QS101: Introduction to Quantitative Methods in Social Science

Week 15: Measures of Association: Correlation

Dr. Florian Reiche
Teaching Fellow in Quantitative Methods
Course Director BA Politics and Sociology
Deputy Director of Student Experience and Progression, PAIS

February 5, 2015

Recap

Correlation

Recap

Queries

- What is χ^2?

Queries

- What is χ^2?
- What are degrees of freedom?

Queries

- What is χ^2?
- What are degrees of freedom?
- What is the p-value?

Numbers and the Media

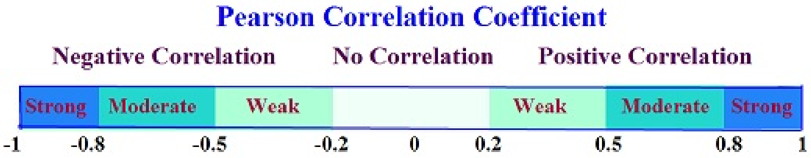
```
https://www.youtube.com/watch?v=oDPCmmZifE8
```

Correlation

Definition

- Correlation is a statistical tool that determines the degree of relationship between two different variables
- If correlation is strong, a person's score on one variable helps us predict the person's score on another variable
- It is limited with the range of -1 and $+1$

Classification of the Relationship



Strong Positive Relationship

- The higher the score on one variable, the higher the score on the other variable
- The lower the score on one variable, the lower the score on the other variable.

Example: Time spent on revision, and exam mark.

Strong Negative Relationship

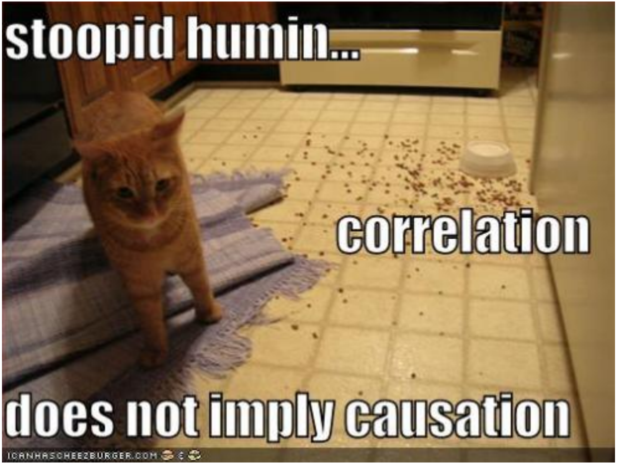
- The higher the score on one variable, the lower the score on the other variable
- The lower the score on one variable, the higher the score on the other variable.

Example: Time spent in the Dirty Duck, and module mark.

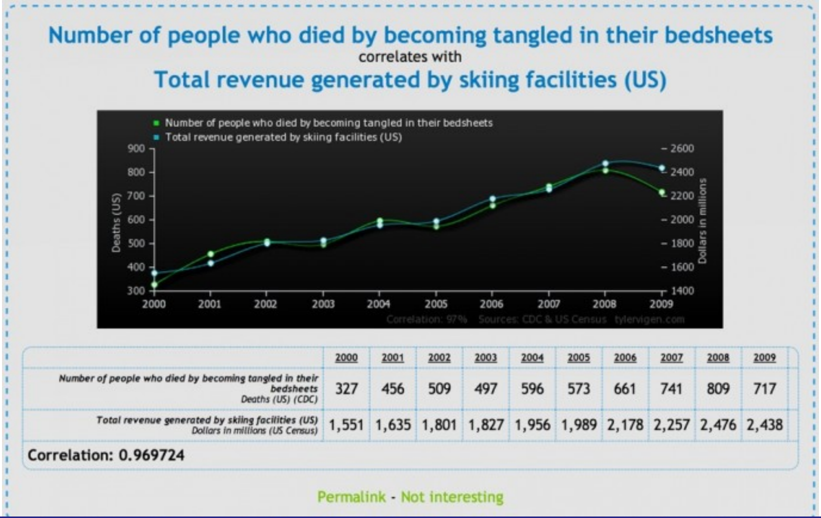
No Relationship

- Correlation coefficient (r) = 0
- Here, the score on one variable tells you nothing about the score on the other variable

Correlation and Causation



Example 1



Number of people who died by becoming tangled in their bedsheets

correlates with

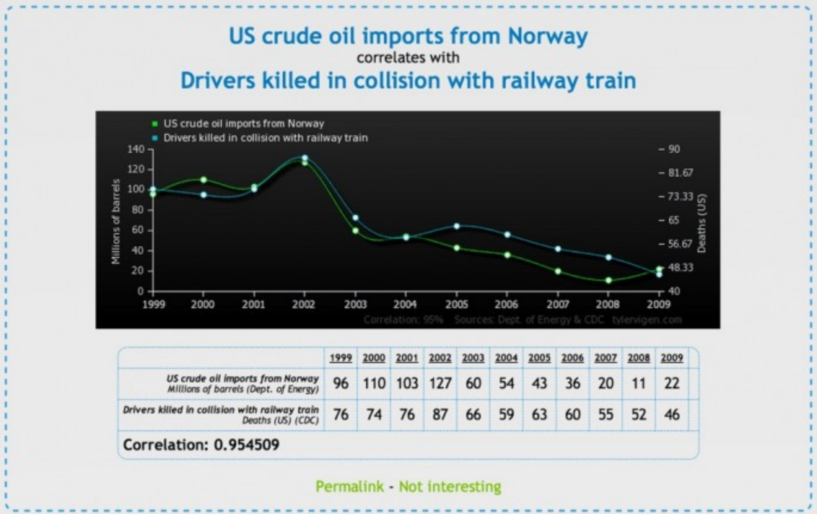
Total revenue generated by skiing facilities (US)

	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Number of people who died by becoming tangled in their bedsheets *Deaths (US) (CDC)*	327	456	509	497	596	573	661	741	809	717
Total revenue generated by skiing facilities (US) *Dollars in millions (US Census)*	1,551	1,635	1,801	1,827	1,956	1,989	2,178	2,257	2,476	2,438

Correlation: 0.969724

Permalink - Not interesting

Example 2



US crude oil imports from Norway

correlates with

Drivers killed in collision with railway train

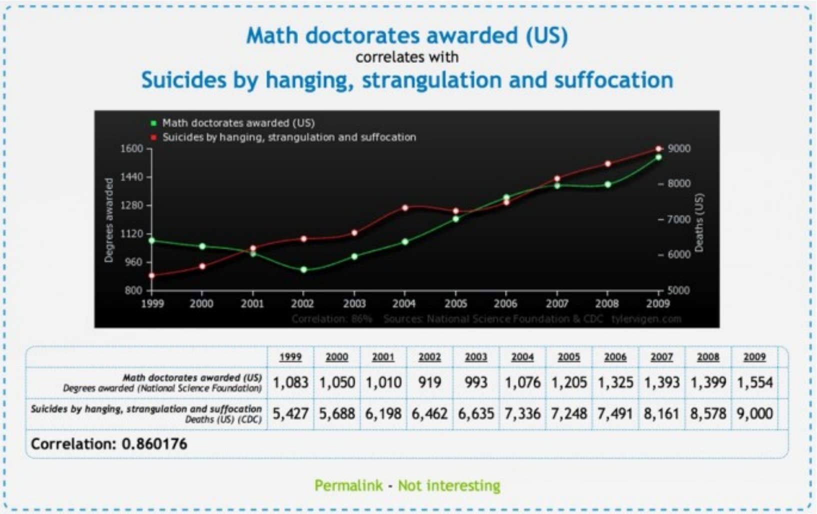
🟩 US crude oil imports from Norway
🟦 Drivers killed in collision with railway train
Millions of barrels
140
120
100
80
60
40
20
0
Deaths (US)
90
81.67
73.33
65
56.67
48.33
40
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
Correlation: 95% Sources: Dept. of Energy & CDC tylervigen.com

	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
US crude oil imports from Norway *Millions of barrels (Dept. of Energy)*	96	110	103	127	60	54	43	36	20	11	22
Drivers killed in collision with railway train *Deaths (US) (CDC)*	76	74	76	87	66	59	63	60	55	52	46

Correlation: 0.954509

Permalink - Not interesting

Example 3



Math doctorates awarded (US)

correlates with

Suicides by hanging, strangulation and suffocation

	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Math doctorates awarded (US) *Degrees awarded (National Science Foundation)*	1,083	1,050	1,010	919	993	1,076	1,205	1,325	1,393	1,399	1,554
Suicides by hanging, strangulation and suffocation *Deaths (US) (CDC)*	5,427	5,688	6,198	6,462	6,635	7,336	7,248	7,491	8,161	8,578	9,000

Correlation: 0.860176

Permalink - Not interesting

Why bother then?

- You can never show a cause-effect relationship with correlation
- Yet, you can get an idea about the data, and patterns within in
- This can help you develop ideas about cause-effect relationships

4 Types of correlation

- Pearson's r: A measure of the strength of a relationship between two continuous variables
- Spearman's r: A measure of the similarity between two ordinal rankings of a single set of data
- Point-biseral r: A measure of the strength of the relationship between one continuous variable and one dichotomous variable (e.g. gender, democracy, etc.)
- Phi (ϕ) correlation: A measure of the strength of the relationship between two dichotomous variables

The Pearson Product-Moment Correlation Coefficient

Is the most commonly employed measure:

$$r = \frac{N\Sigma xv - (\Sigma x)(\Sigma y)}{\sqrt{(N\Sigma x^2 - (\Sigma x)^2)(N\Sigma y^2 - (\Sigma y)^2)}} \tag{1}$$

N: Number of pairs of scores

Example

Subject	Cigarettes	Years Lived
1	25	63
2	35	68
3	10	72
4	40	62
5	85	65
6	75	46
7	60	51
8	45	60
9	50	55

Example contd.

- For example:
  $\Sigma x = 25 + 35 + 10 + 40 + 85 + 75 + 60 + 45 + 50 = 425$
- etc.

$$r = \frac{(9)(24,640) - (425)(542)}{\sqrt{((9)(24,525) - (425)^2)((9)(33,188) - (542)^2)}}$$

$$r = -0.6111$$

$$r = -0.61$$

The Pearson Product-Moment Correlation Coefficient

- Obtained for a sample drawn from the population, denoted r
- The population value is called rho (ρ)
- We are therefore interested in:
  - $H_0 : \rho = 0$
  - $H_a : \rho \neq 0$

Significance Test

- Uses the t-distribution (if you do not know what this is, read up on it, now!)
- The t-test formula for a correlation coefficient is as follows:

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

- This would give you the critical value (just like with χ^2 last week)
- Calculate the degrees of freedom (here: $df = N - 2$)
- You choose a level of significance, find the critical value, compare the values and decide

Example contd.

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

$$t = \frac{-0.6111}{\sqrt{\frac{1-(-0.6111)^2}{9-2}}}$$

$$t = -2.042$$

- We want 95% significance level
- Critical value for $df = 7$: $t = +2.365$ and -2.365
- Is this significant?

Example contd.

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

$$t = \frac{-0.6111}{\sqrt{\frac{1-(-0.6111)^2}{9-2}}}$$

$$t = -2.042$$

- We want 95% significance level
- Critical value for $df = 7$: $t = +2.365$ and $t = -2.365$
- Is this significant? No!

The other 3 Measures

- Spearman: Coolidge, pp. 204-206
- Point Biseral: Coolidge, pp. 208-212
- ϕ correlation: Coolidge, pp. 212-213