



Prifysgol Cymru
Y Drindod Dewi Sant
University of Wales
Trinity Saint David

DSAW7102

Quantitative Research Methods: **Key aspects of methods**

Christine Davies

Trawsnewid Addysg; Trawsnewid Bywydau
Transforming Education; Transforming Lives



www.uwtsd.ac.uk

DSAW7102 timetable

Tues

(9-11am) sampling error; sample size estimation; statistical power

(11.30-1pm) surveys and questionnaires

lunch

(2-3) introduction to learning resources and database access (Dolly Yang)

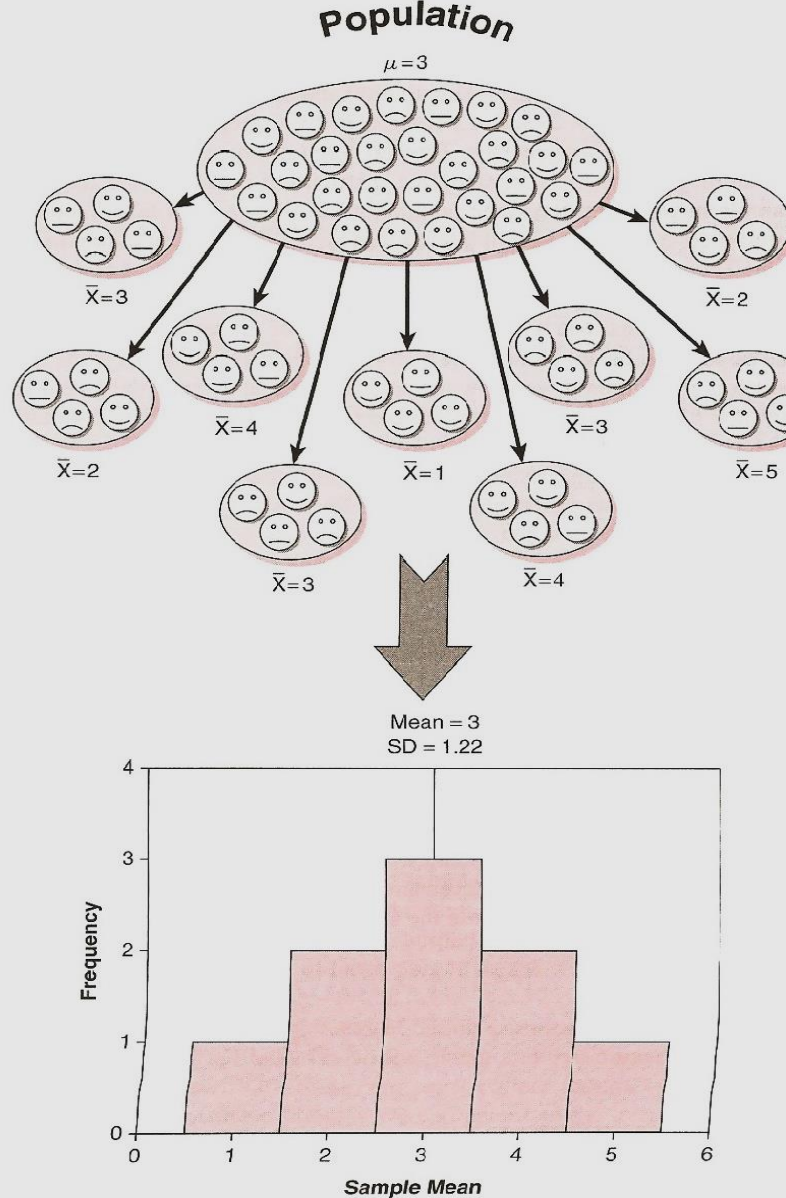
(3.30 -5pm) exploring databases; basic secondary analysis

Sampling distributions, variation, and standard error

population

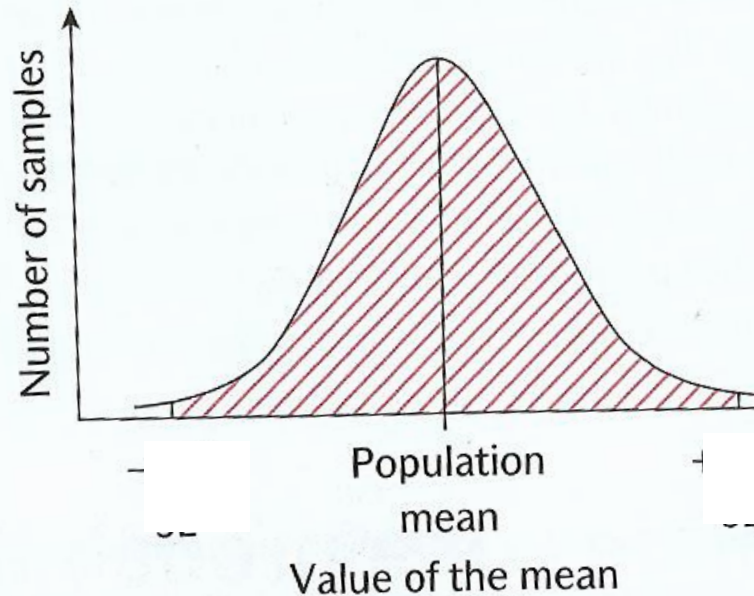
samples

Sample
distribution



From: Field, A. (2005) *Discovering Statistics using SPSS*. Sage, London.

Samples and populations



This graph shows a sample distribution of the means of a large number of samples. This shape is described as a **'normal' distribution**.

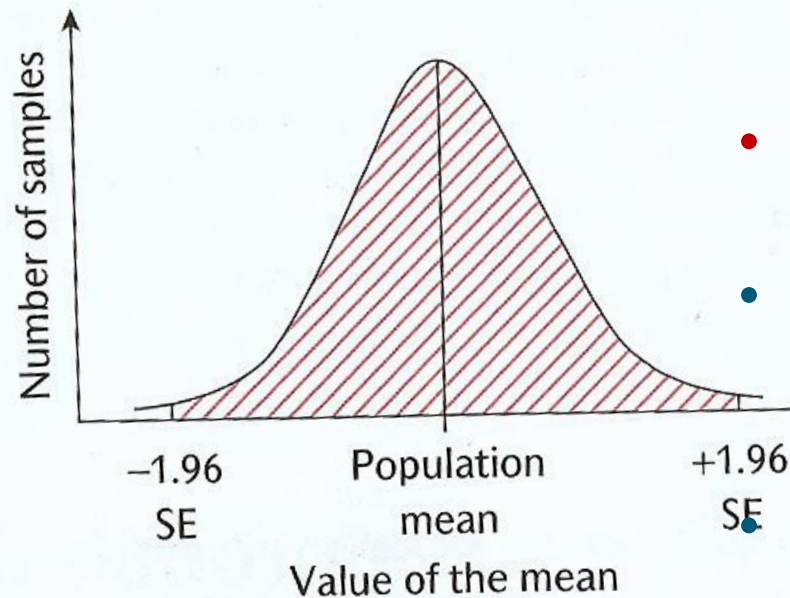
Most sample means are gathered around the population mean (μ), but there is much variation. The extent of the variation can be quantified by the **'standard error'**

From: Bryman, A. (2012) *Social Research Methods* (4th edition).
Oxford University Press, Oxford

What is a 'standard error'?

- A '**standard error**' is a measure of how much variation is present in a distribution of sample means
- It is an estimate of how much a sample mean is likely to differ from the population mean (μ).
- A SE (σ) is like a 'standard deviation' for sample means
- It is equivalent to '**sampling error**'

Standard Error and Confidence Intervals

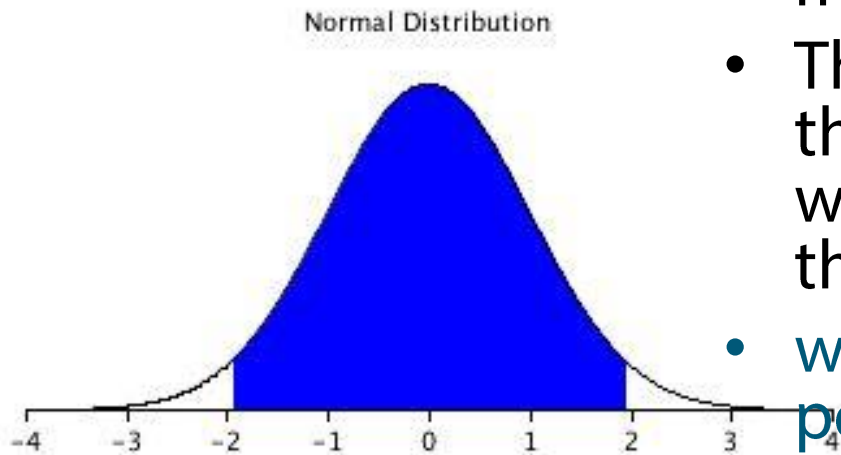


- Most sample means are 'bunched' at the centre of the distribution, around the population mean
- **95%** of all sample means lie within **± 1.96 SE**
- So, we can be 95% confident that the population mean lies within **± 1.96 SE**. This is the **95% confidence interval** or level.
- [we can also look for changes/ effects outside this interval with a 5% level of significance ($p < 0.05$)]

From: Bryman, A. (2012) *Social Research Methods* (4th edition).
Oxford University Press, Oxford

More about confidence intervals

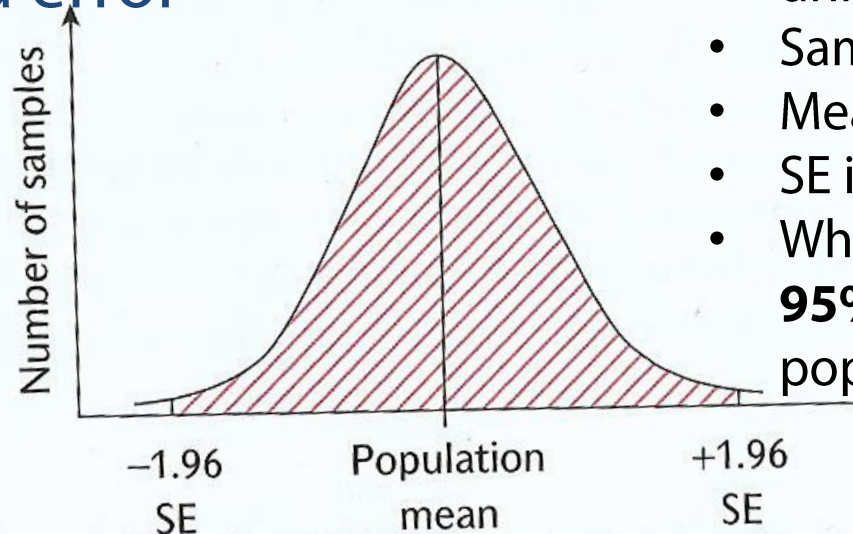
- The confidence interval is the part of the distribution within which we are confident that the population mean is located
- The **95% confidence interval** is the part of the distribution within which we are 95% confident that the population mean is located
- we can be 95% confident that the population mean lies within **± 1.96 SE.**
- This can also be expressed (a priori) as the '**margin of error**'



Estimating standard error

- The standard error provides an indication of the variability of sample means, ie. sampling error
- We can calculate standard error using the formula: $SE = SD / \sqrt{n}$,
- It is also possible to calculate SE from a confidence interval or margin of error decided *a priori*: this is often done in surveys (and in online calculators)

Estimating confidence intervals if we know the Standard error



Example:

- Suppose 450 students are asked how many units of alcohol they drink per week.
- Sample size (n) = 450
- Mean ($\bar{x}$) is 9.7
- SE is 1.3
- What would be our estimate of the **95%** confidence interval for the population mean?

(NB:
 $SE = SD / \sqrt{n}$)

We can be 95% confident that the population mean is within ± 1.96 SEs, ie:

$9.7 + (1.96 \times 1.3)$ and

$9.7 - (1.96 \times 1.3)$

Hence, we can be 95% confident that the population mean is between 12.25 and 7.15 units of alcohol

Now look at your hand-out!

Estimating SE from the 'margin of error'

- The researcher decides an acceptable '**margin of error**' (MRE), eg. $\pm 4\%$
- This represents the confidence interval selected *a priori*
- If the confidence level is 95%, then $\pm 4\%$ represents the 95% confidence interval
- The 95% confidence interval also represents $\pm 1.96SE$, so:
 - $4\% = 1.96SE$, and
 - $SE = 4/1.96 = 2.04$

Two ways to find SE!

- If sample statistics are available, use:
 - **$SE = SD / \sqrt{n}$**
- Decide on a MRE and confidence level. If this is 95%, then MRE represents 95% confidence interval
- Then **$SE = MRE / 1.96$**

Standard error: the need to minimize

- The greater the standard error (or sampling error), the less likely a sample will represent a population
- Hence, we need to minimise sampling error
- This can be done by **increasing sample size** (the greater the sample size, the smaller the standard error)

How do we decide on sample size?

- We need as high a sample size as possible
- But very large samples involve time and money
- So we need a way of estimating an acceptable minimum sample size
- NB: the **response rate** should also be considered, and sample size adjusted accordingly



Estimating sample size for % distributions

1. Decide on **confidence level**: eg. 95% (ie. ± 1.96 SE)
2. Decide on an acceptable **margin of random error (MRE)**, eg. $\pm 3\%$.
3. Find SE. For 95% confidence interval,
$$SE = MRE \div 1.96$$
4. Estimate the expected % result (P). If no clear indication, select 50%
5. Now put values into the formula:

$$n = P(100-P) \div (SE)^2$$

Example of sample size: % calculation

1. Decide on confidence level: eg. 95% (ie. ± 1.96 SE)
2. Decide on an acceptable margin of random error (MRE), eg. $\pm 3\%$.
3. For 95% confidence interval, $SE = MRE \div 1.96 = 1.53$
4. Estimate the expected % result (P) as 50%
5. Now put this into the formula: $n = P(100-P) \div (SE)^2$
6. ie: $50(100-50) \div 2.34 = 2500 \div 2.34$
7. = 1068

Estimating sample size using online calculators

What are you asked for?

- **Confidence level** (usually 95%) [this is needed for SE calculation]
- **margin of error** (eg. $\pm 3\%$) [this is needed for SE calculation]
- **Population size** [this is not especially important; becomes more so if population size is small]
- [the expected % result, P, is generally assumed to be 50%]
- Eg. <http://www.raosoft.com/samplesize.html> [allows change of several variables]
- Eg: <https://www.surveymonkey.com/mp/sample-size-calculator/>

Try out an online calculator

Values to add in:

- **Confidence level** : 95%
- **margin of error** : $\pm 3\%$
- **Population size** : set at 20,000 (this estimates larger populations, also)

- <https://www.surveymonkey.com/mp/sample-size-calculator/>
- Ans: 1014
- <http://www.raosoft.com/samplesize.html>
- Ans: 1014

Now look at your hand-out!

..and eventually, this will need to be adjusted for response rate

More errors!?!

- In *confirmatory* research designs using *inferential statistical models*, there are further possible sources of error
- Researchers who use small samples run the risk of finding an effect that *does not* exist (a '**Type I**' error)
- There is also a risk of *not* finding an effect which *does* exist (a '**Type II**' (ie. **Type 2**) error)



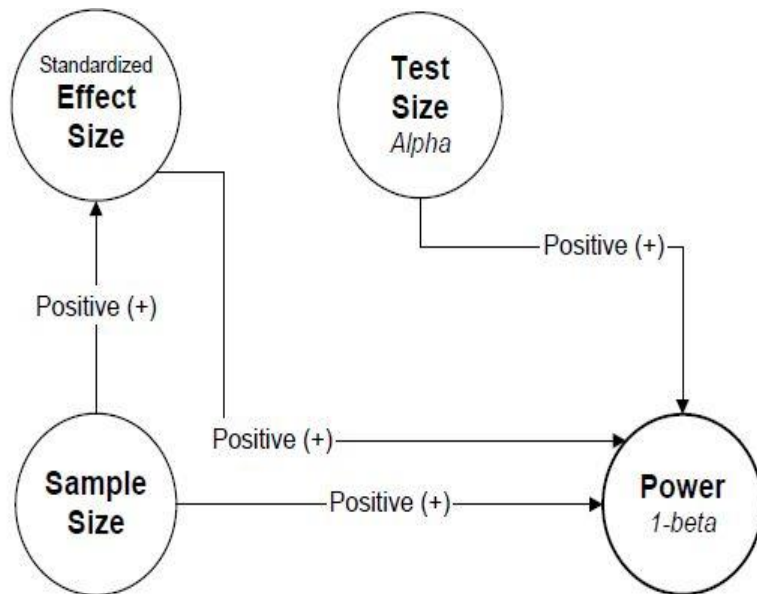
The statistical power of a study

- Studies with **Type II errors** (ie. which don't find a real effect) are said to be '**under-powered**': they don't have enough **statistical power** to detect the effect they were meant to find
- power is usually decided *in advance* (ie. *a priori*) of a study, and set to at least 80%
- The power of past studies can also be calculated (ie. *a posteriori* or *post hoc*)



Factors affecting Statistical power

Several factors influence statistical power, and the greater these are, the greater the statistical power:



- Sample size
- Confidence/Significance level
- Effect size (NB: 'Cohen's d')

Planning a survey

Back to planning a survey!

Tackled so far!

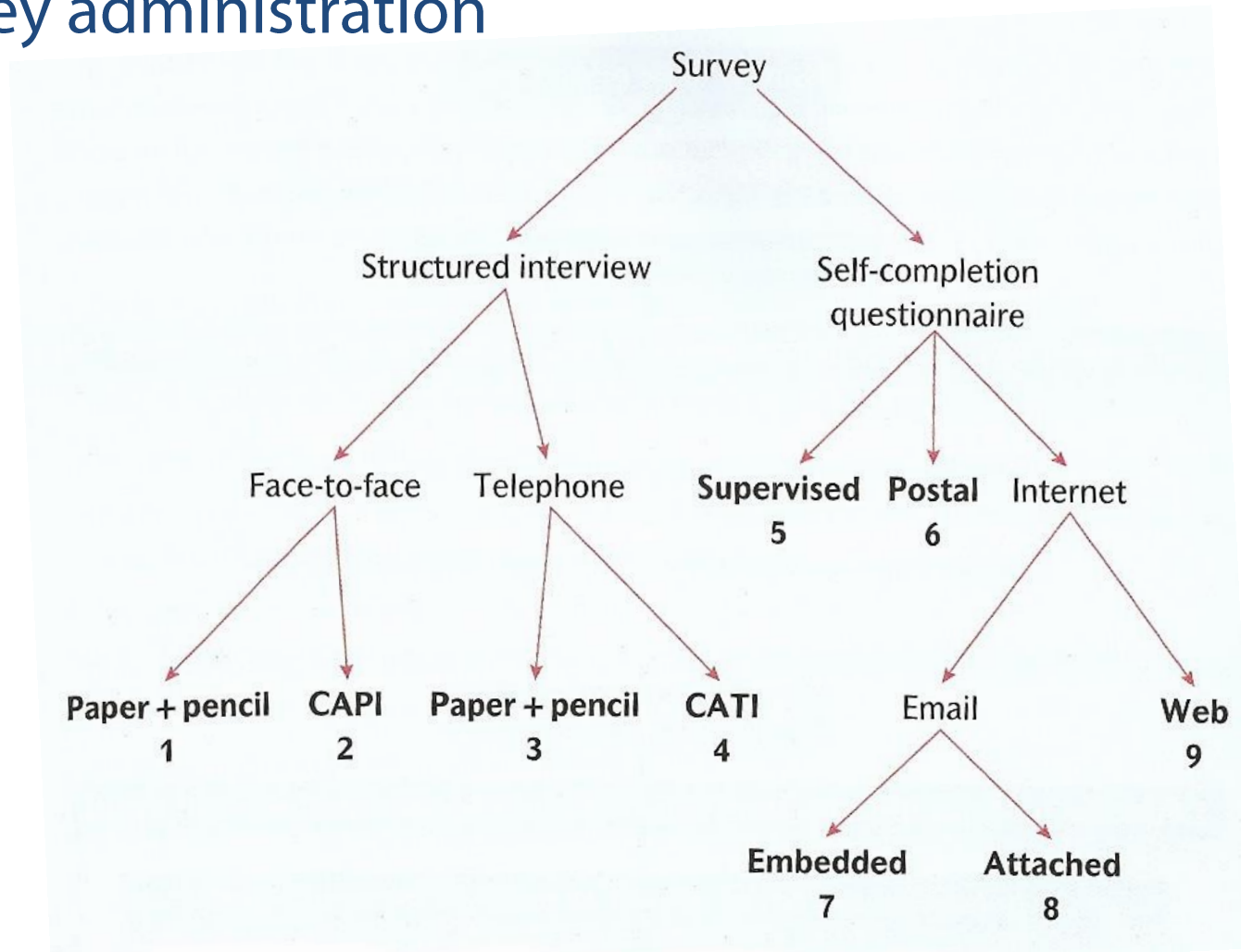
- Sampling frame
- Sampling type
- Sample size



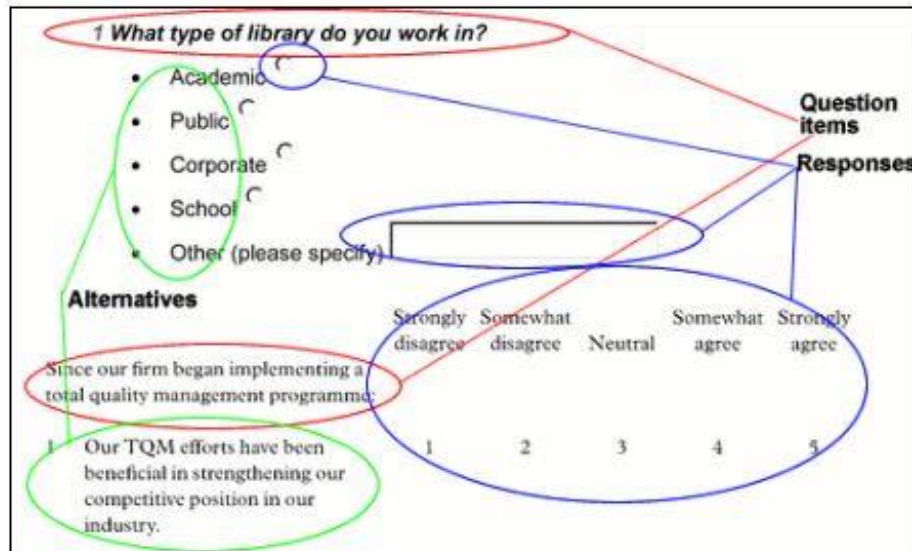
Now we need to find out about...

- Mode of administration
- Questions: style, indicators (coding requirements), number, appearance of questionnaire

survey administration



constructing questions in a questionnaire



3 key aspects:

- Question stem
- Alternatives
- responses

<http://www.emeraldgrouppublishing.com/research/guides/methods/questionnaires.htm?part=4&view=print>

'Closed' questions and indicators

10 On a scale of 1-5, 1 being not important, please rank the importance of data storage for your library staff:

- ☐ 1 not important
- ☐ 2
- ☐ 3
- ☐ 4
- ☐ 5 very important

- Likert scales
- Differential scales
- Checklists
- ranking

See: Leung, W-C. (2001) 'How to design a questionnaire'. *Student BMJ*, 9, p197-189

Validity of questionnaires

- **measurement validity**: indicates whether a particular measure is a valid reflection of the concept
- **external validity**: relates to whether outcomes are **generalisable**, ie. applicable to a wider population
- **internal validity**: relates to validity of relationship between variables
- **ecological validity** relates to validity of outcomes in terms of normal 'environment'

Reliability of questionnaires

The degree to which research outcomes are **consistent**, and do not fluctuate

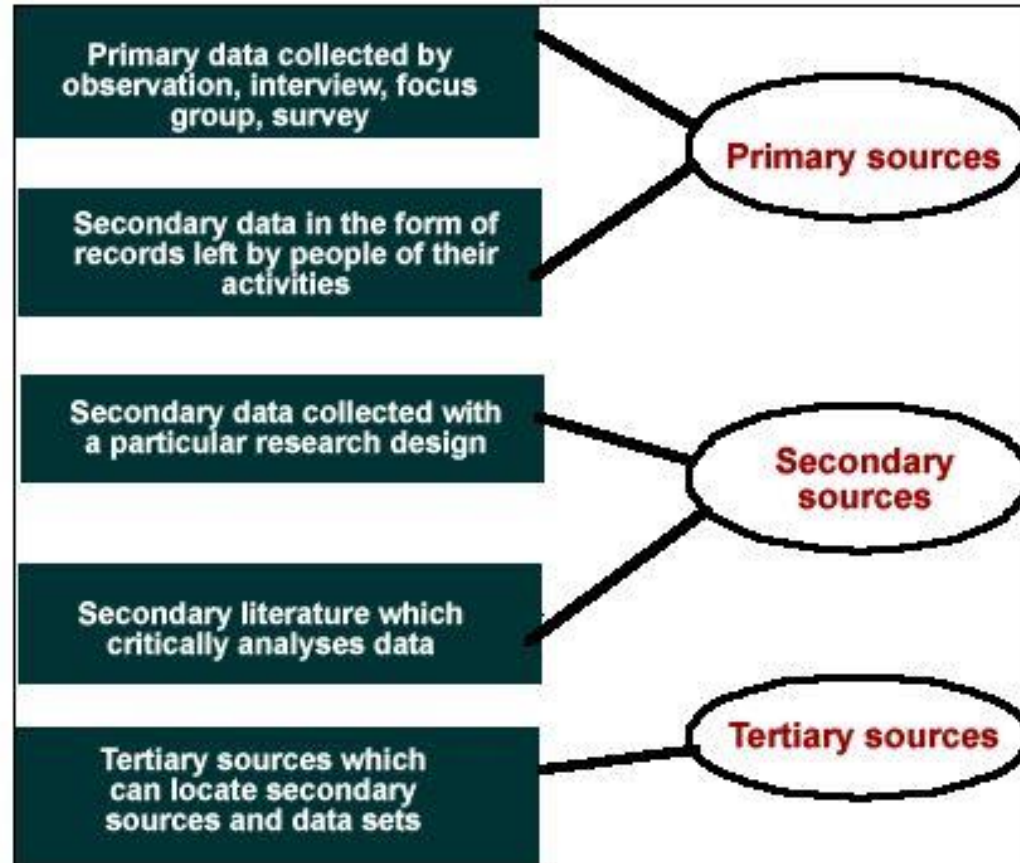
This can be estimated using statistical indices such as '**Cronbach's alpha**' (a type of correlation)

Your questions!

- Using the hand-outs, or other sources of information, think about the nature of the questions you might use in a questionnaire/survey
- What factors will be important for your context?
- Write out a couple of your own exemplar questions
- What other aspects of survey administration do you need to consider?

Finding Secondary data

An overview of secondary data



See:

<http://www.emeraldgrouppublishing.com/research/guides/methods/archival.htm>

Using secondary data: why?

- Efficiencies of time and cost
- Different contexts can be explored
- Source of new ideas
- Large-scale studies from government-related bodies usually provide data with high reliability and validity

[See article](#)

Sources of secondary data

- UK Data Archive (www.data-archive.ac.uk)
- Economic and Social Data Service (www.esds.ac.uk/findingData/majorStudies.asp)
- Office of National Statistics (www.statistics.gov.uk/default.asp)
- British Household Survey
- British crime Survey
- British social Attitudes Survey
- Labour Force Survey

Explore, and see
hand-out!

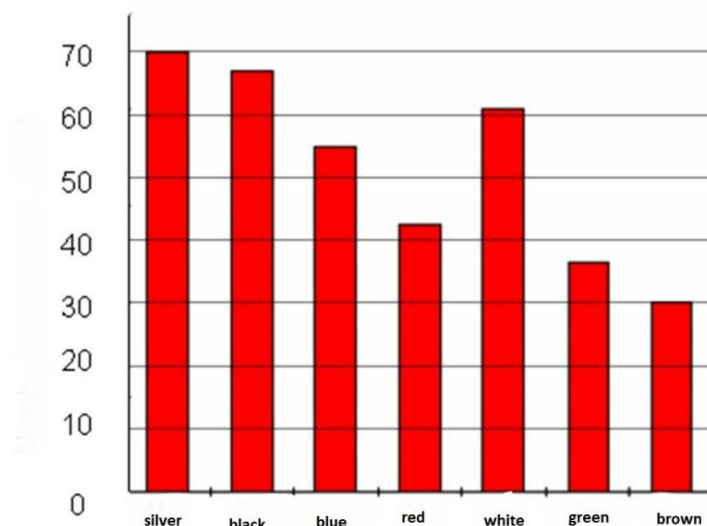
Selecting data

- Which set(s) match your field of interest?
- In huge data-sets, how do you decide which data to use?
- Should you 'correct' data, eg. over-sampling (counting the same data more than once), or under-sampling (reducing the entries in one or more categories)



Using secondary data: secondary analysis

- Graphic summaries: tables, charts
- Descriptive statistics: mean, mode, range, SD,
- Inferential statistics: relationships, differences



thinking about data sets: an example

age	Weekly Hrs on internet	Weekly income, £	M/F
22	10	200	M
44	8	400	F
66	7	200	F
18	12	20	F
65	10	150	M
77	6	250	M
33	4	350	M
82	4	150	F
46	7	450	F
49	9	550	M
52	6	300	M

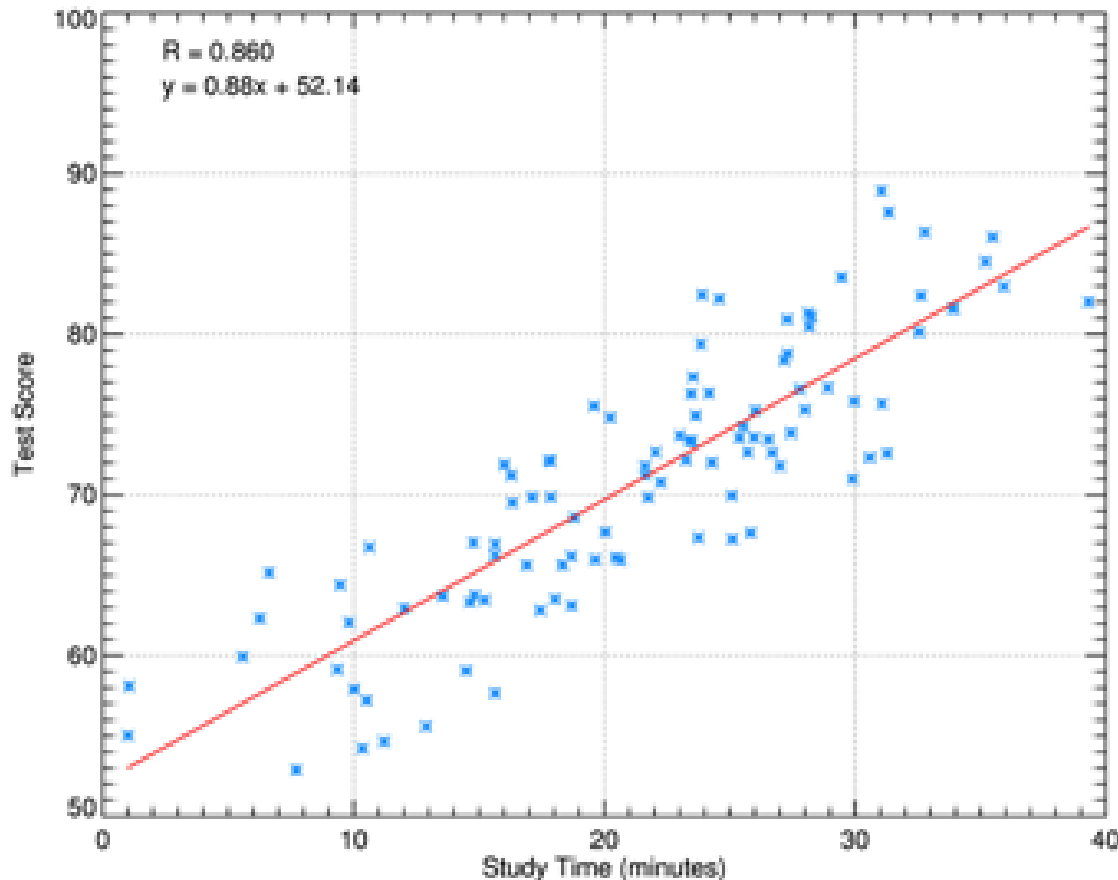
- What is the sample size here, and what do you conclude about it?
- What sort of descriptive statistics could be obtained?
- What sort of inferential statistics could be obtained (given a larger sample)?

Summarising data: Contingency Table ('Cross tab')

Contingency tables summarise variables and help to see trends and relationships

			Hours on internet per week				Total
			0-1	2-4	5-10	>10	
Gender	M	Count	60	139	298	82	579
		%	10.4%	24.0%	51.5%	14.2%	100.0%
	F	Count	43	108	265	53	469
		%	9.2%	23.0%	56.5%	11.3%	100.0%
Total		Count	103	247	563	135	1048
		%	9.8%	23.6%	53.7%	12.9%	100.0%

Relationship between variables (interval/ratio data)



- the 'line of best fit' (regression line) shows the relationship between variables clearly
- This diagram indicates a positive relationship (correlation) between the variables (correlation coefficient, R ; -1 to +1)

Questions??



References

- Bryman, A. (2012) *Social Research methods* (4th edition). Oxford University Press, Oxford. [chapter 7, 14]
- Byrne, D. (2002) *Interpreting Quantitative Data*. Sage, London. [chapter 1]
- Gorard, S. (2001) *Quantitative methods in educational Research*. Continuum, London and New York

To read next!:

- Field, A. (2005) *Discovering Statistics using SPSS*. Sage, London. **(Chapter 2)**