Statistica

Probabilità

Teoria che si riferisce ai possibili risultati di un esperimento.
→ Causa Deterministica [esperimento modellabile tramite precise funzioni matematiche]
→ Causa Probabilistica [esperimento NON modellabile tramite funzioni matematiche]
(Analizzeremo solo la causa probabilistica)

Spazio dei campioni
Insieme di tutti i possibili risultati di un esperimento.

Es. Lancio di una moneta → S = {T, C} (spazio dei campioni)
Es Lancio di due monete → S = {TT, TC, CT, CC}

Modalità
- Con Ordinamento (l'ordine di estrazione ha importanza)
- Senza Ordinamento (l'ordine di estrazione non ha importanza)
- Con Reimmissione (estrazione e successivo reinserimento)
- Senza Reimmissione (estrazione e tolgo l'elemento)

Evento
Insieme sottostante A dell'insieme dei campioni S → $A \subseteq S$
Es. S = {TT, TC, CT, CC}
A_1 = {TT}
A_2 = {TC, CT, CC}

Probabilità
$P:S(A) \rightarrow [0,1] \in R$
S(A) insieme di tutti i possibili eventi (insieme di tutti i possibili sottoinsiemi)

<u>3 assiomi:</u>
1. La probabilità di un qualsiasi evento A è sempre ≥ 0 per ogni A → $\mathbf{P(A) \geq 0 \ \forall A}$
2. La probabilità che si verifichi almeno uno di tutti i possibili risultati è 1 → $\mathbf{P(S) = 1}$
3. Se $A_1, A_2, \dots, A_n$ sono disgiunti => $\mathbf{P(A_1 \cup A_2 \cup \dots \cup A_n) = \sum P(A_i)}$

Equally Likely Model (ELM)
$$P(A) = \frac{\#(A)}{\#(S)} \quad \begin{matrix} \text{(numero di risultati favorevoli)} \\ \text{(numero di risultati possibili)} \end{matrix}$$

Es. lancio di due dadi, probabilità in cui escano numeri ≤ 2
A = {11,12,21,22} → 4/36 = 0.11

Proprietà
1. **$P(A^c) = 1 - P(A)$** [La probabilità del complementare di un evento è 1 – la prob dell'evento]
2. **$P(\{\emptyset\}) = 0$**
3. **Se $A \subseteq B \Rightarrow P(A) \leq P(B)$**
4. **$P(A \cup B) = P(A) + P(B)$ [$- P(A \cap B)$]** (parte tra parentesi solo se gli insiemi NON sono disgiunti)

Metodi di conteggio
1) Principio della moltiplicazione
Per prodotto delle probabilità, in generale di due eventi si intende il verificarsi di entrambe <u>simultaneamente</u>.
Ad esempio, se l'evento A consiste nell'estrazione di un due da un mazzo di carte e l'evento B nell'estrazione di una carta di picche, allora l'evento C=AXB consiste nell'apparizione di un due di picche.

2) Ho **n** elementi totali, ne estraggo **k**

ordered=T replace=T	#(S) = **n^k**
ordered=T replace=F	#(S) = **n(n-1)(n-k+1)**
ordered=F replace=T	#(S) = **(n-1+k)!/((n-1)!k!)**
ordered=F replace=F	#(S) = **n!/(k!(n-k)!)**

Probabilità Condizionata
La probabilità che accada l'evento A, calcolata a condizione che l'evento B si sia verificato o meno, si dice **probabilità condizionata** e si denota con: P(A|B).

A = {La prima carta è cuori} P(A) = 13/52
B = {La seconda carta estratta è cuori} P(A) = 12/52

NOTA: Questo implica che l'evento A sia andato a buon fine.

→ Probabilità di B condizionata ad A

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)}$$

→ Probabilità di A condizionata a B

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

Es. ((13/52)*(12/51))/(13/52) = 12/51

Variabile Aleatoria (o casuale)

E' una funzione, indicata con la lettera maiuscola X, che può assumere valori diversi in dipendenza da qualche fenomeno aleatorio.
$X : S \rightarrow R$
$X(s) = x$

Esempio:
Lancio 2 dadi
S = {{1,1},{1,2},…}
supponiamo che la variabile aleatoria **X** indichi quanti 2 ci sono nelle combinazioni di S
X → numero di 2

X({1,1}) = ∅ (non ci sono 2 in questa combinazione)
X({1,2}) = 1
X({1,3}) = ∅
…
X({2,2}) = 2
…

Esempio:
Lancio 2 monete
S = {{TT},{TC},{CT},{CC}}
X → numero di croci in un lancio

X({TT}) = ∅
X({TC}) = 1
X({CT}) = 1
X({CC}) = 2

! In questi casi la funzione ha un numero **discreto** e **finito**, in quanto i risultati sono un numero finito

Infatti:
Lancio tante volte una moneta
X → numero di lanci prima di ottenere testa
Quali valori può ottenere la mia moneta?

X(S) = ∅, 1, 2, … , ∞ → DISCRETE (in questo caso si ha una variabile aleatoria discreta perché l'insieme delle soluzioni è **infinito** ma **numerabile** → è univocamente associato all'insieme dei numeri finiti, cioè so il precedente e il successivo)

Esempio:
Arrivano dei clienti da un benzinaio
X → tempo di attesa per il primo cliente (potrebbe essere ∞ se il cliente non arriva)
X(S) = [0, T] → l'intervallo va da 0 a ∞ (se T = ∞)

In questo caso sono **infiniti** valori ma non numerabili perché ∈ a R, e **continui**, perché per definizione un intervallo NON è numerabile.

Funzione di densità di probabilità

- Caso **DISCRETO**

Sx → supporto di x (cioè l'insieme di tutti i possibili valori) e finito o numerabile.
fx : Sx → [0, 1]   (Funzione di densità di probabilità: **Probability Mass Function**)

fx(x) = P (X = x) (Un certo valore x della funzione è uguale alla prob. di X che è un certo x∈[0,1])
Sx

Esempio:
Lancio 2 monete
S = {{TT},{TC},{CT},{CC}}
X → numero di teste in un lancio

Di seguito le combinazioni di probabilità che la variabile aleatoria può assumere:
X({TT}) = 2
X({TC}) = 1
X({CT}) = 1
X({CC}) = ∅

Sx = {∅, 1, 2} → valori che può assumere la soluzione
fx : {∅, 1, 2} → [0, 1]
fx(∅) = P(X=∅) = ¼ (è presente 1 volta su 4)
fx(1) = P(X=1) = ½ (è presente 2 volte su 4)
fx(2) = P(X=2) = ¼ (è presente 1 volta su 4)

Proprietà di PMF
1) fx(x) ≥ 0   ∀x ∈ Sx
2) ∑(per x∈Sx) fx(x)   = 1   (La somma di tutti gli elementi di fx è 1)

a questa **si associano**:

- **La MEDIA di un PMF**

μ = EX = ∑(per x∈Sx) x*fx(x)
con E la media

<u>Esempio</u> di prima:
fx(∅) = ¼ , fx(1) = ½ , fx(2) = ¼
MEDIA: μ = ∑ x*fx(x) = 0*1/4 + 1*1/2 + 2*1/4 = **1**

- **La VARIANZA**

σ^2 = ∑(per x∈Sx) $(x - \mu)^2$ * fx(x)

<u>Esempio</u> di prima:
$\sigma^2 = (0 - 1)^2 * \frac{1}{4} + (1 - 1)^2 * \frac{1}{2} + (2 - 1)^2 * \frac{1}{4} = \frac{1}{2}$

- La DEVIAZIONE STANDARD
$\sigma = \sqrt{(\sigma^2)}$

<u>Esempio</u> di prima:
$\sigma = \sqrt{(½)^2} = 1/\sqrt{2}$

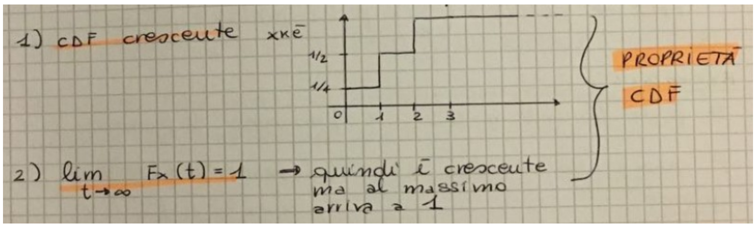
- Funzione di ripartizione o distribuzione di una variabile aleatoria (DISCRETA)
Chiamata **CDF** (**Cumulative Distribution Function**)

$Fx : R \rightarrow [0,1]$

$Fx(t) = P(x \le t)$ (la probabilità è che sia ≤ di t)
con $t \in R$

<u>Esempio</u>:
X{TT} = 2, X{TX} = 1, X{CT} = 1, X{CC} = ∅

$Fx(0.1) = P(x \le 0.1) = ¼$ (mettendo 0.1 che è > 0 ma < 1 la disuguaglianza è valida solo con 0)
$Fx(0.6) = P(x \le 0.6) = ¼$
$Fx(1.6) = P(x \le 1.6) = ½ + ½ = ¾$ (la disuguaglianza è valida sia con 0 che con 1)
$Fx(3) = P(x \le 3) = ¼ + ½ + ¼ = 1$ (la disuguaglianza è valida con tutti: 0, 1, 2)



Ora vediamo alcune particolari funzioni di distribuzione con questi parametri:
- Esperimento
- PMF (forma analitica) μ, σ^2

1 – DISTRIBUZIONE UNIFORME
(Tutti gli esperimenti che hanno la stessa probabilità di accadere)

PMF n di risultati
$fx(x) = 1/n$

$\mu = (n+1)/2$
$\sigma^2 = ((n+1)(2n+1))/6$

2 – DISTRIBUZIONE BINOMIALE
(Descrive il comportamento di una variabile aleatoria → si basa su esperimenti di Bernoulli in cui possono esserci solo 2 risultati (successo o insuccesso)

Esempio:
Nasce un maschio [successo: maschio; insuccesso: femmina]
1) n processi di Bernoulli
2) sono indipendenti
3) ogni processo di Bernoulli ha probabilità P di successo

PMF => fx(x) = (n su p) * p^x * $(1 – p)^{n-x}$ (Per x = 0, 1, … valori numerabili)

μ = n * p = 150/6 = 25
σ^2 = n (n – 1) p^2 = 150*(149)*1/36

Esempio:
5 figli (quindi n = 5) , x = {numero di femmine}
P nasce un maschio 49% (femmina 51%) [dati forniti]

quindi P = 0.51
PMF => fx(x) = (5 su x) * 0.51^x * $(0.49)^{5-x}$

Qual'è la probabilità che nascano 2 femmine? (cioè fx(2) = P(x=2))
PMF => fx(2) = (5 su 2) * 0.51^2 * $(0.49)^{5-2}$

3 – DISTRIBUZIONE DI POISSON
(Eventi rari → n° di accadimenti nell'unità di tempo, es. arrivo di un cliente in una banca)

x = {numero di clienti che arrivano in un ora}
λ = la media di accadimenti nell'unità di tempo (media dei clienti in un ora)

PMF = fx(x) = $e^{-\lambda}$ * λ^x/x! (con x = 0,1,…)

Qual'è la probabilità che arrivino 10 clienti in un'ora? (media λ = 13)
fx(10) = e^{-13} * 13^{10}/10! = P(x=10)
dpois(10, lambda=13) → 0.8587015

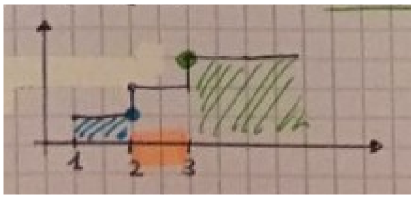
Qual'è la probabilità che arrivi in un'ora un numero di clienti [11,14]?
11 ≤ x ≤ 14
ppois(14, lambda = 13) – **ppois**(10, lambda = 13) Media = λ = 13

Script in R

> Installazione del package '**prob**'
> **dbinom(x, size, prob)** → ritorna PMF (x = punto in cui voglio calcolare la probabilità; size è n; prob = punto in cui voglio andare a calcolare la probabilità)
es. dbinom(2, size=5, p=0.51) → 0.306005
> **pbinom** → ritorna CDF
> **rbinom** → ritorna un insieme di valori casuali con la distribuzione richiesta

<u>Esempio</u>:
Lancio 4 dadi
x = {numero di 2 che escono}
n = 4
p = 1/6
(Bernoulli: uscirà o no un 2?)

$fx(x) = (4 \text{ su } x) * (1/6)^{x} * (5/6)^{4-x}$
Qual'è la probabilità che esca un numero di 2 in [2,3]?
$P(2 \leq x \leq 3)$



$Fx(3) = P(x \leq 3)$
$P(2 \leq x \leq 3) = P(x \leq 3) - P(x \leq 1)$

```
> pbinom(3, size=4, prob=1/6) - pbinom(1, size=4, prob=1/6)
> 0.1311728
```

Questo esempio è valido quando il risultato che vogliamo non è un valore definito <u>ma</u> può essere un intervallo (es. [2, 3])

Con in vettori sarebbe:
> diff(pbinom(c(1,3), size=4, prob=1/6))

rbinom
```
> z = rbinom(1000, size=4, prob=1/6)
> hist(z)                                   istogramma
> hist(z, breaks=20)
```

RIPASSO

Variabili Aleatorie
Insieme DISCRETO: le variabili si possono contare/numerare (finite)
Insieme CONTINUO: le variabili possono assumere valori compresi in un intervallo (infiniti)

<u>DISCRETO</u>
→ Funzione di **massa** (ProbMassFunc)
$fx: Sx \rightarrow [0,1]$ - $fx(x) = P(X = x)$!$[0,1]$ è un insieme numerabile
Sx = supporto di x, insieme di tutti i possibili valori di x
→ Funzione di **distribuzione** (CumulativeDistrFunc)
$Fx: R \rightarrow [0,1]$ - $Fx(t \in R) = P(X <= t)$
1. Distribuzione in forma discreta (un numero finito di risultati)
2. Distribuzione binomiale (Bernulli: ogni evento può avere solo 2 risultati)
3. Distribuzione di Poisson (distribuzione legata al tempo)

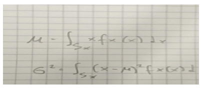
CONTINUO
→ Funzione di **densità di probabilità** (ProbDensitFunc)
fx:Sx → [0,1]   -   P(X∈A) = integrale da a a b(fx(x)dx)   ![0,1] è un intervallo
a e b sono i bound dell'intervallo di A
La prob che X 'stia' in un certo evento A sottinsieme dell'insieme dei campioni S è l'integrale in A di fx(x)dx

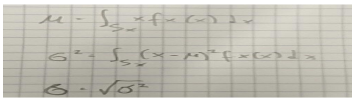
Proprietà:
1. fx(x) >= 0 per ogni x
2. integrale in S (fx(x)dx)   S = insieme dei campioni

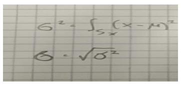
→ Funzione di **distribuzione** (CumulDistrFunc)
Fx:R → [0,1]   -   Fx(t∈R) = P(X<=t) = integrale da -infinito a t di (fx(x)dx)

Proprietà:
1. Fx(t) non decrescente - $t_1 < t_2$ → P(x<=t_1) <= P(X<=t_2)
! Man mano che aumenta t la funzione è sempre non decrescente

Sia discreto che continuo hanno:
a) media
b) varianza
c) deviazione standard







ESPERIMENTI
1) Distribuzione in forma continua
PDF: fx(x) = 1/(b-a)   x∈[a,b] = Sx

Appunti Linguaggio R

Stefania Carapezzi

Salvare i dati creati in R in un file **.Rdata** (salva il workspace)
rm() rimuove i dati nel workspace
load(x,y,file="nomefile.Rdata") ricarica il workspace

Assegnazione variabili (variabile -> valore) oppure (valore <- variabile)
Nei nomi di variabili solo caratteri, numeri, punto e underscore (è case-sensitive)
NOTA: sconsigliato utilizzare l'operatore = per l'assegnazione

Tipi di dato atomici (oggetti di lunghezza 1):
- Numeric
- Complex Number
- Character
- Logical
- Raw (in bytes)

mode(v) e **typeof(v)** ci indicano il tipo di una variabile, oppure **is.numeric(v)** o **is.logical(v)**
Viene restituito come output **na** se c'è un problema nel dato
Viene restituito come output **NaN** se il valore è indefinito

Prove
```
vector1<-c(1,2,3); vector2<-c(3,4,5)
vector1==6
vector1==6; vector2==3
vector1==6 & vector2==3
vector1==6 | vector2==3
vector1[2]<-NA
is.na(vector1); vector2 == 4
any(is.na(vector1)) && vector2 == 4
any(is.na(vector1)) || vector2 == 4
vector1!=3;vector2<=7
xor(vector1!=3, vector2 == 4)
```

Operatori aritmetici

+ - * /	elementi base
^	elevamento a potenza
%%	modulo (resto divisione)
%/%	divisione tra interi
== != < > <= >=	valori base
|| &&	or e and
! | & xor	elementi logici base

Leggere da terminale: var <- scan();

Vettori (creazione)
c(val_1,val_2,…,val_n)
rep(val_1,val_2,…,val_n)
seq(val_1,val_2,…,val_n)

Leggere file scritti in forma di tabella da file

read.table("nome_file", header = T, sep = " ")
! <u>sep</u> è la stringa usata nel file per separare i dati

! header è un booleano T o F indica se la prima riga del file contiene i nomi delle variabili o no
read.csv("nome_file", header = T, sep = " ")
N.B. se nel nome file si mette un indirizzo web vengono presi i dati lo stesso.ù

Scrivere dati su file in forma di tabella

write.table(dati_var, "nome_file", row.names=name_var)
! variabile con i dati da inserire
! stringa per il nome dei file
! variabile con i nomi delle variabili
write.csv(dati_var, "nome_file", row.names=name_var)
Uguale a sopra

N.B. le operazioni aritmetiche con vettori numerici si eseguono applicando gli operatori <u>element-wise</u> (elemento per elemento)
! Se i vettori sono di lunghezze diverse, il più corto è **recycled** (ripetuto)

Creazione di matrici

matrix(val, nrow = nr, ncol = nc, byrow = F)
rbind(matrx, row) → aggiunge una riga alla matrice matrx
cbind(matrx, col) → aggiunge una colonna alla matrice matrx

data.frames(var, names=var_str, row.names=var_str);

Accedere a valori di matrici e vettori
vettori: **v[i]**
matrici: **v[i, j]**

Liste
list()

Appunti Libro (R)

Strutture Dati
→ Vettori
Elemento costituito da un insieme di dati, tutti dello stesso tipo, ognuno con un indice.
In R gli indici partono da 1 (non da 0).
NOTA: le costanti, sono vettori con una sola componente > 23.2 → ## [1] 23.1

Operatori numerici (ordine di priorità) → (), ^, *, /, +, -
Divisione:
  per 0 → ∓ ∞
  tra due valori nulli → NaN

Espansione numero di cifre del risultato (di base sono 7 cifre dopo la virgola)
options(*digits* = X)   [con X da 0 a 22]

Funzioni base
exp(n)   // Esponenziale
log10(n)   // Logaritmo in base 10
log2(n)   // Logaritmo in base 2
log(n)   // Logaritmo in base e
sin(n)   // Seno
cos(n)   // Coseno
sqrt(n)   // Radice quadrata
…

Verificare il tipo di un elemento
→ **str**(elem)   [str sta per structure]
→ **is.<type>**(elem) es. is.integer(2), is.double(3.2)

Operatori logici e di relazione
!, &, |, &&, || (& e | agiscono vettorialmente)
<, >, <=, >=, ==

Vettori atomici (= variabili)
Un vettore atomico è una struttura dati che può contenere n-ple di valori tutti dello stesso tipo.
Per assegnare un valore a una variabile si utilizza il simbolo <- o =
Non è inoltre necessario definire il tipo della variabile

Assegnazione di più valori a un vettore atomico
var ← **c**(1, 3, 40)   [c da concatenate]
[1] 1 3 40
NOTA: con **c()** si può creare anche un vettore vuoto

var ← **seq**(1, 10, 1) [seq da sequence]
[1] 1 2 3 4 5 6 7 8 9 10
NOTA: simile sarebbe stato usare **1:10**

var ← **rep**(2, 5) [rep da replicate]
[1] 2 2 2 2 2

Forzare il tipo degli elementi
as.<type>(var)
Ad esempio **as.integer**(12.4), **as.character**(16)

Stampare un singolo valore all'interno di un vettore
vett[i]

Aggiungere valori a vettori/variabili
var ← **c**(**paste**("x", 1:6, sep=""))
[1] "x1" "x2" "x3" "x4" "x5" "x6"

Fattori
Un fattore (**factor**) permette di rappresentare valori discreti ed è utile per l'analisi di dati qualitativi o ordinali. Es. di dati qualitativi sono M e F oppure T e C, …
factor(**rep**(**c**("T", "C"), **c**(2, 4)))
[1] T T C C C C

Matrici
Strutture bidimensionali di dati, tutti dello stesso tipo.
dim(1:15) ← c(5, 3) [matrice 5 righe per 3 colonne con i valori da 1 a 15]
cbind(1:5, 5:1, 2:6) [unisco questi tre vettori per **c**olonne]
rbind(1:5, 5:1, 2:6) [unisco questi tre vettori per **r**ighe]
matrix(1:15, nrow = 3, byrow = TRUE) { matrice con numeri 1,15 divisa in 3 righe, ordinata per righe]

Stampare un singolo valore all'interno di una matrice
mat[i, j]
mat[,j] [stampa tutta la colonna j]
mat[i,] [stampa tutta la riga i]

NOTE: **rownames**(mat) e **colnames**(mat) definiscono rispettivamente i nomi delle righe e delle colonne della matrice.
t(mat) esegue il trasporto di una matrice (la trasposta)

Array
Gli array sono matrici di dimensione superiore a 2 (es 3x3). Si usa il comando **array**

Dataframe
Struttura dati bidimensionale disomogenei, si usa il comando **data.frame**(vett1, vett2, … , vettn) e per accedere agli elementi/righe/colonne è come per le matrici.
NOTA: mydataframe[[x]] restituisce tutto ciò che contiene x nel dataframe

Liste
Sono una sorta di vettori che possono contenere però elementi di tipo diverso. Si utilizza la keyword **list**(elem1, vettInt2, vettString3, …)

I/O
write.table(dataFrameVar, file = "path/file")
read.table("path/file")
[Simile è **read.csv**]

Controlli
if (condizione) { istruzioni }
else if (condizione) { istruzioni }
else { istruzioni }

switch(var, val1 = 1, val2 = 2, val3 = 3) [se var è == a val1 allora stampa 1]

ifelse(condizione, "successo", "insuccesso")

for(indice **in** intervallo) { istruzioni }