# Informatics 1: Data & Analysis

## Lecture 17: Summary Statistics

Ian Stark

School of Informatics
The University of Edinburgh

Tuesday 21 March 2017
Semester 2 Week 9

THE UNIVERSITY
*of* EDINBURGH

https://blog.inf.ed.ac.uk/da17

# Unstructured Data

## Data Retrieval

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

## Statistical Analysis of Data

- Summary statistics
- Hypothesis testing and $\chi^2$ also *chi-squared*, pronounced “kye-squared”
- Data scales. Correlation and causation.

# Revised Lecture Structure !

Lecture content for the next two weeks will be different to that in previous years: covering slightly more material, in a new order.

## Week 9

| | |
|---|---|
| Tuesday 21 March | Lecture 17: Statistics and correlation coefficients |
| Friday 24 March | Lecture 18: Hypothesis testing and $\chi^2$ |

## Week 10

| | |
|---|---|
| Tuesday 28 March | Lecture 19: Data scales. Correlation and causation |
| Friday 31 March | **No lecture** |

## Week 11

| | |
|---|---|
| Tuesday 4 April | Lecture 20: Course review |
| Friday 7 April | Lecture 21: Past exam questions |

# Analysis of Data

There are many reasons to analyse data. For example:

- To discover implicit structure in the data;

  e.g., finding patterns in experimental data which might in turn suggest new models or experiments.

- To confirm or refute a hypothesis about the data.

  e.g., testing a scientific theory against experimental results.

Mathematical statistics provide a powerful toolkit for performing such analyses, with wide and effective application.

# Analysis of Data

Mathematical statistics provide a powerful toolkit for performing such analyses, with wide and effective application.

This analytic strength is twofold:

- Statistics can sensitively detect information not immediately apparent within a mass of data;
- Statistics can help determine whether or not an apparent feature of data is really there.

Machine assistance is essential for large datasets, and enables otherwise infeasible resampling techniques such as *bootstrapping* and *jackknifing*.

# Learn Statistics

There are lots of books for learning about statistics. The following is recommended as an introduction that is written to be approachable without needing a particularly strong mathematical background.

P. Hinton.
*Statistics Explained: A Guide for Social Science Students*.
Routledge, third edition, 2014.

The third edition is the latest, but earlier editions are also suitable. See the main library for a copy.

# Statistics in Action

Here are two more books, for finding out about how statistics are used and abused. Both are easy reading. The links give full text online.

- M. Blastland and A. Dilnot. http://is.gd/tigerisnt
  *The Tiger That Isn't: Seeing Through a World of Numbers*.
  Profile, 2008.

  "Makes statistics far, far too interesting"

- D. Huff.
  *How to Lie with Statistics*. http://is.gd/huffbook
  W. W. Norton, 1954.

  "The most widely read statistics book in the history of the world"

# Qualitative and Quantitative

What type of statistical analysis we might apply to some data depends on:

- The reason for wishing to carry out the analysis;
- The type of data to hand.

The simplest possible distinction between types of data is *qualitative* vs. *quantitative*.

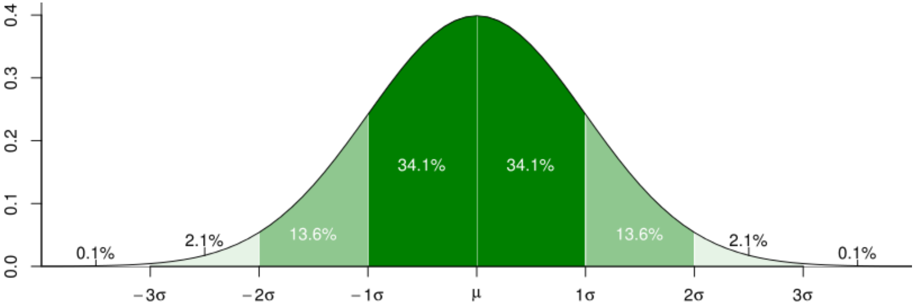
Qualitative: yes/no classification; town of residence; course grade

Quantitative: date; population; length; weight

# Normal Distribution

In the *normal distribution*, numerical data is clustered symmetrically around a central value with a bell-shaped frequency curve.

For sound mathematical reasons, many real-world examples of quantitative data do follow a normal distribution. However, not all do so, and the name "normal" can sometimes be misleading.

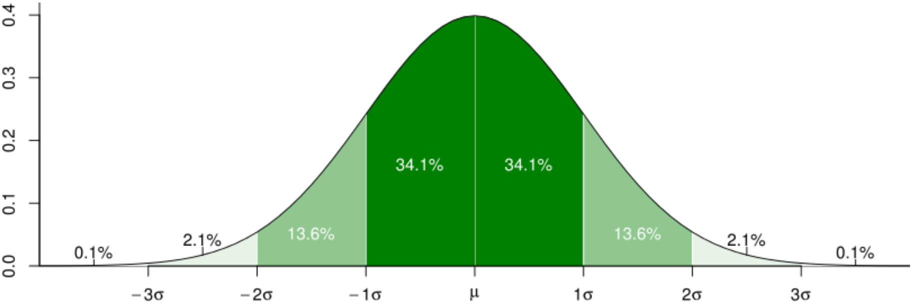


# Normal Distribution

Any normal distribution is described by two parameters.

The *mean* $\mu$ (mu, said "mew") is the centre around which the data clusters.

The *standard deviation* $\sigma$ (sigma) is a measure of the spread of the curve. For a normal distribution, it coincides with the *inflection point* where the curve changes from being convex to concave.



# Statistics

A *statistic* is a single value computed from data that captures some overall property of the data.

For example, the mean of a normal distribution is a statistic that captures the value around which the data is clustered.

Similarly, the standard deviation of a normal distribution is a statistic that captures the degree of spread of the data around its mean.

The notion of mean and standard deviation generalise to any quantitative data, even if it is not normally distributed.

There are also other statistics, the mode and median, that are alternatives to the mean for summarising the "average value" of some data.

# Mode

For any set of data the *mode* is the value which occurs most often.

## Example: Mode

For the data set $\{\text{north}, \text{west}, \text{south}, \text{north}, \text{east}\}$ the mode is north, which is the only value to occur twice.

Data may be *bimodal* (two modes) or even *multimodal* (more than two).

## Example: Bimodal data

For the integer data set $\{6, 2, 3, 6, 2, 5, 1, 7, 2, 5, 6\}$ both 2 and 6 are modes, each occurring three times.

The mode makes sense for all types of data scale. However, it is not particularly informative for quantitative data with real-number values, where it is uncommon for the same data value to occur more than once.

This is an instance of a more general phenomenon: in general it is neither useful nor meaningful to compare real-number values for equality

# Median

Given data values $x_1, x_2, \ldots, x_N$ sorted into in non-decreasing order, the *median* is the middle value $x_{(N+1)/2}$, for $N$ odd. If $N$ is even, then any value between $x_{N/2}$ and $x_{(N/2)+1}$ inclusive is a possible median.

## Example: Median

Given the integer data set $\{6, 2, 3, 6, 2, 5, 0, 7, 2, 5, 6\}$ we can write it in non-decreasing order $\{0, 2, 2, 2, 3, 5, 5, 6, 6, 6, 7\}$ and identify the middle value as 5.

The median exists for any data that can be put in order.

Median is a good summary statistic for data where there is a forced cutoff at one end, or possible distortion by extreme outliers. For example, in reporting average salaries, product lifetimes, cancer survival times.

The median wait for a hospital appointment in Scotland during October–December 2016 was 44 days. Half of patients waited less than that time; half waited longer. NHS Scotland

# Mean

Given data values $\{x_1, x_2, \ldots, x_N\}$, the *mean* is their total divided by the number of values: $(\sum_{i=1}^{N} x_i)/N$.

## Example: Mean

For the integer data set $\{6, 2, 3, 6, 2, 5, 0, 7, 2, 5, 6\}$, the mean is $(6 + 2 + 3 + 6 + 2 + 5 + 0 + 7 + 2 + 5 + 6)/11 = 4$.

As the formula for the mean involves a sum, it only makes sense for quantitative data. Both mean and median can sensibly be called an "average" for data.

A mean incorporates all the data and is a genuine summary; however, it is not always the right choice of summary statistic, and can easily be distorted by extremely high or low values.

> "The mean is like a loaded gun, which in the inexperienced hand can lead to serious accidents, as means can give hopelessly distorted results"
> Karl Pearson, 1857–1936

# Top Hat Course Code 018138

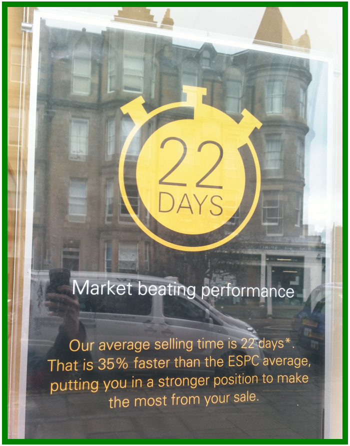
## Market Beating Performance

This poster for *Coulters Property Sales* in Marchmont advertises:

> Our average selling time is 22 days

What do you think is the most suitable average to use here? Which one is the least suitable?

- Mode
- Median
- Mean



Our average selling time is 22 days*.
That is 35% faster than the ESPC average,
putting you in a stronger position to make
the most from your sale.

## Market Beating Performance

This poster for *Coulters Property Sales* in Marchmont advertises:

> Our average selling time is 22 days

I recommend median as the most informative for customers. Some applications might use mean; I can think of no use for the mode here.

1. Median
2. Mean
3. Mode

I don't know what Coulters actually use.

22
DAYS
Market beating performance
Our average selling time is 22 days*.
That is 35% faster than the ESPC average,
putting you in a stronger position to make
the most from your sale.

# Variance and Standard Deviation

Given data values $\{x_1, x_2, \dots, x_N\}$ with mean $\mu$, their *variance* $\sigma^2$ is the mean square deviation from $\mu$:

$$\text{Variance} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 = \left(\frac{1}{N}\sum_{i=1}^{N}x_i^2\right) - \mu^2$$

Not obvious, but the algebra works.

Variance measures the spread of data, but it changes as the square of the data. A more common measure of spread is its square root, known as the *standard deviation* $\sigma$:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2} = \sqrt{\left(\frac{1}{N}\sum_{i=1}^{N}x_i^2\right) - \mu^2}$$

As for the mean, standard deviation makes sense only for quantitative data.

# Example

For the integer data set $\{6, 2, 3, 6, 2, 5, 0, 7, 2, 5, 6\}$ we compute:

$$
\begin{aligned}
\text{Variance} &= \frac{\begin{array}{r}(6-4)^2 + (2-4)^2 + (3-4)^2 + (6-4)^2 + (2-4)^2 + (5-4)^2 \\ + (0-4)^2 + (7-4)^2 + (2-4)^2 + (5-4)^2 + (6-4)^2\end{array}}{11} \\
&= \frac{4+4+1+4+4+1+16+9+4+1+4}{11} \\
&= \frac{52}{11} \\
&= 4.73 \quad \text{to 3 significant figures} \\
\sigma &= \sqrt{\frac{52}{11}} \\
&= 2.17 \quad \text{to 3 significant figures}
\end{aligned}
$$

# Example

For the integer data set $\{6, 2, 3, 6, 2, 5, 0, 7, 2, 5, 6\}$ we compute:

$$\text{Variance} = \frac{6^2 + 2^2 + 3^2 + 6^2 + 2^2 + 5^2 + 0^2 + 7^2 + 2^2 + 5^2 + 6^2}{11} \; - \; 4^2$$

$$= \frac{4 + 4 + 1 + 4 + 4 + 1 + 16 + 9 + 4 + 1 + 4}{11}$$

$$= \frac{52}{11}$$

$$= 4.73 \quad \text{to 3 significant figures}$$

$$\sigma = \sqrt{\frac{52}{11}}$$

$$= 2.17 \quad \text{to 3 significant figures}$$

# Populations and Samples

So far we have seen different statistics for a given set of data, and how to compute them exactly.

Very often, however, data is only a sample drawn from a larger population, and we really want to know — or find out some information about — the statistic on the whole population. For example:

- Experiments in social sciences where one wants to discover information about some section of society — say, university students.
- Surveys and polls — for marketing, opinion gathering, etc.
- In software design when questioning a number of potential users in order to understand general user requirements.

In such cases it is impractical to obtain exhaustive data about the population as a whole; instead, we must work with a sample.

# Sampling

Sampling from a population needs to be done carefully to ensure analysis of the sample is a reliable basis for estimating properties of the whole population.

- The sample should be chosen at random from the population.
- The sample should be as large as is practically possible (given constraints on gathering data, storing data and calculating with data).

These improve the likelihood that a sample is *representative* of the population, reducing the chance of building *bias* into the sample.

Given a sample, we can calculate its statistical properties, and use that to infer information about similar properties of the whole population.

It is a significant topic in statistics, but beyond this course, to work out how to quantify and maximise the reliability of these techniques.

# Estimating Population Statistics

Suppose we have a sample $\{x_1, \ldots, x_n\}$ of size $n$ from a population of size $N$, where $n \ll N$ (i.e., $n$ is much smaller than $N$).

We use the sample $\{x_1, \ldots, x_n\}$ to estimate statistics for the whole population. These estimates may not be correct; but knowing the sample and population size, we can often make estimates about the errors, too.

For mean, the best estimate of the population mean $\mu$ is in fact the sample mean $m$:

$$m = \frac{1}{n}\sum_{i=1}^{n} x_i$$

This is an *unbiased estimator* — its value, given a random sample, is evenly distributed around the mean of the overall population $\mu$.

$$E(m) = \mu$$

# Estimating Population Variance

The variance and standard deviation of a sample are not appropriate estimates for the equivalent statistics on the population from which the sample is drawn; they turn out to be slightly too small, because a sample will be distributed more closely around its own mean than to the population mean.

The best estimate for the variance of the whole population is

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - m)^2 \, .$$

Note the denominator $(n-1)$ rather than $n$. This is known as the *Bessel correction* and gives an unbiased estimator for the variation of the larger population:

$$E(s^2) = \sigma^2 \, .$$

# Estimating Population Standard Deviation

Using this to estimate the variance of a whole population, based on a sample:

$$s^2 \;=\; \frac{1}{n-1}\sum_{i=1}^{n}(x_i - m)^2$$

gives us an estimate for the standard deviation of the whole population:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - m)^2}$$

Again the denominator $n$ for standard deviation has been replaced by $(n-1)$; and the mean $m$ used is that of the sample, not the (unknown) population mean $\mu$.

# Beware

The use of samples to estimate statistics of a larger population is so common that the formula on the previous slide is very often the one needed, rather than the standard deviation of the sample itself.

Its usage is so widespread that sometimes it is wrongly given as **the** definition of standard deviation.

The existence of two different formulas for calculating standard deviations in different circumstances can lead to confusion. So take care.

Often calculators make both formulas available: as $\sigma_n$ for the formula with denominator $n$; and $\sigma_{n-1}$ for the formula with denominator $(n-1)$.

# Data in Multiple Dimensions

So far we have 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 and Causation

## Correlation Does Not Imply Causation

If we do observe a correlation between variables X and Y, it may 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.

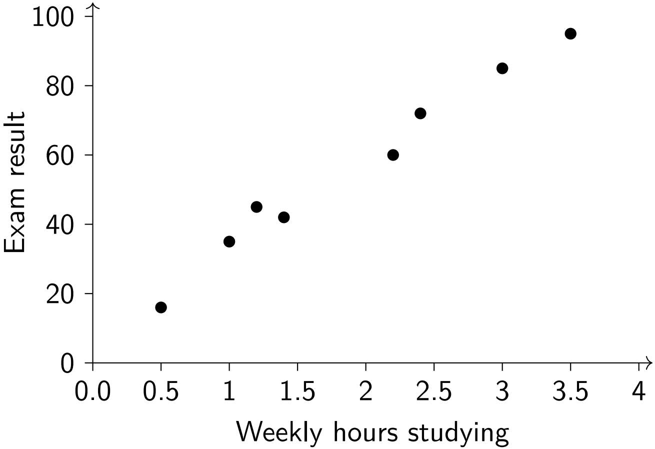
For data like that below, we can draw a *scatter plot* taking one variable as the $x$-axis and one the $y$-axis and plotting a point for each item of data.

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

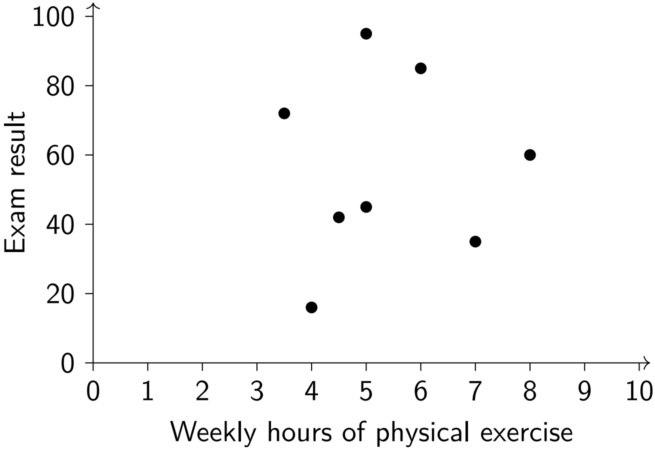
Example Data (NB: Not real students)

| | 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 |

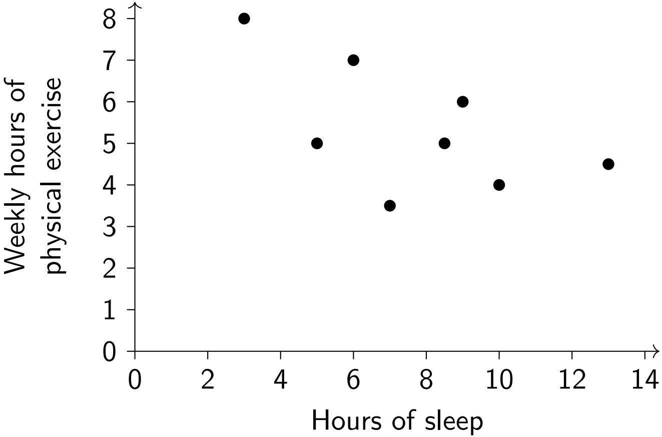
# Studying vs. Exam Results



# Physical Exercise vs. Exam Results



# Hours of Sleep vs. Physical Exercise



# Correlation Coefficient

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

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.

**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 $x_1, \ldots, x_N$ are correlated with another $y_1, \ldots, y_N$.

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.

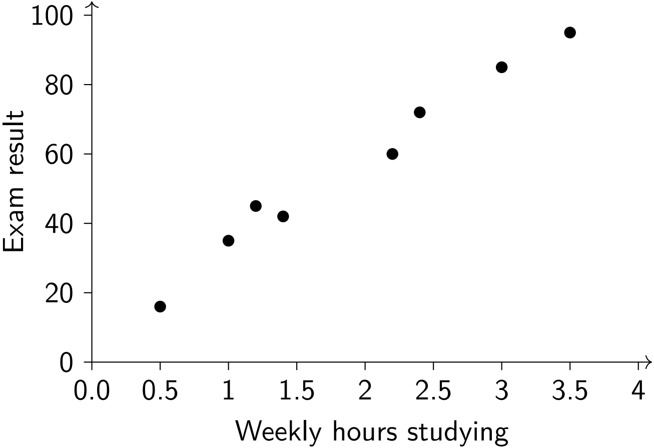
The correlation coefficient $\rho_{x,y}$ is then computed as:

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

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

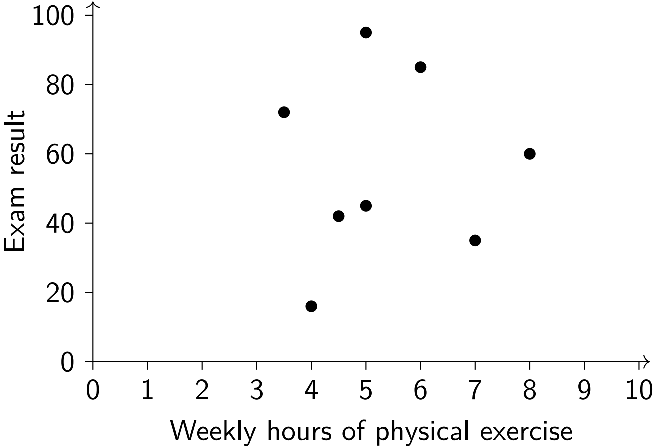
If $\rho_{x,y}$ is close to 0 then this suggests there is no correlation.
If $\rho_{x,y}$ is nearer $+1$ then this suggests $x$ and $y$ are *positively correlated*.
If $\rho_{x,y}$ is closer to $-1$ then this suggests $x$ and $y$ are *negatively correlated*.

# Studying vs. Exam Results



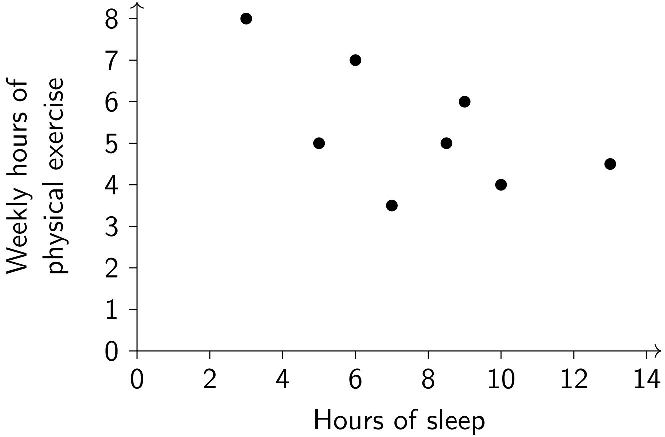
The correlation coefficient is $\rho_{\text{study,exam}} = 0.990$, which suggests a positive correlation: study hours and exam results are high or low together.

# Physical Exercise vs. Exam Results



The correlation coefficient is $\rho_{\text{exercise,exam}} = 0.074$, suggesting no evidence of correlation for these 8 students.

# Hours of Sleep vs. Physical Exercise



The correlation coefficient is $\rho_{\text{sleep,exercise}} = -0.599$. Is this evidence of negative correlation, or just chance?

# Estimating Correlation from a Sample

Suppose that we have sample data $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ 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

Normal Distribution: Bell curve, fixed by mean and standard deviation

Statistic: Single value computed from a set of data

Averages: mean, median, mode

Spread of Data: variance, standard deviation

Sample: Chosen at random from population

Estimates: Population statistics from data about a sample

- Mean: Use mean of the sample
- Variance, Standard Deviation: Use the $(n-1)$ versions.

Correlation: When two datasets vary together. Is not causation.

Correlation Coefficient: Measures correspondence between two datasets.

# What next

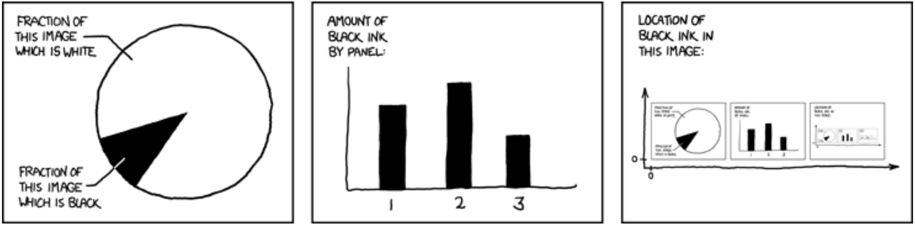
## Do This

Complete the coursework assignment. Write out your solutions to all three questions, staple them together, and post in the box outside the ITO in Forrest Hill before 4pm on Thursday 23 March.

## Read This

The way you're revising may let you down in exams — and here's why
Tom Stafford
The Guardian, 7 May 2016 https://is.gd/ways2learn

# Self-Referential Statistics


FRACTION OF THIS IMAGE WHICH IS WHITE
FRACTION OF THIS IMAGE WHICH IS BLACK
AMOUNT OF BLACK INK BY PANEL:
1
2
3
LOCATION OF BLACK INK IN THIS IMAGE:
0
0

XKCD: 688 Self-Description

Q. Could you construct this?

Can you do it without machine assistance?

Is there more than one way it can come out?