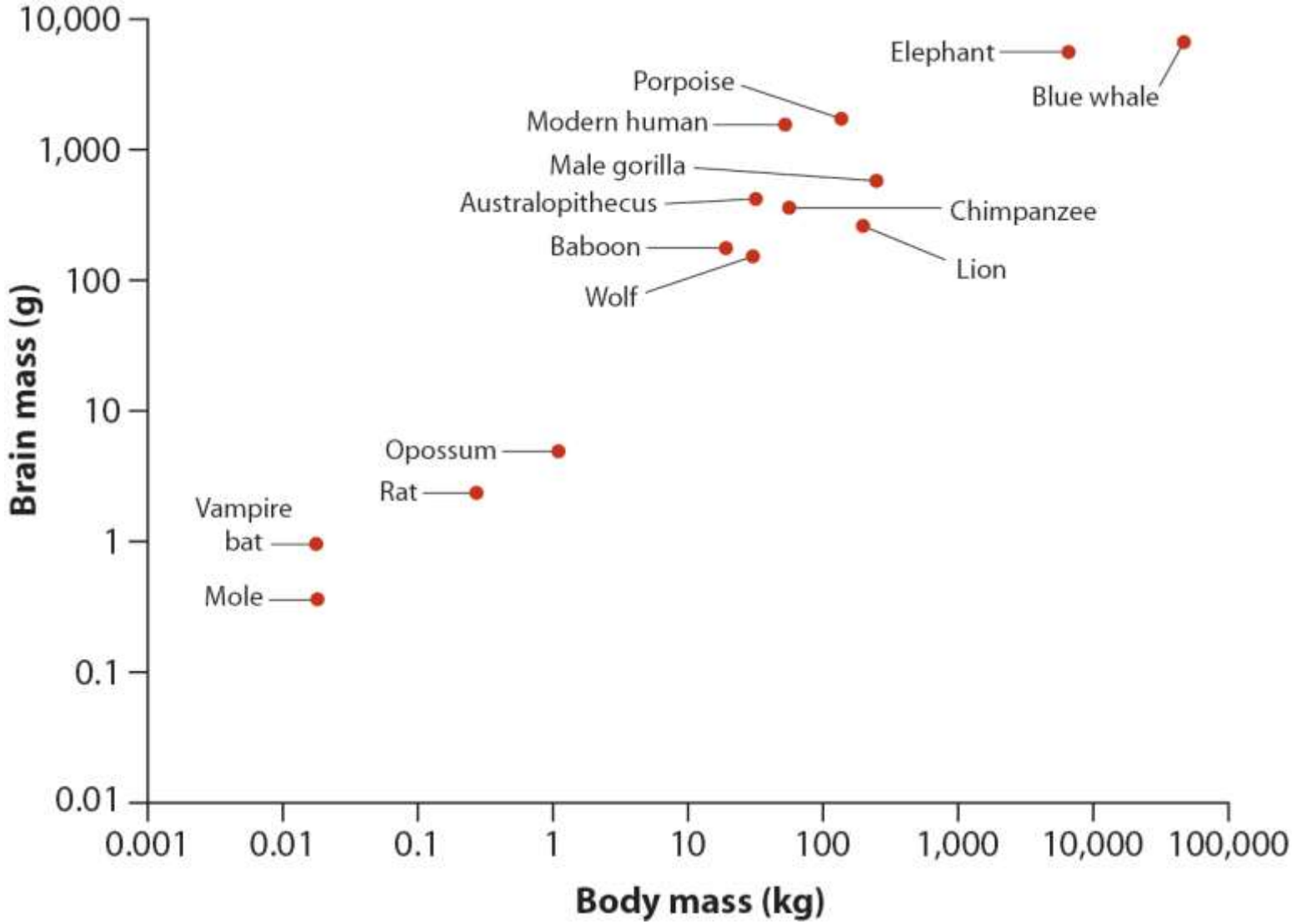
# Correlazione tra variabili numeriche

Capitolo 16

Analisi statistica dei dati biologici

# Come si può misurare l'associazione tra variabili?



# Il coefficiente di correlazione

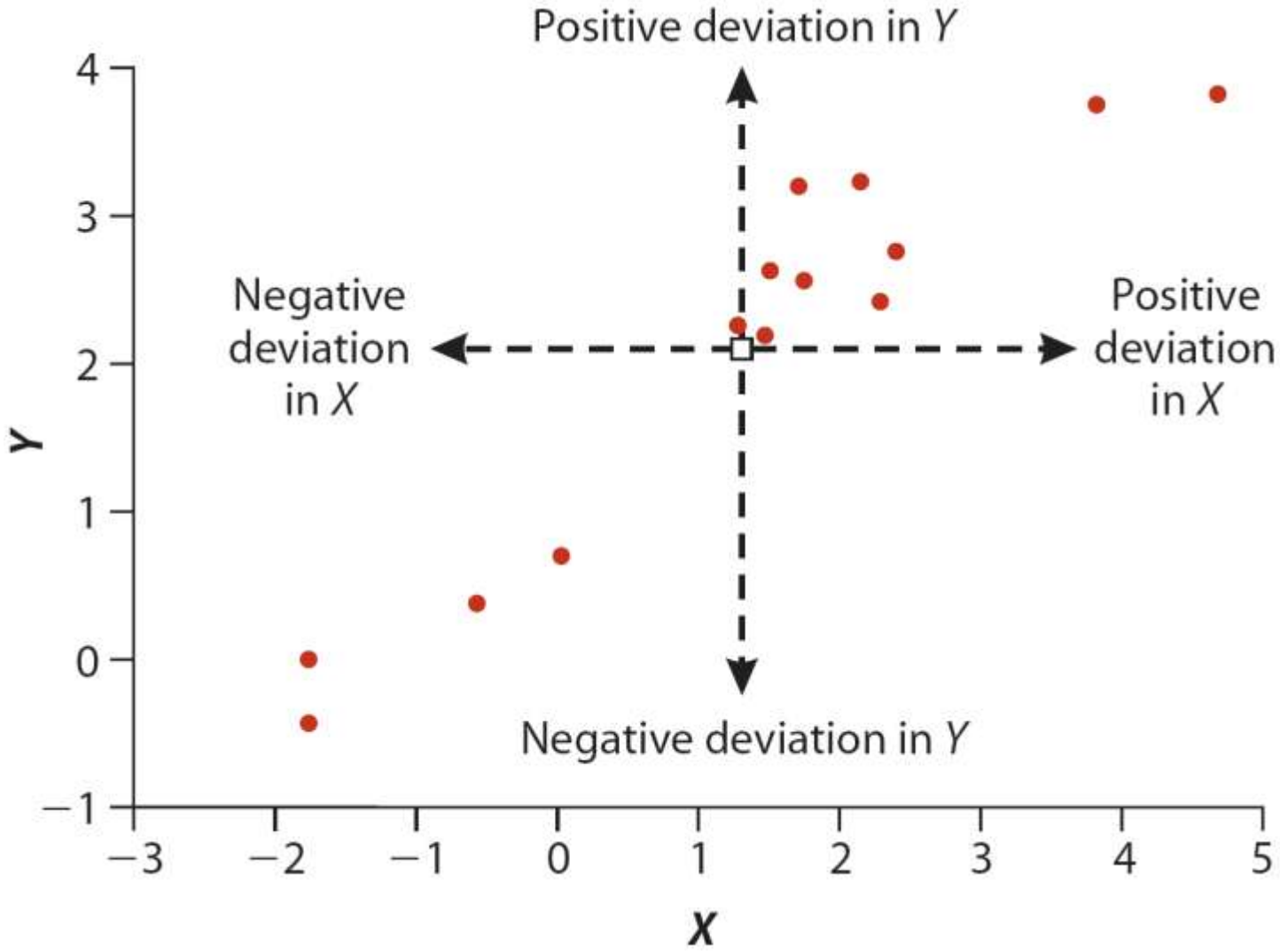
$$r = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum(X - \bar{X})^2}\sqrt{\sum(Y - \bar{Y})^2}}$$

Il coefficiente di correlazione misura l'intensità e la direzione dell'associazione lineare tra due variabili numeriche.

Si indica con «ρ» la correlazione nella popolazione

Si indica con «r» la correlazione nel campione

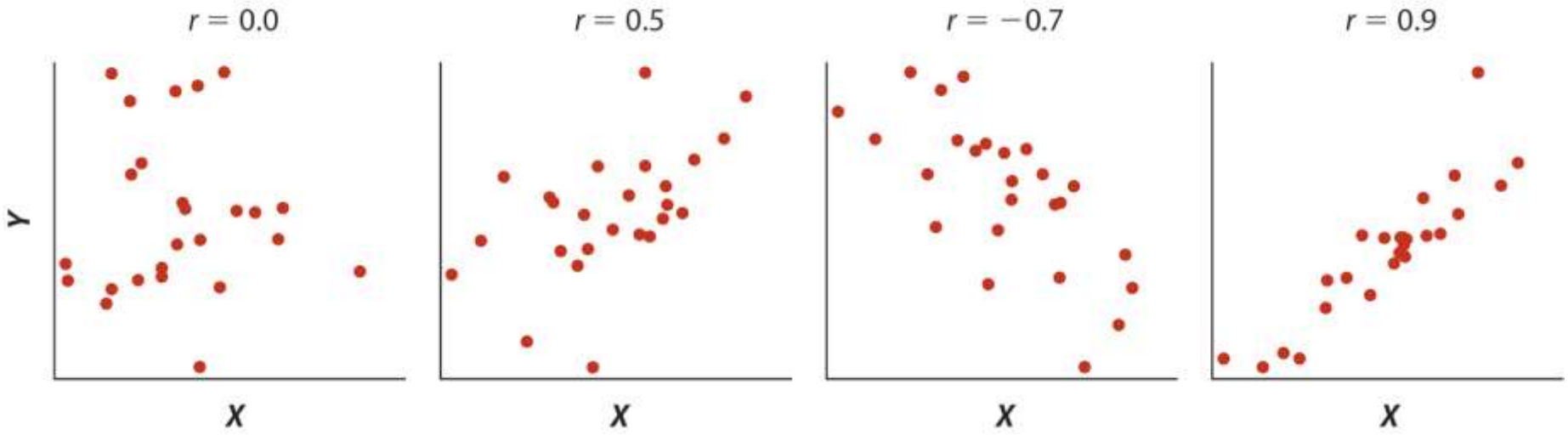
# Posizione delle osservazioni rispetto alle medie di X e Y



# Caratteristiche di r

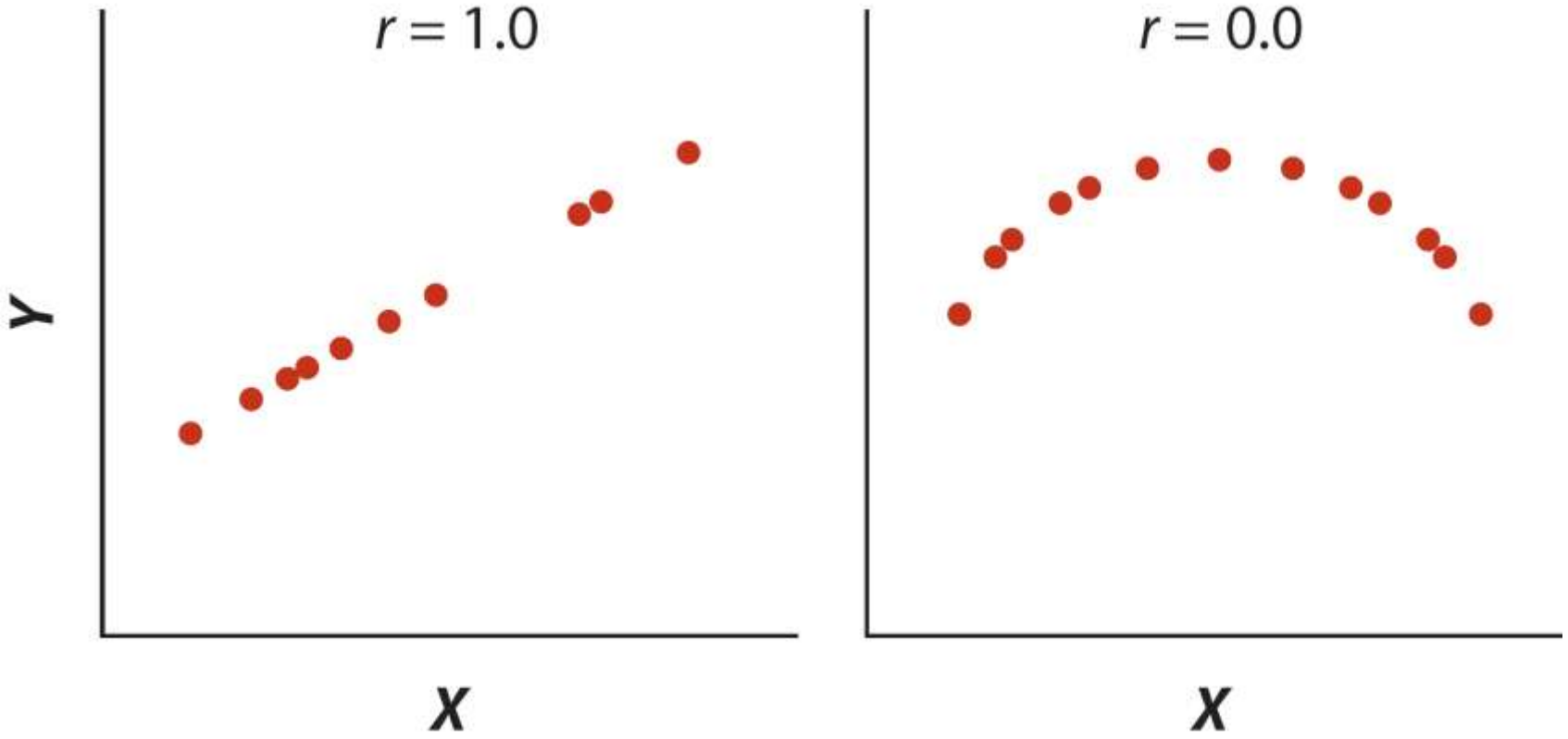
- Indice adimensionale (privo di unità di misura)
- Varia da -1 a 1
- Positivo se la maggior parte delle osservazioni cade nel quadrante superiore destro e inferiore sinistro
- Negativo se la maggior parte delle osservazioni cade nel quadrante superiore sinistro e inferiore destro
- Tenderà a 0 se le osservazioni sono distribuite in modo uniforme nei 4 quadranti

# Direzione




- Correlazione negativa (r<0) significa che quando una variabile cresce l'altra decresce
- Correlazione positiva (r>0) significa che le due variabili crescono (o decrescono) insieme
- Correlazione massima (r=1 o -1) se tutte le osservazioni giacciono su una retta

# Attenzione, l'associazione misurata è lineare

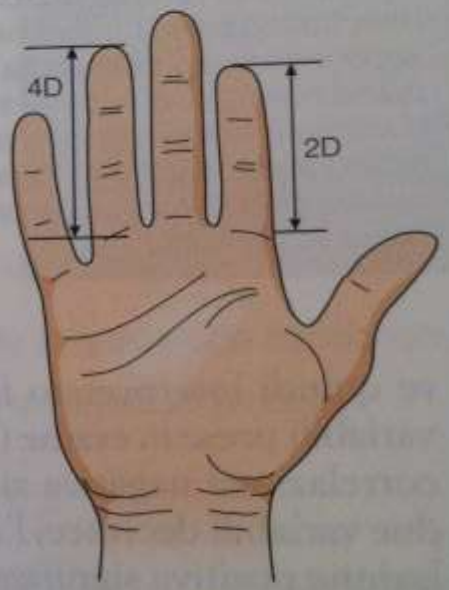


# Esempio 1

Esempio 16.1

## *Dita maschili*

Il rapporto tra la lunghezza del secondo e del quarto dito della mano (il rapporto 2D:4D) è minore negli uomini che nelle donne, e questo effetto è dovuto alla maggiore esposizione dei maschi al testosterone fin da quando sono nell'utero. (Poiché il testosterone influisce anche sullo sviluppo sessuale, è stato ipotizzato che il rapporto 2D:4D potrebbe permettere di prevedere alcuni aspetti del comportamento sessuale degli individui da adulti.) Manning et al. (2003) hanno indagato una possibile connessione tra il rapporto 2D:4D nei soggetti maschi e la struttura del loro gene per il recettore degli androgeni. Il numero di ripetizioni CAG[6] in una regione specifica del gene è correlato con la sensibilità al testosterone. I ricercatori si sono chiesti se il numero di ripetizioni CAG fosse correlato anche con il rapporto 2D:4D. I risultati delle misurazioni eseguite su 46 soggetti sono presentati nella Tabella 16.1-1. ■



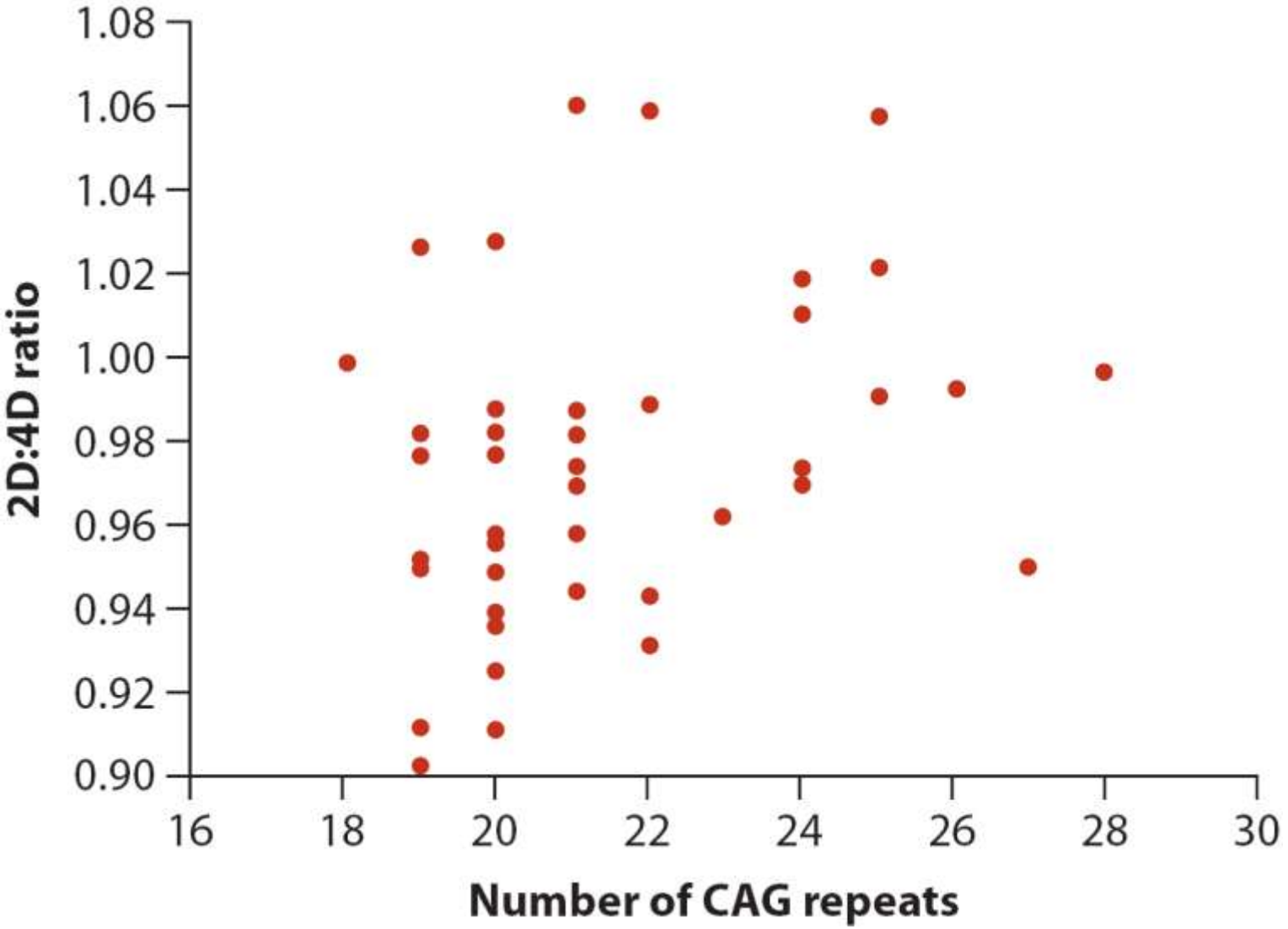
# Dati

| CAGrepeats | finger.ratio |
|---|---|
| 21 | 1.06 |
| 22 | 1.059 |
| 25 | 1.058 |
| 19 | 1.026 |
| 20 | 1.027 |
| 18 | 0.998 |
| 19 | 0.982 |
| 19 | 0.977 |
| 19 | 0.977 |
| 20 | 0.988 |
| 20 | 0.983 |
| 20 | 0.982 |
| 20 | 0.978 |
| 20 | 0.977 |
| 21 | 0.987 |

| CAGrepeats | finger.ratio |
|---|---|
| 21 | 0.987 |
| 21 | 0.982 |
| 21 | 0.974 |
| 21 | 0.973 |
| 21 | 0.97 |
| 24 | 1.019 |
| 24 | 1.01 |
| 25 | 1.022 |
| 22 | 0.988 |
| 25 | 0.99 |
| 26 | 0.992 |
| 28 | 0.996 |
| 19 | 0.952 |
| 19 | 0.95 |
| 20 | 0.958 |

| CAGrepeats | finger.ratio |
|---|---|
| 20 | 0.956 |
| 20 | 0.95 |
| 21 | 0.958 |
| 23 | 0.961 |
| 24 | 0.973 |
| 24 | 0.97 |
| 27 | 0.95 |
| 20 | 0.939 |
| 20 | 0.936 |
| 21 | 0.944 |
| 22 | 0.943 |
| 22 | 0.931 |
| 20 | 0.925 |
| 20 | 0.912 |
| 19 | 0.912 |
| 19 | 0.903 |

# Diagramma di dispersione



# Calcolo di r

$\sum(X - \bar{X})\ (Y - \bar{Y})$= 1,186

$\sum(X - \bar{X})^2$= 250,435

$\sum(Y - \bar{Y})^2$= 0,0633

$$r = \frac{1{,}186}{\sqrt{250{,}435}\sqrt{0{,}0633}} = 0{,}298$$

Il coefficiente di correlazione campionario è pari a 0,298.

# Errore standard di r

L'errore standard di r ($ES_r$) è un modo per valutare quanto sia vicino alla stima il corrispondente parametro ρ nella popolazione.

$$ES_r = \sqrt{\frac{1 - r^2}{n - 2}}$$

Per i dati nell'esempio:

$$ES_r = \sqrt{\frac{1 - 0{,}298^2}{46 - 2}} = 0{,}144$$

# Esempio 1 con R

- Calcolare r e $ES_r$ per l'esempio 1 con R, caricando i dati dal file:

«Correlazione_esercizio_Finger_ratio_CAG.txt»

# Esempio 1 con R: calcolo di r

```
#caricare il file con i dati: «Correlazione_esercizio_Finger_ratio_CAG.txt»
 x<-read.table(choose.files(),header=T)
#calcolare media per le due variabili
Xmed<-mean(x$CAGrepeats)
Ymed<-mean(x$finger.ratio)
#calcolare la somma dei prodotti (numeratore)
sommaprod<-sum((x$CAGrepeats-Xmed)*(x$finger.ratio-Ymed))
#calcolare le due parti del denominatore:
denom1<- sqrt(sum((x$CAGrepeats-Xmed)^2))
denom2<-sqrt(sum((x$finger.ratio-Ymed)^2))
#calcolo r
r<-sommaprod/(denom1*denom2)
r1<-cor(x$CAGrepeats,x$finger.ratio,method="pearson")
```

# Esempio 1 con R: calcolo di $ES_r$

```
#calcolo di ESr:
ES<-sqrt((1-r^2)/(length(x$CAGrepeats)-2))
```

# Intervallo di confidenza approssimato

- Per calcolare IC di ρ dobbiamo trasformare r in una nuova grandezza:

$$z = 0{,}5ln\left(\frac{1+r}{1-r}\right)$$

z segue approssimativamente una distribuzione normale e possiamo usarla per definire un IC secondo la formula classica:

$$z-1{,}96\sigma_z<\zeta<z+1{,}96\sigma_z$$

# IC approssimato

- L’errore standard APPROSSIMATO della distribuzione campionaria di z è:

$$\sigma_z = \sqrt{\frac{1}{n-3}}$$

# IC per l'esempio 1

$$z = 0{,}5ln\left(\frac{1 + 0{,}298}{1 - 0{,}298}\right) = 0{,}307$$

e,

$$\sigma_z = \sqrt{\frac{1}{46 - 3}} = 0{,}1525$$

# IC esempio 1

Sostituendo quanto calcolato alla formula di IC:

0,307-1,96(0,1525)<ζ<0,307+1,96(0,1525)

da cui si ottiene: 0,0081<ζ<0,6059

Questo intervallo NON è nella scala di correlazione originale, per cui bisogna ritrasformare i limiti:

$$r = \frac{e^{2z} - 1}{e^{2z} + 1}$$

# IC esempio 1

Sostituendo i dati dell'esercizio 1 otteniamo:

$$\frac{e^{2(0,0081)}-1}{e^{2(0,0081)}+1} < \rho < \frac{e^{2(0,6059)}-1}{e^{2(0,6059)}+1}$$

da cui:

$$0,0081 < \rho < 0,5412$$

Possiamo concludere che ρ nella popolazione sia maggiore di 0, ma molto minore di 1

# Esercizio nr 3 pag 274 (Analisi statistica dei dati biologici)

In molte specie di uccelli, gli individui mantengono lo stesso partner anno dopo anno. In alcune di queste specie, i due membri di una coppia migrano separatamente e trascorrono l'inverno in luoghi differenti, spesso distanti migliaia di chilometri.

Come fanno a ritrovarsi ogni anno in primavera?

In uno studio condotto sul campo su coppie di pittima reale, Gunnarsson et al 2004 hanno registrato le date di arrivo in primavera dei maschi e delle femmine nel territorio in cui l'anno precedente era stato osservato il loro accoppiamento. I dati per 10 coppie sono riportati nella tabella che accompagna questo problema. La data di arrivo è il numero di giorni trascorsi dopo il 31 marzo.

# Dati

| Data di arrivo del partner femminile | Data di arrivo del partner maschile |
|---|---|
| 24 | 22 |
| 36 | 35 |
| 35 | 35 |
| 35 | 44 |
| 38 | 46 |
| 50 | 50 |
| 55 | 55 |
| 56 | 56 |
| 57 | 56 |
| 69 | 59 |

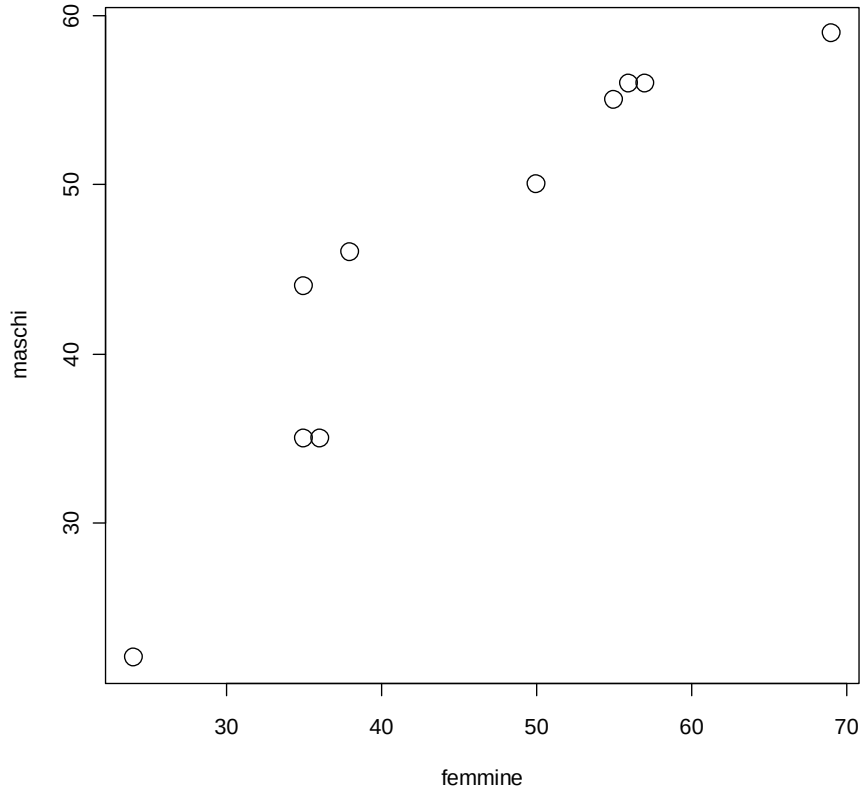
# Domande

a. Visualizzate in un diagramma la relazione fra le date di arrivo dei maschi e delle femmine.

```
#Importare il file «Correlazione_esercizio_pittima_reale.txt»
x<-read.table(choose.files(),header=T)
matplot(x$femdate,x$maldate,type="p",pch=1,xlab="femmine",ylab="maschi")
```

# Domande

b. Descrivete brevemente il pattern identificato nella parte (a). Esiste una relazione? E' positiva o negativa? E' lineare o non lineare? E' debole o forte?



# Domande

c. Calcolate il coefficiente di correlazione tra le date di arrivo dei maschi e delle femmine. Includete l'errore standard nella vostra stima.

```
#calcolo il coefficiente di correlazione
r<-cor(x$femdate,x$maldate,method="pearson")
#calcolo l'errore standard di r
ESr<-sqrt((1-r^2)/(length(x$femdate)-2))
```

# Domande

e. Calcolate un intervallo di confidenza al 95% approssimato di ρ.

```
#trasformo r in z:
z<-0.5*log((1+r)/(1-r))
#calcolo ESz:
ESz<-sqrt(1/(length(x$femdate)-3))
#calcolo IC95% di z
liz<- z-1.96*ESz
lsz<- z+1.96*ESz
#trasformo IC95% di z in IC95% di r
lir<-(exp(2*liz)-1)/(exp(2*liz)+1)
lsr<-(exp(2*lsz)-1)/(exp(2*lsz)+1)
```

# Verifica dell'ipotesi di assenza di correlazione

Nell'analisi di correlazione, la verifica di ipotesi è usata comunemente per verificare l'ipotesi nulla di ASSENZA di correlazione:

$H_0$: $\rho=0$

$H_A$: $\rho\neq0$

# Esempio test d'ipotesi

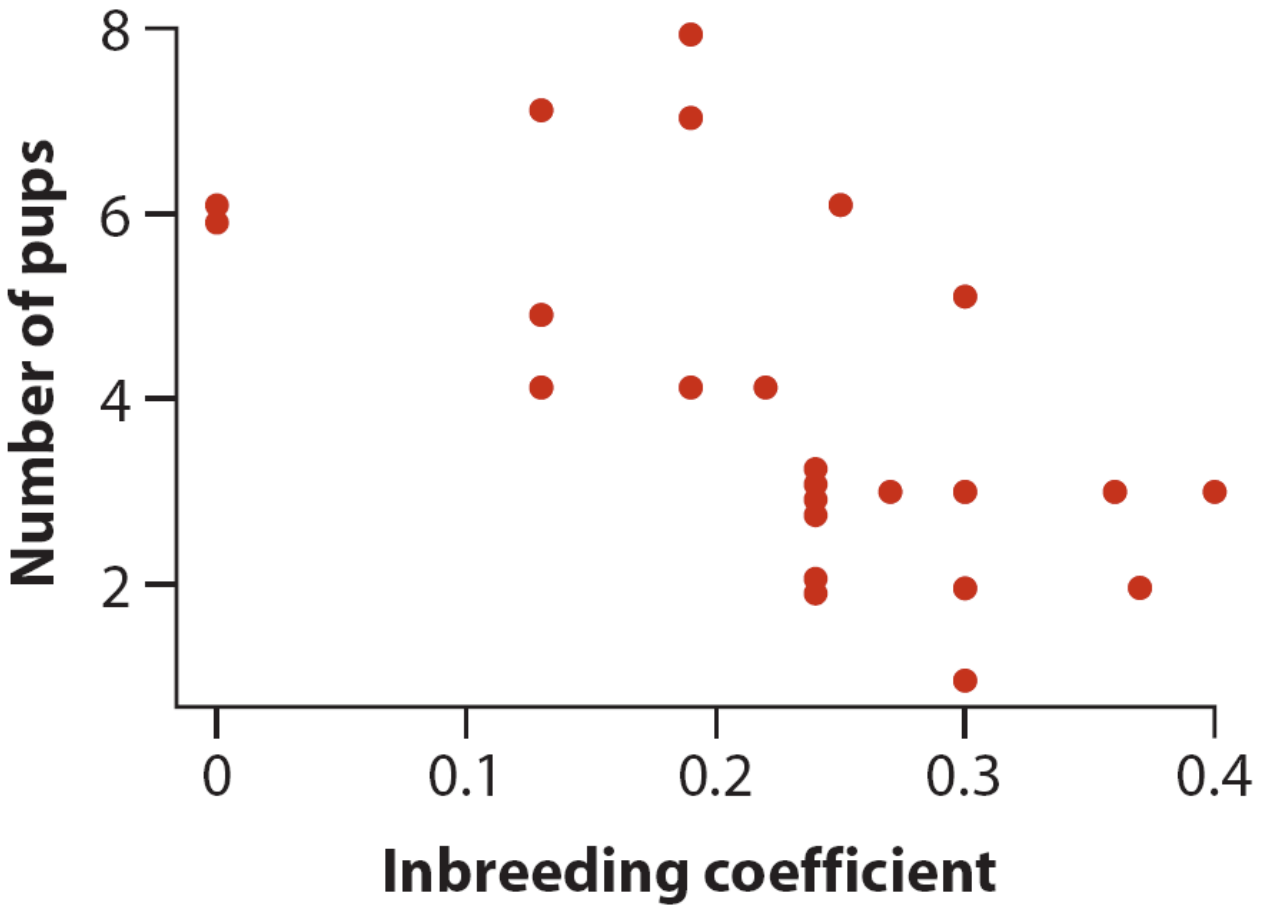
delle ipotesi.

## *Il tuo coefficiente di inincrocio è troppo grande?*

Nel 1970 il lupo (*Canis lupus*) si era estinto in Norvegia e in Svezia, ma intorno al 1980 una coppia di lupi immigrò da una lontana regione orientale e fondò una nuova popolazione. Nel 2002 questa popolazione aveva raggiunto circa 100 individui. Liberg et al. (2005) hanno analizzato molti dati sulla riproduzione raccolti tra il 1983 e il 2002 e hanno costruito l'albero genealogico dei lupi di questa piccola popolazione. I dati in Tabella 16.2-1 riportano i coefficienti di inincrocio nelle figliate e il numero di cuccioli di ogni figliata sopravvissuti al loro primo inverno. Il coefficiente di inincrocio in una figliata dipende dal grado di parentela dei genitori; è pari a zero se i genitori della figliata non sono imparentati, ma è uguale a 0,25, per esempio, se i genitori sono fratello e sorella e i due nonni non sono imparentati. ■

# Dati

| inbreeding.coef. | pups |
|---|---|
| 0 | 6 |
| 0 | 6 |
| 0.13 | 7 |
| 0.13 | 5 |
| 0.13 | 4 |
| 0.19 | 8 |
| 0.19 | 7 |
| 0.19 | 4 |
| 0.25 | 6 |
| 0.24 | 3 |
| 0.24 | 3 |
| 0.24 | 3 |
| 0.24 | 3 |
| 0.24 | 2 |
| 0.24 | 2 |
| 0.27 | 3 |
| 0.3 | 5 |
| 0.3 | 3 |
| 0.3 | 2 |
| 0.3 | 1 |
| 0.36 | 3 |
| 0.4 | 3 |
| 0.37 | 2 |
| 0.22 | 4 |



# La correlazione osservata

$\sum(X - \bar{X})\ (Y - \bar{Y})$= -2,621

$\sum(X - \bar{X})^2$= 0,228

$\sum(Y - \bar{Y})^2$= 80,958

$$r = \frac{-2{,}621}{\sqrt{0{,}228}\sqrt{80{,}958}} = -0{,}608$$

# Distribuzione di riferimento (t)

Nel nostro caso:

$H_0$: Non esiste relazione tra coefficiente di inincrocio e numero di cuccioli sopravvissuti (ρ=0)

$H_A$: Il coefficiente di inincrocio e il numero di cuccioli sopravvissuti sono correlati (ρ≠0)

Per verificare l'ipotesi calcoliamo la statistica t:

$$t = \frac{r}{ES_r}, dove\ ES_r = \sqrt{\frac{1 - r^2}{n - 2}}$$

# Calcoliamo t nell'esempio

$$ES_r = \sqrt{\frac{1-(-0{,}608)^2}{24-2}} = 0{,}169$$

e quindi:

$$t = \frac{-0{,}608}{0{,}169} = -3{,}60$$

Assumendo vera l'ipotesi nulla (assenza di correlazione), la distribuzione campionaria della statistica t è una distribuzione t di student con n-2 gradi di libertà (abbiamo dovuto stimare $\overline{X}$ e $\overline{Y}$)

Il p-value associato è di 0,002 (RIFIUTO L'IPOTESI NULLA)

Il valore critico della distribuzione t con 22 gdl è $t_{0,05;22}$=2,075 (RIFIUTO L'IPOTESI NULLA)

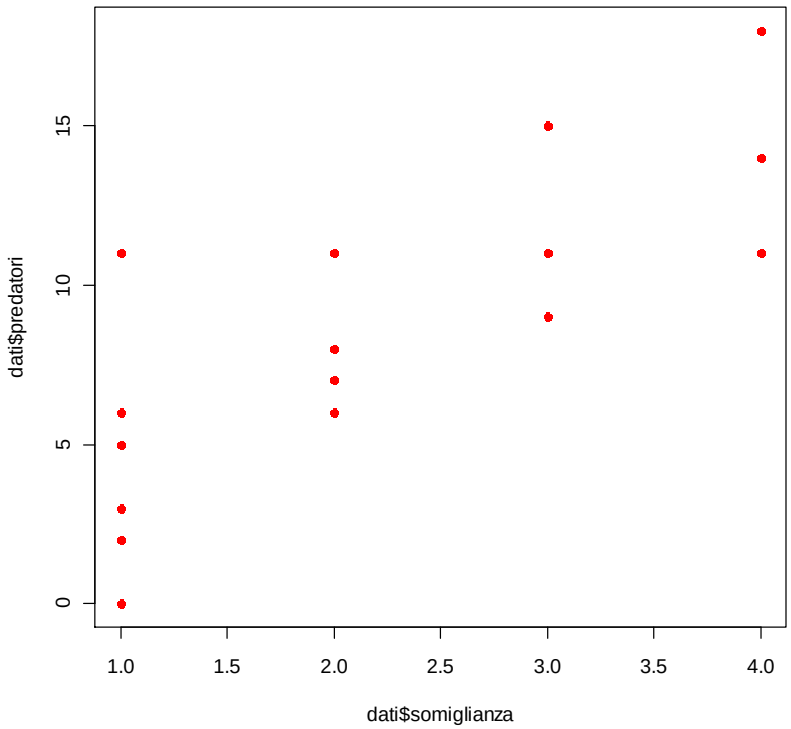
# Esercizio nr 11 pag 275 (Analisi statistica dei dati biologici)

Il pesce lima *Paraluteres prionurus* assomiglia al pesce palla tossico *Canthigaster valentini*. Entrambi sono cumuni nelle stesse scogliere coralline. Per valutare se la somiglianza con il pesce palla rappresenta un vantaggio per al sopravvivenza dei pesci lima, Caley e Schutler (2003) hanno dipinto pesci di plastica con pattern colorati che differivano nel loro livello di somiglianza con il pesce palla. Ai pesci finti è stato assegnato il seguente punteggio: 1 (maggiore somiglianza). 2, 3, 4 (minore somiglianza), a seconda del grado di somiglianza con il pesce palla. Successivamente, i ricercatori hanno collocato i pesci finti in località scelte casualmente sulla scogliera e hanno registrato il numero di pesci predatori che si sono avvicinati a essi durante periodi di osservazione di 5 minuti. Il numero di predatori registrati per ogni pesce finto è riportato nella tabella . La somiglianza con il pesce palla garantisce una certa protezione nei confronti dei predatori?

a. Visualizzate i dati in un diagramma e descrivere brevemente il pattern.
b. Verificate l'ipotesi che il numero di predatori che si avvicinano al pesce finto sia correlato con il grado di somiglianza con il pesce palla.

# Visualizzate i dati in un diagramma e descrivere brevemente il pattern

```
#Carico il file con R e salvo la matrice in «dati»
dati<-read.table(choose.files(),header=T)
#creo il grafico a dispersione
matplot(dati$somiglianza,dati$predatori,type="p",pch=16,col="red")
```



Verificate l'ipotesi che il numero di predatori che si avvicinano al pesce finto sia correlato con il grado di somiglianza con il pesce palla.

## Definizione dell'ipotesi:

$H_0$: Non esiste <u>nessuna relazione</u> tra la somiglianza con il pesce palla e il numero di predatori ($\rho=0$)

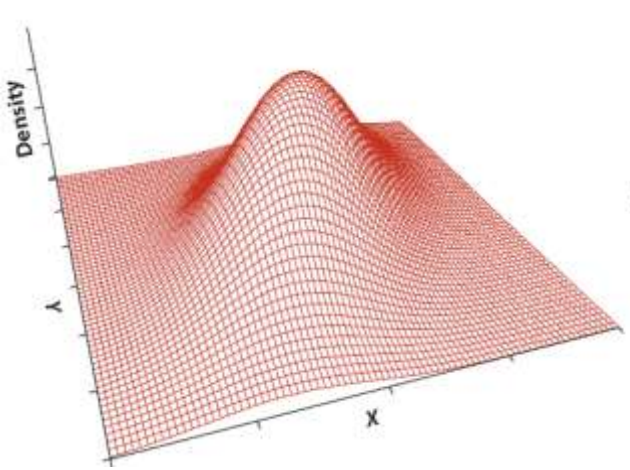
$H_A$: La somiglianza con il pesce palla e il numero di predatori sono correlati ($\rho\neq0$)

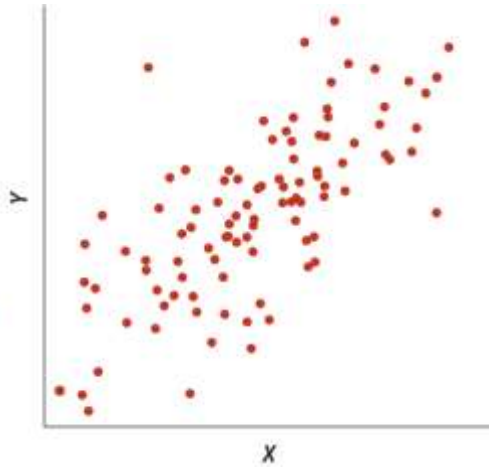
# Test d'ipotesi con R

```
#calcolo il coefficiente di correlazione osservato
r<-cor(dati$somiglianza,dati$predatori,method="pearson")
#calcolo l'errore standard di r
ESr<-sqrt((1-r^2)/(length(dati$somiglianza)-2))
#calcolo t osservato
tcalc<-r/Esr
#calcolo i valori di t critici con un alfa=5% e n-2 gdl
qt(0,975,18)
#eseguo il test d'ipotesi direttamente con la funzione integrata in R
cor.test(dati$somiglianza,dati$predatori, method="pearson", conf.level=0.95)
```

# Le assunzioni principali

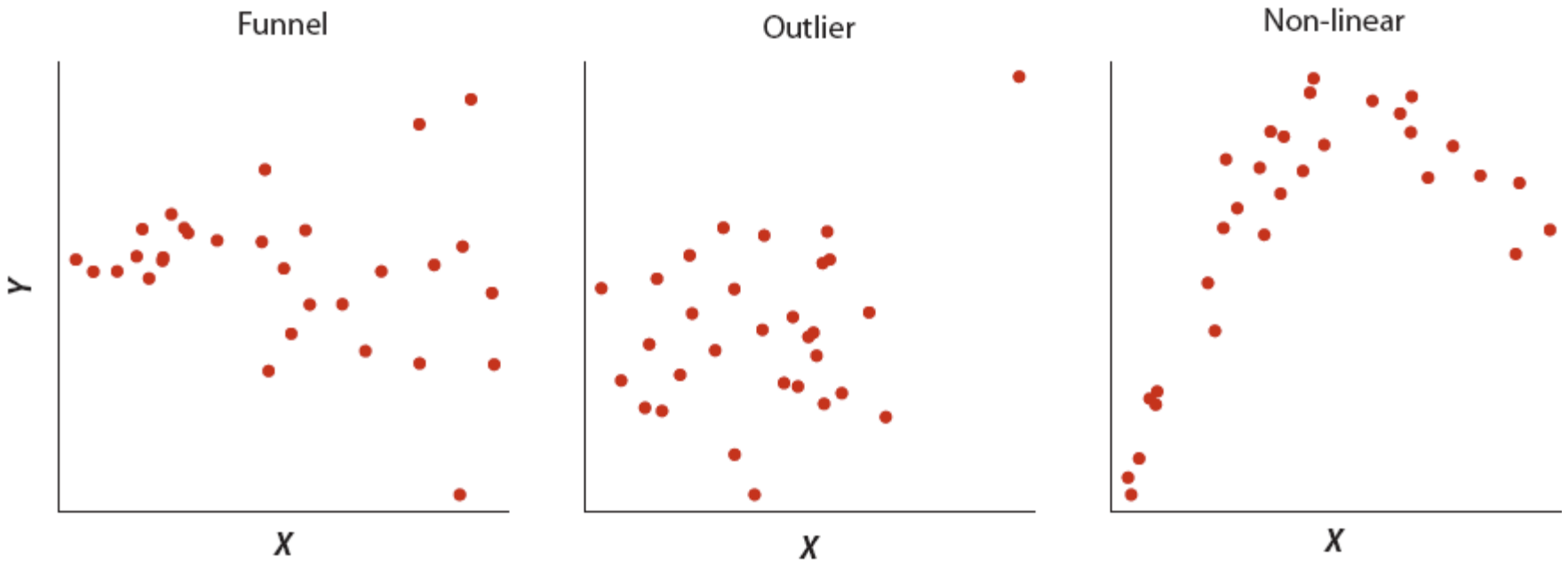
- Campione casuale di individui estratti dalla popolazione
- Le misure (x e y) abbiano una <u>distribuzione normale bivariata</u> nella popolazione.




- Relazione lineare tra X e Y
- Nube di punti circolare o ellittica
- Le distribuzioni di X e Y (separatamente) sono normali

# Scostamenti dalla normale bivariata




Nel caso l'assunzione di normalità bivariata NON sia soddisfatta:

- Trasformazione dei dati
- Utilizzo di metodi non parametrici

# Principali trasformazioni

Trasformazione di X, Y o entrambe in diversi modi per verificare l'assunzione di normalità bivariata in una nuova scala.

Principali trasformazioni:

- Logaritmica (attenzione però a dati negativi o agli zeri)
- Radice quadrata (efficace per dati di conteggi)
- Arcoseno (utile per proporzioni)

Se le trasformazioni non hanno successo, allora è consigliato l'uso di metodi non parametrici

# Correlazione non parametrica: correlazione per ranghi di Spearman

Questo indice di correlazione misura la forza e la direzione dell'associazione tra i ranghi di due variabili (R,S).

$$r_S = \frac{\sum(R - \bar{R})(S - \bar{S})}{\sqrt{\sum(R - \bar{R})^2}\sqrt{\sum(S - \bar{S})^2}}$$

Applicabile nel caso l'assunzione di normalità bivariata di X e Y non sia soddisfatta.

# I ranghi

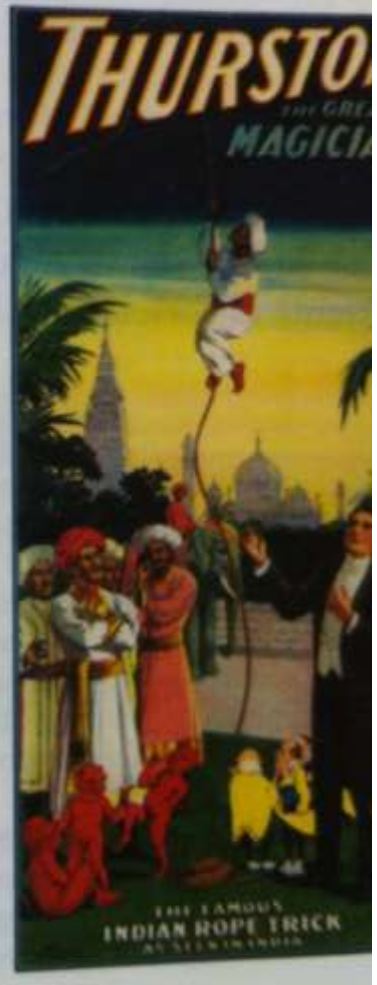
| Rango | X | Rango | X |
|---|---|---|---|
| 1 | 10 | 7(6) | 15 |
| 2 | 11 | 7(7) | 15 |
| 3 | 12 | 7(8) | 15 |
| 4 | 13 | 9 | 16 |
| 5 | 14 | 10 | 17 |
| … | … | … | … |

# Esempio

Esempio 16.5

## I miracoli della memoria

Tra i diversi giochi di illusionismo compiuti dai prestigiatori, il più famoso è sicuramente quello della «corda indiana». Nella versione più sensazionale di questo trucco, un prestigiatore lancia in aria l'estremità di una corda, che assume immediatamente la forma di un'asta rigida. Un bambino si arrampica sulla corda e scompare in alto. Il prestigiatore sgrida il bambino invitandolo a ritornare, ma non riceve risposta; afferra quindi un coltello, si arrampica a sua volta sulla corda e scompare anche lui. In seguito il corpo del bambino cade a pezzi dall'alto in un cesto che si




unteggio di suggestività assegnato ai resoconti del trucco
la corda indiana scritti da testimoni oculari e il numero
anni trascorsi tra l'osservazione dell'evento e la stesura
resoconto. $n = 21$

trova al suolo. Infine, il prestigiatore scende lungo la corda e fa uscire il bambino dal cesto, mostrando che è incolume.

Wiseman e Lamont (1996) hanno raccolto 21 resoconti scritti di prima mano del trucco della corda indiana. I ricercatori hanno assegnato un punteggio a ciascuna descrizione in base al grado di suggestione percepita. Per esempio, hanno assegnato al resoconto il punteggio 1 se l'osservatore ricorda di aver visto soltanto «il bambino arrampicarsi sulla corda e poi ridiscendere». Ai resoconti più impressionanti («il bambino si arrampica sulla corda, scompare in alto, ricompare nel cesto ben visibile agli spettatori») è stato assegnato il punteggio 5, il più alto. Per ogni resoconto, i ricercatori hanno registrato anche il numero di anni trascorsi tra la data in cui è stato visto lo spettacolo e la data in cui è stato redatto il ricordo dell'evento. I punteggi di suggestività e il numero di anni trascorsi sono riportati nello scatter plot in Figura 16.5-1. Esiste un'associazione tra la suggestività dei resoconti dei testimoni e il tempo trascorso dalla performance al momento in cui hanno scritto ciò che ricordavano? In caso affermativo, questo risultato potrebbe rappresentare una prova della fallibilità della memoria umana. ■

# Procedimento

Verifica dell'ipotesi nulla di assenza di correlazione in condizioni di palese violazione dell'assunzione di normalità bivariata (il punteggio di suggestività è una variabile categorica!)

Utilizzo del coefficiente di correlazione per ranghi di Spearman per misurare l'associazione tra i ranghi delle variabili.

Questa misura è definita dal parametro $\rho_S$, stimato da $r_S$

# I dati

Tabella 13.5-1

*Dati grezzi riferiti all'Esempio 16.5-1 e loro ranghi. Il rango è assegnato separatamente a ogni variabile. Ai valori uguali (ties) sono assegnati ranghi medi. n = 21.*

| Anni trascorsi | Rango degli anni trascorsi | Punteggio di suggestività | Rango del punteggio di su |
|---|---|---|---|
| 2 | 1 | 1 | 2 |
| 5 | 3,5 | 1 | 2 |
| 5 | 3,5 | 1 | 2 |
| 4 | 2 | 2 | 5 |
| 17 | 5,5 | 2 | 5 |
| 17 | 5,5 | 2 | 5 |
| 31 | 13 | 3 | 7 |
| 20 | 7 | 4 | 12,5 |
| 22 | 8 | 4 | 12,5 |
| 25 | 9 | 4 | 12,5 |
| 28 | 10,5 | 4 | 12,5 |
| 29 | 12 | 4 | 12,5 |
| 34 | 14,5 | 4 | 12,5 |
| 43 | 17 | 4 | 12,5 |
| 44 | 18 | 4 | 12,5 |
| 46 | 19 | 4 | 12,5 |
| 34 | 14,5 | 4 | 12,5 |
| 28 | 10,5 | 5 | 19,5 |
| 39 | 16 | 5 | 19,5 |
| 50 | 20,5 | 5 | 19,5 |
| 50 | 20,5 | 5 | 19,5 |

# Calcoli

R, rango degli anni trascorsi

S, rango del punteggio di similarità

$\sum(R - \bar{R})\ (S - \bar{S})$= 566

$\sum(R - \bar{R})^2$= 767,5

$\sum(S - \bar{S})^2$= 678,5

$$r_S = \frac{566}{\sqrt{767,5}\sqrt{678,5}} = 0,784$$

# L’ipotesi nulla e alternativa

$H_0$: Non esiste relazionetra la suggestività dei resoconti e il tempo trascorso ($\rho_S=0$)

$H_A$: La suggestività dei resoconti e il tempo trascorso sono correlati ($\rho_S\neq0$)

# Test dell'ipotesi

Se n<=100:

Confrontiamo il $r_S$ calcolato con il valore critico dato dalla tavola statistica G, $r_{S(0,05;21)}$= 0,435

Poiché $r_S$ calcolato è maggiore del $r_S$ critico, rifiuto l'ipotesi nulla.

Se n>100:

Utilizziamo la stessa procedura del coefficiente di correlazione lineare, applicandola ai ranghi.

$$t = \frac{r_S}{ES_{r_S}}, dove\ ES_{r_S} = \sqrt{\frac{1 - {r_S}^2}{n - 2}}$$

# Calcolo del coefficiente di correlazione di Spearman con R

```
#caricare il file «Correlazione_esercizio_corda_indiana.txt»
dati<-read.table(choose.files(),header=T)
#calcolare il rank di x e y
rx<-rank(dati$years)
ry<-rank(dati$impressiveness)
#calcolo il coefficiente di correlazione
num<-  sum((rx-mean(rx))*(ry-mean(ry)))
den1<-sqrt(sum((rx-mean(rx))^2))
den2<-sqrt(sum((ry-mean(ry))^2))
r<-num/(den1*den2)
#calcolo del p-value associato all'ipotesi nulla
cor.test(dati$years,dati$impressiveness,method="spearman")
```