Informatics 1B, 2008

School of Informatics, University of Edinburgh

Data and Analysis

Note 13

Statistical Analysis of Data II

Alex Simpson

Part III — Unstructured Data

Statistical hypothesis testing

Data visualisation gives one 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.

Such hypotheses are corroborated or refuted by means of *statistical tests*.

Example: As in the last lecture, scatter plots give us ways of visualising the existence or absence of (positive or negative) correlation between different variables. In this case, such hypotheses can be corroborated or refuted using *Pearson's correlation coefficient*.

The general form of a statistical test

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 result $\boldsymbol{R}$ 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, So 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 $\boldsymbol{1/20}$ 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 $\boldsymbol{1/100}$ chance)

Case study 1: correlation

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 Pearson correlation coefficient $r_{x,y}$.

We then look the result up in statistical tables to see what the associated probability value p is.

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

Studying vs. exam results

We use the data from 12.18 (see also 12.20), with the study values for $x_1, \ldots, x_N$, and the exam values for $y_1, \ldots, y_N$, where, in this case, $N = 8$.

The relevant statistics are:

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

We now look up significance in a statistical table for Pearson's correlation coefficient. Such tables can be found in statistics books. (They can also be found on the Web.)

Critical values table for Pearson's 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 $r_{x,y} > 0.925$ has probability $p < 0.001$ of occurring (that is less than a $1/1000$ chance of occurring) if the null hypothesis is true.

Our value of 0.985 is indeed (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 (positive) correlation (since $r_{x,y}$ is close to 1 not to -1).

Drinking vs. exam results

We now use the drinking values from 12.18 (see also 12.21), as the values for $x_1, \ldots, x_8$. (The y values are unchanged.)

The new statistics are:

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

From the table, we see that if the null hypothesis were true then the probability p of a value $r_{x,y}$ with $0.834 < |r_{x,y}|$ occurring is $p = 0.01$.

Since $|-0.914| = 0.914 > 8.288$, we can therefore reject the null hypothesis with high confidence ($p < 0.01$). Though the result is less significant than the previous.

This time, the value -0.914 of $r_{x,y}$ is close to -1 so we conclude that the correlation is negative.

Pearsons correlation coefficient — subtleties

Critical value tables for Pearson's correlation coefficient are often given with rows indexed by *degrees of freedom* rather than by N. For Pearson's test, 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 Pearson's test, is too subtle a concept to explain here.)

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 values close to 1 *and* values close to -1 as significant. In a *one-tailed* test, we would be interested in just one of these possibilities.

Case study 2: finding collocations

Recall from note 10 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 statistical tools we can build a corpus-based test to detect whether a given n-gram is a collocation. For simplicity, we focus on bigrams. (N.B. All the examples above are bigrams.)

The statistical tool we use is called the χ^2 *(chi-squared) test*.

The χ^2 test is a general tool for investigating correlations between *categorical data*. (Whereas, the Pearson coefficient we have considered so far applies only to numerical data, both interval and ratio.)

However, we consider χ^2 only in the context of *collocation filtering*.

Finding collocations (continued)

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

The *null hypothesis* is that there is no relationship between occurrences of the the words w_1 and w_2.

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

Once again, if p is signifcantly low, we reject the null hypothesis, and we conclude that there is an atypical relationship between the words.

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

Contingency table

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

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

We shall investigate this as a collocation.

The full contingency table is:

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

Idea of χ^2 test

The observations O_{ij} are the actual frequencies of data in the corpus.

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 the data is likely to be due to 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}	w_1	$\neg w_1$	
w_2	O_{11}	O_{12}	$R_1 = O_{11} + O_{12}$
$\neg w_2$	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 bigram occurrences whose first word is w_1.
- B_2 is the number of bigram occurrences whose first word is not w_1.
- R_1 is the number of bigram occurrences whose second word is w_2.
- R_2 is the number of bigram occurrences whose second word is not w_2.
- N is the total number of bigram occurrences in the corpus.

Given these figures, if there were no relationship between w_1 and w_2, we would expect the number of occurrences of the bigram $w_1\ w_2$ to be

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

Expected frequencies

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

E_{ij}	w_1	$\neg w_1$	
w_2	$E_{11} = B_1 R_1 / N$	$E_{12} = B_2 R_1 / N$	$R_1 = E_{11} + E_{12}$
$\neg w_2$	$E_{21} = B_1 R_2 / N$	$E_{22} = B_2 R_2 / N$	$R_2 = E_{21} + E_{22}$
	$B_1 = E_{11} + E_{21}$	$B_2 = E_{12} + E_{22}$	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}	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 (continued)

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

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

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

$$= 2253.365$$

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

We interpret the table thus. 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 $\chi^2 = 2253.365 > 10.828$,

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 $p \ll 0.001$

We conclude that, at least according to the Dickens corpus, the bigram *strong desire* is (rightly!) identified as a (highly probable) collocation.

χ^2 — 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 13.21 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 is 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 Data & Analysis.)

Populations and samples

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

Often, however, we are 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 requirments of users.
- Many, many other examples.

In such cases it is totally impracticable to obtain exhaustive data about the population as a whole. Sp 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.

To estimate the *mean* of the population, calculate

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

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

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

To estimate the *standard deviation* of the population, calculate

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

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: σ_n for the formula with denominator n; and σ_{n-1} for the formula with denominator $n-1$.

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 very 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, 2007/8