Lecture 8

Correlation
Do taller people earn more?

Dr. Franco Ubaudi

The Nature of Data
Western Sydney University

Spring 2021

Outline

- Difference between observation and experimentation
- Difference between correlation and causation
- Do taller people earn more?
- Correlation defined
- Correlation and dependence
- Correlation and covariance
- Correlation: Pearson / Spearman
- The same old stuff:
  - hypothesis testing and confidence intervals
  - but for correlation

Observation

In an observational study, we are just **observing**.

We are **not** trying to control any variables.

We are **not** trying to exert any influence.

We might be able to determine what variables are being observed, but if retrospective, we probably get what we get.

How do we know which variable(s) determine an observed result? We don't!

Experimentation

In an experiment, we are trying to figure out what variables cause a result of interest.

e.g. whether a new drug cures a disease better than an old drug.

Ideally wish to control as many variables as possible.

e.g. age, gender, pre-existing health issues, genetics.

Goal is to be confident if the difference is only due to the drug!

Correlation versus causation

Only an experiment can tell us if a variable causes the result of interest, since only an experiment controls variables.

In an observational study, we just see patterns between variables.

A pattern or a correlation, does not **imply** a causation.

Does X_7 cause Y?

$Y = f(X_1, X_2, X_3, \ldots X_7, X_8, \ldots)$

Even if it does, is it alone?

Do taller people earn more?

Previously, we looked at the relationship between two variables:

- Between eel species and habitat.
- Between smoking status and birth weight.

In this lecture we are interested in height vs. income.

The data we have is observational and retrospective

$\Longrightarrow$ we are looking for correlation *not* causation.

Do taller people earn more?

An oft-posed question: For example, Steckel, Richard H. 1995. "Stature and the Standard of Living." *Journal of Economic Literature*.

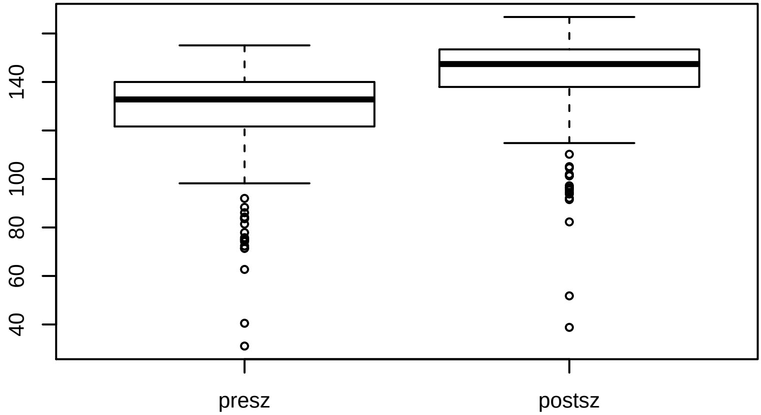
We have a data set of 1376 observations of height and income.

Let's look at a simpler data set first.

Moulting Crabs

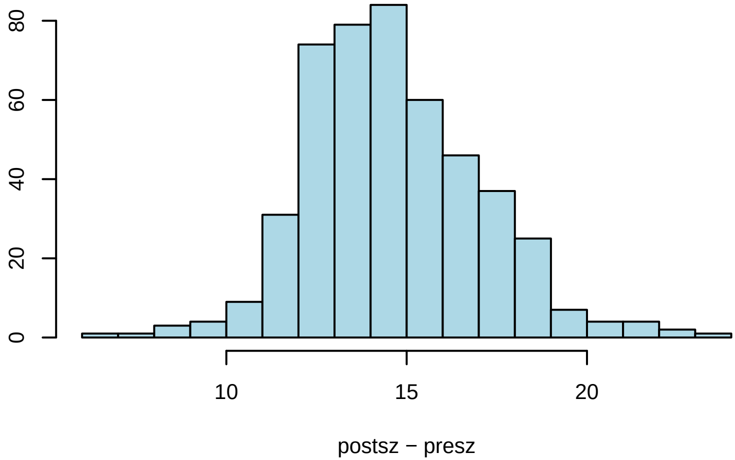
Measurement of crabs before and after moulting from 472 female Dungeness crabs.

Pre-moult size (`presz`) and post-moult size (`postsz`).



Moulting Crabs

The difference in size



Moulting Crabs

Histogram shows that on average crabs grow about 15mm when they moult.

But do small crabs grow more or less than large crabs?

Is the post-size just the pre-size plus a constant, or perhaps, a more complex relationship?

For this we need another type of plot.

Scatter plots

- Boxplots display only each variable seperately.
- The histogram of differences doesn't allow us to see how the pre-size affects the post-size.

Scatter plots display the relationship between two quantitative variables.

Choose one variable on the x-axis variable, and the other on the y-axis.

Each observation is then plotted as a point.

Scatter plots

The larger the crab pre-moulting, the larger it is post-moulting.

How strong is this relationship?

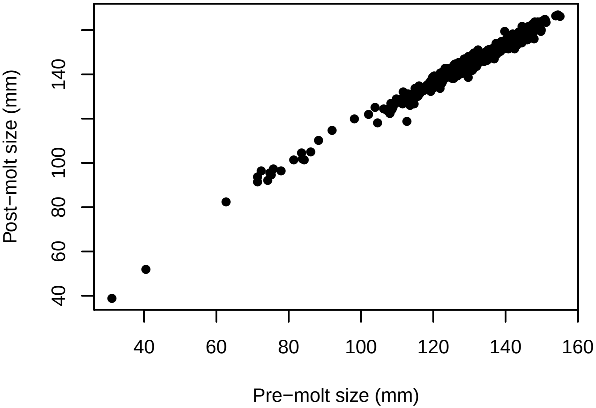


Figure: Relationship between pre and post moulting size.

Scatter plots

Problem

The strength of a relationship measures the information that one variable gives about the other.

Which below x/y relationship seems strongest?

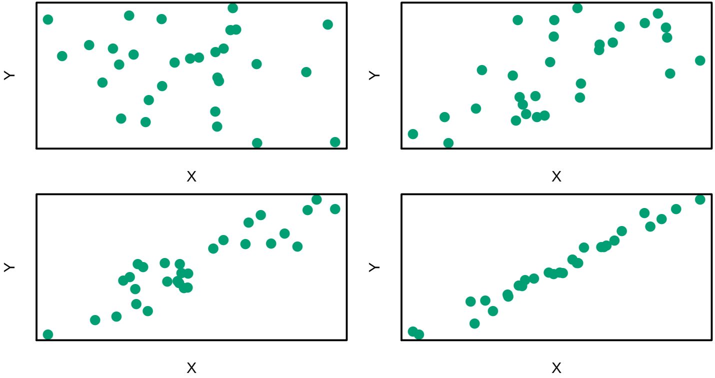
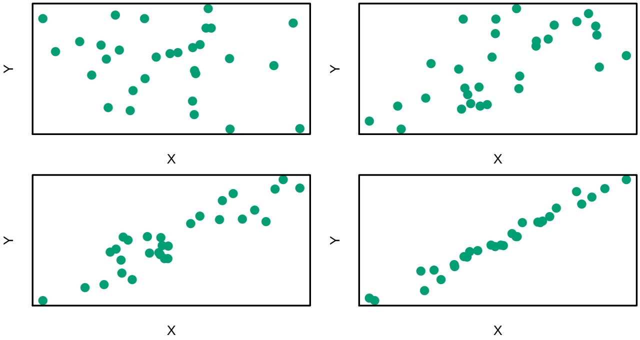


Figure: Relationship of differing strength

Correlation



- Top left panel seems to show no relationship
- Top right has some, but its weak
- Bottom left is stronger
- Bottom right is near *perfect*

How can we quantify this?

Correlation

Suppose we have two quantitative variables x and y measured on the same n individuals.

Let the ith observation be (x_i, y_i)

Now compute the means and standard deviations, $\bar{x}$, $\bar{y}$, s_x and s_y

Pearson Product Moment Correlation Coefficient

$$\rho = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

Correlation

$$\rho = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

ρ *only* take values between -1 and 1. Why?

Negative correlations indicate that when X increases, Y decreases.

- $+1.00$ — Perfect increasing linear relationship
- $+0.70$ — Strong increasing linear relationship
- $+0.50$ — Some increasing linear relationship
- 0.00 — No detectable relationship
- -0.50 — Some decreasing linear relationship
- -0.70 — Strong decreasing linear relationship
- -1.00 — Perfect decreasing linear relationship

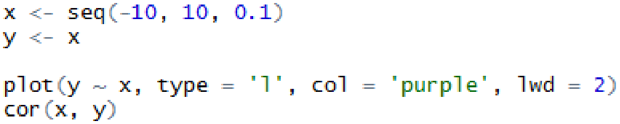
Correlation and dependence

Is correlation and dependence the same thing?

Related but not the same.

- Correlation $\implies$ Dependence
- However dependence may not mean correlation

Note that correlation here is “Linear”

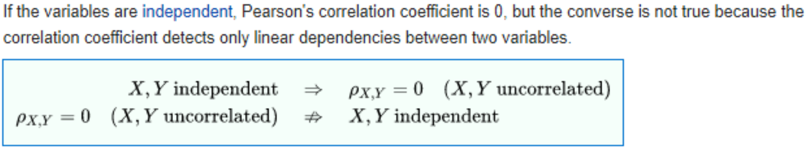


```
x <- seq(-10, 10, 0.1)
y <- x

plot(y ~ x, type = 'l', col = 'purple', lwd = 2)
cor(x, y)
```

Figure: Perfect correlation, since a straight line

Correlation and dependence



If the variables are independent, Pearson's correlation coefficient is 0, but the converse is not true because the correlation coefficient detects only linear dependencies between two variables.

$$\begin{aligned} X, Y \text{ independent} \quad &\Rightarrow \quad \rho_{X,Y} = 0 \quad (X, Y \text{ uncorrelated}) \\ \rho_{X,Y} = 0 \quad (X, Y \text{ uncorrelated}) \quad &\nRightarrow \quad X, Y \text{ independent} \end{aligned}$$

Figure: Wikipedia: Correlation

Correlation and covariance

$$Cor(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}}$$

$$Cov(X, Y) = E[XY] - E[X]E[Y]$$

$if\ Cov(X, Y) = 0$

$$then\ E[XY] = E[X]E[Y]$$

Correlation and variance

$$Cor(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

$$Cor(X,X) = \frac{Cov(X,X)}{\sqrt{Var(X)Var(X)}}$$

$since\ Cor(X,X) = 1$

$$1 = \frac{Cov(X,X)}{\sqrt{Var(X)Var(X)}}$$

$$Cov(X,X) = \sqrt{Var(X)Var(X)}$$

$$\therefore\ Cov(X,X) = Var(X)$$

Sorting Correlation

Problem

Order the the following plots from lowest to highest correlation.

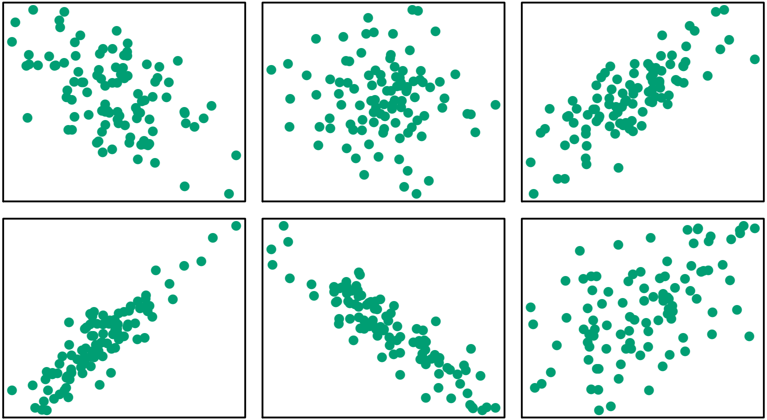


Figure: Relationship of differing strength

Crabs moulting size

For pre and post moult size of the crabs $\rho = 0.99$, showing a strong linear relationship between pre and post moult size.

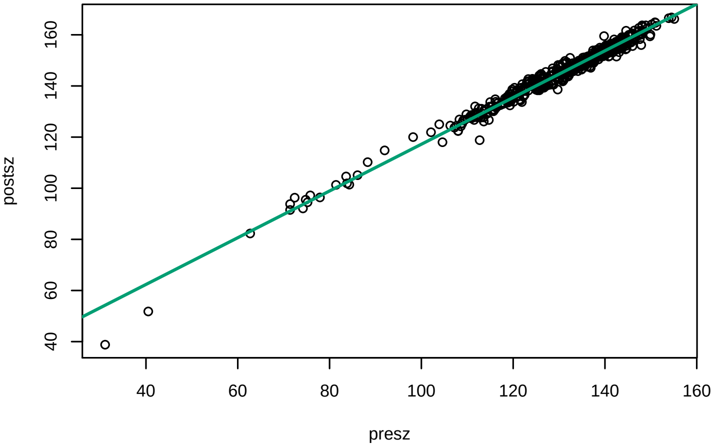


Figure: Change in size after moulting

Anscombe's data sets

Note that (this kind of) correlation only measures a linear relationship

All of the following have a correlation of 0.82

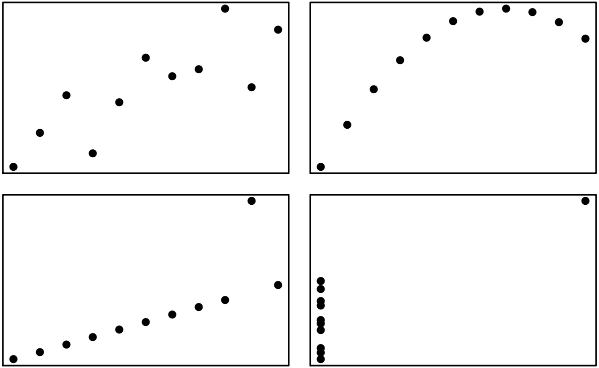


Figure: Anscombe's dataset

Spearman correlation

Spearman correlation tries to alleviate any issues to do with unusual points or curved (non-linear) relationships.

- The variables are individually ranked
- Ranking replaces the smallest value with the rank 1, the second smallest value with rank 2, etc
- After ranking we simply compute the Pearson correlation of the ranks

Spearman Correlation actually measures the extent to which the ranks are linearly related.

Spearman correlation

The crab moult data has Spearman correlation coefficient 0.99

For Anscombe's data sets

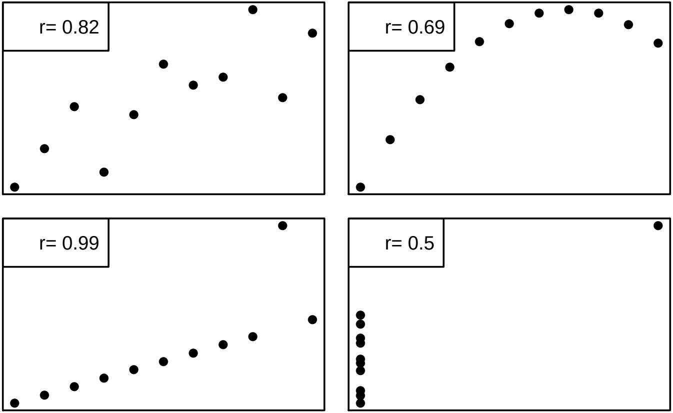


Figure: Anscombe's dataset

Correlation different from zero?

For crab moult data $\rho = 0.99$, but this is the sample correlation.

The important question is whether the population correlation is different to zero.

- Zero correlation means no (linear) relationship
- The observed correlation may have occurred by chance

To test if the population correlation is zero, we must simulate data with the same properties as our sample, but ensure that the population correlation is zero.

Options: permutation-based approach or approximations.

Permutation-based: randomly permute one of the variables and compute the sample correlation, given we have broken the relationship between the two variables.

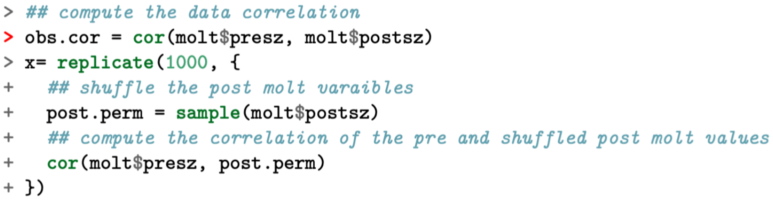
Repeats many times to get sample correlations when the population correlation is zero.

Correlation different from zero?

$\rho =$ population correlation

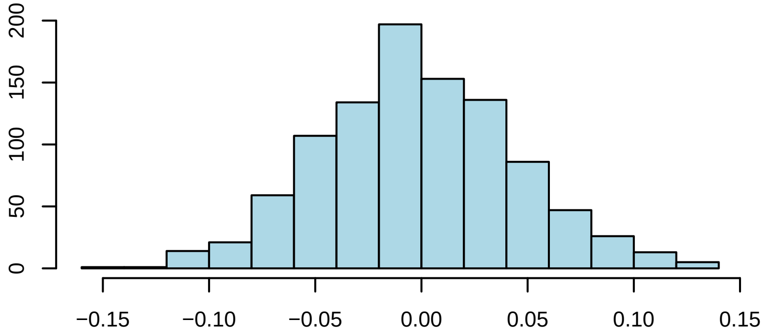
$$H_0 : \rho = 0$$

$$H_1 : \rho \neq 0$$



```
> ## compute the data correlation
> obs.cor = cor(molt$presz, molt$postsz)
> x= replicate(1000, {
+   ## shuffle the post molt varaibles
+   post.perm = sample(molt$postsz)
+   ## compute the correlation of the pre and shuffled post molt values
+   cor(molt$presz, post.perm)
+ })
```

Correlation different from zero?



Correlations computed under the null assumption that population correlation is zero.

Observed data correlation $\rho = 0.99$

Estimate the p value

Conclusion of hypothesis test to determine if: population correlation is not zero

$$\rho = \text{population correlation}$$

$$H_0 : \rho = 0$$

$$H_1 : \rho \neq 0$$

$$CV = 0.05$$

$$p\text{-}value = 0$$

$$CV > p\text{-}value$$
$$TRUE$$

$$\therefore \text{reject } H_0$$

$$\text{hence } \rho \neq 0$$

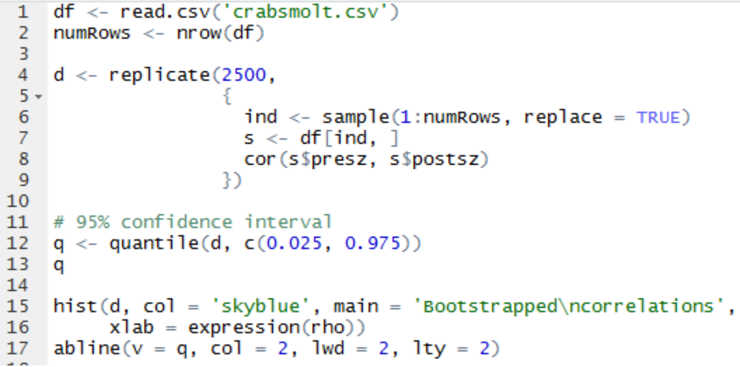
Confidence Intervals for ρ

Using the same resampling/bootstrap method as before.

In this case we must respect the pairs of observations: easiest way to do this is to resample observation numbers.

Giving confidence interval $[0.9882149, 0.9924811]$

Confidence Intervals for ρ



```r
df <- read.csv('crabsmolt.csv')
numRows <- nrow(df)

d <- replicate(2500,
               {
                 ind <- sample(1:numRows, replace = TRUE)
                 s <- df[ind, ]
                 cor(s$presz, s$postsz)
               })

# 95% confidence interval
q <- quantile(d, c(0.025, 0.975))
q

hist(d, col = 'skyblue', main = 'Bootstrapped\ncorrelations',
     xlab = expression(rho))
abline(v = q, col = 2, lwd = 2, lty = 2)
```

Figure: Calculation of confidence interval

Confidence Intervals

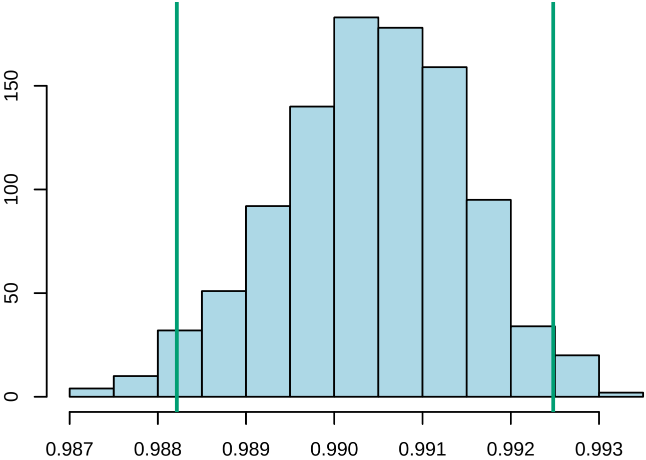


Figure: Bootstrapped sample of moulting dataset

Do Taller people have higher incomes?

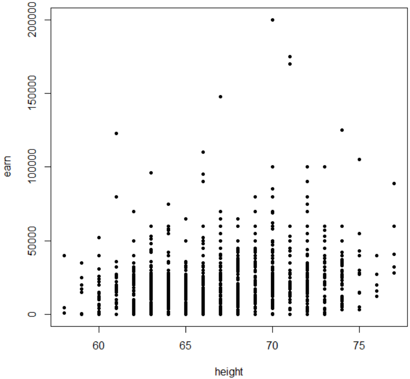
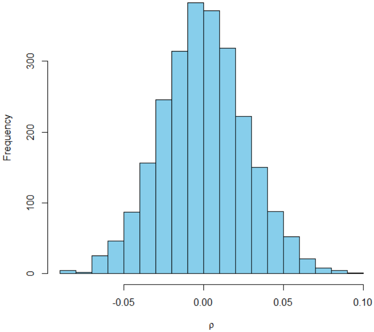


Figure: Height versus income dataset

$\rho = 0.302$ (crab moulting $\rho = 0.99$)

Do Taller people have higher incomes?



Dist. of sample correlations where the population correlation is zero, computed using permutation.

Problem

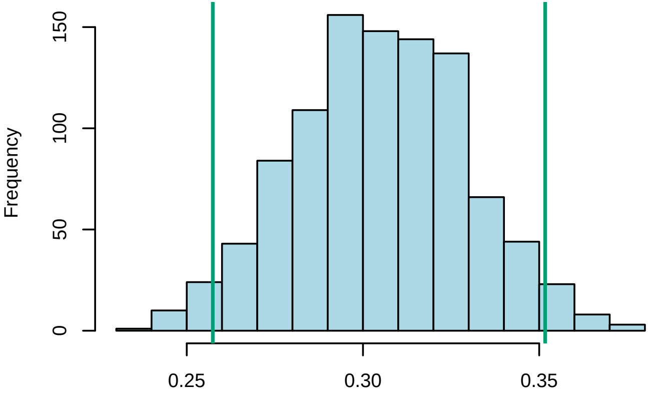
Given that $\rho = 0.302$ for the data, estimate the *p*-value.

Do Taller people have higher incomes?

What is the population correlation between income and height?

Can use the bootstrap distribution (shown below).

95% confidence interval on population correlation [0.257, 0.352]



Caveats

- Correlation measures the extent of a linear relationship between two variables
- Spearman correlation uses ranks to measure possibly non-linear relationships
- Both only look at 2 variables at a time

Correlation doesn't tell us how to predict the post moult size from the pre moult size, for example.

Nor does it allow us to ask, is post-moult size just pre-moult plus a constant.

Straight lines

If post-moult size (y) were exactly pre-moult size (x) $+$ a constant then:

$$y = x + a$$

and correlation would be 1

If post-moult size were pre-moult plus 10%,

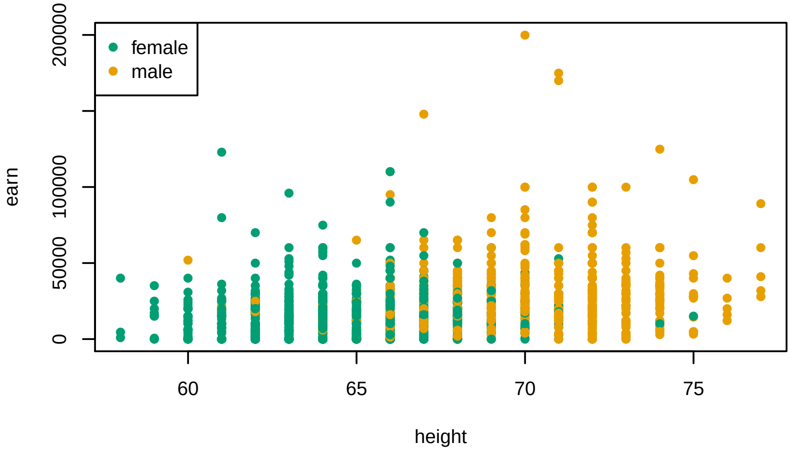
$$y = 1.1x$$

and the correlation would also be 1

We need better tools to look at these relationships.

Correlation and Causation

Let's look at the heights and income again



Correlation and Causation

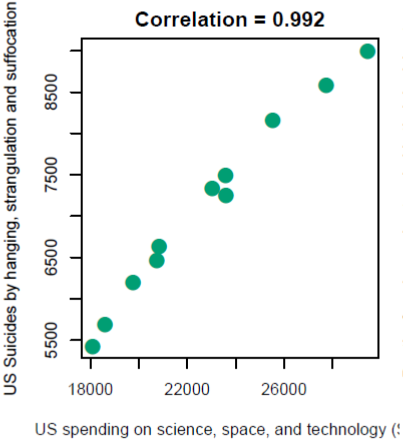
Recall for income and height $\rho = 0.302$

- $\rho = 0.065$ for females
- $\rho = 0.097$ for males

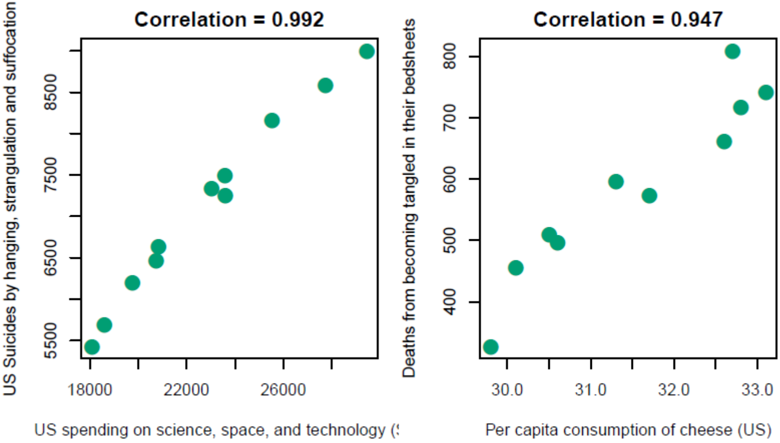
Gender seems a confounding factor

Allowing for it may show no evidence for correlation

Causation or just Correlation



Causation or just Correlation



Summary

- Scatterplots examine the relationship between two quantitative variables
- Strength of the relationship can be measured using correlation
- Pearson correlation measures the linearity relationship, Spearman correlation the linearity of ranks
- Correlation is between $[-1, 1]$, where 0 implies no correlation
- Can test population correlation is not zero using permutations
- Can compute population correlation confidence interval using bootstrapping
- Correlation does not imply causation