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.
Lecture Timetable

This is Teaching Week 10 of Semester 2, next week is Week 11, and the teaching block ends on Friday 7 April

**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

Monday 3 April – Wednesday 5 April: Final tutorial. return of coursework assignment, feedback and discussion on that.
What’s Happened So Far?

Statistics

A statistic is a single value capturing some overall property of a dataset. Given a random sample from a larger population we may be able to compute an estimate of a statistic for the whole population.

Correlation

A multidimensional dataset has multiple data values for each of a series of items or events. These values are correlated if they vary in similar ways. Where there is a causal dependency between data values, they will be correlated, but the reverse is not true: correlation does not imply causation.

Statistical Tests

The use of hypothesis testing can detect correlations in data. For this we: identify a null hypothesis; compute an appropriate statistic; test whether the statistic provides evidence to reject the null hypothesis.
End of Last Lecture

Do This

Find statistically significant results. Analyse 60 years of data on the US economy to see the effect of having Republicans or Democrats in power.

https://projects.fivethirtyeight.com/p-hacking/

Read This

Science Isn’t Broken
Christie Aschwanden
FiveThirtyEight: Science, August 2015
https://fivethirtyeight.com/features/science-isnt-broken/
What Now?

Data Scales

Refining qualitative vs. quantitative.

Appropriate visualizations: bar chart vs. histogram.

Some Bad Ways With Statistics

- What’s the problem with statistical significance?
- Famous bad examples of correlation.
- Big data makes everything worse.
- Problems in reproducibility and the replication crisis.
- What can be done?
Data Scales

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

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

Data may be *qualitative* (descriptive) or *quantitative* (numerical).

We can refine this further into different kinds of *data scale*:

- Qualitative data may be drawn from a *categorical* or an *ordinal* scale;
- Quantitative data may lie on an *interval* or a *ratio* scale.

Each of these supports different kinds of analyses.
Categorical Scales

Data on a **categorical scale** has each item of data being drawn from a fixed number of categories.

**Example: Categorical Scale**

Students graduating from the University of Edinburgh receive their award at one of several ceremonies, depending on the degree subject they have studied. This classification is a categorical scale: the categories are all the different possible degree programmes.

**Example: Categorical Scale**

Insurance companies classify some insurance applications (e.g., home, possessions, car) according to the alphanumeric postcode of the applicant, making different risk assessments for different postcodes. Here the categories are all existing postcodes.

Categorical scales are sometimes called **nominal scales**, particularly where the categories all have names.
Ordinal Scales

Data on an ordinal scale has a recognized ordering between data items, but there is no meaningful arithmetic on the values.

Example: Ordinal Scale

The European Credit Transfer and Accumulation System (ECTS) has a grading scale where course results are recorded as A, B, C, D, E, FX and F. There are no numerical marks. The ordering is clear, but we can’t add or subtract grades.

Example: Ordinal Scale

The Douglas Sea Scale classifies the state of the sea on a scale from 0 (glassy calm) through 5 (rough) to 9 (phenomenal). This is ordered, but it makes no sense to perform arithmetic: 4 (moderate) is not the mean of 2 (smooth) and 6 (very rough).
Interval Scales

An interval scale is a numerical scale (usually with real number values) in which we are interested in relative value rather than absolute value.

Example: Interval Scale
Moments in time are given relative to an arbitrarily chosen zero point. We can make sense of comparisons such as “date X is 17 years later than date Y”. But it does not make sense to say “arrival time P is twice as large as departure time Q”.

Example: Interval Scale
The Celsius and Fahrenheit temperature scales are interval scales, as the choice of zero is externally imposed.

Mathematically, interval scales support the operations of subtraction and average (all kinds, possibly weighted).

Interval scales do not support either addition or multiplication.
Ratio Scales

A ratio scale is a numerical scale (again usually with real number values) in which there is a notion of absolute value.

Example: Ratio Scales

Most physical quantities such as mass, energy and length are measured on ratio scales. The Kelvin temperature scale is a ratio scale. So is age (of a person, for example), even though it is a measure of time, because there is a definite zero origin.

Thus one object can have half the mass of another; or one person can be twice the age of another person.

Like interval scales, ratio scales support subtraction and weighted averages. They also support addition and multiplication by a real number (a scalar).
## Summary of Scales

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Categorical</td>
<td>Qualitative, fixed set of categories, no order, no possible arithmetic.</td>
<td>Postcodes</td>
</tr>
<tr>
<td>Ordinal</td>
<td>Qualitative, fixed set of categories, can be ordered, still no arithmetic.</td>
<td>Exam grades</td>
</tr>
<tr>
<td>Interval</td>
<td>Quantitative, values all relative; can take averages, subtract one value from another; no addition or multiplication.</td>
<td>Dates</td>
</tr>
<tr>
<td>Ratio</td>
<td>Quantitative, absolute values, can take averages, subtract, add, and take scalar multiples of values.</td>
<td>Mass, energy</td>
</tr>
</tbody>
</table>
Visualising data

It is often helpful to visualise data by drawing a chart or plotting a graph of the data. Visualisations may suggest possible properties of the data, whose existence and features we can then explore mathematically with statistics.

- What kind of visualisations are possible depends on the kind of data.

For a data on a categorical or ordinal scale, a bar chart is appropriate, displaying for each category the number of times it occurs in the data.

- Bars in a bar chart are all the same width, and separate.

For data from an interval or ratio scale, collecting data into bands gives a histogram, showing the frequency with which values occur in the data.

- In a histogram the bars are adjacent, and can be of different widths: it is their area, not height, which measures the number of values.
FOX News Chart Fails Math
Matt Bartosik, NBC Chicago (https://is.gd/foxpie)
Bar Chart vs. Histogram

This is a bar chart

Prisoners / 100,000 population (2005)
Credit: Wikipedia, user XcepticZP

This is a histogram

US commuter travel time (2000)
Credit: Wikipedia, user Qwfp
Far out in the uncharted backwaters of the unfashionable end of the Western Spiral arm of the Galaxy lies a small unregarded yellow sun

Orbiting this at a distance of roughly ninety-two million miles is an utterly insignificant little blue green planet...

The Copernican Principle

J. Richard Gott III.
A Grim Reckoning — What has a 16th-century astronomer got to do with the defeat of governments and the possible extinction of the human race?
New Scientist, 15 November 1997
http://is.gd/grimreckoning

Timothy Ferris.
How to Predict Everything
The New Yorker, 12 July 1999, pp. 35–39
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 one causes variation in the other. If this is so, then they will also be correlated. However, the reverse is not true:

Correlation Does Not Imply Causation
Correlation and 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.
Hypothesis testing explores whether data shows evidence of a correlation.

This starts by identifying a *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 evaluates how likely it is that we would see a result like $R$ — just by chance — if the null hypothesis were true.

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

If the p-value is low then this is evidence to *reject* the null hypothesis.

Often we can consult a table of *critical values* for statistic $R$: if the observed $R$ exceeds a critical value then we know that the p-value is low.
Judge & Cable 2004

The Effect of Physical Height on Workplace Success and Income: Preliminary Test of a Theoretical Model. Journal of Applied Psychology 89(3):428–441

In a sample of over 4000 people this meta-analysis observed positive correlation ($r = 0.31$) between height and earnings in data from the US National Longitudinal Survey. The calculated p-value had $p < 0.01$.

What does $p < 0.01$ tell us about the data?

- Earning more money increases your height.
- There is a 99% chance that height and earnings are correlated.
- If height and earnings are in fact unrelated, then the chance of sample data appearing this closely correlated is less than 1%.
- For any two people chosen at random, there is less than 1% chance that the shorter person is paid more.
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 reject the null hypothesis as a poor explanation for the observed data.

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).
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 — *reproducibility* — 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 ($p = 0.09$)  
...  

g vaguely significant ($p > 0.2$)  
verging on being significant ($p = 0.11$)  
verging on significance ($p = 0.056$)  
verging on the statistically significant ($p < 0.1$)  
...

http://is.gd/stillnotsignificant
And When There is a Correlation?

Famous examples of observed correlations which may or may not be causal.

- Salaries of Presbyterian ministers in Massachusetts
- The price of rum in Havana [Huff, 1954]
- Regular smoking
- Lower class grades [US CDC, 2009]
- The quantity of apples imported into the UK
- The rate of divorce in the UK

R. A. Fisher
Correlation and Causation

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)

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)

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

Warning 2

Correlation Still Does Not Imply Causation
In early 2015, the journal *Basic and Applied Social Psychology* banned:

- 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
Abstract
Reproducibility is a defining feature of science, but the extent to which it characterizes current research is unknown. We conducted replications of 100 experimental and correlational studies published in three psychology journals. Ninety-seven percent of original studies had significant results \((p < .05)\). Thirty-six percent of replications had significant results \(\ldots\)
Many scientific studies can’t be replicated. That’s a problem.
Over half of psychology studies fail reproducibility test

Largest replication study to date casts doubt on many published positive results.

Monya Baker

27 August 2015

http://dx.doi.org/10.1038/nature.2015.18248
Psychology’s reproducibility problem is exaggerated – say psychologists

Reanalysis of last year’s enormous replication study argues that there is no need to be so pessimistic.

Monya Baker

03 March 2016

http://dx.doi.org/10.1038/nature.2016.19498
Taking on chemistry's reproducibility problem

BY DALMEET SINGH CHAWLA | 20 MARCH 2017

Efforts to get to grips with the problem have meant new ideas and technologies are now being brought to bear

Not a week passes without reproducibility in science – or the lack of it – hitting the headlines. Although much of the criticism is directed at the

https://is.gd/chemreplicate
What Hope Is There?

Mathematical analysis and statistical testing remain astonishingly powerful and sensitive tools for scientific discovery. However, they don’t magically make results true all by themselves.

- Don’t rely on a single result
- Don’t go p-hacking or dredging for significance
- Don’t switch outcomes
- Don’t HARK: Hypothesise After Results are Known

✓ Hypotheses need justification, not just statistics: from models, mechanisms, even just previous observations.

✓ Repeatability and reproducibility are essential: and genuine underlying causes will continue to give statistically visible results.

✓ Meta-analysis can help boost all this.
Figure 1. Conventional and Cumulative Meta-Analyses of 33 Trials of Intravenous Streptokinase for Acute Myocardial Infarction. The odds ratios and 95 percent confidence intervals for an effect of treatment on mortality are shown on a logarithmic scale. A bibliography of the published trial reports is available from the authors.
http://www.cochrane.org/
The Power of Meta-Analysis

Ben Goldacre.

*Bad Pharma: How Medicine is Broken, and How We Can Fix It.*

Fourth Estate, 2013.

Side effects may include: anger, outrage, action

http://www.alltrials.net
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

http://alltrials.net
Tracking switched outcomes in clinical trials

Outcome switching in clinical trials is a serious problem. Between October 2015 and January 2016, the COMPare team systematically checked every trial published in the top five medical journals, to see if they misreported their findings. We are now submitting the first set of findings from the project as an academic paper, summarising the quantitative results, and the themes of responses from journal editors and trialists in collaboration with a qualitative researcher. Prior to publication, cite our data and methods as per the reference at the bottom of this page. This is our workflow:

http://compare-trials.org
WE ARE HIRING. Here's our lovely @EBMDataLab team in Oxford, here's your next job making clinical trials better. jobs.ac.uk/job/AYG429/res ...