Informatics 1, 2010

School of Informatics, University of Edinburgh

Data and Analysis

Part III

Unstructured Data

Alex Simpson

Part III — Unstructured Data

Data Retrieval:

III.1 Unstructured data and data retrieval

Statistical Analysis of Data:

III.2 Data scales and summary statistics

III.3 Hypothesis testing and correlation

III.4 χ^2 and collocations

Unstructured data — examples

- Plain (unannotated) text

  There is structure, the sequence of characters, but this is *intrinsic* to the data, not imposed.
  We may wish to impose structure by, e.g., annotating (as in Part II).

- Bitmaps for graphics or pictures, digitized sound, digitized movies, etc.

  These again have *intrinsic* structure (e.g., picture dimensions).
  We may wish to impose structure by, e.g., recognising objects, isolating single instruments from music, etc.

- Experimental results.

  Here there may be structure in how represented (e.g., collection of points in n-dimensional space).
  But an important objective is to uncover implicit structure (e.g., confirm or refute an experimental hypothesis).

Topics

We consider two topics in dealing with unstructured data.

1. *Information retrieval*

   How to find data of intrest in within a collection of unstructured data documents.

2. *Statistical analysis of data*

   How to use statistics to identify and extract properties from unstructured data (e.g., general trends, correlations between different components, etc.)

Information Retrieval

The *Information retrieval (IR) task*: given a query, find the documents in a given collection that are relevant to it.

Assumptions:

1. There is a large document collection being searched.
2. The user has a need for particular information, formulated in terms of a query (typically keywords).
3. The task is to find all and only the documents relevant to the query.

Example: Searching a library catalogue. Document collection to be searched: books and journals in library collection. Information needed: user specifies query giving details about author, title, subject or similar. Search program returns a list of (potentially) relevant matches.

Key issues for IR

Specification issues:

- Evaluation: How to measure the performance of an IR system.
- Query type: How to formulate queries to an IR system.
- Retrieval model: How to find the best-matching document, and how to *rank* them in order of relevance.

Implementation issues:

- Indexing: how to represent the documents searched by the system so that the search can be done efficiently.

The goal of this lecture is to look at the three *specification issues* in more detail.

Evaluation of IR

The performance of an IR system is naturally evaluated in terms of two measures:

- *Precision:* What proportion of the documents returned by the system match the original objectives of the search.
- *Recall:* What proportion of the documents matching the objectives of the search are returned by the system.

We call documents matching the objectives of the search *relevant documents*.

True/false positives/negatives

	Relevant	Non-relevant
Retrieved	true positives	false positives
Not retrieved	false negatives	true negatives

- *True positives (TP):* number of relevant documents that the system retrieved.
- *False positives (FP):* number of non-relevant documents that the system retrieved.
- *True negatives (TN):* number of non-relevant documents that the system did not retrieve.
- *False negatives (FN):* number of relevant documents that the system did not retrieve.

Defining precision and recall

	Relevant	Non-relevant
Retrieved	true positives	false positives
Not retrieved	false negatives	true negatives

Precision

$$P = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Recall

$$R = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Comparing 2 IR systems — example

Document collection with 130 documents.
28 documents relevant for a given theory.

System 1: retrieves 25 documents, 16 of which are relevant

$\mathrm{TP}_1 = 16, \quad \mathrm{FP}_1 = 25 - 16 = 9, \quad \mathrm{FN}_1 = 28 - 16 = 12$

$$P_1 = \frac{\mathrm{TP}_1}{\mathrm{TP}_1 + \mathrm{FP}_1} = \frac{16}{25} = 0.64 \qquad R_1 = \frac{\mathrm{TP}_1}{\mathrm{TP}_1 + \mathrm{FN}_1} = \frac{16}{28} = 0.57$$

System 2: retrieves 15 documents, 12 of which are relevant

$\mathrm{TP}_2 = 12, \quad \mathrm{FP}_2 = 15 - 12 = 3, \quad \mathrm{FN}_2 = 28 - 12 = 16$

$$P_2 = \frac{\mathrm{TP}_2}{\mathrm{TP}_2 + \mathrm{FP}_2} = \frac{12}{15} = 0.80 \qquad R_2 = \frac{\mathrm{TP}_2}{\mathrm{TP}_2 + \mathrm{FN}_2} = \frac{12}{28} = 0.43$$

N.B. System 2 has higher precision. System 1 has higher recall.

Precision versus Recall

A system has to achieve both high precision and recall to perform well. It doesn't make sense to look at only one of the figures:

- If system returns all documents in the collection: 100% recall, but low precision.
- If system returns only one document, which is relevant: 100% precision, but low recall.

Precision-recall tradeoff: System can optimize precision at the cost of recall, or increase recall at the cost of precision.

Whether precision or recall is more important depends on the application of the system.

F-score

The *F-score* is an evaluation measure that combines precision and recall.

$$F_\alpha = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}}$$

Here α is a *weighting factor* with $0 \leq \alpha \leq 1$.

High α means precision more important. Low α means recall is more important.

Often $\alpha = 0.5$ is used, giving the *harmonic mean* of P and R:

$$F_{0.5} = \frac{2PR}{P+R}$$

Using F-score to compare — example

We compare the examples on slide III: 10 using the F-score (with $\alpha = 0.5$).

$$F_{0.5}(\text{System}_1) = \frac{2P_1R_1}{P_1 + R_1} = \frac{2 \times 0.64 \times 0.57}{0.64 + 0.57} = 0.60$$

$$F_{0.5}(\text{System}_2) = \frac{2P_2R_2}{P_2 + R_2} = \frac{2 \times 0.80 \times 0.43}{0.80 + 0.43} = 0.56$$

The F-score (with this weighting) rates System 1 as better than System 2.

Query type

We shall only consider *simple queries* of the form:

- Find documents containing *word1*, *word2*, . . . , *wordn*

More specific tasks are:

- Find documents containing all the words *word1*, *word2* . . . *wordn*;
- or find documents containing as many of the words *word1*, *word2* . . . *wordn* as possible.

In real-world applications, queries can be much more complex than this (e.g., they can be combined using boolean operations, one can search for substrings of words or whole phrases, one can match regular expressions, etc.).

A retrieval model

If all documents containing all words of the query are returned this might result in a large number of varying relevance (at least if the document collection is large and the query general).

IR systems need to rank documents according to likely relevance.

There are many such ranking methods.

We focus on one, which uses the *vector space model*.

This model is the basis of many IR applications.

In this course, we shall only use it in one particularly simple way.

The vector space model

Core ideas:

- Treat documents as points in a high-dimensional vector space, based on words in the document collection.
- The query is treated in the same way.
- The documents are ranked according to document-query similarity.

N.B. You do not need to know anything about vector spaces to understand the approach!

The vector associated to a document

Suppose $\text{Term}_1, \text{Term}_2, \ldots, \text{Term}_n$ are all the different words occurring in the entire collection of documents $\text{Doc}_1, \text{Doc}_2, \ldots, \text{Doc}_N$.

Each document, Doc_i, is assigned an n-valued vector:

$$(m_{i1}, m_{i2}, \ldots, m_{in})$$

where m_{ij} is the number of times word Term_j occurs in document Doc_i.

Similarly, the query is assigned an n-valued vector by considering it as a document itself.

Example

Consider the document

Sun, sun, sun, here it comes

and suppose the only words in the document collection are: *comes*, *here*, *it*, *sun*.

The vector for the document is $(\mathbf{1}, \mathbf{1}, \mathbf{1}, \mathbf{3})$

comes	here	it	sun
1	1	1	3

Similarly, the vector for the query *sun comes* is $(\mathbf{1}, \mathbf{0}, \mathbf{0}, \mathbf{1})$

Document matrix

The frequency information for words in the document collection is normally precompiled in a *document matrix*.

This has:

- Columns represent the words appearing the document collection
- Rows represent each document in the collection.
- each entry in the matrix represents the frequency of the word in the document.

Document matrix — example

	Term_1	Term_2	Term_3	...	Term_n
Doc_1	14	6	1	...	0
Doc_2	0	1	3	...	1
Doc_3	0	1	0	...	2
...	...	...	...	...	...
Doc_N	4	7	0	...	5

N.B. Each row gives the vector for the associated document.

Vector similarity

We want to rank documents according to relevance to the query.

We implement this by defining a measure of *similarity* between vectors.

The idea is that the most relevant documents are those whose vectors are most similar to the query vector.

Many different similarity measures are used. A simple one that is conceptually appealing and enjoys some good properties is the *cosine* of the angle between two vectors.

Cosines (from school trigonometry)

Recall that the *cosine* of an angle θ is:

$$\frac{\text{adjacent}}{\text{hypotenuse}}$$

in a right-angled triangle with angle θ.

Crucial properties:

$$\cos(0) = 1 \qquad \cos(90^\circ) = 0 \qquad \cos(180^\circ) = -1$$

More generally, two n-dimensional vectors will have cosine: 1 if they are identical, 0 if they are orthogonal, and -1 if they point in opposite directions.

The value $\cos(x)$ *always* lies in the range from -1 to 1.

Vector cosines

Suppose $\vec{x}$ and $\vec{y}$ are n-value vectors:

$$\vec{x} = (x_1, \ldots, x_n) \qquad \vec{y} = (y_1, \ldots, y_n)$$

Their *cosine* (that is, the cosine of the angle between them) is calculated by:

$$\cos(\vec{x}, \vec{y}) = \frac{\vec{x} \cdot \vec{y}}{|\vec{x}||\vec{y}|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}$$

Vector cosines — example

Continuing the example from slide 11.18, suppose:

$$\vec{x} = (1, 1, 1, 3) \qquad \vec{y} = (1, 0, 0, 1)$$

Then:

$$\vec{x} \cdot \vec{y} = \ 1 + 0 + 0 + 3 \ = 4$$
$$|\vec{x}| = \sqrt{1 + 1 + 1 + 9} = \sqrt{12}$$
$$|\vec{y}| = \sqrt{1 + 0 + 0 + 1} = \sqrt{2}$$

So

$$\cos(\vec{x}, \vec{y}) = \frac{4}{\sqrt{12} \times \sqrt{2}} = \frac{2}{\sqrt{6}} = 0.82$$

to two decimal places.

Ranking documents

Suppose $\vec{y}$ is the query vector, and $\vec{x_1}, \ldots, \vec{x_N}$ are the N document vectors.

We calculate the N values:

$$\cos(\vec{x_1}, \vec{y}), \quad \ldots \quad , \cos(\vec{x_N}, \vec{y})$$

The documents are then ordered so that those with the highest cosine values are counted as most suitable, and those with the lowest cosine values are counted as least suitable.

N.B. On this slide $\vec{x_1}, \ldots, \vec{x_N}$ are N (potentially) different vectors, each with n values.

Discussion of cosine measure

The cosine similarity measure, as discussed here, is very crude.

- It only takes word frequency into account
- It takes all words in the document collection into account (whether very common "stop" words which are useless for IR, or very uncommon words unrelated to the search)
- All words in the document collection are weighted equally
- It ignores document size (just the angles between vectors not their magnitude are considered)

Nevertheless, the cosine method can be refined in various ways to avoid these problems. (This is beyond the scope of this course.)

Other issues

- Precision and recall, as defined, only evaluate the set of documents returned, they do not take *ranking* into account. Other more complex evaluation measures can be introduced to deal with ranking (e.g., *precision at a cutoff*).
- We have not considered the efficient implementation of the search for documents matching a query. This is often addressed using a purpose-built index such as an *inverted index* which indexes all documents using the words in the document collection as keys.
- Often useful ranking methods make use of information extraneous to the document itself. E.g., Google's *pagerank* method evaluates documents according to their degree of *connectivity* with the rest of the web (e.g., number of links to page from other pages).

These are important issues, but are beyond the scope of this course.

Part III — Unstructured Data

Data Retrieval:

III.1 Unstructured data and data retrieval

Statistical Analysis of Data:

III.2 Data scales and summary statistics

III.3 Hypothesis testing and correlation

III.4 χ^2 and collocations

Analysis of data

There are many reasons to *analyse* data.

Two common goals of analysis:

- Discover implicit structure in the data.

  E.g., find patterns in empirical data (such as experimental data).

- Confirm or refute a hypothesis about the data.

  E.g., confirm or refute an experimental hypothesis.

Statistics provides a powerful and ubiquitous toolkit for performing such analyses.

Data scales

The type of analysis performed (obviously) depends on:

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

  For example, the data may be *quantitative* (i.e., numerical), or it may be *qualitative* (i.e., descriptive).

One important aspect of the kind of data is the form of *data scale* it belongs to:

- *Categorical* (also called *nominal*) and *Ordinal* scales (for qualitative data).
- *Interval and ratio* scales (for quantitative data).

This affects the ways in which we can manipulate data.

Categorical scales

Data belongs to a *categorical scale* if each *datum* (i.e., data item) is classified as belonging to one of a fixed number categories.

Example: The British Government (presumably) classifies Visa applications according to the nationality of the applicant. This classification is a categorical scale: the categories are the different possible nationalities.

Example: Insurance companies classify some insurance applications (e.g., home, possessions, car) according to the postcode of the applicant (since different postcodes have different risk assessments).

Categorical scales are sometimes called *nominal scales*, especially in cases in which the value of a datum is a name.

Ordinal scales

Data belongs to an *ordinal scale* if it has an associated ordering but arithmetic transformations on the data are not meaningful.

Example: The *Beaufort wind force scale* classifies wind speeds on a scale from **0** (calm) to **12** (hurricane). This has an obvious associated ordering, but it does not make sense to perform arithmetic operations on this scale. E.g., it does not make much sense to say that scale **6** (strong breeze) is the average of calm and hurricane force.

Example: In many institutions, exam marks are recorded as grades (e.g., A,B,. . . , G) rather than as marks. Again the ordering is clear, but one does not perform arithmetic operations on the scale.

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: Points in time are given relative to an arbitrarily chosen zero point. We can make sense of comparisons such as: moment $\boldsymbol{x}$ is 2009 years later than moment $\boldsymbol{y}$. But it does not make sense to say: moment $\boldsymbol{x}$ is twice as large as moment $\boldsymbol{z}$.

Mathematically, interval scales support the operations of subtraction (returning a real number for this) and weighted average.

Interval scales do not support the operations of addition and 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: Most physical quantities such as mass, energy and length are measured on ratio scales. So is temperature if measured in kelvins (i.e. relative to absolute zero).

Like interval scales, ratio scales support the operations of subtraction and weighted average. They also support the operations of addition and of multiplication by a real number.

Question for physics students: Is time a ratio scale if one uses the Big Bang as its zero point?

Visualising data

It is often helpful to *visualise* data by drawing a *chart* or plotting a *graph* of the data.

Visualisations can help us guess properties of the data, whose existence we can then explore mathematically using statistical tools.

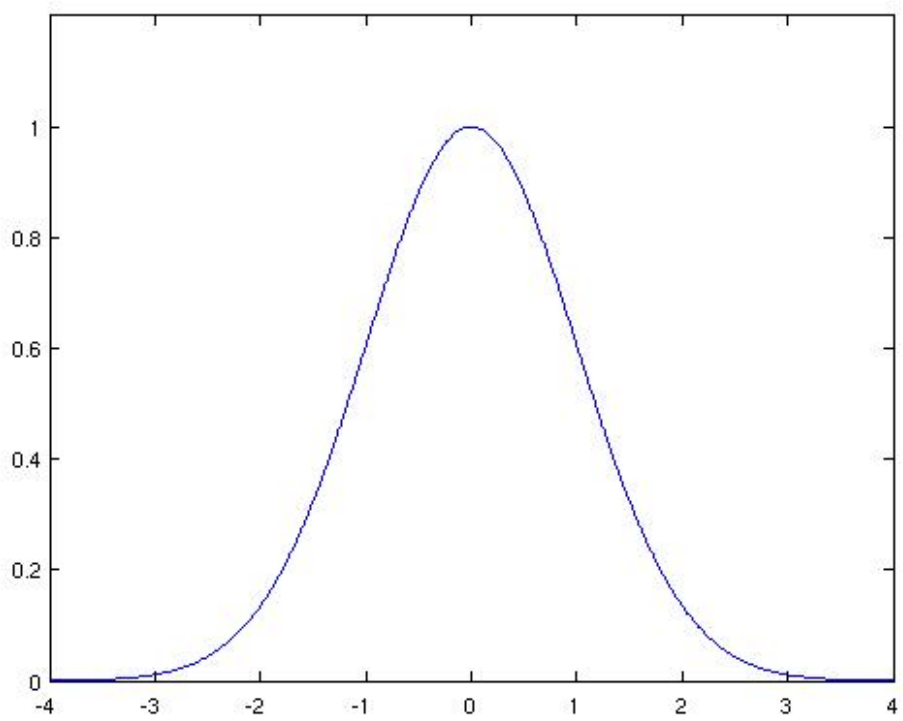
For a collection of data of a categorical or ordinal scale, a natural visual representation is a *histogram* (or *bar chart*), which, for each category, displays the number of occurrences of the category in the data.

For a collection of data from an interval or ratio scale, one plots a *graph* with the data scale as the x-axis and the frequency as the y-axis.

It is very common for such a graph to take a bell-shaped appearence.

Normal distribution

In a *normal distribution*, the data is clustered symmetrically around a central value (zero in the graph below), and takes the bell-shaped appearance below.



Normal distribution (continued)

There are two crucial values associated with the normal distribution.

The *mean*, μ, is the central value around which the data is clustered. In the example, we have $\mu = 0$.

The *standard deviation*, σ, is the distance from the mean to the point at which the curve changes from being *convex* to being *concave*. In the example, we have $\sigma = 1$. The larger the standard deviation, the larger the *spread* of data.

The general equation for a normal distribution is

$$y = c\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

(You do not need to remember this formula.)

Statistic(s)

A *statistic* is a (usually numerical) value that captures some property of 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 data that is not normally distributed.

There are also other, *mode* and *median*, which are alternatives to the mean for capturing the "focal point" of data.

Mode

Summary statistics summarise a property of a data set in a single value.

Given data values $x_1, x_2, \ldots, x_N$, the *mode* (or *modes*) is the value (or values) x that occurs most often in $x_1, x_2, \ldots, x_N$.

Example: Given data: $6, 2, 3, 6, 1, 5, 1, 7, 2, 5, 6$, the mode is 6, which is the only value to occur three times.

The mode makes sense for all types of data scale. However, it is not particularly informative for real-number-valued quantitative data, where it is unlikely for the same data value to occur more than once.
(This is an instance of a more general phenomenon. In many circumstances, it is neither useful nor meaningful to compare real-number values for equality.)

Median

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

Example: Given data: $6, 2, 3, 6, 1, 5, 1, 7, 2, 5, 6$, we write this in non-decreasing order:

$$1, 1, 2, 2, 3, 5, 5, 6, 6, 6, 7$$

The middle value is the sixth value 5.

The median makes sense for ordinal data and for interval and ratio data. It does not make sense for categorical data, because categorical data has no associated order.

Mean

Given data values $x_1, x_2, \ldots, x_N$, the *mean* μ is the value:

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

Example: Given data: $6, 2, 3, 6, 1, 5, 1, 7, 2, 5, 6$, the mean is

$$\frac{6+2+3+6+1+5+1+7+2+5+6}{11} = 4.$$

Although the formula for the mean involves a sum, the mean makes sense for both interval and ratio scales. The reason it makes sense for data on an interval scale is that interval scales support *weighted averages*, and a mean is simply an equally-weighted average (all weights are set as $\frac{1}{N}$).

The mean does *not* make sense for categorical and ordinal data.

Variance and standard deviation

Given data values $x_1, x_2, \ldots, x_N$, with mean μ, the *variance*, written **Var** or σ^2, is the value:

$$\mathbf{Var} = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

The *standard deviation*, written σ, is defined by:

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

Like the mean, the standard deviation makes sense for both interval and ratio data. (The values that are squared are real numbers, so, even with interval data, there is no issue about performing the multiplication.)

Variance and standard deviation (example)

Given data: $\mathbf{6, 2, 3, 6, 1, 5, 1, 7, 2, 5, 6}$, we have $\boldsymbol{\mu = 4}$.

$$\begin{aligned}
\mathbf{Var} &= \frac{\mathbf{2^2 + 2^2 + 1^2 + 2^2 + 3^2 + 1^2 + 3^2 + 3^2 + 2^2 + 1^2 + 2^2}}{\mathbf{11}} \\
&= \frac{\mathbf{4 + 4 + 1 + 4 + 9 + 1 + 9 + 9 + 4 + 1 + 4}}{\mathbf{11}} \\
&= \frac{\mathbf{50}}{\mathbf{11}} \\
&= \mathbf{4.55} \quad \text{(to 2 decimal places)}
\end{aligned}$$

$$\begin{aligned}
\boldsymbol{\sigma} &= \sqrt{\frac{\mathbf{50}}{\mathbf{11}}} \\
&= \mathbf{2.13} \quad \text{(to 2 decimal places)}
\end{aligned}$$

Populations and samples

The discussion of statistics so far has been all about computing various statistics for a given set of data.

Very often, however, one is interested in knowing the value of the statistic for a whole *population* from which our data is just a *sample*.

Examples:

- Experiments in social sciences where one wants to discover some general property of a section of the population (e.g., teenagers).
- Surveys (e.g., marketing surveys, opinion polls, etc.).
- In software design, understanding requirements of users, based on questioning a sample of potential users.

In such cases it is totally impracticable to obtain exhaustive data about the population as a whole. So we are forced to obtain data about a sample.

Sampling

There are important guidelines to follow in choosing a sample from a population.

- The sample should be chosen *randomly* 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 two guidelines are designed to improve the likelihood that the sample is *representative* of the population. In particular, they minimise the chance of accidentally building a *bias* into the sample.

Given a sample, one calculates statistical properties of the sample, and uses these to infer likely statistical properties of the whole population.

Important topics in statistics (beyond the scope of D&A) are *maximising* and *quantifying* the reliability of such techniques.

Estimating statistics for a population given a sample

Tyically one has a (hopefully representative) sample $x_1, \ldots, x_n$ from a population of size N where $n \ll N$ (i.e., n is much smaller that N).

We use the sample $x_1, \ldots, x_n$ to estimate statistical values for the whole population.

Sometimes the calculation is the expected one, sometimes it isn't.

The best estimate m of the *mean* μ of the population is:

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

As expected, this is just the mean of the sample.

Estimating variance and standard deviation of population

To estimate the *variance* of the population, calculate

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

The best estimate s of the *standard deviation* σ of the population, is:

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

N.B. These values are *not* simply the variance and standard deviation of the sample. In both cases, the expected denominator of n has been replaced by $n-1$. This gives a better estimate in general when $n \ll N$.

Caution

The use of samples to estimate statistics of populations is so common that the formula on the previous slide is very often the one needed when calculating standard deviations.

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 the standard deviation in different circumstances can lead to confusion. So one needs to take care.

Sometimes calculators make both formulas available via two buttons: $\boldsymbol{\sigma_n}$ for the formula with denominator $\boldsymbol{n}$; and $\boldsymbol{\sigma_{n-1}}$ for the formula with denominator $\boldsymbol{n-1}$.

Further reading

There are many, many, many books on statistics. Two very gentle books, intended mainly for social science students, are:

P. Hinton
Statistics Explained
Routledge, London, 1995

First Steps in Statistics
D. B. Wright
SAGE publications, 2002

These are good for the formula-shy reader.

Two entertaining books (the first a classic, the second rather recent), full of examples of how statistics are often misused in practice, are:

D. Huff
How to Lie with Statistics
Victor Gollancz, 1954

M. Blastland and A. Dilnot
The Tiger That Isn't
Profile Books, 2008

Part III — Unstructured Data

Data Retrieval:

III.1 Unstructured data and data retrieval

Statistical Analysis of Data:

III.2 Data scales and summary statistics

III.3 Hypothesis testing and correlation

III.4 χ^2 and collocations

Several variables

Often, one wants to relate data in several variables (i.e., multi-dimensional data).

For example, the table below tabulates, for eight students (A–H), their weekly time (in hours) spent: studying for Data & Analysis, drinking and eating. This is juxtaposed with their Data & Analysis exam results.

	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

Thus, we have four variables: study, drinking, eating and exam.
(This is four-dimensional data.)

Correlation

We can ask if there is any *relationship* between the values taken by two 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*.

Caution: A correlation does *not* imply a *causal relationship* between one variable and another. For example, there is a positive correlation between incidences of lung cancer and time spent watching television, but neither causes the other.

However, in cases in which there *is* a causal relationship between two variables, then there often will be an associated correlation between the variables.

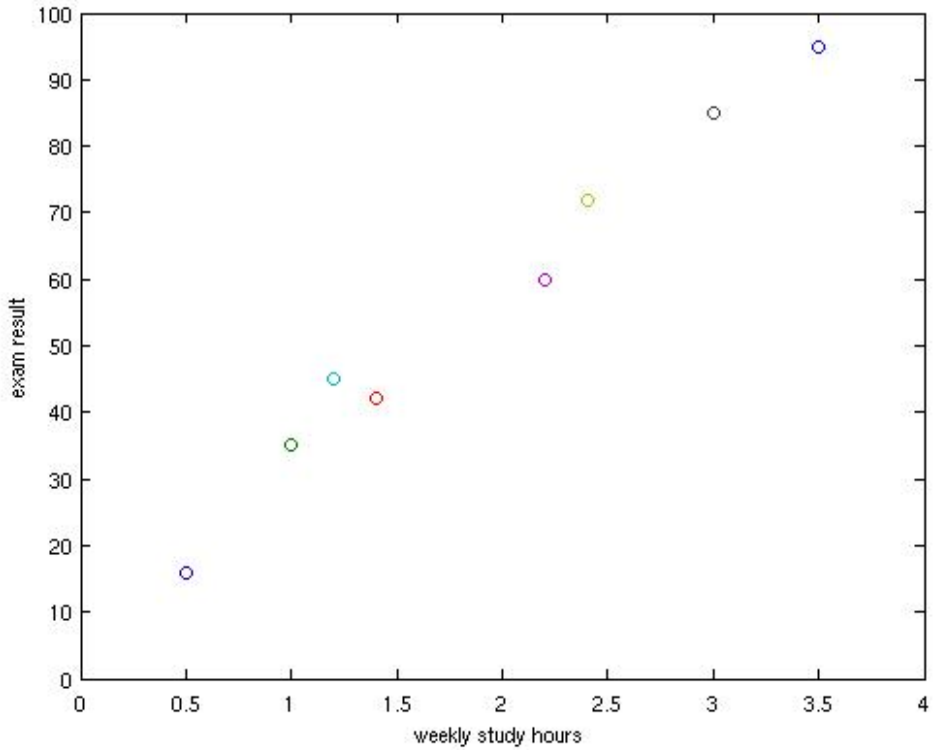
Visualising correlations

One way of discovering correlations is to visualise the data.

A simple visual guide is to draw a *scatter plot* using one variable for the x-axis and one for the y-axis.

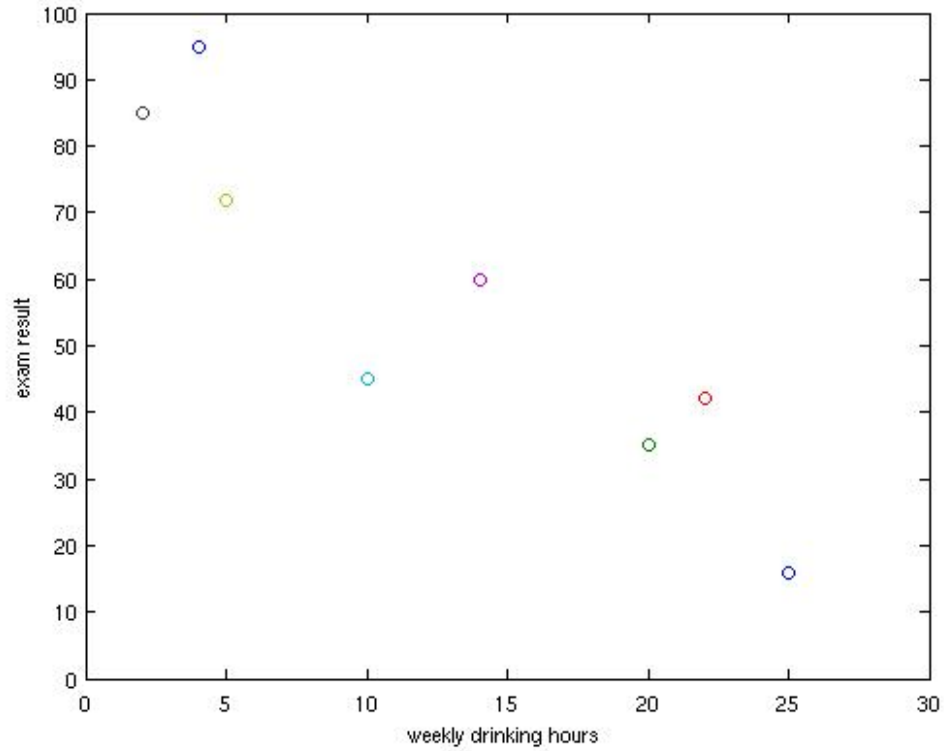
Example: In the example data on Slide III: 51, is there a correlation between study hours and exam results? What about between drinking hours and exam results? What about eating and exam results?

Studying vs. exam results



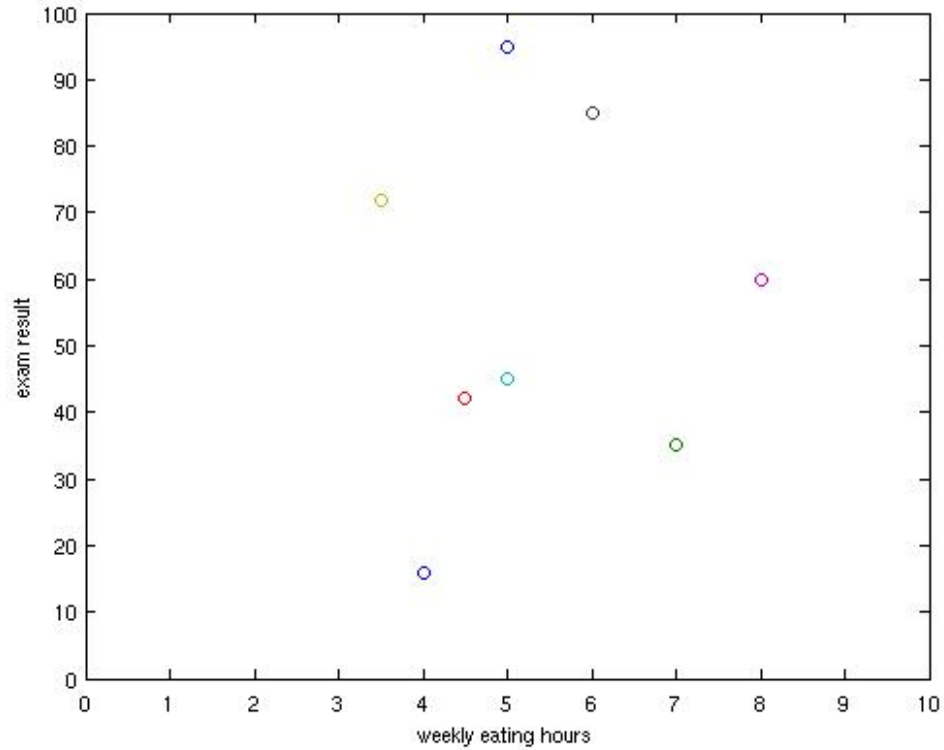
This looks like a *positive* correlation.

Drinking vs. exam results



This looks like a *negative* correlation.

Eating vs. exam results



There is no obvious correlation.

Statistical hypothesis testing

The last three slides use data visualisation as a tool for postulating hypotheses about data.

One might also postulate hypotheses for other reasons, e.g.: intuition that a hypothesis may be true; a perceived analogy with another situation in which a similar hypothesis is known to be valid; existence of a theoretical model that makes a prediction; etc.

Statistics provides the tools needed to corroborate or refute such hypotheses with scientific rigour: *statistical tests*.

The general form of a statistical test

One applies an appropriately chosen statistical test to the data and calculates the result $\boldsymbol{R}$.

Statistical tests are usually based on a *null hypothesis* that there is nothing out of the ordinary about the data.

The result $\boldsymbol{R}$ of the test has an associated *probability value* $\boldsymbol{p}$.

The value $\boldsymbol{p}$ represents the probability that we would obtain a result similar to $\boldsymbol{R}$ if the null hypothesis were true.

N.B., $\boldsymbol{p}$ is *not* the probability that the null hypothesis is true. This is not a quantifiable value.

The general form of a statistical test (continued)

The value $\boldsymbol{p}$ represents the probability that we would obtain a result similar to $\boldsymbol{R}$ if the null hypothesis were true.

If the value of $\boldsymbol{p}$ is *significantly small* then we conclude that the null hypothesis is a poor explanation for our data. Thus we *reject* the null hypothesis, and replace it with a better explanation for our data.

Standard *significance thresholds* are to require $\boldsymbol{p < 0.05}$ (i.e., there is a less than $\mathbf{1/20}$ chance that we would have obtained our test result were the null hypothesis true) or, better, $\boldsymbol{p < 0.01}$ (i.e., there is a less than $\mathbf{1/100}$ chance)

Correlation coefficient

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

Let μ_x and σ_x be the mean and standard deviation of the x values.
Let μ_y and σ_y be the mean and standard deviation of the y values.

The corelation coefficient $\rho_{x,y}$ is defined by:

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

If $\rho_{x,y}$ is positive this suggests x, y are *positively correlated*.
If $\rho_{x,y}$ is negative this suggests x, y are *negatively correlated*.
If $\rho_{x,y}$ is close to 0 this suggests there is no correlation.

Correlation coefficient as a statistical test

In a test for correlation between two variables x, y (e.g., exam result and study hours), we are looking for a correlation and a direction for the correlation (either negative or positive) between the variables.

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

We calculate the correlation coefficient $\rho_{x,y}$.

We then look up significance in a *critical values table* for the correlation coefficient. Such tables can be found in statistics books (and on the Web). This gives us the associated probability value p.

The value of p tells us whether we have significant grounds for rejecting the null hypothesis, in which case our better explanation is that there *is* a correlation.

Critical values table for the correlation coefficient

The table has rows for N values and columns for p values.

N	$p = 0.1$	$p = 0.05$	$p = 0.01$	$p = 0.001$
7	0.669	0.754	0.875	0.951
8	0.621	0.707	0.834	0.925
9	0.582	0.666	0.798	0.898

The table shows that for $N = 8$ a value of $|\rho_{x,y}| > 0.834$ has probability $p < 0.01$ of occurring (that is less than a $1/100$ chance of occurring) if the null hypothesis is true.

Similarly, for $N = 8$ a value of $|\rho_{x,y}| > 0.925$ has probability $p < 0.001$ of occurring (that is less than a $1/1000$ chance of occurring) if the null hypothesis is true.

Studying vs. exam results

We use the data from III: 51 (see also III: 54), with the study values for $x_1, \ldots, x_N$, and the exam values for $y_1, \ldots, y_N$, where $N = 8$.

The relevant statistics are:

$$\mu_x = 1.9 \qquad \sigma_x = 0.981$$
$$\mu_y = 56.25 \qquad \sigma_y = 24.979$$
$$\rho_{x,y} = 0.985$$

Our value of 0.985 is (much) higher than the critical value 0.925. Thus we reject the null hypothesis with very high confidence ($p < 0.001$) and conclude that there is a correlation.

It is a *positive correlation* since $\rho_{x,y}$ is positive not negative.

Drinking vs. exam results

We now use the drinking values from III: 51 (see also III: 55) as the values for $x_1, \ldots, x_8$. (The y values are unchanged.)

The new statistics are:

$$\mu_x = 12.75 \qquad \sigma_x = 8.288 \qquad \rho_{x,y} = -0.914$$

Since $|-0.914| = 0.914 > 0.834$, we can reject the null hypothesis with confidence ($p < 0.01$). This result is still significant though less so than the previous.

This time, the value -0.914 of $\rho_{x,y}$ is negative so we conclude that there is a *negative correlation*

Estimating correlation from a sample

As on slides III: 46–47, assume samples $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ from a population of size N where $n \ll N$.

Let m_x and m_y be the estimates of the means of the x and y values (V: 46)
Let s_x and s_y be the estimates of the standard deviations (V: 47)

The best estimate $r_{x,y}$ of the correlation coefficient is given by:

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

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

Correlation coefficient — subtleties

The correlation coefficient measures how close a scatter plot of x, y values is to a straight line. Nonetheless, a high correlation does not mean that the relationship between x, y is linear. It just means it can be reasonably closely approximated by a linear relationship.

Critical value tables for the correlation coefficient are often given with rows indexed by *degrees of freedom* rather than by N. For the correlation coefficient, the number of *degrees of freedom* is $N - 2$, so it is easy to translate such a table into the form given here. (The notion of degree of freedom, in the case of correlation, is too advanced a concept for D&A.)

Also, critical value tables often have two classifications: one for *one-tailed tests* and one for *two-tailed tests*. Here, we are applying a *two-tailed test*: we consider both positive and negative values as significant. In a *one-tailed* test, we would be interested in just one of these possibilities.

Part III — Unstructured Data

Data Retrieval:

III.1 Unstructured data and data retrieval

Statistical Analysis of Data:

III.2 Data scales and summary statistics

III.3 Hypothesis testing and correlation

III.4 χ^2 **and collocations**

The χ^2 test

While the correlation coefficient, introduced in the previous lecture, is a useful statistical test for correlation, it is applicable only to numerical data (both interval and ratio scales).

The χ^2 *(chi-squared) test* is a general tool for investigating correlations between *categorical data*.

We shall illustrate the χ^2 test with the following example.

> Is there any correlation, in a class of students enrolled on a course, between submitting the coursework for the course and attending the course exam?

General approach

The investigation will conform to the usual pattern of a statistical test.

The *null hypothesis* is that there is no relationship between coursework submission and exam attendance.

The χ^2 test will allow us to compute the probability p that the data we see might occur were the null hypothesis true.

Once again, if p is signifcantly low, we reject the null hypothesis, and we conclude that there is a relationship between coursework submission and exam attendance.

To begin, we use the data to compile a *contingency table* of *frequency observations* O_{ij}.

Contingency table

O_{ij}	sub	¬sub
att	O_{11}	O_{12}
¬att	O_{21}	O_{22}

O_{11} is number of students who submitted coursework and attended the exam.

O_{12} is number of students who did not submit coursework, but attended the exam.

O_{21} is number of students who submitted coursework, but did not attend the exam.

O_{22} is number of students who neither submitted coursework nor attended exam.

Worked example

O_{ij}	sub	¬sub
att	$O_{11} = 94$	$O_{12} = 20$
¬att	$O_{21} = 2$	$O_{22} = 15$

O_{11} is number of students who submitted coursework and attended the exam.

O_{12} is number of students who did not submit coursework, but attended the exam.

O_{21} is number of students who submitted coursework, but did not attend the exam.

O_{22} is number of students who neither submitted coursework nor attended exam.

Idea of χ^2 test

The observations O_{ij} are the actual data frequencies

We use these to calculate *expected frequencies* E_{ij}, i.e., the frequencies we would have expected to see were the null hypothesis true.

The χ^2 test is calculated by comparing the actual frequency to the expected frequency.

The larger the disrepancy between these two values, the more improbable it is that the data could have arisen were the null hypothesis true.

Thus a large discrepancy allows us to reject the null hypothesis and conclude that there is likely to be a correlation.

Marginals

To compute the expected frequencies, we first compute the *marginals* R_1, R_2, B_1, B_2 of the observation table.

O_{ij}	sub	¬sub	
att	O_{11}	O_{12}	$R_1 = O_{11} + O_{12}$
¬att	O_{21}	O_{22}	$R_2 = O_{21} + O_{22}$
	$B_1 = O_{11} + O_{21}$	$B_2 = O_{12} + O_{22}$	N

Here

$$N = R_1 + R_2 = B_1 + B_2$$

Marginals explained

The marginals and N are very simple.

- B_1 is the number of students who submitted coursework.
- B_2 is the number of students who did not submit coursework.
- R_1 is the number of students who attended the exam.
- R_2 is the number of students who did not attend the exam.
- N is the total number of students registered for the course.

Given these figures, if there were no relationship between submitting coursework and attending the exam, we would expect the number of students doing both to be

$$\frac{B_1 R_1}{N}$$

Expected frequencies

The *expected frequencies* E_{ij} are now calculated as follows.

E_{ij}	sub	¬sub	
att	$E_{11} = B_1 R_1 / N$	$E_{12} = B_2 R_1 / N$	$R_1 = E_{11} + E_{12}$
¬att	$E_{21} = B_1 R_2 / N$	$E_{22} = B_2 R_2 / N$	$R_2 = E_{21} + E_{22}$
	$B_1 = E_{11} + E_{21}$	$B_2 = E_{12} + E_{22}$	N

Notice that this table has the same marginals as the original.

The χ^2 value

We can now define the χ^2 value by:

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

$$= \frac{(O_{11}-E_{11})^2}{E_{11}} + \frac{(O_{12}-E_{12})^2}{E_{12}} + \frac{(O_{21}-E_{21})^2}{E_{21}} + \frac{(O_{22}-E_{22})^2}{E_{22}}$$

N.B. It is always the case that:

$$(O_{11}-E_{11})^2 = (O_{12}-E_{12})^2 = (O_{21}-E_{21})^2 = (O_{22}-E_{22})^2$$

This fact is helpful in simplifying χ^2 calculations.

Mathematical Exercise. Why are these 4 values always equal?

Worked example (continued)

Marginals:

O_{ij}	sub	¬sub	
att	**94**	**20**	**114**
¬att	**2**	**15**	**17**
	96	**35**	**131**

Expected values:

E_{ij}	sub	¬sub	
att	**83.542**	**30.458**	**114**
¬att	**12.458**	**4.542**	**17**
	96	**35**	**131**

Worked example (continued)

$$\begin{aligned}
\chi^2 &= \frac{10.458^2}{83.542} + \frac{10.458^2}{30.458} + \frac{10.458^2}{12.458} + \frac{10.458^2}{4.542} \\
&= \frac{109.370}{83.542} + \frac{109.370}{30.458} + \frac{109.370}{12.458} + \frac{109.370}{4.542} \\
&= 1.309 + 3.591 + 8.779 + 24.081 \\
&= 37.76
\end{aligned}$$

Critical values for χ^2 test

For a χ^2 test based on a 2×2 contingency table, the critical values are:

p	0.1	0.05	0.01	0.001
χ^2	2.706	3.841	6.635	10.828

Interpretation of table: If the null hypothesis were true then:

- The probability of the χ^2 value exceeding 2.706 would be $p = 0.1$.
- The probability of the χ^2 value exceeding 3.841 would be $p = 0.05$.
- The probability of the χ^2 value exceeding 6.635 would be $p = 0.01$.
- The probability of the χ^2 value exceeding 10.828 would be $p = 0.001$.

Worked example (concluded)

In our worked example, we have $\chi^2 = 37.76 > 10.828$,

In this case, we can reject the null hypothesis with very high confidence ($p < 0.001$).

In fact since $\chi^2 = 37.76 \gg 10.828$ we have confidence $p \ll 0.001$

We conclude that, at according to our data, there is a strong correlation between coursework submission and exam attendance.

χ^2 test — subtle points

In critical value tables for the χ^2 test, the entries are usually classified by *degrees of freedom*. For an $m \times n$ contingency table, there are $(m-1) \times (n-1)$ degrees of freedom. (This can be understood as follows. Given fixed marginals, once $(m-1) \times (n-1)$ entries in the table are completed, the remaining $m+n-1$ entries are completely determined.)

The values in the table on slide III.79 are those for 1 degree of freedom, and are thus the correct values for a 2×2 table.

The χ^2 test for a 2×2 table is considered unreliable when N is small (e.g. less than 40) and at least one of the four expected values is less than 5. In such situations, a modification *Yates correction*, is sometimes applied. (The details are beyond the scope of this course.)

Application 2: finding collocations

Recall from Part II that a *collocation* is a sequence of words that occurs atypically often in language usage. Examples were: *strong tea*; *run amok*; *make up*; *bitter sweet*, etc.

Using the χ^2 test we can use corpus data to investigate whether a given n-gram is a collocation. For simplicity, we focus on bigrams. (N.B. All the examples above are bigrams.)

Given a bigram $w_1\ w_2$, we use a corpus to investigate whether the words $w_1\ w_2$ appear together atypically often.

Again we shall apply the χ^2-test. So first we need to construct the relevant contingency table.

Contingency table for bigrams

O_{ij}	w_1	$\neg w_1$
w_2	$O_{11} = f(w_1\, w_2)$	$O_{12} = f(\neg w_1\, w_2)$
$\neg w_2$	$O_{21} = f(w_1\, \neg w_2)$	$O_{22} = f(\neg w_1\, \neg w_2)$

$f(w_1\, w_2)$ is frequency of $w_1\, w_2$ in the corpus.

$f(\neg w_1\, w_2)$ is number of bigram occurrences in corpus in which the second word is w_2 but the first word is not w_1. (N.B. If the same bigram appears n times in the corpus then this counts as n different occurrences.)

$f(w_1\, \neg w_2)$ is number of bigram occurrences in corpus in which the first word is w_1 but the second word is not w_2.

$f(\neg w_1\, \neg w_2)$ is number of bigram occurrences in corpus in which the first word is not w_1 and the second is not w_2.

Worked example 2

Recall from note II.5 that the bigram *strong desire* occurred 10 times in the CQP Dickens corpus.

We shall investigate whether *strong desire* is a collocation.

The full contingency table is:

O_{ij}	strong	¬strong
desire	**10**	**214**
¬desire	**655**	**3407085**

Worked example 2 (continued)

Marginals:

O_{ij}	strong	¬strong	
desire	**10**	**214**	**224**
¬desire	**655**	**3407085**	**3407740**
	665	**3407299**	**3407964**

Expected values:

E_{ij}	strong	¬strong	
desire	**0.044**	**223.956**	**224**
¬desire	**664.956**	**3407075.044**	**3407740**
	665	**3407299**	**3407964**

Worked example 2 (continued)

$$\chi^2 = \frac{9.956^2}{0.044} + \frac{9.956^2}{223.956} + \frac{9.956^2}{664.956} + \frac{9.956^2}{3407075.044}$$

$$= \frac{99.122}{0.044} + \frac{99.122}{223.956} + \frac{99.122}{664.956} + \frac{99.122}{3407075.044}$$

$$= 2252.773 + 0.443 + 0.149 + 0.000$$

$$= 2253.365$$

Worked example 2 (continued)

In our worked example, we have $\chi^2 = 2253.365 > 10.828$,

In this case, we can reject the null hypothesis with very high confidence ($p < 0.001$).

In fact since $\chi^2 = 2253.365 \gg 10.828$ we have confidence $p \ll 0.001$

However, all this tells us is that there is a strong correlation between occurrences of *strong* and occurrences of *desire*.

Due to the non-random nature of language, one would expect a strong correlation for *almost any* bigram occurring in a corpus.

Thus the critical values table is not informative for this investigation.

Worked example 2 (concluded)

So how can we tell if *strong desire* occurs atypically often?

One way is to use χ^2 values to *rank* bigrams occurring in a given corpus. A higher χ^2 means that the bigram is more significant.

If a bigram has an *atypically high* χ^2 value for the corpus, then this provides evidence in support of it being a collocation.

We could thus confirm that *strong desire* is a collocation by calculating χ^2 values for many other adjective-noun combinations, and finding that a value of $\mathbf{2253.365}$ is atypically high.

We do not do this, because the main point, that χ^2 values can be used to investigate collocations, has been made.