UNDERSTANDING STATISTICS & EXPERIMENTAL DESIGN

Content

EPFL

1. Basic Probability Theory
2. Signal Detection Theory (SDT)
3. SDT and Statistics I and II
4. Statistics in a nutshell
5. Multiple Testing
6. ANOVA
7. Experimental Design & Statistics
8. Correlations & PCA
9. Meta-Statistics: Basics
10. Meta-Statistics: Too good to be true
11. Meta-Statistics: How big a problem is publication bias?
12. Meta-Statistics: What do we do now?

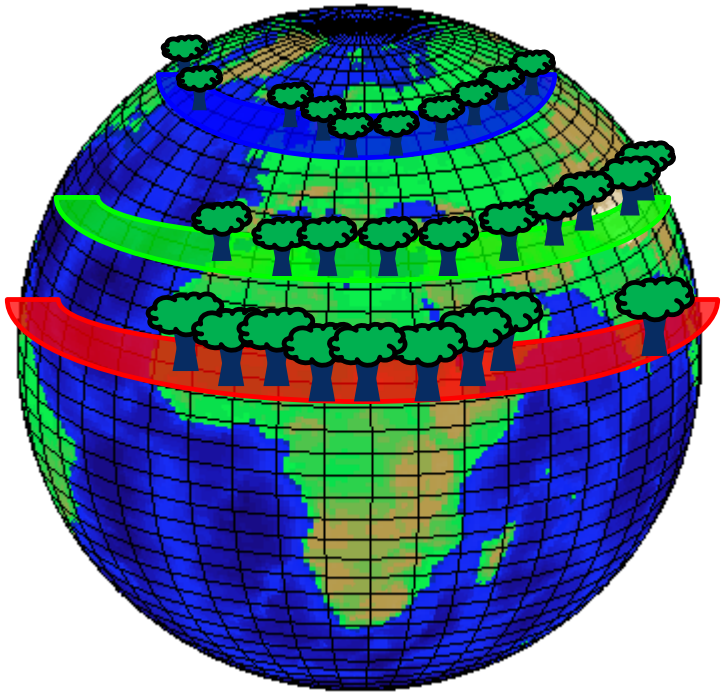
Covariance & Correlation

Correlation & regression
Reduce many data to one value, the r-value

Tree heights

Suppose we want to compare tree heights over the three regions shown on the right in the red, green, and blue bands.

Are the tree heights all the same, or does at least one region contain trees whose heights differ from those in the other regions?

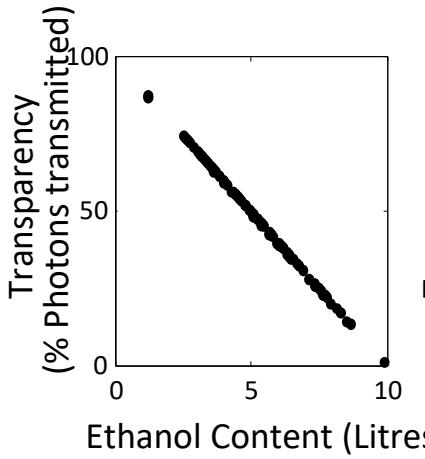


Example

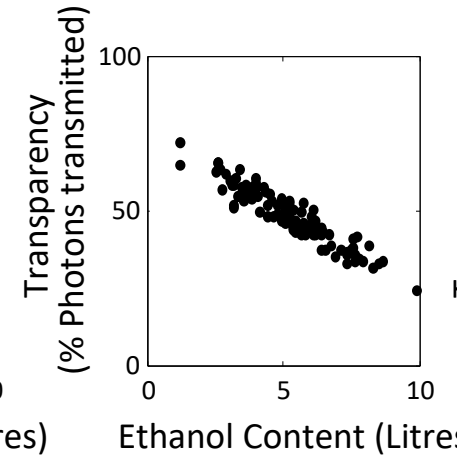
Wonder Woman is interested in determining if there is a relationship between the ethanol content of her fuel and the transparency of the fuel tanks on her invisible jet.

Potential Outcomes

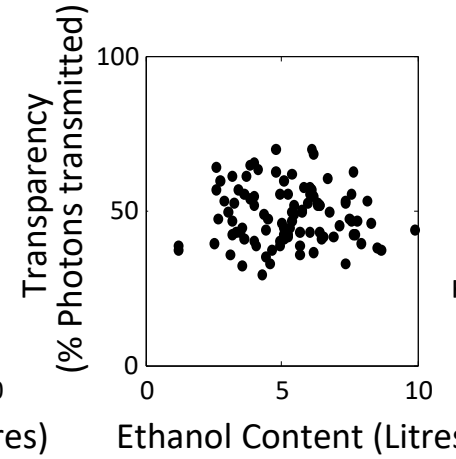
Strong negative relationship



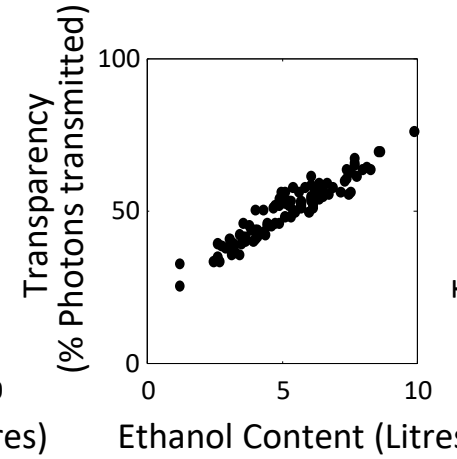
Weak negative relationship



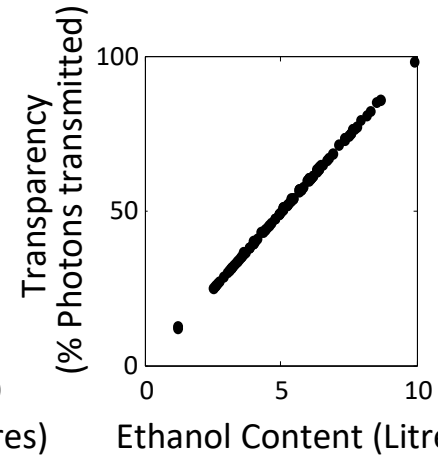
No relationship



Weak positive relationship

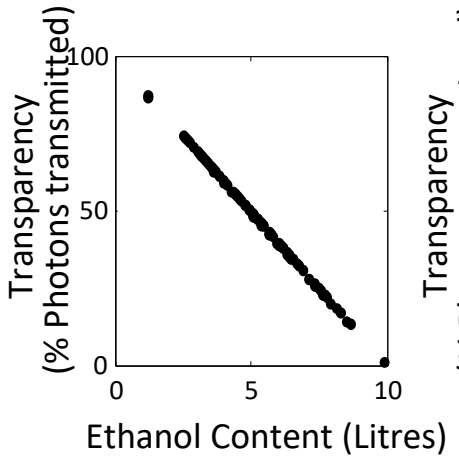


Strong positive relationship

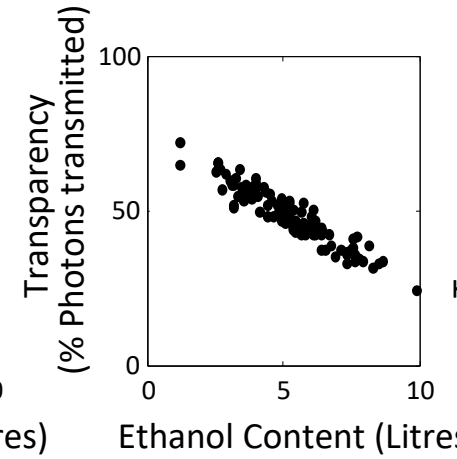


Potential Outcomes

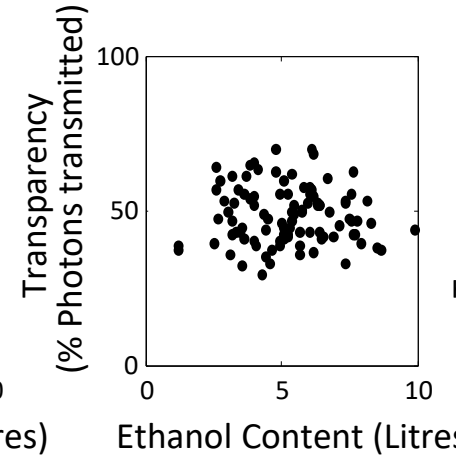
Strong negative relationship



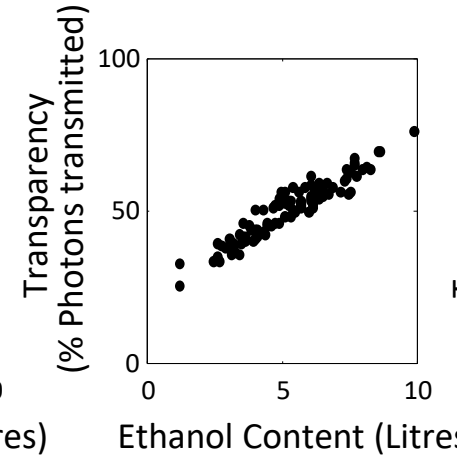
Weak negative relationship



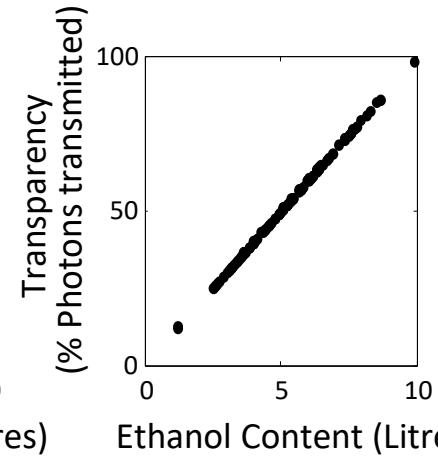
No relationship



Weak positive relationship



Strong positive relationship

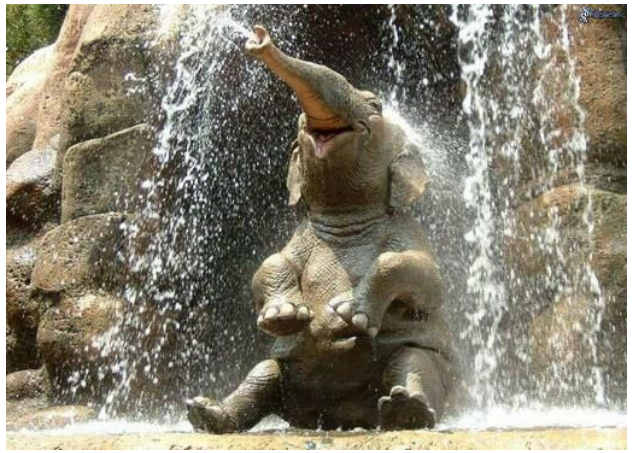


$$Cov(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{n-1}$$

Covariation & Scaling

$$Cov(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{n-1}$$

$$R = \frac{Cov(x,y)}{s_x \cdot s_y}$$

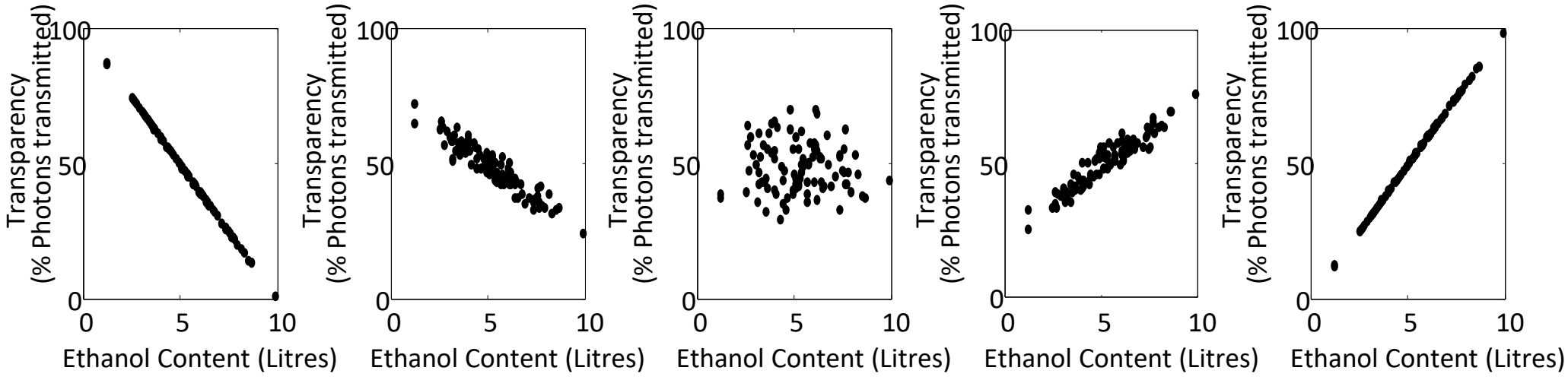


cm & m
Elephant and ant

How to quantify relatedness

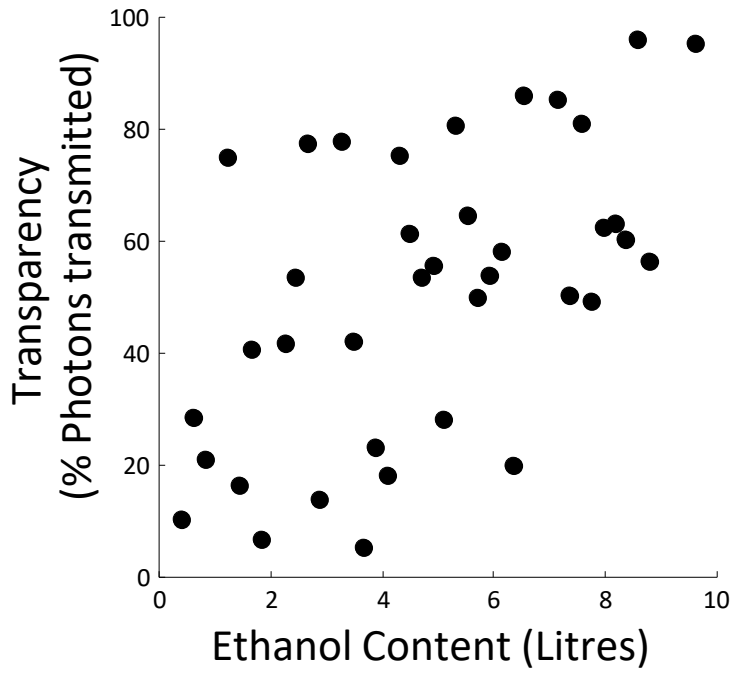
Potential Outcomes

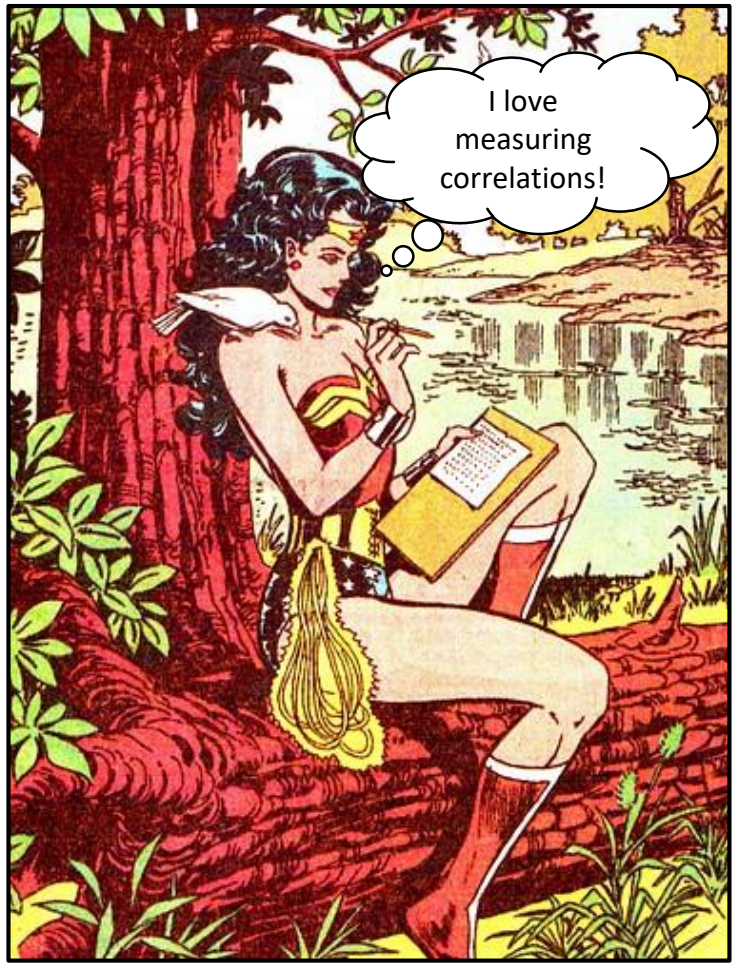
R = -1 R = -0.961 R = 0 R = 0.961 R = 1



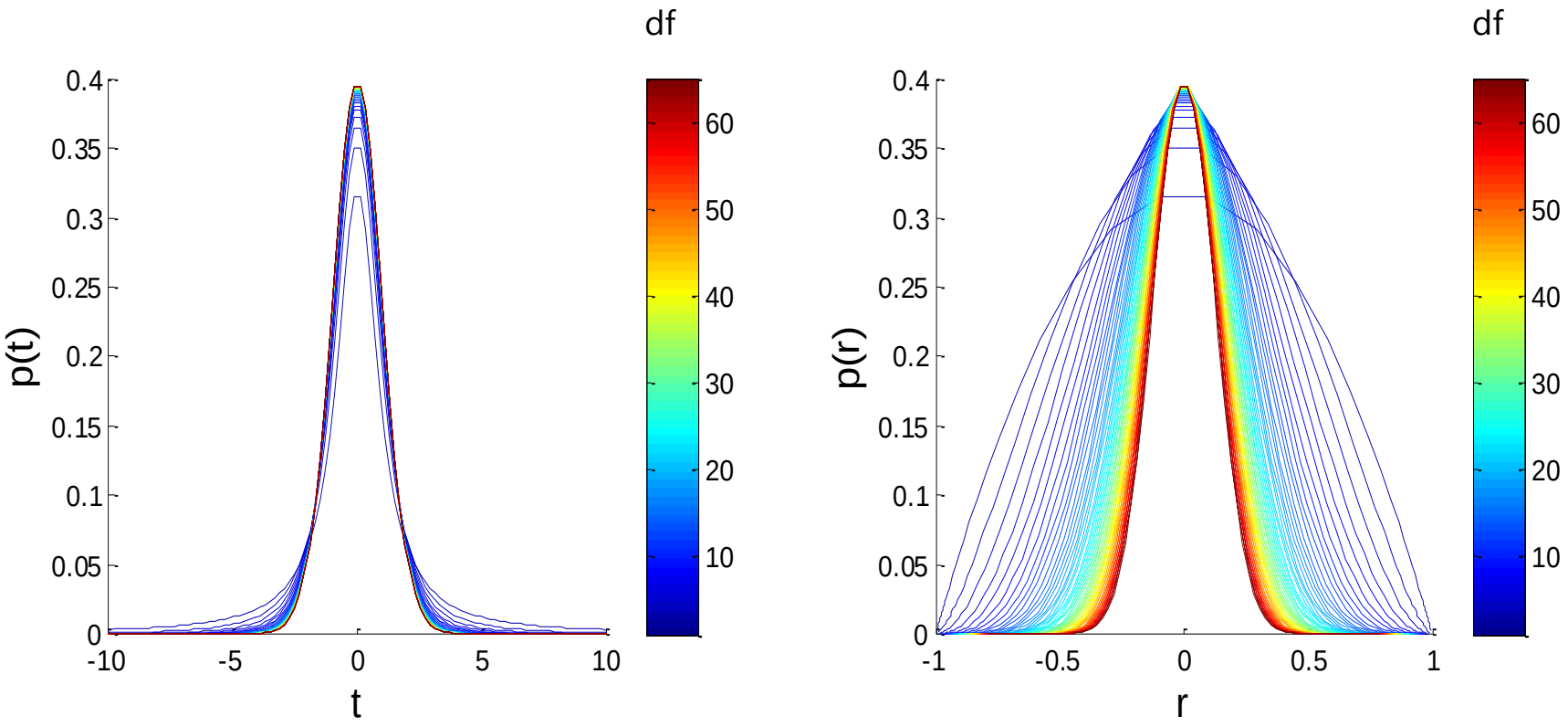
Example

Wonder Woman does some calculations and discovers that with 50 data points R = 0.75.





Test Statistics



Test Statistics

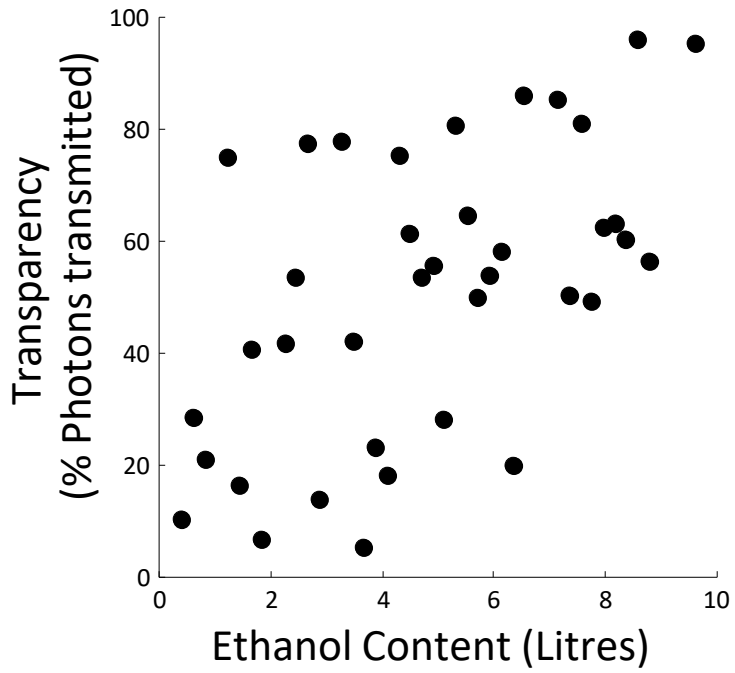
Under the null hypothesis of no correlation (i.e. r = 0)

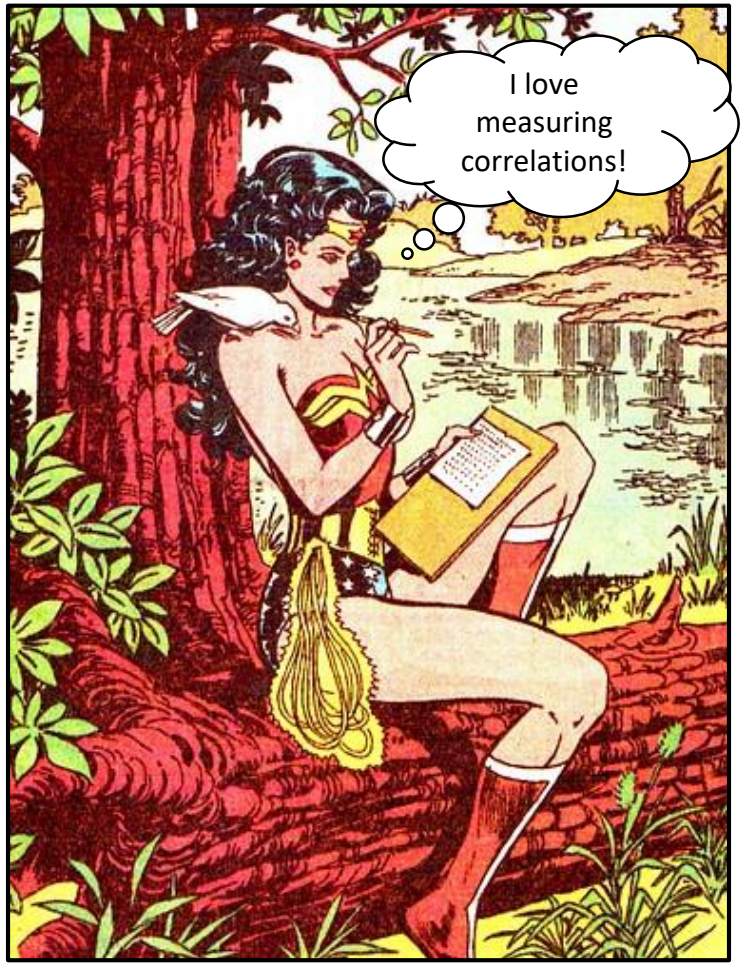
$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

df = n-2

R2 explained variance

For R(48) = 0.75, p = 0.00003

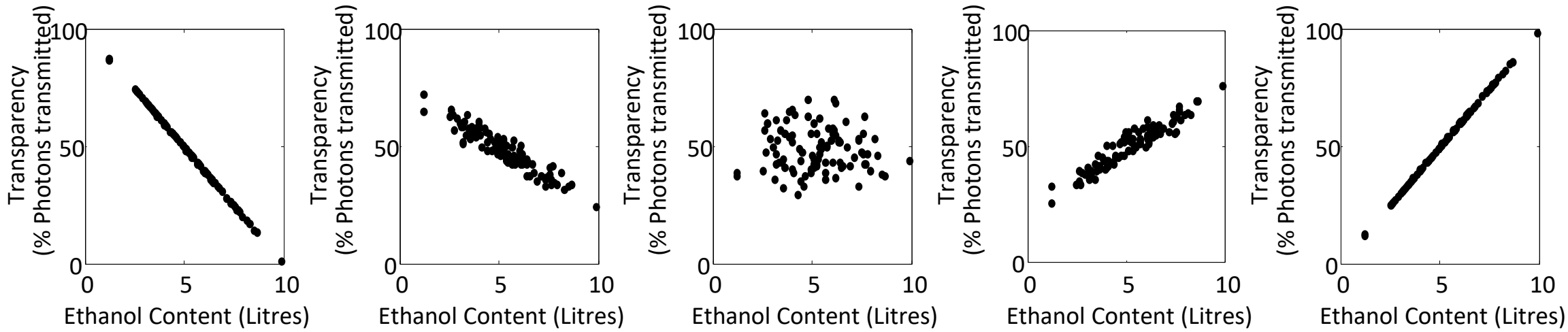




Correlations measure linear dependencies

Suppose there are n data points $\{(x_i, y_i), i = 1, ..., n\}$. The function that describes x and y is:

$$y_i = a + b\,x_i + e_i$$



The goal is to find the equation of the straight line:

$$y = a + b\,x$$

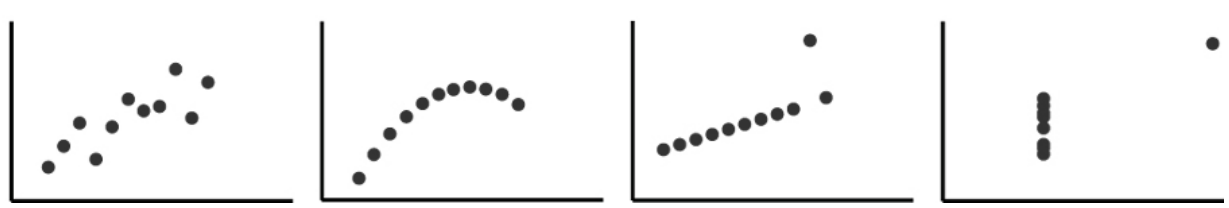
(a) The four sets of numbers that form Anscombe's quartet. (**b**) The highly distinctive graphs that result from plotting the data in **a**.

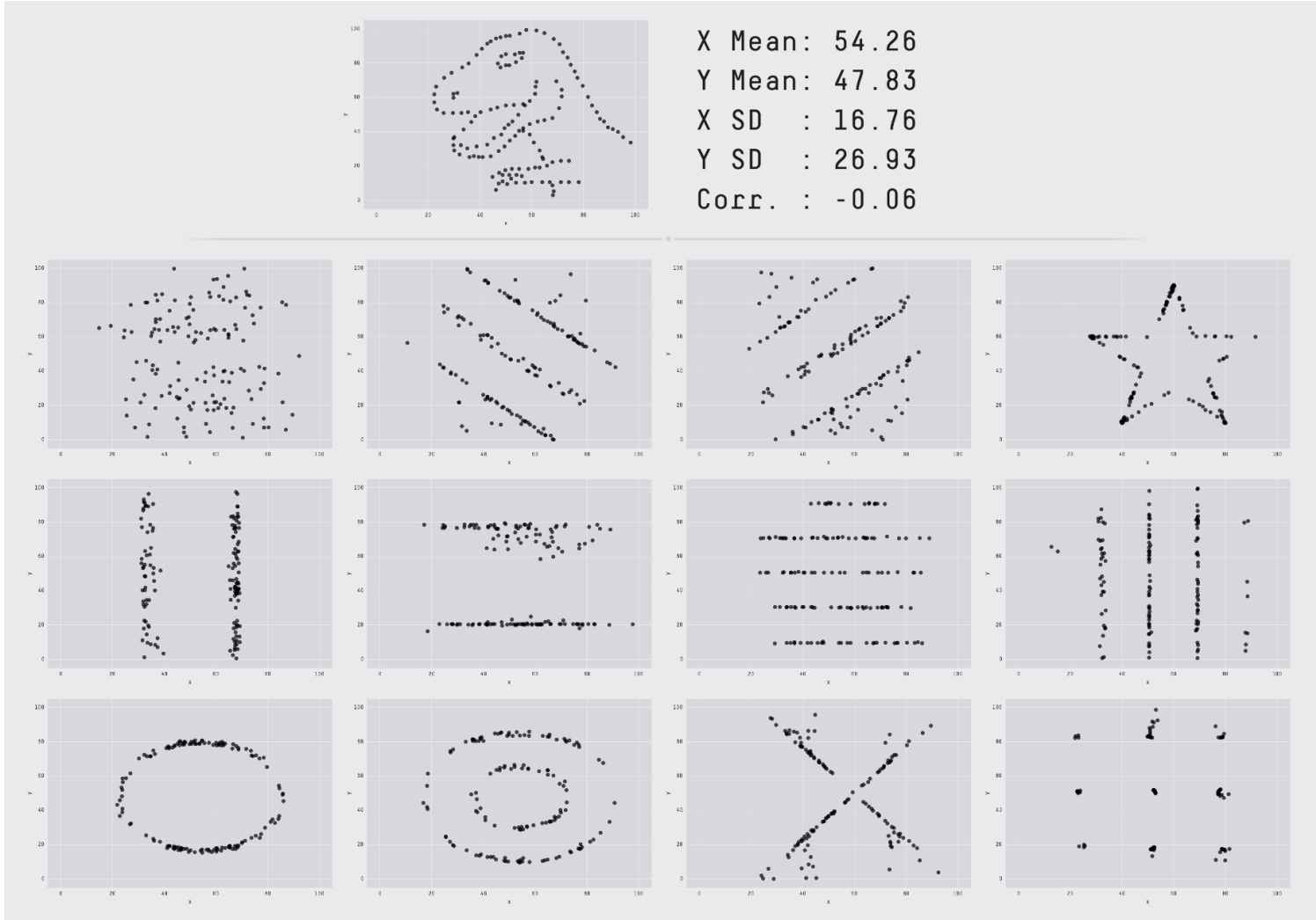
(b) Noam Shoresh, Bang Wong, Nature Methods, 9, 5,(2012)

a

I		II		III		IV	
x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.10	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.10	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

b





https://www.autodeskresearch.com/publications/samestats

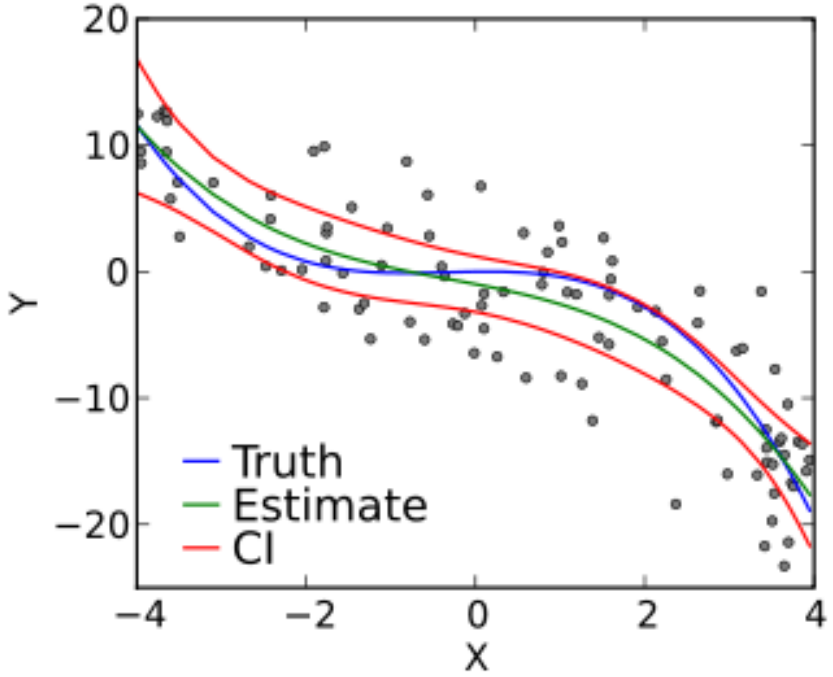
Correlations measure linear dependencies: models

Suppose there are n data points $\{(x_i, y_i), i = 1, ..., n\}$.
The function that describes x and y is:

$$y_i = a + b\,x_i + e_i$$

The goal is to find the equation of the straight line

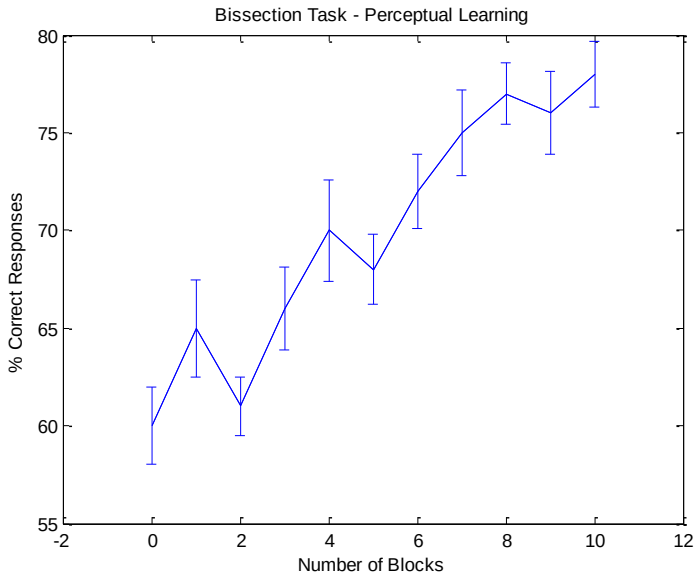
$$y = a + b\,x$$



Example of a cubic polynomial regression, which is a type of linear regression.

Linear fitting: r-square= 0.91

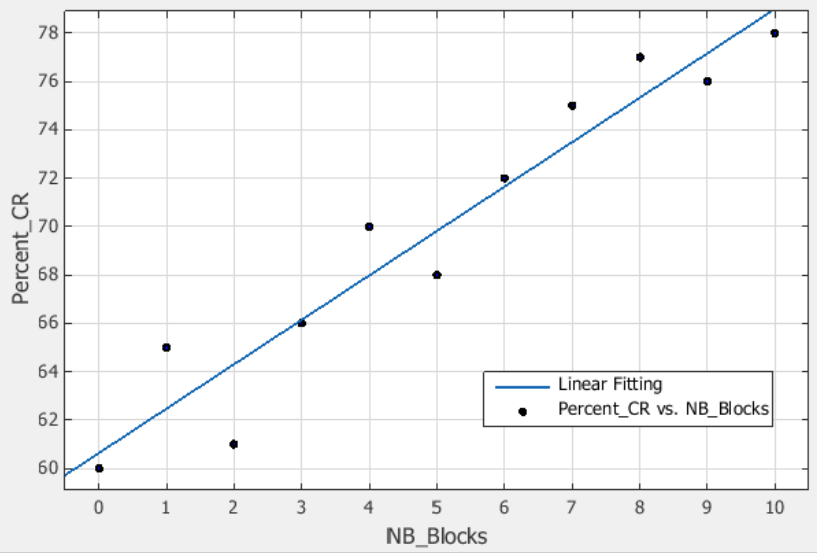
Bisection Task



Results

General model:
f(x) = a*x+b
Coefficients (with 95% confidence bounds):
a = 1.836 (1.425, 2.247)
b = 60.64 (58.2, 63.07)

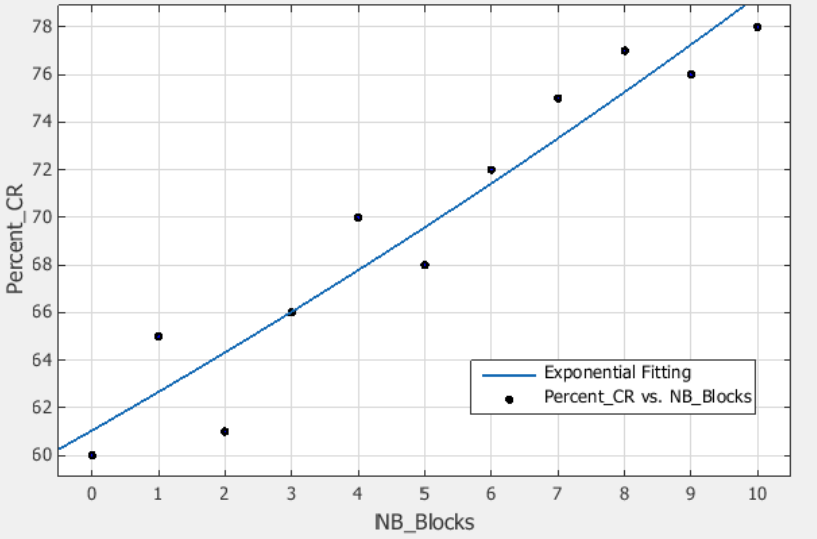
Goodness of fit:
SSE: 32.69
R-square: 0.919
Adjusted R-square: 0.91
RMSE: 1.906



Results

General model Exp1:
f(x) = a*exp(b*x)
Coefficients (with 95% confidence bounds):
a = 61.05 (58.69, 63.4)
b = 0.02617 (0.0201, 0.03225)

Goodness of fit:
SSE: 34.49
R-square: 0.9145
Adjusted R-square: 0.9051
RMSE: 1.958



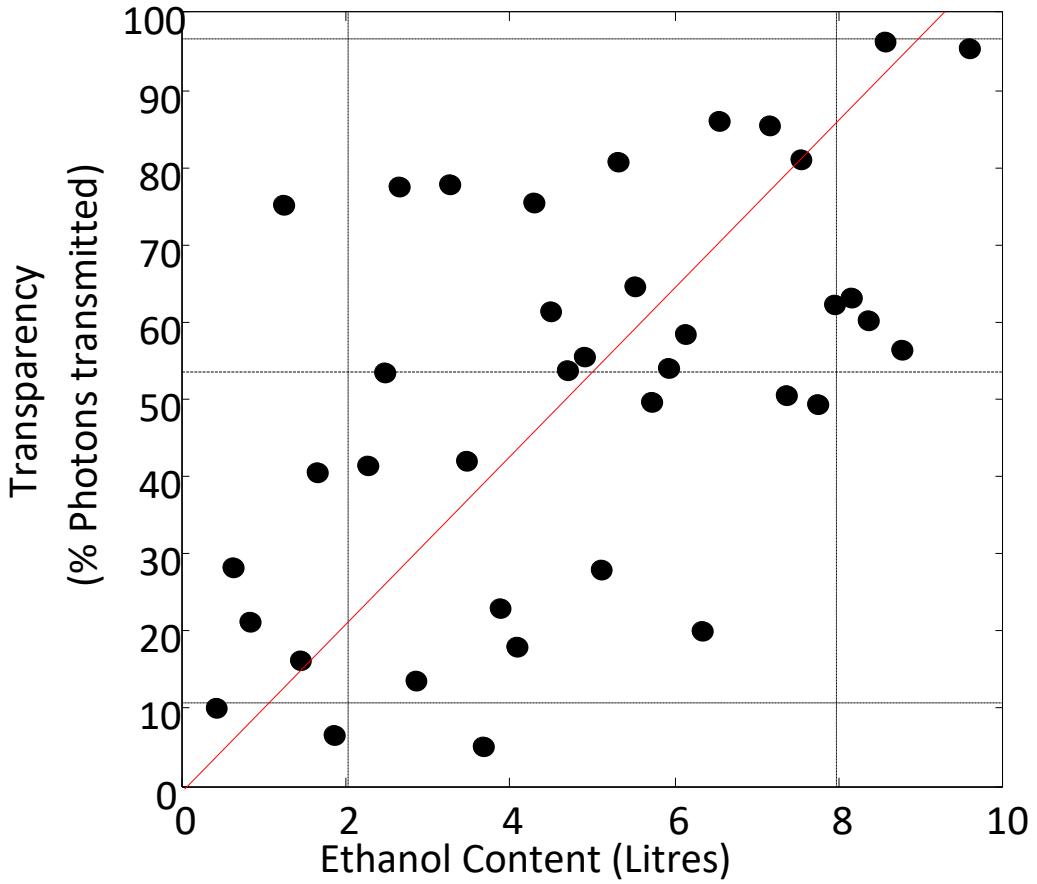
Regression

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n};\ \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

$$s_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

$$s_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}$$

$$m = r\frac{s_y}{s_x};\ b = \bar{y} - m\bar{x}$$



Regression

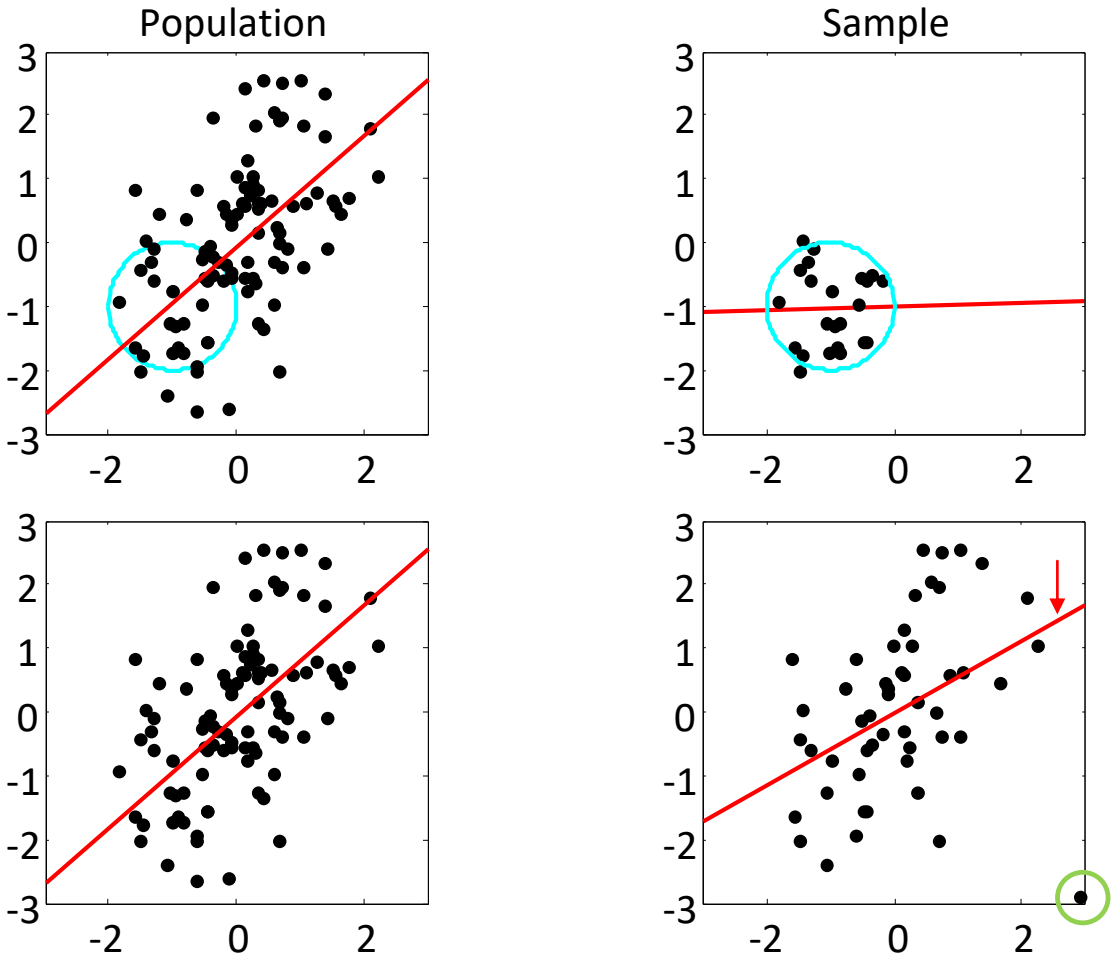
Ethanol	Transparency
1	10
2	40
3	60
4	20
5	70
6	30
7	46.5
8	90
9	70
10	100
$\bar{x} = 5.50$	$\bar{y} = 53.65$
$s_x = 2.87$	$s_y = 28.07$

$$m = r\frac{s_y}{s_x} = 0.75\frac{28.07}{2.87} = 7.33$$

$$b = \bar{y} - m\bar{x} = 53.65 - 7.33 \cdot 5.5 = 13.33$$

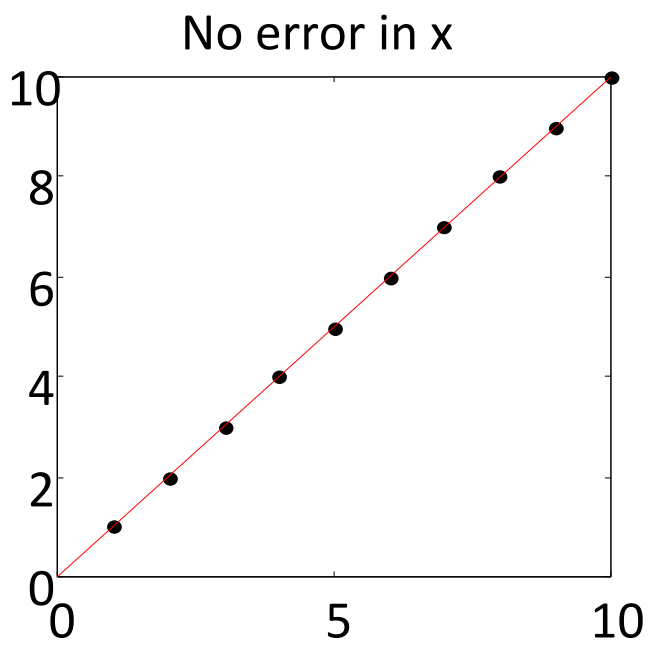
Correlations: assumptions

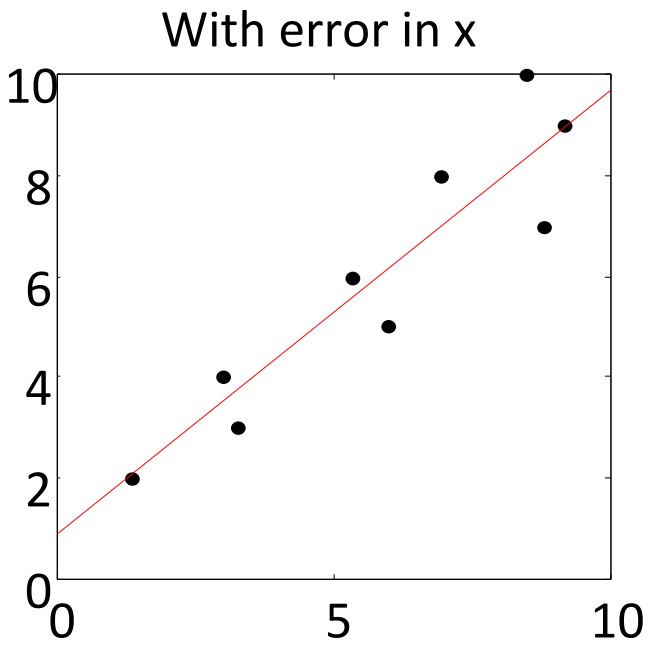
The sample is representative of the population



Correlations: assumptions

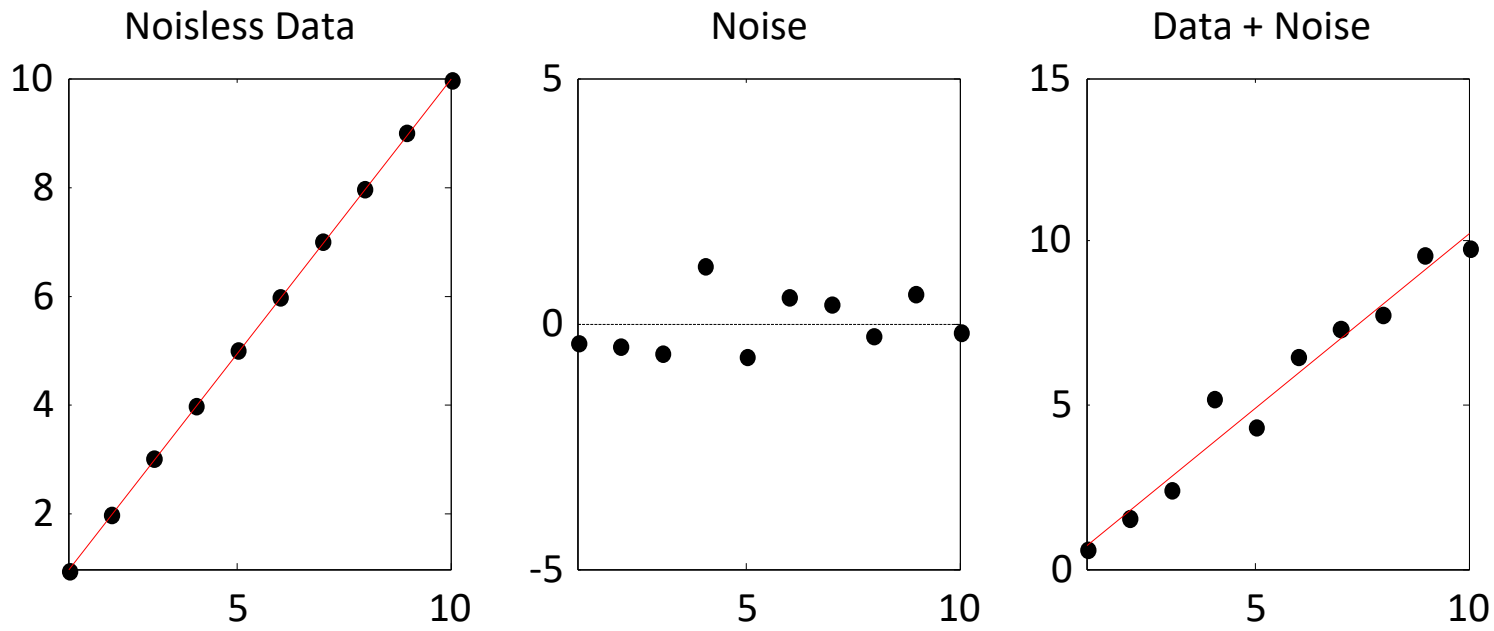
The independent variables are measured with no error





Correlations: assumptions

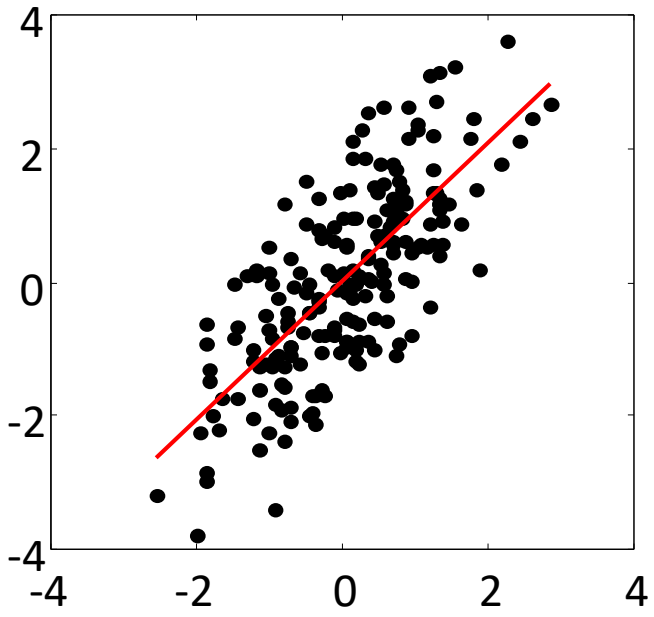
The error, or noise, in the dependent variable has a mean of zero when conditioned on the independent variable(s)



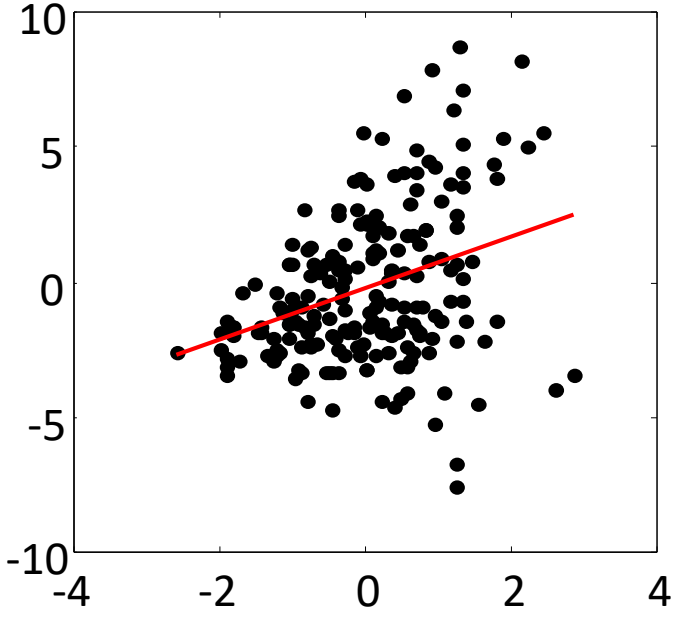
Correlations/regression: assumptions

The variance of the errors is constant across levels of the independent variable

Homoscedastic



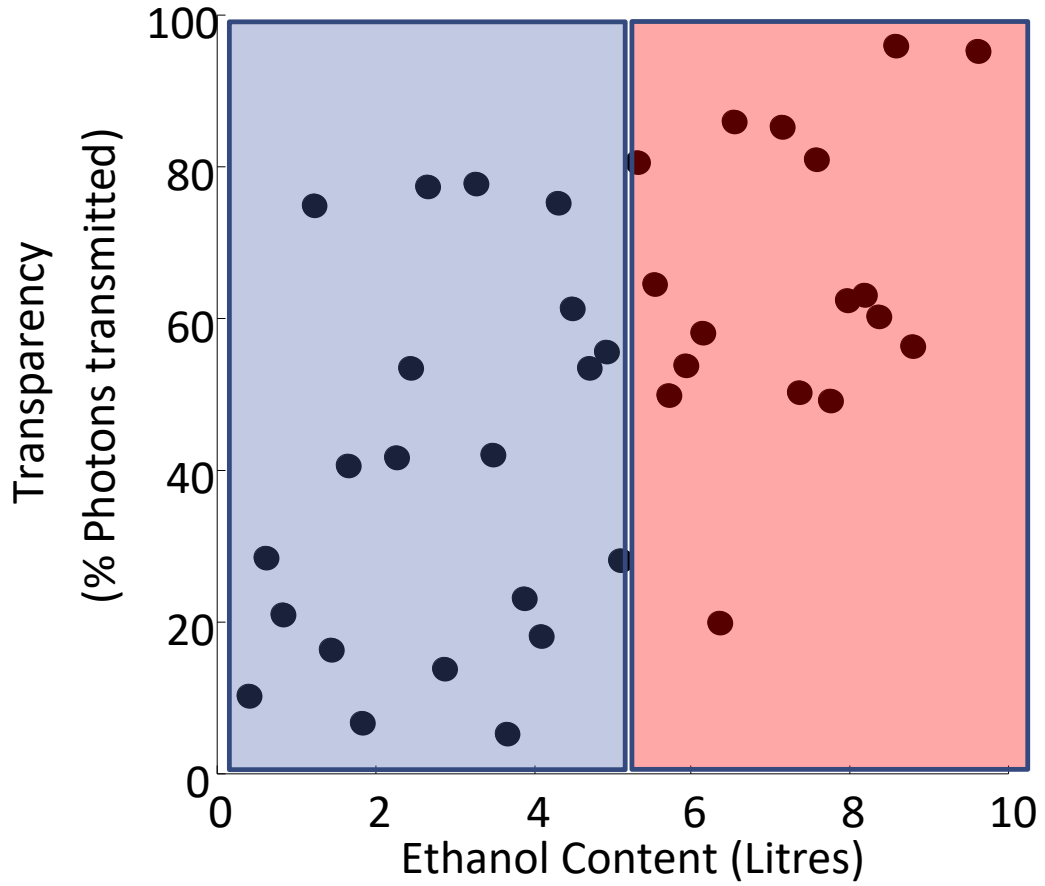
Heteroscedastic



Pearson vs Spearman
Parametric vs. Rank order

Relationship to the t-test

For R(48) = 0.75, p = 0.00003



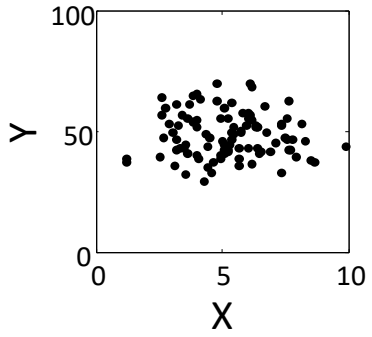
Correlation is not causation

$$Cov(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1} = Cov(y,x)$$

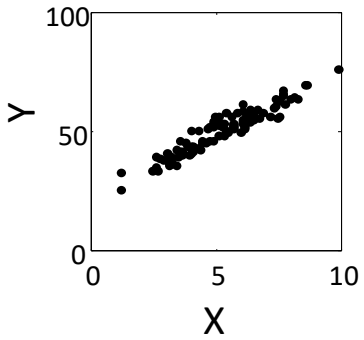
Spurious Correlations & Hidden causes

Covariance & Correlations

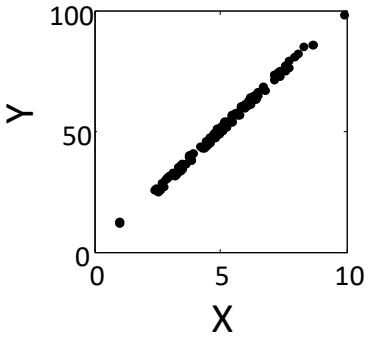
No relationship



Weak positive relationship



Strong positive relationship



$$\mathrm{cov}(X,Y)=\sum_{i=1}^{N}\frac{(x_i-\bar{x})(y_i-\bar{y})}{N-1}$$

$$r=corr(X,Y)=\frac{\mathrm{cov}(X,Y)}{\sigma_x\sigma_y}$$

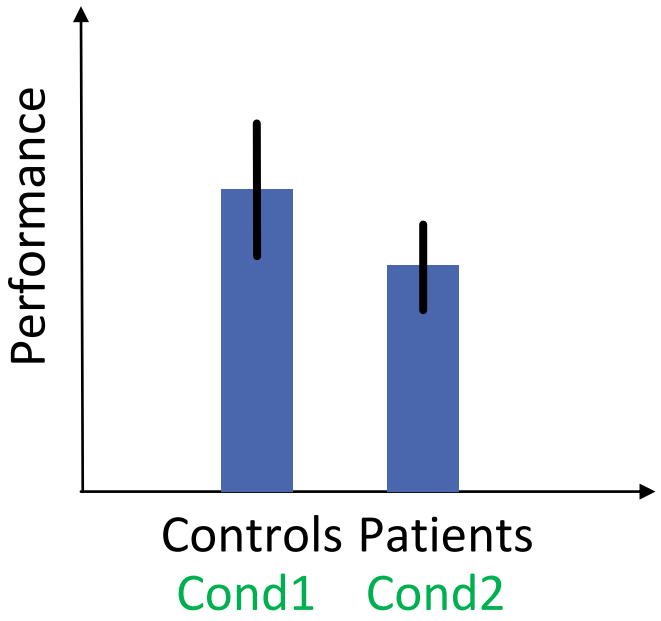
r^2 variance explained

Cohen's rules of thumb for effect size

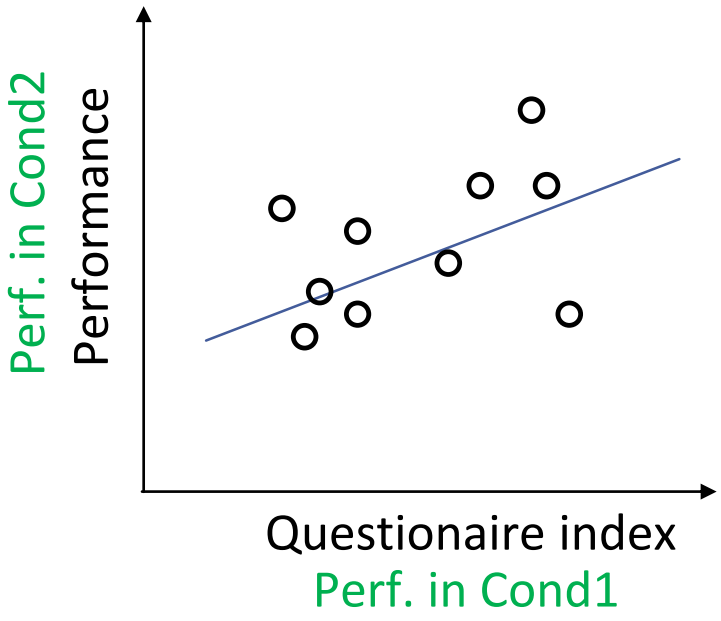
Effect size	Correlation coefficient	Difference between means
“Small effect”	r = 0.1	d = 0.2 standard deviations
“Medium effect”	r = 0.3	d = 0.5 standard deviations
“Large effect”	r = 0.5	d = 0.8 standard deviations

Experimental approach ↔ Correlational approach

Reliability?



Reliability?



Measurement error
Within-subject variance
→ Test/Re-test error

Between-subject variance

Take home messages

Take Home Messages

1. Correlations are the preferred choice if both the x- and y-axis are ratio or interval scaled.
2. Causation and correlation should never be confused.
3. Very different sets of data can lead to the same r.

END Class 8