Informatics 1: Data & Analysis

Lecture 18: Hypothesis Testing and χ^2

Ian Stark

School of Informatics
The University of Edinburgh

Friday 24 March 2017
Semester 2 Week 9

Unstructured Data

Data Retrieval

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

Statistical Analysis of Data

- Summary statistics
- Hypothesis testing and χ^2 also *chi-squared*, pronounced "kye-squared"
- Data scales. Correlation and causation.

What Happened Earlier

Statistics

A statistic is a single value computed from a set of data that captures some overall property: for example the mean, median, mode, variance, or standard deviation.

Given a small random *sample* from a large *population* we can estimate statistics for the population using calculations on the sample.

Correlation

With two sets of data we may look for a correlation between them: if they vary together, with changes in one matching changes in the other.

An observed correlation may be because one thing directly causes another; because both are affected by some other factor; or simply by chance.

Data in Multiple Dimensions

This is the table recording for each of eight imaginary students (A–H) the time in hours they spend each week on studying for Inf1-DA (outside lectures and tutorials) and on physical exercise; how many hours they spent asleep on a particular night; and their performance on the Data & Analysis exam.

There are four variables: study, exercise, sleep and exam results.

Example Data (NB: Not real students)

	A	B	C	D	E	F	G	H
Study	0.5	1	1.4	1.2	2.2	2.4	3	3.5
Exercise	4	7	4.5	5	8	3.5	6	5
Sleep	10	6	13	5	3	7	9	8.5
Exam	16	35	42	45	60	72	85	95

Correlation

Is there any relationship between the values observed for these four variables?

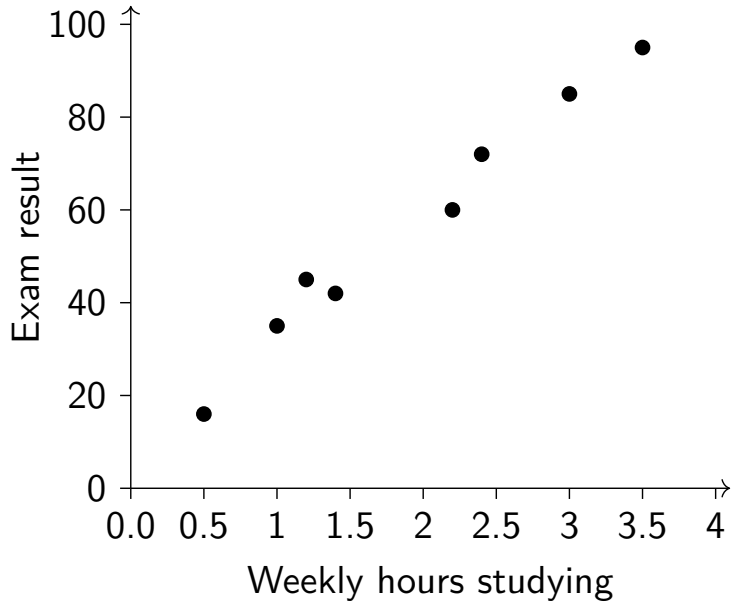
If there is, and the variables change in similar ways to each other, then we say they are *correlated*.

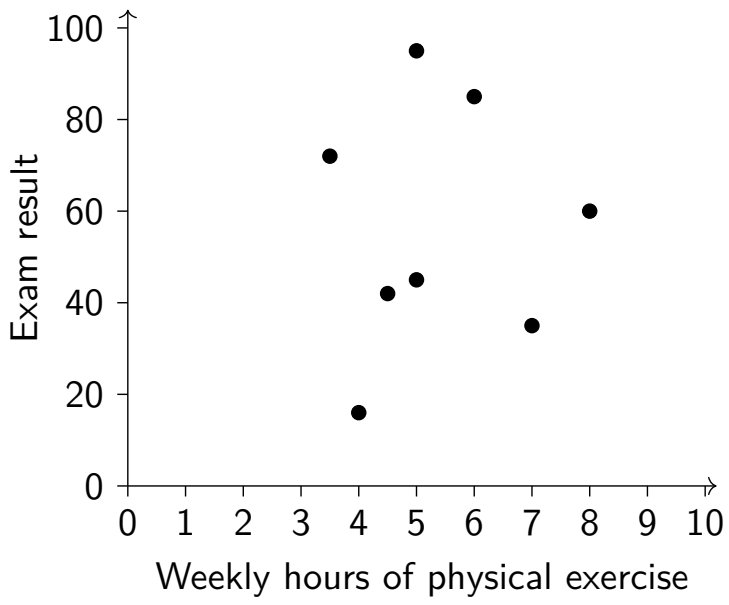
One way to discover correlation is with a *scatter plot* taking one variable for x-axis, another for the y-axis, and plotting a point for each item of data.

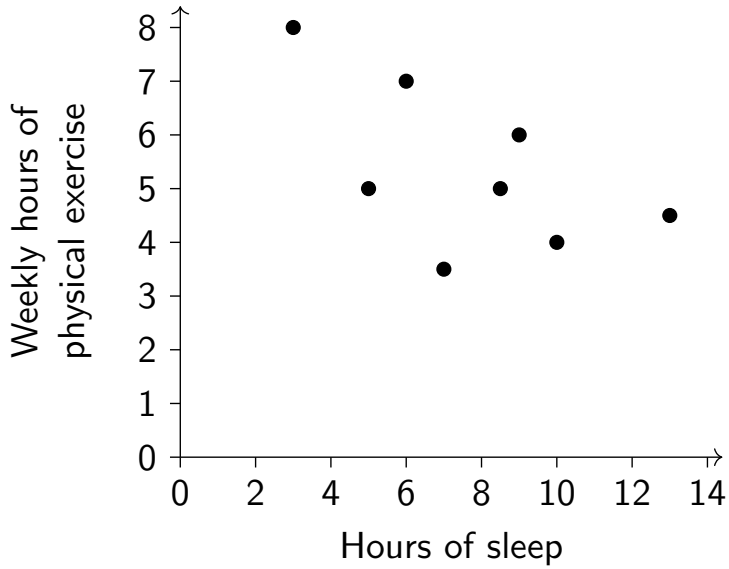
Example Data (NB: Not real students)

	A	B	C	D	E	F	G	H
Study	0.5	1	1.4	1.2	2.2	2.4	3	3.5
Exercise	4	7	4.5	5	8	3.5	6	5
Sleep	10	6	13	5	3	7	9	8.5
Exam	16	35	42	45	60	72	85	95

Scatter Plot







Hypothesis Testing

These scatter plots do suggest possible correlations between variables.

There are other ways to formulate possible correlations. For example:

- From a proposed underlying mechanism;
- Analogy with another situation where some relation is known to exist;
- Based on the predictions of a proposed model for a system.

Any such suggestion of a correlation is a *hypothesis*.

Statistical tests provide the mathematical tools to assess evidence and carry out hypothesis testing.

Statistical Tests

Most statistical testing starts from a specified *null hypothesis*, that there is nothing out of the ordinary in the data: no correlation, no effect, nothing to see.

We then compute some statistic from the data. Call this R.

The *hypothesis test* is then to investigate how likely it is that we would see a result like R if the null hypothesis were true.

This chance is called a *p-value*, with $0 \leqslant \mathrm{p} \leqslant 1$.

Significance

The value p represents the chance that we would obtain a result like R if the null hypothesis were true.

If p is small, then we conclude that the null hypothesis is a poor explanation for the observed data.

Based on this we might *reject* the null hypothesis.

Standard thresholds for "small" are $\mathrm{p} < 0.05$, meaning that there is less than 1 chance in 20 of obtaining the observed result by chance, if the null hypothesis is true; or $\mathrm{p} < 0.01$, meaning less than 1 chance in 100.

An observation that leads us to reject the null hypothesis is described as *statistically significant*.

Correlation Coefficient

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

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 this suggests x and y are *negatively correlated*.

Correlation Coefficient as a Statistical Test

In a test for correlation between two variables x and y — such as study hours and exam results — we are looking to see whether the variables are correlated; and if so in what direction.

The null hypothesis is that there is no correlation.

We calculate the correlation coefficient $\rho_{x,y}$, and then do one of two things:

- Look in a table of *critical values* for this statistic, to see whether the value we have is significant;
- Compute directly the p-value for this statistic, to see whether it is small.

Depending on the result, we may reject the null hypothesis.

Critical Values for Correlation Coefficient

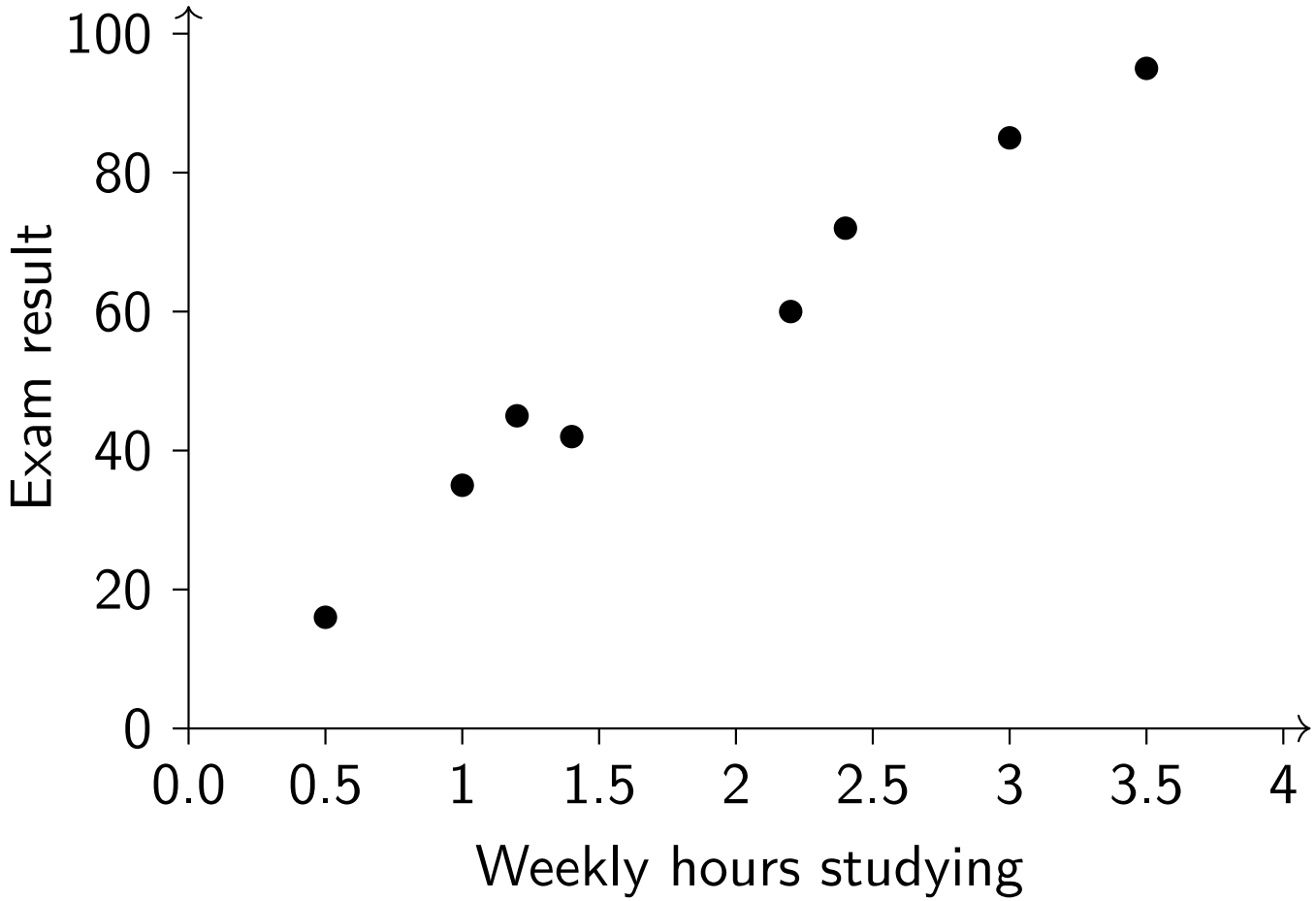
ρ	$p = 0.10$	$p = 0.05$	$p = 0.01$	$p = 0.001$
$N = 7$	0.669	0.754	0.875	0.951
$N = 8$	0.621	0.707	0.834	0.925
$N = 9$	0.582	0.666	0.798	0.898
$N = 10$	0.549	0.632	0.765	0.872

This table has rows indicating the critical values of the correlation coefficient ρ for different numbers of data items N in the series being compared.

It shows that for $N = 8$ data items that are not correlated, there is probability $p = 0.01$ of observing a coefficient $|\rho_{x,y}| > 0.834$.

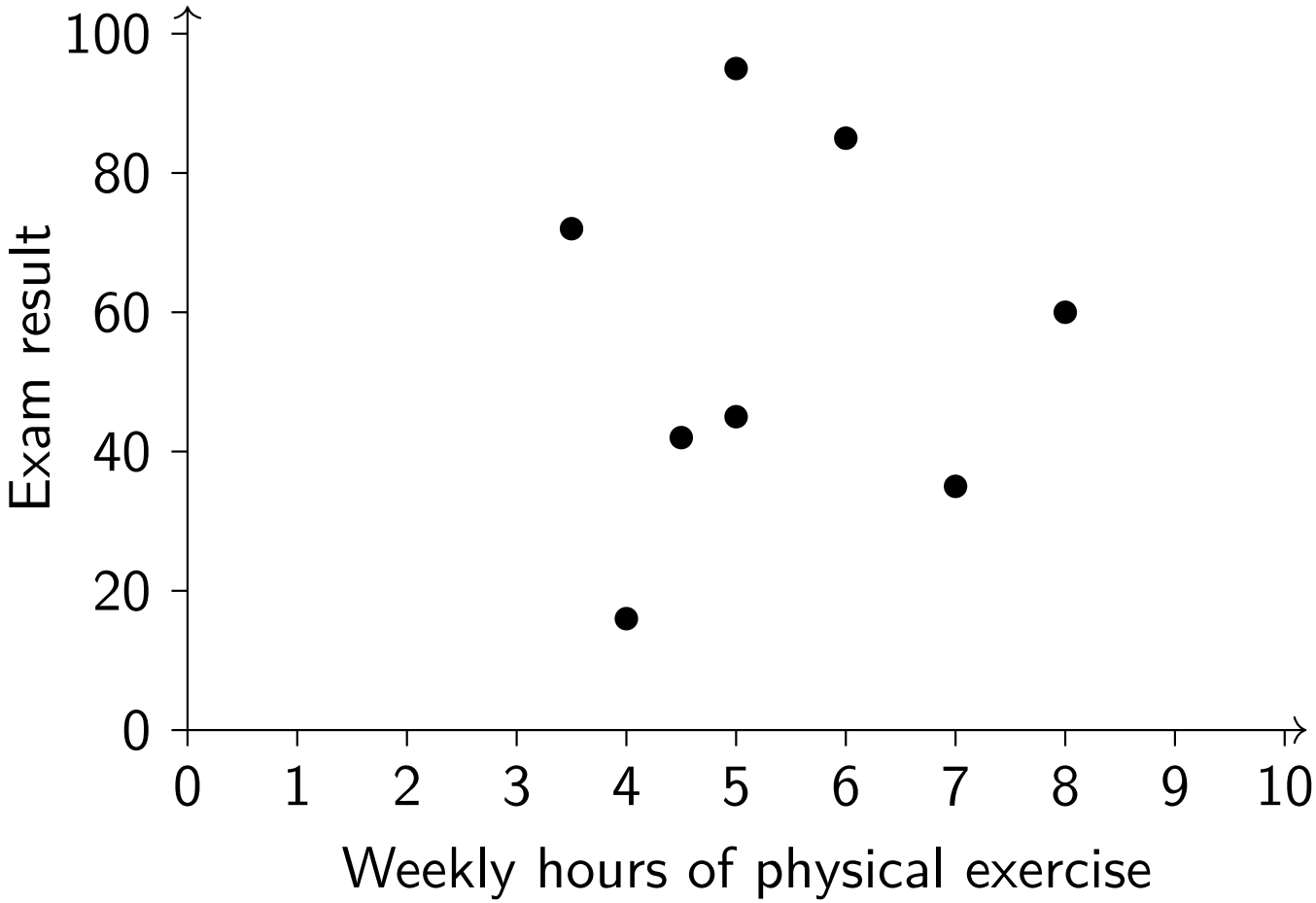
In the same way for $N = 8$ uncorrelated data items a value of $|\rho_{x,y}| > 0.925$ has probability $p = 0.001$ of occurring, only one chance in a thousand.

Studying vs. Exam Results



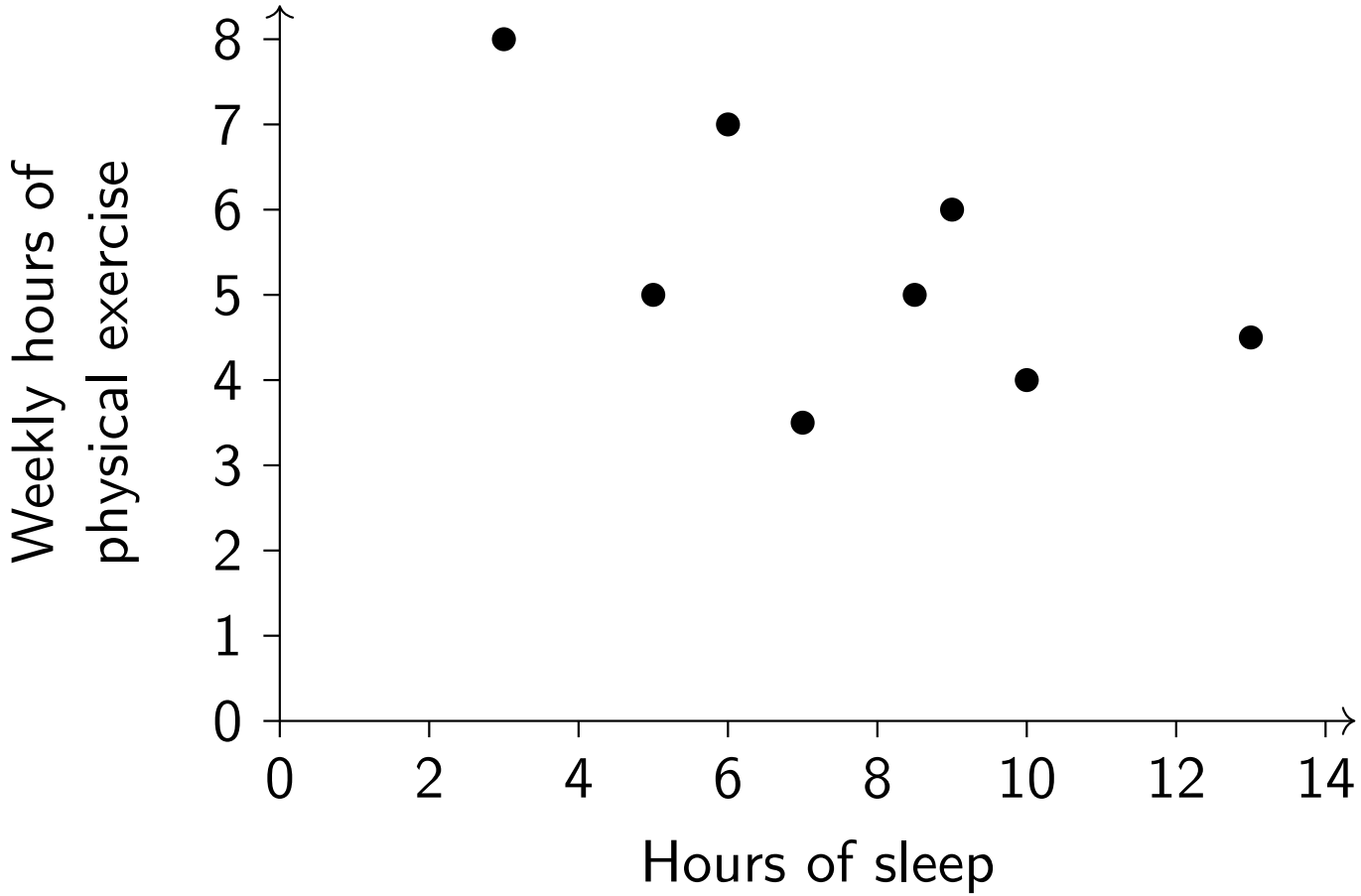
The correlation coefficient is $\rho_{\text{study,exam}} = 0.990$, well above the critical value 0.925 for $p < 0.001$ and strongly indicating positive correlation.

Physical Exercise vs. Exam Results



The correlation coefficient is $\rho_{\text{exercise,exam}} = 0.074$, far less than any critical value and indicating no evidence of correlation for these 8 students.

Hours of Sleep vs. Physical Exercise



The correlation coefficient is $\rho_{\text{sleep,exercise}} = -0.599$, below the critical value of 0.621 for $|\rho_{x,y}|$, so giving no evidence of correlation here.

Judge & Cable 2004

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

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

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

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

The χ^2 Test

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

However, the correlation coefficient only applies to quantitative data.

The χ^2 *test* is statistical tool for assessing correlation in qualitative 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 2015/16;
- 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 last year.

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 the evidence suggests a correlation between coursework submission and exam grade A.

Contingency table

Frequencies

O_{ij}	cw	$\neg$cw
A	O_{11}	O_{12}
$\neg$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	$\neg$cw
A	102	29
$\neg$A	42	34

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

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

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

34 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.

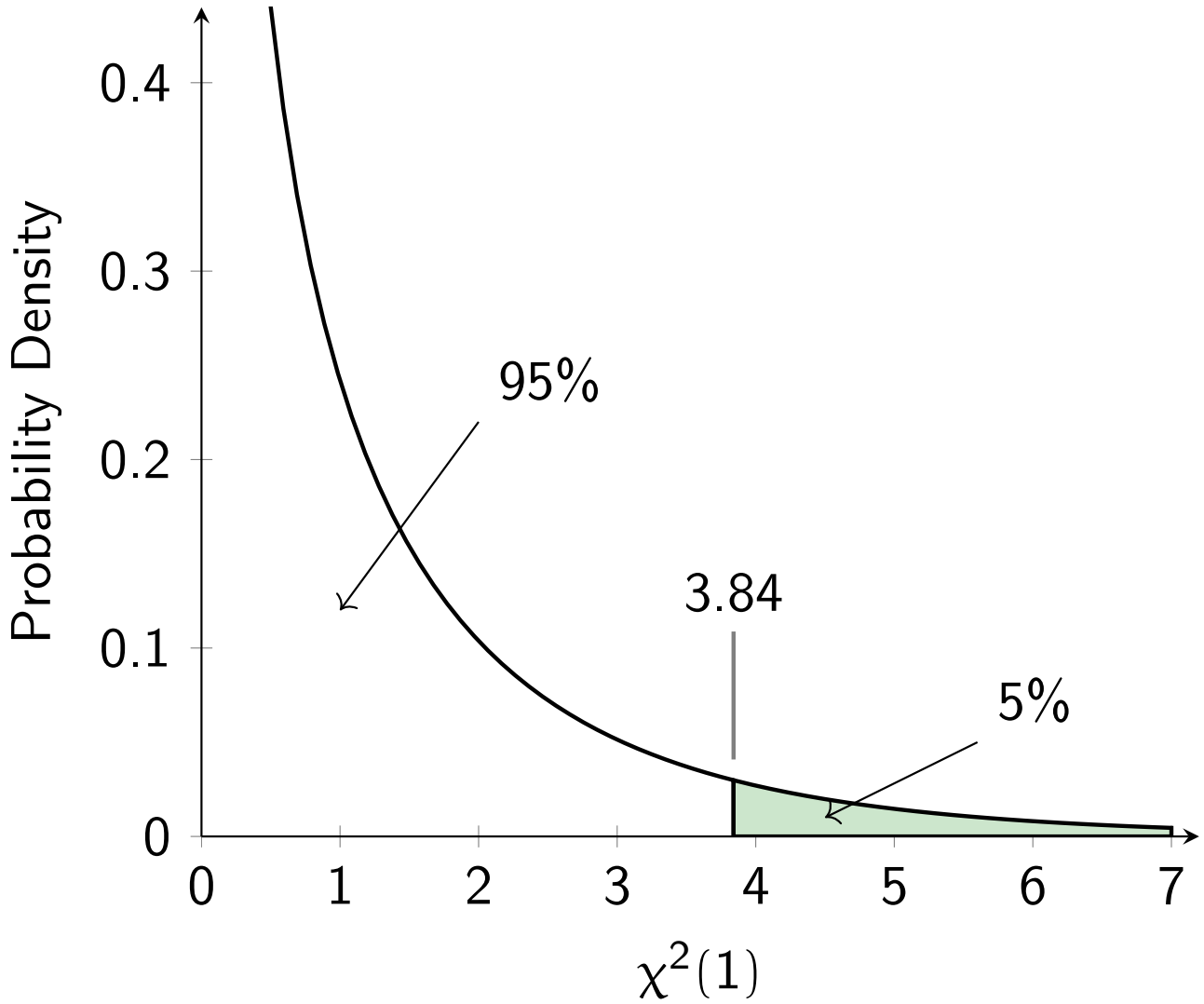
The χ^2 value is calculated by comparing the actual frequencies to the expected frequencies.

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



Marginals

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

$R_1 = O_{11} + O_{12}$ 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	$\neg$cw	
A	E_{11}	E_{12}	R_1
$\neg$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	102	29	131
¬A	42	34	76
	144	63	207

Expected

E_{ij}	cw	¬cw	
A	91.13	39.87	131
¬A	52.87	23.13	76
	144	63	207

The χ^2 statistic for this contingency table is

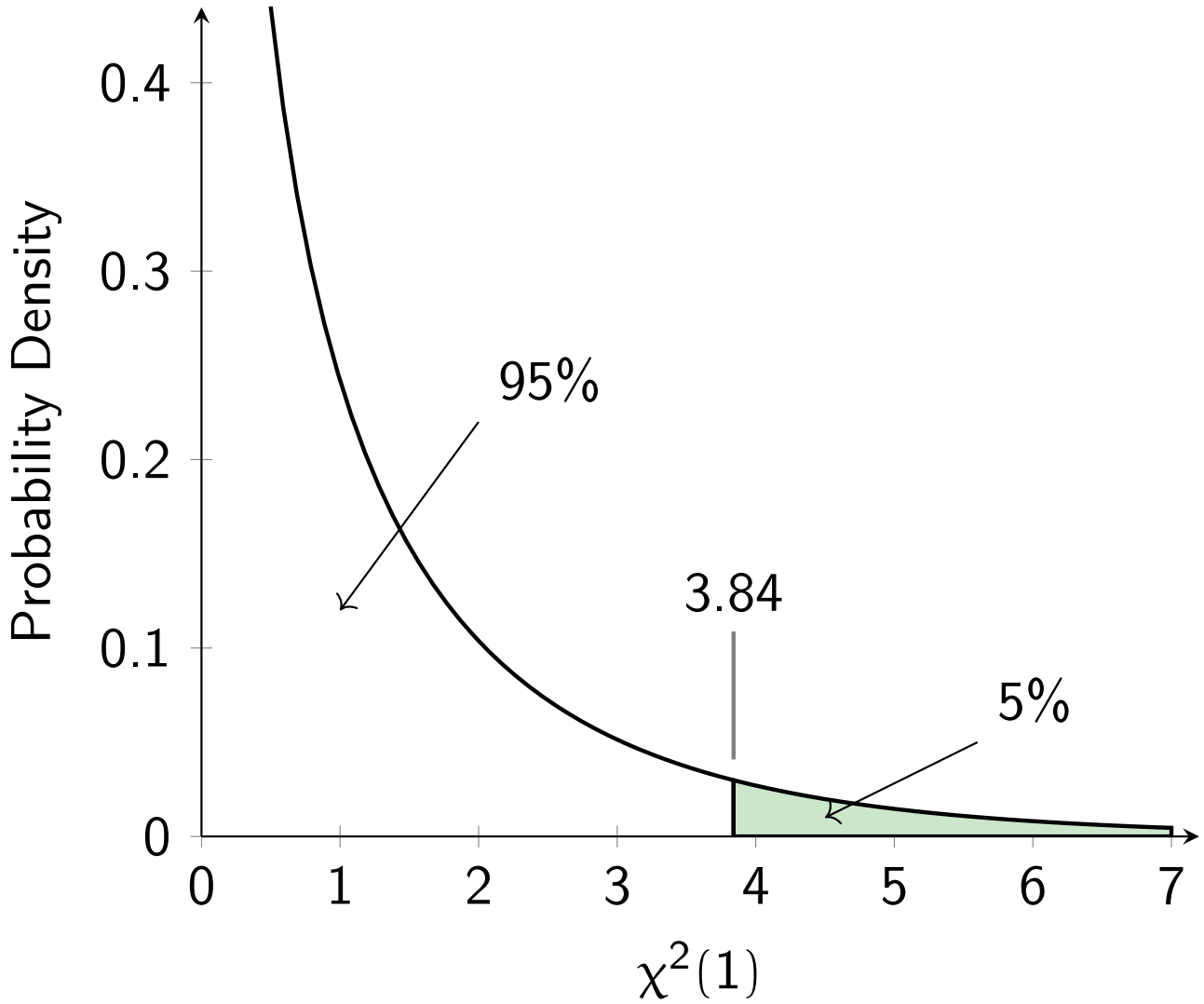
$$\chi^2 = \frac{(O_{11} - E_{11})^2}{E_{11}} + \frac{(O_{12} - E_{12})^2}{E_{12}} + \frac{(O_{21} - E_{21})^2}{E_{21}} + \frac{(O_{22} - E_{22})^2}{E_{22}}$$

$$= \frac{(102 - 91.13)^2}{91.13} + \frac{(29 - 39.87)^2}{39.87} + \frac{(42 - 52.87)^2}{52.87} + \frac{(34 - 23.13)^2}{23.13}$$

$$= \frac{10.87^2}{91.13} + \frac{(-10.87)^2}{39.87} + \frac{(-10.87)^2}{52.87} + \frac{10.87^2}{23.13}$$

$$= 11.60$$

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 a χ^2 value exceeding 2.71 would be $p = 0.1$.
- The probability of a χ^2 value exceeding 3.84 would be $p = 0.05$.
- The probability of a χ^2 value exceeding 6.64 would be $p = 0.01$.
- The probability of a 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 $\chi^2 = 11.60$, meaning $p < 0.001$. This is evidence to suggest that there is a correlation, and we reject the null hypothesis at the 99.9% 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 $(\mathrm{m}-1)\times(\mathrm{n}-1)$ degrees of freedom — given fixed marginals, once there are $(\mathrm{m}-1)\times(\mathrm{n}-1)$ entries in the table the remaining $(\mathrm{m}+\mathrm{n}-1)$ 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

We cannot deduce anything at all from an unreliable test: whatever the χ^2 value, it isn't evidence for anything.

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	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.

What next

Do This

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

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

Read This

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

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	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