AN INTRODUCTION TO STATISTICS

Paul Baker
MSc Dip Stat FRCS (T&O)
StR: Northern Deanery
Associate Clinical Researcher: Newcastle University

Overview

- Type of data and their distributions
- Key statistical principles
- Sensitivity and specificity
- Statistical tests
- Survival analysis

Type of data and their distributions

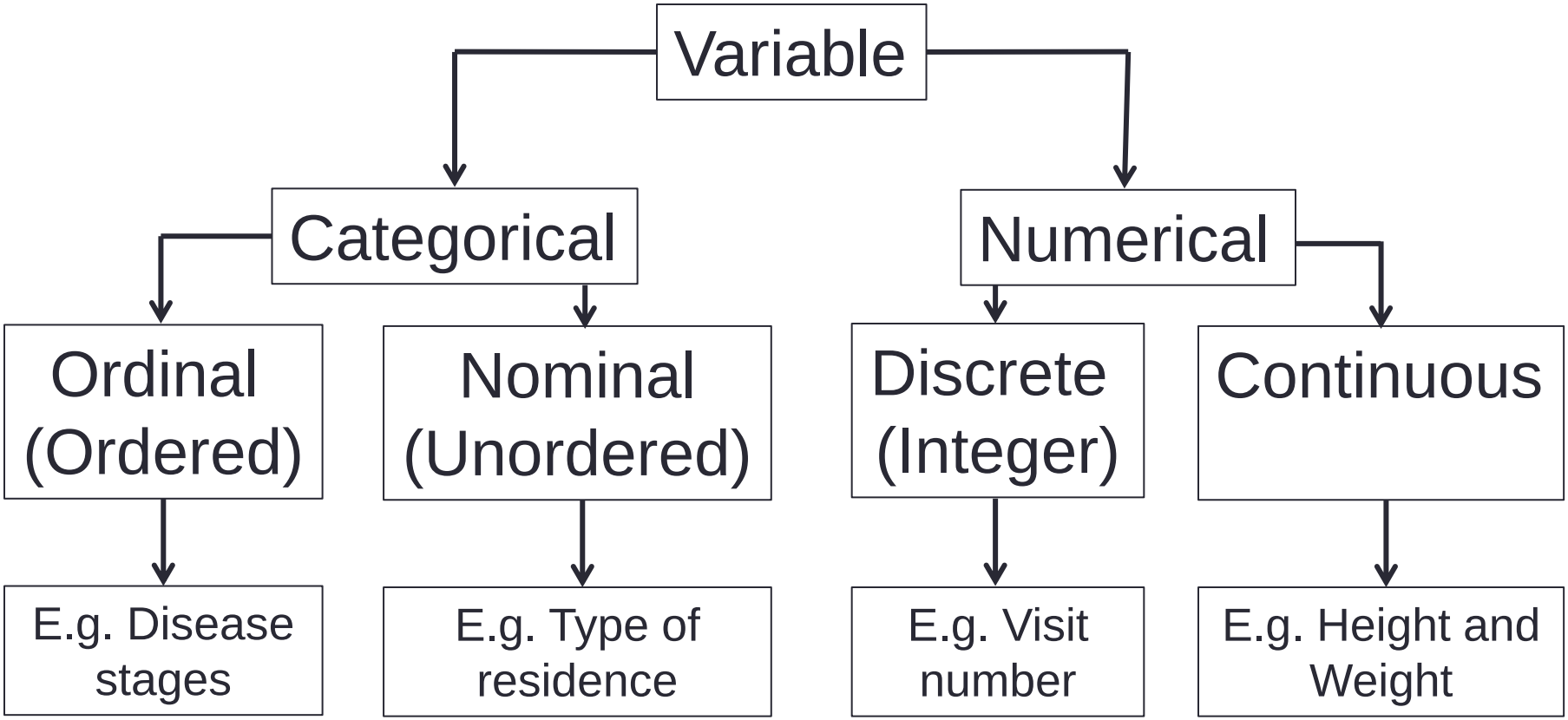
1. Data types
2. Presenting data
3. Summarising data
   - Mean, Median and Mode
   - Standard deviation and variance
   - Interquartile range
4. Normal / Non – normal distributions
5. Box plots
6. Confidence intervals

1. Data types

- A variable is a quantity that can take various values for different individuals

- Categorical: individual belongs to a distinct category
- Numerical: when the values are numeric, either discrete or continuous

1. Data types


Variable
Categorical
Numerical
Ordinal (Ordered)
Nominal (Unordered)
Discrete (Integer)
Continuous
E.g. Disease stages
E.g. Type of residence
E.g. Visit number
E.g. Height and Weight

2. Presenting data

- Categorical data

- Graphical: Bar Chart
- Summary measure: Proportion
- Comparing groups: Chi-Squared

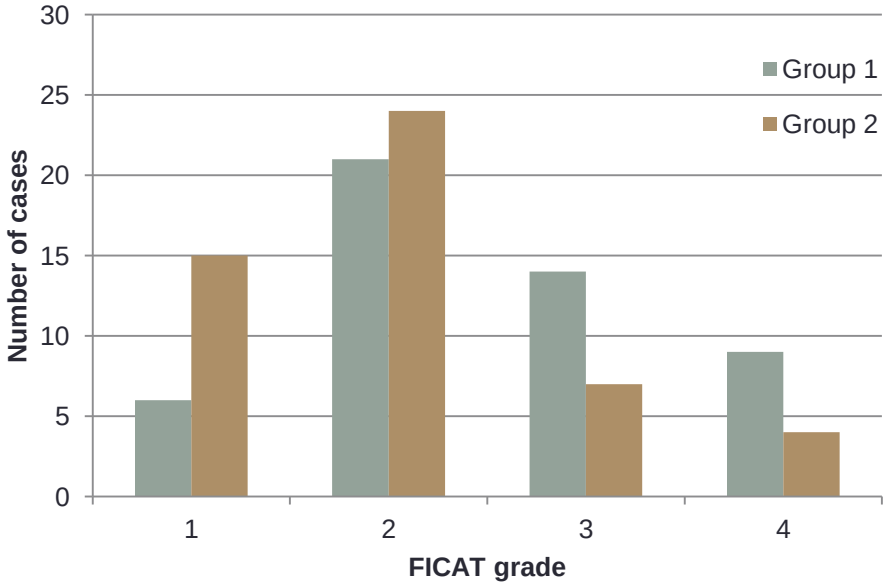
2. Presenting data

- Categorical data
- E.g. Two groups each with 50 patients comparing Ficat grading for AVN (1-4)

- 6/50 grade 1 = 12%
- 13/50 grade 2 = 30%

- Chi-squared test

  p=0.04



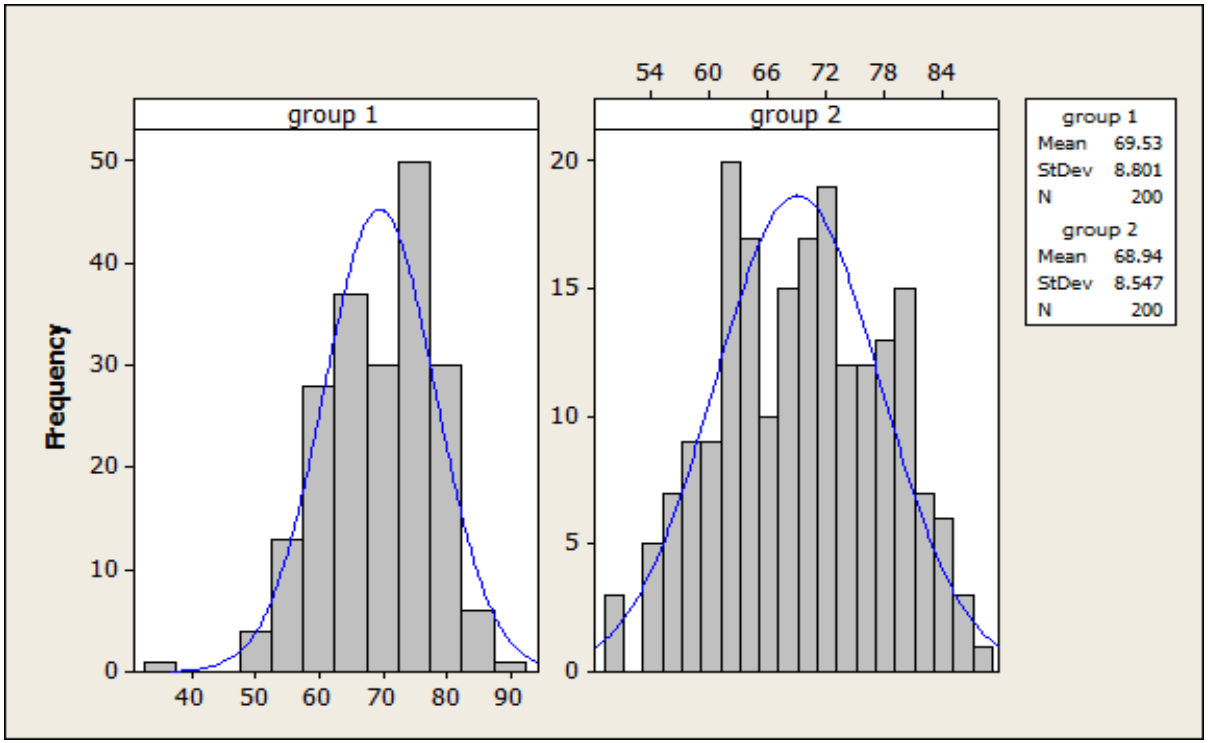
2. Presenting data

- Numerical data

- Graphical: Histogram
- Summary measure: Mean / Median
- Comparing groups: t-test or non-parametric equivalent

2. Presenting data

- Numerical data
- E.g. Two groups each with 200 patients comparing age



T-test:
p=0.50

3. Summarizing data

- Mean
- Median
- Mode
- Standard deviation and variance
- Interquartile Range

3. Summarizing data

- Mean
  - The sum of a collection of numbers divided by the number of numbers in that collection

$$A = \frac{1}{n} \sum_{i=1}^{n} a_i$$

- Median
- Mode
- Standard deviation and variance
- Interquartile Range

3. Summarizing data

- Mean
- Median
  - The numerical value separating the higher half of the data sample from the lower half.
  - I.e. if data is arranged in order it is the middle value
- Mode
- Standard deviation and variance
- Interquartile Range

3. Summarizing data

- Mean
- Median
- Mode
  - The value that appear most often in a set of data
- Standard deviation and variance
- Interquartile Range

3. Summarizing data

- Mean
- Median
- Mode
- Standard deviation and variance
  - Measures of how much variation or dispersion exists from the mean or expected value. Standard deviation = $\sqrt{}$ Variance
  - Low S.D – Data closely clustered
  - High S.D – Data spread over a large range

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2}.$$

- Interquartile Range

3. Summarizing data

- Mean
- Median
- Mode
- Standard deviation and variance
- Interquartile Range
  - Midspread or Middle fifty
  - A measure of dispersion
  - If data is ordered and divided into quarters the IQR is the difference between the upper and lower quartiles

3. Summarizing data

- E.g.
- 12 Hb measurements

  9, 10, 10, 11, 11, 12, 12, 12, 13, 14, 14, 15

- Mean = 1/12 (9+10+10+……+15)

  = 1/12 (143)

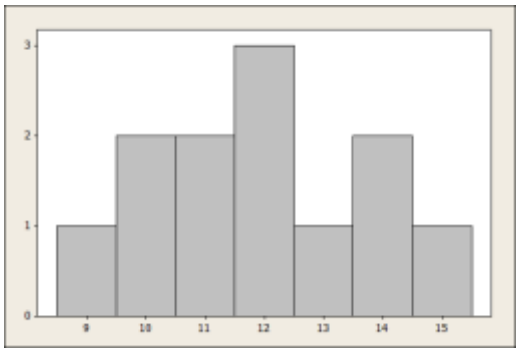
  = 11.9g/dl

3. Summarizing data

- E.g.
- Median = 12g/dl

  9, 10, 10, 11, 11, 12 | 12, 12, 13, 14, 14, 15

- Mode = 12g/dl



- IQR

  9, 10, 10 | 11, 11, 12 | 12, 12, 13 | 14, 14, 15

  = 13.75 – 10.25 = 3.5g/dl

3. Summarizing data

- E.g.
- Standard deviation

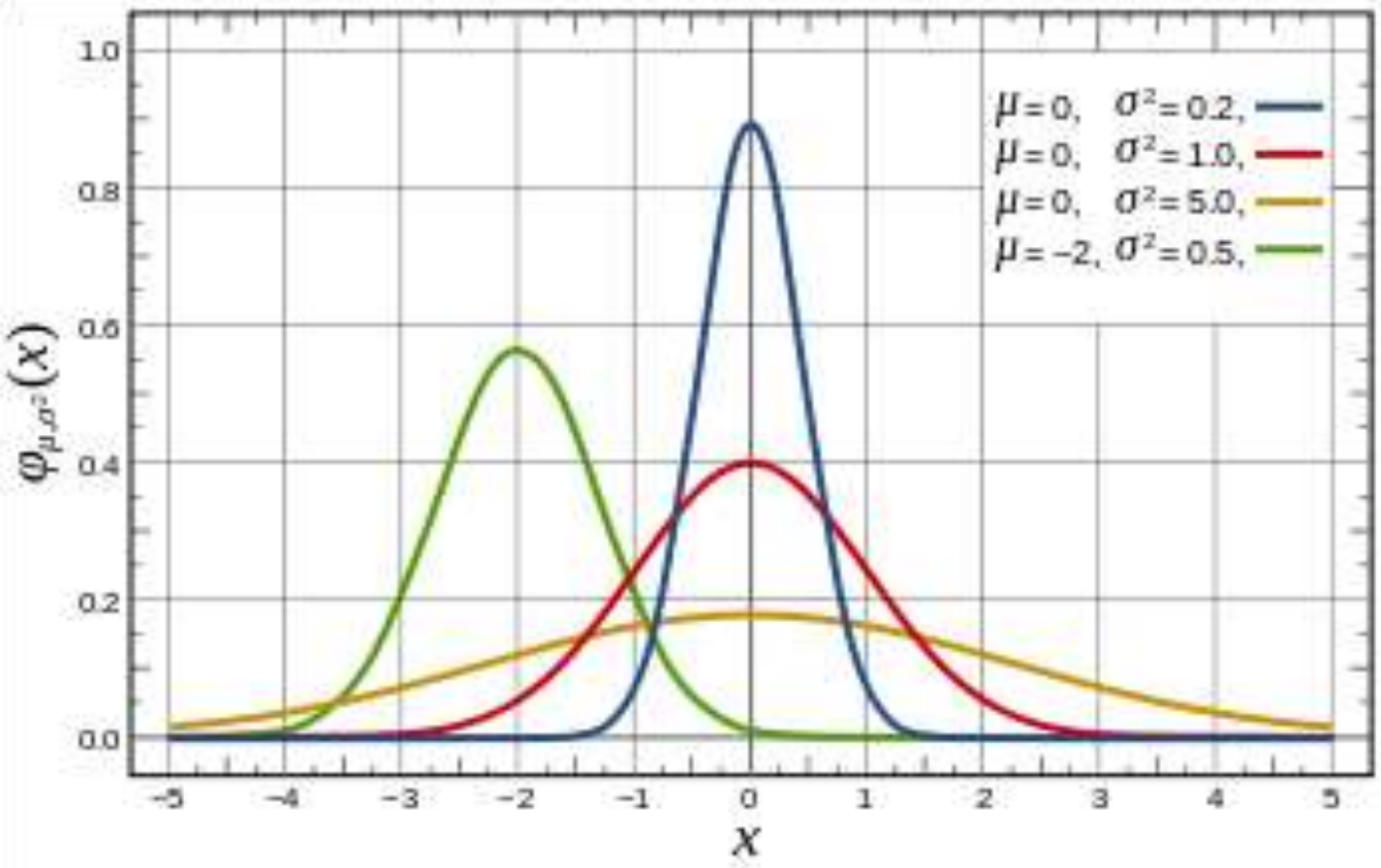
  9, 10, 10, 11, 11, 12, 12, 12, 13, 14, 14, 15

$= \sqrt{(1/11((9-11.9)^2 + (10-11.9)^2 + \ldots + (10-11.9)^2))}$

$= 1.83$

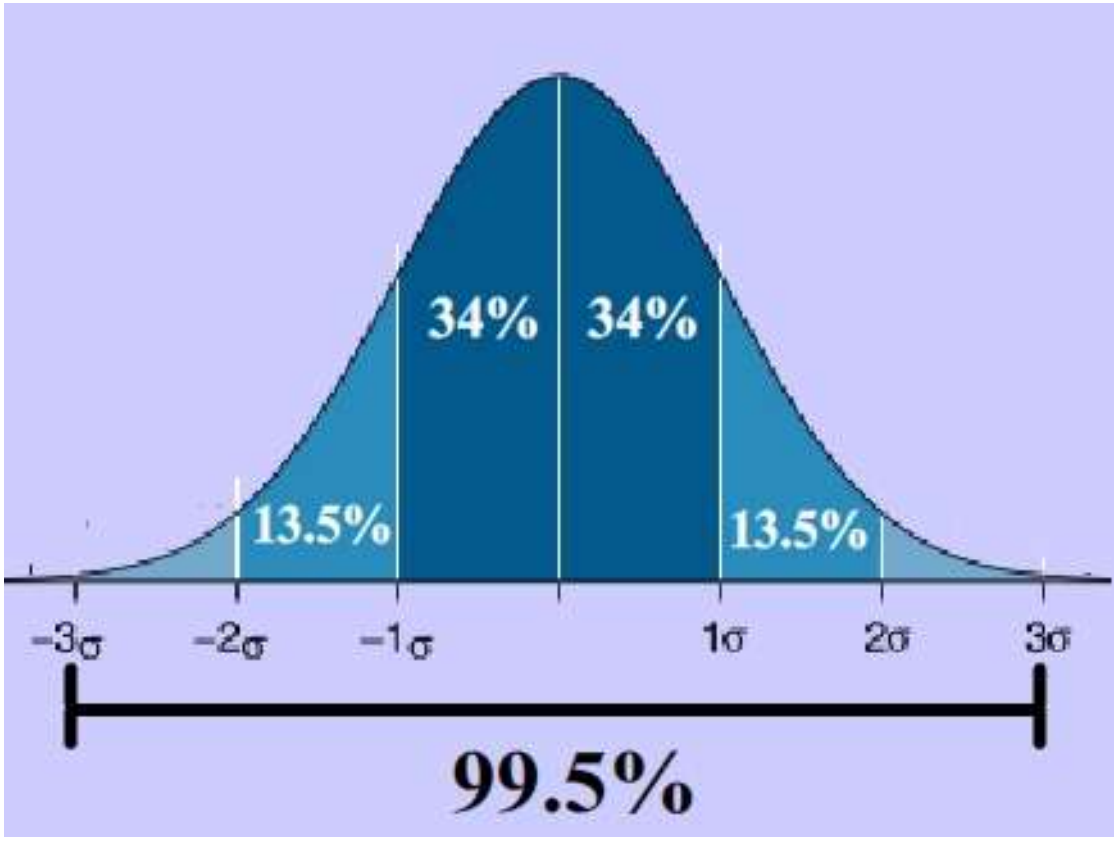
3. Summarizing data

- Relationship between mean and S.D



3. Summarizing data

- Normal “parametric” distribution



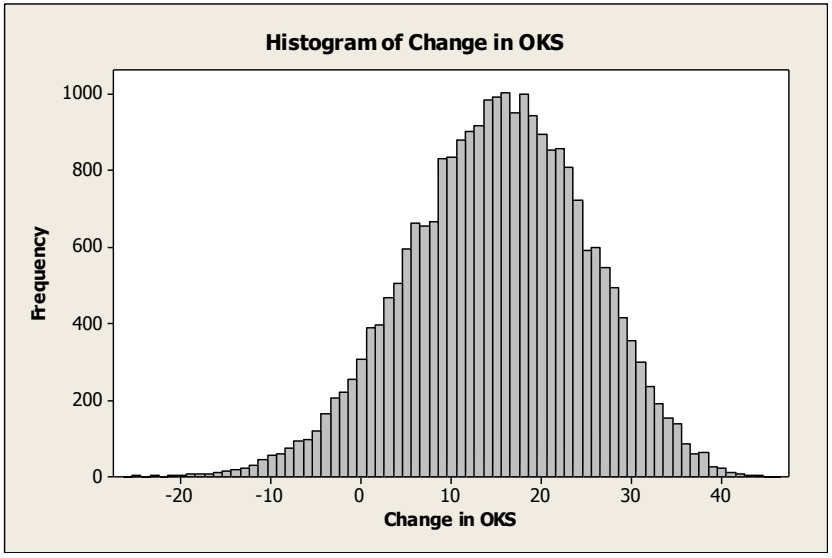
+/- 1SD covers
68% of data

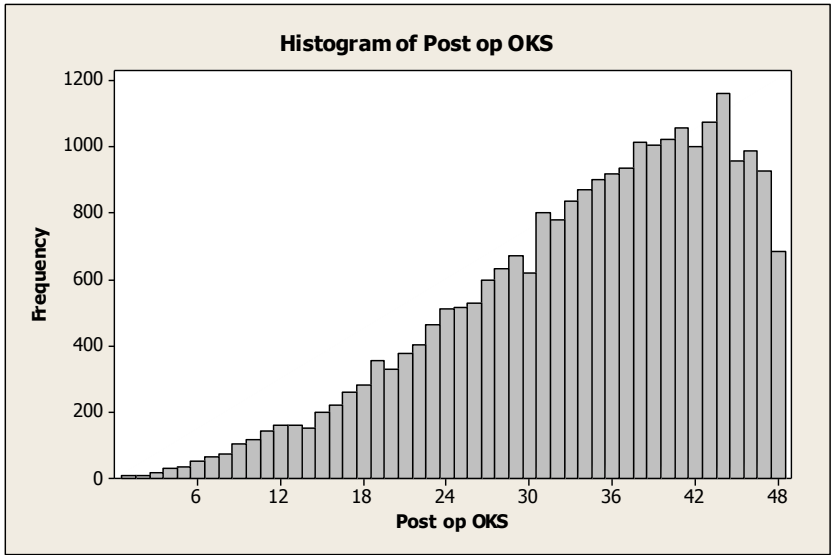
+/- 2 SD covers
95% of data

+/- 3 SD covers
>99% of data

4. Normal/Non–normal distributions

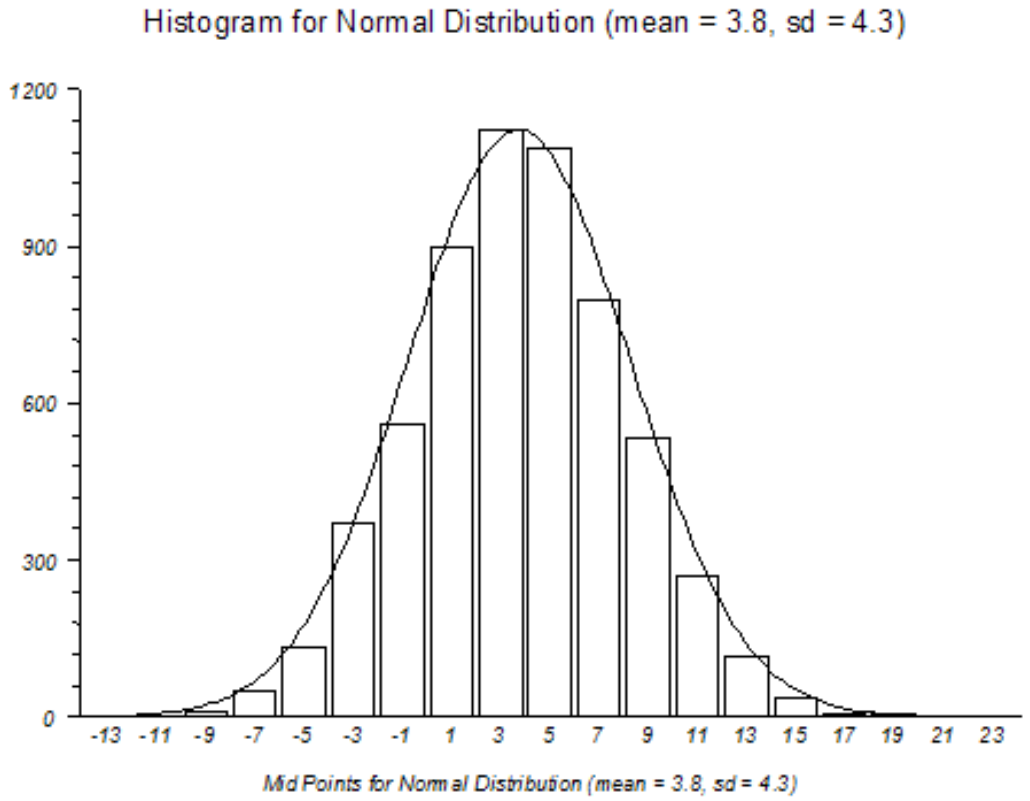
- Graphical summary is the first step in determining distribution type





4. Normal/Non–normal distributions

- Normal “parametric” distribution
- Gaussian

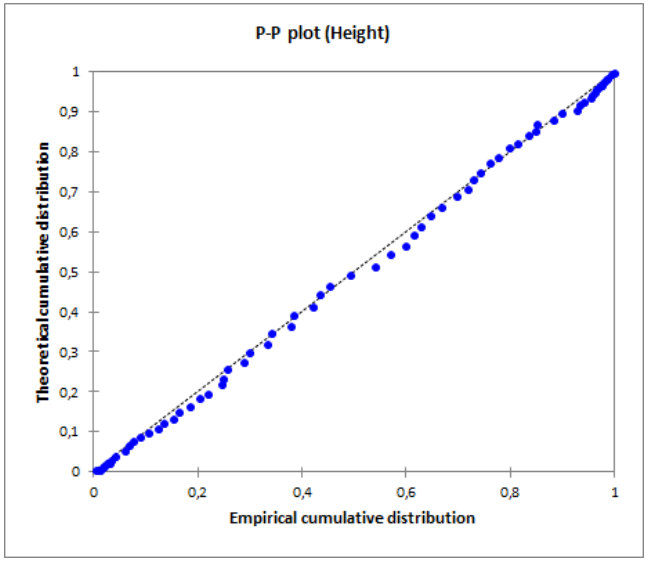


Summarised by the mean and Standard deviation

Median = Mean = Mode

4. Normal/Non–normal distributions

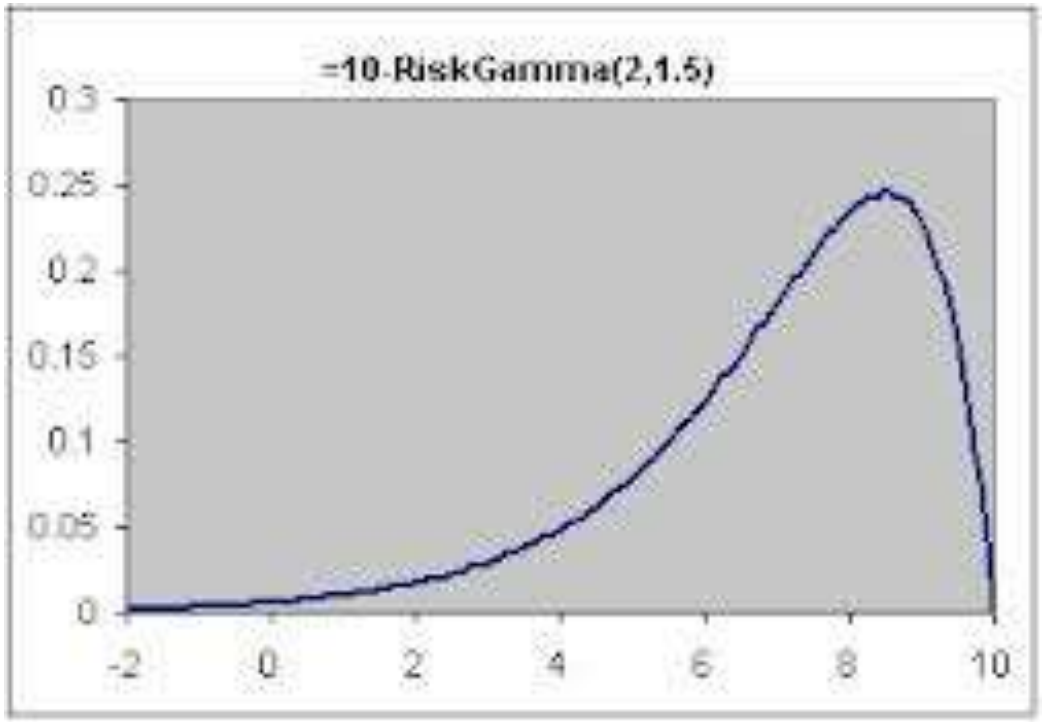
- Normal “parametric” distribution



- Testing for normality
  - Graphs
  - Mean = Median = Mode
  - Q-Q plots
  - Specific tests (eg Kolmogorov-Smirnov)

4. Normal/Non–normal distributions

- Non-Normal “non-parametric” distributions

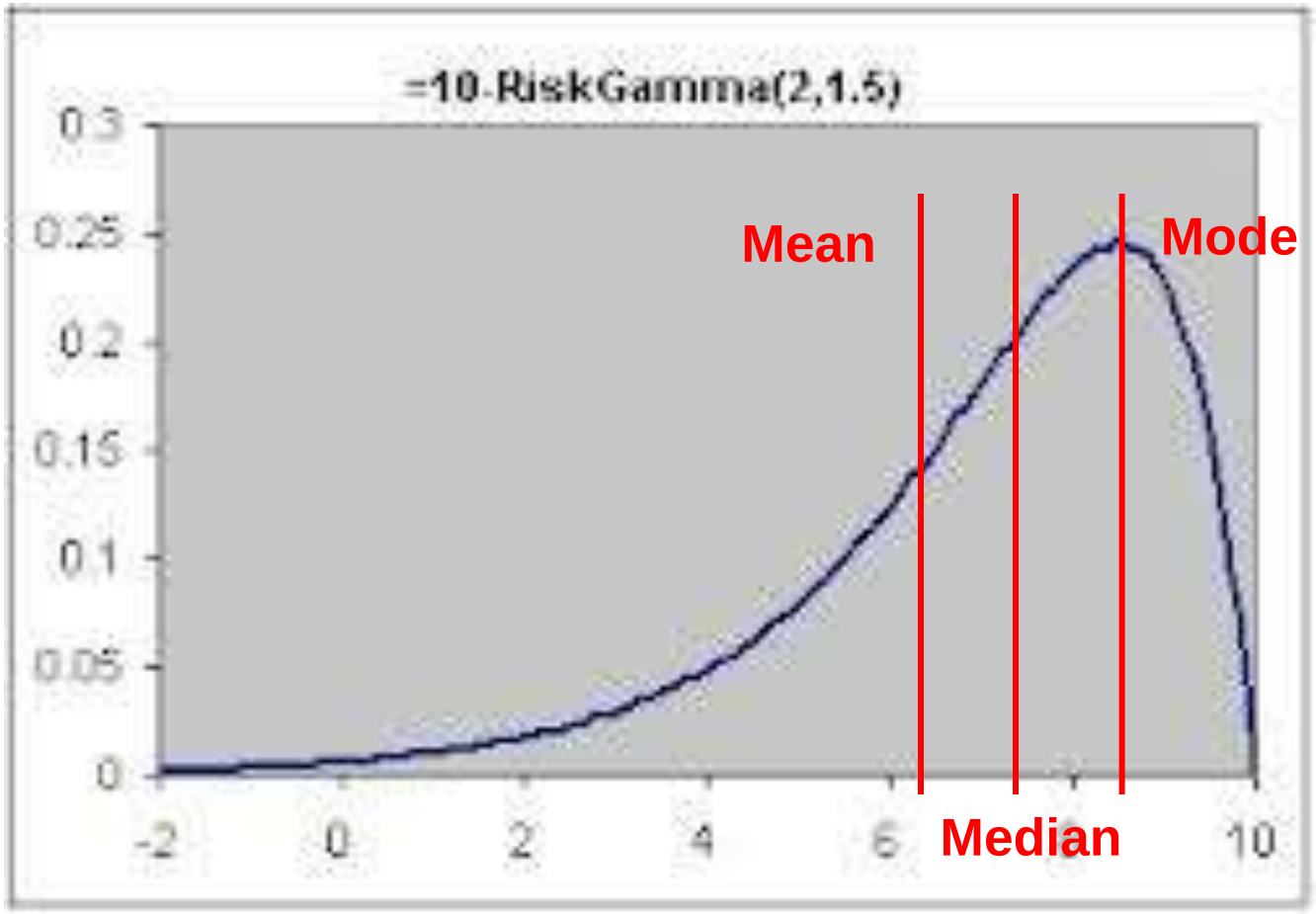


Summarised by the median and IQR

Median ≠ Mean ≠ Mode

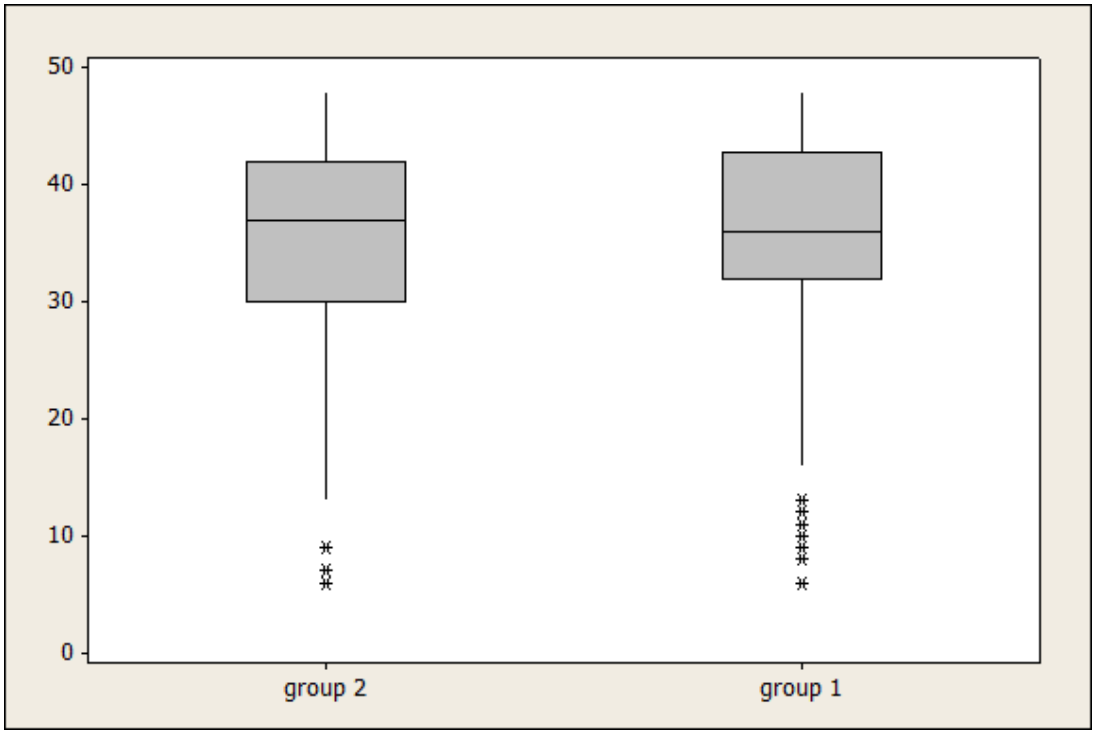
4. Normal/Non–normal distributions

- Non-Normal “non-parametric” distributions



5. Boxplots

- Good way of depicting non-normal data
- Also good for demonstrating “outliers”

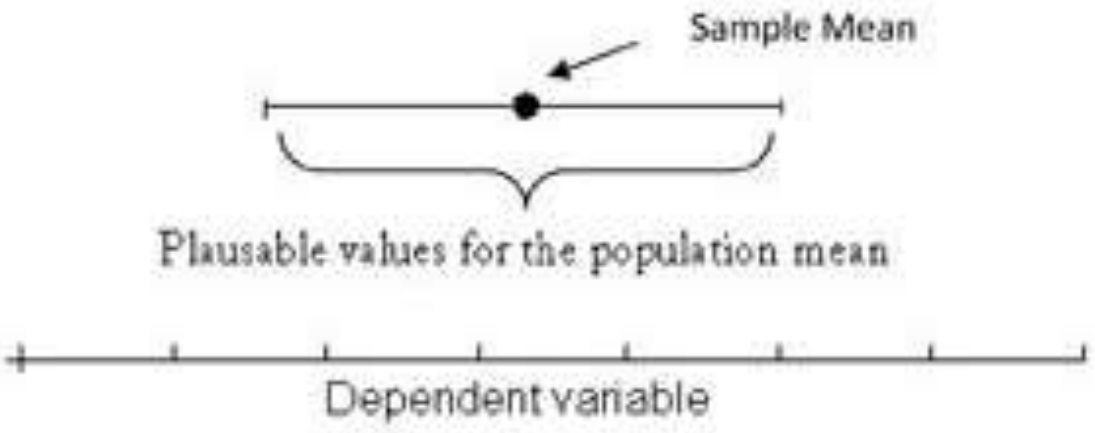


6. Confidence Intervals

- Interval estimate of a population parameter (e.g. mean, proportion, ratio etc)
- Its observed and therefore different from sample to sample
- When we take a sample we are trying to make inferences about the wider population

6. Confidence Intervals

- Interpretation:
- Plausible range (for the population mean)
  - 95% chance of capturing the population mean
- Repeated samples
  - If experiment was repeated then 95% of CI would include the population mean



6. Confidence Intervals

- Based on the Standard error of the mean

$$\mathrm{SE}_{\bar{x}} = \frac{s}{\sqrt{n}}$$

- 95% CI for Mean =

Mean +/- (SE × 1.96)

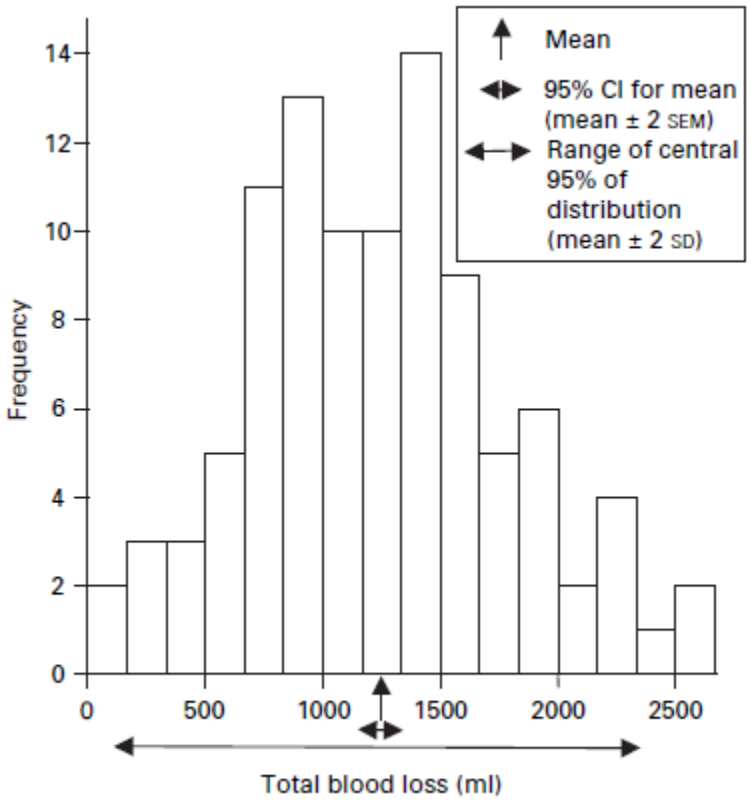


Fig. 5

Key statistical principles

1. Null hypothesis
2. P values and significance
3. Type I and Type II error
4. Power and sample size
5. Precision and accuracy

1. Null hypothesis

- Default position
- States that there is no relationship between two measured variables
- Rejecting or disproving the null hypothesis and demonstrating a measurable difference is one of the keys to modern science
- Alternative hypothesis

2. p values and significance

- p value is the estimated probability of rejecting the null hypothesis when the null hypothesis is true

- E.g. p=0.04 means that there is a 4% chance that the findings are due to random chance and not an effect of the intervention

2. p values and significance

- The significance level (alpha) refers to a pre-chosen probability that the investigator is happy to accept
- Conventionally either <0.05 or <0.01
- The p value is the probability calculated after the study
- When doing sample size calculations a smaller significance level means a larger sample size

3. Type 1 and Type 2 error

	H_0 True	H_0 False
Reject H_0	Type I Error	Correct Rejection
Fail to Reject H_0	Correct Decision	Type II Error

3. Type 1 and Type 2 error

Type 1:

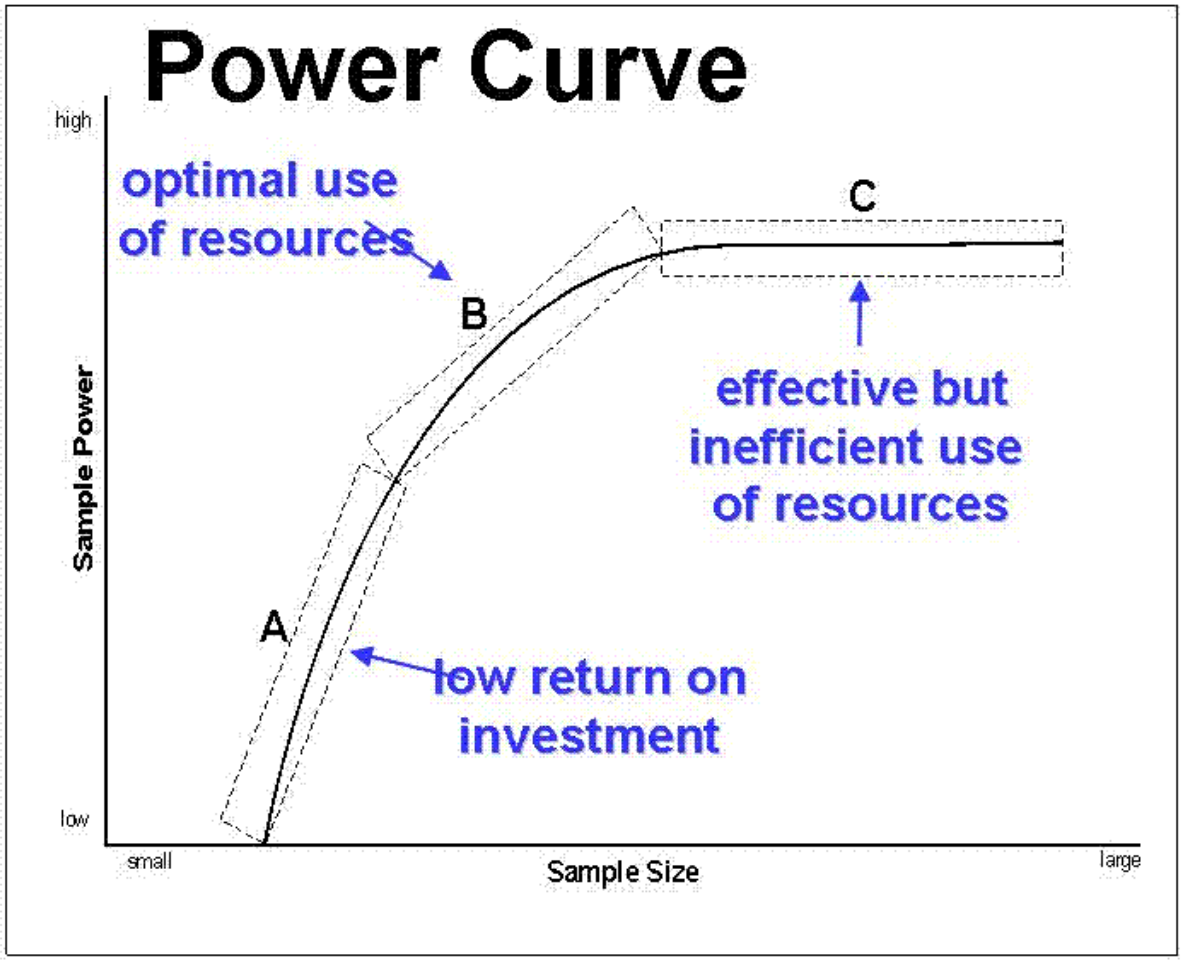
- Rejection of the null hypothesis when it is true
- False Positive
- Controlled by alpha – the significance level
- Alpha = 0.05 mean 5% of tests will result in type 1 error

3. Type 1 and Type 2 error

Type 2:

- Failure to reject the null hypothesis that is false
- False negative
- Controlled by beta – the power or sensitivity
- Usually set at 0.8 or 0.9 – ie if a difference exists then we will find it in 80-90% of occasions

4. Power and sample size

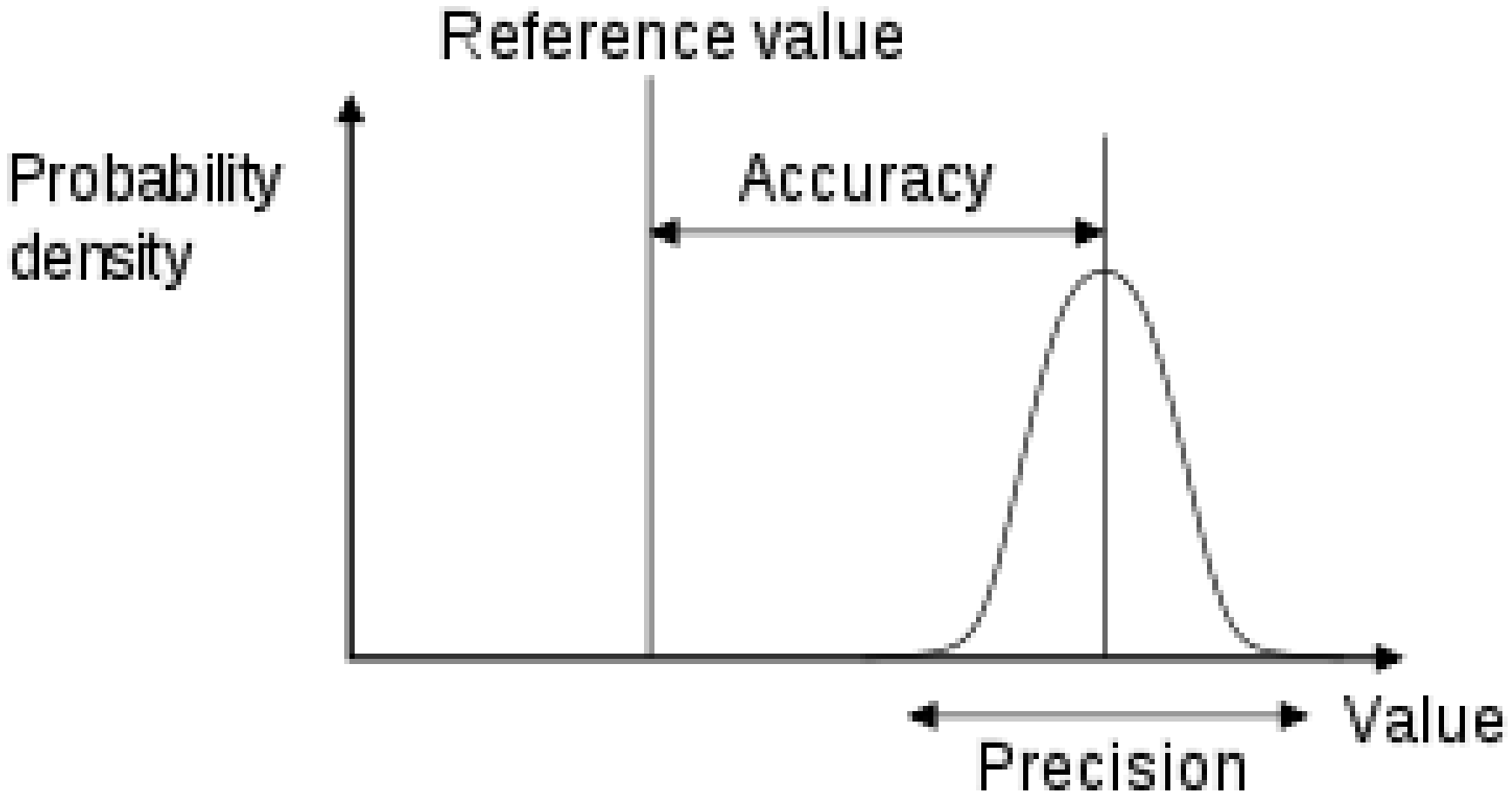


4. Power and sample size

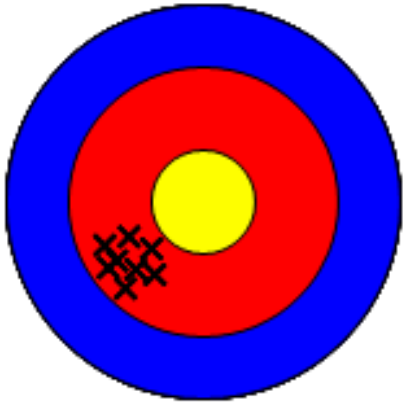
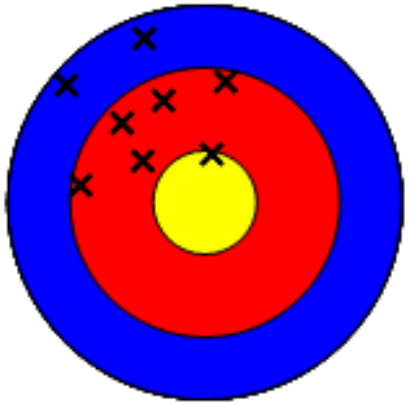
- Power calculations are designed to tell us how many patients are required to avoid a type 1 or 2 error

- Combined with other key information to determine sample size
  - Precision/variance of the sample
  - Clinically important difference
  - Type of statistical test we will be performing

5. Precision and accuracy



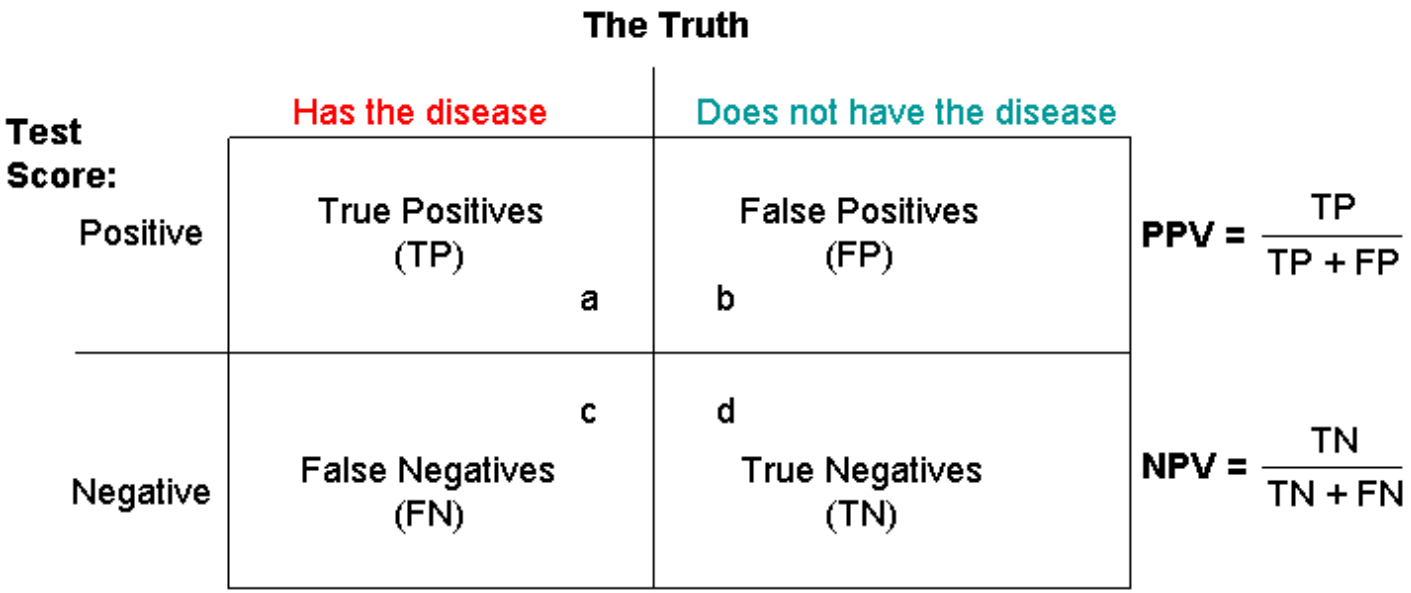
5. Precision and accuracy

	Accurate	Inaccurate (systematic error)
Precise		
Imprecise (reproducibility error)		

Sensitivity and Specificity

	Sensitivity	Specificity
Definition	Proportion of patients with a disease who test positive	Proportion of patients without the disease who test negative
100% (1.0) Means	The test correctly identify every person who has the target disorder	The test correctly identify every person who does not have the target disorder

Sensitivity and Specificity



Sensitivity

$$\frac{TP}{TP + FN}$$

Or,

$$\frac{a}{a + c}$$

Specificity

$$\frac{TN}{TN + FP}$$

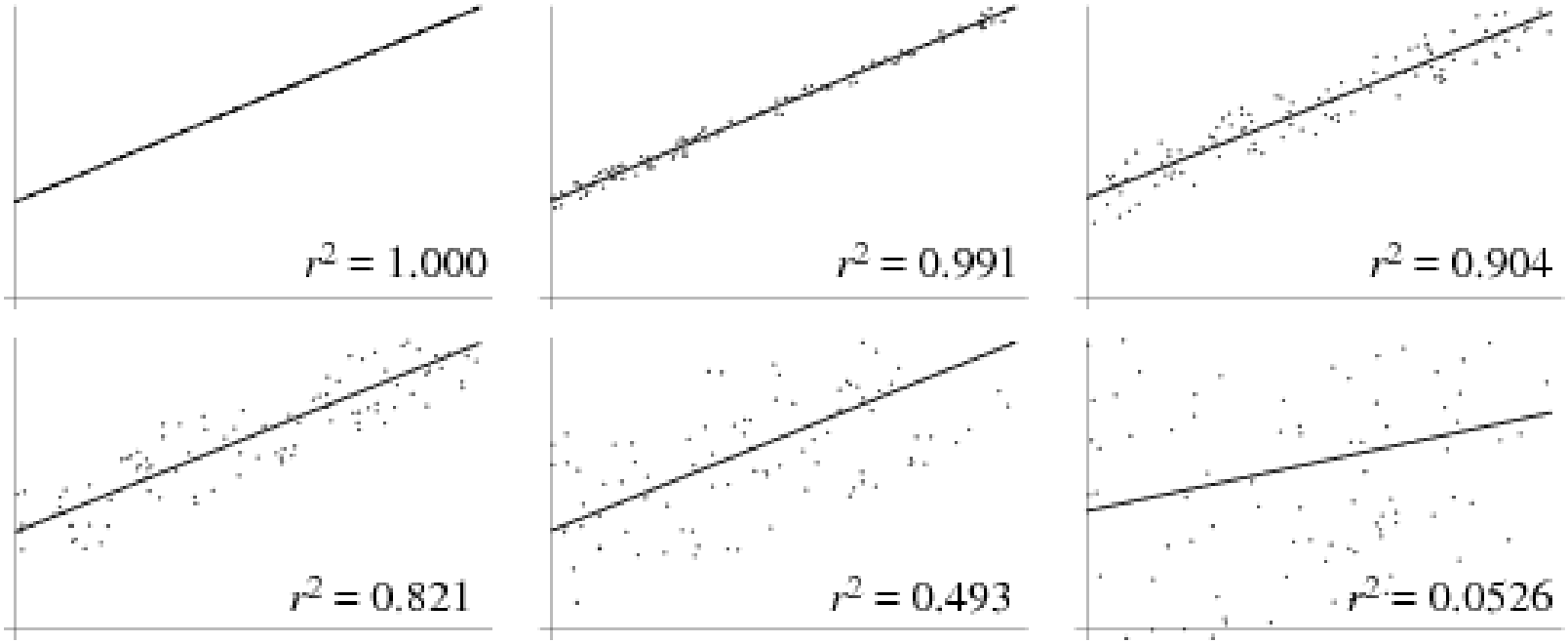
$$\frac{d}{d + b}$$

Statistical tests

1. Correlation co-efficients
2. Regression
3. Parametric and non- parametric tests

1. Correlation co-efficients

- Measure of the strength and direction of the linear relationship between two variables



2. Regression

- Not the same as correlation
- Correlation quantifies the degree to which two variables are related
  - Usually when both variables are known
- Regression finds the best line that predicts X from Y
  - X = response variable / Y = explanatory variable
  - Looking for cause and effect

3. Parametric/non-parametric tests

Data type	One group	Two groups	More than two groups
Binary variable Eg complication Y/N	Sign test	**- Independent** Chi-Squared or Fishers **- - Paired** McNemars or Binomial	**- Independent** Chi-squared **- Paired** Q test or Exact test
Numerical Eg Height	***Parametric:*** One sample t-test ***Non parametric:*** Sign test	**- Independent** ***Parametric:*** Independent t-test ***Non parametric:*** Wilcoxon or Mann Whitney U **- Paired** ***Parametric:*** Paired t-test ***Non parametric:*** Sign or Wilcoxon test	**- Independent** ***Parametric:*** ANOVA ***Non parametric:*** Kruskal-Wallis **- Paired** ***Parametric:*** Repeated measures ANOVA ***Non parametric:*** Friedman two way ANOVA

Survival analysis

1. Life table survival analysis
2. Kaplan Meier
3. Cox regression

Survival Analysis

- Survival data
  - Plotted over time
  - Allows for variable dates of entry
  - Variable length of follow up

- Care must be given to the definition of failure and how loss to FU is handled

1. Life Table Survival analysis

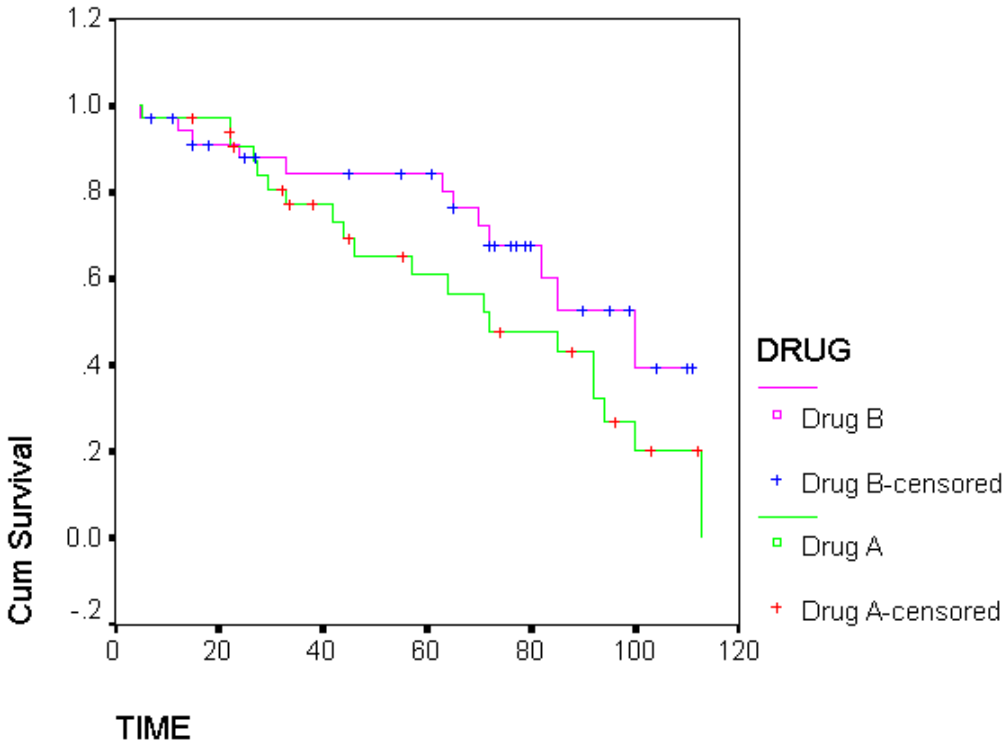
- Survival is typically calculated annually
- Multiple failures between time points

Table 2. Life table analysis for the uncemented ABG I prosthesis with revision of the acetabular component as the end point

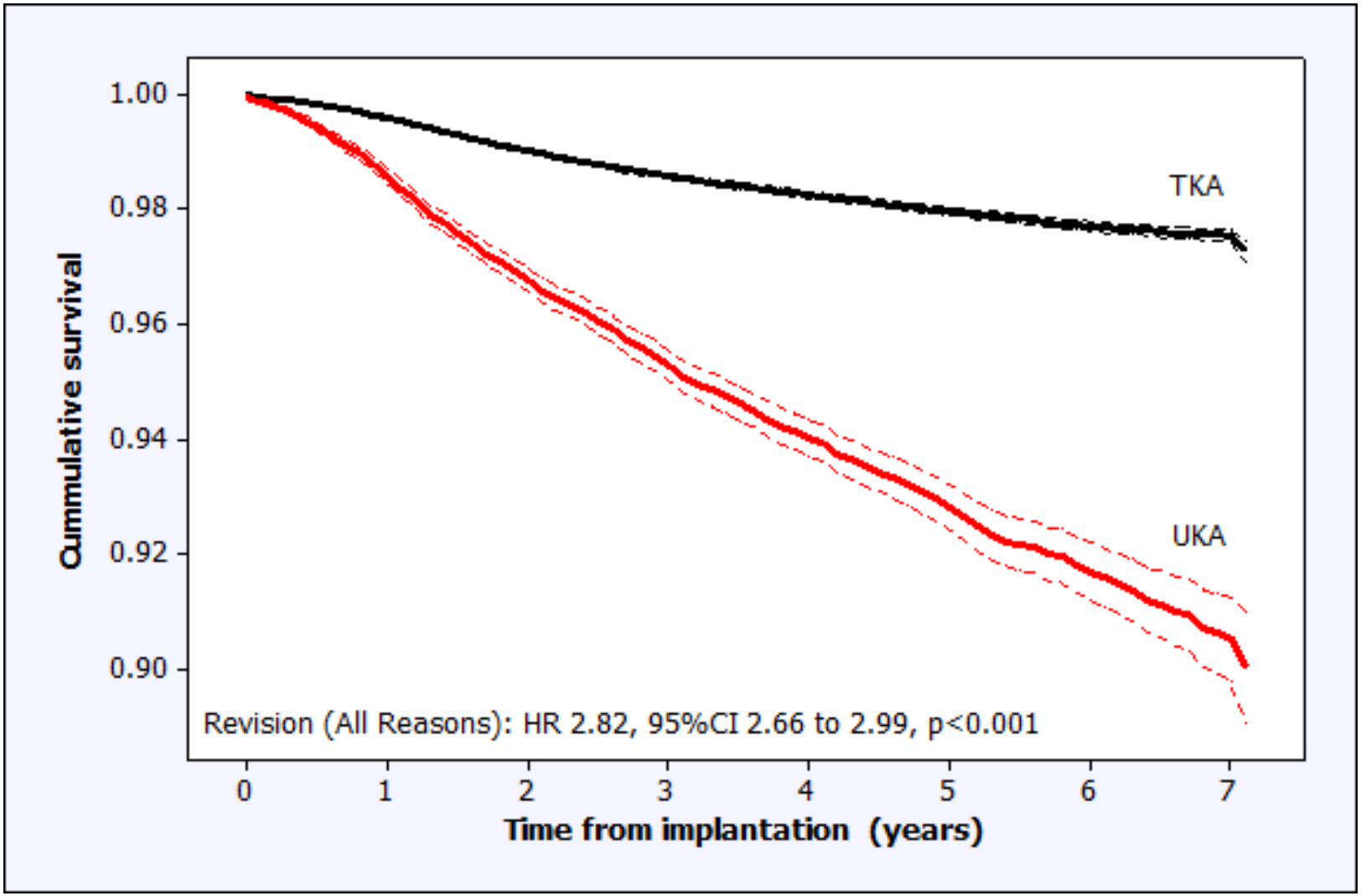
Years since operation	Number at start	Number revised	Withdrawn	Death	Loss to followup	Number at risk	Annual failure rate (%)	Annual success rate (%)	Survival rate (%)	95% confidence interval
0–1	69	0	0	0	0	69	0	100.0	100.0	94.7–100.0
1–2	69	0	0	0	0	69	0	100.0	100.0	94.7–100.0
2–3	69	0	0	0	0	69	0	100.0	100.0	94.7–100.0
3–4	69	0	0	0	0	69	0	100.0	100.0	94.7–100.0
4–5	69	0	0	0	0	69	0	100.0	100.0	94.7–100.0
5–6	69	0	0	0	0	69	0	100.0	100.0	94.7–100.0
6–7	69	0	0	0	0	69	0	100.0	100.0	94.7–100.0
7–8	69	0	0	0	0	69	0	100.0	100.0	94.7–100.0
8–9	69	0	0	0	0	69	0	100.0	100.0	94.7–100.0
9–10	69	0	0	0	0	69	0	100.0	100.0	94.7–100.0
10–11	69	4	0	0	0	69	5.8	94.2	94.2	85.8–98.1
11–12	65	0	0	0	0	65	0	100.0	94.2	85.5–98.3
12–13	65	0	0	0	0	65	0	100.0	94.2	85.5–98.3
13–14	65	1	21	0	0	54.5	1.8	98.2	92.4	82.0–97.7
14–15	43	2	12	1	0	36.5	5.5	94.5	86.9	71.7–96.0
15–16	28	1	11	0	0	22.5	4.4	95.6	82.5	61.1–96.8
16–17	16	0	7	1	0	12	0	100.0	82.5	51.7–100.0

2. Kaplan Meier survival

- Survival calculated at times of failure
- Censored data represented by a tick

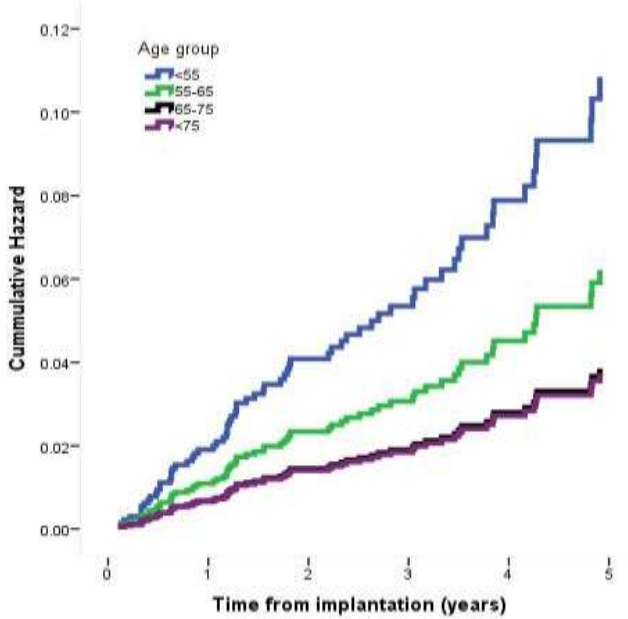


2. Kaplan Meier survival



3. Cox proportional hazards

- Assesses the effect of specific variables upon the hazard of failure
- Interpreted as a ratio relative to a baseline group



	Unadjusted (Univariate)			Adjusted (Multivariate)*		
	Hazard ratio	95.0% CI for HR	p value	Hazard ratio	95.0% CI for HR	p value
Total volume						
-1 to 25	2.72	2.12 to 3.50	<0.001	2.54	1.97 to 3.27	<0.001
-26 to 50	2.20	1.70 to 2.85	<0.001	2.11	1.63 to 2.72	<0.001
-51 to 100	1.61	1.24 to 2.09	<0.001	1.56	1.21 to 1.63	<0.001
-101 to 200	1.25	0.96 to 1.64	0.10	1.24	0.95 to 1.63	0.11
->200	Ref	Ref	-	Ref	Ref	-

Any Questions?

- Good Reference
  - A Petrie: Statistics in Orthopaedic papers
  - JBJS (Br) 2006;88(B):1121-36.