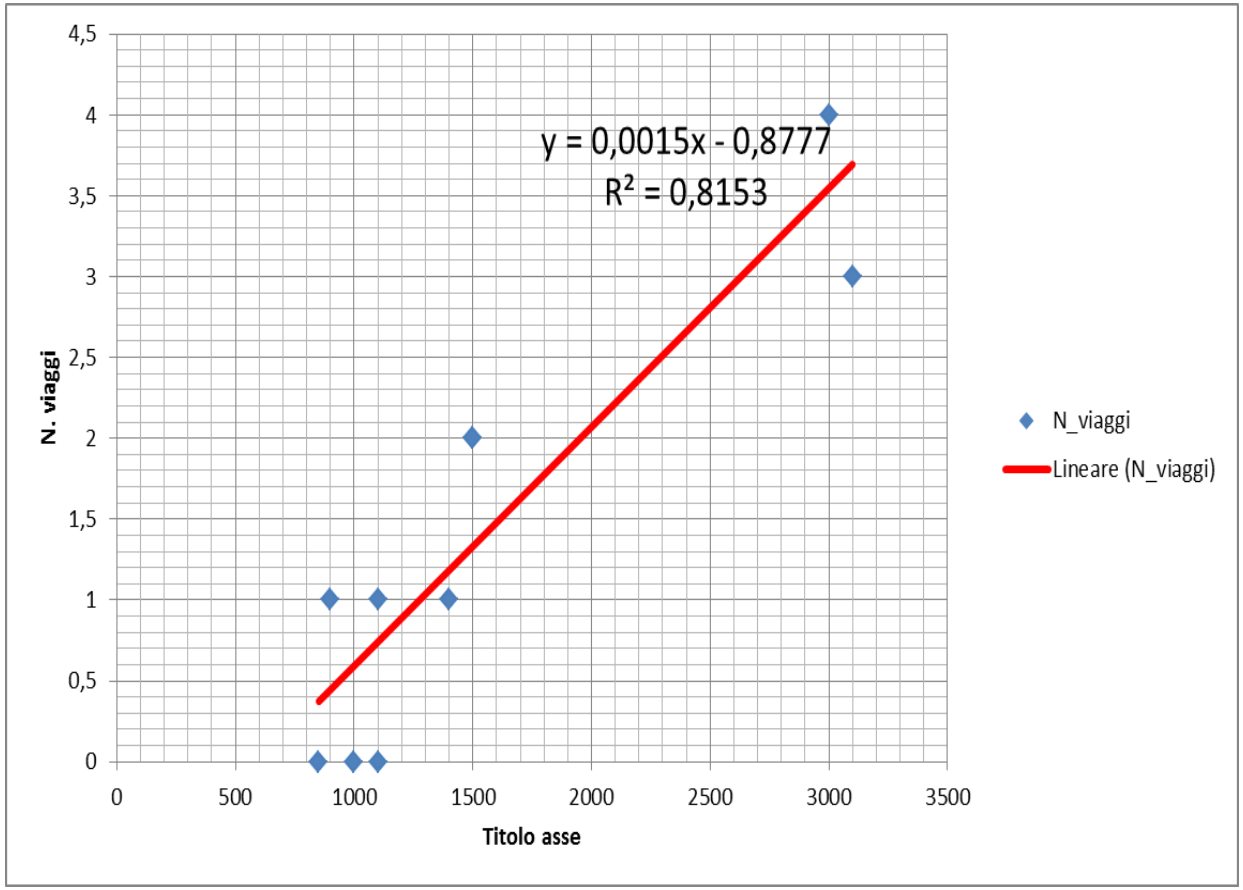
Introduzione alla Regressione Logistica

Contenuto

- regressione lineare semplice e multipla
- regressione logistica lineare semplice
  - La funzione logistica
  - Stima dei parametri
  - Interpretazione dei coefficienti
- Regressione logistica Multipla
  - Interpretazione dei coefficienti
  - Codifica delle variabili
- Esempi in R
- Modellare i propri dati

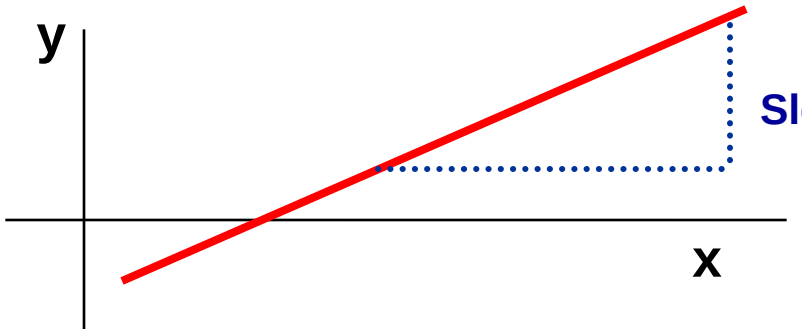
Regressione lineare semplice

N_viaggi	Red
0	1000
2	1500
1	900
4	3000
1	1100
1	1400
0	850
2	1500
3	3100
0	1100



Regressione lineare Semplice

- Relazione tra 2 variabili quantitative (numero viaggi e reddito)



$$y = \alpha + \beta_1 x_1$$

- **coefficiente di Regressione β_1**
  - **Misura l'associazione tra y ed x**
  - **Valore del cambiameto di y in media quando x cambia di una unità**
  - **Metodo dei minini quadrati**

Regression lineare Multipla

- Relazione tra una variabile continua ed un a set di variabili continue

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_i x_i$$

- coefficienti di regressione Parziale β_i
  - Valore del cambiamento di y in media quando x_i cambia di una unità e tutte le altre x_J , per $j \neq i$, rimangono costanti
  - Misura l'associazione tra x_i ed y corretta per tutte le altre x_J

- Esempio
  - Numero viaggi *verso* età, reddito, n. componenti famiglia etc

Regressione lineare Multipla

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_i x_i$$

Dipendente	Variabili indipendenti
Predetta	Variabili predittive
Variabile Risposta	Variabili esplicative
Variabile Esito	Covariate

Analisi Multivariata

Modello	Risultato.
Regressione Lineare	quantitativo continuo.
Regressione di Poisson	conteggi.
Cox model	sopravvivenza.
Regressione Logistica	binomiale.
.....	

- Scelta del modello secondo lo studio, gli obiettivi, e le variabili.
  - Controllo del confondimento.
  - Costruzione di un modello, predizione.

Regressione logistica

- Modella la relazione tra un set di variabili x_i
  - dicotomiche (mangiare : si/no)
  - categoriche (classe sociale, ...)
  - continue (eta', ...)

e

  - Variabile dicotomica Y
  - I modelli di regressione logistica costituiscono una forma particolare dei modelli lineari generalizzati. Sono, in sostanza, una variante dei modelli di regressione lineare.
  - Come è noto, sui dati qualitativi possiedono una elevata autonomia semantica e NON SI POSSONO COMPIERE OPERAZIONI ALGEBRICHE.

Cosa posso fare con i dati qualitativi?

Posso associare le diverse probabilità con cui si manifestano le modalità del carattere Y

ESEMPIO: Se consideriamo 100 individui e 60 hanno acquistato un volo low cost, possiamo fare una lettura in termini probabilistici. Estraendo a caso un soggetto abbiamo una probabilità 0.6 che abbia acquistato un volo low-cost e 0,4 che non l'abbia acquistato

ESEMPIO

DATI: campione di 100 individui

Soddisfazione (0= NO, 1=Si)
Sesso (M=23; F=27)
Età

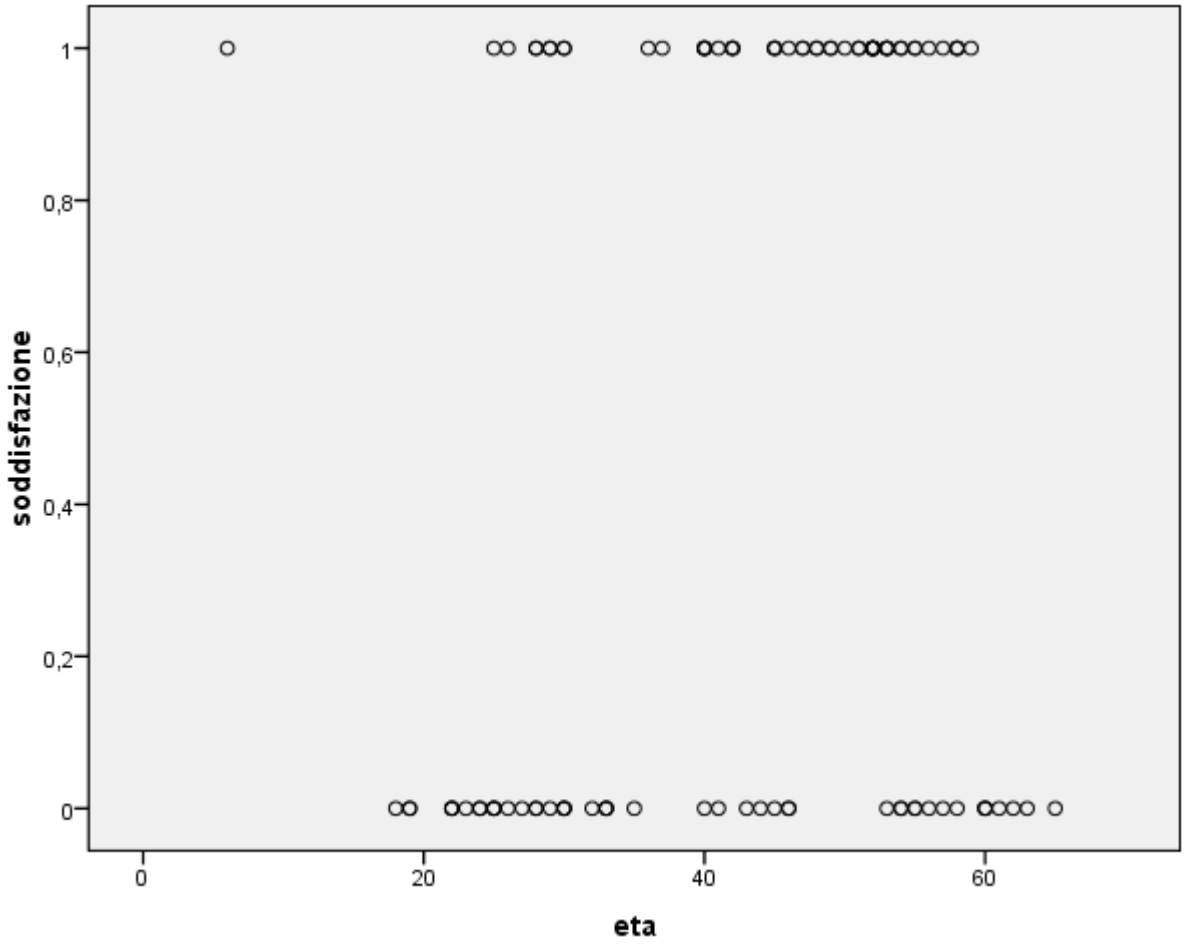
Come possiamo analizzare questi dati ?

- Confronto di Età media degli individui soddisfatti e non soddisfatti

  - Non soddisfatti: 39,15 anni
  - Soddisfatti: 45,40 anni

- Regressione Lineare?

Plot a punti: Dati di Tabella

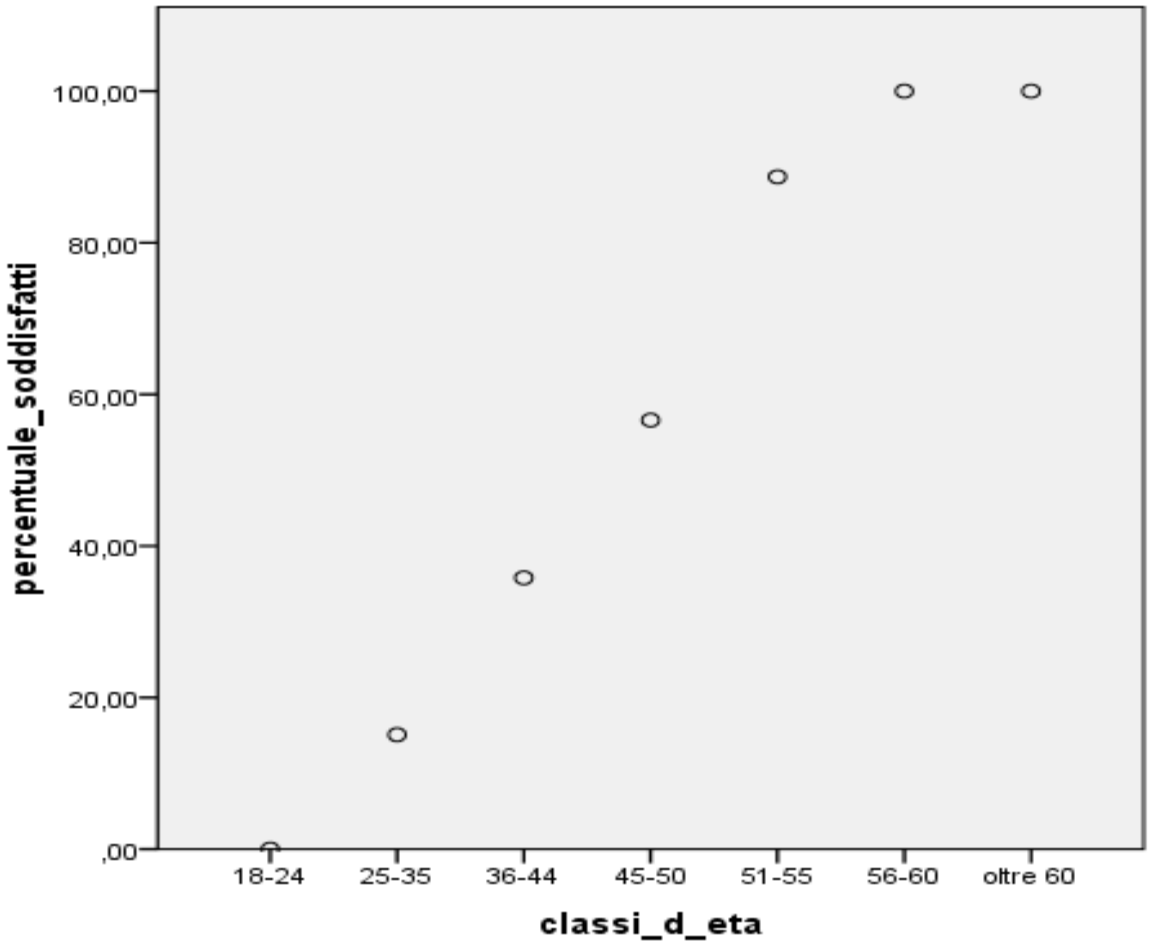


Soddisfazione per classi d'età

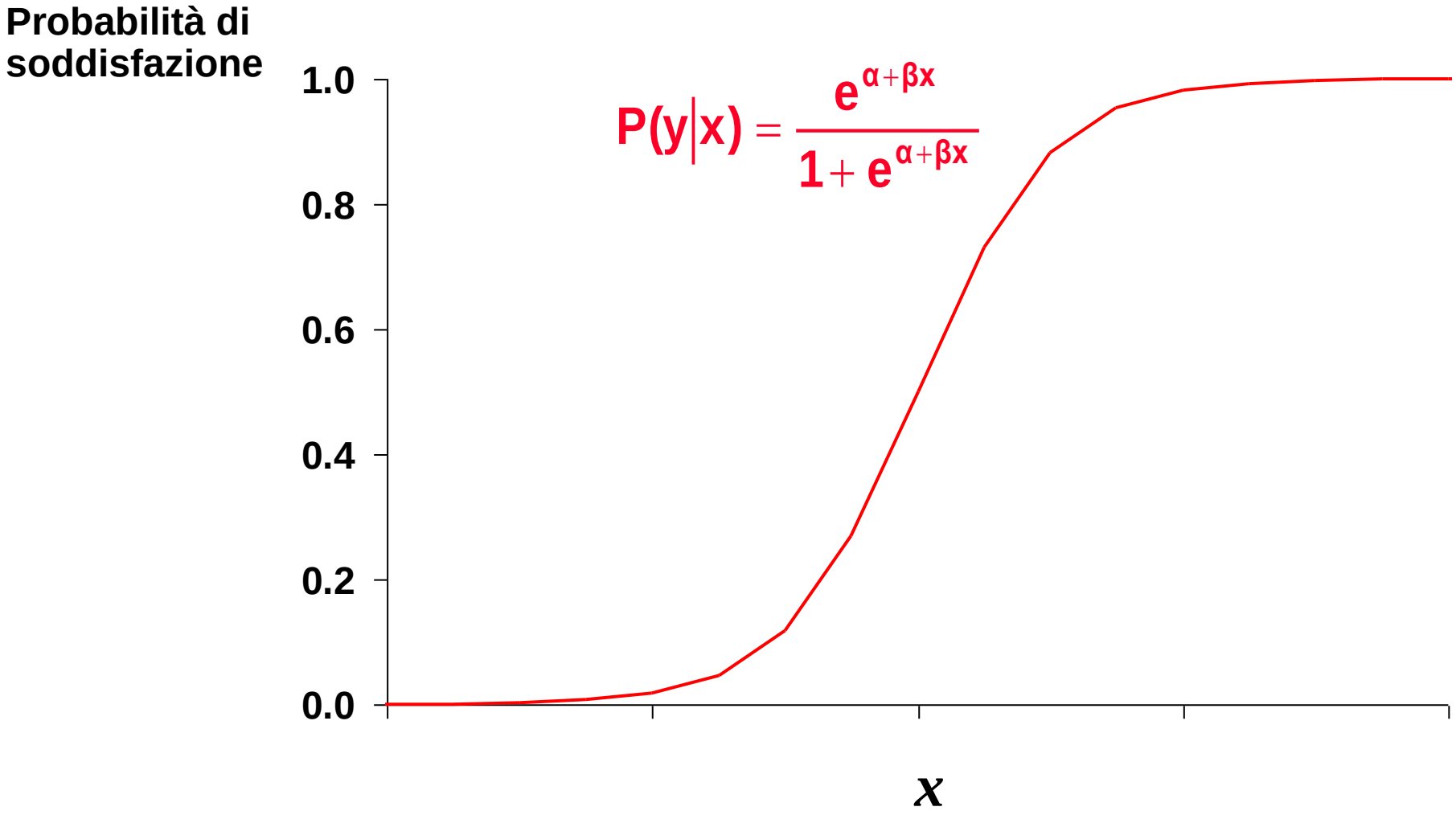
soddisfazione	eta_Classi							
	18-24	25-35	36-44	45-50	51-55	56-60	oltre 60	Total
no	9	16	4	3	5	6	4	47
si	0	8	11	11	17	6	0	53
	9	24	15	14	22	12	4	100

classi d'età	nj	Nj	Fj	PJ
18-24	0	0	0,0	0,0
25-35	8	8	0,2	15,1
36-44	11	19	0,4	35,8
45-50	11	30	0,6	56,6
51-55	17	47	0,9	88,7
56-60	6	53	1,0	100,0
oltre 60	0	53	1,0	100,0

Dot-plot: Dati di Tabella



La funzione logistica (1)



La funzione logistica (2)

$$P(y|x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$$

$$\ln\left[\frac{P(y|x)}{1-P(y|x)}\right] = \alpha + \beta x$$

logit di *P(y|x)*

La funzione logistica(3)

- Vantaggi del logit
  - transformazione semplice di P(y|x)
  - relazione lineare con x
  - Può essere continua (Logit tra - ∞ to + ∞)
  - E' nota la distribuzione binomiale (P tra 0 ed 1)
  - Diretto legame con la nozione di odds di malattia

$$\ln\left(\frac{P}{1-P}\right) = \alpha + \beta x \qquad \frac{P}{1-P} = e^{\alpha+\beta x}$$

Interpretazione di β (1)

	(x)	
oddisfazione (y)	Si	No
SI	$P(y\|x=1)$	$P(y\|x=0)$
No	$1-P(y\|x=1)$	$1-P(y\|x=0)$

$$\frac{P}{1-P} = e^{\alpha+\beta x}$$

$$Odds_{d|e} = e^{\alpha+\beta}$$

$$Odds_{d|\bar{e}} = e^{\alpha}$$

$$OR = \frac{e^{\alpha+\beta}}{e^{\alpha}} = e^{\beta}$$

$$\ln(OR) = \beta$$

Incrocio fra fascia d'età e soddisfazione

soddisfazione	SI	NO	TOTALE
giovani	104	6	110
adulti	405	35	440
TOTALE	509	41	550

soddisfazione	SI	NO	TOTALE
giovani	94,5	5,5	100
adulti	92,0	11,7	100
TOTALE	92,5	7,5	550

- **Fonte: De Lillo et al, 2007**

Calcolo dell'odds ratio

$$odds = \frac{a \times d}{b \times c} = \frac{104 \times 35}{405 \times 6} = 1.5$$

Come va letta questa misura?

La probabilità di dichiararsi soddisfatti per un adulto è di una volta e mezza superiore a quella di un giovane

Calcoliamo l'ODDS RATIO

soddisfazione * eta_2 Crosstabulation

Count

		eta_2		
		GIOVANI	ADULTI	Total
soddisfazione	0	25	22	47
	1	9	44	53
Total		34	66	100

- chisq.test(soddisfazione, eta_2)

=(25*44)/(22*9)=5,5

- Pearson's Chi-squared test with Yates' continuity correction

- data: soddisfazione and eta_2
- X-squared = 12.9862, df = 1, p-value = 0.0003138

Interpretazione di β (2)

- β = incremento del log-odds per incremento unitario di x
- Test d'ipotesi H_0 β=0 (test di Wald)

$$\chi^2 = \frac{\beta^2}{\text{Varianza}(\beta)} \quad (1\ \text{df})$$

- Intervallo di confidenza

$$95\%\ \text{CI} = e^{(\beta \pm 1.96 \text{SE}_\beta)}$$

Adattamento dell'equazione ai dati

- regressione lineare: minimi quadrati
- regressione logistica: massima verosimiglianza
- funzione di verosimiglianza
  - I parametri stimati α e β hanno reso massima la verosimiglianza (probabilità) dei dati osservati rispetto ad ogni altro valore
  - In pratica è più semplice lavorare con log-verosimiglianza

$$L(\mathrm{B}) = \ln\left[l(\mathrm{B})\right] = \sum_{i=1}^{n} \left\{ y_i \ln\left[\pi(x_i)\right] + (1 - y_i) \ln\left[1 - \pi(x_i)\right] \right\}$$

La componente stocastica

Nei modelli logistici vengono applicate principalmente due forme di distribuzione: bernoulliana e multinomiale.

Variabile dipendente dicotomica: distribuzione bernoulliana

$$y_i \in Y_i \approx Bernoulli(\pi_i)$$

Variabile dipendente composta da più di due categorie: distribuzione multinomiale. La componente stocastica può essere considerata una generalizzazione del modello binomiale, dove le k categorie della variabile osservata sono associabili a k variabili casuali di tipo bernoulliano

$$y_i \in Y_{i1},...,Y_{ik} \approx Multinomiale(\pi_{i1},...,\pi_{ik})$$

Distribuzione Bernoulliana

- **Una v.c. Bernoulliana, descrive una prova in cui possono comparire due soli eventi: successo/insuccesso:**

$$P(x)=\begin{cases}1 & \pi \\ 0 & 1-\pi\end{cases} \qquad 0\le\pi\le 1 \qquad P(X)=\pi^{x}(1-\pi)^{1-x}$$

- **parametro**

- **La v.c. assume valori X=0, 1**

- **Valore atteso:** $E(X)=1\cdot\pi+0\cdot(1-\pi)=\pi$

- **Varianza:** $V(X)=(1-\pi)^{2}\cdot\pi+(0-\pi)^{2}\cdot(1-\pi)=\pi(1-\pi)$

Distribuzione Binomiale

Una v.c. Binomiale, rappresenta il numero di successi che si presentano in una sequenza di n sottoprove bernoulliane indipendenti nelle quali è costante la probabilità di successo π.

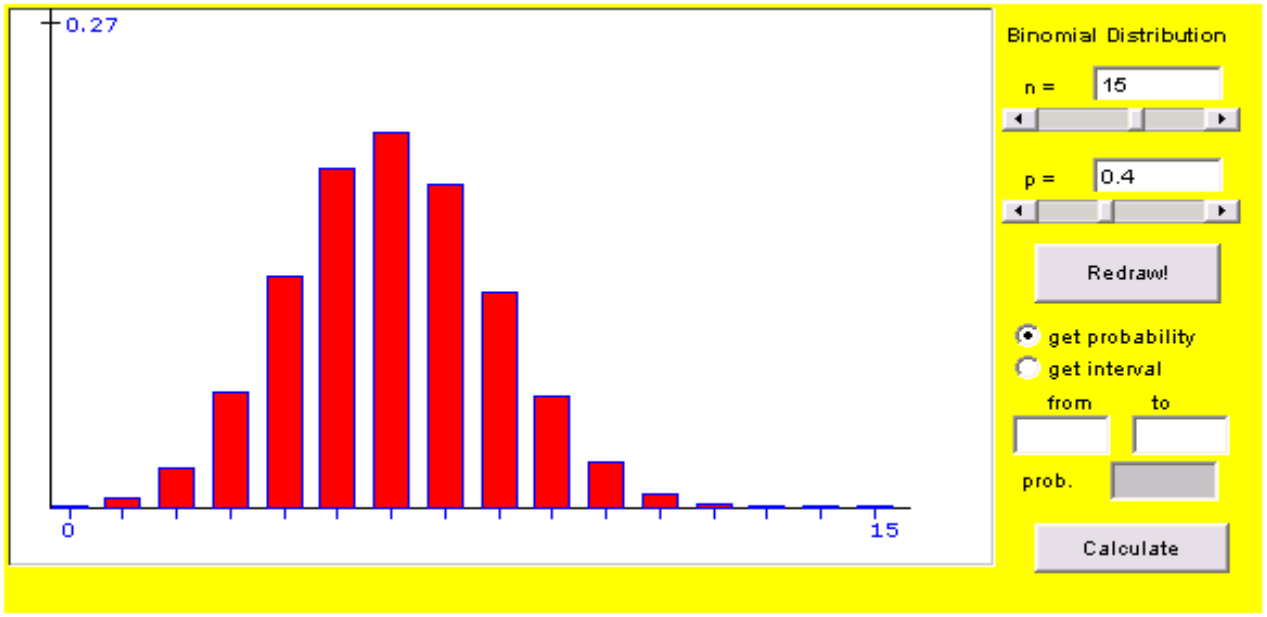
La funzione di probabilità è definita come:

$$P(X=k)=\binom{n}{k}\pi^{k}(1-\pi)^{n-k}$$

- **X=0,1,2,…,n 0<π<1**

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

Distribuzione Binomiale



Media e varianza della distribuzione Binomiale

$E(X) = n \cdot \pi$ • **Media**

$V(X) = n \cdot \pi \cdot (1 - \pi)$ • **Varianza**

- **Proprietà della distribuzione Binomiale**

1. **Il valore atteso e la varianza crescono al crescere di n;**
2. **La distribuzione è simmetrica rispetto al valor atteso (n/2) per π=0,5;**
3. **Per n$\rightarrow$+∞ la distribuzione tende ad essere simmetrica rispetto al valor medio.**

- **Un test con 50 domande vero/falso, qual è la probabilità che rispondendo a caso si rispondi correttamente a 25 domande?**

- **n=50** $\text{a caso} \Rightarrow \pi = 0{,}5$

$$P(X = 25) = \binom{50}{25} (0{,}5)^{25} (1 - 0{,}5)^{25} \approx 0{,}11$$

CALCOLIAMO LA FUNZIONE IN R

dbinom(x,size,prob)

X=25

Size=50

Prob=0,5

dbinom(25, 50, 0.5)

[1] 0.1122752

La regressione logistica in R

```
load("C:/Users/Stella/Desktop/reg.rda")
attach(reg)
mylogit<- glm(soddisfazione~eta_2,family=binomial)
mylogit
```

```
Call:  glm(formula = soddisfazione ~ eta_2, family = binomial(link =
"logit"))

Coefficients:
(Intercept)  eta_2ADULTI
     -1.022        1.715

Degrees of Freedom: 99 Total (i.e. Null);  98 Residual
Null Deviance:      138.3
Residual Deviance: 123.3        AIC: 127.3
```

summary(mylogit)

```
summary(mylogit)

Call:
glm(formula = soddisfazione ~ eta_2, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4823  -0.7842   0.9005   0.9005   1.6304

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.0217     0.3887  -2.628 0.008584 **
eta_2ADULTI   1.7148     0.4683   3.662 0.000250 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Number of Fisher Scoring iterations: 4
```

```
confint(mylogit)
Waiting for profiling to be done...
                             2.5 %      97.5 %
(Intercept)                -1.839729 -0.2959947
as.factor(eta_2)ADULTI  0.826796  2.6758129
exp(mylogit$coefficients)
          (Intercept) as.factor(eta_2)ADULTI
            0.360000               5.555556
```

Massima verosimiglianza

- Calcolo terativo
  - scelta di un valore arbitrario per i coefficienti (usualmente 0)
  - Calcolo della log-verosimiglianza
  - Variazione dei valori dei coefficienti
  - Reiterazione fino alla massimizzazione (plateau)

- Resultati
  - stime di massima verosimiglianza (MLE) per α e β
  - stime di P(y) per a assegnato valore di x

Regressione logistica multipla

- Piu' di una variabile indipendente
  - dicotomica , ordinale, nominale, continua …

$$\ln\left(\frac{P}{1-P}\right)=\alpha+\beta_1 x_1+\beta_2 x_2+\ldots\ \beta_i x_i$$

- Interpretazione di b_i
  - Incremento del log-odds per un Incremento unitario di x_i con tutte le altre x_i constanti
  - misure di associazione tra x_i e log-odds corretta per tutte le altre x_i

Regressione logistica Multipla

- Modifica dell'effetto
  - Puo' essere modellato includendo termini di interazione

$$\ln\left(\frac{P}{1-P}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 \cdot x_1$$

Test dell'ipotesi Statistica

- Domanda
  - Il modello che include una variabile indipendente assegnata fornisce più informazione circa la variabile dipendente del modello in cui tale variabile è assente ?
- Tre test
  - statistica rapporto di verosimiglianza [statistica (LR)]
  - Wald test
  - Score test