STATISTICA E CALCOLO DELLA PROBABILITA'

MinchiaSoft

2021-27-03

“So anyway, what exactly do you think that means . . . a 50% chance of rain?” “Perhaps that there is a 75% chance that it will rain in 66.6% of the places?” (From Inspector Morimoto and the Japanese Cranes, by Timothy Hemion)

”La media e la varianza non danno la conoscenza anche se e’ un dato di fatto che Hayter e’ uno stronzo”

”Lezione chiarissima, grazie mille per le delucidazioni sugli esercizi. Comunque sono incredibili alcune coincidenze, non saprei calcolare la probabilità che possa accadere (forse lei si AAHAHHA) ma ho anch’io un orologio a cucù. Le auguro una buona serata con una frase di un antico statistico: Alea iacta est ;)” -cit un moderno statistico ”Ti do il palo” cit Assistente

Contents

Dispense relative al corso di Probabilità e Statistica de Piccioni a ingegneria informatica/automatica della Sapienza.
Si parla di calcolo della probabilità e di statistica inferenziale.
Le fonti utilizzate sono principalmente il manuale di Hayter e quello di Ross.
Se ci sono errori contattateci a questo link: reclami@minchiasoft.it (non esiste davvero non farlo bruh)

Ringraziamo della ispirazione Ananas Molesto.
Siamo lieti che il nostro lavoro possa essere utile a qualcun altro.
Forza Roma.

Chapter 1

PROBABILITÀ

1.1 Assiomi della Probabilità:

1. La probabilità di un evento è un reale non negativo.
2. La probabilità dello spazio dei campioni è 1.
3. Qualsiasi sequenza numerabile di insiemi disgiunti soddisfa $P(\bigcup_\infty) = \sum_\infty p_i$
4. La funzione probabilità è monotona e definita da 0 a 1.
5. La probabilità di $\emptyset$ è 0.

1.2 Combinazione di eventi:

$$P(A \cap B) + P(A \cap B') = P(A) \tag{1.1}$$
$$P(A \cap B) + P(A' \cap B) = P(B) \tag{1.2}$$

Se eventi disgiunti :

$$p(a \cup b) = p(a) + p(b) \tag{1.3}$$

senno':

$$p(a \cup b) = p(a) + p(b) - p(a \cap b) \tag{1.4}$$

Unione tre eventi:

$$P(A \cup B \cup C) = [P(A) + P(B) + P(C)] - [P(A \cup B) + P(A \cup C) + P(B \cup C)] + P(A \cup B \cup C) \tag{1.5}$$

1.3 Leggi di de morgan:

$(a \cap b)' = a' \cup b'$

$(a \cup b)' = a' \cap b'$

1.4 Bayes e condizionata:

$$P(A|B) = P(A \cap B)/P(B) \tag{1.6}$$

$$P(Ai|B) = \frac{P(Ai) * P(B|Ai)}{P(B)} \tag{1.7}$$

$$P(B) = \sum_{i=0}^{n} P(Ai) * P(B|Ai) \tag{1.8}$$

1.5 Eventi Indipendenti:

If two events are independent, then the probability that they both occur can be calculated by multiplying their individual probabilities. : VERO

It is always true that $P(A \mid B) + P(A' \mid B) = 1$. : VERO

It is always true that $P(A \mid B) + P(A \mid B_i) = 1$. : FALSO

It is always true that $P(A \mid B) \leq P(A)$. : FALSO

1.6 Combinazioni, Permutazioni, Disposizioni:

There are n! ways in which n objects can be arranged in a line. If the line is made into a circle and rotations of the circle are considered to be indistinguishable, then there are n arrangements of the line corresponding to each arrangement of the circle. Consequently, there are $(n-1)!$ ways to order the objects in a circle.

Consider 5 blocks, one block being Andrea and Scott and the other four blocks being the other four people. At the cinema these 5 blocks can be arranged in 5! ways, and then Andrea and Scott can be arranged in two different ways within their block, so that the total number of seating arrangements is 2 * 5! = 240. Similarly, the total number of seating arrangements at the dinner table is 2*4! = 48. If Andrea refuses to sit next to Scott then the number of seating arrangements can be obtained by subtraction. The total number of seating arrangements at the cinema is 720-240 = 480 and the total number of seating arrangements at the dinner table is 120 - 48 = 72.

1.6.1 Permutazioni e basta

Modi in cvi posso ordinare n oggetti, che sarebbe
$n!$

1.6.2 Permutazioni con ripetizioni

Modi in cui posso ordinare n oggetti, di cui alcuni sono uguali(interscambiambili)
n_i è il numero di oggetti uguali del tipo i

$Volte = \frac{n!}{n_1! * ... * n_k!}$

1.6.3 Disposizioni e basta:

Modi in cui posso ordinare una selezione di k oggetti da un totale di n oggetti, in cui l'ordine è importante

$$modi = \frac{n!}{(n-k)!} \tag{1.9}$$

1.6.4 Disposizioni con ripetizioni

$$modi = n^k \tag{1.10}$$

1.6.5 Combinazioni e basta

Modi in cui posso ordinare una selezione di k oggetti da un totale di n oggetti, ma l'ordine NON è importante

$$modi = \frac{n!}{(n-k)!k!} = \binom{n}{k} \tag{1.11}$$

1.6.6 Combinazioni con ripetizioni

Modi in cui posso ordinare una selezione di k oggetti da un totale di n oggetti, ma l'ordine NON è importante e certi so uguali

$$modi = \binom{n+k-1}{k} \tag{1.12}$$

1.7 Studio delle variabili aleatorie

Stalin	Discreta	Continua
F(x)	cumulative distribution function	cumulative distribution function
f(x)	probability mass function	probability density function
E(x)	$\sum p_i x_i$	$\int_S xf(x)$
mediana	non c'e'	$x \mid F(x) = 0.5$
Varianza:	$E(x^2) - (E(x))^2$	$E(x^2) = \int x^2 f(x)$

In una variabile continua ho:

$f(x) = dF(X)/dx$

$F(x) = \int_a^x f(x)$

$\sigma = \sqrt{VAR(x)}$

1.7.1 VARIABILI ALEATORIE CONGIUNTE:

NB integrale $\int_S f(x,y)dxdy = 1$ SEMPRE!!!

Dis Marginale:

ae $f_x(X) = \int_a^b f(x,y)dy$

senno con le discrete e' la somma delle righe o delle colonne

Calcolo covarianza:

$COVAR(X,Y) = E(XY) - (E(X)E(Y))$

E(XY)=

D: $\sum x_i * y_j * p_{ij}$ (ae 1*1*1/4 + 2*1*2/3+...)

C: $\int_a^b \int_c^d xy * f(x,y)dxdy$

Calcolo correlazione:

$Corr(X,Y) = \frac{Covar(X,Y)}{\sqrt{Var(X)Var(Y)}}$

Btw, la correlazione è un valore compreso tra 1 e -1 Se è 0 si dice che le variabili non sono correlate (o anche "incorrelate" secondo piccioni)

1.8 Combinazioni lineari e non di variabili aleatorie

Lineari: Se ho $Y = aX + b$ avro':

$E(Y) = aE(x) + b$

$Var(Y) = a^2 * Var(X)$

(se ci sono sottrazioni la media si sottrae, invece la var si somma sempre)

Se sono non lineari devo fare l'integrale della f(x) e poi sostituire con una Y che sia minore della X che mi sono calcolato

esempio:

Ho $f(x) = 1$ per $0 \leq x \leq 1$ e $Y = e^x$

Quindi F(X)=x

$P(e^x <= y) \longrightarrow P(X <= ln(y)) \longrightarrow Fx(ln(y)) \longrightarrow$

ln(y)

$$f_y = \frac{dF_y(y)}{dy} = \frac{1}{y}$$

a) The variance of a random variable is measured in the same units as the random variable. F
(b) In a diving competition, the scores awarded by judges for a particular type of dive have an expected value of 78 with a standard deviation of 5. If the scores are doubled so that they can be compared with scores from an easier type of dive, the new scores will have an expected value of 156 and a standard deviation of 10. V
(c) The variance of the difference between two independent random variables cannot be smaller than the larger of their two variances. V
(d) If a continuous random variable has a symmetric probability density function, then the mean and the median are identical. V
(e) If X is a continuous random variable, then $P(X \geq x) = P(X > x)$ for any value of x. V
(f) If X is a discrete random variable, then $(X \geq x) = P(X > x)$for any value of x. F

1.9 DISTRIBUZIONI DISCRETE

RICORDA: BINOMIALE, BINOMIALE NEGATA E GEOMETRICA CALCOLANO COSE MOLTO CORRELATE!!!!!

1.10 Distribuzione Binomiale:

n bernoulli trials, misura i successi su n esperimenti indipendenti

$$P(X = x) = \mathrm{C}_x^n \; p^x (1-p)^{n-x} \tag{1.13}$$

$E(x) = np$
$Var(x) = np(1-p)$

1.11 Proporzione successi:

$Y = \frac{X}{n}$
$E(x) = p$
$Var(x) = \frac{p(1-p)}{n}$

1.12 Distribuzione Geometrica:

Data un sequenza bernoulli trials la con prob. successo p, la distribuzione geometrica misura il numero di tentativi fatti fino al primo successo.

$P(X = x) = (1-p)^{x-1}p$

$E(x) = \frac{1}{p}$

$Var(x) = \frac{1-p}{p^2}$

P. cumulativa $F(x) = 1 - (1-p)^x$

1.13 Distribuzione Binomiale Negativa:

numero di tentativi fino al successo resimo

$$P(X = x) = \binom{x-1}{r-1}(1-p)^{x-r}p^r \tag{1.14}$$

$E(x) = \frac{r}{p}$
$Var(x) = \frac{r(1-p)}{p^2}$

1.14 Distribuzione Ipergeometrica:

Tipo la binomiale, ma senza replacement (la p di scegliere un oggetto con prob iniziale r/N diminuisce via via).

con ad eesempio:
- N tot palline
- r palline rosse
- n numero di palline prese a caso
- x il num di palline rose prese (avendo preso n palline a caso)

$$P(X = x) = \frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}} \tag{1.15}$$

$E(X) = \frac{nr}{N}$
$Var(X) = \frac{N-n}{N-1}\ \frac{nr}{N}\ (1 - \frac{r}{N})$

1.15 Distribuzione di Poisson: (pesce)

Viene usata per modellare il numero di eventi che avviene in una certa unità di tempo, distanza o volume e ha media e varianza pari al parametro lambda λ.

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!} \tag{1.16}$$

$E(X) = Var(X) = \lambda$

La poissoniana con lambda $\lambda = np$ puo' approsimare una binomiale con n grande e p piccolo, quindi, dato che un binomiale puo approssimare un'ipergeometrica con N grande e r piccolo, LA POISSONIANA PUO' APPROSSIMARE L'IPERGEOMETRICA

1.16 Distribuzione Multinomiale:

generalizzazione della binomiale, ci sono n trials che possono avere k outcomes

X1,..,Xk misurano quante volte deve avvenire un determinato outcome

$$P(X_1 = x_1, ..., X_n = x_n) = \frac{n!}{x_1! * ... * x_k!} * (p_1^{x_1} * ... * p_k^{x_k}) \tag{1.17}$$

ogni Xi ha la propria media e la propria varianza:

$E(X_i) = n * p_i$
$Var(X_i) = n * p_i(1 - p_i)$

D : What is the probability that four sets of right-handed clubs are sold before four sets of left-handed clubs are sold? R: Mi calcolo con la binomiale negata la p di ne vendo 4 in 4 tenativi + p di 5 tentativi + + p di 6 + p di 7 (a 8 significa che le altre hanno gia venduto 4 pezzi)

1.17 DISTRIBUZIONI CONTINUE

1.17.1 Distribuzione Uniforme:

La distr uniforme e una retta orizzontale in un intervallo a,b
tutti gli eventi sono equiprobabili

$$f(x) = \frac{1}{b-a} \qquad a \leq x \leq b \tag{1.18}$$

$$F(x) = \frac{x-a}{b-a} \tag{1.19}$$

$$E(x) = \frac{a+b}{2} \tag{1.20}$$

$$Var(x) = \frac{(b-a)^2}{12} \tag{1.21}$$

1.17.2 Distribuzione Esponenziale:

utile per modellare tassi di rottura e tempi di attesa

$$f(x) = \lambda * e^{-\lambda x} \tag{1.22}$$

$$F(x) = 1 - e^{-\lambda x} \tag{1.23}$$

$E(x) = \sigma f(x) = \frac{1}{\lambda}$
$Var(x) = \frac{1}{\lambda^2}$

The implications of the memoryless property can be rather confusing when first encountered. Suppose that you are waiting at a bus stop and that the time in minutes until the arrival of the bus has an exponential distribution with = 0.2. The expected time that you will wait is consequently 1/ = 5 minutes. However, if after 1 minute the bus has not yet arrived, what is the expectation of the additional time that you must wait? Unfortunately, it has not been reduced to 4 minutes but is still, as before, 5 minutes. This is because the additional waiting time until the bus arrives beyond the first minute during which you know the bus did not arrive still has an exponential distribution with = 0.2. In fact, as long as the bus has not arrived, no matter how long you have waited, you always have an expected additional waiting time of 5 minutes! This is true right up until the time you first spot the bus coming.

$P(X >= x_0) = P(X \geq x_0 + y) = e^{-\lambda x}$

la probabilita che un oggetto funzionera' per un ulteriore periodo di tempo e la stessa sia che l'oggetto sia nuovo sia che sia gia utilizzato da un po

1.17.3 Processo di Poisson:

data una serie di intervalli di tempo in sequenza la distribuzione di N(t) (del tempo tra un avvenimeto e un altro) e una distr esp intorno a lambda, e il numero di eventi che avvengono in un detemtinato intervallo di tempo e una distr poissoniana intorno a λt

Quindi se N(t) e un processo di Poisson di intensita λ:

1) il numero di eventi che avvengono in un intervallo di tempo [0,t] e modellato da una distr di Poisson intorno a λt

2) i tempi che separano gli eventi di un processo di Poisson di intensita' λ sono una successione di var aleatorie esponenziali di intensita' λ tra di loro indipendenti

1.17.4 Distribuzione Gamma:

E' una generalizzazione della dist esponenziale e ha importanti applicazioni nello studio della teoria dell'affidabilita' e dei processi di Poisson

la funzione gamma(k) $\Gamma(k)$ è una generalizzazione del concetto di fattoriale:
$\Gamma(k) = \int_0^{\inf} x^{k-1} * e^{-x} dx$
$\Gamma(k) = (k-1)\Gamma(k-1)$
in particolare per k intero positivo e $k > 1$:
$\Gamma(n) = (n-1)!$
non ci sono altre modi noti per saltare l'integrale negli altri casi
casi particolari:
$\Gamma(1) = 1$
$\Gamma(1/2) = \sqrt{\pi}$
k=forma e lambda=scala
La distr gamma con parametri $k\ maggiore0\ e\ \lambda\ maggiore0$ ha :

$$f(x) = \frac{\lambda^k\ x^{k-1}\ e^{-\lambda * x}}{\Gamma(k)}$$

(1.24)

$E(x) = \frac{k}{lambda}\ Var(x) = \frac{k}{lambda^2}$
In un processo di Poisson il tempo che viene impiegato per fare k eventi ha una distrizione gamma la gamma e la somma di piu distr esponenziali

1.18 Distribuzione Normale o Gaussiana:

E' figa perche' modella natruralmente la distribuzione degli errori e grazie al teorema del limite centrale appossima decentemente il comportamento di un numero notevole di fenomeni casuali

ha come parametri la media μ e la dev std σ e la pdf e' una curva a campana (curva di Gauss) simmetrica intorno alla media $N(\mu, \sigma^2)$

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \tag{1.25}$$

$E(x) = \mu\ Var(x) = \sigma^2$

Se si fa una trasformaxione lineare di una distr gaussaina X tipo Y=aX+b Y e' a sua volta una gaussiana con media $\mu + b$e varianza $a^2 * sigma^2$

1.18.1 Normale Standard

Quindi se ho

$$Z = \frac{X - \mu}{\sigma} \tag{1.26}$$

Ho la cosidetta normale standard, ovvero con media 0 e var 1, di cui la cumulativa è FI(x) ($\Phi(x)$) che e' molto utile per calcolare i valori di probabilita cumulativa nelle distribuzioni dato che abbiamo le tabelle dove controllare i valori. ae

$$P(X \leq b) = \Phi(\frac{b - \mu}{\sigma}) \tag{1.27}$$

da cio' segue che la prob di un intervallo si calcola:
$P(a \leq x \leq b) = \Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})$

$P(X \leq -x) = 1 - P(X \leq x)$

La distanza interquartile di una gaussiana qualsiasi e' pari a $1.3490 * \sigma$

In general, if $\bar{X}$ is the average of n of these random variables, what is the smallest value of n for which $P(|\bar{X}|0.5) \geq 0.99$?
sol:
$0.5 * \sqrt{n}$ = valore della std con $p = 0,995 = 2{,}5...$
$n \geq 27$
(si risolve coi punti critici della gaussiana)

1.19 Disuguaglianza di Markov

La probabilità che x sia maggiore uguale di alpha è minore uguale della media di x su alpha

$$P(X \geq \alpha) \leq \frac{E(x)}{\alpha} \tag{1.28}$$

1.20 Diseguaglianza di Chebichev:

$$P(\mu - c\sigma \leq x \leq \mu + c\sigma) >= 1 - \frac{1}{c^2} \tag{1.29}$$

Dimostrazione: Si mette il complementare, si eleva tutto al quadrato per togliere il valore assoluto e si fa la Disuguaglianza di Markov

$P(\mid X - \mu \mid \geq c\sigma) \rightarrow P((X-\mu)^2 \geq (c\sigma)^2) \rightarrow P(Y) \leq \frac{E((X-\mu)^2)}{(c\sigma)^2}$

$$P(Y) \leq \frac{1}{c^2} \tag{1.30}$$

Problema stronzo: Probabilita' tra mi -c e mi + c
$P(\mu - c \leq x \leq \mu + c) = K$
1) normalizzando si ottiene che : $p(c/sigma) = k + \frac{1-k}{2}$
2) occorre vedere sulla tabella della gaussiana a che valore di z=z0 corrispone
3) c=sigma * z0

1.20.1 Distribuzione chi quadro:

La distribuzione chi quadro e la somma di n quadrati di distribuzioni normali standard:

$X = Z_1^2 + ... + Z_n^2$
X di dice variabile aleatoria chi quadro con n gradi di liberta'
La distribuzione chi quadro puo' essere approssimata da una distr Gamma con parametri:
lambda $\lambda = \frac{1}{2}$

e kappa $k = \frac{n}{2}$

e con $E(X) = n$ e $Var(X) = 2n$

Se X_n è una chi quadro con n gradi di libertà, a è un numero reale tra 0 e 1 si definisce la quantità $X^2_{a,n}$ come la qta per cui vale:

$P(X \geq (X_{a,n})^2) = a$

1.20.2 Distribuzione T:

$$f(x) = t_n = \frac{N(0,1)}{\sqrt{\frac{X^2_n}{n}}} \tag{1.31}$$

Per valori molto grandi di n coverge alla gaussiana standard:

$E(x) = 0$, $n >= 2$

$Var(X) = \frac{n}{n-2}$, $n \geq 3$

In maniera analoga alla Chi quadro si puo definire un qta ta,n per cui:
$P(T_n >= t_{a,n}) = a$
poi $a = 1 - P(T_n \geq -t_{a,n})$
Per la simmetria della dist intorno allo 0
e quindi:
$P(T_n \geq -t_{a,n}) = 1 - a$
quindi $-t_{a,n} = t_{1-a,n}$

Chapter 2

STATISTICA

2.1 Media Campionaria:

Ho n varilibili aleatorie indipendenti (una popolazione) con la stessa distribuzione e media mi e var sigma^2

la media campionaria $\bar{X}$ è definita come:
$\bar{X} = \frac{X1+...+Xn}{n}$
$E(\bar{X}) = \mu$
$Var(\bar{X}) = \frac{\sigma^2}{n}$
$\bar{X}$ è centrata in mi e la sua varibilità si riduce con l'aumentare di n.

2.2 Teorema del Limite Centrale:

"La somma di un numero elevato di variabili indipendenti tende ad avere una distribuzione approssimativamente normale"

Siano $X_1, ..., X_n$ variabili aleatorie indipendenti ed identicamente distribuite, tutte con media μ e varianza σ^2:

allora $X_1 + ... + X_n$ con un n molto elevato e' approssimativamete una Normale con
media $n\mu$ e varianza $n\sigma^2$

quindi:
$\frac{X1+...+Xn-n*mi}{\sigma\sqrt{n}}$ è approssimativamente distribuita come una Normale standard Z.

2.3 Legge dei grandi numeri

Leggi sulle bernulliane(eventi che possono accadere o no).

Debole

La probabilità che la distanza tra la media campionaria e la media sia maggiore di un numero arbitrariamente piccolo tende a 0 con l'aumentare del numero di osservazioni fatte.

$$n \to \infty \Longrightarrow P(\mid \bar{X}_n - \mu \mid > \varepsilon) \to 0 \tag{2.1}$$

Forte

All'aumentare del numero di osservazioni fatte, la probabilità che media campionaria e la media vera coincidano tende a 1

$$P((\lim_{n\to\infty} \bar{X}_n) = \mu) = 1 \tag{2.2}$$

Dimostrazione

Dato che la combinazione de Distribuzioni co media μ e varianza σ^2 tende a una Gaussiana, so che la media campionaria va a $\frac{n\mu}{n}$, ma tanto le n si semplificano e diventa μ sisi.
La dimostrazione è finta ma forse se la imbastisci al piccioni non ci fa caso.

2.4 Distribuzione approssimata di una Media Campionaria:

$\bar{X} = (1/n)\sum_1^n x_i$
E quindi $\sqrt{n}\frac{(\bar{X}-\mu)}{\sigma}$ che si comporta come un normale standard per il TLC

IL MIO CAMPIONE DEVE ESSERE ALMENO DI 30 PER COMPORTARSI SUFFICIENTEMENTE COME UNA GAUSSIANA.

2.5 Varianza Campionaria:

$$S^2 = \frac{1}{n-1}\sum_1^n (X_i - \bar{X})^2 \tag{2.3}$$

$S = \sqrt{S^2}$:= dev std campionaria

$E(S^2) = \sigma^2$ (cioe' la var delle distribuzioni considerate)
La distribuzione congiunta di $\bar{X}$ e S^2 e' data da:

$\frac{(n-1)*S^2}{sigma^2}$ che per motivi matematicamente validi si comporta come una Chi quadnro con n-1 gradi di liberta'
"Se $X_1, ..., X_n$ e' un campione proveniente da una distr normale con media μ e var σ^2 allora $\bar{X}$ ed S^2 sono variabili
aleatorie indipendenti, inoltre $\bar{X}$ e' normale con media μ e var $\frac{\sigma^2}{n}$
, e $\frac{(n-1)*S^2}{\sigma^2}$ e' una Chi quandro con n-1 gradi di liberta'"

Chapter 3

Inferenza Statistica

"L'inferenza statistica (o statistica inferenziale) è il procedimento per cui si inducono le caratteristiche di una popolazione dall'osservazione di una parte di essa (detta "campione"), selezionata solitamente mediante un esperimento casuale (aleatorio)"

Campione:

Scegliendo un campione casuale di intervistati si ottiene una buona approssimazione dei dati da calcolare sull'intera popolazione a causa di fenomeni matematici incomprensibili che leggeremo un giorno, forse, (come le doc di MongoDB) (la distribuzione all'interno del campione che prendo di individui con una determinata caratteristica che si manifesta nella popolazione con probabilità p può essere approssimata da una binomiale su popolazioni molto grandi)

3.1 Stima,Parametri,Statistiche e Stimatori

La stima è quando ricavi i parametri di una popolazione dalle osservazioni che fai sul campione.
Parametri sono le caratteristiche della popolazione, tipo media, varianza, o proporzione di capelli blu.

La statistica è una funzione in cui ho solo variabili già note.
Lo stimatore è la funzione che mi perfette di stimare un parametro(il valore atteso dellla funzione stimatore è il parametro).

3.2 Teoria della stima

Le stime possono essere:

3.2.1 Puntuali:

stimo un valore preciso := una stima corretta (o "non distorta") ha lo stesso valore del parametro e quella piu efficiente e' quella con una deviazione standard minore. Di solito si usano la Media Campionaria e la Varianza Campionaria per stimare la media e la varianza

3.2.2 Intervallari:

Viene stimato un intervallo di confidenza all'interno del quale si dovrebbe trovare un parametro con un livello di confidenza pari a (1-a)

3.3 Stime intervallari della Media:

3.3.1 Intervallo -t a due lati:

Devo stimare l'intervallo di confidenza della media della mia distribuzione non sapendo esattamente quale sia la mia varianza (uso quella campionaria)
quindi dato il mio intervvallo con livello di confidenza $(1-a) = 0.95$
la mia media μ si trova

$$\bar{x} - t_{\frac{\alpha}{2},n-1}\frac{S}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{\frac{\alpha}{2},n-1}\frac{S}{\sqrt{n}} \tag{3.1}$$

È un intervallo centrato in μ dove $t_{\frac{\alpha}{2},n-1}$ è il valore critico (critical point) della distribuzione t in a/2 con n-1 gradi di liberta'

la larghezza dell'intervallo è pari a
$L = 2t_{\frac{\alpha}{2},n-1}\frac{S}{\sqrt{n}}$ = 2*punto critico*errore semplice(standard error)

3.3.2 Intervallo -t a un lato:

Avendo:

intervallo di confidenza:=$(1-\alpha)$
population mean := μ
numero di "continuous data observations":= n
sample mean := $\bar{x}$
sample standard deviation s

Gli intervalli t a un lato sono:

$$\mu \in (-\infty, x + t_{\alpha,n-1}\frac{S}{\sqrt{n}}) \tag{3.2}$$

$$\mu \in (x - t_{\alpha,n-1}\frac{S}{\sqrt{n}}, +\infty) \tag{3.3}$$

3.3.3 Intervallo -z a due lati:

Devo stimare l'intervallo di confidenza della media della mia distribuzione ma CONOSCO LA VARIANZA σ^2

$$\bar{x} - \frac{z_{\frac{\alpha}{2}}\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} - \frac{z_{\frac{\alpha}{2}}\sigma}{\sqrt{n}} \tag{3.4}$$

Con z_α che è il punto critico della normale std per a
Cioè z_α è l'analogo di t_α quando la standard deviation σ è conosciuta

3.3.4 Intervallo -z a un lato:

$$\mu \in (-\infty, \bar{x} + \frac{z_a\sigma}{\sqrt{n}}) \tag{3.5}$$

$$\mu \in (\bar{x} - \frac{z_a\sigma}{\sqrt{n}}, +\infty) \tag{3.6}$$

3.4 Test d'ipotesi

definisco a partire da un'analisi di una certa popolazione delle ipotesi derivate dai dati.

3.4.1 Ipotesi nulla e ipotesi alternativa:

L'ipotesi nulla AFFERMA che la mia distribuzione sulla popolazione soddisfi una certa proprietà
L'ipotesi alternativa e' costruita NEGANDO quella nulla

3.4.2 P-value:

il p-value e' un valore compreso tra 0 e 1.
Se assume un valore minore di 0,01 l'ipotesi nulla viene respinta.
Se assume un valore piu' grande di 0,10 l'ipotesi nulla viene considerata ma non e' necessariamente vera.
Se assume un valore tra 0,01 e 0,10 non ho informazione sull'ipotesi nulla,

3.4.3 Potenza del test

Si definisce potenza del test 1- la probabilità di Rigettare H0 quando H0 è falsa. Questo parametro esprime quindi l'efficacia del test di individuare l'ipotesi alternativa. 1-B = P(Rigettare H0—H0 è falsa)

3.5 Errori:

3.5.1 di tipo 1:

Definito α il livello di significatività scelto per il test di ipotesi, esso esprime anche la probabilità che si manifesti un errore detto di tipo 1. Tale errore consiste nel rigettare H0 pur essendo H0 è vera α = P(Rigettare H0—H0 è vera) Se diminuisco α diminuisco la possibilità dell'errore di tipo 1.

3.5.2 di tipo 2:

β esprime la probabilità che si manifesti un errore di tipo 2, ovvero un errore in cui non si rigetta H pur essendo H_0 falsa.
β = P(Non rigettare H0—H0 falsa)
Per diminuire β occorre aumentare la potenza del test. Questo può essere fatto o diminuendo il livello di confidenza del test (aumentando la probabilità di un errore di tipo 1 oppure aumentando la dimensione del campione.

3.5.3 t Test a due lati

Ipotesi Nulla: H0: ho media $\mu = \mu_0$ (quella stimata)

Ipotesi Alternativa: H1: $\mu \neq \mu_0$

$p - value = 2 * P(x \geq |t|)$

t è la distanza tra il valore della media campionaria e la nostra media stimata diviso la dev std campionaria

$$t = \frac{\sqrt{n}(x/ - mi0)}{S} \tag{3.7}$$

Rifiuto l'ipotesi nulla se $|t| > t_{\frac{a}{2},n-1}$

Accetto l'ipotesi nulla se $|t| \leq t_{\frac{a}{2},n-1}$

3.5.4 T test a un lato:

$\mu \leq \mu_0$

Ipotesi Nulla: H0: ho media $\mu \leq \mu_0$ (quella stimata)

Ipotesi Alternativa: H1: $\mu > \mu_0$

$p - value = P(x \geq t)$
t è la distanza tra il valore della media campionaria e la nostra media stimata diviso la dev std campionaria

$$t = \frac{\sqrt{n} * (x/ - \mu_0)}{S} \tag{3.8}$$

Rifiuto l'ipotesi nulla se $|t| > t_{\frac{a}{2},n-1}$

Accetto l'ipotesi nulla se $|t| \leq t_{\frac{a}{2},n-1}$

3.5.5 INVECE PER VEDERE SE MI $\geq$ MI0:

Ipotesi Nulla: H0: ho media $\mu \geq \mu_0$ (quella stimata)

Ipotesi Alternativa: H1: $\mu < \mu_0$

$p - value = P(x \leq t)$
t è la distanza tra il valore della media campionaria e la nostra media stimata diviso la dev std campionaria $t = \frac{\sqrt{n} * (\bar{x} - \mu_0)}{S}$

Rifiuto l'ipotesi nulla se $|t| < t_{\frac{a}{2},n-1}$

Accetto l'ipotesi nulla se $|t| \geq t_{\frac{a}{2},n-1}$

3.6 z test a due lati

Ipotesi Nulla: H0: ho media $\mu = \mu_0$ (quella stimata)

Ipotesi Alternativa: H1: $\mu \neq \mu_0$

$p - value = 2 * \Phi(-|z|)$
z è la distanza tra il valore della media campionaria e la nostra media stimata diviso la dev std della distribuzione

$$z = \frac{\sqrt{n}(\bar{x} - \mu_0)}{\sigma} \tag{3.9}$$

Rifiuto l'ipotesi nulla se $|z| > z_{\frac{a}{2}}$

Accetto l'ipotesi nulla se $|z| \leq z_{\frac{a}{2}}$

3.7 Z test a un lato:

3.7.1 Verificare che $\mu \leq \mu_0$:

Ipotesi Nulla: H0:= ho media $\mu \leq \mu_0$ (quella stimata)

Ipotesi Alternativa: H1:= $\mu > \mu_0$

$p - value = 1 - \Phi(z)$
z è la distanza tra il valore della media campionaria e la nostra media stimata diviso la dev std della distribuzione

$$z = \frac{\sqrt{n}(\bar{x} - \mu_0)}{\sigma} \tag{3.10}$$

Rifiuto l'ipotesi nulla se $|z| > z_a$

Accetto l'ipotesi nulla se $|z| \leq z_a$

3.7.2 INVECE PER VEDERE SE $\mu \geq \mu_0$:

Ipotesi Nulla: H0: ho media $\mu \geq \mu_0$ (quella stimata)

Ipotesi Alternativa: H1: $\mu < \mu_0$

$p - value = \Phi(z)$
z è la distanza tra il valore della media campionaria e la nostra media stimata diviso la dev std della distribuzione

$$z = \frac{\sqrt{n}(\bar{x} - \mu_0)}{\sigma} \tag{3.11}$$

Dunque:

Rifiuto l'ipotesi nulla se $|z| < -z_a$

Accetto l'ipotesi nulla se $|z| \geq -z_a$

3.8 Stima della Proporzione di popolazione

La proporzione della popolazione è uno stimatore della probabilità di un evento bernoulliano
$\hat{p}$ = p cappello
Intervalli di confidenza (inferenza) a due lati per un proporzione di popolazione:
$p \in (\hat{p} - z_{\frac{\alpha}{2}} * \sqrt{\frac{\hat{p}*(1-\hat{p})}{n}}, \hat{p} + z_{\frac{\alpha}{2}} * \sqrt{\frac{\hat{p}*(1-\hat{p})}{n}})$
con $\hat{p} = \frac{x}{n}$

3.8.1 Intervalli di confidenza a un lato per una proporzione di popolazione:

$p \in (\hat{p} - z_a \sqrt{\frac{\hat{p}*(1-\hat{p})}{n}}, 1)$

$p \in (0, \hat{p} + z_a \sqrt{\frac{\hat{p}*(1-\hat{p})}{n}})$

3.9 Test di ipotesi a due lati:

Ipotesi Nulla: H0: $p = p_0$

Ipotesi Alternativa: H1: $p \neq p_0$

p-value Due valori possibili: (X crediamo sia una binomiale coi parametri stimati)

$$\hat{p} = \frac{\bar{x}}{n} > p_0 \Rightarrow p - value = 2 * P(X \leq \bar{x}) \tag{3.12}$$

$$\hat{p} = \frac{\bar{x}}{n} < p_0 \Rightarrow p - value = 2 * FI(-|z|) \Rightarrow z = \frac{\bar{x} - np_0}{\sqrt{(np_0(1-p_0))}} \tag{3.13}$$

Dunque:

Accetto l'ipotesi nulla se $|z| \leq z_{\frac{\alpha}{2}}$

Rifiuto l'ipotesi nulla se $|z| > z_{\frac{\alpha}{2}}$

3.10 Test del Chi quadro di Pearson

Questo test verifica se la mia distribuzione segue effettivamente la distribuzione con cui l'ho modellata (bonta' dell'adattamento)

Ipotesi Nulla H_0: $P(Y = i) = p_i \quad 1 \leq i \leq k$

Ipotesi Alternativa H_1: $P(Y = i) \neq p_i \quad 1 \leq i \leq k$

Ciascuna delle var Y_j assume il valore i con probabilita' p_i, quindi X_i e binomiale di parametri n e p_i e il suo valore atteso e' np_i
T: $\sum_1^{K-1} \frac{(X_i - np_i)^2}{np_i}$

L'ipotesi nulla va rifiutata per T troppo grande. Quando T e' troppo grande dipende dal livello di significativita' a del test. La regione critica la calcolo con il valore c per cui:

P H0(T $\geq$ c)= a

Ovvero quando H0 è falsa, T è superiore a c con probabilità a Il valore critico si trova sfruttando il fatto che con n molto grande la distr. t si comporta come una chi quadro con k-1 gradi di libertà. Quindi c va come $\chi^2_{a,k-1}$

Dunque:

Accetto l'ipotesi nulla se $T > \chi^2_{a,k-1}$

Rifiuto l'ipotesi nulla se $T \leq \chi^2_{a,k-1}$

3.10.1 Test Chi quadro per due campioni indipendenti:

Tabella di contingenza:

Le tabelle di contingenza sono un particolare tipo di tabelle a doppia entrata (cioè tabelle con etichette di riga e di colonna), utilizzate in statistica per rappresentare e analizzare le relazioni tra due o più variabili. In esse si riportano le frequenze congiunte delle variabili.

Questa variante del test del chi quadro verifica l'ipotesi che due campioni siano indipendenti e derivino dalla stessa popolazione. Cioè

Ipotesi Nulla H_0: le variabili sono indipendenti.

Ipotesi Alternativa H_1: le variabili non sono indipendenti.
Organizzati i dati in una tabella di contingenza g*2:

χ^2 (chi quadro) va come : $\sum_{i=1}^{g}\sum_{j=1}^{2}\frac{(n_{ij}-E_{ij})^2}{E_{ij}} == \sum_{i=1}^{g}\sum_{j=1}^{2}\frac{n_{ij}^2}{E_{ij}} - n$
dove:

- nij = numero casi osservati nel campione j che corrispondono alla iesima modalita
- Eij = numero di casi attesi nel campione j e per la iesima modalita' del caso se H0 fosse vera
- g = numero di modalita' nella quale si esprime la variabile nominale
- n = la numerosita' dei due campioni messi insieme

per via dell'ipotesi dell'indipendenza dei campioni si ha:

$$E_{ij} = \frac{n_i n_j}{n}$$

con n_j = numerosita' di ciascun campione n_i= la frequenza marginale per ciascuna delle g modalita'

Quindi con campioni sufficientemente grandi la nostra funz va come la χ^2_{g-1} e tutte le osservazione fatte sopra sono vere anche in questo caso.

Chapter 4

Regressione Lineare

La regressione è una tecnica statistica che si usa per lo studio della correlazione di una o più variabili indipendenti:

4.1 Regressione Lineare Semplice:

si cercano i parametri di una funzione lineare che leghi Y a X. In particolare studiamo il coefficiente angolare della retta: È positivo quando Y cresce all'aumentare di X È negativo quando Y decresce all'aumentare di X È 0 se Y non varia al variare di X
Il coefficiente angolare è la covarianza di x e y diviso la varianza campionaria di X, cioè
$b_1 = \frac{COV(X,Y)}{s_x^2}$ Invece l'intercetta è $b_0 = \bar{y} - b_1\bar{x}$

Chapter 5

Domande Orale

l' uniforme , un esercizio sull'uniforme simile all'esame poi media campionaria, varianza e media della media campionaria (anche nel caso uniforme che avevo calcolato prima), legge debole dei grandi numeri, chebishev
dimostrazione della legge debole dei grandi numeri
l'esercizio del compito quello con XY e poi una domanda strana, tipo avevo delle variabili aleatorie X1... Xn e Y1...Yn e dovevo dirgli i vari legami
multinomiale

CDF, PMF, PDF, valore atteso varianza le due diseguaglianze e la legge debole dei grandi numeri, varianza, media, varianza campionaria e media campionaria, limite centrale, e poi le varie distribuzioni, geometrica Poisson e il suo processo, esponenziale

5.1 Legame tra la Binomiale Negativa e la Normale

Normal approximation to the Negative Binomial is valid when the number of required successes, s, is large, and the probability of success, p, is neither very small nor very large. This approximation can be justified via Central Limit Theorem, because the NegBin(s, p) distribution can be thought of as the sum of s independent NegBin(1, p) distributions. In practice, some difficulty is knowing whether the values for s and p fall within the bounds for which the Normal distribution is a good approximation. The smaller the value of p, the longer the tail of a NegBin(1,p) distribution would be.

5.2 Approssimare una variabile aleatoria a cazzo a una Normale standard passando dal limite centrale

La mia variabile aleatoria X ha media μ e varianza σ^2 Io col teorema del limite centrale determino la media e la varianza della variabile X_n data dalla somma di X con se stessa per un numero n molto grande di volte,
Mi ritrovo con $X_n = \sum^n X_i$
Che ha parametri $n\mu$ e $n\sigma^2$, siccome questa somma va come una gaussiana per il teorema del limite centrale, applicando la formula $\bar{X} = \frac{X_n - n\mu}{\sigma\sqrt{n}}$ so che $\bar{X}$ è una normale standard.

5.3 Intervalli di confidenza

GLi intervalli di confidenza definiscono zone in cui posso trovare il parametro Θ che mi interessa. Sono o a due lati o a un lato. Tipo considerando la t-distribution l'intervallo di confidenza per la media di una popolazione μ si basa su un campione di n dati continui, una media campionaria $\hat{x}$ e una deviazione standard campionaria S: con un certo livello di confidenza $1-\alpha$ l'intervallo a due lati della media di una popolazione è dato da $\mu \in (\bar{x} - \frac{t_{\frac{\alpha}{2},n-1}S}{\sqrt{n}}, \bar{x} + \frac{t_{\frac{\alpha}{2},n-1}S}{\sqrt{n}})$, invece a un lato è uguale ma con α invece che $\frac{\alpha}{2}$ poichè definisce il limite superiore o inferiore mentre l'altro estremo è $-\infty$ o $+\infty$.
Per un punto critico fissato, la lunghezza dell'intervallo è L inversamente proporzionale alla radice della dimensione del numero dei campioni

5.4 Teorema del limite centrale

Dato un numero sufficientemente grande di variabili aleatorie indipendenti identicamente distribuite(stessa media e varianza) , la loro somma tende a essere una normale con media $n\mu$ e varianza $n\sigma^2$

5.5 Fundamental bridge:

http://www.hcs.harvard.edu/cs50-probability/fundamentalbridge.php(Blitzstein) ha detto che è una formula
$E(I_J) = P(J)$
Vuol di che se faccio una variabile I che può esse 1 o 0, alla fine il valore medio di questa variabile è pari alla probabilità dell'evento J

5.6 Cosa è un indicatore

È quella variabile che c'ha valore 0 o 1 a seconda che becca un successo o no (una variabile di Bernoulli)(una Booleana)

5.7 Cioè basta che gli dimostri chebishev

1: Ma che cazzo vuol dire "basta" ma che è chebishev
2: Nella statistica descrittiva la disuguaglianza di Chebishev afferma che un valore a caso della distribuzione in esame ha probabilità di almeno $\frac{1}{\lambda^2}$ di essere situato nell'intervallo compreso tra $\mu - \lambda\sigma$ e $\mu + \lambda\sigma$
3: Per variabili dotate di media e varianza non si può trovare una disuguaglianza migliore di quella di Chebishev

5.8 Cos'è il test di verifica di un'ipotesi?

Per esempio dato un valore μ_0 rimandiamo al manga delle scienze volume 05, Statistica. (che lo spiega molto bene con l'ausilio delle marionette)

5.9 Errori di Tipo 1 e 2

L'errore di tipo 1 consiste nel rigettare H0 pur essendo H0 è vera. Definito α il livello di significatività scelto per il test di ipotesi, esso esprime anche la probabilità che si manifesti un errore detto di tipo 1.

L'errore di tipo 2 è quando non si rigetta H pur essendo H_0 falsa.

Chapter 6

Ringraziamenti

Grazie Tomonobu Itagaki
"Questi appunti sono una merda, non li ho scritti io. Ci sarebbero state più tette" - Nobu
Grazie Lindo Ferretti.
"Lode a piccioni e a Tomonobu.
Tu devi scomparire anche se non ne hai voglia, e contare solo su Hayter".
Grazie autore del manga delle scienze Shin Takahashi (non abbiamo capito se è l'autore o il detentore dei diritti)
Di nuovo grazie Ananas Maldestro.