Statistica I, Laurea triennale in Ingegneria Gestionale, a.a. 2017/18

December 18, 2017

1 Registro delle lezioni

27/09/2017. Introduzione al corso. Vengono illustrati alcuni elementi di statistica descrittiva (Capitolo 2 del S. Ross), come le proporzioni, la media aritmetica di un campione sperimentale, il range e la deviazione standard (misure della dispersione), l'istogramma.

3/10/2017. Prima esercitazione con R. Comandi R su media (mean), deviazione standard (sd), istogramma (hist) con varianti (numero classi, FALSE), applicati ad una stringa scritta a mano.

Generazione di una stringa di numeri casuali normali (rnorm), ripetizione delle analisi precedenti. Come varia il risultato con la numerosità. Come varia l'istogramma con la numerosità delle classi.

Plot di strighe (successioni) e di funzioni; comando seq, plot di una densità gaussiana. Sovrapposizione della densità gaussiana ad un istogramma (comando lines).

Definizione di densità, esempi delle densità uniformi, esponenziali di parametro λ, gaussiana standard. Ricerca della costante di normalizzazione o verifica della proprietà di area uno, nei tre casi.

4/10/2017. Densità gaussiana generica, cenno alla verifica della proprietà di area uno, ruolo grafico dei parametri. Aree e quantili, sia z_α che q_α, uso delle tavole dei quantili gaussiani standard, regola $q_\alpha = -q_{1-\alpha}$.

Esempio di soglia per motivare l'interesse per i quantili.

Definizione di valor medio associato ad una densità e suo calcolo per la densità esponenziale. Verifica, per le densità simmetriche, che il valor medio coincide col punto di simmetria. Quindi per gaussiane $N\left(\mu, \sigma^2\right)$ coincide col parametro μ.

Insistenza sulla differenza tra modello e dati, tra parametri del modello (a volte chiamati "parametri veri") e stime sperimentali.

10/10/2017. Varianza associata ad una densità. Suo calcolo per uniformi, esponenziali e gaussiane, verificando che per quest'ultime si tratta di σ^2. Definizione di deviazione standard σ. Interpretazione grafica (a differenza di σ^2), stessa unità di misura dei dati.

Concetto di variabile aleatoria attraverso esempi. Densità f di una v.a. X: probabilità che una v.a. X assuma valori un un intervallo $[a,b]$ (cioè $P\left(X \in [a,b]\right)$) uguale a $\int_a^b f\left(x\right) dx$. Calcolo di $P\left(X > x_0\right)$ nel caso di una v.a. esponenziale e nel caso di una gaussiana standard tramite le tavole.

Col software R: calcolo di $P\left(X > 70.5\right)$ (ed anche > 70.2, > 70.1) quando $X \sim N\left(70, \left(0.1\right)^2\right)$. Uso del comando pnorm. Intuizione numerica circa deviazioni dalla media pari a $\sigma, 2\sigma, ..., 6\sigma$.

11/10/2017. Definizione di universo degli eventi, regole insiemistiche e corrispondenza con operazioni logiche su affermazioni; esempi del dado e della semiretta. Definizione di probabilità, sue regole, caso di uno spazio finito (probabilità dei singoletti che determinano la probabilità degli eventi), eventi equiprobabili in uno spazio finito, esempio del dado; probabilità definita da una densità sui numeri reali. Probabilità condizionale, motivazione della formula, verifica in un esempio con due dadi.

17/10/2017. Esercitazione con R dedicata a grafici di varie densità. Sono state trattate le densità normale, gamma, esponenziale, Weibull, log-normale e beta. Abbiamo rintracciato in rete una breve illustrazione (wikipedia), il comndo di R con l'illustrazione dei parametri, abbiamo poi visualizzato con rnorm, rgamma, rexp, rweibull, rlnorm, rbeta ed il comando hist alcuni istogrammi, esplorando il ruolo dei parametri tramite hist(r...). Abbiamo poi visualizzato i grafici tramite il comando curve, ad esempio curve(dnorm(x,0,1),from=-4,to=4). C'è il problema del range, che va un po' cercato per tentativi, se non si conosce ancora bene la densità. E' stato anche usato il comando curve(...,add=TRUE,col="...") che sovrappone un grafico all'altro. In questo modo abbiamo anche verificato che una gamma con shape 1 e scale λ è un'esponenziale di parametro λ. E' stato usato anche il comando polygon per colorare delle parti del sottografico.

18/10/2017. Coi comandi m=rnorm(1,10,10); X=rnorm(100,m,1), m.emp=mean(X) si ottiene la media vera? Si può conoscere, nei problemi concreti, la media vera e l'errore commesso da quella empirica? Si può dire che l'errore tra media vera e media empirica è minore di una quantità calcolabile? No. Si può però calcolare la probabilità di commettere un certo errore $P\left(\left|\overline{X} - m\right| \leq \delta\right)$. Qui $\overline{X}$ è costruito a partire da un campione di v.a. $X_1, ..., X_n$. Come calcolare quella probabilità? In casi semplici, come per le gaussiane, riusciremo a svolgere un calcolo esatto. Per altre distribuzioni, possiamo utilizzare un metodo simulativo che vedremo a breve.

Esercizi preliminari su aree e quantili gaussiani, standard e non standard. Enunciato di alcuni teoremi sulle gaussiane (standardizzazione e media aritmentica di gassiane resta gaussiana; come si trasformano i parametri). Esempio di problema preliminare su $\overline{X}$: calcolare $P\left(\overline{X} > a\right)$.

Infine, è stato illustrato il ciclo for di R, da sperimentare ad esercitazione, utilizzato per costruire un istogramma della media empirica; si vedrà come usare questo metodo per esaminare uno stimatore.

24/10/2017. Soluzione esercizio per casa del 10/10, allargato (cioè anche con den-

sità Gamma, con soluzione non-parametrica, e con domanda sul quantile). Per la stima dei parametri di shape e scale di una gamma, si usi il "metodo dei momenti" illustrato ad esempio nelle dispense del prof. Ricci intitolate "fitting distributions...", pag 12 (shape=m^2/s^2, scale=m/s^2).

Formula per l'intervallo di confidenza per la media per le gaussiane; abbiamo visto l'illustrazione concettuale e grafica della formula, non la dimostrazione.

31/10/2017. Soluzione esercizi dei compiti. Sottolineatura della distinzione tra la grnadezza aleatoria esaminata e la media aritmetica di un suo campione; se si cerca un intervallo, nel primo caso è un intervallo che riguarda i valori della grandezza esaminata, nel secondo caso si sta studiando la precisione della stima della media. Compito di settembre 2017, domande 1.1, 1.2 e, dopo alcune premesse sulle regole di valor medio e varianza, domanda 1.5. [Vengono introdotti i simboli $E[X]$ e $Var[X]$ ed enunciate alcune regole; poi applicate al caso gaussiano.]

7/11/2017. Esercizi dal compito d'esame del 19 luglio 2017. Le prime due ore sono annullate per assemblea.

8/11/2017. Proprietà della probabilità condizionale. Eventi indipendenti. Variabili aleatorie indipendenti. Proprietà dei valori attesi di v.a. indipendenti. Proprietà delle gaussiane. Proprietà di $\overline{X}$. Dimostrazione della formula dell'intervallo di confidenza per la media. Inquadramento dei temi del compitino.

14/11/2017. Uso del ciclo for, esempio della media empirica, percezione della precisione della stima al variare della numerosità del campione, indipendente dalla numerosità di Monte Carlo; gaussianità della distribuzione empirica; percezione anche della debolezza del miglioramento al variare della numerosità n del campione e confronto con la formula dell'errore $\delta = \sigma q_{1-\alpha/2}/\sqrt{n}$.

Primo compitino.

15/11/2017. (con recupero dalle 10:30 alle 11:30) Test statistici: ipotesi nulla e alternativa, regione di rifiuto, conclusione del test (cenno al concetto di errore di prima specie); esempio dal compito del 28/6/2017, problema 1.3. Per la creazione delle regioni di rifiuto abbiamo usato, in questa prima fase, la teoria degli intervalli di confidenza (bilateri) ed una loro versione unilatera. Soggettività nella scelta di α e riformulazione tramite il p-value (valore p). Questi primi elementi si trovano nei paragrafi 8.1, 8.2 e 8.3.1 del libro.

21/11/2017. Analisi del compito con R. Alcuni elementi sono simili alla scheda del corso Statistica II del 5/10/2015.

In sintesi: caricamento della stringa dei voti totali, sue prime analisi statistiche (media, deviazione, mediana, range, istogrammi, sovrapposizione della gaussiana). Trovare un intervallo bilatero che contenga area 0.9, in modo non parametrico. Prese come vere media e dev st totali, calcolare l'intervallo di confidenza al 90% e controllare se la media empirica del gruppo 1 cade in tale intervallo. Se non ci cade, è come se avessimo eseguito un test statistico, con ipotesi nulla "la media del campione è quella totale", e la rifiutiamo.

Concetto di coefficiente (o indice) di correlazione, indipendente dall'unità di misura. Suo calcolo con R per le stringhe voti-peso-es1-matricola. Istogramma delle correlazioni e

test sui valori trovati.

Regressione lineare sempice: comando lm e abline (dopo plot).

28/11/2017. Esercizio su test statistici (dal 16/12/2011): formulazione delle ipotesi, regione di rifiuto, esecuzione del test; calcolo della probabilità di riuscita del test, quando la nuova media è cambiata di una certo valore rispetto alla vecchia. Calcolo del valore p partendo dalla definizione come "probabilità che la grandezza statistica usata per fare il test ($\overline{X}$ nell'esercizio) superi ($>$) il valore sperimentale di quella grandezza statistica ($\overline{x} = 51.8$ nell'esercizio)".

Esame, con R, dei dati su altezza, peso ecc. (terza tabella di dati in rete). Caricamento di una tabella ed estrazione delle colonne. Correlazione e plot dell'intera tabella o di singole coppie. Esercizio per casa.

29/11/2017. (con recupero dalle 10:30 alle 11:30) Errori di prima e seconda specie, loro probabilità, funzione $\beta(\mu)$. Potenza di un test. Ripresa dell'esercizio della lezione precedente.

Valore p da diversi punti di vista. Diversi test (unilaterali e bilaterale). Regioni di rigetto scritte tramite z (standardizzazione); valore p calcolato nei due modi per ciascuna.

Uso pratico della t di Student.

Esercizi dal compito d'esame del 21/12/2011, domande 3, 4 e soprattutto 5.

5/12/2017. Esercizi sulla somma di gaussiane; Teorema Limite Centrale ed esercizi sulla somma di v.a. indipendenti con altre distribuzioni, es. Bernoulli di parametro p (interpretazione della loro somma come numero di accadimenti). Si veda ad es. il compito dell'11 gennaio 2017, es. 1.5, 17 febbraio es. 1.3.

Revisione dei primi due compiti del 2017, per identificare le cose da approfondire. Ad es. 11 gennaio 2017, es. 1.2 (soglie con peggioramento delle stime), 31 gennaio 2017, es. 1.3 sul valore p risolto in due modi diversi (partendo dalle due definizioni), e naturalmente esercizi sui test, che costituiranno il blocco principale. Mentre nel compitino non saranno presenti esercizi su chi quadro, come 11 gennaio 2017, es. 1.3.

6/12/2017. (con recupero dalle 10:30 alle 11:30) T di Student, teorema sul fatto che $\frac{\overline{X}-\mu}{S}\sqrt{n}$ è t di Student a $n-1$ gradi di libertà; in intervalli di confidenza e test basati sulla t di Student.

Esempio del controllo di qualità: carte di controllo, calcolo del limiti LCL, UCL (quindi test bilaterale); calcolo della potenza del test (probabilità di accorgersi di una certa variazione, che può risultare invalidante per il prodotto).

Utilizzo in modo inverso della potenza, in due modi: per trovare la numerosità che garantisce una certa potenza (ad es. in un test con regione di rifiuto $\overline{x} > \mu_0 + \frac{\sigma q_{1-\alpha}}{\sqrt{n}}$, si deve trovare il più piccolo n tale che $P^{\mu_0+\Delta\mu,\sigma}\left(\overline{X} > \mu_0 + \frac{\sigma q_{1-\alpha}}{\sqrt{n}}\right)$ ha un valore della potenza $\geq$ del valore preassegnato), oppure per trovare le variazioni rilevate dal test con una certa potenza (nell'esempio precedente, si deve trovare $\Delta\mu$ tale che $P^{\mu_0+\Delta\mu,\sigma}\left(\overline{X} > \mu_0 + \frac{\sigma q_{1-\alpha}}{\sqrt{n}}\right)$ ha il valore preassegnato della potenza). Si suggerisce di esercitarsi a partire dalla domanda

1.4 del 31 gennaio 2017.

12/12/2017. Esercizi con R: qqnorm per esaminare la gaissianità, confronto tra varie distribuzioni; istogrammi e valutazione della gaussianità per somme di variabili aleatorie partendo da distribuzioni non gaussiane (verifica numerica del TLC); varianza empirica, sua non gaussianità.

Secondo compitino.

13/12/2017. Cenni alle variabili aleatorie discrete (Bernoulli, binomiale, Poisson). Legame tra Bernoulli e binomiali; legame tra binomiale e Poisson (teorema degli eventi rari); legame con le gaussiane (TLC). Legge dei grandi numeri.

Funzione di ripartizione, legame con quantile e densità, uso per esaminare trasformazioni.

Assenza di memoria dell'esponenziale.

18/12/2017 (recupero dalle 10:30 alle 13:30).

Stimatori di massima verosimiglianza e metodo dei momenti; esempio della densità esponenziale.

Chi quadro, in intervalli di confidenza e test.

Simulazione orale con R.

2 Esercizi suggeriti per casa

2.1 Esercizi per i compiti scritti

(Suggerito il 24/10) Compito d'esame del 11 gennaio 2017, es 1.1, 1.2, 2.1.

(Suggerito il 24/10) Compito d'esame del 31 gennaio 2017, es Es 1.1, 1.2, 2.1.

(Suggerito il 31/10) Compito d'esame del 17 febbraio 2017, es Es 1.1, 1.2, 2.1.

(Suggerito il 31/10) Compito d'esame del 19 aprile 2017, es Es 1.1, 1.4, 2.1.

(Suggerito il 7/11) Compito d'esame del 10 giugno 2017, es Es 1.1, 2.1.

(Suggerito il 7/11) Compito d'esame del 28 giugno 2017, es Es 1.1, 2.1.

(Suggerito il 29/11) Esercizi 1.3 e 1.4 del 31/1/2017; 1.4 e 1.5 del 17/2/2017; 1.3 e 1.5 del 19/4/2017.

2.2 Esempi di esercizi su R

Esercizio (suggerito il 10/10). i) Produrre, 100 numeri gaussiani $N\left(7, 2^2\right)$.

ii) Scordatevi da dove provengono. Voi avete solo quei 100 numeri. Stimare la probabilità che la vostra grandezza (quella da cui provengono i numeri) superi il valore 8. Svolgere l'esercizio usando più di una densità, ed anche in modo non parametrico.

iii) Calcolate la stessa probabilità nell'ipotesi che la vostra grandezza sia una $N\left(7, 2^2\right)$.

iv) Riprendiamo il problema della domanda ii. Trovare la soglia λ tale che la probabilità di superare λ ($\geq \lambda$) è 0.1 (oppure 0.001). Come per il punto ii, svolgere l'esercizio usando più di una densità, ed anche in modo non parametrico.

Esercizio (suggerito il 24/10)**.** Esplorare i grafici possibili della distribuzione gamma con parametro di shape pari a 0.5, 1, 2, 5, 10, 100, e vari parametri di scale. Eseguire l'esercizio sia tramite istogramma, sia tramite grafico della densità. Scelto il caso shape=5, scale=1, determinare il valore sull'asse delle x che ha, alla sua destra, area 0.1 (il quantile di ordine 0.9), e, sul grafico della densità, colorare di azzurro l'area da tale x in poi verso destra (cioè l'area di ampiezza 1).

Esercizio (suggerito il 24/10)**.** Prendere il campione sperimentale messo in rete, scegliere una classe di densità osservando l'istogramma, stimare i parametri, osservare la densità in sovrapposizione all'istogramma come verifica, calcolare la soglia a tale che $P(X < a) = 0.1$.

Esercizio (suggerito il 14/11)**.** Al grafico della densità gaussiana sovrapposta all'istogramma realizzato nella lezione del 14/11, sovrapporre una colorazione rossa della zona corrispondente all'intervallo di confidenza al 90%.

Esercizio (suggerito il 21/11). Analizzare i dati in rete denominati "dati per esercizio 2", caricando le varie colonne (dopo aver trasformato virgole in punto), calcolando le varie correlazioni, medie e visualizzando istogrammi come nella lezione del 21/11. Calcolare e visualizzare intervalli di confidenza. Esaminare l'entità della correlazione a confronto dell'istogramma di correlazione.

Esercizio (suggerito il 28/11). Analizzare i dati in rete denominati "dati per esercizio 3", caricando la tabella (dopo aver trasformato virgole in punto), isolando le varie colonne, calcolando la matrice di correlazione ed alcune correlazioni singole per confronto, plottando la matrice ed alcune coppie, cercando inoltre, per quelle più correlate, la retta di regressione.

Esaminare l'entità della correlazione a confronto dell'istogramma di correlazione.

Esercizio (suggerito il 12/11). Esaminare la gaussianità di un campione estratto da varie distribuzioni visualizzando l'istogramma con la gaussiana sovrapposta, e visualizzando il qqnorm con la bisettrice (indicativa, trovata per tentativi) sovrapposta (usare ad esempio abline(...,...,col="red")).

Esercizio (suggerito il 12/11).

3 Materiale utile per R

3.1 Comandi

Nota: se può essere utile, visionare le prime schede su R del corso Statistica II del docente, in rete.

X=c(...,...,...)
mean(X)
sd(X)
range(X)
cor(X,Y) se X e Y sono vettori
cor(A) se A è una matrice

hist(X), eventualmente con numero di classi e FALSE

rnorm(n), rnorm(n,m,s)

dnorm, pnorm, qnorm

plot(X), plot(X,Y), event. con type="..." (l, h, s, n, p), event. con col="..." (es. red, skyblue, ecc.)

plot(A) se A è una matrice

(secondario: lines(X))

seq(a,b,h); es. X=seq(-4,4,0.01);Y=dnorm(X,0,1);plot(X,Y)

curve; es. curve(dnorm(x,0,1), from=-4, to=4); event. con add=TRUE e colori

*norm, *gamma, *exp, *weibull, *lnorm, *beta

polygon, esempio:

```
curve(dnorm(x,0,1), from=-4, to=4)
x=seq(-4,0,0.01)
y=dnorm(x,0,1)
polygon(c(-4,x,0),c(0,y,0),col="red")
```

X=scan("clipboard") ; utenti mac: provare X=scan(pipe("pbpaste"))

A=read.table("clipboard")

data una matrice o tabella A, per estrarre righe e colonne: X=A[k,]; Y=A[,k]

round(X,k)

il separatore di comandi è ;

quantile(X,0.9)

sort(X)

for (... in 1:...) {...} (vedi esempio sotto)

lm(Y~X) (regressione lineare semplice) (cenno al comando summary)

abline(lm(Y~X),col="...") dopo aver plottato plot(X,Y)

qqnorm

3.2 Esempi di problemi affrontati

Generazione di un campione casuale secondo una certa distribuzione, disegno dell'istogramma con area 1, disegno ad esso sovrapposto della densità.

Esplorazione del grafico possibile di una certa classe (es. Weibull) al variare dei parametri, sia con hist sia con curve, colorazione di sotto-aree.

Verifica grafica che una gamma con shape 1 è un'esponenziale.

Construzione dell'istogramma della media empirica:

```
m=1:100
for (k in 1:100000) {
X=rgamma(25,7,10)
m[k]=mean(X)
}
hist(m)
```

Dato un campione sperimentale, si sceglie una classe di densità osservando l'istogramma, si stimano i parametri, si osserva la densità in sovrapposizione all'istogramma come verifica, poi si calcolano grandezze di interesse, esempio probabilità (es. prob(X>a)) o quantili (es. la soglia l tale che prob(X>l)=0.1). Soluzione non-parametrica dello stesso problema.

Problema del 14/11:

N.montecarlo=100000; n.campione=200; m.empirica=1:N.montecarlo

for (k in 1:N.montecarlo) {X=rnorm(n.campione); m.empirica[k]=mean(X)}

hist(m.empirica,100,FALSE); curve(dnorm(x,0,1/sqrt(200)), from=-2, to=2, add=TRUE)

la seguente è sbagliata: curve(dnorm(x,0,1), from=-2, to=2, add=TRUE)

Problema del 12/12:

n = 10; N = 10000; S =1:N

for (i in 1:N) {X = rweibull(n,2,3); S[i] = sum(X)}

hist(S); qqnorm(S)

Problema del 12/12:

n = 100; N = 10000; Var =1:N

for (i in 1:N) {X = rnorm(n); Var[i] = var(X)}

hist(Var); qqnorm(Var)