1) Parliamo della funzione cumulativa empirica(a cosa serve e proprietà).

$$F_n(x)=\frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,x)}(x_i)$$

E' una funzione di variabile reale che rappresenta la funzione di ripartizione della misura empirica di un campione. Difatto e' una stima della vera funzione di ripartizione che ha generato il campione. E' uno stimatore corretto e consistente della funzione di ripartizione.

2) Teorema delle probabilita' totali

L'equazione afferma che P(E) si puo' ricavare come media pesata delle probabilita' condizionali di E sapendo che F si sia vericato e che non si sia verificato

$$P(E)=P(E\vee F)\,P(F)+P(E\vee \overline{F})\,P(\overline{F})$$

$$P(A)=\sum_{i=1}^{n} P(A\vee B_i)P(B_i)$$

3) Variabile aleatoria normale,panoramica su ciò che sai.

Leggete sul libro o videolezioni

4) Specificazioni possibili che può assumere una variabile aleatoria normale.

$$X : R\mapsto R^{+¿¿}$$

5) Abbiamo due variabili normali X,Y entrambe con valore atteso mu, X ha dev.std =1, e Y ha mu = 3. disegna il grafico di densità di entrambe le variabili.

Si fanno due campane centrate a μ con area della campana sempre a 1 e la campana appiattita proporzionalmente alla deviazione standard.

6) Abbiamo campione di dati, vogliamo trasformarli in modo che il più piccolo sia uguale a zero. Se poi voglio standardizzarli? Dimostrare

Per il punto 1) si sottrae ai campioni il valore minore (k)

Z = X - k

Per il punto 2)

$$Y=\frac{Z-\varepsilon(Z)}{\sigma}=\frac{(X-k)-(\varepsilon(X)-k)}{\sigma}=\frac{X-\mu}{\sigma}$$

7) V.A X e vogliamo calcolare il valore atteso del log(X)

$$\varepsilon(log(x))=\int_{-\infty}^{+\infty} log(x) f_x(x)\,dx$$

8) Esempio di VA discreta con numero infinito di specificazioni.

Una qualsiasi variabile aleatoria distribuita secondo Poisson

9) Abbiamo un campione di n elementi estratto da popolazione di valore atteso mu. Che stimatore utilizzeresti per mu? Vogliamo garantire che l'errore che stiamo facendo nell'approssimare il valore atteso con il valore della media compionaria sia al più 0.1. Immaginiamo n fissato. Cosa possiamo dire della probabilità che l'errore compiuto sia al più 0.1?

$$P\left(\left|\overline{X_n}-\mu\right|\leq 0.1\right)=P\left(\left|\frac{\overline{X_n}-\mu}{\frac{\sigma}{\sqrt{n}}}\right|\leq\frac{0.1\sqrt{n}}{\sigma}\right)=P\left(|Z|\leq\frac{0.1\sqrt{n}}{\sigma}\right)=2\phi\left(\frac{0.1\sqrt{n}}{\sigma}\right)-1$$

10) Abbiamo un campione di dati, vogliamo tracciare un boxplot di questi dati. Come si fa?

```
import matplotlib.pyplot as plt

plt.box
```

Il box plot e' composto dalla scatola che ha i dati al primo e terzo quartile e una linea che rappresenta la mediana. I baffi sono i massimi e i minimi. Eventuali puntini sono gli outlier.

11) Concetto di quantile.

In un campione ordinato si dice quantile di ordine α con $\alpha \in [0,1)$ un valore q_α che divide la popolazione in due parti proporzionali caratterizzate dai valori > o < di q_α

I percentili sono di ordine n/100

La mediana e' un quantile di ordine 1/2

12) Abbiamo urna che contiene 10 palline, 3 bianche e 7 nere. Faccio 3 estrazioni. Probabilità che la terza palla sia bianca? Consideriamo il caso con reimmissione. Caso senza reimmissione?

P(con reimmissione) = $\frac{3}{10}$

P(senza reimmissione) =
$P\left(B_3 \vee B_1 \cap B_2\right)P\left(B_1 \cap B_2\right)+P\left(B_3 \vee \overline{B_1} \cap B_2\right)P\left(\overline{B_1} \cap B_2\right)+P\left(B_3 \vee \overline{B_1} \cap \overline{B_2}\right)P\left(\overline{B_1} \cap \overline{B_2}\right)+P\left(B_3 \vee B_1 \cap \overline{B_2}\right)P\left(B_1 \cap \overline{B_2}\right)$

DA VEDERE

13) Cos'è uno stimatore, cosa serve e quali caratteristiche ci piacerebbe avesse.

Uno stimatore puntuale e' una funziona che associa ad ogni possibile campione un valore del parametro da stimare. Vorremmo che fosse non distorto o per lo meno consistente.

14) Campione di coppie di attributi, valutare se tra queste quantità c'è relazione.

Coefficiente di correlazione lineare o inverso

$$R=\frac{cov(x,y)}{\sigma_x \sigma_y}=[-1;1)$$

Se $R \to 1$ si ha una relazione diretta, se $R \to -1$ la relazione e' inversa, se $R \to 0$ non c'e' relazione lineare

15) Due eventi A e B. Probabilità loro unione?

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$ altrimenti

$P(A \cup B) = P(A) + P(B)$ nel caso di eventi DISGIUNTI

16) Variabile aleatoria Bernoulliana di parametro s. Grafico funzione di ripartizione.

Vedi Wikipedia

17) Parliamo della funzione cumulativa empirica.

Vedi domanda 1)

18) Perchè abbiamo introdotto la funzione di ripartizione empirica,ovvero, a cosa serve?

Vedi domanda 1)

19) Come è definita la funzione di ripartizione nel caso di una variabile aleatoria?

Vedi 1) + in generale la prima e' una stima non distorta mentre la seconda e' una funzione vera e propria

20) Differenze tra funzione di ripartizione empirica e funzione di ripartizione di una distribuzione

$P(X \leq x)$

21) Posso usare questo strumento (punto 20) come alternativa al qq-plot?

Si perche' il qq-plot serve ugualmente a rappresentare una funzione cumulativa

22) Definizione deviazione standard campionaria

$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i -\overline{x_n})^2} $

Viene usata quando si vuole ottenere un valore che abbia la stessa unità di misura dei valori osservati.

23) Cosa è la non distorsione?

Il valore atteso di uno stimatore non distorto e' sempre uguale al valore effettivo del parametro da stimare

24) Definizione di stimatore non deviato per una certa quantità

Vedi 23)

25) Parliamo in generale di indici di centralità e di dispersione nell'ambito della statistica descrittiva

Gli indici in generale servono per descrivere le proprieta' di una distribuzione statistica.

Come indici di centralità abbiamo:

- Media
- Mediana
- Moda

Come indici di dispersione abbiamo:

- Varianza/Deviazione Standard
- Covarianza
- Coefficiente di variazione
- Range interquartile

26) Cosa sono il primo e il terzo quartile?

Sono i quantili a 0.25 e 0.75

27) A partire da un insieme di dati si vuole descrivere questi dati fornendo indice di centralità e di dispersione. Come ti comporteresti nello scegliere un indice di centralità e un indice di dispersione da accoppiare insieme?

Media come indice di centralità e deviazione standard come indice di dispersione

28) Perchè è più sensato usare la deviazione standard da accoppiare alla media, piuttosto che usare la varianza?

Perchè la deviazione standard ha la stessa unita' di misura della media mentre la varianza e' al quadrato.

29) Regola empirica, cosa dice?

La regola empirica dice che in una distribuzione simmetrica a forma di campana (tipo normale) il 68% dei valori si discosta da $\varepsilon(x)$ di $\pm\sigma$, del 95% di $\pm 2\sigma$, del 99.7% di $\pm 3\sigma$

30) Nella statistica inferenziale la media campionaria è stimatore. Cosa stima, proprietà?

Stima il valore atteso e come proprietà e' non distorto e consistente, come conseguenza del teorema della legge dei grandi numeri, per n abbastanza grande e' approssimabile a una distribuzione normale di parametri μ e $\frac{\sigma}{\sqrt{n}}$

31) Abbiamo insieme di dati, voglio trasformarli in modo che il più grande diventi uguali a 1.

Divido tutti i dati per il piu' grande

32) Definisci la trasformazione di scalatura. Dimostra che l'operazione di scalatura si può applicare direttamente alla media campionaria.

$\forall\, i\ y_i = k\, x_i$

$\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} x_i$

$$\overline{Y_n} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n}(k\,x_i) = \frac{k}{n}\sum_{i=1}^{n}(x_i) = k\,\overline{X_n}$$

33) Enunciami il teorema delle probabilità totali

Vedi 2)

34) Cosa descrive il modello binomiale

Il modello binomiale descrive una ripetizione di n eventi bernoulliani. Puo' essere usato in pratica per delle estrazioni con reimmissione

35) Popolazione descritta da modello esponenziale di parametro lambda. Proporre stimatore per lambda.

Il modello esponenziale descrive il tempo di attesa prima che si verifichi un determinato evento casuale. La distribuzione esponenziale e' priva di memoria (il fenomeno non invecchia). Un esempio e' la durata della vita di una particella radioattiva prima di decadere.

$$\overline{X_n} = \varepsilon(X) = \frac{1}{\lambda} \rightarrow \lambda = \frac{1}{\overline{X_n}}$$

Il reciproco della media campionaria e' quindi uno stimatore non distorto di λ

36) Dato un insieme di dati,con quali strumenti puoi valutare se seguono una distribuzione normale?

Si puo' capire tracciando il grafico della funzione di densita' e vedendo se e' simmetrica, ha una forma a campana ed e' centrata nel valore atteso, oppure si puo' usare un istogramma per le frequenze relative e un box plot

37) Cosa dice il teorema centrale del limite?

Avendo una successione di n variabili aleatorie indipendentie identicamente distribuite la loro somma si puo' approssimare ad una distribuzione normale di parametri $n\,\mu$ e $\sigma^2 n$. Si puo' applicare il processo di normalizzazione per passare ad una normale standard. Piu' grande e' n migliore e' l'approssimazione

38) Variabili aleatorie X e Y. Valore atteso di X+Y?

$\varepsilon(X+Y) = \varepsilon(X) + \varepsilon(Y) $

Per linearità

39) Concetto di eterogenità

Nel caso di eterogenità massima abbiamo tutte le occorrenze con la stessa frequenza

Nel caso di eterogenità minima (o omogeneita' massima) una sola occorrenza di frequenza massima

L'eterogenità si puo' calcolare con l'indice di Gini

$$I = 1 - \sum_{i=1}^{n} f_i^2$$

Con f_i frequenze relative. L'indice di Gini e' compreso tra 0 e 1 e si ha il massimo dell'eterogenità a $\frac{n-1}{n}$, minima a 0

40) Alberi di decisione

Gli alberi di decisione si usano per classificare dei dati in delle classi basandosi sui dati a disposizione. Partendo dalla radice, si definiscono dei nodi con delle condizioni binarie, andando a diminuire l'eterogeneita' mano a mano che si percorre l'albero fino ad arrivare a dei nodi foglia nei quali l'eterogeneita' e' 0.

41) La varianza è operatore lineare?

No, infatti
$var(X+a) = \varepsilon\left((X+a)^2\right) - \varepsilon(X+a)^2 = \varepsilon\left(X^2 + 2aX + a^2\right) - \left(\varepsilon(X)+a\right)\left(\varepsilon(X)+a\right) = \varepsilon\left(X^2\right) + 2a\varepsilon(X) + a^2 - \left(\varepsilon(X)^2 + 2a\varepsilon(X) + a$

Oppure

$$var(Xk) = k^2 var(X)$$

42) Dimostra che il valore atteso di una costante è la costante stessa

$$\varepsilon(a) = \sum_{i=1}^{1} (a)P(a)$$

Siccome la probabilita' di una costante e' sempre 1 allora $\varepsilon(a) = a$

43) Varianza del prodotto di due cose (Esempio di una costante per una variabile aleatoria)

$$var(Xa) = a^2 var(X)$$

44) Varianza del prodotto di due variabili aleatorie

$$var(XY) = \varepsilon(X)^2 var(Y) + \varepsilon(Y)^2 var(X) + var(X) var(Y) = \varepsilon\left(X^2\right)\varepsilon\left(Y^2\right) - \varepsilon(X)^2 \varepsilon(Y)^2$$

45) Normalizzazione variabile aleatoria Bernoulliana. Quello che ottengo è una variabile aleatoria normale standard?

No per standardizzare una variabile alleatoria X facciamo semplicemente $\frac{X-\mu}{\sigma}$ questo non cambia la distribuzione, rimane una bernoulliana

46) A cosa serve un classificatore binario e come ne valuto la sua performance

A classificare dei dati in due classi (esempio falso/vero positivo/negativo).

Ne valuto la perfomance usando la matrice di confusione, valutando specificita', sensibilita' e la curva ROC.

47) Abbiamo due variabili aleatorie X e Y e conosciamo solo il grafico della loro funzione di ripartizione. Disegnami il grafico di una funzione di ripartizione di una v.a. discreta. Sopra al grafico disegna il grafico di una funzione di ripartizione di una v.a. continua che deve essere definita sull'intervallo che va da 0 a 3. Guardando il grafico sai dirmi la relazione che c'è tra il valore atteso di queste due variabili aleatorie?

Come esempio di una discreta si puo' fare la binomiale, per la continua una uniforme continua. La relazione e' l'area sopra la F_x che puo' essere piu' grande per una delle due

48) Abbiamo campione aleatorio, stimare valore atteso della popolazione usando la varianza campionaria.

E' una cosa che normalmente non avrebbe senso fare. In alcuni casi potrei usare la varianza come stimatore non distorto del valore atteso (ad esempio poisson dove $\mu=\sigma^2$. Funziona solo per relazioni lineari.

49) Grafico ripartizione Bernoulliana

Andatevelo a vedere sfaticati

50) Ti ho fornito un dataframe di pandas e un attributo contenga dei dati di cui ti chiedo quanto è sensata l'ipotesi che provengano da un modello normale, cosa potresti fare?

Faccio un istogramma, NON GRAFICO A BARRE, in quanto il modello normale è modello continuo. Oppure calcolo media e mediana e vedo se sono simili.

51) Probabilità intersezione di due eventi

$P(A \cap B)=P(A)P(B)$ se sono indipendenti

$P(A \cap B)=P(A \vee B)\,P(B)$ se A dipende da B

$P(A \cap B)=P(B \vee A)\,P(A)$ se B dipende da A

52) Grafico funzione di ripartizione di una Bernoulliana di parametro m evidenziami probabilità x>1/2

$$F(x)=P(X \leq x)=P\left(X \leq \frac{1}{2}\right)=1-m$$

$$P\left(X \geq \frac{1}{2}\right)=1-P\left(X \leq \frac{1}{2}\right)=m$$

53) Abbiamo due eventi A e B. Probabilità loro unione e intersezione.

Per l'intersezione vedi 51)

Per l'unione vedi 15)

54) Grafico funzione di densità di due variabili aleatorie normali che hanno stesso valore atteso e diversa dev.std.

Hanno entrambe il centro in μ ma quella con la dev.standard maggiore ha la campana piu' schiacciata

55) Proprietà funzione di densità

La funzione di densita' p_x rappresenta il valore della probabilita' intorno ad x.

Si puo' vedere come $P(X \in A) = \int_A p_x(x)\,dx$

Vale per le distribuzione continue e $\int_{-\infty}^{+\infty} p_x(x)\,dx = 1$

56) Abbiamo dataframe pandas, e che abbia colonne con peso e altezza. Verificare se esiste relazione, cosa puoi fare?

Plot.scatter(attributo1,attributo2)

```
import matplotlib.pyplot as plt

plt.scatter
```

57) Urna di 10 palle, 7 bianche e 3 nere. Estraggo 3 palle, cosa posso dire della probabilità di estrarre palle bianche alla terza estrazione

Vedi 12)

58) Variabile aleatoria X e vogliamo standardizzarla. Come si fa? Dimostra che dopo la standardizzazione il valore atteso e' uguale a 0.

Se $X \sim N(\mu, \sigma)$:

$$Y = \frac{x - \mu}{\sigma}$$

$$\varepsilon(Y) = \varepsilon\left(\frac{X - \mu}{\sigma}\right) = \frac{\varepsilon(X) - \mu}{\sigma} = \frac{0}{\sigma}$$

$$var(Y) = var\left(\frac{X - \mu}{\sigma}\right) = \frac{var(X - \mu)}{\sigma^2} = \frac{var(X)}{\sigma^2} = 1$$

Altrimenti:

$$X = \sum_{i=1}^{n} X_i$$

$$\varepsilon(X_i) = \mu$$

(Teorema centrale del limite)

$$\frac{X - n\mu}{\sigma\sqrt{n}} \dot{\sim} N(0,1)$$

$$\varepsilon\left(\frac{X - n\mu}{\sigma\sqrt{n}}\right) = \frac{\sum_{i=1}^{n} \varepsilon(X_i) - n\mu}{\sigma\sqrt{n}} = \frac{n\mu - n\mu}{\sigma\sqrt{n}} = \frac{0}{\sigma\sqrt{n}}$$

59) Immagina di avere un dataframe su pandas con un certo attributo e volerlo standardizzare.

d # DataFrame con colonna altezza

s = (d.altezza - d.altezza.mean())/d.std()

60) Applica la standardizzazione su una variabile X e chiamare il risultato Y. Che proprieta' ha? Sai dimostrarmelo?

Vedi 58)

61) Dimostra che $\varepsilon(X+Y) = \varepsilon(X) + \varepsilon(Y)$

$$\varepsilon(X+Y) = \sum_x \sum_y (x+y)\, p_{x,y}(x,y) = \sum_x \sum_y x\, p_{x,y}(x,y) + y\, p_{x,y}(x,y) = \sum_x \sum_y x\, p_{x,y}(x,y) + \sum_x \sum_y y\, p_{x,y}(x,y) = \sum_x x$$

62) Che proprieta' vorremo che avesse uno stimatore?

Correttezza e coerenza

63) Cos'e' il bias?

$$b_T(\tau(\theta)) = \varepsilon(T) - \tau(\theta)$$

In altre parolle: $ b_T(parametro da stimare) = \varepsilon(Stimatore) - parametro$

Se bias = 0 → stimatore è non distorto

64) Cos'e' l'MSE?

$$MSE_T(\tau(\theta)) = var(T) + b_t(\tau(\theta))^2$$

In altre parolle: $ MSE_T(parametro) = var(Stimatore) + bias(parametro)^2$

Se per n → +∞, $MSE_T \to 0$ allora stimatore ha proprieta di consistenza quadratica

65) Cos'e' una curva ROC? (Piu' altre cose riguardo matrice di confusione e specificita'/sensibilita')

66) Abbiamo un dataframe di pandas e vogliamo valutare se I valori contenuti in una colonna sono compatibili con distr normale

Vedi 50)

67) Stimare il parametro di una esponenziale

Vedi 35)

68) Teorema centrale del limite

Vedi 37)

69) Dataframe pandas, vogliamo selezionare tutte le righe per cui l'attributo eta' >18

res = df[df.eta > 18]

70) Rispetto alla 69) vogliamo vedere se ci sia una relazione tra prima apparizione ed eta'

71) Combinazioni, permutazioni, disposizioni

Permutazioni $n!$ disposizioni con $k=n$

Disposizioni $\frac{n!}{(n-k)!}$ senza ripetizioni e con ordine

Combinazioni $\binom{n}{k}$ senza ripetizioni e senza ordine

72) L'area sopra la funzione di ripartizione e' sempre il valore atteso?

No, non e' il caso della normale perche' il dominio non e' >0

73) Potrei utilizzare varianza campionaria per stimare mu ?

No. Ci sono però dei casi particolari in cui potrebbe essere una buona scelta, ed è il caso in cui la media coincide con la varianza(esempio poissoniana).

74) Hai un dataframe pandas traccia il box plot di una sua colonna.

```
x = df['colonna']
x.plot.box()

import matplotlib.pyplot as plt
x=df["colonna"]
plt.boxplot(x)
```

75) Cos'è un box plot? La media è sempre al centro della scatola?

Il boxplot è un grafico che rappresenta i valori "divisi" per quantili, in quanto è rappresentato come una scatola, i cui estremi corrispondono al primo quantile e al terzo quantile; la linea in mezzo alla scatola corrisponde al secondo quantile, cioè la mediana, il minimo e il massimo corrispondono ai baffi (in alcune implementazioni di boxplot, questi baffi possono non raggiungere il massimo in quanto considerato outlier).

Non è detto che la media sia al centro della scatola

76) Come verifico che una dataset sia normale?

Traccio il grafico della funzione di massa di probabilità e vedo se ho la campana ovvero che il grafico abbia un picco al centro e sia simmetico. Chiaramente accettiamo un certo errore sperimentale, la curva non sarà, ad esempio, **perfettamente** simmetrica.

Altrimenti posso usare il qqplot (presente in scipy.stats.api) per confrontare i quantili reali del dataset con quelli teorici.

77) Se il mio dataset è molto preciso e quindi tutti i valori sono unici? (rispetto alla 76)

Considero degli intervalli e faccio un istogramma

78) Come decido se valutare specificazioni singole o intervalli?

Dipende se la distribuzione è continua o discreta

79) Cosa sai della distribuzione geometrica?

La distribuzione geometrica rappresenta quanti fallimenti di una variabile aleatoria bernoulliana bisogna fare prima che avvenga il primo successo, X~G(p), gli esperiementi devono essere indipendenti fra di loro,

$$p_X(x)=p(1-p)^x, P_X(x)=1-(1-p)^{x+1}, E(X)=\frac{1-p}{p}, V\,A\,R(X)=\frac{1-p}{p^2}$$

80) Hai un urna con 10 palline bianche e 10 nere, ne estrai 3, qual'è la probabilita che la terza sia bianca?

Dipende se c'è reimmisione della pallina pescata o meno, nel caso che la pallina sia reinserita, la probabilità di pensare la terza pallina bianca è pari a P(3 pallina bianca)=1/2.

Nel caso non ci fosse reimissioni, allora devo vedere che palline ho pescato nella prime due pescate, per vederla risolta vedete la domanda 12

81) Teorema centrale del limite?

Il teorema centrale del limite mi dice che se ho una distribuzione di variabili aleatorie indipendenti e identicamente distribuite,con $E(X)=\mu$ e varianza pari a $V\,A\,R\,(X)=\sigma^2$, allora tale distribuzione si può approssimare a una variabile aleatoria normale di paramentro

$$X \sim N(n\mu, \sqrt{n}\sigma)$$

82) Taglia del campione

Dato una distribuzione normale(o una tale che, attraverso il teorema centrale del limite, sia approsimabile a lei), mi dice quanti valori devo avere affinché l'errore dato sia minore di un errore dato.

83) Data una variabile alleatoria X, standardizzandola ottengo Y. Che propietà ha Y?

Y in tal caso ha $E(Y)=0$ e $\sigma^2=1$

84)Teorema di Bayes

$$P(B \vee A) = \frac{P(A \vee B) P(B)}{P(A)}$$

85) Classificatori binari

86) Il concetto di consistenza

Abbiamo due tipi di consistenza, la consistenza debole, che mi dice che

$$\lim_{n\to \infty}P(|T_n(X)-\tau(\theta)|\leq\epsilon)=1 \space\space \forall\epsilon>0$$

La consistenza forte mi dice che

$$\lim_{n \to \infty} P\left(T_n(X) = \tau(\theta)\right) = 1$$

Si noti che la legge dei grandi numeri è un caso particolare di queste formule.

Bonus 2): Come faccio a vedere se un campione e' compatibile con una distribuzione teorica

Si normalizza e poi si sovrappongono i grafici