statistica e
calcolo delle probabilità

Rina Camporese

la statistica è SINTESI ...

- osserva un insieme di **manifestazioni individuali**
- per analizzare **fenomeni collettivi**
- servendosi di **sintesi quantitative**

- *n. di sigarette fumate da ogni singolo studente*
- *abitudine al fumo tra gli studenti dell'Università di Padova*
- *n. medio di sigarette fumate al giorno*

... descrive AGGREGATI

“se Toni mangia un pollo e Bepi non mangia niente
per la statistica hanno mangiato mezzo pollo per ciascuno”

credete ancora a questi luoghi comuni?

dire
“Toni e Bepi hanno mangiato in media mezzo pollo a testa”

è molto diverso dal dire

“sia Toni che Bepi hanno gustato mezzo pollo per ciascuno”

la sintesi è RINUNCIA ...

calcolando sintesi statistiche
(ad esempio medie)

SI PERDONO INFORMAZIONI SULLE SINGOLE UNITA' OSSERVATE

Toni	***Bepi***	***Media***
1 pollo	*0 polli*	*1/2 pollo*
1/2 pollo	*1/2 pollo*	*1/2 pollo*
1/3 pollo	*2/3 pollo*	*1/2 pollo*
....	*.....*	*1/2 pollo*
0 polli	*1 pollo*	*1/2 pollo*

… ma è anche GUADAGNO

calcolando sintesi statistiche
(ad esempio medie)

SI GUADAGNANO INFORMAZIONI SULL'INSIEME DELLE UNITA' OSSERVATE

Chiara fuma 3 sigarette al dì, Mario 4, Antonio 6, Giovanna 12, Antonello 0, Luca 0, Mark 5, ... Federica 0

Tra queste 27 persone ci sono 15 fumatori, che consumano in media 5 sigarette al giorno

analizza la VARIABILITA’

in statistica non si è interessati ad appiattire i fenomeni studiati attorno a pochi valori sintetici
es. media

l’obiettivo principale è descrivere con indici sintetici il modo in cui il fenomeno studiato si manifesta nella sua varietà (variabilità)
es. perché alcuni valori si discostano dalla media? come si distribuiscono le unità nel campo di variazione?

la statistica è INDUZIONE ...

- osserva un **insieme ridotto** di unità
- per ottenere informazioni su un **collettivo più ampio**

- *consumo di alcool di un campione di studenti della facoltà di ingegneria*
- *consumo di alcool degli studenti della facoltà di ingegneria*

… estrapola e prevede

Un aereo ha 4 motori
Si rompe il primo: stimati 30 minuti di ritardo in atterraggio

“	*2°*	*1 ora*	*“*
“	*3°*	*2 ore*	*“*
“	*4°*	*4 ore*	*“*

In passato vi siete bevuti anche questa?

l’induzione è un **processo delicato**
deve tener conto del **contesto**
e dei **vincoli naturali** del fenomeno studiato

LE FORMULE MATEMATICHE
NON DEVONO FAR DIMENTICARE IL BUON SENSO

es. stima tasso di accrescimento grandi felini africani
es. previsioni di popolazione a 5,10, 20, 50 anni

la statistica SBAGLIA ...

... ma ne è consapevole

la serietà di una statistica si misura dalla cura riposta nel **calcolare l'errore**

le conclusioni statistiche significative devono essere sempre accompagnate da un certo **grado di incertezza**

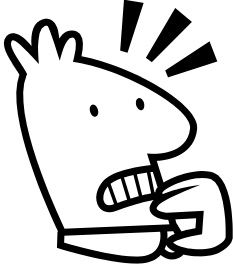
l'abilità nell'usare le tecniche statistiche sta anche nello scegliere quella che comporta l'errore minore

es. campioni di diversa numerosità

es. taratura strumenti di misura

es. diversi modi di fare le domande ottengono risposte diverse

ma la parola statistica significa anche …



Una statistica S è una funzione h
indipendente da parametri incogniti che, riferita alle variabili campionarie $X_1, X_2, \ldots, X_n$ *genera una variabile casuale*

$$S = h(X_1, X_2, \ldots, X_n)$$

Applicata ai valori $x_1, x_2, \ldots, x_n$ *del campione la funzione h assume un valore numerico* $s = h(x_1, x_2, \ldots, x_n)$

Vi siete spaventati?

in parole povere, si intende per STATISTICA un valore calcolato a partire da una serie di dati statistici

ad esempio, una media, una percentuale, …

Ho inserito la definizione formale perché possiate cominciare a osservare la terminologia con la quale farete conoscenza durante questo corso

Non arriverete a studiare la definizione citata sopra, ma cercheremo di farvi acquisire dimestichezza con formule matematiche e concetti statistici

E ORA SI COMINCIA

per prima cosa

bisogna imparare un po' di
LESSICO SETTORIALE

parole nuove o
parole "vecchie" con nuovi significati

CARATTERE e MODALITA’

CARATTERE:

è una caratteristica di interesse

viene rilevata / misurata / osservata

MODALITA’:

manifestazioni del carattere

possono essere **numeriche o non** numeriche

devono essere **esaustive** e **non sovrapposte**

es. il carattere PESO assume la modalità 72 per Maria

a Cagliari diremmo che Maria è grassa? e a Trieste?

es. il carattere COLORE DEGLI OCCHI ha modalità VERDE per Jo

UNITA' STATISTICA

E' L'UNITA' ELEMENTARE SU CUI VENGONO OSSERVATI I CARATTERI

es. il carattere COLORE DEGLI OCCHI ha modalità VERDE per Jo

Chi è l'unità statistica? Jo

può essere "naturale" *es. un essere umano, un'auto*

o convenzionale *es. la famiglia*

POPOLAZIONE

E' L'INSIEME DELLE UNITA' STATISTICHE DI INTERESSE, omogenee rispetto a uno o più caratteri

può essere **statica** se è esattamente definita

es. il parco mezzi dell'ACAP al 31.12.2001

o **in movimento** se varia nel tempo

es. gli studenti della facoltà di ingegneria

può essere **empirica** se tutte le unità che la compongono sono osservabili *es. gli iscritti al circolo del golf nel 1999*

o **teorica** se non tutte le unità sono osservabili

es. i potenziali malati di influenza dell'inverno 2002-2003

può essere **finita** *es. le schede madre prodotte da X nel dic. 2001*

o **infinita** *es. lampadine prodotte da Y nel corso dei prossimi 10 anni*

un esempio ... poco chiaro

Auto rubate per marca	
Fiat	534
Volkswagen	241
Ford	120
Audi	115
Lancia	111
Mercedes	111
Opel	101
Renault	96
Autobianchi	84
BMW	83

Fonte: Quattroruote n. 556 Febbraio 2002

Popolazione: auto rubate
quando? dove?
Unità: singola auto
Carattere: marca
Modalità: Fiat, ..., BMW
e le altre?
Come è stato rilevato il carattere?
Come possiamo usare questi dati?
La graduatoria è falsata da qualche altro carattere?
(es. share di mercato delle varie marche)

un esempio ... più chiaro

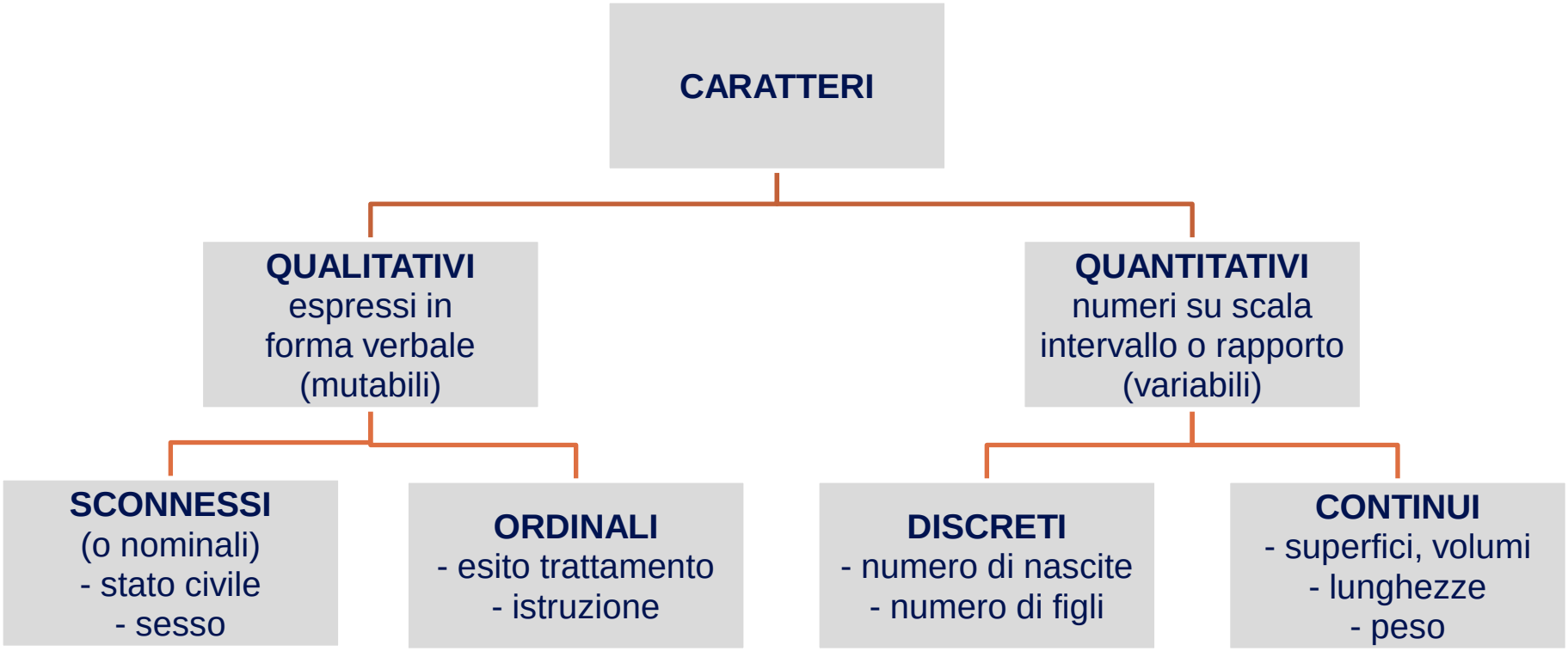
Così le Case si dividono il mercato italiano

	Immatricolazioni Dic. 2001
Italiane	46.046
Estere	85.854
Totale	131.900

Fonte: Quattroruote n. 556 Febbraio 2002

Popolazione: auto immatricolate in Italia nel 2001
Unità: ogni auto immatricolata
Carattere: marca
Modalità: italiana, estera

tanti tipi di CARATTERI



approfondimenti sui CARATTERI (2)

I **QUANTITATIVI** POSSONO ESSERE

discreti: insieme delle modalità in corrispondenza biunivoca con sottoinsieme dei ***numeri interi***

es. n. di viaggi all'estero

continui: insieme delle modalità in corrispondenza biunivoca con sottoinsieme dei ***numeri reali***

es. peso, altezza

approfondimenti sui CARATTERI (3)

scala nominale: date due modalità è possibile solo dire se sono uguali o diverse *es. colore degli occhi*

scala ordinale: è possibile solo dare un ordine (naturale o convenzionale) alle modalità (<>) *es. titolo di studio*

scala a intervalli: non esiste uno zero assoluto non arbitrario *es. temperatura in gradi centigradi*

scala di rapporti: esiste uno zero assoluto “naturale” *es. peso, età*

relazioni tra CARATTERI e SCALE

Tipo di carattere	Esempio	Scala	Operazioni tra modalità
qualitativo sconnesso	*sesso* *colore occhi*	nominale	conteggio uguale / diverso
qualitativo ordinabile	*titolo di studio*	ordinale	conteggio uguale / diverso
quantitativo discreto o senza zero assoluto	*anni calendario* *temperatura*	a intervalli	conteggio uguale / diverso somma / sottrazione
quantitativo continuo o con zero assoluto	*peso* *età* *tempo*	di rapporti	conteggio uguale / diverso somma / sottrazione moltiplic. / divisione

… attenzione ai tranelli …

non tutti i numeri sono quantità !
i CODICI sono dei traditori

indossano la maschera dei numeri

ma il loro significato è qualitativo

se il codice della provincia di Verona è 23
e quello di Rovigo è 29
che provincia sarà la 26 = (23+29) / 2 ?

forse una che si trova a metà strada tra le due? No, è Treviso

I programmi a computer non sanno distinguere
quali siano le operazioni appropriate per i “numeri” memorizzati
sommano, sottraggono, elaborano … tritano qualsiasi cosa abbia aspetto di cifra

la responsabilità di scegliere quali operazioni fare sui “numeri”
non può essere delegata al computer
rimane sulle spalle della persona che calcola le statistiche

DISTRIBUZIONI DI FREQUENZA (1)

si costruiscono raggruppando in classi le **n unità** statistiche secondo le **k modalità** del carattere

ogni unità statistica

- deve poter essere **classificata** (classi esaustive)
- deve appartenere ad **una sola classe** (classi disgiunte)

es. colore dei capelli
es. presenza di una malattia

DISTRIBUZIONI DI FREQUENZA (2)

Così le Case si dividono il mercato italiano

	Immatricolazioni Dic. 2001	*Valori percentuali*
Italiane	46.046	34,9
Estere	85.854	64,1
Totale	*131.900*	*100,0*

modalità

K=2

frequenza **assoluta**

$\mathbf{n_1}+\mathbf{n_2}=n$

frequenza **relativa**

$\mathbf{f_1}=n_1/n$ $\quad \mathbf{p_1}=f_1*100$

$\mathbf{f_2}=n_2/n$ $\quad \mathbf{p_2}=f_2*100$

$f_1+f_2=1$ $\quad \mathbf{p_1}+\mathbf{p_2}=100$

RAPPRESENTAZIONI GRAFICHE (1)

caratteri qualitativi

diagrammi a barre
diagrammi a torta

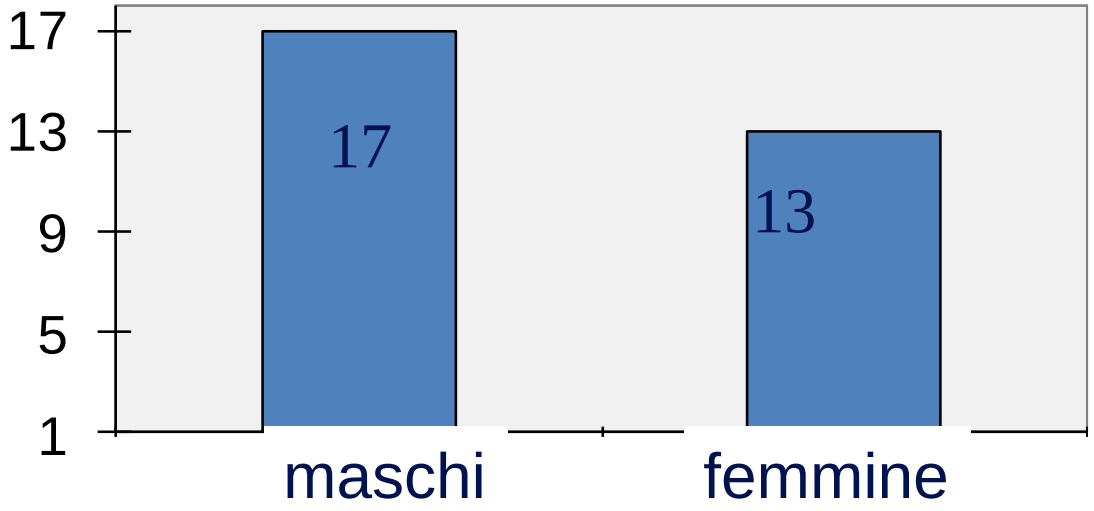
caratteri quantitativi

diagrammi a bastoncini
istogrammi di frequenza

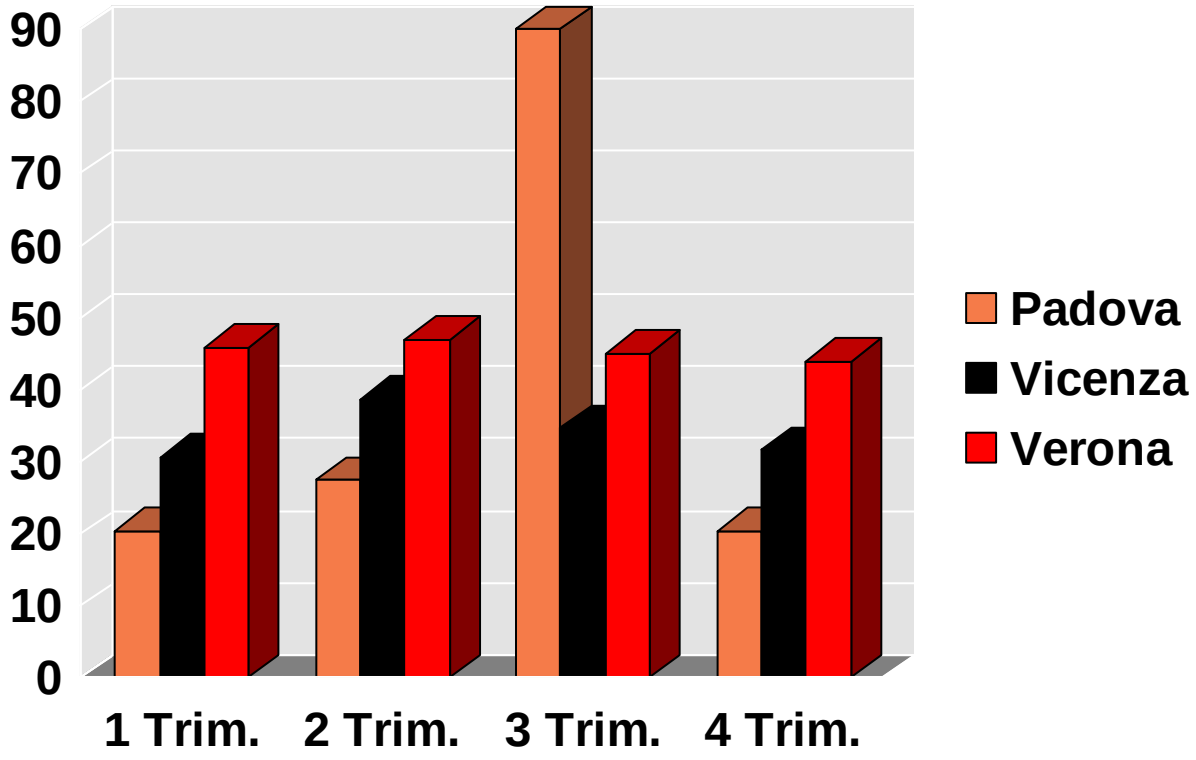
……

UN ESEMPIO SEMPLICE *ma non troppo*

Sesso	***n(i)***	***f(i)***	***p(i)i***
Maschi	17	.57	56.67
Femmine	13	.43	43.33
Totale	*30*	*1.00*	*100.00*



E UNO UN PO’ PIU’ COMPLESSO *ma brutto*



RILEVAZIONE DEI DATI

come si reperiscono le informazioni?

DA CHI SONO DATI I “DATI”?

prima di calcolare delle statistiche sui DATI è necessario
conoscerne la **fonte** e capire
da che **tipo di rilevazione** sono stati prodotti

ci sono tecniche e metodi statistici diversi per le varie tipologie di DATI

rilevazione SPERIMENTALE

LA PRODUZIONE DEI DATI E' **CONTROLLATA** DA CHI GUIDA L'ESPERIMENTO

sono sotto controllo:

- le condizioni dell'esperimento e

(es. coltura batteri a temperatura di 18°)

- le caratteristiche delle unità statistiche scelte

(es. malati di epatite C, maschi)

il controllo avviene:

- direttamente, tramite il **disegno sperimentale** o

(es. esperimenti chimici in laboratorio)

- indirettamente, tramite la **randomizzazione**

(es. somministrazione di due farmaci a due gruppi di pazienti)

rilev. OSSERVAZIONALE

NON SI POSSONO CONTROLLARE LE CONDIZIONI
IN CUI VENGONO “PRODOTTI I DATI”
(es. n. di incidenti stradali a Padova nel 2001)

non sempre si possono controllare le caratteristiche delle unità statistiche

l’osservatore ha un **ruolo passivo**
si limita a osservare / misurare i fenomeni che accadono senza poterli modificare

INDAGINE STATISTICA (1)

E' UNA SITUAZIONE OSSERVAZIONALE

obiettivo: studiare un fenomeno

(es. propensione all'acquisto dei visitatori di un sito internet)

passi da compiere:

- individuare la **popolazione** (chi, quando, dove, ...)

  (es. visitatori del sito nel mese di marzo 2002)

- individuare il **carattere** d'interesse e le sue **modalità**

  (es. acquisto o meno di un certo prodotto SW)

- individuare altri **caratteri collegati**

  (es. sesso, età, reddito, ...)

- individuare il **metodo di misura**

  (es. intervista CAWI)

.... cioè **PROGETTARE L'INDAGINE**

INDAGINE STATISTICA (2)

può essere:

- **TOTALE** o CENSUARIA
  si rilevano i caratteri di **tutte le unità** della popolazione

- **CAMPIONARIA**
  si rilevano i caratteri di un **sottoinsieme di unità** della popolazione
  per **induzione** si ottengono informazioni su tutta la popolazione

INDAGINE CAMPIONARIA

vantaggi:

- **riduzione di tempi e costi**
- **miglior controllo della rilevazione**
- **informazioni più approfondite**

es. censimento della popolazione
es. popolazioni con numerosità molto elevate
es. indagini di mercato sui potenziali acquirenti di un bene
es. casi in cui l'osservazione distrugge l'unità statistica
es. fenomeni sociali, interviste lunghe e dettagliate, ...

LA CASUALITA’

un campione è casuale quando

LA SELEZIONE DELLE UNITA’ E’ AFFIDATA A REGOLE PROBABILISTICHE

il “caso” è un alleato prezioso:

- consente di applicare tecniche di **INFERENZA STATISTICA**
- **EVITA LE DISTORSIONI SISTEMATICHE** causate da selezioni “soggettive”

le DUE STATISTICHE

STATISTICA DESCRITTIVA

sintetizza
il modo di manifestarsi di un fenomeno
nella **popolazione osservata**

STATISTICA INFERENZIALE

attraverso l'osservazione di una sottopopolazione (poche unità)
consente di ottenere informazioni
su una **popolazione più ampia**

distribuzione e rappresentazione dei caratteri

un dettaglio sulle MODALITA'

per le variabili discrete, le modalità si lasciano individuare abbastanza facilmente

ma, per le variabili continue, occorre attuare un processo di

DISCRETIZZAZIONE

cioè individuare delle **CLASSI** di valori assunti dal carattere che ci interessa

es. *distribuzione degli studenti di ingegneria per altezza. Padova. A.a. 2001-2002*

meno di 180 cm	*40*	*studenti*
180 |- 190 cm	*32*	*studenti*
190 |- 200 cm	*8*	*studenti*
200 cm o più	*1*	*studente*

Notate:
dove classifico uno studente alto esattamente 180 cm?

… classi aperte o chiuse?

* nei raggruppamenti in classi deve essere chiaro come sono stati classificati i valori agli estremi, inferiore e superiore

  soprattutto quando si rappresentano caratteri continui attraverso classi con estremi espressi in numeri interi

  es. tonnellate di rifiuti raccolti

  …

  10-20 dove si colloca

  20-30 il valore 20,5 ?

  …

* talvolta, quando l'insieme dei valori del carattere non è esattamente definibile, si tende a non precisare il limite inferiore e il limite superiore, rispettivamente, della prima e dell'ultima classe

  es. fino a 10 addetti

  65 anni o più

qualche informazione sulle CLASSI

il processo di **CLASSIFICAZIONE** deve rispettare certe regole

* **il numero di classi deve essere equilibrato né troppe, né troppo poche**

* **le classi devono, in genere, avere la stessa ampiezza**

* **evitare il più possibile le classi aperte**

es. .....

ritorniamo alle FREQUENZE

per calcolare le frequenze serve la

distribuzione unitaria del carattere

cioè l'elencazione delle modalità osservate unità per unità, nella popolazione di interesse

es. *distribuzione degli studenti di ingegneria per mano dominante Padova. A.a. 2001-2002*

Mario R.	*destrorso*
Luca V.	*destrorso*
.....	*.....*
Nadia G.	*mancina*
.....	*.....*
Andrea S.	*destrorso*

notate: il titolo esplicita spazio e tempo e descrive: popolazione, carattere e unità

le modalità del carattere si deducono dalla lista

ABBIAMO BISOGNO DI SINTESI

partendo dalla distribuzione unitaria calcoliamo le

frequenze assolute delle modalità del carattere
cioè il numero di volte che ciascuna modalità
viene osservata

es. *distribuzione degli studenti di ingegneria per mano dominante Padova. A.a. 2001-2002*

destrorsi	*87*
mancini	*13*
totale	*100*

$$\sum_{j=1}^{k} n_j = n$$

K=2
n1=87, n2=13
n=n1+n2=100

notate: la prima colonna della tavola non contiene più le unità ma le modalità del carattere

e ci servono sintesi SIGNIFICATIVE

partendo dalle frequenze assolute calcoliamo le

frequenze relative cioè il rapporto tra la frequenza j-esima e il totale delle unità

es. *distribuzione degli studenti di ingegneria per mano dominante Padova. A.a. 2001-2002*

destrorsi	*0,87*	f1=0,87 f2=0,13 f1+f2=1
mancini	*0,13*	p1=87% p2=13% p1+p2=100%
totale	*1,00*	

$$\sum_{j=1}^{k} f_j = 1 \qquad \sum_{j=1}^{k} p_j = 100$$

perchè si calcolano le **frequenze relative**?

perchè le frequenze assolute dipendono da **n**, hanno significati diversi a seconda del numero totale di unità

es. *n(j)=13* *è alta se si tratta di giocatori di calcio tra i 15 bambini di una classe elementare (p(j)=0,87)*

n(j)=13 *è bassa se si tratta di vincitori al superenalotto tra i 209.000 cittadini di Padova (p(j)=0,00006)*

il fatto è rilevante soprattutto quando si confrontano le distribuzioni di frequenza di 2 o più gruppi di unità

es.

	Gruppo 1		*Gruppo 2*	
	n(j)	*f(j)*	*n(j)*	*f(j)*
A	*2*	*0,333*	*12*	*0,207*
B	*4*	*0,667*	*46*	*0,793*
Tot.	*6*	*1,000*	*58*	*1,000*

come si rappresentano le FREQUENZE?

attraverso la **distribuzione di frequenze**

X	Frequenza assoluta	Frequenza relativa	Frequenza relativa percentuale
A_1	n_1	f_1	p_1
A_2	n_2	f_2	p_2
.....	.....	.....	.....
Aj	**nj**	**fj**	**pj**

N.B. *questa simbologia vi deve diventare familiare, osservatela nel dettaglio, con attenzione ... cos'è X? cos'è Aj ?*

una frequenza nuova: la CUMULATA

partendo dalle frequenze semplici si calcolano le

frequenze cumulate

sommando le frequenze della modalità j
a tutte le frequenze delle modalità precedenti

es. *carattere X*

	n(j)	*N(j)*	*F(j)*	*P(j)*
modalità 1	*46*	*46*	*0,46*	*46%*
modalità 2	*13*	*59*	*0,59*	*59%*
modalità 3	*41*	*100*	*1,00*	*100%*
totale	*100*	*-*	*-*	*-*

$$N_j = \sum_{i=1}^{j} n_j$$

$N_1=n_1$ $F_1=f_1$ $P_1=p_1$

$N_k=n$ $F_k=1$ $P_k=100$

quando si usa la frequenza CUMULATA?

ha senso quando le modalità del carattere sono

ORDINABILI

altrimenti, possiamo anche calcolarla, ma non ci dice un gran che

es. *Attesa in fila prima di poter entrare al museo dell'Accademia di Venezia. Estate 2001*

	P(j)
meno di 10 min.	*46*
da 10 a 20 min.	*58*
da 20 a 30 min.	*76*
più di 30 minuti	*100*

il 76% dei visitatori attende meno di 30 minuti prima di entrare al museo

domanda ...

sapreste intuire cos’è una frequenza

RETROCUMULATA?

la FUNZIONE DI RIPARTIZIONE

quando un carattere è quantitativo le formule si complicano, ma la sostanza rimane la stessa

dato un carattere quantitativo (discreto o continuo) X

suddiviso in k classi (X_0-X_1, X_1- X_2 , ... , X_{k-1}-X_k)
le frequenze relative cumulate si possono rappresentare tramite la funzione di ripartizione

$$F(x) = \begin{cases} 0 & x < x_0 \\ F_1 & x_0 <= x < x_1 \\ .... & \\ F_k & x_{k-1} < x_k \\ 1 & x >= x_k \end{cases}$$

F_1 ... F_k
sono funzioni di X

non frequenze cumulate

PROPRIETA' della funzione di ripartizione

*** F(x) = 0 per $x < x_1$**
*** F(x) = 1 per $x > x_k$**
*** F(x) è una funzione non decrescente**

Funzione di ripartizione di una variabile continua

X	frequenza relativa cumulata
0	0
10	0
20	0,65
30	0,85
40	0,95
50	1

frequenza relativa cumulata

Variabile X

a che può servire ?

la funzione di ripartizione

può servire a capire come si distribuiscono le frequenze, se si concentrano verso le modalità più basse o più alte del carattere analizzato

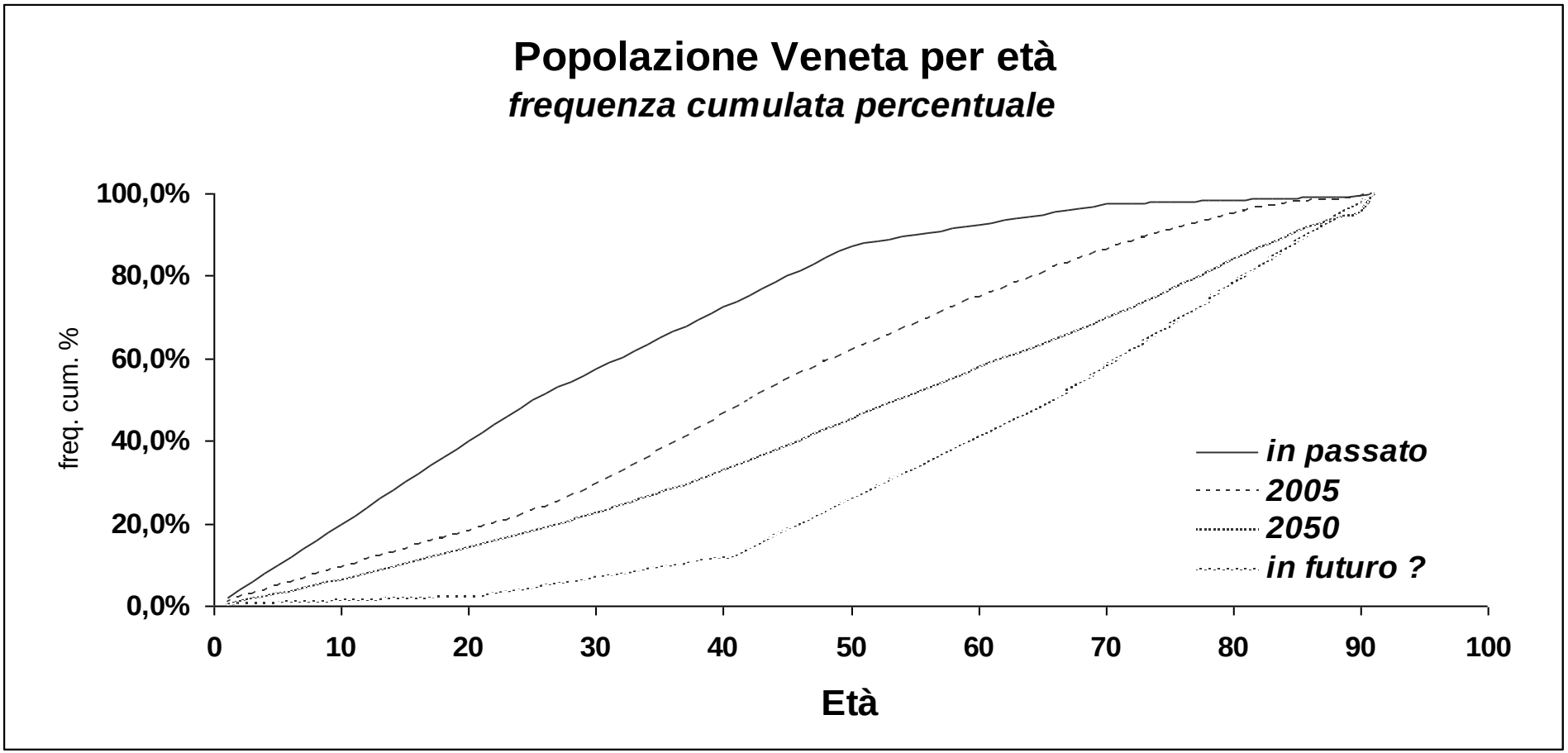
sicuramente la funzione di ripartizione del numero di componenti le famiglie italiane è cambiata nel tempo

sono diminuite le famiglie numerose e aumentate quelle composte di una sola persona

la si può anche rappresentare in grafico

e il suo andamento ci dice qualcosa sul fenomeno rappresentato

ci torneremo quando affronteremo i grafici



e se le variabili sono 2 ?

per ciascuna unità della popolazione si rilevano due caratteri *(es. sesso ed età)*

dati due caratteri, si definisce

TABELLA A DOPPIA ENTRATA

l'insieme delle frequenze delle unità che presentano congiuntamente

la modalità i-esima del primo carattere e
la modalità j-esima del secondo carattere

una generica
TABELLA A DOPPIA ENTRATA

				Y			
		b1	….	bj	….	bk	**tot.**
	a1	n11	….	n1j	….	n1k	**n1.**
	….	….	….	….	….	….	**….**
X	ai	ni1	….	nij	….	nik	**n1.**
	….	….	….	….	….	….	**….**
	ah	nh1	….	nhj	….	nhk	**nh.**
	tot.	**n.1**	**….**	**n.j**	**….**	**n.k**	**n..**

N.B. *chi ha dimestichezza con le matrici non dovrebbe aver problemi per gli altri serve qualche esempio*

le distribuzioni MARGINALI

sono due, ai MARGINI della tavola

sono la riga e la colonna dei TOTALI

rappresentano le DUE DISTRIBUZIONI SEMPLICI

dei caratteri X e Y presi singolarmente

distribuzioni CONDIZIONATE

sono le RIGHE e le COLONNE INTERNE alla tabella

* **alcune sono CONDIZIONATE dalla modalità i-esima che individua la riga**

  fissata la modalità i-esima del carattere X osservo la distribuzione del carattere Y

* **le altre sono CONDIZIONATE dalla modalità j-esima che individua la colonna**

  fissata la modalità j-esima del carattere Y osservo la distribuzione del carattere X

esempi ...

un esempio ...

es. Individui per sesso e tipo di musica preferita

	Uomini		***Donne***		***Totale***	
	n	*p*	*n*	*p*	***n***	***p***
nessuna	*3*	*5,9*	*1*	*4,0*	***4***	***5,3***
solo classica	*2*	*3,9*	*0*	*-*	***2***	***2,6***
solo rock	*36*	*70,6*	*16*	*64,0*	***52***	***68,4***
entrambe	*10*	*19,6*	*8*	*32,0*	***18***	***23,7***
totale	***51***	***100,0***	***25***	***100,0***	***76***	***100,0***

che valore ha n (1.)?
e p (2,2)?

… tradotto in struttura di una tabella

Titolo della tabella		
	Carattere X	
Carattere Y	Riga madre	Totale
Colonna madre	Valori della variabile doppia	Colonna marginale
Totale	Riga marginale	Totale generale

come si calcolano le marginali ?

se **X** a **h** modalità e **Y** ha **k** modalità
le frequenze marginali si calcolano come segue:

$$n_{i.} = \sum_{j=1}^{k} n_{ij}$$ **per i = 1, ..., h**

per j = 1, ..., k $$n_{.j} = \sum_{i=1}^{h} n_{ij}$$

e il numero totale di unità ?

se **X** a **h** modalità e **Y** ha **k** modalità
il numero totale di unità è definito da:

$$n_{..} = \sum_{i=1}^{h} \sum_{j=1}^{k} n_{ij} = \sum_{i=1}^{h} n_{i.} = \sum_{j=1}^{k} n_{.j}$$

sommo per riga e per colonna tutte le celle

sommo tutte le h marginali di riga

sommo tutte le k marginali di colonna

medie
analitiche e di posizione

analisi della DISTRIBUZIONE di un carattere

*** attraverso “sintesi”**
che evidenzino le caratteristiche essenziali

*** le sintesi [matematiche] servono**
a “far parlare i dati”
a estrarne il succo

*** le statistiche più conosciute sono i**

VALORI CENTRALI

VALORI CENTRALI

min <= VALORE CENTRALE <= max

*** il più famoso è la media**

*** ne esistono molti altri**

*** ognuno con le proprie peculiarità**

es. qual è il colore medio dei vostri capelli ?

MEDIE ANALITICHE

(dette anche algebriche)

*** si applicano solo a caratteri quantitativi**

*** si lavora sulle frequenze (n_i)**

*** e sulle modalità del carattere (x_i)**

MEDIE DI POSIZIONE

*** si adattano anche a caratteri di tipo qualitativo**

*** si lavora sulle frequenze (n_i)**

e NON sulle modalità del carattere (x_i)
le modalità del carattere non sono trattate algebricamente

es. anno di calendario

MEDIA ARITMETICA **semplice**

* **è analitica**
* **dato un insieme di n valori x_1, x_2, ..., x_n di un carattere quantitativo X**

$$M_a = \frac{1}{n}(x_1 + x_2 + \ldots\ldots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

si fanno 2 operazioni:

1. **sintesi** — **somma degli x_i**
2. **significativa** — **si divide per n**
   ci si rende indipendenti da n

perchè si calcola la **media aritmetica**?

perchè **dobbiamo fare dei confronti**

* **tra due o più gruppi**

es. gli studenti di Feltre sono più alti di quelli di Rovigo?

* **o con un modello di riferimento**

es. gli studenti di Treviso sono alti?

problema ....

come mai gli studenti hanno vita più breve dei suonatori di jazz?

droga e alcol sono più salutari dei libri?

no! stiamo confrontando medie di due gruppi che non si possono confrontare, perché per diventare jazzisti famosi bisogna essere sopravvissuti almeno fino ai trent'anni di età

non si può confrontare l'età media alla morte di vari gruppi di popolazione selezionati per sopravvivenza

(è difficile morire da studenti a 50 anni di età)

attenti agli scivoloni …

si calcola per un insieme di valori
di una caratteristica qualitativa

non si calcola la media di codici numerici associati a
significati qualitativi

individui	*stato civile*	
Mario	*1*	*celibe*
Antonio	*2*	*coniugato*
Giulia	*1*	*nubile*
Roberto	*4*	*vedovo*
Anna	*3*	*divorziata*

stato civile medio = (1+2+1+4+3) / 5 = 2,2 = un po' più che coniugato

media aritmetica in TABELLE DI FREQUENZA

$$M_a = \frac{1}{n}\sum_{j=1}^{k} x_j n_j$$

k = n. di modalità del carattere X

es. *n. di figli per donna*

k = 4

x_j		n_j		$x_j * n_j$	
0	x_1	*10*	n_1	*0*	$n_1 * x_1$
1		*20*		*20*	
2		*15*		*30*	
3	x_4	*3*	n_4	*9*	$n_4 * x_4$
Tot.		*48*	n	*59*	$\sum_{j=1}^{k} x_j n_j$

quale strana mutazione ha prodotto 1/3 di figlio?

Ma = 59 / 48 = 1,2

media aritmetica per frequenze IN CLASSI

$$M_a \cong \frac{1}{n}\sum_{j=1}^{k} c_j n_j$$

k = n. di classi
c(j) = valore centrale della classe

es.

Età	n(j)	c(j)	c(j)n(j)
0-9	10	5	50
10-19	20	15	300
20-29	10	25	250
Tot.	40		600

è meglio 4,5 ? cioè (0+9) / 2

Ma = 600/40 = 15 circa

es.

Età	n(j)	c(j)n(j)
0	10	0
10	20	200
20	10	200
Tot.	40	400

Ma = 400/40 = 10 esattamente

attenzione agli estremi …

Se dovete calcolare il numero medio giornaliero di utenti della biblioteca comunale, potete contare per una settimana quanti utenti vi entrano

giorno	*utenti*	*giorno*	*utenti*
Lun	*15*	*Ven*	*16*
Mar	*13*	*Sab*	*45*
Mer	*11*	*Dom*	*9*
Gio	*12*		

la media è (15 + 13 + 11 + 12 + 16 + 45 + 9) / 7 = 17,3

ma il sabato è arrivata una comitiva di un Paese vicino e molti dei visitatori hanno voluto visitare la biblioteca

il valore 45, che spicca tra gli altri, può essere considerato un valore "anomalo", che falsa un po' la media dei visitatori "normali"

è preferibile, allora, escluderlo dal calcolo della media che diventa

M = (15 + 13 + 11 + 12 + 16 + 9) / 6 = 12,6

le cose cambiano molto se state cercando di valutare quanto spazio, tavoli e sedie mettere a disposizione degli utenti

prima di effettuare qualsiasi calcolo sui vostri dati, osservateli e cercate di capire se contengono VALORI ANOMALI

PROPRIETA' della media aritmetica (1)

$$\sum_{i=1}^{n} x_i = n \cdot M_a$$

ridividendo in parti uguali la somma del carattere, tutte le unità riceverebbero una quantità pari alla media aritmetica

da questa proprietà è scaturita la barzelletta del pollo, che fa proprio ridere perchè NON TUTTI I CARATTERI SONO REDISTRIBUIBILI

PROPRIETA' della media aritmetica (2)

$$\sum_{i=1}^{n} (x_i - M_a) = 0$$

la somma algebrica degli scarti dalla loro M_a è uguale a zero: PROPRIETA' di EQUILIBRIO

* **gli scarti dalla media si compensano alcuni sono positivi, altri negativi**

* **per questo la media aritmetica si considera un VALORE CENTRALE**

PROPRIETA' della media aritmetica (3)

$$\sum_{j=1}^{k} (x_j - M_a)^2 n_j \leq \sum_{j=1}^{k} (x_j - A)^2 n_j \qquad \forall A \neq M_a$$

la somma algebrica
dei quadrati degli scarti
è minima
quando gli scarti sono calcolati da M_a

PROPRIETA' dei MINIMI QUADRATI

perchè si calcolano i **quadrati degli scarti** ?

*** la potenza al quadrato rende positivi**
anche gli scarti negativi

*** ed enfatizza gli scarti maggiori**

es. misure ripetute di uno stesso fenomeno fisico

PROPRIETA' della media aritmetica (4)

$$M_{tot} = \frac{\sum_{i=1}^{L} M_i n_i}{\sum_{i=1}^{L} n_i}$$

la media artimetica complessiva di più gruppi parziali è uguale alla media artimetica ponderata delle medie parziali

es. numero medio di sigarette fumate dai 4 gruppi di studenti

	PD	*RO*	*TV*	*FE*	*tot*
n.medio sig.	*14*	*15*	*20*	*18*	
n. studenti	*20*	*18*	*24*	*16*	*78*

*M tot = (14 * 20 + 15 * 18 + 20 * 24 + 18 * 16) / 78 = 17*

PROPRIETA' della media aritmetica (5)

*** dal punto di vista matematico**
la media aritmetica può essere considerata come un OPERATORE LINEARE con le seguenti proprietà:

$$M_a(kx) = kM(x)$$

k costante condizione di omogeneità

$$M_a(x+y) = M(x) + M(y)$$

condizione di additività

PROPRIETA' della media aritmetica (6)

*** se y è una trasformazione lineare di x:**

$$y = a + bx \qquad M_a(y) = a + bM(x)$$

*** la trasformazione inversa di una trasformazione lineare è anch'essa lineare:**

$$x = \frac{y - a}{b} \qquad M_a(x) = \frac{M(y) - a}{b}$$

perchè si calcolano **le trasformazioni** ?

ad esempio, per lavorare più agevolmente con dati “scomodi”

es. fatturato medio in migliaia di euro

x(i)	*150.000*	*.....*	*850.000*
y(i)=x(i)/1.000	*150*	*.....*	*850*

M(y) = 350
*M(x) = M(y) * 1.000 = 350.000*

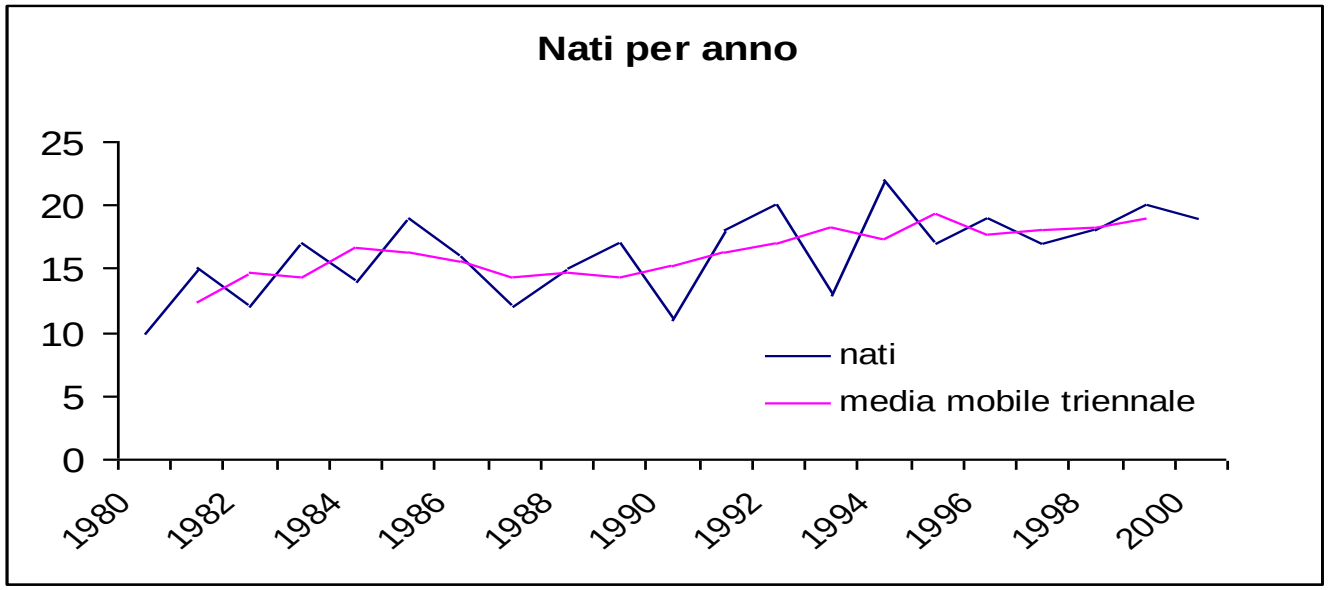
le MEDIE MOBILI

quando si analizzano serie storiche di dati

talvolta si pone il problema di renderle stabili

perché le misure possono fluttuare molto nel tempo e

può essere difficile leggerne l'andamento sottostante



come si calcolano …

vediamolo con un esempio

Anni	*Nati*	*Media mobile triennale*	
1980	*10*	*-*	
1981	*15*	*12*	*= (10 + 15 + 12) / 3*
1982	*12*	*15*	
1983	*17*	*14*	
1984	*14*	*17*	
1985	*19*	*16*	
1986	*16*	*…*	
…	*…*	*…*	

MEDIA GEOMETRICA

*** è algebrica** $M_g = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}$

$$M_g = \sqrt[n]{x_1^{n_1} \cdot x_2^{n_2} \cdot \ldots \cdot x_k^{n_k}}$$

*** si chiama geometrica perché**
quando si applica ad una progressione geometrica
con un numero dispari di termini
fornisce il termine centrale della progressione stessa

es. Mg (3,6,12,24,48) = 12

quando si usa **la media geometrica** ?

*** per grandezze il cui comportamento è moltiplicativo**

*** es. in statistica economica e demografia**

es. tassi medi di incremento della popolazione o dei prezzi

N.B. gli x(i) non assumono valori “assoluti”
ma esprimono il rapporto tra due grandezze

PROPRIETA' della media geometrica (1)

$$\prod_{i=1}^{n} x_i = \left[M_g \right]^n$$

è analoga alla proprietà
della media aritmetica
ma
anzichè la funzione somma
qui troviamo la funzione prodotto

PROPRIETA' della media geometrica (2)

$$\sum_{i=1}^{n} (\log x_i - \log M_g) = 0$$

anche questa
è analoga alla proprietà
della media aritmetica

PROPRIETA' della media geometrica (3)

$$\log M_g = \frac{1}{n}\sum_{i=1}^{n} \log x_i$$

* è comoda quando n è elevato
  la radice n-ma diventa complicata

* la base dei logaritmi è scelta a piacere

PROPRIETA' della media geometrica (4)

*** se y è una trasformazione di x:**

$$y = a \cdot x^{b}$$ **a e b costanti, a>0**

$$M_g(y) = a \, M_g(x)^{b}$$

a volte fa comodo lavorare su trasformate dei dati originali

PROPRIETA' della media geometrica (5)

$$M_g\left(\frac{a_1}{b_1}, \frac{a_2}{b_2}, \ldots, \frac{a_n}{b_n}\right) = \frac{M_g(a_1, a_2, \ldots, a_n)}{M_g(b_1, b_2, \ldots, b_n)}$$

la media geometrica di una serie di rapporti
è uguale al rapporto fra
la media geometrica dei numeratori e
la media geometrica dei denominatori

che ci crediate o no,
le formule complicate rendono la vita più semplice

TRIMMED MEAN

*** la Tm al 50% di un carattere quantitativo è la media aritmetica del 50% dei valori più centrali di un insieme di osservazioni**

*** medie aritmetica e geometrica sono molto sensibili ai valori estremi**

*** talvolta i valori estremi sono errori o vengono deliberatamente trascurati**

es. se producessi giacconi per i maestri di sci di Cortina
e tra loro ci fosse un pigmeo,
farei prendere freddo alla pancia a tutti gli altri?

MEDIE di POSIZIONE

*** si possono calcolare anche per caratteri qualitativi**

*** non si basano su calcoli algebrici**

*** cercano le modalità del carattere che occupano certe particolari posizioni nella distribuzione**

MEDIANA: è la modalità centrale

*** può essere calcolata solo per caratteri ordinabili**

M_e è la <u>modalità</u> dell'unità centrale

cioè quell'unità che divide le n unità
in due gruppi di pari numerosità

ovvero

si chiama mediana in una successione ordinata di dati
il termine che è preceduto e seguito
dallo stesso numero di dati

COME SI CALCOLA la mediana

*** ordinare i dati in senso crescente** *(3,7,8,10,12,14)*

*** verificare se n è pari o dispari** *(n pari)*

*** individuare la posizione centrale** *(terza e quarta)*

*** - se n è dispari** $M_e = x_{\frac{(n+1)}{2}}$

- se n è pari $M_e = x_{\frac{n}{2}}$ **e** $M_e = x_{\frac{(n+1)}{2}}$

-- se x è quantitativo

$$M_e = \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{(n+1)}{2}} \right)$$

e per una distribuzione di frequenze ?

si utilizzano le frequenze cumulate percentuali e ci si ferma quando si individua il 50%

Età mediana dei bimbi che frequentano i campi estivi

Età	*n(i)*	*N(i)*	*P(i) = N(i) / n * 100*
6	*7*	*7*	*14,9 %*
7	*11*	*18*	*38,3 %*
8	*13*	*31*	*66,0 %*
9	*12*	*43*	*91,5 %*
10	*3*	*46*	*97,9 %*
11	*1*	*47*	*100 %*
Tot	*47*		

l'età mediana dei bimbi è 8 anni
cioè, se potessimo ordinare i 47 bimbi per età in anni, mesi e giorni
il bimbo a metà della fila avrebbe 8 anni
(sarebbe il 24-mo)

e per classi di modalità ?

si segue lo stesso sistema

Aziende agricole per numero di capi bovini

Classi di capi	n(i)	N(i)	P(i) = N(i) / n * 100
1-2	284	284	9,1%
3-5	341	625	20,0%
6-9	226	851	27,3%
10-19	338	1.189	38,1%
20-49	694	1.883	60,4%
50-99	610	2.493	79,9%
100-499	548	3.041	97,5%
500-999	65	3.106	99,6%
1000-1999	12	3.118	99,9%
2000+	2	3.120	100,0%
	3.120		

la classe mediana di bovini è 20-49

se ordinassimo le 3.120 aziende in base al numero di bovini le due aziende a metà avrebbero tra 20 e 49 capi (sarebbero la 1.560-ma e la 1.561-ma)

“più della metà delle aziende ha meno di 50 capi bovini”

CARATTERISTICHE della mediana

*** è meno sensibile ai valori estremi**

*** per le distribuzioni con valori anomali ed estremi la mediana è preferibile alla media aritmetica**

es. redditi, investimenti
casi in cui i numeri sono “grandi” e molto diversi tra loro

PROPRIETA' della mediana (1)

$$\sum_{i=1}^{n} |x_i - M_e| \leq \sum_{i=1}^{n} |x_i - c| \qquad \forall\, c \neq M_e|$$

*** la mediana rende minima la somma dei valori assoluti degli scarti**

come scegliere i punti in cui collocare le fermate dell'autobus?

MODA: è la modalità più frequente

*** può essere calcolata per qualsiasi tipo di carattere anche se qualitativo sconnesso**

*** è la MODALITA' (o la classe di modalità) PIU' FREQUENTE**

NON la frequenza più alta

*** non dice nulla sulle altre modalità e frequenze**

es.

tipo di gonna	*mini*	*corta*	*midi*	*longuette*	*lunga*
donne che la portano	*10%*	*10%*	*40%*	*10%*	*30%*

sono di moda le gonne al ginocchio?

distribuzioni BIMODALI o plurimodali

tipo di gonna	*mini*	*corta*	*midi*	*longuette*	*lunga*
donne che la portano	*10%*	*10%*	*40%*	*10%*	*30%*

ma la moda è varia, le signore preferiscono le gonne al ginocchio
mentre le ragazze amano quelle lunghe

*** la presenza di due mode di solito indica che**
la popolazione osservata non è omogenea
ma composta da due gruppi che differiscono
per un carattere (non osservato)
molto legato alla variabile osservata

es. malattia di Hodgkin

media, moda e mediana vanno d'accordo? in distribuzioni unimodali

*** se c'è simmetria** $M_a = M_e = M_o$

*** se c'è asimmetria positiva** $M_o <= M_e <= M_a$

*** se c'è asimmetria negativa** $M_a <= M_e <= M_o$

se c'è più di una moda, la media aritmetica e la mediana possono non essere significative

es. Tempo impiegato per andare a scuola o al lavoro

minuti	*n(i)*	
meno di 15	*25*	←
15-29	*2*	
30-44	*5*	
45-59	*21*	←
60 +	*12*	⇠

qui la moda è "meno di 15"
ma anche la classe "45 – 59" non scherza

questa tabella potrebbe descrivere la situazione di un paese piccolo e un po' isolato
in cui i bimbi impiegano poco tempo ad andare a scuola
mentre gli adulti che lavorano fuori comune
hanno bisogno di almeno tre quarti d'ora per arrivare al lavoro

la media di 37 minuti non rappresenta quasi nessuno degli individui osservati

prima di calcolare misure statistiche
cercate di osservare come si distribuiscono le frequenze
per far questo i grafici sono utilissimi, li vedremo più avanti …

un paio di esempi …

Aziende agricole per numero di capi bovini

Classi di capi	*n(i)*
1-2	284
3-5	341
6-9	226
10-19	338
20-49	694
50-99	610
100-499	548
500-999	65
1000-1999	12
2000+	2
	3.120

la classe modale
è 20-49
perché vi corrisponde
la frequenza più
elevata

la moda non è 694 !

Abitanti per titolo di studio

Titolo di studio	*n(i)*
Laurea	43
Media sup.	241
Media inf.	563
Elementare	338
Nessuno	93

il titolo di studio
modale è la
licenza media inf.

QUARTILI

*** dividono la distribuzione in 4 parti di uguale numerosità**

*** il 2° quartile è la mediana**

*** il 1° e il 3° quartile sono, rispettivamente, la mediana della prima e della seconda metà dei dati**

*** il 4° quartile è l'ultimo termine**

i quartili NON sono EQUIDISTANTI

*** dipendono dalle frequenze**

es.

x(i):	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
f(i):	*0,1*	*0,1*	*0,1*	*0,3*	*0 ,0*	*0,1*	*0,1*	*0,2*	*0,0*
F(i):	*0,1*	*0,2*	*0,3*	*0,6*	*0,6*	*0,7*	*0,8*	*1,0*	*1,0*

NO	*SI*
Q1 = 3	*Q1 = 3*
Q2 = 5	*Q2 = 4*
Q3 = 7	*Q3 = 7*
Q4 = 9	*Q4 = 9*

una domanda …

chi indovina cosa sono i DECILI ?

e i CENTILI ?

e i QUANTILI ?

le VARIAZIONI PERCENTUALI

si utilizzano per valutare quanto siano variate nel tempo le entità di interesse

anni	utenti in biblioteca	var. % rispetto all'anno precedente	
1995	*65*	-	
1996	*72*	*+11% **	*= (72 – 65) / 65 * 100*
1997	*63*	*-12% ***	*= (63 – 72) / 72 * 100*
1998	*75*	*+16%*	*= (75 – 63) / 63 * 100*
1999	*80*	*…*	*…*
2000	*82*	*…*	
2001	*86*	*…*	

** nel '96 si è osservato un aumento dell' 11% rispetto al 1995*

*** nel '97 c'è stato un calo del 12% rispetto al 1996*

* hanno un segno (+ o -)
che indica aumento o diminuzione

i numeri INDICI (a base fissa)

si utilizzano per evidenziare i percorsi di crescita o diminuzione di misure che variano nel tempo

in questi casi il confronto è più facile se tutte le misure sono espresse in indici
calcolati a partire da una base comune
che si pone pari a 100 (è un punto di riferimento comprensibile)

anni	utenti in biblioteca	indice con base 1995	
1995	*65*	*100*	*= 65 / 65 * 100*
1996	*72*	*111 **	*= 72 / 65 * 100*
1997	*63*	*97 ***	*= 63 / 65 * 100*
1998	*75*	*115*	*= 75 / 65 * 100*
1999	*80*	*123*	*= 80 / 65 * 100*
2000	*82*	*126*	*= 82 / 65 * 100*
2001	*86*	*132*	*= 86 / 65 * 100*

avete notato come in questo caso variazione percentuale e indice conducano allo stesso risultato?

** nel '96 si è osservato un aumento dell' 11% rispetto al 1995*

*** nel '97 c'è stato un calo del 3% rispetto al 1995*

su che base costruire gli indici ?

* l’anno di riferimento non deve essere troppo lontano nel tempo, nel frattempo potrebbero essere intervenuti cambiamenti profondi del fenomeno analizzato

* l’anno di riferimento non deve avere avuto eventi anomali per quanto riguarda il carattere oggetto di studio

nell’esempio precedente
non si può usare come base
l’anno in cui la biblioteca è stata
chiusa per tre mesi per ristrutturazione

i RAPPORTI

si utilizzano per confrontare coppie di misure che sono molto influenzate le une dalle altre

ad esempio, la spesa comunale per i servizi scolastici e il numero di studenti

anni	studenti	spesa in Euro	spesa procapite	
1995	*65*	*1.090*	*17*	*= 1.090 / 65 * 100*
1996	*72*	*1.200*	*17*	*= 1.200 / 72 * 100*
1997	*63*	*1.100*	*17*	*= 1.100 / 63 * 100*
1998	*75*	*1.350*	*18*	*= 1.350 / 75 * 100*
1999	*80*	*1.470*	*18*	*= 1.470 / 80 * 100*
2000	*82*	*1.475*	*18*	*= 1.475 / 82 * 100*
2001	*86*	*1.560*	*18*	*= 1.560 / 86 * 100*
	+32%	*+43%*	*+6%*	

tra il 1995 e il 2001

le DIFFERENZE

si utilizzano per confrontare coppie di misure che derivano dal “bilanciamento” di componenti positive e negative

ad esempio, il saldo naturale, il saldo migratorio e il saldo complessivo di popolazione

non li spiego perché credo che
vi escano ormai dalle orecchie
e perché li sentirete nominare spesso durante il corso

… generalizziamo …

in generale, le misure che abbiamo appena descritto
e che molti di voi già conoscono
appartengono al gruppo dei

RAPPORTI STATISTICI

che consentono di sintetizzare in un’unica misura
due fenomeni diversi
legati tra loro da un nesso logico
mettendoli a confronto
tramite l’operazione di divisione

variabilità e concentrazione

la VARIABILITA’

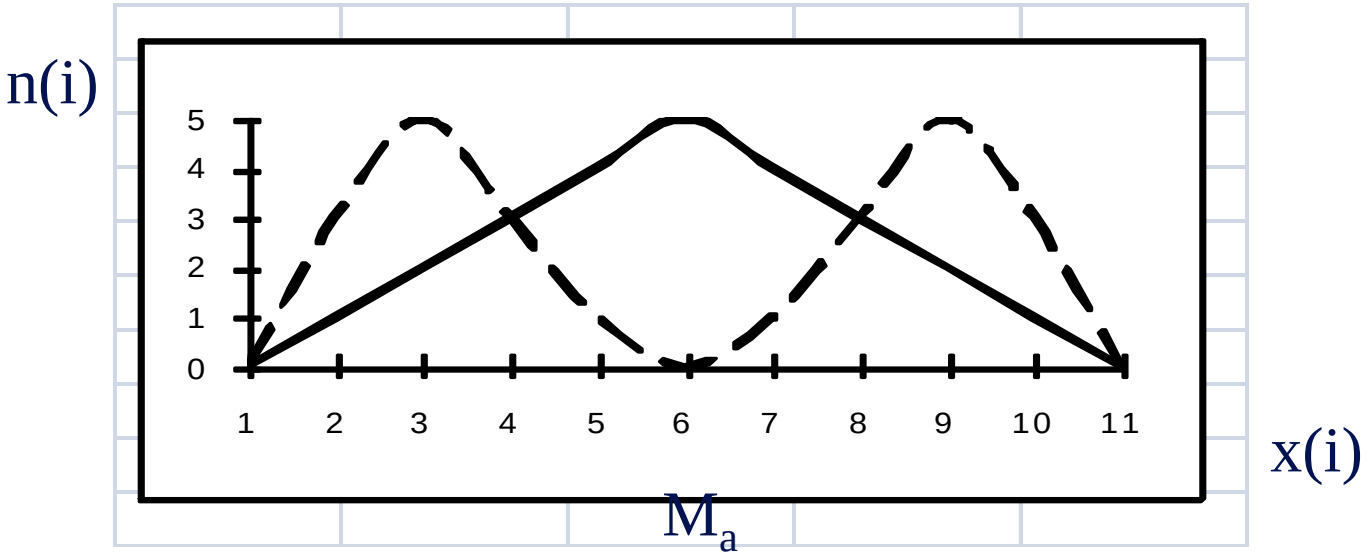
una MEDIA NON BASTA a dare un’immagine della distribuzione di un carattere

è come fotografare un essere umano da due metri di distanza centrando l’obiettivo sull’ombelico;
se la persona è piccola potremo rappresentarla tutta in foto,
se è alta, cioè se le estremità (i valori estremi) sono lontane dal centro (media),
la fotografia non ci farà capire chi è la persona fotografata

servono MISURE di DISPERSIONE

che descrivono la DIVERSITA' delle unità osservate, cioè la loro ETEROGENEITA'

es. due distribuzioni con media aritmetica uguale, ma molto diverse tra loro



la VARIABILITA’

Def. **la VARIABILITA’ di una distribuzione esprime la tendenza delle unità ad assumere diverse modalità del carattere**

Def. **un INDICE DI VARIABILITA’**
deve assumere il suo valore minimo
se e solo se tutte le unità della distribuzione presentano uguale modalità del carattere,
l’indice deve aumentare
all’aumentare della diversità tra le modalità assunte dalle varie unità

queste def. valgono
per caratteri qualitativi e quantitativi

CAMPO DI VARIABILITA’
o di variazione o Range

dato un insieme di n valori x_1, x_2, ..., x_n ordinati in senso crescente, R è la differenza tra il più grande e il più piccolo dei valori

$$R = x_1 - x_n$$

min: 0

max: dipende dal carattere

*** si adatta a caratteri quantitativi**
*** semplice da calcolare**
*** molto sensibile ai valori estremi (poco affidabile)**

un esempio ...

campioni da tre catene di produzione di pacchi di zucchero da mezzo Kg

Catena 1	Catena 2	Catena 3
Ma 495 gr	Ma 500 gr	Ma 505 gr
R 10 gr	R 100 gr	R 10 gr

quale catena è “migliore”?

e per chi? per il produttore o per il consumatore?

DIFFERENZA INTERQUARTILE

Def. **dato un insieme di n valori x_1, x_2, ..., x_n la DIFFERENZA INTERQUARTILE è la differenza tra il terzo e il primo quartile**

$$W = Q_3 - Q_1$$

min: 0

max: dipende dal carattere

* **è il campo di variazione per il 50% delle unità centrali, quelle più vicine alla mediana**

* **è meno sensibile ai valori estremi**

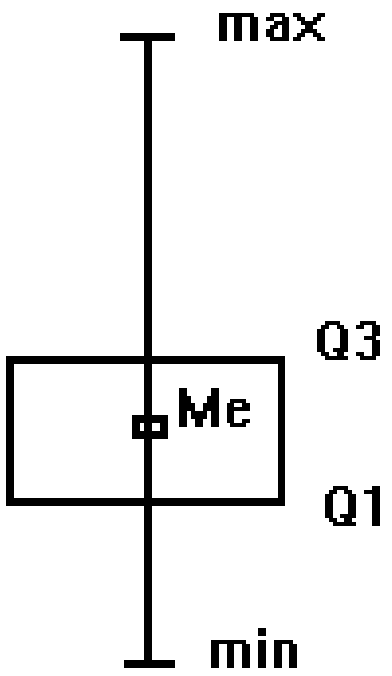
BOX PLOT

è un GRAFICO caratterizzato da 3 elementi:

a. un PUNTO che indica la posizione centrale (di solito la mediana)

b. un RETTANGOLO (box) di altezza legata alla variabilità dei valori "prossimi alla media" (scarto interquartile)

c. 2 SEGMENTI che partono dai lati del rettangolo e i cui estremi sono determinati in base ai valori estremi della distribuzione



indici basati sullo SCOSTAMENTO DA UNA MEDIA

**** gli indici visti finora si basavano su indicatori di POSIZIONE (valore minimo e massimo, mediana, quartili)***

*** per caratteri quantitativi gli indici più "famosi" si basano sugli SCARTI DALLA MEDIA ARITMETICA M_a**

*** ci fanno capire quanto sono lontani i valori osservati dalla loro media**

la DEVIANZA

*** la distanza tra ogni valore osservato x(i)**
e la media dell'insieme delle unità M_a
è misurata attraverso la differenza $(x(i) - M_a)$

*** una sintesi di tutti questi scostamenti è**
la DEVIANZA

$$DEV = \sum_{i=1}^{n} (x_i - M_a)^2$$

*** elevo al quadrato per rendere positivi gli scarti negativi**
e per enfatizzare gli scarti maggiori

la VARIANZA è l'indice più noto

*** la devianza dipende dal numero di unità osservate n:**
bisogna trovare una misura indipendente da n

$$VAR = \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - M_a)^2$$

min: 0
max: dipende da X

*** la varianza è una devianza media,**
risponde a questa domanda:
in media, quanto sono lontani / devianti
i singoli valori x(i) dalla loro media aritmetica?

PROPRIETA' della varianza

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^{\,2} - M_a^{\,2} = M_a(x_i^{\,2}) - \overline{M_a(x_i)}^{\,2}$$

... la media dei quadrati meno il quadrato della media ...

*** consente calcoli più rapidi e agevoli**

VARIANZA COMPLESSIVA
di più gruppi parziali

$$\sigma^2_{Tot} = \frac{\sum_{i=1}^{L} \sigma_i^2 n_i}{n} + \frac{\sum_{i=1}^{L} (M_i - M_{Tot})^2 n_i}{n}$$

... media delle varianze più varianza delle medie ...

variabilità all'interno dei gruppi e variabilità tra gruppi

varianza come operatore non lineare: OPERATORE QUADRATICO

$$\sigma^2(k) = 0$$

$$\sigma^2(kx) = k^2\sigma^2(x)$$

$$\sigma^2(k + x) = \sigma^2(x)$$

*** una costante non varia (!?!?!)**

*** una costante che moltiplica tutti gli x(i) si riproduce nella varianza con effetto quadratico**

*** la varianza non risente di cambiamenti d'origine**

*** sono comode se i valori di x(i) sono molto "grandi" cioè composti da molte cifre**

VARIANZA DI UNA TRASFORMAZIONE LINEARE

$$y = a + bx$$

$$\sigma^2(y) = b^2\sigma^2(x)$$

*** il termine moltiplicativo viene elevato al quadrato**

*** il termine additivo sparisce**

la VARIAZIONE STANDARD

*** l'unità di misura della varianza è il quadrato dell'unità di misura dei dati originali**

es. se l'unità di misura sono i minuti,
la varianza è espressa in minuti al quadrato; cosa sono?

*** per avere una misura di variabilità più "comprensibile" è opportuno che abbia la stessa unità di misura dei dati ... quindi ...**

$$\sigma = \sqrt{\sigma^2}$$

min: 0
max: dipende dal carattere

esempi ...

es. peso medio degli studenti: 78 Kg
variabilità: sigma=3 Kg

in media ogni studente pesa 3 Kg in più o in meno rispetto alla media del gruppo

es. se le x(i) sono durate in minuti
sigma è espresso in minuti

es. se le x(i) sono durate in ore
sigma è espresso in ore

la variabilità ... A SPANNE

*** una stima grossolana della deviazione standard è**

$$\sigma \approx \frac{Range}{4}$$

*** è basato sul principio in base al quale, in moltissimi insiemi di dati, la gran parte degli x(i) è contenuta nell'intervallo [M_a - 2sigma ; M_a + 2sigma]**

... lo capirete meglio fra qualche lezione ...

*** può essere un sistema per verificare se il calcolo dettagliato di sigma è stato eseguito in modo corretto, ma non è sempre infallibile**

DIFETTI della DEVIAZIONE STANDARD

*** sigma risente dell'unità di misura dei dati**

*** non consente confronti tra fenomeni misurati con scale diverse**

es. cm e grammi

*** risente dell'ordine di grandezza dei dati**

*** non consente confronti tra caratteri con la stessa unità di misura, ma valori medi molto diversi**

es. peso di adulti e peso di bambini

vediamo un esempio …

N. giornate in cui si superano i limiti di inquinante nell'aria per mese

	Com. A	Com. B
gen	**2**	**1**
feb	**4**	**1**
mar	**3**	**2**
apr	**2**	**2**
mag	**4**	**4**
giu	**3**	**5**
lug	**2**	**5**
ago	**3**	**6**
set	**2**	**4**
ott	**4**	**2**
nov	**3**	**2**
dic	**4**	**2**
Totale	**36**	**36**

in entrambi i Comuni il numero medio di giornate "nere" per mese è pari a 3

ma le distribuzioni sono ben diverse

nel Comune A i valori sono abbastanza simili nei vari mesi: variano tra 2 e 4

nel Comune B, invece, ci sono alcuni mesi in cui l'inquinamento sembra sotto controllo, mentre d'estate aumenta notevolmente

le decisioni e i comportamenti da adottare nei due Comuni per limitare l'inquinamento saranno sicuramente diverse

un esempio …

. COM A	x(i)	x(i) - M	$(x(i) - M)^2$
gen	2	-1	1
feb	4	1	1
mar	3	0	0
apr	2	-1	1
mag	4	1	1
giu	3	0	0
lug	2	-1	1
ago	3	0	0
set	2	-1	1
ott	4	1	1
nov	3	0	0
dic	4	1	1
Totale	36		8

devianza

varianza nel Comune A
dev / n = 8 / 12
cioè 0,67

per il Comune B
il valore è 2,67

è maggiore

significa che
nel Comune B
i mesi sono
più diversi tra di loro
in termini di
giornate inquinate

COEFFICIENTE di VARIAZIONE

$$CV = \gamma = \frac{\sigma}{M_a} \cdot 100$$ **M_a diverso da 0**

*** esprime la percentuale della variabilità per ogni unità di valore medio**

*** dividendo per M_a si "neutralizza" l'effetto della scala di misura**

*** CV è un numero puro, adimensionale**

come al solito, vogliamo fare dei confronti

*** in genere, due distribuzioni di frequenza hanno medie e variabilità diverse;**
il confronto tra le due è "disagevole"

es. confronto le reazioni ad un trattamento di due gruppi di malati

*** bisogna eliminare l'influenza dei due diversi valori medi e delle due diverse dispersioni,**
che "confondono" i nostri confronti

es. come sono variate nel tempo
l'altezza dei pigmei e l'altezza dei watussi?

VARIABILI STANDARDIZZATE

*** si elimina l'influenza di M_a e sigma attraverso la STANDARDIZZAZIONE**

*** si sottrae la M_a :**
cambiamento d'origine della scala di misura

*** si divide per sigma:**
si "neutralizza" l'effetto della variabilità

$$Z = \frac{X - M_a}{\sigma}$$

cioè si calcolano
i valori standardizzati z(i)

$$z_i = \frac{x_i - M_a}{\sigma} \qquad i = 1, \ldots, n$$

... ce ne dovremo ricordare in seguito ...

CARATTERISTICHE degli z(i)

$$M_a(Z) = 0$$

$$\sigma^2(Z) = 1$$

es.	*x(i):*	*3*	*5*	*7*	*11*	*14*	*M(x)=8 sigma/x)=4*
	z(i):	*-1,25*	*-0,75*	*-0,25*	*0,75*	*1,5*	*M(z)=0 sigma(z)=1*

*** gli z(i) consentono confronti che altrimenti non si potrebbero fare**

un esempio in centimetri ...

Michael Jordan è "più alto" di Rebecca Lobo?

x_J = 195 cm ***x_L = 190 cm***

ovvio, 195 > 190 ... direbbe qualcuno ...

ma, M. Jordan è più alto tra i maschi di quanto lo sia R. Lobo tra le femmine?

$$z_J = \frac{x_J - M_m}{\sigma_m} = \frac{195 - 172{,}5}{7} = 3{,}21$$

$$z_L = \frac{x_L - M_f}{\sigma_f} = \frac{190 - 159}{6{,}25} = 4{,}96$$

$$z_J < z_L$$

R. Lobo è "relativamente" più alta fra le donne di quanto lo sia M. Jordan tra gli uomini

****** la cosa interessante è che funziona anche in inches ******

lo stesso esempio in inches ... 1 inch = 2,5 cm

$$x_J = 78 inches \qquad x_L = 76 inches$$

$$z_J = \frac{x_J - M_m}{\sigma_m} = \frac{78 - 69{,}0}{2{,}8} = 3{,}21$$

$$z_L = \frac{x_L - M_f}{\sigma_f} = \frac{76 - 63{,}6}{2{,}5} = 4{,}96$$

in linguaggio statistico si dice che:

M.J. è 3,21 deviazioni standard oltre la media degli uomini americani
R.L. è 4,96 deviazioni standard oltre la media delle donne americane

ancora una volta ragioniamo ... a spanne

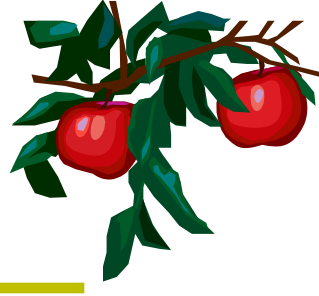
*** “di solito” vale la regola per cui**

- se $-2 \leq z_i \leq +2$ **siamo in presenza di valori ordinari, <u>normali</u>**

- se $|z_i| > 2$ **siamo in presenza di valori inusuali, <u>fuori norma</u>**

in seguito vedremo da dove deriva questa “regola”

la CONCENTRAZIONE



una mela al giorno toglie il medico di torno

ma

se mangio 30 mele tutte insieme

sarò sana per un mese?

o farò un indigestione?

un esempio …

Unità locali per numero di addetti

dove c'è più concentrazione?

in un comune con 100 aziende tutte con circa 5-6 addetti

o

in un comune con

60 aziende con 5 addetti	*(300 addetti in totale)*
39 aziende con 10 addetti	*(390 addetti in totale)*
1 azienda con 1.200 addetti	*(1.200 addetti in totale)*

nel secondo caso perché
in una sola azienda lavora il 63,5 % degli addetti del Comune

esistono alcune misure statistiche per valutare la CONCENTRAZIONE

Le variabili casuali

un sentiero attraverso la statistica …

DOVE SIAMO ARRIVATI:

* **dato un esperimento/procedimento**

lancio ripetuto di un dado

* **la statistica descrittiva consente di descrivere gli esiti osservati con frequenze, medie, varianze, grafici, …**

quante volte è comparso il 3?

* **la teoria della probabilità consente di calcolare la probabilità di ogni possibile risultato**

P(faccia pari)

DOVE STIAMO ANDANDO:

* **lo studio delle variabili causali consente di descrivere ciò che può "probabilmente" accadere anziché ciò che è già di fatto accaduto**

se lancio 100 volte, quante volte mi aspetto che esca il 3?

un esempio di percorso statistico attraverso la probabilità

* **sperimentare: lancio ripetuto di un dado**

* **descrivere il campione:**

x_i	**1**	**2**	**3**	**4**	**5**	**6**	$\bar{x} = 3,6$
n_i	***8***	***10***	***9***	***12***	***11***	***10***	$s = 1,7$

* **trovare la probabilità di ogni possibile esito**

  P(1)=1/6 P(2)=1/6 … P(6)=1/6

* **creare un modello teorico che descriva il comportamento del processo sperimentale e calcolarne i parametri d'interesse: (VARIABILI CASUALI)**

x_i	**1**	**2**	**3**	**4**	**5**	**6**	$\mu = 3,6$
P(x)	**1/6**	**1/6**	**1/6**	**1/6**	**1/6**	**1/6**	$\sigma = 1,7$

perché ci serve un MODELLO TEORICO?

perché la realtà non è un dado!

* **perché vogliamo rappresentare i fenomeni attraverso simboli matematici**

  e perché i fenomeni si articolano in eventi

  es. tempo atmosferico: sole, pioggia, nebbia, neve, …

* **i possibili eventi della realtà non sono equiprobabili (come le facce di un dado)**
  ma ciascuno accade con una sua propria probabilità

* **associando a ciascun evento una probabilità che esso accada si costruisce una VARIABILE CASUALE**

VARIABILE CAUSALE o aleatoria

definizione formale:

si definisce variabile casuale (v.c.)
una funzione X
che associa ad ogni evento elementare
dello spazio campionario S_x
uno e un solo numero reale

ma cosa significa? *un esempio*

il risultato di una prova è in molti casi un numero che è casuale (o aleatorio) perché incerto prima della prova

il numero che ci interessa a volte è generato direttamente dalla prova *(peso di una persona)***, altre volte si ottiene elaborando il risultato della prova** *(calcolo dell'indice di massa corporea: BMI)*

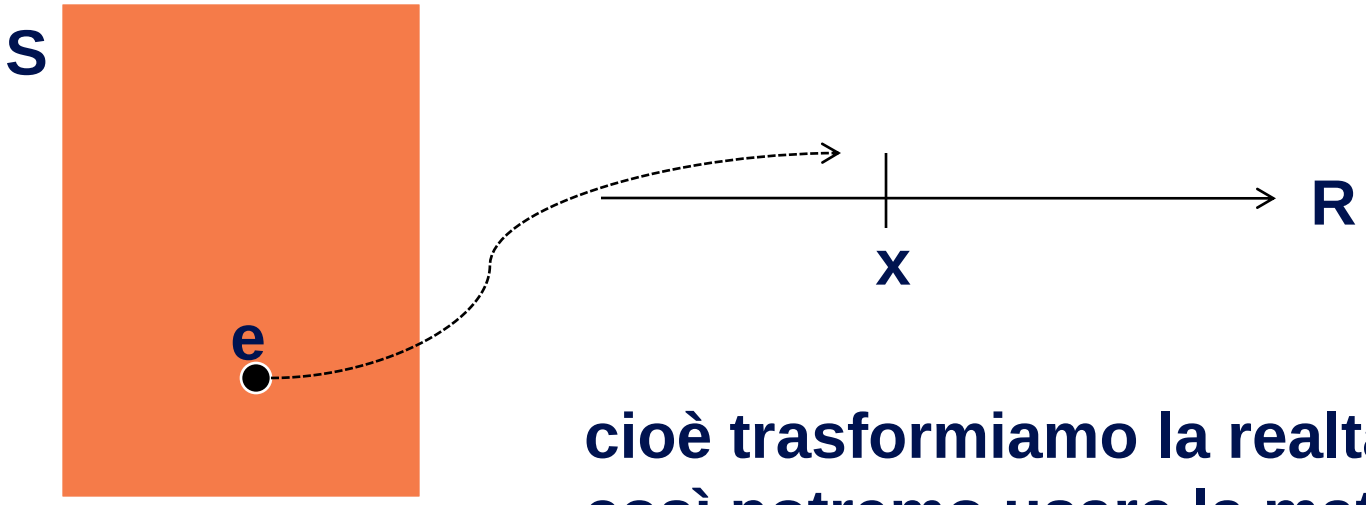
la funzione del risultato della prova è una variabile casuale

si studia la v.c. X per determinare la probabilità con cui X assume un certo valore *P(X=89 kg)* **o appartiene ad un insieme B** *obesi: BMI>30 P(X>30)*

trasformiamo le cose in numeri

semplifichiamo lo spazio campionario

trasformando gli eventi elementari in S
in punti dell'asse reale R
attraverso il concetto di variabile casuale X



cioè trasformiamo la realtà in numeri
così potremo usare la matematica
per studiare i fenomeni

un esempio …

uno studio consiste nel selezionare casualmente 14 neonati e contare il numero di bimbe

X = n. di bimbe tra i 14 neonati
X è una v.c. perché dipende dal caso, dalla natura

si può fare l'elenco di tutti i possibili valori di X e calcolare la probabilità che si verifichino

x	**0**	**1**	**2**	**…**	**13**	**14**
P(x)	***0,000***	***0,001***	***0,006***	***…***	***0,001***	***0,000***

DISTRIBUZIONE DI PROBABILITA'
fornisce la probabilità di ciascun valore della v.c.

altri esempi …

**** pressione arteriosa sistolica***
**** durata di vita di una valvola termoionica***
**** resistenza alla rottura per compressione di un conglomerato***

molti fenomeni possono essere descritti tramite le v.c. e il loro comportamento può essere analizzato e compreso applicando formule matematiche alla loro distribuzione di probabilità

** quanta scorta di lampadine devo tenere in casa per essere sicuro di poter sostituire quelle che eventualmente si rompono?*
** quanto rischio di aspettare in coda alla posta? prenderò il prossimo autobus?*
** vale la pena di installare un terzo motore di scorta su un aereo in rotta transoceanica?*

v.c. DISCRETE

def.

la v.c. è DISCRETA se è definita in uno spazio campionario discreto, cioè può assumere un numero finito o un'infinità numerabile di valori

notazioni: **X** **v.c.**

S_x **spazio campionario della v.c. X**

la DISTRIBUZIONE DI PROBABILITA'

def.

se S_x è composto da un n. finito (o infinità numerabile) di punti disgiunti (cioè l'unione di tutti i punti è pari a S_x)

ad ogni punto si può assegnare una probabilità

valori di X	x_1	x_2	…	x_i	…
probabilità	p_1	p_2	…	p_i	…

$$p_i = P(X=x_i) \qquad i=1,2,\ldots$$

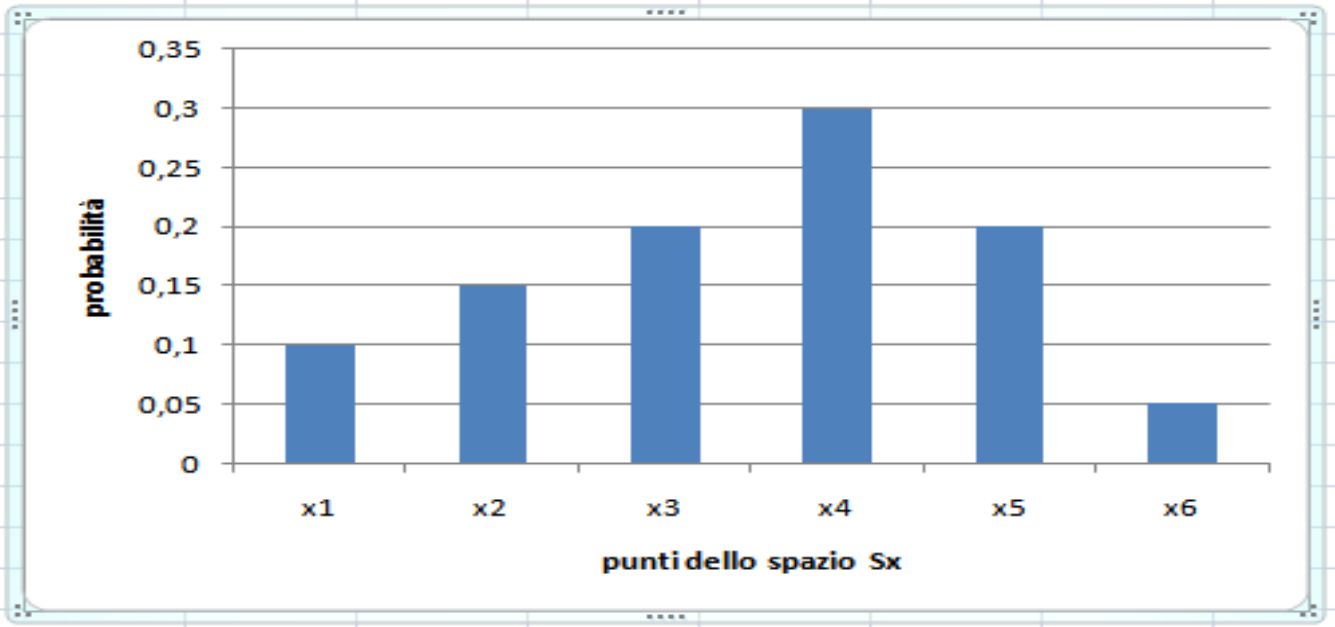
è una DISTRIBUZIONE DI PROBABILITA'

per analogia …

per analogia con le distribuzioni di frequenza relativa

alle distribuzioni di probabilità possono essere applicate le tecniche di statistica descrittiva

ad esempio,
si può tracciare l'ISTOGRAMMA DI PROBABILITA'



in ascissa i valori di X, in ordinata le probabilità corrispondenti

PROPRIETA'
della funzione di probabilità

$$\sum_x p(x) = 1 \qquad \forall x \in S_x$$

$$0 \leq p(x) \leq 1 \qquad \forall x \in S_x$$

la prima significa che la somma delle probabilità di tutti gli eventi possibili è pari a 1, cioè che siamo certi che uno di quegli eventi sicuramente accade

quando si descrive un fenomeno attraverso una v.c. bisogna essere sicuri che tutte le possibili sfaccettature della realtà possano essere rappresentate, quindi

*** attenzione alla probabilità che non accada nulla**
*** attenzione alle ipotesi in campo**

es. dado in bilico su un terreno sconnesso

un esempio

X = n. di uova deposte in un giorno da una gallina

x_1=0 x_2=1 x_3=3 …
ma
nessuna gallina depone 8.436 uova in un giorno

N.B. quando una v.c. descrive un fenomeno concreto occorre concentrare l'attenzione sui valori (eventi) concretamente realizzabili (possibili), cioè quelli con probabilità > 0

ciò da una forma particolare alle funzioni di probabilità di una v.c.

un esempio

$$\begin{cases} p(x) = x/5 & x = 0,1,2,3 \\ p(x) = 0 & \textit{altrove} \end{cases}$$

è una funzione di probabilità?

x	**0**	**1**	**2**	**3**
p(x)	***0/5***	***1/5***	***2/5***	***3/5***

$0 \leq p(x) \leq 1 \qquad \forall x \in S_x$ **ok**

$\sum_x p(x) = p(0) + p(1) + p(2) + p(3) = 6/5 \neq 1$ **ko**

un esempio

$$\begin{cases} p(x) = x/3 & x = 0,1,2 \\ p(x) = 0 & \textit{altrove} \end{cases}$$

è una funzione di probabilità?

x	**0**	**1**	**2**
p(x)	***0/3***	***1/3***	***2/3***

$0 \leq p(x) \leq 1 \qquad \forall x \in S_x$ **ok**

$\sum_x p(x) = p(0) + p(1) + p(2) = 1$ **ok**

un esempio in stile Las Vegas

X = lancio di un dado p(x)=1/6 x=1,2,3,4,5,6
Y = lancio di due dadi, si considera la somma delle facce

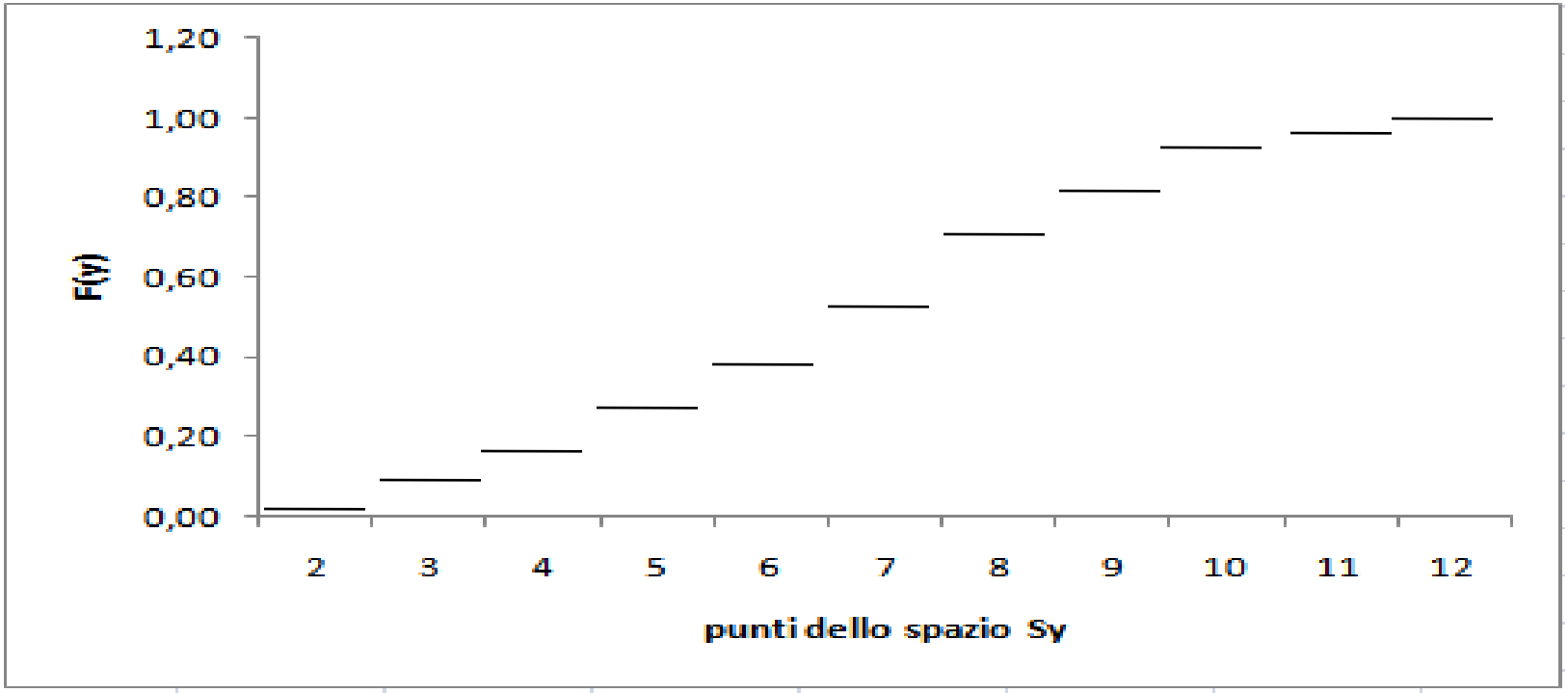
S degli eventi originali è composto dalle 36 coppie (i,j) con i,j=1,2,3,4,5,6

Y(coppia)=i+j y=2,3,4,5, …,12 S_y è {2,3,4,…,12}

p(2) = p(1,1)=1/36
p(3) = p(1,2)+p(2,1) = 2/36
…
p(7) = p(1,6)+p(2,5)+p(3,4)+p(4,3)+p(5,2)+p(6,1) = 6/36
…
p(12) = p(6,6) = 1/36

graficamente …

F(y) è una funzione costante a tratti per v.c. discrete



es. lancio di due dadi e somma dei risultati

y	*2*	*3*	*4*	*5*	*6*	*…*	*12*
F(y)	*1/36*	*3/36*	*6/36*	*10/36*	*15/36*	*…*	*36/36*

PROPRIETA' della f.r.

*** è una funzione limitata e il suo campo di variazione è l'intervallo [0,1]** $\lim_{x \to -\infty} F(x) = 0$ $\lim_{x \to +\infty} F(x) = 1$

*** è non decrescente se a<b F(a)<=F(b)**

*** è continua a destra** $\lim_{h \to 0^+} F(x+h) = F(x)$

***** P(a<X≤b) = F(b) – F(a)**

ciò consente di passare dalla funzione di ripartizione alla funzione di probabilità

passiamo alle v.c. CONTINUE

def.

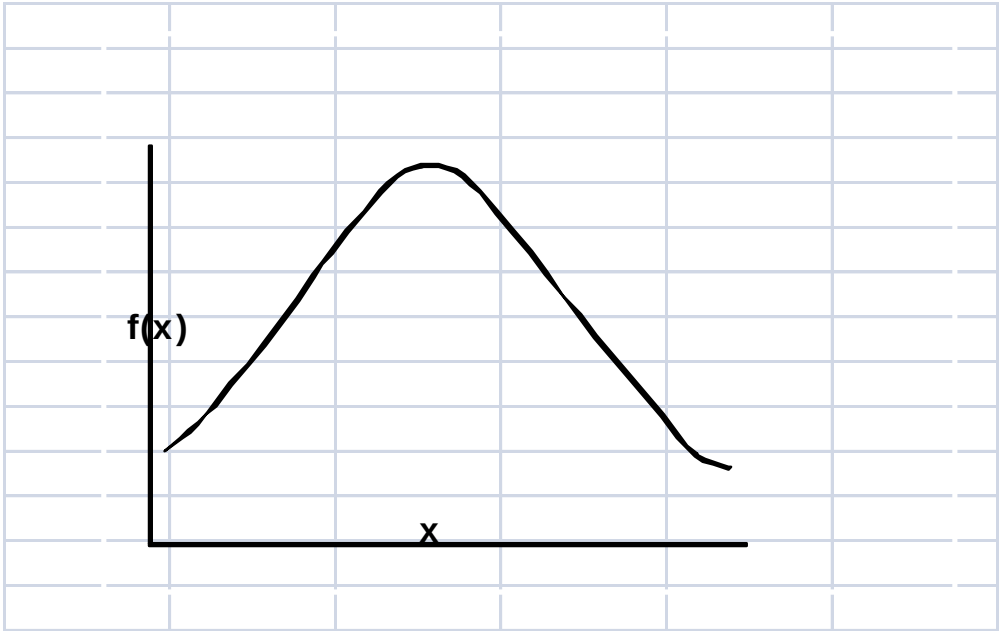
la v.c. è CONTINUA se è definita in uno spazio campionario continuo, cioè può assumere un numero infinito di valori su scala continua

notazioni: X **v.c.**

S_x **spazio campionario della v.c. X**

la FUNZIONE DI DENSITA’

di una v.c. X CONTINUA definita in un intervallo (l,L)
è la funzione f(x) tale per cui



$$f(x) \geq 0$$

$$\int_{l}^{L} f(x)\,dx = 1$$

*** non è una probabilità (prob. di un punto = 0)**
*** attenzione all’intervallo (l,L) (valori possibili / realistici)**

la FUNZIONE DI RIPARTIZIONE

della v.c. X CONTINUA è definita come

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\,dt$$

proprietà

$$\lim_{x \to -\infty} F(x) = 0 \qquad \lim_{x \to +\infty} F(x) = 1$$

*** limitata**

*** non decrescente se a<b** **F(a) ≤ F(b)**

***** P(a ≤ x ≤ b) = F(b) – F(a)**

***** assolutamente continua** $\dfrac{d}{dx} F(x) = f(x)$

esempi

X = quantità di latte prodotto quotidianamente da una mucca

X = misura del voltaggio di una batteria

X = tempo impiegato per riempire una lattina di coca-cola

attraverso le v.c. continue si può rispondere a domande del tipo:

** qual è la probabilità che mio figlio adolescente riattacchi la cornetta prima che scattino i 30 minuti di telefonata?*

** quale probabilità ho di attendere da 10 a 15 minuti prima che arrivi l'autobus?*

** quante delle 850 postazioni PC della mia azienda mi aspetto di dover rimpiazzare quest'anno perché rotte?*

le v.c. sono come le sfere di cristallo

*** permettono di prevedere il futuro, prevedendo la probabilità di accadimento degli eventi**

** quanta pioggia mi aspetto che cada in Sardegna quest'estate?*
** che probabilità ho di non poter salire in aereo a causa dell'overbooking?*

*** o di reinterpretare il passato, confrontando la probabilità teorica degli eventi con il loro accadere reale**

** la pioggia caduta l'estate scorsa in Sardegna è stata normale o si è trattato di un anno particolare ?*
** la compagnia aerea Z "ha fatto la furba" e ha giocato in modo irregolare con l'overbooking?*

distribuzioni di PROBABILITA' CONGIUNTE

di una coppia di v.c.
o di una v.c. bivariata o v.c. doppia

un esito di un processo / esperimento viene misurato con due strumenti

es. di una vite vengono misurati il diametro e il peso

ad ogni evento elementare originale viene fatta corrispondere una coppia di numeri (x,y)

x=diametro y=peso

si passa a considerare un nuovo spazio S_{xy}
costituito dai punti nel piano reale R^2

distribuzione DOPPIA di probabilità

		x					
		0	1	2	…	i	…
	0	p(0,0)	p(0,1)	p(0,2)	…	p(0,i)	…
	1	p(1,0)	p(1,1)	p(1,2)	…	p(1,i)	…
y	…	…	…	…	…	…	…
	j	p(j,0)	p(j,1)	p(j,2)	…	p(j,i)	…
	…	…	…	…	…	…	…

- **notate l'analogia con la distribuzione di frequenza doppia**

- **ma in quel caso si trattava di frequenze osservate**

- **qui parliamo di probabilità teoriche**

FUNZIONE DI PROBABILITA' CONGIUNTA

di due variabili DISCRETE X e Y

$$p(x, y) = P(X = x, Y = y)$$

contemporaneamente X=x e Y=y

ancora una volta, le proprietà

$$p(x, y) \geq 0 \qquad \sum_x \sum_y p(x, y) = 1$$

FUNZIONE DI RIPARTIZIONE CONGIUNTA

di due variabili DISCRETE X e Y

$$F(x,y) = P(X \le z, Y \le y) = \sum_{t \le x} \sum_{v \le y} p(t,v)$$

contemporaneamente X ≤ x e Y ≤ y

proprietà:

*** è limitata**

$$\lim_{x \to -\infty} F(x,y) = 0 \qquad \lim_{y \to -\infty} F(x,y) = 1 \qquad \lim_{\substack{x \to +\infty \\ y \to +\infty}} F(x,y) = 1$$

*** non decrescente rispetto a ciascuna delle variabili**

se x ≤ x’ F(x,y) ≤ F(x’,y)

se y ≤ y’ F(x,y) ≤ F(x,y’)

funzioni di probabilità MARGINALI

data p(x,y) *probabilità di superare al primo colpo l'esame di statistica e telerilevamento*

$$p_1(x) = \sum_y p(x, y)$$

probabilità di superare al primo colpo statistica

cioè la probabilità che X=x per un qualsiasi valore di Y

$$p_2(y) = \sum_x p(x, y)$$

probabilità di superare al primo colpo telerilevamento

cioè la probabilità che Y=y per un qualsiasi valore di X

funzioni di probabilità CONDIZIONATE

data p(x,y)

probabilità di superare al primo colpo l'esame di statistica e telerilevamento

$$p_1(y \mid x) = \frac{p(x, y)}{p_1(x)}$$

dato che ho superato statistica che probabilità ho di superare anche telerilevamento

$$p_2(x \mid y) = \frac{p(x, y)}{p_2(y)}$$

dato che ho superato telerilevamento che probabilità ho di superare anche statistica

un esempio

		x			
		0	1	2	$p_2(y)$
y	0	0,1	0,1	0,2	0,4
	1	0,2	0,3	0,4	0,6
	$p_1(x)$	0,3	0,3	0,4	1

y	0	1
$p_1(x=1)$	0,1/0,3 = 0,333	0,2/0,3 = 0,667

relazioni tra v.c.
l'INDIPENDENZA

due v.c. X e Y si dicono indipendenti se e solo se

$$p(y,x) = p_1(x) \cdot p_2(y) \qquad \forall (x,y)$$

oppure $p_1(y \mid x) = p_2(y)$ **e**

$$p_2(x \mid y) = p_1(x)$$

in caso di indipendenza si ha che F(x,y)=F(x) F(y)
cioè la f.r. congiunta è il prodotto delle due singole f.r.

nelle applicazioni reali l'indipendenza non è facilmente verificabile

e per finire:
distrib. congiunte per v.c. CONTINUE

FUNZIONE DI DENSITA' CONGIUNTA f(x,y)

è tale per cui

$$f(x, y) \geq 0$$

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x, y)\,dxdy = 1$$

FUNZIONE DI RIPARTIZIONE CONGIUNTA F(x,y)

$$F(x, y) = \int_{-\infty}^{y}\int_{-\infty}^{x} f(u, v)\,dudv$$

marginali, condizionate e indipendenza nel CASO CONTINUO

marginali

$$f_1(x) = \int_{-\infty}^{+\infty} f(x,y)\,dy \qquad f_2(y) = \int_{-\infty}^{+\infty} f(x,y)\,dx$$

condizionate

$$f_1(y \mid x) = \frac{f(x,y)}{f_1(x)} \qquad f_2(x \mid y) = \frac{f(x,y)}{f_2(y)}$$

indipendenza

$$f(x,y) = f_1(x) \cdot f_2(y) \qquad \forall (x,y)$$

Valore atteso e varianza

delle variabili casuali

… ripetiamo un po' …

FUNZIONI DI PROBABILITA' di una v.c. DISCRETA

valori di X	**x1**	**x2**	**…**	**xi**	**…**
p(x)	**p(x1)**	**p(x2)**	**…**	**p(xi)**	**…**

es. *lanciando 4 volte una moneta*
sia X la v.c. numero di "teste"

valori di X	*0*	*1*	*2*	*3*	*4*
p(x)	*1/16*	*4/16*	*6/16*	*4/16*	*1/16*

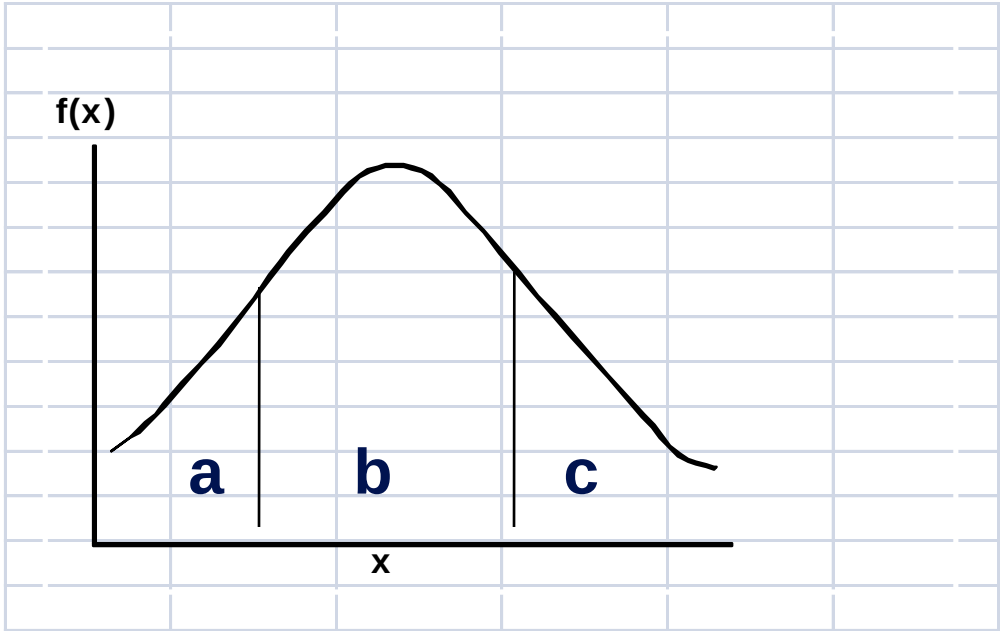
provate a verificare da soli
se le probabilità sono proprio queste

… ripetiamo un po' …

FUNZIONE DI DENSITA' di una v.c. CONTINUA

f(x)

esempio:
v.c. X altezza dei convocati alla visita di leva



a: i più bassi
b: i "normali" o "medi"
c: i più alti

… ribattiamo sempre sullo stesso chiodo …

le funzioni di probabilità o densità descrivono in DETTAGLIO tutte le probabilità associate ai modi di manifestarsi di un fenomeno

consentono di definire un qualsiasi evento e di calcolare le probabilità che si verifichi

MA in statistica SERVE SINTESI

cioè misure che descrivano con pochi numeri le caratteristiche salienti delle variabili casuali (fenomeni)

(il modo più immediato di far sintesi è la media) il VALORE ATTESO

un concetto simile a quello di media per le v.c. è quello di VALORE ATTESO

def. formale:
si chiama valore atteso o media della v.c. X DISCRETA

$$E(x) = \mu = \sum_{x} x \cdot p(x)$$

somma estesa a tutti i valori di X

il simbolo E deriva da EXPECTATION
infatti, E(X) può essere letto come

l'esito medio teorico atteso
in caso di prove ripetute all'infinito

VALORE ATTESO per v.c. CONTINUE

$$E(x) = \mu = \int_{-\infty}^{+\infty} x \cdot f(x)\,dx$$

dove f(x) è la funzione di densità

da dove nasce l'analogia tra MEDIA e VALORE ATTESO?

è chiara se adottiamo l'approccio frequentista alla probabilità:

* **immaginiamo di ripetere "tante" volte un esperimento**

* **costruiamo la tabella delle frequenze relative osservate per ciascuno dei possibili risultati**

* **immaginiamo che il numero di prove cresca a dismisura (n tende all'infinito)**

* **secondo la legge dei grandi numeri, le frequenze relative degli esiti osservati approssimano bene le probabilità che quegli esiti si manifestano**

la probabilità letta come LIMITE DELLA FREQUENZA RELATIVA

per n che tende all'infinito

$$\bar{x} = \sum_i x_i \cdot \frac{n_i}{n} \longrightarrow E(X) = \mu = \sum_x x \cdot p(x)$$

media
dei valori osservati

valore teorico atteso
secondo il modello
matematico
associato al fenomeno

il valore atteso è una MISURA CENTRALE

E(X) **può essere letto come**

una MEDIA PONDERATA dei valori di X

usando come PESI le PROBABILITA'

(anziché le frequenze relative)

cioè mantiene il suo significato di

VALORE CENTRALE della funzione p(x) o f(x)

un esempio …

$$\begin{cases} p(x) = x/3 & x = 0,1,2 \\ p(x) = 0 & \textit{altrove} \end{cases}$$

valori di X	**0**	**1**	**2**
p(x)	**0/3**	**1/3**	**2/3**

$$E(X) = \mu = x_1 \cdot p(x_1) + x_2 \cdot p(x_2) + x_2 \cdot p(x_2) = 0 \cdot \frac{0}{3} + 1 \cdot \frac{1}{3} + 2 \cdot \frac{2}{3} = \frac{1}{6}$$

PROPRIETA' del valore atteso

*** E(c)=c** **il valore atteso di una costante è la costante stessa**

che esito mi aspetto da un fenomeno che non cambia mai?
sempre lo stesso!

*** $E(c_1 \cdot X_1 + \ldots + c_n \cdot X_n) = c_1 \cdot E(X_1) + \ldots + c_n \cdot E(X_n)$**
linearità della media: la media di una combinazione lineare di v.c. è uguale alla combinazione lineare delle medie

*** se $X_1, X_2, \ldots, X_n$ sono v.c. INDIPENDENTI**
$E(X_1 \cdot X_2 \cdot \ldots \cdot X_n) = E(X_1) \cdot E(X_1) \cdot \ldots \cdot E(X_n)$

… un’altra proprietà

$$E(X - \mu) = 0$$

il valore atteso degli scarti di una variabile casuale dalla propria media è nullo

*** **conferma che E(X) ha un significato di tendenza centrale della distribuzione di X**

E(X) è una COSTANTE DI POSIZIONE

TRASFORMATE di v.c.

sia g(x) una funzione della v.c. X

$$E[g(X)] = \begin{cases} \sum_{x} g(x) \cdot p(x) & \text{per v.c. discrete} \\ \int_{-\infty}^{+\infty} g(x) \cdot f(x)\,dx & \text{per v.c. continue} \end{cases}$$

queste formule sono utili quando dovete trasformare i valori di X per renderli più "comodi" da trattare

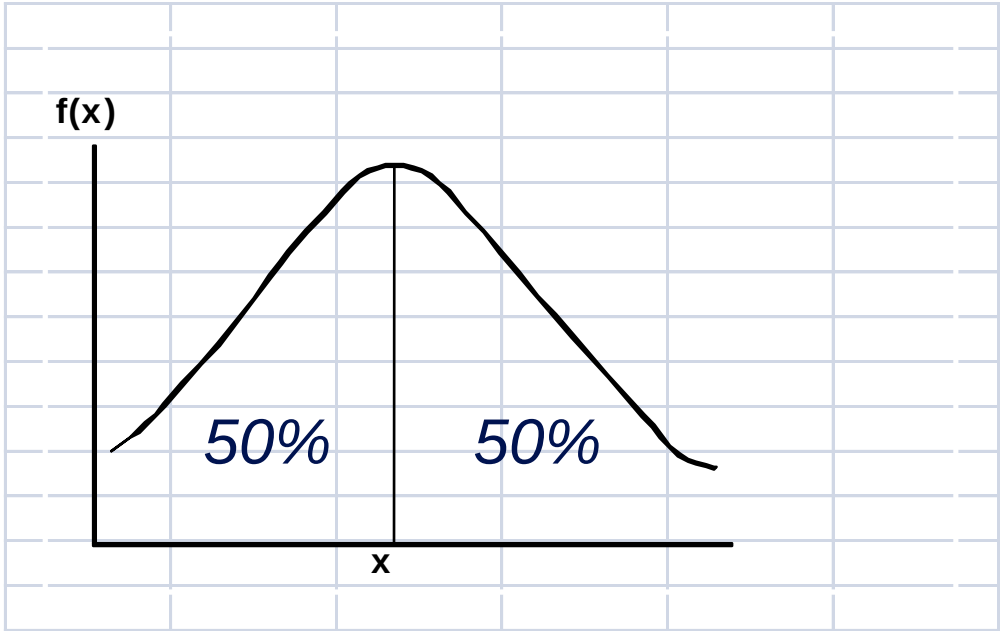
es. diminuire le cifre da elaborare dividendo tutto per 1.000

altri indici di posizione di una distribuzione di probabilità

la MEDIANA è il valore x_m tale per cui

$P(X < x_m) \leq ½$ **e**

esempio:



altri indici di posizione di una distribuzione di probabilità

la MODA è quel valore di X tale per cui è massima la funzione di probabilità / densità

esempio:

valori di X	*1*	*2*	*3*	*4*	*5*	*6*
p(x)	*0,1*	*0,1*	*0,4*	*0,2*	*0,1*	*0,1*

$x_{mo} = 3$

come si distribuiscono i valori della v.c. X?
MISURE DI DISPERSIONE

la misura più comune è la VARIANZA

data una v.c. X e dato il suo valore atteso E(X)

si definisce varianza Var(X) o $\sigma_X^2 = E(X-\mu)^2$ **.**

come in statistica descrittiva
la varianza fornisce informazioni sul grado di concentrazione della distribuzione intorno alla media, cioè ci dice di quanto si allontanano "in media" i valori di X dal loro valore atteso

come si calcola la varianza?

grazie alle proprietà di E(X) si ottiene una formula abbreviata per il calcolo della varianza

$$\sigma_X^2 = E(X^2) - [E(X)]^2$$

nel caso di v.c. discrete

$$\sigma_X^2 = \left[\sum_x x^2 \cdot p(x)\right] - [\mu]^2$$

è utile se si fanno i calcoli a mano

SCOSTAMENTO QUADRATICO MEDIO o DEVIAZIONE STANDARD

è la radice quadrata della varianza

$$\sigma_X = \sqrt{\sigma_X^2}$$

ancora analogie con la statistica descrittiva …

$\overline{x}$	μ
S^2	σ^2
S	σ
campione di dati osservati	**valori teorici secondo la distribuzione di probabilità**

attenzione alla simbologia: σ^2_X **è usato anche quando si descrive la varianza osservata**

l'importante è aver capito che concettualmente sono due cose diverse

PROPRIETA' della varianza

$$\sigma^2(a \cdot X + b) = a^2 \cdot \sigma^2_X$$ **a e b costanti**

già descritta e commentata in precedenza

*** se due v.c. X e Y sono INDIPENDENTI**

$$\sigma^2_{X+Y} = \sigma^2_X + \sigma^2_Y$$

la v.c. STANDARDIZZATA

def.

$$Z = \frac{X - \mu}{\sigma}$$

ha proprietà interessanti: **E(Z) = 0**
Var(Z) = 1

vi ricordate il confronto tra l'altezza di M. Jordan e di R. Lobo?

un esempio …

una compagnia aerea applica l'overbooking, la v.c. X rappresenta il numero di passeggeri che non possono salire perché non ci sono posti disponibili

x	p(x)	x · p(x)	x^2	$x^2 \cdot p(x)$
0	0,805	0	0	0
1	0,113	0,113	1	0,113
2	0,057	0,114	4	0,228
3	0,009	0,027	9	0,081
4	0,002	0,008	16	0,032
tot.		0,262		0,454

$$\sum_x x \cdot p(x) \qquad \sum_x x^2 \cdot p(x)$$

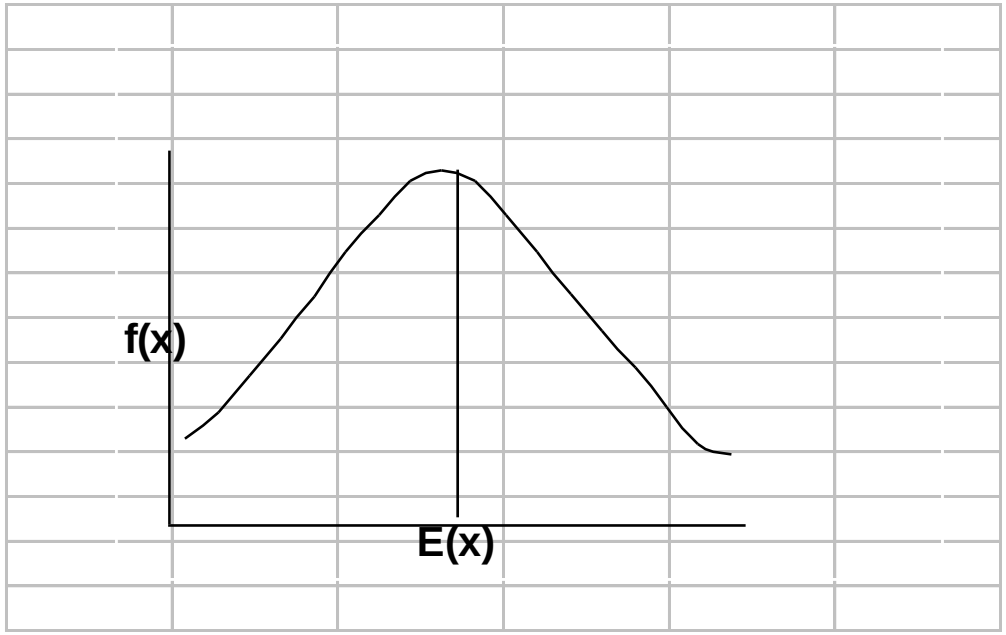
E(X) = 0,262 *ci si aspetta che in media non si possa sedere circa ¼ di passeggero*
chi dovrà rimanere a terra dalle ginocchia in giù?

$E(X) = 0{,}454 - (0{,}262)^2 = 0{,}385$

distribuzioni congiunte
e
variabili casuali discrete

... ripetiamo un po' ...

data una distribuzione di probabilità / densità



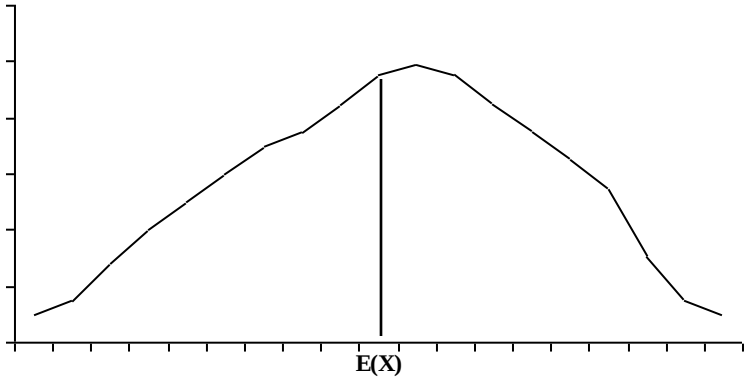
E(X) ci parla del CENTRO

VAR(X) ci parla della DISPERSIONE intorno al centro

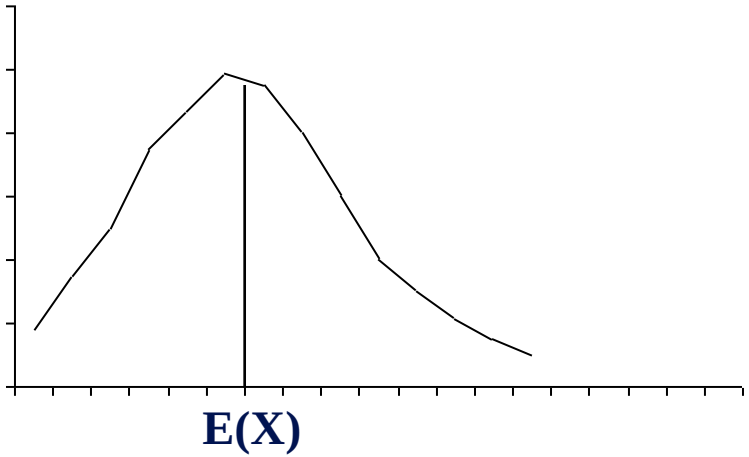
cioè di come i valori di X si dispongono intorno al loro valor medio

due esempi …

VAR(X) alta



VAR(X) più bassa



la disuguaglianza di Tchebycheff

data una variabile casuale X con media μ e varianza σ

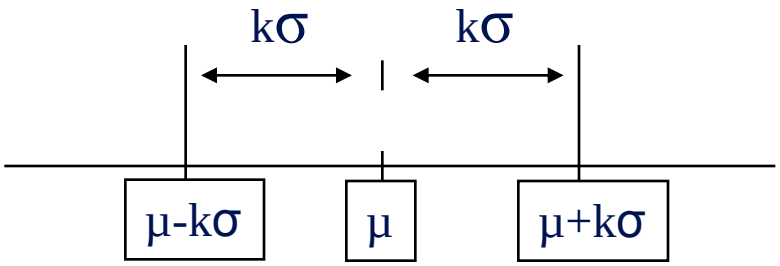
si ha che, per ogni reale k > 0

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

fornisce un limite superiore alla variazione della funzione di ripartizione su intervalli simmetrici intorno a μ

VALE PER QUALSIASI DISTRIBUZIONE DI PROBABILITA'

… proviamo a disegnare gli eventi considerati e facciamo un esempio



es. ***sia X una v.c. con media μ = 100 e s.q.m sigma = 2***
posto k = 1,5
grazie a Tchebycheff si può dire che

$$P(|X - 100| \geq 3) \leq \frac{1}{1,5^2} = 0,44$$

cioè che i valori esterni all'intervallo (97,103) hanno una probabilità complessiva non superiore a 0,44

VARIABILI CASUALI DOPPIE
… ripetiamo

es. distribuzione di probabilità congiunta nel caso discreto

		x					
		0	1	2	...	i	...
	0	p(0,0)	p(0,1)	p(0,2)	...	p(0,i)	...
y	1	p(1,0)	p(1,1)	p(1,2)	...	p(1,i)	...
	...	...	...	...	...	...	...
	j	p(j,0)	p(j,1)	p(j,2)	...	p(j,i)	...
	...	...	...	...	...	...	...

$$p(x,y) = P(X = x, Y = y)$$

COVARIANZA

siano X e Y due v.c. e μ_X,μ_y i rispettivi valori attesi

si chiama COVARIANZA COV(X,Y) il valore atteso della funzione $g(X,Y) = (X-\mu_X) \cdot (Y- \mu_Y)$

la co-varianza ci dice se e quanto X e Y tendono a variare nella stessa direzione, cioè se CO-VARIANO
cioè se vanno a braccetto, in CO-ppia

cioè misura il grado in cui due variabili sono LEGATE LINEARMENTE

a grandi valori di X tendono ad associarsi grandi valori di Y
piccoli valori di X si associano piccoli valori di Y

es. X=peso Y=altezza
chi più è alto, più pesa (di solito)

… perché è fatta proprio così ?

* **serve una misura indipendente dall'origine delle scale di X e Y: si sottrae a entrambi la media, cioè si traslano i valori al centro della loro distribuzione**

* **si considera il prodotto delle due nuove coordinate $[X-E(X)] \cdot [Y-E(Y)]$**

* **se a valori di X sopra la media corrispondono valori di Y sopra la media, il prodotto sarà positivo**
  se a valori di X sotto la media corrispondono valori di Y sotto la media, il prodotto sarà positivo

* **il prodotto è negativo se i valori di X e Y non si muovono nella stessa direzione (es. valori grandi di X corrispondono a valori piccoli di Y)**

perché è fatta proprio così !

* **se effettuiamo la somma di tutti i prodotti per tutte le possibili coppie di valori X e Y e li ponderiamo ciascuno con la propria probabilità**
  otteniamo una misura della variazione congiunta di X e Y

infatti, la formula teorica della covarianza è: *(es. nel caso discreto)*

COV(X,Y) = E [(X-μX) · (Y- μY)] o $\sigma_{XY} = \sum_x \sum_y (x - \mu_X) \cdot (y - \mu_Y) \cdot p(x, y)$

- COV(X,Y) > 0 le due variabili tendono a muoversi insieme

- COV(X,Y) < 0 le due variabili tendono a muoversi in direzioni opposte

- COV(X,Y) = 0 le due variabili sono INCORRELATE

come si calcola la covarianza

COV (X,Y) = E (X·Y) – E(X) · E(Y)

cioè calcolando

il valor medio dei prodotti

meno

il prodotto dei valori medi

anche questa filastrocca aiuta a ricordare

un esempio …

date due variabili casuali

			Y	
		0	1	2
X	0	0,1	0,3	0,1
	1	0,2	0,1	0,2

provate a verificare se sono indipendenti tramite il calcolo della covarianza

per l'illustrazione del calcolo vi rinvio ad un libro di statistica

attenzione alle parole e ai significati !

Se due variabili sono indipendenti si ha che
$E(X \cdot Y) = E(X) \cdot E(Y)$
quindi
$COV(X,Y) = E(X \cdot Y) - E(X) \cdot E(Y) = E(X) \cdot E(Y) - E(X) \cdot E(Y) = 0$
cioè sono anche incorrelate

ma

se due variabili sono incorrelate
cioè se $COV(X,Y)=0$
non è necessariamente vero che sono indipendenti

la covarianza ci dice pochino …

* la COV non ha limiti
  cresce in valore assoluto anche in base alla scala di valori usata per X e Y

* possiamo interpretarne il segno, ci dice se la correlazione è positiva o negativa, ma non è facile interpretarne la "quantità", la forza della correlazione

* serve una misura che ci faccia capire se la correlazione è "tanta" o "poca", se è forte o debole

* quindi:
  serve un indice che abbia una valore minimo e massimo con cui confrontarci

il COEFFICIENTE di CORRELAZIONE

$$\rho = \frac{Cov\ (X\ ,Y\ )}{\sigma_X \cdot \sigma_Y}$$

il coeff. di correlazione è compreso tra -1 e +1

ρ = -1 se Y è funzione lineare di X

Y = a+bX b<0 (al crescere di una v.c. l'altra cala)

ρ = +1 se X è funzione lineare di Y

Y = a+bX b>0 (al crescere di una v.c. cresce anche l'altra)

cosa ci dice ρ ?

* **è una misura della RELAZIONE LINEARE tra X e Y**

* **fornisce il grado di precisione**
  con cui possiamo prevedere Y a partire da una trasformazione lineare di X
  e/o
  con cui possiamo prevedere X a partire da una trasformazione lineare di Y

informazione “simmetrica”
non necessariamente “causale” o “unidirezionale”

* **se tra X e Y esiste una relazione molto stretta**
  ma non lineare
  ρ può assumere ugualmente valori molto vicini allo zero

chi pensasse che ρ≈0 significa nessuna relazione tra X e Y
potrebbe sbagliarsi di grosso!

distribuzioni di probabilità particolari

*** ci sono alcuni modelli matematici che ben si adattano a descrivere dei fenomeni e/o esperimenti**

*** cioè alcune funzioni di probabilità / densità particolarmente importanti perché descrivono l'andamento di fenomeni naturali o l'esito di esperimenti**

*** si tratta di formule particolari, parametriche**

*** delle quali si conosce il comportamento (valore atteso, varianza, forma, limiti, …)**

*** che si usano frequentemente in statistica**

la distribuzione BINOMIALE

* **si adatta a situazioni in cui gli ESITI OSSERVABILI sono DUE**

es. successo/fallimento, vero/falso, testa/croce

* **soddisfa i seguenti requisiti**

  - **è definito un numero fisso di tentativi / prove**
  - **le prove sono indipendenti l'una dall'altra**
  - **ciascuna prova ha solo due possibili esiti**
  - **le probabilità dei due esiti rimangono costanti per ogni prova**

es. n lanci di una moneta
es. 100 neonati, quanti i maschietti? e le femminucce?

… la distribuzione binomiale

* **per convenzione si definisce un esito SUCCESSO e l'altro INSUCCESSO**

ma non sempre si gioisce dei "successi"
in epidemiologia, ad esempio, i successi sono spesso malattie

* **notazioni:**

p **probabilità di successo in una prova**
q **probabilità di insuccesso in una prova** $q = (1-p)$
n **numero di prove**
x **numero di successi** $0 \leq x \leq n$
P(x) **probabilità di ottenere esattamente x successi in n prove**

p(x) = P(X=x) è la funzione di probabilità

es. probabilità di indovinare tutte le risposte
a una serie di 20 quesiti a risposta multipla

… la formula della distribuzione binomiale

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x} = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$

per x = 0,1,2, … ,n

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

n. modi in cui si possono osservare x successi in n prove

$$p^x$$

x volte si osserva il successo associato a probabilità p
p*p*….*p x volte
si moltiplica perché le prove sono indipendenti

$$(1-p)^{(n-x)}$$

(n-x) volte si osserva l'insuccesso associato a probabilità (1-p)

… le caratteristiche della binomiale

VALORE ATTESO	**$E(X) = n\,p$**
VARIANZA	**$VAR(X) = n\,p\,(1- p)$**
FORMA	**simmetrica**

un esempio ...

considerate la nascita di 14 neonati

quanti maschietti vi aspettate di trovare tra 14 neonati?

E(X) = n · p = 14 · 0,5 = 7 *è piuttosto intuitivo, no?*

qual è la probabilità che vi siano esattamente 2 maschietti tra di loro?

p = 0,5
n = 14
x = 2

$$p(2) = \frac{14!}{2!(14-2)!} 0{,}5^{2} (1-0{,}5)^{14-2} = 0{,}005$$

la distribuzione IPERGEOMETRICA

* **si adatta sempre a situazioni in cui gli ESITI OSSERVABILI sono DUE**

* **soddisfa tutti i requisiti della binomiale**
  - **è definito un numero fisso di tentativi / prove**
  - **ciascuna prova ha solo due possibili esiti**

* **tranne che le PROVE NON SONO INDIPENDENTI cioè man mano che si effettuano le prove la probabilità di successo cambia**

es. n estrazioni senza reinserimento da un'urna contenente M palline bianche e (N-M) nere
tra le n palline estratte, quante sono le bianche?

… la formula dell' ipergeometrica

$$p(x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$$

M	**unità di "successo"**
x	**successi tra le n prove**
	max (0, n-(N-M)) <= x <= min (n,M)
N	**n. totale di unità**
n	**n. di prove**

es. estrazione di un campione di 100 studenti tra i 400 iscritti al diploma di ingegneria informatica sapendo che ci sono 25 studentesse e 375 studenti qual è la probabilità che del campione facciano parte 2 studentesse ?

spesso non è utile calcolare la probabilità di un preciso valore x ma di un intervallo (x>3, x<2, …)
per far questo si utilizzano le proprietà della probabilità

es. prob. di avere nel campione almeno una studentessa = 1 - prob. di estrarre zero studentesse

… le caratteristiche dell' ipergeometrica

VALORE ATTESO $E(X) = n\,p$

VARIANZA $VAR(X) = n\,p\,(1-p)\,(N-n)/(N-1)$

dove $p=M/N$

LIMITE **fissati x e n**
per N che tende all'infinito assieme ad M
la distribuzione ipergeometrica
tende alla distribuzione binomiale

es. estraendo un campione da popolazioni molto numerose
(N grande, es. la popolazione italiana)
anche se il campione è senza reinserimento
(cioè non è possibile che un individuo finisca nel campione due volte)
le probabilità di "successo" possono essere calcolate
usando la distribuzione binomiale

un esempio ...

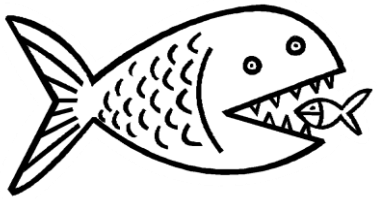
considerate un gruppo di 14 neonati
di cui 4 sono nati sottopeso e gli altri sono di peso regolare

estratto un campione di 5 neonati
qual è la probabilità di selezionarne 1 sottopeso?

N = 14
M = 4
n = 5
x = 1

… fate voi i calcoli …

la distribuzione di POISSON



* **si adatta a situazioni in cui si conta il numero di volte in cui accade un evento all'interno di un intervallo**

* **l'intervallo può essere di tempo, di spazio (distanza, area), …**

es. quante persone si aggiungono in coda in dieci minuti ?
quante auto passano al casello in un'ora ?
quanti studenti accedono alla biblioteca al giorno ?
quante persone si ammalano di tumore in un anno ?
quanti utenti si connettono ad un sito internet ?

… la formula della Poisson

$$p(x) = \frac{\lambda^x}{x!} e^{-\lambda}$$

x = 0,1,2, …

λ costante positiva

*** gli eventi devono accadere in modo indipendente gli uni dagli altri**

es. gli avventori che entrano in un negozio

*** l'accadere degli eventi deve avvenire in modo casuale, non sistematico**

es. l'afflusso degli operai in una fabbrica che lavora a turni non è il nostro caso

*** gli eventi devono essere uniformemente distribuiti all'interno dell'intervallo**

es. non devono accadere tutti all'inizio dell'intervallo
es. in un negozio di souvenir entrano i turisti di un pullman tutti insieme e poi più nessuno per tre ore: non va

… le caratteristiche della Poisson

VALORE ATTESO $E(X) = \lambda$

VARIANZA $VAR(X) = \lambda$

la Poisson approssima la Binomiale

* se la probabilità che accada l'evento di interesse è molto piccola

  e se il numero n di prove è molto elevato

* **il modello Binomiale può essere approssimato con la Poisson in cui il parametro $\lambda = n\,p$**

- è comodo quando n cresce e il calcolo del coefficiente binomiale diventa scomodo

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

Regola empirica: si usa la Poisson per approssimare la Binomiale se

n p <= 10
n >= 50

un esempio ...

si sa che, durante l'ora di ricevimento di un docente si presentano in media 2,3 studenti

qual è la probabilità che, scelta a caso un'ora di ricevimento di un certo giorno, non si presenti nessuno studente?

$\lambda = 2,3$

$$p(0) = \frac{\lambda^0}{0!} e^{-2,3} = 0,1$$

qual è la probabilità che si presenti almeno uno studente?

…. fate voi i calcoli ….

variabili casuali

continue

... ripetiamo un po' ... le v.c. continue

funzione di densità di probabilità

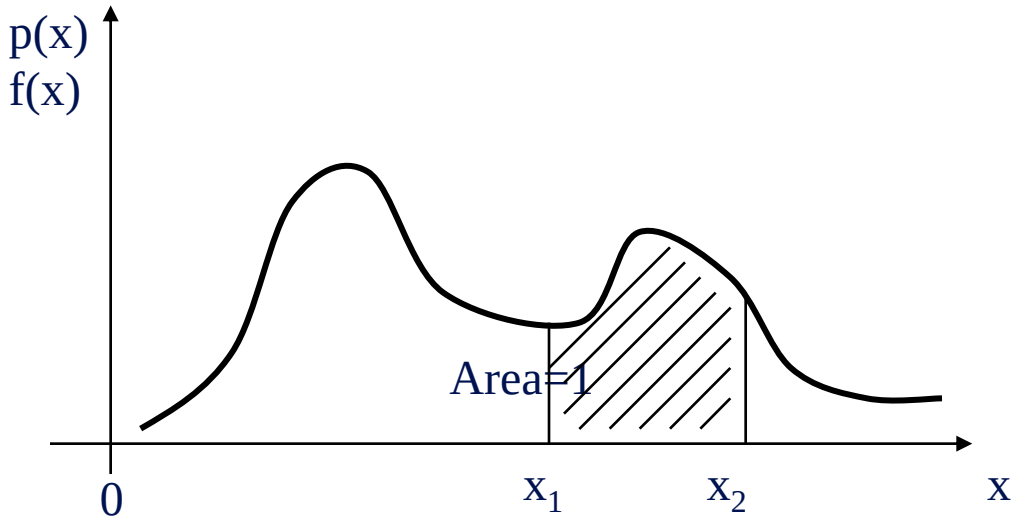
$$f(x) = \frac{d}{dx} F(x)$$

dove F(X) è la funzione di ripartizione

tale per cui

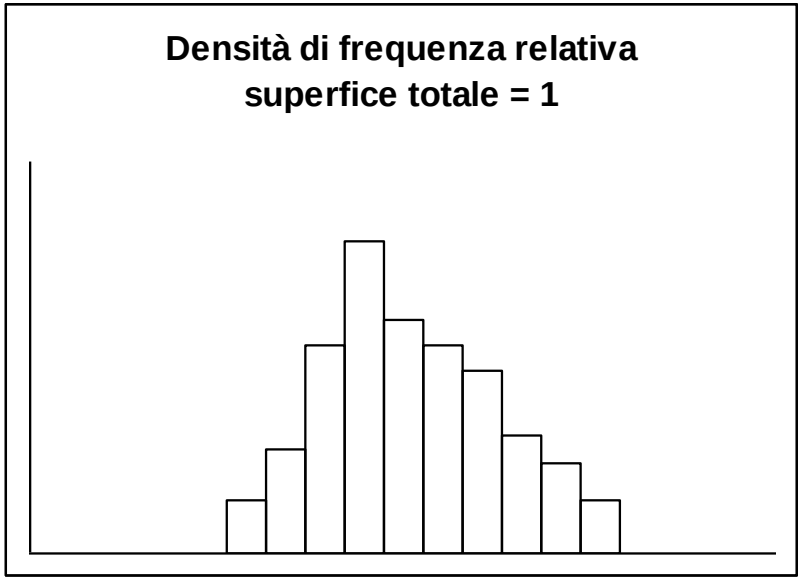
$$f(x) \geq 0$$

$$\int_{-\infty}^{+\infty} f(x)\,dx = 1$$



una "fetta" di area esprime la probabilità che vengano osservati valori di X contenuti nell'intervallo che si trova in ascissa "alla base della fetta"

dall’istogramma alla funzione di densità



se aumenta n e diminuisce l’ampiezza delle classi

le frequenze relative si dispongono come le probabilità
e
la densità di frequenza relativa tende ad approssimarsi alla curva che esprime la funzione di densità

valori caratteristici di una v.c. continua

$$\mu = \int_a^b x \cdot p(x)\,dx$$

a e b sono
gli estremi inf. e sup.
della v.c. X

$$\sigma^2 = \int_a^b (x-\mu)^2 \cdot p(x)\,dx$$

la v.c. RETTANGOLARE o UNIFORME

è continua

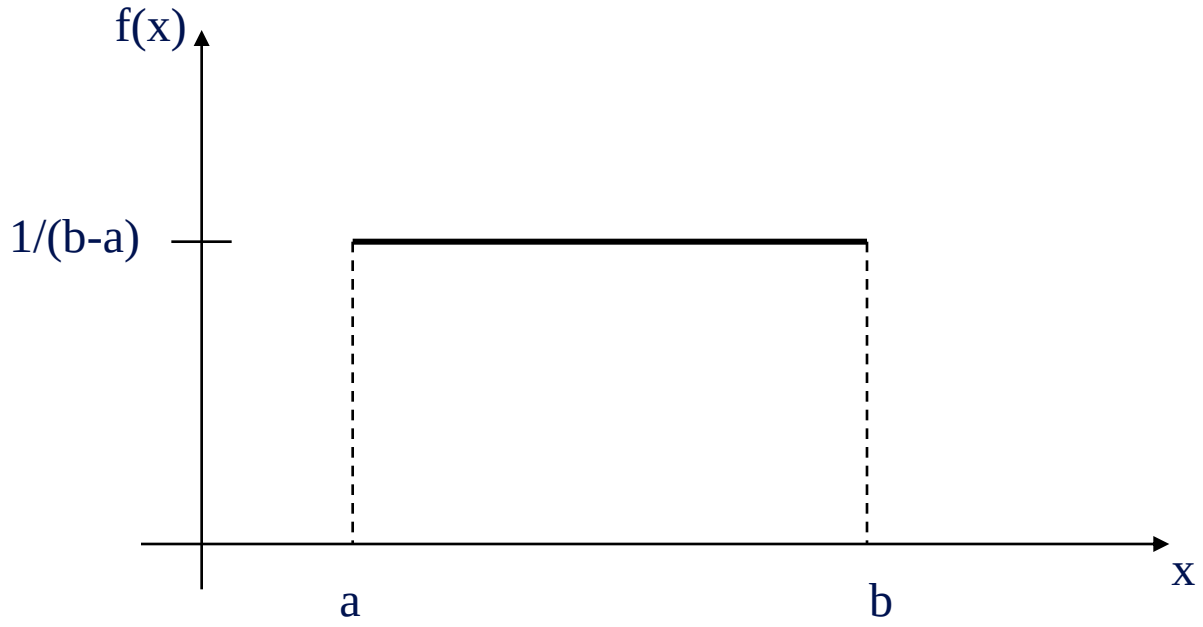
$$X \sim R(a,b)$$

si legge

“X si distribuisce come una v.c. rettangolare con parametri a e b”

$$f(x) = \frac{1}{b-a} \qquad a \le x \le b \qquad -\infty < a < b < +\infty$$

forma, media e varianza di R(a,b)

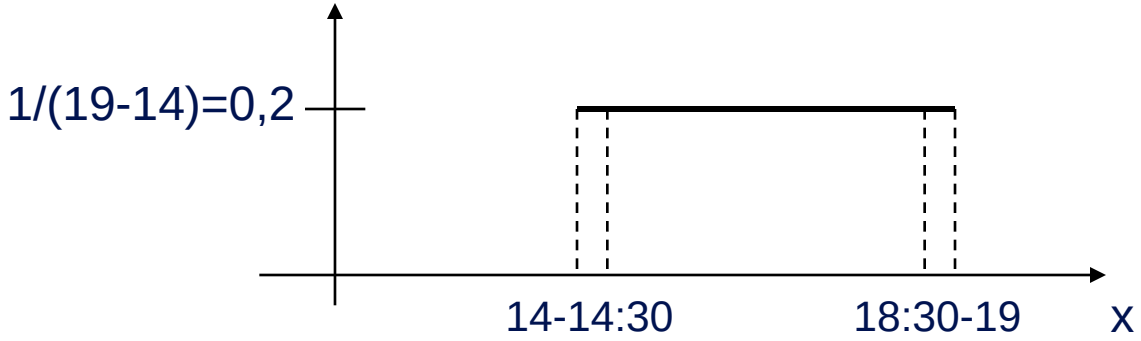


$$E(X) = \frac{a+b}{2} \qquad Var(X) = \frac{(b-a)^2}{12}$$

un’esercitazione antincendio

il responsabile della sicurezza del British Museum deve stabilire il momento in cui far scattare l’allarme antincendio per effettuare un’esercitazione fissata per il pomeriggio del 17 maggio 2002, tra le 14 e le 19

se il momento iniziale viene scelto casualmente nell’intervallo di tempo, questa situazione può essere descritta tramite la v.c. R(14,19)



qual è probabilità che l’allarme scatti nella prima o nell’ultima mezzora?

*area A = base per altezza = 0,5 * 0,2 = 0,1*
area B idem

probabilità = 0,2

la v.c. NORMALE o di GAUSS

* è la più utile e usata in statistica

* molti fenomeni si distribuiscono “normalmente”

è, ad esempio, un modello descrittivo
di caratteri antropometrici e
di caratteri sociali

* altre funzioni di probabilità (ad es. la binomiale) possono essere approssimate dalla normale

formula di N

$$X \sim N(\mu, \sigma^2)$$

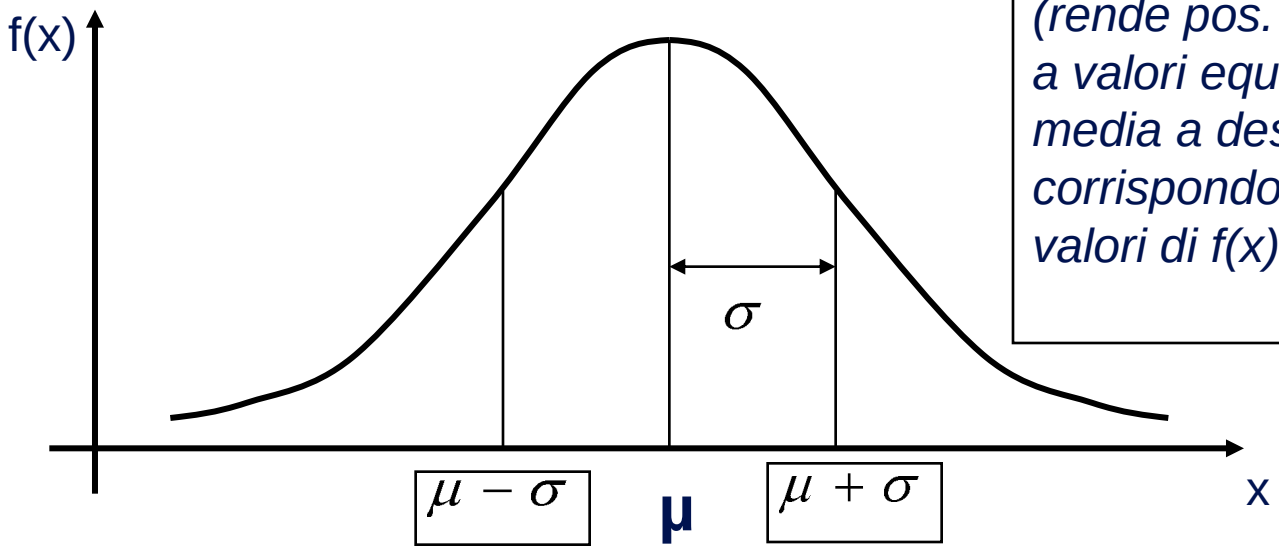
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \qquad -\infty < x < +\infty$$

$\frac{1}{\sigma\sqrt{2\pi}}$ **è un fattore di scala che rende l'area totale = 1**

π **3,14** e **2,718**

dalla formula alla forma di N

- **forma a campana**
- **simmetrica**
- **centrata sulla media**
- **il max si ha sulla media μ**
- **due flessi nei punti** $\mu - \sigma$ **e** $\mu + \sigma$



man mano che ci si allontana da μ: $(x-\mu)^2$ aumenta e poiché compare come esponente negativo: f(x) diminuisce

poiché (x-μ) è al quadrato (rende pos. gli scarti neg.) a valori equidistanti dalla media a destra e a sinistra corrispondono valori di f(x) uguali

valore atteso e varianza di N

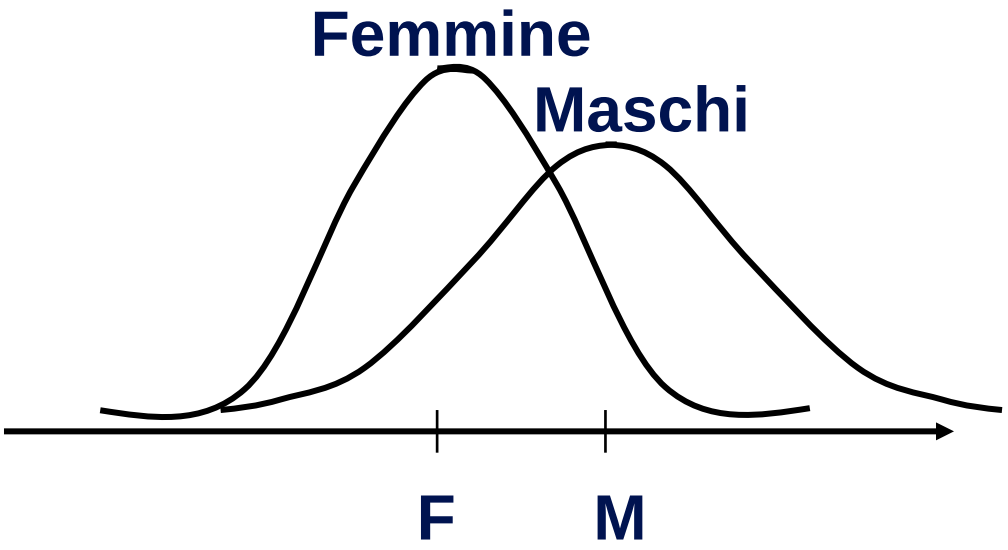
$$E(X) = \mu$$

$$Var\ (X) = \sigma^2$$

μ determina la posizione della curva sull'asse delle ascisse

sigma^2 determina la maggiore o minore concentrazione della curva intorno a μ

esempio: altezza di uomini e donne



la curva che descrive l'altezza delle donne è più alta di quella degli uomini: significa che le donne sono più alte degli uomini?

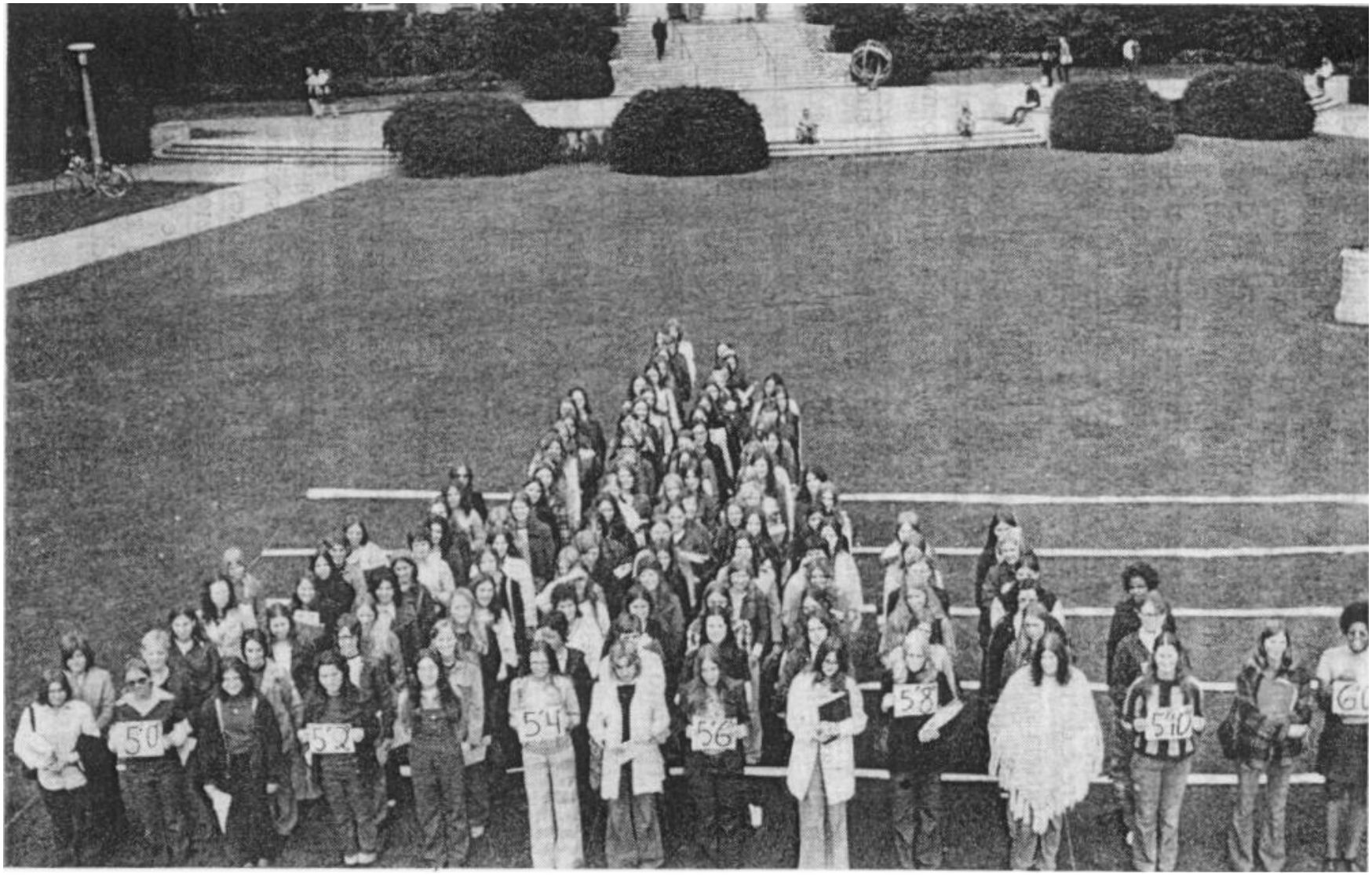
NO!

gli uomini sono in media più alti delle donne perché la loro curva è più a destra, cioè verso valori di X (altezza) maggiori

poiché le altezze dei maschi hanno una variabilità maggiore, la curva dei maschi è più bassa e larga (entrambe le aree sottese valgono 1)

proviamo a visualizzarlo …

Studentesse americane classificate secondo la statura



esempio: errori di misura casuali (non sistematici)

siano $g_1, g_2, \ldots, g_n$ n misure strumentali
di una grandezza incognita g *es. altezza di un monte*

siano $x_1=g_1-g, x_2=g_2-g, \ldots, x_n=g_n-g$

gli errori da cui ciascuna misura è affetta
a causa dell'imperfezione dello strumento di misura
e di altre cause accidentali di contesto

la v.c. X = "errore di misura" si distribuisce come una v.c. normale

cioè ci si aspettano errori distribuiti simmetricamente intorno alla loro media e con valori lontani da μ via via sempre meno probabili

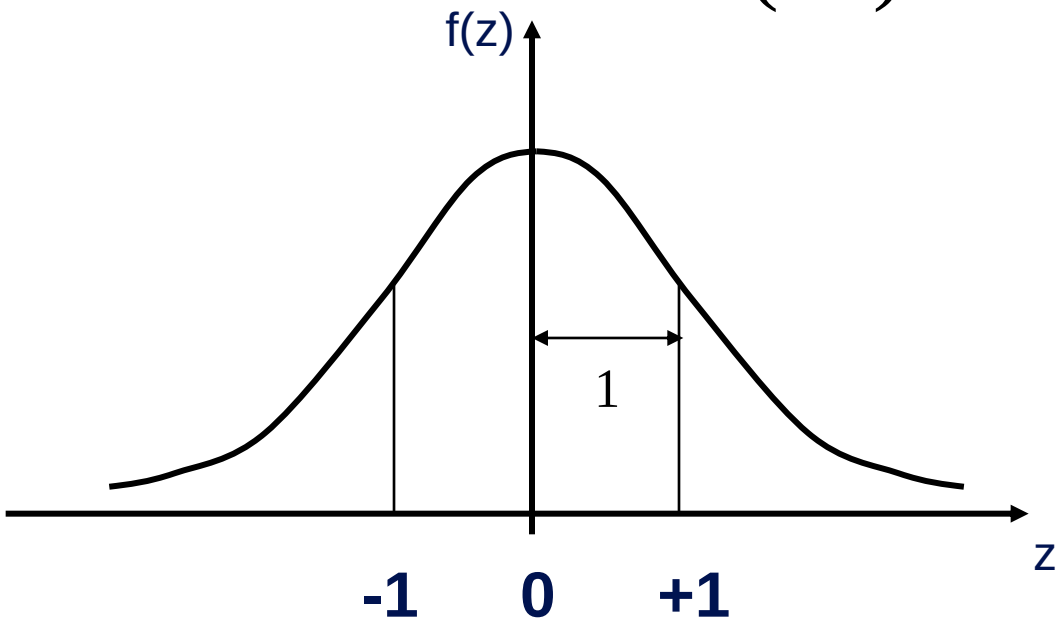
se μ = 0 si può sperare (e molti statistici lo fanno …)
che gli errori positivi e negativi si compensino e
nel complesso tendano ad annullarsi

la v.c. NORMALE STANDARDIZZATA

$$X \sim N(0,1)$$

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-1/2 \cdot z^2}$$

$$E(X) = 0 \quad Var(X) = 1$$



infinite normali = infiniti calcoli ?
e con quella formula così “brutta” ?

per fortuna NO!

una qualsiasi v.c. X normale
di media e varianza qualsiasi
può essere trasformata nella sua forma standardizzata

$$Z = \frac{X - \mu}{\sigma}$$

Z è una trasformazione lineare di X

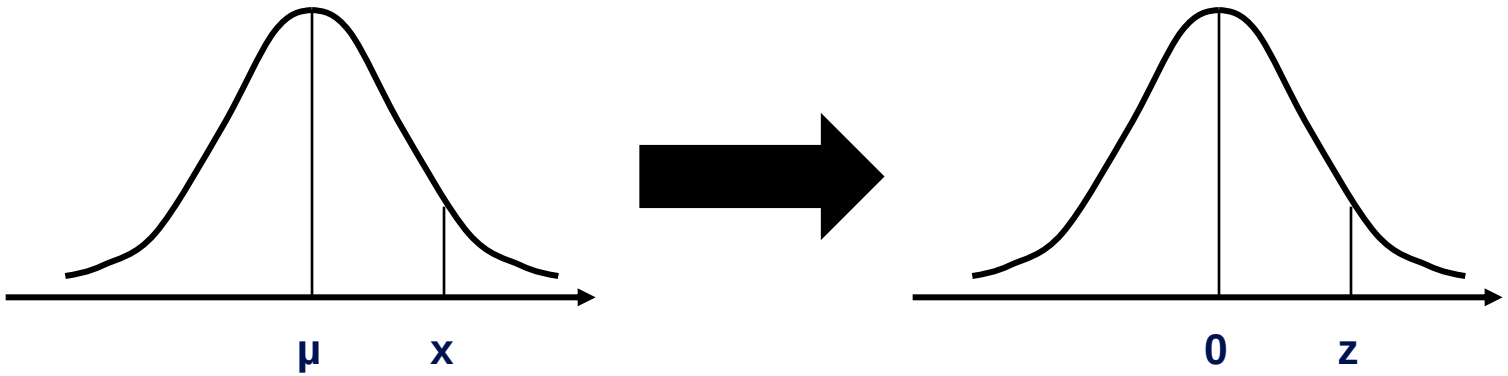
si cambia l’origine centrando i dati sul loro valore medio
si cambia l’unità di misura in modo che la varianza valga 1

mentre μ e sigma di X possono assumere infiniti valori
μ e sigma di Z valgono sempre 0 e 1

passando da X a Z …

le probabilità degli intervalli (eventi) rimangono inalterate

$$P(X \geq a) = P\left(\frac{X-\mu}{\sigma} \geq \frac{a-\mu}{\sigma}\right)$$



un esempio …

es. X è una v.c. $N(100,5^2)$

P(X>=110) = ?

$$P(X \geq 110) = P\left(\frac{X - 100}{5} \geq \frac{110 - 100}{5}\right) = P(Z \geq 2)$$

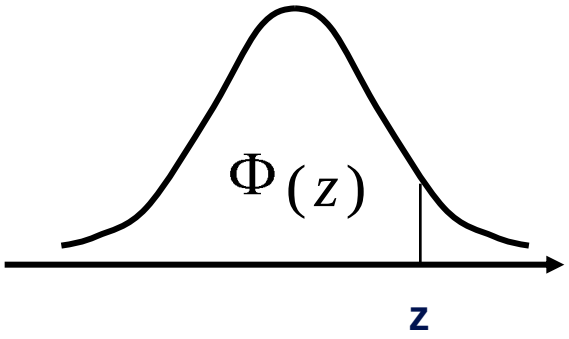
cioè, ci viene chiesto di calcolare qual è la probabilità di ottenere un valore distante dalla media aritmetica di oltre 2 sigma

ma allora bisogna calcolare le probabilità usando la formula della normale standard ?

per fortuna non serve fare neanche questo!

esistono della TAVOLE che forniscono i valori di probabilità associati a certi intervalli legati ai valori di Z
cioè forniscono i valori di porzioni di area sottese alla curva normale standardizzata

esiste una tavola che fornisce i valori della funzione di ripartizione di N (0,1) per valori di Z>0



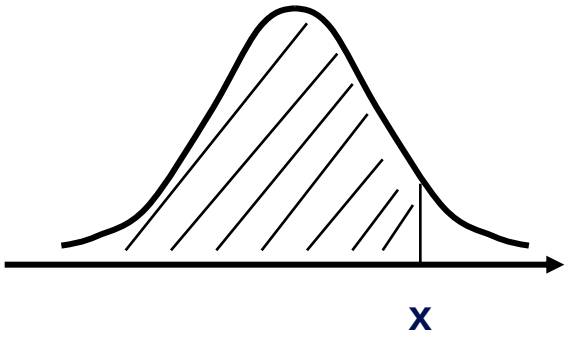
anche Excel ha una funzione analoga

f(x)

statistiche

DISTRIB.NORM

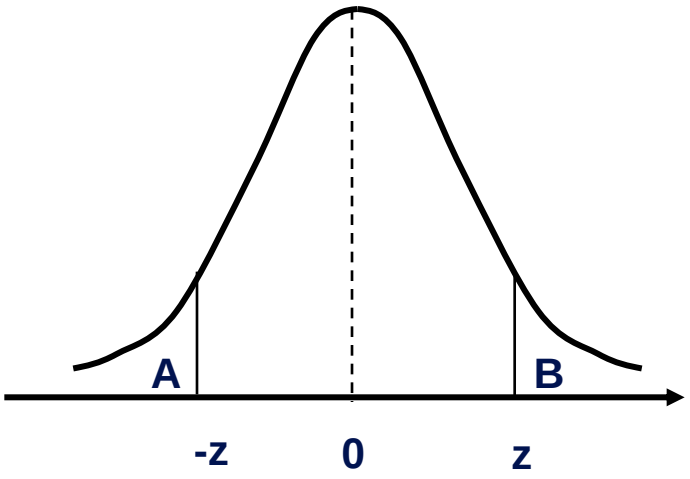
digitate il valore di X, la media, la deviazione standard e "VERO" e otterrete l'area in figura



e se ho un valore di z < 0 ?

si usa la proprietà di simmetria della curva e

si osserva che $\Phi(-z) = 1 - \Phi(z)$



Area A = Area B

Area B = $1 - \Phi(z)$

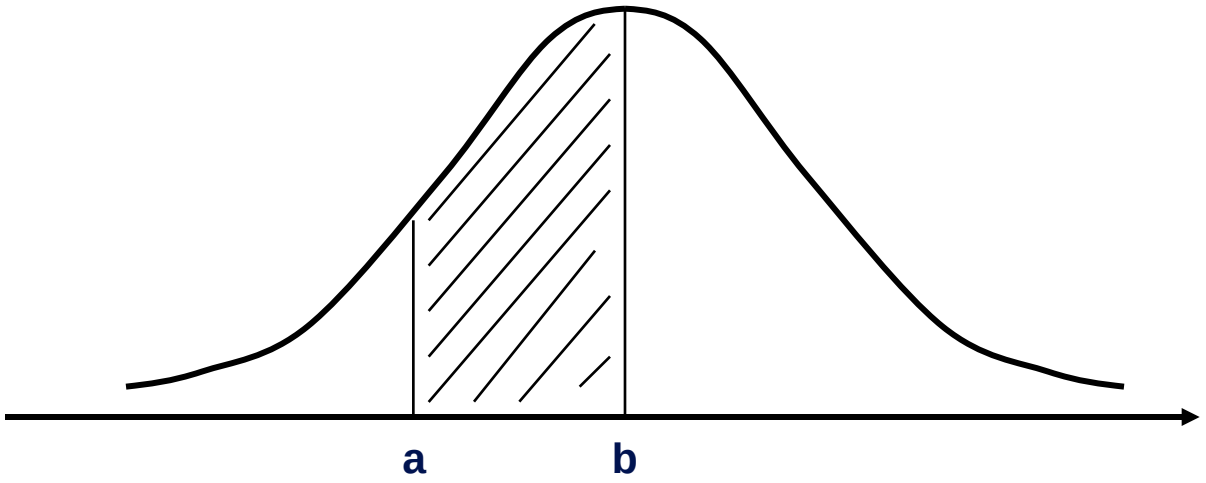
consultando le tavole è importante ricordare che i valori z sono sull'asse delle ascisse e non confonderli con le aree

i valori di z possono essere negativi le aree non lo sono mai

… giocando con le aree …

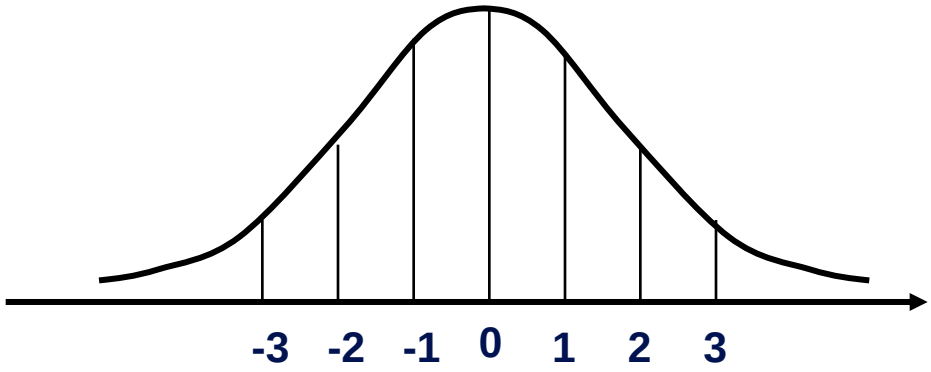
si calcolano le probabilità di un intervallo di valori di X

$$P(a \leq x \leq b) = \Phi(b) - \Phi(a)$$



perché sono considerati “strani” i valori lontani dalla media più di 2 volte sigma ?

se un fenomeno si distribuisce secondo una v.c. N si ha che:



- **circa il 68% di tutti i valori cade nell’intervallo di + e – 1 deviazione standard dalla media**

$$P(\mu - \sigma \leq x \leq \mu + \sigma) = 0,6826$$

- **circa il 95% dei valori cade nell’intervallo di + e – 2 deviazioni standard dalla media**

$$P(\mu - 2\sigma \leq x \leq \mu + 2\sigma) = 0,9544$$

- **e per 3 sigma …**

$$P(\mu - 3\sigma \leq x \leq \mu + 3\sigma) = 0,9974$$

come si procede per calcolare le probabilità nel caso di una v.c. Normale?

- si definiscono i valori di X, μ e sigma e l'evento di interesse
- si calcola il valore standardizzato z
- si disegna la curva normale individuando sul grafico l'area di interesse
- si usano tavole, simmetria e probabilità dell'evento complementare (1- …) per calcolare il valore della probabilità (area) che si desidera

come se la cavano le donne alla guida dei jet militari?

i seggiolini da espulsione dei jet militari americani "una volta" erano stati progettati per un pilota che pesasse tra le 140 e le 211 libbre
un pilota con peso < 140 o > 211 libbre potrebbe subire gravi danni fisici al momento dell'espulsione

X è la v.c. "peso medio delle donne", è noto che X è N (143,29^2)

qual è la prob. che una donna pilota rischi danni fisici al momento dell'espulsione?

piuttosto bene, visto che …

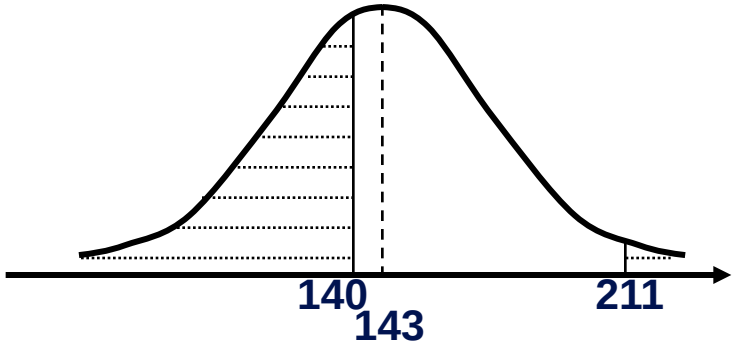
$$P(x \leq 140 \cup x \geq 211) =$$

$$P\left(z \leq \frac{140 - 143}{29}\right) + P\left(z \geq \frac{211 - 143}{29}\right) =$$

$$\Phi(-0{,}10) + \left[1 - \Phi(2{,}34)\right] =$$

$$\left[1 - \Phi(0{,}10)\right] + \left[1 - \Phi(2{,}34)\right] =$$

$$0{,}4602 + 0{,}0096 = 0{,}4698$$



cioè circa il 47% delle donne, quasi la metà, ha un peso al di fuori dei limiti consentiti dai seggiolini di espulsione dei jet

circa la metà delle donne pilota rischia di subire seri danni fisici in caso di espulsione: così non va!

i seggiolini sono stati modificati

la distribuzione della v.c. GAMMA

definiamo prima la FUNZIONE GAMMA

$$\Gamma(a) = \int_0^{+\infty} x^{a-1} e^{-x} dx = (a-1)! \qquad a > 0$$

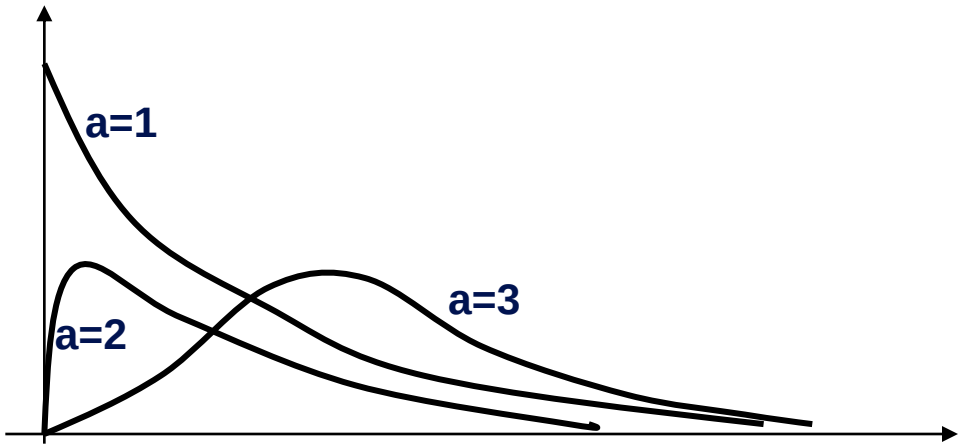
e poi la VARIABILE CASUALE GAMMA

$$X \sim G(\lambda, a) \qquad f(x) = \frac{1}{\Gamma(a)} \lambda^a x^{a-1} e^{-\lambda x}$$

forma, valor medio e varianza di G

la forma cambia abbastanza al variare dei parametri

ed esempio, per $\lambda = 1$
e a = 1,2,3



$$E(X) = \frac{a}{\lambda}$$

$$Var(X) = \frac{a}{\lambda^2}$$

utilizzo della Gamma

- **è un modello di sopravvivenza nello studio di apparecchi soggetti a logoramento e nella teoria delle file d'attesa**

- **la somma dei quadrati di variabili normali standardizzate ha distribuzione Gamma**

proprietà della Gamma

- $a \longrightarrow +\infty \qquad G(\lambda, a) \longrightarrow N\left(\frac{a}{\lambda}, \frac{a}{\lambda^2}\right)$

- **siano x1, X2, …, Xn n v.c. indipendenti tali per cui** $X_1 \sim G(\lambda, a_1)$ $X_2 \sim G(\lambda, a_2)$ … $X_n \sim G(\lambda, a_n)$

  allora la v.c.

$$Y = \sum_{i=1}^{n} X_i \sim G(\lambda, \sum_{i=1}^{n} a_i)$$

la distribuzione della v.c. ESPONENZIALE

è la distribuzione Gamma con parametro a = 1

$$X \sim E(\lambda) \qquad f(x) = \lambda e^{-\lambda x} \qquad 0 \leq x \leq +\infty \qquad \lambda > 0$$

$$E(X) = \frac{1}{\lambda} \qquad Var(X) = \frac{1}{\lambda^2}$$

da dove deriva la v.c. ESPONENZIALE ?

nasce a partire dal processo di Poisson

- **il n. di realizzazioni di un evento in un intervallo di tempo segue la legge di Poisson con parametro** $\lambda = \mu \cdot t$

$$P(x) = \frac{\lambda^x}{x!} e^{-\lambda}$$

- **supponiamo che un evento sia appena accaduto**
- **qual è la distribuzione di probabilità della v.c. X="tempo di attesa fino alla prossima realizzazione"?**
- **si definisce l'evento (x>t), cioè non accade nulla nel lasso di tempo di durata t**

… dalla Poisson

la funzione di ripartizione di X è $P(x \leq t) = F(t) = 1 - e^{-\mu \cdot t}$

la funzione di densità di X è $f(t) = \left. \frac{dF(v)}{dv} \right|_{v=t} = \mu \cdot e^{-\mu \cdot t}$

che è la funzione di densità esponenziale con parametro $\lambda = \mu$

teorema del limite centrale
e
distribuzioni collegate
alla normale

che cosa dobbiamo aspettarci
da un campione casuale
estratto da una popolazione nota?

le v.c. aiutano a rispondere a questa domanda

e perché ci poniamo questa domanda?

- **per calcolare la dimensione del campione (n)**
- **per scegliere il miglior modo di costruire il campione**
- **per stimare l'errore campionario che stiamo per commettere**
- **per verificare se le unità estratte si comportano come ci aspettiamo o sono "strane", particolari**
- **…**

un campione CASUALE

- **ciascuna unità della popolazione ha uguale probabilità di essere estratta**

- **può essere "senza reinserimento", cioè le unità estratte non vengono riesaminate, non possono essere estratte di nuovo**
  man mano che si estraggono le unità campione, cambiano le probabilità di estrarre le altre

- **può essere "con reinserimento", cioè ogni elemento campionato ritorna a far parte della popolazione e può essere estratto di nuovo**
  la probabilità di estrarre un elemento rimane costante

alcuni esempi …

- **un campione di lampadine estratto da un processo di produzione per verificarne la qualità**
- **un campione di studenti di ingegneria informatica per studiare il loro utilizzo e possesso di mezzi tecnologici moderni per la comunicazione**
- **un campione di cittadini aventi diritto al voto per prevedere i risultati delle prossime elezioni amministrative**
- **un campione di casalinghe per analizzare il fenomeno degli incidenti domestici**
- **…**

eventi rari, eventi frequenti, …, estrazioni costose o distruttive, …, variabili "sensibili", …, variabili soggettive, …

le unità campionarie come “palline”: *un esempio*

- ***immaginiamo un’urna gigantesca che contenga tutti gli italiani aventi diritto al voto (unità campionarie) e un enorme bambino bendato che estragga un votante***
- ***il primo estratto ha un suo orientamento al voto a noi sconosciuto, questa è la v.c. X_1***
- ***l’orientamento al voto del secondo estratto è la v.c. X_2***
- ***… e così via***
- ***ad un campione di n unità corrispondono n v.c. X_i***
- ***ciascuna di queste v.c. ha la stessa funzione di probabilità che è uguale alla funzione di probabilità della popolazione di origine***

es. se noi conosciamo le percentuali di voto delle elezioni amministrative passate, possiamo usarle per avere un’idea (cioè stimare) la funzione di prob. delle v.c. che ci interessano, cioè degli orientamenti al voto attuale

diamo forma matematica a questi concetti

dato un campione di n elementi

siano $X_1, X_2, \ldots, X_n$ le n v.c. che rappresentano il carattere di interesse posseduto dalle n unità

queste v.c. hanno la stessa funzione di probabilità

$$p(x_1) = p(x_2) = \ldots = p(x_n)$$

ciò vale sia per campioni casuali con reinserimento sia per campioni casuali senza reinserimento

ma è proprio vero che $p(x_1) = p(x_2) = \ldots = p(x_n)$?

per tutti i campioni casuali, con o senza reinserimento?

si perché:

nel campione senza reinserimento ciò che cambia è la probabilità condizionata $p(x_1|x_2)$

ma noi ora stiamo parlando della probabilità marginale $p(x_2)$ nel caso non si abbia nessuna conoscenza di X_1

e poi, quando la popolazione è molto numerosa (N grande) e si estrae un campione relativamente poco numeroso (n piccolo), il campionamento senza reinserimento è *praticamente* uguale a quello con reinserimento

l'indipendenza delle n v.c. X_i

nel campione casuale con reinserimento
le n v.c. $X_1, X_2, \ldots, X_n$ sono indipendenti l'una dall'altra

ma, se una popolazione è finita e
il campionamento è senza reinserimento,
le X_i sono dipendenti perché
la distribuzione condizionata di una X_i
dipende dai precedenti valori di $X_1, X_2, \ldots, X_{(i-1)}$

per quasi tutta la durata di questa lezione ragioneremo
nell'ipotesi che
le X_i siano tutte indipendenti

la SOMMA CAMPIONARIA

consideriamo la somma delle osservazioni campionarie

$$S = X_1 + X_2 + X_3 + \ldots + X_n$$

es. campione di barbabietole da un campo
v.c. X_i peso della i-ma barbabietola campione
quale sarà, in kg, il peso totale del campione?

S è una trasformazione lineare delle X_i e, per le proprietà viste nelle lezioni precedenti

$$E(S) = E(X_1) + E(X_2) + E(X_3) + \ldots + E(X_n)$$

il suo valor medio è uguale alla trasformazione lineare dei valori medi delle v.c. originali

il valore atteso della somma campionaria

sapendo che

*** μ è la media della popolazione**

*** ogni X_i ha distrib. uguale a quella della popolazione**

$$E(S) = \mu + \mu + \ldots + \mu \qquad \mu_S = n\mu$$

cioè la media di una somma campionaria è uguale alla media della popolazione originale moltiplicata per la dimensione del campione

es. so che in media una barbabietola pesa 0,5 Kg e
so che nel mio campo ci sono 10.000 barbabietole
se estraggo un campione di 100 barbabietole
*mi aspetto che il loro peso totale sia 50 kg (=100*0,5)*
e se verifico che le mie barbabietole campione pesano in totale solo 23 kg?
calcolo la probabilità che questo evento accada e
se la probabilità è bassa mi preoccupo e chiamo un agronomo!

la varianza della somma campionaria

$$Var\,(S) = Var\,(X_1 + X_2 + \ldots + X_n)$$

se $X_1, X_2, \ldots, X_n$ sono indipendenti

$$Var\,(S) = Var\,(X_1) + Var\,(X_1) + \ldots + Var\,(X_n)$$

poiché le X_i hanno tutte la stessa distribuzione, avranno anche tutte la stessa varianza sigma^2, quindi

$$Var\,(S) = \sigma^2 + \sigma^2 + \ldots + \sigma^2 \qquad \sigma^2{}_S = n\sigma^2$$

$$\sigma_S = \sqrt{n}\sigma$$

la variabilità della somma di n elementi è maggiore della variabilità campionaria di un singolo elemento, ma lo s.q.m. della somma campionaria non aumenta di n volte

perché, se il campione è casuale, ci si aspetta di osservare alcune unità campione sovradimensionate e altre sottodimensionate e quindi ci si aspetta che gli errori dei vari elementi che compongono la somma si compensino a vicenda

un esempio …

una macchina produce anelli di catena
la lunghezza media di un anello è μ = 0,4 cm e lo s.q.m. sigma=0,02 cm

una catena è composta di 100 anelli

la v.c. somma campionaria S descrive la lunghezza della catena

$$\mu_S = n\mu = 100 \cdot 0,4 = 40\,cm$$

possiamo ipotizzare che le lunghezze degli anelli siano indipendenti le une dalle altre, quindi

$$\sigma_S = \sqrt{n}\sigma = \sqrt{100} \cdot 0,02 = 0,2\,cm$$

la MEDIA CAMPIONARIA

consideriamo la media delle osservazioni campionarie

$$\overline{X} = \frac{1}{n}(X_1 + X_2 + \ldots + X_n) = \frac{1}{n}S$$

è una trasformazione lineare di S somma campionaria
e quindi

$$\mu_{\overline{X}} = \frac{1}{n}\mu_S = \frac{1}{n}(n\mu) = \mu$$

valore atteso

$$\mu_{\overline{X}} = \mu$$

$$\sigma^2_{\overline{X}} = \left(\frac{1}{n}\right)^2 \sigma^2_S = \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n}$$

varianza
e
s.q.m.

$$\sigma^2_{\overline{X}} = \frac{\sigma^2}{n}$$

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$$

osserviamo che …

le due formule

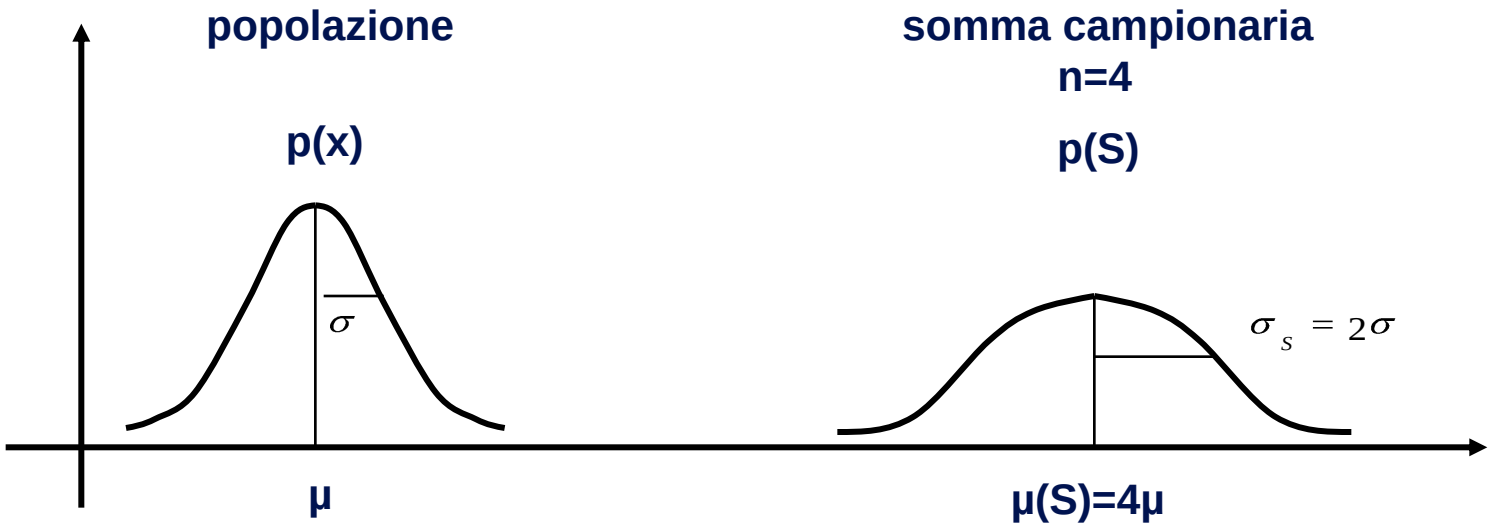
$$\mu_{\bar{X}} = \mu \qquad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

significano che man mano che n aumenta

* **la media rimane invariata**
* **la sua distribuzione si concentra intorno a μ**
  perché la variabilità diminuisce
  (n compare a denominatore)

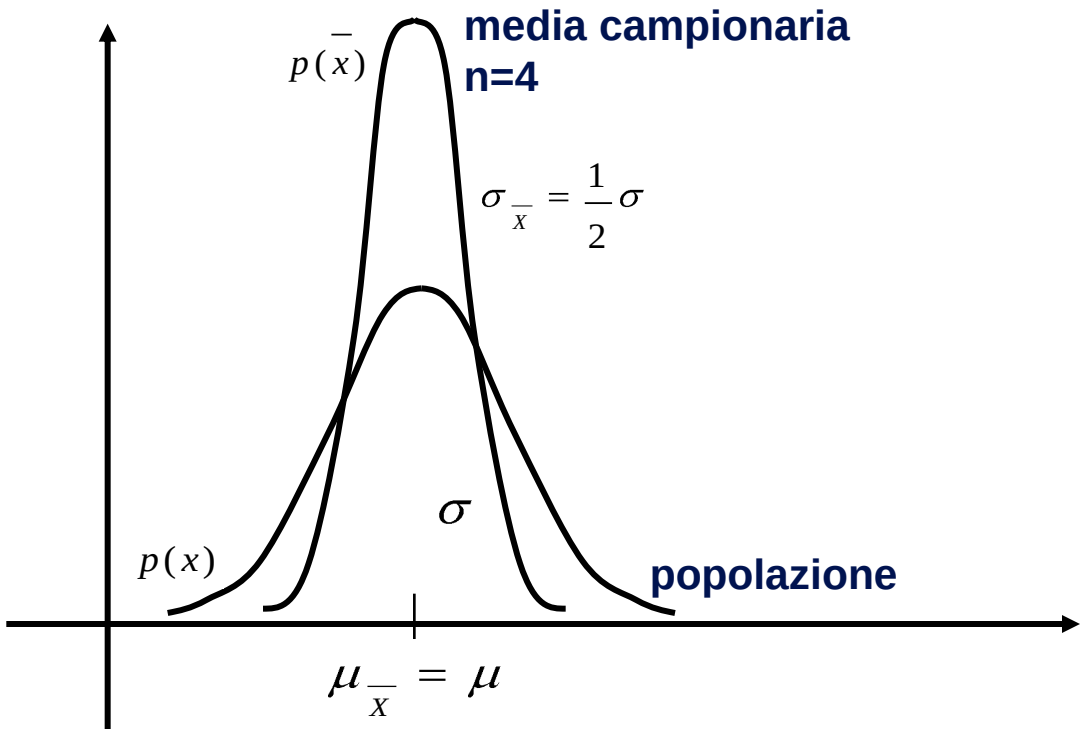
cioè più aumenta la dimensione del campione
meno errore si commette nel calcolare
la media campionaria dei dati

vediamolo dal punto di vista grafico …



… vediamolo dal punto di vista grafico …

cosa accade alla media campionaria se, ad esempio, se n=4



TEOREMA DEL LIMITE CENTRALE

descrive la forma della distribuzione
*** della somma e**
*** della media campionaria**

e dice che

aumentando la dimensione n del campione
la distribuzione della somma e
della media di un campione
estratto da una QUALSIASI popolazione
si approssima alla distribuzione NORMALE

somma

$$N(n\mu, \sqrt{n}\sigma)$$

media

$$N(\mu, \frac{\sigma}{\sqrt{n}})$$

accade sempre che …

al crescere di n

la funzione di densità della media campionaria si concentra su μ e tende ad assumere una forma campanulare

qualsiasi sia la forma delle v.c. originali
non solo se le v.c. originali sono normali

in pratica, per campioni in cui $n > 30$
si può considerare che la distribuzione della v.c. media campionaria sia normale

è molto comodo poter contare su questo teorema

**** <u>è il teorema cardine su cui si poggia la teoria statistica della stima e della verifica d'ipotesi</u>**

vediamone la formulazione matematica

siano X_1, X_2, … X_n n variabili casuali indipendenti e identicamente distribuite con media $E(X_i)=\mu$ e varianza $Var(X_i)=sigma^2$ entrambe finite

sia $S_n = X_1 + X_2 + \ldots + X_n$ la loro somma avente media $E(S_n)=n\cdot\mu$ e $Var(S_n)=n\cdot sigma^2$

si ha allora che

$$\lim_{n\to\infty} P\left(\frac{S_n - n\mu}{\sqrt{n}\sigma} \leq z \right) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du$$

cioè la successione Z_n (di v.c. standardizzate) converge in distribuzione alla normale N(0,1)

sia per v.c. discrete, sia per v.c. continue

una formulazione alternativa

salve le condizioni già citate, il teorema può essere formulato anche come segue

$$\lim_{n\to\infty} P\left(z_1 \le \frac{S_n - n\mu}{\sqrt{n}\sigma} \le z_2 \right) = \frac{1}{\sqrt{2\pi}} \int_{z_1}^{z_2} e^{-u^2/2} du$$

cioè
non appena n cresce a sufficienza
senza fare troppa fatica
abbiamo a disposizione la formula per calcolare
la probabilità che la somma campionaria sia contenuta
all'interno di un intervallo di valori

un’altra formulazione alternativa

salve le condizioni già citate, il teorema può essere formulato anche come segue

$$\lim_{n\to\infty} P\left(a \le S_n \le b\right) = \frac{1}{\sqrt{2\pi}} \int_{\frac{a-n\mu}{\sqrt{n}\sigma}}^{\frac{b-n\mu}{\sqrt{n}\sigma}} e^{-u^2/2}\,du$$

cioè

il teorema può essere riformulato per la v.c. somma non standardizzata

ma quando si può considerare buona l'approssimazione ?

autori diversi danno versioni diverse

ma, per essere cauti

con n>30 si va abbastanza sul tranquillo

soprattutto se la distribuzione originale delle v.c. X_i pur non essendo normale
è comunque abbastanza simmetrica

****

se le v.c. originali sono normali,
il teorema è verificato per ogni n

un esempio …

grazie alle nozioni acquisite finora potremo risolvere problemi del tipo

un ascensore ha portata massima di 1.000 kg e una capacità di 10 persone

se i pesi delle persone che lo usano abitualmente sono distribuiti normalmente con una media di 90 kg e una deviazione standard di 11 kg

qual è la probabilità che un gruppo di 10 persone ecceda il limite di portata dell'ascensore?

un esempio … per chi è paziente e curioso

considerate una popolazione di cifre 0,1,2,3,4,5,6,7,8,9
estratte casualmente con reinserimento

singola prova: estrarre casualmente una cifra
variabile casuale: valore della cifra estratta X {0,1,2,3,4,5,6,7,8,9}
distribuzione di probabilità: P(x) = 1/10 0 <= x <= 9

supponiamo di estrarre casualmente tanti campioni, ciascuno di numerosità 4

per ogni campione calcoliamo il valore atteso della v.c. X

il teorema del limite centrale ci dice che la v.c. media delle X_i è di tipo Normale

con pazienza, potreste generare in excel "n" insiemi di 4 numeri casuali compresi tra 0 e 9 e calcolare per ciascun insieme il relativo valor medio
poi, potreste rappresentare con un istogramma la distribuzione di frequenza di questi n valori medi

al crescere di n potreste verificare che l'istogramma tende ad assumere una forma campanulare del tipo normale

un esempio …

è noto che il peso degli abitanti di una nazione X segue una distribuzione normale con media 64,35 e s.q.m 13,04

si estragga un campione di 36 individui

qual è la probabilità che il peso medio di questi 36 individui sia maggiore di 67,5 kg ?

usiamo il teorema del limite centrale:
la v.c. peso medio campionario si distribuisce come un normale di media 64,35 e s.q.m. 2,17 = (13,04 / 6)

calcoliamo il valore standardizzato

$z = (67,5 – 64,35) / 2,17 = 1,45$

e usiamo le tavole della normale standard per calcolare la probabilità che ci interessa

$P (z > 1,45) = P (x \text{ medio} > 67,5) = 0,0735$

… continuiamo l'esempio …

calcoliamo la prob. che il peso di un solo individuo sia > 67,5 kg usando il solito metodo della standardizzazione troveremo che

A. $P(x > 67{,}5) = 0{,}405$

prima avevamo calcolato che per n=36

B. $P(x\ \text{medio} > 67{,}5) = 0{,}0735$

possiamo osservare che la prob. A è molto maggiore della prob. B

cioè, è molto più facile (probabile) per un individuo singolo deviare dalla media di quanto non accada per un gruppo di 36 individui

cioè, calcolando la media di n misure, si attutisce l'effetto di valori lontani dalla media, perché ci si aspetta che alcuni valori siano sopra la media, altri sotto la media e che, quindi, in qualche modo si compensino

correzione per popolazioni finite

quando si effettua un campionamento senza reinserimento da popolazioni finite è necessario introdurre un coefficiente correttivo nel calcolo della varianza di somma e media campionaria

somma

$$\sigma_S = \sqrt{n}\sigma \cdot \left[\frac{N-n}{N-1}\right]$$

media

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \cdot \left[\frac{N-n}{N-1}\right]$$

lo si applica nel caso di campionamento senza reinserimento in cui la dimensione del campione n superi il 5% della popolazione totale

cioè se n > 0,05 N

che effetto ha il coefficiente correttivo?

$$\left[\frac{N-n}{N-1}\right] \leq 1$$

il coef. correttivo è minore di 1
quindi riduce il valore della varianza

e, in effetti, il campionamento senza reinserimento comporta una media campionaria più significativa con minor varianza
dato che i valori estremi della popolazione possono essere campionati una volta sola e non possono tornare a far parte del campione, aumentandone la variabilità

la PROPORZIONE CAMPIONARIA

consideriamo una situazione rappresentabile da una distribuzione binomiale
e una popolazione bernoulliana

- **deve essere fissato un numero n di prove**
- **le prove devono essere indipendenti**
- **ogni prova deve avere soltanto due possibili esiti (0,1)**
- **le probabilità dei due eventi 0 e 1 devono rimanere costanti per ogni prova**

notate che siamo veramente lontani da una situazione adatta a essere descritta tramite una v.c. normale

distribuzione di probabilità per una popolazione bernoulliana

x	p(x)	x·p(x)
0	(1-p)	0
1	p	p

E(x) = p **Var(x) = p · (1-p)**

es. X=fare o meno sport
0=non fare sport 1=fare sport

in queste condizioni la somma campionaria esprime il numero di "successi" nel campione

es. S = n. di persone che fanno sport nel campione

la v.c. somma S può essere approssimata tramite la distribuzione Normale

$$N\left(np\,,\sqrt{np\,(1-\,p)}\right)$$

dalla somma S alla proporzione P

la proporzione campionaria è pari a

$$P = \frac{S}{n} = \overline{X}$$

es. P = percentuale di persone che fanno sport nel campione

poiché P può essere interpretata come una media campionaria, anche la v.c. proporzione P può essere approssimata tramite la distribuzione Normale

$$N\left(p, \sqrt{\frac{p(1-p)}{n}} \right)$$

formalizzazione dei concetti appena esposti

siano $X_1, X_2, \ldots, X_n$ v.c. indipendenti di tipo bernoulliano
le X_i assumono valori 0 e 1 con probabilità (1-p) e p

la v.c. S somma campionaria
che ha distribuzione binomiale
con $E(S_n)=np$ e $Var(S_n)=np(1-p)$

è tale per cui

$$\lim_{n\to\infty} P\left(\frac{S_n - np}{\sqrt{np(1-p)}} \le z \right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-u^2/2} du$$

quando si può considerare buona questa approssimazione ?

il ritmo di convergenza è rapido se la binomiale è simmetrica, cioè se p è vicina a 0,5

ma, in pratica, se

$$np > 5 \quad \text{e} \quad n(1-p) > 5$$

l'approssimazione è accettabile

come procedere per usare questo risultato

- **stabilite se è il caso di usare l'approssimazione normale cioè verificate se np>5 e n(1-p)>5**

  es. campione di 50 bimbi tra i neonati di un ospedale

- **trovate i valori di media e varianza con (np) e [np(1-p)]**

  es. attesi in media 25 maschietti con s.q.m. 3,53

- **identificate l'evento e il valore di x (n. di successi) che vi interessa**

  es. meno di 10 maschietti nel campione

- **standardizzatelo e calcolate z**

  es. meno di [(10-25)/3,53] maschietti standard

- **usate le solite tavole per calcolare la probabilità che cercate**

è bene utilizzare la CORREZIONE PER CONTINUITA’

quando si usa la distribuzione normale (continua)
per approssimare la binomiale (discreta)

serve una correzione per continuità

per tener conto che il numero intero x nella binomiale

viene rappresentato dall’intervallo [x-0,5 ; x+0,5] nella scala continua della v.c. normale

la correzione consiste nel sommare o sottrarre 0,5 al numeratore della variabile standardizzata

si sottrae o si somma a seconda della probabilità a cui si è interessati

quando sommare e quando sottrarre 0,5 ?

spieghiamolo con degli esempi:

evento	**area di interesse**
almeno 520 (include 520 e più)	*a destra di 519,5 (-0,5)*
più di 520 (non include 520)	*a destra di 520,5 (+0,5)*
al massimo 520 (include 520 o meno)	*a sinistra di 520,5 (+0,5)*
meno di 520 (non include 520)	*a sinistra di 519,5 (-0,5)*
esattamente 520	*tra 519,5 e 520,5*

ma la dobbiamo proprio usare questa correzione di continuità ?

SI !

quando utilizzerete la statistica nel vostro lavoro

o

nel compito se vorrete ottenere la lode

ma non sarete penalizzati se non la userete

e ora un po’ di teoria

propedeutica alle prossime nozioni

combinazione lineare di v.c. normali

siano X_1, X_2, …, X_n v.c. indipendenti aventi distribuzioni

$$N(\mu_1, \sigma^2_1) \quad N(\mu_2, \sigma^2_2) \quad \dots \quad N(\mu_n, \sigma^2_n)$$

allora la v.c.

$Y = a_1X_1 + a_2X_2 + \dots + a_nX_n$ **$a_1, a_2, \dots$ costanti reali**

ha distribuzione normale $$X \sim N\left(\sum_{i=1}^{n} a_i \mu_i, \sum_{i=1}^{n} a_i^2 \sigma_i^2\right)$$

la distribuzione CHI-QUADRATO

la v.c. continua X ha distribuzione chi-quadrato se

$$X \sim \chi^2(r) \qquad f(x) = \frac{1}{2^{r/2}\Gamma\left(r/2\right)} x^{r/2-1} e^{-x/2} \qquad 0 \le x \le \infty$$

dove r è chiamato “n. dei gradi di libertà”

$$E(X) = r \qquad Var(X) = 2r$$

un teorema che lega Normale e Chi-quadrato

siano X_1, X_2, …, X_n v.c. indipendenti e identicamente distribuite secondo una $N(\mu, \sigma^2)$

allora la v.c.

$$Y = \left(\frac{X_1 - \mu}{\sigma}\right)^2 + \left(\frac{X_2 - \mu}{\sigma}\right)^2 + \ldots + \left(\frac{X_n - \mu}{\sigma}\right)^2$$

ha distribuzione chi-quadrato con n gradi di libertà

la distribuzione *t* di Student

la v.c. continua X ha distribuzione *t* di Student se

$$X \sim t(r) \qquad f(x) = \frac{\Gamma\left((r+1)/2\right)}{\sqrt{\pi r}\,\Gamma\left(r/2\right)} \left(1 + \frac{x^2}{r}\right)^{-\frac{(r+1)}{2}} \qquad -\infty \le x \le +\infty$$

dove r, intero positivo, indica i gradi di libertà (g.d.l.)

$$E(X) = 0 \qquad Var(X) = \frac{r}{(r-2)}$$

al crescere dei g.d.l.

t **tende alla normale standardizzata N(0,1)**

un teorema che lega Normale, chi e *t*

siano U e V due v.c. indipendenti tali che

$$U \sim \mathrm{N}(0,1) \qquad V \sim \chi^2(r)$$

allora la v.c. $$t = \frac{U}{\sqrt{V/r}}$$

ha distribuzione *t* di Student con r gradi di libertà

la distribuzione F di Snedecor

la v.c. continua X ha distribuzione F di Snedecor se

$$X \sim F(r_1, r_2) \qquad f(x) = \frac{\Gamma\left((r_1 + r_2)/2\right)}{\Gamma\left(r_1/2\right)\Gamma\left(r_2/2\right)} \left(\frac{r_1}{r_2}\right)^{\frac{r_1}{2}} x^{\frac{r_1}{2}-1} \left(1 + \frac{r_1}{r_2} x\right)^{-\frac{(r_1+r_2)}{2}} \qquad 0 \le x \le \infty$$

dove r_1 e r_2, interi positivi, indicano i gradi di libertà (g.d.l.)

$$E(X) = \frac{r_2}{(r_2 - 2)} \qquad Var(X) = \frac{2 r_2^2 (r_1 + r_2 - 2)}{r_1 (r_2 - 2)^2 (r_2 - 4)}$$

un teorema che lega chi e F

siano U e V due v.c. indipendenti tali che

$$U \sim \chi^2(r_1) \qquad V \sim \chi^2(r_2)$$

allora la v.c.

$$F = \frac{U / r_1}{V / r_2}$$

ha distribuzione F di Snedecor con r_1 e r_2 gradi di libertà

grazie al cielo

qualcuno ha pensato bene di tabulare

i valori di chi-quadrato, *t* e F

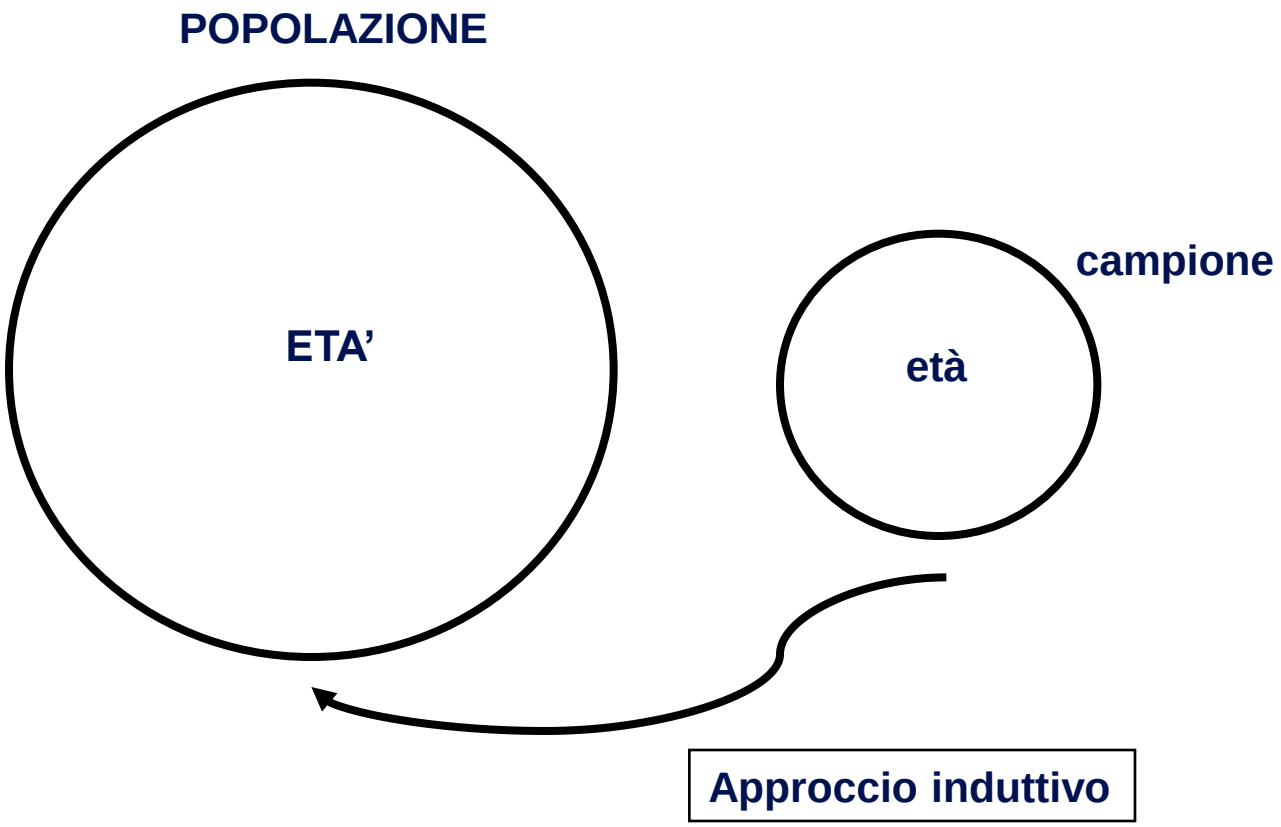
popolazione, campione e inferenza statistica & distribuzioni campionarie

popolazione e campione

- **lo studio dei fenomeni si attua spesso attraverso dei campioni, cioè una parte di tutta la popolazione**
- **dallo studio del campione, attraverso tecniche statistiche, si cerca di avere informazioni sull'intera popolazione**
- **questo procedimento si chiama**

  INFERENZA INDUTTIVA

Inferenza induttiva



un altro significato della parola STATISTICA

una statistica S è una funzione h
indipendente da parametri incogniti
che, riferita alle variabili campionarie $X_1, X_2, ..., X_n$
genera una variabile casuale

$S = h(X_1, X_2, ..., X_n)$

applicata ai valori $x_1, x_2, \ldots, x_n$ del campione
la funzione h assume un valore numerico

$s = h(x_1, x_2, \ldots, x_n)$

Inferenza statistica

a che serve?

ad acquisire informazioni su di una popolazione usando le osservazioni di un campione da essa estratto

come funziona?

le informazioni si ottengono attraverso il calcolo di una statistica campionaria
(elaborazione sui dati del campione)

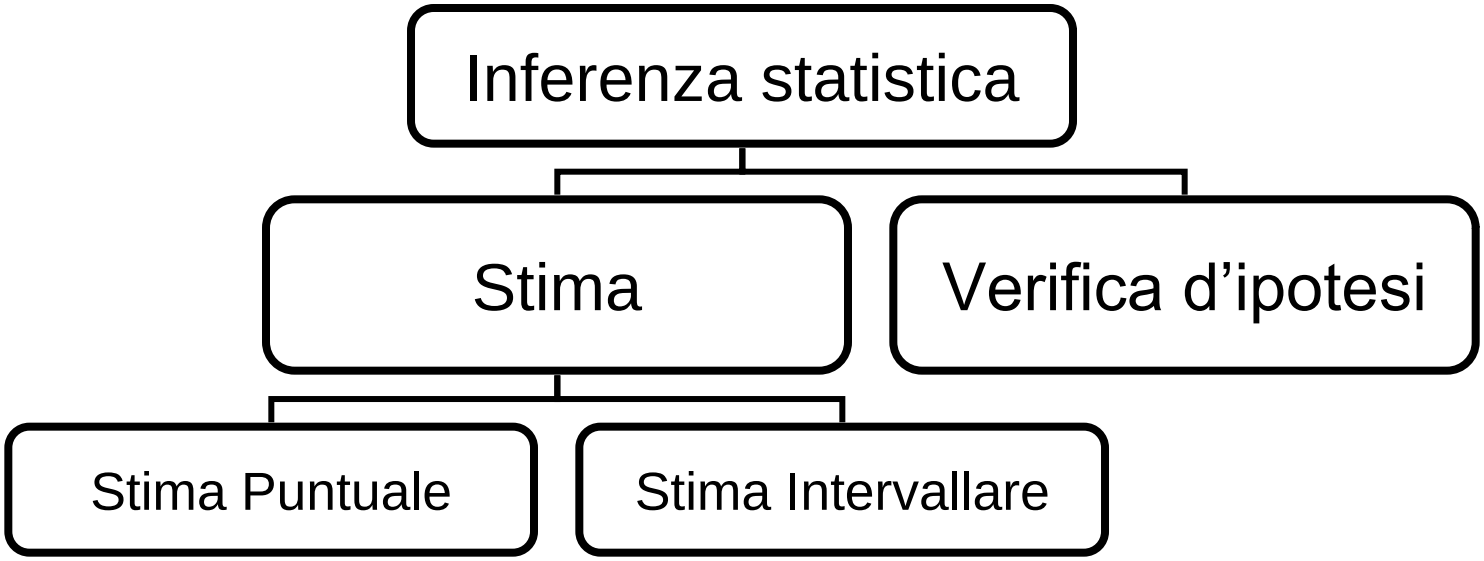
ma la statistica è un numero o una variabile casuale ?

?!?!?!?!?

- **per ogni campione si ottiene un solo risultato un solo numero reale (s)**
- **ma da una popolazione si possono estrarre tanti campioni diversi e calcolare per ciascuno di essi un diverso valore di s**
- **al variare del campione, il valore della statistica varia: è una variabile casuale S**

in realtà, gli statistici lavorano quasi sempre con un campione solo e immaginano, con la fantasia, di poterne estrarre tanti

vari modi di fare inferenza statistica



la STIMA

- **consiste nell'attribuire il valore "più appropriato" ad un parametro sconosciuto della popolazione**

  es. si vuole conoscere il peso medio degli studenti di ingegneria, attraverso un campione degli iscritti all'A.a. 2001-2002

- **la si calcola**

  - **attraverso i risultati campionari**

    es. media dei pesi degli studenti estratti nel campione

  - **nel rispetto di certi criteri di ottimalità**

    es. devo calcolare un valore preciso al grammo o al chilo?

  - **con prefissata probabilità di errore**

    es. data la numerosità campionaria e il tipo di fenomeno studiato (variabilità, …) mi potrò sbagliare di circa X grammi

la STIMA PUNTUALE

il parametro (caratteristica) sconosciuto della popolazione viene rappresentato con il valore puntuale espresso dalla statistica campionaria

$$t\,(x_1, x_2, \ldots, x_n)$$

es. dato un campione di 100 studenti iscritti nel 2001 stimo che l'altezza media degli studenti di ingegneria sia 174,5 cm (pari all'altezza media dei 100 studenti campione)

la STIMA INTERVALLARE

l'informazione sul parametro (caratteristica) sconosciuto della popolazione viene rappresentato con due statistiche campionarie

$$t_1 (x_1, x_2, \ldots, x_n) \qquad t_2 (x_1, x_2, \ldots, x_n)$$

dove $t_1 < t_2$ sono tali per cui (t_1, t_2) costituisce un intervallo per il quale si può determinare a priori la probabilità che contenga l'ignoto parametro

es. dato un campione di 100 studenti iscritti nel 2001
stimo che, con buona probabilità (95%),
l'altezza media degli studenti di ingegneria
è compresa tra 172 e 176 cm
non vuol dire che il 95% degli studenti è alto tra 172 e 176 cm

la VERIFICA d'IPOTESI

- **consiste nell'accettare o respingere un'ipotesi formulata su un parametro della popolazione**

  es. l'esercito deve verificare se i militari di oggi continuano ad avere lo stesso fisico (altezza, peso, torace, …) di qualche decennio fa per valutare se cambiare le percentuali di produzione delle divise secondo le varie taglie

- **la si effettua**

  - **attraverso i risultati campionari**

    es. cfr media di un campione con valore di interesse

  - **con prefissata probabilità di errore**

    es. data la numerosità campionaria e il tipo di fenomeno studiato (variabilità, …) mi potrò sbagliare con probabilità molto bassa (5%)

un po’ di linguaggio settoriale

STIMATORE

è una formula o un procedimento di calcolo da applicare ai dati campionari per stimare un parametro di popolazione

STIMA

uno specifico valore (o intervallo di valori) usato per stimare / approssimare un parametro sconosciuto di popolazione

es. come viene spontaneo pensare,
la media campionaria è la migliore stima puntuale
per la media della popolazione

come deve essere uno stimatore per fornire delle buone stime ?

CRITERI DI OTTIMALITA'

- **CORRETTEZZA**
  il valore medio della v.c. Statistica
  è uguale al parametro da stimare

$$E(\overline{X}) = \mu$$

- **CONSISTENZA**
  all'aumentare della dimensione campionaria
  il valore della statistica
  tende a quello del parametro

$$\lim_{n \to +\infty} \Pr(|\ \overline{X} - \mu\ | < \varepsilon) = 1$$

- **EFFICIENZA**
  a parità di dimensione campionaria
  tra diversi stimatori
  conviene scegliere quello con varianza minore

$$Var\ (S_1) < Var\ (S_2)$$

l'intuito e il buon senso sono sempre d'aiuto …

una buona regola per scegliere gli stimatori è quella di prediligere gli

STIMATORI NATURALI

cioè quelli che hanno lo stesso significato dei parametri incogniti della popolazione

$$\bar{x} = \sum \frac{x_i}{n} \quad \Rightarrow \quad \mu$$

media campionaria

$$S_c^2 = \sum \frac{(x_i - \bar{x})^2}{(n-1)} \quad \Rightarrow \quad \sigma^2$$

varianza campionaria corretta

la DISTRIBUZIONE CAMPIONARIA

- **è la distribuzione di tutti i possibili valori che possono essere assunti da una statistica S**

  calcolata a partire da campioni della stessa dimensione n

  estratti casualmente dalla stessa popolazione

- **permette di rispondere a questioni probabilistiche su statistiche campionarie**

- **fornisce la teoria necessaria per utilizzare procedure valide di inferenza statistica**

una Statistica importante: la MEDIA CAMPIONARIA

sia (X_1, X_2, …, X_n) un campione casuale e sia

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_1$$

la media del campione

prima di estrarre il campione $\overline{X}$, funzione di v.c., è anch'essa una v.c.

e avrà quindi una sua distribuzione di probabilità, un suo valore atteso e una sua varianza

$$E(\overline{X}) = \mu \qquad Var(\overline{X}) = \frac{\sigma^2}{n}$$

è uno stimatore interessante perché

$$E(\overline{X}) = \mu$$

cioè la media campionaria
si distribuisce intorno al
parametro sconosciuto μ
che è proprio il suo valore atteso

$$Var(\overline{X}) = \frac{\sigma^2}{n}$$

al crescere di n (numerosità campionaria)
la variabilità della media campionaria
diminuisce
significa che le fluttuazioni
delle medie calcolabili da campioni diversi
si riducono

se la popolazione di partenza è Normale

se (X_1, X_2, …, X_n) sono v.c. i.i.d.* che descrivono un campione casuale estratto da una popolazione Normale con media μ e varianza σ^2

la media campionaria, combinazione lineare delle n v.c. originali ha esattamente distribuzione

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

* i.i.d. = indipendenti e identicamente distribuite

e se non lo è ?

confidiamo nel teorema del limite centrale

in base al quale
per n sufficientemente elevato
la media campionaria si distribuisce asintoticamente come una Normale

$$\bar{X} \xrightarrow[n \to \infty]{} N\left(\mu, \frac{\sigma^2}{n}\right)$$

da studi empirici si è visto che
per n > 50 la distribuzione della popolazione non ha quasi più influenza sulla distribuzione della media campionaria

ma perché si studia la distribuzione della media campionaria?

dopo aver estratto un campione e
dopo aver calcolato il valore campionario della media

grazie alla distribuzione di probabilità della v.c. media campionaria

possiamo calcolare la probabilità di osservare proprio il valore che abbiamo trovato e verificare se è “strano”

oppure

possiamo calcolare un intervallo di valori intorno alla media osservata che “quasi certamente” contiene il vero valore μ che ci interessa

queste riflessioni valgono anche per le altre statistiche campionarie

un esempio …

da studi di mercato, è noto che il costo medio di riparazione delle Honda Odyssey dopo un incidente stradale è pari a 2.053 Euro con uno s.q.m. di 1.077 Euro

se un'officina convenzionata effettua riparazioni con costo sempre superiore a 4.000 Euro, qual è la probabilità che stia esagerando con i prezzi?

qual è l'intervallo di prezzi che ci si può aspettare di spendere?

un esempio …

estratto un campione di 60.000 famiglie italiane

e calcolato il reddito medio campionario

qual è l'intervallo di redditi che quasi certamente contiene il vero reddito medio delle famiglie italiane?

un'altra Statistica importante: la VARIANZA CAMPIONARIA

sia (X_1, X_2, …, X_n) un campione casuale e sia

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

la varianza del campione

prima di estrarre il campione S^2, funzione di v.c., è anch'essa una v.c.

e avrà quindi una sua distribuzione di probabilità, un suo valore atteso e una sua varianza

$$E(S^2) = \frac{n-1}{n}\sigma^2 \qquad Var(S^2) =$$ vedi il Cicchitelli

(n-1) o n a denominatore ? questo è il problema !

$$E(S^2) = \frac{n-1}{n}\sigma^2$$

il valore atteso di S^2 non è proprio uguale al parametro da stimare

ma se calcolo la varianza in modo leggermente diverso

$$S_c^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

ottengo uno stimatore corretto che è preferibile

$$E(S_c^2) = \sigma^2$$

se la popolazione di partenza è Normale

se (X_1, X_2, …, X_n) sono v.c. i.i.d. che descrivono un campione casuale estratto da una popolazione Normale con media μ e varianza σ^2

una trasformazione della varianza campionaria corretta ha distribuzione

$$\frac{(n-1)\cdot S_c^2}{\sigma^2} \sim \chi^2_{(n-1)}$$

un esempio …

chi produce flaconi di antigelo per motori deve garantire che ogni flacone contenga in media 1 kg di prodotto e deve contenere entro certi limiti la variabilità da flacone a flacone

estratto un campione di flaconi da una linea di produzione

qual è lo s.q.m. della popolazione infinita di flaconi in uscita dallo stabilimento di produzione?

come lo posso stimare a partire dalla varianza campionaria?

che succede della media campionaria se non conosciamo la varianza ?

se la distribuzione della popolazione è Normale (o quasi tale)

la statistica

$$t = \frac{\overline{X} - \mu}{\frac{S_c}{\sqrt{n}}} \sim t_{n-1}$$

ha distribuzione t di Student con (n-1) g.d.l.

cosa vi ricorda questa Statistica?
non vi sembra un valore standardizzato?
solo che a denominatore non poniamo il vero valore (sconosciuto) dello s.q.m., ma la sua stima campionaria

la DIFFERENZA tra DUE MEDIE

da una pop. $N(\mu_1, \sigma_1^2)$ **estraiamo n_1 unità campione**
e
da una pop. $N(\mu_2, \sigma_2^2)$ **estraiamo n_2 unità campione**

i due campioni sono indipendenti

la statistica

$$Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

si distribuisce come una N(0,1)

che succede se non conosciamo la varianza ?

da una pop. $N(\mu_1, \sigma_1^2)$ **estraiamo n_1 unità campione**
e
da una pop. $N(\mu_2, \sigma_2^2)$ **estraiamo n_2 unità campione**

la statistica

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2} \cdot \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \sim t_{(n_1 + n_2 - 2)}$$

si distribuisce come una t di Student con (n_1+n_2-2) g.d.l.

un esempio …

pesa di più una lattina di Coca-cola normale o una di Coca-light ?

estraggo due campioni di ciascun tipo e confronto le medie campionarie

sono diverse?

e lo sono così tanto da farmi pensare che, in genere, nelle lattine di coca-light ci sia meno prodotto?

un esempio …

considerati due gruppi di allergici alla parietaria

a un gruppo viene somministrato il vaccino A

all'altro gruppo il vaccino B

quale dei due vaccini funziona meglio?

come si comportano le due medie campionarie dei valori sanguigni legati alle reazioni allergiche?

il RAPPORTO tra DUE VARIANZE

da una pop. $N(\mu_1, \sigma_1^2)$ **estraiamo n_1 unità campione**
e
da una pop. $N(\mu_2, \sigma_2^2)$ **estraiamo n_2 unità campione**

i due campioni sono indipendenti

la statistica

$$F = \frac{S_{1c}^2}{S_{2c}^2} \sim F_{(n_1-1, n_2-1)}$$

si distribuisce come una F di Snedecor

stima puntuale
e
stima intervallare

ripetiamo un po' …

- **stiamo studiando una popolazione fissa**
  - **es. con media μ e varianza σ^2 costanti, incognite chiamate parametri**
  - **media campionaria $\overline{X}$ e varianza campionaria S^2 sono variabili casuali che variano al variare del campione secondo la loro distribuzione di probabilità**
  - **una volta estratto un campione si possono calcolare le statistiche campionarie per $\overline{X}$ e S^2**

il campione è affetto da errore

- **quando si estrae un campione si commette un errore poiché si misura solo un sottoinsieme di tutte le unità della popolazione**
  - **es. la stima $\overline{X}$ non sarà proprio identica al parametro sconosciuto μ**
- **si commetterà un errore, E**
  - **es. è difficile pensare che μ sia esattamente uguale al valore campionario di $\overline{X}$**
  - **possiamo, però, pensare che μ sia compreso in un intervallo intorno a $\overline{X}$**

INTERVALLO DI CONFIDENZA

- **il parametro sconosciuto sarà uguale alla statistica campionaria più o meno un certo errore**

$$\mu = \overline{X} \pm E$$

- **il calcolo del valore campionario della statistica di interesse è piuttosto semplice** $\overline{X}$
- **la bravura e la fatica degli statistici è proprio quella di calcolare con precisione il margine di errore legato al processo di campionamento** E
- **è possibile calcolare E se si conosce la distribuzione di** $\overline{X}$

quanto CONFIDANO gli statistici nelle loro stime?

che grado di fiducia si vuole avere sulla bontà della propria stima intervallare?

di quanto ci possiamo sbagliare?

di solito si sceglie un grado di fiducia del 95% e

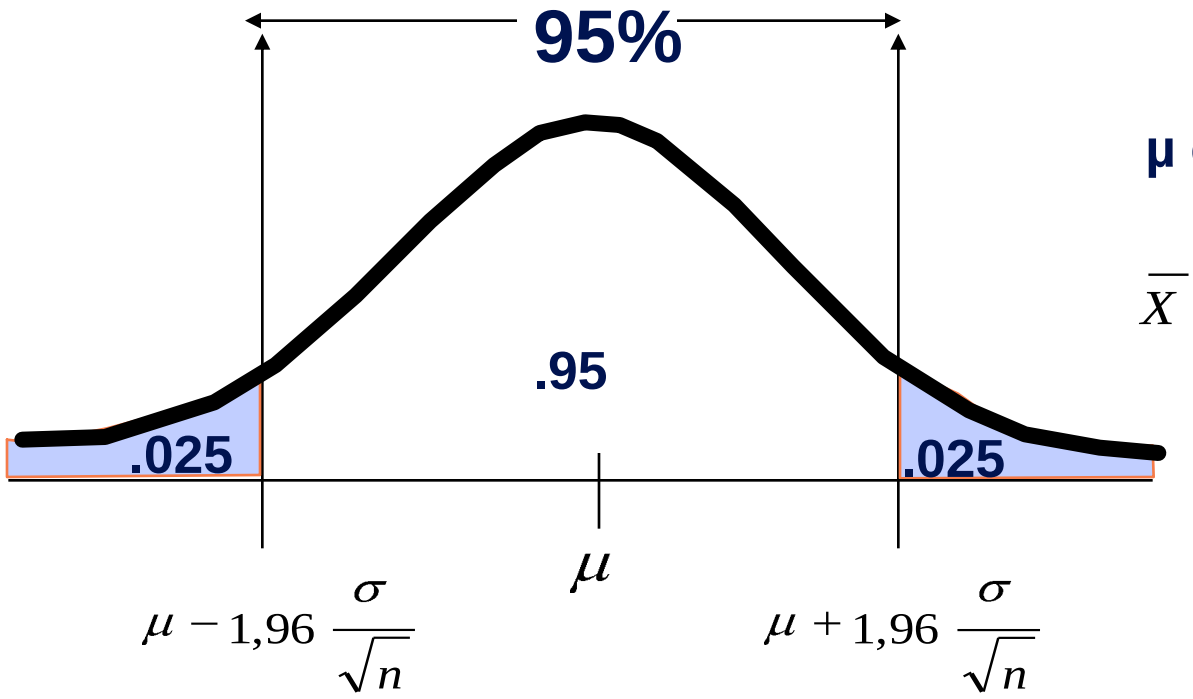
ad esempio, si sceglie il più piccolo intervallo intorno a $\overline{X}$

tale da comprendere il 95% della sua distribuzione di probabilità

in formula e graficamente

es. nell'ipotesi che $\overline{X}$ sia normale

$$P\left(\mu - 1{,}96\,\frac{\sigma}{\sqrt{n}} < \overline{X} < \mu + 1{,}96\,\frac{\sigma}{\sqrt{n}}\right) = 0{,}95$$



μ è una costante ignota

$\overline{X}$ si distribuisce intorno a μ

qual è la costante ?
qual è la variabile casuale ?

μ rimane costante, è il vero valore sconosciuto

mentre la media campionaria varia al variare dei campioni che si possono estrarre dalla popolazione

la formula

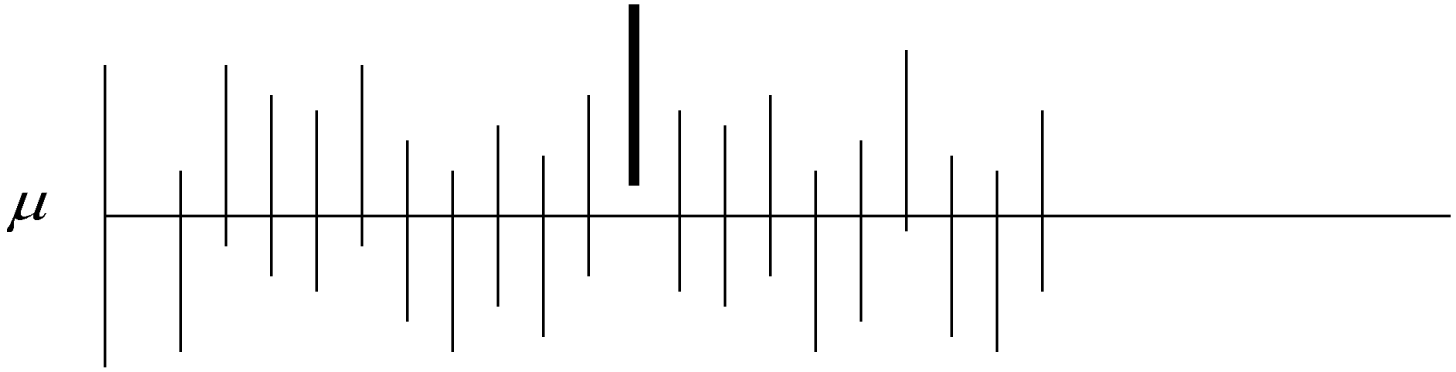
$$P\left(\mu - 1{,}96 \frac{\sigma}{\sqrt{n}} < \overline{X} < \mu + 1{,}96 \frac{\sigma}{\sqrt{n}} \right) = 0{,}95$$

esprime una relazione di probabilità per $\overline{X}$ **non per μ**

quindi è l'intervallo che varia al variare dei campioni, non μ

il sogno degli statistici:
avere a disposizione tanti campioni

- **se si potessero estrarre da una stessa popolazione normale di media μ e varianza σ^2**
- **per ciascuno si potrebbe calcolare la stima puntuale e l'intervallo di confidenza**
- **la formula precedente ci dice che il 95% degli intervalli di confidenza calcolati a partire dai nostri campioni ripetuti conterrebbe il valore vero e sconosciuto di μ**



la realtà degli statistici:
un solo campione a disposizione

quell'unico campione estratto ha il 95% di probabilità di contenere il vero valore sconosciuto del parametro μ

cioè si lavora con un metodo che ha il 95% di probabilità di successo

- **dicendo che** $\bar{x} - 1{,}96\,\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 1{,}96\,\frac{\sigma}{\sqrt{n}}$

  si ha il 95% di probabilità di dire una cosa vera

  cioè, essendo elevata la probabilità che l'intervallo contenga il valore incognito di μ, si ha fiducia che il particolare intervallo osservato per il campione contenga effettivamente il valore incognito di μ

un errore da evitare accuratamente

interpretazione ERRATA dell'intervallo di confidenza:

"c'è il 95% di probabilità
che il valore di μ cada all'interno
dell'intervallo di confidenza (t_1, t_2)"

interpretazione CORRETTA dell'intervallo di confidenza:

"c'è il 95% di probabilità
che l'intervallo di confidenza (t_1,t_2)
contenga il valore μ"

perché ciò che varia è solo l'intervallo di confidenza
μ è fisso, non si muove, anche se noi non lo conosciamo

se avete capito questa sottile differenza, siete a cavallo

come varia l'intervallo di confidenza al variare della numerosità campionaria ?

al crescere della dimensione del campione
la distribuzione di $\overline{X}$ tende a concentrarsi intorno a μ

perché $\sigma / \sqrt{n}$ diminuisce al crescere di n

quindi, l'intervallo di confidenza si restringe

e la stima diventa più precisa

***** **l'ampiezza dell'intervallo determina la precisione di stima: quanto minore è l'ampiezza, tanto maggiore è la PRECISIONE DI STIMA**

es. mi è di qualche utilità sapere che
l'altezza media di un gruppo di individui
è compresa tra 10 cm e 10 metri?
eppure questo è un intervallo di confidenza al 100%!

la realtà degli statistici:
non esistono solo gli errori campionari, purtroppo

le formule che vedremo oggi ci permettono di calcolare gli

ERRORI CAMPIONARI

cioè quelli legati al solo fatto che studiamo una "parte" per trarre informazioni sul "tutto"

ma, nel fare indagini ed esprimenti si compiono molti

ERRORI NON CAMPIONARI

e per questi la teoria della probabilità aiuta poco

***un consiglio da zia*: prima di elaborare i dati con formule complicate, controllatene la qualità, verificate che siano stati raccolti con metodi statistici seri, che siano state rispettate le regole della casualità e che le misure siano state fatte in modo corretto**

altrimenti: potreste perdere il vostro tempo e prendere decisioni errate

alcuni esempi …

se le bilance per pesare le unità che vi interessano sono mal tarate e indicano sempre qualche grammo in meno del peso effettivo, avrete sicuramente una stima distorta per difetto

se metà delle persone estratte nel vostro campione non rispondono ai quesiti che avete progettato di porre loro
avrete un bel da fare a far calcoli, ma io non mi fiderei delle vostre stime, qualsiasi sia il numero di decimali usato nei calcoli

se farete un campione di confezioni di sale prodotto da una catena per stimarne il peso medio
proprio in un giorno in cui c'è un elevato tasso di umidità
sarete indotti a pensare che il processo di produzione produca sacchetti troppo pieni e pesanti

torniamo un po' indietro …

e diamo definizioni formali ai concetti esposti:

dato un campione casuale $X_1, X_2, \ldots, X_n$ estratto da una popolazione con funzione di probabilità $f(x;\theta)$

date le statistiche $t_1(X_1, X_2, \ldots, X_n)$ e $t_2(X_1, X_2, \ldots, X_n)$ con $t_1 < t_2$ e

$$P(t_1 < \theta < t_2) = 1 - \alpha$$

l'intervallo CASUALE (t_1, t_2) è un intervallo di confidenza per θ

con grado di fiducia pari a $(1-\alpha)$

qualunque sia il parametro θ, fissato il livello α, in caso di ripetizione del campione, l'intervallo (t_1, t_2) conterrebbe il parametro nel $(1-\alpha)$% dei casi e lo escluderebbe nel α % dei casi

il GRADO DI FIDUCIA

ogni intervallo di confidenza (t_1, t_2)

è sempre associato ad un grado di fiducia $(1-\alpha)$

che è la probabilità che l'intervallo contenga il vero e sconosciuto valore del parametro da stimare

i valori scelti più di frequente sono:

$(1-\alpha) = 0{,}90$

$$(1-\alpha) = 0{,}95$$

$(1-\alpha) = 0{,}99$

è il più usato perché costituisce un buon compromesso tra precisione e affidabilità

il VALORE CRITICO

ogni intervallo di confidenza (t_1, t_2)

è sempre associato ad un valore critico $z_{\alpha/2}$

che è il valore in ascissa della distribuzione di probabilità che lascia alla sua destra un'area pari ad $\alpha/2$

ad esempio, per la media campionaria, lo si può scrivere come

$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

se la distribuzione è Normale

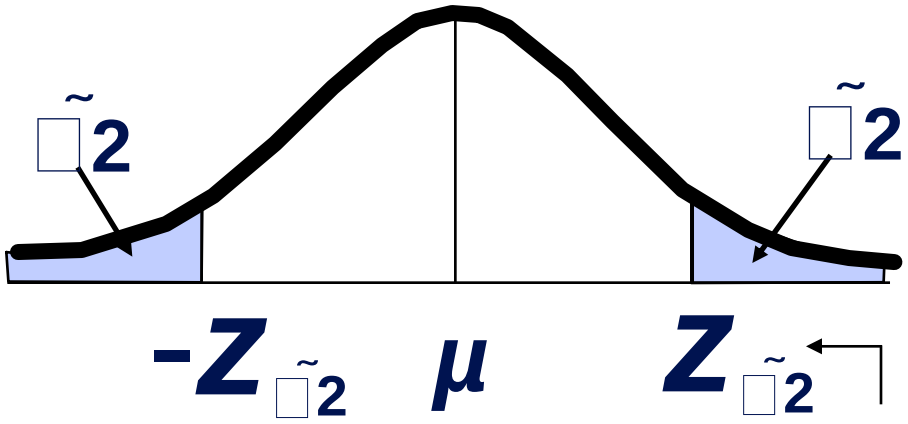


grafico per normale standardizzata

si trovano nelle tavole

$(1-\alpha) = 0{,}90 \qquad z_{\alpha/2} = 1{,}645$

valori per qualsiasi normale

$(1-\alpha) = 0{,}95 \qquad z_{\alpha/2} = 1{,}96$

$(1-\alpha) = 0{,}99 \qquad z_{\alpha/2} = 2{,}575$

il MARGINE DI ERRORE

ogni intervallo di confidenza (t_1, t_2)

è sempre associato ad un margine di errore

che è legato alla distanza della stima puntuale dagli estremi dell'intervallo di confidenza

ad esempio, per la media campionaria, è pari a

$$E = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

è la quantità che tolgo e aggiungo alla media campionaria per ottenere l'intervallo di confidenza

intervallo di confidenza per la MEDIA CAMPIONARIA, grandi campioni

n > 30 qualsiasi sia la distribuzione delle v.c. originali
o n > 50 per v.c. originali molto lontane dalla normale

COME SI COSTRUISCE ?

- **si stabilisce il grado di fiducia** $(1-\alpha)$
- **si confida nel teorema del limite centrale**
- **si trova il valore critico** $z_{\alpha/2}$
- **si calcola il margine di errore** $E = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$
- **si calcolano i due estremi dell'intervallo**

$$\left(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}; \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

e se il campione è piccolo ?
e se non conosco sigma ?

intervallo di confidenza per la media campionaria
se n <= 30 o sigma ignoto

è necessario che sia verificata un'altra condizione:

il campione deve essere stato estratto da una popolazione Normale

e allora si può ricorrere alla distribuzione *t* di Student

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}}$$

e ai suoi valori critici $t_{\alpha/2}$
che variano a seconda dei gradi di libertà, pari a (n-1)

iIntervallo di confidenza per la MEDIA CAMPIONARIA, piccoli campioni e/o sigma ignoto

n <= 30 e/o sigma ignoto
v.c. originali approssimativamente normali

COME SI COSTRUISCE ?

- **si stabilisce il grado di fiducia** $(1-\alpha)$
- **si confida in un teorema sulla *t* (che abbiamo visto)**
- **si stima sigma a partire dai dati campionari: s**
- **si trova il valore critico per (n-1) g.d.l.** $t_{\alpha/2}$
- **si calcola il margine di errore**

$$E = t_{\alpha/2}\frac{s}{\sqrt{n}}$$

- **si calcolano i due estremi dell'intervallo**

$$\left(\bar{x} - t_{\alpha/2}\frac{s}{\sqrt{n}}; \bar{x} + t_{\alpha/2}\frac{s}{\sqrt{n}}\right)$$

per fortuna …

anche se sigma è ignoto

al crescere di n la *t* di Student tende alla Normale

e, quindi,

se il campione è sufficientemente grande, anche quando si usa la statistica campionaria s per stimare sigma

per calcolare l'intervallo di confidenza si ricorre alla distribuzione Normale

cioè si usano gli $z_{\alpha/2}$ **che si imparano a memoria facilmente**

anziché i $t_{\alpha/2}$ **che sono più complicati da trovare**

un esempio …

la neve provoca infarti?
(Journal of the American Medical Association)

In concomitanza con abbondanti nevicate si è notato un incremento dei ricoveri per infarto

10 persone sono state messe a spalare la neve e si è misurato il loro numero massimo di battiti al minuto

n = 10 - media campionaria = 175 - sigma campionario = 15

è stato calcolato l'intervallo di confidenza al 95% del n. max di battiti al minuto

...

n = 10 - media campionaria = 175 - s = 15

grado di fiducia = 0,95

campione piccolo e sigma ignoto: non si usa la normale, ma la t

g.d.l. = (n-1) = 9

valore critico = $t_{\alpha/2}$ *= 2,262*

margine di errore = $E = t_{\alpha/2} \frac{s}{\sqrt{n}}$ *= 10,7296*

limite inferiore = 175 – 10,7296 *limite superiore = 175 + 10,7296*

I.C. = (164 , 186)

...

I.C. = (164 , 186)

cioè,

sulla base del campione osservato e
con un livello di fiducia del 95%
si può dire che

il valore medio del numero massimo di battiti cardiaci al minuto
di chi spala la neve è compreso tra 164 e 186

è noto che

il valore medio del numero massimo di battiti cardiaci al minuto
della popolazione in situazione di riposo è pari a 100

quindi

spalare la neve sovraccarica il cuore e aumenta il rischio di infarto

intervallo di confidenza per la DIFFERENZA TRA DUE MEDIE

si è interessati a calcolare un I.C. per il valore ignoto

$$\mu_1 - \mu_2$$

che è la differenza delle medie di due popolazioni

lo si fa estraendo due campioni indipendenti

calcolando la stima campionaria

$$\bar{x}_1 - \bar{x}_2$$

e riflettendo sulla distribuzione di probabilità dello

stimatore $$\bar{X}_1 - \bar{X}_2$$

… differenza tra due medie

sappiamo che la media campionaria si distribuisce come una Normale $\overline{X}_1 \sim N(\mu_1, \sigma_1^2/n_1)$ $\overline{X}_2 \sim N(\mu_2, \sigma_2^2/n_2)$

e che campioni indipendenti rendono le v.c.

$\overline{X}_1$ e $\overline{X}_2$ **indipendenti tra loro**

quindi la v.c. $\overline{X}_1 - \overline{X}_2 \sim N(\mu_1 - \mu_2, \sigma_1^2/n_1 + \sigma_2^2/n_2)$

e l' I.C. al 95% sarà

$$\mu_1 - \mu_2 = (\bar{x}_1 - \bar{x}_2) \pm 1{,}96 \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

e se il campione è piccolo ?
e se non conosco sigma ?

tirate un sospiro di sollievo perché

questa parte ve la risparmio

intervallo di confidenza per una PROPORZIONE

si è interessati a calcolare un I.C. per il valore ignoto

$$p$$

che è la proporzione di popolazione che possiede una certa caratteristica

es. persone che praticano sport

abbiamo bisogno di verificare alcune condizioni:

- **campione casuale**
- **condizioni per l'uso della distribuzione binomiale**
- **è possibile usare l'approssimazione normale alla binomiale** *np >= 5 e n(1-p) >= 5*

notazioni

p **proporzione ignota nella popolazione**

$\hat{p} = \frac{x}{n}$ **proporzione di successi osservata in un campione di n unità**

$1 - \hat{p} = \frac{n - x}{n}$ **proporzione di insuccessi**

proporzione, percentuale o frequenza ? **cambia poco !**

percentuale = proporzione per 100

frequenza = proporzione * N (numerosità della popolazione)

es. 0,4 proporzione di sportivi
40% percentuale di sportivi
1.200 sportivi in una popolazione di 3.000 persone

… I.C. per una proporzione

il margine di errore è

$$E = z_{\alpha/2}\sqrt{\frac{\hat{p}\cdot(1-\hat{p})}{n}}$$

e l'intervallo di confidenza

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}\cdot(1-\hat{p})}{n}}$$

un esempio …

la gente imbroglia quando risponde alle indagini sul voto ?

un'indagine ha riguardato 1.002 persone e di queste 701 hanno dichiarato di essersi recati a votare alle ultime elezioni

però la percentuale di votanti registrata ai seggi è stata pari al 61%

troviamo un intervallo di confidenza al 95% per la proporzione di persone intervistate che dichiarano di essere andate a votare

n = 1.002 *x = 701*

stima puntuale della proporzione $\hat{p} = \frac{701}{1002} = 0{,}6996$

margine di errore $E = 1{,}96\sqrt{\frac{(0{,}6996)\cdot(1-0{,}6996)}{1002}} = 0{,}0283855$

I.C. $0{,}6996 - 0{,}0283855 < p < 0{,}6996 + 0{,}0283855$

...

$$0{,}671 < p < 0{,}728 \qquad 67{,}1\% < p < 72{,}8\%$$

$$p = 0{,}7 \pm 0{,}028 \qquad p = 70\% \pm 2{,}8\%$$

si stima che il 70% degli elettori abbia dichiarato di essersi recato a votare, con un margine di errore di più o meno 2,8 punti percentuali

sulla base dei risultati campionari si può ritenere, con un grado di fiducia del 95%, che la vera percentuale di persone che rispondono di essersi recate a votare sia compresa tra 67,1% e 72,8%

è noto che la percentuale effettiva di votanti è stata del 61%

poiché il valore 61% non è compreso all'interno dell'intervallo stimato

si può concludere che …..

qualcuno non la racconta giusta

intervallo di confidenza per la VARIANZA

si è interessati a calcolare un I.C. per il valore ignoto

$$\sigma^2$$

che è la varianza di una popolazione

es. variabilità nei tempi di attesa
per ottenere una prestazione sanitaria

abbiamo bisogno di verificare alcune condizioni:

- **campione casuale**
- **le v.c. originali DEVONO essere Normali anche per campioni grandi**

utilizziamo la distribuzione chi-quadrato

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

statistica campionaria chi-quadrato
n = numerosità campionaria
(n-1) gradi di libertà
s^2 = varianza campionaria
sigma^2 = varianza nella popolazione (ignota)

l' I.C. si calcola come segue

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}$$

verifica
delle
ipotesi

qual è il nostro problema ?

- **stiamo studiando una popolazione fissa**
  - ***es. con media*** μ ***e varianza*** σ^2 ***costanti, incognite chiamate parametri***
- **definiamo un parametro** θ **di cui conosciamo la funzione di densità** $f(x;\theta)$
- **abbiamo in testa un'ipotesi sul parametro**
- **estraiamo un campione per verificare se l'ipotesi può essere considerata corretta o errata**

cos'è un'IPOTESI ?

in statistica un'ipotesi è un'affermazione su un parametro della popolazione considerata

esempi:

i medici affermano che la temperatura del corpo umano degli adulti sani è diversa da 38 gradi (ipotesi su una media)

gli automobilisti che usano il cellulare in auto hanno una probabilità di avere un incidente superiore del 13% rispetto a quelli che non telefonano finché sono alla guida (ipotesi su una proporzione)

un nuovo sistema di produzione di altimetri da aeroplano fornisce altimetri con una minore variabilità, le misure delle altezze sono più consistenti (ipotesi su una varianza)

vi ricordate ?

REGOLA DEGLI EVENTI RARI
IN STATISTICA INFERENZIALE

se, sotto certe ipotesi

la probabilità di un evento osservato è particolarmente bassa

lo statistico conclude che

probabilmente, l'ipotesi non è corretta

es. ipotesi: l'altezza media dei Watussi è 165 cm
da un campione numeroso di Watussi si stima altezza media pari a 180cm
se è vera l'ipotesi (165 cm)
la probabilità di ottenere un campione con media 180 è molto bassa
quindi, molto probabilmente l'ipotesi è sbagliata

a ben pensarci,
è una regola dettata dal buon senso

un vecchio libro di ricette dice che la torta della nonna deve cuocere in forno per 1 ora a 200 gradi

voi provate la prima volta e la torta si brucia
provate la seconda e la torta si brucia
…
provate la n-ma volta e la torta si brucia

quante torte fareste bruciare prima di farvi venire il dubbio che 60 minuti di cottura siamo troppi e che il libro contenga un refuso?

ipotesi: tempo di cottura 60 minuti
campione: n torte cotte e bruciate
probabilità di bruciare n torte se l'ipotesi è vera: bassissima
o è accaduto un insieme di eventi altamente improbabili
(sfi…? legge di Murphy?)
o l'ipotesi di partenza è falsa, va rifiutata

come usiamo questa regola del buon senso per verificare statisticamente le ipotesi?

per verificare delle ipotesi statistiche:

si analizza un campione

per cercare di individuare i risultati che posso verificarsi con buona probabilità

distinguendoli dai risultati che sono altamente improbabili

e se i fatti non sono così lampanti ?

es. cos'avreste pensato del tempo di cottura
se aveste bruciato due torte su tre?
e se ne aveste bruciate una su tre?

qual è un risultato campionario così SIGNIFICATIVO da far rifiutare l'ipotesi considerata?

le procedure statistiche che vedremo oggi ci aiuteranno a capire
quali sono i risultati campionari SIGNIFICATIVI

a tal punto da farci rifiutare un'ipotesi

partiamo da un esempio …

alcuni ricercatori americani affermano che la temperatura corporea umana media per adulti sani è pari a 98,6 F

altri ricercatori hanno studiato un campione di 106 adulti sani trovando che la loro temperatura media è 98,2 F con s.q.m 0,62F

i risultati di questo campione sono così significativi da farci rifiutare l'ipotesi che la temperatura media sia pari a 98,6 F?

se la media della popolazione fosse proprio 98,6 F, la probabilità di ottenere una media campionaria pari a 98,2 F è 0,0002

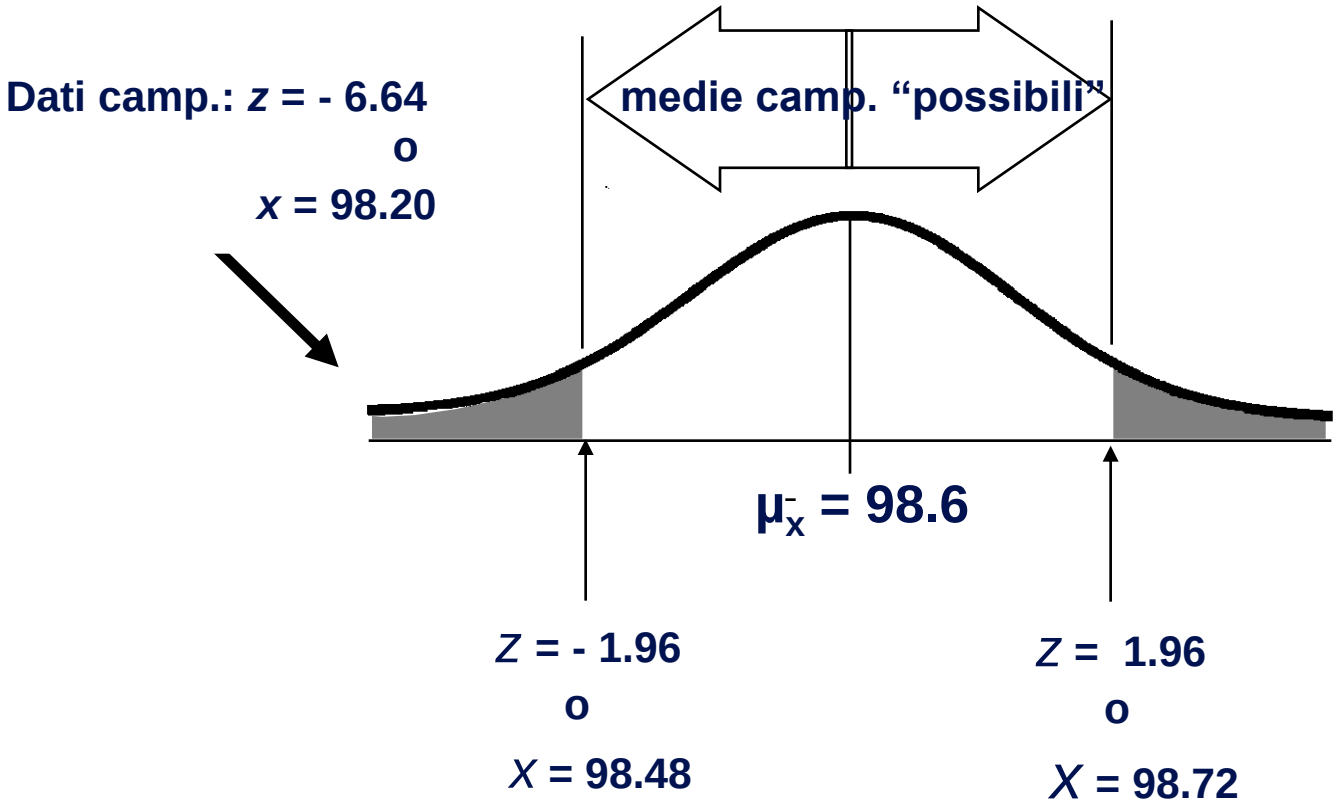
(non sapete come calcolare questa prob., ma fidatevi)

0,0002 è una probabilità molto bassa, quindi …

siamo portati a ritenere che l'ipotesi temperatura media = 98,6 F sia da rigettare

l'esempio continua …
grazie al teorema del limite centrale

distribuzione attesa della media campionaria nell'ipotesi □ = 98.6



l'esempio continua …
per definire gli elementi chiave

elementi chiave e filo del ragionamento:

- ***è parere comune in America che la temperatura corporea media sia 98,6 F***
- ***il campione ha prodotto media campionaria 98,2 F***

- ***avendo considerato***
  - ***la distribuzione della media campionaria***
  - ***la dimensione del campione***
  - ***la distanza tra 98,6 e 98,2***

- ***si è trovato che la media campionaria 98,2 F è poco probabile (meno del 5% di probabilità) se l'ipotesi è vera***

- ***ci sono due spiegazioni:***
  - ***o è accaduto un evento molto raro (campione sfortunato)***
  - ***o la media non è 98,6 F***

e ora qualche definizione:
IPOTESI NULLA

indicata con H_0

- **è un'affermazione sul valore di un parametro della popolazione**

- **deve contenere una condizione di uguaglianza/disuguaglianza secondo tre possibili forme: ad esempio, per la media**
  - **H_0: μ = valore**
  - **H_0: μ <= valore**
  - **H_0: μ >= valore**

es. H_0: μ = 98,6F

- **si verifica la plausibilità dell'ipotesi nulla, cioè**
  - **la si rifiuta**
  - **o non la si rifiuta**

IPOTESI ALTERNATIVA

indicata con H_1

- **è un'affermazione sul valore di un parametro della popolazione, che deve essere vera se H_0 è falsa**

- **deve contenere una condizione di uguaglianza/disuguaglianza secondo tre possibili forme: ad esempio, per la media**
  - **H_1: μ not= valore**
  - **H_1: μ > valore**
  - **H_1: μ < valore**

es. H_0: μ not = 98,6F

- **è praticamente l'opposto dell'ipotesi nulla**

STATISTICA TEST

- **è un valore calcolato a partire dal campione**

  es. media campionaria

- **usato per prendere la decisione (rigettare o meno H_0)**
- **viene tradotto in forma standardizzata**

  es. valore z $z = \dfrac{\bar{x} - \mu}{\sigma / \sqrt{n}}$

- **e confrontato con valori standard**

  es. tavole della normale

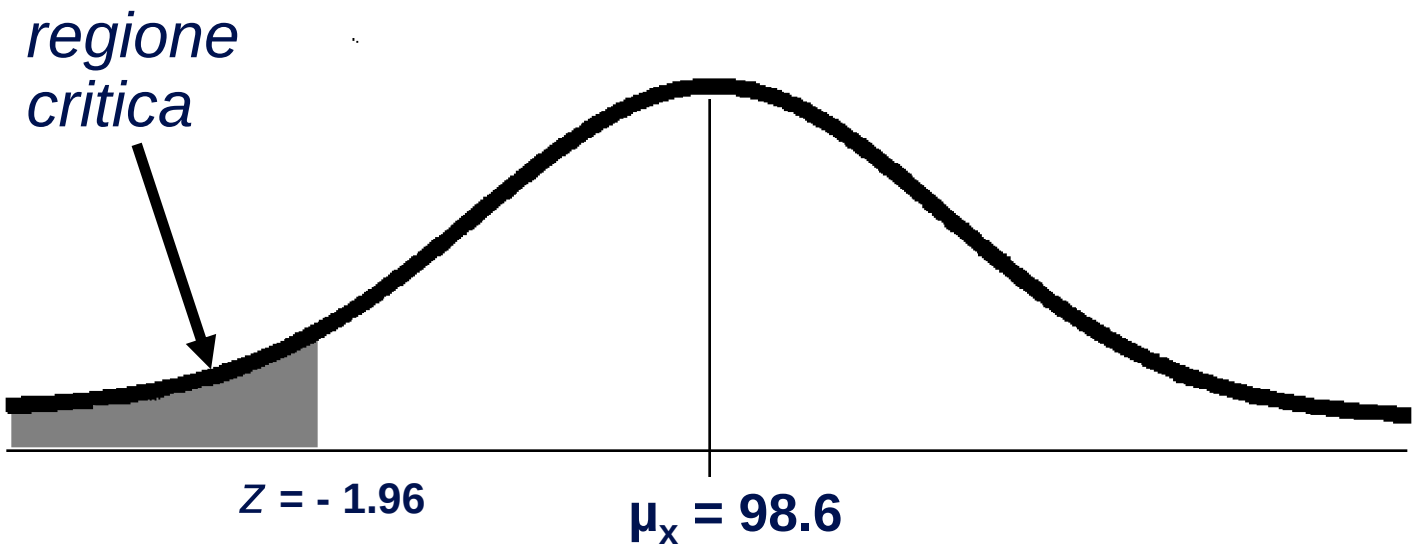
- **l'esito del confronto porta a rifiutare o meno H_0**

  es. H_0: μ=98,6

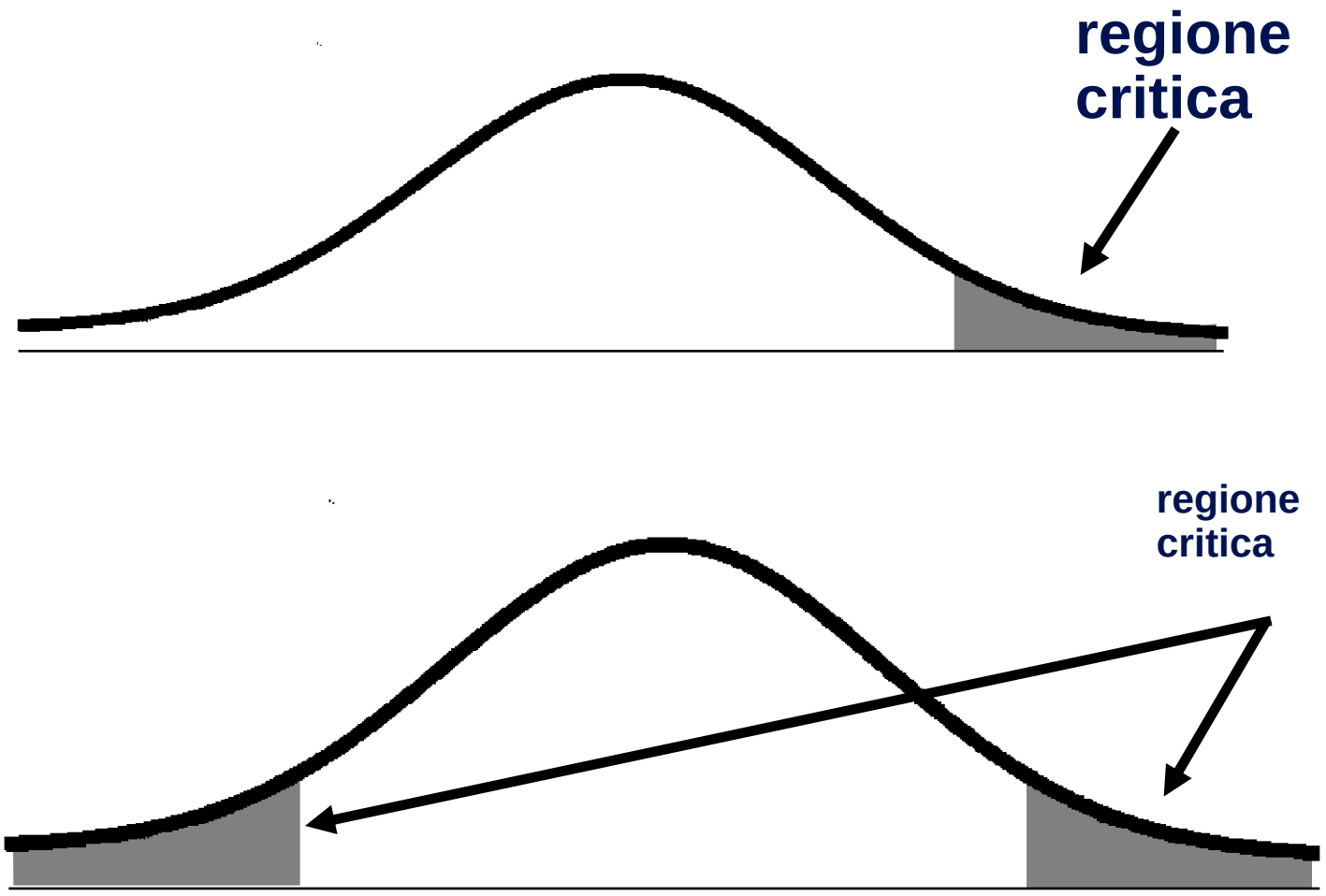
$$z = \frac{98,2 - 98,6}{0,62 / \sqrt{106}} = -6,64$$

REGIONE CRITICA

- **è l'insieme di valori della statistica test che conduce a rifiutare l'ipotesi nulla**



… regione critica … ha varie forme

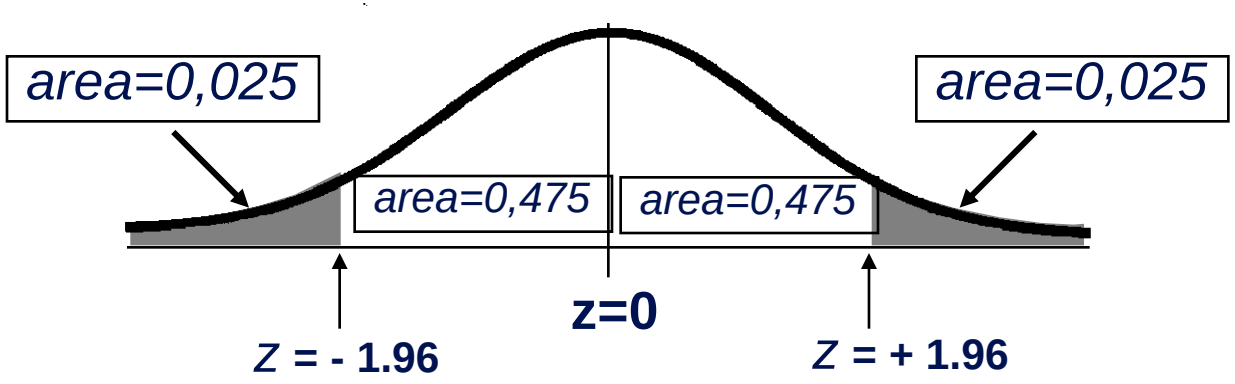


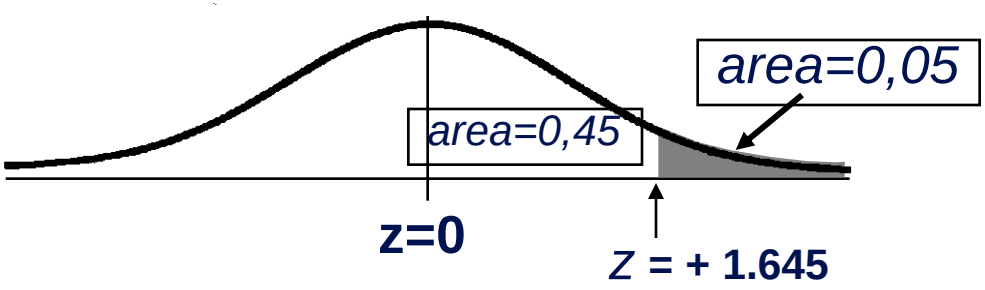
LIVELLO DI SIGNIFICATIVITA’

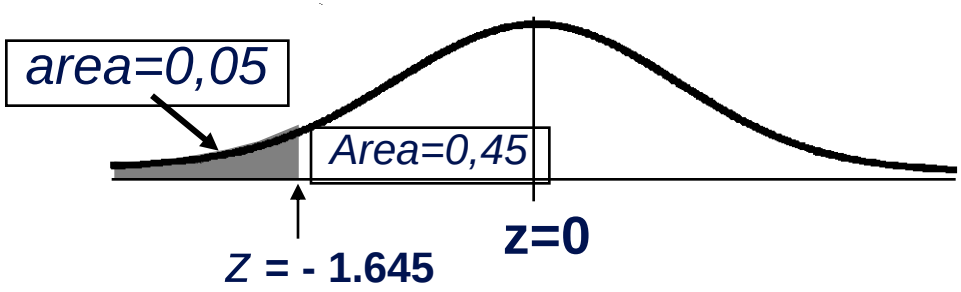
indicato con alpha: α

- **è la probabilità che la statistica test cada nella regione di rifiuto quando l’ipotesi nulla è vera**
- **se la statistica test cade nella zona di rifiuto noi saremo indotti a rifiutare l’ipotesi nulla anche se questa è vera**
- **quindi alpha è la probabilità di commettere un errore cioè di rifiutare l’ipotesi nulla quando è vera**
- **è lo stesso alpha degli intervalli di confidenza**
- **i valori più frequenti sono 0,05 0,01 (e 0,001)**

esempi classici …

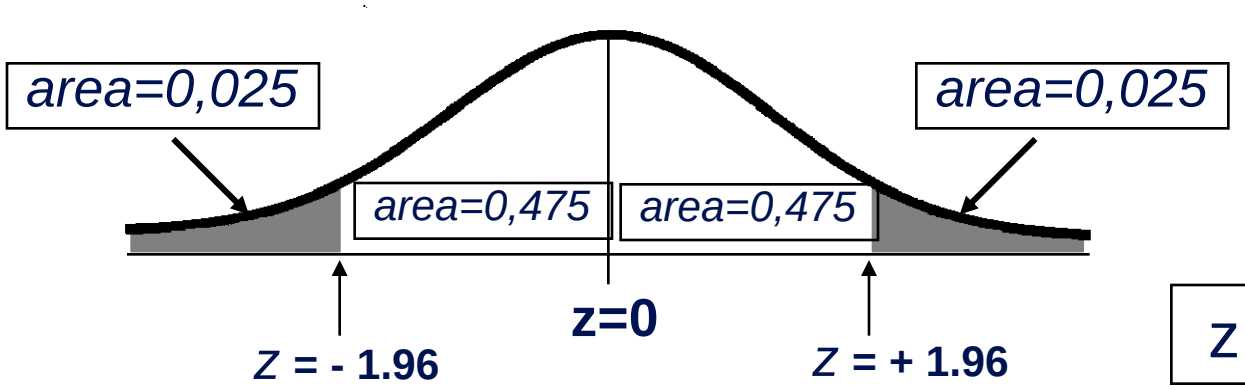






VALORE CRITICO

- **è il valore sull'asse delle ascisse che individua la regione critica**
- **dipende dalla forma dell'ipotesi nulla**
- **dipende dalla forma della variabile casuale test**
- **dipende dal livello di significatività alpha**

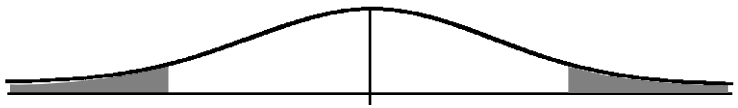


z = valori critici

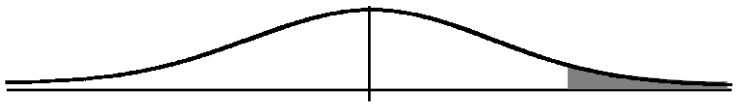
test a DUE CODE, a UNA CODA a DESTRA, a UNA CODA a SINISTRA

- le CODE sono le regioni critiche agli estremi della distribuzione di probabilità della statistica test
- individuano le regioni di valori in conflitto con H_0

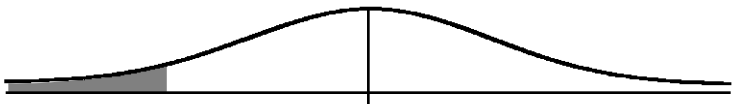
- TEST A DUE CODE la regione critica ha due code



- TEST A UNA CODA DESTRA



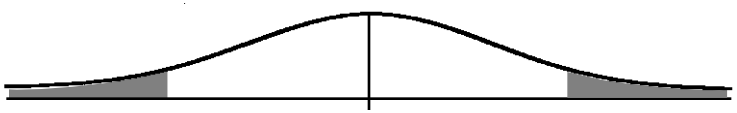
- TEST A UNA CODA A SINISTRA



le code dipendono dalla forma dell'ipotesi nulla

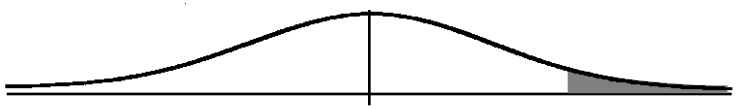
ad esempio, per una verifica d'ipotesi su una media

- **TEST A DUE CODE**



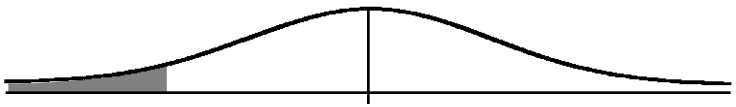
H_0: μ = valore

- **TEST A UNA CODA DESTRA**



H_0: μ <= valore

- **TEST A UNA CODA A SINISTRA**

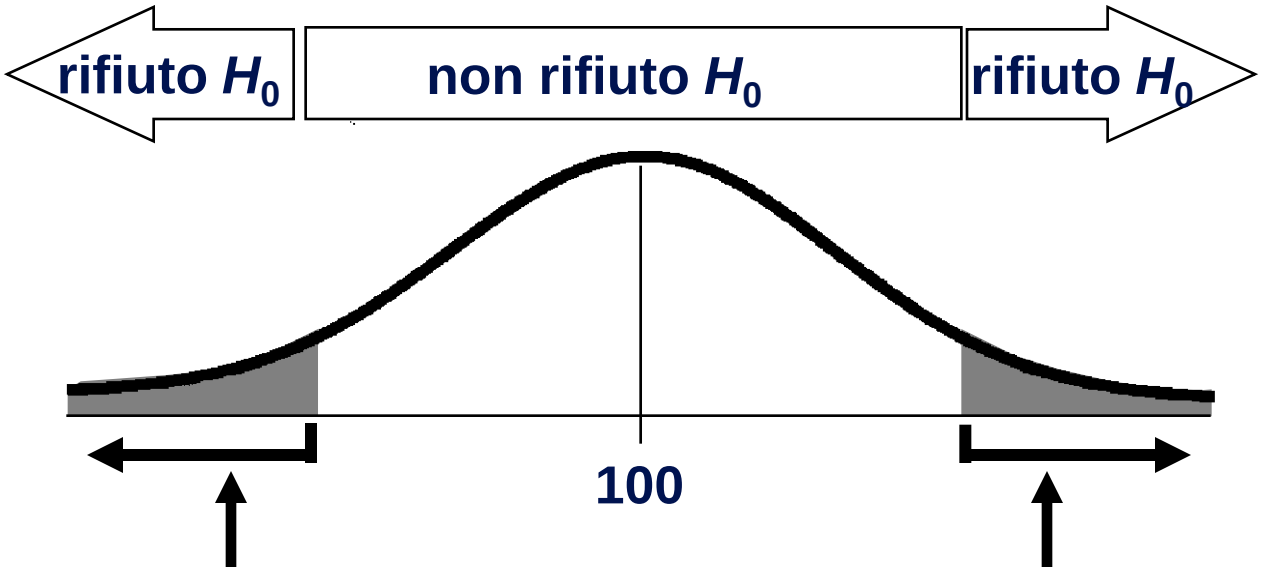


H_0: μ >= valore

TEST A DUE CODE esempio

- $H0$: $\mu = 100$
- $H1$: $\mu \neq 100$

significa minore di o maggiore di

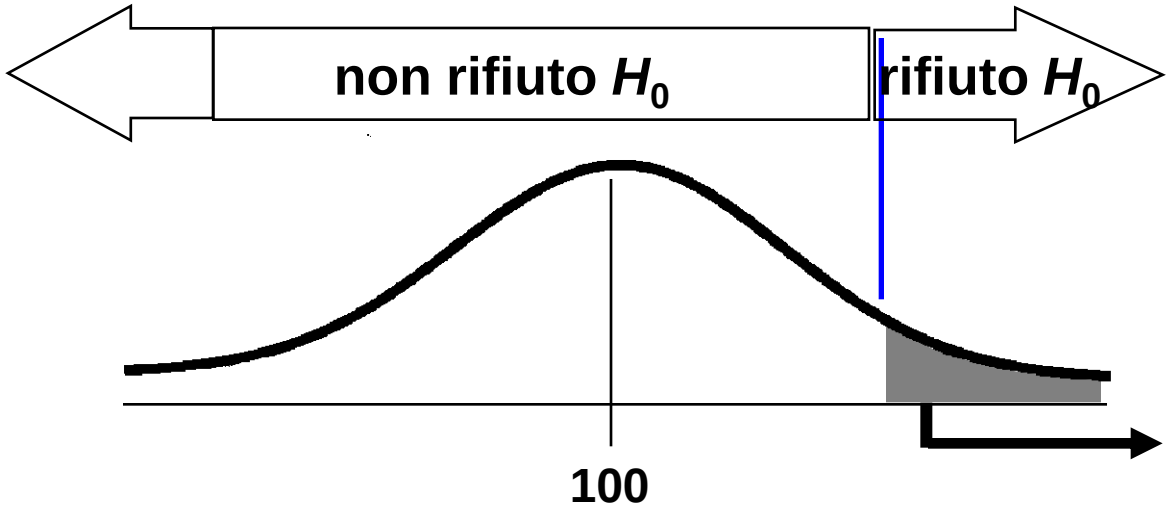


valori che differiscono significativamente da 100

TEST A UNA CODA A DESTRA esempio

- ***H*0: μ ≤ 100**
- ***H*1: μ > 100**

regione a destra

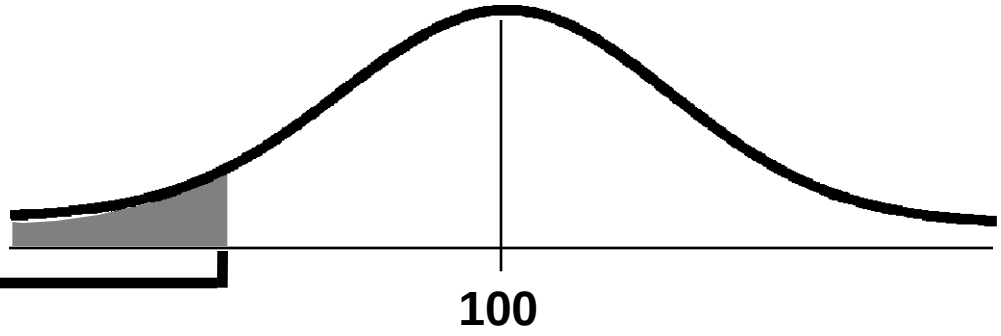


valori che differiscono significativamente da 100

TEST A UNA CODA A SINISTRA esempio

- $H0$: $\mu \geq 100$
- $H1$: $\mu < 100$

regione a sinistra



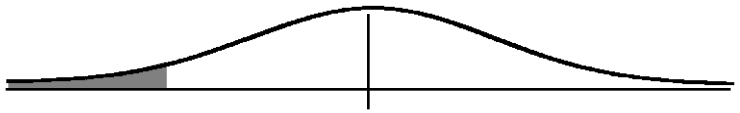
valori che differiscono significativamente da 100

… attenzione …

la coda va in direzione opposta rispetto alla disuguaglianza espressa nell'ipotesi nulla

ad esempio, per una verifica d'ipotesi su una media

- **TEST A UNA CODA A SINISTRA**



H_0: μ >= valore

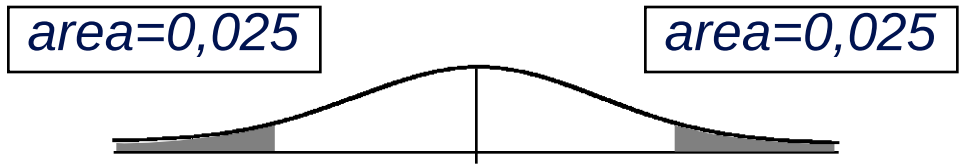
- **vogliamo verificare se la media è maggiore o uguale ad un certo valore**
- **un campione con media "molto" minore (a sx) di quello definito nell'ipotesi nulla ci porterà a rifiutarla**
  cioè la regione di rifiuto è nella parte sx della distribuzione

… attenzione …

il livello di significatività alpha nei test a due code si divide equamente in due

ad esempio, per una verifica d'ipotesi su una media

- **TEST A DUE CODE con alpha = 0,05**



H_0: μ = valore

- **vogliamo verificare se la media è diversa da un certo valore**
- **un campione con media "molto" minore o maggiore di quello definito nell'ipotesi nulla ci porterà a rifiutarla**
  cioè la zona di rifiuto è un po' a dx un po' a sx della distribuzione

… ma come si fa a fare un test d'ipotesi ?

a partire dai dati a disposizione si definiscono tutti i valori finora descritti:

	Esempio:
• ipotesi nulla	H0: μ = 12
• ipotesi alternativa	H1: μ not = 12
• statistica test	media campionaria
• livello di significatività	alpha = 0,05
• regione di rifiuto	
• valori critici	

si calcola il valore campionario della statistica test

se questo cade nella regione di rifiuto
si rigetta l'ipotesi nulla

altrimenti la si accetta *(o non la si rigetta ?)*

e siamo proprio sicuri della decisione presa ?

NO !

però ci consoliamo calcolando la probabilità di prendere una decisione sbagliata!

a causa della variabilità campionaria potremmo essere sfortunati e imbatterci in un campione che ci fa prendere una decisione sbagliata

ad esempio, ci fa rifiutare un'ipotesi nulla che è vera

o ci fa accettare un'ipotesi alternativa che è falsa

ERRORE DI PRIMA SPECIE α

si commette quando si rifiuta un'ipotesi nulla vera

cioè quando il valore campionario della statistica test cade nella regione di rifiuto per puro effetto della variabilità campionaria
ma, in realtà, l'ipotesi nulla è vera

alpha, livello di significatività del test, esprime la probabilità di commettere questo errore

se, ad esempio alpha = 0.05, stiamo adottando un sistema che ha il 95% di probabilità di condurre a decisioni corrette nel caso l'ipotesi nulla sia vera
cioè il 5% dei campioni ci condurrà, per pura sfortuna, a rifiutare l'ipotesi nulla vera (DECISIONE ERRATA)
mentre il 95% dei campioni ci condurrà a non rifiutare l'ipotesi nulla vera (DECISIONE CORRETTA)

ERRORE DI SECONDA SPECIE β

si commette quando non si rifiuta un'ipotesi nulla falsa

cioè quando il valore campionario della statistica test non cade nella regione di rifiuto
ma, in realtà, l'ipotesi nulla non è vera

beta esprime la probabilità di commettere questo errore

I DUE TIPI DI ERRORE

		situazione vera: l'ipotesi nulla è vera	situazione vera: l'ipotesi nulla è falsa
decisione	decidiamo di rifiutare l'ipotesi nulla	errore di I specie (rifiutare H_0 vera) alpha	decisione corretta
	non rifiutiamo l'ipotesi nulla	decisione corretta	errore di II specie (rifiutare H_0 falsa) beta

un esempio …

in un processo una giuria deve decidere fra

l'ipotesi nulla H_0 che l'accusato sia innocente
e
l'ipotesi alternativa H_1 che l'accusato sia colpevole

si commette un errore di prima specie
se un innocente viene condannato

si commette un errore di seconda specie
se un colpevole è lasciato libero

la POTENZA DI UN TEST

$1 - \beta$

è la probabilità di rifiutare un'ipotesi nulla falsa
è una misura della bontà del test

è una funzione
calcolabile a partire da un livello prefissato di alpha
e da una particolare forma dell'ipotesi alternativa

la funzione varia al variare dei possibili valori compatibili
con l'ipotesi alternativa

es. H_1: μ > 100 **si può calcolare la potenza**
per tutti i valori
100, 101, 102, …

come controllare i due tipi di errore ?

- **alpha e beta sono in competizione**
- **se diminuisce alpha, cresce beta**
- **se diminuisce beta, cresce alpha**
- **l'unico modo per farli diminuire entrambi è aumentare la numerosità campionaria**

TEST per una MEDIA (grandi campioni)

IPOTESI IN CAMPO:

- **il campione è casuale**
- **il campione è abbastanza grande (n > 30)**
  - **possiamo confidare nel teorema del limite centrale**
  - **e usare la distribuzione normale**
- **se sigma è ignoto, possiamo usare la sua stima campionaria S**

… TEST per una MEDIA (grandi campioni)

OBIETTIVO:

- **identificare un risultato campionario che è SIGNIFICATIVAMENTE diverso dall'ipotesi nulla**

STATISTICA TEST: $$Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}}$$

VALORE CAMPIONARIO TEST: $$z = \frac{\overline{x} - \mu}{\sigma / \sqrt{n}}$$

… TEST per una MEDIA (grandi campioni)

CONFRONTO TRA:

- **valore della statistica campionaria e un valore di riferimento tratto dalle tavole della Normale**

RIFIUTO DELL'IPOTESI NULLA

- **se il valore campionario della statistica test è compreso nella regione di rifiuto**

un esempio …

dato un campione di 106 temperature di adulti sani con media 98,2 F e s.q.m. 0,62 F

verifichiamo se la temperatura media è pari a 98,6 F

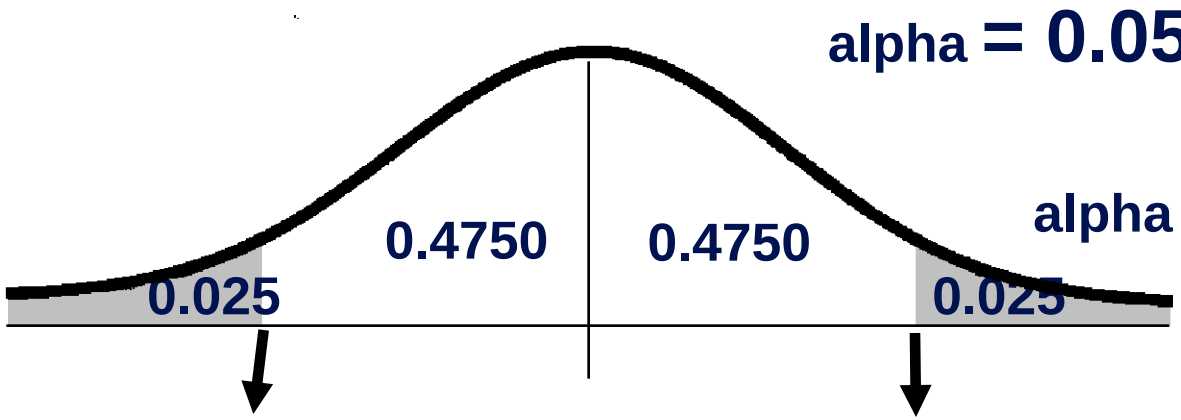
- *μ = 98,6* H_0*: μ = 98,6*
  H_1*: μ not = 98,6*

- *selezioniamo il livello di significatività 0,05*

- *calcoliamo la statistica test* $z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{98,2 - 98,6}{0,62 / \sqrt{106}} = -6,64$

… un esempio …

- *determiniamo le regioni critiche e i valori critici*



… un esempio

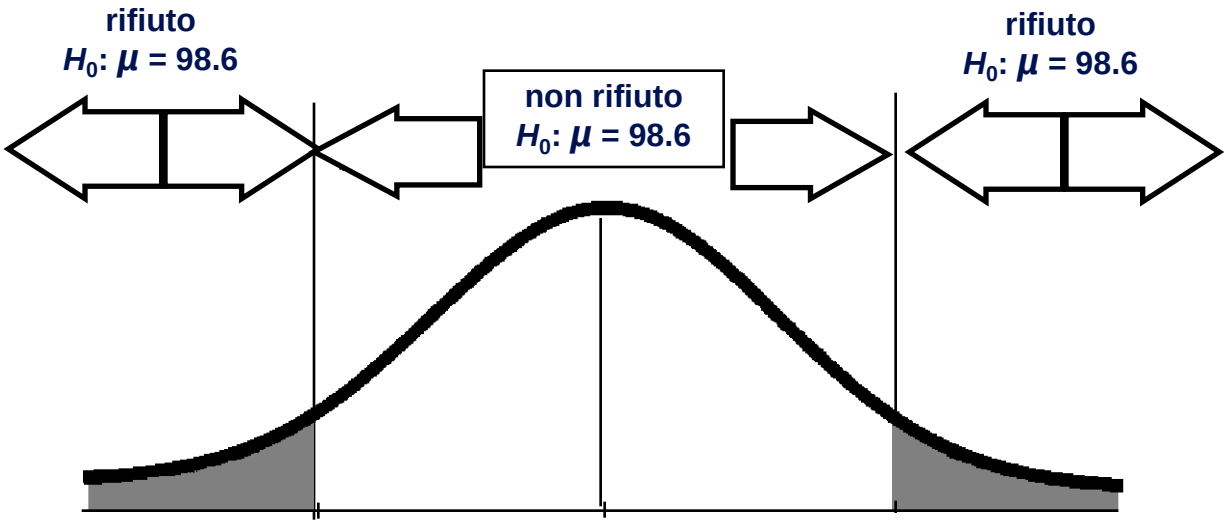
- *confrontiamo il valore campionario con i valori critici*
- *rifiutiamo H_0 se il valore campionario cade nella regione di rifiuto*

due modi di esprimere I dati campionari:

X = 98.2

o

Z = - 6.64



RIFIUTIAMO H_0

in modo analogo si possono costruire

i test per

- **media campionaria e sigma ignoto (t di Student)**
- **differenza tra due medie campionarie**
- **proporzione (binomiale approssimata con la normale)**
- **varianza (chi-quadrato)**

collegamento con l'INTERVALLO DI CONFIDENZA

un intervallo di confidenza per la stima di un parametro

contiene i valori che sono più probabili per quel parametro

quindi,

dovremo rifiutare l'ipotesi che il parametro abbia un valore esterno all'intervallo di confidenza

cioè,

UN INTERVALLO DI CONFIDENZA PUO' ESSERE CONSIDERATO ESATTAMENTE COME L'INSIEME DELLE IPOTESI ACCETTABILI

il legame è stretto !

esempio:

la temperatura media è pari a 98,6 F $H_0: \mu = 98,6$
campione: n = 106 media camp. = 98,2 s.q.m. = 0,62

I.C. al 95%: $98,08 < \mu < 98,32$

98,6 non è compreso nell'intervallo

quindi, con un livello di significatività del 5%

possiamo rifiutare l'ipotesi nulla

stime di massima verosimiglianza & co.

ripetiamo: la STIMA

- **consiste nell'attribuire il valore "più appropriato" ad un parametro sconosciuto della popolazione**

  es. si vuole conoscere il peso medio degli studenti di ingegneria, attraverso un campione degli iscritti all'A.a. 2001-2002

- **la si calcola**

  - **attraverso i risultati campionari**
  - **nel rispetto di certi criteri di ottimalità**
  - **con prefissata probabilità di errore**

ad esempio: la STIMA PUNTUALE

il parametro (caratteristica) sconosciuto della popolazione viene rappresentato con il valore puntuale espresso dalla statistica campionaria

$$t(x_1, x_2, \ldots, x_n)$$

es. dato un campione di 100 studenti iscritti nel 2001
stimo che
l'altezza media degli studenti di ingegneria sia 174,5 cm
(pari all'altezza media dei 100 studenti campione)

vi ricordate un po' di linguaggio settoriale ?

STIMATORE

è una formula o un procedimento di calcolo da applicare ai dati campionari per stimare un parametro di popolazione

STIMA

uno specifico valore (o intervallo di valori) usato per stimare / approssimare un parametro sconosciuto di popolazione

es. come viene spontaneo pensare,
la media campionaria è la migliore stima puntuale
per la media della popolazione

come deve essere uno stimatore per fornire delle buone stime ?

CRITERI DI OTTIMALITA'

- **CORRETTEZZA**
  il valore medio della v.c. Statistica
  è uguale al parametro da stimare

$$E(\overline{X}) = \mu$$

- **CONSISTENZA**
  all'aumentare della dimensione campionaria
  il valore della statistica
  tende a quello del parametro

$$\lim_{n \to +\infty} \Pr(|\ \overline{X} - \mu\ | < \varepsilon) = 1$$

- **EFFICIENZA**
  a parità di dimensione campionaria
  tra diversi stimatori
  conviene scegliere quello con varianza minore

$$Var\ (S_1) < Var\ (S_2)$$

proprietà: CORRETTEZZA

Def: Uno stimatore T = t (X1, X2, …, Xn) di θ si dice CORRETTO se e solo se

$$E(T) = \theta \qquad \forall\, \theta \in \Theta$$

dove Θ è lo spazio parametrico

uno stimatore non corretto è detto DISTORTO e la differenza

$$B(T) = E(T) - \theta$$

è detto ERRORE SISTEMATICO o DISTORSIONE

es. se $E(T) - \theta < 0$ in media le stime saranno errate per difetto
se $E(T) - \theta > 0$ “ “ “ per eccesso

Proprietà: CONSISTENZA

Def: uno stimatore T = t (X1, X2, …, Xn) di θ si dice CONSISTENTE quando

$$\lim_{n \to +\infty} \Pr(|\ T - \theta\ | \leq \varepsilon) = 1$$

Def: uno stimatore T = t (X1, X2, …, Xn) di θ si dice CONSISTENTE IN MEDIA QUADRATICA quando

$$\lim_{n \to +\infty} E\ (T - \theta)^2 = 0$$

proprietà: **uno stimatore consistente in media quadratica è consistente e asintoticamente corretto**

proprietà: EFFICIENZA

Def: sia T = t (X1, X2, …, Xn) uno stimatore di θ, si chiama ERRORE QUADRATICO MEDIO di T

$$MSE\ (T) = E\left[T - \theta\right]^2$$

tra più stimatori dello stesso parametro θ è da preferire, a parità di altre condizioni, quello che ha errore quadratico medio minimo

$$se \quad MSE\ (T') < MSE\ (T) \quad \forall\, \theta$$

$$T' \quad è \quad preferibile \quad a \quad T$$

MSE Mean Square Error

SUFFICIENZA

Def: T = t (X1, X2, …, Xn), stimatore di θ, si definisce SUFFICIENTE se fornisce tutte le informazioni sul parametro oggetto di stima, cioè

se la distribuzione di probabilità condizionata di (X1, X2, …, Xn) dato T = t non dipende da θ, per ogni t

cioè: **T è una statistica, compie una sintesi dei dati**
da n valori xi si arriva ad un unico valore t

nel processo di sintesi è bene non perdere informazioni

se T è sufficiente, il suo contenuto informativo sul parametro θ è lo stesso dell'insieme dei singoli valori campionari

come deve essere lo stimatore ideale?

corretto, consistente, efficiente, sufficiente, …

è come trovare quest'araba fenice?

in precedenza noi abbiamo usato per lo più il

METODO DEGLI STIMATORI NATURALI

cerco uno stimatore per la media di una popolazione?
Uso la media campionaria!

cerco uno stimatore per la proporzione di una popolazione?
Uso la proporzione campionaria!

ma esistono altri metodi più sofisticati …

metodo della MASSIMA VEROSIMIGLIANZA

ci aiuta a individuare quali sono le statistiche adatte a essere dei buoni stimatori

si basa sulla FUNZIONE DI VEROSIMIGLIANZA

$L(\theta; x_1, x_2, \ldots, x_n)$ **detta anche** $L(\theta)$

e punta a cercare lo stimatore che genera la stima più plausibile, cioè quello che offre la migliore spiegazione dei dati osservati

è il metodo più usato

FUNZIONE DI VEROSIMIGLIANZA e STIMA DI MASSIMA VEROSIMIGLIANZA (MLE)

L(θ) è la funzione di probabilità congiunta di un campione di numerosità n è funzione di θ

se (x1, x2, …, xn) è un campione casuale, la stima di max verosimiglianza di θ è

il valore $\hat{\theta} = \hat{\theta}(x_1, x_2, \ldots, x_n)$

che rende massima la funzione di verosimiglianza

MLE = Maximum Likelihood Estimate

PROPRIETA' delle stime di max verosimiglianza

sotto condizioni non molto restrittive, le MLE sono

EFFICIENTI

CONSISTENTI

ASINTOTICAMENTE NORMALI

con media = θ

varianza = $\dfrac{\sigma^2(\theta)}{n}$

un esempio …

campionamento da una popolazione bernoulliana

> ***abbiamo lanciato 10 volte una moneta truccata***
> ***abbiamo ottenuto 4 teste***

e vogliamo stimare la proporzione di teste p

avete notato il cambio di prospettiva rispetto ad alcune riflessioni precedenti?

quando studiavamo le v.c. immaginavamo di non aver ancora estratto il campione e volevamo fare delle previsioni sui campioni che ci potevamo attendere

ora, invece, abbiamo già estratto il campione e vogliamo cercare il modo migliore di stimare un parametro sconosciuto

… esempio …

abbiamo ottenuto 4 teste con 10 lanci

ci chiediamo:

se p fosse uguale a 0,1 qual è la probabilità di ottenere 4 teste su 10 lanci?

$$\binom{n}{x} p^{x} (1-p)^{n-x} = \binom{10}{4} \cdot 0{,}1^{4} \cdot (0{,}9)^{6} = 0{,}011$$

e se p fosse 0,2, con che probabilità avrei ottenuto il campione che ho?

e se p fosse 0,3 o 0,4 ?

possiamo farci questa domanda per tutti i possibili valori di p

cioè possiamo definire una funzione di probabilità al variare di p

… esempio …

calcoliamo i valori della probabilità di osservare proprio il campione che abbiamo sotto il naso al variare dei possibili valori di p

p	$L(p) = \binom{10}{4} \cdot p^4 \cdot (1-p)^6$
0	0
0,1	0,011
0,2	0,088
0,3	0,200
0,4	0,251
0,5	0,205
0,6	0,111
0,7	0,37
0,8	0,006
1,0	0

è poco "verosimile" che p = 0,1 produca il campione che abbiamo osservato

MAX della funzione di verosimiglianza

p, in realtà, non è una variabile casuale perché è fissa per la nostra popolazione

ma, poiché non ne conosciamo il valore preciso, possiamo ipotizzare che assuma vari valori (0,1 0,2 …)

… esempio …

abbiamo considerato una funzione di un'unica variabile (p) perché il campione per noi è completamente noto dopo averlo estratto (sono fissati i valori 10 e 4)

il nostro obiettivo è: ***STIMARE p***

e per raggiungerlo: ***scegliamo il valore di p che con maggiore probabilità avrebbe prodotto proprio il campione osservato***

in termini matematici: ***abbiamo calcolato il massimo della funzione di verosimiglianza***

… esempio …

possiamo generalizzare per qualsiasi valore di n e x:

avuti x successi in n prove

la funzione di probabilità congiunta
o funzione di verosimiglianza
se le prove sono indipendenti è $L(p) = \binom{n}{x} p^x (1-p)^{n-x}$

grazie al calcolo infinitesimale possiamo cercare il massimo di questa funzione e lo otteniamo proprio per il valore $p = \frac{x}{n}$

cioè, la proporzione campionaria è lo stimatore di massima verosimiglianza della proporzione di una popolazione

torniamo alle definizioni formali

Def. sia (x1, x2, …, xn) un campione di n osservazioni e sia

$$\hat{\theta} = \hat{\theta}(x_1, x_2, \ldots, x_n) \in \Theta$$

una funzione dei dati del campione tale che

$$L(\hat{\theta}; x_1, x_2, \ldots, x_n) \geq L(\theta; x_1, x_2, \ldots, x_n), \forall \theta \in \Theta$$

allora $\hat{\theta}$ **è una stima di massima verosimiglianza di** θ

****

la ricerca di uno stimatore, quindi, si trasforma nella ricerca del massimo di una funzione

come si cerca un massimo di una funzione?

(in modo veloce, visto che gli statistici sono un po' pigri)

anziché cercare il massimo di $L(\hat{\theta})$

spesso è conveniente cercare il massimo della

LOG-VEROSIMIGLIANZA $\log L(\hat{\theta})$

poiché
se la funzione di verosimiglianza è continua e possiede derivate prima e seconda, per trovare lo stimatore MLE si può risolvere l'EQUAZIONE DI VEROSIMIGLIANZA

$$\frac{d}{d\theta} \log L(\hat{\theta}) = 0$$

e verificare se

$$\frac{d^2}{d\theta^2} \log L(\hat{\theta}) < 0$$

un esempio …

supponiamo di aver estratto un campione di n osservazioni da una popolazione Normale

vogliamo determinare la stima di massima verosimiglianza per la media della popolazione μ

se le n osservazioni sono indipendenti la funzione congiunta di densità di probabilità è

$$p(x_1, x_2, \ldots, x_n) = \frac{1}{(\sqrt{2\pi})^n} e^{-\sum (x_i - \mu)^2 / 2\sigma^2}$$

risolvendo la log-verosimiglianza:

$$\frac{d}{d\mu} \log L(\mu) = \frac{1}{\sigma^2} \sum (x_i - \mu) = 0$$

$$\frac{d^2}{d\mu^2} \log L(\mu) \Big|_{\mu=\hat{\mu}} = -\frac{n}{\sigma^2} < 0$$

si ottiene proprio la media campionaria

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{x}$$

un esempio … per uno stimatore generico

supponiamo di aver estratto un campione di n osservazioni da una popolazione con funzione di densità di probabilità $p(x;\theta)$

vogliamo determinare la stima di massima verosimiglianza per un qualunque parametro θ

le Xi sono indipendenti e identicamente distribuite con funzione di densità di probabilità $p(x_i;\theta)$

la funzione congiunta per l'intero campione si ottiene mediante il prodotto:

$$p(x_1, x_2, \ldots, x_n;\theta) = p(x_1;\theta)\cdot p(x_2;\theta)\cdots p(x_n;\theta)$$

$$L(\theta) = \prod_{i=1}^{n} p(x_i;\theta)$$

la stima di max verosimiglianza sarà il valore di θ che massimizza la funzione di verosimiglianza

per la varianza, però …

MLE per la varianza è pari a $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$

ma abbiamo già visto che questo stimatore non è corretto (in senso statistico)

quello corretto è $\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \mu)^2$ ***che ha valore atteso pari al parametro da stimare***

il metodo della massima verosimiglianza non è l'unico modo per trovare degli stimatori

e non è sempre il migliore

ce ne sono altri

che vi risparmio !

e adesso un bel salto

dalla teoria

alla pratica

per allenare un po' il vostro spirito critico e di osservazione

lo sapevate che …

secondo un'indagine svolta nel 2000
solo il 28,6% delle persone di 16 anni e più usa il Pc
e solo il 18,3% usa internet

però la diffusione dei nuovi media è molto rapida: rispetto al 1995 il numero degli utenti di Pc a casa è praticamente raddoppiato

analizziamo i dati:

solo il 28,6% delle persone di 16 anni e più usa il PC

28,6% **è una percentuale di "successi"**
popolazione: italiani di 16 anni e più
unità: individui
variabile: uso del Pc
modalità: si / no

lo sapevate che …

secondo la stessa indagine svolta nel 2000
circa un terzo delle famiglie (28,1%) possiede almeno un Pc, ma solo il 15,4% ha un accesso a internet

analizziamo i dati:

solo il 15,4% ha un accesso a internet

15,4% **è una percentuale di "successi"**
popolazione: famiglie italiane
unità: famiglie
variabile: avere la connessione a internet
modalità: si / no

qui l'unità è la famiglia e non l'individuo, perché l'accesso a internet è una caratteristica condivisa da tutti i componenti la famiglia, non è personale

lo sapevate che …

secondo la stessa indagine svolta nel 2000
la quota di bambini e ragazzi che usano il computer aumenta progressivamente al crescere del titolo di studio dei genitori

che variabile è il "titolo di studio dei genitori"?

se i genitori sono due, i titoli di studio sono due

per definire una variabile unica, si deve scegliere il titolo di studio più alto tra i due genitori (o il più basso)

sembra una stupidaggine, ma …

qualche anno fa, per formulare una legge che prevedeva contributi alle coppie giovani è stata chiesta all'Istat la stima del numero di nuove coppie coniugate con età minore di 35 anni
e nella prima bozza del testo di legge si citavano proprio le coppie "di età inferiore ai 35 anni"

ma che senso ha? finché entrambi gli sposi hanno meno di 35 anni tutto va bene, ma se lui ha 36 anni e lei 34 che succede?

lo sapevate che …

secondo la stessa indagine svolta nel 2000
tra la popolazione italiana
l'82,8% ascolta musica
86,6% rock e pop
39,1% classica
29,9% dance e house
29,5% jazz e blues

la somma non fa 100
perché ogni % è
tratta da una variabile diversa
perché una persona può
ascoltare due o più tipi di musica

analizziamo i dati:

sono tutte percentuali, ma dov'è il 100?

82,8% è una % sulla popolazione
su 100 persone quasi 83 ascoltano musica

86,6% è una % sulla popolazione che ascolta musica
su 100 persone che ascoltano musica quasi 87 ascoltano rock e pop

lo sapevate che …

secondo la stessa indagine svolta nel 2000
il 20,2% della popolazione di 3 anni e più pratica sport con continuità
il 9,8% lo pratica saltuariamente
il Nord-est è la ripartizione geografica con la quota più elevata di persone che praticano sport con continuità (25,6%)

analizziamo i dati:

vi è del tutto chiaro il significato di questi dati?

cosa vuol dire "con continuità"?
almeno due volte a settimana per tutto il 2000?
almeno una volta a settimana per almeno 6 mesi?

e come la mettiamo con gli sport stagionali, ad esempio lo sci?

per capire bene il significato di queste percentuali bisognerebbe conoscere il questionario somministrato alle persone selezionate nel campione

lo sapevate che …

secondo un'indagine del 1999
nel corso dell'anno 3 milioni e 48 mila persone hanno subito un incidente domestico (il 53‰ della popolazione)

avete notato che il valore tra parentesi è per 1.000 e non per cento?

è una rappresentazione usata per eventi rari (come gli incidenti domestici) per i quali i valori per 100 sarebbero molto bassi

più di tre quarti (il 79,1%) di tutti gli incidenti vengono subiti da donne

qui le unità di riferimento sono gli EVENTI, non gli individui
(un individuo può aver subito più di un incidente)

cioè la percentuale viene calcolata ponendo a

	denominatore	**tutti gli incidenti domestici**
e a	**numeratore**	**tutti gli incidenti subiti da donne**
unità: incidenti		**variabile: sesso di chi lo subisce**

Io sapevate che …

persone di 6 anni e più secondo il linguaggio abitualmente usato in diversi contesti relazionali. anno 2000 (valori percentuali)

in famiglia	*solo o prevalentemente italiano*	*44,1*
	solo o prevalentemente dialetto	*19,1*
	sia italiano sia dialetto	*32,9*
	altra lingua	*3,0*
con gli amici	*solo o prevalentemente italiano*	*48,0*
	solo o prevalentemente dialetto	*16,0*
	sia italiano sia dialetto	*32,7*
	altra lingua	*2,4*
con gli estranei	*solo o prevalentemente italiano*	*72,7*
	solo o prevalentemente dialetto	*6,8*
	sia italiano sia dialetto	*18,6*
	altra lingua	*0,8*

la somma fa 99,1

dov'è il 100% ?
non c'è perché queste % sono al netto di errori e mancate risposte

questa tavola riproduce una variabile doppia **X = contesto relazionale**
Y = lingua utilizzata

notate il titolo, indica:
popolazione, variabili, periodo di riferimento e tipo di dati (%)

Io sapevate che …

persone di 6 anni e più secondo il linguaggio abitualmente usato in diversi contesti relazionali. Italia e Veneto. Anno 2000 (valori percentuali)

		Italia	*Veneto*
in famiglia	*solo o prevalentemente italiano*	*44,1*	*22,6*
	solo o prevalentemente dialetto	*19,1*	*42,6 (*)*
	sia italiano sia dialetto	*32,9*	*29,8*
	altra lingua	*3,0*	*3,9*
con gli amici	*solo o prevalentemente italiano*	*48,0*	*23,7*
	solo o prevalentemente dialetto	*16,0*	*38,2 (*)*
	sia italiano sia dialetto	*32,7*	*34,4*
	altra lingua	*2,4*	*2,7*
con gli estranei	*solo o prevalentemente italiano*	*72,7*	*52,4*
	solo o prevalentemente dialetto	*6,8*	*14,2 (*)*
	sia italiano sia dialetto	*18,6*	*32,0*
	altra lingua	*0,8*	*0,2*

qui le variabili sono tre: Z = area geografica

(*) in Veneto siamo affezionati al dialetto

lo sapevate che …

secondo un'indagine del 2001

l'86,4% delle imprese con 10 e più addetti è dotata di attrezzature informatiche (Pc o terminali)

tra le imprese informatizzate, l'84% utilizza Internet come tecnologia di rete e il 40,1% dispone di un sito Web

il ricorso al commercio elettronico è ancora limitato: l'11,6% delle imprese ha effettuato nel 2000 acquisti on-line e il 3,6% delle imprese ha venduto on-line

è sottinteso che si tratta sempre di imprese con 10 o più addetti

lo sapevate che …

secondo un'indagine del 1998
il 35,3% delle famiglie italiane possiede almeno un cellulare
nel 1997, la stessa percentuale era pari al 21%

avete notato come la percentuale è cambiata nel giro di due anni?
è passata dal 21 al 35,3 per cento (incremento pari al 68,1%)

l'incremento è calcolato come segue:

35,3 / 21 = 1,681
1,681 - 1 = 0,681
0,681 * 100 = 68,1%

se il fenomeno ha mostrato un incremento così elevato nel giro di un anno, cosa pensate sia accaduto tra il 1998 e oggi?
sicuramente, data la velocità di evoluzione, i dati del 1998 (pur essendo gli ultimi a disposizione) non sono affidabili; sono obsoleti, vecchi

Io sapevate che …

nell'anno accademico 2000/2001 si sono registrate 38.023 iscrizioni al primo anno dei corsi di diploma universitario

per 100 immatricolati all'università, 12 hanno scelto i corsi di diploma

immatricolati a corsi di diploma per sesso. Anno accademico 2000/2001

	Iscritti	% Maschi	% Femmine
totale diplomi	38.023	45,4	54,6
- Ingegneria	3.524	88,0	12,0
- Ingegneria informatica	1.014	-	-

Io sapevate che …

esiste un'indagine Istat sull'inserimento professionale dei laureati e diplomati universitari

dai dati 1999 raccolti su un campione di diplomati nel 1996 è risultato che

- ***i diplomati universitari hanno possibilità di inserimento lavorativo più agevoli rispetto a quelle dei laureati***

- ***tra i diplomi, non tutti garantiscono le stesse opportunità; quelli del gruppo di ingegneria offrono le possibilità migliori:***
  - ***quasi l'80% dei diplomati in ingegneria dopo tre anni svolge un lavoro continuativo iniziato dopo la conclusione degli studi (la media per tutti i diplomi è 54,5%)***

Io sapevate che …

diplomati del 1996 che nel 1999 lavorano e si dichiarano soddisfatti rispetto ad alcuni aspetti del lavoro svolto

	Tutti i diplomi	*Ingegneria e architettura*
trattamento economico	*72,7*	*79,0*
possibilità di carriera	*56,6*	*72,6*
stabilità del posto di lavoro	*68,8*	*84,5*
utilizzo delle conoscenze acquisite	*68,4*	*72,1*

fonte: Istat, Indagine sull'inserimento professionale dei diplomati universitari del 1996