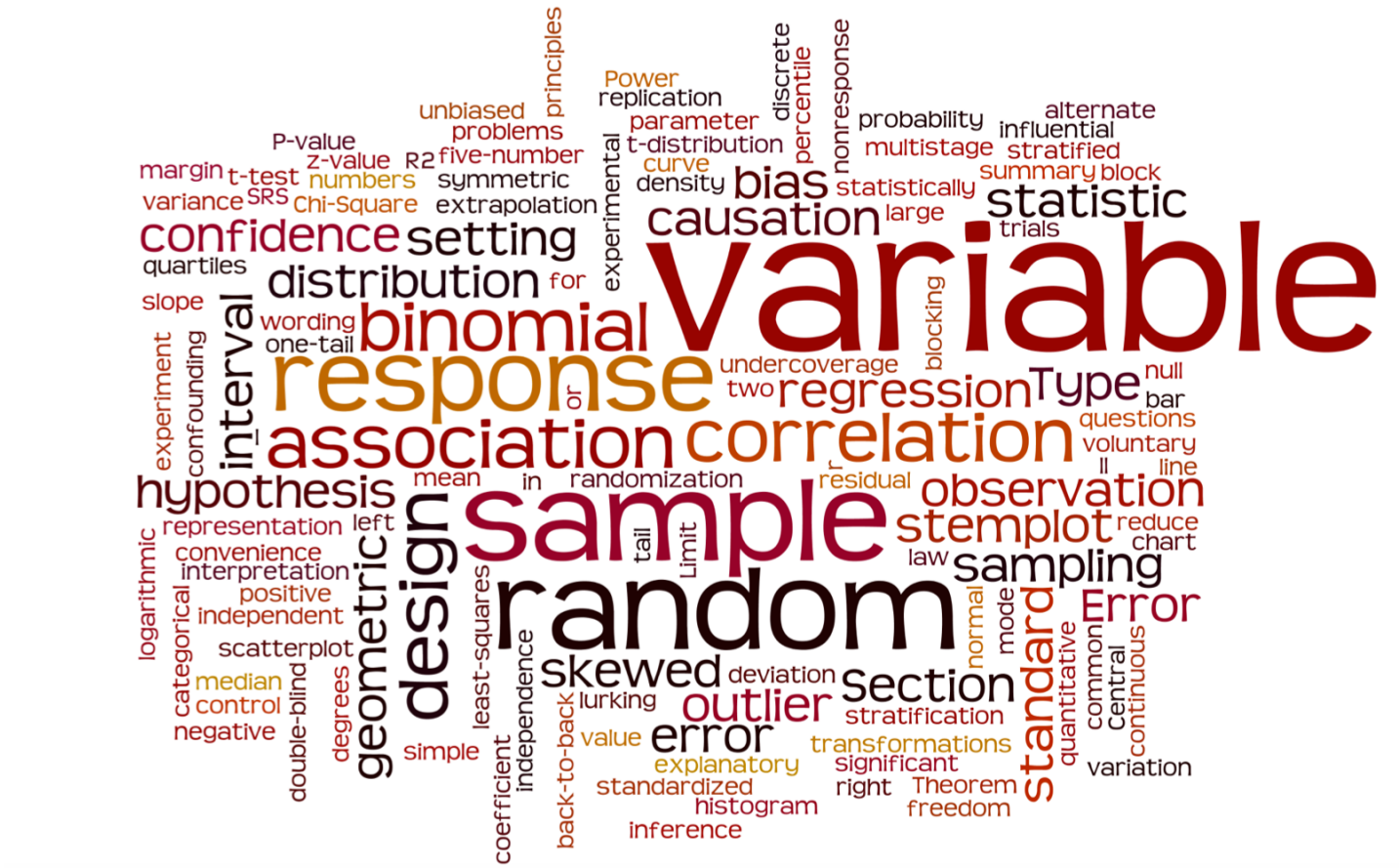
Introduction to Quantitative Methods



Kristin Eccles

Outline

- Related/associated variables
- Continuous variables
  - Scatterplots
  - Covariance
  - Correlation
- Categorical variables
  - Contingency tables and Chi-squared test

Relational statistics

- In statistics we look to:
  - Describe data
  - Compare data
  - Relate data
- Relational statistics:
  - Explore the relationship between variables
  - Are two variables associated with each other?
  - *Technically:* To what extent is the variation in one variable associated with the variation in another variable?

From descriptive to relational statistics

- Variables have variation (dispersion, spread)
  - Range, IQR, variance, standard deviation
  - These are descriptive

- Wondering if tree height is related to latitude?

- This is not comparing the variables
  - To assess relationships need a different graph!

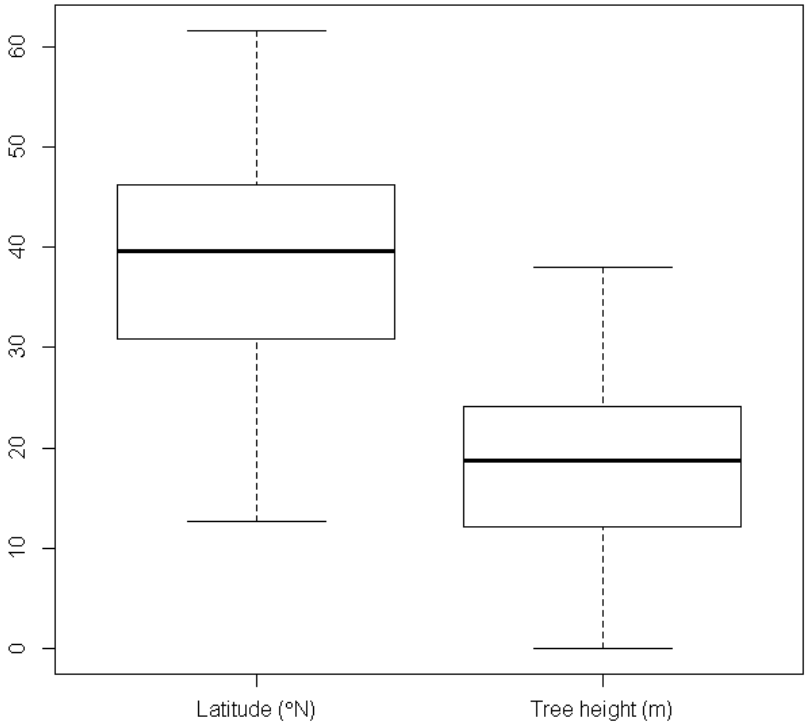
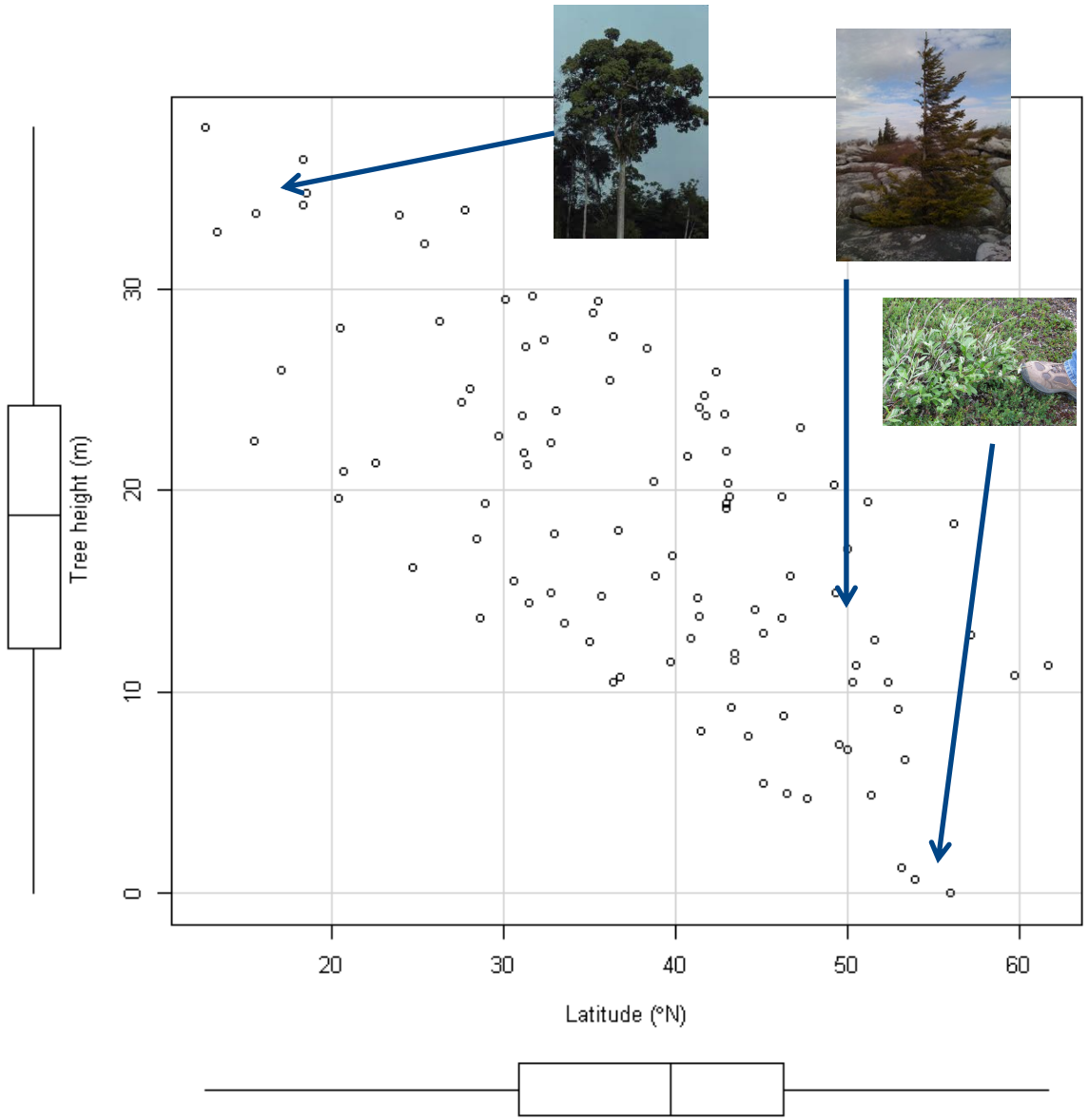


Figure 2. Boxplots of tree height and latitude. Source: GEOG/ENST 2006

Scatterplot

- Use to show relationships between variables
- x axis = horizontal
  - Latitude
- y axis = vertical
  - Tree height
- Look at:
  - Direction
  - Form
  - Strength



- Note: boxplots not typically included

Scatterplot – x '*explains*' y

- By convention: look for *explanatory relationship* between x and y
- This relationship may or may not exist

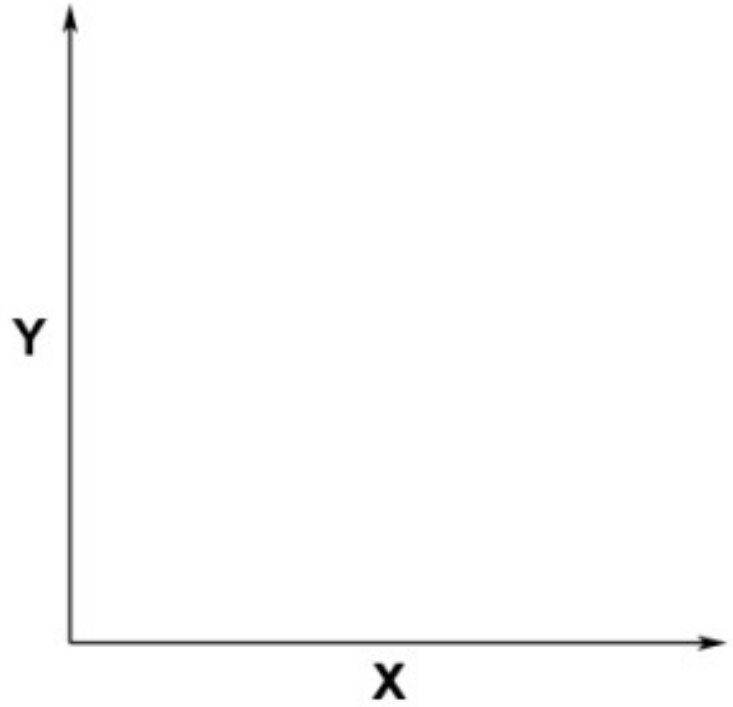
x variable
independent variable
predictor variable
explanatory variable

y variable
dependent variable
response variable
response variable

- Sometimes which variable explains the other is not so apparent (chicken and egg)
- May be a **direct** or an **indirect relationship**

X vs Y

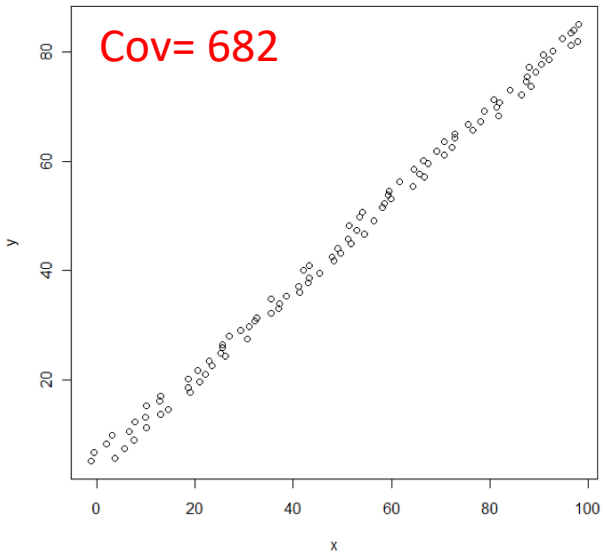
1. Road density and air pollution
2. Fish body mass and mercury concentration
3. Pollution concentration and distance from some stack
4. Housing density and population density
5. Smoking and lung cancer

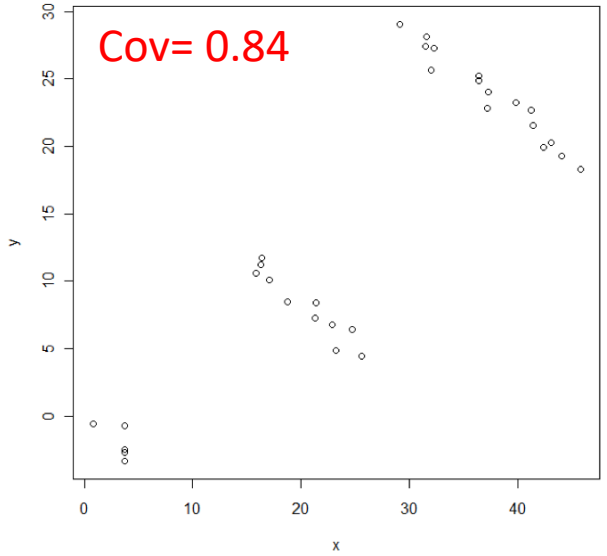


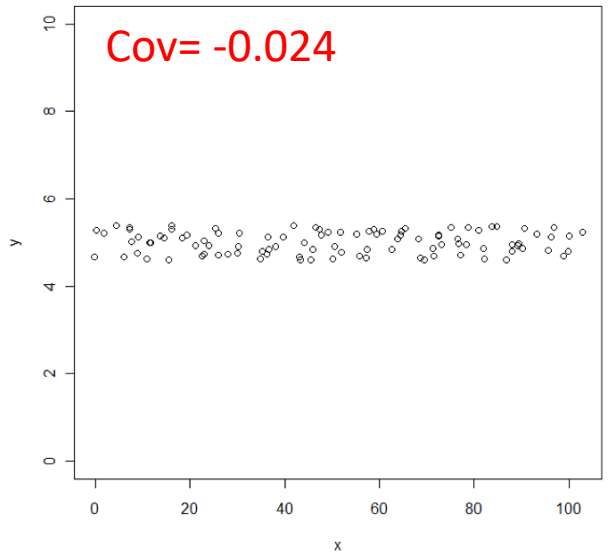
Relationships

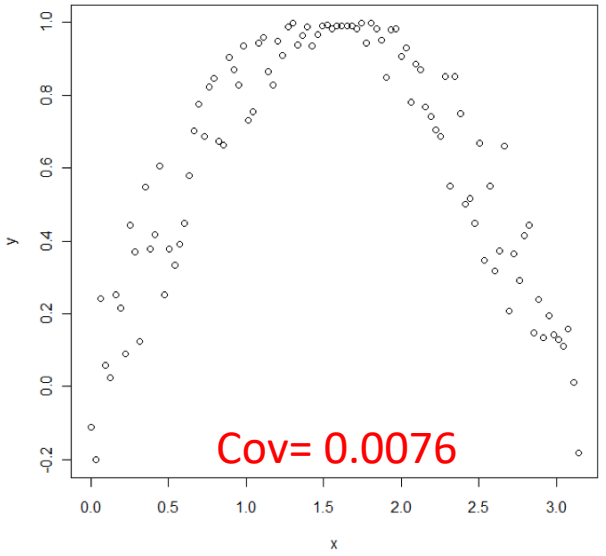
- An association between variables
  - How they are related in a general sense/ overall (not individual pairs)
  - We cant tell much individually
  - There is uncertainty and randomness in real life
- Direction – positive, negative
- Form – linear, curved, other
- Strength – is there a very tightly-constrained association or are the data 'noisy'

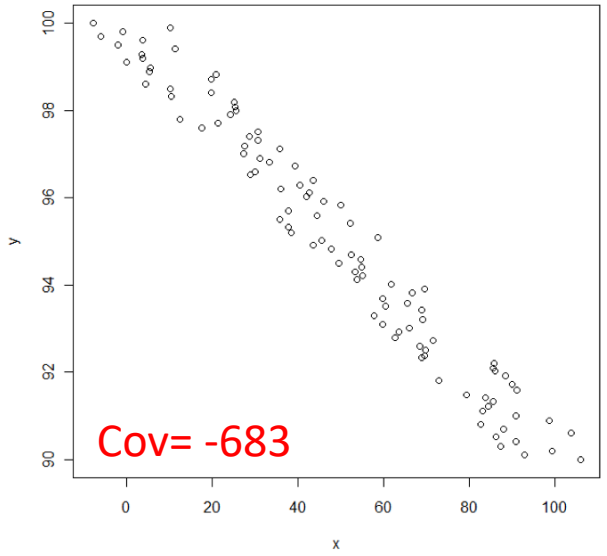
Relationship: Direction, form, strength?

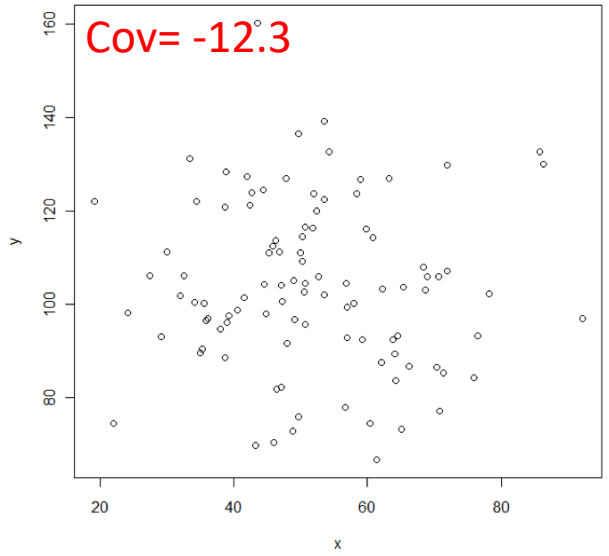












Covariance

- How much to the variables co-vary?
- The higher the covariance, the more this is the case
- Covariance changes with the magnitude of x and y
  - Just like variance
- Need a way to standardize this measure of association

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

Comparing Deviation, Variance, Covariance

Deviation: $$d_i = x_i -$$

Variance: $$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Covariance: $$cov(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{n-1}$$

Covariance Example

	x	y			
ID	lat	tree	x-xbar	y-ybar	x-xbar*y-ybar
1	36.2547644	23.6698751	-2.1062786	10.2742943	-21.640526
2	28.2633464	10.9598841	-10.097697	-2.4356967	24.5949267
3	39.514523	4.77753775	1.15348001	-8.6180431	-9.9407405
4	28.4476717	10.8928541	-9.9133712	-2.5027267	24.8104592
5	47.5079317	17.3147954	9.14688876	3.91921456	35.8486196
6	42.7278257	15.9834944	4.36678272	2.58791351	11.300856
7	45.6674563	17.8544669	7.30641334	4.458886	32.5784642
8	29.4170507	16.7071343	-8.9439923	3.31155346	-29.618509
	….	….			
Mean	38.361043	13.3955809		Sum	-3615.0495
N	100			Sum/n-1	-36.515652

Pearson's correlation coefficient

- AKA Pearson's r or Pearson's product moment correlation coefficient
- Measures the amount of correlation between 2 variables
- Large r (- or +) means a strong linear association
- r is a number from -1 to 1
  - -1: perfect negative correlation
  - 0: no correlation
  - 1: perfect positive correlation
- Formula: same as the covariance, but use z-scores for x and y

$$r = \frac{\sum_{i=1}^{n} z_{xi} \cdot z_{yi}}{n-1}$$

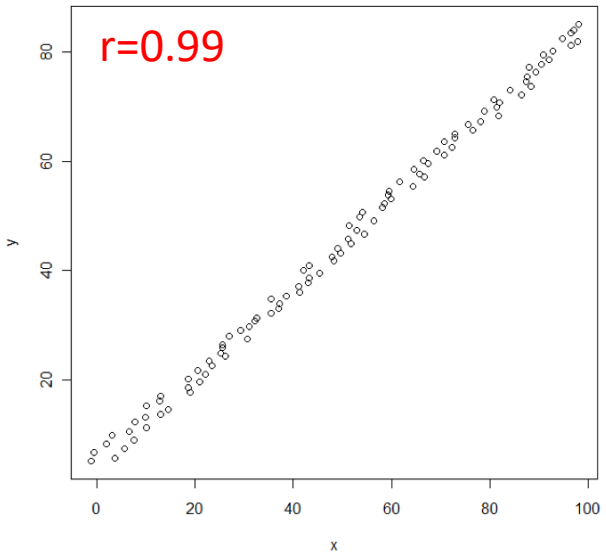
$$z = \frac{X - \bar{X}}{s}$$

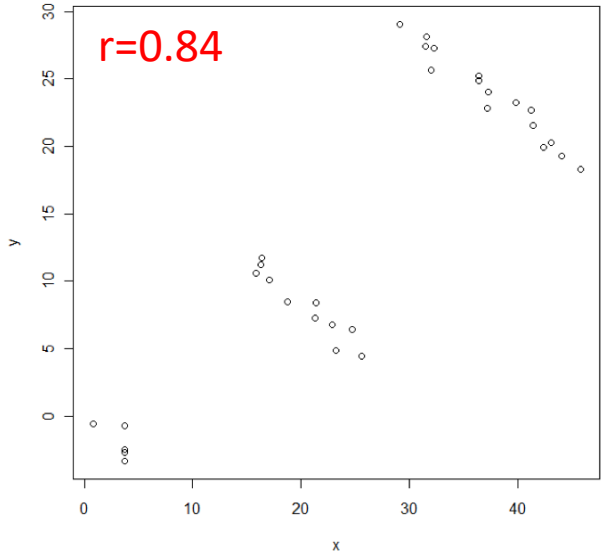
- Note that correlation is only one type of association
  - it is possible to have r=0 and yet have a strong relationship

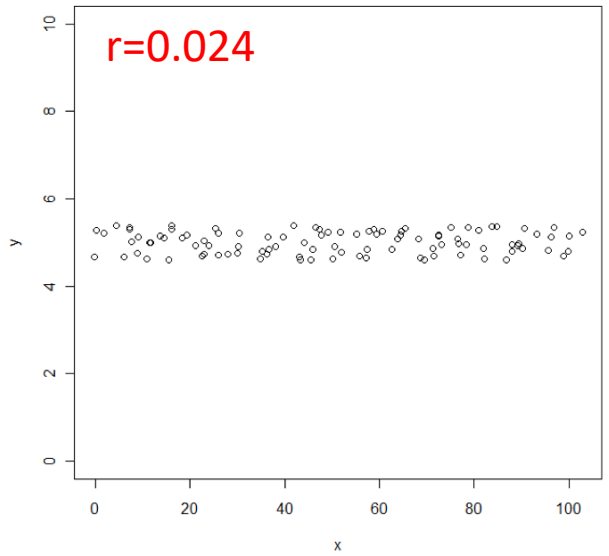
Calculating R

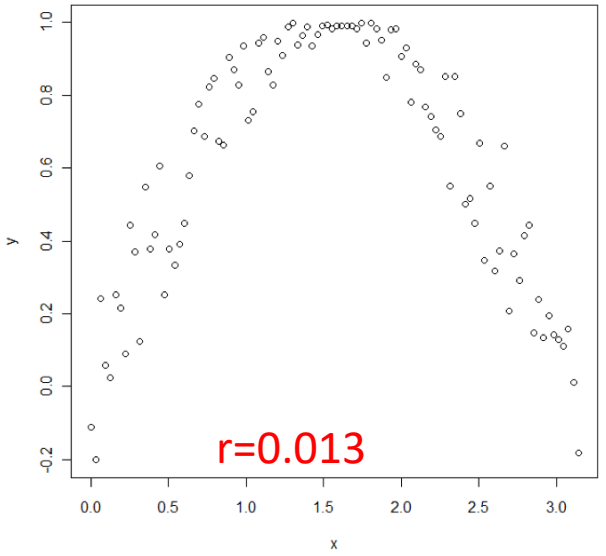
	x	y			
ID	lat	tree	z-zscore	y-zscore	z-zscore*y-zscore
1	36.2547644	23.6698751	-0.19276	1.55980914	-0.3006688
2	28.2633464	10.9598841	-0.9241095	-0.3697794	0.34171664
3	39.514523	4.77753775	0.10556288	-1.3083626	-0.1381145
4	28.4476717	10.8928541	-0.9072407	-0.3799556	0.34471121
5	47.5079317	17.3147954	0.83709459	0.5950021	0.49807304
6	42.7278257	15.9834944	0.39963427	0.39288841	0.15701167
7	45.6674563	17.8544669	0.66866005	0.67693322	0.4526382
8	29.4170507	16.7071343	-0.8185261	0.50274902	-0.4115132
9	47.4175746	12.930071	0.82882539	-0.0706722	-0.0585749
10	36.439535	18.9610242	-0.1758504	0.84492708	-0.1485808
	…	…			
Mean	38.361043	13.3955809		Sum	-50.226724
Standard Deviation	10.9269476	6.58689196		Sum/n-1	-0.5073406
N=100					

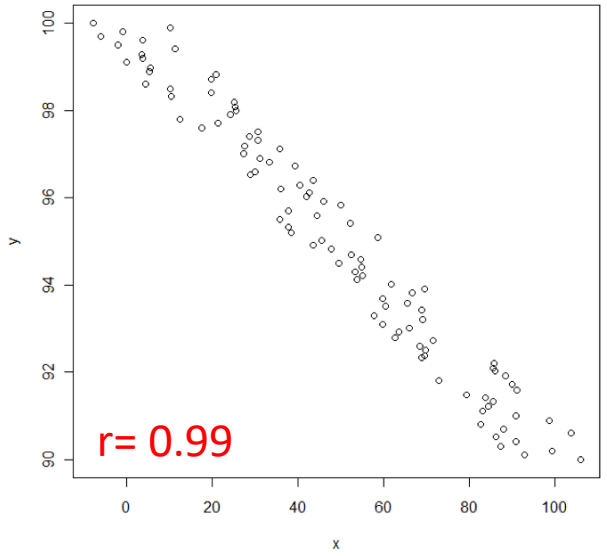
Correlations? r=?

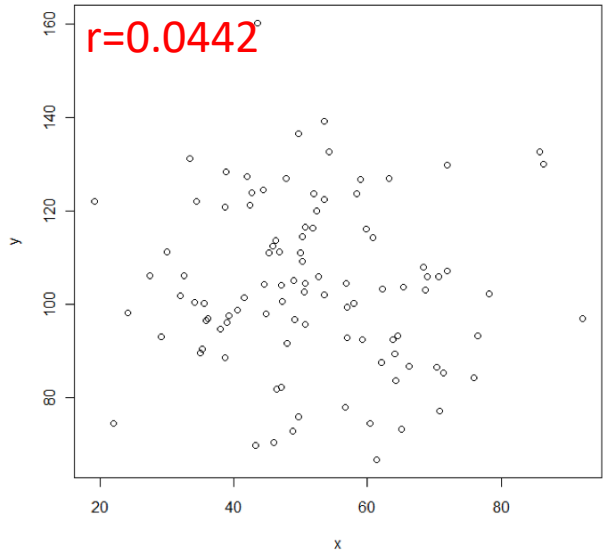












Interpreting r

- 0.00-0.19: very weak
- 0.20-0.39: weak
- 0.40-0.59: moderate
- 0.60-0.79: strong
- 0.80-1.0: very strong

Is the correlation significant?

- Say your sample has an r of 0.2
- Could this be due to random chance or is there an underlying correlation in the population?
- A *t*-test will reveal if the population correlation coefficient ρ (Greek letter rho) is different from 0 (no correlation)
- H_0: ρ = 0
- H_1: ρ ≠ 0
- Can also do a one-tailed test: H_1: ρ<0 or H_1: ρ>0
- Test statistic is *t,* calculated as follows:

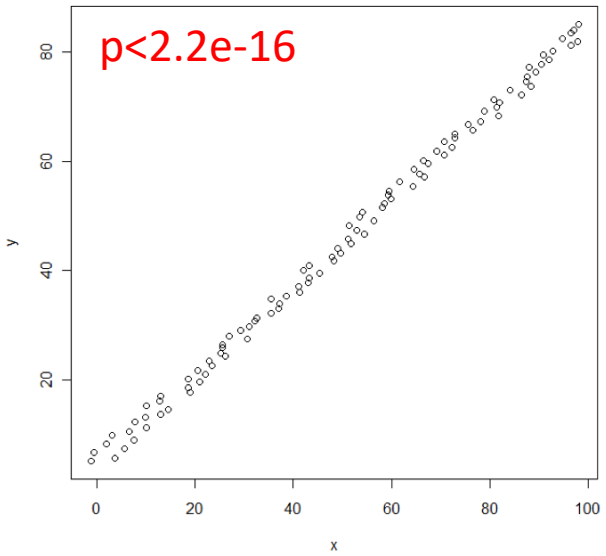
$$t = \frac{r}{\sqrt{\frac{1 - r^2}{n - 2}}}$$

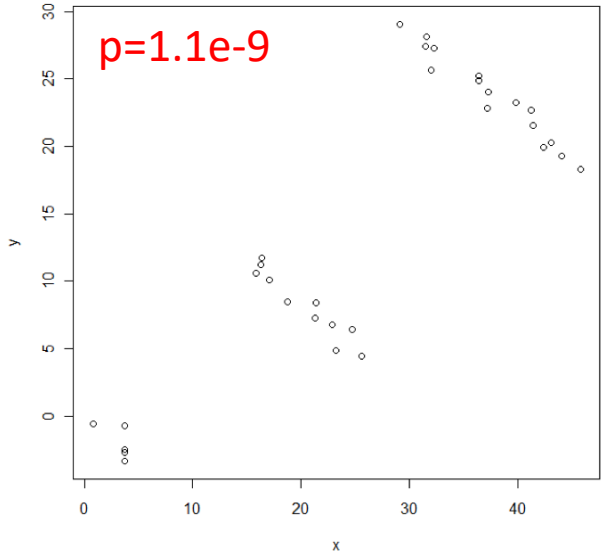
Ex. $$t = \left(\frac{-0.507}{\sqrt{\frac{1 - (-0.507)^2}{100 - 2}}} \right)$$

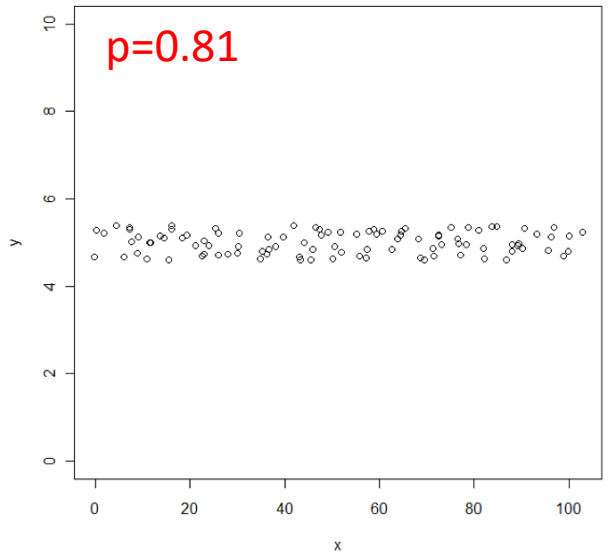
t=-5.822921

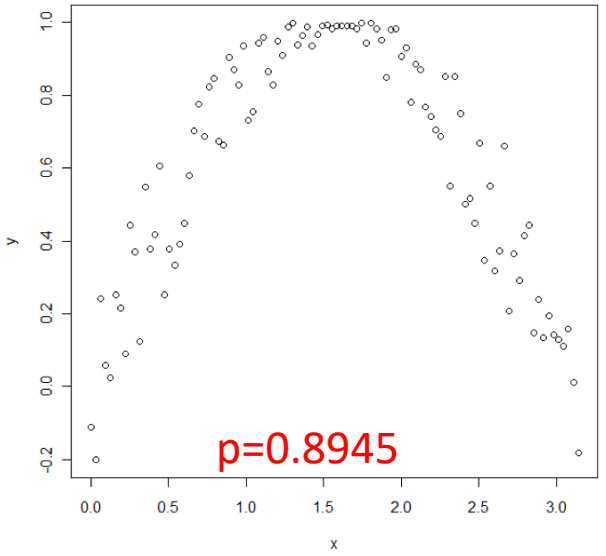
Check your answers: www.mathpapa.com/algebra-calculator.html

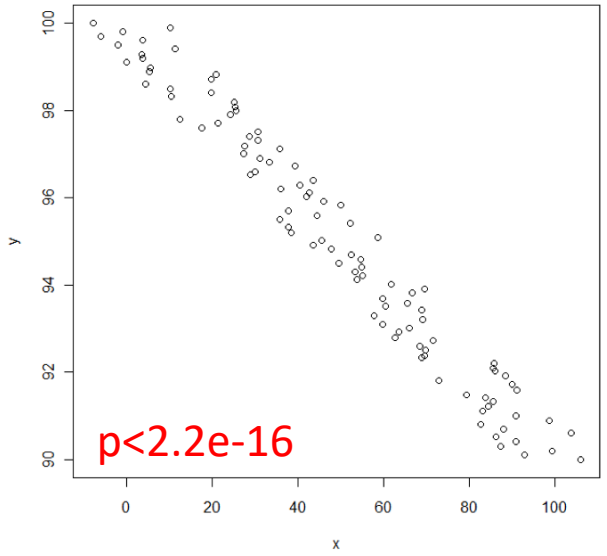
Correlations? p-value=?

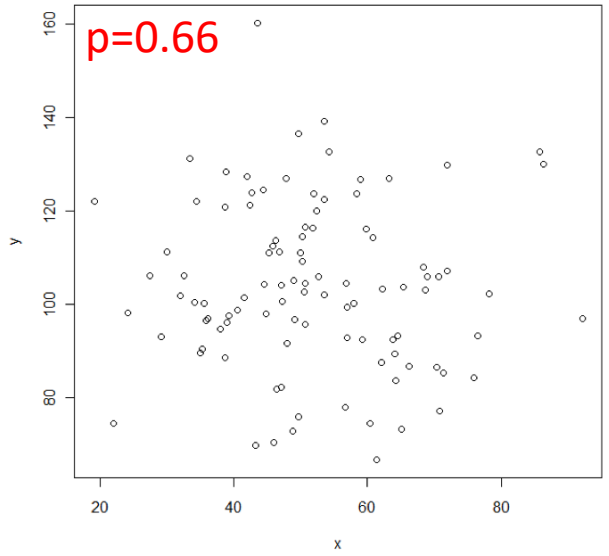












R Output

```
        Pearson's product-moment correlation

data:  lat and tree
t = -5.8282, df = 98, p-value = 7.167e-08
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.6399838 -0.3453355
sample estimates:
       cor
-0.5073406
```

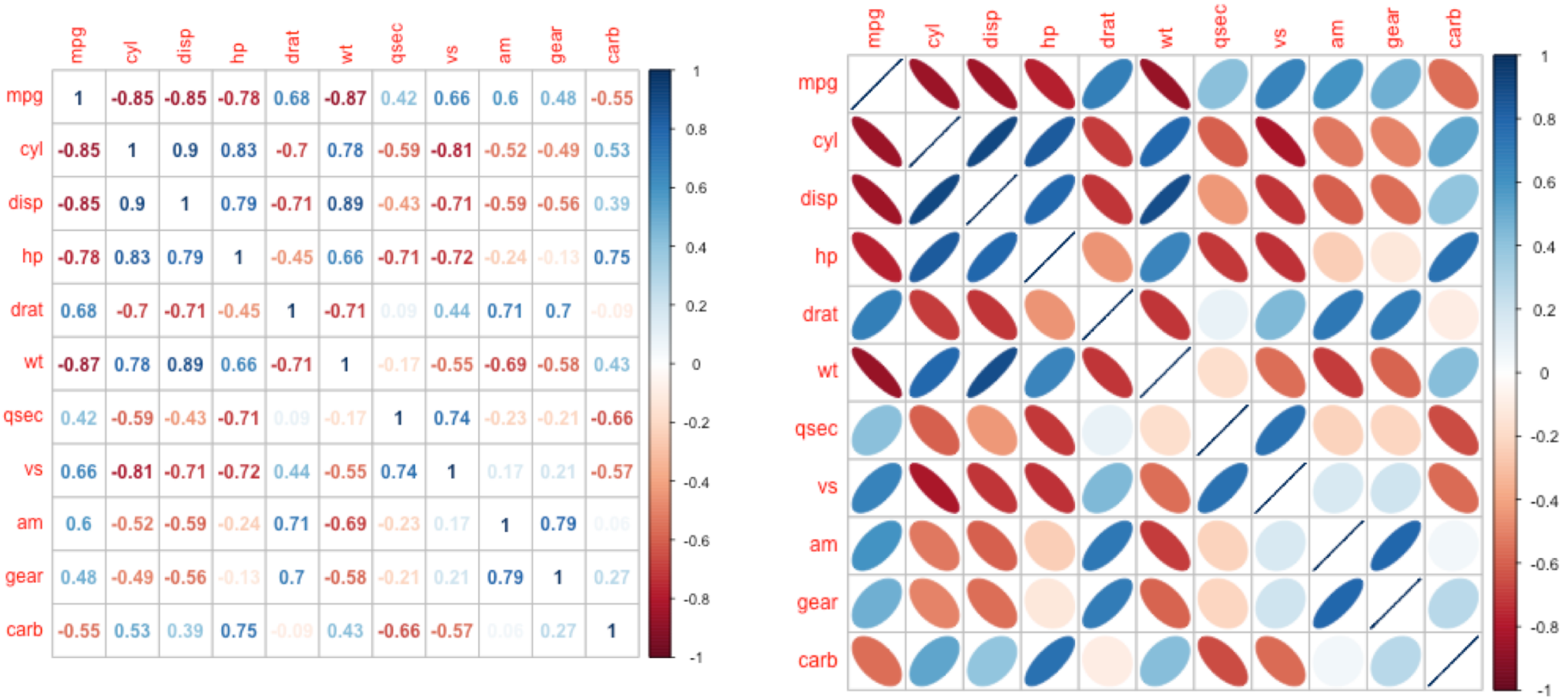
Correlation Assumptions

- Assumptions:
  - Random/probabilistic sample of paired variables
    - Paired by subject, site, location
  - Variables at interval/ratio scale
  - Variables have a linear association
    - As x increases 1 unit so will y
  - Outliers that don't unduly influence the relationship
  - Variables should be normally distributed

Examples- Correlation Matrix

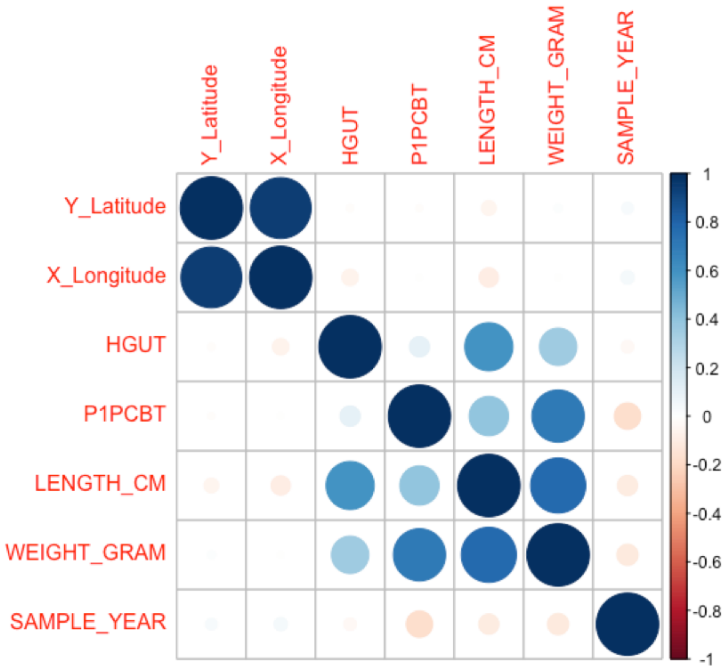
library(corrplot)

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
mpg	1	-0.85	-0.85	-0.78	0.68	-0.87	0.42	0.66	0.6	0.48	-0.55
cyl	-0.85	1	0.9	0.83	-0.7	0.78	-0.59	-0.81	-0.52	-0.49	0.53
disp	-0.85	0.9	1	0.79	-0.71	0.89	-0.43	-0.71	-0.59	-0.56	0.39
hp	-0.78	0.83	0.79	1	-0.45	0.66	-0.71	-0.72	-0.24	-0.13	0.75
drat	0.68	-0.7	-0.71	-0.45	1	-0.71	0.09	0.44	0.71	0.7	-0.09
wt	-0.87	0.78	0.89	0.66	-0.71	1	-0.17	-0.55	-0.69	-0.58	0.43
qsec	0.42	-0.59	-0.43	-0.71	0.09	-0.17	1	0.74	-0.23	-0.21	-0.66
vs	0.66	-0.81	-0.71	-0.72	0.44	-0.55	0.74	1	0.17	0.21	-0.57
am	0.6	-0.52	-0.59	-0.24	0.71	-0.69	-0.23	0.17	1	0.79	0.06
gear	0.48	-0.49	-0.56	-0.13	0.7	-0.58	-0.21	0.21	0.79	1	0.27
carb	-0.55	0.53	0.39	0.75	-0.09	0.43	-0.66	-0.57	0.06	0.27	1



https://cran.r-project.org/web/packages/corrplot/vignettes/corrplot-intro.html

Examples




```
        Pearson's product-moment correlation

data:  HGUT and P1PCBT
t = 4.2216, df = 986, p-value = 2.65e-05
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.07147045 0.19400147
sample estimates:
      cor
0.1332451
```

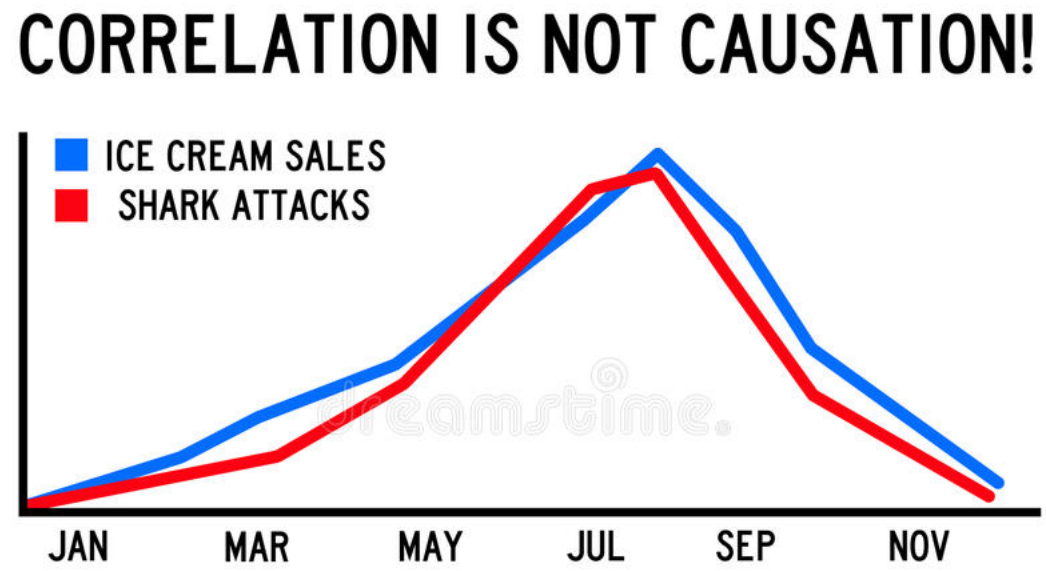
```
        Pearson's product-moment correlation

data:  HGUT and SAMPLE_YEAR
t = -9.9563, df = 2331, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.2405775 -0.1627193
sample estimates:
       cor
-0.2019675
```

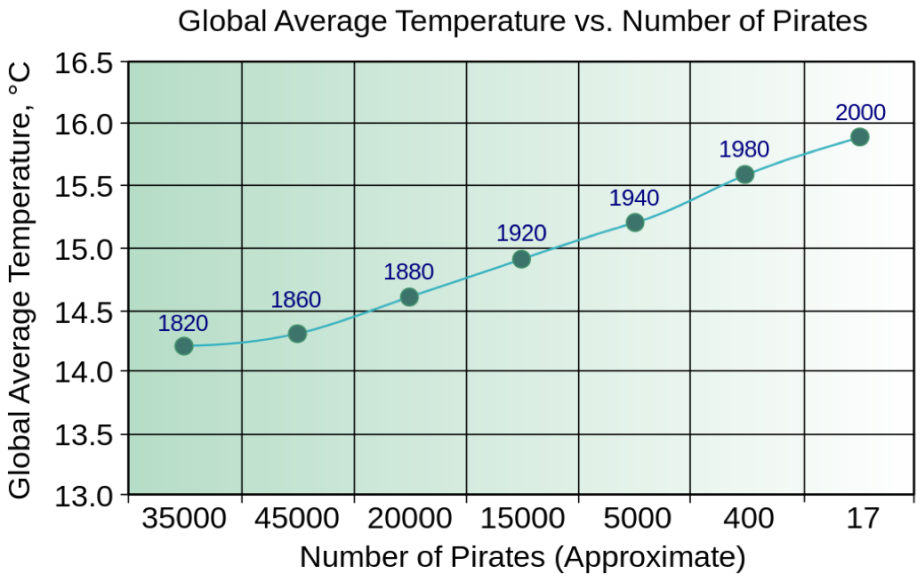
```
        Pearson's product-moment correlation

data:  HGUT and WEIGHT_GRAM
t = 21.032, df = 2047, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3852678 0.4565090
sample estimates:
      cor
0.4215386
```

Warning: Correlation vs. Causation



Both ice cream sales and shark attacks increase when the weather is hot and sunny, but they are not caused by each other (they are caused by good weather, with lots of people at the beach, both eating ice cream and having a swim in the sea)



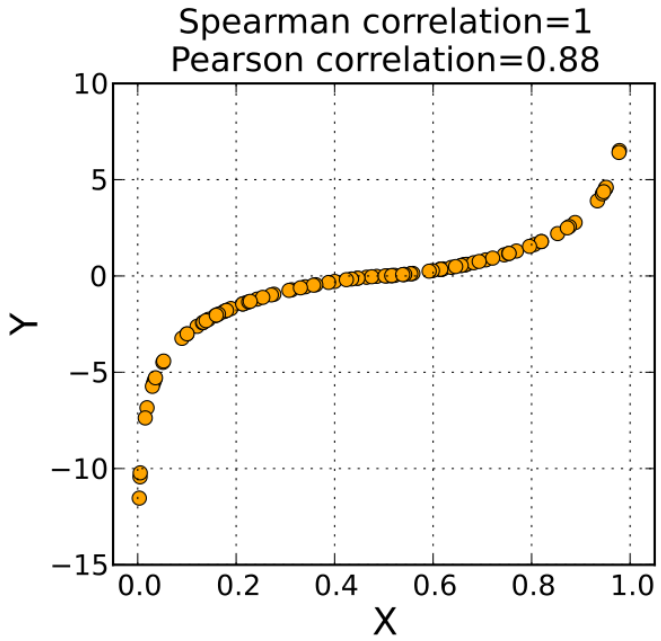
Non-Parametric Correlations

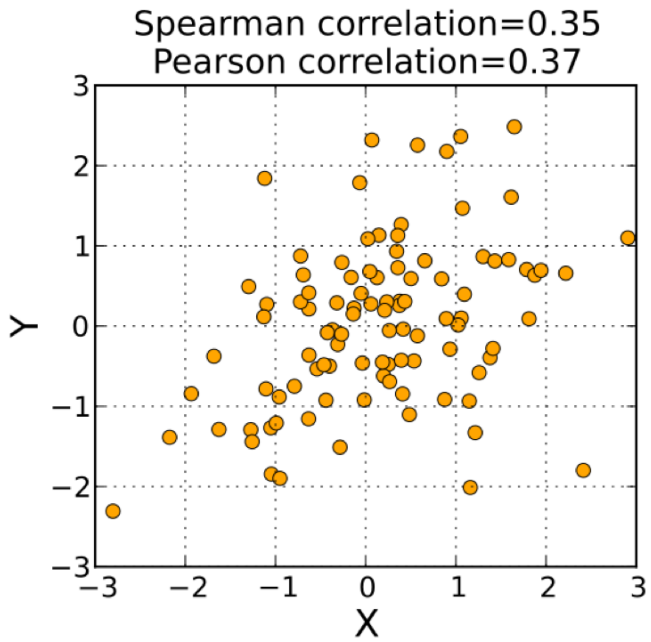
- When you violate the assumptions of the parametric test you must look to non-parametric options
  - i.e. **not** normally distributed or a non-linear association
  - Pearson's r is not appropriate, since it is a parametric test
- Spearman's rank correlation
- Contingency tables

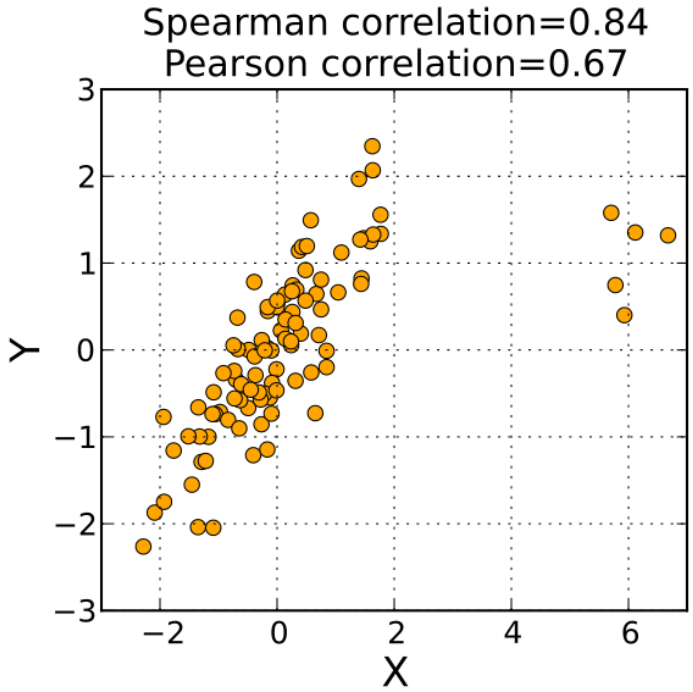
Spearman's rank correlation

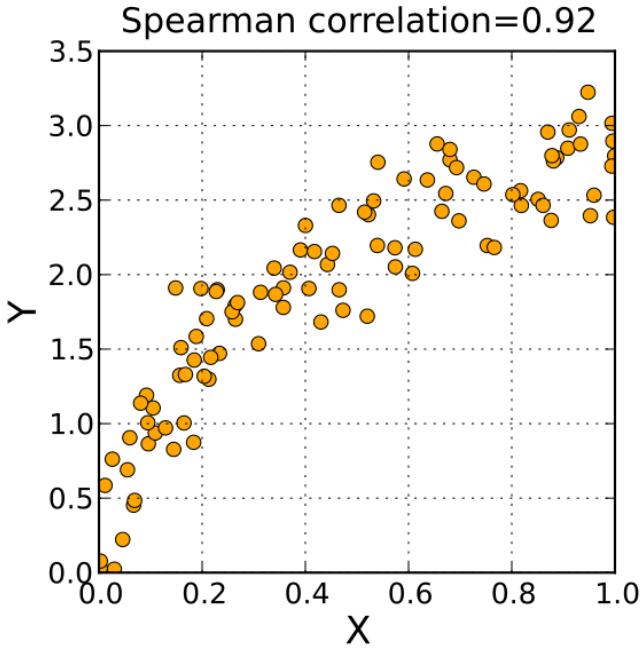
- A non-parametric alternative is: Spearman's rank correlation (r_s)
  - Actually tests for a **monotonic** relationship
  - r_s from -1 to +1 (0 is no correlation)

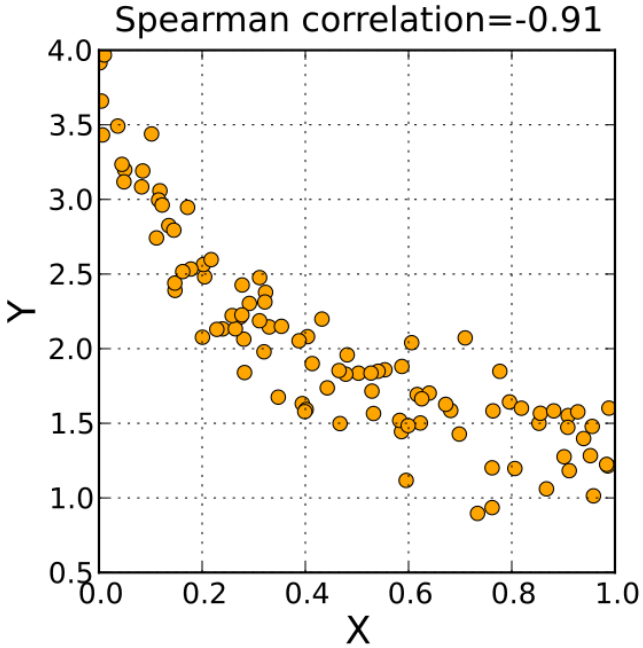
  - Works on ordinal data, and ranked non-normal interval/ratio data
  - Looks at the difference in ranks for x and y
  - The significance of r_s can be tested as well











Source: https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient

Contingency table

- Say you are interested in women in BSc. Science and Engineering
- 2 nominal (or ordinal variables)
- Frequency counts go in a **contingency table**

	Life Sci.	Eng. & Comp Sci.	Math & Phys. Sci.	**Total**
Males	20	47	11	78
Females	30	10	7	47
Total	50	57	18	**125**

Row totals

Grand total

Column totals

Column and row totals often called the **marginal totals**

Source: http://www.nserc-crsng.gc.ca/_doc/Reports-Rapports/Women_Science_Engineering_e.pdf

What do you expect?

What would you expect if there was no relationship between prof type and answer?

Expected value: *E* = (row total x column total) ÷ grand total

	Life Sci.	Eng. & Comp Sci.	Math & Phys. Sci.	**Total**
Males	20 (31.2)	47 (35.6)	11 (11.2)	78
Females	30 (18.8)	10 (21.4)	7 (6.8)	47
Total	50	57	18	**125**

Expected value in parentheses

Formula:

Where E_{ij} is the expected value at the

i^{th} row and j^{th} column,

R is row total, *C* is column total and *N* is grand total

$$E_{ij} = \frac{R_i \times C_j}{N}$$

Compare observed and expected

- So compare what you *observe* vs. what you *expect* in each cell of the table
- If the difference is great, then you may have a significant association between variables
- **Observed** minus **Expected**

$$\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

- For example in the first cell: 20 (31.2)
- $(20 - 31.2)^2 / 31.2 = 4.02$

Contingency analysis

- Is there an association between the variables?
- H_0: Variables are statistically independent (there is no relationship/association between them)
- H_1: Variables are statistically dependent (there is a relationship/association between them)

- Test statistic: χ^2 'chi-squared'
- Degrees of freedom = (*r*-1) x (*c*-1)
  - Why do we care about this?
  - Its not in the equation!

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

- Basically as you add up the values from the previous slide in all the cells to get χ^2
- Note: Greek letter chi; say: *k-eye*

Review: χ^2-distribution

- As observed-expected increases, χ^2 gets larger and more improbable
- Find the probability of getting a larger χ^2
- In our example:
  - $X^2 = 20.5$
  - df = 2
  - *p*-value = 3.6e-05

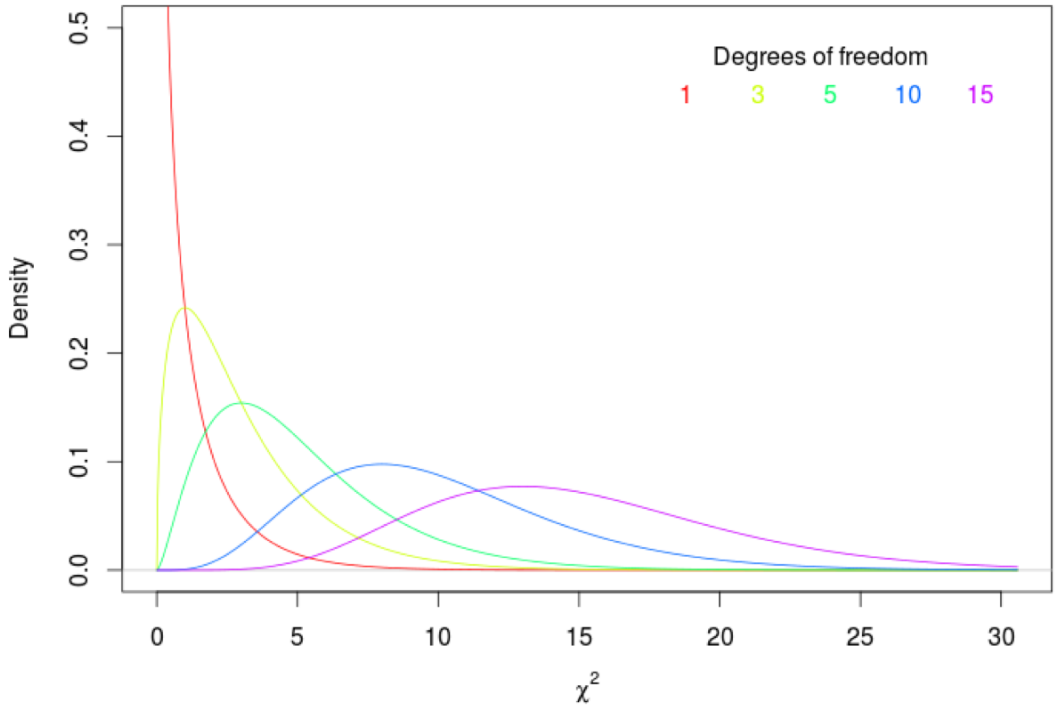


Figure 1: Chi squared distribution with 1,3,5,10 and 15 degrees of freedom

Contingency analysis assumptions

- Test assumptions:
  - Categorical variables (nominal or ordinal) – enter frequency counts into the contingency table
  - Single random/probabilistic sample
  - No more than 20% of expected values should be <5
  - No expected values should be <2

Summary

- Identifying associations/relationships (even indirect ones) helps to understand how the world works
- Can be done with categorical data (contingency analysis/Spearman Rank) and with continuous data (correlation)
- Beware the pitfalls:
  - always look at the scatterplot
  - association vs. correlation (not the same)
  - Correlation vs. causation
  - Pay attention to the assumptions!!
- Test for significance, especially when r or rs are close to zero
- **Other Resources:**
  - Correlation: https://www.khanacademy.org/math/probability/scatterplots-a1/creating-interpreting-scatterplots/e/correlation-coefficient-intuition