

Informatics 1: Data & Analysis

Lecture 18: Hypothesis Testing and Correlation

Ian Stark

School of Informatics
The University of Edinburgh

Friday 18 March 2016
Semester 2 Week 9



THE UNIVERSITY
of EDINBURGH

<http://blog.inf.ed.ac.uk/da16>

Data Retrieval

- The information retrieval problem
- The vector space model for retrieving and ranking

Statistical Analysis of Data

- Data scales and summary statistics
- Hypothesis testing and correlation
- χ^2 tests and collocations also *chi-squared*, pronounced “kye-squared”

Data in Multiple Dimensions

The previous lecture looked at **summary statistics** which give information about a single set of data values. Often we have multiple linked sets of values: several pieces of information about each of many individuals.

This kind of **multi-dimensional** data is usually treated as several distinct **variables**, with statistics now based on several variables rather than one.

Example Data

(NB: Not real students)

	A	B	C	D	E	F	G	H
Study	0.5	1	1.4	1.2	2.2	2.4	3	3.5
Exercise	4	7	4.5	5	8	3.5	6	5
Sleep	10	6	13	5	3	7	9	8.5
Exam	16	35	42	45	60	72	85	95

Data in Multiple Dimensions

The table below presents for each of eight imaginary students (A–H), the time in hours they spend each week on studying for Inf1-DA (outside lectures and tutorials) and on physical exercise; and how many hours they spent asleep on a particular night. This is juxtaposed with their Data & Analysis exam performance as a percentage.

We have four variables: study, exercise, sleep and exam results.

Example Data

(NB: Not real students)

	A	B	C	D	E	F	G	H
Study	0.5	1	1.4	1.2	2.2	2.4	3	3.5
Exercise	4	7	4.5	5	8	3.5	6	5
Sleep	10	6	13	5	3	7	9	8.5
Exam	16	35	42	45	60	72	85	95

Correlation

We can ask whether there is any observed **relationship** between the values of two different variables: do they vary up and down together?

If there is no relationship, then the variables are said to be **independent**.

If there is a relationship, then the variables are said to be **correlated**.

Two variables are **causally** connected if variation in the first causes variation in the second. If this is so, then they will also be correlated. However, the reverse is not true:

Correlation Does Not Imply Causation

Correlation Does Not Imply Causation

If we do observe a correlation between variables X and Y , it may be due to any of several things.

- Variation in X causes variation in Y , either directly or indirectly.
- Variation in Y causes variation in X , either directly or indirectly.
- Variation in X and Y is caused by some third factor Z .
- Chance: we just happen to have some values that look similar.

Visualizing Correlation

One way to discover correlation is through human inspection of some data visualisation.

For data like that below, we can draw a *scatter plot* taking one variable as the x-axis and one the y-axis and plotting a point for each item of data.

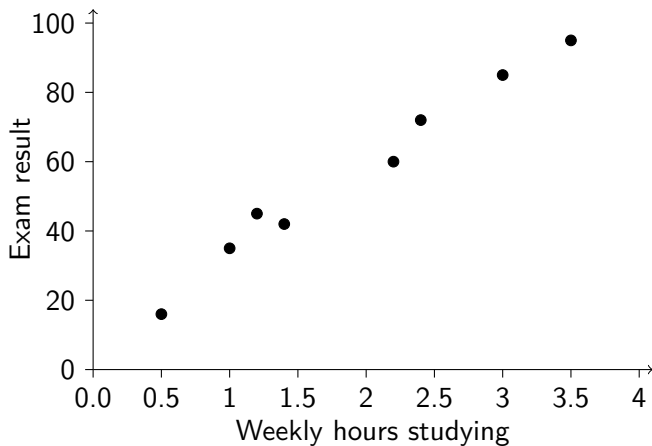
We can then look at the plot to see if we observe any correlation between variables.

Example Data

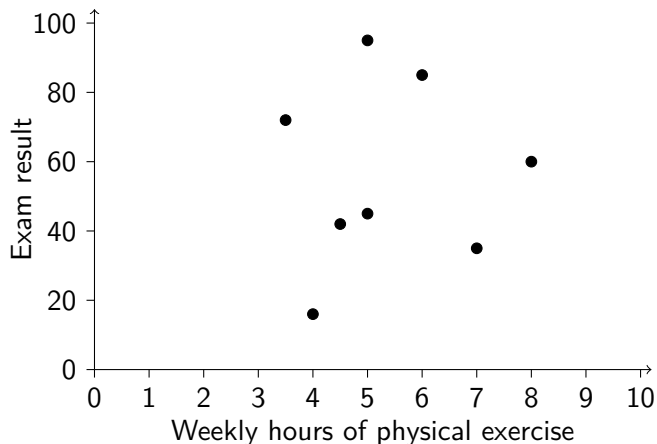
(NB: Not real students)

	A	B	C	D	E	F	G	H
Study	0.5	1	1.4	1.2	2.2	2.4	3	3.5
Exercise	4	7	4.5	5	8	3.5	6	5
Sleep	10	6	13	5	3	7	9	8.5
Exam	16	35	42	45	60	72	85	95

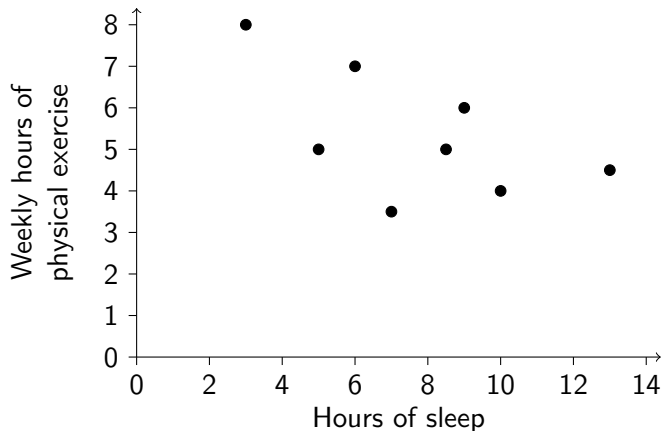
Studying vs. Exam Results



Physical Exercise vs. Exam Results



Hours of Sleep vs. Physical Exercise



Hypothesis Testing

The previous visualisations of data suggested possible correlations between variables.

There are many other ways to formulate possible correlations. For example:

- From a proposed underlying mechanism;
- Analogy with another situation where some relation is known to exist;
- Based on the predictions of a proposed model for a system.

Any such suggestion of a correlation is a *hypothesis*.

Statistical tests provide the mathematical tools to assess evidence and carry out **hypothesis testing**.

Statistical Tests

Most statistical testing starts from a specified *null hypothesis*, that there is nothing out of the ordinary in the data: no correlation, no effect, nothing to see.

We then compute some statistic from the data. Call this R .

The *hypothesis test* is then to investigate how likely it is that we would see a result like R if the null hypothesis were true.

This chance is called a *p-value*, with $0 \leq p \leq 1$.

Significance

The value p represents the chance that we would obtain a result like R if the **null hypothesis** were true.

If p is small, then we conclude that the null hypothesis is a poor explanation for the observed data.

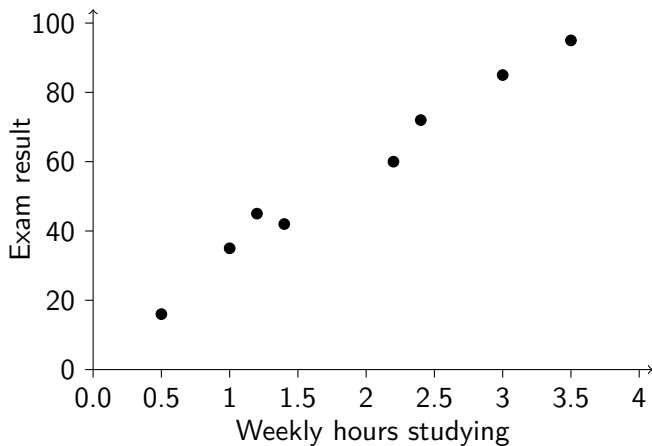
Based on this we might **reject** the null hypothesis.

Standard thresholds for “small” are $p < 0.05$, meaning that there is less than 1 chance in 20 of obtaining the observed result by chance, if the null hypothesis is true; or $p < 0.01$, meaning less than 1 chance in 100.

An observation that leads us to reject the null hypothesis is described as **statistically significant**.

This idea of testing for significance is due to R. A. Fisher (1890–1962).

Studying vs. Exam Results



Correlation Coefficient

The *correlation coefficient* is a statistical measure of how closely one set of data values x_1, \dots, x_N are correlated with another y_1, \dots, y_N .

Take μ_x and σ_x the mean and standard deviation of the x_i values.

Take μ_y and σ_y the mean and standard deviation of the y_i values.

The correlation coefficient $\rho_{x,y}$ is then computed as:

$$\rho_{x,y} = \frac{\sum_{i=1}^N (x_i - \mu_x)(y_i - \mu_y)}{N\sigma_x\sigma_y}$$

Values of $\rho_{x,y}$ always lie between -1 and 1 .

Correlation Coefficient

The *correlation coefficient* is a statistical measure of how closely one set of data values x_1, \dots, x_N are correlated with another y_1, \dots, y_N .

Take μ_x and σ_x the mean and standard deviation of the x_i values.

Take μ_y and σ_y the mean and standard deviation of the y_i values.

The correlation coefficient $\rho_{x,y}$ is then computed as:

$$\rho_{x,y} = \frac{\sum_{i=1}^N (x_i - \mu_x)(y_i - \mu_y)}{N\sigma_x\sigma_y}$$

Values of $\rho_{x,y}$ always lie between -1 and 1 .

Bonus non-examinable observation:

The correlation coefficient $\rho_{x,y}$ turns out to be the cosine similarity of the two datasets when rebased around their means and treated as high-dimensional vectors.

Correlation Coefficient

The *correlation coefficient* is a statistical measure of how closely one set of data values x_1, \dots, x_N are correlated with another y_1, \dots, y_N .

Take μ_x and σ_x the mean and standard deviation of the x_i values.

Take μ_y and σ_y the mean and standard deviation of the y_i values.

The correlation coefficient $\rho_{x,y}$ is then computed as:

$$\rho_{x,y} = \frac{\sum_{i=1}^N (x_i - \mu_x)(y_i - \mu_y)}{N\sigma_x\sigma_y}$$

Values of $\rho_{x,y}$ always lie between -1 and 1 .

If $\rho_{x,y}$ is close to 0 then this suggests there is no correlation.

If $\rho_{x,y}$ is nearer $+1$ then this suggests x and y are *positively correlated*.

If $\rho_{x,y}$ is closer to -1 then this suggests x and y are *negatively correlated*.

Correlation Coefficient as a Statistical Test

In a test for correlation between two variables x and y — such as study hours and exam results — we are looking to see whether the variables are correlated; and if so in what direction.

The **null hypothesis** is that there is no correlation.

We calculate the correlation coefficient $\rho_{x,y}$, and then do one of two things:

- Look in a table of **critical values** for this statistic, to see whether the value we have is significant;
- Compute directly the **p-value** for this statistic, to see whether it is small.

Depending on the result, we may reject the null hypothesis.

Critical Values for Correlation Coefficient

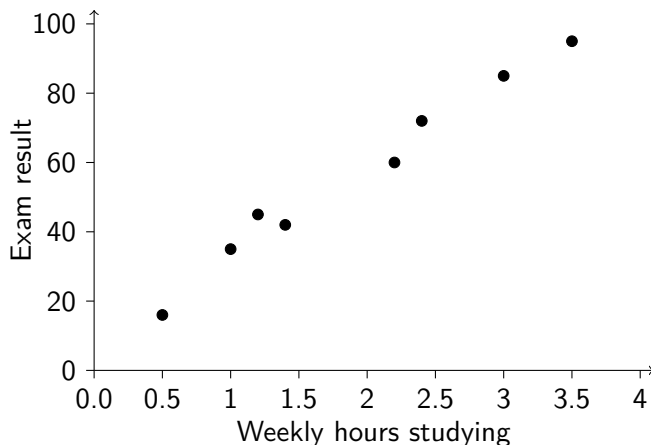
ρ	$p = 0.10$	$p = 0.05$	$p = 0.01$	$p = 0.001$
$N = 7$	0.669	0.754	0.875	0.951
$N = 8$	0.621	0.707	0.834	0.925
$N = 9$	0.582	0.666	0.798	0.898
$N = 10$	0.549	0.632	0.765	0.872

This table has rows indicating the critical values of the correlation coefficient ρ for different numbers of data items N in the series being compared.

It shows that for $N = 8$ data items that are not correlated, there is probability $p = 0.01$ of observing a coefficient $|\rho_{x,y}| > 0.834$.

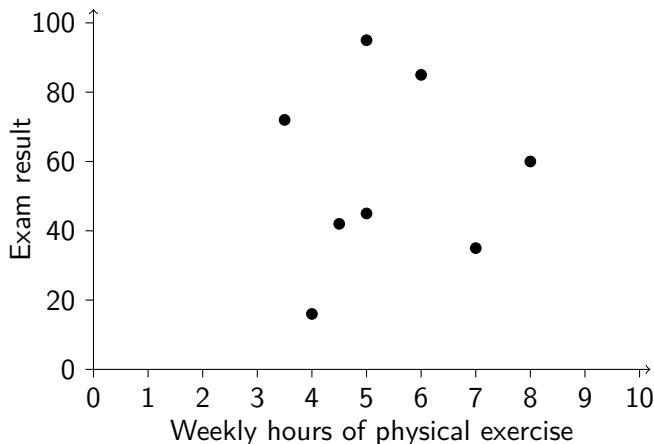
In the same way for $N = 8$ uncorrelated data items a value of $|\rho_{x,y}| > 0.925$ has probability $p = 0.001$ of occurring, only one chance in a thousand.

Studying vs. Exam Results



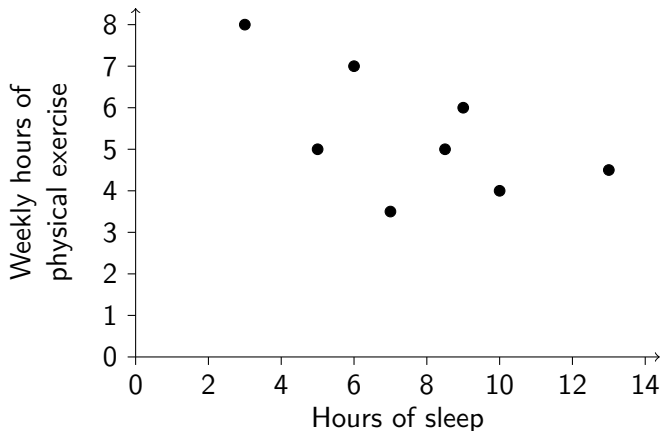
The correlation coefficient is $\rho_{\text{study,exam}} = 0.990$, well above the critical value 0.925 for $p < 0.001$ and strongly indicating **positive correlation**.

Physical Exercise vs. Exam Results



The correlation coefficient is $\rho_{\text{exercise,exam}} = 0.074$, far less than any critical value and indicating **no evidence of correlation** for these 8 students.

Hours of Sleep vs. Physical Exercise



The correlation coefficient is $\rho_{\text{sleep}, \text{exercise}} = -0.599$, below the critical value of 0.621 for $|\rho_{x,y}|$, so giving **no evidence of correlation** here.

Estimating Correlation from a Sample

Suppose that we have sample data x_1, \dots, x_n and y_1, \dots, y_n drawn from a much larger population of size N , so $n \ll N$.

Calculate m_x and m_y the estimates of the population means.

Calculate s_x and s_y the estimates of the population standard deviations.

Then an estimate $r_{x,y}$ of the correlation coefficient in the population is:

$$r_{x,y} = \frac{\sum_{i=1}^n (x_i - m_x)(y_i - m_y)}{(n-1)s_x s_y}$$

Not that, as with estimating the variation in a larger population, we use $(n-1)$ in the denominator.

The correlation coefficient is sometimes called *Pearson's correlation coefficient*, particularly when it is estimated from a sample using the formula above.

Summary

So far, we have the following procedure to identify correlation between two series of data.

- Draw a **scatter plot**. Does it look as though there is a relationship between the two variables?
- Calculate a **correlation coefficient** R .
- Look in a table of **critical values** to see whether R is large, given the number of data points.
- If R is above the critical value for some chosen p , say 0.05 or 0.01, then this may be judged **statistically significant** and lead us to **reject the null hypothesis**.

However, there are two major warnings associated with this approach.

Beware

Warning 1

The arrangement of **null hypothesis** and **significance testing** is enticing and convenient, but *very very* slippery in practice.



John P. A. Ioannidis.

Why Most Published Research Findings Are False.

PLoS Medicine 2005 2(8):e124. DOI: [10.1371/journal.pmed.0020124](https://doi.org/10.1371/journal.pmed.0020124)

Warning 2

Correlation Still Does Not Imply Causation

What's Wrong With Significance?

- The value p is the probability of seeing certain results if the null hypothesis were true.
It is **not** the probability that the null hypothesis is true.
- It doesn't say whether an observed variation is actually large or small (that's measured by "effect size").
It is really about whether it is statistically *detectable*.
- Events with $p < 0.05$ happen all the time. Well, 1 time in 20.
Seeing a low p -value is perhaps evidence to suggest an effect. It's a reason to do another experiment, or make a prediction.
Only if we see this evidence again and again can we say with confidence that we have a result.

What if p is close to 0.05?

If p is not below the chosen threshold, then you have no result. No evidence of anything. It's not “nearly significant” — it's noise. It isn't:

- a certain trend toward significance ($p = 0.08$)

- a considerable trend toward significance ($p = 0.069$)

- a distinct trend toward significance ($p = 0.07$)

- a favorable trend ($p = 0.09$)

- ...

- verging on significance ($p = 0.056$)

- verging on the statistically significant ($p < 0.1$)

- verging-on-significant ($p = 0.06$)

- very close to approaching significance ($p = 0.060$)

<http://is.gd/stillnotsignificant>

In early 2015, the journal *Basic and Applied Social Psychology* banned the use of:


- Hypothesis testing;
- p-values;
- significance;
- confidence intervals; and
- all related statistical techniques.

So far *BASP* is unique in this, and the issue is a discussion point online among statisticians and social scientists.

<http://is.gd/significancebanned>

+ AllTrials

All Trials Registered | All Results Reported

[Home](#)[Find out more](#)[Get involved](#)[Supporters](#)[News](#)[Sign the petition](#)[Donate](#)[Q](#)

Around half of clinical trials have never been reported.
This is the story of the campaign to find them—
and to fix medicine.

[Read the AllTrials story](#)[DONATE](#)[GET INVOLVED](#)[LATEST NEWS](#)

Correlation Does Not Imply Causation

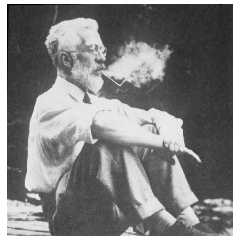
If we do observe a correlation between variables X and Y , it may be due to any of several things.

- Variation in X causes variation in Y , either directly or indirectly.
- Variation in Y causes variation in X , either directly or indirectly.
- Variation in X and Y is caused by some third factor Z .
- Chance: we just happen to have some values that look similar.

Examples?

Famous examples of observed correlations which may not be causal.

- Salaries of Presbyterian ministers in Massachusetts
- The price of rum in Havana
- Regular smoking
- Lower grades at university
- The quantity of apples imported into the UK
- The rate of divorce in the UK



R. A. Fisher

Nonetheless, statistical analysis can still serve as evidence of causality:

- Postulate a causative mechanism, propose a hypothesis, **make predictions**, and then look for a correlation in data;
- Propose a hypothesis, repeat **experiments** to confirm or refute it.

Polio epidemics in 1950s USA

<http://is.gd/poliocorrelation>

The Daily Mail Oncological Ontology Project

<http://kill-or-cure.herokuapp.com>

Spurious Correlations

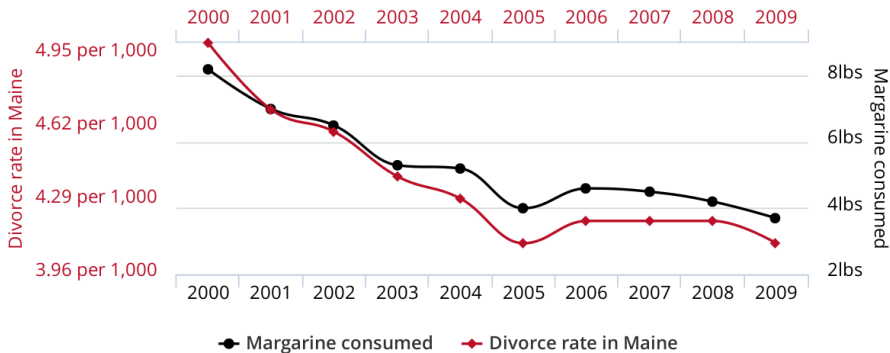
<http://tylervigen.com/spurious-correlations>

Divorce rate in Maine

correlates with

Per capita consumption of margarine

Correlation: 99.26% ($r=0.992558$)



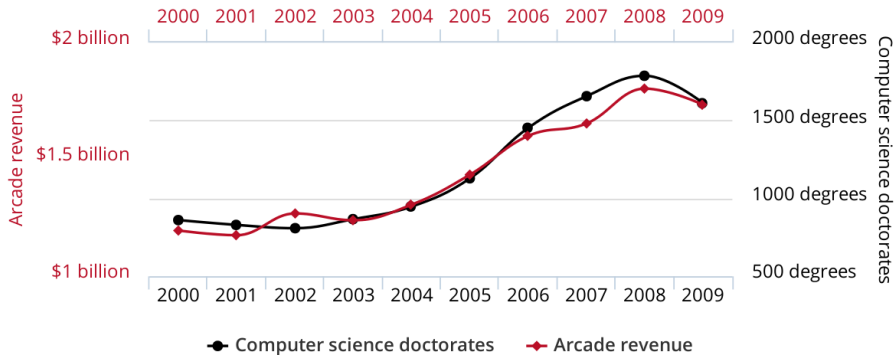
tylervigen.com

Data sources: National Vital Statistics Reports and U.S. Department of Agriculture



Total revenue generated by arcades correlates with Computer science doctorates awarded in the US

Correlation: 98.51% ($r=0.985065$)



tylervigen.com

Data sources: U.S. Census Bureau and National Science Foundation

Read Wikipedia on [The German Tank Problem](#)

Month	Statistical Estimate	Intelligence Estimate	German Records
June 1940	169	1000	122
June 1941	244	1550	271
August 1942	327	1550	342

<http://is.gd/tankstats>

If you like that, then try this.



T. W. Körner

The Pleasures of Counting

Cambridge University Press, 1996

<http://is.gd/da16counting>

Borrow a copy from the Murray Library, King's Buildings, [QA93 Kor](#).