Informatics 1: Data & Analysis

Lecture 19: χ^2 Testing on Categorical Data

Ian Stark

School of Informatics
The University of Edinburgh

Tuesday 22 March 2016
Semester 2 Week 10

http://www.inf.ed.ac.uk/teaching/courses/inf1/da

Unstructured Data

Data Retrieval

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

Statistical Analysis of Data

- Data scales and summary statistics
- Hypothesis testing and correlation
- χ^2 tests and collocations also *chi-squared*, pronounced "kye-squared"

Unstructured Data

Data Retrieval

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

Statistical Analysis of Data

- Data scales and summary statistics
- Hypothesis testing and correlation
- χ^2 tests and collocations

also *chi-squared*, pronounced "kye-squared"

Timetable !

This is Teaching Week 10 of Semester 2, next week is Week 11, and the teaching block ends on Friday 1 April

Inf1-DA has the following events remaining:

- Friday 25 March: Lecture 20. Review of exam arrangements; summary of topics covered in the course; revision of specific topics.
- Tuesday 27 March: Final Lecture. Review of past exam questions.
- Monday 26 March – Wednesday 28 March: Final tutorial. return of coursework assignment, feedback and discussion on that.

Which topics to cover in revision lectures?

Speak your brains: http://is.gd/da16revision

Review: Correlation Coefficient

The correlation coefficient ρ is a way to measure how closely two datasets $x_1, \ldots, x_N$ and $y_1, \ldots, y_N$ are related to each other.

Take μ_x and σ_x the mean and standard deviation of the x_i values.
Take μ_y and σ_y the mean and standard deviation of the y_i values.

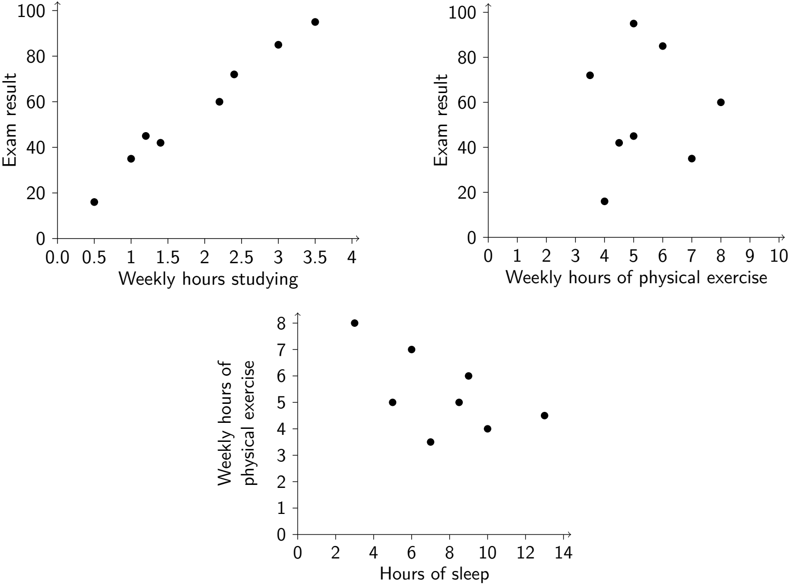
The correlation coefficient $\rho_{x,y}$ is then computed as:

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

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*.

Review: Correlation Coefficient



Visualisation and Anscombe's Quartet (1973)

Data set 1

x	y
10.0	8.04
8.0	6.95
13.0	7.58
9.0	8.81
11.0	8.33
14.0	9.96
6.0	7.24
4.0	4.26
12.0	10.84
7.0	4.82
5.0	5.68

Data set 2

x	y
10.0	9.14
8.0	8.14
13.0	8.74
9.0	8.77
11.0	9.26
14.0	8.10
6.0	6.13
4.0	3.10
12.0	9.13
7.0	7.26
5.0	4.74

Data set 3

x	y
10.0	7.46
8.0	6.77
13.0	12.74
9.0	7.11
11.0	7.81
14.0	8.84
6.0	6.08
4.0	5.39
12.0	8.15
7.0	6.42
5.0	5.73

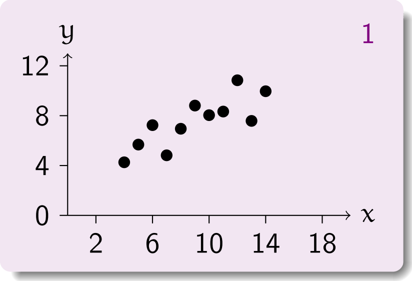
Data set 4

x	y
8.0	6.58
8.0	5.76
8.0	7.71
8.0	8.84
8.0	8.47
8.0	7.04
8.0	5.25
19.0	12.50
8.0	5.56
8.0	7.91
8.0	6.89

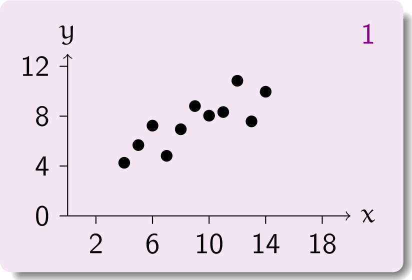
$\mu_x = 9 \quad \mu_y = 7.04 \quad \sigma_x = 3.16 \quad \sigma_y = 1.94 \quad \rho_{x,y} = 0.82 \quad \hat{y} = 3.00x + 0.50$

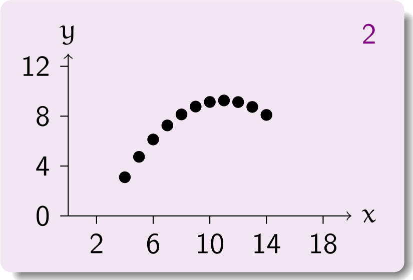
Visualisation and Anscombe's Quartet (1973)

+

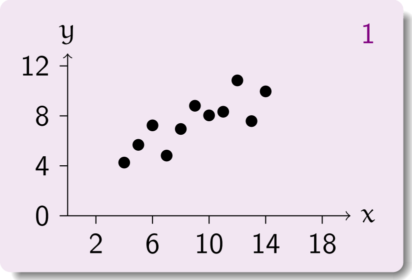


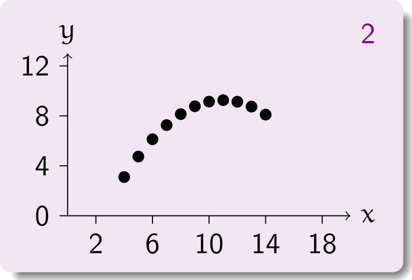
Visualisation and Anscombe's Quartet (1973)

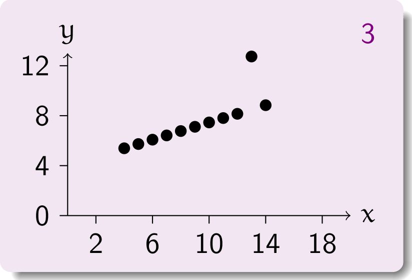




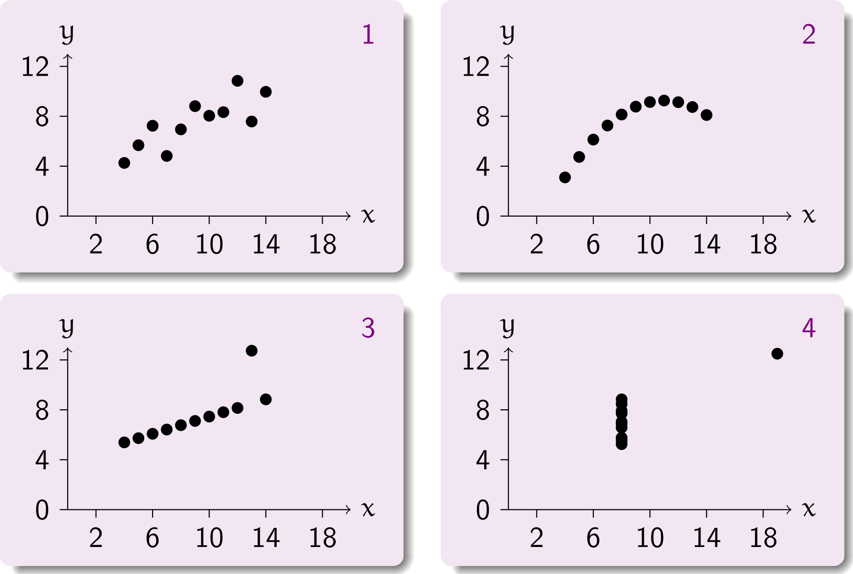
Visualisation and Anscombe's Quartet (1973)







Visualisation and Anscombe's Quartet (1973)



Review: Hypothesis Testing for Correlation

So far, we have the following procedure to identify correlation between two series of data.

- Draw a scatter plot. Does it look as though there is a relationship between the two variables?
- Calculate a correlation coefficient R.
- Look in a table of critical values to see whether R is large, given the number of data points.
- If R is above the critical value for some chosen p, say 0.05 or 0.01, then this may be judged statistically significant and lead us to reject the null hypothesis.

Review: Problems 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 “significant” (that's measured by “effect size”).

  It is really about whether it is statistically *detectable*.

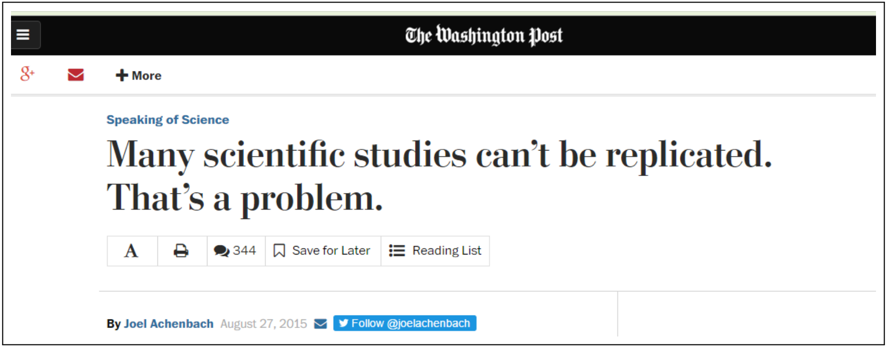
- Events with $p < 0.05$ happen all the time. Well, 1 time in 20.

  Seeing a low p-value is perhaps evidence to suggest an effect. It's a reason to do another experiment, or make a prediction.

Despite this, hypothesis testing can still serve as evidence of correlation and even causality:

- Postulate a mechanism, propose a hypothesis, make predictions, then carry out repeated experiments to confirm or refute the hypothesis.

The Reproducibility Problem / Replication Crisis



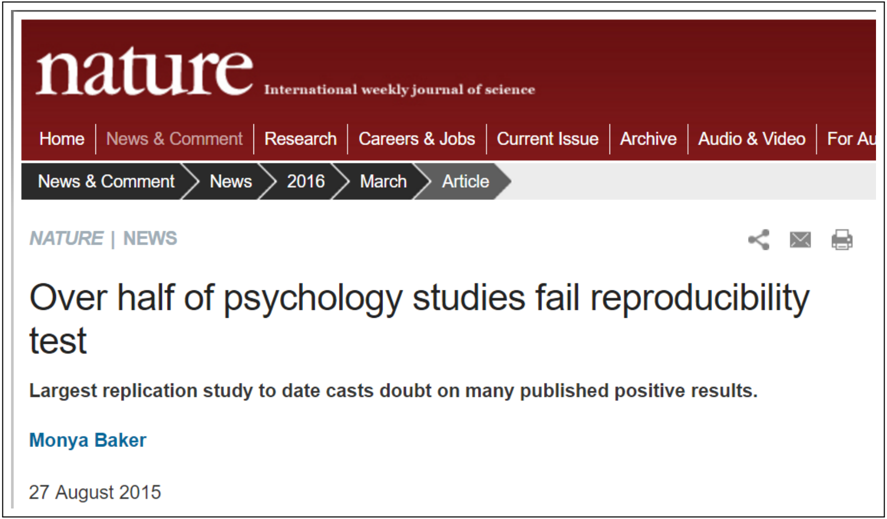
Speaking of Science

Many scientific studies can’t be replicated. That’s a problem.

By Joel Achenbach August 27, 2015

http://is.gd/waporeplicate

The Reproducibility Problem / Replication Crisis



nature International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | For Au
News & Comment > News > 2016 > March > Article

NATURE | NEWS

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

The Reproducibility Problem / Replication Crisis +

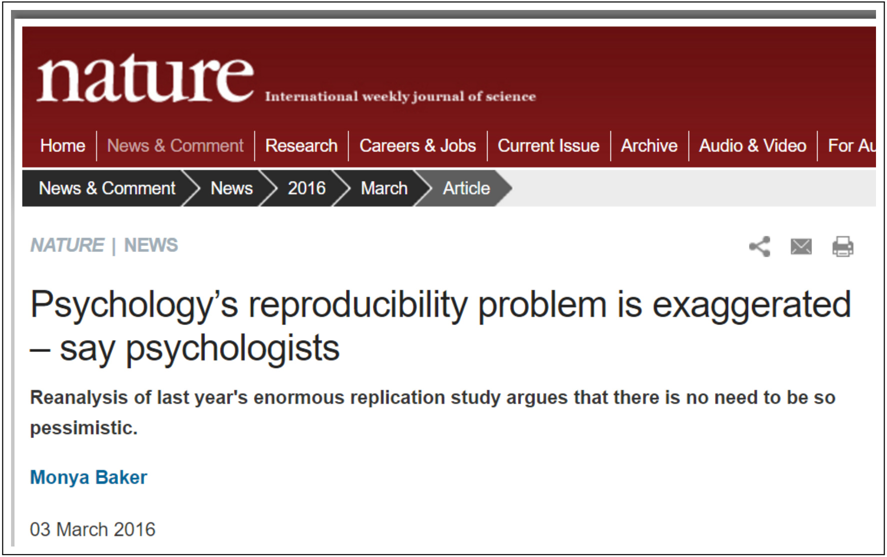
Open Science Collaboration.
Estimating the Reproducibility of Psychological Science.
Science, 349(6521), 2015
DOI: 10.1126/science.aac4716

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 . . .

https://osf.io/ezcuj/wiki

The Reproducibility Problem / Replication Crisis



NATURE | NEWS

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

The χ^2 Test

We have just seen the correlation coefficient, a useful test to identify whether or not an apparent correlation between variables is statistically significant.

However, the correlation coefficient is only applicable to quantitative data. (A variant, the Spearman rank correlation coefficient, can also be applied to ordinal data.)

The χ^2 *test* is statistical tool for assessing correlations in categorical data.

This rest of this lecture will go through the calculations for a χ^2 test, using three example sets of data:

- Student results for Inf1-DA in 2010/2011;
- Bigram frequency in the British National Corpus;
- Student admissions to the University of California, Berkeley in 1973.

Example: Student Exam Results

Question

Is there any correlation, in a class of students enrolled on a course, between submitting the coursework assignment and obtaining grade A (70% or higher) on the exam for that course?

The data we will use is the actual performance of those students who took the Informatics 1: Data & Analysis exam in May 2011.

Example: Student Exam Results

Question

Is there any correlation, in a class of students enrolled on a course, between submitting the coursework assignment and obtaining grade A (70% or higher) on the exam for that course?

Our analysis follows the usual pattern of a statistical test:

- The null hypothesis here is that there is no relationship between coursework submission and exam grade A.
- The χ^2 test indicates the probability p that data of the kind we actually see would turn up if the null hypothesis were true.
- If p is low, then we reject the null hypothesis and conclude that there is a correlation between coursework submission and exam grade A.

Contingency table

Frequencies

O_{ij}	cw	¬cw
A	O_{11}	O_{12}
¬A	O_{21}	O_{22}

O_{11} is the number of students who submitted coursework and obtained an A grade.

O_{12} is the number of students who did not submit coursework and obtained an A grade.

O_{21} is the number of students who submitted coursework and did not obtain an A grade.

O_{22} is the number of students who did not submit coursework and did not obtain an A grade.

Contingency table

Frequencies

O_{ij}	cw	¬cw
A	42	7
¬A	49	19

42 is the number of students who submitted coursework and obtained an A grade.

7 is the number of students who did not submit coursework and obtained an A grade.

49 is the number of students who submitted coursework and did not obtain an A grade.

19 is the number of students who did not submit coursework and did not obtain an A grade.

χ^2 Test Intuition

We have a table of observed frequencies O_{ij}, and from these we calculate *expected frequencies* E_{ij} — the numbers we would expect to see if the null hypothesis were true.

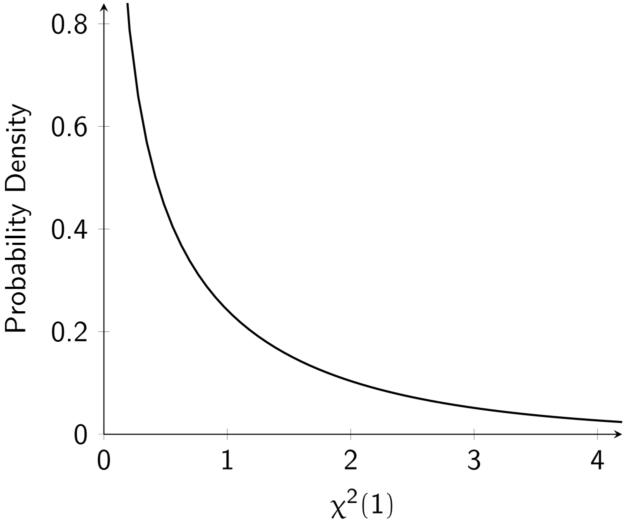
The χ^2 value is calculated by comparing the actual frequencies to the expected frequencies.

The larger the discrepancy between these two, the less probable it is that observations like this would occur were the null hypothesis true.

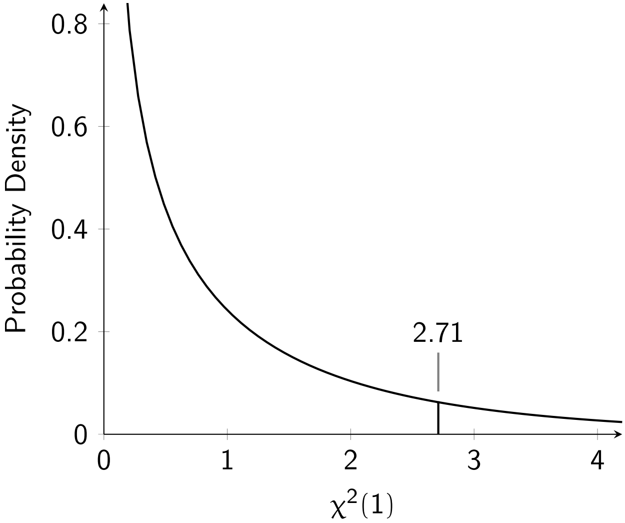
More precisely, if the null hypothesis were true, then the χ^2 value would vary according to the distribution shown on the next slide.

If the χ^2 is significantly large then we reject the null hypothesis.

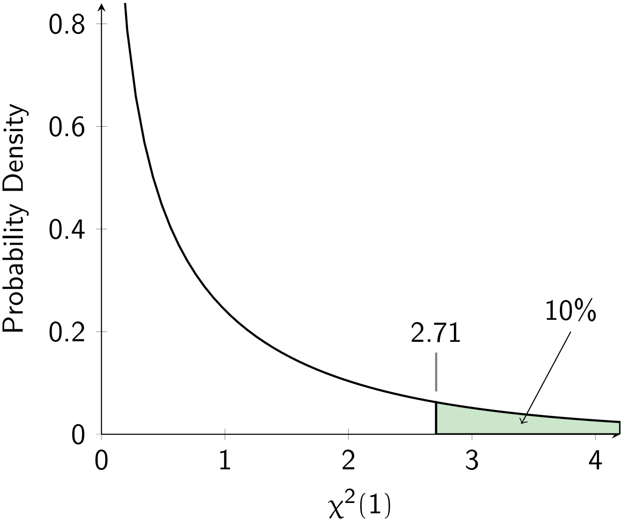
Graph of χ^2 Distribution



Graph of χ^2 Distribution



Graph of χ^2 Distribution



Marginals

Observed

O_{ij}	cw	¬cw	
A	O_{11}	O_{12}	R_1
¬A	O_{21}	O_{22}	R_2
	C_1	C_2	N

$R_1 = O_{11} + O_{12}$ is the number of students who obtained an A grade.

$R_2 = O_{21} + O_{22}$ is the number of students who did not obtain an A grade.

$C_1 = O_{11} + O_{21}$ is the number of students who submitted coursework.

$C_2 = O_{21} + O_{22}$ is the number of students who did not submit coursework.

N is the total number of students in the data set.

Expected Frequencies

Expected

E_{ij}	cw	¬cw	
A	E_{11}	E_{12}	R_1
¬A	E_{21}	E_{22}	R_2
	C_1	C_2	N

If there were no relationship between coursework submission and exam grade A, then we would expect to see the number of students with both being

$$E_{11} \ = \ \frac{R_1}{N} \times \frac{C_1}{N} \times N \ = \ \frac{R_1 C_1}{N}$$

and similarly for other values

$$E_{12} = \frac{R_1 C_2}{N} \qquad E_{21} = \frac{R_2 C_1}{N} \qquad E_{22} = \frac{R_2 C_2}{N} \; .$$

Computing χ^2

Observed

O_{ij}	cw	$\neg$cw	
A	O_{11}	O_{12}	R_1
$\neg$A	O_{21}	O_{22}	R_2
	C_1	C_2	N

Expected

E_{ij}	cw	$\neg$cw	
A	E_{11}	E_{12}	R_1
$\neg$A	E_{21}	E_{22}	R_2
	C_1	C_2	N

The χ^2 statistic for a contingency table in general is defined as

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

which for a 2 × 2 table expands to

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

For a 2 × 2 table the four numerators are always equal. Why?

Worked Example

Observed

O_{ij}	cw	¬cw	
A	42	7	
¬A	49	19	

Expected

E_{ij}	cw	¬cw	
A			
¬A			

The χ^2 statistic for this contingency table is

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

Worked Example

Observed

O_{ij}	cw	¬cw	
A	42	7	49
¬A	49	19	68
	91	26	117

Expected

E_{ij}	cw	¬cw	
A			
¬A			

The χ^2 statistic for this contingency table is

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

Worked Example

Observed

O_{ij}	cw	¬cw	
A	42	7	49
¬A	49	19	68
	91	26	117

Expected

E_{ij}	cw	¬cw	
A			49
¬A			68
	91	26	117

The χ^2 statistic for this contingency table is

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

Worked Example

Observed

O_{ij}	cw	¬cw	
A	42	7	49
¬A	49	19	68
	91	26	117

Expected

E_{ij}	cw	¬cw	
A	38.11		49
¬A			68
	91	26	117

The χ^2 statistic for this contingency table is

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

Worked Example

Observed

O_{ij}	cw	¬cw	
A	42	7	49
¬A	49	19	68
	91	26	117

Expected

E_{ij}	cw	¬cw	
A	38.11	10.89	49
¬A	52.89		68
	91	26	117

The χ^2 statistic for this contingency table is

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

Worked Example

Observed

O_{ij}	cw	¬cw	
A	42	7	49
¬A	49	19	68
	91	26	117

Expected

E_{ij}	cw	¬cw	
A	38.11	10.89	49
¬A	52.89	15.11	68
	91	26	117

The χ^2 statistic for this contingency table is

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

Worked Example

Observed

O_{ij}	cw	¬cw	
A	42	7	49
¬A	49	19	68
	91	26	117

Expected

E_{ij}	cw	¬cw	
A	38.11	10.89	49
¬A	52.89	15.11	68
	91	26	117

The χ^2 statistic for this contingency table is

$$\begin{aligned}\chi^2 &= \frac{(O_{11}-E_{11})^2}{E_{11}} + \frac{(O_{12}-E_{12})^2}{E_{12}} + \frac{(O_{21}-E_{21})^2}{E_{21}} + \frac{(O_{22}-E_{22})^2}{E_{22}} \\ &= \frac{(42-38.11)^2}{38.11} + \frac{(7-10.89)^2}{10.89} + \frac{(49-52.89)^2}{52.89} + \frac{(19-15.11)^2}{15.11}\end{aligned}$$

Worked Example

Observed

O_{ij}	cw	¬cw	
A	42	7	49
¬A	49	19	68
	91	26	117

Expected

E_{ij}	cw	¬cw	
A	38.11	10.89	49
¬A	52.89	15.11	68
	91	26	117

The χ^2 statistic for this contingency table is

$$\begin{aligned}\chi^2 &= \frac{(O_{11}-E_{11})^2}{E_{11}} + \frac{(O_{12}-E_{12})^2}{E_{12}} + \frac{(O_{21}-E_{21})^2}{E_{21}} + \frac{(O_{22}-E_{22})^2}{E_{22}} \\ &= \frac{(42-38.11)^2}{38.11} + \frac{(7-10.89)^2}{10.89} + \frac{(49-52.89)^2}{52.89} + \frac{(19-15.11)^2}{15.11} \\ &= \frac{3.89^2}{38.11} + \frac{-3.89^2}{10.89} + \frac{-3.89^2}{52.89} + \frac{3.89^2}{15.11}\end{aligned}$$

Worked Example

Observed

O_{ij}	cw	¬cw	
A	42	7	49
¬A	49	19	68
	91	26	117

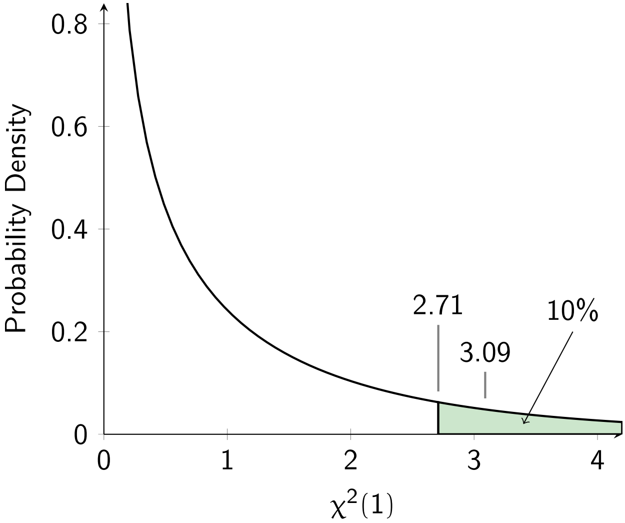
Expected

E_{ij}	cw	¬cw	
A	38.11	10.89	49
¬A	52.89	15.11	68
	91	26	117

The χ^2 statistic for this contingency table is

$$\begin{aligned}
\chi^2 &= \frac{(O_{11} - E_{11})^2}{E_{11}} + \frac{(O_{12} - E_{12})^2}{E_{12}} + \frac{(O_{21} - E_{21})^2}{E_{21}} + \frac{(O_{22} - E_{22})^2}{E_{22}} \\
&= \frac{(42 - 38.11)^2}{38.11} + \frac{(7 - 10.89)^2}{10.89} + \frac{(49 - 52.89)^2}{52.89} + \frac{(19 - 15.11)^2}{15.11} \\
&= \frac{3.89^2}{38.11} + \frac{-3.89^2}{10.89} + \frac{-3.89^2}{52.89} + \frac{3.89^2}{15.11} \\
&= 3.09
\end{aligned}$$

Graph of χ^2 Distribution



Critical Values for χ^2

These are the critical values for different significance levels of the χ^2 distribution for a 2×2 table.

p	0.10	0.05	0.01	0.001
χ^2	2.71	3.84	6.64	10.83

This means that if the null hypothesis were true then:

- The probability of the χ^2 value exceeding 2.71 would be $p = 0.1$.
- The probability of the χ^2 value exceeding 3.84 would be $p = 0.05$.
- The probability of the χ^2 value exceeding 6.64 would be $p = 0.01$.
- The probability of the χ^2 value exceeding 10.83 would be $p = 0.001$.

Critical Values for χ^2

These are the critical values for different significance levels of the χ^2 distribution for a 2×2 table.

p	0.10	0.05	0.01	0.001
χ^2	2.71	3.84	6.64	10.83

In this case $\chi^2 = 3.09$, meaning $0.10 > p > 0.05$. This is evidence to suggest that there is a correlation, and we reject the null hypothesis at the 90% level. The result is statistically significant.

It appears that in this data there is a correlation between submitting the coursework and achieving an A grade in the exam. Of course, this does not tell us whether there is any causal link, either between these outcomes or from some third factor. What it does do is give a hypothesis that we could explore in further data.

Additional Features of χ^2 Tests

Degrees of Freedom

In tables of critical values for the χ^2 distribution, entries are usually classified by *degrees of freedom*. An m by n contingency table has $(m-1) \times (n-1)$ degrees of freedom — given fixed marginals, once there are $(m-1) \times (n-1)$ entries in the table the remaining $(m+n-1)$ entries are forced.

A 2 by 2 table has only one degree of freedom, and the table on the previous slide gave the critical values for a χ^2 distribution with one degree of freedom.

Additional Features of χ^2 Tests

Low Frequencies

The statistics underlying the χ^2 test become inaccurate when expected frequencies are small.

Reasons include: inevitable differences up to 0.5 as observed values can only be whole numbers; and that χ^2 is only an approximation to the exact (but computationally more expensive) distribution.

The test is usually considered unreliable for a 2×2 table if any cell has expected value below 5; or for a larger table, if more than 20% of cells have expected value below 5.

That's really just a rule of thumb: opinions vary
on what are appropriate limits here

For these cases there are more refined methods such as *Fisher's Exact Test*.

Example: Collocations

Recall that a collocation is a sequence of words that occurs atypically often in a language. For example: “run amok”, “strong tea”, “make do”.

So far, we haven’t looked at what exactly “atypically often” might mean.

The χ^2 test is one way to approach this, and we shall use it to assess whether the bigram “make do” appears atypically often in the 10^8 words of the British National Corpus (BNC).

The null hypothesis will be that the two words “make” and “do” appear together just as often as would be expected by chance, given their individual frequencies in the corpus.

If we reject this hypothesis, then we might take this as evidence of “make do” being a collocation.

Contingency table

Bigram Frequencies

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 the frequency of $w_1 w_2$ in a corpus, the number of times that bigram appears.

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

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

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

Worked Example

Observed

O_{ij}	make	¬make	
do	230	270546	
¬do	77162	111833081	

Expected

E_{ij}	make	¬make	
do			
¬do			

Worked Example

Observed

O_{ij}	make	¬make	
do	230	270546	270776
¬do	77162	111833081	111910243
	77392	112103627	112181019

Expected

E_{ij}	make	¬make	
do			
¬do			

Worked Example

Observed

O_{ij}	make	¬make	
do	230	270546	270776
¬do	77162	111833081	111910243
	77392	112103627	112181019

Expected

E_{ij}	make	¬make	
do			270776
¬do			111910243
	77392	112103627	112181019

Worked Example

Observed

O_{ij}	make	¬make	
do	230	270546	270776
¬do	77162	111833081	111910243
	77392	112103627	112181019

Expected

E_{ij}	make	¬make	
do	186		270776
¬do			111910243
	77392	112103627	112181019

Worked Example

Observed

O_{ij}	make	¬make	
do	230	270546	270776
¬do	77162	111833081	111910243
	77392	112103627	112181019

Expected

E_{ij}	make	¬make	
do	186	270589	270776
¬do	77205		111910243
	77392	112103627	112181019

Worked Example

Observed

O_{ij}	make	¬make	
do	230	270546	270776
¬do	77162	111833081	111910243
	77392	112103627	112181019

Expected

E_{ij}	make	¬make	
do	186	270589	270776
¬do	77205	111833038	111910243
	77392	112103627	112181019

Worked Example

Observed

O_{ij}	make	¬make	
do	230	270546	270776
¬do	77162	111833081	111910243
	77392	112103627	112181019

Expected

E_{ij}	make	¬make	
do	186	270589	270776
¬do	77205	111833038	111910243
	77392	112103627	112181019

The χ^2 statistic for this table is 10.02, which is significant at the 99% level.

Example: Berkeley Admissions

Following the fall admissions round of students to graduate school at the University of California, Berkeley in 1973, the University was sued for bias against women.

Admission statistics showed that men applying were significantly more likely to be admitted than women applying.

The following table is based on some of those admission statistics.

Berkeley Admissions

	Accepted	Rejected	Applied	Rate
Men	1122	1005	2127	53%
Women	511	590	1101	46%
Total	1633	1595	3228	51%

The χ^2 statistic for this table is 11.66, significant at the 99.9% level.

Not So Simple

One obvious action is to break down these figures to identify which departments are the source of this bias.

Faculty Group “S”

	Accepted	Rejected	Applied
Men	864	521	1385
Women	106	27	133
Total	970	548	1518

Faculty Group “A”

	Accepted	Rejected	Applied
Men	258	484	742
Women	405	563	968
Total	663	1047	1710

Not So Simple

One obvious action is to break down these figures to identify which departments are the source of this bias.

Faculty Group "S"

	Accepted	Rejected	Applied	Rate
Men	864	521	1385	62%
Women	106	27	133	80%
Total	970	548	1518	64%

Faculty Group "A"

	Accepted	Rejected	Applied	Rate
Men	258	484	742	35%
Women	405	563	968	42%
Total	663	1047	1710	39%

Not So Simple

One obvious action is to break down these figures to identify which departments are the source of this bias.

Faculty Group "S"

	Accepted	Rejected	Applied	Rate
Men	864	521	1385	62%
Women	106	27	133	80%
Total	970	548	1518	64%

$\chi^2 = 15.77$

Faculty Group "A"

	Accepted	Rejected	Applied	Rate
Men	258	484	742	35%
Women	405	563	968	42%
Total	663	1047	1710	39%

$\chi^2 = 8.84$

Not So Simple

This curious behaviour is known as *Simpson's Paradox*. It turns up occasionally in a range of real-life cases; and it is not easily resolved. Judea Pearl argues that the resolution lies in identifying the causal networks in any given situation.

In the Berkeley case, the disparity arose because:

- Subject choice was correlated with gender;
- Competition for places varied substantially between departments.

More detailed investigation suggested no significant bias in admissions committees; but that the bias in aggregated data was linked to real bias in wider cultural expectations and social pressures.

P. J. Bickel, E. A. Hammel, and J. W. O'Connell.
Sex bias in graduate admissions: Data from Berkeley.
Science, 187(4175):398–404, 1975.
DOI: 10.1126/science.187.4175.398

http://is.gd/berkbias