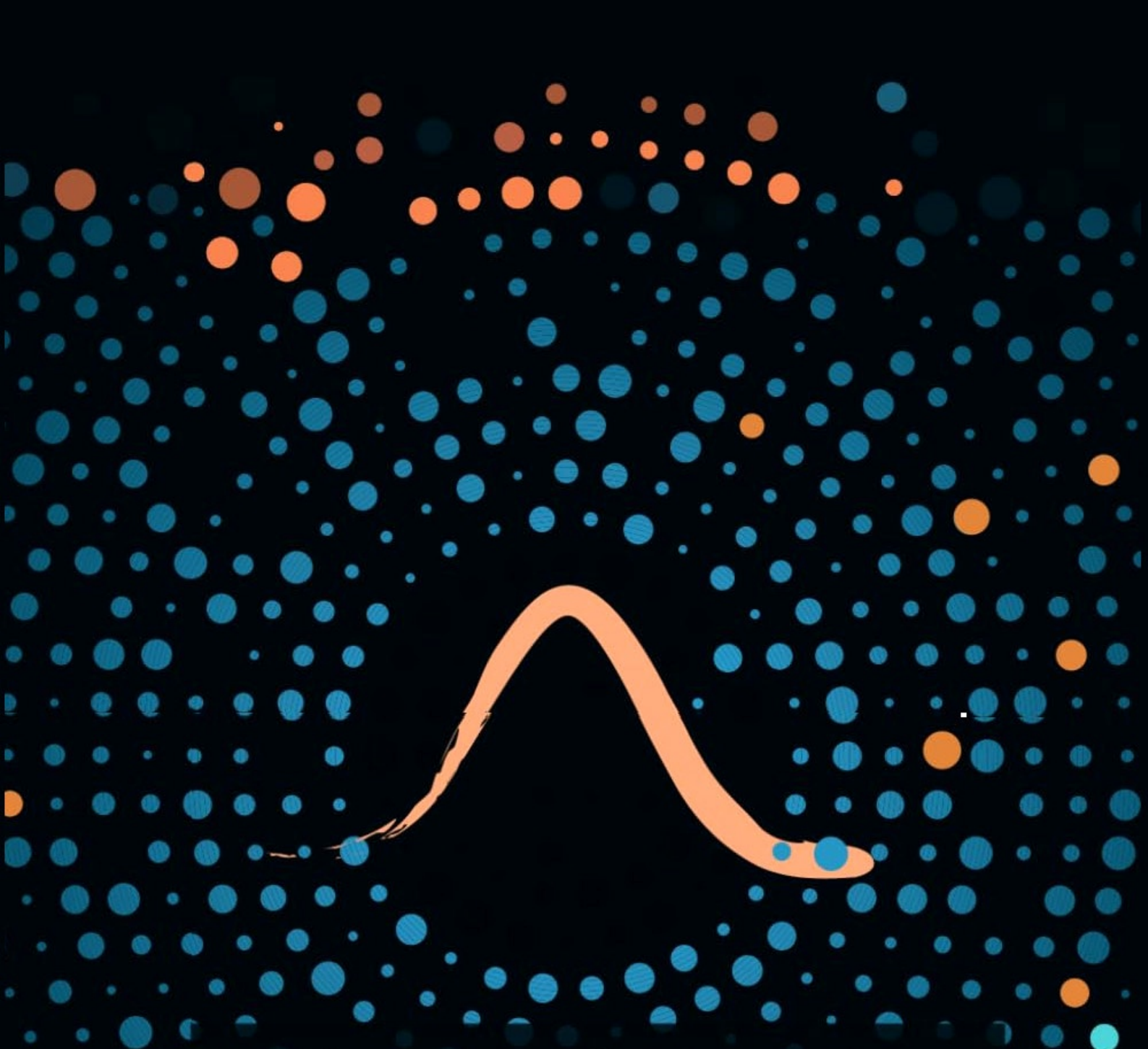


Modern Statistics

Intuition, Math, Python, R

Mike X Cohen

sincXpress

7- Anes and Improve Data Quality

7.1 - Data quality matters

ben organizzati // poco rumore // pochi artefatti // pochi dati mancanti /

- si usano tecniche di filtraggio, trasformazioni
- fulire i dati richiede meticolosità; critico valore, basso rapporto e universale, conoscenze statistiche, esperienza pratica, solida comprensione dell'argomento, familiarità con le caratteristiche uniche dei propri dati.
- La qualità è fondamentale, su queste analisi sono prese decisioni operative.

Garbage in, garbage out (GIGO)

Dati scadenti → risultati scadenti

Dati di qualità sorprendenti → risultati sorprendenti non fenomenali

↳ necessari, ma non sufficienti

- i dati sono sempre imperfetti. La domanda è: ho bisogno di ulteriori dati?

7.2 - Data cleaning phases

4 finestre temporali x data quality

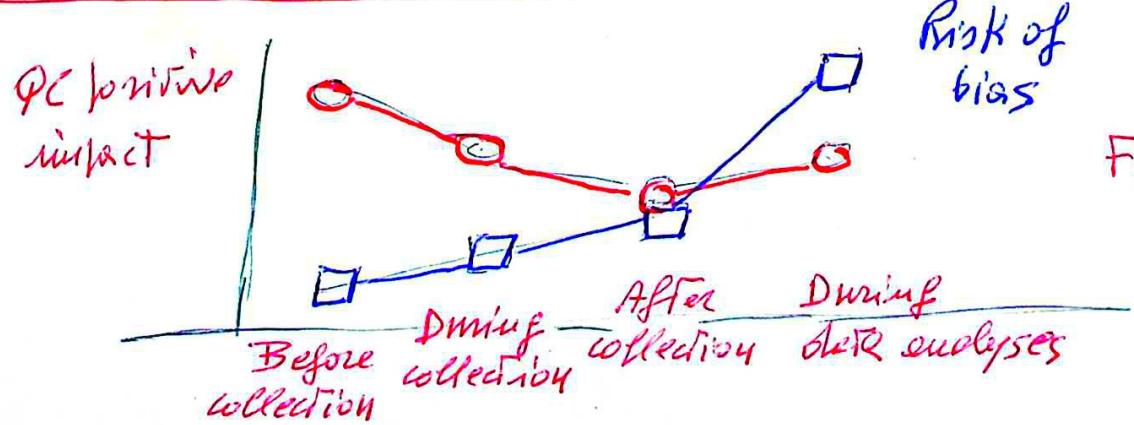


Fig. 7.1

P. 240

Before getting data

- familiarizzare con pubblicazioni scientifiche, relazioni tecniche, blog, youtube
- condurre uno studio pilota, x intercettare problemi imprevisti

During data collection

- ispezionare i dati in tempo reale x verificare presenza di artefatti, errori di calibrazione dei sensori -
- provare a risolvere un problema

After data collection

- pulizia dei dati, 2 strategie / — trasformazione dei dati
  — rimozione dei dati errati
  — questo è il momento perfetto x aumentare la qualità!

During data analysis

possono essere introdotti bias, che cambiano il risultato -

- meglio evitare la trasformazione dei dati - se possibile

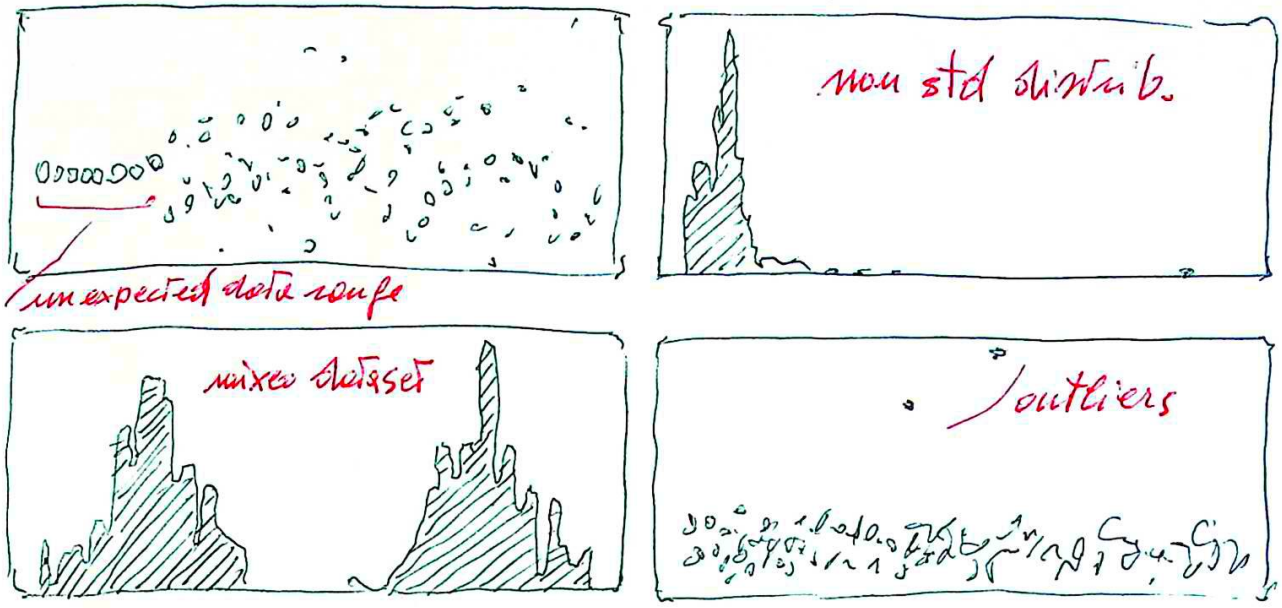
7.3 Assessing data quality

- ci sono 2 famiglie di approcci x fare questo (NON sono mutuamente esclusive)

Qualitative quality assessment

- visualizzare i dati in modalità differenti, cosa cercare in queste visualizzazioni?
  — l'intervallo dei dati, tenere che i valori siano quelli attesi
- la forma della distribuzione

— facciamo qualche esempio con l'aiuto di grafici



Quantitative quality assessment

- le metriche + utili tendono a essere specifiche dell'applicazione e talora specifiche del set di dati. Vediamone un elenco limitato:
  - sample size – molti campioni → risultati statistici + affidabili; molti campioni → diminuire numero parametri della popolazione + accurati. Il paragone a migliore qualità non è comunque automatico. Quando un campione è grande? → cap. 17
  - Error rates – qui ci stiamo riferendo ai valori ... , però → misurare la quantità dei rimossi → semplicità → troppi errori → dubbio sulla qualità dell'esperimento
  - Data range – ci riferiamo agli estremi. Nella pratica difficilmente i click di un set di dati può essere nullo.
  - Variance – se troppo piccolo o troppo grande → problema – troppa varianza → potrebbe oscurare i dati reali; curare la normalizzazione x permettere confronti

7,4 - Improving data quality through transformations

- distribuzioni fortemente tagliate da un lato → diventano simili a Gauss con una trasformazione log
- la conoscenza del dominio aiuta a scegliere la trasformazione
- testare prima e dopo x vedere se ci sono miglioramenti

7,5 - What are outliers? = valore relativamente insolito

- decidere se "outlier" ha una componente di soggettività.

Il valore anomalo è un NO non valido

- ci può essere rumore, errori, malfunzionamento dei sensori (es. 0,0254 → può diventare 254 se la virgola non è gestita bene)
- una data sbagliata (errore di battitura)

Il valore anomalo è valido es. dati sul patrimonio netto di ogni maschio negli USA, con nome di battesimo Jeff. Più incontriamo Jeff Bezos con un patrimonio > 100 e9

es. assicurazioni contro le inondazioni → in Texas >> Utah
es. carta da regalo - bene tutto l'anno, non in inverno a Natale
es. transazioni con carte di credito relativamente piccole in una data area → un addebito particolarmente elevato potrebbe essere fraudolento

- non-rappresentativo, anomalo, estremo, deviante
- sono difficili da gestire

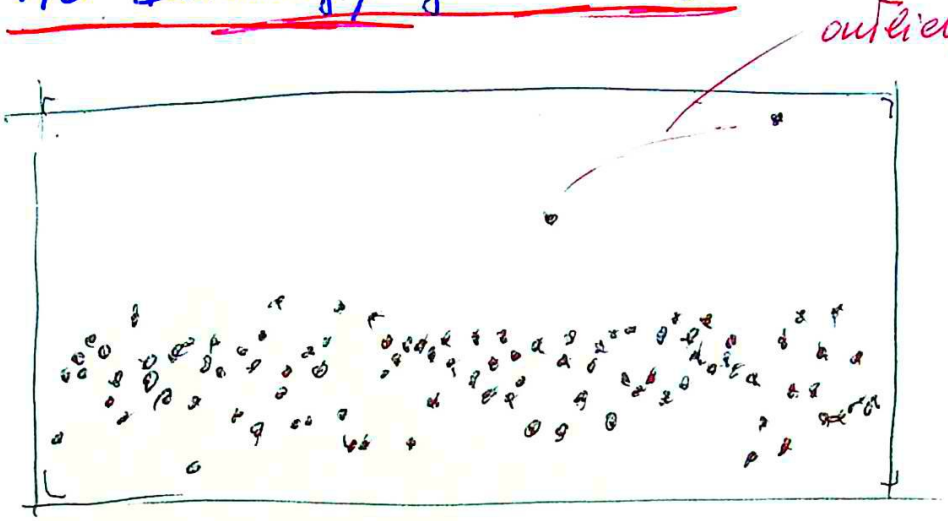
Outliers richiedono decisioni

- Rimozione outliers - se non riflettono il sistema in oggetto
- Tenere outliers
  1) applicare trasformazione x ridurne loro impatto (es. log)
  2) usare analisi robuste rispetto a essi (statistiche non parametriche, regressioni robuste (ponderate)

→

3) analizzare i sottogruppi → separazione dei dati in gruppi con caratteristiche simili, poi analisi statistiche sui singoli gruppi

7,6 - Identifying outliers



La sola ispezione visiva non è un metodo ottimale.

1) non è scalabile, funziona bene su dati piccoli o universali;
2) rischio di soggettività: bias soggettivi, strategia analizzatore
3) se persone diverse → probab. scelte diverse → risultati diversi

Absolute threshold detection – ciò che è fuori dallo certo intervallo è escluso. Qui si vede molta conoscenza del fenomeno. È una scelta personalizzata x quel set di dati.

The z-score method — z-score

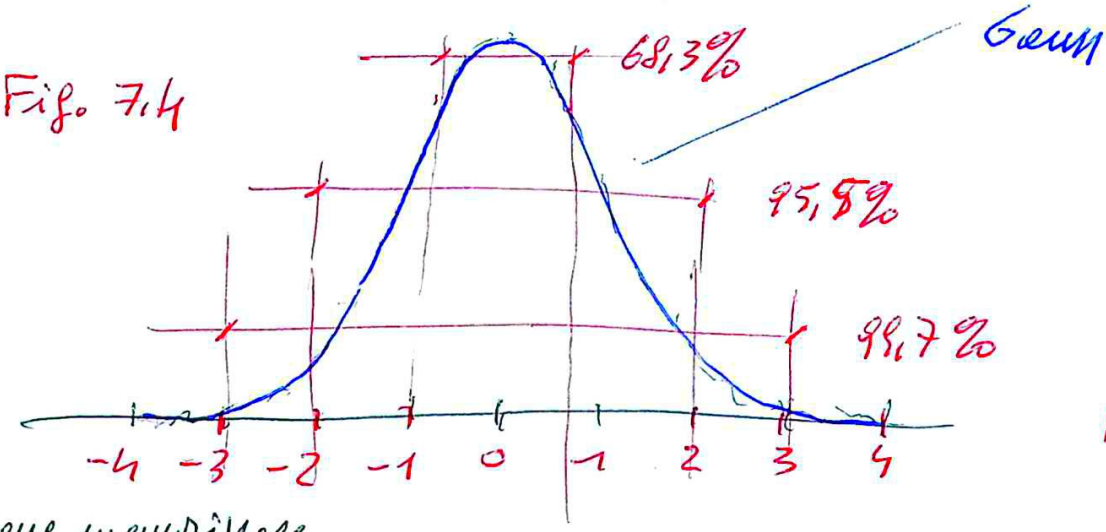
Per rimediare il con precedente → soglia relativa, ecco x che')

z-scores and data proportions — z-score

- possiamo valutare la quantità di dati entro diversi limiti di')

Vediamo meglio con una figura:

Fig. 7.4



p. 251

- è bene misurare

σ →	68,3 %
2σ	95,5 %
3σ	99,7 %

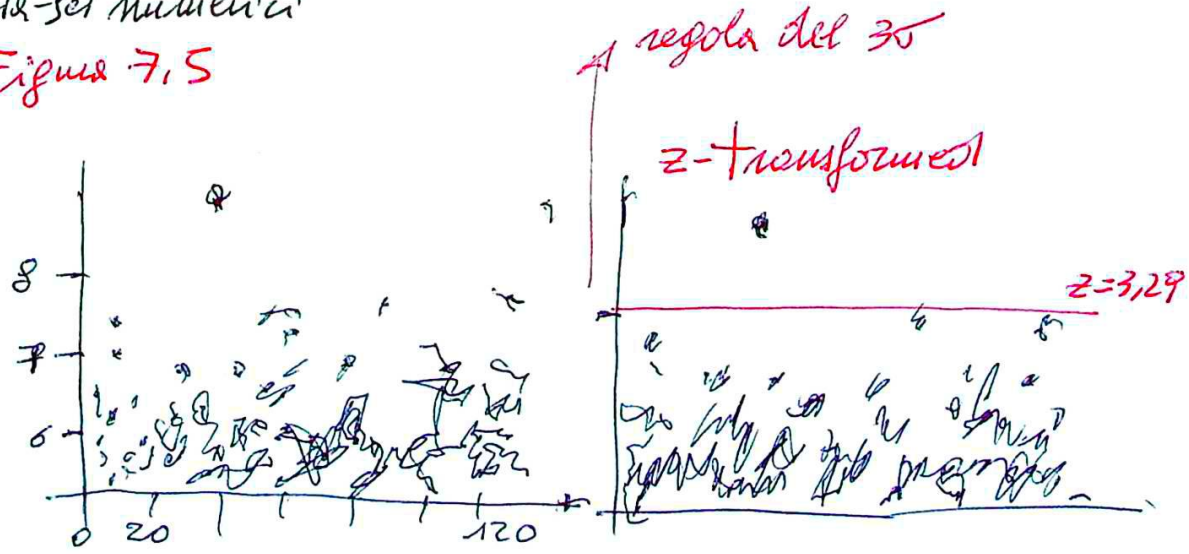
X calcolo di dettaglio → Cap. 10

- possiamo usare z-score x identificare dati relativamente grandi come outliers

The z-score method

- E' il metodo + comune x identificare outliers. Si convertono i dati con z-score, si considerano quelle che superano data soglia.
- Il punto di forza è che il metodo funziona x un minimo di data-sets – indipendentemente dalle unità di misura
- Vale x i data-set numerici

Esempio – Figura 7.5



più in generale: $|z| > 3,29$

- z-score è riproducibile
- il metodo è "relativo"
- ha un parametro la soglia (una certa soggettività)
  - Valori + comuni: 2.3, 3, 3.29

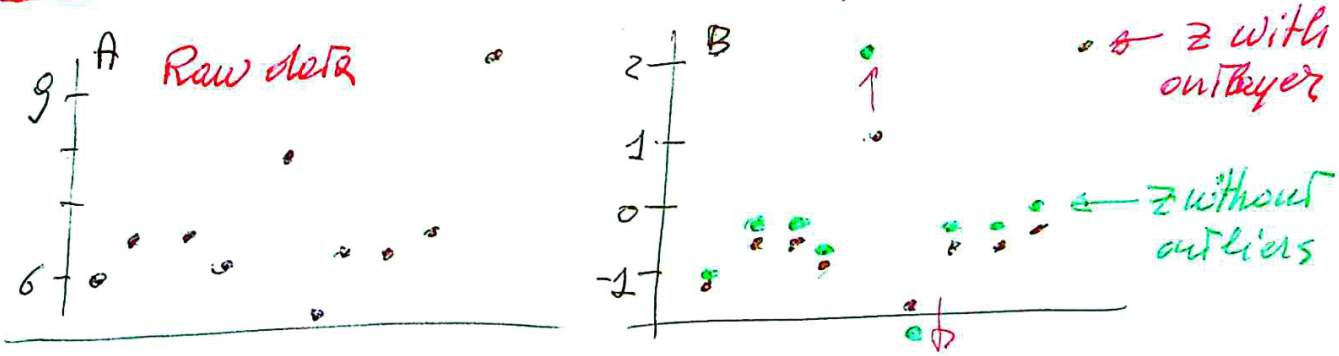
– Se il fenomeno in oggetto è di per sé molto disperso → anche 5, 8

– i dati "non-normali" potrebbero richiedere soglie molto alte → z-modificato

Modified-z

- usa la mediana e la MAD, sono le + appropriate x tendenza centrale e dispersione. Poi approccio similare (soglia)

Iterative z-score method – vediamo problemi con fig. 7.6

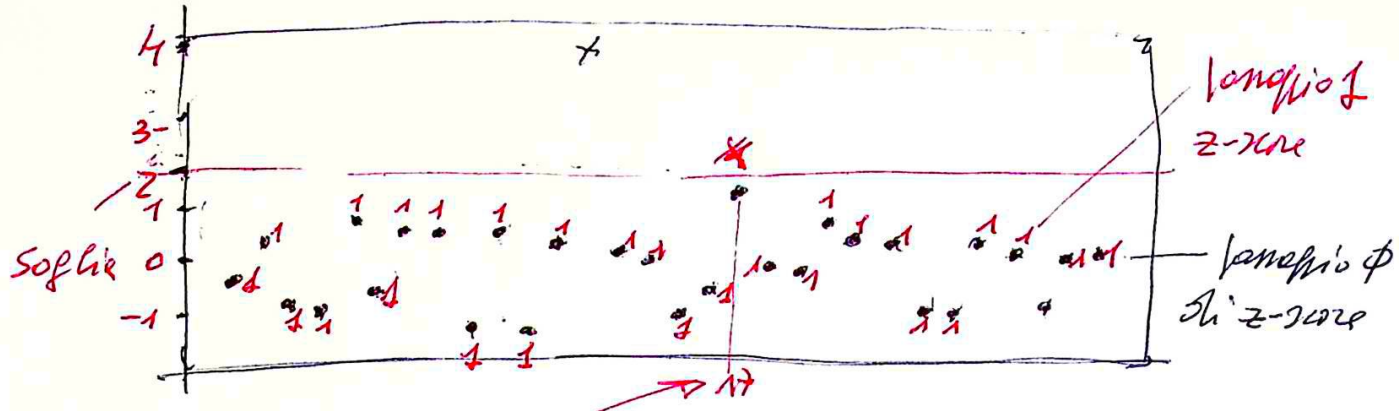


z-transform modifica i dati: sempre outlier si espande rispetto allo zero (e' cambiata la media)

---

1) Implementiamo z-score come prima descritto
   - diamo una etichetta ai dati + estremi che superano la soglia classificandoli come outliers
2) Ricalcoliamo z-scores senza gli outliers. Segnamo con una label i nuovi dati che superano la soglia
3) Ripetere fino a che non ci sono + outliers

Vedi fig. 7.7



Nota - fatto uno z-score, elimino outliers, → ricalcolo

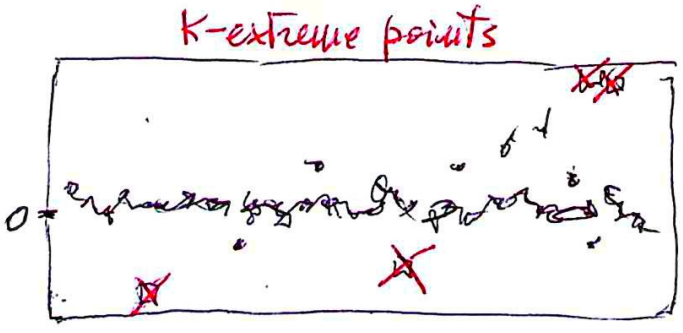
- che dire del fg 17?
  non sembra irrelativamente grande (soglia $z=2$) ⇒
- i metodi iterativi possono rimuovere troppo
- fidarsi ciecamente è sconsigliabile - ispezionare i dati
- se ci sono outliers grandi e piccoli → quelli grandi possono oscurare quelli piccoli
- non è di utilizzo comune questo metodo

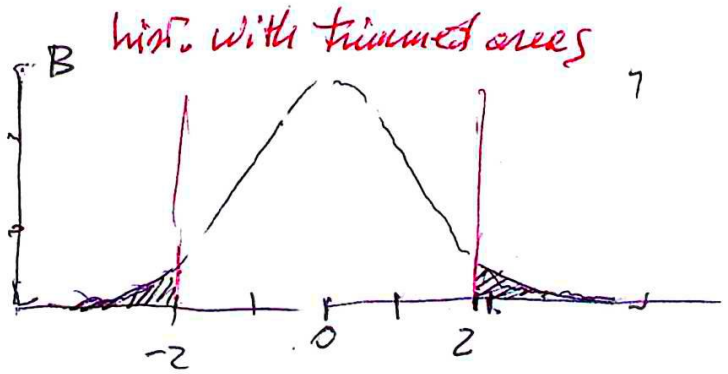
È utile sapere che ∃ questa possibilità

Removing data by trimming

Simile al metodo della "soglia", qui però si rimuovono i punti estremi, oppure si rimuovono le "code" della distribuzione.

- Es. poniamo un soglia $z=2$ x ogni coda Fig. 7.8
- L'autore non è un fan di questo metodo
- i vantaggio sono che il metodo è semplice, riflettivo





- è bene conoscere questo metodo

Manual, automatic, and semi-automatic cleaning

- se automatico → implica fiducia nel metodo
- se manuale → ispezione visiva. Esperienza del dominio
- semi-automatico → mix, magari si osservano in automatico i possibili outliers, si procede a mano nella eliminazione
  - alla fine del lavoro, si può dire (bisogna avere esperti) ok all'uso in automatico in quelle circostanze
  - migliora il problema della scalabilità - rende efficiente la analisi in termini di tempo - ci si può concentrare solo sui dati rilevanti
  - molto utilizzato nella elaborazione di dati geospaziali e immagini (es. può rilevare un gran versamento su un deposito, la specialità viene informata x eventuali azioni)

What happens to rejected outliers?

- potrebbe essere fattibile rimuoverli (come se ∄)
- si possono presentare in una "feature" e non in altre, potrebbe essere stesso cancellare una intera riga, solo perché ~ lo contiene (es. outlier in una sola colonna)
- possono essere "accettati", allora devono essere abbinati alle corrispondenti righe una in colonne diverse, o in data set separati
- possono essere rimpiazzati con "NaN" = not a number - Nella computazione sono ignorati.

%

Hybrid methods

- si possono mescolare e abbinare diverse strategie di rilevamento
  - es. soglia assolute (si eliminano quelli eclatanti) / metodo iterativo (x identificare outliers non-enormi) // z-modificato (se i dati non sono Gauss)
- molti metodi ottimizzati sono specifici

7.7 - Analysis-based solutions to outlayers

- Talora gli outliers sono significativi, li vuoi mantenere, ma non vuoi che alterino i risultati
- ci sono metodi di analisi progettati x minimizzare il loro impatto

Nonparametric analyses

- molti metodi es. t-test, correlazione, ANOVA hanno alternative non parametriche che implicano la trasformazione in base al rango o al test delle mediane. Sono robuste rispetto a outliers.
- tuttavia queste analisi non-parametriche possono avere sensibilità ridotta nel rivelare effetti sottili.

Permutation-based tests

- creano distribuzioni statistiche "null hypothesis"
  - si ipotizza che ∄ relazione tra due serie di dati o variabili analizzate.
    - se vera → ∀ effetto osservato è dovuto al caso

- questa "null hypothesis" distribution si ottiene ~~ogni~~ eseguendo numeri casuali – I valori anomali in una condizione vengono eseguiti casualmente ed altre condizioni –
- ciò significa che i valori anomali vengono incorporati nel processo di inferenza statistica. → Cap. 16

ESEMPIO – H_0 = ipotesi nulla

- se $(x_1 - x_n)$ è di un certo tipo → rifiuto H_0
- l'ipotesi alternativa è H_1

– H_0 la si ritiene vera fino a prova contraria

- il Test di ipotesi ci darà una regola di decisione (o rifiuto)

- Se H_0 = True → molto improbabile che il campione ci porti al rifiuto

REGIONE di RIFIUTO

NON COMPLETO

$R = x_1 - x_n$ porta a rifiutare se

statistica ottimale

$$R = \{(x_1 - x_n) : t(x_1 - x_n) \in I\}$$

intervallo regione di rifiuto

- dare un Test = dare una regione di rifiuto

L ogni Test è associato alla sua R

$I \subset \mathbb{R}^n$ → meglio specificato nel Cap. 16

Weighted analyses

- se il problema e' che gli enomali hanno un impatto sproporzionato sui risultati → perche' non de-ponderarli?
- de-ponderare = valore $\times$ i $\quad 0<i<1$ in base al loro impatto

Es. [1, 2, 3, 40] → vettore di ponderazione [1, 1, 1, ...] →

→ [1, 2, 3, 4]

- e' un metodo complicato → Cap. 15

Subgroups analysis

- se ci sono molti enomali → potrebbe essere che rappresentino un gruppo qualitativamente distinto

Es. consumo di carburante di 100 auto

80 = 18 ÷ 25 miles/gallon

20 = 40 ÷ 60 ——————

(questi appaiono come outliers, in realta' sono auto ibride electric-gas

↳ l'analisi statistica si fara' x gruppi separati.

7.8 - Missing data - Fig. 7.9

- i dati mancanti possono avere un grande impatto. Possono mancare x vari motivi:
  - Drop out - es. se traccio gli stessi individui nel tempo (studio longitudinale). Alcuni partecipanti possono interrompere il test → dati parziali:
    - i dati parziali sono utili in alcune analisi, ma non in altre

- Malfunzionamento attrezzatura
  - perdita irrimediabile, corruzione, inutilizzabili
- Errori umani

Che fare con i dati mancanti?

Vediamo varie modalità

- Rimozione x riga (appropriata x dati accoppiati)
  - es. misure prestazioni sportive prima e dopo una settimana di allenamento
- se ci sono dati mancanti → nessun dato di quell'individuo è utilizzabile

- Analysis-specific row removal
  - però i dati di quella riga potrebbero essere validi, utilizzabili
    - Lo mantenere la riga nel data set, escluderla dalle analisi che necessitano proprio di quei fattori
  - utile nel caso di dati limitati
  - si può sostituire il dato con NaN

- Imputation (replacement)
  - sostituire i dati con valori stimati, basati sui dati disponibili
    - spesso si usa la media

Time 1	Time 2
5	7
6	6
4	?
8	7

Fig. 7.8

6,67 = media di Time 2

- Predictive modeling
- concettualmente simile a sostituzione
- "fitting" the model parameters richiede molti dati

7.8 – Exercises

1) implementare "iterative z-score" outlier removal → → ottenere Fig. 7.7

2) data trimming – creare un dataset casuale $y = \exp^{\sin(x)}$ dove $x \sim SN(0,1)$ con $N = 10'000$ – Fare un copia, poi fare un outlier originale con Trimmed.

Ttrim = 4% dei dati (2% di ogni coda)

– sostituire i dati rifiutati con NaN → 9600 dati validi

– calcolare media, mediana, trim e dopo troncamento →

$\bar{x} = 1,222$	$\tilde{x} = 0,989$	np.mean
$\bar{x}_{trimmed} = 1,054$	$\tilde{x}_{trim} = 0,989$	np.argsort()

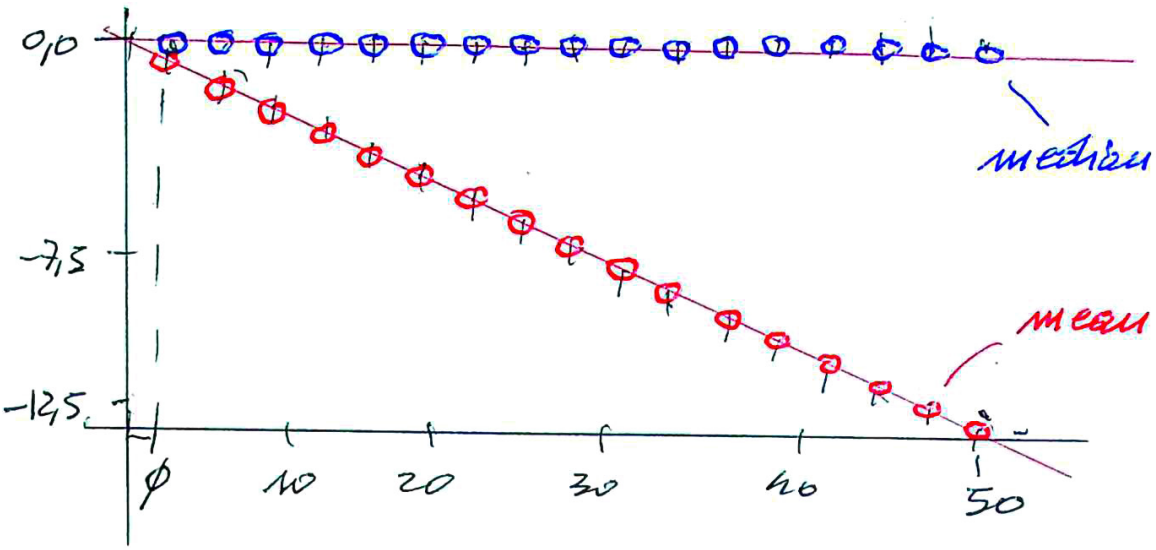
3) mettiamo codice 2) in un loop con K = 1% → 50%

– calcolare media, mediana, e quantile

– utilizzare lo stesso set di dati casuali (→ descrittivi confrontabili)

– Testare K es. se K = 12 → 8800 punti validi

– visualizzare la caratteristica descrittiva con un grafico Fig. 7.10 dove si mostra la variazione % di media, mediana vs. K% to trim.



Commento

- la mediana è imperturbata (n° di significati è uguale nella coda di destra e in quella di sinistra)
- la mediana cala
- x questo set il trimming a due code non è appropriato
- provare con trim su una sola coda – In questo caso come influisce su stat?
- i risultati cambiano qualitativamente con una diversa distribuzione di dati?

p. 263

4) Vediamo l'impatto della rimozione degli outliers sulla forma della distribuzione – valutare se tale risultato dipende dall'algoritmo di calcolo del n° dei bin –

N = 1000 casuali, estratti da una distribuzione F, utilizzando 5 e 100 gradi di libertà → cap. 14

```
np.random.f(5, 100, size=N)
```

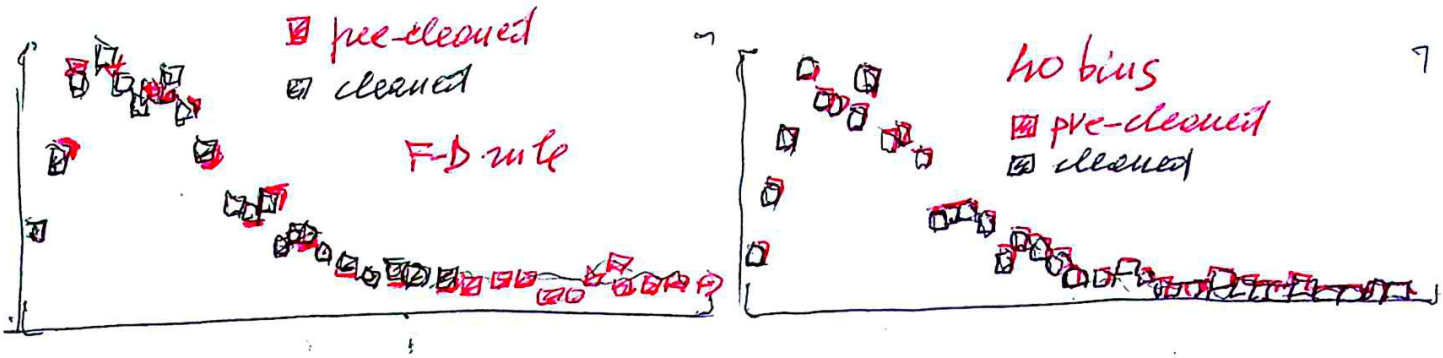
– usiamo z-score e poniamo la soglia z=3

Calcolare % di eliminati

N = 1000
removed = 883
% removed = 1,70%

%

– disegniamo hist. di pre e post cleaning
↳ Freedman-Diaconis Rule – ho bins sulla base degli estremi



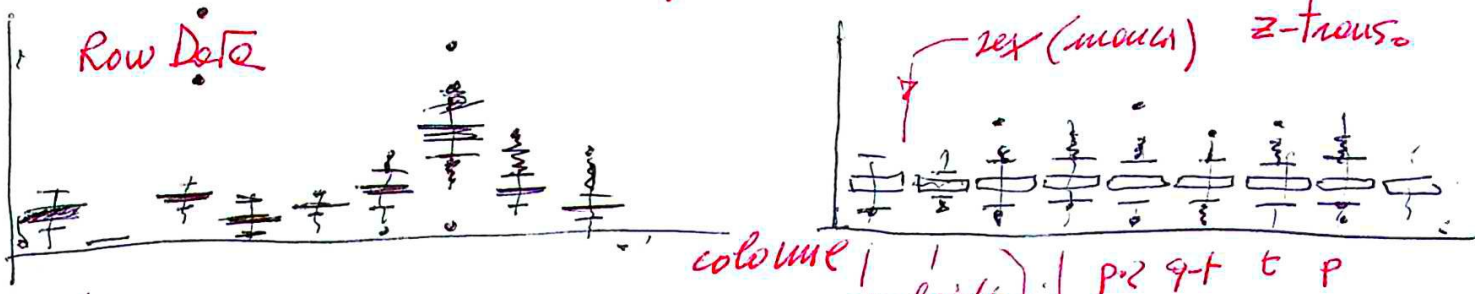
- la forma esatta di un ist. può essere influenzata da scelte di encoding, missing – gli ist. dovrebbero essere interpretati x le loro caratteristiche generali, non x i dettagli e piccole fine–

5) i dati simulati sono una grande cosa, ma quelli reali sono meglio – (es. http://archive.ics.uci.edu/ml)
Avvertimento: questo esercizio e il prossimo si basano su "pandas" e "seaborn" di Python.

- useremo un set di dati medici sulle aritmie cardiache – scopo: prevedere aritmie sulla base di ECG (encefalogrammi) – dobbiamo qui trovare i valori anomali –
- importiamo i dati da Web come pandas dataframe
https://archive.ics.uci.edu/ml/datasets/Arrhythmia
importiamo solo le prime nove colonne

	età	sesso	alt.	peso	qrs	p-r	q-t	t	p
0	75	0	190	80	91	193	371	174	121
1	56	1	165	64	81	174	401	149	39
2	54	0	172	95	138	163	386	185	102
3	55	0	175	94	100	202	380	179	143
4	75	0	190	80	88	181	360	177	103
5									

- facciamo una copia del dataset, applichiamo z-score a tutte le colonne (eccetto "zero"), creiamo un "boxplot"



vediamo che ci sono molti outliers

L x ogni variabile ci vuole diversi

L bisogna creare cardiologi x poter scegliere

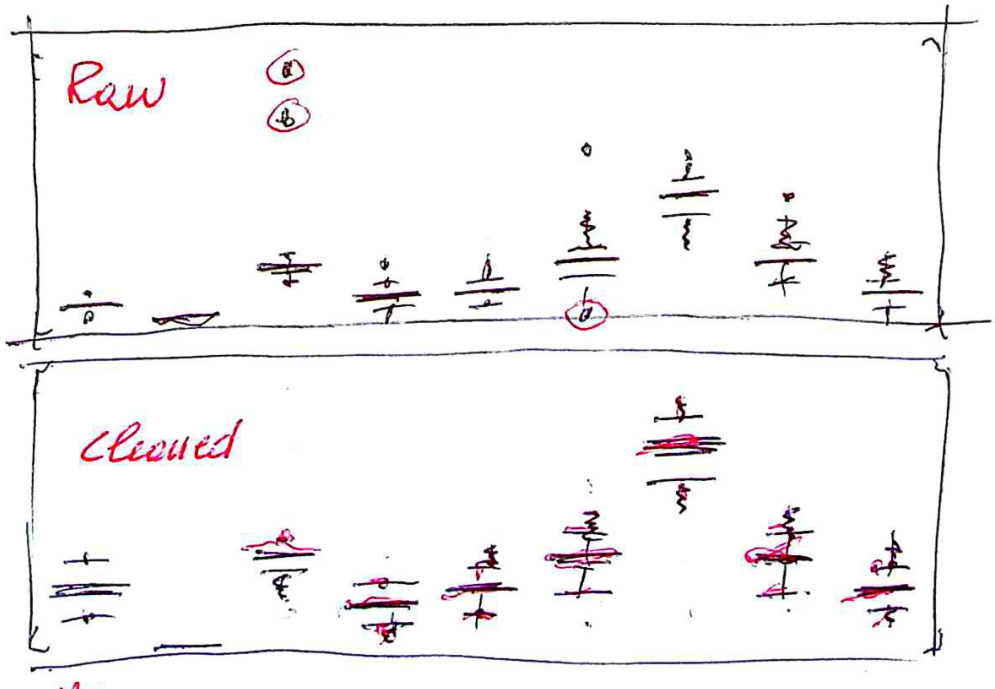
NOTA

qui è sufficiente una soglia applicabile a tutte le variabili

---

6) Riconoscimento outliers e loro rimozione

- soglia $z = 3{,}29$
- nel raw data, gli outliers sono diventati NaN



age

- outliers positivi e negativi sono stati eliminati

- stampiamo le medie di raw e cleaned

age	46,47	46,47
sex	0,55	0,55
height	166,19	163,84
weight	68,17	68,33
qrs	88,82	87,48
p-2	155,15	160,38
q-t	367,21	368,05
t	169,95	168,28
p	90,00	91,14

prende un po' il femminile

ora ne facciamo un grafico in %

$$\frac{\text{cleaned} - \text{raw}}{\text{raw}} \cdot 100$$

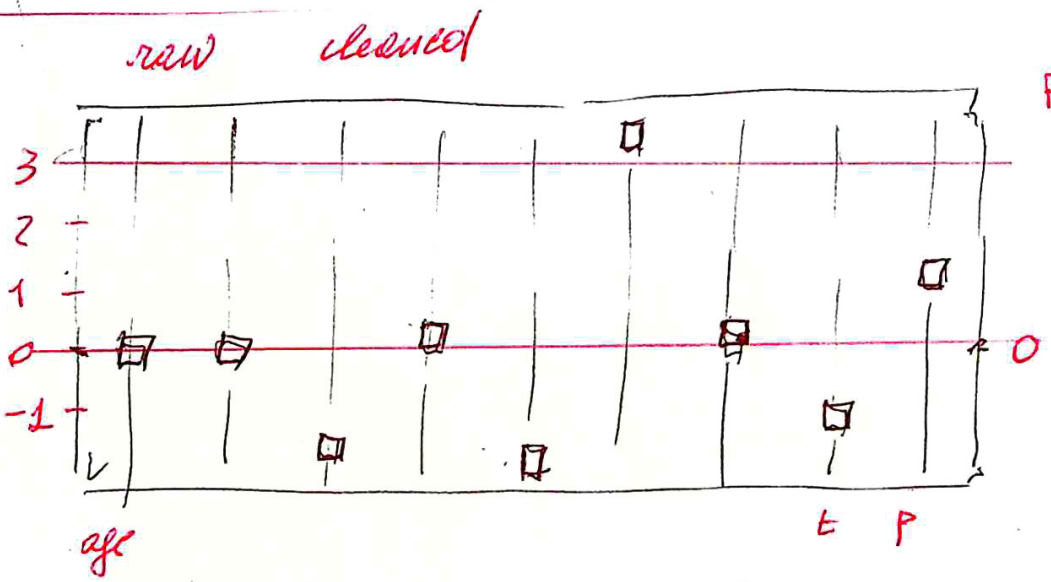


Fig. 7,15

- qui abbiamo visto che i dati anomali sono presenti nei dati reali -

p. 268

— FINE CAPITOLO 7 —

8 - Probability Theory

8.1 - From descriptive to inferential statistics

- Questo è un po di molte
  - le statistiche inferenziali usano i dati x fare previsioni "pensate" (educated guesses) sul mondo al di là del mio dataset -

es. Occhiali (riflessione
- descrittive: dimensione, forma, colore, caratteristiche di
- inferenziali: con quegli occhiali guardare il mondo)

L x capire meglio partiamo dalla probabilità
L sono un aiuto a prendere decisioni su un mondo incerto
L la probabilità quantifica l'incertezza

"Pure" vs. "computer" probabilities

- La teoria della probabilità è molto orientata matematicamente
- ∃ però la statistica applicata → quel pezzo di rigore non è più indispensabile - Questo sarà il taglio del libro -
- Loro riferimento ad approccio "computer", "empirici" -
- La statistica computazionale comporta imprecisioni di calcolo, errori di arrotondamento, domini ristretti - Tutte queste problematiche possono essere affrontate con Python e R -

8.2 - What is probability?

Nella realtà concreta nulla è certo. La probabilità varia tra $\phi \div 1$, può essere espressa come percentuale

8.2.1 Il problema con la probabilità

- Talora la probabilità è anche controintuitiva - Vediamo di riconoscere frasi corrette e non corrette:
  1) pioverà x il 40% della giornata
  2) —— nel 40% della città, il rimanente sarà asciutto
  3) c'è una probabilità su 5 che piova oggi in un certo punto, — ok da qualche parte a Osaka
  4) i meteorologi sono sicuri al 40% che pioverà oggi

  L no, la confidenza è differente dalla probabilità (40%)

  L sono ragionevolmente al 99% sicuri di una probabilità di pioggia del

- Esempi di probabilità:
  - lancio della moneta - ogni faccia ha probabilità = 0,5
  - lancio del dado - ha 6 facce → probabilità = $1/6 \simeq 16,66\%$
  - gioco delle carte - 52 carte → —— = $1/52 = 1,92\%$
    - prob. di una carta rossa = 0,5
    - —— —— regina (di ogni seme) = $4/52 = 7,69\%$
  - la probabilità di pioggia nel deserto non è 50, anche se 2 sono le possibilità.

- la probabilità totale somma = 1 - proviamo visualizzare come? Il tutto 1

L sul PC potrebbe non tornare esattamente

L vedremo nel seguito esempi di ciò.

When do we need probabilities?

- Alcuni eventi sono così probabili che non serve il concetto di probabilità.
- Alcune situazioni (es. un referto medico) richiedono invece la probabilità.

8.3 - Probability vs. proportion

- Probabilità e proporzione sono concetti correlati, ma distinti.

Es. lanciamo 10 monete

L'esperimento dà 6 monete con testa →

probabilità = 0,5 = 50%
proporzione = 0,6 = 60% →

PROBABILITÀ = la possibilità che un evento accada

PROPORZIONE = una frazione del tutto

Es. (1020 min / 11)

- impiego 5,1 min/day a lavarmi i denti su 17 ore di veglia.
- proporzione della parte di veglia usata x lavarsi i denti:
  51/1020 = 0,5%
- probabilità che "un minuto nella mia vita" includa il lavaggio dei denti = 0,5%
- in questo caso numericamente uguali

- accade che il valore della probabilità dipenda da come è posta la domanda.
- la probabilità riguarda un evento futuro
- la proporzione è una semplice conta.

%

8.1 - Computing probabilities

Dove nasce la probabilità?

Intuizione della conoscenza semantica

Conoscenza semantica = l'informazione sul mondo che conosciamo.

- La maggior parte delle probabilità che conosciamo sono inutili nell'analisi statistica formale. Possiamo inoltre avere opinioni che in realtà sono fallaci.

Una formula matematica

- Impareremo la meccanica della determinazione della significatività statistica =
  - prob. che una statistica descrittiva di un campione rifletta le caratteristiche di una popolazione con parametri noti
  - L deriva da una formula

Misure empiriche

p. 277

La probabilità si stima da misure empiriche.
Nei casi + semplici, matematicamente:

$$p(x_i) = \frac{N(X = x_i)}{N(X)} \qquad (8,1)$$

$N(X = x_i)$ ← occorrenza di x_i

$N(X)$ ← set degli eventi possibili

Es. Barattolo con 90 biglie / 40 blu, 30 gialle, 20 arancione

quale la probabilità di fissare un colore?

- questa risposta può essere soddisfatta anche senza dati empirici

$$b\&y = \frac{40}{90} = 44{,}4\%$$

$$giallo = \frac{30}{90} = 33{,}3\% \rightarrow 44{,}4 + 33{,}3 + 22{,}2 = 99{,}9\%$$

~~22,2%~~

More complicated analytical probabilities

- Se prendo un generico set di dati empirici → ho bisogno di calcolo numerico

Calcolare probabilità empiriche

Vale sempre la 8.1 – però consideriamo meno, es. supponiamo che nel barattolo ci sono 40 biglie

Requisiti sui dati di cui compilare le probabilità

- due sono i requisiti
  - i dati possono essere numerici – discreti; ordinali; categoriali Es.
  - supponiamo di voler conoscere la probabilità empirica che un pinguino abbia un certo peso
    N=100 devo creare un intervallo di peso
    Es. tra 3 e 3,1 kg
  - i dati di intervallo devono essere convertiti in intervalli discreti
  - bisogna dunque contare gli eventi favorevoli (il numeratore della 8.1)

%

— i dati devono avere etichette o contenitori mutuamente esclusivi — Es.

se un ginnasio registra 3 e 3,1, non può pesare tra 2,8 e 2,9

Contro-esempio

- dove le persone prendono le misure? — cap. 3
- possiamo calcolare la probabilità di quella tabella? NO

L x che la Σprob. > 100%

Interpretare le probabilità empiriche

Si basano sui campioni / campioni diversi possono avere probab. empiriche diverse —

Per superare questa difficoltà: misurazione di campioni + grandi / calcolo di intervalli di confidenza

— le probabilità spesso sono uguali alla proporzione, che è un concetto di cose già accadute

8.5 - Probability functions, mass, and density

Probabilità = la possibilità che un evento accada

Funzione di probabilità = una espressione matematica che collega ogni elemento di un set a un valore di probabilità numerico

Mappa di probabilità = una funzione che descrive le probabilità per un set di eventi discreti esclusivi

Densità di probabilità = una funzione che descrive le probabilità per eventi continui esclusivi

in formule:

$$f(x) = p(X = x) \qquad (8,5)$$

set di valori — in particolare valore

permette di eseguire a quell'evento una probabilità.

$$f(x_i) = p(X = x_i)$$

8,6 / 8,7 / 8,8 / 8,9

$$p(X = x_i) \geq 0 \;/\; p(X \neq x_i) = 0 \;/\; \sum_{i=1}^{n} p(X = x_i) = 1$$

- Massa di probabilità x eventi discreti (si visualizzano con barre,
- Densità di probabilità x ——— continui con linee)

Es. (non reale) — barre — in Trimpeco

Tipi di auto - sono categoriali, non hanno ordinamento

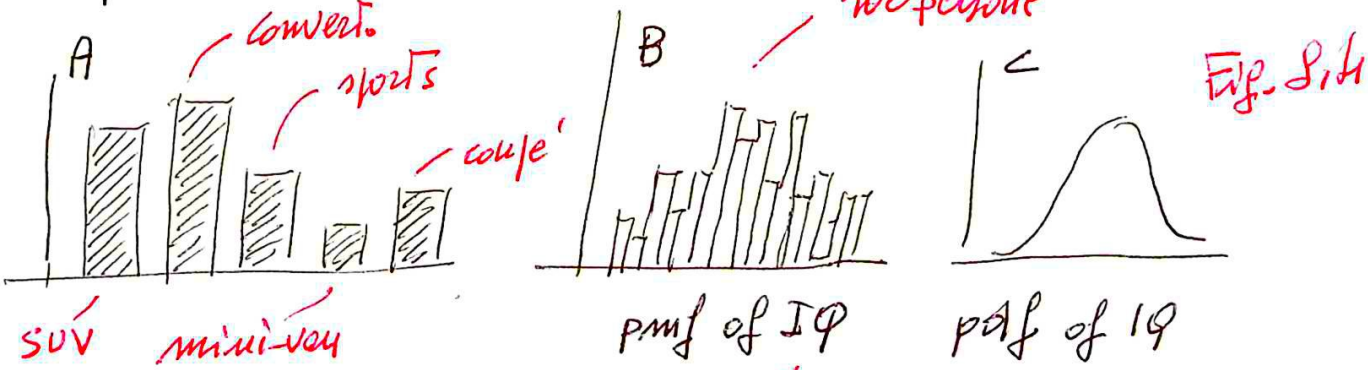


Fig. 8.14

pmf di tipologie di auto — quoziente di Intelligenza

- Consideriamo B
  - se proporzione → stai caratterizzando quel particolare campione di dati
    - non puoi fare inferenze sul resto della popolazione
  - se massa di probabilità → si possono fare inferenze su persone al di fuori del tuo campione

deve essere casuale, e rappresentativo

p. 287

%

col computere le grandezze sono solitamente discretizzate (binning)

→ funzioni di densità empiriche → essendo discretizzate, sono tecnicamente rappresentate come masse (non come densità)_

- è utile pensare alle masse di probabilità come a stime delle densità di probabilità_ Es.
  - ho viaggiato in Antartide x pesare pinguini
  - ero dotato di una bilancia ad altissima risoluzione (misura da 10 a 16 Kg)
  - voglio calcolare la funzione di probabilità empirica che descrive i pesi dei pinguini
  - discretizzo in contenitori di larghezza 0,25 Kg ⇒

⇒ dal mondo reale sono passato a valori numerici discreti

- l'analogo matematico è che non ha senso calcolare l'integrale in un punto_ Necessito di un intervallo_ corrisponde ai bins dell'istogramma_ ⇒

⇒ le funzioni di probabilità sono continue

```
penguin = np.arctanh(np.random.uniform(size=473)*1.8-.9)*2+4.5
bin_edge = np.arange(np.min(penguin), np.max(penguin), step=0.25
```

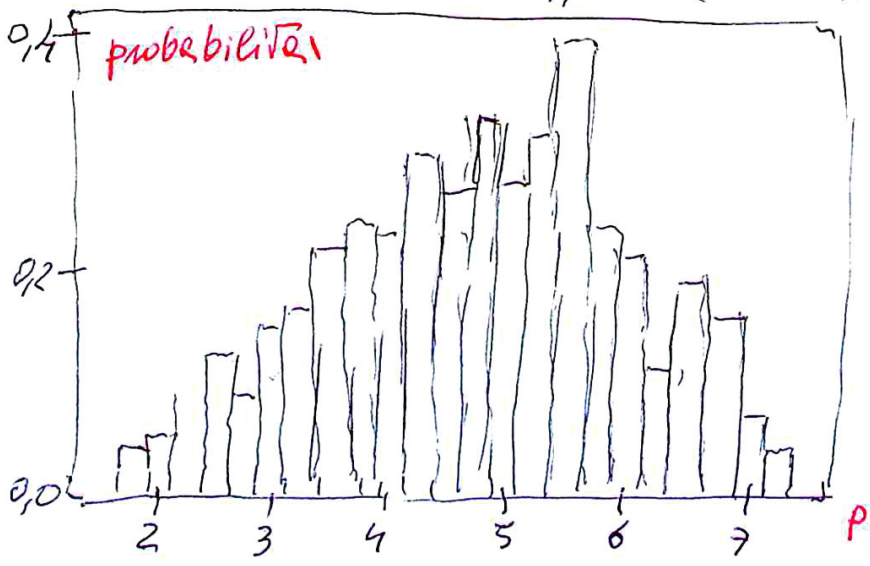


Fig. 8.5

8.6 - Cumulative distribution function (cdf)

Sono la somma cumulativa della massa di probabilità o della funzione di densità (qui invece di $\Sigma \to \int$)

- Ogni pto della cdf dice la probabilità di avere un valore minore o uguale al suo valore. In formule

$$F(x) = p(X \leq x) \qquad (8,10)$$

← prob. di essere $\leq x$

$$f(x) = p(X = x) \qquad (8,5)$$

← prob. di un valore specifico

- Notare la sottile distinzione tra le due

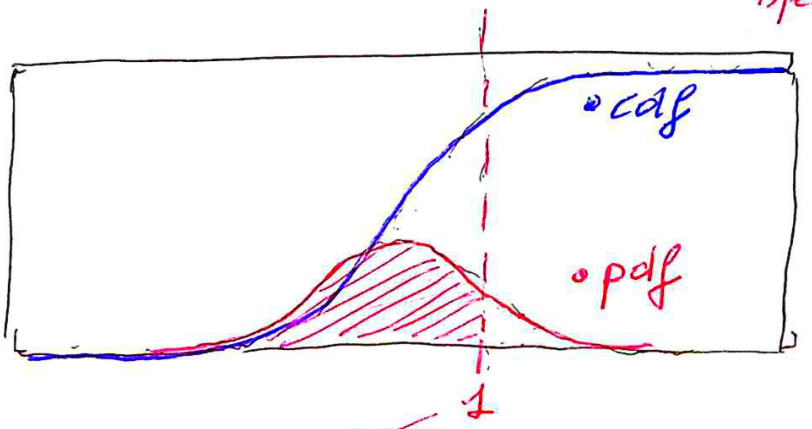


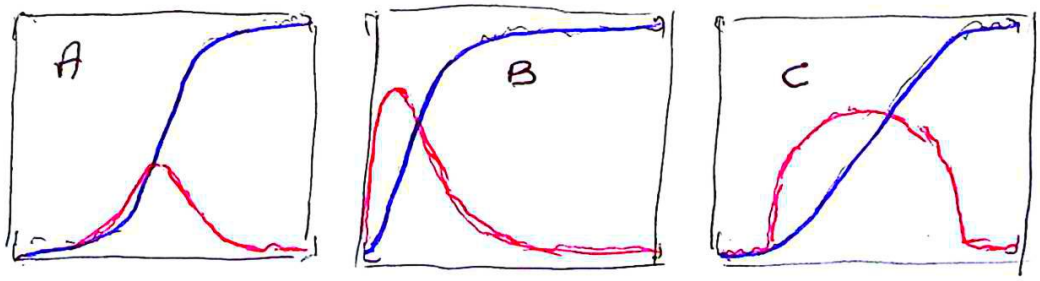
Fig. 8.6

il valore della cdf x $x = 1$ = area a sinistra della pdf

Nella Fig. 8.7 alcuni esempi di pdf e corrispondenti cdf

↳ cdf = somma cumulativa → sempre crescente

Fig. 8.7



- cdf è utile x calcolare un valore p mentre

↳ interpretabile solo x dati che fanno capo a ordinati numerici in modo significativo

8,7 – Expected value

- Valore atteso, come si ottiene, come lo si interpreta, come lo si relaziona con la media e la varianza?

$$\bar{X} = \sum_{i=1}^{n} \frac{x_i}{n} \qquad (8,11)$$

$$E[X] = \sum_{i=1}^{n} x_i p_i \qquad (8,12)$$

p_i → probabilità di x_i

Valore atteso → è una media pondereta

Se $p = \frac{1}{n}$ → le due equazioni sono identiche

- il valore atteso è analitico – calcolo teorico di una distribuzione
- la media è empirica – statistica descrittiva empirica di un campione di dati

Es. dado "pesato"

faccia	prob
1	1/4
2	1/4
3	1/8
4	1/8
5	1/8
6	1/8

Fig. 8,8

$$E(x) = \frac{1}{4} + \frac{2}{4} + \frac{3}{8} + \frac{4}{8} + \frac{5}{8} + \frac{6}{8} = 3$$

x un dado non pesato

$E(x) = 3,5$

- calcoliamo il valore medio facendo lanci, supponiamo di aver ottenuto:

$$x = [1,3,4,4,4,3,2,5] \rightarrow \bar{x} = 3,25$$

sappiamo che non coincide con $E(x)$ — questa è la differenza tra i due

Continuing expected value – nel caso di distrib. uniforme

$$E(x) = \int_a^b \frac{x}{b-a}\,dx \longrightarrow \frac{(b+a)}{2}$$

8,13 → 8,18

Stima dei valori attesi

Molte distribuzioni di dato e pdf sono sconosciuti →

- per avere $E[\,]$ possiamo esaminare l'integrando, fare una ipotesi con la somiglianza a una distribuzione nota → calcolare il valore atteso
- es. molte distribuzioni empiriche sono ≈ Gauss possibile
- per avere stime empiriche sicure, e.g., $N \to \infty$, in pratica non

→ ∀ campione con N finito → una semplice stima

Valori attesi e momenti statistici

$E[X]$ = 1st moment

$E[X^2]$ = 2nd "

$E[X^k]$ = k-mo momento

possiamo calcolare la varianza attesa, se il suo 2nd moment è centrato sulla media →

→ varianza = 2nd mean-centered moment:

$$\sigma_X^2 = E\left[(X - E[X])^2\right] \qquad (8.19)$$

$$= E[X^2] - (E[X])^2 \qquad (8.20)$$

- Esaminiamo il caso della distrib. uniforme

con la 8.18 $E(x) = \int_a^b \frac{x}{b-a}\,dx$ ⟹

poco cambia x^2

$$E[X^2] = \int_a^b \frac{x^2}{b-a}\,dx = \frac{1}{b-a}\,\frac{b^3 - a^3}{3}$$

- combiniamo i due momenti x ottenere la varianza (8.18)

$$\sigma_X = \frac{b^3 - a^3}{3(b-a)} - \left(\frac{b+a}{2}\right)^2 \to \text{calcolando} \to \frac{b^2 - a^2}{12} \qquad (8.28)$$

ripetiamo $\sigma_x = \frac{b^2 - a^2}{12}$ distrib. uniforme (8,28)

- **Softmax** = trasformazione matematica,

converte un insieme di numeri in una mappa di probabilità

Layer di apprendimento automatico

Layer che è direttamente una probabilità

Layer fornisce una distribuzione di probabilità che codifica previsioni utili x prendere una decisione

$$\sigma(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}} \qquad (8,29)$$

Softmax

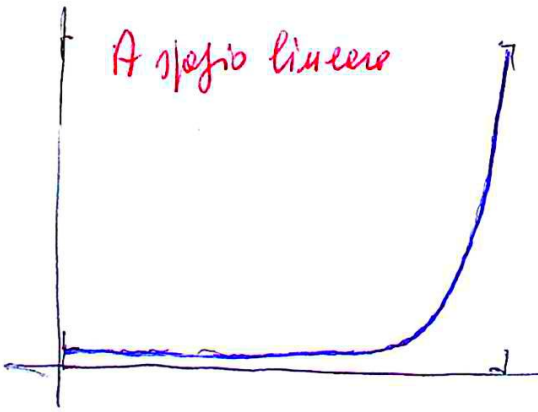
non confondere con deviazione std $= \sigma_x$ pedice

Es. numerico

Raw	Softmax
4	0,042
5	0,114
7	0,844

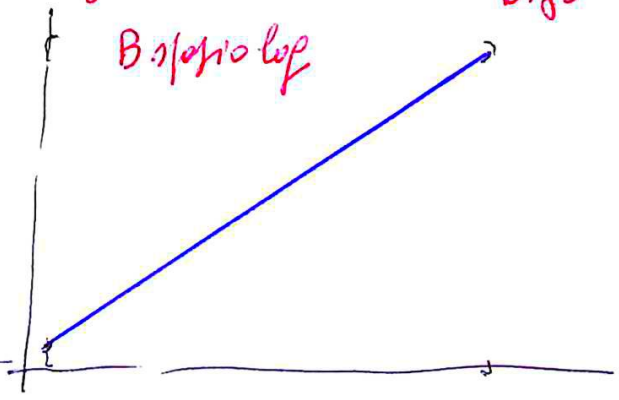
prendendo un intervallo + ampio, graficando: Fig. 8,9

Softmax



Softmax

Fig. 8,10



$\sigma(x)$ produce veramente una funzione di massa?

Sì perché soddisfa i requisiti di una distribuzione di probabilità:

- eventi reciprocamente esclusivi
- $\sigma(x_i)$ / $0 < \sigma(x_i) < 1$
- $\sum \sigma(x_i) = 1$ ——————— (8,30)

A cosa serve softmax?

- classificazione in cui bisogna apprendimento automatico

Es. foto cane-gatto

L trasforma il set di pixel dell'immagine in una label

[cane, gatto]

questi valori sono trasformati in una serie di probabilità' mediante l'uso di softmax

L in base ai valori di probabilità' → si seleziona etichetta

Es. predire se: no Tumore, Tumore benigno, Tumore maligno

- per un particolare paziente ..... una tabella simile a Fig 8.9 — viene processata con softmax

4	0,042	— 4% no tumore
5	0,114	— 11% benigno
7	0,844	— 84% maligno

%

8.9 - Exercises

1) esploriamo le unità dei pdf offerti da "scipy"
— distrib. Gauss

— creiamo pdf a partire da Gauss

```
x4 = np.linspace(-4,4,400)
pdf4 = stats.norm.pdf(x4)
```

normalizzo

```
pdf4N = pdf4 * (x4[1] - x4[0])
np.sum(pdf4N)  ------ = 1
```

2) come prima, ma differente intervallo

linspace(-2,2,300)

[rifatto, verifico che = 1

— in realtà ottengo .955

Provo una diversa normalizzazione

```
pdf2U = pdf2 / np.sum(pdf2)
np.sum(pdf2U)  ------ = 1
```

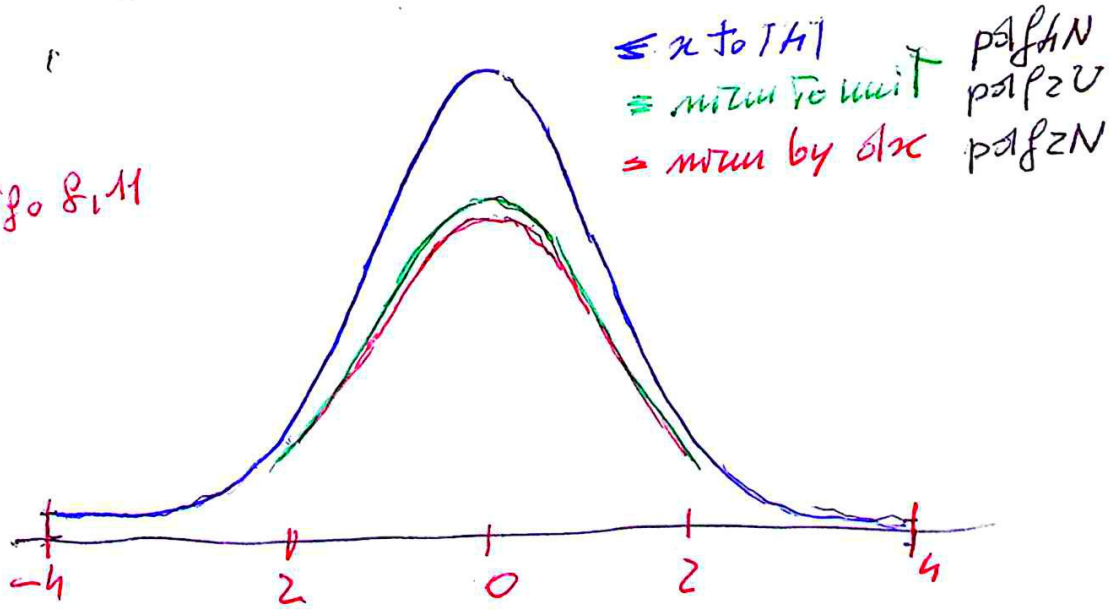


Fig. 8.11

– le tre distribuzioni sono tutte corrette!
L semplicemente abbiamo diviso x numeri diversi

3) due pdf, entrambi $-4 \div 4$ ⟨ 100 punti / 1000 punti

uso un ciclo for (2 steps) x grafico

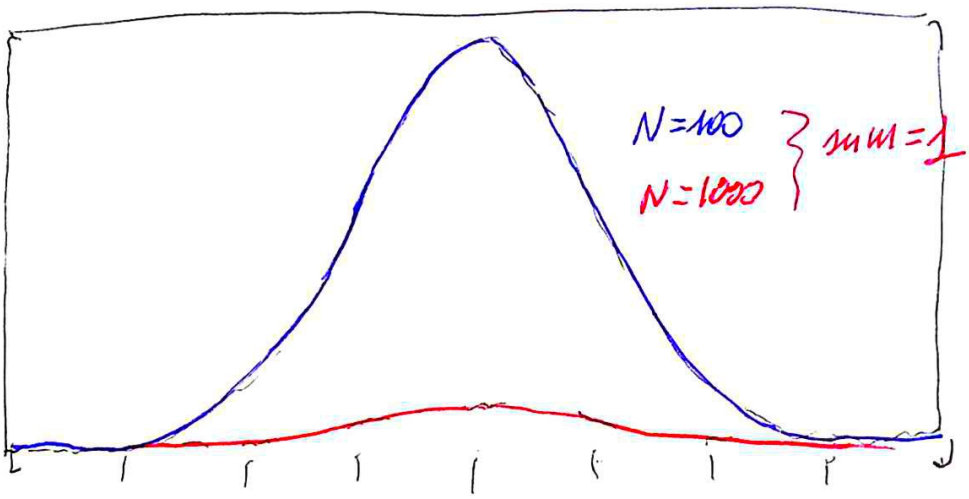


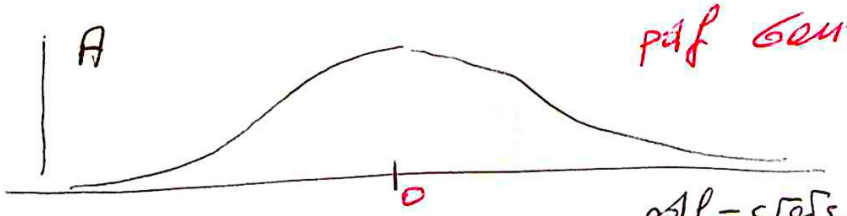
Fig. 8.12

4) partiamo a "cdf"

tra -4 e 4, 300 steps – creiamo pdf

– facciamo la somma cumulativa (1)

– ——————— sulla normalizzata (2)



pdf Gaus (raw)

Fig. 8.13

pdf = stats.norm.pdf(x)

pdf_sp = stats.norm.cdf(x)

calcolo manuale

cdf_my = np.cumsum(pdf) ——————— B

cdf_myN = np.cumsum(pdf) * (x[1] − x[0]) —— C

grafico:

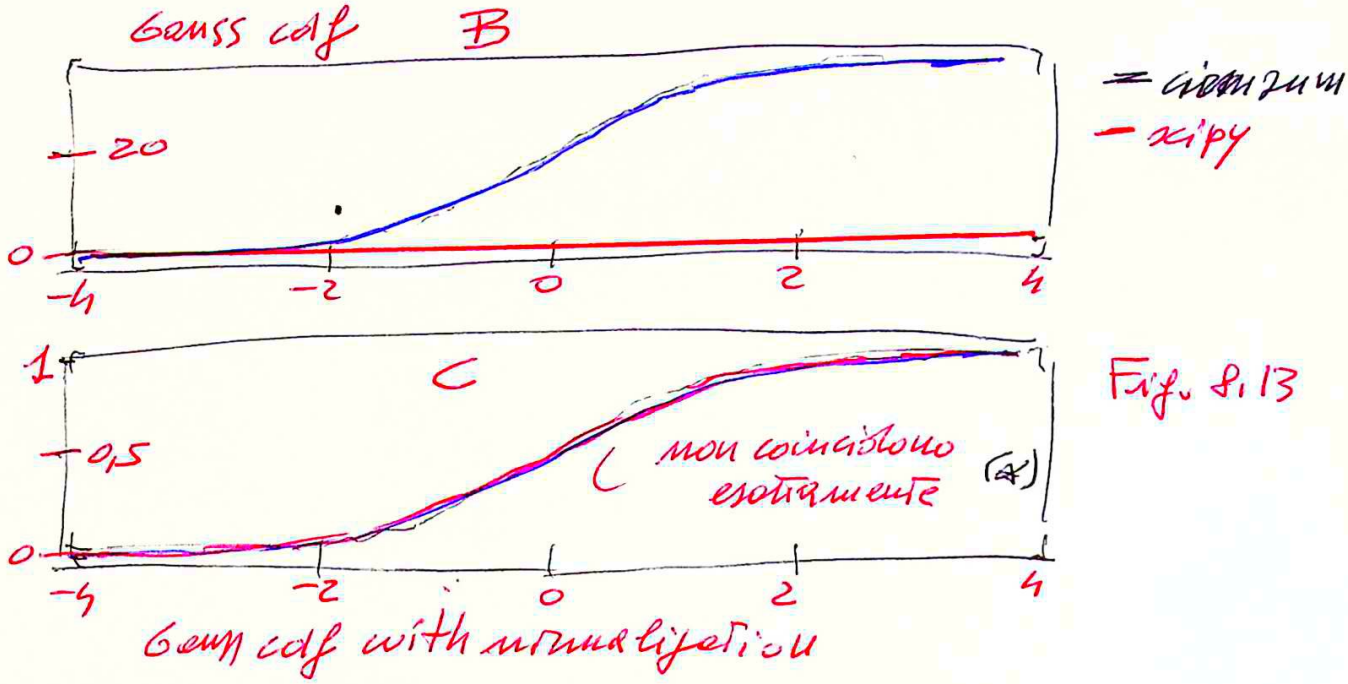


Fig. 8.13

5) per capire meglio (*) usiamo una risoluzione minore (20 punti) — ricostruiamo il grafico C —

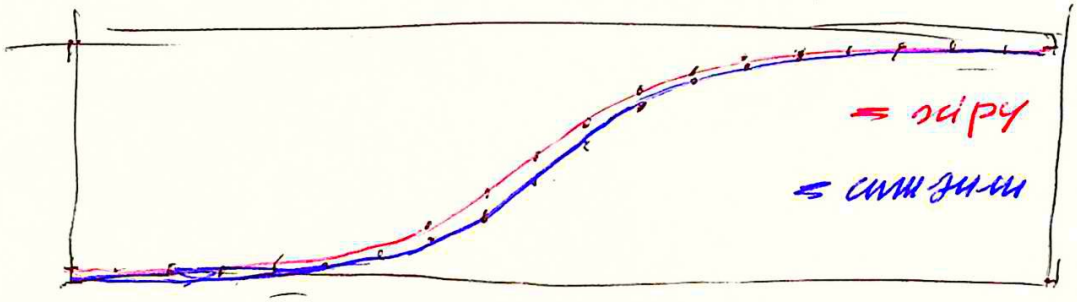


Fig. 8.14

la verità è che "scipy" è + accurata — Le differenze sono tanto maggiori quanto maggiore è Δx

6) Calcoliamo cdf empirica non usando PyThon

```
x = np.linspace(-6,6,1001)
pdf = stats.norm.pdf(x-2.7) + stats.norm.pdf(x+2.7)
```

normalizzo

```
cdf = np.cumsum(pdf) * np.mean(np.diff(x))
```

migliore normalizzazione

```
pdfN = pdf / np.sum(pdf)
cdfN = np.cumsum(pdfN)
```

grafico

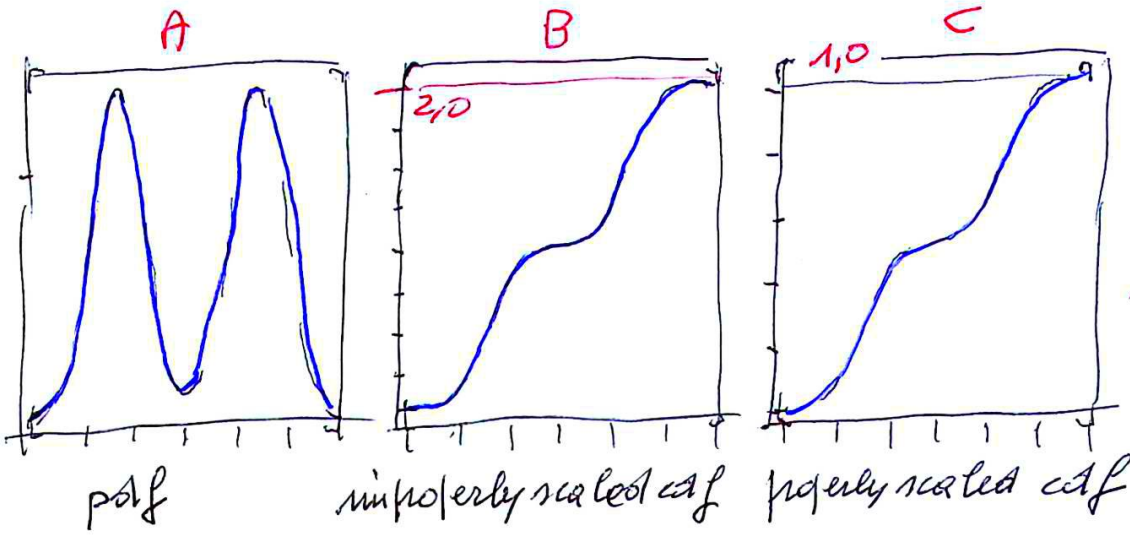


Fig. 8.15

Avendo sommato 2 pdf, il risultato non è un vero pdf, anche se scalato con dx –

– meglio quindi preventivamente calcolare pdfN, e solo dopo calcolare cdfN

7) ancora cdfs – TEST: pdf può essere calcolata come derivata di cdfs –

```
x = np.linspace(0,10,200)
cdf = stats.lognorm.cdf(x,1,1/2)
pdfD = np.diff(cdf)          ——— da derivata
pdfA = stats.lognorm.pdf(x,1,1/2)   ] analitica
pdfA *= x[1]-x[0]
```

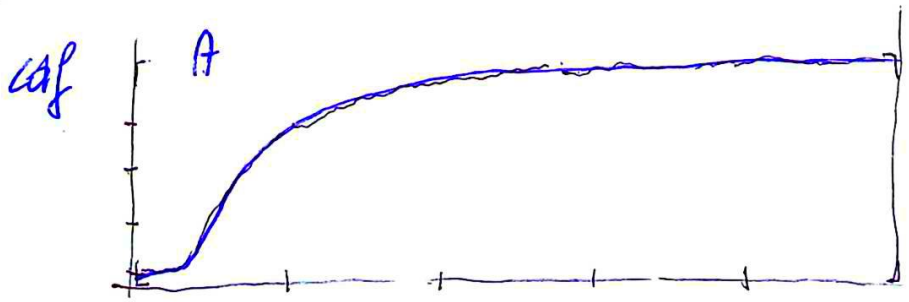
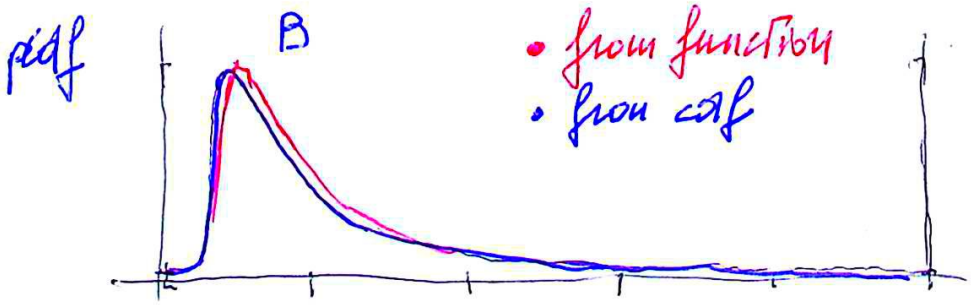


Fig. 8.16



%

8) differenza tra probabilità empiriche e analitiche -

40 biglie blu / 30 gialle / 20 arancioni / $\Sigma = 90$

- creiamo un vettore di 90 elementi
  40 elem. = 1 / 30 elem = 2 / 20 elem = 3
- scegliamo con estrazione casuale una biglia nell'array
- ripetiamo l'estrazione 500 (con sostituzione)
  ↳ registriamo i totali x ogni colore
- convertiamo in proporzione
- tracciamo la proporzione empirica e quella analitica

Fig. 8.17

9) nel precedente probabilità e proporzioni mostrano valori simili - Aumentando i campioni convergerebbero (legge dei grandi numeri) -
Vediamo di misurare con successo variando n° campioni
↳ x quantificare questo usiamo la formula della root mean square — una misura fisica di concordanza in statistica. p = probab. analitica / $\hat{p}$ = probab. empirica

$$RMS = \sqrt{\frac{1}{c}\sum_{i=1}^{c}(p_i - \hat{p}_i)^2} \qquad (8.31)$$

c = n° di colori → $c = 3$

RMS = 0 concordanza perfetta

- prendiamo il codice dell'esercizio precedente e lo mettiamo in un loop "for", variando la dimensione del campione da 20 a 2000 in passi di 10
- memorizzare ogni RMS x ogni dimensione del campione
- graficare

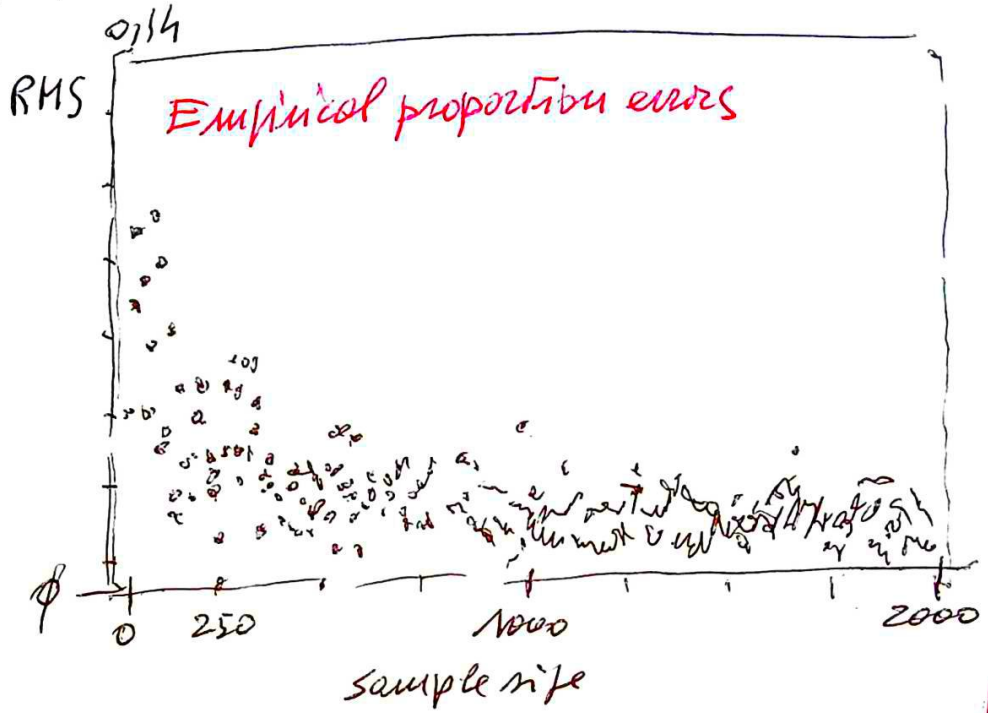


Fig. 8.18

- prova con 20 ÷ 20'000 a passi di 100
- aumentare il n° di bifle di un fattore 10

legge grandi numeri

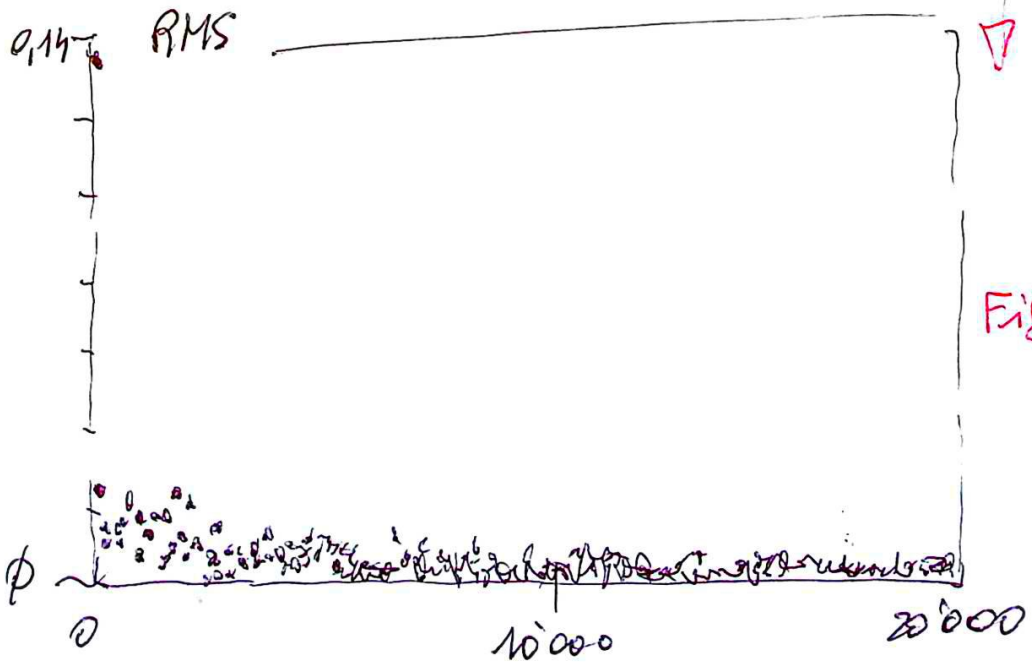


Fig. 8.18-bis

10) calcolare funzioni di proporzione cumulative empiriche, utilizzando dati simulati -

- copiamo i valori "p" e sue code — Capo 10

N = 1000 casuali gaussiani

creare vettore di 41 numeri separati linearmente, tra −3, 3 → serve usare come "cut-off" boundaries

chiamiamo ad esso con la lettera greca zeta ζ

- essendo Gauss) → la maggior parte dei valori sarà sopra ζ = −3 // la metà dei dati sopra ζ = 0 // la maggior parte dei dati ζ < 3
- facciamo un ciclo sopra ogni valore di ζ (1) calcolando la proporzione di dati sotto ζ
- calcoliamo la proporzione di dati sopra ζ (2)
- Visualizzare Figg 8/9

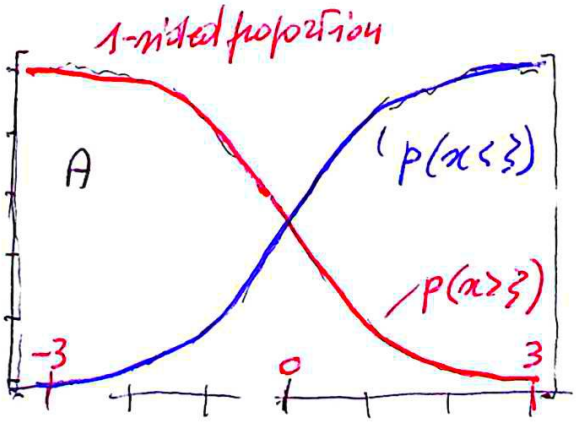
N = 1000
K = 41 — bins
data = np.random.randn(N)
bounds = np.linspace(−3, 3, K)

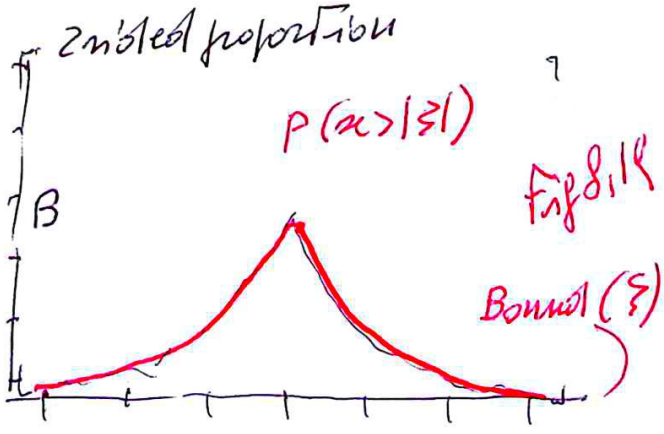
for i, bi in enumerate(bounds):

emp props GT = np.sum(data > bi) / N

~~LT =~~ ~~(data < bi) / N~~

emp prop 2tail [bi] = np.sum(data > np.abs(bi)) / N





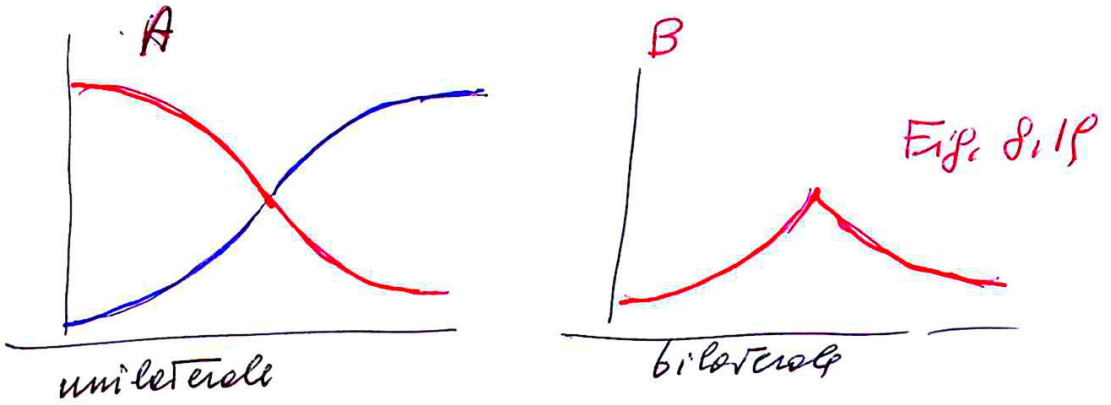


Fig. 8.16

ogni funzione di ripartizione unilaterale è esemplificativa di una cdf (o del suo opposto) (A)

ogni funzione di ripartizione bilaterale non sembra una cdf, ne sembra il pdf di una Gauss

11) vedere la 8.17

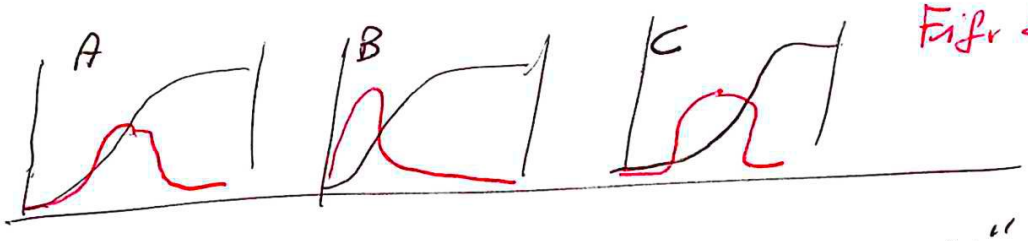


Fig. 8.17

usando distribuzioni prese dalla libreria di "scipy" e che offrono riferimenti. Ne troviamo un elenco qui

https://docs.scipy.org/doc/scipy/reference/stats.html#

continuous distributions

FINE CAPITOLO 8

9 - Sampling and distributions

9.1 - Sampling variability and its annoyance

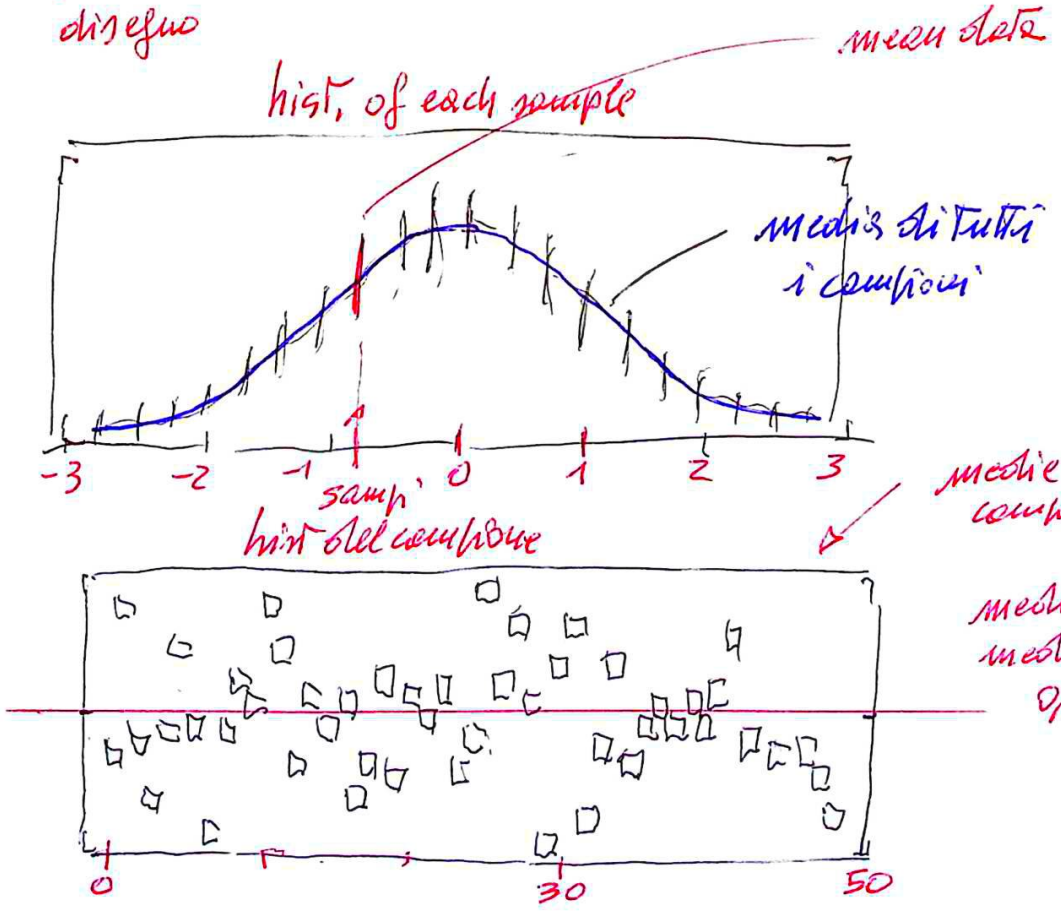
- Sei uno sviluppatore dei dati, una azienda di fitness vuole sapere se uomini o donne sono più propensi a fare esercizio in determinati momenti della giornata.
- le scelte qui sarebbero troppo personali, quindi non generalizzabili
- allora decidi di campionare 1000 uomini, 1000 donne... ma perdi i dati → ricampioni altri 1000+1000 ma su persone diverse
- ci possiamo aspettare che le due distribuzioni siano simili, ma non identiche
- quindi abbiamo una variabilità nei campioni (non è la variabilità all'interno di 1 campione)
  L questa distinzione ha impatto su statistiche inferenziali, su intervalli di confidenza   p. 310
- la differenza tra le due variabilità è la base di ANOVA.
- La variabilità del campionamento è fastidiosa x che riflette la variabilità a un livello diverso da quello che abbiamo considerato prima.
- la motivazione x misurare un campione - invece che un individuo - è che quest'ultimo potrebbe non essere rappresentativo della popolazione. Ma il campione ha una sua variabilità → non è garantito che sia un parametro affidabile

- perfetto ci fidiamo di più della media di un campione, che non di un singolo No-
- questo è vero ma... - vedi variabilità del campione - meno
- se i campioni sono grandi es. N=10'000, il problema si riduce)

Un esempio con dati casuali

N=50, ciascuno dei quali comprende 500 valori casuali

```
N=500
nSamples=50
kHistBins=20
edges = np.linspace(-3,3, kHistBins)   fisso x istog.
dichiaro le matrici
allHistY  (50,20)
allMeans  (50)
for sampi in range(nSamples):
    data = np.random.normal(loc=0, scale=1, size=N)
    faccio histo, calcolo la media  x,y
    disegno
```



A cosa è dovuta la variabilità del campione?

È utile capire x tentare di ridurre.

Natural variation

È comune nei dati biologici, nella neuroimmagine cellulare, nella psicologia, nei trattamenti medici, nelle pratiche culturali.

Es. persone diverse reagiscono in modo diverso a un stimolo →

(1) è probabile che campioni diversi abbiano caratteristiche diverse

(2) è ancora + probabile che campioni di età, sesso, stato economico, background culturali diversi, siano diversi

Nota Variabilità anche in sistemi non biologici:
magnitudo di terremoti, n° di stelle in una galassia, altezza degli edifici in una città.

Rumore di misurazione i sensori sono imperfetti

Dinamiche e cambiamenti la natura è dinamica - stesse misure in tempi diversi possono variare Es. cambiamenti geopolitici

Sistemi complessi la maggior parte degli oggetti di indagine scientifica sono complessi Es. con retroazioni

Stocasticità (casualità) c'è casualità nell'universo che non possiamo misurare, che non comprendiamo Es. moto browniano, eventi

In concreto, limitare la variabilità, senza compromettere la generalità è un processo complicato.

9.2 - Creating sample estimate distributions

Es. K campioni indipendenti, presi da una popolazione

S = i singoli campioni nella loro interezza

S_i = i-mo campione

- ciascun campione ha il suo valore medio = $\bar{S}_i$ →

K campioni producono K medie - Fig. 9,2 A

ciascun S_k ha la sua distrib -

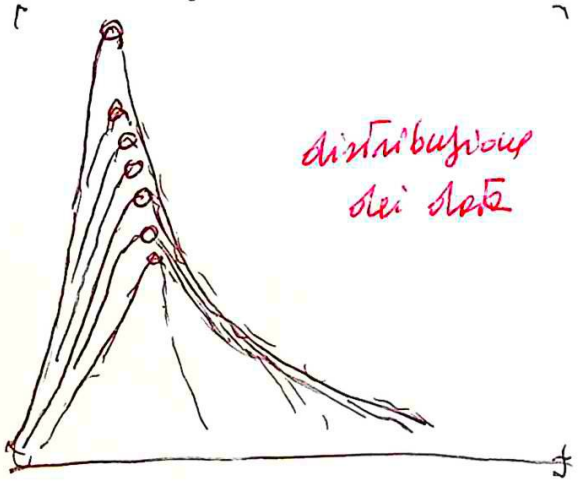
A) Data & means

data points

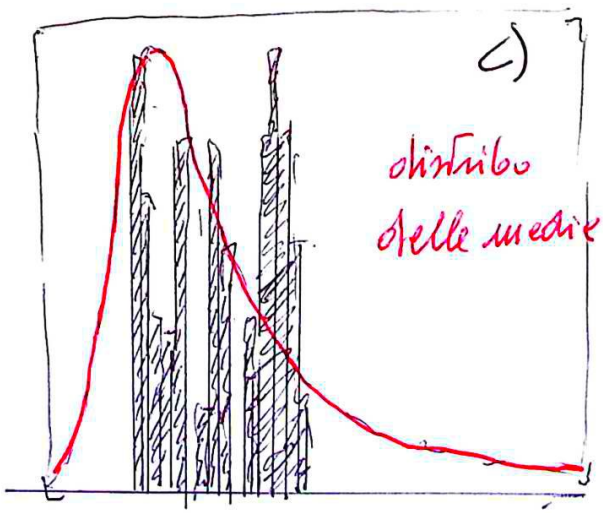
S_1	S_2	---	S_K
o	o		o
o	o		o
o	⋮		⋮
⋮	o		⋮
⋮			o
o			
⊗	⊗	--	⊗

medie

B)



Le medie dei campioni sono una collezione di K valori, che producono un loro istogramma. Questo non è calcolato dei "raw data", ma viene dalla statistica descrittiva che abbiamo calcolato -



Importante la distribuzione delle medie da K campioni non è la stessa della media degli info. campione (vedi Teorema del limite centrale)

Perché le distribuzioni di stima del campione sono importanti?

- ene rivelano la variabilità intrinseca della popolazione
- ene consentono di calcolare gli intervalli di confidenza

Si possono calcolare le distribuzioni di stima del campione x ogni statistica descrittiva (es. std, mediana, coeff. di correlazione ---)

9,3 - Standard error of the mean

- la "sample mean" è una stima della media di una popolazione

↳ se faccio tre "sample mean" ottengo valori diversi → statistica descrittiva

Quanto è precisa la mia stima?

lo dice l'"errore std della media" = SEM

SEM = quantità di variabilità che ci si può aspettare nelle ("medie di campioni casuali" riferiti con la stessa dimensione, ricavate dalla popolazione.

- viene anche detto deviazione std della popolazione

in formule

$$SEM = \frac{\sigma}{\sqrt{N}} \qquad (9,1)$$

σ = deviazione std della popolazione
N = campione

- in concreto però non conosciamo σ, al suo posto mettiamo la std del campione

$$SEM \approx \frac{s}{\sqrt{N}} \qquad (9,2)$$

$$SEM \approx \frac{s}{\sqrt{N}} \qquad (9.2)$$

s = std del campione e' $\approx \sigma$ Se:

- il campione e' random
- ——— e' rappresentativo
- N e' sufficientemente grande

Std error of the mean vs. std deviation

Entrambi quantificano la variabilità associata a un set di dati del campione
Anche matematicamente sono simili, ma riflettono aspetti diversi dell'incertezza → interpretazioni distinte

Conceptual meaning

std deviation = dispersione dei valori attorno alla media del campione

SEM = precisione della media del campione, intesa come stima della media della popolazione
= quanto è probabile che s sia vicino a σ

Calculation

std = si ottiene utilizzando ogni dato in un campione

Applications

- come misura della variabilità in un "dato semplice"
- in trasformazioni di tipo z-scoring
- x generare intervalli di confidenza
- x valutare la significatività statistica delle statistiche di test (es. t-values, coeff. di regressione)

Impact of sample size

La deviazione std non varia con la dimensione del campione

SEM diminuisce all'aumentare della ——————,
ma std no.

9.4 - Random and representative sampling

Lo scopo delle statistiche inferenziali è valutare se le caratteristiche di un campione si generalizzeranno a una popolazione.

Campioni rappresentativi:

Es. Abbiamo dati su quanti soldi delle famiglie americane finiscono in cibo spazzatura.
Possiamo da qui prevedere la mappa delle stelle? No, il campione non è rappresentativo

Es. dati sulle opinioni degli studenti universitari sul reddito di base universale. Possiamo da qui inferire come la popolazione concepisce UBI? No xché gli studenti univers. non rappresentano la popolazione.
Forse potrebbero esserci correlazioni con studenti univ. di altre nazioni.

- Consideriamo ancora quegli studenti.
Ma ci riferiamo a un compito di percezione visiva in cui bisognava abbinare due forme 3D mostrate. Lo possiamo generalizzare? Potrebbe essere ragionevolmente sì.

%

Campioni casuali

È una delle modalità x avere campioni rappresentativi_
Riformiamo a UBI // supponiamo che gli intervistati siano tutti iscritti a "equità economica"_ Le loro opinioni saranno diverse da quelle di altri studenti_
Gli studenti iscritti a "Fare e capitale" anch'essi avranno idee diverse_

→ raccogliere dati casualmente su tutta la popolazione universitaria
→ dati raccolti casualmente da tutta la popolazione sotto studio

Attenzione offri denaro a chi ti lascia intervistare! questo crea una selezione di campo_
offri crediti formativi? // stai chiedendo prima al supermercato? probabilmente selezioni femminucci o senza lavoro // compili numeri telefonici? comunque un sotto insieme

Raccogliere campioni casuali è cosa complicata_

Se possibile confrontare le statistiche descrittive del campione con caratteristiche note della popolazione_ Es. età, equilibrio di genere, anni di istruzione del tuo campione con quelli ISTAT_
Se qui c'è corrispondenza → + fiducia che i dati possano essere generalizzati_

Buona notizia se i campioni individuali hanno un bias, ma differenti campioni hanno un bias diverso — mediando le medie — in qualche misura si compensano (legge grandi numeri)_

Independent and identically distributed data

IID (iid, i.i.d.) è un termine tecnico usato nella matematica stats.

IID = fornisce una definizione più approfondita di "random part" di dati casuali e rappresentativi. Secondo IID:

- Indipendenti, cioè "scorrelati" tra loro Es.
  - campionare in base all'ultima cifra del n° di previdenza sociale garantisce indipendenza
  - campionare nello stesso nucleo familiare = dipendenza
- Identica distribuzione nei dati empirici questo è comunque approssimativo

9.5 - Law of Large Numbers

Il campionamento è una necessità pratica. È ragionevole che campioni più grandi riflettano + accuratamente la popolazione.

$$LLN = \lim_{n \to \infty} P(|\bar{x}_n - \mu| > \varepsilon) = \phi \qquad (9,3)$$

sample size / sample mean / true population mean

LLN and sample size (LLN demo #1)

Facciamo una dimostrazione usando i dati simulati, con che possiamo riprodurre tutta la popolazione, quindi conoscere la vera media

- ripetiamo [1, 2, 3, 4] molte volte → 4'194'304 data points

→ media esatta = 2,5

– prendo campioni casuali da 1 a 1500 Fig. 9,3

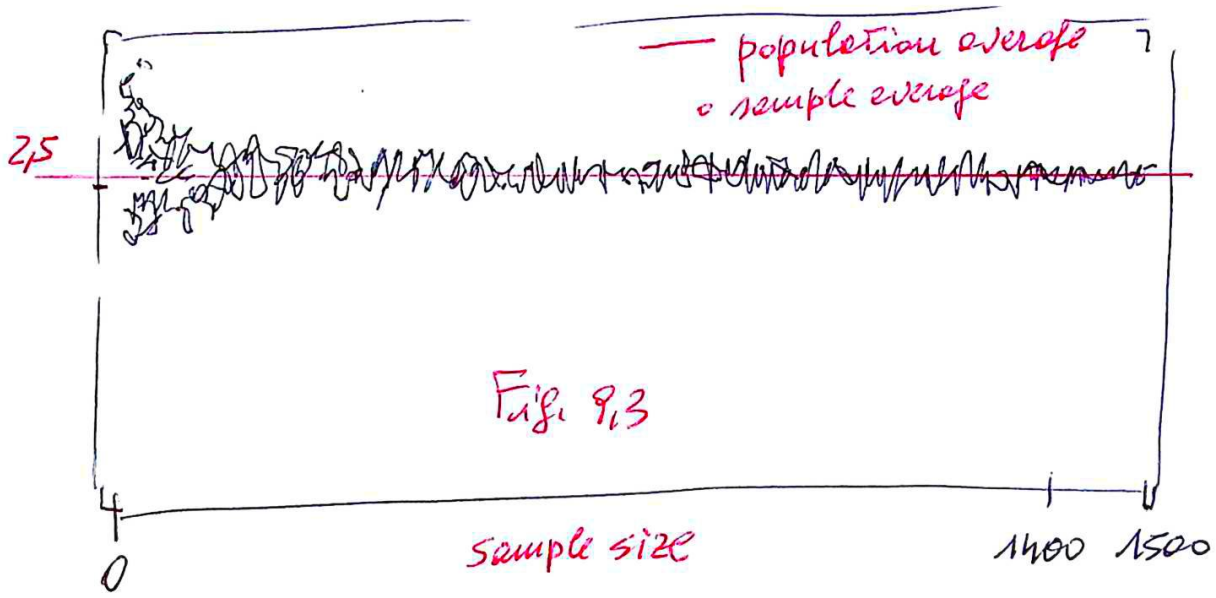


Fig. 9,3

(1) le medie del campione non sono esattamente = μ – sembrano vicine

(2) sembra che non ci siano bias

(3) la la variabilità diminuisce, aumentando le dimensioni

(4) anche a 1500 non sono esattamente uguali

LLN and repeated samples (LLN demo #2)

L'idea è che ogni campione (ogni esperimento) è sensibile alla variabilità del campione (rumore, altre variazioni non sistematiche) → È improbabile che un campione possa fornire una buona stima di μ. → consideriamo la media delle medie dei campioni.

Per mostrare la forza di questa procedura, facciamo questa simulazione: Dalla popolazione precedente prendo 50 campionamenti, ciascuno di 30 elementi. Calcolo la "cumulative average" Fig. 9,4 B.

I campioni vanno da 1 → 1500

In un ciclo for memorizzo tutte le medie – Grafico medie e medie cumulative –

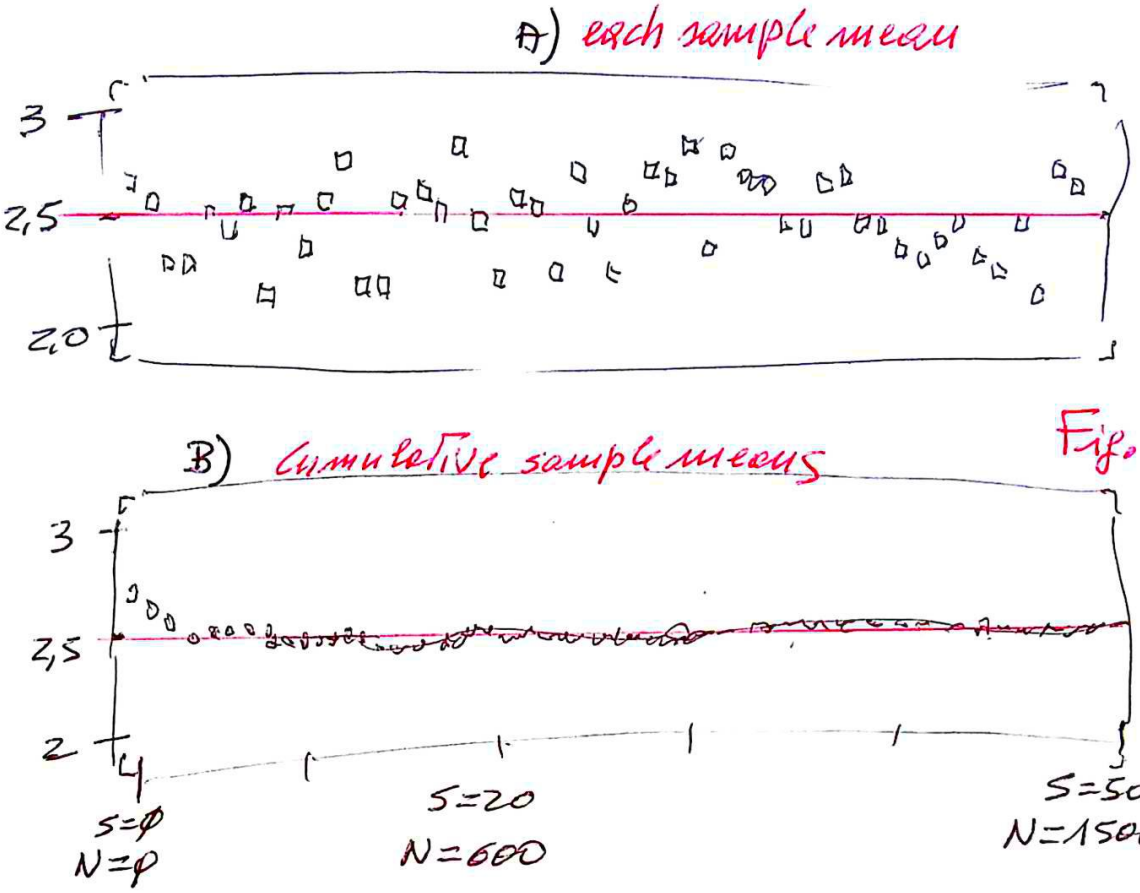


Fig. 9.4

(1) Le medie dei campioni individuali non hanno bias (✓)

(2) La media cumulativa s'è approssima a μ solo dopo un piccolo numero di campioni → i campioni individuali hanno una variabilità maggiore di quella delle loro medie

---

Consideriamo ora la Fig. 9.5 che rappresenta una popolazione di dati su due variabili – Da questi dati estraiamo due campionature, la rossa e la blu – Entrambe hanno la media fornita al centro della figura – Questa è una rappresentazione visiva del concetto (✓) –

Nella realtà i campioni sono sempre pochi (Tempi e costi) → combinare i dati di più studi individuali porta a una valutazione + accurata della realtà (meta-analisi) – E' particolarmente importante nella ricerca medica –

%

- i campioni dei diversi gruppi possono contenere distorsioni (bias) dovuti a campioni non rappresentativi, cattiva progettazione, strumenti tendenziosi ---
- Se le distorsioni non sono le stesse → LLN comunque fa una stima + accurata - Es.

Immaginiamo di raccogliere 6 campioni, ma ogni campione ha un bias - La media delle medie è + accurata -

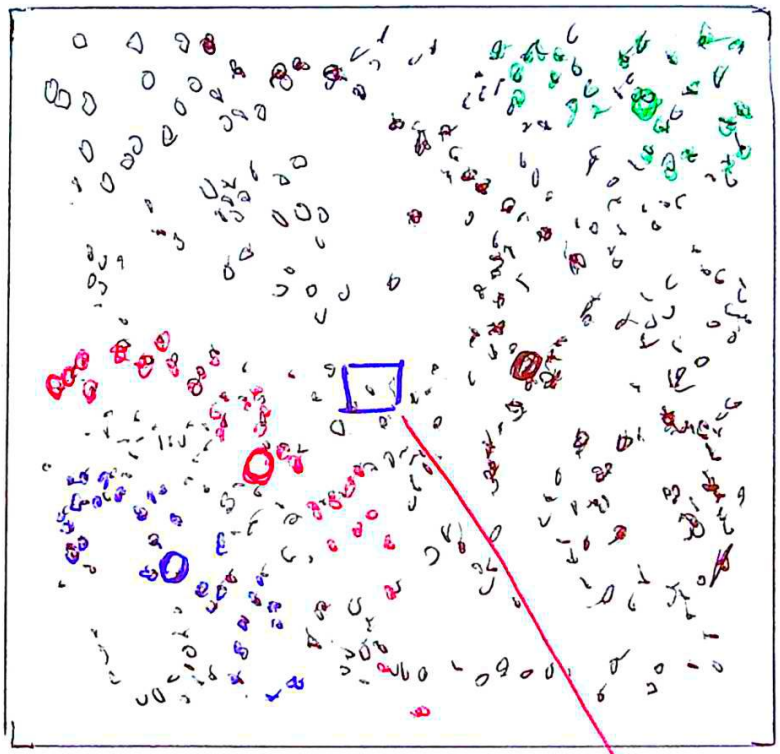


Fig. 9.6

9.6 - The Central Limit Theorem

CLT = una distribuzione di medie di campioni → Gaun anche se le distribuzioni originali non lo sono

CLT = aiuta a garantire la validità delle ipotesi che usiamo nelle statistiche inferenziali

CLT part 1: sampling distributions

Popolazione di interi (1÷6) multiplando il dado puderoso del Cap. 8 — 1,2 → p=1/4 // gli altri p=1/8

Fig. 9.7A mostra istogramma della popolazione

— contiamo 500 campioni casuali, ciascuno con N=30

Fig. 9.7B mostra istogramma delle medie (Gauss)

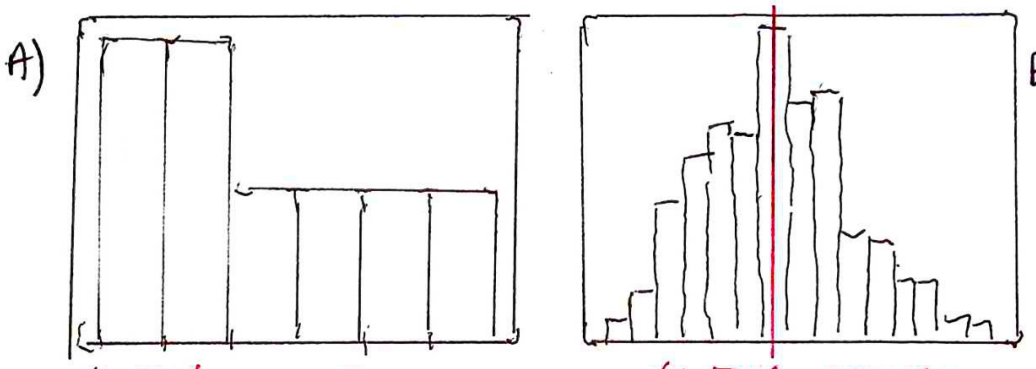


Fig. 9.7

Es.

Dadi con legge di potenza — Fig 9.8A — Estraggo 500 campioni con N=30

Faccio istogramma delle medie Fig 9.8B

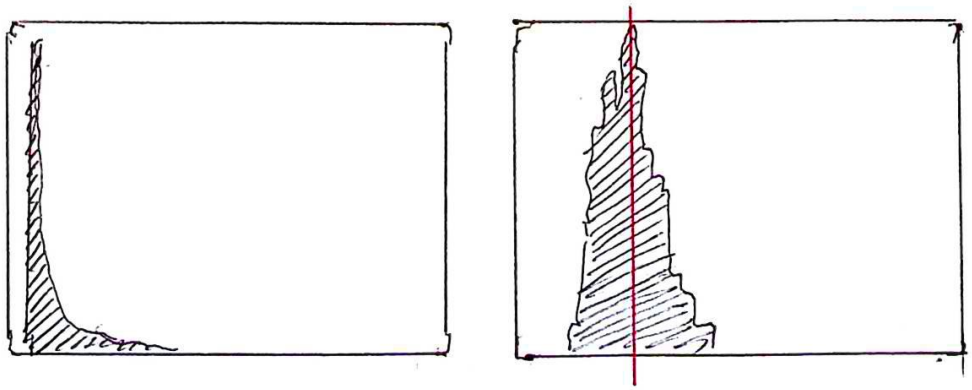


Fig. 9.8

LLN e CLT formano un duo potente

↳ miglioriamo la predizione, diventiamo Gauss

CLT part 2: mixing variables

CLT = miscele casuali di variabili → tende a CLT

L scegli cose casuali nell'universo e mettile in pila → vera' Gauss

Fig. 8.8 illustra questo concetto

- due variabili — un'onda sinusoidale Dati 1
- un rumore uniforme Dati 2 P. 329

i due set di dati sommati insieme hanno una distrib. ≈ Gauss

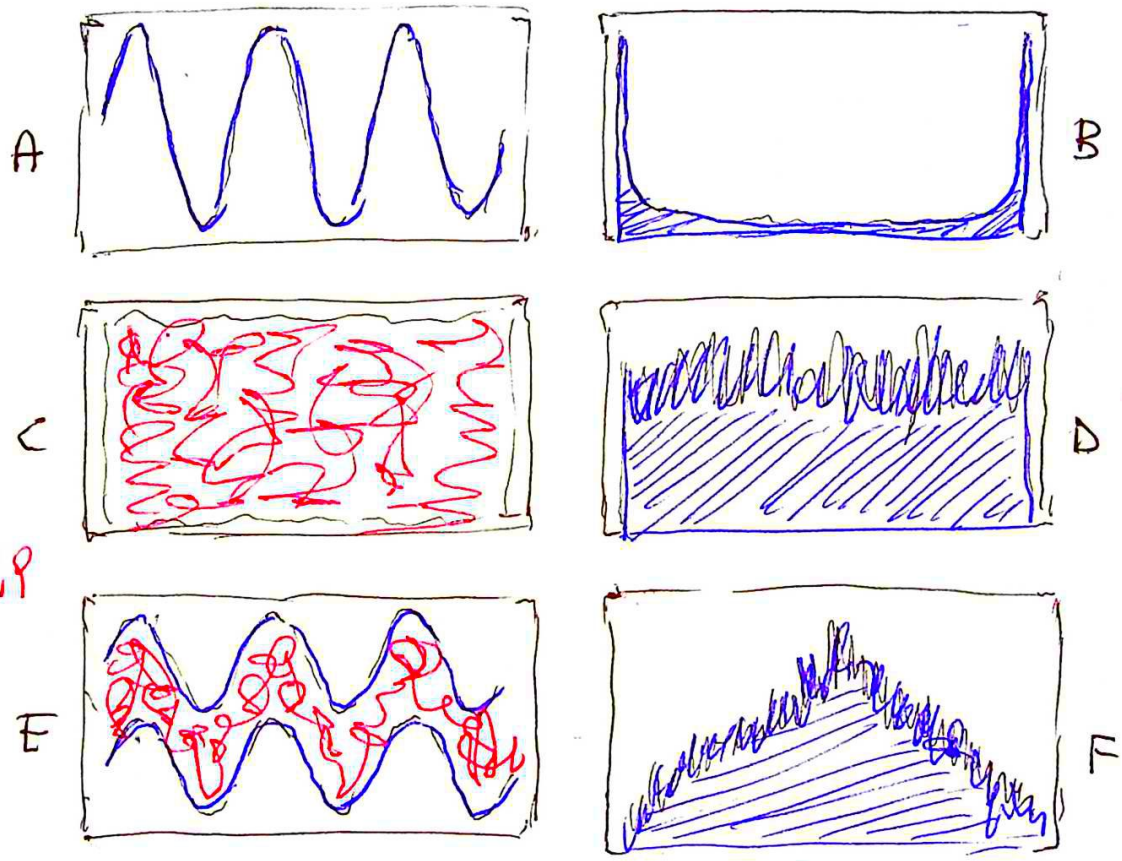
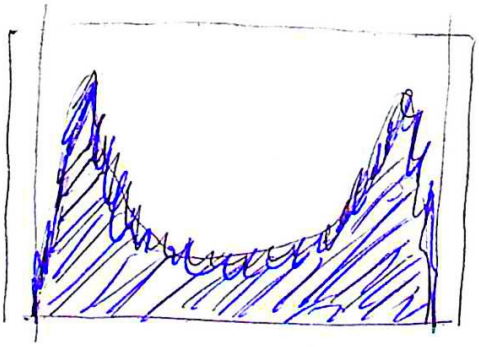


Fig. 8.8

≈ Gauss

Il risultato non e' banalmente gerarchico - Es. la similarità di scala delle variabili - Se ridisco il numero di un fattore 10 →



F' Fig. 9.10

The distributions of sample means

CLT = le medie dei campioni sono Gauss, media μ, std = SEM
in formule

$$\bar{x} \sim N\left(\mu, \frac{\sigma}{\sqrt{N}}\right) \qquad (9,4)$$

$\bar{x}$ = media del campione, μ = media della popolazione, σ = std della popolazione, N = numerosità del campione

se $N \to \infty \rightarrow SEM \to 0$

NOTA μ è indipendente dal valore di N

Implications of the CLT

- **Estimating population means from samples**

  CLT in combinazione con LLN dice che raggruppando i dati su molti campioni, otteniamo una buona stima della popolazione

- **Assumption of normal distribution**

  Molte inferenze statistiche si basano su Gauss. Ma nel mondo reale non sempre si ha Gauss. → medie dei campioni come base della inferenza statistica → intervalli di confidenza.

- **Deriving multivariate signals**

  con immagine e segnali, spesso i sorgenti sono mischiate. ∃ un metodo di analisi x demixare, chiamato metodo delle componenti principali indipendenti, basato sul presupposto che i segnali siano non-gaussiani.

9,7 - Exercises

1) Esplorazione della LLN

- manipoliamo il numero
- 200 punti estratti casualmente da Gauss $x \sim N(\phi, \tau^2)$

Varia da 0.1 a 10 in 10 steps

- iteriamo sui valori di τ^2
  - L calcoliamo, memorizziamo medi, std del campione
  - L x generare una distribuzione di numeri, estraiamo 20 volte x ogni τ^2

```
tau2 levels = np.linspace(0.1, 10, 10)
sample size = 200
num samples = 20
results = np.zeros((num samples, len(tau2 levels), 2))
```

loop over multi samples, tau

```
for mi, tau2 in enumerate( . . . ):
    for sampi in range(num samples):
        data = np.random.normal(0, np.sqrt(tau2), size = sample size)
        results [sampi, mi, 0] = np.mean(data)
        ~~1~~ = np.var(data, ddof = 1)
```

for grafico

```
stampo results / mean of results [:, :, 0]   ] 1° grafico
                 mean ~~of results [:, :, 1]~~
                 var of results [:, :, 0]    ] 2° grafico
```

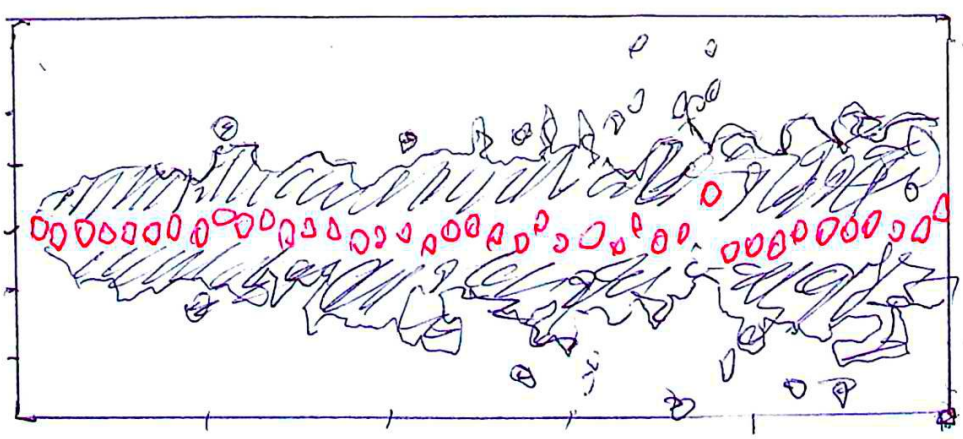
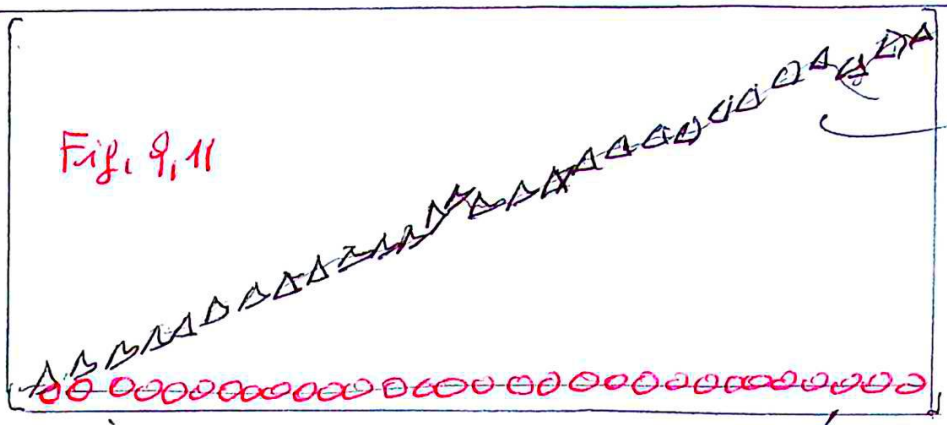


Fig. 9.11



2) Modifichiamo il codice della Fig. 9.3 (campioni e medie al variare di N)

- calcoliamo std (invece che le medie)
- usiamo random.seed con $\sigma = 2{,}4$
- LLN è formalmente definita x le medie, ma si applica anche ad altre statistiche descrittive

```
samplesizes = np.arange(10, 1001)
pop_std = 2.4
populationN = 1 000 000
population = np.random.random(populationN)        → std = 1
_____ = population / np.std(population, ddof=1)
_____ = _____ * pop_std  — fixo std
samplestds = np.zeros(len(samplesizes))
for sampi in range(len(samplesizes))
    pick a random sample
    sample = np.random.choice(population, size=samplesizes[sampi])
    samplestds[sampi] = np.std(sample, ddof=1)
plot
    samplestds
la riga viola è calcolata
```

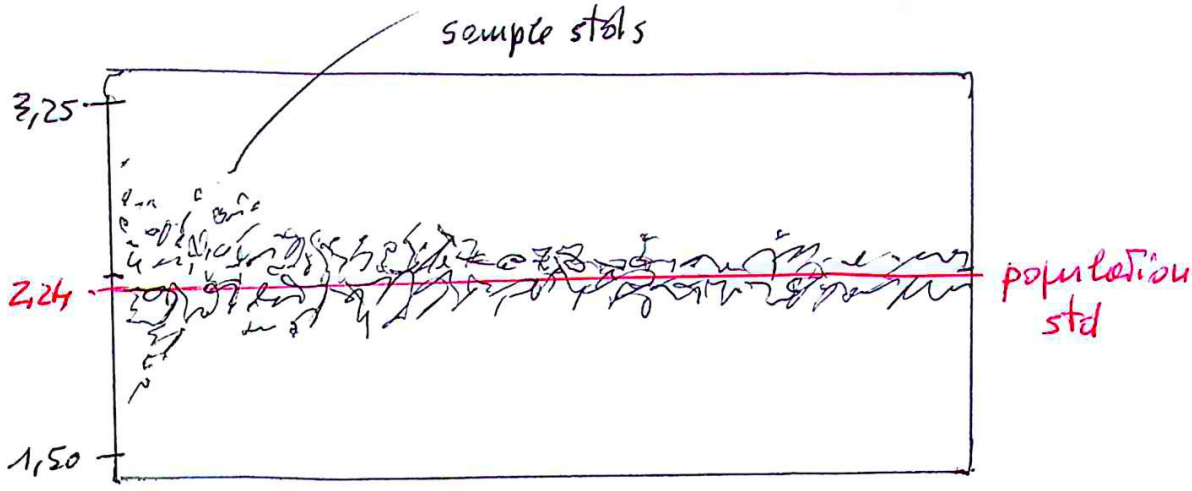


Fig. 9,12

3) Calcoliamo qui la differenza media tra due campioni, tratti da distribuzioni uguali. È un esempio che percorre le statistiche inferenziali (calcoleremo il numeratore di t-Test)

popMean1 = 3

———— 2 = 3,2

population1 = np.random.random(populationN) / = 10^6

———— = population1 − np.mean(population1) + popMean1

analogamente × population2

un campione

s1 = np.mean(np.random.choice(population1, size=30)

s2 = ———————————— 2

stampo la differenza popMean1 − popMean2 — −0200

s1 − s2 — −0,595

———

Ripetiamo la procedura

sample diffs = np.zeros(len(samplesize))

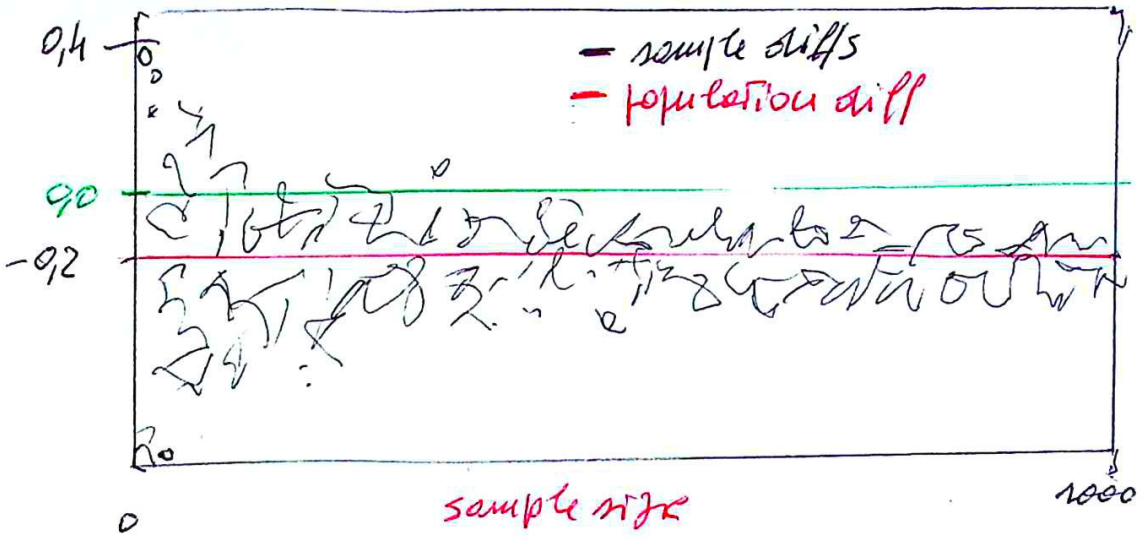
for sampi in range(len(samplesize)):

pick a random sample

s1 = np.random.choiche(population1, size = samplesizes[sampi])

s2 = ———————————— 2

sample diffs[sampi] = np.mean(s1) − np.mean(s2)



4) Scegliamo due interi casuali tra 0 e 100, calcoliamo la media
Ripetiamo 1200 volte – Ad ogni passo memorizziamo questi
Tre numeri – Facciamo istogrammi –

5) CLT funziona se i campioni sono abbastanza grandi –
Quando è "abbastanza grande"?
Riprendiamo il codice di Fig. 8.8 (dati con legge di potenza)
N=5 (con n° campioni = 500) poi
N=100 cosa osserviamo?

6) $N_{pop} = 10^6$ $\quad x^2$ / $x \sim \mathcal{N}(0,1)$

```
population = np.random.randn(Npop)**2
samplesizes = np.arange(5, 500, 8)
numberOfsamps = 1000
samplemeans = np.zeros(numberOfsamps)
fwhms = ——— (len(samplesizes))
peakvals = ———————————————
```

line colors

```
c = np.linspace((.9,.9,.9), (0,0,0), len(samplesizes))
for Ns in range(len(samplesizes)):
    calcolo la media di molti campioni
    for expi in range(numberOfsamps):
        samplemeans[expi] = np.mean(np.random.choice(
            population, size = samplesizes[Ns]
    fatto un istogramma x ogni media
    yy, xx = np.histogram(samplemeans, np.linspace(.4, 1.6, 41))
    yy = yy/np.sum(yy)
    calcolo FWHM
    ym = yy/np.max(yy)  normalizzo
    i0 = np.argmax(ym)  trovo peak index
    peakvals[Ns] = xx(i0)
    fwhms[Ns] = xx[i0-1+np.argmin(np.abs(ym[i0:] - .5))] -
      - xx[np.argmin(np.abs(ym[:,i0] - .5))]
    grafico xx, yy
```

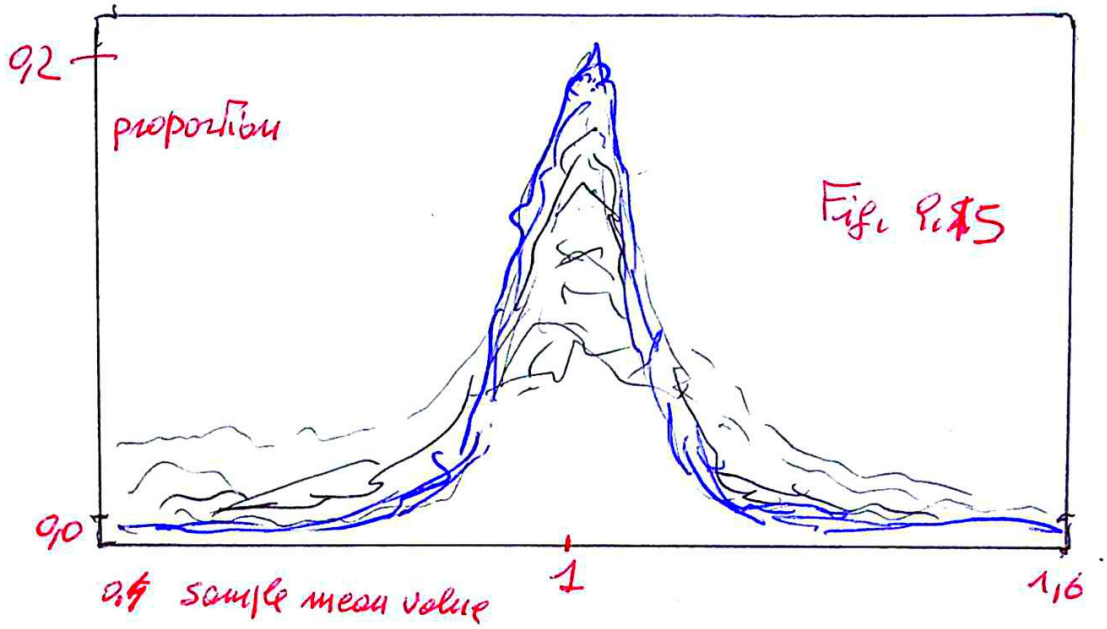


Fig. 9.15

- più spesse le linee → mostrare il n° di campioni (da 4 a 500)
- FWHMs empirici, e la distribuzione dei picchi possono essere visualizzati con grafici native -

```
ax s[0].plot(samplesizes, fwhms, 'ks', ———)
ax s[1].plot(———, peakvals, ———)
```

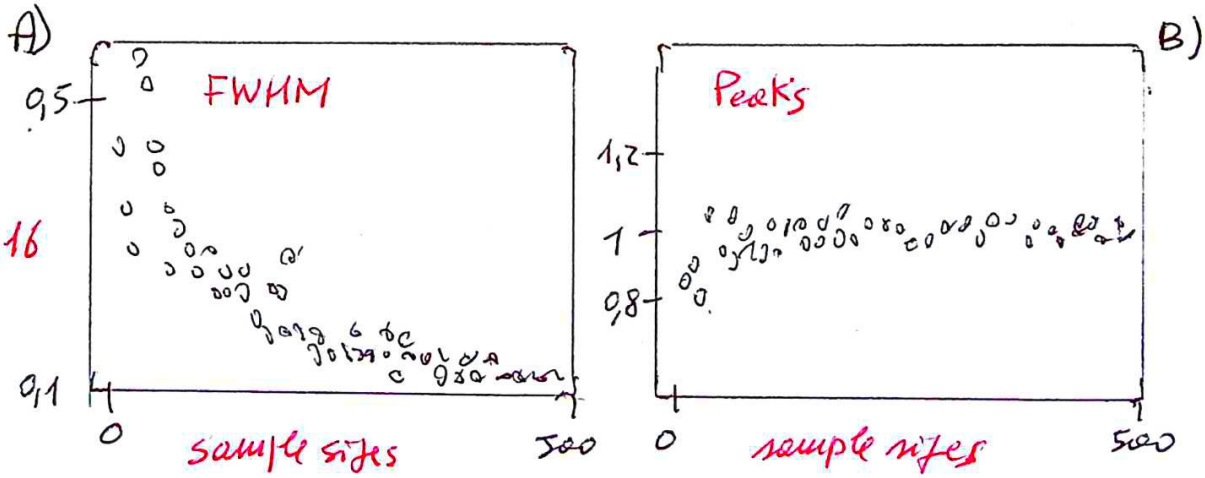


Fig. 9.16

7) confronto di SEM analitico con due stime empiriche

$$SEM = \frac{\sigma}{\sqrt{N}} \approx \frac{s}{\sqrt{N}}$$

calcoliamo il vero SEM x 25 dimensioni del campione, spaziate log tra $N=10$ e $N=10\%$ della popolazione.

Successivamente stimiamo SEM in due modi

(1) std di campione casuale / $\sqrt{\text{sua dimensione}}$

(2) std di più medie campionarie empiriche

↓ 50 campioni casuali x ogni size di campione, calcolare SEM del campione — calcolare media del campione

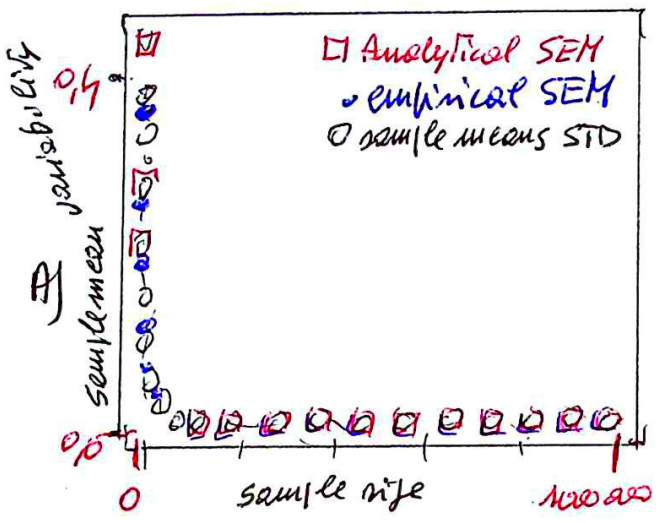
Grafico:

```
Npop = 10^6
population = np.random.random(Npop) * 2
samplesizes = np.logspace(np.log10(10), np.log10(Npop/10), 25, dtype=int)
numExps = 50
theory = np.std(population)/np.sqrt(samplesizes)
standerror = np.zeros((numExps, len(samplesizes)))
samplemeans = ————————————————————
for expi in range(numExps):
    for idx, ssize in enumerate(samplesizes):
        generate a random sample
        rsample = np.random.choice(population, size=ssize)
        compute std error and samplemean
        standerror[expi, idx] = np.std(rsample, ddof=1)/np.sqrt(ssize)
        samplemeans[expi, idx] = np.mean(rsample)
Plot axs[0].plot(samplesizes, theory ———— Analytical SEM
————————, np.std(samplemeans, axis=0, ddof=1)
———— Sample means STD
```

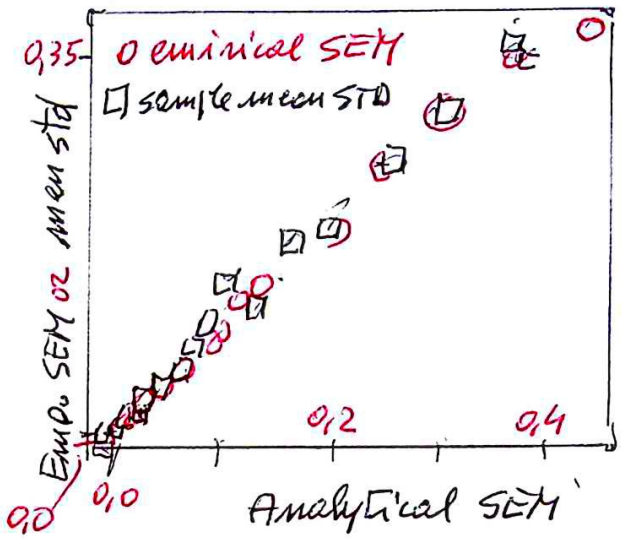
axs[1].plot(theory, np.mean(standerr, axis=0)) — empirical SEM

———— (np.std(samplemeans—), np.mean(standerr—))

sample mean std



A)



B)

Fig. 9,17

Estimates by sample size

Consistency of estimates

- La variabilità delle medie campionarie diminuisce drasticamente all'aumentare del campione Fig. 9,17A
- Le stime teoriche e empiriche di SEM corrispondono bene Fig. 9,17B
  I valori di SEM grandi → campioni piccoli

p. 339

— FINE CAPITOLO 9 —

10 - Hypothesis Testing

10.1 - Hypotheses

How to specify a hypothesis

Spesso si parte da una domanda → si pone una ipotesi

- Il caffè aiuta a studiare? → 1-2 tazze/die in esperimento
  - alcune ipotesi possono essere simulate o pensate matematicamente
- A noi interessano ipotesi che possono essere valutate con i dati

(Why) do we need hypotheses?

1) le ipotesi migliorano la progettazione degli esperimenti
2) da idee vaghe partiamo a idee concrete
3) aiutano a sviluppare nuove teorie
4) consentono una valutazione quantitativa
5) portano alla comprensione dei meccanismi sottostanti
   - forse la voce + importante

Strong and weak hypotheses

L'ipotesi deve essere chiara, specifica, falsificabile

- il farmaco X aiuta i pazienti? → solo una domanda
- _____ ha un effetto → ipotesi debole
- _____ riduce i sintomi della malattia Y in base alla patologia
  - ipotesi forte
- l'ipotesi debole non fornisce spiegazioni
- usiamo la debole se ne sappiamo troppo poco → una nuova legge: solo dopo qualche anno di applicazione possiamo formulare ipotesi forte

%

Esempi veri, da valutare

1) ∃ altri universi con leggi fisiche — non falsificabile
2) indossare biancheria intima viola migliora l'umore — F
3) le fiamme ascono in modo diverso nell'acqua zuccherata — W
4) i libri di Max Cohen sono fantastici — opinione
5) lavarsi le mani x 20 secondi riduce la diffusione di malattie infettive — F
6) una mela al giorno toglie il medico di torno — proverbio / domanda
7) le persone sono + creative dopo aver guardato una "stand-up comedy"?

10,2 - IVs, DVs, models, and other sorts bingo

Dependent variables (DVs)

- DV include l'esito di una malattia in uno studio medico
- il peso perso dopo un regime dietetico
- il n° di click in uno studio di marketing
- si possono avere anche + DV

Independent variables (IVs)

- è una variabile che usa x spiegare le variazioni nelle DV
  - alcune possono essere manipolate sperimentalmente
  - altre possono essere misurate
- variabili regressive, esplicative, predittive

Ricerca	DV	IVs
cancro	mortalità	età, esercizio, tipo cancro
perdita peso	Δ peso	tipo dieta, esercizio, ore di sonno
E-commerce	profitt	overtime, budget / Twitter google Facebook
psicologia	numero riportato	ore sui social, genere

- alcune variabili sono numeriche, altre categoriche

Residuals

- IV raramente spiega completamente DV
  - la differenza tra ciò che spieghiamo e il dato effettivo = residuo
    - chiamato anche errore, deviazione, innovazione

Model

è un framework x interpretare i dati - I modelli vengono in profondità, specificità, capacità esplicativa, obiettività matematica -
- in questo libro Model = formula matematica

Es. Altezza persone dipende da altezza madre e padre, da alimentazione
x_1 x_2 x_3
infantile (insieme di ϕ e w) →

$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x + \varepsilon$$

con analisi di regressione calcolo β_i, la variabile residua il sistema con ε
- vedremo come trasformare una ipotesi → modello

Test statistic

È una statistica descrittiva di un set di dati, associata a un'ipotesi -

statistica inferenziale → determinazione probabilità che la statistica di test sia stata osservata a caso (errore...), invece che essere un effetto della popolazione -

statistiche di test → coeff. di correlazione, t-test, F-value, chi-square value

Null and alternative hypotheses

Le statistiche inferenziali includono l'adattamento ("fit") di un modello ai dati -
In molti casi un risultato importante (es. p-value) riflette la comparazione tra due modelli derivanti da due ipotesi - Le due ipotesi sono definite **null** e **alternative** -

Una "ipotesi nulla" è spesso abbreviata in H_0 oppure $H\phi$.

- la *null* dice che nulla di interessante accade (DV correlato da IV).
  L l'equazione matematica di H_0 ha alcuni dei β_i nulli.
- la *alternative hypothesis* è quella con cui specifichiamo la ricerca.
  È una alternativa alla *null*.
  L la si dice anche *effect hypothesis*, nel senso che specifica che c'è un effetto
- si usano anche H_1, H_2 quando ci sono + alternative

H_A = le persone acquisteranno + widgets dopo aver visto la pubblicità X rispetto a quella Y

H_0 = il tipo di pubblicità non ha effetto sui widgets

10,3 - Can you prove a hypotheses?

Confutare una ipotesi è molto semplice. Basta un caso contrario.

- Molte ipotesi in biologia non possono essere confutate in pratica - Questo perché la comprensione della realtà è limitata.
- accettiamo queste imperfezioni, dati non coerenti; possono ↑ essere utili x il progresso della scienza.
- Raccolgo i dati e scopro che H_A si adatta meglio ai dati di H_0, questo non implica che H_A sia dimostrata. Potrebbero ∃ modelli migliori → con una certa prob. H_A è migliore di H_0.

Es. supponiamo di avere $x_1=1$, $x_2=2$ e $y=3$
quale modello ti lega? certamente $y=x_1+x_2$, ma ∃ un altro modello? Sì, es. $-5x_1+3x_2$.
Quale modello scegliere? x esempio quello + semplice (Rasoio di Occam).

Non riusciamo a dire se abbiamo scelto il "modello vero", possiamo solo valutare H_A vs. H_0.

W.4 - Sample distributions under H_0 and H_A

La miglior dispersione SEM può fornire supporto statistico a favore di una ipotesi.

Es. La caffeina migliora l'umore

H_A: La Tesa di caffè aumenta lo stato emotivo auto-riportato dopo un'ora del consumo

H_0: La Tesa di caffè non ha impatto

- per testare ipotesi conduciamo un esperimento:

200 partecipanti (100 partecipanti in doppio cieco), entrambi dopo un'ora riportano lo stato emotivo. migliore

- lo stato emotivo è un costrutto complesso (lo valutiamo da 1–10)

modello $\delta = \beta c + \varepsilon$ ——— ——— (-10, 1)

$c = \phi$ se decaffeinato
$= 1$ " con caffeina

β = impatto della caffeina

ε = cambiamenti di umore dovuti ad altre cause

$\bar{\delta}$ = variazione auto-dichiarata / pedice C, D / caffeina - stecca

H_A: $\bar{\delta}_C > \bar{\delta}_D$

H_0: $\bar{\delta}_C = \bar{\delta}_D$

questi i termini dell'esperimento

- abbiamo bisogno di un gruppo di controllo
  - [- può essere che umore ↑ dopo aver bevuto bevande calde
  - [senza questo gruppo $\bar{\delta}_C = \phi$ diventa un'ipotesi debolissima
- "cambia l'umore" potrebbe anche essere scritto come $\bar{\delta}_C \neq \bar{\delta}_D$ (differenza tra ipotesi unilaterale e bilaterale)
- la variazione media è riferita a 100 camion. Non è necessario che tutti abbiano umore ↑ - Basta che la media ↑
- riscriviamo le due ipotesi:

$\bar{\delta}_C - \bar{\delta}_D > \phi$ ——— $\bar{\delta}_C - \bar{\delta}_D = \phi$

- in relazione al più precedente, in alcune analisi come la regressione, le ipotesi sono + direttamente collegate al modello-

Ricordiamo la 10,1 : $\delta = \beta c + \varepsilon$

$H_0 : \beta = \phi$ $H_A : \beta > \phi$ diciamo $\Delta = \bar{S}_0 - \bar{S}_D$ →

$H_0 : \Delta = \phi$ $H_A : \Delta > \phi$

- Come testiamo l'ipotesi? di Δ?

Supponiamo che la verità sia H_0 → quale il valore atteso Δ in fisica non sarà esattamente ϕ

- ripetiamo + volte l'esperimento con 200 partecipanti:

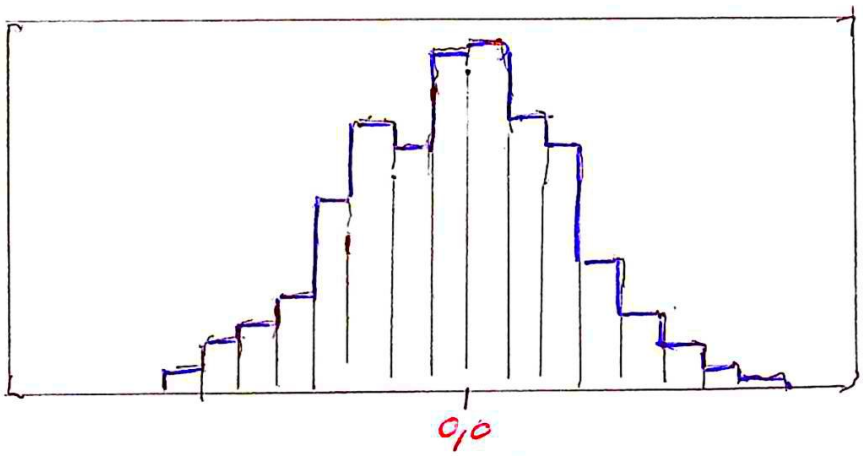
L vogliamo la distribuzione di Δ

LLN ci dice che questo valore converge a ϕ, se l'ipotesi è corretta-

Visualizziamo una ipotetica distribuzione di Δ-

Esperimento ripetuto 1000 volte-

Fig. 10,2



teorico

Questo grafico dice che è valida H_0

- Supponiamo un altro risultato

"A" ⇒ $\Delta = 0,1$ // "B" → $\Delta = 0,7$ tracciamo grafico

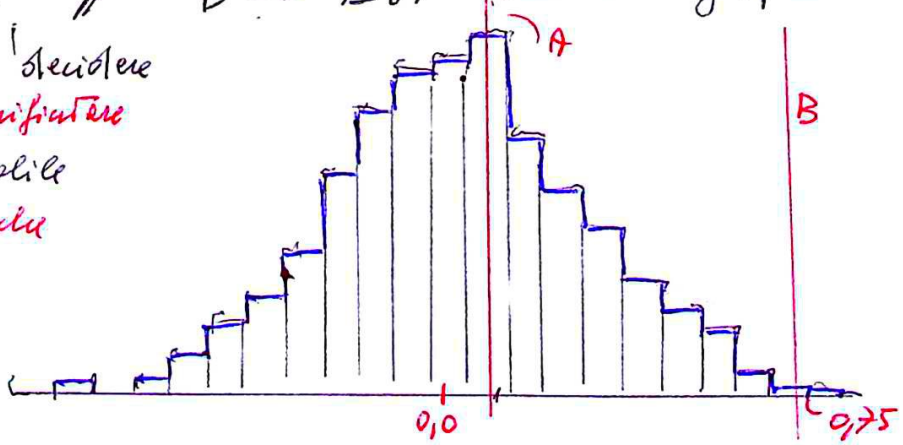
A da poca info x decidere

L con A non posso rifiutare

B è con poca probabilità

L posso decidere che

H_0 è vera



• vediamo di tradurre in forma quantitativa

10.5 - Where do H_0 distributions come from?

Come si crea un'ipotesi nulla? Necessitiamo di una distrib. di H_0 per valutare la "p" che non sia dovuta al caso.

• H_0 distribuzione analitica → distribuzione ottenibile senza dati

• H_0 empirica → randomizziamo i dati x creare un set di dati artificiali (mescolati, permutati, sumofati) che potrebbero essere sotto H_0 → creiamo dati falsi con le stesse caratteristiche di quelli reali, ma in cui siamo che $H_0 = \phi$ è vero

• H_0 empirica richiede di avere più dati → Cap. 16

10.6 - P-values: definition and misinterpretations

p = probabilità di ottenere per vera H_0
- p piccolo = poco probabile

p = probabilità di ottenere una statistica del test estrema, ... o + estrema della statistica di test osservata
- supponendo che sia vero H_0

NB p piccolo non implica che H_A sia vero

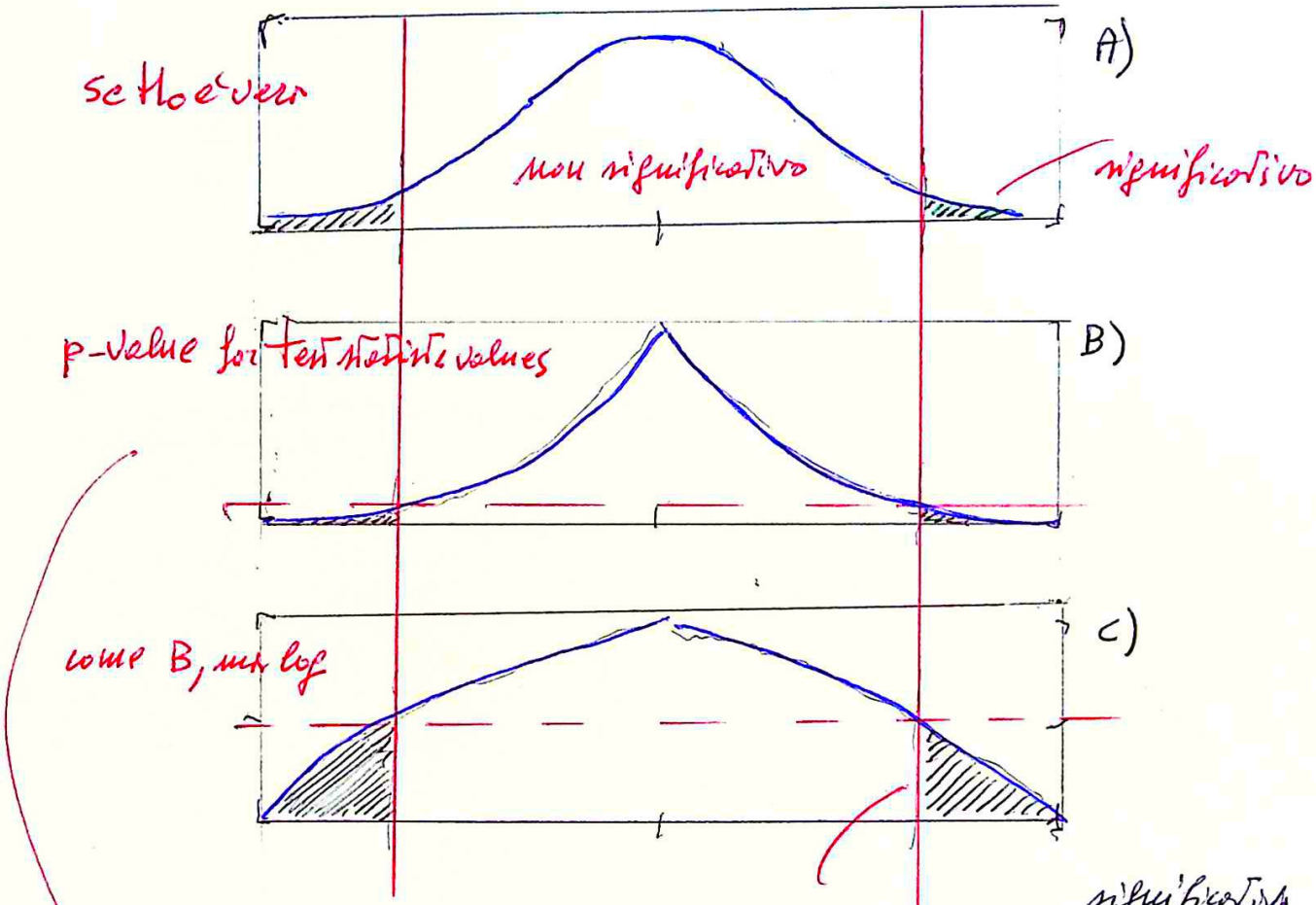
possiamo calcolare la probabilità della statistica ipotizzando H_A
- se p piccolo → dati più coerenti con H_A
  - non significa che H_A sia vero

• i valori di "p" sono sfumati → vedi più avanti

P-values and statistical significance

significanza = rifiutiamo H_0 sulla base di "p".
i valori di "p" sono continui da $\phi \to 1$

- bisogna prendere una decisione → dobbiamo binarizzare
  es. seguendo come soglia .5
  L si usano anche altre soglie .01 / .001
  - ho statistica di test se p < soglia → statistica significativa
  - di solito la soglia si indica con α
    es. $\alpha = 0{,}05$ o $\alpha = 5\%$
- dividiamo quindi in "regioni significative"



una statistica con p associato < 0,05 è statisticamente significativa

Nei grafici "test a due code"

qui si vede meglio ora

il grafico B) può sembrare a un esempio, ma viene dal fatto che cdf di Gauss è invertito per $z > 0$ — ancora

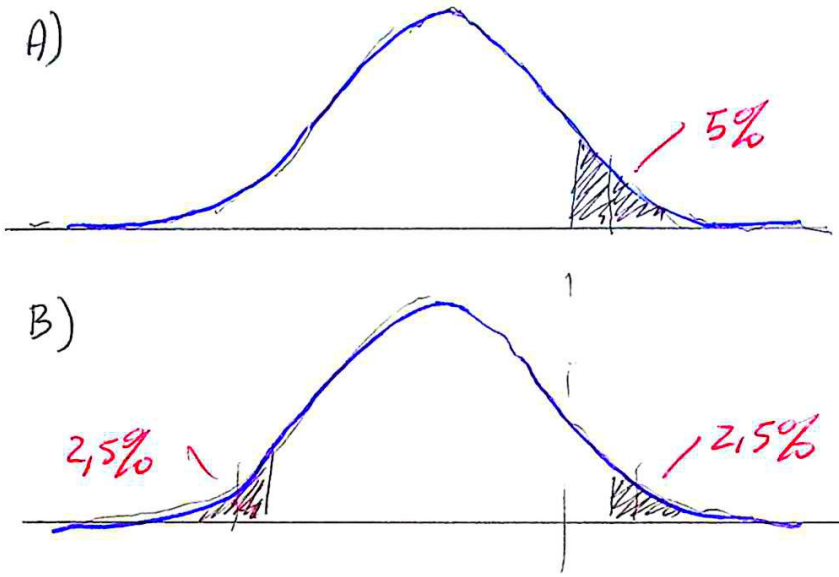
se $z < 0$ allora p corrisponde a cdf
se $z < 0$ " p " " 1 − cdf

P-values and distribution tails

X specificare una ipotesi: effetto positivo, negativo, $\neq \phi$

strong una coda $\begin{cases} \Delta < \phi \\ \Delta > \phi \end{cases}$ due code $\Delta \neq \phi$ weak

le due code mappano x una ipotesi debole



weak

strong

Utile x rigettare e' la strong - In fisica si usa sempre la bilaterale -

Talora anche una coda - Es. F test nelle ANOVA e nelle regressioni -

Oppure con distribuzioni di forma

p si origina dagli stati che definiscono H_0

p-values ottenuti mediante formula

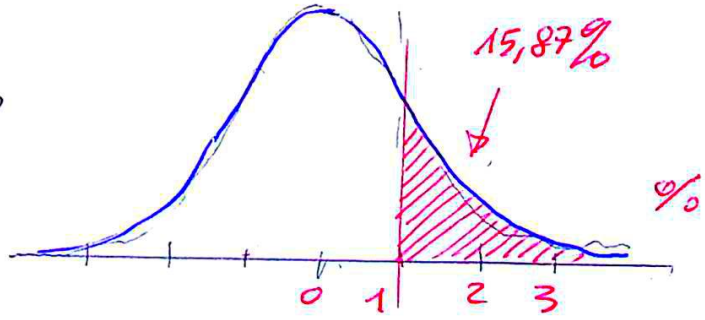
Esempio

punteggi z, come ottenerli in p? z sono Gauss, media ϕ

quale la probabilità di $z > 1$? → dobbiamo calcolare l'area a destra di $z = 1$ nella pdf corrispondente a Gauss -

Come si vede da figura:

$z > 1 \longrightarrow p = 0,158$



in unità di σ

empirical p-values

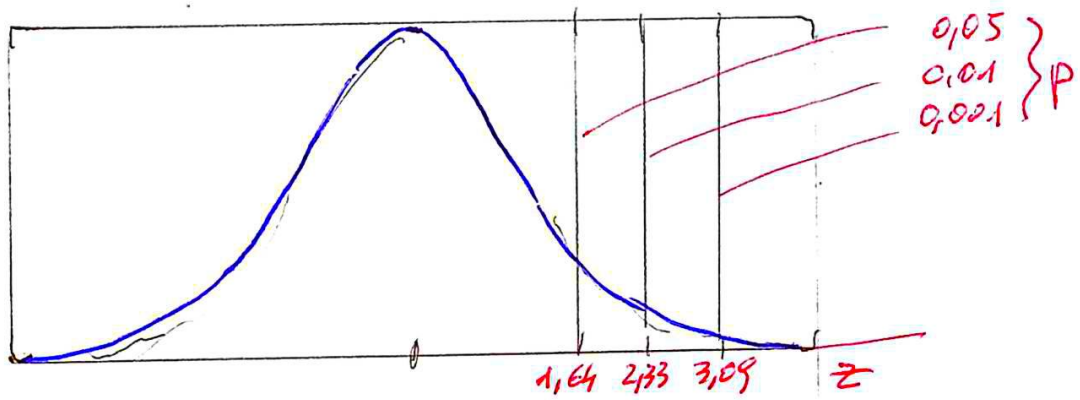
Se usiamo statistiche basate su permutazioni x avere H0 empirico

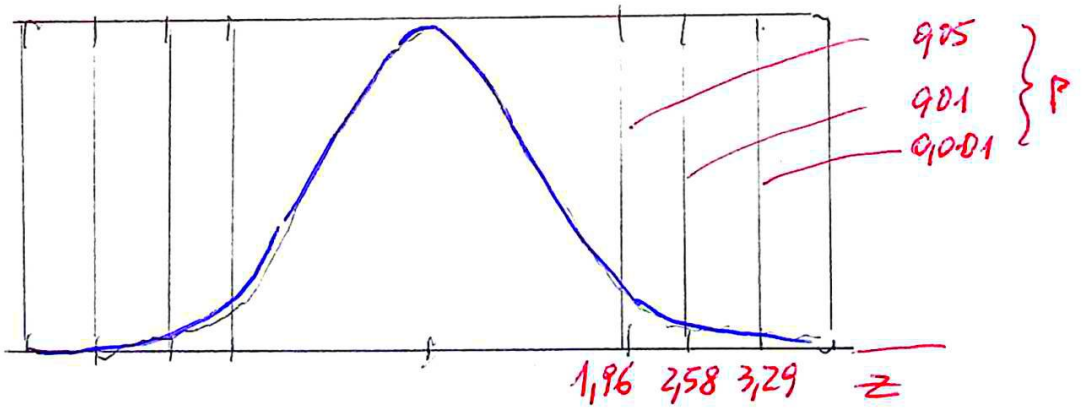
→ p-value = distanza normalizzata rispetto al centro della distrib. di H0 → Cap. 16

Precision of p-values - si può calcolare con precisione arbitraria. In pratica i numeri sono troncati. Per valori di p piccoli si usa la notazione scientifica.

P-z combinations to memorize

alcune di queste combinazioni meritano la memorizzazione





Misinterpretations riferite a p=0,02 → il p=0,02%

1) l'effetto è presente x il 98% della popolazione (popolazione)
2) 98% di probabilità che la mia statistica di test = parametro della)
3) p < soglia → l'effetto è reale

4) $p <$ soglia → una variabile causa l'altra

5) se $p = 0{,}08$ → $p >$ soglia → H_0 è vera

Risposte e correzioni

1) c'è una $p = 2\%$ che non ci sia alcun effetto / è la mia statistica era dovuta a variabilità, rumore, piccola dimensione del campione, distorsione sistematica.
2) c'è una $p = 2\%$ che la mia statistica sia dovuta a rumore
3) $p <$ soglia → è improbabile che l'effetto sarebbe stato osservato
4) $p <$ soglia → variabili correlate
5) $p >$ soglia → l'effetto avrebbe potuto essere osservato se H_0 vera
è improbabile che mio H_A specifico sia vero — vedi ipotesi alternative che siano migliori di H_0.

Discussione sui 5 punti

1) $p \approx$ effetto medio all'interno del gruppo in relazione alla variabilità; non dice nulla sulla presenza o meno dell'effetto x ogni individuo. Si può determinare se un effetto è significativo per ciascun individuo, ma non è questo che indica p.

2) p non dice nulla sulla relazione tra campione e le caratteristiche della popolazione. È semplicemente la probabilità che la statistica di test avrebbe potuto essere osservata se H_0 = vera.
Si possono usare gli intervalli di confidenza x quantificare la relazione tra le caratteristiche del campione e i parametri della popolazione.

3) $p <$ soglia non dimostra che l'effetto è reale, dice che l'effetto osservato è improbabile data H_0 = vera.
L'uso colloquiale lo diciamo, ma non è appropriato)

4) p che solo non stabilisce la causalità - Le interazioni causali sono state determinate tramite manipolazioni sperimentali o con altri tipi di analisi -

p significativo → relazione tra le variabili, non causalità -

p > soglia → non dimostra H_0 → dice che l'effetto è probabile se H_0 = vera -

Problems with p-values

p non dimostra una ipotesi, non dimostra causalità, non dice la proporzione di popolazione che prova l'effetto -

p < 0,001 oppure p > 0,3 è relativamente facile da interpretare, perché siamo su valori estremi → elevata confidenza su presenza/assenza dell'effetto -

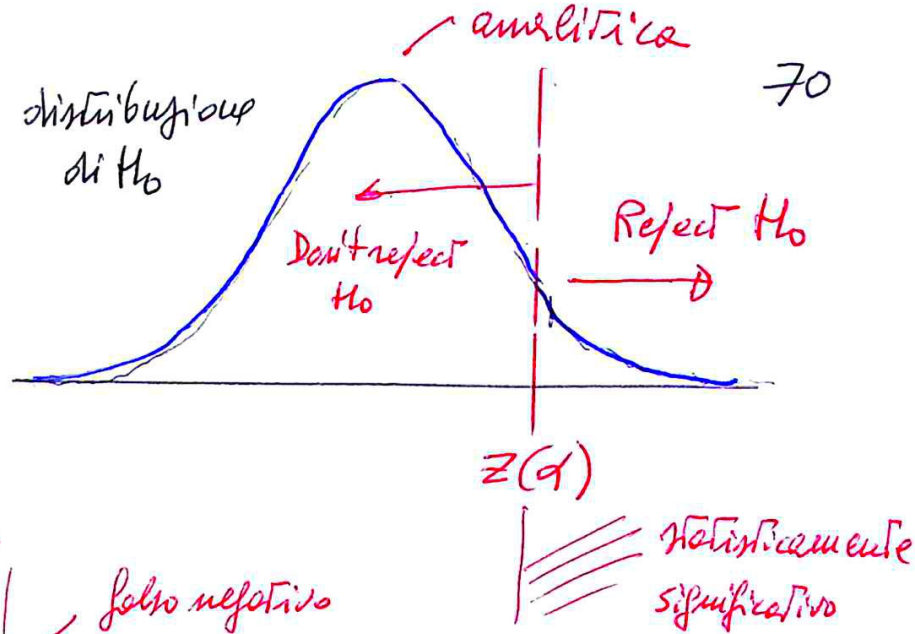
p prossimo a 0,05 è + difficile da interpretare, anche xché siamo prossimi alla soglia di 0,05%, e un po' di rumore potrebbe cambiare tutto -

10.7 - P-values and significance categorization

Per capire ciò che ci circonda raccogliamo dati, e valutiamo se una ipotesi è in linea con essi - Il risultato è statistico - → → ci affidiamo a p x decidere in modo informato -
Come si passa allo spazio bidimensionale della decisione?

- lo vediamo con una figura:

Vediamo come tabella delle decisioni statistiche corrette



Realtà su Ho

Vera	Falsa	
Vero < ϕ (1-α)	(β)	Non rigetto Ho
(α)	Vero > ϕ (1-β)	Rigetto Ho

falso negativo

falso positivo

- poi rifiuto Ho quando è davvero falsa
- non rifiuto Ho " " " vera

– la decisione implica incertezza

10,8 – Type I and Type II errors – un giudice:

– dice a un innocente sei colpevole
– dice a un criminale sei innocente

Where do errors come from?

– non possiamo provare che H_A vera
– rifiutiamo H_A se il nostro test è < 5%

%

la decisione è statistica → possono essere fatte errori

Type I errors rifiuto H_0 quando è vero

la probabilità di questo è la "p-value significance threshold"

L sono detti α errors → falso negativo

Type II errors non rifiuto H_0 quando è falso

la probabilità di questo è correlata alla potenza statistica del test

$(1-\beta)$ → Cap. 17

L migliora con campioni + grandi

The balance of Type I and Type II errors

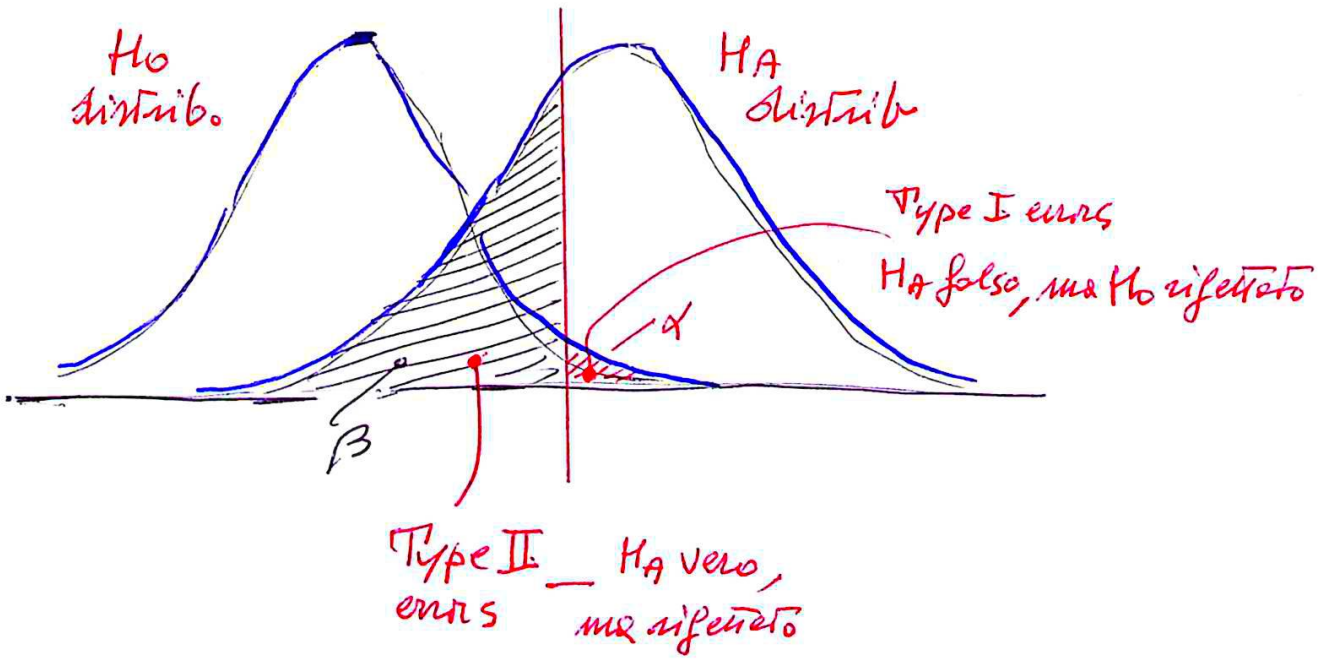
I due Tipi di errori sono correlati. Se diminuisco l'uno, l'altro aumenta

Esperimento mentale

H_A vero al 100% — possiamo ripetere molti esperimenti →

avremo un'intera distribuzione di statistiche di test in cui H_A = vero

Possiamo visualizzare le distribuzioni sotto H_0 e H_A



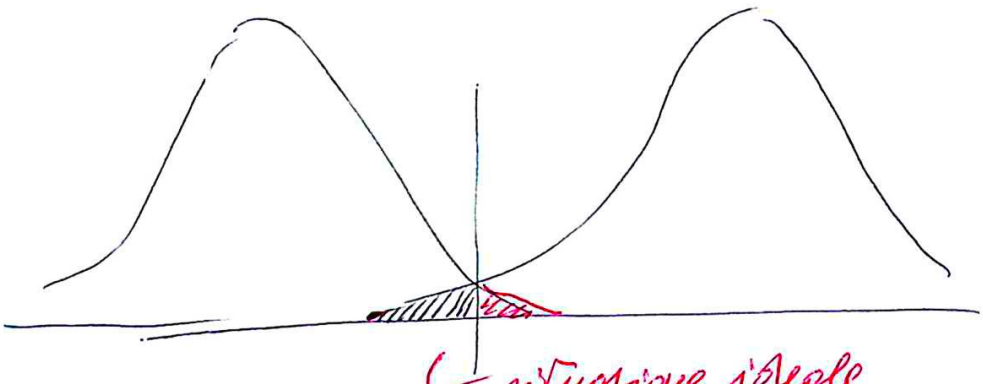


Fig 10.12 C

C situazione ideale

L o si diminuisce la variabilità
o si aumenta la dimensione dell'effetto } → Cap. 11

– però non conosciamo la realtà → non conosciamo la distrib. di H_A – Abbiamo un solo valore di H_A, e ci domandiamo se quel valore della statistica di test empirica sia un valore insolitamente alto – tratto da H_0 – o se sia un valore tratto da H_A.

10.9 - Various interpretations of "significant"

Se p abbastanza piccolo → statisticamente significativo
L non è l'unica interpretazione

praticamente

Significante: statisticamente, teoricamente, clinicamente, socialmente

Statistical significance = prob. che una statistica di test empirica venga ottenuta con H_0 = vero

Theoretical significance – ci stiamo riferendo a una teoria scientifica
L non c'entra nulla con quella statistica

Clinical significance – risultato rilevante x la pratica clinica
L non correlato con statistica

Pratical, societal, educational, etc.

la statistica vive el confine tra scienze dure e soft →
come utilizza "p" calcolato e' compito che richiede competenze specifiche

10.10 - Multiple comparisons

= Testare più ipotesi x uno stesso set di dati
- con Test ognuno di Type I
- non si deve pensare ai singoli Test, ma al loro insieme
  L + membri → + probabile che qualcuno sia etichettato erroneamente

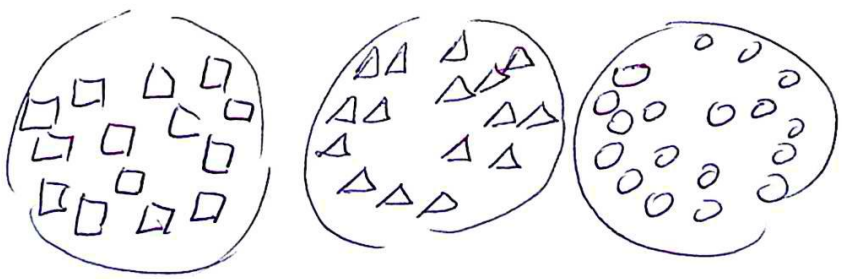
Esempio - soglia $p < 0{,}05$ - facciamo 3 Test → (FWE)
→ probabilità combinata di Type I = 0,15 = 15% →
ogni singolo 0,05 → $\sum_{1}^{3} = 0{,}15$ = "family-wise error rate" accurato
in formula $FWE = m\alpha$ anche non esattamente

Dipendenza tra i dati: può diminuire FWE

altra formula $1-(1-\alpha)^m$

Una dipendenza può essere insita confrontando / A vs B
- A vs C

L qui i Test sono correlati x che' entrambi contengono A



H_1 □ vs. △ H_2 ○ vs. △

H_3 □ vs ○

Solutions to the multiple comparison problem

- bisogna regolare la soglia di significatività / corrisponde a FWE
  - è non alla soglia individuale
  - ∃ metodi x fare questo → Ch. 14 in ANOVA

Bonferroni = imposto FWE = soglia desiderata = es. 0,05

e la soglia dei singoli test = α/m = 0,0167

1) questa correzione può essere troppo rigorosa (presuppone la indipendenza dei test)
2) si basa solo su m e non sulle caratteristiche dei dati
3) aumenta la probabilità di TypeII → riduce la identificazione di risultati veri

Conclusione: buona se m non troppo grande, se i test sono indipendenti

False discovery rate (FDR)

utile se molti valori di p da test correlati (serie temporali, dati geografici, genetici) - FDR fornisce soglia oltre la quale viene calcolata la distribuzione di p -

$$q = \frac{\text{n° previsti di falsi positivi}}{\text{veri}}$$ → Esercizio 4

q = f(p) è un limite → un risultato potrebbe essere etichettato come significativo o non significativo solo in funzione di p

Cluster correction = buona x dati che hanno struttura di correlazione

Test significativo se e' all'interno di un insieme di test contigui significativi

↳ Fig. 10,14 = griglia — colore corrisponde a valore di p

- la soglia minima del cluster puo' essere determinata in base a considerazioni a priori (es. un cluster deve contenere almeno 100 ms di punti contigui, oppure almeno 6 pixel contigui significativi)
- la soglia minima in base a "permutation testing" x definire una dimensione empirica del cluster sotto Ho_

Questi metodi controllano Type I, ma Type II ↗

↳

10,11 - Degrees of freedom

elem3 = 10
elem1 = 2
— 2 = 3

Se dico: 3 elementi — media = 5

↳ quanti gradi di libertà?

df = n° gradi di libertà oppure ν (nu)

una statistica descrittiva viene calcolata dai dati — potrebbe essere distorta

df è usato come parametro nelle distribuzioni H_0 →

df utilizzato x fare inferenze sulla popolazione sottostante

df = utile x controllare errori, correggere interpretazioni dei risultati nelle regressioni e nelle ANOVA

~~∃~~ unica formula x calcolare df — p. 380

Possiamo pensare al df di una analisi come il n° delle osservazioni meno il n° di parametri in formula

$df = N - K$ — è una semplificazione

N: n° di elementi o il n° di gruppi o le condizioni di esperimento

Esempio df associato a t-test → $\begin{cases} N-1 \\ N-2 \end{cases}$ dipende se c'è un solo gruppo oppure due — K = n° di IVs — vedi dopo

I gradi di libertà possono essere anche non interi —

%

10,2 - Exercises

1) nello Fig. 10.5 il codice emette $p < 0,05$ - Modificare in modo da utilizzare un arbitrario p.

In seguito modificare codice

2) valutare il tono di fastidiosi - → uso t-Test x determinare se la media di un campione deriva significativamente da un valore prefissato - La media con attesa = ϕ, ma un campione casuale può avere un po' diversa in modulo: "statisticamente significativa" - Questo sarebbe un Type I → la vera media = ϕ, la media del campione $\neq \phi$ con $p < 0.05$ -

```
N = 20
X = np.random.randn(N)
p = stats.ttest_1samp(X, φ)[-1]        p-value from t-Test = 0,0185
media = np.mean(X)  ⟶ media = 0,44
```

simulo i dati e t-test

```
alphaRange = np.linspace(.001, .9, 15)

M = 100
type1errors = np.zeros(M, len(alphaRange))
meansAndPvals = np.zeros(M*len(alphaRange), 2)
```

loop sopra alpha

```
for ai, alpha in enumerate(alphaRange):
  for expi in range(M):
    X = np.random.randn(20)          loop su esperimento
    p = stats.ttest_1samp(X, φ)[-1]
    memorizzo
    type1errors[expi, ai] = p < alpha
    estraggo media e p
    meansAndPvals[expi*len(alphaRange)+ai, :] = [np.mean(X), p]
```

Grafico ⟹

ax s[1].plot(meansAndPvals[:,0], meansAndPvals[:,1], 'k')

sample means — Pvalues

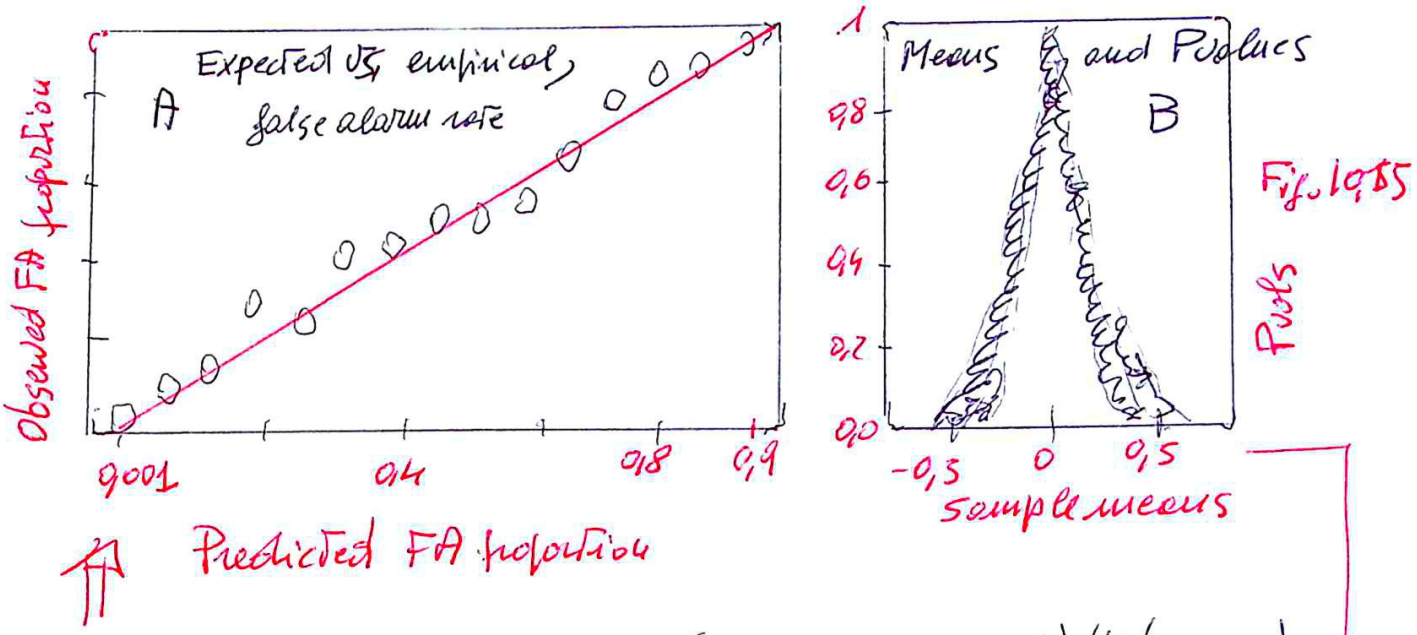


Fig. 10.15

ax s[0].plot(alphaRange, np.mean(typeIerrs, axis=0), 'ks' ——)

ax s[0].plot(——, alphaRange, ——)

Ho ripetuto t-test 100 volte, registriamo il n° di volte che p è sotto soglia (cioè quando il test è significativo) (1500?)

Con una soglia $\alpha = .05$ → si trova di circa 5 campioni

∀ risultato con $p < .05$ = Type I

Metto un secondo ciclo con $\alpha = .001 \div .9$ → rappresento graficamente il tasso di falsi allarmi empirici in funzione di α

FIG. 10.15A.

Nella FIG. 10.15B vediamo che – più la media è vicina a 0 → maggiore è il valore di p – Non è una relazione perfetta anche t-test dipende anche da std – In ogni caso è una relazione ben consolidata – Più leptocurtica rappresenta il numeratore della statistica t e il denominatore è ≃ cost in tutte le simulazioni – → Cap. 11

3) la distribuzione Ho di questo corvolo sono le pdf Gaus_

Gaus sono usate spesso x visualizzare le distrib. Ho x che' comuni in stats, e non richiedono parametri_

Ma molte altre distrib. x Ho sono importanti_

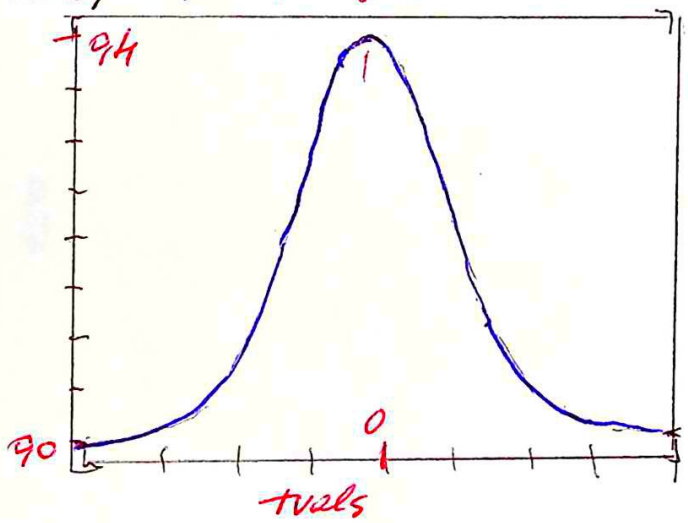
Esploriamo la distribuzione t Ho di un t-values_

La distribuzione t richiede un parametro: i gradi di libertà

La forma della t-distrib. dipende da df

- Prendiamo una "famiglia" di distribuzioni supponendo Ho=vero_

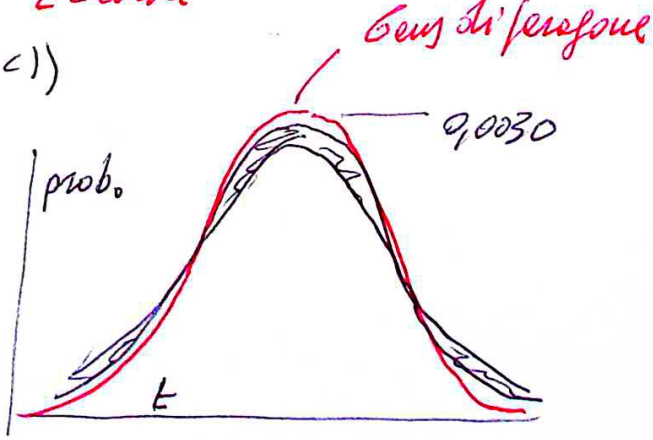
```
tvals = np.linspace(-4,4,1001)
calcolo pdf                          = df
tpdf = stats.t.pdf(tvals, 20)        per grafico
per.plot(tvals, tpdf)
```



Ora vogliamo usare df come parametro

```
dfs = np.arange(4,41)
for df in dfs:
    calcolo pdf
    tpdf = stats.t.pdf(tvals, df) * np.diff(tvals[:2])
    devo colore x ogni curva, stampo   c = colore
    plt.plot(tvals, tpdf, color=(c,c,c))
```

df ha un impatto blando_ t ha sempre code + spesse di Gaus



4) Due modi di usare la correzione FDR

1) aggiustare i valori "raw" considerando significativo $\forall p$ / $p_{adj} < \alpha$

2) derivare una soglia α aggiustata e considerare significativo $\forall p$ empirico $< \alpha$ aggiustato

Es. FDR derivato da α aggiustato, corrispondente a $p < .05$ potrebbe essere $q < .008$

$\forall$ p-value $< .008$ sarà considerato significativo

$\forall$ ——— $> .008$ " " non-significativo

importo fdrcorrections

```
from statsmodels.stats.multitest import fdrcorrections
```

valore di soglia corretto

```
pThresh = .05
```

pvalues

```
K = 40
pvals = np.random.uniform(low=.001, high=.3, size=K)**2
```

ordino p-values

```
pvalsSort = np.sort(pvals)
```

linear interpolated distribution

```
pvalsInterp = np.arange(1, K+1)/K
```

adjusted p-values

```
pvals_adjusted = pvalsSort / pvalInterp
```

questa funzione dà in tuple: 0 rifiutare, 1 adjusted p-values

```
qq = fdrcorrections(pvalsSort, pThresh)[1]
```

grafico

```
plt.plot(qq, 'ko' ———) fdrcorrections()
——— (pvals_adjusted, 'k^' ———) p-values
——— (pThresh * pvalsInterp, color='gray' ——— rejection line
```

vedi Plot a pagina successiva

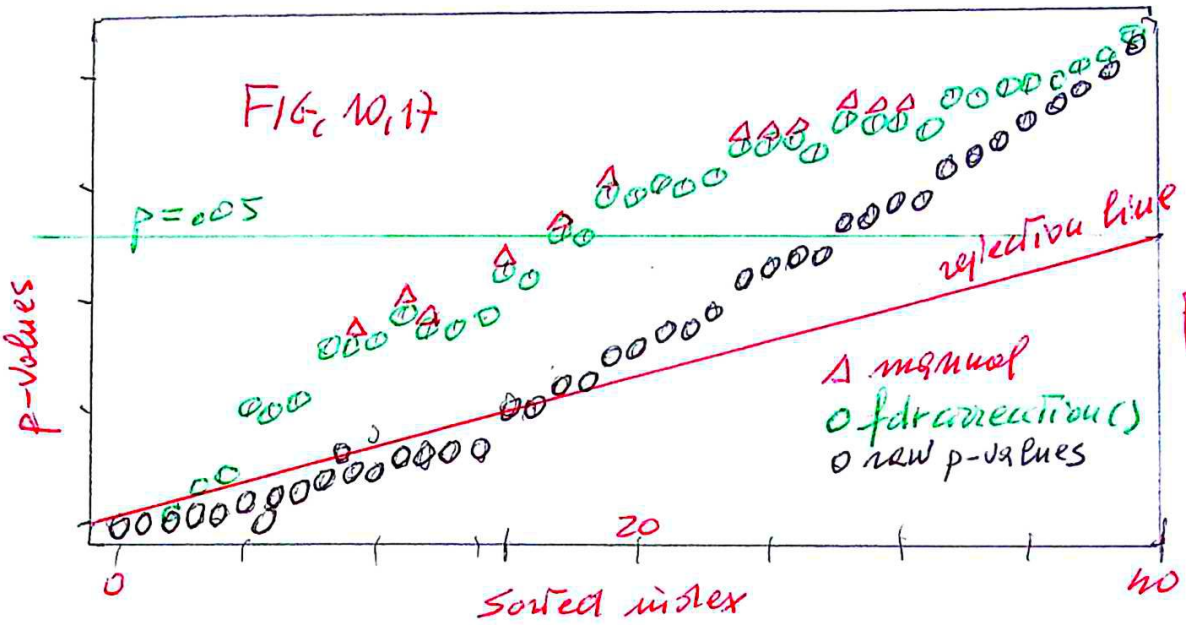


Fig. 10,17

abbiamo creato 40 random p-values come u^2 con $u \in U(.001, .3)$

$K = 40$ Ordinati, interpolazione lineare con v/K

dove v = vettore da $1 \to K$

calcolati adjusted p-values dividendo p-values / pvalsInterp

abbiamo usato la funzione fdrcorrections()

- FDR correction considera che $\forall$ raw p-value sotto la "rejection line" siano significativi.
- FDR correction " " $\forall$ adjusted p-values sotto la linea orizzontale $p = q = .05$ siano significativi
- con FDR i valori di p sono diversi da quelli raw

— Uso complementare: determinare di aggiustare senza modificare p. → esercizio successivo

p. 385

Esempio - t Student test

Vogliamo fare il test x il confronto di due dati appaiati - tra medie

6 coppie di gemelli, un gemello fuma, l'altro no
andiamo ad esaminare una variabile distribuita Gauss, relativa alle particelle inspirate -

	F	$\bar{F}$
1	60,6	47,5
2	12,0	13,3
3	56,0	33,0
4	75,2	55,2
5	12,5	21,9
6	29,7	27,9

$\bar{F}$ → non fumatori $\bar{F}$=gx-oF1SφyzU YouTube

Ipotesi $H_0 \rightarrow \mu_F = \mu_{\bar{F}} \rightarrow \mu_F - \mu_{\bar{F}} = \phi$

Ipotesi altern. $H_1 \rightarrow \mu_F - \mu_{\bar{F}} \neq \phi$

supponiamo $\alpha \rightarrow 0,10$ $n = 6$

I dati non sono indipendenti x che i soggetti sono gemelli → non possiamo usare il t-test x gruppi indipendenti -

→ uso: differenza tmedie = media delle differenze →

scrivo in modo diverso l'ipotesi nulla:

$H_0 \rightarrow \mu_d = \phi$ (differenza) $H_1 \rightarrow \mu_d \neq \phi$

calcolo le differenze

	d
1	13,1
2	-1,3
3	23
4	20
5	-9,4
6	1,8

Prima mi sono ridotto al test su un solo gruppo → test su una media

→ uso t-Test

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

che sotto H_0 è distribuito come t (n-1) gradi di libertà

→ nel nostro caso

→ $$t = \frac{\bar{d} - \phi}{\frac{s_d}{\sqrt{n}}}$$ radice della varianza campionaria

	d
1	13,1
2	−1,3
3	23
4	20
5	−8,4
6	1,8

$\bar{d} = 7{,}87$

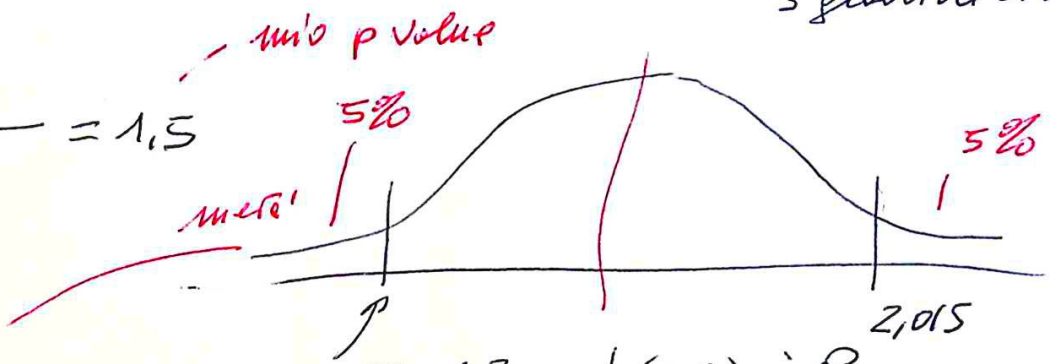
$gdl = n - 1 = 5$

$s_d^2 = 164{,}5 \rightarrow s_d = 12{,}8$

il nostro valore di t vale

$$t = \frac{7{,}87}{\frac{12{,}8}{\sqrt{6}}} = 1{,}5$$

t-Student con 5 gradi di lib.



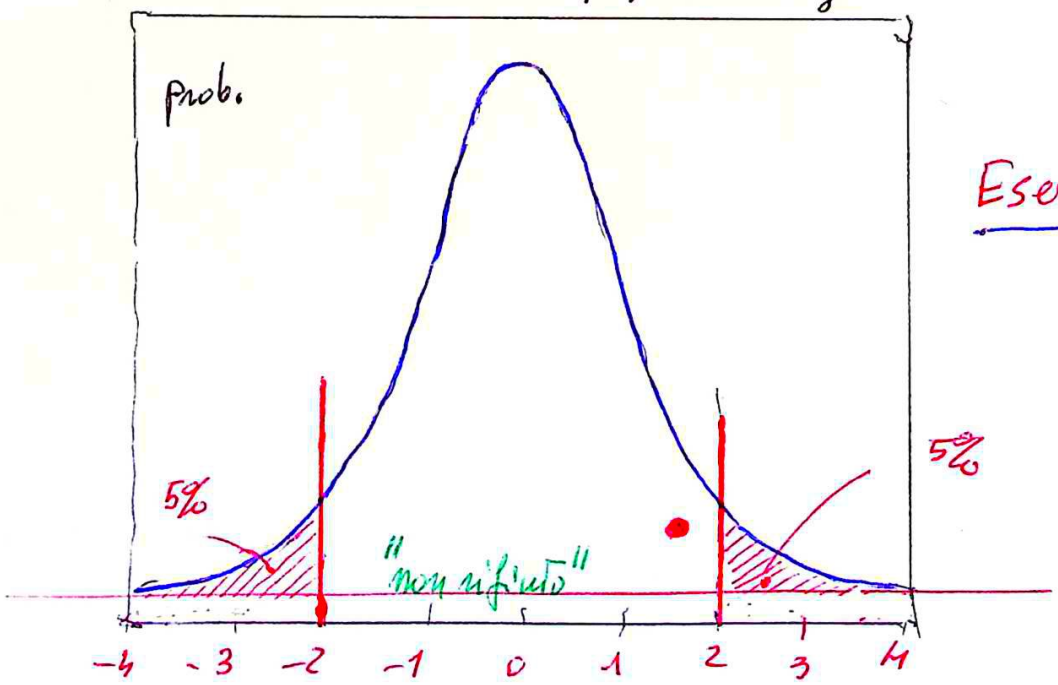
se $\alpha = 0{,}1$

quantovale? qt(p,8) in R

ESERCIZIARIO di STATISTICA Vol. 1 — soluzioni ragionate

L'100 esercizi svolti di statistica descrittiva

t-Student 5 gradi di libertà



Esercizio 39

5) determiniamo una soglia corretta da applicare ai p-value zero-

vedi Fig. 10,17 di esercizio 4-

E' il p-value grezzo + grande che si trova sotto la soglia di rifiuto

Nel nostro grafico e' il n° 18 ordinato, e il suo valore e' $p = 0,0181$

Quelli sotto soglia sono i p-value di "non rifiuto"-

Scrivere codice x identificare questo p-

(usare fdrcorrection())

6) compariamo tra loro Bonferroni e FDR (usa fdrcorrection())-

Dobbiamo creare un set di valori p, contare quanti di questi sono significativi, usando vari metodi di correzione-

N=100 casuali come $p = u^2$ per $u \in [0.001, .25)$ con $\alpha = .05$

Calcoliamo la percentuale dei p che sono sotto soglia-

```
pvals = np.random.uniform(.001, .25, N)**2          (*)
bon_thresh = .05/N
q = fdrcorrection(pvals, .05)
```

FDR significativi, Bonferroni significativi

```
print(100*np.mean(q[0]))  → 86%
print(100*np.mean(pvals < bon_thresh)) → 11%
```

Facciamo un loop di 100 sets di p-values, calcoliamo media, std dei p-values sotto soglia

```
sigTest = np.zeros((100, 2))
for expi in range(100):
    pvals = ____________ (*)
    bon_thresh = .05/N
    q = ____________
    sigTest[expi, 0] = 100*np.mean(q[0])
    sigTest[expi, 1] = ____________(pvals < bon_thresh)
```

FDR + indulgente di Bonferroni

esperimento x scegliere

FDR — 80,45% — significativi

Bonferroni — 8,89% — "

%

⇒) Bonferroni si basa sul n° di test N
FDR si basa sulla distribuzione di p

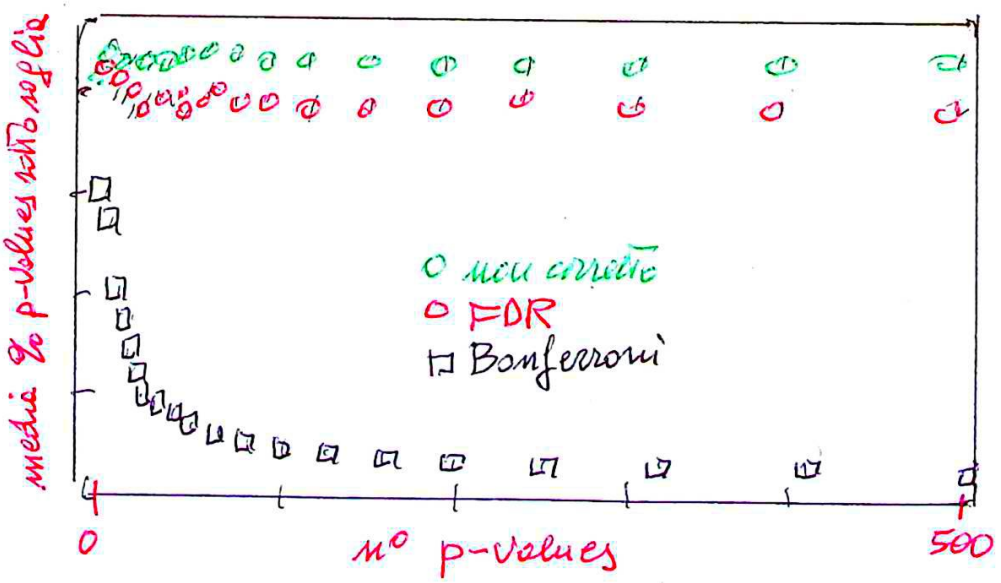
Nell'esempio 6 in un altro loop, che varia i valori di p da 2→500 a passi di 25 spaziati log-

```
Ns = np.logspace(np.log10(2), np.log10(500), 25, dtype=int)
```
distribuzione log dei p ↗

```
nRepetitions = 100
```
matrice del risultato finale (una x ogni ripetizione)

```
sigTests = np.zeros((len(Ns), 3))
for ni, n in enumerate(Ns):
    for _ in range(nRepetitions):
        p-values & corrections
        pvals = np.random.uniform(.001, .25, n) ** 2
        bon_thresh = .05/n
        q = fdrcorrection(pvals, .05)
        sigTests[ni, 0] += 100 * mean(pvals < .05)   non corretti
        ———————— 1 += ———— (q[0])                    FDR corretti
        ———————— 2 += ———— (pvals < bon_thresh)      Bonferroni corretti
```

La distribuzione dei p non viene al variare di N → FDR ~ costante



8) Dimostriamo "empiricamente" il rischio di "errore flessibile", cosa che può succedere eseguendo ripetutamente un test stat. inferenziale, aumentando ogni volta il campione. → Cap. 18

È una prassi non etica, che aumenta il rischio di ottenere $p<.05$ anche quando non c'è alcun effetto.

5 numeri da $N(0,1)$ — eseguiamo t-test contro $H_0: \mu = 0$

naturalmente il test può dare risultato non significativo.

Ad ogni iterazione aggiungo 3 numeri casuali al campione.

L quindi ricalcolo t-test

Fare grafico di p vs. dimensione campione

NOTA l'errore flessibile varia in funzione della casualità dei p

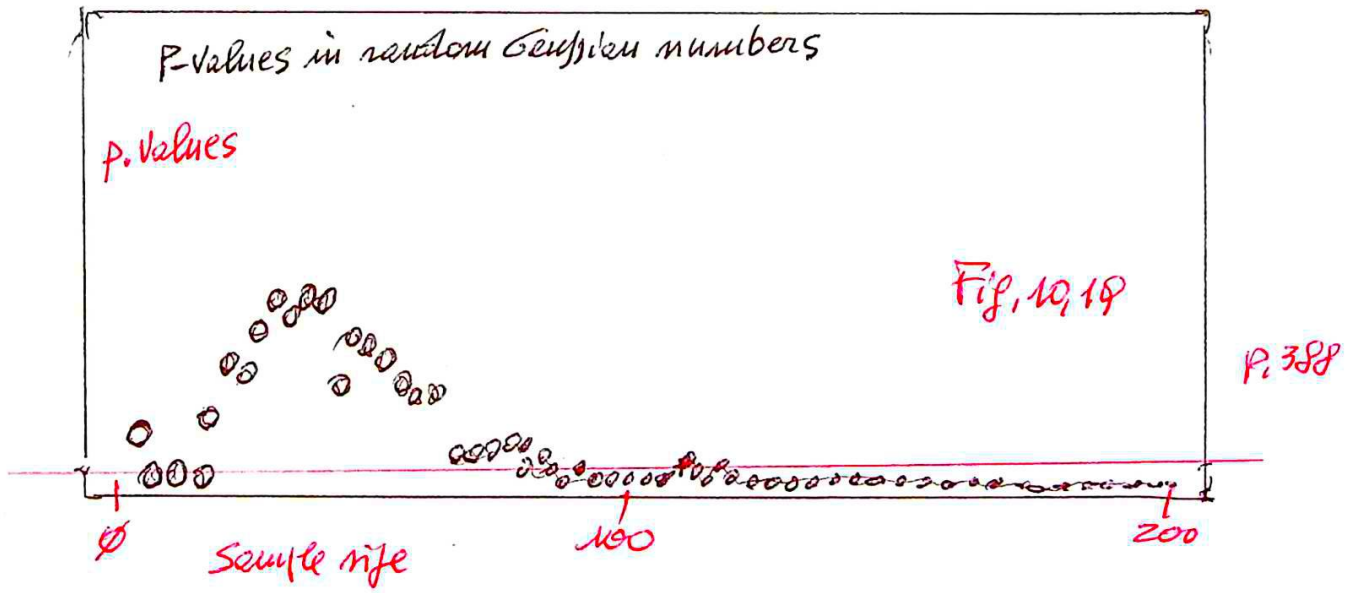


Fig. 10.19

p. 388

Stiamo generando numeri casuali da una distribuzione casuale con $\mu = 0$

Tuttavia il test diventa significativo solo $\simeq$ a 100

Se aumento i campioni, alla fine confermo l'ipotesi, ma ciò non corrisponde alla realtà.

9) incoraggiamento

— FINE CAPITOLO 10 —