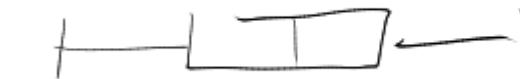


Primo studente:

Alla prima domanda: il parametro lambda può essere negativo? La risposta "no perché X sarebbe negativo ed è impossibile": mi scriva la funzione di densità del modello esponenziale (domanda retorica: ha sbagliato risposta). La funzione di densità ad esempio $f(3)$ non indica la probabilità che X valga 3 (ragionare bene: forse perché siamo nel continuo e non nel discreto). Mi disegni il grafico della funzione di densità del modello esponenziale: attenzione quando x vale 0 il grafico vale lambda. Qual è il dominio di una variabile aleatoria: attenzione alla differenza tra variabile continua e discreta: nel caso discreto sono ad esempio i valori 1, 2, 3, 4... Nel caso continuo (nell'esempio esponenziale) vale $[0, +\infty)$. Alla domanda "qual è il dominio della variabile aleatoria esponenziale?" vuole sentirsi dire "il dominio della funzione di densità".

Immagina di avere due attributi, uno è scalare e uno è categorico. Ha senso fare lo scatter plot di questi attributi? IL tempo di guasto è scalare, modello è categorico: se uno in questo esercizio fa lo scatter plot non è il grafico adatto al contesto. Ora a partire dallo scatter plot che ha fatto lo studente all'esame: puoi disegnare i 3 boxplot? Dato un boxplot cosa notiamo? Che la mediana doveva essere più a sinistra. Stimatore non deviato: definizione formale. Ora dalla definizione: perché l'assenza di deviazione è desiderabile? Il valore atteso indica la centralità e vogliamo che il suo valore "centrale" sia quello che vogliamo stimare. Ora dato che T_n è anch'essa una variabile aleatoria, varia.



attributi, uno è scalare e senso fare lo scatter plot tempo di guasto è categorico: se uno in scatter plot non è il

grafico adatto al contesto. Ora a partire dallo scatter plot che ha fatto lo studente all'esame: puoi disegnare i 3 boxplot? Dato un boxplot cosa notiamo? Che la mediana doveva essere più a sinistra. Stimatore non deviato: definizione formale. Ora dalla definizione: perché l'assenza di deviazione è desiderabile? Il valore atteso indica la centralità e vogliamo che il suo valore "centrale" sia quello che vogliamo stimare. Ora dato che T_n è anch'essa una variabile aleatoria, varia.

Nell'esame: "c'è uno stimatore non distorto per la dev st e varianza". Perché la dev st campionaria è uno stimatore non distorto per la dev st? La risposta era "lo stimatore non distorto per la deviazione standard è la media campionaria" (questo perché sia la media che la dev std per la esponenziale è $1/\lambda$). Perché calcolare il valore atteso della dev standard campionaria è difficile? Perché nella formula c'è una radice che non è lineare. Lo stimatore e una stima sono cose diverse.

Studente 2

Data una var al X, se le applichiamo una trasformazione per ottenere una nuova var Y (dove $Y = cX$) come calcolare la funzione di densità di Y conoscendo quella di X?

X , $Y = Xc$ è cont, conosc. $f_X(x)$, come si calcolò $f_Y(x)$?

Risposta dello studente(sbagliata):

$$f_Y(x) = c \cdot f_X(x)$$

Perché vale? (è sbagliata come cosa: domanda retorica), in questo modo raddoppia la densità, la risposta era:

$$f_Y(x) = f_X\left(\frac{x}{c}\right)$$

Nell'esame si dice che un attributo segue una legge esponenziale. L'idea dell'esame era: confuta l'ipotesi che segue una legge normale e avvalora l'ipotesi che segue la legge esponenziale. Grafico QQ: attenzione: confronta un campione con un (teorico) e non con un altro campione. Quindi

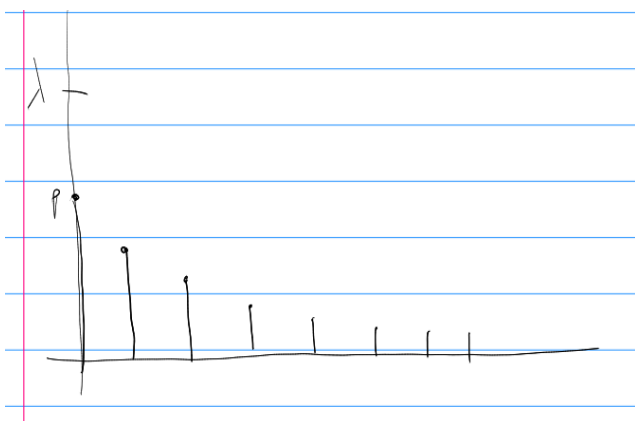
confronta quantili campionari e quantili teorici. Praticamente come si disegna un diagramma QQ? A partire dai quantili. Il modello teorico scelto da confrontare con un certo campione rende più semplice o complesso il confronto? No. Ora si mettono tutti i quantili nel grafico? Ogni elemento di un campione è un quantile di che livello? Se ho 1 elemento e prendo il secondo: che livello corrisponde? A 0.3. Quasi sempre si usa il QQ plot con la normale standard. E se i dati non seguono la distribuzione standard, come confrontarlo (nel QQ) con il modello teorico normale standard. Cosa posso fare? Si può standardizzare. In cosa consiste la standardizzazione? Standardizzare una variabile aleatoria e standardizzare un campione non sono la stessa cosa: attenzione ai termini. Qual è l'effetto della standardizzazione di una v.a.? Rende media = 0, e std = 1. E se standardizzo una normale?

Terzo studente: teorema centrale del limite: attenzione: ha detto "dati n campioni..." doveva dire "date n variabili aleatorie iid, la sommatoria...". È un risultato approssimato. Cosa misura la bontà di questa approssimazione? (boh), quando n va a infinito quel "quasi distribuito secondo un modello normale": vuole sentire la parte in cui si standardizza: si toglie la sommatoria delle medie e si divide per radice di n per sigma. Diventa esattamente distribuita secondo una normale standard. Perché si può fare solo con il limite standardizzato e non prima? Perché avrei avuto i parametri infiniti e non ha senso $N(\infty, \infty)$ non ha senso.

Proprietà di assenza di memoria: cosa ci dice? Mi dimostri che vale per il modello esponenziale (vuole che sappiamo a memoria tutti passaggi che ha fatto lui a lezione). Si ottiene un risultato simile se X è distribuita in modo normale? (boh). Quale altro modello ha la proprietà di assenza di memoria? Geom.

Se dobbiamo sovrapporre il grafico di densità di una geometrica e di una esponenziale con stesso valore atteso come si comportano? Prima però: mi disegni i due grafici separatamente: mi disegni la geometrica (densità). Attenzione: la geometrica non ha una curva continua (il numero di successi / insuccessi è discreto). Ora mi sovrapponga la esponenziale con stessa media.

$$\frac{1}{\lambda} = \frac{1-p}{p} \Rightarrow \lambda = \frac{p}{1-p}$$



Quindi lambda è maggiore di p. Quindi a questo punto basta disegnare stando attenti che sull'asse y lambda (punto di partenza del grafico esponenziale) sia sopra p. Si aspetta una cosa del genere:

Si aspetta che uno sappia risolvere una equazione esponenziale (rigirandola col logaritmo). (grafico a sx in cui manca l'esponenziale)

Parlami di ANOVA. Sapere bene SST SSW SSB. Una volta definite queste in cosa consiste l'analisi della varianza? Formula $SST = SSW + SSB$ (non ho scritto i denominatori): cosa succede se $SSB = 0$?

Fammi un esempio pratico in cui possiamo applicare ANOVA. Perché uno dovrebbe fare tutta sta cosa di ANOVA? Cosa vogliamo capire? (boh).

Terzo studente: una volta assodato che il tempo di guasto di un modello o gruppo di modelli è estratto da una esponenziale, una volta fatta questa ipotesi, abbiamo poi stimato il parametro e possiamo fare ipotesi sul futuro, nell'esame: "qual è la prob che in futuro funzioni?", data X distribuita secondo un modello esponenziale e rappr il numero di mesi prima che si guasti un HDD, qual è la prob che duri più di un anno? Dato che X conta i mesi: $P(X > 12)$, maggiore o maggiore uguale? Non è importante in quanto l'esponenziale è un modello continuo,

Ora se
$$P(X > 12) = 1 - P(X \leq 12) = 1 - F_X(12) = e^{-\lambda 12} = p$$

abbiamo 2 hard disk? Come calcolare che almeno uno sarà funzionante? Calcoliamo la prob dell'evento complementare (1 meno entrambi funzioneranno), ora i due eventi sono indipendenti (potrebbero non esserlo)

$$\begin{aligned} & 1 - P(\text{entr. sono guasti}) \\ & 1 - P(P \text{ guasto} \cap S \text{ guasto}) \\ & 1 - P(P \text{ guasto}) P(S \text{ guasto}) \quad \text{ind} \\ & 1 - P(\text{guasto})^2 = 1 - (1 - e^{-\lambda 12})^2 \end{aligned}$$

Dove P e S sono i due hard disk.

Ora vogliamo calcolare la prob che S funzioni dato che p è guasto: è una prob condizionata:

$$P(S \text{ funz} \mid P \text{ guasto}) = \frac{P(X_S > 12 \mid X_P \leq 12)}{P(X_P \leq 12)}$$

Ora non servono questi calcoli: se sono indipendenti: se uno è rotto non influenza.

Ora parliamo di statistica inferenziale: vogliamo stimare la varianza di una popolazione, come rappresentiamo il campione? In questo caso lo rappresentiamo come un insieme di variabili aleatorie (non è sempre così). La popolazione è iid? No: le variabili del campione sono iid come la variabile aleatoria che rappresenta il campione. Noi vogliamo stimare la varianza, il prof propone questo stimatore:

$$t(x_1, \dots, x_n) = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2$$

Questo è uno stimatore per la varianza? Certo che può esserlo (non ha chiesto se deviato o non deviato). È deviato o no? È deviato: (andava messo $n-1$ sotto): mi mostri che lo stimatore così definito (con $1/n$) è deviato. Definizione di stimatore: non solo ha come argomenti gli elementi dal campione ma non deve dipendere da valori ignoti: perché sarebbe un problema se ciò avvenisse? Se uso come stimatore $T_n = \mu$, ho una variabile che non conosco (μ), quindi il suo valore atteso è sempre μ . Quindi tornando alla domanda di prima: in immagine c'è uno stimatore: no. Sarebbe uno stimatore se avessi ad esempio 42 al posto di μ .

Alberi di decisione: ... noi vogliamo che ci sia una alta omogeneità di cosa? (dei sottogruppi generati). Ora esempio pratico: voglio costruire un albero binario: devo decidere la prima condizione: come la scelgo? Ora supponiamo che scelgo una soglia e calcolo l'eterogeneità dei due sottogruppi generati, come gestisco i due valori? Minimizzo cosa? Si fa la media pesata con il numero di individui. Come si calcola l'eterogeneità? Abbiamo Gini e entropia. Gini, formula, è compreso tra quali valori? Indice di Gini normalizzato. Ora "convincimi che questo numero misura l'indice di eterogeneità": vuole che si provi l'indice nei casi estremi: massima omogeneità e massima eterogeneità ($0, k-1/k$).