Experimental Design, Probability and Statistics

Helen Pain, Alyssa Alcorn and Henry Thompson

Scatter Plots

Allow you to look at data in more than one dimension at once, e.g. age, height and weight.

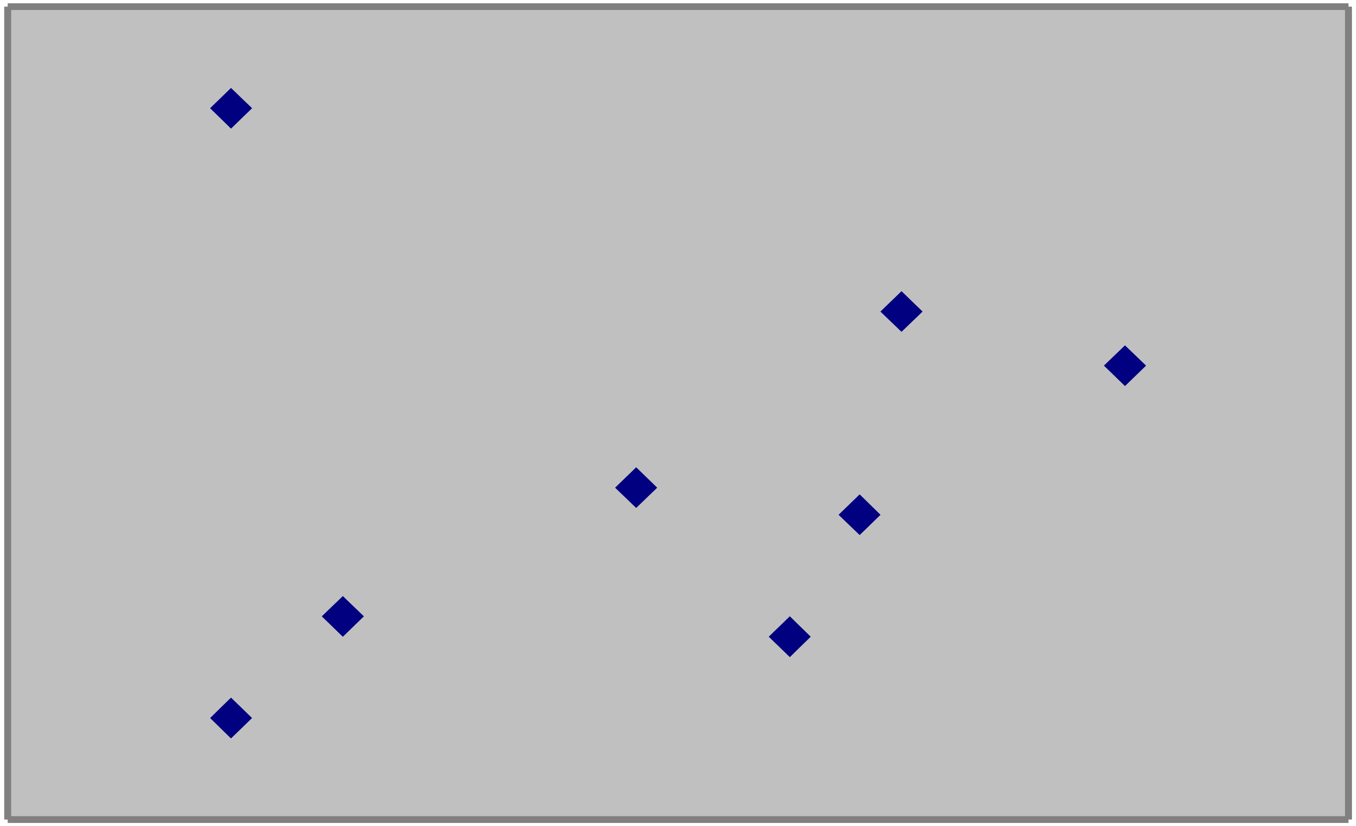
-a large set of triples of numbers, plot these numbers as points in space

- three axes at right angles to each other, triples specify co-ordinates of a point

If there is no relationship between the individual measurements, the points ought to be scattered randomly

If there is an effect, the points will be clustered more densely

Consider scatter plots...

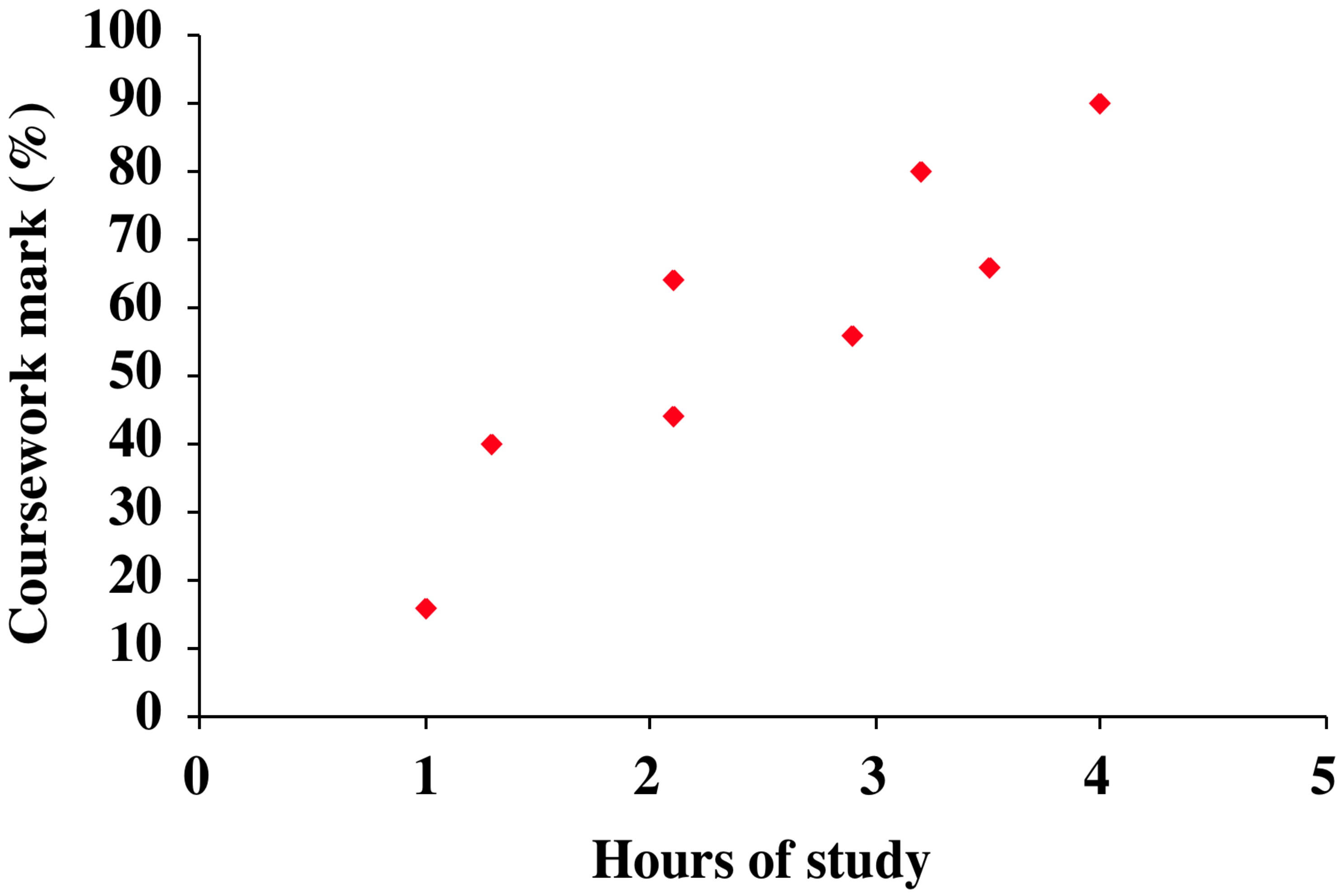


Where would you draw a line?

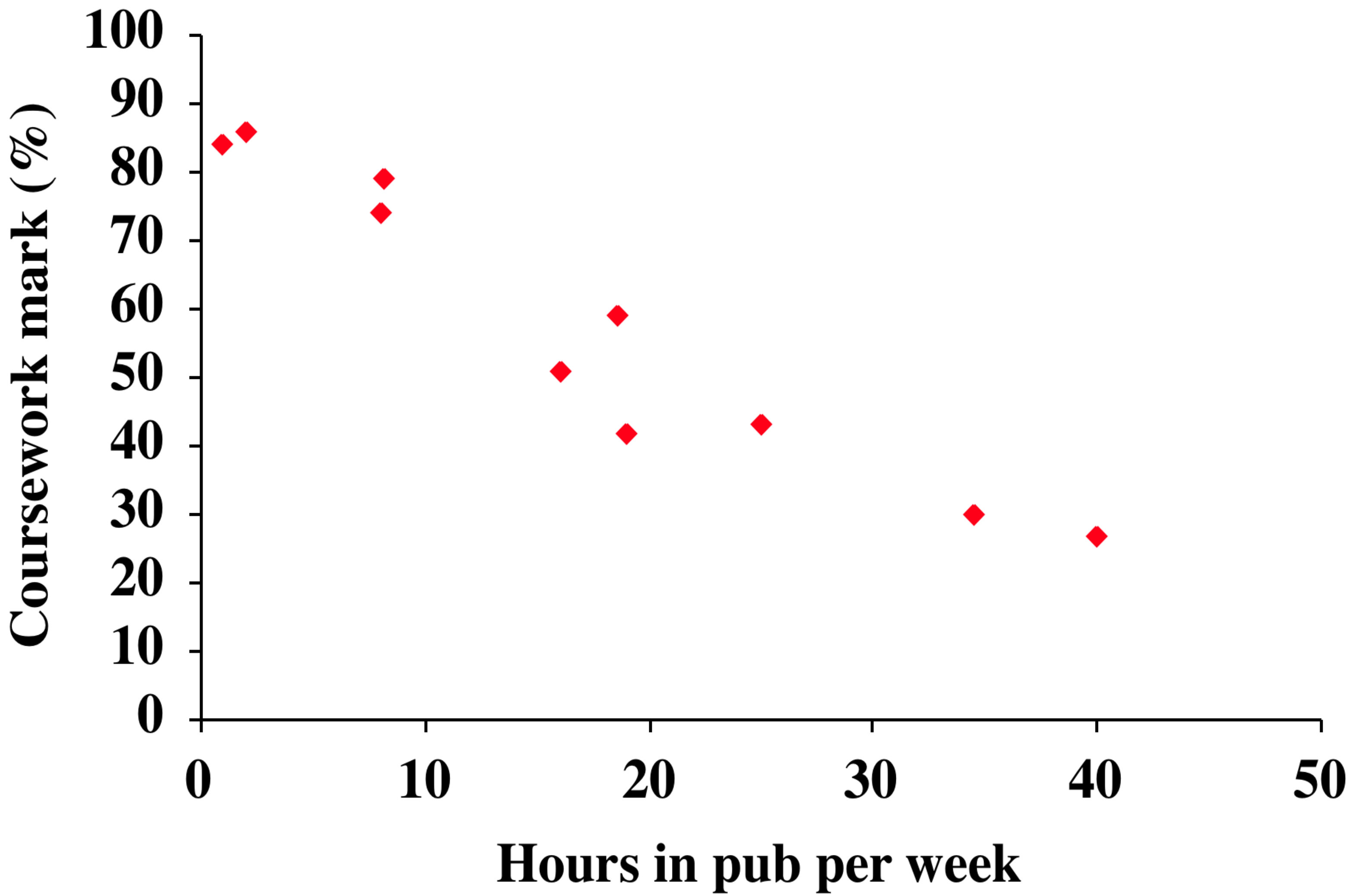
Hours spent v coursework mark

STUDENT								
	a	b	c	d	e	f	g	h
Hours spent	1	1.3	2.1	2.1	3.2	2.8	3.5	4
% on cwork	16	40	44	64	80	56	66	90

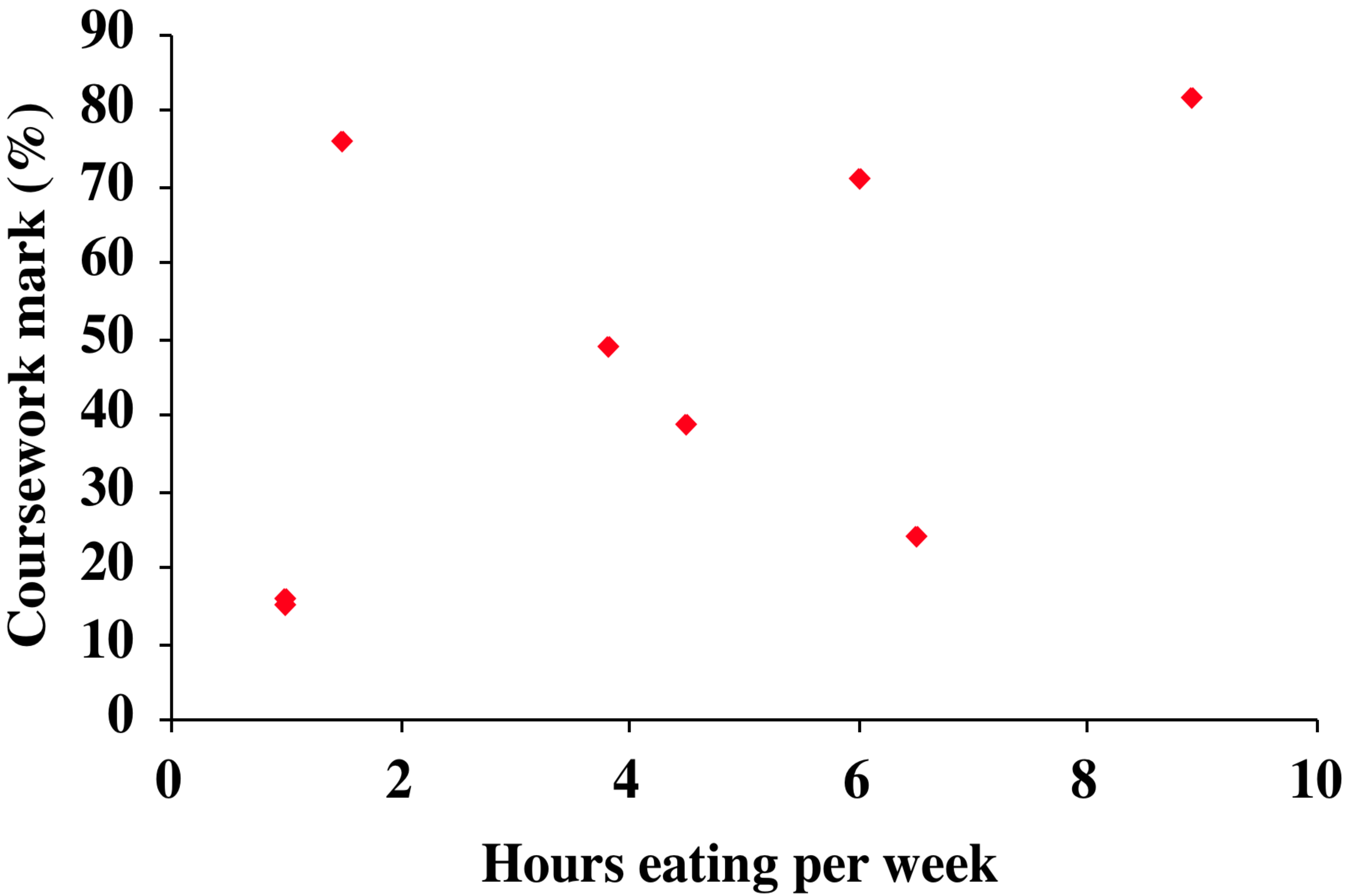
Here we have a Positive Correlation....



What about other factors: may indicate a Negative Correlation



or No correlation…..



Linear correlation

Linear correlation measures ***how well the data fit the model of a straight line*** relationship.

1. **Compute the means** of the x and y data from the scatter plot separately.
2. For each point in the scatter plot (pair of data) **calculate the deviation** of each datum from its mean and multiply, that is:

   compute (x - mean(x))*(y - mean(y))
3. **Sum these products** for all the data pairs **and divide by N-1** for N data.
4. **Work out the standard deviation of x and y separately**, and **divide the sum from step 3. by the product of these** standard deviations.

Pearson's Correlation Coefficient

Measures how well the data fit the straight line model it assumes:

$$\text{correlation} = \frac{\sum \{(x - \mu x)(y - \mu y)\}}{(N-1)\ \sigma x\ \sigma y}$$

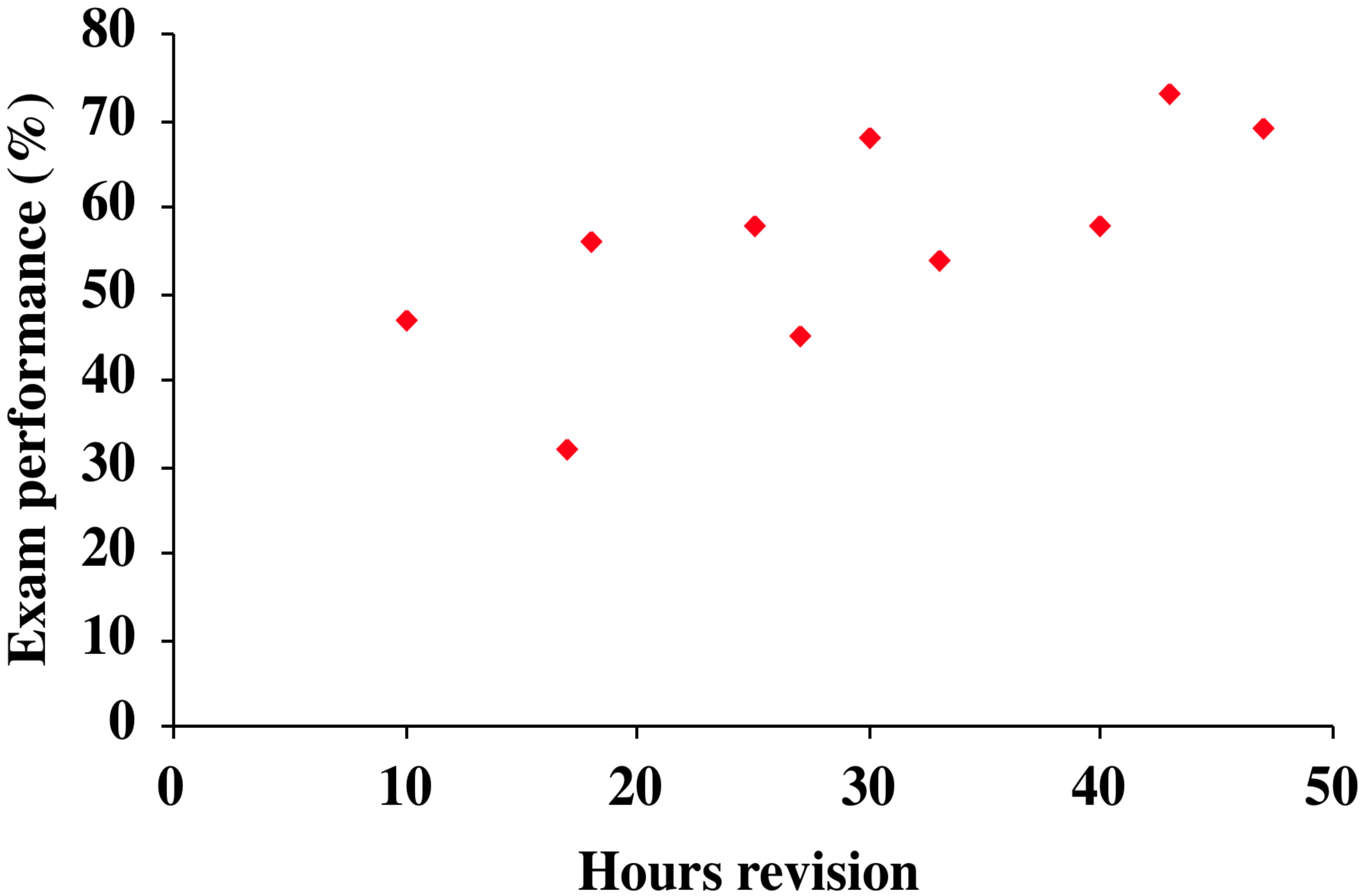
Lies between -1 (low X means high Y) and +1 (high X means high Y) with 0 meaning no correlation

Revision v exam performance

(example from Hinton, 1995)

STUDENT										
	a	b	c	d	e	f	g	h	i	j
Hours studied	40	43	18	10	25	33	27	17	30	47
% on exam	58	73	56	47	58	54	45	32	68	69

Revision v Exam performance



Using Pearson's Correlation Coefficient

correlation = average product of z-scores =

$$\frac{\sum \{(x - \mu x)(y - \mu y)\}}{(N-1)\ \sigma x\ \sigma y} = 0.72$$

To see if this might be due to chance, we need to know the **degrees of freedom** = n-2 = 8

One-tailed test - is correlation +ve or -ve?

Two-tailed test - is there a significant correlation?

Here, +ve correlation predicted, so one-tailed

From tables of probability for one tailed = 0.05, for 8 d.f. r = 0.5494

0.72 is greater than that, so correlation *is* significant with probability it is due to chance less than 5% ($p < .05$)

Interpretation of the size of a correlation

Cohen (1988) suggests guidelines for interpretating correlation coefficient :

Correlation	Negative	Positive
Small	-0.29 to -0.10	0.10 to 0.29
Medium	-0.49 to -0.30	0.30 to 0.49
Large	-1.00 to -0.50	0.50 to 1.00

Criteria are somewhat arbitrary, should not be observed too strictly.

Interpretation depends on the context and purposes.

Correlation of 0.9:

- may be very low if verifying physical law with high-quality instruments
- may be very high in social sciences where there may be a greater contribution from complicating factors.

Comments on Correlation…

A **high positive correlation** between two variables **doesn't mean that one causes the other**……

Say we get a correlation of 0.8 between exam performance and hours of study:

Does this mean that the longer you study the better your exam results will be?

- or the better the exam results the more you will study?
- or some other variable influencing both (you are conscientious and bright)

Or ***time spent watching television and incidence of lung cancer are correlated***, but neither causes the other:

- both are caused by economic factors providing people with leisure time and money to buy cigarettes…

Statistical dependence is not the same thing as causal dependence.

χ^2 Test

The χ^2 (chi-squared) test:

- compares n *frequency distributions*, each with m values;
- tests the *null hypothesis* that the distributions are the same;
- takes as its input an $n \times m$ *contingency table*.

Example

Compare performance of boys and girls in an exam with marks A, B, C, and D. Data: 4×2 contingency table, with marks on x-axis and distribution on y-axis.

χ^2 Test

Example: exam data

O_{ij}	A	B	C	D	$\sum_i O_{ij}$
Boys	3	23	43	10	79
Girls	6	34	31	4	75
$\sum_j O_{ij}$	9	57	74	14	154

Compute χ^2 statistic by comparing:

- *observed frequencies:* frequencies that have been observed experimentally, and
- *expected frequencies:* frequencies that would be expected if the null hypothesis was true (no difference between the distributions).

χ^2 Test

Equation for χ^2:

$$(1)\quad \chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

i: ranges over rows of the contingency table; j: ranges over its columns; O_{ij}: observed frequency for cell (i, j); E_{ij}: the expected frequency for cell (i, j)

Equation for *expected frequencies:*

$$(2)\quad E_{ij} = \frac{\sum_j O_{ij} \sum_i O_{ij}}{N}$$

N: overall number of observations; $\sum_j O_{ij}$ and $\sum_i O_{ij}$: *marginals* of contingency table.

Example

Example: exam data

Calculate the expected frequencies for the exam data:

E_{ij}	A	B	C	D
Boys	4.62	29.24	37.96	7.18
Girls	4.38	27.76	36.04	6.82

Now compute χ^2 and compare it against the *critical value:* if it exceeds it, the null hypothesis can be rejected, test is *significant.*

Example: exam data

Plug the expected frequencies into (1): $\chi^2 = 7.55$. This doesn't exceed critical value of 7.82 (get this from a stats book): exam performance of boys and girls not significantly different.

Notes on tables of critical values

Spreadsheet and statistical packages have means of accessing critical values for many distributions, including χ^2

Tables of critical values can also be found online, for example http://www.itl.nist.gov/div898/handbook/eda/section3/eda3674.htm

Visualisation Techniques

Visualisation techniques - used for exploratory data

- make patterns in data apparent to human analyst,
- display visually relationships between different data variables

Tools for this include MATLAB, matrix manipulation system with excellent graphical display abilities.

Apparent effects can be confirmed by simple statistical techniques

- allows us to determine extent to which anticipated effect is present in data from experiment

Histograms (bar charts)

Shows how many data fall into each of a number of classes

Record temperature at noon each day for a year, then count how many days between 16 and 17 Celsius, 17 and 18, and so on.

- plot of ***the number of days*** (vertical axis) against the various ***temperature categories*** (horizontal axis)
- shows the ***distribution*** of the data

Multiple peaks indicate something going on

Split set of data into clusters associated with peaks

Investigate whether members of the clusters differ from each other in consistent ways.

e.g. peaks around 25 and 16 with trough in between;

days in **25** cluster **'bright'**, but in **16** cluster **'cloudy'.**

Infer bright days are hotter than cloudy ones

Evaluating Usability Example

We have developed an interface for a variety of users to use

We ask users to rate the usability of the interface as:

1. easy to use

2. average

3. difficult to use

We test it on different groups of users, recording how many users select each rating, for each of:

a. children (under 12 years)

b. teenagers (13 to 18 years)

c. adults (over 18 years)

If there is no consistency of usability then the ratings should be equally spread across 1 to 3 ratings.

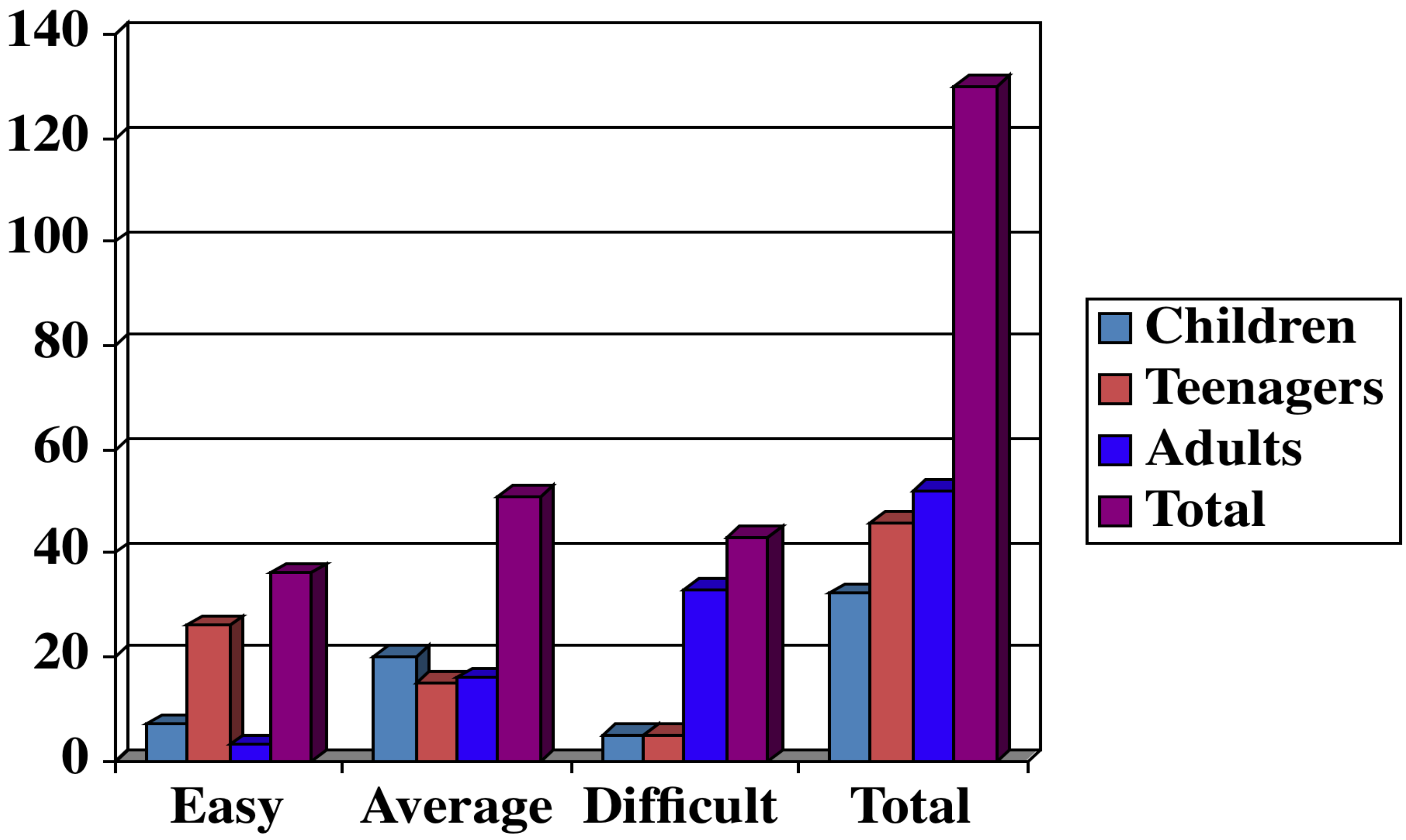
Is there a difference between different users?

Usability: by age group and ease of use

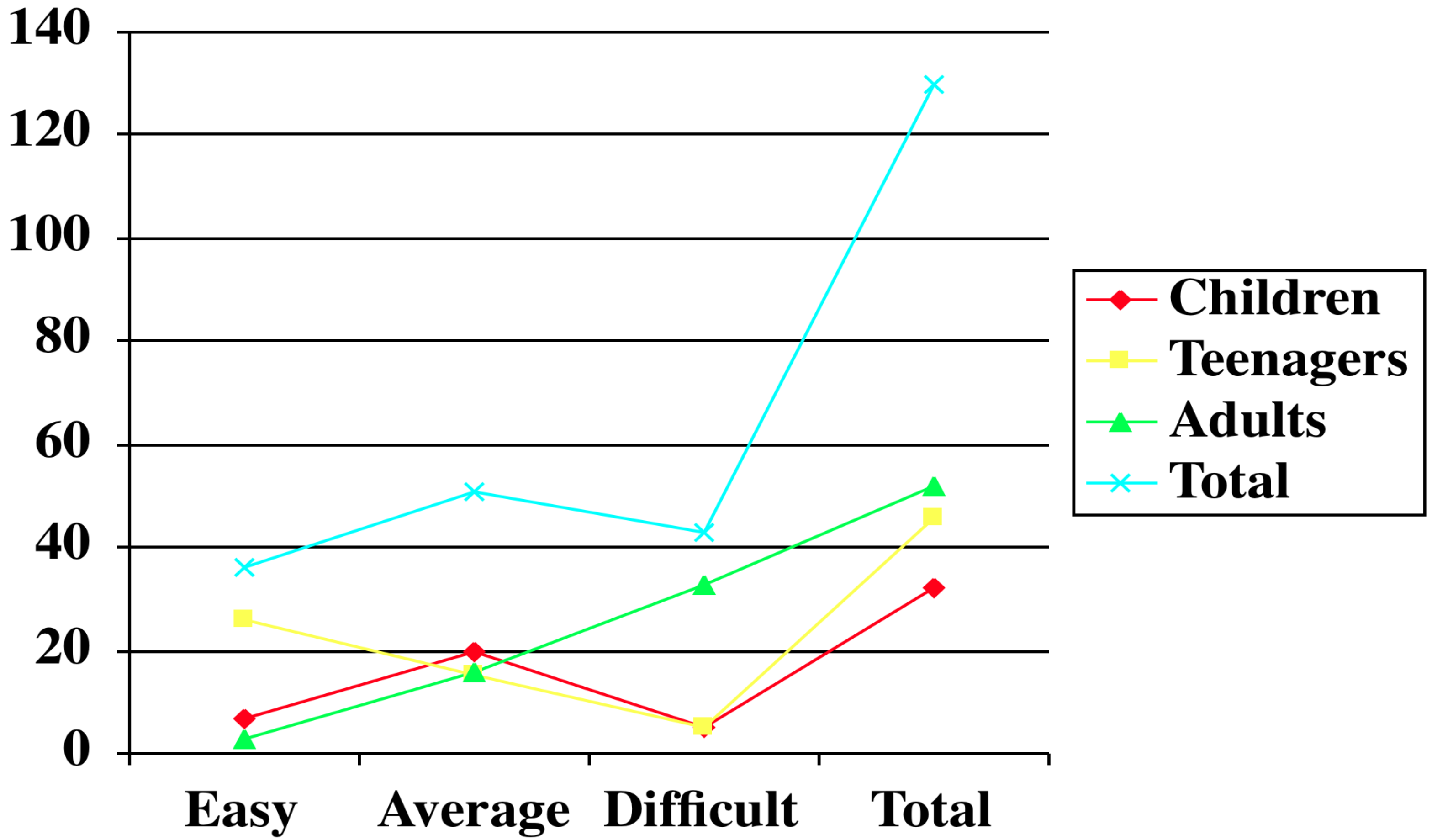
Ratings:	easy	average	difficult	Totals
Children	7	20	5	32
Teenagers	26	15	5	46
Adults	3	16	33	52
Total	36	51	43	130

There are various ways that we an visualize this data – some more helpful or appropriate than others

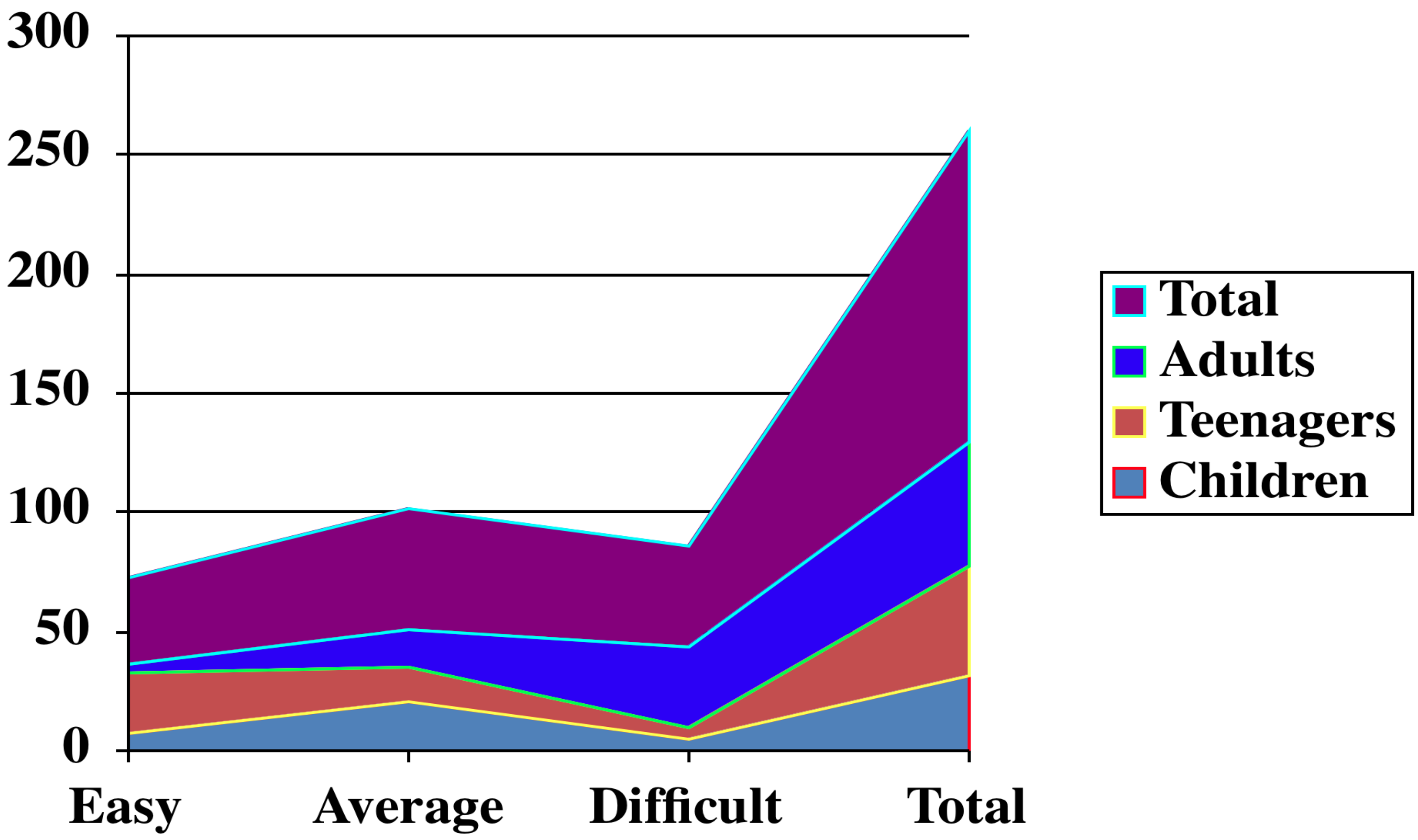
Age group v ease of use: Bar chart



Age group v ease of use: graph



Age group v ease of use: area chart



More information

Statistical graphics, also known as **graphical techniques**, are information graphics in the field of statistics used to visualize quantitative data.

http://en.wikipedia.org/wiki/Statistical_graphics

Gallery of Data Visualization
The Best and Worst of Statistical Graphics

http://www.datavis.ca/gallery/index.php

Writing-up empirical studies 1

There are standard ways of writing up empirical studies – you will have seen a number of examples in the research papers that you have read...

Title:

The shortest description of the study

Abstract:

Short summary of the problem, the results and the conclusion.

Introduction:

What is the problem? What related work have other people done?

[Should go from general statement of the problem to a succinct and testable statement of the hypothesis].

Writing-up empirical studies 2

Method:

Participants: state number, background and any other relevant details of participants

Materials: exactly what test materials, teaching materials, etc. were used, giving examples

Procedure: clear and detailed description of what happened at each stage in the experiment

[Someone reading should be able to duplicate it from this information alone. Should also clearly indicate what data was collected and how.]

Writing-up empirical studies 3

Results:

Give actual data, or a summary of it.

Provide an analysis of data, using statistical tests where/if appropriate.

Use tables and graphs to display data clearly.

*[Interpretation of results does **not** go here, but in discussion section].*

Writing-up empirical studies 4

Discussion:

Interpretation of results; restating of hypothesis and the implications of results; discussion of methodological problems such as weaknesses in design, unanticipated difficulties, confounding variables, etc.

Wider implications of the work should also be considered here, and perhaps further studies suggested.

Conclusion:

Statement of overall conclusion of the study.

References:

All publications cited in the text should be listed here using standard formats

Useful sources

Choosing tests – search on ‘statistical tests for research’

Look at:

http://www.socr.ucla.edu/Applets.dir/ChoiceOfTest.html

see also

http://www.wadsworth.com/psychology_d/templates/student_resources/workshops/stat_workshp/chose_stat/chose_stat_05.html

General stats workshop

http://www.wadsworth.com/psychology_d/templates/student_resources/workshops/stats_wrk.html

in particular “choosing the correct statistical test”

References

Cohen, P. (1995) *Empirical Methods for Artificial Intelligence*, MIT Press, 1995.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.) Hillsdale, NJ: Lawrence Erlbaum Associates.

Corbett, A.T. and Anderson, J.R., (1990) The Effect of Feedback Control on Learning to Program with the Lisp Tutor, *Proceedings of the 12th Annual Conference of the Cognitive Science Society*, LEA, New Jersey, 1990

Dix, A., Finlay, J., Abowd, R. and Beale, R. (2004) *Human-Computer Interaction.* Prentice Hall