MiRoR
Methods in
Research on Research

A review of basic statistical concepts: variability, uncertainty and confidence intervals

Jamie Kirkham

Statistics

Descriptive Statistics

- Describe the characteristics of a sample
- Graphical methods
  - E.g. bar chart
- Summary measures of location and spread

Variability in the sample

Measure of 'central tendency'

Inferential Statistics

- Use the sample to draw conclusions (or make inferences) about the sampled population
  - hypothesis testing
  - confidence interval estimation

Types of data

- Categorical
  - nominal: sex (M,F), diagnostic group (A, B)
  - ordinal: response to therapy (worse, none, improved), social class (I, II, III, IV, V)

- Continuous
  - E.g.: age, height, weight, blood pressure

Summarising categorical data

- Most commonly use counts, proportions, percentages
- Sometimes rates are more helpful

  e.g. 2 out of a sample of 5000 are carriers of a rare disease:

  proportion = 2/5000 = 0.0004

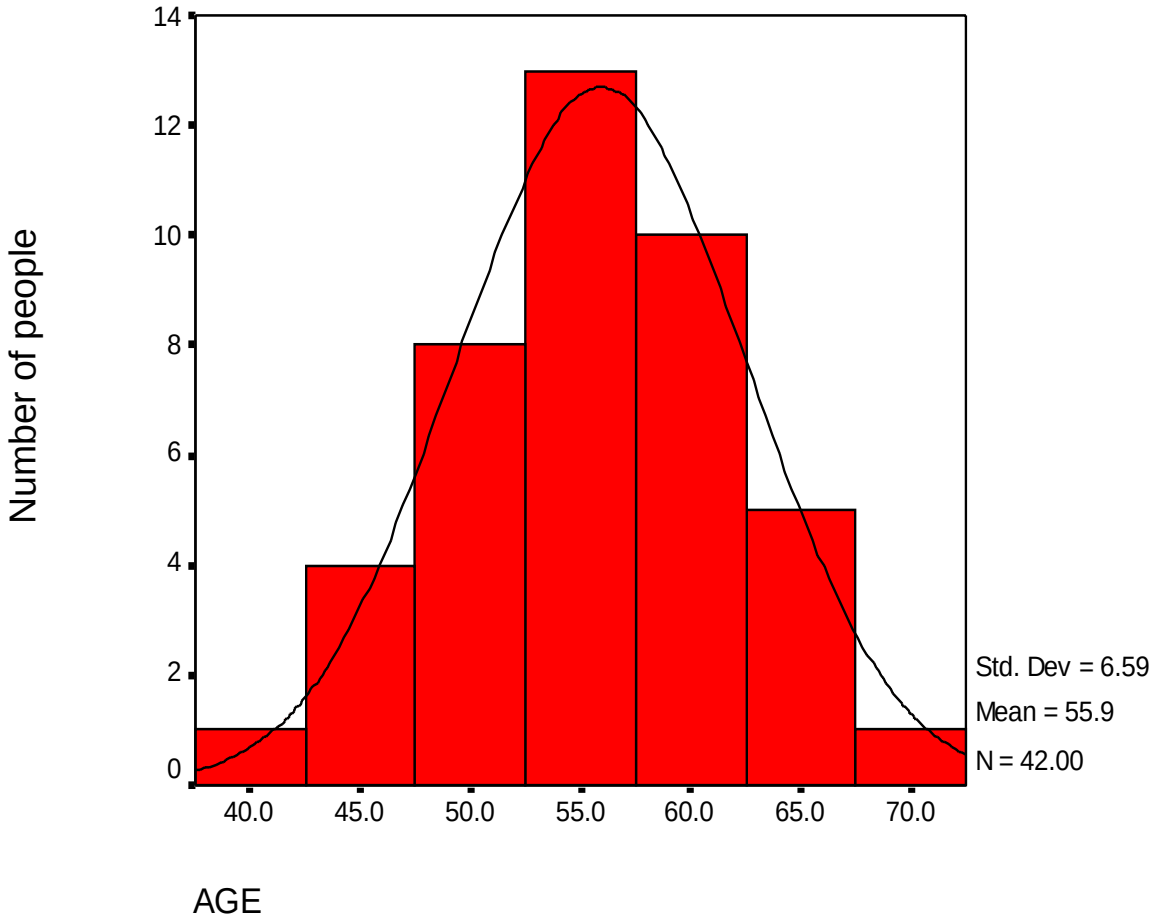
  percentage = 2/5000 $\times$ 100 = 0.04%

  rate = 2/5000 x 10000 = 4 per 10,000

- Use the format most appropriate for the data

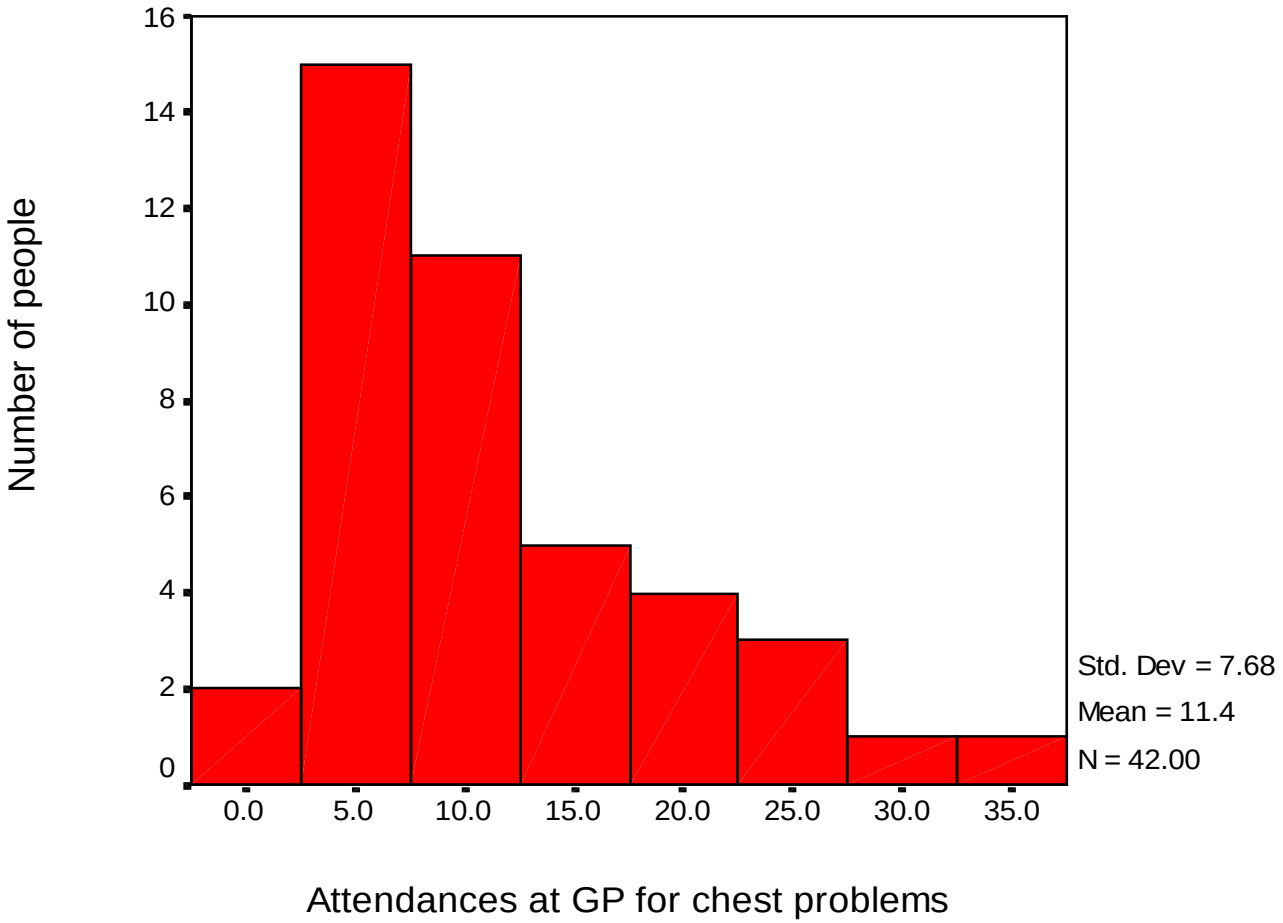
Presenting continuous data

Normal distribution



Presenting continuous data

Skewed distribution



Summarising numerical data

Symmetrical data (normally distributed)

- The mathematical way of defining the mean and variance (as seen in text books) is as follows:
- Mean $(\bar{x})$:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

where $\sum$ = 'the sum of'

and x_1, x_2, x_3, ….x_n are the data values, and there are n of them

- Variance (s^2):

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

Summarising numerical data

Symmetrical data

Example: Seven men were weighed in kilogram's as

57.0 62.9 63.5 64.1 66.1 67.1 73.6

- Using the **mean** as a measure of location:

$$\bar{x} = \frac{57.0 + 62.9 + 63.5 + 64.1 + \ldots + 73.6}{7} = 64.9\,\text{kg}$$

- Using the **variance** as a measure of 'spread':

$$s^2 = \frac{(57.0 - 64.9)^2 + (62.9 - 64.9)^2 + \ldots + (73.6 - 64.9)^2}{6}$$

$$= 25.16\,\text{kg}^2$$

Summarising numerical data

Symmetrical data

- It is more common to quote the square root of the variance, called the **standard deviation (*s*)**:

$$s = \sqrt{\text{variance}}$$

- In our example, $s = \sqrt{25.16\,\text{kg}^2} = 5.02\,\text{kg}$

- This gives a measure of variation on the original scale of measurement

Summarising numerical data

Skewed data

- Use the **median** as a measure of location ie. the value which divides the *ordered* data in half
- Use the **interquartile range** (IQR) as a measure of spread:

The **lower quartile** and the **upper quartile** are the values below which and above which one-quarter of the data fall, respectively.

IQR = upper quartile - lower quartile

Summarising numerical data

Skewed data (not normally distributed)

Example: The number of children in 19 families was

5, 3, 3, 8, 0, 1, 3, 4, 1, 7, 2, 6, 2, 4, 2, 2, 2, 6, 10

- Order the data from smallest to largest

  0, 1, 1, 2, 2, 2, 2, 2, 3, **3**, 3, 4, 4, 5, 6, 6, 7, 8, 10

- Median is the $1/2 \times (n+1)^{th}$ value

  i.e. $1/2 \times (19 + 1) = 10$th value = 3 children

NB. For an even number of observations, the median is the average of the middle two values. So, for instance, the median for 20 observations would be the average of the 10th and 11th values.

Summarising numerical data

Skewed data

Example (cont'd):

0, 1, 1, 2, **2**, 2, 2, 2, 3, 3, 3, 4, 4, 5, **6**, 6, 7, 8, 10

lower quartile = $1/4 \times (19 + 1)$ = 5th value = 2 children

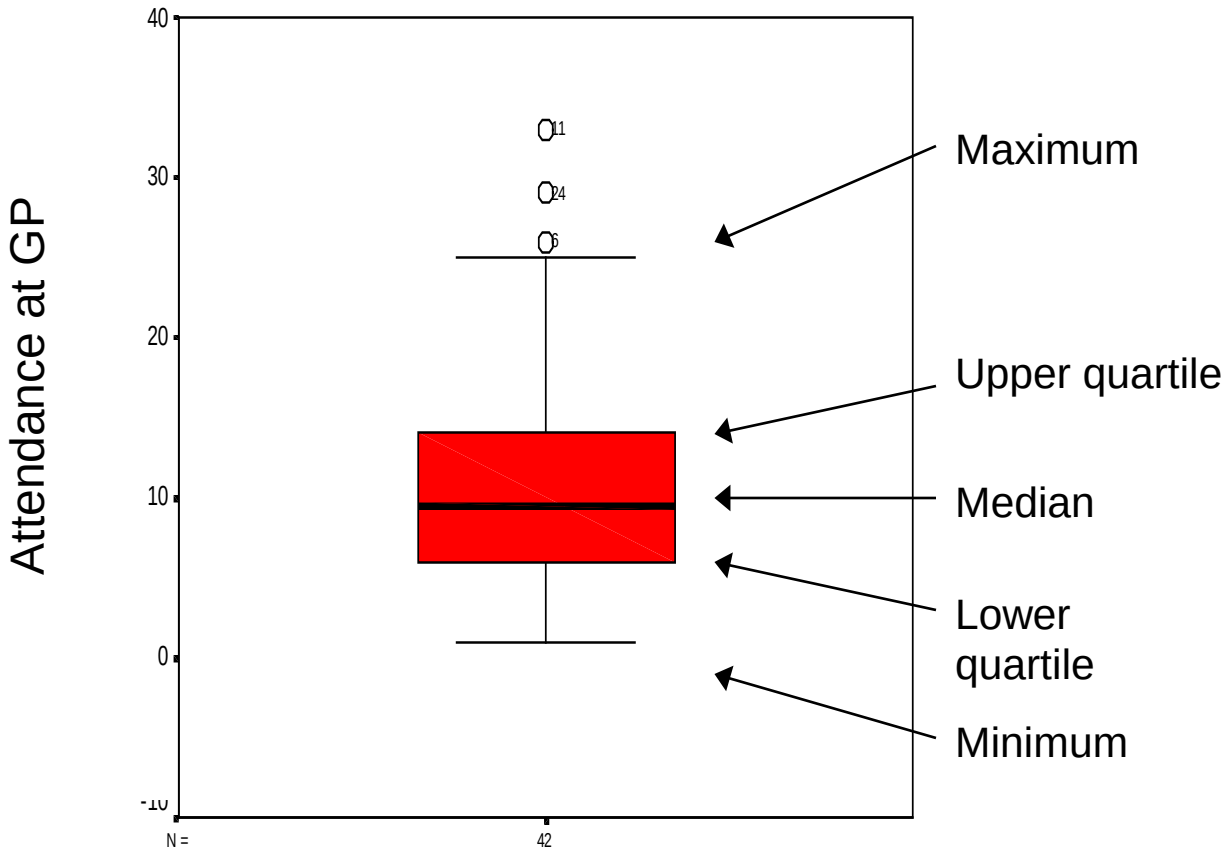
upper quartile = $3/4 \times (19 + 1)$ = 15th value = 6 children

Inter-quartile range = 6 - 2 = 4 children

NB. If there were an even number of observations e.g. 20, the lower quartile would be the average of the 5th and 6th values, and the upper quartile the average of the 15th and 16th values

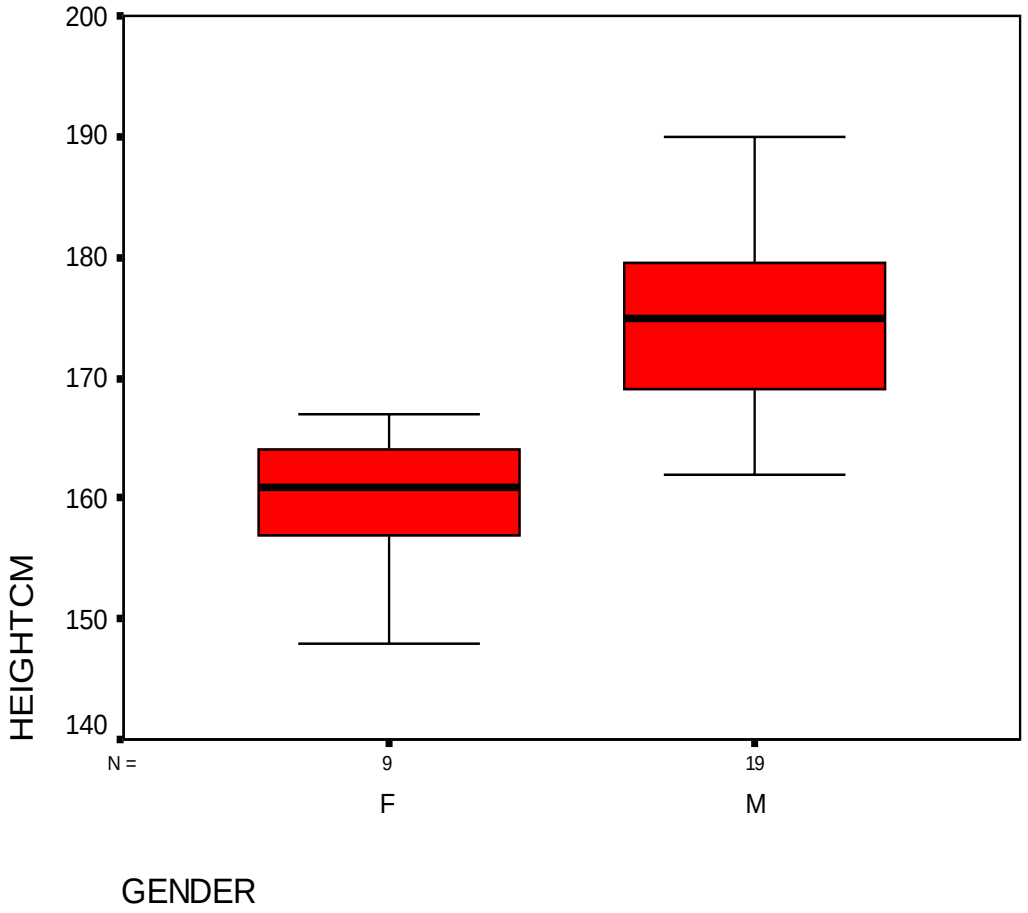
Presenting continuous data

Box and Whisker plot - skewed distribution



Presenting continuous data

Box plots are useful for comparing groups

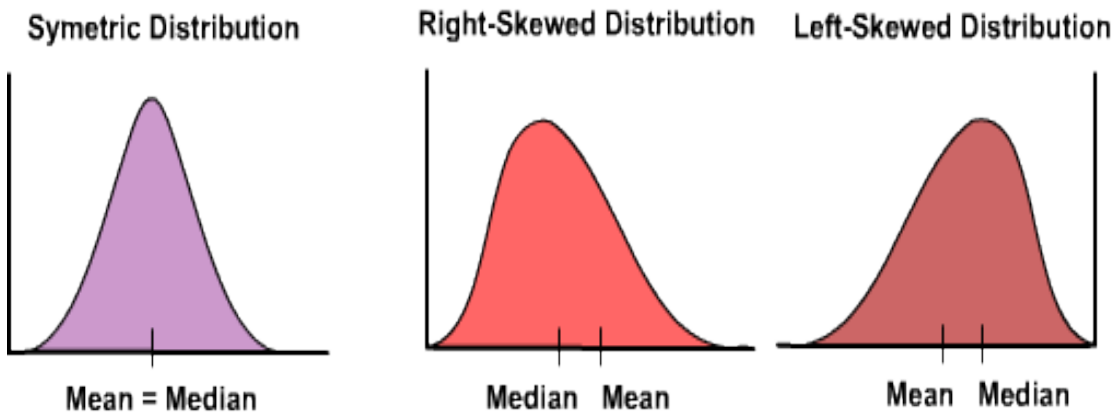


Type of data and distribution

Type of Data

Categorical	Continuous Normal	Continuous Skewed
Frequency Percentage Rate	Mean SD	Median IQR

Summary measures:



Exercise: summary measures

- Pregnant women (within month 4) who are being followed-up by a nutritionist had weights (kg) equal to

64.3; 65.2; 70.0; 54.5; 58.8; 81.5; 61.0; 62.0

- What was:
  - A) the mean
  - B) the standard deviation
  - C) the median

- Do the data suggest a strong skewness of the distribution of the weight?

Confidence Intervals

- Statistical inference involves making estimates from the data.

- Such estimates have uncertainty, and may differ from the true underlying value of interest by chance.

- One way of expressing uncertainty is a **confidence interval, which gives us a range within which the true underlying value is likely to lie.**

- **A 95% confidence interval:** if we repeatedly take independent samples from the same population and compute a confidence interval for each sample, then 95% of the confidence intervals calculated will include the true underlying value of interest.

Confidence interval for a mean

- A 95% confidence interval for a mean is given by:

Estimate of mean $\pm$ 1.96 x SE of mean

SE = standard error = $\frac{St.Dev}{\sqrt{n}}$

Example: The mean and St. dev. of the weights of a random sample of 40 men are 64.9 and 12.02kg.

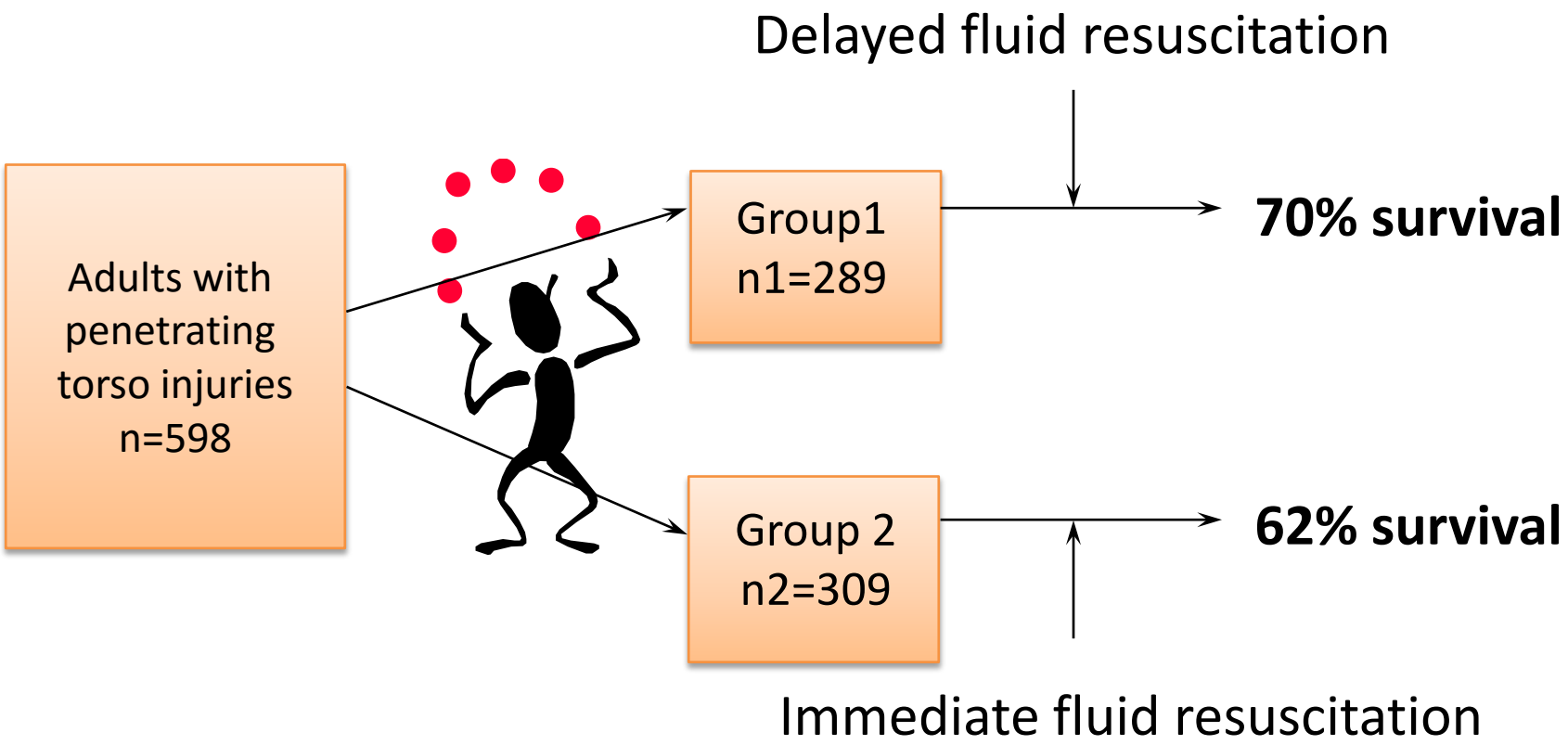
SE = $\frac{12.02}{\sqrt{40}}$ = 1.9 kg

95% CI 64.9 – 1.96x1.9, 64.9 + 1.96x1.9

(61.2, 68.6) kg

➢ 95% Confident that true population mean lies within this interval.

Difference in proportions



Risk of death is reduced by 8%

Confidence interval for a proportion

- Difference in survival: p_1=70% p_2=62% → p_1-p_2=8%
- A 95% confidence interval is a range of values which we are 95% confident includes the true difference.
- 95% CI for the true difference in survival between the two groups is (0.4%, 15.6%)

Is this clinically important?

95% Confidence Interval (CI)

- Confidence Interval for a difference in survival rates
- Difference in survival: p_1=70% p_2=62% → p_1-p_2=8%
- Standard error:

$$\sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}}=$$

$$\sqrt{\frac{0.7(1-0.7)}{289}+\frac{0.62(1-0.62)}{309}}=0.039=3.9\%$$

- General formula: estimate $\pm$ 1.96 · standard error
- 8% $\pm$ 1.96 · 3.9% → 95%CI for the true difference in survival between the two groups = (0.4%, 15.6%)

Confidence interval for a mean difference

- A confidence interval can also be constructed for a mean difference, e.g. When assessing the difference between two groups.

Note: When assessing the significance of a mean difference the important value is zero.

- If CI crosses 0, e.g. (-0.5, 3.4) then no significant difference - implies true difference could be zero.
- If CI does not contain zero, e.g. (2.3, 5.7) then can conclude that difference exists.

Confidence Interval for a ratio

- In a similar manner confidence intervals can be constructed for relative risks / odds ratios and hazard ratios

- Note: When assessing the significance of a mean difference the important value is one.

Odds ratio example: case-control study

Association between Lung Cancer and Smoking reported by Doll and Hill (BMJ 1952)

Smoker	Lung Cancer	
	Cases	Control	Total
Yes	1350	1296	2646
No	7	61	68
Total	1357	1357	2714

Odds Ratio $\hat{\varphi} = \dfrac{1350 \times 61}{7 \times 1296} = 9.08$

95% Confidence Interval (4.12, 20.02)

Odds ratio example

- Study suggests that smokers are 9 times more likely than non-smokers to develop lung cancer.
- Confidence interval suggests that true risk between 4 and 20 times higher for smokers compared to non-smokers.
- Sure about association between smoking and lung cancer less sure about the magnitude of association.

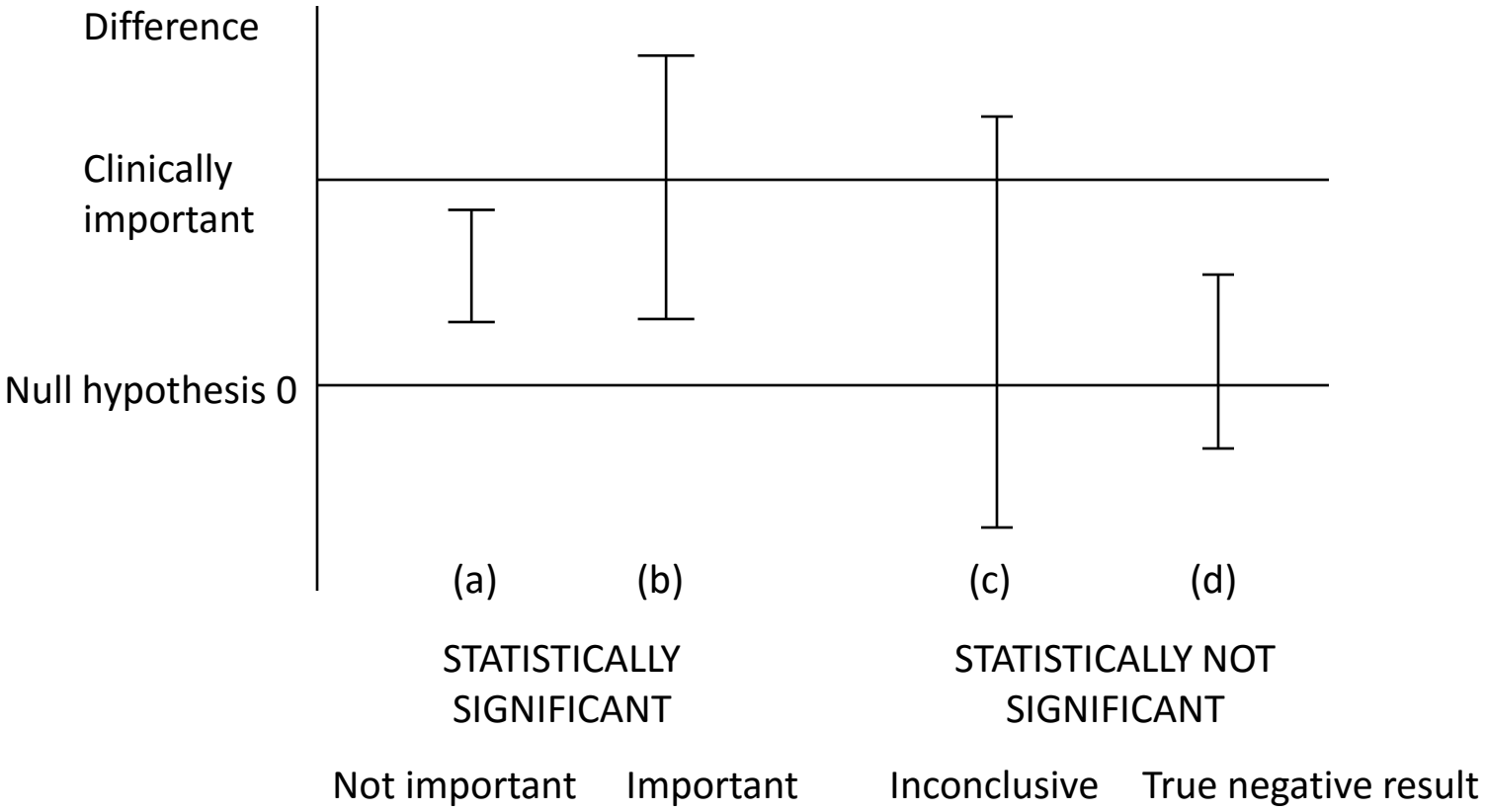
Note: $\hat{\varphi}$ = 1 Odds the same for both cases and controls

$\hat{\varphi}$ > 1 Cases more at risk than controls

$\hat{\varphi}$ < 1 Controls more at risk than cases

Statistical and clinical significance

(from Berry G. (1986), *Med. J. Aust*, **144**: 618-619)



NB. Presentation of confidence intervals is **good practice**

Group Discussion

What factors might influence the width of a confidence interval?

- Change in confidence level:
  - 90% (critical value; 1.65)
  - 95% (critical value; 1.96)
  - 99% (critical value; 2.58)
- Standard deviation decreases
- The sample size increases

In general the narrower the confidence interval the more precise the estimate.

Exercise (1) - Confidence Intervals

Estimate Type	Confidence Interval	Conclusion
Difference in Proportions	(-5%, 7%)	Not significant
Odds Ratio	(2.1, 2.7)	
Difference in Means	(0.7, 1.4)	
Relative Risk	(-0.9, 2.8)	
Hazard Ratio	(1.01,1.03)	

Exercise (2) - Confidence Intervals

No. (%) of successes

Time (weeks)	Wait and See	Injection	Injection - wait and see. Relative risk difference (99% CI)
3	9/57 (16)	47/63 (75)	0.7 (0.4 to 0.9)
12	35/59 (59)	29/65 (45)	0.3 (-0.1 to 0.6)
52	56/62 (90)	44/65 (68)	0.3 (0.04 to 0.4)

The table shows results of a randomised controlled trial comparing corticosteroid injection with a policy of wait and see for tennis elbow. Follow up was for 52 weeks and the results at each follow up visit are shown.

Select from the list below the single true statement with respect to relative risk reduction.

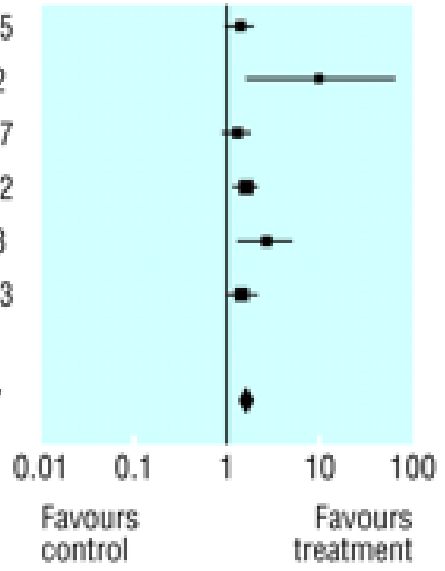
A. There is no significant difference at any time between injection and wait and see
B. At 3 weeks injection is significantly better than wait and see
C. At 12 weeks wait and see is significantly better than injection
D. At 12 weeks injection makes some people worse
E. At 52 weeks there is no significant difference between injection and wait and see

Exercise (3) - Confidence Intervals

Study	Treatment group (No of events/ No of patients)	Control group (No of events/ No of patients)	Relative risk (fixed effect) (95% CI)	Weight (%)	Relative risk (fixed effect) (95% CI)
Cederlof 1966[7]	20/26	14/25		20.12	1.37 (0.92 to 2.06)
Dieppe 1980[8]	10/12	1/12		1.41	10.00 (1.51 to 66.43)
Friedman 1980[9]	15/17	12/17		16.92	1.25 (0.88 to 1.78)
Gaffney 1995[10]	33/42	21/42		29.60	1.57 (1.12 to 2.21)
Ravaud 1999[11]	16/25	7/28		9.31	2.56 (1.26 to 5.18)
Smith 2003[12]	25/38	15/33		22.64	1.45 (0.93 to 2.24)
Total (95% CI)	160	157		100.00	1.66 (1.37 to 2.01)

Test for heterogeneity: χ^2=8.70, df=5, P=0.12, I^2=42.5%

Test for overall effect: z=5.14, P<0.00001



Select which studies:

A. Shows no significant difference between control and treatment
B. Significantly favours control
C. Significantly favours treatment

What can we say about the width of the confidence intervals?

Methods in Research on Research

A presentation delivered at the

first MiRoR training event
October 19-21, 2016
Ghent, Belgium

This project has received funding from the EU Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie Grant Agreement #676207