

# Informatics 1: Data & Analysis

## Lecture 18: Hypothesis Testing and Correlation

Ian Stark

School of Informatics  
The University of Edinburgh

Tuesday 26 March 2013  
Semester 2 Week 10



## 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
- $\chi^2$  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

	A	B	C	D	E	F	G	H
Study	0.5	1	1.4	1.2	2.2	2.4	3	3.5
Drinking	25	20	22	10	14	5	2	4
Eating	4	7	4.5	5	8	3.5	6	5
Exam	16	35	42	45	60	72	85	95

# Data in Multiple Dimensions

The table below presents for each of eight hypothetical students (A–H), the time in hours they spend each week on: studying for Inf1-DA (outside lectures and tutorials), drinking and eating. This is juxtaposed with their Data & Analysis exam results.

Here we have four variables: studying, drinking, eating and exam results.

## Example Data

	A	B	C	D	E	F	G	H
Study	0.5	1	1.4	1.2	2.2	2.4	3	3.5
Drinking	25	20	22	10	14	5	2	4
Eating	4	7	4.5	5	8	3.5	6	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.

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  $Z$ .
- Coincidence

# Examples?

Famous examples of 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

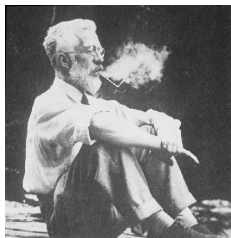
Nonetheless, statistical analysis can still serve as evidence of causality:

- 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.

# Examples?

Famous examples of 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



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.



# Visualizing Correlation

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

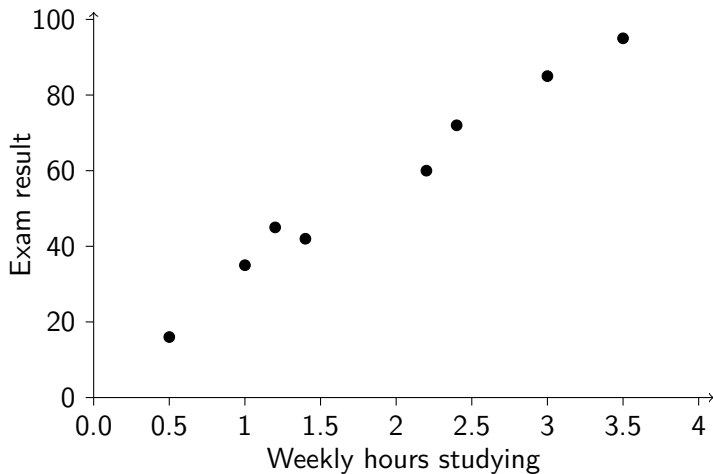
For data like that below, we can draw a *scatter plot* taking one variable as the  $x$ -axis and one the  $y$ -axis and plotting a point for each item of data.

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

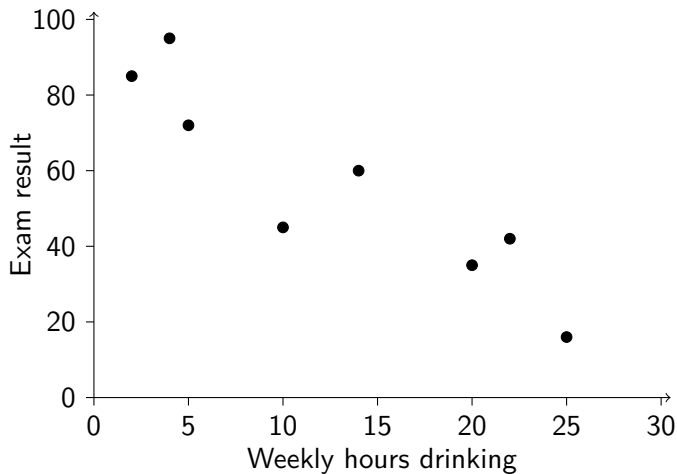
## Example Data

	A	B	C	D	E	F	G	H
Study	0.5	1	1.4	1.2	2.2	2.4	3	3.5
Drinking	25	20	22	10	14	5	2	4
Eating	4	7	4.5	5	8	3.5	6	5
Exam	16	35	42	45	60	72	85	95

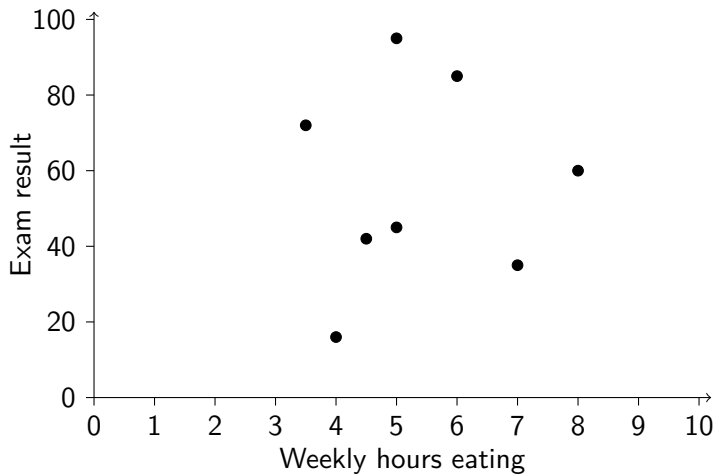
# Studying vs. Exam Results



# Drinking vs. Exam Results



# Eating vs. Exam Results



# Hypothesis Testing

The previous visualisations of data suggested hypotheses about possible correlations between variables.

There are many other ways to formulate hypothesis. For example:

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

**Statistical tests** provide the mathematical tools to confirm or refute such hypotheses.

# Statistical Tests

Most statistical testing starts from a specified *null hypothesis*, that there is nothing out of the ordinary in the data.

We then compute some statistic from the data, giving result  $R$ .

For this result  $R$  we calculate a *probability value*  $p$ .

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

**Note:**  $p$  is not a probability that the null hypothesis is true. That is not a quantifiable value.

# Significance

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

If the value of  $p$  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.

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

# The Power of Meta-Analysis



Ben Goldacre.

*Bad Pharma: How Drug Companies Mislead Doctors and Harm Patients.*

Fourth Estate, 2012.



# 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  $\mu_x$  and  $\sigma_x$  the mean and standard deviation of the  $x_i$  values.

Take  $\mu_y$  and  $\sigma_y$  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 the **probability value**  $p$  for this statistic, to see whether it is small.

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

## Critical Values for Correlation Coefficient

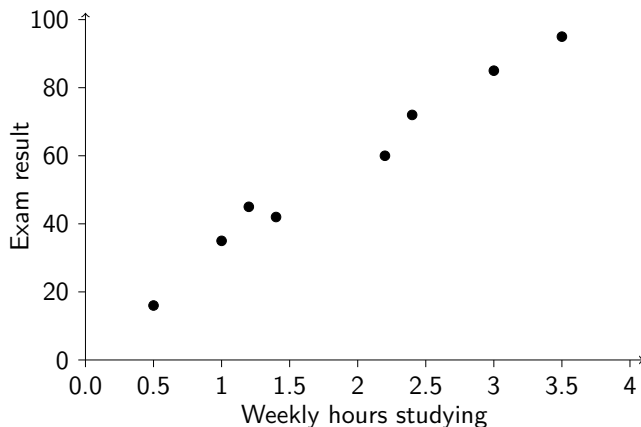
$\rho$	$p = 0.10$	$p = 0.05$	$p = 0.01$	$p = 0.001$
$N = 7$	0.669	0.754	0.875	0.951
$N = 8$	0.621	0.707	0.834	0.925
$N = 9$	0.582	0.666	0.798	0.898
$N = 10$	0.549	0.632	0.765	0.872

This table has rows indicating the critical values of  $p$  for depending on the number of data items  $N$  in the series being compared.

It shows that for  $N = 8$  uncorrelated data items a value of  $|\rho_{x,y}| > 0.834$  has probability  $p < 0.01$  of occurring.

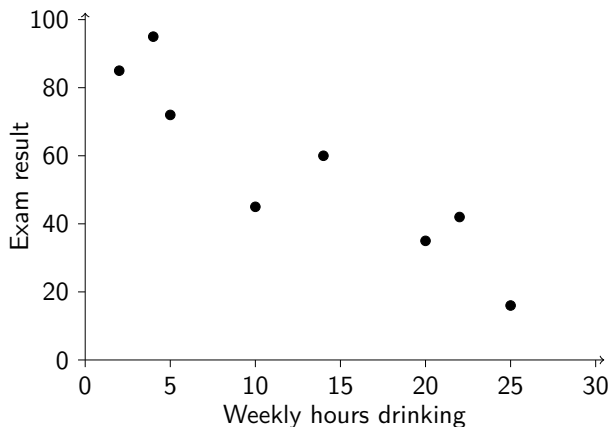
In the same way for  $N = 8$  uncorrelated data items a value of  $|\rho_{x,y}| > 0.925$  has probability  $p < 0.001$  of occurring, less than one chance in a thousand.

# Studying vs. Exam Results



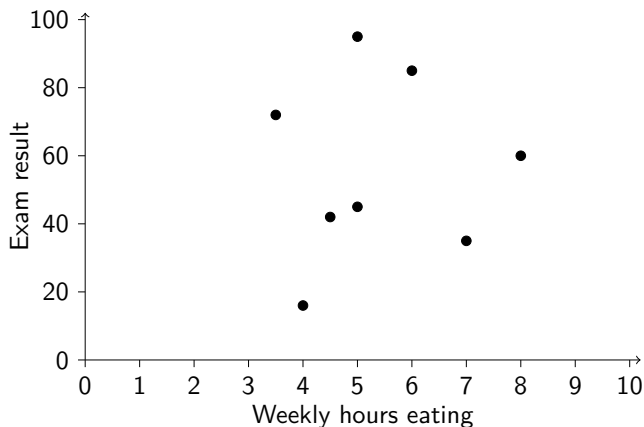
The correlation coefficient is  $\rho_{\text{study,exam}} = 0.990$ , well above the critical value 0.925 for  $p < 0.001$  and strongly indicating **positive correlation**.

# Drinking vs. Exam Results



The correlation coefficient is  $\rho_{\text{drink,exam}} = -0.914$ , beyond the critical value 0.834 for  $p < 0.01$  and indicating **negative correlation**.

# Eating vs. Exam Results



The correlation coefficient is  $\rho_{\text{eat,exam}} = 0.074$ , far less than any critical value and indicating no significant correlation.

# 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:

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

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

# One-Tail and Two-Tail Tests

There are two subtly different ways to apply the correlation coefficient.

- **Two-tailed test:** Looking for a correlation of any kind, positive or negative, either is significant.
- **One-tailed test:** Looking for a correlation of just one kind (say, positive) and only this is significant.

We have been using **two-tailed tests**. The choice of test affects the critical value table: in general, a one-tailed test requires a lower critical value for significance.

two-tail	$p = 0.10$	$p = 0.05$	$p = 0.01$	$p = 0.001$
one-tail	$p = 0.05$	$p = 0.025$	$p = 0.005$	$p = 0.0005$
$N = 7$	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



# On Using Statistics to Find Things Out

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

If you like that then try this.



[T. W. Körner.](#)

*The Pleasures of Counting.*

[Cambridge University Press, 1996.](#)

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