Slide 8

Gli area chart di ***McCundless*** con cui ha fatto vedere la distribuzione delle linee di trend.

Si è creata la linea ed è stato colorato sotto la linea.

Slide 9

Se si aggiunge l’area si aggiunge un ulteriore attributo visivo che aiuta la percezione visiva umana.

Slide 10

Come si può fare una **stima** delle distribuzioni delle età?

Una distribuzione bisogna vederla come una funzione che può essere plottata e resa in forma di grafico, che dice per ogni età, qual è la probabilità che trovi una persona di quella età.

La somma della distribuzione di probabilità è uguale a 1.

Fatte 100 tutte le persone di una nazione, se si parla di 20% 30% … si sono identificate delle popolazioni.

La probabilità fornisce delle quantità in percentuale.

Slide 11

SI tratta di ditribuzioni non normalizzate.

Sull’asse x si ha la quantità idi persone che hanno una data età.

Sul’asse y le età delle persone.

Si tratta di distribuzioni non normalizzate quindi non le si possono assimilare a probabilità.

Alla fine queste distribuzioni sono istogrammi.

Bisogna normalizzare l’istogramma (si ha un punto dell’istogramma diviso tutti i punti che sono stati considerati per creare l’istogramma stesso).

Come si stima la distribuzione delle età di tutti i cinesi?

Vado in Cina e campiono un po’ di persone, ad esempio in una città cinese come Pechino.

Si prendono tutte le persone di Pechino registrate e si parte dall’età 0.

Si conta quante persone hanno 0 anni e si fa un punto per ognuna di esse.

E così via, si va avanti a barrette fino a plottare per ogni età il numero di persone che hanno quelle età.

Questo è un campionamento non esaustivo, in quanto potrebbe esserci a Pechino un’azienda che magari causa tante morti per inquinamento.

Per fare un buon campionamenteo bisogna farne uno che copra tutta la variabilità dei casi, in modo da avere un campionamento abbastanza esaustivo della Cina.

Se si vuole una probabiltià di trovare una persona di 80 anni in Cina, bisogna normalizzare l’istogramma.

Intuitivamente perché non si ha normalizzato ogni singola distribuzione (Italia, Cina e USA)?

Se si fossero normlizzate tutte e tre le distribuzioni, si avrebbe avuto una sola scala (tutte percentuali), quindi un’unica informazione, ad esempio che in Cina l’età media è di 30 anni.

Quindi ci sono scale totalmente diverse.

La normalizzazione porta ad avere tutte le distribuzioni con la stessa scala.

Con la distribuzione di probabilità normalizzata non si avrebbe avuto la quantità di persone in più che ci sono ad esempio in Italia rispetto alla Cina.

Si sono volute tenere le distribuzioni non normalizzate per avere più informazioni a disposizione.

Tufte quando parla di ***data ink ratio*** non pensava al fatto che l’area è un ulteiore attributo visivo che migliora la percezione visiva.

Per fare una stima di una distribuzione, si fa un campionamento del fenomeno (in questo caso età delle persone) e si calcola l’istogramma delle loro età; se si vuole la distribuzione normalizzata (quinidi stimare la distribuzione) si normalizza l’isotgroamma, altrimenti si lascia così.

Questo vale per variabili continue.

Per variabili categoriche (ad esempio la distribuzione delle marche di auto) si fa un’isotgramma che indica la frequenza media della presenza di Fiat/BMW... in Italia.

Questa è la stima di trovare un auto BMW in Cina, USA o Italia.

Questi istogrammi sono “scalettati” in quanto si sta creando un’istogramma con un dato numero di bin.

Spesso si vuole avere una distribuzione nel continuo.

Quindi si fa un ***density plot***.

Slide 12

Più si interpola (si limita la “scalettatura”), più la distribuzione non rispecchia il campionamento reale.

Meglio se non rispecchia il campionamento reale perché altrimenti dei picchi potrebbero essere dovuti aun un “problema” del sistema di campionamento.

Che si tratti di ***density plot*** o ***area chart*** che esprimono un trend, spesso queste funzioni le si vogliono comparare: orizzontalmente o verticalmente a seconda se i valori sono più alti o più bassi.

In slide si tratta di istogrammi orizzontali (in verticale sarebbe stato tutto schiacciato).

Slide 13

Spesso l’***area chart*** sopra dà un’idea però non sono facilmente confrontabili due ***area chart*** in quanto l’asse x non è comune.

Quindi si è pensato di trasformare gli ***area chart*** in *stream graph*.

Se si vule permettere un confronto più semplice, si fa uno *stream graph* in cui tutte le aree, che si vedono sopra, si distribuiscono intorno all’asse x.

Usano le aree, quindi intuitivamente queste aree aiutano a vedere la differenza in quantità.

Ciò che non piace è che anche con gli stream graphs non è semplice il confronto tra aree.

Possono quindi anche loro risultare difficili da comprendere.

Slide 15

Oppure si può decidere di organizzare il tutto in uno ***small multiple***.

Slide 16

Se si hanno tante aree da confrontare, come si può organizzare un ***small multiple*** per distinguere bene i punti che sono negativi ed i punti che sono positivi.

Slide 19

Si pongono vicini e li si prova a confrontare.

Indicativamente quale va bene? Il nostro continua a salire, il problema è che sono su scale diverse.

Quindi cerchiamo di non mentire.

Mettiamo tutto in un plot con la medesima scala.

Slide 21

Se si allarga o si restinge il plot sembra che il bench vada meglio o peggio→ ma è solo percezione!

Indipentemente da tutto, se quando il ***benchmarck*** arriva a 70, quanto è la distanza da noi?

Non c’è un’asse delle x, è difficile, non si riesce a capire la differenza tra noi ed il *benchmark*.

Slide 22

Si fa in modo che il *benchmark* sia in ogni momento l’asse x e si sottrae ad ogni valore, il valore del benchmark: si trasla il risulato in base al valore corrispondente nel *benchmark*.

Si può vedere quando l’algoritmo va peggio del benchmark e si può vedere anche subito quando si va meglio e di quanto si va meglio.

Se si vuole dare la percezione di negativo e positivo si possono usare due colori (ad esempio rosso e azzurro) come in slide.

Il *benchmark* è stato fatto diventare 0.

Slide 23

Si possono guardare i valori (dividerli in blocchetti) e magari si può sottolineare un dato aspetto con una griglia oppure con dei colori di intensità maggiore per sottolinearne un conceto.

Slide 24

Ad esempio:

magari non ci interessa quanto in soldi guadagniamo rispetto al concorrente.

Bisogna fare il calcolo che faceva Tufte: si prende il valore, si sottrae quello del concorrente e lo si divide per i propri punti.

Quando si ha rosso intenso è perché si sta perdendo tanto rispetto al concorrente.

All’inizio il concorrente va meglio di noi e poi va peggio.

Con colori di intensità diversa si è indicato l’andamento positivo o negativo, al crescere del valore assoluto.

In questo modo si aiuta la percezione a vedere i valori tanto positivi e tanto negativi.

(In slide si ha un concorrente, come si può fare per far vedere più concorrenti?)

Slide 26

1. Visto che si è data una colorazione diversa al positivo ed al negativo si potrebbero invertire positivi e nagativi.

Se c’è un rosso più intenso od un blu più intenso vince lui.

Se si riesce a comprimere così il grafico e si vuole confrontare con tanti concorrenti, si può fare uno ***small multiple*** in cui si hanno tanti concorrenti.

Slide 27

Per far capire la visualizzazione a chi la vedrà, si mette una legenda che chiarificherà la scala dei rossi e dei blu.

Alla fine questa visualizzazione riesce a comprimere tanti *line chart*. Dà un’idea di come si sta andando rispetto agli altri.

Si sta utilizzando in effetti un po’ di ***data ink ratio***: in quanto si potevano mettere delle linee (***line chart***) anziché ***area chart*** → è vero, ma non avrebbe inteso bene l’idea.

L’area dà più la percezione della quantità.

Una volta spiegata la visualizzazione, il lettore deve solo capire come approcciarsi ad ogni linea.

A patto che non si inganni; la cosa bella degli ***small mutiple*** è che si inserisce lo stesso tipo di grafico: quindi l’utente capisce un solo grafico e applica lo stesso schema interpretativo agli altri plot.

Slide 28

Si vuole vedere in totale l’età dei cinesi.

Si prendono tutte le età dei cinesi e si vuole vedere se la distribuzione delle età è maggiore della propria.

A SX distribuzione età cinesi e a DX italiani.

Come si fa a vedere se una distribuzione ha valori più alti di un’altra distribuzione?

Si usa un ***box plot***.

Di tutta la distribuzione viene calcolato il valore mediano, visualizzato come un trattino orizzontale.

(La traslazione dell’orientazione del ***box plot*** dipende dalla lunghezza delle label sull’asse x.

Se si inclinano le label si rende difficile la visualizzazione, meglio scegliere un’orientamento orizzontale o verticale.)

Il ***box plot*** espirme i valori sotto cui sta il 25 percentile della distribuziona e sopra cui sta il 75% della distribuzione.

Ciò che corrisponde al 25° punto è il 25 percentile.

I punti che sono fuori dal *box plot* sono gli **outlier** ossia quei punti che si distaccano dalla distribuzione.

Ogni *box plot* è carattereizato da un *notch* (strettoia) che può essere più piccola o più larga e indica il 95% intervallo di confidenza della mediana.

Un conto è dire 101 pazienti ed un conto è dire tutta la popolazione.

Se i dati sono 101 si può stimare una mediana per pazienti morti con Covid e no Covid.

Il concetto di intervallo di confidenza è realtivo ad una statistica, non è la realtà dei fatti.

Fornisce una regola per dire quanti sono i potenziali valori che la mediana assume.

QUINDI il *box plot* in una visualizzazione fornisce tantissime informazioni.

Slide 29

Il *box plot* si può anche rendere orizzontale.

**IQR** → *Inter-Quartile Range*

I *box plot* lavorano su distribuzioni gaussiane (dette anche normali).

Teorema del limite centrale: al crescere del numero di campioni, le distribuzioni approssimano una normale.

I punti fuori dalla distribuzione soo ritenuti *outlier*.

Slide 30

Si supponga di avere un campionamento come quello sopra (*histogram of normal data*) in slide che evidenzi il comportamento di una **Guassiana**.

La visualizzazione di tale visualizzazione è quella che si può vedere a SX.

Siccome si tratta di una distribuzione abbastanza **simmetrica** (quella di SX), lo sarà anche il *box plot*.

Non è lo stesso per la distribuzione di DX, che è **asimmetrica**; il *box plot* dice che tutti i punti che sono dopo la mediana sono un po’ sparsi, inoltre quella distribuzione ha tanti *outlier*.

Slide 31

Se si fa un po’ di parallelismo con la statistica, vedendo i 5 *box plot*:

si può dire che la distribuzione D ha valori alti rispetto alle altre, la E ha tanti *outlier* ed è schiacciata rispetto alle altre.

Una cosa che si può andare a vedere è la sovrapposizione degli intervalli di confidenza della mediana.

L’intervallo di confidenza della mediana di B si sovrappone all’intervallo di confidenza della mediana di A.

Se si hanno due intervalli di confidenza che si svorappongono nel *box plot* non si può sapere se sono state campionate due distribuzioni che sono realmente differenti.

Non è transitiva come cosa: il fatto che l’intervallo di confidenza di D intersechi l’intevallo di confidenza di C vuole dire che sono stati fatti due campionamenti.

La stima delle due distribuzioni esprime una differenza che rispecchia la differenza tra le due distribuzioni reali?

Quando due intervalli di confidenza si sovrappongono non si riesce a dire se le due distribuzioni differiscono.

Slide 32

Il fatto che i due *notch* non si sovrappongano significa che si hanno distribuzioni di cui non esiste una differenza statisticamente rilevante.

Questi test forniscono un numero che è la probabilità che due distribuzioni non abbiano una differenza statisticamente significativa.

Slide 33

Se per una variabile si ha più del 50% dei valori mancanti (***missing data***) allora non la si considera, altrimenti tutte le stime che si fanno su quella variabile sarebbero ***biased*** (parziali).

Blocco di slide “HypothesistestingNotes”

Slide 1

TEST DI IPOTESI

Si suppone di avere dei bambini che fanno parte della 1° e 5° elementare.

Che fanno parte della 1° e della 5° non lo sappiamo ma si vuole vedere se a partire dai dati si riescono a distinguere bambini di 1° e di 5°.

Lo si fa considerando due variabili: l’altezza e l’IQ, per cui si hanno due distribuzioni differenti per variabile.

Per quanto riguarda il coefficiente di intelligenza (IQ) le distribuzioni sono più o meno le stesse.

Slide 2

Si raccolgono le variabili per i bambini e si raccolgono le età dei bambini.

Ogni pallino è un bambino.

Il campione azzurro è di 5° e il campione rosso di 1°.

Slide 4

La distribuzione che si stima è diversa da quella reale (confrontare infatti con la slide 1).

Questo lo si giustifica col fatto che non potremmo mai avere una conoscenza assoluta dei fatti (non potremmo mai sapere perfettamente come sono le cose).

Tendenzialmente i problemi multiclasse (si hanno bambini di 1°, 2°, 3°, 4° e 5°) si riconducono a problemi binari.

Slide 5

L’ipotesi è che si assume l’ipotesi nulla ossia si assume che le distribuzioni siano molto simili se non addirittura la stessa (distribuzioni altezze asse y), ovvero non differiscano.

Slide 6

Si sta considerano la distribuzione quindi tratteggiata: ossia che le due distribuzioni delle altezze dei bambini di 1° e di 5° siano molto simili se non addirittura la stessa.

Qual è la probabiltà che valga l’ipotesi nulla visto il campionamento fatto, quanto è probabile che le distribuzioni sottostanti siano le medesime?

Se il test di ipotesi dà una probabilità sotto lo 0.05 vuole dire che l’ipotesi nulla vale al 5%, quindi si può scartre l’ipotesi nulla (H0) e si può considerare l’ipotesi alternativa, ossia che le due distribuzioni siano differenti, ovvero esiste una differenza statiscamente significativa tra le due distribuzioni.

Slide 9

Il problema è che si può solo dire che sono diverse, ma non QUANTO.

Si può solo stimare la distribuzione reale e non si può dire di quanto si stimi la distribuzione reale.

Ci sono dei test di ipotesi nulla che permettono di andare a vedere che le distribuzioni siano uguali oppure altri test di ipotesi che fanno vedere se una distribuzione è maggiore di un’altra.

Slide 10

Si hanno due distribuzioni che si sovrappongono.

Non si può scartare l’ipotesi nulla.

Si tratta di test di ipotesi nel continuo.

Ci sono anche test di ipotesi per variabili categoriche.

Ad esempio il totale di macchine BMW rispetto a tutte le auto in Italia.

Slide 11

Ci sono test di ipotesi differenti a seconda del tipo di variabile con cui si sta lavorando.

(Tutte le distribuzioni viste in precedenza sono delle Gaussiane, dette anche normali.

Hanno la caratteristica forma a campana.)