Laboratorio 2

analisi statistica dei dati sperimentali

Laboratorio 2 – analisi statistica dei dati sperimentali

introduzione

parte (4 CFU) del corso di Laboratorio 2 che verrà svolta in parallelo con la parte di laboratorio di elettromagnetismo

consiste di lezioni e di ‘’laboratori’’

lezioni: parte teorica ma anche introduzioni alle attività di laboratorio e discussione dei risultati

laboratori: svolgimento di esercizi (obbligatori per sostenere l’esame) strettamente legati a quanto fatto a lezione, seguito direttamente anche grazie all’aiuto di A. Kerbizi e A. Moretti

scopo del corso: acquisire gli elementi di base per l’analisi statistica dei dati: come estrarre informazioni quantitative dai dati disponibili

argomento vastissimo e difficile: solo le basi, senza pretesa di completezza, mediando tra tempo disponibile e rigore: non tutto verrà dimostrato, ma non sarà un ‘’ricettario’’

fondamentale verificare e provare ad applicare la parte teorica

→ molto importanti, oltre agli esercizi a lezione, i laboratori

statistica

Treccani

‘’**statistica**

Scienza che ha per oggetto lo studio dei fenomeni collettivi suscettibili di misura e di descrizione quantitativa: basandosi sulla raccolta di un grande numero di dati inerenti ai fenomeni in esame, e partendo da ipotesi più o meno direttamente suggerite dall’esperienza o da analogie con altri fenomeni già noti

mediante l’applicazione di metodi matematici fondati sul calcolo delle probabilità, si perviene alla formulazione di leggi di media che governano tali fenomeni, dette leggi statistiche

spesso la raccolta dei dati viene limitata a un campione più ristretto, opportunamente predeterminato in modo da rappresentare fedelmente le caratteristiche generali.

Concepita inizialmente come attività descrittiva di certi fatti sociali e in particolare come attività amministrativa dello Stato, la s. ha via via ampliato i suoi confini, fino a diventare una vera e propria ‘scienza del collettivo’, disciplina con finalità non solo descrittive dei fenomeni sociali e naturali, ma orientata **anche** a finalità di ricerca nei vari ambiti scientifici’’

statistica

Wikipedia

La **statistica** è una disciplina che ha come fine lo studio quantitativo e qualitativo di un particolare fenomeno collettivo in condizioni di incertezza o non determinismo, cioè di non completa conoscenza di esso o parte di esso.

Strumento del metodo scientifico, si avvale della **matematica** per studiare i modi in cui un fenomeno collettivo può essere sintetizzato e compreso e ciò avviene attraverso la raccolta e l'analisi delle informazioni relative al fenomeno studiato

La scienza statistica è comunemente suddivisa in due branche principali:

- statistica **descrittiva**: ha come scopo quello di sintetizzare i dati attraverso … strumenti grafici (diagrammi a barre, a torta, istogrammi,…) e indici (indicatori statistici, indicatori di posizione, di dispersione, di forma, …) che descrivono gli aspetti salienti dei dati osservati
- statistica **inferenziale**. ha come obiettivo quello di stabilire delle caratteristiche dei dati e dei comportamenti delle misure rilevate (variabili statistiche) con una possibilità di errore predeterminata. Le inferenze possono riguardare la natura teorica (la legge probabilistica) del fenomeno che si osserva. La conoscenza di questa natura permetterà poi di fare una previsione (…ad esempio, …"l'inflazione il prossimo anno avrà una certa entità" … esiste un modello dell'andamento dell'inflazione derivato da tecniche inferenziali). La statistica inferenziale è fortemente legata alla **teoria della probabilità**….

La statistica inferenziale si suddivide poi in altri capitoli, di cui i più importanti sono la teoria della stima (stima puntuale e stima intervallare) e la verifica delle ipotesi.

statistica

Wikipedia

La **statistica** è una disciplina che ha come fine lo studio quantitativo e qualitativo di un particolare fenomeno collettivo in condizioni di incertezza o non determinismo, cioè di non completa conoscenza di esso o parte di esso.

Strumento del metodo scientifico, si avvale della **matematica** per studiare i modi in cui un fenomeno collettivo può essere sintetizzato e compreso e ciò avviene attraverso la raccolta e l'analisi delle informazioni relative al fenomeno studiato

La scienza statistica è comunemente suddivisa in due branche principali:

- statistica **descrittiva**: ha come scopo quello di sintetizzare i dati attraverso … strumenti grafici (diagrammi a barre, a torta, istogrammi,…) e indici (indicatori statistici, indicatori di pos di forma, …) che descrivono gli aspetti salienti dei dati osservati

Inferenza statistica

- statistica **inferenziale**. ha come obiettivo quello di stabilire delle caratteristiche dei dati e dei comportamenti delle misure rilevate (variabili statistiche) con una possibilità di errore predeterminata. Le inferenze possono riguardare la natura teorica (la legge probabilistica) del fenomeno che si osserva. La conoscenza di questa natura permetterà poi di fare una previsione (…ad esempio, …"l'inflazione il prossimo anno avrà una certa entità" … esiste un modello dell'andamento dell'inflazione derivato da tecniche inferenzia nziale è fortemente legata alla t à….

stima dei parametri

test d'ipotesi

La statistica renziale si suddivide poi in altri capitoli, di cui i piu portanti sono la teoria della stima (stima puntuale e stima intervallare) e la verifica delle ipotesi.

statistica

inferenza statistica.

- **stima dei paramenti** (puntuali o di intervalli)

  → estrarre da un campione rappresentativo informazioni sull'intera popolazione
  può essere la stima dei parametri della distribuzione di una variabile casuale (grandezza fisica, misurabile)
  oppure dei parametri che compaiono in una relazione tra variabili casuali (grandezze fisiche)
  a partire da un campione (insieme di misure)

- **test d'ipotesi**

  → estrarre informazioni quantitative sulla validità di un'ipotesi statistica (funzione di distribuzione di variabili casuali, relazioni tra variabili casuali) a partire da un campione rappresentativo

argomenti molto vasti, di cui vedremo solo alcuni metodi e loro applicazioni

saranno trattati nella seconda parte del corso: per affrontarli servono più nozioni di **teoria della probabilità**

darò per scontati i concetti fondamentali di fenomeni statistici, probabilità, variabili casuali, distribuzioni di probabilità e funzioni di distribuzione,….

anche se molte cose le rivedremo

programma sintetico del corso

- fenomeni statistici caratterizzati da una variabile casuale
  - rapido ripasso
  - indicatori statistici, momenti, funzioni generatrici dei momenti
  - alcune funzioni di distribuzione
- più variabili casuali
  - funzioni di distribuzione congiunte
  - funzione di distribuzione multinormale, distribuzione multinomiale
- funzioni di una o più variabili casuali
  - considerazioni generali
  - alcune funzioni importanti: media, somma di quadrati,…
- stima dei parametri
  - metodo del maximum likelihood
  - metodo dei minimi quadrati
  - intervalli di confidenza
- test d'ipotesi
  - test parametrici
  - test di χ^2

con molti esempi, esercizi/esercitazioni, applicazioni

organizzazione

lezioni:

- in aula; se possibile molto alla lavagna
  la vostra partecipazione attiva è molto importante
- lunedì e mercoledì dalle 14 alle 16 (con quarto d'ora accademico)

laboratori:

- consistono nell'analizzare dati reali o generare dati e analizzarli con il computer (ma non è un ''laboratorio di calcolo''…)
- impegno minimo: 3 ore alla settimana, con settimane di recupero

organizzazione

attività durante i laboratori:

- per ogni laboratorio ci sarà un problema specifico da risolvere, generalmente illustrato e discusso prima a lezione
- potrete chiedere eventuali chiarimenti sul problema
- a rotazione tra i presenti, verranno discusse eventuali difficoltà e soprattutto i risultati ottenuti
- verrà compilato l'elenco degli studenti che hanno correttamente completato il lavoro, cosa necessaria per poter sostenere l'esame

- per chi partecipa attivamente ai laboratori, se utile, ci saranno ulteriori spazi per discussione negli altri pomeriggi

organizzazione

laboratori

- stiamo cercando un modo per svolgerne almeno parte in presenza, per una discussione più diretta di risultati e problemi incontrati;
  se non possibile, useremo solo Teams
  e gli strumenti che avete usato a Laboratorio di Calcolo (mobaXterm, più compilatore fortran, editor, gnuplot)
  problemi specifici da segnalare, PC? **da definire !**
  verifica: esercizio preliminare
- per poter seguire tutti bisogna, tentativamente (*quanti studenti?*)
  - formare due gruppi, che faranno laboratorio in due pomeriggi diversi: dovete organizzarvi e scegliere il giorno in modo che non ci siano interferenze interferenze con gli altri laboratori
  - dividere ogni gruppo in due sottogruppi, in due ‘’classi’’ diverse, seguite in parallelo: massimo 20 studenti per sottogruppo
  - possibile orario: lunedì e mercoledì 16-19

  importante avere lo stesso numero di studenti per sottogruppo; e non cambiare (sotto)gruppo per tutto il semestre

  elenco via mail (anna.martin@ts.infn.it) entro questa settimana, specificando gruppo, sottogruppo, giorno

altro

materiale didattico

- appunti
- pagina web del corso: https://wwwusers.ts.infn.it/~martin/univ/didattica/lab2/lab2.html
  - dispense sintetiche
  - elenco libri di consultazione

esame

- necessario aver fatto e discusso tutti gli esercizi proposti durante il corso (verifica durante il corso)
- necessario presentare con una settimana di anticipo una relazione su un lavoro di analisi individuale che includa stima di parametri e test d'ipotesi

  dati: possibilmente da misure in (altri) laboratori

  solo per la prima sessione di esami: relazione sull'ultimo esercizio proposto
- all'orale: domande sulla parte teorica, sugli esercizi, sulla relazione

- un unico esame per Laboratorio 2, con un unico voto

prima di iniziare il corso vero e proprio

1. un esercizio di **teoria della probabilità**
   applicazione del teorema di Bayes

2. formule per la **stima dei parametri di una retta**
   utili da subito per le misure

3. **esercizi preliminari** da fare subito
   necessari per lavoro successivo

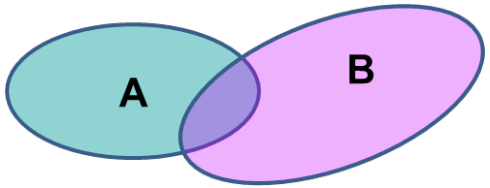
prima di iniziare il corso vero e proprio

1. un esercizio di **teoria della probabilità**
   applicazione del teorema di Bayes

   noi statistica classica

prima di iniziare 1

teorema di Bayes $$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$



diagnostica di laboratorio: popolazione s e m, test p o n

sensibilita' di un test: probabilita' risultato p se m $S_e = P(p|m) \sim 0.95$

specificita': probabilita' risultato n se s $S_p = P(n|s) \sim 0.95$

misurate su un campione di S e M (definizione frequentista)

a noi interessa: probabilita' m se p $P(m|p)$

$$P(m|p) = \frac{P(p|m)P(m)}{P(p)}$$

$$P(p) = P(p|s)P(s) + P(p|m)P(m) = \big(1 - P(n|s)\big)P(s) + P(p|m)P(m)$$

serve la **prevalenza** $P_r = P(m) \sim 0.90, 0.50, 0.20, 0.10, \ldots$

$$P(m|p) = \frac{S_e}{(1 - S_p)(1 - P_r) + S_e P_r} \cdot P_r$$

$$P(s|n) = \frac{S_p}{S_p(1 - P_r) + (1 - S_e)P_r} \cdot (1 - P_r)$$

prima di iniziare 1

$$\mathbf{P(m|p)} = \frac{\mathbf{S_e}}{\mathbf{(1-S_p)(1-P_r)+S_eP_r}} \cdot \mathbf{P_r} \qquad = \frac{0.95}{0.05 \cdot 0.90 + 0.95 \cdot 0.10} \cdot 0.10 = 0.68$$

$$\mathbf{P(s|n)} = \frac{\mathbf{S_p}}{\mathbf{S_p(1-P_r)+(1-S_e)P_r}} \cdot \mathbf{(1-P_r)} \qquad = \frac{0.95}{0.95 \cdot 0.90 + 0.05 \cdot 0.10} \cdot 0.90 = 0.99$$

P_r	0.10	0.20	0.50	0.90
$P(m\|p)$	0.68	0.83	0.95	0.99

P_r	0.10	0.20	0.50	0.90
$P(s\|n)$	0.99	0.99	0.95	0.68

importanza della scelta / comoscenza del “compione”

due test?

prima di iniziare il corso vero e proprio

1. un esercizio di **teoria della probabilità**
   applicazione del teorema di Bayes

2. formule per la **stima dei parametri di una retta**
   utili da subito per le misure

prima di iniziare 2

anticipando quello che faremo nella seconda metà del corso,

formule per la stima dei parametri di una retta

ipotesi: $Y = mX + q$

dati: n coppie di valori misurati (x_i, y_i)
con incertezze statistiche trascurabili su x_i, e σ_i su y_i

metodo dei minimi quadrati: stima migliore di m e di q valori che minimizzano

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - y_i^t)^2}{\sigma_i^2}, \qquad y_i^t = mx_i + q$$

$$\begin{cases} \dfrac{\partial X^2}{\partial m} = 0 \\ \dfrac{\partial X^2}{\partial q} = 0 \end{cases} \Rightarrow \quad \hat{m} = \frac{1}{D}(S_{00}S_{11} - S_{10}S_{01}) \qquad \hat{q} = \frac{1}{D}(S_{01}S_{20} - S_{11}S_{10})$$

$$D = S_{00}S_{20} - S_{10}^2 \qquad S_{jk} = \sum_{i=1}^{n} \frac{x_i^j y_i^k}{\sigma_i^2}$$

e, con la legge di propagazione della varianza $\sigma_{\hat{m}}^2 = \sum_i \left(\frac{\partial \hat{m}}{\partial y_i}\right)^2 \sigma_i^2$, ...

$$\sigma_{\hat{m}}^2 = \frac{S_{00}}{D} \qquad \sigma_q^2 = \frac{S_{20}}{D}$$

verificare

prima di iniziare 2

note

a) 1. se $Y = mX$ **formule simili, più semplici, facili da ricavare**

b) se è l'incertezza su y_i trascrurabile, **basta scambiare X e Y**

c) se **entrambe le incertezze sono non trascurabili**

al primo ordine
si stimamo una proma volta i parametri come se le incertezze su x_i fossero trascurabili → $\hat{m}_0,\ \hat{q}_0$

si propaga l'errore $\sigma^2_{y_i,x_1} = \hat{m}_0^2\ \sigma^2_{x_i}$, $\sigma'^2_i = \sigma_i^2 + \sigma^2_{y_i,x_i}$
si stimano di nuovo i parametri

….fino ad ottenere risultati stabili

prima di iniziare 2

note

a) 1. se $Y = mX$ **formule simili, più semplici, facili da ricavare**

b) se è l'incertezza su y_i trascrurabile, **basta scambiare X e Y**

c) se **entrambe le incertezze sono non trascurabili**

d) se **la relazione tra X e Y non è lineare**…

spesso di può ''linearizzare''

ad es $V(t) = V_0 e^{-t/c}$ ''campione'' (t_i, V_i)

però $\ln V = \ln V_0 - t/c$

e quindi, introducendo $Y = lnV$ e $X = t$, ci si riconduce al caso precedente

NB: $\sigma_i^2 = \cdots \sigma_{V_i}^2$

prima di iniziare il corso vero e proprio

1. un esercizio di **teoria della probabilità**
   applicazione del teorema di Bayes

2. formule per la **stima dei parametri di una retta**
   utili da subito per le misure

3. alcuni **esercizi preliminari** da fare subito
   necessari per lavoro durante le esercitazioni

   faremo la prossima volta