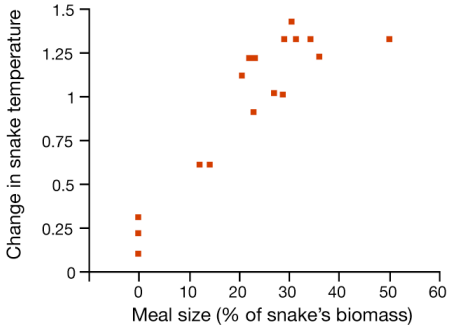
Two variables: Which test?

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

Scatter plot

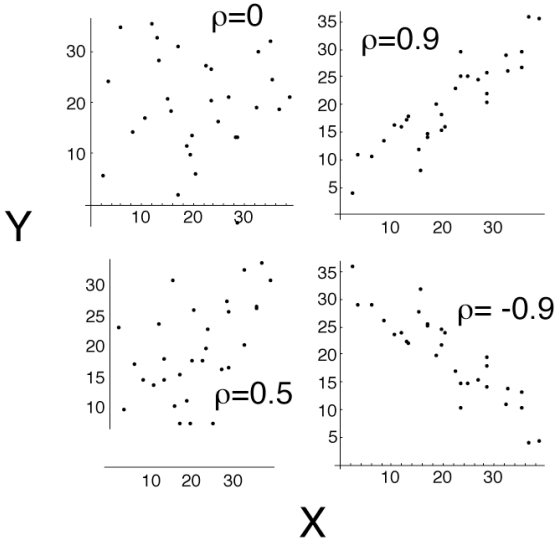

Tropical Rattlesnake *(Venomous)*



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*

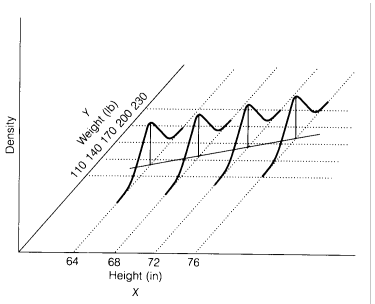


Correlation assumes...

•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*



Correlation coefficient facts

- $-1 < \rho < 1$
- Coefficient of determination: r^2: Describes the proportion of variation in one variable that can be predicted from the other

Estimating the correlation coefficient

"Sum of cross products"

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

"Sum of squares"

Standard error of *r*

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

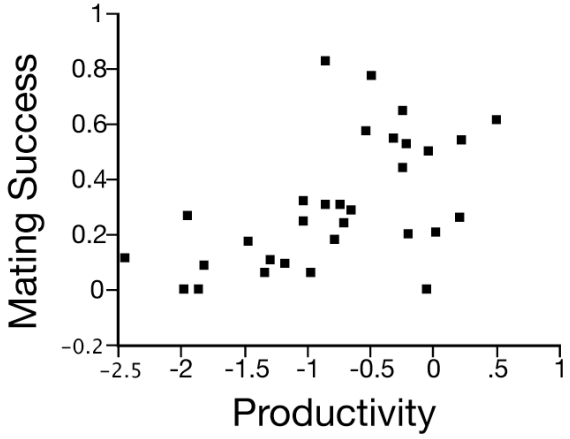
If $\rho = 0$,...

r is normally distributed with mean 0

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

Example

- Are the effects of new mutations on mating success and productivity correlated?
- Data from various visible mutations in *Drosophila melanogaster*



Hypotheses

H_0: Mating success and productivity are not related ($\rho = 0$).

H_A: Mating success and productivity are correlated ($\rho \neq 0$).

X is productivity,
Y is the mating success

$$\sum X = -24.228 \quad \sum Y = 9.498$$

$$\sum X^2 = 35.1808 \quad \sum Y^2 = 4.5391$$

$$\sum XY = -4.62741 \quad n = 31$$

Shortcuts

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

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

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

Finding *r*

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

$$= -4.627 - \frac{(-24.228)(9.4982)}{31} = 2.796$$

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum\left(X_i^2\right) - \frac{\left(\sum X_i\right)^2}{n} = 35.1808 - \frac{(-24.228)^2}{31} = 16.245$$

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum\left(Y_i^2\right) - \frac{\left(\sum Y_i\right)^2}{n} = 4.539 - \frac{(9.49824)^2}{31} = 1.6289$$

$$r = \frac{2.796}{\sqrt{(16.245)(1.6289)}} = 0.5435$$

$$SE_r = \sqrt{\frac{1 - r^2}{n - 2}} = \sqrt{\frac{0.7045}{29}} = 0.1558$$

$$t = \frac{0.5435}{0.1558} = 3.49$$

df= n-2=31-2=29

t=3.49 is greater than $t_{0.05(2),\,29} = 2.045$, so we can reject the null hypothesis and say that productivity and male mating success are correlated across genotypes.

Spearman's rank correlation

- An alternative to correlation that does not make so many assumptions

Example: Spearman's r_s

VERSIONS:

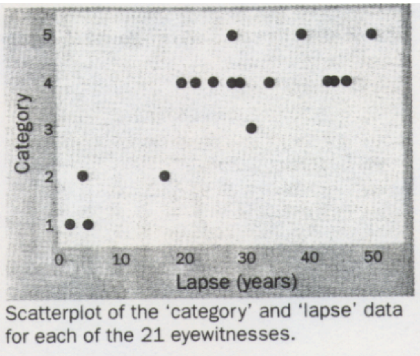
1. Boy climbs up rope, climbs down again
2. Boy climbs up rope, seems to vanish, re-appears at top, climbs down again
3. Boy climbs up rope, seems to vanish at top
4. Boy climbs up rope, vanishes at top, reappears somewhere the audience was not looking
5. Boy climbs up rope, vanishes at top, reappears 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

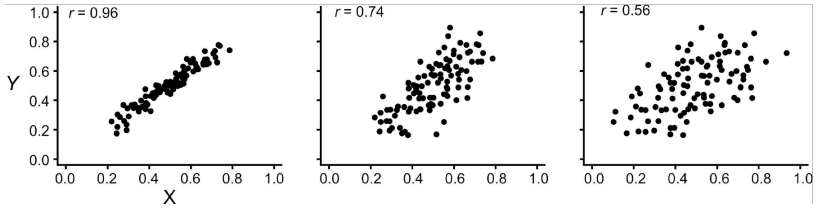


Scatterplot of the 'category' and 'lapse' data for each of the 21 eyewitnesses.

$$r_s = 0.712$$

$P<0.05$

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

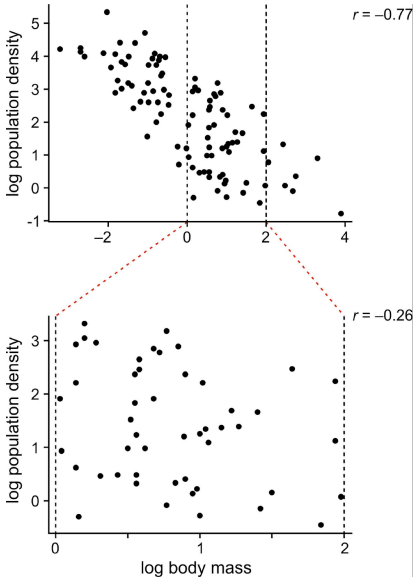


Real correlation

Y estimated with measurement error

X and *Y* estimated with measurement error

Correlation depends on range



Species are not independent data points