Statistica, qualche istruzione per un buon uso

Brunero Liseo
Sapienza Università di Roma
brunero.liseo@uniroma1.it

Sapienza, aprile 2014

Outline

Perché la statistica è importante

La statistica è oggi parte essenziale del bagaglio culturale di ogni studioso.

- Si usa la statistica quando prendiamo decisioni, sia individuali che a livello politico o pubblico.
- usiamo la statistica per comprendere il mondo, per capire quali connessioni logiche esistono tra diversi fenomeni.
- Gran parte delle scelte politiche – che influenzano la vita di ognuno di noi – hanno dietro una giustificazione in termini statistici e per cmprenderne l'impatto e l'importanza occorre conoscere il linguaggio statistico

Quali differenze tra data mining e statistica

Alcuni pareri:

Witten and Franke (2000) Data Mining: Practical Machine Learning Tools with Java

> *What is the difference between machine learnig and statistics? Cynics, looking wryly at the explosion of commercial interest and hype in this area, equate data mining to statistics plus marketing.*

Ripley (2004)

> *. . . Machine learning is statistics minus any checking of models and assumptions . . .*

Cosa serve

Uno statistico deve conoscere

- l'analisi matematica
- il calcolo delle probabilità
- l'algebra lineare
- uno o più linguaggi di programmazione [**R**, C++, Phyton, Matlab, SAS, etc.]

ma soprattutto . . .

- la materia oggetto di studio!

Una storia vera

Ricevi una e-mail con il testo seguente:

> Salve, ti scriviamo a nome della società di investimenti 'Previsioni Perfette!' Affidaci i tuoi risparmi, e ti garantiremo guadagni sicuri!
> Vuoi una prova?
> Domenica prossima il Napoli vincerà 2-1!

Lancio dell'amo...

La settimana successiva, dopo che il Napoli ha effettivamente vinto la partita per 2-1, ricevi un nuovo messaggio:

Lancio dell'amo...

La settimana successiva, dopo che il Napoli ha effettivamente vinto la partita per 2-1, ricevi un nuovo messaggio:

> Salve, siamo sempre noi della 'Previsioni Perfette!'
> Hai visto che precisione?
> Vuoi una prova ulteriore?
> Domenica prossima il Napoli pareggerà 1-1!

Lancio dell'amo...

La settimana successiva, dopo che il Napoli ha effettivamente vinto la partita per 2-1, ricevi un nuovo messaggio:

> Salve, siamo sempre noi della 'Previsioni Perfette!'
> Hai visto che precisione?
> Vuoi una prova ulteriore?
> Domenica prossima il Napoli pareggerà 1-1!

Anche stavolta ...previsione corretta!!

Dopo qualche settimana . . .

Dopo 5 o 6 settimane di previsioni esatte (il risultato viene previsto anche nel punteggio) cominci a vacillare.

Dopo qualche settimana . . .

Dopo 5 o 6 settimane di previsioni esatte (il risultato viene previsto anche nel punteggio) cominci a vacillare.

- È un caso o davvero riescono a prevedere il futuro??!
- Mi conviene rischiare e affidar loro i miei risparmi?

Dopo qualche settimana . . .

Dopo 5 o 6 settimane di previsioni esatte (il risultato viene previsto anche nel punteggio) cominci a vacillare.

- È un caso o davvero riescono a prevedere il futuro??!
- Mi conviene rischiare e affidar loro i miei risparmi?

No . . .

Nel modo più assoluto . Nessuno, fino ad oggi, ha fornito prove scientifiche di essere in grado di prevedere il futuro

Dopo qualche settimana . . .

Dopo 5 o 6 settimane di previsioni esatte (il risultato viene previsto anche nel punteggio) cominci a vacillare.

- È un caso o davvero riescono a prevedere il futuro??!
- Mi conviene rischiare e affidar loro i miei risparmi?

No . . .
Nel modo più assoluto . Nessuno, fino ad oggi, ha fornito prove scientifiche di essere in grado di prevedere il futuro
Ma allora . . . come hanno fatto?!!

È un trucco semplice

La prima settimana la Previsioni perfette invia un milione di e-mail:

È un trucco semplice

La prima settimana la Previsioni perfette invia un milione di e-mail:

- Nelle prime 100 mila c'è scritto che il Napoli vincerà $1 - 0$

È un trucco semplice

La prima settimana la Previsioni perfette invia un milione di e-mail:

- Nelle prime 100 mila c'è scritto che il Napoli vincerà $1 - 0$
- Nelle seconde 100 mila c'è scritto che il Napoli vincerà $2 - 0$

È un trucco semplice

La prima settimana la Previsioni perfette invia un milione di e-mail:

- Nelle prime 100 mila c'è scritto che il Napoli vincerà $1-0$
- Nelle seconde 100 mila c'è scritto che il Napoli vincerà $2-0$
- Nelle terze 100 mila c'è scritto che il Napoli vincerà $2-1$

È un trucco semplice

La prima settimana la Previsioni perfette invia un milione di e-mail:

- Nelle prime 100 mila c'è scritto che il Napoli vincerà $1-0$
- Nelle seconde 100 mila c'è scritto che il Napoli vincerà $2-0$
- Nelle terze 100 mila c'è scritto che il Napoli vincerà $2-1$
- ...

È un trucco semplice

La prima settimana la Previsioni perfette invia un milione di e-mail:

- Nelle prime 100 mila c'è scritto che il Napoli vincerà $1 - 0$
- Nelle seconde 100 mila c'è scritto che il Napoli vincerà $2 - 0$
- Nelle terze 100 mila c'è scritto che il Napoli vincerà $2 - 1$
- ...
- Nelle ... 100 mila c'è scritto che il Napoli pareggerà $0 - 0$

È un trucco semplice

La prima settimana la Previsioni perfette invia un milione di e-mail:

- Nelle prime 100 mila c'è scritto che il Napoli vincerà $1-0$
- Nelle seconde 100 mila c'è scritto che il Napoli vincerà $2-0$
- Nelle terze 100 mila c'è scritto che il Napoli vincerà $2-1$
- ...
- Nelle ... 100 mila c'è scritto che il Napoli pareggerà $0-0$
- Nelle ... 100 mila c'è scritto che il Napoli pareggerà $1-1$
- ...

È un trucco semplice

La prima settimana la Previsioni perfette invia un milione di e-mail:

- Nelle prime 100 mila c'è scritto che il Napoli vincerà $1-0$
- Nelle seconde 100 mila c'è scritto che il Napoli vincerà $2-0$
- Nelle terze 100 mila c'è scritto che il Napoli vincerà $2-1$
- ...
- Nelle ... 100 mila c'è scritto che il Napoli pareggerà $0-0$
- Nelle ... 100 mila c'è scritto che il Napoli pareggerà $1-1$
- ...
- Nelle ultime 100 mila c'è scritto che il Napoli perderà $0-2$

È un trucco semplice

La prima settimana la Previsioni perfette invia un milione di e-mail:

- Nelle prime 100 mila c'è scritto che il Napoli vincerà $1-0$
- Nelle seconde 100 mila c'è scritto che il Napoli vincerà $2-0$
- Nelle terze 100 mila c'è scritto che il Napoli vincerà $2-1$
- ...
- Nelle ... 100 mila c'è scritto che il Napoli pareggerà $0-0$
- Nelle ... 100 mila c'è scritto che il Napoli pareggerà $1-1$
- ...
- Nelle ultime 100 mila c'è scritto che il Napoli perderà $0-2$

Ovviamente si scelgono i dieci risultati più verosimili....

La seconda settimana

Ci si concentra sui 100 mila che hanno ricevuto una previsione corretta, abbandonando gli altri 900mila indirizzi e-mail...

- Nelle prime 10mila c'è scritto che il Napoli vincerà 1-0
- Nelle seconde 10mila c'è scritto che il Napoli vincerà 2-0
- Nelle terze 10mila c'è scritto che il Napoli vincerà 2-1
- ....
- Nelle ... 10mila c'è scritto che il Napoli pareggerà 0-0
- Nelle ... 10mila c'è scritto che il Napoli pareggerà 1-1
- ...
- Nelle ultime 10mila c'è scritto che il Napoli perderà 0-2

Anche stavolta, si scelgono i dieci risultati più verosimili....

Dopo 5 settimane...

.....ci saranno 10 indirizzi e-mail che hanno ricevuto un pronostico corretto per cinque volte di fila e tra quelli ci siete anche voi....

Dopo 5 settimane...

.....ci saranno 10 indirizzi e-mail che hanno ricevuto un pronostico corretto per cinque volte di fila e tra quelli ci siete anche voi....

Tutto ciò non ha nulla a che vedere con la capacità divinatorie della 'Previsioni Perfette'....

Il paradosso di Simpson: un esempio

- Ad Alberto e Barbara piace giocare a basket e si sfidano in una gara di tiri.

Il paradosso di Simpson: un esempio

- Ad Alberto e Barbara piace giocare a basket e si sfidano in una gara di tiri.
- Ognuno prova 200 tiri con i seguenti risultati:

Il paradosso di Simpson: un esempio

- Ad Alberto e Barbara piace giocare a basket e si sfidano in una gara di tiri.
- Ognuno prova 200 tiri con i seguenti risultati:

	Alberto	Barbara
Centri	100	80
Fuori	100	120
Totale	200	200

Il paradosso di Simpson: un esempio

- Ad Alberto e Barbara piace giocare a basket e si sfidano in una gara di tiri.
- Ognuno prova 200 tiri con i seguenti risultati:

	Alberto	Barbara
Centri	100	80
Fuori	100	120
Totale	200	200

Chi è più bravo??

Il paradosso di Simpson: un esempio

- Ad Alberto e Barbara piace giocare a basket e si sfidano in una gara di tiri.
- Ognuno prova 200 tiri con i seguenti risultati:

	Alberto	Barbara
Centri	100	80
Fuori	100	120
Totale	200	200

Chi è più bravo?? Alberto, senza dubbio!
Ha una percentuale del 50% contro il 40% di Barbara.

Un momento però....!

- C'è chi è più bravo nel tiro da fuori e chi nel tiro da sotto....

Un momento però....!

- C'è chi è più bravo nel tiro da fuori e chi nel tiro da sotto....
- Controlliamo meglio i risultati...

Un momento però....!

- C'è chi è più bravo nel tiro da fuori e chi nel tiro da sotto....
- Controlliamo meglio i risultati...

	Alberto			Barbara		
	da fuori	da sotto	Totale	da fuori	da sotto	Totale
Centri	10	90	100	50	30	80
Fuori	30	70	100	100	20	120
Totale	40	160	200	150	50	200

Un momento però....!

- C'è chi è più bravo nel tiro da fuori e chi nel tiro da sotto....
- Controlliamo meglio i risultati...

	Alberto			Barbara		
	da fuori	da sotto	Totale	da fuori	da sotto	Totale
Centri	10	90	100	50	30	80
Fuori	30	70	100	100	20	120
Totale	40	160	200	150	50	200

Chi è più bravo allora ?!?

Facciamo i calcoli!

	Alberto			Barbara		
	da fuori	da sotto	Totale	da fuori	da sotto	Totale
Centri	10	90	100	50	30	80
Fuori	30	70	100	100	20	120
Totale	40	160	200	150	50	200

Facciamo i calcoli!

	Alberto			Barbara		
	da fuori	da sotto	Totale	da fuori	da sotto	Totale
Centri	10	90	100	50	30	80
Fuori	30	70	100	100	20	120
Totale	40	160	200	150	50	200

Nei tiri da fuori

- Alberto ha una percentuale del 25% (10/40)

Facciamo i calcoli!

	Alberto			Barbara		
	da fuori	da sotto	Totale	da fuori	da sotto	Totale
Centri	10	90	100	50	30	80
Fuori	30	70	100	100	20	120
Totale	40	160	200	150	50	200

Nei tiri da fuori

- Alberto ha una percentuale del 25% (10/40)
- Barbara ha una percentuale del 33% (50/150)

Almeno nei tiri da fuori, è meglio Barbara!

Facciamo i calcoli!

	Alberto			Barbara		
	da fuori	da sotto	Totale	da fuori	da sotto	Totale
Centri	10	90	100	50	30	80
Fuori	30	70	100	100	20	120
Totale	40	160	200	150	50	200

Facciamo i calcoli!

	Alberto			Barbara		
	da fuori	da sotto	Totale	da fuori	da sotto	Totale
Centri	10	90	100	50	30	80
Fuori	30	70	100	100	20	120
Totale	40	160	200	150	50	200

Nei tiri da sotto

- Alberto ha una percentuale del 56.25% (90/160)

Facciamo i calcoli!

	Alberto			Barbara		
	da fuori	da sotto	Totale	da fuori	da sotto	Totale
Centri	10	90	100	50	30	80
Fuori	30	70	100	100	20	120
Totale	40	160	200	150	50	200

Nei tiri da sotto

- Alberto ha una percentuale del 56.25% (90/160)
- Barbara ha una percentuale del 60% (30/50)

Facciamo i calcoli!

	Alberto			Barbara		
	da fuori	da sotto	Totale	da fuori	da sotto	Totale
Centri	10	90	100	50	30	80
Fuori	30	70	100	100	20	120
Totale	40	160	200	150	50	200

Nei tiri da sotto

- Alberto ha una percentuale del 56.25% (90/160)
- Barbara ha una percentuale del 60% (30/50)

Anche nei tiri da sotto, Barbara è più brava!

Come è possibile?

- Accade spesso che delle conclusioni apparentemente ovvie siano ribaltate quando si analizzino i dati con maggiore attenzione
- Nel nostro esempio Alberto ha tirato molto più da sotto che da lontano, al contrario di Barbara ... e i tiri da sotto sono, semplicemente, più difficili da realizzare.

Il paradosso di Simpson

Una forma di associazione statistica che vale per tutti i sottogruppi di una popolazione può cambiare direzione allorché i dati vengono aggregati a formare un solo gruppo

Simpson a Berkeley

- Nel 1973 vennero ammessi ai programmi di Ph.D. della prestigiosa università di Berkeley il 44% dei ragazzi che avevano fatto domanda ed il 35% delle aspiranti ragazze.

Simpson a Berkeley

- Nel 1973 vennero ammessi ai programmi di Ph.D. della prestigiosa università di Berkeley il 44% dei ragazzi che avevano fatto domanda ed il 35% delle aspiranti ragazze.
- Le associazioni femministe non solo protestarono vivacemente, ma fecero causa all'università, accusandola di discriminazione sessista.

Simpson a Berkeley

- Nel 1973 vennero ammessi ai programmi di Ph.D. della prestigiosa università di Berkeley il 44% dei ragazzi che avevano fatto domanda ed il 35% delle aspiranti ragazze.
- Le associazioni femministe non solo protestarono vivacemente, ma fecero causa all'università, accusandola di discriminazione sessista.
- L'università si difese (con successo) dimostrando che il risultato cumulativo dipendeva dal paradosso di Simpson.

Simpson a Berkeley

Un'occhiata ai dati

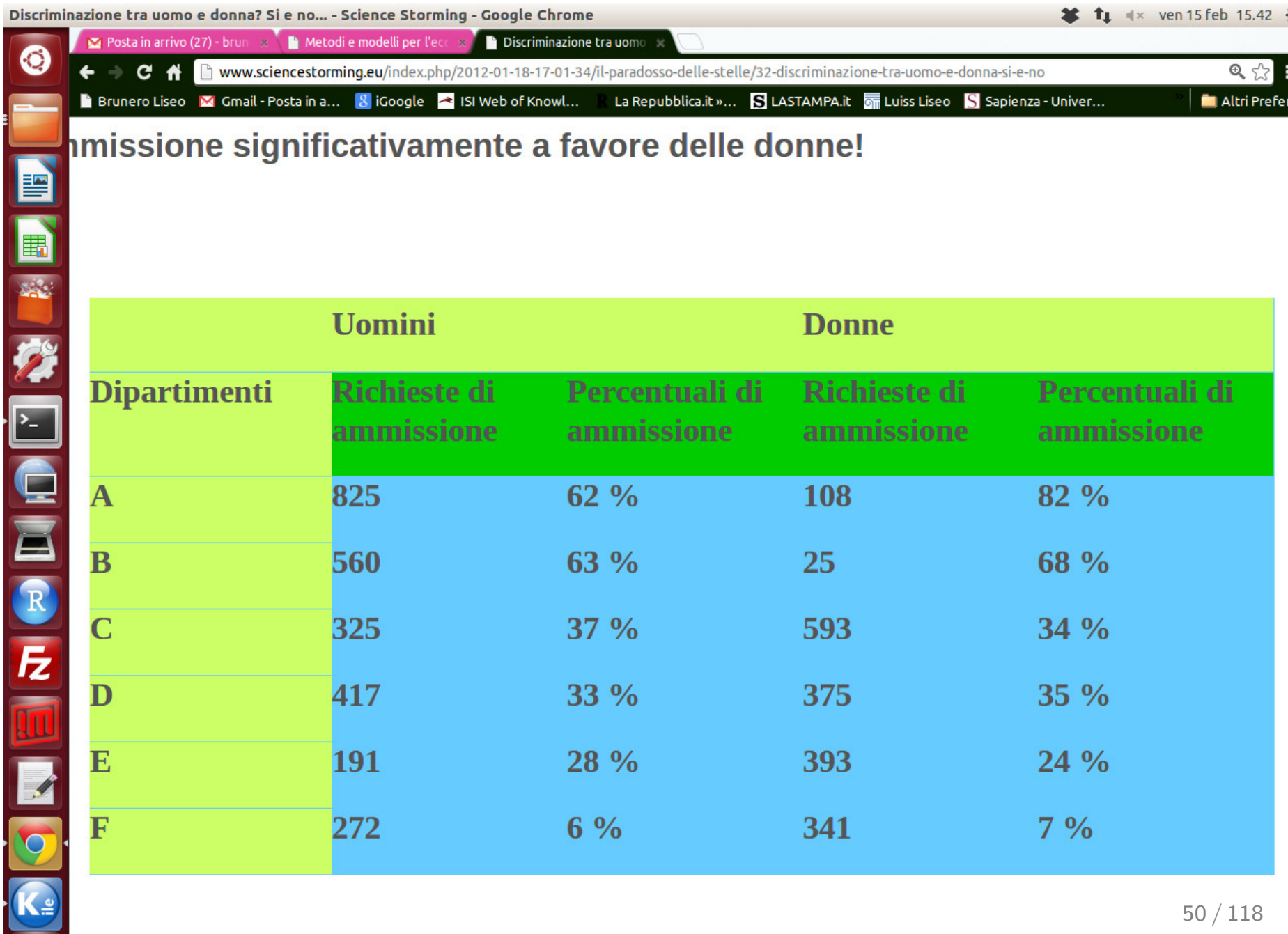
	Richieste di ammissione	Percentuali di ammissione
Uomini	8442	44%
Donne	4321	35%

Simpson a Berkeley

- Analizzando i dati relativi ai singoli dipartimenti, si poteva notare come le ammissioni erano sostanzialmente equilibrate, spesso anzi con una prevalenza a favore delle ragazze.

Simpson a Berkeley

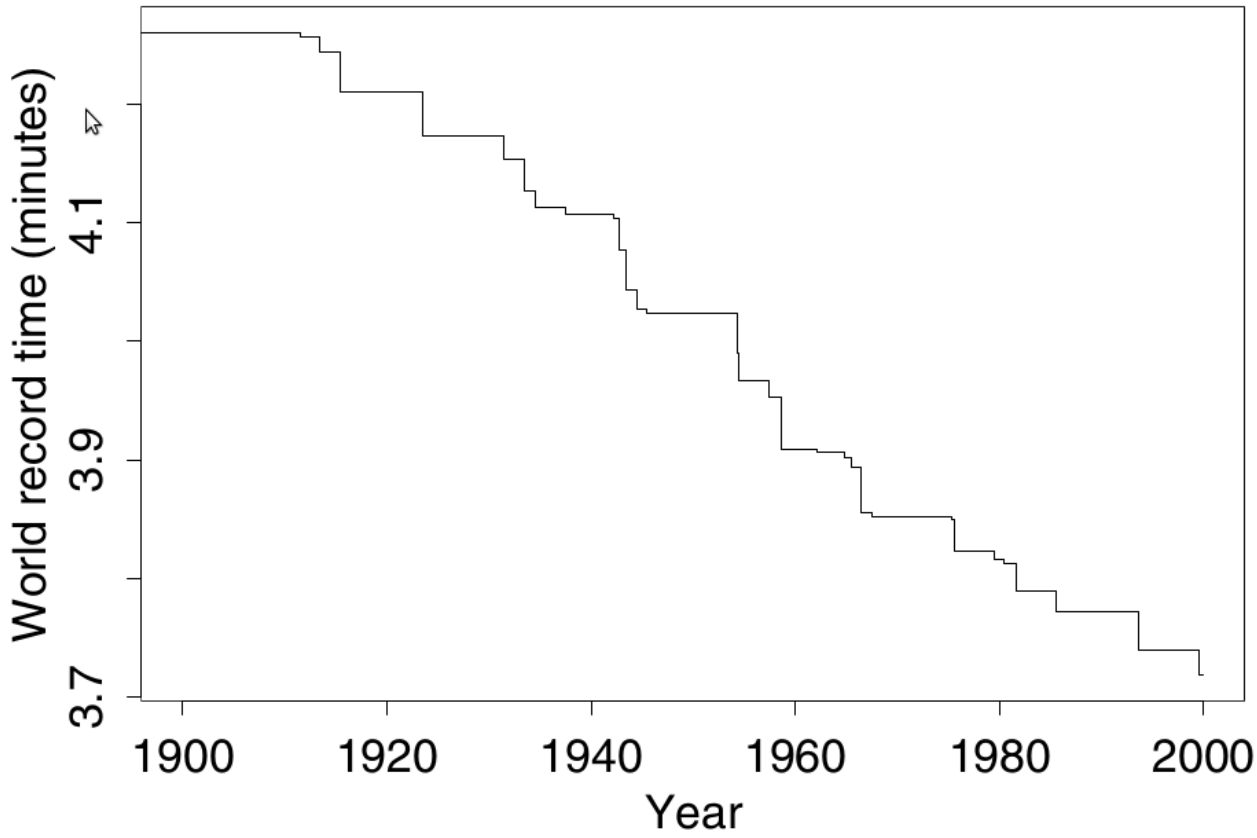
- Analizzando i dati relativi ai singoli dipartimenti, si poteva notare come le ammissioni erano sostanzialmente equilibrate, spesso anzi con una prevalenza a favore delle ragazze.
- Qui la variabile nascosta è rappresentata dalle diverse preferenze manifestate da maschi e femmine.



	Uomini		Donne	
Dipartimenti	Richieste di ammissione	Percentuali di ammissione	Richieste di ammissione	Percentuali di ammissione
A	825	62 %	108	82 %
B	560	63 %	25	68 %
C	325	37 %	593	34 %
D	417	33 %	375	35 %
E	191	28 %	393	24 %
F	272	6 %	341	7 %

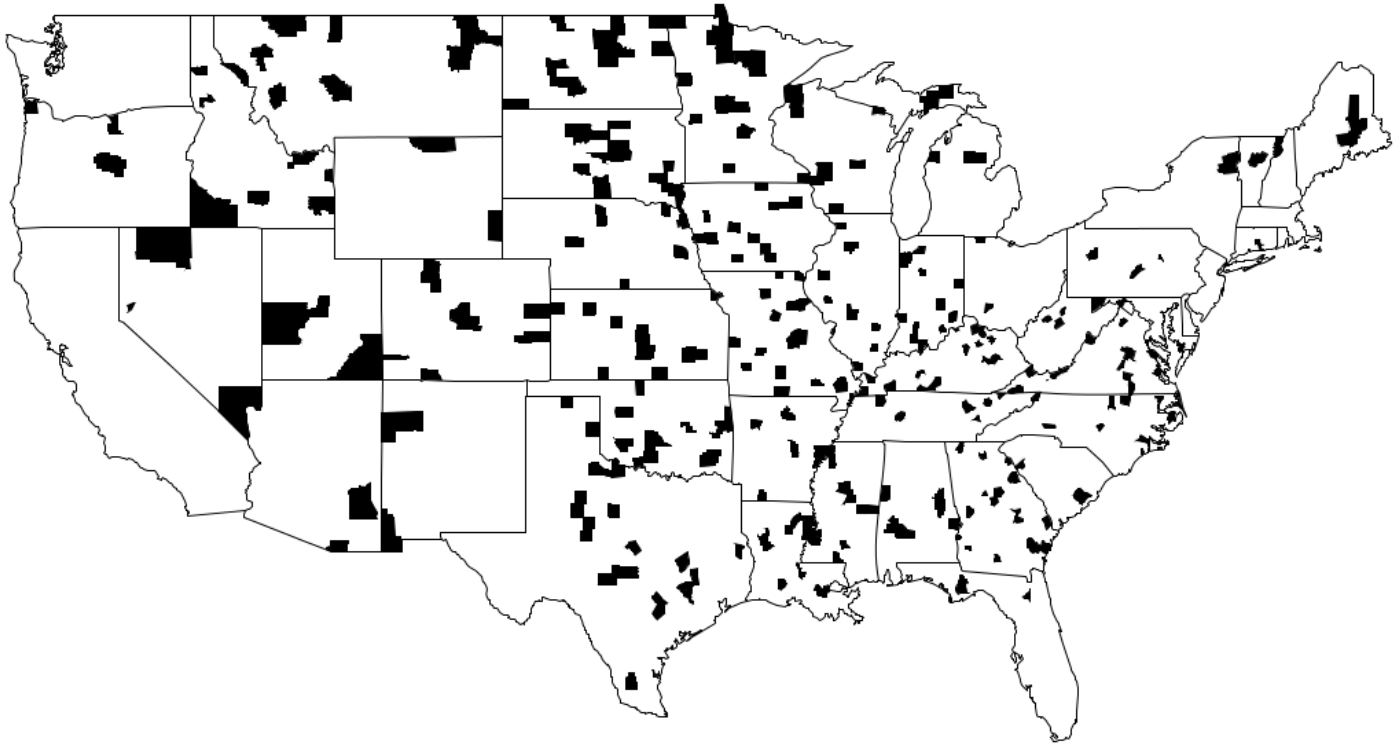
Sapere leggere un grafico

I grafici aiutano a comprendere le relazioni ma occorre saperli interpretare.

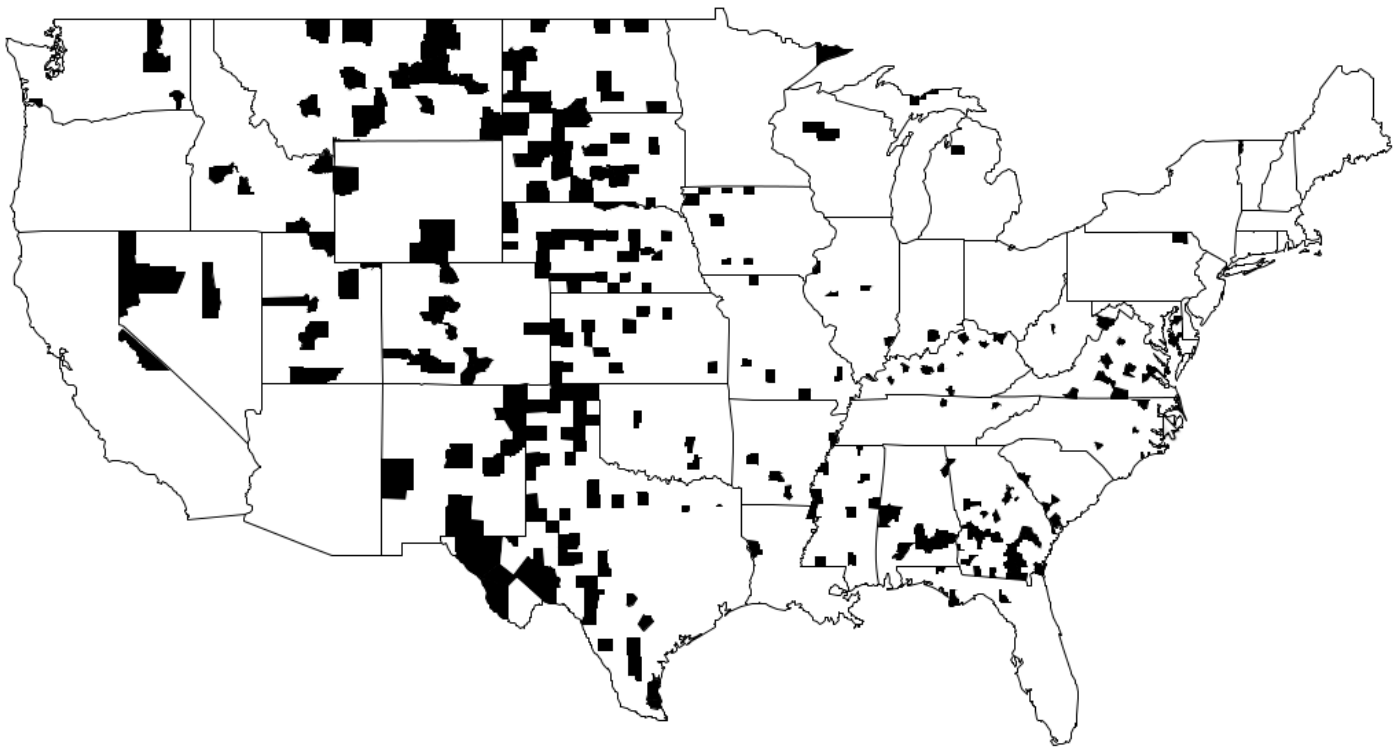


Sapere leggere un grafico

Highest kidney cancer death rates



Quali sono le contee più a rischio?



La regressione

- Esiste un legame tra due fenomeni?

La regressione

- Esiste un legame tra due fenomeni?
- come scoprirlo?

La regressione

- Esiste un legame tra due fenomeni?
- come scoprirlo?
- come descriverlo?

La regressione

- Esiste un legame tra due fenomeni?
- come scoprirlo?
- come descriverlo?
- può aiutarmi a fare previsioni?

La regressione

- Esiste un legame tra due fenomeni?
- come scoprirlo?
- come descriverlo?
- può aiutarmi a fare previsioni?

La regressione

- La retta di regressione è una tecnica di adattamento di un modello teorico ai dati

La regressione

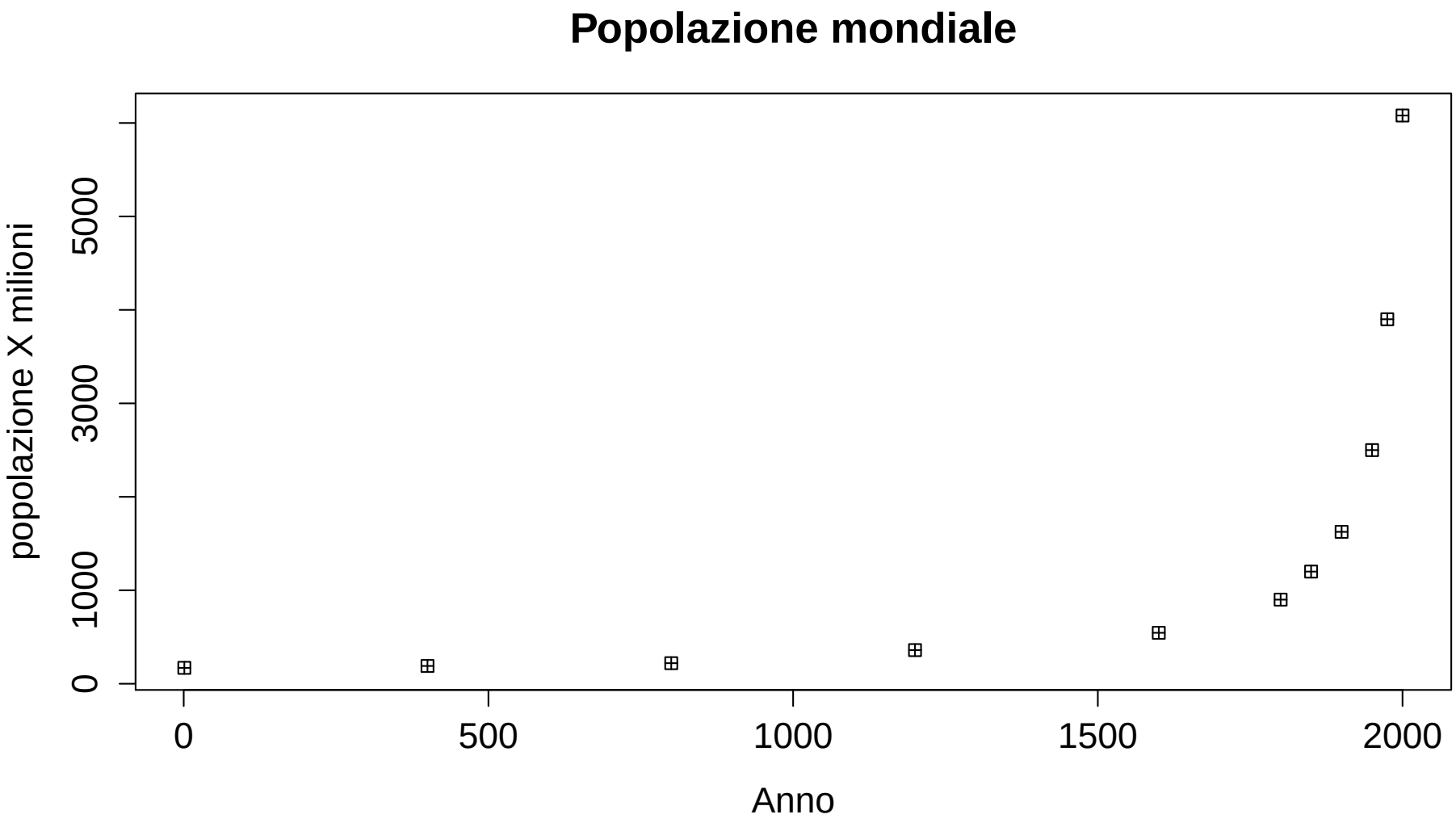
- La retta di regressione è una tecnica di adattamento di un modello teorico ai dati

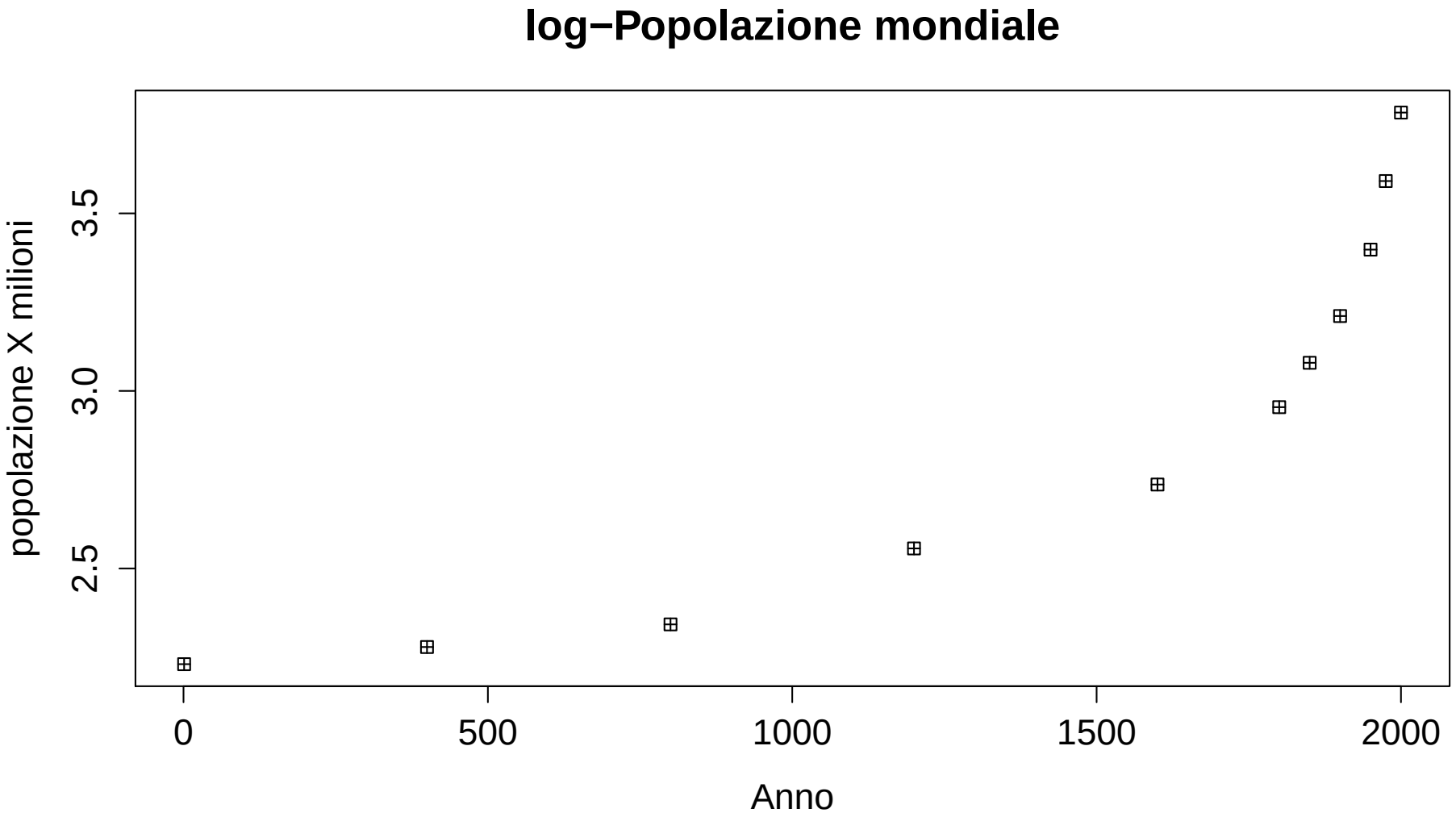
A partire dall'osservazione di due fenomeni $((X, Y)$ su n unità statistiche, assumiamo

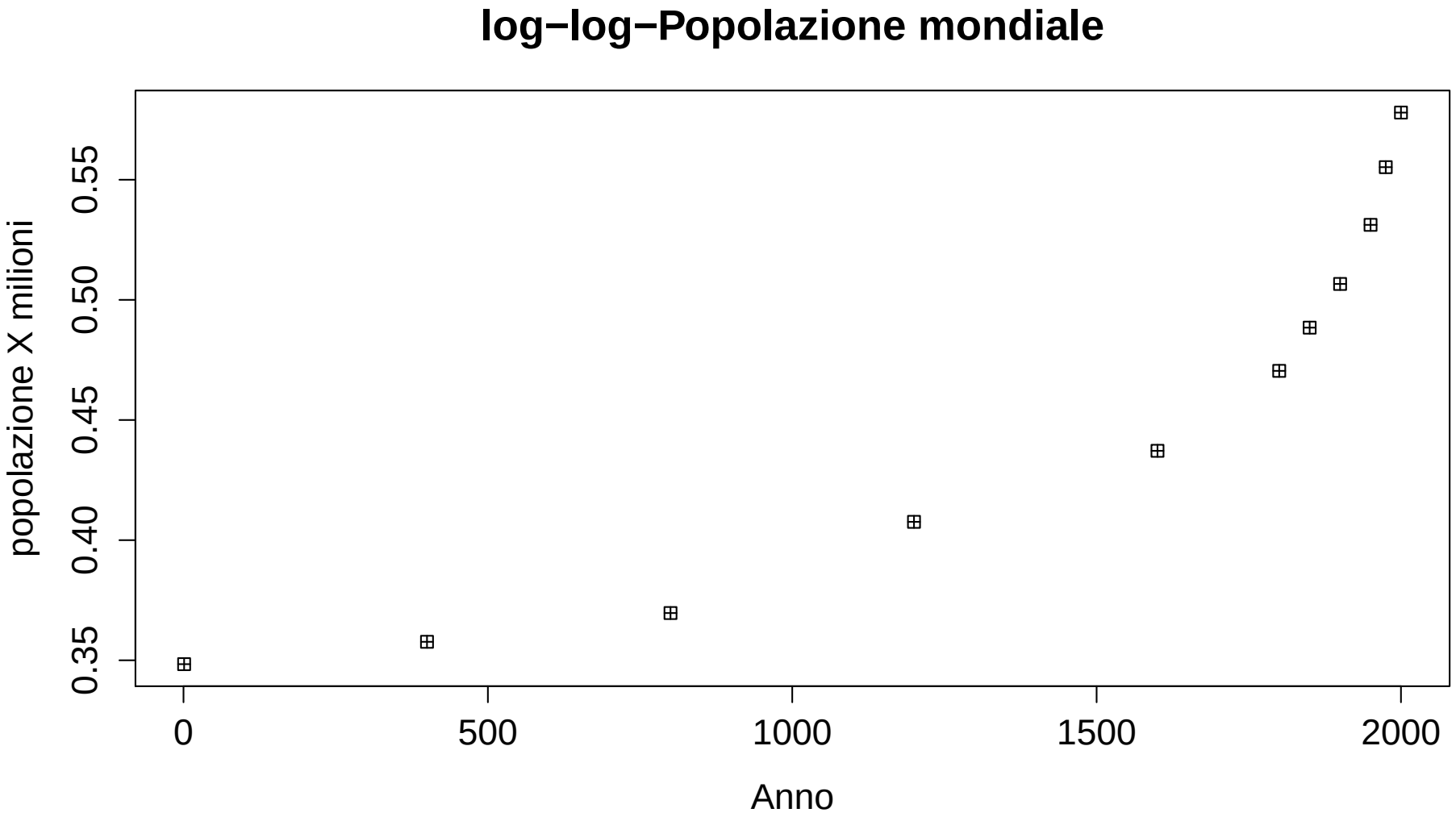
$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \qquad \varepsilon \overset{\texttt{iid}}{\sim} N(0, \sigma^2)$$

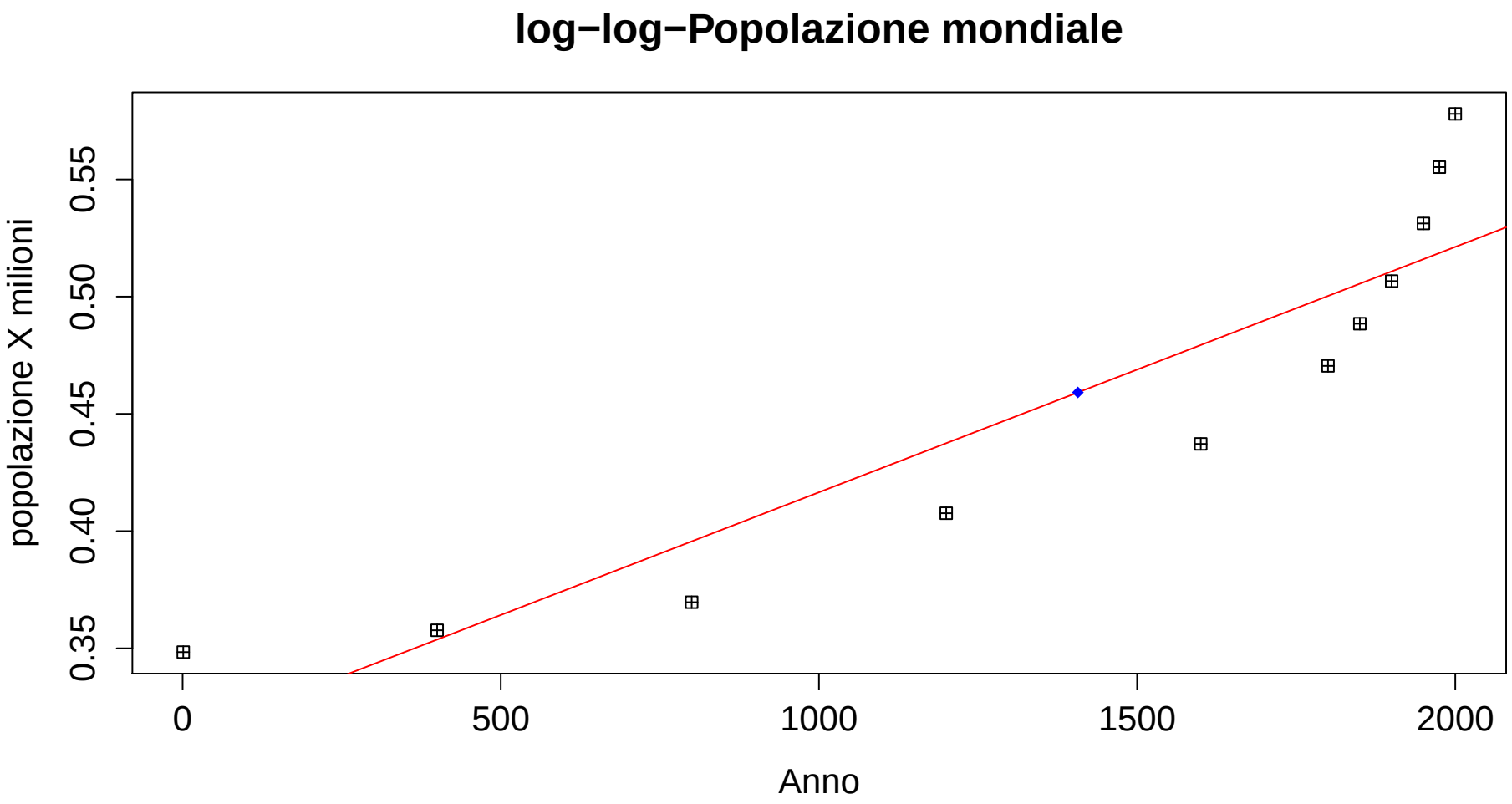
Si cerca di determinare una stima di $(\beta_0, \beta_1, \sigma^2)$ mediante la tecnica dei minimi quadrati o della massima verosimiglianza.

Evoluzione della popolazione mondiale




log−Popolazione mondiale
popolazione X milioni
2.5
3.0
3.5
0
500
1000
1500
2000
Anno


log−log−Popolazione mondiale
popolazione X milioni
0.35
0.40
0.45
0.50
0.55
0
500
1000
1500
2000
Anno


log−log−Popolazione mondiale
popolazione X milioni
0.35
0.40
0.45
0.50
0.55
0
500
1000
1500
2000
Anno

Regressione sulla log-log-Popolazione

```
lm(formula = y ~ anno)
Residuals:
      Min         1Q     Median         3Q        Max
-0.042162 -0.027884 -0.004121  0.025830  0.056731

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.3118680  0.0240242  12.981 3.93e-07 ***
anno        0.0001047  0.0000154   6.797 7.93e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.03444 on 9 degrees of freedom
Multiple R-squared:  0.837, Adjusted R-squared:  0.8189
F-statistic: 46.21 on 1 and 9 DF,  p-value: 7.928e-05
```

Il disastro dello Shuttle (1986)

I dati che seguono si riferiscono alle temperature in gradi Farenheit registrate al momento del lancio nei giorni dei 24 lanci precedenti il disastro del 28 gennaio 1986 ed il numero di avarie verificatesi agli O-rings.
Gli O-rings di gomma sono speciali guarnizioni, che impediscono la fuoriuscita dei gas di combustione al momento del decollo. Essi riprendono subito la loro forma a **temperature calde**, ma non a temperature fredde.

I dati

	1	2	3	4	5	6	7	8	9	10	11
temperature	53	57	58	63	66	67	67	67	68	69	70
num. avarie	5	1	1	1	0	0	0	0	0	0	1

	12	13	14	15	16	17	18	19	20	21	22	23
temperature	70	70	70	72	73	75	75	76	76	78	79	81
num. avarie	0	1	0	0	0	0	1	0	0	0	0	0

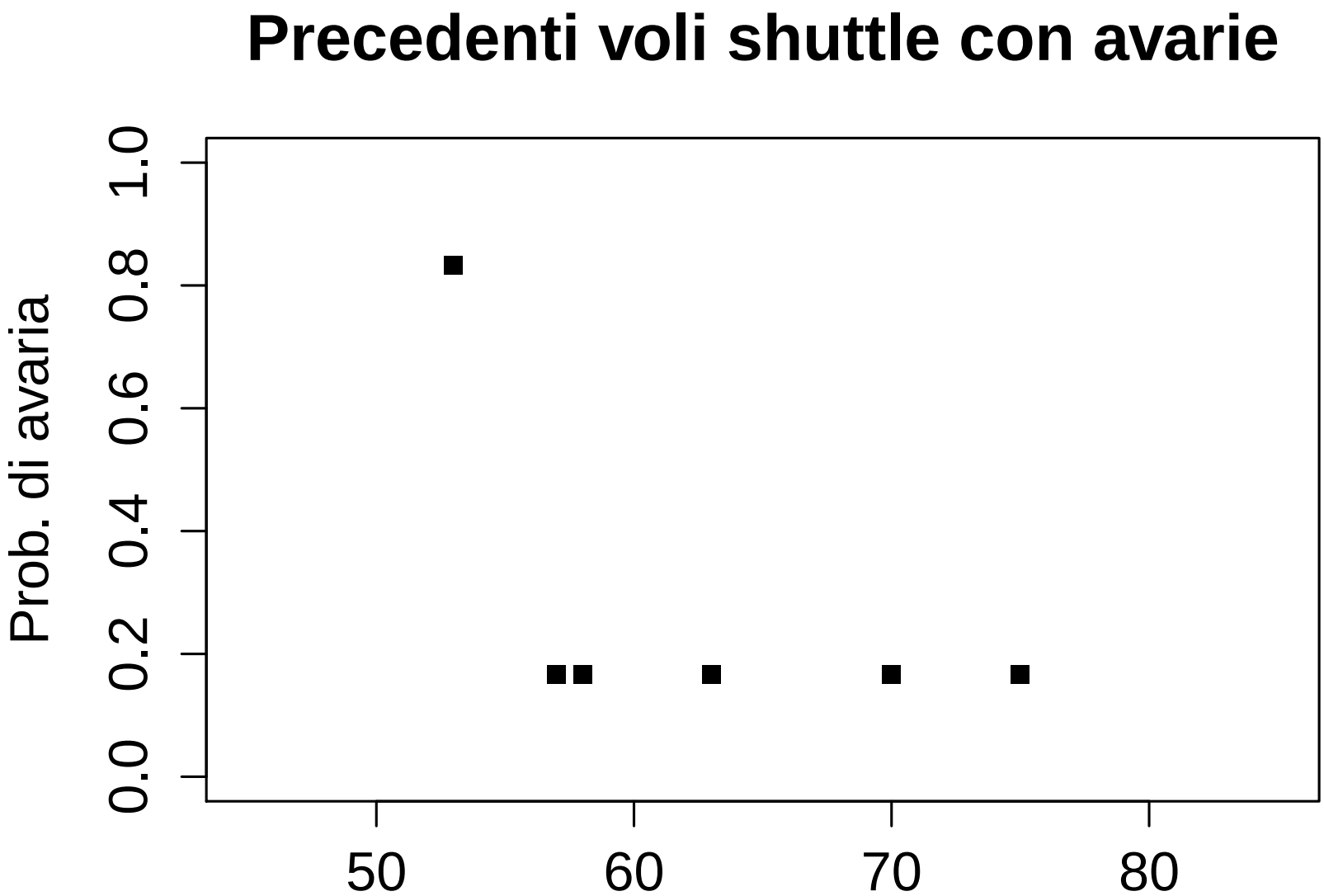
La temperatura prevista per il giorno del lancio, 28 gennaio '86, era 31 gradi F^o
La decisione di effettuare il lancio si basò sul plot delle temperature rispetto al numero di avarie (Y = 1,2,3), eliminando i dati in cui non si erano verificate avarie. Si ebbe così $n = 7$ e

	1	2	3	4	5	6	7
temperature	70	57	63	70	53	75	58
num. avarie	1	1	1	1	5	1	1

In questo plot non si vede nessun trend ovvio e la correlazione tra temperature e numero di avarie e' pari a -0.61
In realtà , considerando tutti i dati (anche i casi con zero avarie il plot che si ottiene è diverso, il trend discendente è più marcato.
C'è anche un dato anomalo in (75,1).
La correlazione, includendo il dato anomalo è pari a -0.64 ed escludendo il dato anomalo è -0.67

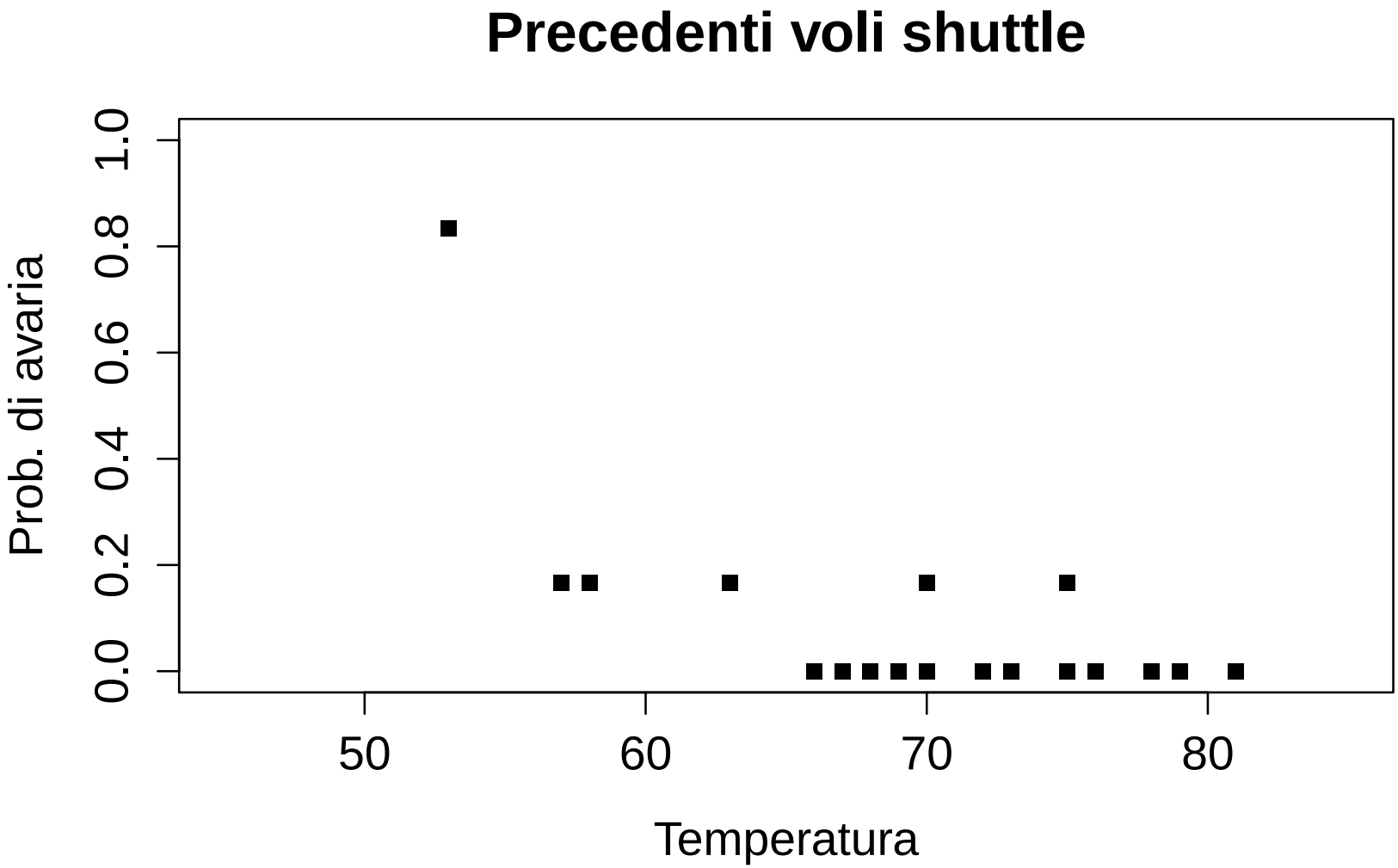
Con **R**

```
library(faraway)
data(orings)
attach(orings)
or<-orings[orings$damage>0,]
plot(or$damage/6 ~ or$temp, orings, xlim=c(45,85),
     ylim = c(0,1), xlab="Temperatura",
     ylab="Prob. di avaria",
     main="Precedenti voli shuttle con avarie",pch=15)
```


Precedenti voli shuttle con avarie
Prob. di avaria
0.0
0.2
0.4
0.6
0.8
1.0
50
60
70
80

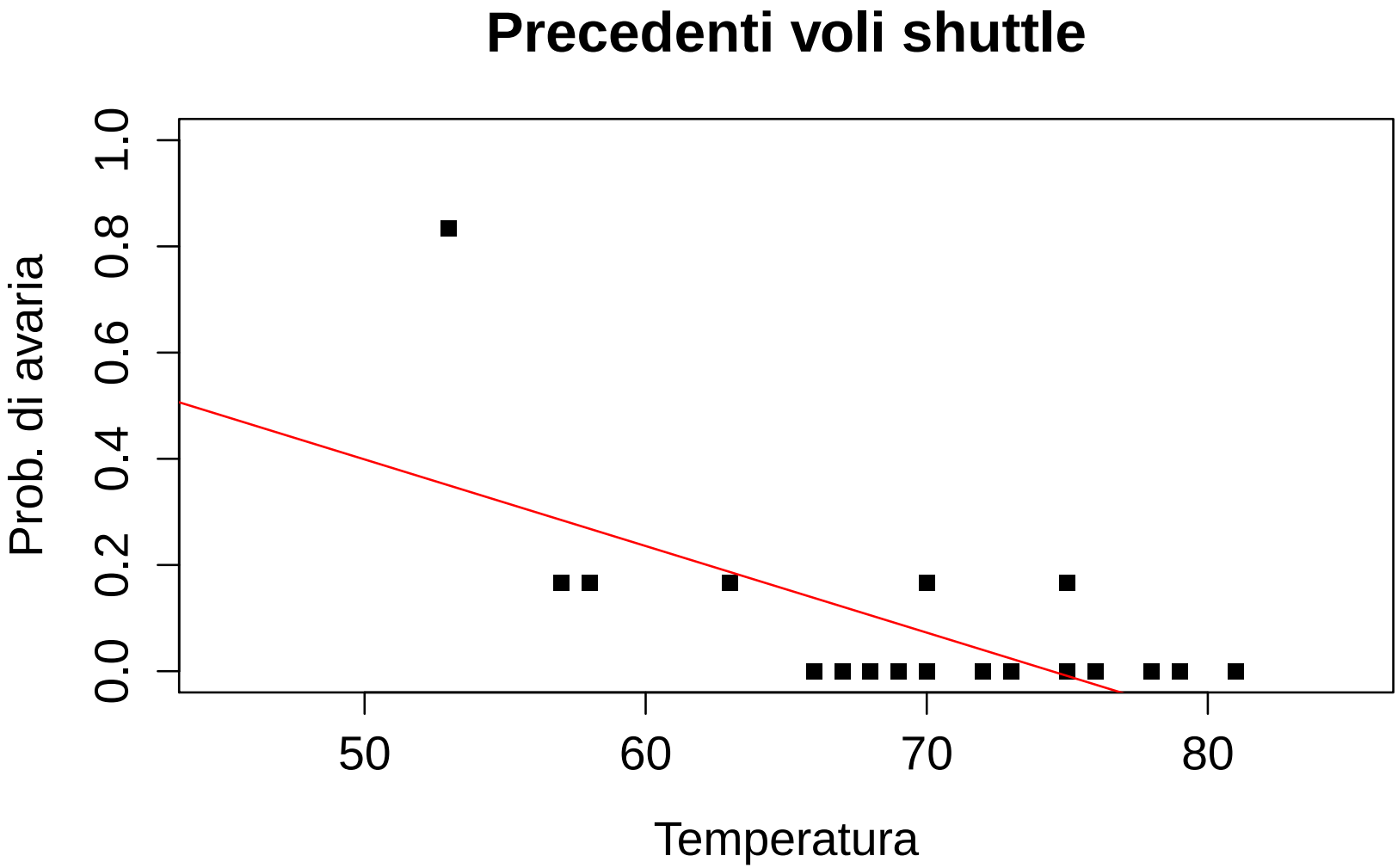
Con **R**

```
plot(damage/6 ~ temp, orings, xlim=c(45,85),
     ylim = c(0,1), xlab="Temperatura",
     ylab="Prob. di avaria",
     main="Precedenti voli shuttle",pch=15)
```


Precedenti voli shuttle
Prob. di avaria
0.0
0.2
0.4
0.6
0.8
1.0
50
60
70
80
Temperatura

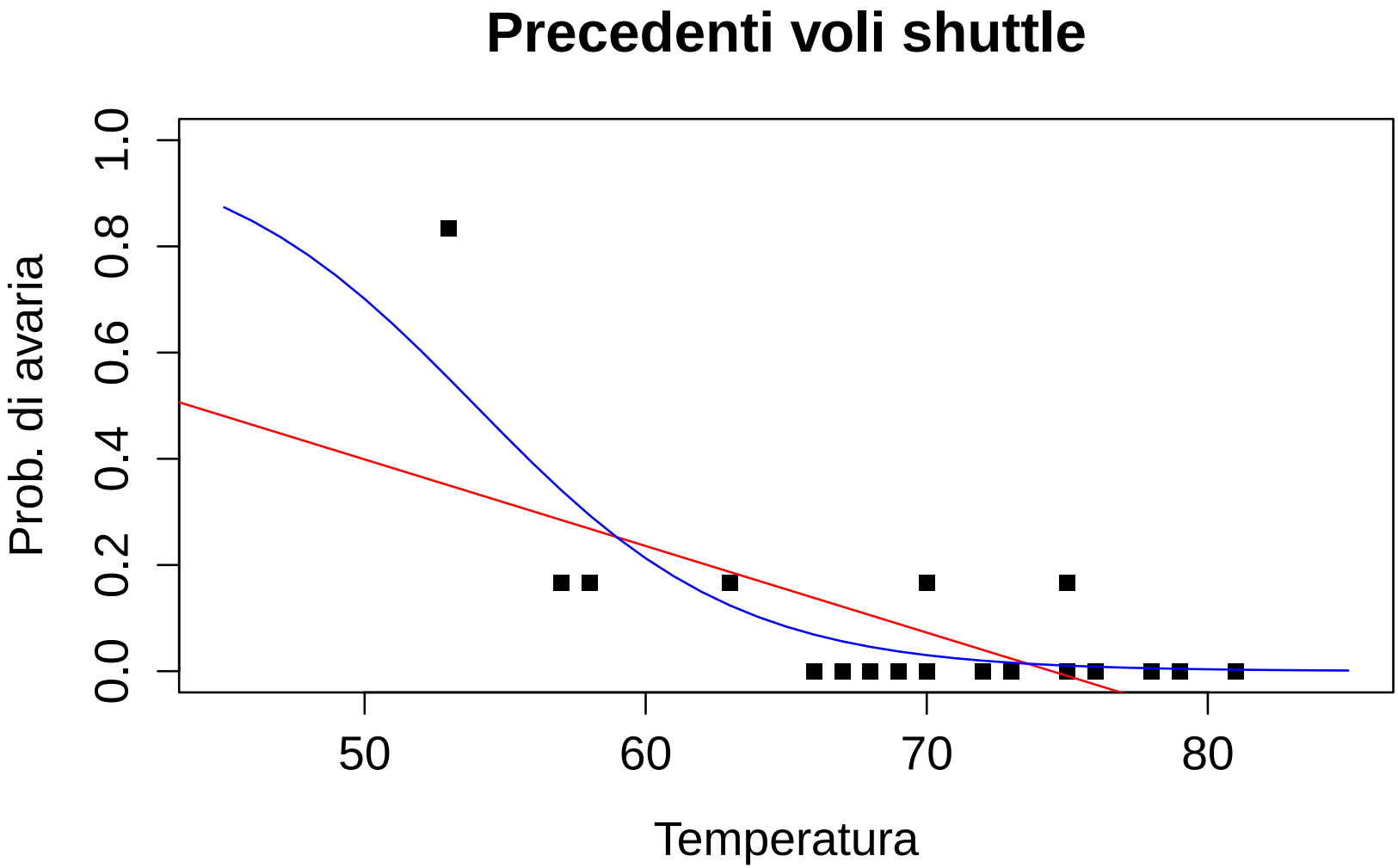
Con **R**

```
m1<-lm(damage/6~temp)
abline(m1$coeff, col="red")
```


Precedenti voli shuttle
Prob. di avaria
0.0
0.2
0.4
0.6
0.8
1.0
50
60
70
80
Temperatura

Con **R**

```
logit1<-glm(cbind(damage, 6-damage)~temp,
            family=binomial(link=logit),
            data= orings)
summary(logit1)
x<-seq(45,85,1)
lines(x,exp(11.663-0.2162 * x)/(1+exp(11.663-0.2162 * x)),
      lty=1, col="blue")
```


Precedenti voli shuttle
Prob. di avaria
0.0
0.2
0.4
0.6
0.8
1.0
50
60
70
80
Temperatura

Convivere con l'incertezza

L'incertezza ha un ruolo in gran parte delle nostre azioni.

- quanto tempo impiegherò per arrivare a scuola in autobus?

Convivere con l'incertezza

L'incertezza ha un ruolo in gran parte delle nostre azioni.

- quanto tempo impiegherò per arrivare a scuola in autobus?
- la ragazza del secondo banco accetterà di uscire con me sabato prossimo?

Convivere con l'incertezza

L'incertezza ha un ruolo in gran parte delle nostre azioni.

- quanto tempo impiegherò per arrivare a scuola in autobus?
- la ragazza del secondo banco accetterà di uscire con me sabato prossimo?
- quanto varranno le mie azioni tra un mese?

Convivere con l'incertezza

L'incertezza ha un ruolo in gran parte delle nostre azioni.

- quanto tempo impiegherò per arrivare a scuola in autobus?
- la ragazza del secondo banco accetterà di uscire con me sabato prossimo?
- quanto varranno le mie azioni tra un mese?
- .....

Il linguaggio con cui cerchiamo di fare chiarezza in questo sistema complesso è la probabilità

La probabilità soggettiva

In altri termini

la probabilità di un evento A è il prezzo che si è disposti a pagare per ricevere 1 se A si verifica e 0 se A non si verifica, e inoltre si è disposti a scommettere, indifferentemente, su A oppure su A^c. (de Finetti, 1931).

Probabilità e Odds

Gli scommettitori e gli anglosassoni preferiscono parlare di odds (quote) piuttosto che di probabilità.

Probabilità e Odds

Gli scommettitori e gli anglosassoni preferiscono parlare di odds (quote) piuttosto che di probabilità.
Se un evento A ha probabilità $P(A)$, le odds in suo favore sono pari a

$$O(A) = \frac{P(A)}{1 - P(A)}; \qquad P(A) = \frac{O(A)}{1 + O(A)}$$

Probabilità e Odds

Gli scommettitori e gli anglosassoni preferiscono parlare di odds (quote) piuttosto che di probabilità.
Se un evento A ha probabilità $P(A)$, le odds in suo favore sono pari a

$$O(A) = \frac{P(A)}{1 - P(A)}; \qquad P(A) = \frac{O(A)}{1 + O(A)}$$

Nel linguaggio delle scommesse si utilizzano le odds contro il verificarsi di A, cioè in favore del verificarsi di A^c.

$$O(A^c) = \frac{1 - P(A)}{P(A)}; \qquad P(A) = \frac{1)}{1 + O(A^c)}$$

Probabilità e Odds: un esempio

Nella prossima partita la squadra del Napoli è data 5 a 2

Probabilità e Odds: un esempio

Nella prossima partita la squadra del Napoli è data 5 a 2
Questo implica che l'evento $A^c = \{\text{NON vince il Napoli}\}$ ha odds $O(A^c) = 5/2$

Probabilità e Odds: un esempio

Nella prossima partita la squadra del Napoli è data 5 a 2
Questo implica che l'evento $A^c = \{\text{NON vince il Napoli}\}$ ha odds $O(A^c) = 5/2$
e dunque

$$P(A) = \frac{1}{1 + O(A^c)} = \frac{1}{1 + 2.5} = 0.285$$

Odds e scommesse

1X2 FINALE, U/O 2,5, GOL NO GOL	1	X	2	UNDER	OVER	GOAL	NOGOAL
PALERMO - GENOA	2,15	3,30	3,40	1,73	1,95	1,75	1,90
24/02 12:30							
SAMPDORIA - CHIEVO	1,95	3,20	4,25	1,65	2,05	1,80	1,85
24/02 15:00							
ATALANTA - ROMA	3,00	3,40	2,30	1,83	1,83	1,70	2,00
BOLOGNA - FIORENTINA	3,15	3,30	2,25	1,78	1,88	1,70	2,00
CAGLIARI - TORINO	2,25	3,10	3,40	1,63	2,10	1,80	1,85
JUVENTUS - SIENA	1,20	6,00	15,00	2,20	1,57	2,10	1,63
PARMA - CATANIA	2,10	3,25	3,60	1,70	2,00	1,73	1,95
24/02 20:45							
INTER - MILAN	2,85	3,40	2,40	1,85	1,80	1,68	2,02
25/02 19:00							
UDINESE - NAPOLI	3,00	3,30	2,35	1,73	1,95	1,73	1,95
25/02 21:00							
LAZIO - PESCARA	1,25	5,50	12,00	2,15	1,60	2,00	1,70

cosa vuol dire

partita	1	X	2
Atalanta Roma	3	3.40	2.30
togli 1 euro	2	2.40	1.30

cosa vuol dire

partita	1	X	2
Atalanta Roma	3	3.40	2.30
togli 1 euro	2	2.40	1.30

Nell'ultima riga ci sono le odds "contro" i tre eventi da cui

cosa vuol dire

partita	1	X	2
Atalanta Roma	3	3.40	2.30
togli 1 euro	2	2.40	1.30

Nell'ultima riga ci sono le odds "contro" i tre eventi da cui

$$P(1) = \frac{1}{1+2} = 0.333; \qquad P(X) = \frac{1}{1+2.4} = 0.294; \qquad P(2) = \frac{1}{1+1.3} = 0.438;$$

cosa vuol dire

partita	1	X	2
Atalanta Roma	3	3.40	2.30
togli 1 euro	2	2.40	1.30

Nell'ultima riga ci sono le odds "contro" i tre eventi da cui

$$P(1) = \frac{1}{1+2} = 0.333; \qquad P(X) = \frac{1}{1+2.4} = 0.294; \qquad P(2) = \frac{1}{1+1.3} = 0.438;$$

Nota che

$$P(1) + P(X) + P(2) = 1.065 > 1!!$$

Perché?

Decisioni in condizioni di incertezza

Breve elenco di esempi sui ruoli e contesti in cui si è chiamati a prendere decisioni.

- Il Manager: controllo di un processo produttivo, strategie di marketing, segmentazione del mercato, sconti e offerte.

Decisioni in condizioni di incertezza

Breve elenco di esempi sui ruoli e contesti in cui si è chiamati a prendere decisioni.

- Il Manager: controllo di un processo produttivo, strategie di marketing, segmentazione del mercato, sconti e offerte.
- Il Politico e l'amministratore pubblico: interventi di politica economica, costruzione di strutture pubbliche, modifica dei tassi di interesse.

Decisioni in condizioni di incertezza

Breve elenco di esempi sui ruoli e contesti in cui si è chiamati a prendere decisioni.

- Il Manager: controllo di un processo produttivo, strategie di marketing, segmentazione del mercato, sconti e offerte.
- Il Politico e l'amministratore pubblico: interventi di politica economica, costruzione di strutture pubbliche, modifica dei tassi di interesse.
- Il Consumatore: quale prodotto scelgo.

Decisioni in condizioni di incertezza

Breve elenco di esempi sui ruoli e contesti in cui si è chiamati a prendere decisioni.

- Il Manager: controllo di un processo produttivo, strategie di marketing, segmentazione del mercato, sconti e offerte.
- Il Politico e l'amministratore pubblico: interventi di politica economica, costruzione di strutture pubbliche, modifica dei tassi di interesse.
- Il Consumatore: quale prodotto scelgo.
- Lo scienziato sperimentale (Fisico, Biologo, Medico). Quante osservazioni mi occorrono, in quali condizioni sperimentali.

Decisioni in condizioni di incertezza

Breve elenco di esempi sui ruoli e contesti in cui si è chiamati a prendere decisioni.

- Il Manager: controllo di un processo produttivo, strategie di marketing, segmentazione del mercato, sconti e offerte.
- Il Politico e l'amministratore pubblico: interventi di politica economica, costruzione di strutture pubbliche, modifica dei tassi di interesse.
- Il Consumatore: quale prodotto scelgo.
- Lo scienziato sperimentale (Fisico, Biologo, Medico). Quante osservazioni mi occorrono, in quali condizioni sperimentali.
- L'ingegnere: costruzione una diga (previsione di eventi estremi)

Tre contesti tipici di decisioni statistiche in condizioni di incertezza

- Problemi di Classificazione
  - Questo regalo piacerà a mia moglie/figlia/sorella o compagna?
  - Il cliente risulterà insolvente?
  - La terapia sarà efficace?

Tre contesti tipici di decisioni statistiche in condizioni di incertezza

- Problemi di Classificazione
  - Questo regalo piacerà a mia moglie/figlia/sorella o compagna?
  - Il cliente risulterà insolvente?
  - La terapia sarà efficace?
- Previsione
  - Quali conseguenze sulla temperatura nei prossimi dieci anni per l'effetto serra?
  - Quanti immigrati arriveranno nei prossimi 5 anni?
  - Quanto varrà un'azione FIAT fra un anno?
  - Quante ambulanze devo dislocare nel territorio comunale?
  - Quanto costa una riforma fiscale?

Tre contesti tipici di decisioni statistiche in condizioni di incertezza

- Problemi di Classificazione
  - Questo regalo piacerà a mia moglie/figlia/sorella o compagna?
  - Il cliente risulterà insolvente?
  - La terapia sarà efficace?
- Previsione
  - Quali conseguenze sulla temperatura nei prossimi dieci anni per l'effetto serra?
  - Quanti immigrati arriveranno nei prossimi 5 anni?
  - Quanto varrà un'azione FIAT fra un anno?
  - Quante ambulanze devo dislocare nel territorio comunale?
  - Quanto costa una riforma fiscale?
- Quali informazioni sono rilevanti/utili e quali no
  - Ingredienti di un farmaco
  - Domande da inserire in un test psico-attitudinale

Tre contesti tipici di decisioni statistiche in condizioni di incertezza

- Problemi di Classificazione
  - Questo regalo piacerà a mia moglie/figlia/sorella o compagna?
  - Il cliente risulterà insolvente?
  - La terapia sarà efficace?
- Previsione
  - Quali conseguenze sulla temperatura nei prossimi dieci anni per l'effetto serra?
  - Quanti immigrati arriveranno nei prossimi 5 anni?
  - Quanto varrà un'azione FIAT fra un anno?
  - Quante ambulanze devo dislocare nel territorio comunale?
  - Quanto costa una riforma fiscale?
- Quali informazioni sono rilevanti/utili e quali no
  - Ingredienti di un farmaco
  - Domande da inserire in un test psico-attitudinale

La regola di Bayes

La regola di Bayes consente di formalizzare il comportamento razionale dell'individuo di fronte all'incertezza.

Rev T. Bayes (1701-1761)

Regola di Bayes

Supponiamo che

- un evento E possa essere stato 'causato' da uno e uno solo degli eventi $(A_1, A_2, \ldots, A_k)$.
- gli eventi $(A_1, A_2, \ldots, A_k)$ sono mutuamente incompatibili.

Regola di Bayes

Supponiamo che

- un evento E possa essere stato 'causato' da uno e uno solo degli eventi $(A_1, A_2, \ldots, A_k)$.
- gli eventi $(A_1, A_2, \ldots, A_k)$ sono mutuamente incompatibili.

Allora,

$$P(A_1 \mid E) = \frac{P(A_1)P(E \mid A_1)}{P(A_1)P(E \mid A_1) + P(A_2)P(E \mid A_2) + \ldots P(A_k)P(E \mid A_k)}$$

Regola di Bayes

Supponiamo che

- un evento E possa essere stato 'causato' da uno e uno solo degli eventi $(A_1, A_2, \ldots, A_k)$.
- gli eventi $(A_1, A_2, \ldots, A_k)$ sono mutuamente incompatibili.

Allora,

$$P(A_1 \mid E) = \frac{P(A_1)P(E \mid A_1)}{P(A_1)P(E \mid A_1) + P(A_2)P(E \mid A_2) + \ldots P(A_k)P(E \mid A_k)}$$

Dalle probabilità a priori, $P(A_1)$, $P(A_2)$, $\ldots$,$P(A_k)$, di k eventi disgiunti A_1, A_2, $\ldots$,A_k, e grazie al verificarsi dell'evento E, è possibile calcolare le probabilità condizionate (dette 'a posteriori' rispetto al verificarsi di E) $P(A_1 \mid E)$, $P(A_2 \mid E)$, $\ldots$,$P(A_k \mid E)$.

Mille applicazioni

- Classificazione
- scelta di portafoglio
- Sistemi esperti - tecniche anti spam

L'intera disciplina statistica, e forse qualcosa di più, può essere riletta come una grande applicazione del metodo bayesiano.

Un esempio per tutti

Quando si ha il sospetto di soffrire di una certa patologia ci si può spesso sottoporre ad un test.
Nessun test è infallibile poichè può produrre i cosiddetti

- Falsi positivi (il test risulta positivo ma l'individuo è sano)

Un esempio per tutti

Quando si ha il sospetto di soffrire di una certa patologia ci si può spesso sottoporre ad un test.
Nessun test è infallibile poichè può produrre i cosiddetti

- Falsi positivi (il test risulta positivo ma l'individuo è sano)
- Falsi negativi (il test risulta negativo ma l'individuo è malato)

Spieghiamoci con un esempio.

Test AIDS

- Ci si sottopone ad un Test per verificare la sieropositività (AIDS). Il test ha
  - una sensibilità del 98% ovvero produce un 2% di falsi positivi

Test AIDS

- Ci si sottopone ad un Test per verificare la sieropositività (AIDS). Il test ha
  - una sensibilità del 98% ovvero produce un 2% di falsi positivi
  - una specificità del 95% ovvero produce un 5% di falsi negativi

Test AIDS

- Ci si sottopone ad un Test per verificare la sieropositività (AIDS). Il test ha
  - una sensibilità del 98% ovvero produce un 2% di falsi positivi
  - una specificità del 95% ovvero produce un 5% di falsi negativi
- Nella popolazione di riferimento la percentuale di malati di AIDS è un decimo dell'1% (ovvero 0.001, ovvero l'1 per mille)

Test AIDS

- Ci si sottopone ad un Test per verificare la sieropositività (AIDS). Il test ha
  - una sensibilità del 98% ovvero produce un 2% di falsi positivi
  - una specificità del 95% ovvero produce un 5% di falsi negativi
- Nella popolazione di riferimento la percentuale di malati di AIDS è un decimo dell'1% (ovvero 0.001, ovvero l'1 per mille)
- Il medico si trova di fronte ad un paziente, un elemento della popolazione, che è risultato positivo al test e deve fornire la sua diagnosi:
  il paziente è sieropositivo oppure no?

Formalizziamo...

Il medico deve calcolare la probabilità a posteriori:

$$P(\text{malato} \mid \text{ test positivo}) = P(M \mid T^+)$$

Tizio si sottopone ad un test per una certa malattia **M**.
Il test ha le seguenti caratteristiche

	M	NM
T^+	0.95	0.02
T^-	0.05	0.98
Tot.	1	1

Sappiamo poi che il test, applicato a Tizio, risulta positivo (T^+).

Formalizziamo

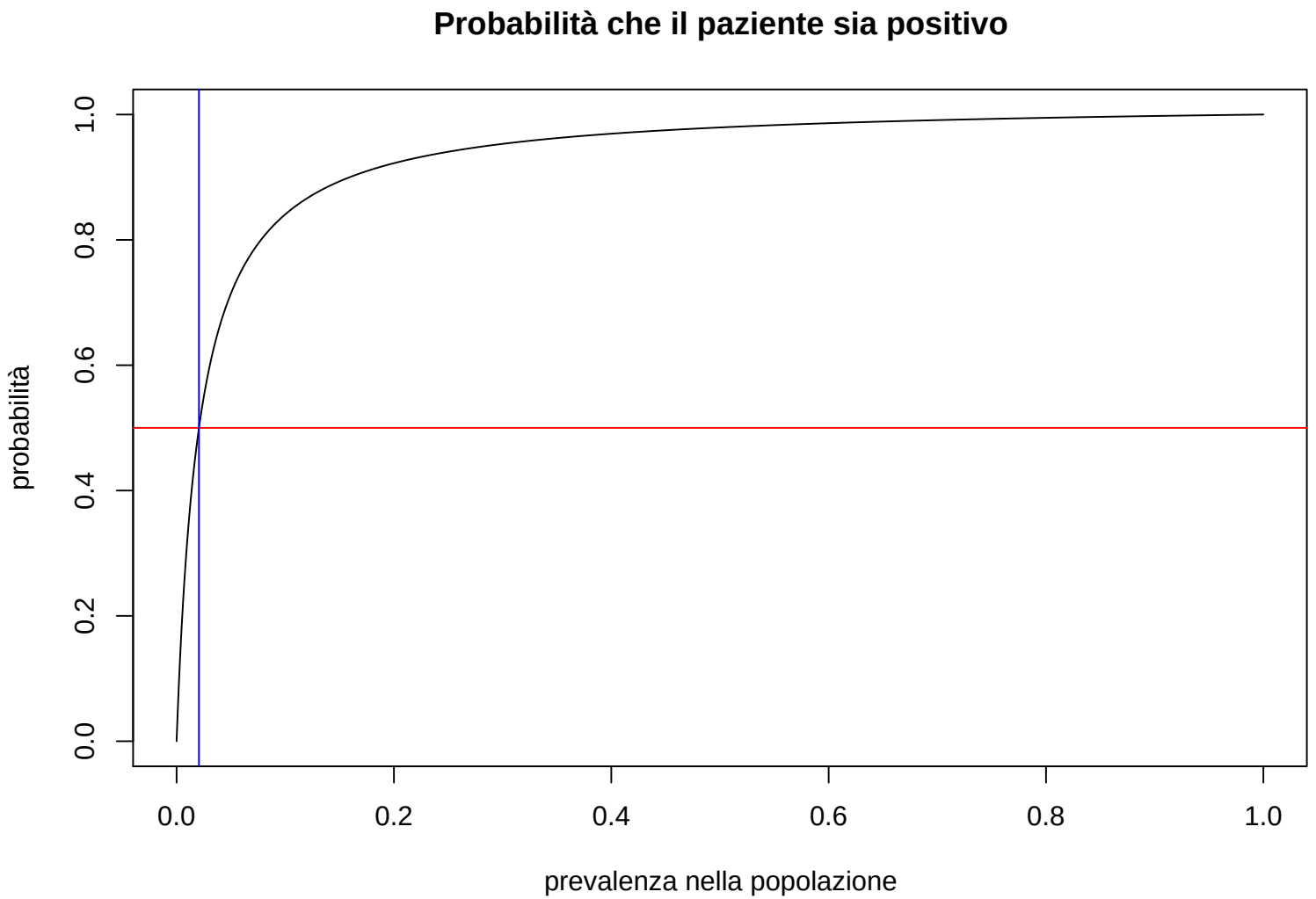
Occorre calcolare la probabilità condizionata che Tizio sia affetto da M ovvero $\mathrm{P}(M \mid T^+)$.
Detta $\pi = P(M)$, si ha

$$\begin{aligned}
\pi^* &= P(M \mid T^+) = \frac{P(M)P(T^+ \mid M)}{P(T^+)} \\
&= \frac{P(M)P(T^+ \mid M)}{P(M)P(T^+ \mid M) + P(NM)P(T^+ \mid NM)} \\
&= \frac{0.95 \times \pi}{0.95 \times \pi + 0.02 \times (1-\pi)} = \frac{95\pi}{93\pi + 2}
\end{aligned}$$

Dunque la risposta al quesito B dipende dalla probabilità a priori π, che in questo caso corrisponde a conoscere la prevalenza della malattia in quel dato contesto geografico.
Ad esempio

$P(M)$	0.001	0.01	0.1	0.2
$P(M \mid T^+)$	0.045	0.324	0.841	0.922

Graficamente



A volte è sufficiente stabilire delle semplici disuguaglianze.
Per quale livello di prevalenza $\pi = P(M)$ la probabilità finale π^*risulterà maggiore di una certa soglia?

Ad esempio, per avere $\pi^* > 0.5$ occorre

$$\frac{95\pi}{93\pi + 2} > \frac{1}{2} \Leftrightarrow 190\pi > 93\pi + 2 \Leftrightarrow \pi > 0.0206$$