Esercizio Binomiale

- Una casa di cura privata convenzionata con il S.S.N. possiede 20 letti disponibili per interventi di day hospital. Tuttavia, di solito accade che il 10% dei pazienti già prenotati non si presenta all'appuntamento. Per tale motivo il CUP propone di accettare fino ad un massimo di 22 prenotazioni al giorno. E' una buona scelta oppure è rischiosa? In altri termini, qual è la probabilità di ritrovarsi con almeno un paziente che non trova un letto pronto ad accoglierlo?

Impostazione

- Chi sono n, k e p ?
- n e' il numero totale degli esperimenti di Bernoulli che fanno parte della variabile binomiale.
- Nel nostro caso sono le 22 prenotazioni. Ogni singola prenotazione può andare a buon fine o il paziente non si presenta con probabilità p. Il numero di successi di interesse k e' 20. Più esattamente l'evento di interesse e' E{k=<20}
- Controllo assunti modello binomiale:Dal punto di vista del CUP la probabilità e' la stessa per tutte le prenotazioni e sicuramente un evento di successo o insuccesso non influenza il risultato della prossima prenotazione

Svolgimento

- La funzione di probabilità cumulativa risponde alla domanda P(X=<k) ed e' esattamente quello che l'esercizio chiede
- Dunque pbinom e' la funzione da usare
- pbinom(20,22,0.9)
- In alternativa si può usare la formula della probabilità binomiale $Cp^k(1-p)^{n-k}$ e facendo la somma di

$$P(k \leq 20) = \sum_{k=0}^{20} C_k^n p^k (1-p)^{n-k} = 1 - \sum_{k=21}^{22} C_k^n p^k (1-p)^{n-k}$$

I Test (parte I)

saverio.vicario@ba.itb.cnr.it

Test sulla media

- Vogliamo verificare se il valore della media μ di una variabile aleatoria X e' uguale o diverso da un dato valore μ_0. Quindi vogliamo verificare le seguenti ipotesi
- H0: $\mu=\mu_0$ H1:$\mu\neq\mu_0$
- In questo caso e' un test bilaterale
- Se la coppia di ipotesi e' la seguente
- H0: $\mu=\mu_0$ H1:$\mu>\mu_0$ oppure H0: $\mu=\mu_0$ H1:$\mu<\mu_0$
- Il test e' detto unilaterale destro o sinistro, rispettivamente

Test sulla media

- Per svolgere il test useremo il Teorema del Limite Centrale che descrive il comportamento della media aritmetica calcolata su un campione con una distribuzione arbitraria.
- Basandosi sul TLC la media del campione e' uno stimatore non distorto della media della popolazione
- Sempre basandosi sul TLC, si distingue inoltre un test a varianza nota da usare quando il campione e' basato sulla totalità della popolazione di interesse o su precedenti studi e un test in cui la varianza campionaria e' la nostra unica informazione disponibile sulla varianza.

Test della media con varianza nota

- Dal TLC possiamo affermare che se la variabile X : $N(\mu,\sigma^2)$ la media artimetica $\overline{X}$ di X segue una distribuzione $N(\mu, \sigma^2/n)$ in cui n e' la dimensione del campione.
- La variabile trasformata $Z= (X-\mu)/(\sigma/n^{1/2})$ avra' una media $\overline{Z}$ che seguira' una distribuzione N(0,1)
- La media $\overline{Z}$ si può ottenere direttamente da $\overline{X}$ usando la stessa trasformazione.
- Il test della media cosi costruito e' anche detto Z test.
-

Impostazione del test bilaterale

- La probabilità di osservare $\overline{Z}$ in una distribuzione N(0,1) risulta equivalente a quella di osservare μ in una distribuzione $N(\mu_0,\sigma^2)$ e dunque rappresenta la significatività del test.
- Infatti P($|\overline{X}| > |\mu|$ |H0) e' la probabilità di interesse ovverossia la probabilità di osservare valori più estremi dell'osservato sotto l'ipotesi nulla
- Trasponendo la domanda nella variabile Z la probabilità di interesse diventa P($|\overline{Z}| > 0$ |H0).
- La funzione di ripartizione segue eventi del tipo X≤a, dunque si puo' scrivere $1\text{-}F_{N(0,1)}(|\overline{Z}|)$

Impostazione in R

- Dunque data una v.a. X i comandi R sono:

>Z=(H0-mean(X))/(sd(X)/length(X)^0.5)

>Pvalue=1-pnorm(Z)

Il valore Pvalue o significativita va poi paragonato con la soglia prescelta α per decidere se va rigettata l'ipotesi nulla oppure no

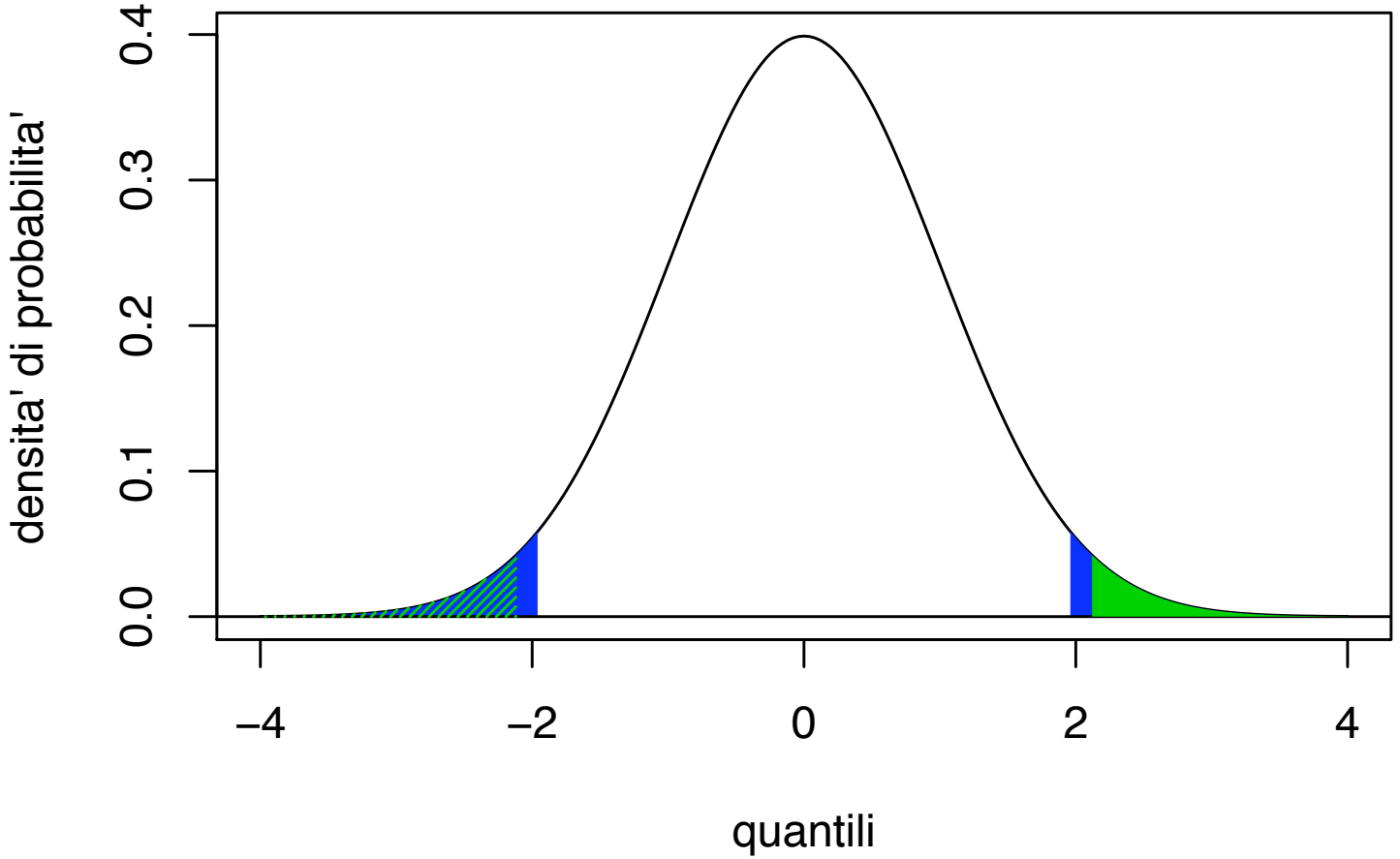
Esempio

- Vogliamo stabilire se l'età media di una popolazione e' diversa da 30 anni, nell'ipotesi che $\sigma^2=20$.
- Allora estraiamo un campione casuale di n=10 individui e otteniamo un valore di $\overline{X}$ =27. Le ipotesi da verificare sono :
- H0:μ=30 H1:μ≠30 con α=0.05 essendo un test bilaterale il valore da tenere in conto e' α/2=0.025
- Calcoliamo il valore della statistica Z
- nel caso H0 vera: $\overline{Z}$ =27−30√20/√10=−2.12

```
>1-pnorm(abs(2.12))
[1] 0.01700302
```

Allora rigettiamo H0 e concludiamo che l'età media della popolazione e' diversa da 30anni.

Perche’ α/2



In verde intero la probabilità calcolata, le barre verdi indicano la quantità implicita essendo la curva simmetrica si può paragonare un solo tipo di evento estremo contro meta della probabilità di soglia

Test media con varianza nota unilaterale

- H0: $\mu=\mu_0$ H1:$\mu>\mu_0$ L'evento H2:$\mu<\mu_0$ deve essere impossibile o comunque non compreso negli eventi S presi in considerazione

- In questa impostazione la probabilità di interesse e'
- P($\overline{Z}>0$ |H0) e la soglia e' esattamente uguale ad α
- nel caso H0 vera: $\overline{Z}$ =27−30√20/√10=−2.12
- >1-pnorm(abs(2.12))
- 0.01700302
- Questo valore va dunque raffrontato con 0.05

Test media con varianza non nota

- Assumendo sempre che la variabile segue una distirbuzione X : $N(\mu,\sigma^2)$ e dunque la media artimetica $\overline{X}$ di X segue una distribuzione $N(\mu, \sigma^2/n)$ in cui n e' la dimensione del campione. Non conoscendo la varianza dobbiamo usare la varianza campionaria come stimatore di σ^2.

Varianza campionaria

- La varianza campionaria invece di essere esattamente il momento centrale di grado secondo
- >momentoC<-function(x,k){mean((x-mean(x))^k)}.
- Con k=2
- E' la varianza corretta dal fattore n/(n-1)

```
> n=10
> X=rnorm(n)
> momentoC(X,2)
[1] 1.180407
> var(X)
[1] 1.311564
> (10/(10-1))*momentoC(X,2)
[1] 1.311564
```

Test media con varianza non nota

- Si applica sempre la stessa trasformazione per riportare ad una distribuzione standard
- $T=(X-\mu)/(\sigma/n^{1/2})$
- Solo che la distribuzione di riferimento non e' più la distribuzione normale ma la distribuzione T di student
- Questa distribuzione tiene conto dell'incertezza della stima della varianza.
- T(df)
- Il parametro df indica i gradi di libertà della distribuzione df=n-1
- df esprimono il numero di dati effettivamente disponibili per valutare la quantità d'informazione contenuta nella statistica

Distribuzione T di student

- Densità di probabilità:
- >dt
- Funzione di ripartizione
- >pt
- Inverso della funzione di ripartizione:
- >qt

```
>plot(-100:100/25,dnorm(-100:100/25), type='l')
>lines(-100:100/25,dt(-100:100/25, df=10), col=2)
```

Distribuzione T di student

- Al crescere di n la distribuzione converge sulla distribuzione normale perché la stima della varianza e' sempre più corretta.
- La distribuzione di studenti inoltre ha:
- Moda=0
- Varianza= n/(n-2) se n > 2
  Simmetria= 0 se n > 3
  Curtosi= 6/(n-4)se n > 4
- Essendo simmetrica si può applicare lo stesso trucco quando si vuole calcolare le probabilità dei valori più estremi di una certa soglia che con la distribuzione normale

Esempio

- Sia dato l'indice di massa corporeo BMI=peso(Kg)/ statura^2(m^2) di14 soggetti. Vogliamo stabilire se la media della popolazione da cui il campione´e stato estratto´e diversa da 35.
- X=23 25 21 37 39 21 23 24 32 57 23 26 31 45
- H0:$\mu=35$ H1:$\mu\neq35$ con $\alpha=0.05$
-

Svolgimento

```
>H0=35
>X=c(23, 25, 21, 37, 39, 21, 23, 24, 32, 57, 23, 26, 31, 45)
>n=length(X)
>T=(mean(X)-H0)/(sd(X)/n^0.5)
>pvalue=pt(T,df=n-1,lower.tail=F)
>alpha=0.05/2
```

E' possibile rigettare l'ipotesi zero?

```
>pvalue<alpha
```

NO

In alternativa si può scrivere

```
t.test(X, mu=35)
```

Differenze tra i due svolgimenti

- Il valore di pvalue calcolato con la funzione pt e' esattamente la metà di quello calcolato dalla funzione t.test.
- Questo perché t.test calcola la probabilità globale delle due code in un test bilaterale e dunque il valore dato non va paragonato con $\alpha/2$ ma con α

Paragone fra le medie di due gruppi

- Vogliamo confrontare due v.a. X e Y e scegliamo di usare la media come statistica di confronto.
- La domanda formale diventa dunque

$$H0 : \overline{X} = \overline{Y}$$

$$H1 : \overline{X} \neq \overline{Y}$$

- Bisogna inoltre distinguere fra due modalità di generazione dei campioni accoppiati e casuali.

Tipo di campionamento

- Se l'unita' statistica su cui gli attributi sono misurati e' unica le misure di X e Y devono essere considerate accoppiate.
- Es. Misura di un parametri fisiologico (es. pressione arteriosa) prima e dopo l'assunzione di un farmaco in un gruppo di 10 pazienti
- X_1 e Y_1 sono misure di pressione arteriosa sul paziente 1 prima e dopo l'assunzione del farmaco

Tipo di campionamento

- Se l'unita statistica su cui le variabili sono misurate e' diversa il campionamento e' detto casuale e i campioni indipendenti.
- Es. Effetto di un insetticida sulla presenza di insetti nei campi. Le misure sono fatte su due campi diversi uno trattato con l'insetticida e uno no. Si contano gli insetti presenti in 10 e 12 piante a caso nei due campi.
- Nota che in questo caso il numero di osservazioni può essere diverse nelle due variabili

Nel caso accoppiato

- Si lavora con la variabile derivata W in cui $W_i=X_i-Y_i$ per ogni unita statistica i.
- Le ipotesi si formulano come segue:
- H0: $\mu=\mu_0$ H1:$\mu\neq\mu_0$
- Con generalmente $\mu_0=0$ per provare che non c'e' differenza fra X e Y.
- Il testo cosi e' riportato al caso precedente del test sulla media.
- Se il la varianza e' nota si usa uno Z test
- Se la varianza e' non nota si usa un T test

Paragone fra le medie di due gruppi campioni accoppiati varianza nota

- La variabile W viene trasformata nella variabile Z con la nota trasformazione

Esempio di accoppiati

- Richiamare i dati pair65
- X=pair65$heated
- Y=pair65$ambient
- 9 elastici divisi in due pezzi il gruppo heated e' stato scaldato a 65° per tutte e due i gruppi si e' poi proceduto ad una misura di elasticità
- Se assumiamo che la popolazione di interesse e' l'insieme dei nove elastici possiamo assumere che la varianza del campione sia nota

Svolgimento

- Ci aspettiamo che il campione X siccome scaldato sia quello più elastico dunque facciamo un test unilaterale con X>Y e W>0
- Più formalmente
- Assumendo W: $N(\mu, \sigma^2)$ H0: $\mu=0$ H1:$\mu>0$ alpha=0.05
- W=X-Y
- Z=(mean(W)-0)/(momentoC(W,2)/length(W))^0.5
- pvalue=1-pnorm(Z, lower.tail=T)
- Se pvalue<α allora si può rigettare l'ipotesi zero

Svolgimento in caso di varianza non nota

- Se assumiamo che i nove elastici sono rappresentativi della popolazione di tutti gli elastici prodotti dalla ditta da cui provengono dobbiamo assumere che la varianza non e' nota dobbiamo utilizzare la varianza campionaria e come distribuzione per l'ipotesi nulla la distribuzione t di student
- >alpha=0.05
- >T=(mean(W)-0)/(var(W)/length(W))^0.5
- >n=length(W)
- >pvalue=1-pt(T, df=n-1, lower.tail=T)
- >pvalue<alpha
- >t.test(W, alternative='greater')
- >t.test(X,Y, alternative='greater', paired=TRUE)

Nel caso indipendente

- Siano $X=(x_1,\ldots,x_n)$ e $Y=(y_1,\ldots,y_m)$ distribuiti rispettivamente $N(\mu_x,\sigma^2_x)$ e $N(\mu_y,\sigma^2_y)$ e vogliamo verificare se $\mu_x=\mu_y$
- Anche qui dobbiamo distinguere tra il caso in cui σ^2_x e σ^2_y sono note e quello che invece vanno stimate dai dati stessi.
- In ambedue due i casi si calcola la media di ogni campione separatamente

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Caso varianze note

- $\overline{X}$ =N(μ_x,σ^2_x/n) e $\overline{Y}$ =N(μ_y,σ^2_y/m)
- Quindi la differenza delle medie

$$\overline{X} - \overline{Y} : N(\mu_x - \mu_y, \frac{\sigma_x^{\ 2}}{n} + \frac{\sigma_y^{\ 2}}{m})$$

- Ne consegue che la variabile

$$\overline{Z} = \frac{(\overline{X} - \overline{Y}) - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^{\ 2}}{n} + \frac{\sigma_y^{\ 2}}{m}}} : N(0,1)$$

Caso varianze note

- Se H0: $\mu_x=\mu_y$ allora Z si semplifica come segue

$$\overline{Z} = \frac{(\overline{X} - \overline{Y})}{\sqrt{\frac{\sigma_x^{\,2}}{n} + \frac{\sigma_y^{\,2}}{m}}} : N(0,1)$$

Caso varianze non note

- La corretta stima della varianza deve tener conto che i due campioni possono avere diversa ampiezza (n≠m)

$$V^2 = \frac{1}{n+m-2}\left[\sum_{i=1}^{n}(x_i - \overline{X})^2 + \sum_{i=1}^{m}(y_i - \overline{Y})^2\right]$$

$$= \frac{(n-1)S_x^{\ 2} + (m-1)S_y^2}{n+m-2}$$

V^2 e' detta varianza combinata ed e' una media pesata della varianza campionaria dei due campioni.

Varianze non note

- Usando lo stimatore V^2 si applica la stessa trasformazione precedente ottenendo stavolta una variabile distribuita come una distribuzione t

$$\overline{T} = \frac{(\overline{X} - \overline{Y}) - (\mu_x - \mu_y)}{V\sqrt{\frac{1}{n} + \frac{1}{m}}} : T(df = n + m - 2)$$

- df=n+m-2 rispecchia i gradi di libertà della stima della varianza. Infatti per calcolarla bisogna calcolare prima le due medie consumando due gradi di libertà

Esempio con campioni indipendenti

Dati PlantGrowth

>PlantGrowth

>summary(PlantGrowth)

>?PlantGrowth

Sono tre gruppi di 10 piante cresciute con in tre condizioni di crescita di cui due sperimentali. Ci concentreremo sui trattamenti ‘ctrl’ e ‘trt1’

X=PlantGrowth$weight[PlantGrowth$group=='trt1']

Y=PlantGrowth$weight[PlantGrowth$group=='ctrl']

L’attributo misurato e’ il peso al raccolto ed essendo le piante diverse nei due gruppi i due campioni sono indipendenti

Impostazione e svolgimento: varianza nota

- Assumendo che la varianza sia nota e uguale alla varianza del campione. Non abbiamo aspettative sugli effetti del trattamento trt1 rispetto a ctrl dunque il test sarà bilaterale
- H0:$\mu_x=\mu_y$ H1:$\mu_x\neq\mu_y$ e alpha=0.05

```
> n=length(X)
> m=length(Y)
> sigmax=momentoC(X,2)
>sigmay=momentoC(Y,2)
>Z<-(mean(X)-mean(Y))/(sigma2y/(m)+ sigma2x/(n))^0.5
>pvalue=1-pnorm(abs(Z), lower.tail=F)
```

Impostazione e svolgimento: varianza non nota

- Assumendo che la varianza sia non nota useremo la varianza campionaria come stima della varianza. Non abbiamo aspettative sugli effetti del trattamento trt1 rispetto a ctrl dunque il test sarà bilaterale
- H0:$\mu_x=\mu_y$ H1:$\mu_x\neq\mu_y$ e alpha=0.05

```
>v2=((n-1)*var(X)+(m-1)*var(Y))/(n+m-2)
>T=(mean(X)-mean(Y))/(((1/n)+(1/m))*v2)^0.5
>pvalue=1-pt(abs(T), df=m+n-2)
>pvalue<alpha/2
```

Impostazione e svolgimento: varianza non nota

- t.test(X,Y, var.equal=TRUE)
- Anche qui nota che il pvalue e' doppio e va dunque paragonato con α e non $\alpha/2$

Alternative

- Vedremo nelle prossime lezioni che in caso non ci senta confidenti di applicare il teorema del limite centrale (pochi campioni, stratificazione dei dati, estrema non normalita' della distribuzione) ci sono alternative
- >wilcox.test(X,Y)
- >wilcox.test(X,Y, paired=TRUE)

Test multipli

- Se l'errore accettato nel rifiutare un ipotesi nulla e' **α** (es. 0.05) se si eseguono 100 test e in tutti i casi H0 e' vera ci si aspetta che in media 100*alpha test daranno un risultato significativo sotto il la soglia **α**.
- Dunque se una domanda coinvolge diversi test il livello **α** deve essere modificato.
- Quando si eseguono tanti test per esplorare tanti effetti.
- Per esempio si esegue 100 differenza tra media per verificare l'efficacia di 100 farmaci candidati.

Correzione di Bonferroni

- Una soluzione semplice anche se conservativa (rigetta l'ipotesi zero con più difficoltà di quanto richiesto) e' la correzioni di Bonferroni.
- Lo scopo e' che la probabilità di di erroneamente rigettare l'ipotesi zero sia globalmente per tutti i test eseguiti di livello **α**. Dunque ci si domanda quale modifica bisogna fare ad **α** per ottenere in ogni singolo test per ottenere un livello di rischio globale **α**
- Dunque se la probabilità globale di ottenere dei risultati significativi se tutte le ipotesi zero sono vere in n test e' esattamente n***α**, la soglia per il singolo test che consentirebbe di avere un valore di **α** globale deve essere $alpha_i = \alpha/n$

Correzione Bonferroni sequenziale

- Questo valore di alpha/n e' pero' molto restrittiva.
- Infatti e' utilizzabile solo se ci si aspetta un solo test significativo su tutta la serie di n tests
- In alternativa bisogna operare una procedura sequenziale.
- >pvalues=runif(100)
- >pvalues=sort(pvalues)
- >alphas=0.05/100:1
- > sum(pvalues<alphas)
- [1] 0
- > sum(pvalues<0.05)
- Vedi articolo allegato sulla correzioni test multipli