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 significantly 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 discrepancy 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 $\mathbf{4}$ 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)

$$\chi^2 = \frac{10.458^2}{83.542} + \frac{10.458^2}{30.458} + \frac{10.458^2}{12.458} + \frac{10.458^2}{4.542}$$

$$= \frac{109.370}{83.542} + \frac{109.370}{30.458} + \frac{109.370}{12.458} + \frac{109.370}{4.542}$$

$$= 1.309 + 3.591 + 8.779 + 24.081$$

$$= 37.76$$

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 our data provides strong support for a 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.80 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)

$$\begin{aligned}
\chi^2 &= \frac{\mathbf{9.956^2}}{\mathbf{0.044}} + \frac{\mathbf{9.956^2}}{\mathbf{223.956}} + \frac{\mathbf{9.956^2}}{\mathbf{664.956}} + \frac{\mathbf{9.956^2}}{\mathbf{3407075.044}} \\
&= \frac{\mathbf{99.122}}{\mathbf{0.044}} + \frac{\mathbf{99.122}}{\mathbf{223.956}} + \frac{\mathbf{99.122}}{\mathbf{664.956}} + \frac{\mathbf{99.122}}{\mathbf{3407075.044}} \\
&= \mathbf{2252.773 + 0.443 + 0.149 + 0.000} \\
&= \mathbf{2253.365}
\end{aligned}$$

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 **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.

Berkeley Sex Bias

	Accepted	Rejected	Applied	Success
Male	**1122**	**1005**	**2127**	**53%**
Female	**511**	**590**	**1101**	**46%**
Total	**1633**	**1595**	**3228**	**51%**

$\chi^2 = 11.66$

Simpson's Paradox

FG S	Accepted	Rejected	Applied	Success
Male	**864**	**521**	**1385**	**62%**
Female	**106**	**27**	**133**	**80%**
Total	**970**	**548**	**1518**	**64%**

$\chi^2 = 15.77$

FG A	Accepted	Rejected	Applied	Success
Male	**258**	**484**	**742**	**35%**
Female	**405**	**563**	**968**	**42%**
Total	**663**	**1047**	**1710**	**39%**

$\chi^2 = 8.84$