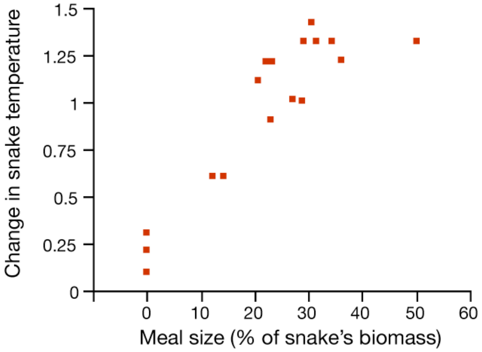
Two variables: Which test?

		Explanatory variable	
		Categorical	Numerical
Response variable	Categorical	Contingency analysis	Logistic regression Survival analysis
	Numerical	*t*-test Analysis of variance	Regression Correlation

Correlation

Chapter 16

Scatter plot



Tattersall et al. (2004) *Journal of Experimental Biology* 207:579-585

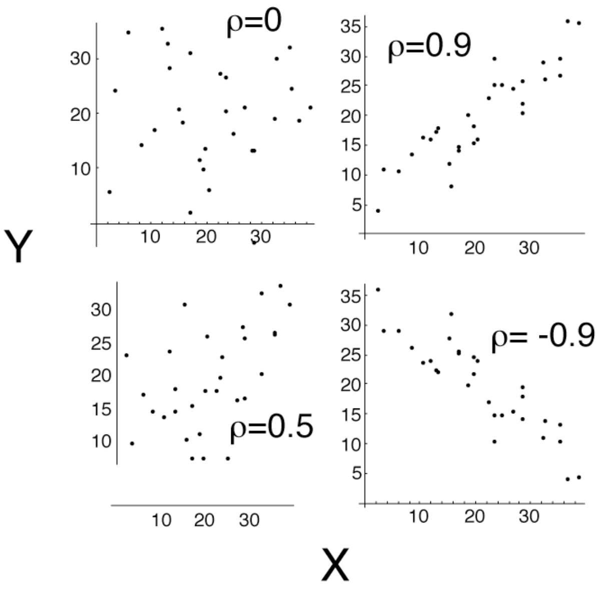
Correlation: *r*

r is called the "correlation coefficient"

Describes the relationship between two numerical variables

Parameter: ρ (rho) Estimate: *r*

$-1 < \rho < 1$



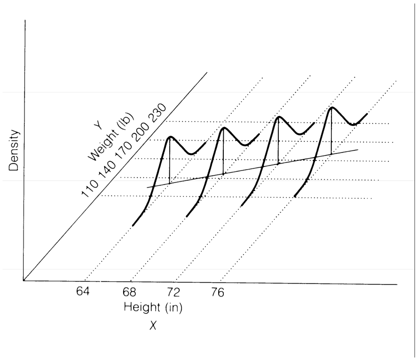
https://shiney.zoology.ubc.ca/whitlock/Guessing_correlation/

Correlation assumes...

Random sample

X is normally distributed with equal variance for all values of Y

Y is normally distributed with equal variance for all values of X



Coefficient of determination

r^2

Describes the proportion of variation in one variable that can be predicted from the other variable

Correlation coefficient

$$r = \frac{\text{Covariance}(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

Estimating the correlation coefficient

"Sum of cross products"

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

"Sum of squares"

Covariance

$$\text{Covariance}(X,Y) = \frac{\sum\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{n-1}$$

Standard error of *r*

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

If ρ = 0,...

r is normally distributed with mean 0

$$t = \frac{r}{SE_r}$$ with $df = n - 2$

Example

Black-tailed godwits are migratory and socially monogamous.

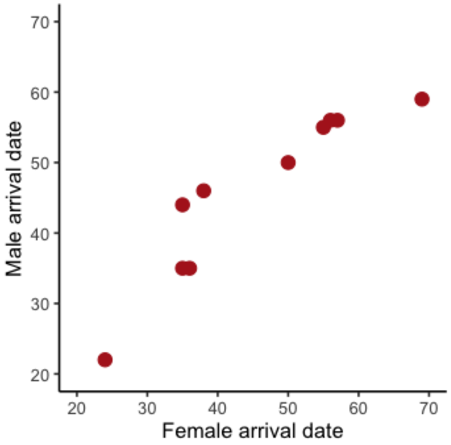
Are the males and females in a pair correlated in their arrival dates after migration?

© Andreas Trepte, www.photo-natur.net, CC-BY-SA

Godwit arrival time data

(units: days after March 31)

Female arrival date (*X*)	Male arrival date (*Y*)
24	22
36	35
35	35
35	44
38	46
50	50
55	55
56	56
57	56
69	59
$\sum X = 455$	$\sum Y = 458$



Gunnarsson, T. G., J. A. Gill, T. Sigurbjörnsson, and W. J. Sutherland. 2004. Arrival synchrony in migratory birds. *Nature* 431: 646.

Hypotheses

H_0: Arrival date of female and arrival date of male are not related ($\rho = 0$).

H_A: Arrival date of female and arrival date of male are correlated ($\rho \neq 0$).

Godwit arrival time data

(units: days after March 31)

Female arrival date (X)	Male arrival date (Y)	$X - \bar{X}$	$Y - \bar{Y}$	$(X - \bar{X})^2$	$(Y - \bar{Y})^2$	$(X - \bar{X})\times(Y - \bar{Y})$
24	22	-21.5	-23.8	462.25	566.44	511.7
36	35	-9.5	-10.8	90.25	116.64	102.6
35	35	-10.5	-10.8	110.25	116.64	113.4
35	44	-10.5	-1.8	110.25	3.24	18.9
38	46	-7.5	0.2	56.25	0.04	-1.5
50	50	4.5	4.2	20.25	17.64	18.9
55	55	9.5	9.2	90.25	84.64	87.4
56	56	10.5	10.2	110.25	104.04	107.1
57	56	11.5	10.2	132.25	104.04	117.3
69	59	23.5	13.2	552.25	174.24	310.2
	Sum	0	0	1734.5	1287.6	1386

Finding *r*

$$\sum (X - \bar{X})\,(Y - \bar{Y}) = 1386$$

$$\sum (X - \bar{X})^2 = 1734.5$$

$$\sum (Y - \bar{Y})^2 = 1287.6$$

$$r = \frac{\sum (X - \bar{X})\,(Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2 \sum (Y - \bar{Y})^2}} = \frac{1386}{\sqrt{1734.5 \times 1287.6}} = 0.927$$

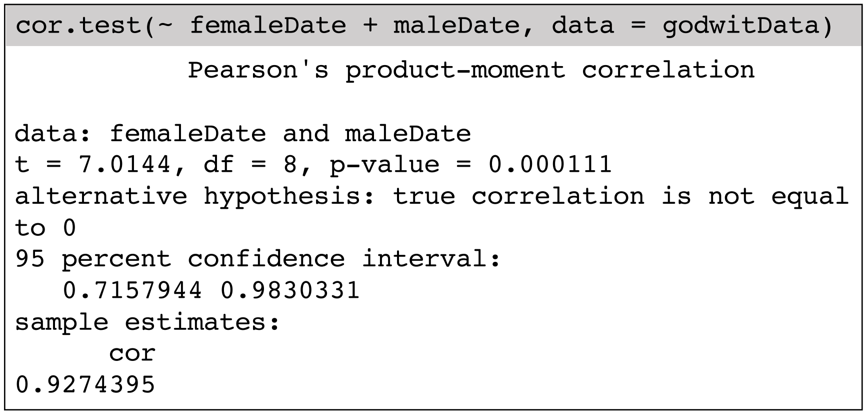
$$r = 0.927$$

$$SE_r = \sqrt{\frac{1-r^2}{n-2}} = \sqrt{\frac{1-0.927^2}{8}} = 0.1322$$

$$t = \frac{r}{SE_r} = \frac{0.927}{0.1322} = 7.01$$

df = n – 2 = 10 – 2 = 8

t=7.01 is greater than $t_{0.05(2),\,8}$ = 2.31, so we can reject the null hypothesis and say that female and male arrival times are correlated.



```
cor.test(~ femaleDate + maleDate, data = godwitData)
               Pearson's product-moment correlation

data: femaleDate and maleDate
t = 7.0144, df = 8, p-value = 0.000111
alternative hypothesis: true correlation is not equal
to 0
95 percent confidence interval:
   0.7157944 0.9830331
sample estimates:
      cor
0.9274395
```

Shortcuts

$$\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})=\left(\sum X_iY_i\right)-\frac{\sum X_i\sum Y_i}{n}$$

$$\sum_{i=1}^{n}(X_i-\bar{X})^2=\sum\left(X_i^2\right)-\frac{\left(\sum X_i\right)^2}{n}$$

$$\sum_{i=1}^{n}(Y_i-\bar{Y})^2=\sum\left(Y_i^2\right)-\frac{\left(\sum Y_i\right)^2}{n}$$

X is female arrival date,
Y is male arrival date

$$\sum X = 455 \qquad \sum Y = 458$$

$$\sum X^2 = 22437 \qquad \sum Y^2 = 22264$$

$$\sum XY = 22225 \qquad n = 10$$

Gunnarsson, T. G., J. A. Gill, T. Sigurbjörnsson, and W. J. Sutherland. 2004. Arrival synchrony in migratory birds. *Nature* 431: 646.

Spearman's rank correlation

An alternative to correlation that does not make so many assumptions

Example: Spearman's r_s

VERSIONS:

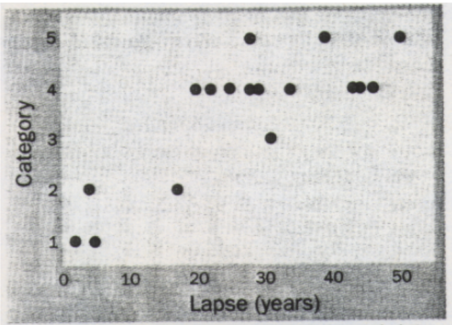
1. Boy climbs up rope, climbs down again
2. Boy climbs up rope, seems to vanish, re-appears at top, climbs down again
3. Boy climbs up rope, seems to vanish at top
4. Boy climbs up rope, vanishes at top, reappears somewhere the audience was not looking
5. Boy climbs up rope, vanishes at top, reappears in a place which has been in full view

Hypotheses

H_0: The difficulty of the described trick is not correlated with the time elapsed since it was observed.

H_A: The difficulty of the described trick is correlated with the time elapsed since it was observed.

Example: Spearman's r_s

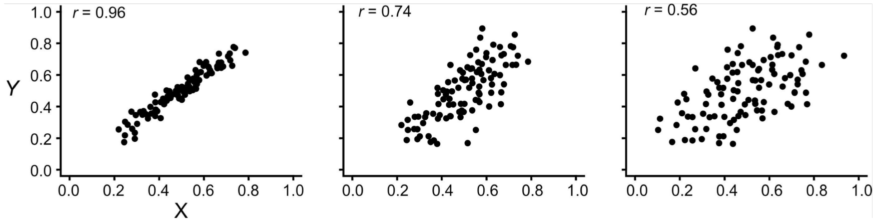


Scatterplot of the 'category' and 'lapse' data for each of the 21 eyewitnesses.

$$r_s = 0.712$$

P<0.05

Attenuation:
The estimated correlation will be lower if *X* or *Y* are estimated with error

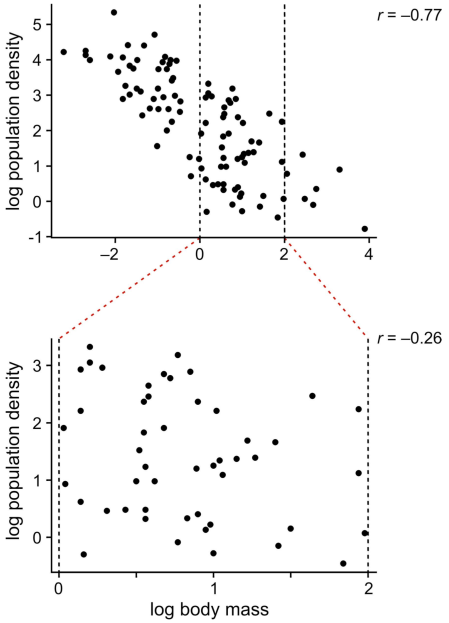


Real correlation

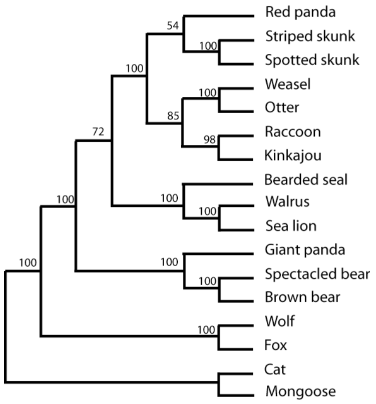
Y estimated with measurement error

X and *Y* estimated with measurement error

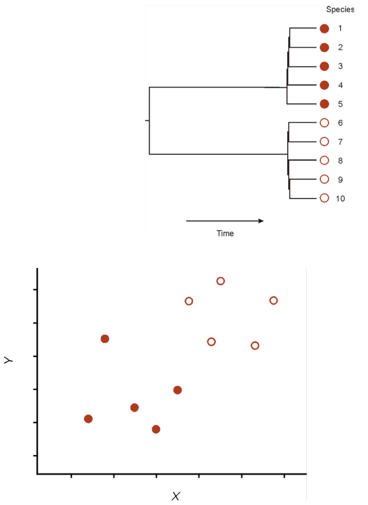
Correlation depends on range



Species are not independent data points



Red panda
Striped skunk
Spotted skunk
Weasel
Otter
Raccoon
Kinkajou
Bearded seal
Walrus
Sea lion
Giant panda
Spectacled bear
Brown bear
Wolf
Fox
Cat
Mongoose
54
100
100
100
85
98
72
100
100
100
100
100
100
100
100


Species
Time
Y
X