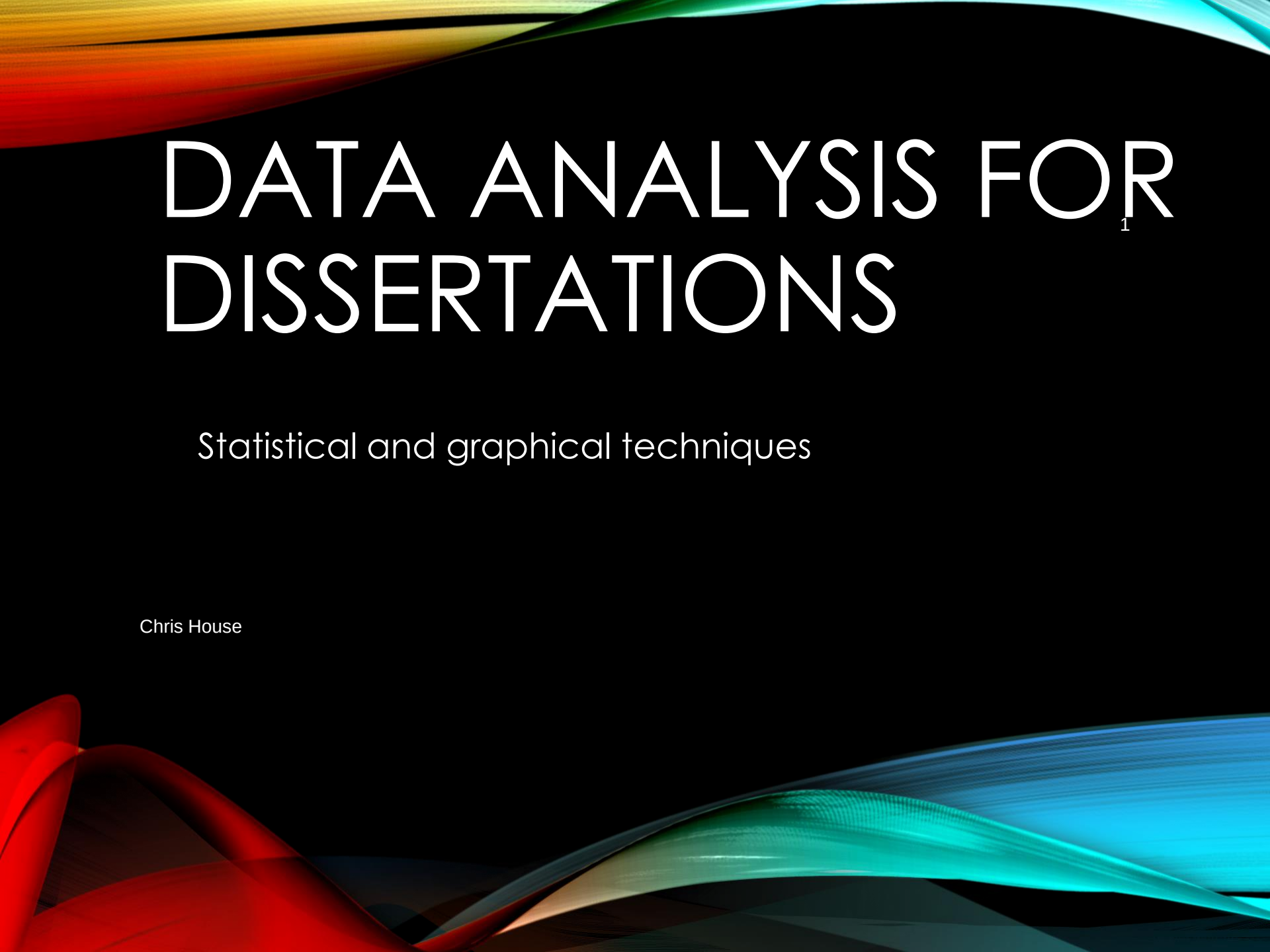




DATA ANALYSIS FOR DISSERTATIONS₁

Statistical and graphical techniques

Chris House

PLOTTING DATA²

- ▣ Introduction
- ▣ Bigwood, S. & Spore, M. (2003) Presenting numbers, tables and charts. Oxford University Press (Classmark 808.021 PRE)
 - ▣ This is an excellent summary text of how to present data and is a very important text for your dissertations.
- ▣ Imagine this hypothetical set of data.
 - ▣ Asbestos (in milligrams per air sample) at Wembley construction site
 - 0.9 0.9 0.7 0.8 0.7 1.0 0.8 0.8 0.9 0.7 1.2 0.6 0.5 0.8 1.0 1.0 0.8 0.6 0.8
 - ▣ This data can be presented in a tabular or graphical form to make its interpretation easier. Statistical tests can then be used to reduce the data into a smaller but representative form.
- ▣ **Frequency tables**
 - ▣ x = data value
 - ▣ f = frequency
 - ▣ n = Sample size
 - ▣ Spread combine data into classes of equal size (Boundaries should be mutually exclusive (don't overlap))
 - ▣ *Ex 1 Construct a FT of the above values and one of classes.*

ANSWER EXERCISE 1

x	f	x	f
0.5	1	0.40-0.49	0
0.6	2	0.50-0.59	1
0.7	4	0.60-0.69	1
0.8	6	0.70-0.79	2
0.9	3	0.80-0.89	5
1.0	3	0.90-0.99	5
1.1	0	1.00-1.09	5
1.2	1	1.10-1.19	0
Chris House		1.20-1.19	1

FREQUENCY HISTOGRAMS

- Graphically presents the data most frequent is the mode. Helps assess symmetry and when plotted is called a frequency distribution. Symmetry increases (generally) with sample size (normal distribution). Many stats assume ND for their tests. Positive and negative skew. If there are 2 peaks it is called bimodal.
- *Construct a histogram and frequency curve of the data.*

BAR CHART

- ▣ Gaps between columns and is used for **discrete data** (4.5 people, abstract) (and therefore a bar chart i.e. gaps) and not continuous data (4.5 metre tree/timber) (histogram).
- ▣ Number of places in a site with Japanese knotweed identified at different planning approved development sites.



Site Frequency

1	3
2	4
3	8
4	5
5	2
6	4
7	5
8	1

- ▣ *Construct a frequency distribution of the above data (Bar chart).*

MEASURES OF CENTRAL TENDENCY

- ▣ Distributions are not always appropriate descriptions therefore single values can be more useful
- ▣ **Mode:**
 - ▣ Most frequently occurring (particularly useful in nominal data)
(How 100 conservation or construction workers travel to work)
- ▣ **Arithmetic mean**
 - ▣ Calculate the sample mean and use it to estimate the population mean ('ex bar'). NB only useful if you have a normal distribution.
- ▣ **Median**
 - ▣ Middle data point in order. Particularly useful when you have outliers since the value is not influenced by these outliers.

MEASURES OF CENTRAL TENDENCY 2

- If the distribution is symmetrical then all 3 will be the same but any other pattern will give different figures for each form of central tendency. Always include (n) because it demonstrates the level of representation of your measure of CT
 - $N+1/2$
 - e.g. cost of renewable energy projects in millions £
 - 2 4 3 5 6 6 5 4 3
 - Ascending 2 3 3 4 4 5 5 6 6
 - $9+1/2=5$
 - Then the median is the middle value 4 (i.e. fifth point)
 - When 10 values $10+1/2=5.5$
 - 2 3 3 4 4 5 5 6 6 6 median 4.5

EXERCISE

8

- *What are the Mode, Mean and Median of the following?*
 - Tree heights in metres or
 - Meters of wall rendered per hour
- 15 24 28 25 18 24 27 16 20 22 23 18 22 28 19 16 22
26 15 26 24 21 19 27 16 23 26 25 18 27 17 20 19 25
23 24



Chris House



CUMULATIVE FREQUENCY CURVES (OGIVE) (HAMMOND & MCCULLAGH, 1998:7)

9

- ▣ Another method of depicting distribution.
- ▣ e.g. population is normally presented in a pyramid format but a CF curve has this adv.
 - ▣ Visual comparison between age structures
 - ▣ Pattern is not influenced by unequal age bands
 - ▣ Statistics can be easily read from curve
- ▣ *From the handout assess what points ABC represent.*
- ▣ *Draw an Ogive for the age structure of England & Wales 1971 (page 8) and compare it to the 1861 curve (p.9). What does this demographic mean for housing or environmental policy in the UK?*

THE LORENZ CURVE

- ▣ Used with nominal scales (category) and helps measure dispersion. For example, industrial employment structure (industrial diversification) or Bird species visiting garden feeding stations. How it differs from a hypothetical even distribution.
- ▣ Based on percentages therefore comparable.
- ▣ Standard Industrial classification orders.
- ▣ Can work out the area of the curve to give a clear measurement, but this is time-consuming. However, if you add the cumulative frequencies a crude index is identified. Nottingham 1156 shows more diversification of industry than Coventry 1311.
- ▣ ***Construct a LC of the following data set***

RECYCLING

11

What does the LC tell us about recycling strategies in housing association sites?

Recycled	2006 Kg	%	2016 Kg	%
Glass	3,900	39	3,550	20
Cans	3,000	30	3,400	18
Paper	2,350	23.5	4,700	26
Plastic	550	5.5	3,000	16
Green waste	50	0.5	2,500	14
Other	150	1.5	1,050	6
total	10,000	100	18,200	100

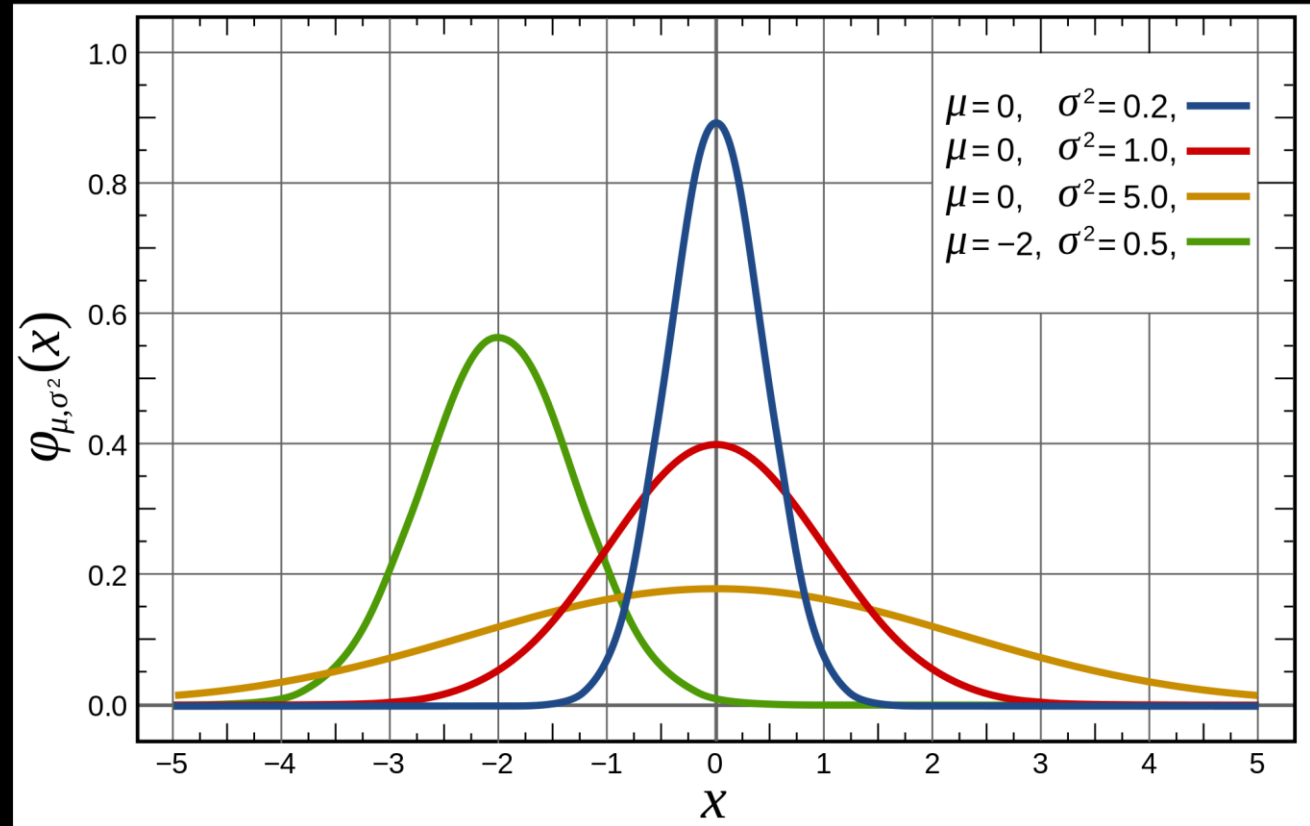
MEASUREMENT OF VARIATION

- ▣ Central T does not describe the variation in the data. The range is a simple measure of variation but outliers misrepresent the data set. E.g.
- ▣ Age of carrion Crows in Brecon Beacons National Park or age of labourers on site:
- ▣ 16, 17, 18, 18, 19, 20, 20, 20, 21, 21, 22, 23, 23, 24, 26, 29, 55. The mean and range would be biased due to the 55 year old. Hence we need a form of measurement that reduces the influence of outliers. Standard Deviation is one such method for normally distributed data.

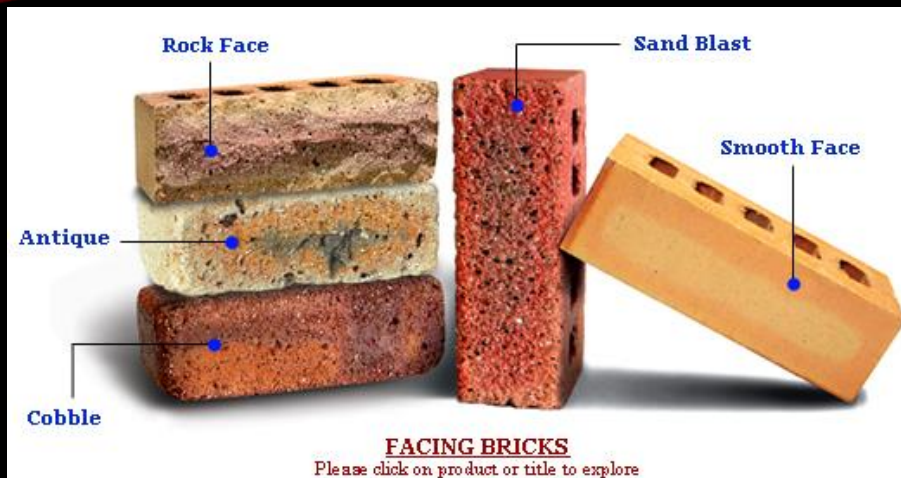
VARIABILITY AND SD

μ -Mean, Median, Mode

Variance - σ^2



STANDARD DEVIATION



Little Bustard

- ▣
$$S = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$$
- ▣ Refer to handout p. 32-37
- ▣ Using the data on the next slide:
 - ▣ Calculate the mean
 - ▣ Note $\sum x$ is the sum of the x values, and is given at the base of the column
 - ▣ Calculate the square of each of the values of x
 - ▣ Then follow the formula.
 - ▣ The result should be presented one decimal place more than the mean.



WORKED EXAMPLE 2.1 Calculating the standard deviation (s) for sulphur dioxide levels in rainfall

x (mg)	x^2	Calculation
0.7	0.49	Calculate the mean ($\sum x$ is the sum of the x values, and is given at the base of the x column; n is the sample size of 20):
0.9	0.81	
0.9	0.81	
0.7	0.49	
0.8	0.64	Calculate the square of each of the values of x (shown in the second column). From the two columns, use $\sum x$ and $\sum x^2$ and the sample size in the following formula. Notice the distinction between $\sum x^2$ and $(\sum x)^2$: the former is the sum of the squared values (those in the second column) and the latter is the sum of the x values (those in the first column) which is then squared.
0.7	0.49	
1.0	1.00	
0.8	0.64	
0.8	0.64	
0.9	0.81	
0.7	0.49	
1.2	1.44	
0.6	0.36	
0.5	0.25	
0.8	0.64	
1.0	1.00	
1.0	1.00	
0.8	0.64	
0.6	0.36	
0.8	0.64	
$\sum x = 16.2$		$\bar{x} = \frac{\sum x}{n} = \frac{16.2}{20} = \underline{0.81 \text{ mg}}$ Mean
$\sum x^2 = 13.64$		$s = \sqrt{\frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}} = \sqrt{\frac{13.64 - \frac{16.2^2}{20}}{19}} = \sqrt{\frac{0.518}{19}}$$= \sqrt{0.0272631} = 0.165115414 \text{ mg}$ Note that we report the standard deviation to one decimal place more than the mean, so in this example $s = \underline{0.165}$. Remember that s^2 is the variance. In this example: $s = \sqrt{0.0272631}$ and: $s^2 = 0.0272631$

STANDARD DEVIATION 1 ¹⁶

Work out the SD for **each** these hypothetical data sets

Sparrow eggs Or QS wages SA1	Sparrow eggs Or QS wages world	Bird (g) Or Timber lengths	Birds (g) Or Timber lengths
2	1	1	100
2	2	100	102
2	3	200	104
2	4	300	100
2	5	400	101
	6	500	103
	7	600	102
Answers Below	8	700	100
Chris House	2.4	244.75	1.51

STANDARD DEVIATION 2¹⁷

1. Compressive strength of cheap bricks (N/mm²) OR
2. Weight of Great Bustard chicks *Otis tarda* (Salisbury Common)

0.7	0.9	0.9	0.7	0.8	0.7	1.0	0.8
0.7	1.2	0.6	0.5	0.8	1.0	1.0	0.8

1. Calculate the Mean
2. Do a SD of the above data set?
3. Work out the +/- 1 SD
4. Evaluate your results
5. SD 0.175

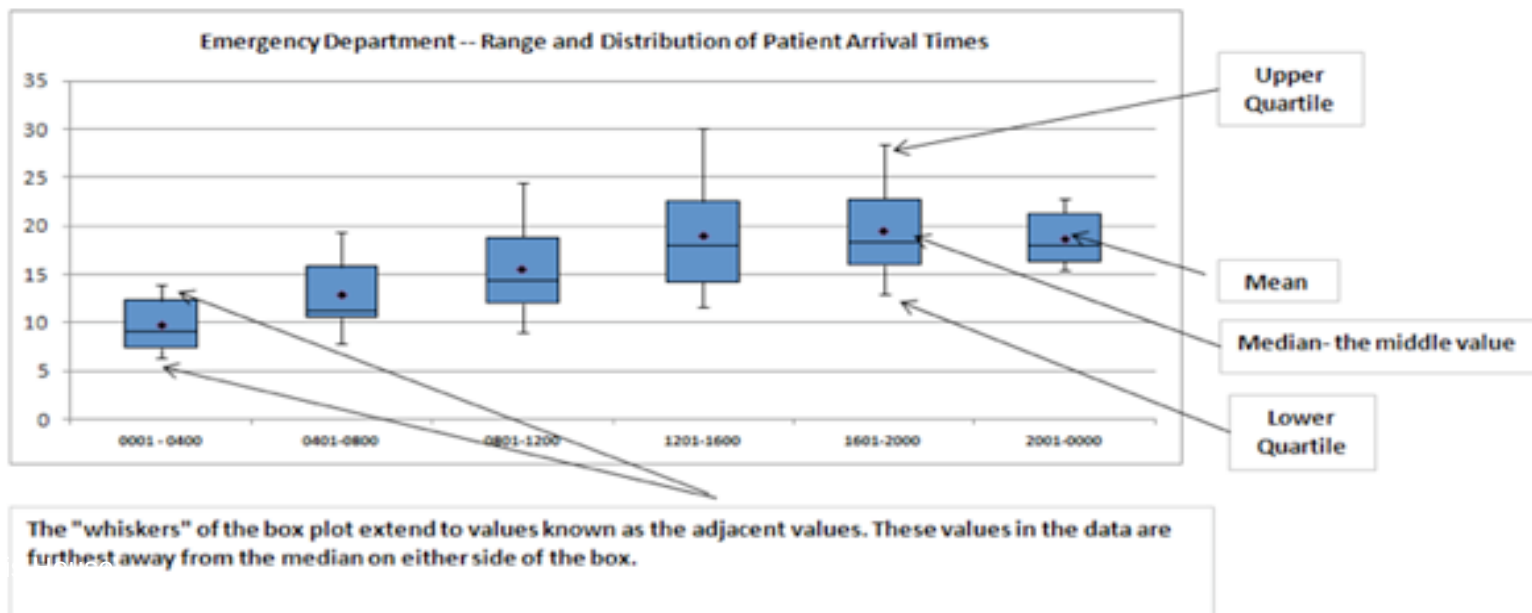
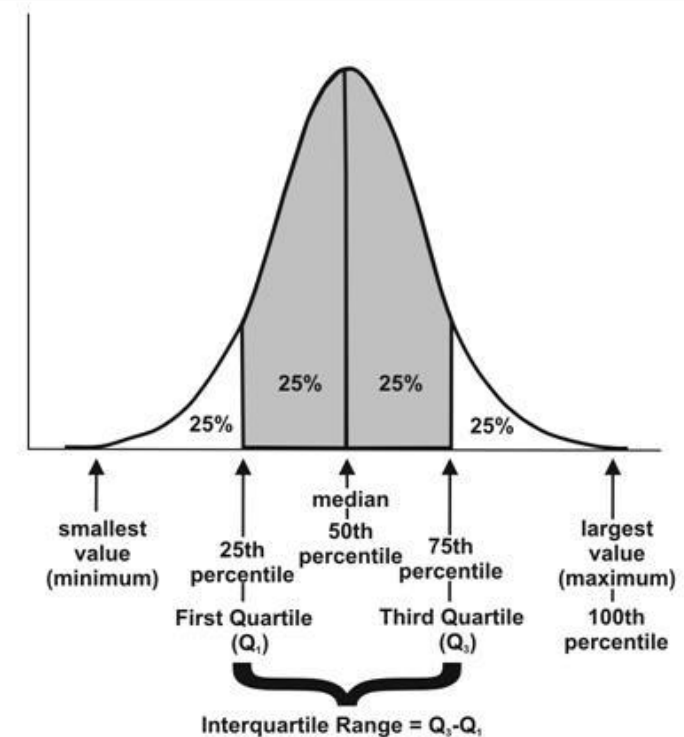
STANDARD DEVIATION: ERROR

- ▣ The SD gives a measure of variability. Larger SD (relative to the mean) show a larger variation in the data.
- ▣ Mean 0.81 (Handout data sulphur)
- ▣ SD 0.165
- ▣ Mean plus 1 SD = 0.975 N/mm^2 (*Newton per square millimetre*)
- ▣ Mean minus 1 SD = 0.645 N/mm^2
- ▣ Therefore within a normal distribution 68.27% of the data points are within these boundaries. Probability (see handout)

- ▣ **Standard Error**
- ▣ Reliability of sample mean: Divide the sample variance by n then square root

INTERQUARTILE RANGE

- **Variation in ordinal data**
 - Interquartile range (handout)
 - $Q^1 (n+1)/4$
 - $Q^2 3(n+1)/4$
 - **Box and Whisker**
 - What is the IR for the Great Bustard data/Brick data ?



HOW SHOULD WE PRESENT THE DATA?

- Depends on the data and the **reasons** for display.
- Most important results should be presented in a graphical form (**Figures**).
- **However, be selective.**
- **Tables** help summarise a series of results. Tables: Variables horizontally, classes vertically.
- Additional information can be included in footnotes.
- Labelled 2.2 etc. Referred to in the text.
- Clear headings. **Figures on top/ Tables on bottom**
- Located directly **after** the point they have been referred to.
- Other methods include box and whisker plots (handout)
- Label **both axis**
- Note units on axis
- Scale should be appropriate
- If there are 2 variables **scatter plots** are useful and with nominal data **cross tabulation** tables are appropriate (handout). No line required unless mathematically constructed. Regression causal relationships.

SCATTERPLOTS

Annual Gas bill	654	589	693	799	753	666	682	400	775	503
Age	52	21	30	36	39	28	42	64	41	58

1. Draw a scatter plot of the above data set.
2. Evaluate the graph.
3. Do a SD of age and evaluate
4. X axis independent
5. Y axis dependent

OTHER GRAPHICAL PRESENTATIONS

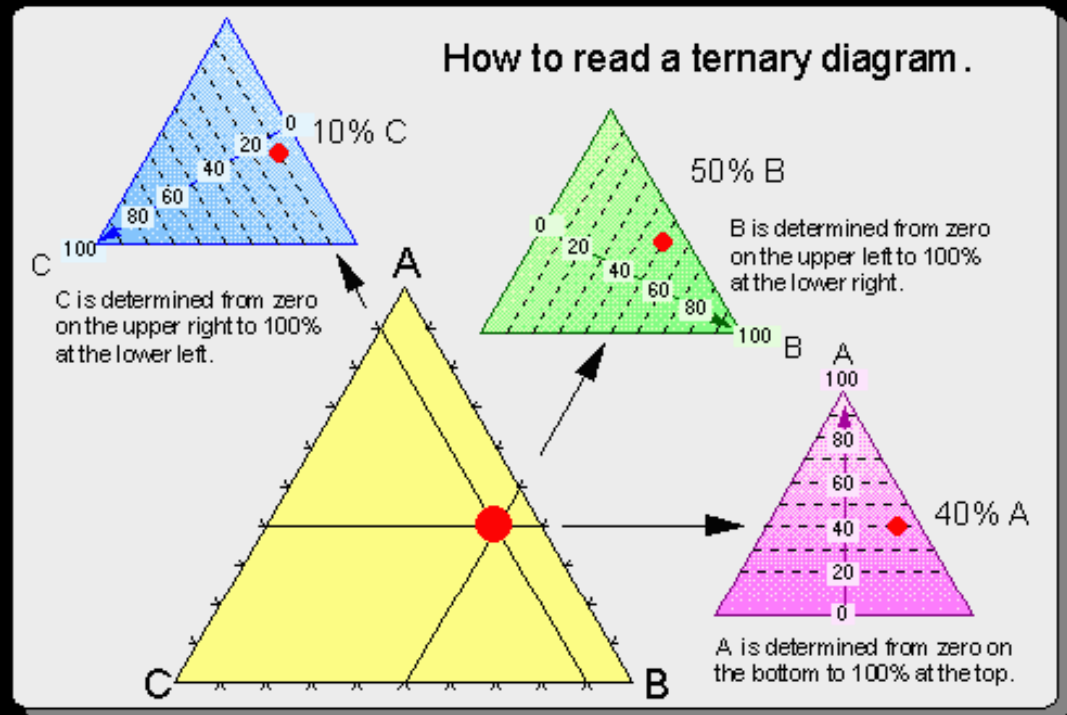
- **Ternary** Diagrams (Handout)

- Are scatter plots for 3 variables? Interpret the Ternary in the handout and assess how numbers 1-10 could influence planning deliberations for a wind farm construction.

- **Pie charts** are especially used when you want to compare frequencies with the whole data set.
- Selection of graph types

[Link](#)

Chris House



HYPOTHESIS TESTING (HYPOTHETICAL DEDUCTIVISM)

23

Specify the null hypothesis

Transform data
(If applicable)

Calculate the
test statistic

Note P Value

$P > 0.05$

Accept Null hypothesis

$P < 0.05$

Reject null hypothesis

Accept the Alt. H. and find
direction of relationship
/difference

- For the data set on the following slide
 - Construct a null and alternative hypothesis
 - Calculate the Means
 - Graphically present the data
 - Calculate the SD
 - Evaluate the results



Chris House

REVISION

24



COST OF RENEWABLE ENERGY VS NON-RENEWABLE HOUSING PROJECTS²⁵

Aim: Are unit costs different for these methods of construction?
Unmatched data set.

Non Renewable costs £000's	Renewable projects £
564	505
521	557
445	465
560	458
480	530
540	480
585	665
426	525
530	605
475	495

RESULTS

26

- Construct a null and alternative hypothesis
 - No difference between unit costs of renewable and non-renewable construction.
- Calculate the Means
 - Non renew. 512
 - Renewable 528
- Graphically present the data
 - Discrete
- Calculate the SD
 - Non renew.
 - Sum $x=5126$,
 - Sum $x^2 = 2,653,388$
 - SD=53.54
 - Renew.
 - Sum $x=5285$
 - Sum $x^2 = 2,831,463$
 - SD=65.27
- Evaluate the results



HOW TO USE STATS TO ADDRESS RESEARCH QUESTIONS

Scientific conclusions that are needed

1. Differences between categories (e.g. Soil types)
2. Relationship between variables (e.g. Distance and cost)
3. Associations between frequency distributions (proportions of people choosing sustainable homes)

STAGES OF INFERENCEAL STATS

- Differences, Relationships and Associations are tested using inferential stats. The extent to which the data is by chance
 1. Hypothesis testing (NB different to scientific hypotheses)
 - Null 'no difference'
 - Alternative 'is a difference'
 2. Test statistic
 - Single number and associated probability value
 - They have known mathematical properties
 - Ratio of explained versus unexplained
 3. Statistical significance
 - $P=0.05$ (95%) & $p=0.01$ (99%) (computers)

CHOOSING A STATISTICAL TEST

- There are many types of stat test
- They normally have specific criteria and need to be carefully selected for purpose
- Selection of statistical tests
- **Parametric** (normally distributed)
- **Nonparametric** (distribution free)
- Assumptions
- Parametric tend to be more powerful
- Assess your data for 'normal distribution'
 - Bell shaped curve
 - 68% data is within mean $\pm$ SD
 - Data can often be made normal through data transformation (see p. 60-62 Wheater & Cook)

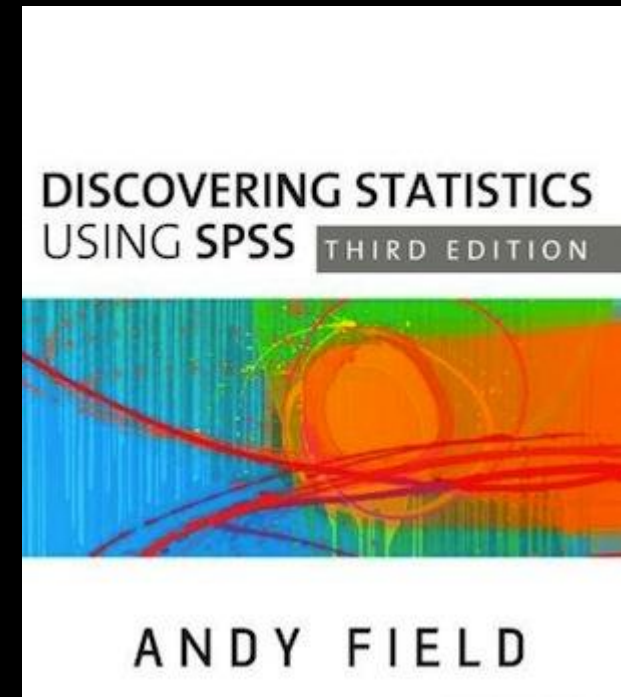
NON PARAMETRIC AND PARAMETRIC TESTS

30

	Parametric	Non-parametric
Assumed distribution	<u>Normal</u>	Any
Assumed variance	<u>Homogeneous</u>	Any
Typical data	<u>Ratio</u> or <u>Interval</u>	<u>Ordinal</u> or <u>Nominal</u>
Data set relationships	Independent	Any
Usual central measure	<u>Mean</u>	<u>Median</u>
Benefits	Can draw more conclusions	Simplicity; Less affected by outliers
Tests		
Choosing	<u>Choosing parametric test</u>	<u>Choosing a non-parametric test</u>
Correlation test	<u>Pearson</u>	<u>Spearman</u>
Independent measures, 2 groups	<u>Independent-measures t-test</u>	<u>Mann-Whitney test</u>
Independent measures, >2 groups	One-way, independent-measures <u>ANOVA</u>	Kruskal-Wallis test
Repeated measures, 2 conditions	<u>Matched-pair t-test</u>	Wilcoxon test
Repeated measures, >2 conditions	One-way, repeated measures <u>ANOVA</u>	Friedman's test

TESTS OF DIFFERENCE

- When we assess the difference between 2 samples
 - **Unpaired** (unmatched) (Independent)
 - **Paired** (matched)
 - E.g. comparing results of students on 2 different modules.

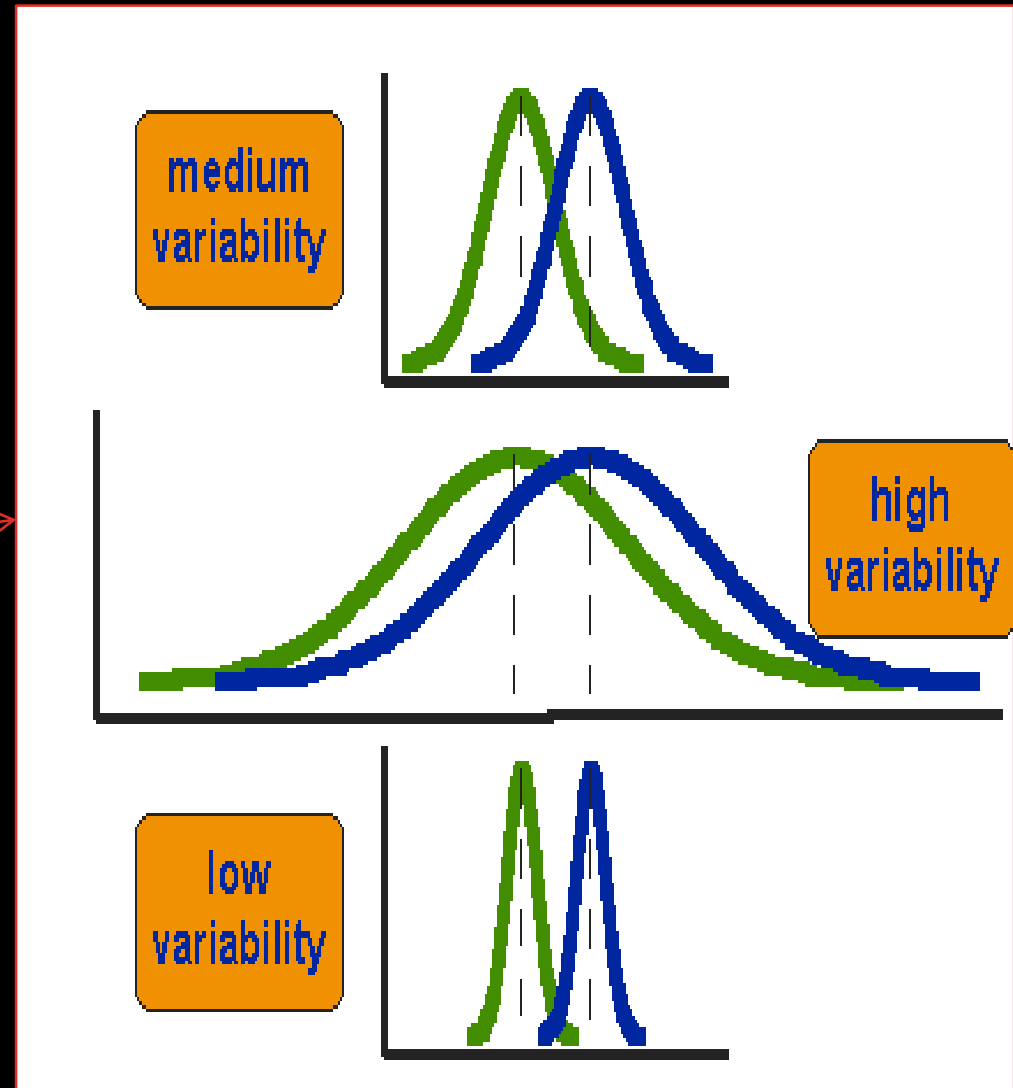


UNMATCHED STATISTICAL TESTS

32

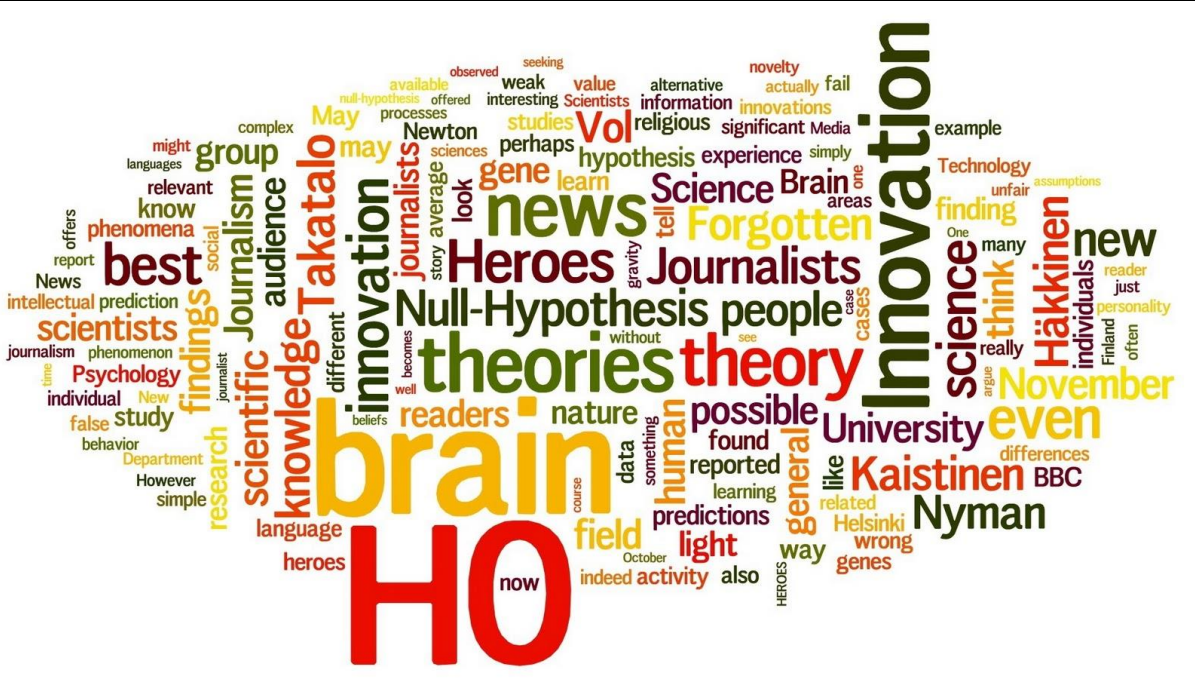
1. Mann Whitney U (non parametric) non normally distributed and Ordinal data
2. T-Test (parametric) normally distributed

- Both establish the difference between 2 samples and if it is **significant**
- The diagram shows that the means are exactly the same but the reasons are very different (variability).
- This leads us to a very important conclusion: when we are looking at the differences between scores for two groups, we have to judge the difference between their means **relative to the spread or variability** of their scores (Trochim, 2013)



33

- # Reject level 0.05



T TEST

- Compares the **absolute** difference of 2 means (degree of difference between the means)
- Purpose: To determine if 2 sets of data are significantly different from one another (O'Mahony, 1986)
- If (assumptions)
 1. **Samples are ND (approx.)**
 2. **Sample sizes are small (generally < 30 if $>$ use Z test)**
 3. **SD of the populations are equal**

Worked example from Wheater & Cook, 2012: 72)

WHY DO A T-TEST?

- Introduction to T Tests 10mins



Hmm. The girls are significantly better than the boys at mathematics!



WORKED EXAMPLE 4.1 Calculating a *t* test statistic for strontium-90 activity in milk from farms near to two types of nuclear plant

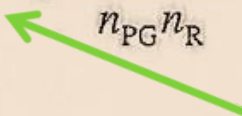
Power Gush plants		Rayzoom plants	
$(n_{PG} = 10)$		$(n_R = 9)$	
x_{PG}	x_{PG}^2	x_R	x_R^2
0.25	0.0625	0.43	0.1849
0.34	0.1156	0.35	0.1225
0.26	0.0676	0.42	0.1764
0.20	0.0400	0.39	0.1521
0.21	0.0441	0.45	0.2025
0.15	0.0225	0.49	0.2401
0.27	0.0729	0.34	0.1156
0.26	0.0627	0.39	0.1521
0.31	0.0961	0.44	0.1936
0.38	0.1444		
$\Sigma x_{PG} = 2.63$		$\Sigma x_R = 3.70$	
$\Sigma x_{PG}^2 = 0.7333$		$\Sigma x_R^2 = 1.5398$	
$\bar{x}_{PG} = 0.263$		$\bar{x}_R = 0.411\ 111\ 11$	
$s_{PG}^2 = 0.004\ 623\ 333$		$s_R^2 = 0.002\ 336\ 111\ 1$	

Variance from SD Calculation

First calculate the mean ($\bar{x}$) and variance (s^2) of the two samples (these are given below the columns above: see Boxes 2.1 and 2.3 for how to calculate these). Use the mean, variance and the sample sizes in the formula below to calculate t :

$$t = \frac{|\bar{x}_{PG} - \bar{x}_R|}{\sqrt{\frac{(n_{PG} - 1)s_{PG}^2 + (n_R - 1)s_R^2}{(n_{PG} + n_R - 2)} \times \frac{n_{PG} + n_R}{n_{PG}n_R}}}$$

Means 

Variance 

$$= \frac{|0.236 - 0.41111111|}{\sqrt{\frac{(10 - 1) \times 0.0046233 + (9 - 1) \times 0.0023361}{(10 + 9 - 2)} \times \frac{10 + 9}{10 \times 9}}}$$

Remember
to square
root

$$= \frac{0.148111}{\sqrt{0.003546811765 \times 0.211111111}}$$

$$= \frac{0.148111}{0.027364} = 5.413$$

WORKED EXAMPLE INTERPRETATION³⁸

- $T = 5.413$
- $df = 17$ ($n_1 + n_2 - 2$) ($10 + 9 - 2 = 17$)
- Tabulated value 0.05 = 2.110, 0.01 = 2.898
- If T is $>$ tab. Value we reject the Null (H_0) i.e. There is a significant difference . . .
- *Milk from near the Rayzoom plants has significantly higher strontium 90 activity than does milk from farms near to Power Gush plants* ($T = 5.413$, $df = 17$, $P < 0.01$)
- NB The t-value will be positive if the first mean is larger than the second and negative if it is smaller (Trochim, 2013)
- The data in the handout compares water samples above and below a sewerage treatment plant. Determine if there is a significant difference between the samples.



Data included in your booklet

EXERCISE 1: SEWAGE TREATMENT WORKS POLLUTION MONITORING

39

Test Sample	χ^2 Above χ^2 Below	
	Above ↓	Below ↓
1	5.6 31.36	5.1 26.01
2	5.8 33.64	5.2 27.04
3	6.2 38.44	4.5 20.25
4	5.4 29.16	4.6 21.16
5	4.6 21.16	5.3 28.09
6	5.0 25	

$$n = 6$$

$$n = 5$$

Determine using the T Test within 95% and 99% confidence limits, whether the outfall has had a significant impact on the dissolved oxygen content of the river.

STAGE 1

40

Test Sample	Above	Below
1	5.6 31.36	5.1 26.01
2	5.8 33.64	5.2 27.04
3	6.2 38.44	4.5 20.25
4	5.4 29.16	4.6 21.16
5	4.6 21.16	5.3 28.09
6	5.0 25	

$$n = 6$$

$$n = 5$$

Determine using the T Test within 95% and 99% confidence limits, whether the outfall has had a significant impact on the dissolved oxygen content of the river.

Sum of
above
treatment
works

$$\sum x_{ab} = 32.6$$

$$\sum x_b = 24.7$$

Chris House

$$\sum x_{ab}^2 = 178.76$$

$$\sum x_b^2 = 122.55$$

$$\bar{x}_{ab} = 5.433$$

$$\bar{x}_b = 4.94$$

Means

STAGE 2

SD ab

Standard Deviations

SD below

$$\sqrt{\frac{178.76 - \frac{(32.6)^2}{6}}{5}}$$

$$\sqrt{\frac{122.55 - \frac{(24.7)^2}{5}}{4}}$$

$$178.76 - 177.126 =$$

$$122.55 - 122.018$$

$$\frac{1.634}{5} \quad \text{Variance} \quad \sqrt{0.3268}$$

$$\frac{0.532}{4} \quad \sqrt{0.133}$$

$$SD \ 0.57166$$

$$SD \ 0.3647$$

USE  Use Variance which is the figure before you square root in SD

STAGE 3 & INTERPRETATION

$$T = \frac{5.433 - 4.94}{\sqrt{\frac{6-1 \times 0.326 + 5-1 \times 0.133}{6+5-2} \times \frac{6+5}{6 \times 5}}}$$

Means

$$\sqrt{\frac{0.493}{1.63 + 0.532 \times 0.36}} = 0.294$$

Variance

$$\frac{0.493}{\sqrt{0.0864}} = 1.67$$

$$Df \ 6 + 5 - 2 = 9 \quad Tab \ 0.05 = 2.262$$

$$0.01 = 3.250$$

Tabulated

- Therefore < tabulated value
- Accept Null- there is no difference
- Need more testing because even though the trend is lower it is not significant.
- Hence, 'eyeballing' not scientifically acceptable

- If T is $>$ tab. Value we reject the Null (H_0) i.e. There is a significant difference .

Degrees of freedom	Significance level					
	20% (0.20)	10% (0.10)	5% (0.05)	2% (0.02)	1% (0.01)	0.1% (0.001)
1	3.078	6.314	12.706	31.821	63.657	636.619
2	1.886	2.920	4.303	6.965	9.925	31.598
3	1.638	2.353	3.182	4.541	5.841	12.941
4	1.533	2.132	2.776	3.747	4.604	8.610
5	1.476	2.015	2.571	3.365	4.032	6.859
6	1.440	1.943	2.447	3.143	3.707	5.959
7	1.415	1.895	2.365	2.998	3.499	5.405
8	1.397	1.860	2.306	2.896	3.355	5.041
9	1.383	1.833	2.262	2.821	3.250	4.781
10	1.372	1.812	2.228	2.764	3.169	4.587
11	1.363	1.796	2.201	2.718	3.106	4.437
12	1.356	1.782	2.179	2.681	3.055	4.318
13	1.350	1.771	2.160	2.650	3.012	4.221
14	1.345	1.761	2.145	2.624	2.977	4.140
15	1.341	1.753	2.131	2.602	2.947	4.073
16	1.337	1.746	2.120	2.583	2.921	4.015
17	1.333	1.740	2.110	2.567	2.898	3.965
18	1.330	1.734	2.101	2.552	2.878	3.922
19	1.328	1.729	2.093	2.539	2.861	3.883
20	1.325	1.725	2.086	2.528	2.845	3.850
21	1.323	1.721	2.080	2.518	2.831	3.819
22	1.321	1.717	2.074	2.508	2.819	3.792
23	1.319	1.714	2.069	2.500	2.807	3.767
24	1.318	1.711	2.064	2.492	2.797	3.745
25	1.316	1.708	2.060	2.485	2.787	3.725

EXERCISE 1

44

- **Research Project objective**

Have construction site labour costs changed between pre and post recession? The case of Glasgow

OR

Did ecologists labour costs change between pre and post recession. The case of CCW/NRW?

1. *Perform a T Test on the following data set.*
2. *What is the Null and alternative hypothesis?*
3. *Determine using the T Test within 95% and 99% confidence limits, whether the labour costs are significantly different pre and post recession.*
4. *Interpret the results*

Table 1 Labour costs pre and post recession in millions £

Site pre recession	Site 2 recession
$N_1 = 6$	$N_2 = 8$
Mean = 74.8	Mean = 72.99
SD = 1.04	SD = 1.48
$SD^2 = 1.08$	$SD^2 = 2.20$



$$\begin{aligned}
 t &= \frac{(74.8 - 72.99)}{\sqrt{\left[\frac{(6-1)1.08 + (8-1)2.20}{(6+8-2)} \right] \times \left(\frac{6+8}{6 \times 8} \right)}} \\
 &= \frac{1.81}{\sqrt{\left[\frac{5.4 + 15.4}{12} \right] \times 0.2917}} \\
 &= \frac{1.81}{\sqrt{1.733 \times 0.2917}} = \frac{1.81}{0.711} \\
 &= 2.55
 \end{aligned}$$

- The calculated figure exceeds the tabulated value (2.179 at $p=0.05$) ($14-2=12df$)
- Therefore we reject the null hypothesis and conclude there is a statistically significant difference between the means.
- Hence the recession did have an impact on . . .
- However, what does this say about the results? ($p=0.01=3.055$).

(Chris House from Fowler et al 1998)

EXERCISE 2

T TEST

SAND DUNE

46

- Does sand grain type influence sand dune angle?
- Go through the exercise in the hand-out



- Ottawa
 - Mean = 32.6
 - S2 (variance) = 0.102
- Beach
 - Mean 35.1
 - S2 (variance) = 0.303
- T Test 12.5
- Ttest 10 mins video how to use Excel

Sand dune T-Test

$$T = \frac{35.1 - 32.6}{\sqrt{\frac{(10-1)0.303 + (10-1)0.102}{(10+10-2)}} \times \frac{10+10}{10 \times 10}}$$

Variance

$$= \frac{2.5}{\sqrt{\frac{3.645}{18} \times \frac{20}{100}}}$$

$$\frac{2.5}{\sqrt{0.0405}}$$

$$\frac{2.5}{0.2013} = \underline{\underline{12.419}}$$

$$df = 18 \quad (n^1 + n^2 - 2) \quad 10 + 10 - 2 = 18$$

Tabulated $0.05 = 2.101$ $\therefore$ Reject null
 $0.01 = 2.878$ Is a significant
 $T = 12.419, df 18, P < 0.01$ difference angle/grain

NONPARAMETRIC DIFFERENCES BETWEEN SAMPLES

- Mann-Whitney U test nonpara. Equivalent of the T Test
- Can be applied to ND but is less powerful than T Test
- Median rather than Mean measure of Central T
- Ranked (ordinal) (loses info.) Unmatched
- Stage 1 rank data (ranked value) of BOTH samples in a single list
- Calculate the U value (refer to handout p.75)
- Which ever U value is LOWER is compared to the tabulated value
- Sample sizes do not need to be equal
- Go through worked example
- Smaller the U score the smaller the probability (i.e. opposite to T Test)

MANN- WHITNEY U STAGE 1

Box 4.3 Formula for the Mann-Whitney U statistic

Two values of U are calculated as follows:

$$U_1 = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - \sum R_1$$

and

$$U_2 = n_1 n_2 + \frac{n_2(n_2 + 1)}{2} - \sum R_2$$

where:

n_1 and n_2 are the numbers of data points in samples 1 and 2 respectively;

$\sum R_1$ and $\sum R_2$ are the sums of the ranks in samples 1 and 2.

If the arithmetic has been correctly performed then

$$U_1 + U_2 = n_1 n_2 \text{ (check this is so)}$$

The smaller of the two U values is chosen for comparison with the appropriate statistical table (Table D.3 in Appendix D).

TIED RANKS

$$7+8+9+10+11+12/6=9.5$$

Box 4.4 Calculating tied ranks

If we have a data set where several points have the same value, then we must allow for this in the ranking procedure by calculating the tied rank. In the 20 data points below, several have the same value. The data are first ordered to give values from 1 to 20. Data with the same value are given their mean rank order:

Score	1	2	2	2	3	3	4	4	4	4	4	4	5	5	5	6	6	6	6	6
Order	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Ranks	1	3	3	3	5.5	5.5	9.5	9.5	9.5	9.5	9.5	9.5	14	14	14	18	18	18	18	18

You can see that a score of 1 is the lowest value so this is given a rank of 1. There are three scores of 2 so these all receive the rank of 3, calculated as the next three ranks (2 + 3 + 4) divided by the number of items (3). Next we have two values of 3 which are given ranks of 5.5 (5 + 6 divided by 2). So in general, tied ranks are given the mean rank of the group of ranks which would have applied had there been no tied values. You can check your accuracy in ranking the data by summing the ranks ($\sum R$) and checking that:

$$\sum R = \frac{n(n+1)}{2}$$

where n is the number of items being ranked.

Note that when using a computer, the program automatically ranks the data before carrying out the test.

WORKED EXAMPLE STAGE 2

Tree condition based on a qualitative measure of leaf 'health' for a EIA on a proposed construction site

Whole data set.
Polluted and
cleaned in rank
order



WORKED EXAMPLE 4.2 Calculating a Mann-Whitney U test statistic for tree condition (using ranked colour scores) in clean and polluted sites

Site type	Condition score	Ascending order	Calculating the rank value	Rank value ^a
Polluted	1	1	1	1
Polluted	2	2		3
Polluted	2	3	$\frac{2+3+4}{3} = 3$	3
Polluted	2	4		3
Clean	3	5	$\frac{5+6}{2} = 5.5$	5.5
Polluted	3	6		5.5
Polluted	4	7		9.5
Clean	4	8		9.5
Polluted	4	9		9.5
Clean	4	10	$\frac{7+8+9+10+11+12}{6} = 9.5$	9.5
Polluted	4	11		9.5
Clean	4	12		9.5
Polluted	5	13		14
Clean	5	14	$\frac{13+14+15}{3} = 14$	14
Clean	5	15		14
Clean	6	16		18
Clean	6	17		18
Polluted	6	18	$\frac{16+17+18+19+20}{5} = 18$	18
Clean	6	19		18
Clean	6	20		18

^a Ranked values from polluted sites are shaded to discriminate between the two site types.

n_{Clean} and n_{Polluted} both equal 10, and the sums of ranks for the two samples are

$$\sum R_{\text{Clean}} = 134 \text{ and } \sum R_{\text{Polluted}} = 76$$

Next check the ranking accuracy by calculating the sum of all possible ranks in two ways and checking that they agree:

$$\sum R_{\text{Total}} = \sum R_{\text{Clean}} + \sum R_{\text{Polluted}} = 134 + 76 = 210$$

and

$$\sum R_{\text{Total}} = \frac{(n_{\text{Clean}} + n_{\text{Polluted}}) \times (n_{\text{Clean}} + n_{\text{Polluted}} + 1)}{2} = \frac{(10 + 10) \times (10 + 10 + 1)}{2} = 210$$

Since these both agree, we are sure that we have ranked our data correctly. We can now calculate both values of U :

$$U_{\text{Clean}} = n_{\text{Clean}} n_{\text{Polluted}} + \frac{n_{\text{Clean}} (n_{\text{Clean}} + 1)}{2} - \sum R_{\text{Clean}} = (10 \times 10) + \left(\frac{10 (10 + 1)}{2} \right) - 134 = 21$$

$$U_{\text{Polluted}} = n_{\text{Clean}} n_{\text{Polluted}} + \frac{n_{\text{Polluted}} (n_{\text{Polluted}} + 1)}{2} - \sum R_{\text{Polluted}} = (10 \times 10) + \left(\frac{10 (10 + 1)}{2} \right) - 76 = 79$$

The smaller of these ($U = 21$) is then compared with the critical value given in appropriate tables (see Table 4.3).

Also check that $U_{\text{Clean}} + U_{\text{Polluted}} = n_{\text{Clean}} n_{\text{Polluted}}$ ($21 + 79 = 10 \times 10$) to confirm that the U value calculations are accurate.

TABULATED SCORE

Table 4.3 Selected values of U for the Mann-Whitney U test. For significance, calculated U values must be less than or equal to the table value. Shaded lines indicate the critical values for the example referred to in the text. A more comprehensive table of U values is given in Table D.3 (Appendix D). The upper table values (in bold) are for $P = 0.05$, while the lower values are for $P = 0.01$.

		Larger <i>n</i> value					
		7	8	9	10	11	12
Smaller <i>n</i> value	7	8 4	10 6	12 7	14 9	16 10	18 12
	8		13 7	15 9	17 11	19 13	22 15
	9			17 11	20 13	23 16	26 18
	10				23 16	26 18	29 21
	11					30 21	33 24
	12						37 27

- Calculated value (21) is between the tab. Value (23 = 0.05, 16=0.01)
- There is a significant difference ($P < 0.05$) in the condition of trees
- i.e. Reject null

HOW TO PRESENT THE RESULTS

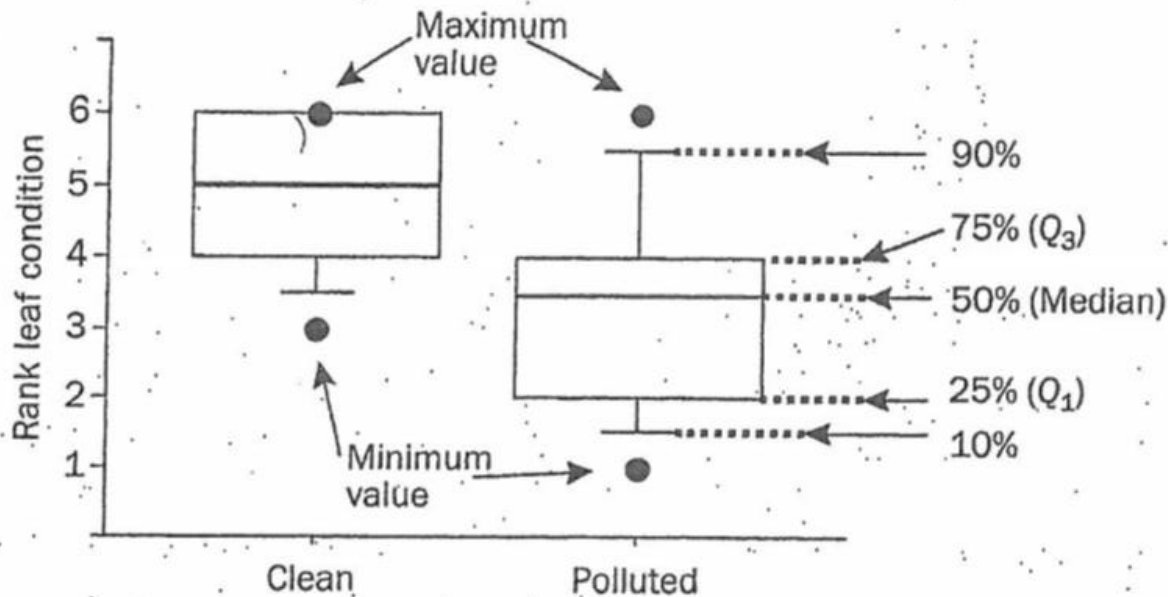


Figure 2.11 Rank leaf condition of trees found on polluted and clean sites ($n = 10$ for each). Higher scores indicate leaves in better condition.

- Display the medians and interquartile ranges e.g. A box and whisker plot

EXERCISE 1⁵⁵

- Twelve environmentally friendly building blocks were cast from the same mix and six were cured properly whilst the remaining 6 were left exposed to the elements. Results from the 28 day cube crushing test produced the following results:
- Using the Mann-Whitney U test, determine whether curing makes any significant difference to the results and hence the sustainability of the block.
- What conclusions can you draw regarding the effects of curing environmentally friendly building blocks?
- A 9 min video from medical science if you are still unclear

DATA SET⁵⁶

Test Sample	Cured N/mm2	Exposed N/mm2
1	26	23
2	30	21
3	28	26
4	27	27
5	29	26
6	27	24

Results

Chris House

REVISION



- A House of Commons Select Committee wanted to assess the public sector expenditure on cost of renewable energy projects. Hence a MWU and T Test were used on the data set of Public and Private sector construction projects in Wales between 2007 and 2012. Average per capita labour costs were compared to assess if there was any statistical difference.
- 1. Perform a MWU test on the following data set (include Null etc.) and comment on the significance of the results for renewable energy projects in the future. $P 0.05$ & $P 0.01$.
- 2. Perform a T Test on the data set and comment on the significance of the results for sector construction projects in the future. $P 0.05$ & $P 0.01$

EXERCISE 1

MWU

COST OF RENEWABLE ENERGY PROJECTS IN THE PRIVATE AND PUBLIC SECTOR

Sector	£'000s	Ascending order	Ascending rank
Private	14		58
Public	18		
Private	18		
Private	18		
Public	19		
Private	19		
Private	19		
Public	21		
Private	22		
Public	26		
Private	26		
Public	29		
Private	29		
Private	30		
Public	31		
Public	32		
Public	34		
Public	36		

RESULTS

Sector	£'000s	Ascending order	Ascending rank
Private		14	1
Public		18	3
Private		18	3
Private		18	3
Public		19	6
Private		19	6
Private		19	6
Public		21	8
Private		22	9
Public		26	10.5
Private		26	10.5
Public		29	12.5
Private		29	12.5
Private		30	14
Public		31	15
Public		32	16
Public		34	17
Public Chis House		36	18

WORKINGS

$$\Sigma \text{ pub.} = 106$$

$$\Sigma \text{ priv.} = 65$$

$$\Sigma \text{ Total} \quad \Sigma \text{ priv} + \Sigma \text{ pub} = 65 + 106 = \underline{171}$$

$$\Sigma \text{ Total} \quad \frac{(9+9) \times (9+9+1)}{2}$$

$$\frac{18 \times 19}{2} = \underline{171}$$

$$U_{\text{pub}} \quad \frac{9 \times 9 + 9 \times (9+1)}{2} \quad - 106$$

$$81 + 45 - 106 = 20$$

$$U_{\text{priv}} \quad 81 + 45 - 65 = 61$$

$$U \text{ value } 20 > \text{tab } \frac{17}{11} \begin{matrix} 0.05 \\ 0.01 \end{matrix}$$

Check

$$\begin{aligned} U_{\text{pub}} + U_{\text{priv}} &= n \times n \\ 20 + 61 &= 9 \times 9 \\ \therefore \text{Correct } 81 \end{aligned}$$

$\therefore$ Accept Null No difference

EXERCISE 2 T TEST DATA SET

61

Public £'000s	Pub ²	Private £'000s	Priv ²
18		30	
21		29	
19		22	
26		19	
32		18	
31		26	
29		18	
36		14	
34		19	

RESULTS

Public £'000s	Pub2	Private £'000s	Priv2
18	324	30	900
21	441	29	841
19	361	22	484
26	676	19	361
32	1024	18	324
31	961	26	676
29	841	18	324
36	1296	14	196
34	1156	19	361

STAGE 2

63

$$\Sigma_{\text{public}} \\ 246$$

$$\Sigma_{\text{public}}^2 \\ 7080$$

$$\Sigma_{\text{priv.}} \\ 195$$

$$\Sigma_{\text{priv}}^2 \\ 4467$$

$$\bar{x}_{\text{pub}} = 27.3 \quad \checkmark$$

$$\bar{x}_{\text{priv}} = 21.7 \quad \checkmark$$

$$n_{\text{pub}} = 9$$

$$n_{\text{priv}} = 9$$

$$\sqrt{\frac{7080 - 246^2}{9}}$$

$$\sqrt{\frac{4467 - 195^2}{9}}$$

$$\sqrt{\frac{7080 - 6724}{8}}$$

$$\sqrt{\frac{4467 - 4225}{8}}$$

$$\sqrt{44.5}$$

$$6.67$$

*

Variance

$$\sqrt{30.25}$$

$$5.5$$

*

STAGE 3/4



T Test Public Expenditure

$$\frac{27.3 - 21.7}{}$$

$$T = \sqrt{\frac{((9-1) \times 44.5) + ((9-1) \times 30.25)}{9+9-2}} \times \frac{9+9}{9 \times 9}$$

$$\frac{5.6}{\frac{356 + 242}{16}} \times \frac{18}{81}$$

$$\frac{5.6}{37.375 \times 0.222}$$

$$\frac{5.6}{37.375 \times 0.222}$$

$$\frac{5.6}{8.297}$$

$$\frac{5.67}{2.88} = \underline{\underline{1.968}}$$

$1.968 < \text{Tab } 2.92 (p=0.01)$
 $\therefore$ Accept null - No difference
 Df $n+n-2 = 16$

WHY STATS?

PERCEPTION VERSUS FACT

- Individual exercise
- Consider different types of transport and risk
- Write down how far you can travel before you get a statistical 1 in a million chance of death
- E.g. Walking is 27 miles
- Write your answers for Motorbike, Car, train and passenger planes
- Answers
 - Motorbike 15 miles
 - Car 175 miles
 - Train 700 miles
 - Plane 7,000 miles
- You need to base decisions on statistical fact not perceptions and this is why we use statistical tests in research

WILCOXON

NONPARAMETRIC MATCHED

- Difference of matched data without a ND
It is the Nonpara. Equivalent to T Test
- Calculate the differences between pairs of data
- Rank differences in order of size
 - Ignore +/- and 0 difference
 - N is the number of scored data (i.e. Not 0)
- Sign of the difference is attached to the rank
- No difference would result in a balanced score
- Compare the lowest to the tab. Value
 - If figure is < tab then reject null

WILCOXON MATCHED PAIRS

There are 5 number
ones (ignore – sign)
therefore:

$$1+2+3+4+5=15/5=3$$

$$6+7+8=21/3=7 \text{ etc}$$

NB **n** ignores
zeros
therefore $n=9$
not 10



WORKED EXAMPLE 4.4 *Calculating a Wilcoxon matched-pairs test statistic for attitudes to environmental taxation before and after watching an educational video*

Respondent	Score		Differences (score after minus score before)	Rank of the differences (ignoring any zeros and the sign)	Signed ranks
	before	after			
A	2	1	-1	3	-3
B	5	5	0	—	—
C	4	5	+1	3	+3
D	2	4	+2	7	+7
E	3	5	+2	7	+7
F	4	5	+1	3	+3
G	3	4	+1	3	+3
H	2	5	+3	9	+9
I	3	4	+1	3	+3
J	2	4	+2	7	+7

From these data we can find the absolute values of the sum of the positive ranks (T_+) and the sum of the negative ranks (T_-):

$$T_+ = 3 + 7 + 7 + 3 + 3 + 9 + 3 + 7 = 42$$

$$T_- = 3$$

The lowest value (irrespective of sign) is 3 so it is this which is compared with the values in Table 4.4.

It is possible to check that T_+ and T_- have been correctly calculated since:

$$T_- = \frac{n(n+1)}{2} - T_+ = \frac{9 \times (9+1)}{2} - 42 = 45 - 42 = 3$$

and:

$$T_+ = \frac{n(n+1)}{2} - T_- = \frac{9 \times (9+1)}{2} - 3 = 45 - 3 = 42$$

Because differences of zero are ignored in this test, ranking begins at the lowest difference greater than 0. If zeros occur, then n is the number of items with differences of more than zero.

TABULATED VALUE

Reject Null at P 0.05

Therefore subjects
become more
committed to
environmental
taxation after
watching the video
($T=3$, $n=9$, $P<0.05$)

Table 4.4 Selected values of T for the Wilcoxon matched-pairs test. Reject the null hypothesis if the calculated T value is less than or equal to the table value. Shaded lines indicate the critical values for the example referred to in the text. A more comprehensive table of T values is given in Table D.4 (Appendix D)

n (number of items
differences) minus any zero

	T	
	P = 0.05	P = 0.01
7	2	—
8	3	0
9	5	2
10	8	3

DATA FOR LOCAL RESPONDENTS

69

1. VIEWS OF LNR BEFORE (PLAN) AND AFTER **OR**
2. VIEWS OF THE OLYMPIC STADIUM BEFORE (PLAN) AND AFTER CONSTRUCTION DESIGNATION

Responden †	Before	After	difference	Rank	Signed Rank
A	4	6			
B	6	7			
C	5	6			
D	4	7			
E	5	8			
F	6	5			
G	7	6			
H	5	7			
I	6	8			
J	9	9			

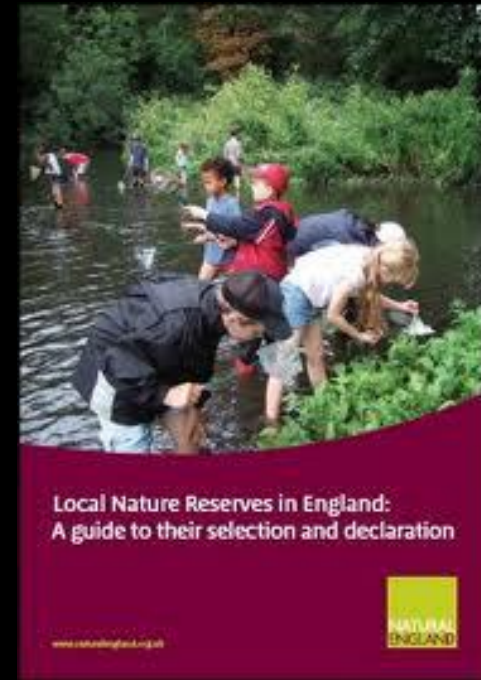
RESULTS ⁷⁰

Responden †	Before	After	difference	Rank	Signed Rank
A	4	6	2	6	6
B	6	7	1	2.5	2.5
C	5	6	1	2.5	2.5
D	4	7	3	8.5	8.5
E	5	8	3	8.5	8.5
F	6	5	-1	2.5	-2.5
G	7	6	-1	2.5	-2.5
H	5	7	2	6	6
I	6	8	2	6	6
J	9	9	0	—	—

RESULTS

71

- $T_+ = 6 + 2.5 + 2.5 + 8.5 + 8.5 + 6 + 6 = 40$
- $T_- = 2.5 + 2.5 = 5$
- Check
 - $9 * (9 + 1) / 2 - 40 = 5$
 - $9 * (9 + 1) / 2 - 5 = 40$
- Tab 0.05=5 therefore the test result is equal to the tabulated and we can reject the null hypothesis.
 - There is a statistical difference between residents views pre and post LNR designation/2012.
 - ($T=5$, $n=9$, $P<0.05$)
 - Their views improved. Why?
- Complete exercises on handout

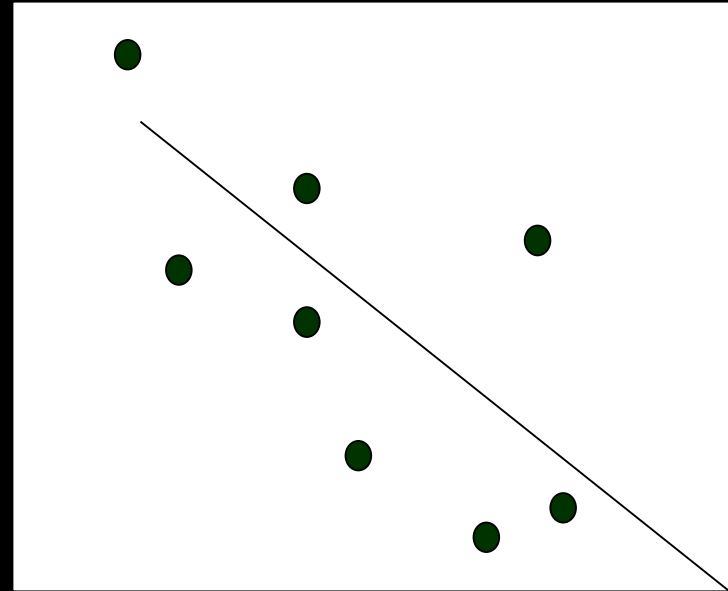
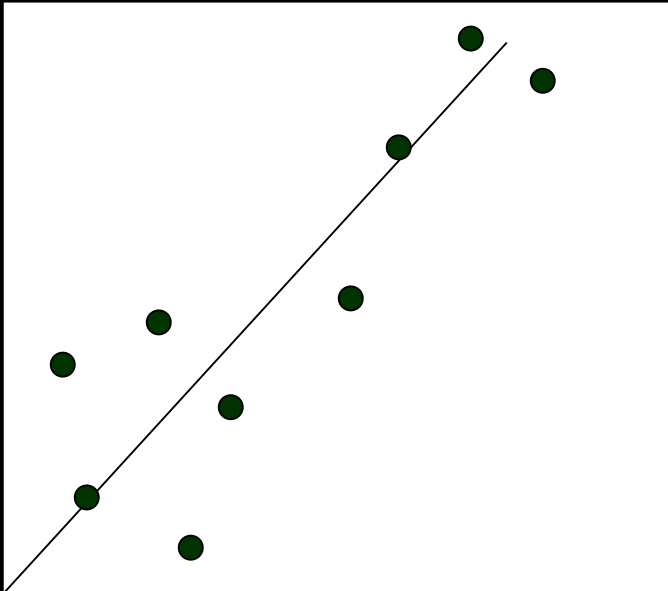


HOW DO WE MEASURE THE RELATIONSHIP BETWEEN 2 VARIABLES?

- With only one variable predicting/modelling your data set can be done using the mean. But if you have 2 variables (Foltz, 2014, <http://www.youtube.com/watch?v=ZkjP5RJLQF4>) This is a good video of the basics 23 mins)
- Quantify the relationship between variables: *“as the magnitude of 1 var. increases does the other increase or decrease?”* (Wheater & Cook, 2000: 89)
- Why is it important within research?
- e.g. Pollution and cost of remediation in the lower Swansea valley

THE RELATIONSHIP

- Pos+ relationship Neg- relationship



- Individually think of an example of each

CORRELATION

- Measures the degree to which 2 variables vary (strength of relationship).
- *Pearson's Product Correlation coefficient (Parametric)*
- *Spearman's Rank Cor. Coe. (Non parametric)*
- *Present in a scatter plot*
- - Y Measured variables
 - (dependent, predicted)
- X controlled variable e.g pH
- (Independent)



CORRELATION

- Perfect Pos+ Relationship

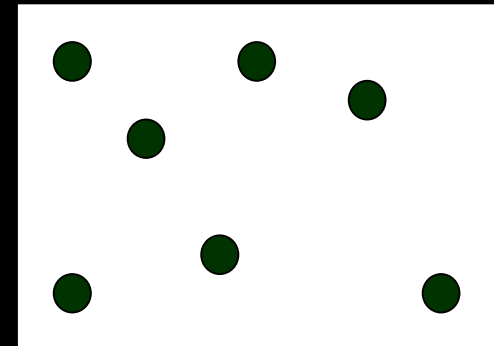
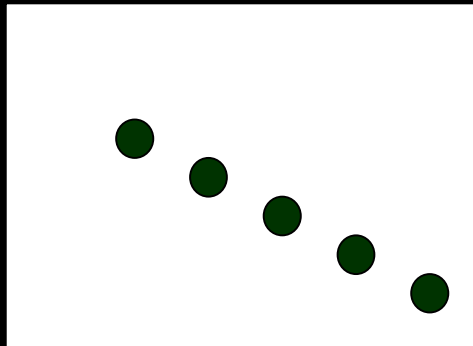
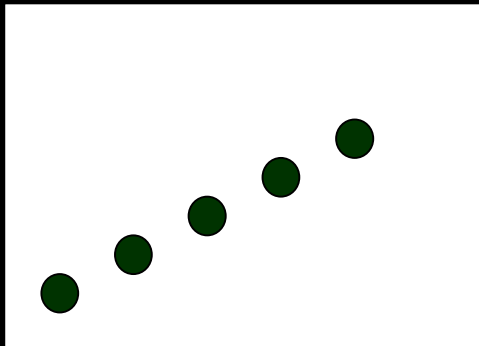
Perfect Neg-

No relationship

Correlaton+1

-1

0



SOME CONSIDERATIONS

76

- Sample size is important because of chance
- Relationship does **NOT** indicate cause and effect: disease and doctors, Rum and Priests
- Plausible, repeatable, proven mechanism, predictable) (Bowers, 2010)



SPEARMAN'S RANK CORRELATION⁷⁷ COEFFICIENT (GO THROUGH HANDOUT)

Location River Ebbw Pollution event	Distance from point source pollution metres	Ammoniacal nitrogen (NH ₃ -N) mg/L (milligram per litre). Ammonium from fertilisers
1	50	1.8
2	175	1.2
3	270	2.0
4	375	1.0
5	425	1.0
6	580	1.2
7	710	0.8
8	790	0.6
9	890	1.0
10	980	0.85

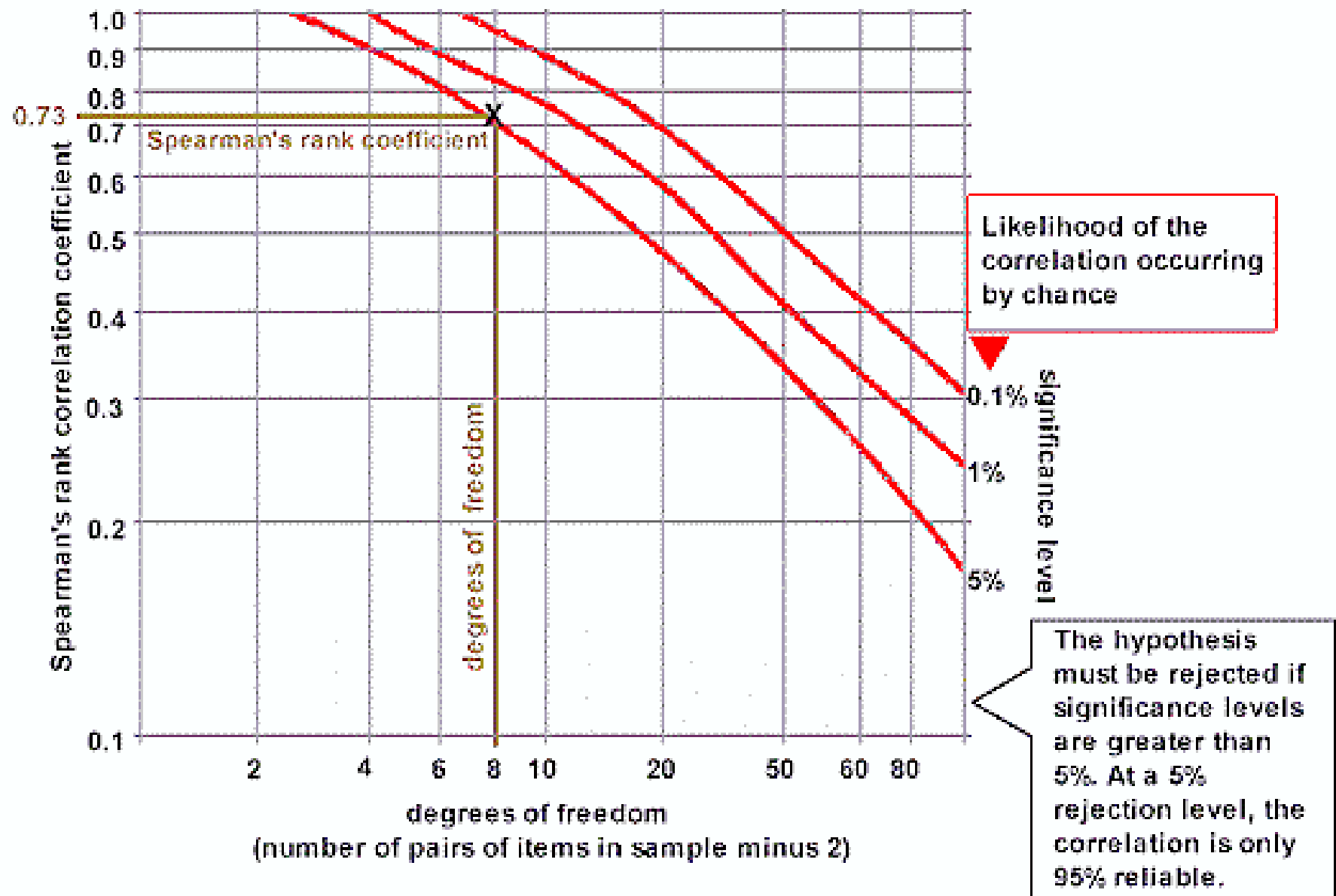
ANSWER

- $\sum d^2 = 285.5$
- $T = -0.73$
- The closer T is to $+1$ or -1 , the stronger the likely correlation. A perfect positive correlation is $+1$ and a perfect negative correlation is -1 . The T value of -0.73 suggests a fairly strong negative relationship.

ANSWERS CONT'D

- In the example, the value 0.73 gives a significance level of slightly less than 5%. That means that the probability of the relationship you have found being a chance event is about 5 in a 100. You are 95% certain that your hypothesis is correct. The reliability of your sample can be stated in terms of how many researchers completing the same study as yours would obtain the same results: 95 out of 100.
- The fact two variables correlate cannot prove anything - only further research can actually prove that one thing affects the other.
- Data reliability is related to the size of the sample. The more data you collect, the more reliable your result.

The significance of the Spearman's rank correlation coefficients and degrees of freedom



- Calculates the line of 'best fit' which can potentially have the power of **prediction**.
- If we expect a cause/effect/linear.
- e.g sound levels at distances from a landfill site.



TWO PROCEDURES

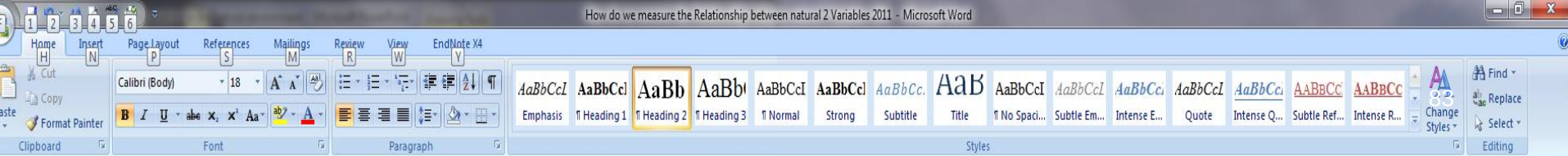
1. Fitting the line

2. Calculating the significance of regression

1. Fitting the line

- Formula for a straight line
- $y = a + bx$
- x and y coordinates
- Best fit passes through the **mean** of x and y so that the residual distance is as small as possible.



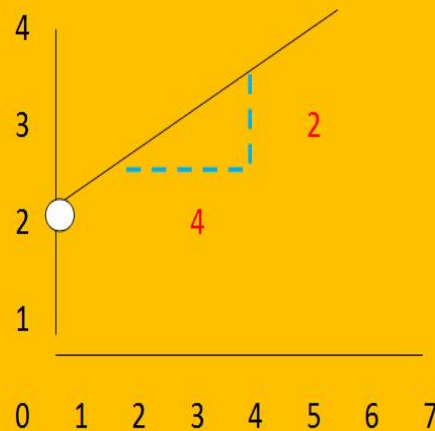


Simple linear regression

a (intercept) 2.0

b (gradient) 0.5

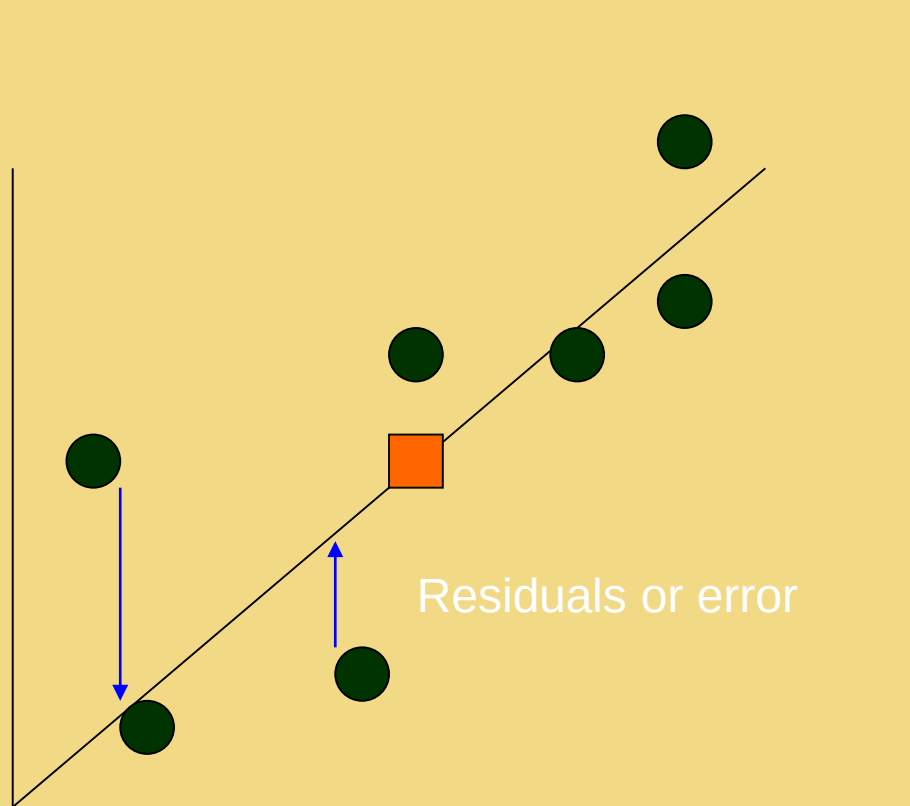
= line of best fit



$$\frac{2}{4} = 0.5$$

BEST FIT

84



Means

Squaring the residuals

Because:

1. Makes them pos+
2. Emphasises larger deviations

BEST FIT LINE CONTINUED

- *Line of lowest total squared residuals (least squares)*
- *Gradient is called the regression coefficient*
- *Should only plot within the known data.*
- Overview of stage 1 (5 mins)
- **EXAMPLE HANDOUT 5.3A (Go through handout)**



WORKED EXAMPLE 5.3a Regression of traffic noise level (y) on distance from road (x): calculation of the equation of the regression line

Distance (m)		Noise (dB_A)		
x	x^2	y	y^2	xy
20	400	90	8 100	1 800
40	1 600	90	8 100	3 600
60	3 600	86	7 396	5 160
80	6 400	81	6 561	6 480
100	10 000	82	6 724	8 200
120	14 400	75	5 625	9 000
140	19 600	74	5 476	10 360
$\Sigma x = 560$		$\Sigma y = 578$	$\Sigma y^2 = 47 982$	$\Sigma xy = 44 600$

$$b = \frac{\Sigma xy - \frac{\Sigma x \Sigma y}{n}}{\Sigma x^2 - \frac{(\Sigma x)^2}{n}} = \frac{44\,600 - \frac{560 \times 578}{7}}{56\,000 - \frac{560^2}{7}} = \frac{-1\,640}{11\,200} = -0.146$$

$$a = \bar{y} - b\bar{x} = \frac{578}{7} - \left(-0.146 \times \frac{560}{7}\right) = 94.29$$

The regression equation is thus

$$y = 94.29 - 0.146x$$

(the negative sign of b tells us that the slope is negative).

To fit the regression line by hand we take one low value of x (20) and one high value (140) and calculate the value of y for each. When $x = 20$,

$$y = 94.29 - 0.146 \times 20 = 91.37$$

When $x = 140$,

$$y = 94.29 - 0.146 \times 140 = 73.85$$

These two points can then be plotted and the line drawn between them.

EXERCISES

87

Environmental exercises

- A. Sustainability and London 2012 (Handout)
- B. Grass yield in East Anglia (Handout)
- C. Soil fertility and number of saplings per m² (predict for 11 & 43)
- D. Length of time it takes flies to hatch and levels of humidity (saturation deficit)

Built environment exercises

- A. Supply of materials and London 2012 (Handout)
- B. Accidents and cube strength on construction sites (Handout)
- C. Additional construction costs due to rainfall on houses on a Bovis estate (predict for 11 & 43 cm)

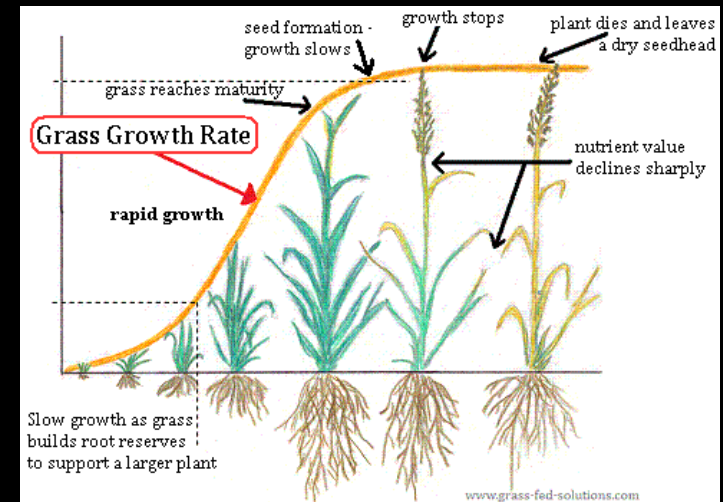
EXERCISE A (OLYMPICS)

- $B = 0.5$
- $A = 2.0$
- Y at
 - 3.5 miles = £3.75,
 - 9.89 miles = £6.94
 - PLOT



EXERCISE B (GRASS YIELD/ACCIDENTS)

- $B = 0.9897$
- $A = -0.252$
- Y at
 - 11 = 10.63,
 - 38 = 37.36,
 - 43 = 42.30
 - PLOT

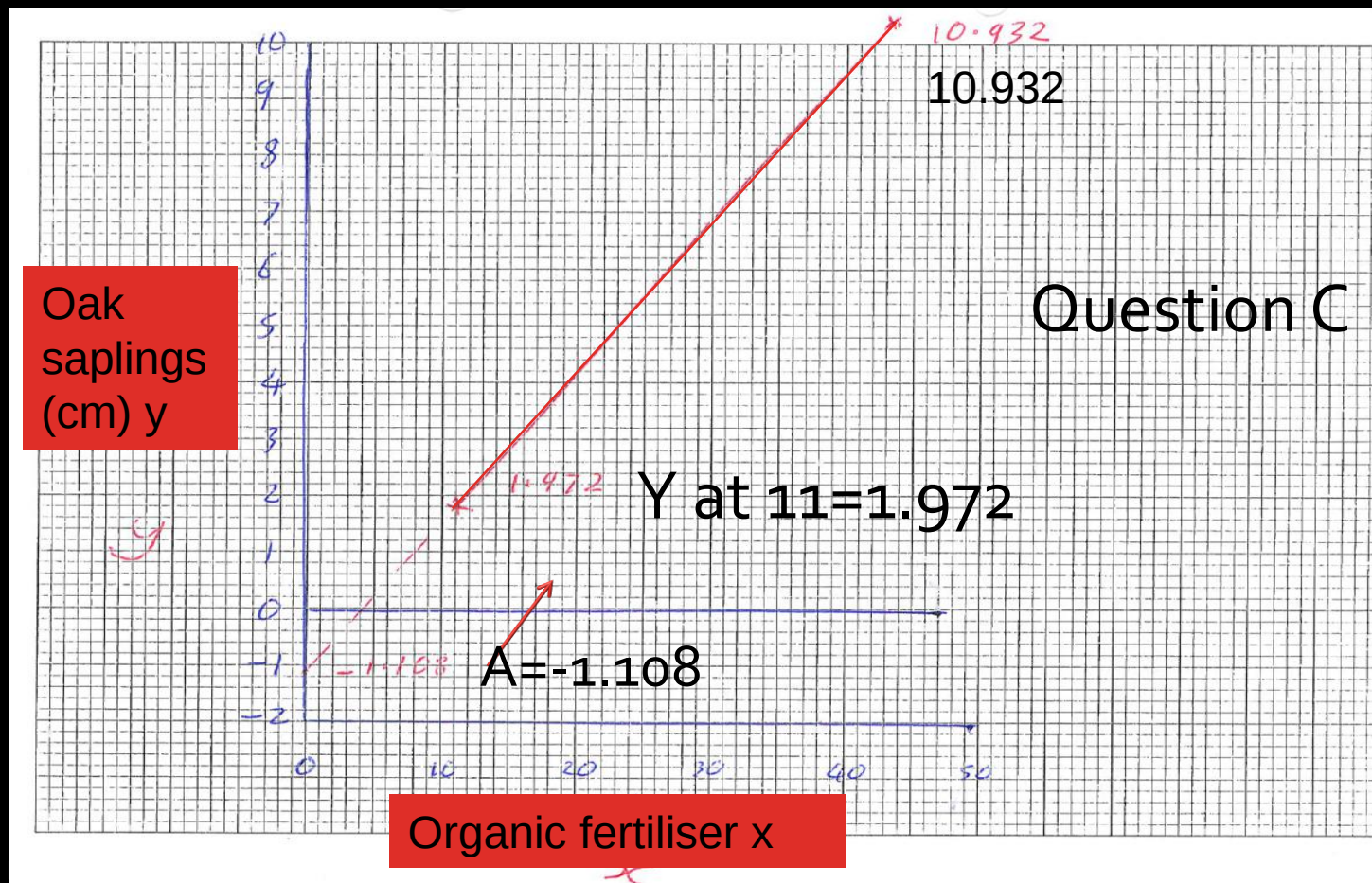


EXERCISE C

FERTILISER AND SAPLING GROWTH/ADDITIONAL CONSTRUCTION COST N=8

Fertiliser x	x ²	sapling s y	y ²	xy
11	121	1	1	11
15	225	2	4	30
16	256	3	9	48
22	484	5	25	110
18	324	6	36	108
28	784	8	64	224
36	1296	9	81	324
43	1849	10	100	430
Σx 189	Σx^2 5339	Σy 44	Σy^2 320	Σxy 1285

- Question C (Oak saplings)
 - $B=0.28$
 - $A=-1.108$
 - Y at $11=1.972$, $43=10.932$ PLOT



EXERCISE D (FLIES HUMIDITY)

- $B = 3.82$
- $A = 16.46$
- Y at
 - $0.3 \text{ kPa} = 6 \text{ hours}$
 - $1.7 \text{ kPa} = 34.4 \text{ hours}$
 - PLOT



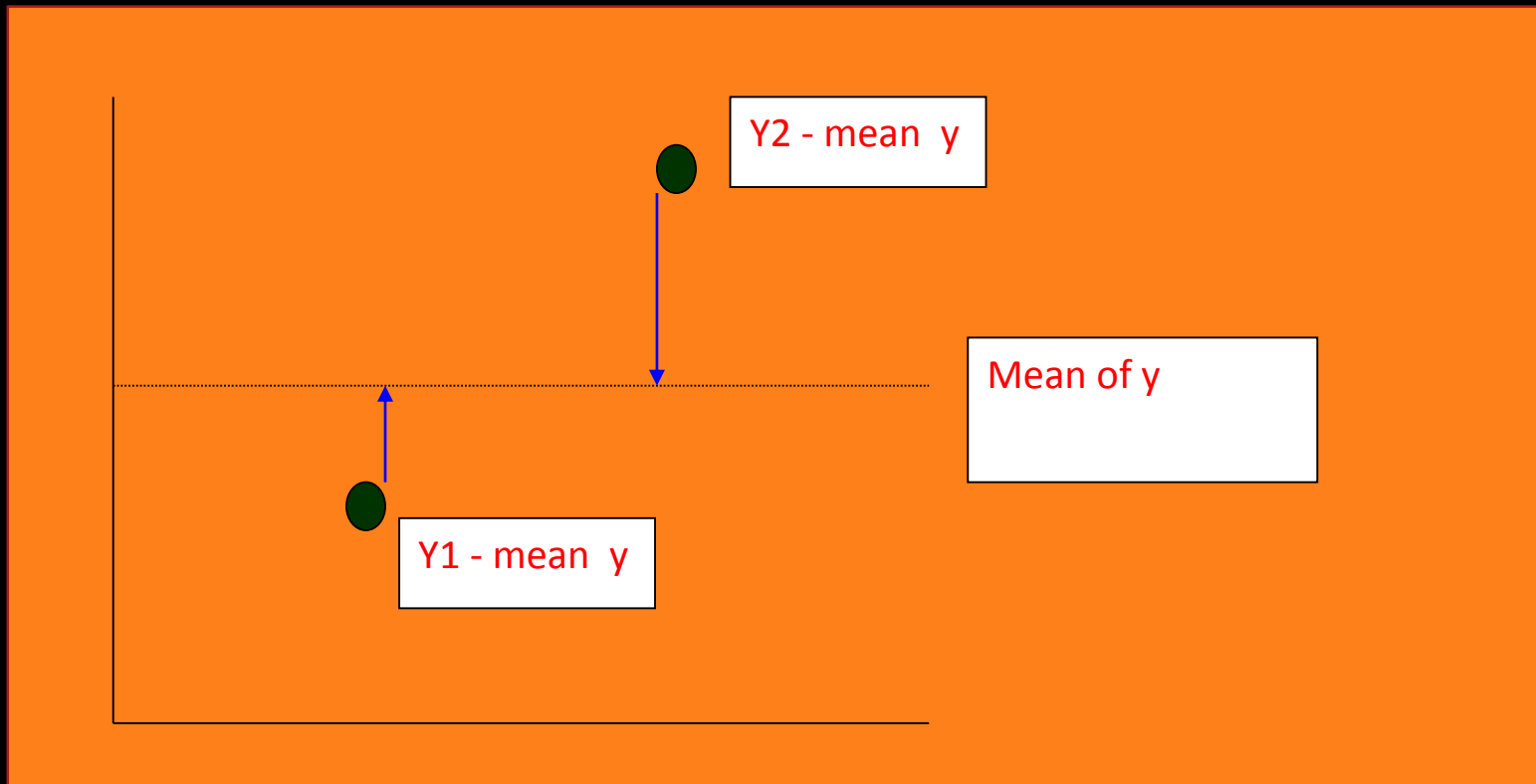
2. CALCULATING THE SIGNIFICANCE OF REGRESSION

- *Is there a relationship between noise and distance from a landfill site? How significant is the relationship?*
- *Called ANOVA Analysis of Variance*
 - *Calculate the total variation of Y Called the total sum of squares SS_{total} and how much can be explained by the variation in x.*
 - *Hence.*

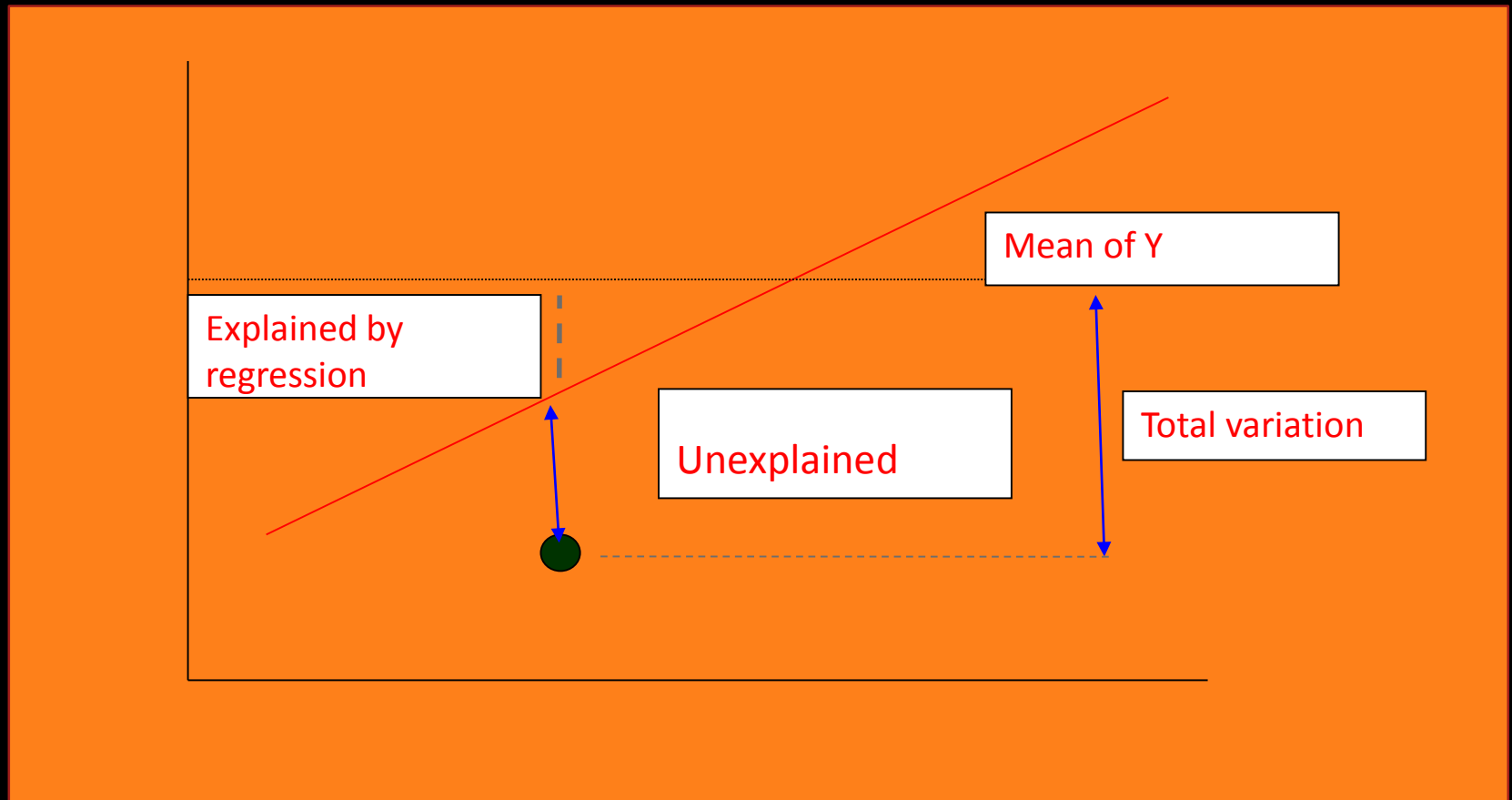
To what extent is the variation in y a product of x ?
 - *Subtract the mean from each figure (see following slides)*
 - *The calculation is part of the SD equation*

MEAN

- Subtract the mean of y from each value

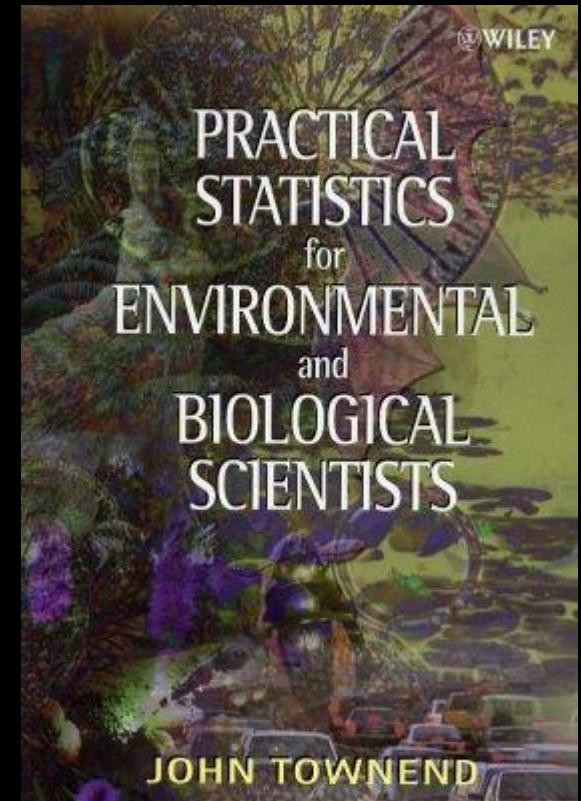


- The variation is a product of
 - 1. Explained from regression answers ($SS_{\text{regression}}$)
 - 2. Unexplained (SS_{residual})



CALCULATING F ⁹⁶

- Therefore $SS_{total} = SS_{regression} + SS_{residual}$
- Read through 109 & 110 for prediction.
- Worked example
- In this case $F_{1,5}$. The first figure is df from regression second is the residual
- Reject null if F is greater than the tabulated values.
- Results should include coefficient of determination r^2 . Amount to which the variation can be explained. (see box 5.7 / 5.3c)





WORKED EXAMPLE 5.3B Regression of traffic noise level on distance from road: testing the significance of the regression line

First we calculate the sums of squares. Note that the components used in the calculations below have already been calculated in Worked Example 5.3a. The total sum of squares is:

$$SS_{\text{total}} = \sum y^2 - \frac{(\sum y)^2}{n} = 47\,982 - \frac{578^2}{7} = 255.714\,29$$

The regression sum of squares is calculated as:

$$SS_{\text{regression}} = \frac{\left(\sum xy - \frac{\sum x \sum y}{n} \right)^2}{\sum x^2 - \frac{(\sum x)^2}{n}} = \frac{\left(44\,600 - \frac{560 \times 578}{7} \right)^2}{56\,000 - \frac{560^2}{7}} = \frac{(-1\,460)^2}{11\,200} = 240.142\,86$$

$$\begin{aligned} -1640^2 &= \\ -2,689,600 \end{aligned}$$

Now subtract the latter from the former to give the residual sum of squares:

$$SS_{\text{residual}} = SS_{\text{total}} - SS_{\text{regression}} = 255.714\,29 - 240.142\,86 = 15.571\,43$$

Substituting these into an ANOVA table, we get:

	df	SS	MS	F	P
Regression	1	240.14	$\frac{SS_{\text{regression}}}{1} = 240.14$	$\frac{MS_{\text{regression}}}{MS_{\text{residual}}} = 77.11$	<0.01
Residual	$n - 2 = 5$	15.57	$\frac{SS_{\text{residual}}}{n - 2} = 3.11$		
Total	$n - 1 = 6$	255.71			

In Worked Example 5.3b, the test statistic (F) is calculated to be 77.11. We then look up the value of F in a table of critical values (see the shaded area of Table 5.5). Critical values of F have two values for the degrees of freedom. In this case, we look up F at $df = 1$ and 5 (sometimes written as $F_{1,5}$). The first figure for the degrees of freedom (1) comes from the regression sum of squares. For simple linear regression, the degrees of freedom for the regression sum of squares are always 1. The second figure for the degrees of freedom (5) comes from the residual sum of squares ($n - 2$). Looking at the shaded section of Table 5.5 we find that the critical value at $F_{1,5}$ is 6.61 (at $P = 0.05$). Since our calculated value of F (77.11) is much higher than either this or the value at $P = 0.01$ (16.3), we can reject the null hypothesis and conclude that a

NB
Book is
wrong

Table D.7 Values of F for use in regression analysis, one-way ANOVA and two-way ANOVA: The upper table values (in bold) are for $P = 0.05$, while the lower values are for $P = 0.01$. Look up the table values using the df for the explained and unexplained variation. For regression analysis (see Table 5.4), the df for the explained variation are the regression df (1) and those for the unexplained variation are the residual df ($n - 2$). For one way ANOVA (see Table 7.2), the df for the explained variation are the between df ($k - 1$), while those for the unexplained variation are the within df ($[n_T - 1] - [k - 1]$). For two-way ANOVA (see Table 7.9), the df for the explained variation are the df associated with the variable ($r - 1$ or $c - 1$) or the interaction between the variables ($[r - 1][c - 1]$), while those for the unexplained variation are the within df ($r \times c[n - 1]$). For two-way ANOVA with single observations in each cell (see bottom of Worked Example 7.7), the df for the explained variation are the df associated with the variable ($r - 1$ or $c - 1$), while those for the unexplained variation are the remainder df ($[r - 1][c - 1]$). The results are significant if the calculated value of F is higher than the table value

		df for the explained variation (e.g. regression df or between df)							df for the explained variation (e.g. regression df or between df)				
		1	2	3	4	5			1	2	3	4	5
df for the unexplained variation (e.g. residual df or within df)	1	161 4052	200 5000	216 5403	225 5625	230 5764	df for the unexplained variation (e.g. residual df or within df)	16	4.49 8.53	3.63 6.23	3.24 5.29	3.01 4.77	2.85 4.44
	2	18.5 98.5	19.0 99.0	19.2 99.2	19.2 99.2	19.3 99.3		17	4.45 8.40	3.59 6.11	3.20 5.18	2.96 4.67	2.81 4.34
	3	10.1 34.1	9.55 30.8	9.28 29.5	9.12 28.7	9.01 28.2		18	4.41 8.29	3.55 6.01	3.16 5.09	2.93 4.58	2.77 4.25
	4	7.71 21.2	6.94 18.0	6.59 16.7	6.39 16.0	6.26 15.5		19	4.38 8.18	3.52 5.93	3.13 5.01	2.90 4.50	2.74 4.17
	5	6.61 16.3	5.79 13.3	5.41 12.1	5.19 11.4	5.05 11.0		20	4.35 8.10	3.49 5.85	3.10 4.94	2.87 4.43	2.71 4.10
	6	5.99 13.7	5.14 10.9	4.76 9.78	4.53 9.15	4.39 8.75		21	4.32 8.02	3.47 5.78	3.07 4.87	2.84 4.37	2.68 4.04
	7	5.59 12.2	4.74 9.55	4.35 8.45	4.12 7.85	3.97 7.46		22	4.30 7.95	3.44 5.72	3.05 4.82	2.82 4.31	2.66 3.99
	8	5.32 11.3	4.46 8.65	4.07 7.59	3.84 7.01	3.69 6.63		23	4.28 7.88	3.42 5.66	3.03 4.76	2.80 4.26	2.64 3.94
	9	5.12 10.6	4.26 8.02	3.86 6.99	3.63 6.42	3.48 6.06		24	4.26 7.82	3.40 5.61	3.01 4.72	2.78 4.22	2.62 3.90
	10	4.96 10.0	4.10 7.56	3.71 6.55	3.48 5.99	3.33 5.64		25	4.24 7.77	3.39 5.57	2.99 4.68	2.76 4.18	2.60 3.85
	11	4.84 9.65	3.98 7.21	3.59 6.22	3.36 5.67	3.20 5.32		30	4.17 7.56	3.32 5.39	2.92 4.51	2.69 4.02	2.53 3.70
	12	4.75 9.33	3.89 6.93	3.49 5.95	3.26 5.41	3.11 5.06		35	4.12 7.42	3.27 5.27	2.87 4.40	2.64 3.91	2.49 3.59
	13	4.67 9.07	3.81 6.70	3.41 5.74	3.18 5.21	3.03 4.86		40	4.08 7.31	3.23 5.18	2.84 4.31	2.61 3.83	2.45 3.51
	14	4.60 8.86	3.74 6.51	3.34 5.56	3.11 5.04	2.96 4.69		45	4.06 7.23	3.20 5.11	2.81 4.25	2.58 3.77	2.42 3.45
	15	4.54 8.68	3.68 6.36	3.29 5.42	3.06 4.89	2.90 4.56		50	4.03 7.17	3.18 5.06	2.79 4.20	2.56 3.72	2.40 3.41

• Question B

- SS Total 873.87
- SS Regression 807.97
- SS Residual 65.9
- MS 10.98
- F 73.58
- R 0.925
- R^2 92.5%
- F1,6 5.99 & 13.7 Our calculated is higher (73.58) therefore reject null. A significant amount of the variation in Y is accounted for by x
- Refer to page 109 to note how to present this result



• Question C

- SS Total 78
- SS Regression 68.96
- SS Residual 9.04
- MS 1.506
- F 45.79
- R 0.88
- R^2 88%
- F1,6 5.99 & 13.7 Our calculated is higher (45.79) therefore we reject null. A significant amount of the variation in Y is accounted for by x. Refer to page 109 to note how should present this result



PREDICTION

- If the result is high and we can explain a lot of the variation then we can predict. If the variation is more complex then multiple regression maybe more appropriate.
- Our gradient and intercept are estimates and therefore we need to place confidence limits to our prediction (NB Predictions of y should only be within limits of x) box 5.8



WORKED EXAMPLE 5.3D Regression of traffic noise level on distance from road: predicting a value of y from the regression line (e.g. when $x = 65$ m)

Using the equation obtained from Worked Example 5.3a, the new value of y is estimated at $x = 65$ m:

$$y = 94.29 - 0.146 \times 65 = 84.8 \text{ dB}_A$$

Next, calculate the sums of squares of x (note that the components are already calculated in Worked Example 5.3a):

$$SS_x = \sum x^2 - \frac{(\sum x)^2}{n} = 56\,000 - \frac{560^2}{7} = 11\,200$$

and then, using the value of SS_{residual} calculated earlier (15.57: Worked Example 5.3b), and inserting the value of x' (65) for which we wish to estimate y into the formula from Box 5.8:

$$\begin{aligned} SE \text{ of prediction} &= \sqrt{\frac{SS_{\text{residual}}}{n-2} \times \left[1 + \frac{1}{n} + \frac{(x' - \bar{x})^2}{SS_x} \right]} \\ &= \sqrt{\frac{15.57}{5} \times \left[1 + \frac{1}{7} + \frac{(65 - 80)^2}{11\,200} \right]} = \sqrt{3.114 \times 1.162\,946\,4} = 1.903\,001\,6 \end{aligned}$$

Since $n < 30$, the 95% confidence limits are calculated by:

$$95\% \text{ confidence limits} = \text{estimate} \pm (t \times SE)$$

The degrees of freedom are $n - 2$, and the value of t at $P = 0.05$ and $df = 5$ is 2.571. Our estimated value of y was 84.8 dB_A. Therefore:

$$95\% \text{ confidence limits} = 84.8 \pm (2.571 \times 1.903\,001\,6) = 84.8 \pm 4.89 \text{ dB}_A$$

So

$$\text{Upper limit} = 89.69 \text{ dB}_A$$

$$\text{Lower limit} = 79.91 \text{ dB}_A$$

This means that if we were to take a single reading at 65 m from the road, we would be 95% confident that the sound level lay between 79.91 dB_A and 89.69 dB_A.

In Figure 5.13a there is an increase in unexplained variation with increase in x value. In Figure 5.13b the curved relationship between the residuals and x suggests that a curved relationship may exist between y and x . In either case we should reconsider our use of regression, and if we do continue, the data should be transformed. It is also important to look out for clusters of residuals (which may indicate an autocorrelation with a third, unmeasured variable) and for single unusually large residuals (single

EXERCISE D: FLIES AND SATURATION DEFICIT

102

$$\sum x$$

9.0

$$\sum x^2$$

11.4

$$\sum y$$

199.9

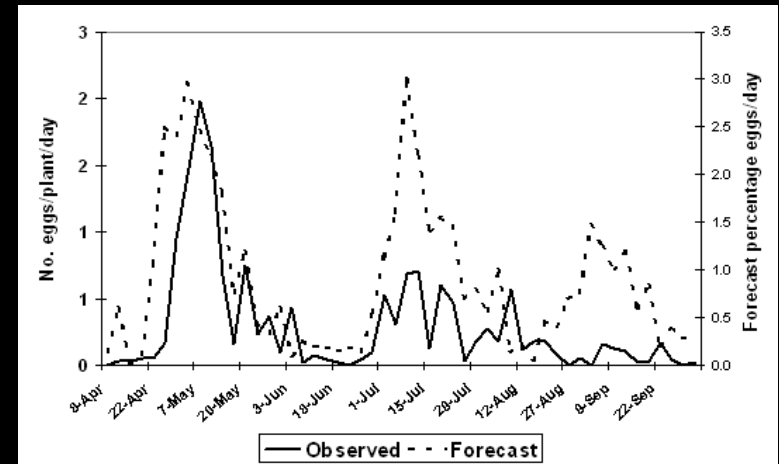
$$\sum y^2$$

4045.07

$$\sum xy$$

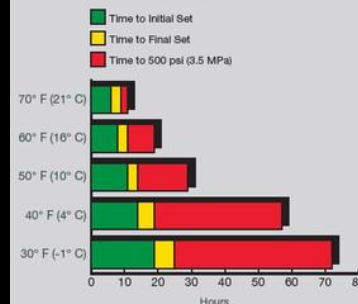
192.52

- Refer to handout
 - B= 3.82
 - A= 16.6
 - Y and PLOT
 - SS Total 49.069
 - SS Regression 48.185
 - SS Residual 0.884
 - F 436.3
 - R 0.982
 - R² 98.2%
 - F1,8 5.32 & 11.26 Our calculated is higher (436.3) therefore reject null. A significant amount of the variation in Y is accounted for by x

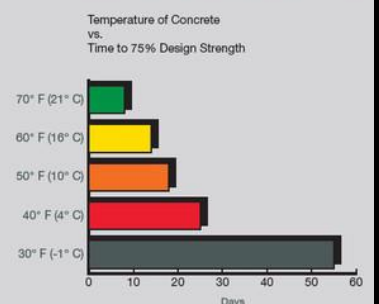


How temperature affects the concrete curing process:

Temperature of concrete vs. time



Long-term strength development



INTERPRETATION

- Martin & Bateson (1993:142-145) (Measuring behaviour)
- A small correlation can be statistically significant. Sample size is important

Value of r	Informal interpretation
<0.2	Slight; almost negligible relationship
0.2-0.4	Low correlation; definite but small
0.4-0.7	Moderate; substantial relationship
0.7-0.9	High correlation; marked relationship
0.9-1.0	Very high correlation; very dependable relationship

FREQUENCY DATA ANALYSIS

The case for Chi Square

Chris House



INTRODUCTION

- We have analysed ordinal data but what about nominal?
- Very common and hence this is a widely applied test
- 2 types of frequency analysis
 - Tests of association
 - Tests of goodness of fit
- Are we testing a measured FD against a 'theoretical distribution'?

CHI SQUARE

- The frequency we **observed** is compared to what we would **expect** on the basis of a null hypothesis
- If the difference between O and E is big then the calculated will exceed the tabulated/critical value and we can reject null
- Video 12mins (Paul Andersen, 2012) **Stop the video after the heads /tails example and do the same test in class.**
 - Chi square
 - Try the heads and tails exercise with a result of 42 heads and 8 tails.
 - P.S. how to get the symbol in word 03c7 Alt x

• χ^2

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

O = the frequencies observed

E = the frequencies expected

$\sum$ = the 'sum of'

ANOTHER DICE EXAMPLE

RESULT ($P \leq .05$)

In APA format:

The observed frequency of each number on the die differed significantly from what would be expected on a theoretically fair die, $\chi^2(5, N = 600) = 12.26, p \leq 0.05$.

Our problem χ^2 was 12.26. Our critical $\chi^2(0.05, 5)$ was 11.07. Therefore our variation was too great to be explained by chance alone. Therefore we must reject H_0 (the die IS FAIR) and accept H_1 , the die is NOT FAIR.

We are 95% confident that you have the loaded die.

EXAMPLE 2

108

- Public access agreements between local authorities and landowners
- Is an access agreement independent of authority?
- 2 variables recorded access agreements present and type of LA (Extract below)

Data GOES TO 287	Access Agreements present or absent	Type LA District (D), Metropolitan (M), London Borough (L)
1	A	M
2	A	M
3 etc	P	D

BASIC SUMMARY OF DATA

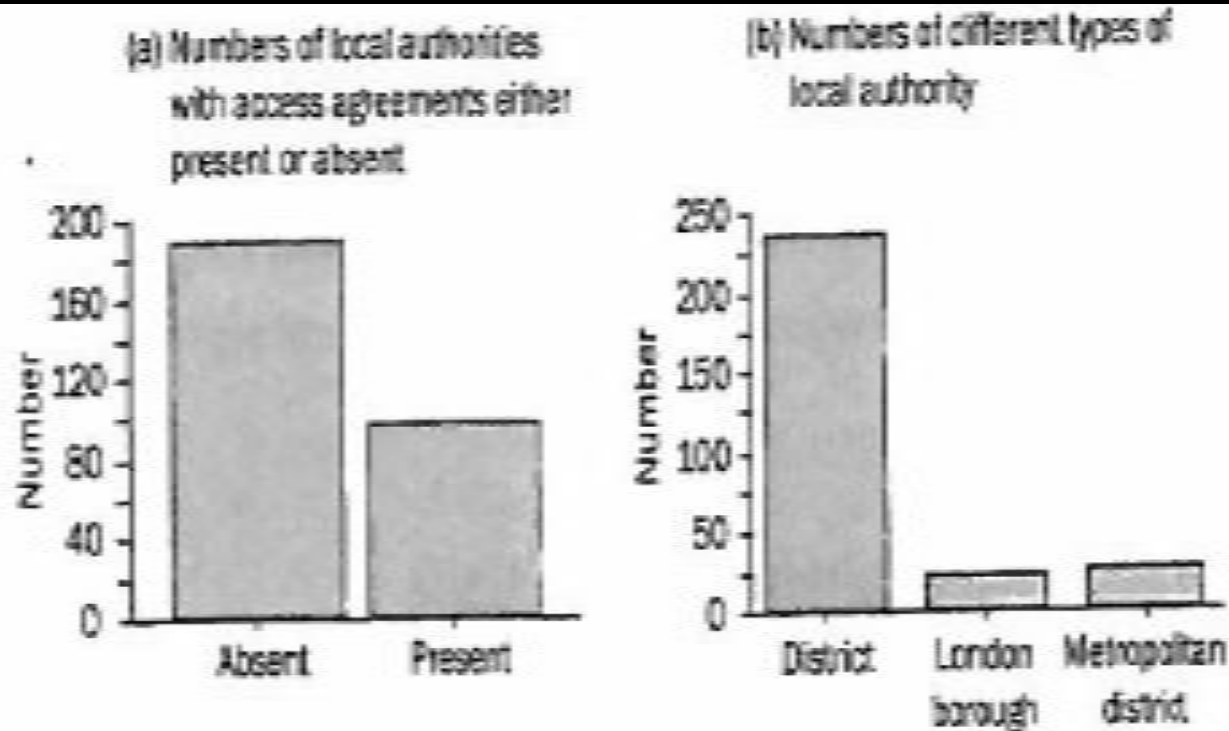


Figure 6.1 Frequency distribution of types of local authority with access agreements either present or absent ($n = 287$)

RESULTS

110

- Two frequency distributions
 - 1. for the number of LA with /without access agreements
 - 2. Numbers of different types of LA
- The results can be summarised in a contingency table

Authority	Agreements Present	Agreements Absent	Totals
District	68	169	237
Metro.	20	7	27
London Bor.	10	13	23
TOTALS	98	189	287

EXPECTED VALUE

- NB

Don't need to use this formula for simple one row Chi (e.g. your first exercise.) Just divide the total by the number of .columns

Box 6.1 Formula for the expected values for testing associations between frequency distributions

The expected values for each cell of a contingency table are calculated as:

$$\text{Expected value (E)} = \frac{\text{Row total} \times \text{Column total}}{\text{Grand total}}$$

This formula can be arrived at from basic probability theory (see Box 2.4). From the data in Worked Example 6.1a, there are 287 authorities in total. Of these, 237 are districts. If the null hypothesis is true, an authority randomly drawn from the 287 would have a 237/287 chance of being a district. Similarly with access agreements, 98 have access agreements, so there is a 98/287 chance of randomly drawing an authority with such agreements. If authority type and having access agreements are independent events, then the two probabilities are multiplied to find the probability of being a district *and* having access agreements:

$$\frac{237}{287} \times \frac{98}{287} = 0.282$$

This is a probability (and thus it lies between 0 and 1); to turn this into an expected number of districts with access agreements, we multiply the probability by the grand total (287):

$$\frac{237}{287} \times \frac{98}{287} \times 287 = 80.927$$

Note that this calculation simplifies to the general formula by cancelling 287 on the top and bottom of the equation:

$$\frac{237 \times 98}{287} = 80.927$$

EXPECTED CALCULATION

- Null no association/no difference (the distributions are independent of each other)
- Expected value (NB Check the totals should be the same as the observed)
- **Expected value** = $\frac{\text{Row total} * \text{Column total}}{\text{Grand Total}}$

$$\frac{237 * 98}{287} = 80.927$$

CALCULATION

Example

$$237 \times 98 = 80.927$$

287

- Calculate the other expected values
- Suggested layout for calculations
- Answers in 2 slides



	Agreement	Present	Chi	Agreement	Absent	Chi	Row Totals
	Observed	Expected	Sum	Observed	Expected	Sum	
D	68	80.927	2.0649	169	156.073	1.0707	237
M	20	9.220	12.6058	7	17.780	6.5363	27
L	10	7.854	0.5866	13	15.146	0.3042	23
Column total	98			189			<u>Total</u> <u>287</u>

by looking at the individual cell X^2 components (see

Box 6.2 Formula for the test statistic χ^2

The components of the test statistic are first calculated for each cell of the contingency table:

$$\text{Cell } X^2 = \frac{(O - E)^2}{E}$$

where: O is the observed (measured) value and E the expected (calculated) value.

The individual cell X^2 values are then summed to give the total X^2 :

$$X^2 = \sum \frac{(O - E)^2}{E}$$

The total X^2 value is then compared with a table of chi-square values (χ^2 : an extract from which is shown in Table 6.2) at $df = (r - 1)(c - 1)$, where r is the number of rows and c the number of columns: see Box 6.3 for an explanation of the logic behind this calculation of the degrees of freedom.

CHI SQUARE

- So is there an association between our observed and expected results?
- See worked example in booklet & next slide p125
- At 2 degrees of freedom our calculated value is >than the tabulated $p=0.05$ (5.99) and hence there is a significant association (Reject null/ no association)

WORKED EXAMPLE OF CHI

116

Use this table to interpret what the results mean (see p.126)



WORKED EXAMPLE 6.1C *Test for association between local authority types and presence or absence of access agreements: calculating the test statistic (X^2). Observed and expected values are taken from Worked Examples 6.1a and 6.1b, respectively. The formula for the cell X^2 values is $(O - E)^2/E$ (see Box 6.2)*

	Agreements present		Agreements absent	
	Calculating cell X^2 values	Cell X^2	Calculating cell X^2 values	Cell X^2
Districts	$(68 - 80.926)^2/80.926 =$	2.0649	$(169 - 156.073)^2/156.073 =$	1.0707
Metropolitan districts	$(20 - 9.220)^2/9.220 =$	12.6058	$(7 - 17.780)^2/17.780 =$	6.5363
London boroughs	$(10 - 7.854)^2/7.854 =$	0.5866	$(13 - 15.146)^2/15.146 =$	0.3042

$$\text{Total } X^2 = 2.0649 + 12.6058 + 0.5866 + 1.0707 + 6.5363 + 0.3042 = 23.168$$

This value is then compared with values in a chi-square table (see the shaded area of Table 6.2) at degrees of freedom calculated as:

$$df = (r - 1)(c - 1) = (3 - 1)(2 - 1) = 2$$

where r is the number of rows and c the number of columns.

HOW TO USE THE CRITICAL VALUE TABLES

117

Table 6.2 Selected values of chi-square (χ^2). Reject the null hypothesis if the calculated value of χ^2 is greater than the table χ^2 value. Shading indicates the critical values for the example referred to in the text. A more comprehensive table of χ^2 values is given in Table D.8 (Appendix D)

df	χ^2	
	P = 0.05	P = 0.01
1	3.841	6.635
2	5.991	9.210
3	7.815	11.345
4	9.488	13.277
5	11.070	15.086

RECYCLING EXERCISE

- 100 people to see if they recycle or not. Is there a significant difference based on gender?
- 4 scenarios
 1. Male recycles
 2. Male doesn't
 3. Female recycles
 4. Female doesn't

Gender	Recycles	Doesn't	Total
Females	18	7	25
Male	42	33	75
Total	60	40	100

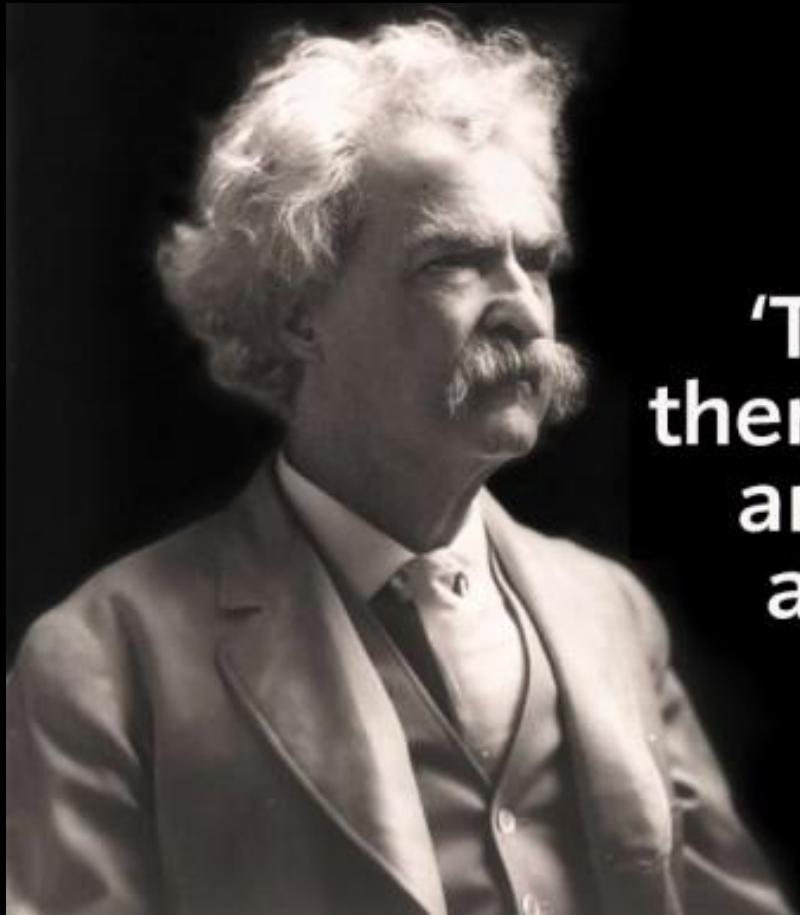
RECYCLING ANSWERS ¹¹⁹

- Set hypotheses

Gender	Recycles	Expected	Chi	Doesn't	Expected	Chi	Total
Females	18	15	0.6	7	10	0.9	25
Male	42	45	0.2	33	30	0.3	75
Total	60			40			100

- Answer $0.6+0.9+0.2+0.3=2.0$
- 1 degree of freedom tabulated 3.841 Complete the exercises from the hand-out/booklet
 - Exercise 1 Dixella pupae
 - Exercise 2 woodland management

GOOD LUCK



**'There are lies,
there are damn lies
and then there
are statistics.'**

MARK TWAIN