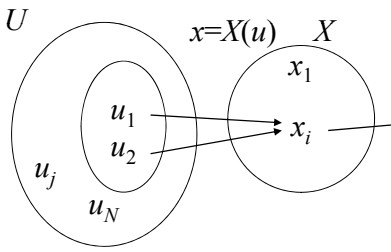
Sesso	M	F	*Tot*
Freq	5	3	*8*

$x_1, x_2, \ldots x_i \ldots x_k$ modalità

$n_1, n_2, \ldots n_i \ldots n_k$ frequenze assolute

Distribuzioni di frequenza

n_i = frequenza assoluta, cioè numero di unità statistiche che presentano la modalità $X = x_i$



Modalità	Frequenze
x_1	n_1
.	.
x_i	n_i
.	.
x_k	n_k
Totale	N

numerosità della popolazione $N = n_1 + n_2 + \ldots + n_k$

Frequenze relative

Si dice FREQUENZA RELATIVA di una modalità x_i, o di una classe di modalità $(x_{i-1}; x_i)$ e si indica con f_i, la frazione o proporzione di u.s. che presentano tale modalità.

$$f_i = \frac{n_i}{\sum_{i=1}^{k} n_i} = \frac{n_i}{N} \qquad i = 1, 2, \ldots, k$$

Proprietà: $0 \le f_i \le 1 \qquad i = 1, 2, \ldots, k$

$$\sum_{i=1}^{k} f_i = 1$$

Frequenze relative: perché?

Facilitare la percezione del PESO delle modalità

Sesso	Freq. Assoluta	Freq. Relativa	Freq. Rel. %
M	1750	0.583	58.3
F	1250	0.417	41.7
Totale	3000	1	100

Facilitare CONFRONTI tra popolazioni

Sesso	Freq. Assoluta		Freq. Rel. %	
	Pop. A	Pop. B	Pop. A	Pop. B
M	1750	850	58.3	85.0
F	1250	150	41.7	15.0
Totale	3000	1000	100	100

Frequenze cumulate

- **Frequenze cumulate**: somma delle frequenze sino alla modalità considerata (variabili ordinali o quantitative)
  - **Assolute**: numero di u.s. con valore di X minore o uguale a x_i

$$N_i = n_1 + n_2 + \ldots + n_i$$

  - **Relative**: proporzione di u.s. con valore di X minore o uguale a x_i

$$F(x_i) = \Pr(X \le x_i) = f_1 + f_2 + \ldots + f_i$$

F( ) è chiamata **funzione di ripartizione**

Esempio: soddisfazione dei clienti

Carattere qualitativo sconnesso

Soddisfatto	Frequenza assoluta	Frequenza relativa (percentuale)
Sì	330	61.1%
No	210	38.9%
Totale	540	100%

Non sa / Non risponde: 20 (3.6% degli intervistati)

Esempio: soddisfazione dei clienti

Carattere qualitativo ordinale

Soddisfatto	Frequenza assoluta	Frequenza relativa (percentuale)	Frequenza cumulata (percentuale)
Poco	120	22.2%	22.2%
Abbastanza	240	44.4%	66.6%
Molto	180	33.3%	100.0%
Totale	540	100%	

Non sa / Non risponde: 20 (3.6% degli intervistati)

Distorsione da dati mancanti

- In un quartiere con 100 abitanti, 50 hanno fiducia nel sindaco e 50 no (percentuale favorevoli 50%). Tutti vengono contattati, ma il 20% non risponde (tasso di risposta 80%)

Scenario A

Modalità	Freq	Tasso risposta	Risposte
Sì	50	80%	40
No	50	80%	40

Tasso di risposta 80%

Percentuale favorevoli 50%

Scenario B

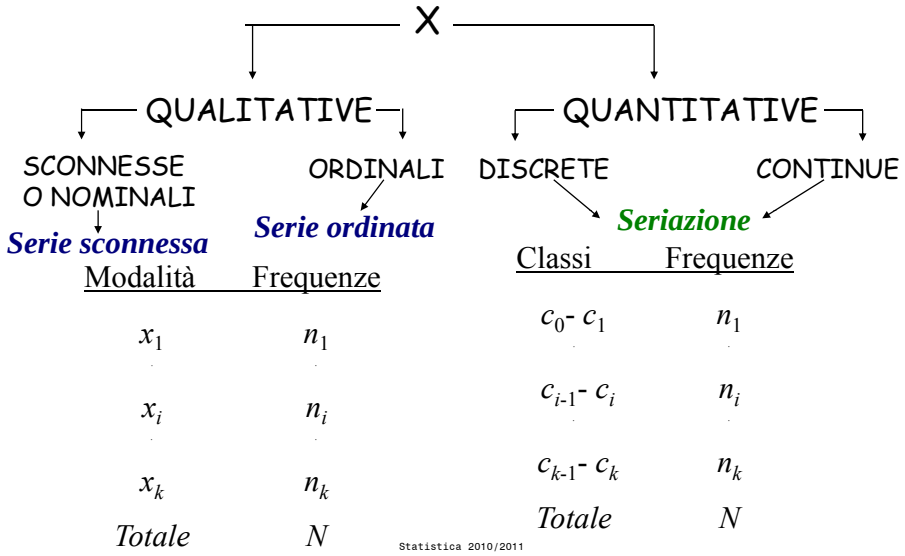
Modalità	Freq	Tasso risposta	Risposte
Sì	50	60%	30
No	50	100%	50

Tasso di risposta 80%

Percentuale favorevoli 37.5%

Mod.tà	Freq.	Fr.rel.	Freq.cum	F(x)
x_1	n_1	f_1	n_1	f_1
x_2	n_2	f_2	n_1+n_2	f_1+f_2
…	…	…	…	…
x_i	n_i	f_i	$n_1+n_2+\ldots+n_i$	$f_1+f_2+\ldots+f_i$
…	…	…	…	…
x_k	n_k	f_k	N	1
Totale	N	1		

Classificazione delle distribuzioni



Seriazioni

- Variabile discreta con poche modalità → riportare in tabella tutte le modalità con le corrispondenti frequenze
- Variabile discreta con molte modalità oppure variabile continua → raggruppare le modalità in classi e calcolare le frequenze delle classi

Obiettivo: sintetizzare i dati per facilitare l'interpretazione

Determinare gli estremi delle classi

- Ciascuna classe di intervallo ha la stessa ampiezza
- Determinare l'ampiezza di ciascuna classe nel seguente modo:

$$w = \text{Ampiezza dell'intervallo} = \frac{\text{Valore massimo} - \text{Valore minimo}}{\text{Numero di classi}}$$

- Usare almeno 5 ma non più di 15-20 intervalli
- Gli intervalli non si sovrappongono mai
- Arrotondare l'ampiezza dell'intervallo per ottenere gli estremi delle classi

Esempio di seriazione

- Esempio: Un produttore di isolante seleziona a caso 20 giorni invernali e registra la temperatura massima giornaliera (°F)

  24, 35, 17, 21, 24, 37, 26, 46, 58, 30,
  32, 13, 12, 38, 41, 43, 44, 27, 53, 27

Esempio di seriazione

- Ordina i dati grezzi in ordine crescente: 12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58
- Trova il campo di variazione: 58 - 12 = 46
- Seleziona il numero di classi: 5 (solitamente fra 5 e 15)
- Calcola l'ampiezza dell'intervallo: 10 (46/5 poi arrotonda per eccesso)
- Determina i limiti dell'intervallo: 10 ma meno di 20, 20 ma meno di 30, . . . , 60 ma meno di 70
- Conta le osservazioni & assegnale alle classi

Esempio di seriazione

Dati in sequenza ordinata:

12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58

Intervallo	Frequenza	Frequenza Relativa	Percentuale
10 ma meno di 20	3	.15	15
20 ma meno di 30	6	.30	30
30 ma meno di 40	5	.25	25
40 ma meno di 50	4	.20	20
50 ma meno di 60	2	.10	10
Totale	20	1.00	100

Indagine sulla fecondità (INF/2, 1995)

Sotto-insieme delle donne coniugate o conviventi residenti nelle regioni del centro Italia

- Alcune delle caratteristiche rilevate
- Anno di nascita, nella forma aa
- Titolo di studio alla data dell'intervista
- Anno di nascita del primo figlio
- Anno di nascita del secondo figlio
- Numero totale di figli
- Ha mai lavorato? (1=no, 2=in passato, 3=attualmente)

Indagine sulla fecondità (INF/2, 1995)

ID	ANNONASC	TITSTUD	FIGLIO1	FIGLIO2	NFIGLI	MAILAV
4252	59	6	84	87	2	2
4262	46	5	73	74	2	1
4272	53	2	74	75	3	3
4287	47	7	71	76	2	3
4290	51	4	75	76	2	3
4297	58	2	78	79	3	3
4303	66	4	90	.	1	3
4307	50	2	69	.	1	2
4322	56	5	79	.	1	3
4323	61	5	86	.	1	3

Indagine sulla fecondità (INF/2, 1995)

Dalla matrice dei dati alle tabelle

Numero di figli alla data dell'intervista

NFIGLI	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	71	12.10	71	12.10
1	181	30.83	252	42.93
2	264	44.97	516	87.90
3	61	10.39	577	98.30
4	9	1.53	586	99.83
5	1	0.17	587	100.00

Indagine sulla fecondità (INF/2, 1995)

Anno di nascita della donna

ANNONASC	Frequency	Percent	Cumulative Frequency	Cumulative Percent
46	33	5.62	33	5.62
47	22	3.75	55	9.37
48	21	3.58	76	12.95
49	21	3.58	97	16.52
50	29	4.94	126	21.47
51	33	5.62	159	27.09
52	26	4.43	185	31.52
53	26	4.43	211	35.95
54	20	3.41	231	39.35
...				
73	1	0.17	586	99.83
75	1	0.17	587	100.00

Per una migliore lettura: raggruppare le modalità in classi!

Indagine sulla fecondità (INF/2, 1995)

Anno di nascita della donna (dati raggruppati in classi)

ANNONASC	Frequency	Percent	Cumulative Frequency	Cumulative Percent
46-50	126	21.47	126	21.47
51-55	127	21.64	253	43.10
56-60	125	21.29	378	64.40
61-65	124	21.12	502	85.52
66-70	77	13.12	579	98.64
71-75	8	1.36	587	100.00

Indagine sulla fecondità (INF/2, 1995)

Titolo di studio alla data dell'intervista

TITSTUD	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	2	0.34	2	0.34
2	105	17.89	107	18.23
3	183	31.18	290	49.40
4	53	9.03	343	58.43
5	177	30.15	520	88.59
6	7	1.19	527	89.78
7	60	10.22	587	100.00

Per una migliore lettura: decodifica delle modalità!

1=licenza elementare; 2=licenza media

3-5=diploma

6=diploma universitario; 7=laurea

Indagine sulla fecondità (INF/2, 1995)

Titolo di studio alla data dell'intervista (uso delle etichette)

TITSTUD	Freq	Percent	Cumulative Frequency	Cumulative Percent
lic. elementare	2	0.34	2	0.34
lic. media	105	17.89	107	18.23
diploma	413	70.36	520	88.59
diploma univ.	7	1.19	527	89.78
laurea	60	10.22	587	100.00

Indagine sulla fecondità (INF/2, 1995)

Condizione lavorativa alla data dell'intervista

LAV	Frequency	Percent
1	134	22.83
2	163	27.77
3	290	49.40
Totale	587	100.00

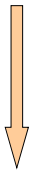
Per una migliore lettura: decodifica delle modalità!

LAV	Frequency	Percent
mai lavorato	134	22.83
lavorato in passato	163	27.77
lavora attualmente	290	49.40
Totale	587	100.00

Tabelle di frequenza bivariate

Distribuzione doppia disaggregata

(M,2),(M,4),(F,4),…,(F,2)



Distribuzione doppia di frequenza

SESSO	*LIVELLO DI ISTRUZIONE*			*TOTALE*
	Lic. Elem.	Lic. Media	Laurea	
M	1	3	1	5
F	1	1	1	3
TOTALE	2	4	2	8

Un esempio di matrice dei dati

Unità u	Etichetta	SESSO	ETA'	LIVIST	DIST
Alpio	1	M	28	2	5
Caio	2	M	17	4	7.5
Prima	3	F	20	4	12
Velio	4	M	32	2	3.2
Rufa	5	F	16	1	-
Sesto	6	M	34	2	12.3
Beowful	7	M	18	1	25
Sebaste	8	F	25	2	7.7

Tabella a doppia entrata o tabella di contingenza

Tabelle di frequenza bivariate

Esempio: numero figli e condizione lavorativa delle donne (INF/2, 1995)

Frequency	mai lavorato	lavorato in passato	lavora attualmente	Total
0	15	14	42	71
1	32	49	100	181
2	63	82	119	264
3	20	16	25	61
4	4	2	3	9
5	0	0	1	1
Total	134	163	290	587

Frequenze assolute

Tabelle di frequenza bivariate

Esempio: numero figli e condizione lavorativa delle donne (INF/2, 1995)

Percent	mai lavorato	lavorato in passato	lavora attualmente	Total
0	2.56	2.39	7.16	12.10
1	5.45	8.35	17.04	30.83
2	10.73	13.97	20.27	44.97
3	3.41	2.73	4.26	10.39
4	0.68	0.34	0.51	1.53
5	0.00	0.00	0.17	0.17
Total	22.83	27.77	49.40	100.00

Frequenze relative

Tabella di contingenza: obiettivi e rischi di 121 fondi di investimento (FREQUENZE ASSOLUTE)

	Livello di rischio			
Obiettivo	Alto	Medio	Basso	Totale
Crescita	14	23	12	49
Valore	3	23	46	72
Totale	17	46	58	121

Tabella di contingenza: obiettivi e rischi di 121 fondi di investimento (PERCENTUALI TOTALI)

	Livello di rischio			
Obiettivo	Alto	Medio	Basso	Totale
Crescita	11.57	19.01	9.92	40.50
Valore	2.48	19.01	38.02	59.50
Totale	14.05	38.02	47.93	100.00

Tabella di contingenza: obiettivi e rischi di 121 fondi di investimento (PERCENTUALI DI RIGA)

	Livello di rischio			
Obiettivo	Alto	Medio	Basso	Totale
Crescita	28.57	46.94	24.49	100.00
Valore	4.17	31.94	63.89	100.00
Totale	14.05	38.02	47.93	100.00

Tabella di contingenza: obiettivi e rischi di 121 fondi di investimento (PERCENTUALI DI COLONNA)

	Livello di rischio			
Obiettivo	Alto	Medio	Basso	Totale
Crescita	82.35	50.00	20.69	40.50
Valore	17.65	50.00	79.31	59.50
Totale	100.00	100.00	100.00	100.00

Soccorso con elicottero o con ambulanza?

	Elicottero	Ambulanza	*Totale*
Morti	64	260	*324*
Soprawissuti	136	840	*976*
Totale	*200*	*1100*	*1300*

Pazienti morti:

- Elicottero: 64/200 = 32%
- Ambulanza: 260/1100 = 24%

Soccorso con elicottero o con ambulanza?

Incidenti gravi

	Elicottero	Ambulanza	*Totale*
Morti	48	60	*108*
Soprawissuti	52	40	*92*
Totale	*100*	*100*	*200*

Pazienti morti:

- Elicottero: 48/100 = 48%
- Ambulanza: 60/100 = 60%

Incidenti non gravi

	Elicottero	Ambulanza	*Totale*
Morti	16	200	*216*
Soprawissuti	84	800	*884*
Totale	*100*	*1000*	*1100*

Pazienti morti:

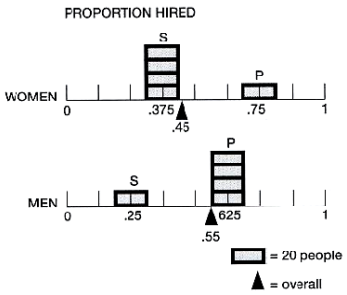
- Elicottero: 16/100 = 16%
- Ambulanza: 200/1000 = 20%

E' un esempio del paradosso di Simpson!

Simpson, E. H. 1951. The interpretation of interaction in contingency tables. *J. Roy. Statist. Soc. Ser. B* 13: 238-241.

Discriminazione sessuale

	Social Sciences		Physical Science		Overall	
	Male	Female	Male	Female	Male	Female
Hired	5	30	50	15	55	45
Denied	15	50	30	5	45	55
Total applied	20	80	80	20	100	100



Un altro esempio del paradosso di Simpson!

Razza e pena di morte

Imputato bianco		
	Vittima bianca	Vittima nera
A morte	19	0
No	132	9
	12.6%	**0%**

Imputato nero		
	Vittima bianca	Vittima nera
A morte	11	6
No	52	97
	17.5%	**5.8%**

Esercizio:

Costruire la tabella bivariata con “Esito (a morte vs no)” e “Imputato (bianco vs nero)”. Commentare il paradosso.

Rappresentazioni grafiche per variabili qualitative

Cicchitelli Cap. 4

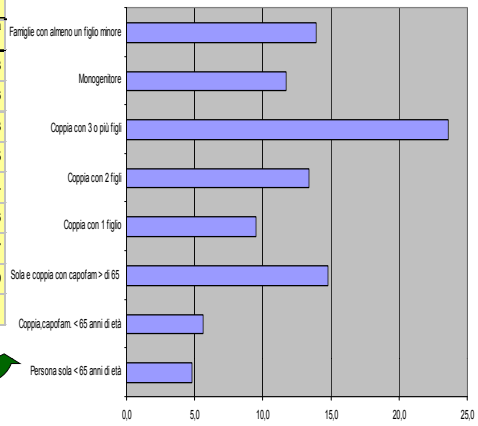
Rappresentazioni grafiche

dalla tabella … alla rappresentazione grafica

Tabella 1 - Famiglie povere per caratteristiche della famiglia Anno 1998 (Istat)

TIPOLOGIE FAMILIARI	1998	
	Numero (migliaia)	Incidenza *Incidence*
Persona sola < 65 anni di età	98	4,8
Coppia,capofam. < 65 anni di età	127	5,6
Sola e coppia con capofam > di 65	738	14,8
Coppia con 1 figlio	420	9,5
Coppia con 2 figli	521	13,4
Coppia con 3 o più figli	252	23,6
Monogenitore	190	11,7
Famiglie con almeno un figlio minore	851	13,9
TOTALE	**2.557**	**11,8**

Figura 1 - Incidenza della povertà per caratteristiche della famiglia Anno 1998 (Istat)

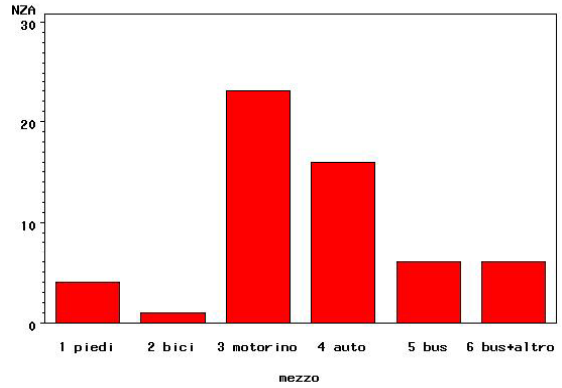


Migliore percezione dell’informazione

Diagrammi a barre per variabili qualitative

- Barre verticali, categorie lungo l’asse orizzontale
- Altezze proporzionali alle frequenze (assolute o relative)
- In alternativa: barre orizzontali

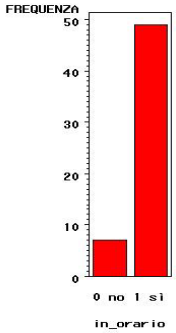
Diagramma a barre mezzo di trasporto

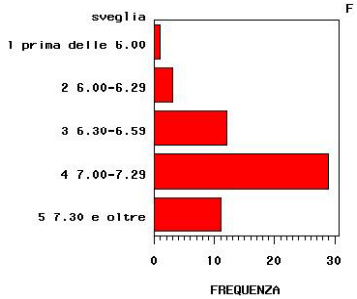


NB. Le barre sono separate: la variabile non può assumere valori tra una categoria e l’altra

Ordine delle barre

VARIABILI SCONNESSE: nel diagramma per il mezzo di trasporto o in quello dell'arrivo a scuola in orario l'ordine delle barre è completamente arbitrario

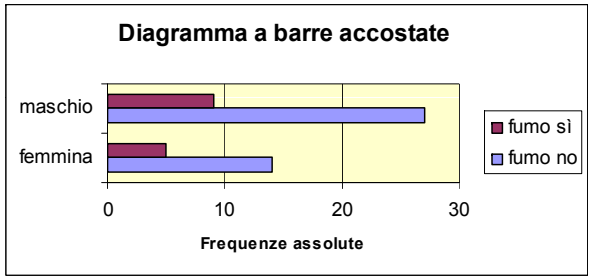




VARIABILI ORDINALI: nel diagramma dell'ora in cui ci si alza le categorie sono ORDINATE e quindi devono essere rappresentate nell'ordine giusto per vedere l'andamento delle frequenze

Diagramma a barre accostate

	femmina	maschio	
fumo no	14	27	41
fumo sì	5	9	14
	19	36	55

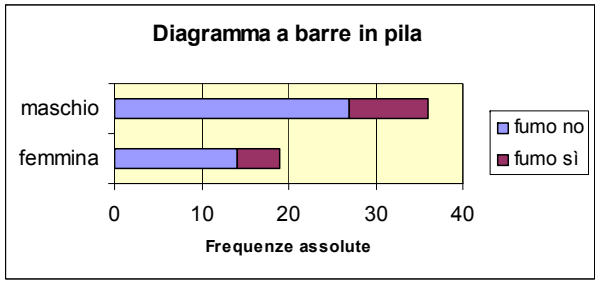


Le barre del diagramma rappresentano le frequenze congiunte: ci sono 14 ragazze non fumatrici

- Confrontando le barre adiacenti possiamo vedere che sia tra i maschi che tra le femmine di questa classe è più probabile essere non fumatori che fumatori
- Mentre confrontando le due barre viola, possiamo vedere che tra i fumatori ci sono più maschi che femmine

Diagramma a barre in pila

	femmina	maschio	
fumo no	14	27	41
fumo sì	5	9	14
	19	36	55

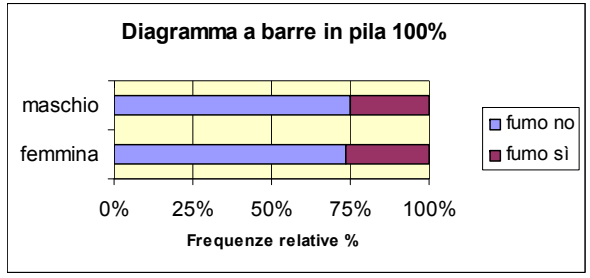


Per capire qual è la proporzione di fumatori tra i maschi e le femmine, conviene impilare le barre

Diagramma a barre in pila 100%

	femmina	maschio	
fumo no	14	27	41
fumo sì	5	9	14
	19	36	55

	femmina	maschio
fumo no	73.7%	75.0%
fumo sì	26.3%	25.0%
	100.0%	100.0%



Per confrontare le proporzioni di fumatori tra i maschi e le femmine, conviene impilare le barre usando le percentuali di colonna anziché le frequenze

Barre o torta?

Tab. 2- Forze lavoro per condizione anno 1999 (migliaia)

Condizione	*TOTALE*
Occupati	20435
disoccupati	996
in cerca di 1a occup.	1152
altri	596
TOTALE	23179

Fonte: Istat, Rapporto sull'Italia 2001

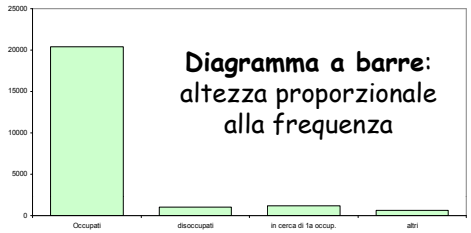


Diagramma a barre: altezza proporzionale alla frequenza

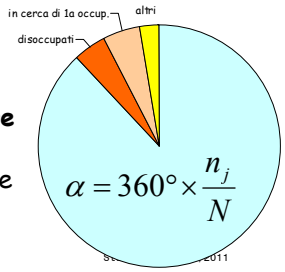


Diagramma circolare (torta): angolo al centro proporzionale alla frequenza

Diagramma a barre → migliore percezione delle differenze

Diagramma a torta → migliore percezione della composizione

"Doesn't matter where they're posted, those are not *BAR* graphs."

http://www.causeweb.org

Rappresentazioni grafiche per variabili quantitative

Cicchitelli Cap. 4

Grafici per variabili quantitative

- Per capire come sintetizzare la distribuzione di un carattere quantitativo è utile conoscere la sua forma
- La forma di una distribuzione può essere vista attraverso un grafico
- Grafici più utilizzati
  - Dotplot
  - Istogramma
  - Boxplot [verrà presentato più avanti, dopo gli indici di forma]
  - Ramo-foglia (Steam and leaf)
  - Diagramma a bastoncini

Diagramma a bastoncini

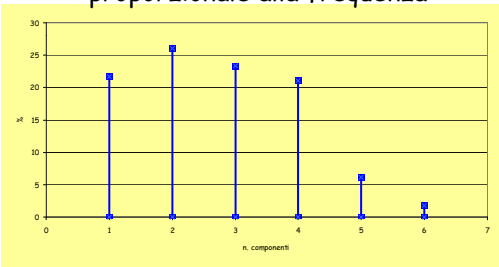
Quando la variabile è **discreta con poche modalità**

Tab. 3 – Famiglie per numero di componenti. Italia 1998 (v.a e %)

Componenti	*v.a.*	*%*
1	4594130	21.65
2	5527810	26.05
3	4954870	23.35
4	4466810	21.05
5	1294420	6.1
6 e più	381960	1.8
Totale	21220000	100

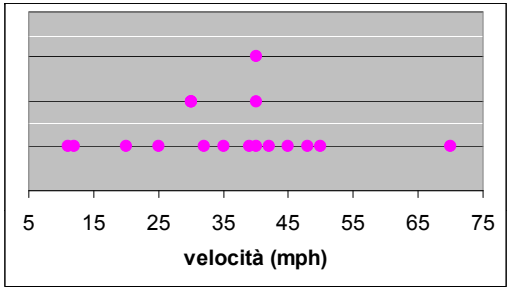
Fonte: Istat, Rapporto sull'Italia 2001

Diagramma a bastoncini: altezza proporzionale alla frequenza



Dotplot

- mostra i singoli casi osservati come punti
- dal dotplot possiamo vedere la forma, il centro e la dispersione dei dati



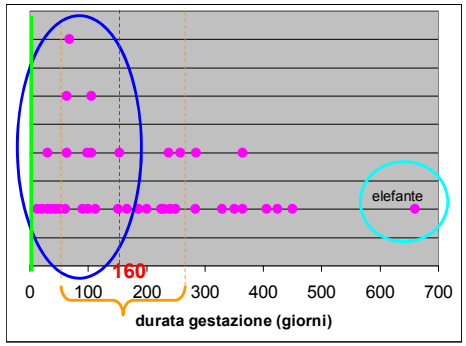
Il dotplot è utile quando:

- si hanno pochi casi
- si vogliono vedere i singoli valori

Attenzione. Software diversi fanno dotplot diversi: a volte 1 punto rappresenta 1 singolo caso, a volte 2 o più casi, a volte i valori vengono arrotondati

Dotplot: durata gestazione di alcuni mammiferi

- La distribuzione è centrata verso i valori più bassi, senza gruppi o buchi particolari
- C'è una sorta di '**muro**' a 0 giorni, perché nessun mammifero può avere un periodo di gestazione più piccolo!



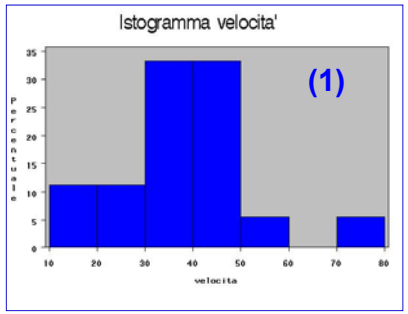
- L'elefante è l'unico mammifero fuori norma (**outlier**)
- Circa la **metà** dei mammiferi hanno un periodo di gestazione superiore a 160 giorni e la metà hanno un periodo più breve
- La **metà centrale** ha un periodo di gestazione che varia tra i 63 e i 284 giorni.

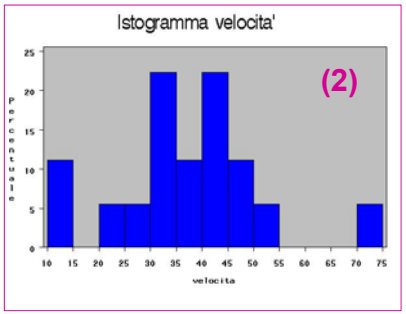
Istogramma

- L'istogramma rappresenta un insieme di casi (raggruppati in classi) come rettangoli
- *Nel caso più semplice le classi sono di uguale ampiezza*: in tal caso l'altezza del rettangolo è proporzionale alla frequenza della classe
- L'istogramma mostra la forma, il centro e la dispersione dei dati
- Rappresenta la distribuzione sotto la seguente ipotesi: *in ogni classe le frequenze sono uniformemente distribuite nell'intervallo*

Istogramma

- Cambiando l'ampiezza delle barre dell'istogramma (classi) a volte si ha un'impressione diversa della forma della distribuzione
- Per esempio, l'istogramma (1) per la velocità dei mammiferi ha meno barre ma più ampie rispetto all'istogramma (2) e mostra una forma a campana più simmetrica, con un solo picco invece di due
- Se ci sono pochi valori è difficile identificare i picchi, in questi casi è meglio utilizzare grafici che mostrano i singoli dati, come il dotplot o il ramo-foglia





Istogramma

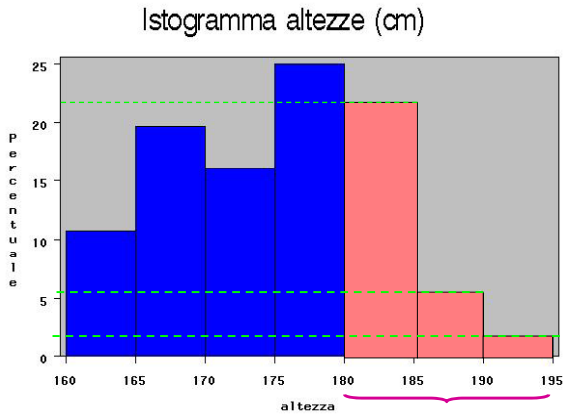
- Non c'è una regola per trovare qual è l'ampiezza di classe migliore per disegnare l'istogramma, proprio come per un fotografo non c'è una regola che gli dica quando e come usare lo zoom!
- Versioni diverse del grafico mettono in luce caratteristiche differenti della distribuzione: il lavoro dello statistico è trovare quella versione che mostra le caratteristiche più importanti!
- Un istogramma è una buona rappresentazione dei dati quando:
  - Ci sono molti valori da rappresentare
  - Non interessa conoscere la posizione di ciascun valore
  - Si è interessati a mostrare la forma generale della distribuzione

Esempio

- Quale proporzione degli studenti ha un'altezza di 180 cm o più?

Soluzione

- Individuare **l'intervallo di valori** >180 sull'asse X
- Quale **proporzione dell'area** totale corrisponde alle barre su questo intervallo?

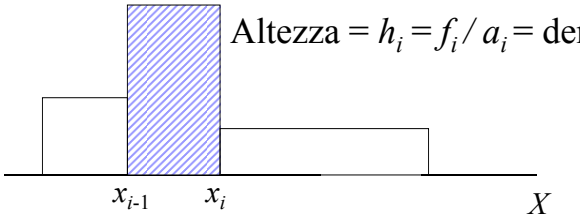


- A occhio questa proporzione è **circa 1/3** → circa 1/3 degli studenti di questa classe hanno un'altezza>180
- In maniera più precisa: possiamo sommare le altezze delle 3 barre dell'istogramma alla destra di 180, cioè 22+6+2 = 30
- Se le classi non hanno uguale ampiezza: sommare le aree!

Istogramma

- L'istogramma può essere costruito utilizzando sia le frequenze assolute che relative (o %)
- Attenzione: se le classi non hanno ampiezza costante, come negli esempi fatti, la costruzione dell'istogramma è più complicata!
  - Base = ampiezza della classe
  - Altezza = densità della classe = frequenza/ampiezza
  - Area = frequenza della classe

Istogramma



Altezza = $h_i = f_i / a_i$ = densità classe

Base = $x_i - x_{i-1} = a_i$ = ampiezza classe

Area = $a_i \times h_i = f_i$ = frequenza classe

Istogramma

Classi	Freq.rel.	Ampiezza	Densità
x_0-| x_1	f_1	a_1	h_1
...	...	...	...
x_{i-1}-| x_i	f_i	a_i	h_i
...	...	...	...
x_{k-1}-| x_k	f_k	a_k	h_k
Totale	1		

Densità di frequenza: $h_i = f_i / a_i$

Ampiezza di classe: $a_i = x_i - x_{i-1}$

Istogramma con poligono di frequenza

segmenti che uniscono **i punti centrali** dei lati superiori dei rettangoli che definiscono l'istogramma, comprese due classi terminali con frequenza zero e ampiezza pari all'ampiezza della classe adiacente

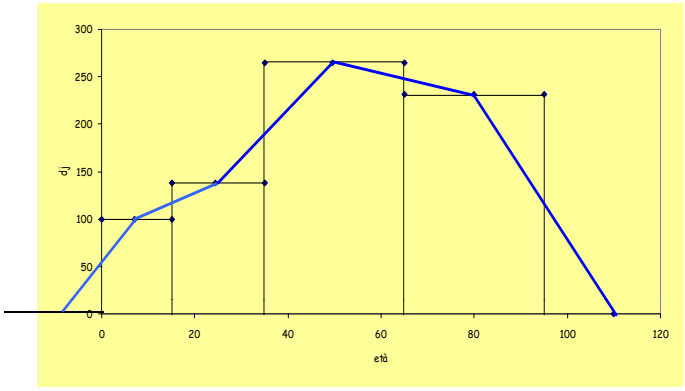


Diagramma Ramo-Foglia

Un modo semplice per vedere i dettagli della distribuzione di un set di dati

METODO: Separare la serie di dati ordinata in

- cifre più significative (i **rami**)
- cifre meno significative (le **foglie**)

Esempio Ramo-Foglia

Dati ordinati:

21, 24, 24, 26, 27, 27, 30, 32, 38, 41

Qui usiamo le decine come unità per i rami:

- 21 è mostrato come → Ramo 2, Foglia 1
- 38 è mostrato come → Ramo 3, Foglia 8

Ramo	Foglia
2	1 4 4 6 7 7
3	0 2 8
4	1

La funzione di ripartizione

Data una v.s. quantitativa X si dice *funzione di ripartizione* $F(x)$ la frequenza relativa (proporzione) dei valori minori o uguali a x:

$$F(x) = pr\left(u : X(u) \le x\right) = pr\left(X \le x\right)$$

Proprietà:
- $F(x)=0$ per $x < x_{min}$
- $F(x)=1$ per $x \ge x_{max}$
- $F(x)$ non decrescente

Insieme delle modalità ordinate di X:
X v.s. discreta $\{x_{min},\ldots, x_j,\ldots, x_{max}\}$
X v.s. continua $[x_{min}, x_{max}]$

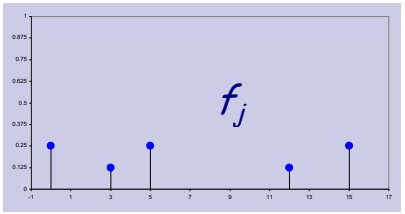
Vediamo 2 tipi di funzione di ripartizione: quella empirica e quella dedotta dall'istogramma

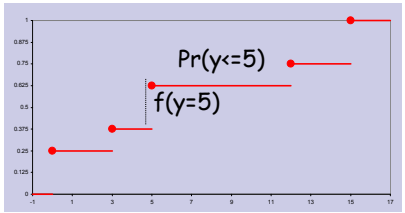
Funzione di ripartizione empirica

Data una successione di dati grezzi $x_1, x_2, \ldots, x_n$ di una v.s. X, la $F(X)$ calcolata a partire da tali dati è detta funzione di ripartizione empirica.

X: {0,0,3,5,5,12,15,15}

x_j	n_j	f_j	$F(x)$
0	2	0.250	0.250
3	1	0.125	0.375
5	2	0.250	0.625
12	1	0.125	0.750
15	2	0.250	1.000
tot	8	1.000	

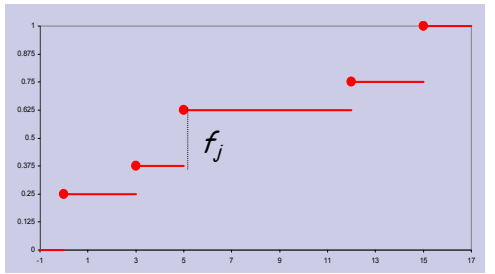




Funzione di ripartizione empirica

Proprietà:

- $F(X<x_{min})=0$; $F(X\ge x_{max})=1$; non decrescente
- Funzione 'a gradini': costante in $[x_{j-1}; x_j)$
- In $X=x_j$ $F(x)$ "salta" di f_j (freq. rel,. di x_j)



Funzione di ripartizione dedotta dalla densità (variabili continue)

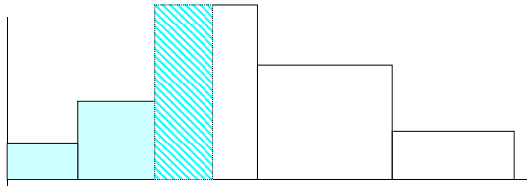
$$pr(X \le x) = F(x) = \int_{-\infty}^{x} f(t)dt$$

densità

Ipotesi dell'istogramma

$$F(x) = F(x_{j-1}) + h_j(x - x_{j-1}), \quad x \in \left(x_{j-1}; x_j\right]$$

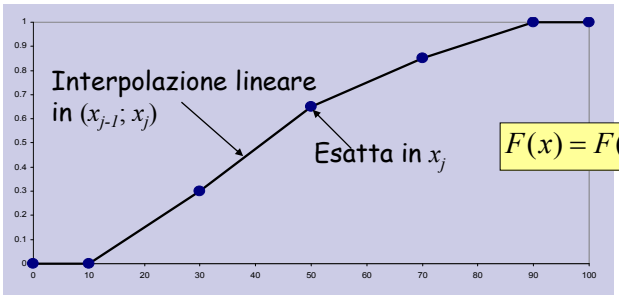
$$h_j = \frac{f_j}{x_j - x_{j-1}}$$



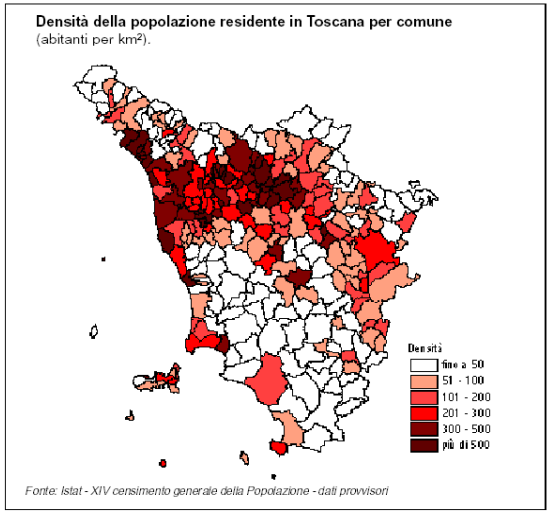
Funzione di ripartizione dedotta dalla densità (variabili continue)

Proprietà:

- $F(X<x_{min})=0$; $F(X\ge x_{max})=1$; non decrescente
- Funzione lineare in $[x_{j-1}; x_j)$
- la derivata prima rappresenta la pendenza dei segmenti di retta che uniscono due estremi di classe successivi



Cartogrammi



Rappresentazione di serie territoriali

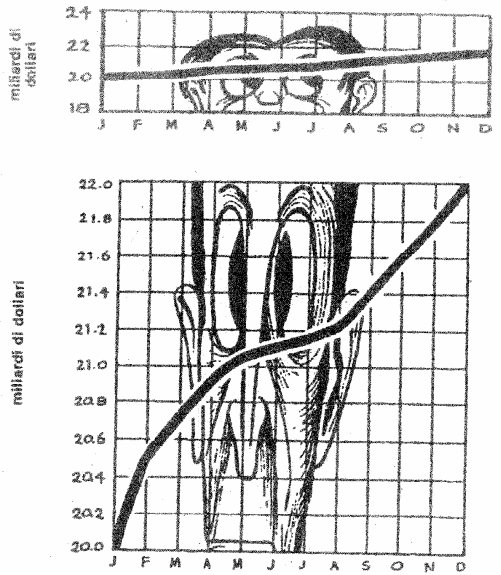
Aree geografiche: comuni

Carattere: densità della popolazione

Errori nella Presentazione dei Dati

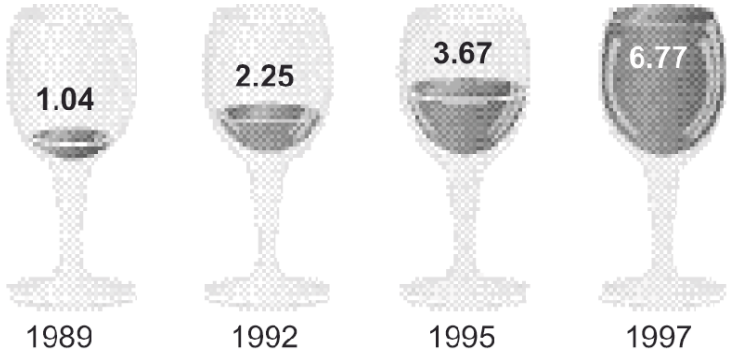
- Compressione o distorsione dell'asse verticale
- Omissione dello zero sull'asse verticale
- Non fornire una base di riferimento per il confronto di dati di diversi gruppi

Fig. V.1 - Due diverse rappresentazioni dello stesso fatto: l'incremento del reddito nazionale in un anno (da HUFF (1954), p. 62).



Un grafico inutile

Noi stiamo bevendo di più...
Esportazioni di vino australiano verso gli Stati Uniti
in milioni di galloni



Fonte: adattato da S. Watterson: "Liquid Gold-Australians Are Changing the World of Wine. Even the French Seem Grateful", *Time*, 22 novembre 1999, 68.

Rapporti statistici

Cicchitelli Cap. 2 (escluso 2.5)

Rapporti statistici

$$R=A/B$$

- R indica quanta parte dell'intensità di A compete, in media, ad ogni unità di B
- Almeno una delle due grandezze, A o B, deve riferirsi ad un fenomeno collettivo
- Tra A e B deve intercorrere un nesso logico
- A seconda della relazione che intercorre tra A e B si hanno diversi tipi di rapporti statistici

Tipologie di rapporti statistici

- Rapporti di composizione (di parte al tutto)
- Rapporti di coesistenza
- Rapporti di derivazione
- Rapporti di densità
- Rapporti di incremento
- Numeri indice (non li studieremo qui!)

Rapporti di composizione

R=100xA/B

- A è una parte di B
- Esempio: A= n. pernottamenti per vacanza
  B= n. totale pernottamenti
  R= Tasso di turismo proprio

Città	*Pernottamenti al 1991*			**TTP**
	A: Vacanza	Altri motivi	B: Totale	**100xA/B**
Lione	222872	2563028	2785900	**8**
Roma	9494633	2523890	12018523	**79**

R esprime quanti *pernottamenti turistici in senso proprio* si hanno ogni 100 pernottamenti

Altro esempio: frequenze relative

Rapporti di coesistenza

R=100xA/B

- A + B = totale
- Esempio: A= n. presenze italiani, B=n. presenze stranieri
  R= rapporto composizione italiani/stranieri

CATEGORIE	*Italiani*	*Stranieri*	**Ita/str**
Alberghi di 5 stelle e 5 stelle lusso	1279328	2724885	46.95
Alberghi di 4 stelle	27369943	30772225	88.94
Alberghi di 3 stelle	67517341	43144304	156.49
Alberghi di 2 stelle	19991704	10832685	184.55
Alberghi di 1 stella	8014643	4450227	180.10
Residenze turistico alberghiere	9122097	5912841	154.28
Esercizi alberghieri	**133295056**	**97837167**	136.24

R : *presenze di turisti italiani* ogni 100 presenze di stranieri

Rapporti di densità

- Particolari rapporti di derivazione
- relazione logica tra A e B tipo 'affollamento'

- Esempio: A= popolazione residente B= superficie

Ripartiz	**A pop**	**B superf (kmq)**	**R=A/B**	**I=B/A**
NO	14984766	57950.05	259	3.87
NE	10694115	61981.40	173	5.80
Centro	10946174	58379.55	188	5.33
Sud isole	20532353	123024.98	167	5.99
TOT	57157408	301335.98	190	5.27

Fonte:ISTAT 2003

R=A/B
abitanti per kmq

I=1000*B/A
Kmq per 1000 abitanti

Rapporti di incremento

- X_t dato al tempo t, X_{t-1} dato al tempo t-1
- $R=(X_t - X_{t-1})/ X_{t-1}$ → variazione nell'intervallo di tempo

- Esempio: arrivi in Europa per anno (Fonte: WTO)

t Anno	X Arrivi (milioni)	R Tasso incremento	tasso medio annuo
1975	153.86		
1980	186.11	20.96	4.19
1985	212.11	13.97	2.79
1990	282.88	33.36	6.67
1995	335.60	18.64	3.73

$R= (X_t - X_{t-5})/ X_{t-5}$
Variazione nel quinquennio

$T= (X_t - X_{t-5})/ (5*X_{t-5})$
Variazione media annua

Rapporti di derivazione

$R=100xA/B$

- A deriva logicamente da B (B *produce* A)
- Esempio: A= n. vacanze 1-3 gg B= popolazione residente

R : *vacanze brevi ogni 1000 residenti*

RIPARTIZIONE GEOGRAFICA	Popolazione residente (migliaia)	viaggi 1-3 notti (migliaia)	viaggi*100 residenti
Nord	25,910	20,399	78.7
Centro	11,046	7,540	68.3
Sud	20,581	10,131	49.2
Italia	**57,537**	**38,069**	**66.2**

Rapporti di derivazione

- A fenomeno di movimento (flusso)
- B fenomeno di stato (stock)
- Problemi:
  - Scegliere B
  - Calcolare media di B
  - Tener conto di altri fattori oltre a B

Ma è proprio vero che le donne guidano meglio? di Enzo Ballatori

http://sis-statistica.it/magazine/