

# Orali del 11 settembre 2020

Alessandro Di Gioacchino

5 febbraio 2021

Cos'è un diagramma QQ? A cosa serve? Come lo generiamo? I dati si allineano su una qualunque retta o deve essere parallela alla bisettrice del primo-terzo quadrante? Perché il fatto che esista una relazione lineare tra il quantile del primo campione ed il quantile del secondo campione per ogni fissato livello avvalora l'ipotesi che i due campioni siano estratti dalla stessa distribuzione? Perché posso dire che la distribuzione è uguale solo se i quantili si allineano sulla bisettrice del primo-terzo quadrante? Se i percentili sono uguali indipendentemente dai valori, cosa posso dire?

Che differenza c'è tra combinazione e disposizione? Perché ne esistono due tipi ciascuna? Cosa cambia tra la versione con ripetizione e senza? A cosa serve il vincolo  $k < n$ ? Come conteggio il numero di disposizioni di  $n$  oggetti su  $k$  posti con e senza ripetizione? Perché  $\frac{n!}{(n-k)!}$  è il numero di disposizioni senza ripetizione di  $n$  oggetti in  $k$  posti? Perché calcolando  $n!$  ottengo il numero di permutazioni di  $n$  oggetti? Quante scelte ho per l'ultimo oggetto? Solo una. Perché moltiplicando le possibili scelte ottengo il numero di permutazioni? Possiamo ragionare in questo stesso modo per calcolare il numero di disposizioni: al primo posto abbiamo  $n$  scelte, al secondo  $n - 1$ , e così via fino al  $k$ -esimo dove ne ho  $n - k + 1$ ; la moltiplicazione di questi valori coincide con  $\frac{n!}{(n-k)!}$

Statistica inferenziale. Cosa vuol dire che uno stimatore è non distorto rispetto ad una quantità ignota? Perché è desiderabile che uno stimatore goda di questa proprietà?  $\tau(\Theta)$  è la stima del parametro la quantità che voglio stimare? Perché ci sono  $\Theta$  e  $\tau(\Theta)$ ? Il primo è il parametro ignoto della distribuzione. Quando uno stimatore  $T$  è non deviato per  $\Theta$ ? (Ci stiamo momentaneamente dimenticando della funzione  $\tau$ , c'è solo il parametro)

Le  $x$  in input allo stimatore sono variabili aleatorie o valori numerici? Cosa ottengo da questo calcolo?  $s$ , la stima della quantità.  $\tau(\Theta)$  è quanto voglio stimare. Quindi ci aspettiamo che  $s$  sia una buona approssimazione per  $\tau(\Theta)$

Cosa vuol dire che uno stimatore è non distorto? Perché ci piace che lo sia? Cosa indicano  $\Theta$  e  $\tau(\Theta)$ ? Esempio pratico.  $\mu$ , il valore atteso di una popolazione, è un numero reale? Dimentichiamoci del fatto che esista  $\tau$  per stimare solo  $\Theta$ , quindi i valori dello stimatore sono stime per  $\Theta$ : quando posso dire che lo stimatore  $T$  è non distorto? Formalmente come catturiamo il fatto che lo stimatore varia intorno a quanto vogliamo stimare? Supponiamo che la distribuzione sia un'esponenziale, e vogliamo sempre stimarne il valore atteso: cos'è  $\Theta$ ? Cos'è  $\tau(\Theta)$ ?  $\Theta$  è un parametro, cioè  $\lambda$  (in questo caso): siccome non sempre quanto vogliamo stimare è esattamente il valore di un parametro, ma una quantità che dipende da esso, il fatto di non conoscere quel parametro implica di non conoscere neanche quella quantità. Come troviamo uno stimatore per il valore atteso? Avrebbe senso usare la varianza campionaria per stimare il valore atteso di una popolazione esponenziale? Se calcolassimo il valore atteso dello stimatore, che risultato troveremmo supponendo che sia non distorto? È un calcolo che abbiamo già visto: quanto otteniamo come risultato? La varianza della popolazione. Cosa stima la varianza campionaria? 'Calcolare' la varianza campionaria significa "calcolare uno scarto medio:" ma tale scarto non è fatto rispetto al valore atteso delle singole variabili, perché nella varianza campionaria non c'è  $\mu$  (cioè il valore atteso delle singole variabili) ma  $\bar{X}$

Stiamo quindi calcolando una sorta di scarto medio rispetto al valore della media campionaria. Il numero che otteniamo in questo modo può essere usato per stimare qualcosa: cosa?

Distribuzione normale: aspetti rilevanti per descrivere questo modello. Il massimo della distribuzione rappresenta il valore atteso? Un punto è individuato da due valori, mentre il valore atteso è un numero solo (cioè l'ascissa del massimo). La regola empirica deriva dalla simmetria della distribuzione? No, è una proprietà tipica della normale. Una distribuzione normale può assumere valori negativi? Cosa individuano  $\mu + \sigma$  e  $\mu - \sigma$ ? I punti di flesso. Cerca di disegnare un buon grafico ...

A cosa serve l'indice di correlazione? Stiamo parlando di statistica descrittiva, quindi cosa rappresentano  $x$  ed  $y$ ? I campioni. Al denominatore abbiamo le deviazioni standard campionarie. Che tipo di relazione individua questo indice? Cos'è la covarianza? Ci permette di valutare la presenza di una relazione diretta o inversa: come ne interpretiamo il valore ottenuto? Come interpretiamo il valore ottenuto dall'indice di correlazione lineare? È vero che se è maggiore di zero la retta ha un coefficiente angolare positivo e se è minore di zero un coefficiente

angolare negativo, ma possiamo dire di più. Come è legata l'intercetta della retta (cioè  $q$  in  $y = mx + q$ ) all'indice di correlazione? In nessun modo: la retta su cui si dispongono due campioni può toccare l'asse delle ordinate in un punto qualsiasi, sia se la relazione è lineare diretta che se la relazione è lineare inversa. L'indice di correlazione ha una proprietà interessante, che mi permette di interpretarlo in una maniera più efficace della covarianza. Quale valore ci aspettiamo di ottenere calcolando questo indice a partire da due campioni legati da un forte grado di relazione lineare diretta? È importante dire che i valori dell'indice di correlazione vanno tra  $-1$  ed  $1$ , perché in questo modo possiamo capire quanto sia forte l'eventuale relazione lineare.

Standardizzazione di variabili aleatorie: in cosa consiste? A cosa serve? Non c'è alcun valore, ma solo un modello; la standardizzazione non dipende dalle specificazioni, ma da qualcosa di più generale. Nel caso dei campioni, non riguarda il fatto di 'uniformare' la scala e non riguarda campioni diversi ma solo uno: effettuata la standardizzazione, il campione ha alcune proprietà interessanti.

Problema della stima parametrica e concetto di consistenza.

Calcoliamo la mediana di un modello esponenziale. Come si declina il concetto di mediana quando lo calcolo su una variabile aleatoria (o su una distribuzione)? Come definiamo la mediana di una variabile aleatoria? Conoscere il valore atteso di un'esponenziale potrebbe darmi una buona informazione sulla mediana? Come faremmo per calcolare la mediana di una distribuzione normale? È esattamente  $\mu$ , ma la legge empirica non c'entra niente. Come calcoliamo il quantile di una distribuzione? Vediamo graficamente dove possiamo aspettarci di trovare la mediana dell'esponenziale (tracciando la funzione di densità). La distribuzione esponenziale mi permette di calcolare analiticamente il valore della mediana: come calcoliamo l'area sottesa dal grafico della funzione di densità tra  $0$  e la mediana? Usiamo la funzione di ripartizione. Cambia qualcosa se cominciamo ad integrare da  $-\infty$ ? A cosa è uguale l'integrale della funzione di densità da  $-\infty$  alla mediana  $m$ ? Alla funzione di ripartizione calcolata in  $m$ . Qual è la forma analitica della funzione di ripartizione di un'esponenziale? Se  $m$  è la mediana, quanto deve valere l'integrale calcolato? Poniamo il tutto uguale a  $\frac{1}{2}$  e risolviamo rispetto ad  $m$ .

Il logaritmo di una frazione è uguale alla differenza dei logaritmi, quindi  $\ln \frac{1}{2} = -\ln 2$

Quindi la mediana di un'esponenziale è sempre  $\frac{\ln 2}{\lambda}$

Confrontiamola con il valore atteso. Chi è più grande?  $\ln 2$  è leggermente più piccolo di  $1$ , quindi la mediana è più piccola del valore atteso.

Abbiamo un campione estratto da una popolazione esponenziale di parametro  $\lambda$  incognito: come potrei trovare uno stimatore per la mediana? Perché potremmo usare  $\ln 2\bar{X}$  (dove  $X$  è la popolazione)? Se utilizzassi la media campionaria, otterrei  $\frac{1}{\lambda}$  facendo che cosa? Calcolandone il valore atteso. Cosa preserva l'assenza di deviazione della media campionaria? La linearità del valore atteso: per capirlo basta applicare la definizione di stimatore non distorto. La linearità del valore atteso mi permette di portare fuori una costante moltiplicativa quale  $\ln 2$ .

Poiché c'è una relazione lineare tra la mediana ed il valore atteso di un'esponenziale, posso sfruttare la linearità del valore atteso partendo da uno stimatore non deviato per il valore atteso della popolazione per ottenere uno stimatore non deviato per la mediana.

Quale altra proprietà degli stimatori abbiamo visto? Perché la consistenza è una proprietà desiderabile? Lo stimatore proposto gode di consistenza? Com'è definito lo scarto quadratico medio? Il motivo per cui la consistenza è una proprietà desiderabile diventa più chiaro usando la vera definizione dello scarto quadratico medio, e non la differenza tra varianza dello stimatore e bias rispetto alla quantità ignota. Il bias è una variabile aleatoria o un numero? Un numero, quindi non servirebbe calcolarne il valore atteso. Cosa calcola lo scarto quadratico? Calcolare il valore atteso della differenza tra quello che voglio stimare ed il valore atteso dello stimatore è una cosa; calcolare il valore atteso tra quello che voglio stimare e lo stimatore è un'altra.

Avendo dimostrato che lo stimatore è non distorto, come possiamo calcolare lo scarto quadratico medio? Come calcoliamo la varianza della media campionaria? Perché possiamo scrivere la varianza della somma delle  $X_i$  come la somma delle singole varianze? Perché sono indipendenti.

A cosa serve una curva ROC? Per visualizzare la bontà di un classificatore binario. Un classificatore fissato rispetto ad un campione individua un singolo punto nel grafico in cui tracciamo la curva ROC: dobbiamo prima introdurre il concetto di 'soglia.'

Variabile aleatoria uniforme continua. Quanto vale il grafico al di fuori dell'intervallo su cui la variabile aleatoria è definita? Vale zero. Calcoliamo il valore atteso di questa variabile. Perché quanto otteniamo è ragionevole? Qual è l'interpretazione geometrica del risultato ottenuto? Dove piazziamo  $\frac{\alpha+\beta}{2}$  nel grafico di densità? ( $\alpha$  e  $\beta$  sono gli estremi dell'intervallo in cui la variabile aleatoria è definita)

Perché la densità è costante, quindi ha senso che il baricentro dell'intervallo corrisponda al valore atteso.

Come stimiamo la mediana di una popolazione uniforme continua? Come calcoliamo la mediana della distribuzione

uniforme continua vista finora? A cosa è uguale? Al valore atteso, quindi usiamo la media campionaria.

Ho una popolazione descritta da una variabile aleatoria  $X$  distribuita secondo un modello uniforme di parametri  $\alpha$  e  $\beta$ , entrambi ignoti; ho un campione casuale  $X_1, \dots, X_n$  estratto da questa popolazione, e voglio trovare uno stimatore per la mediana della popolazione. Cosa si indica con  $\Theta$ ? Il parametro ignoto, quindi in questo caso sarebbe più la coppia  $(\alpha, \beta)$

A cosa è uguale la mediana? Che ragionamento si può fare per stimare la mediana? Cosa abbiamo scoperto di interessante circa la mediana di questa popolazione? Che è uguale al valore atteso: come ci può aiutare questa osservazione? Possiamo usare il valore atteso della distribuzione per stimare la mediana? No, perché è funzione di parametri sconosciuti quali  $\alpha$  e  $\beta$ : in generale, come definiamo uno stimatore?

Abbiamo un campione numerico: quali strumenti possiamo utilizzare per capire se è stato estratto da una popolazione normale? Ragionando in termini di distribuzioni approssimativamente simmetriche, non basta richiedere che l'istogramma presenti una barra centrale più alta e sia simmetrico rispetto ad essa. Lo 'skew' indica una leggera asimmetria verso sinistra o destra. Bisogna imporre la presenza di una forma a campana, che è quanto accade se vale la regola empirica. Quale altro strumento grafico abbiamo visto, oltre all'istogramma? Il diagramma QQ: abbiamo un solo campione, e vogliamo capire se è compatibile con una popolazione normale. Tuttavia, per calcolare i quantili teorici della normale, devo conoscerne i parametri: come faccio?