UNIVERSITY OF LIVERPOOL

The Royal Liverpool and Broadgreen University Hospitals NHS Trust

NHS

Methods and Statistics

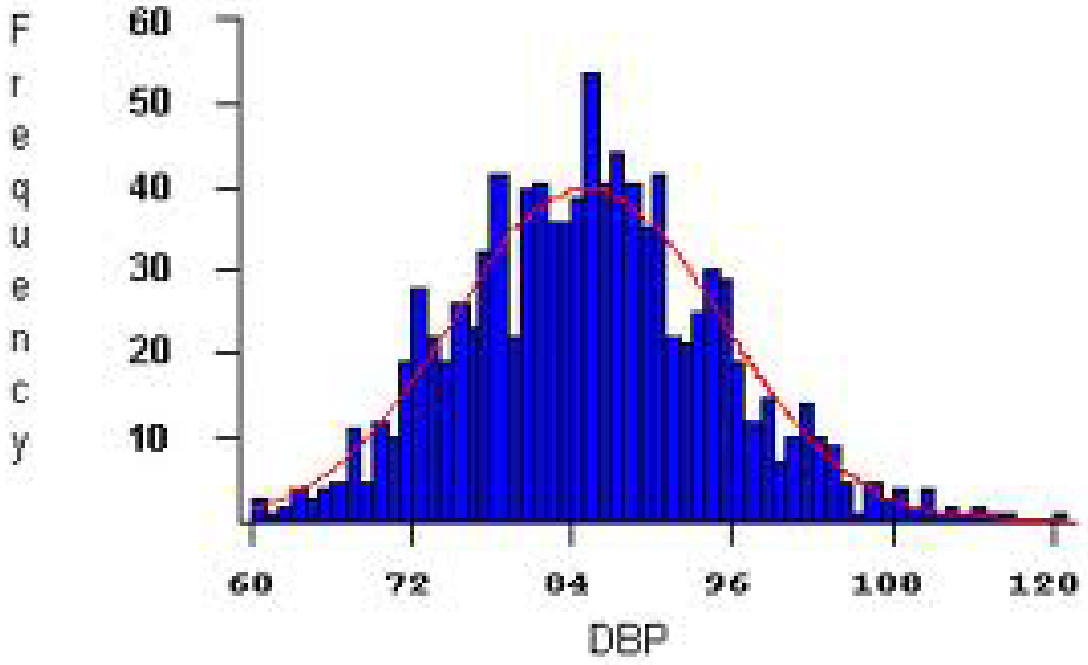
Professor Azzam Taktak

Consultant Clinical Scientist

Types of Data

- Scalar
  - Age, weight, height, BP
- Ordinal
  - BIS index, GCS, TNM stage
- Categorical
  - Sex, ethnic group, marital status

Histograms – Scalar Data



Blood pressure measurement can be reasonably approximated by a normal distribution with a mean of 85 mmHg and a standard deviation of 20 mmHg

Numerical Summaries

- Mean: The arithmetic average of the data values

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Median: the middle value of the data when sorted in ascending (or descending) order. If the number of samples is even then the median is halfway between the middle two values
- Mode: The most occurring value

Example

- Consider the following dataset:
  - {3, 4, 9, 1, 5, 3, 2, 3, 2, 4}

Example

- Consider the following dataset:
  - {3, 4, 9, 1, 5, 3, 2, 3, 2, 4}

- Mean = (3+4+9+1+5+3+2+3+2+4)/10 = 3.6

Example

- Consider the following dataset:
  - {3, 4, 9, 1, 5, 3, 2, 3, 2, 4}

- Mean = (3+4+9+1+5+3+2+3+2+4)/10 = 3.6

- {1, 2, 2, 3, 3, 3, 4, 4, 5, 9}

  Median = 3+3/2 = 3

Example

- Consider the following dataset:
  - {3, 4, 9, 1, 5, 3, 2, 3, 2, 4}

- Mean = (3+4+9+1+5+3+2+3+2+4)/10 = 3.6

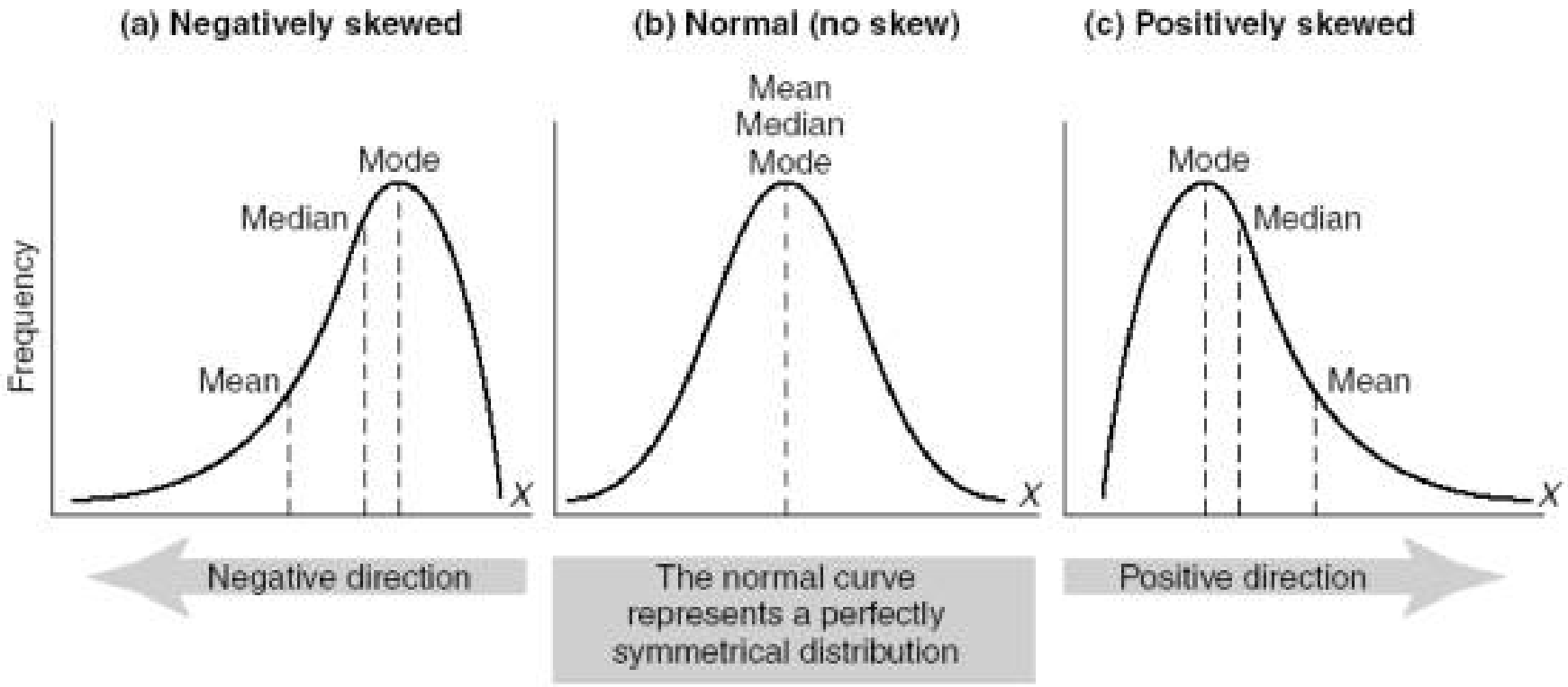
- {1, 2, 2, 3, 3, 3, 4, 4, 5, 9}

  Median = 3+3/2 = 3

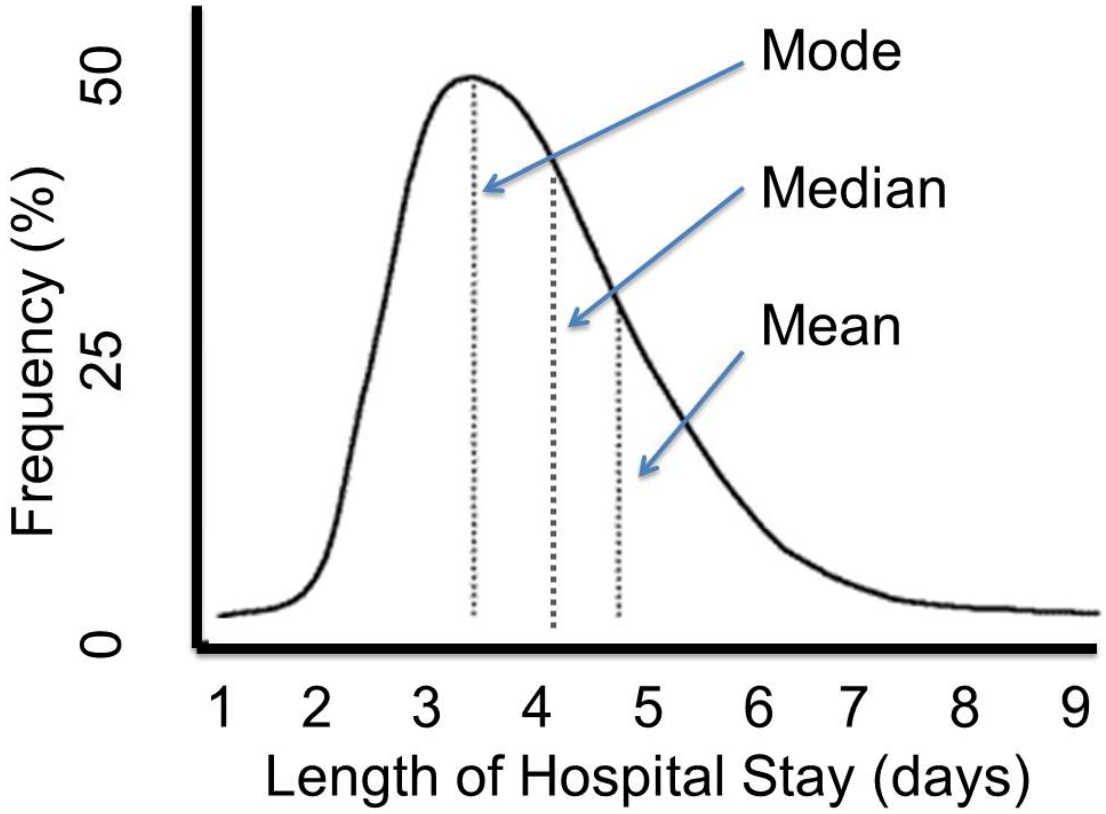
- {1, 2, 2, 3, 3, 3, 4, 4, 5, 9}

  Mode = 3

Skewness

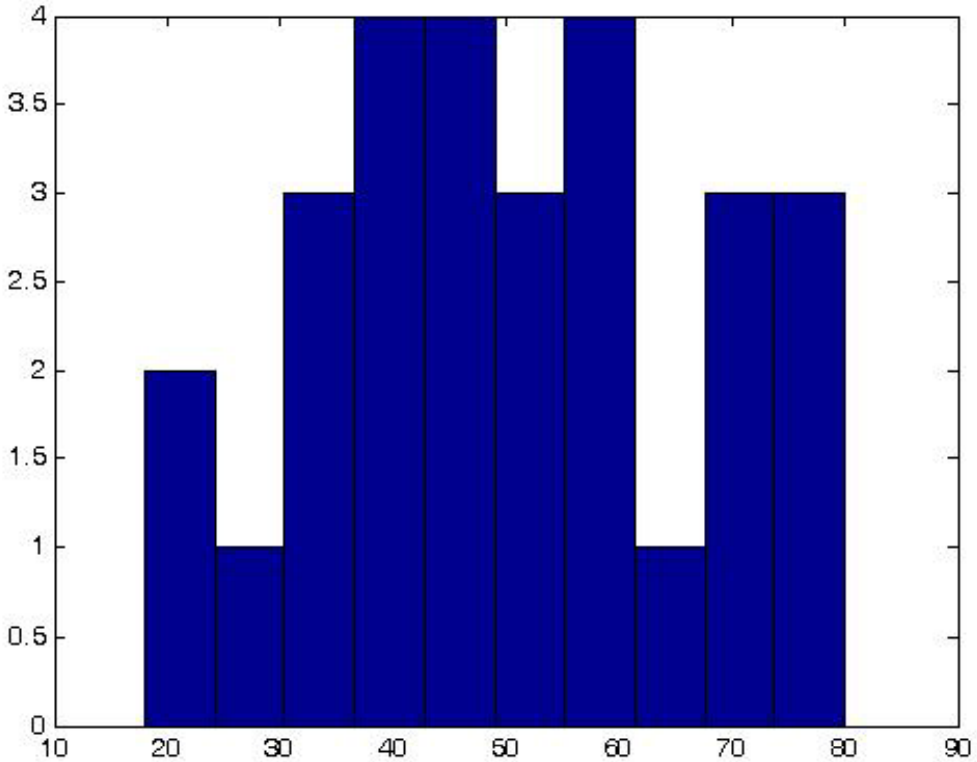

(a) Negatively skewed
(b) Normal (no skew)
(c) Positively skewed
Frequency
Mode
Median
Mean
Mean
Median
Mode
Mode
Median
Mean
X
X
X
Negative direction
The normal curve represents a perfectly symmetrical distribution
Positive direction

Skewed Distribution Example



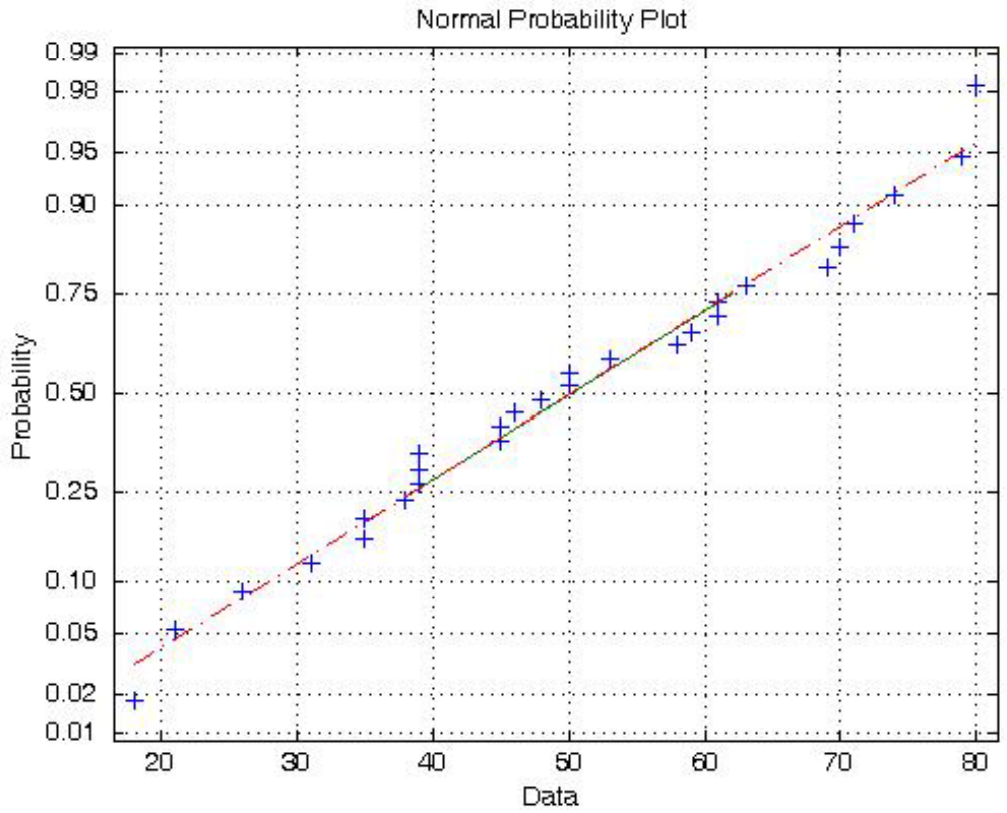
Assessing Normality

Histogram

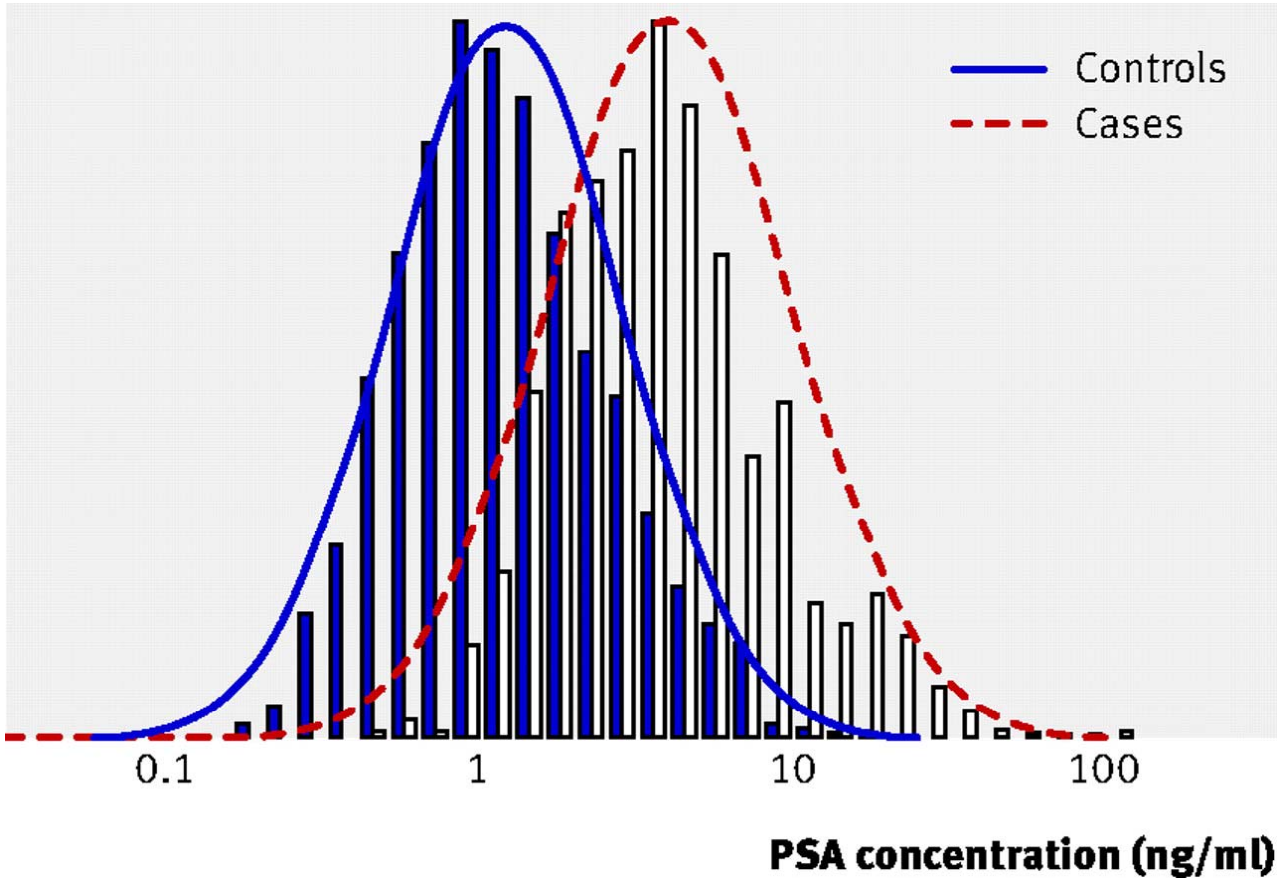


Assessing Normality

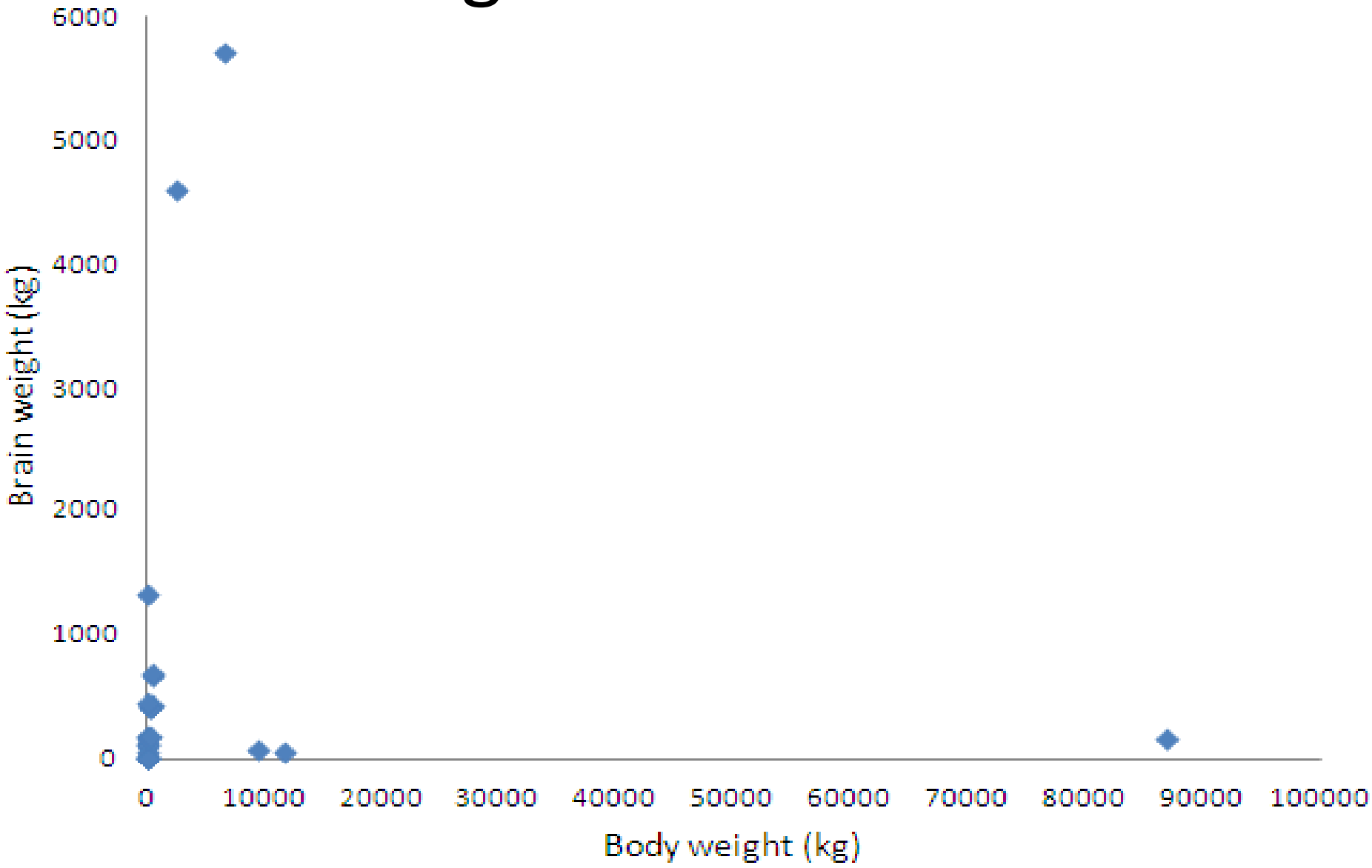
Normal (Q-Q) Plot



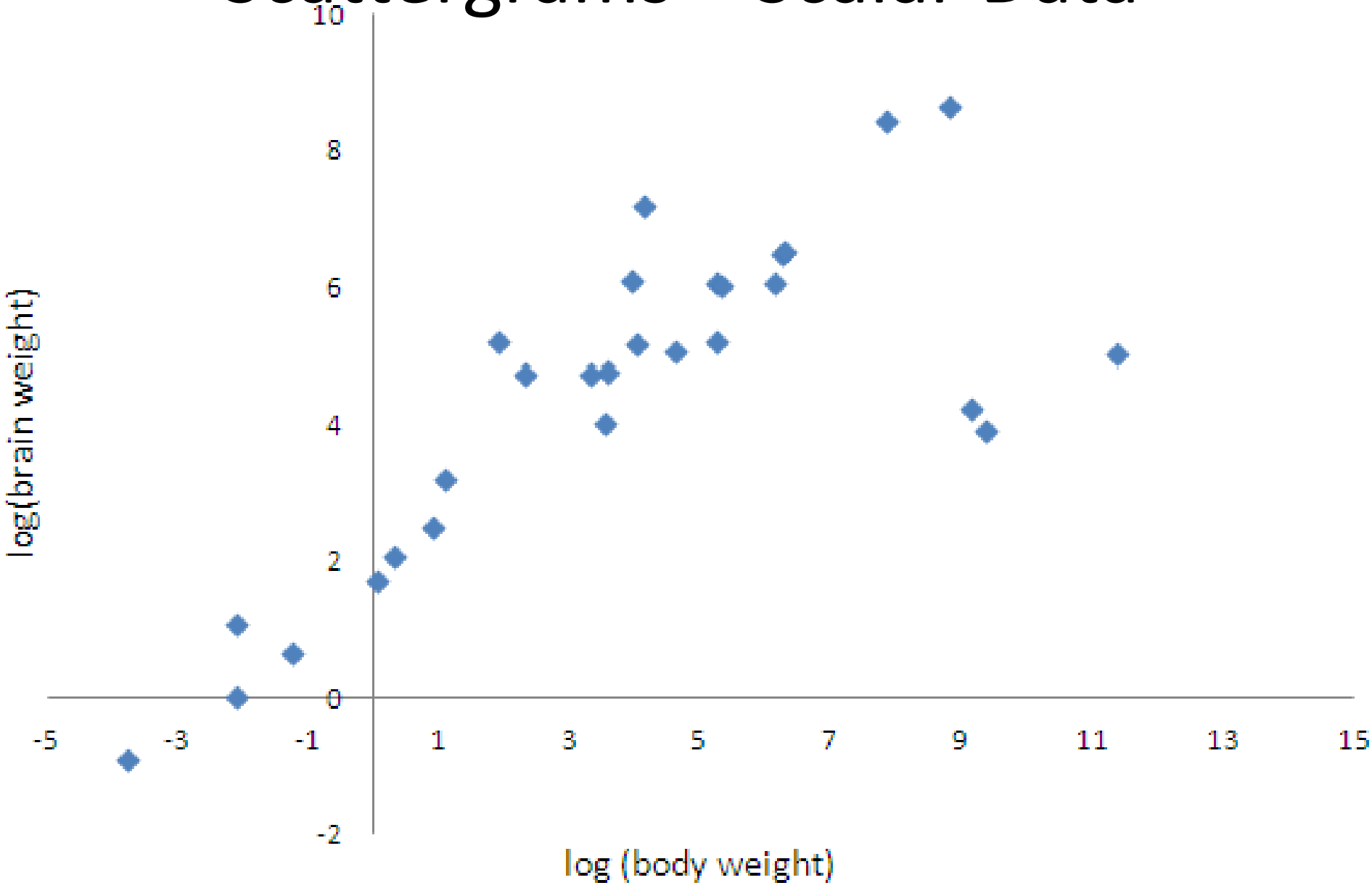
Transforming Data



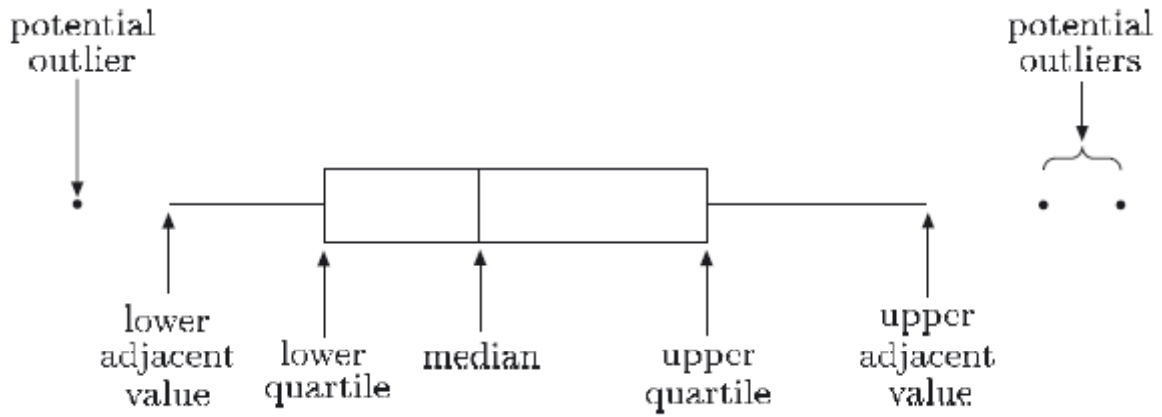
Scattergrams – Scalar Data



Scattergrams – Scalar Data



Box Plots – Ordinal Data



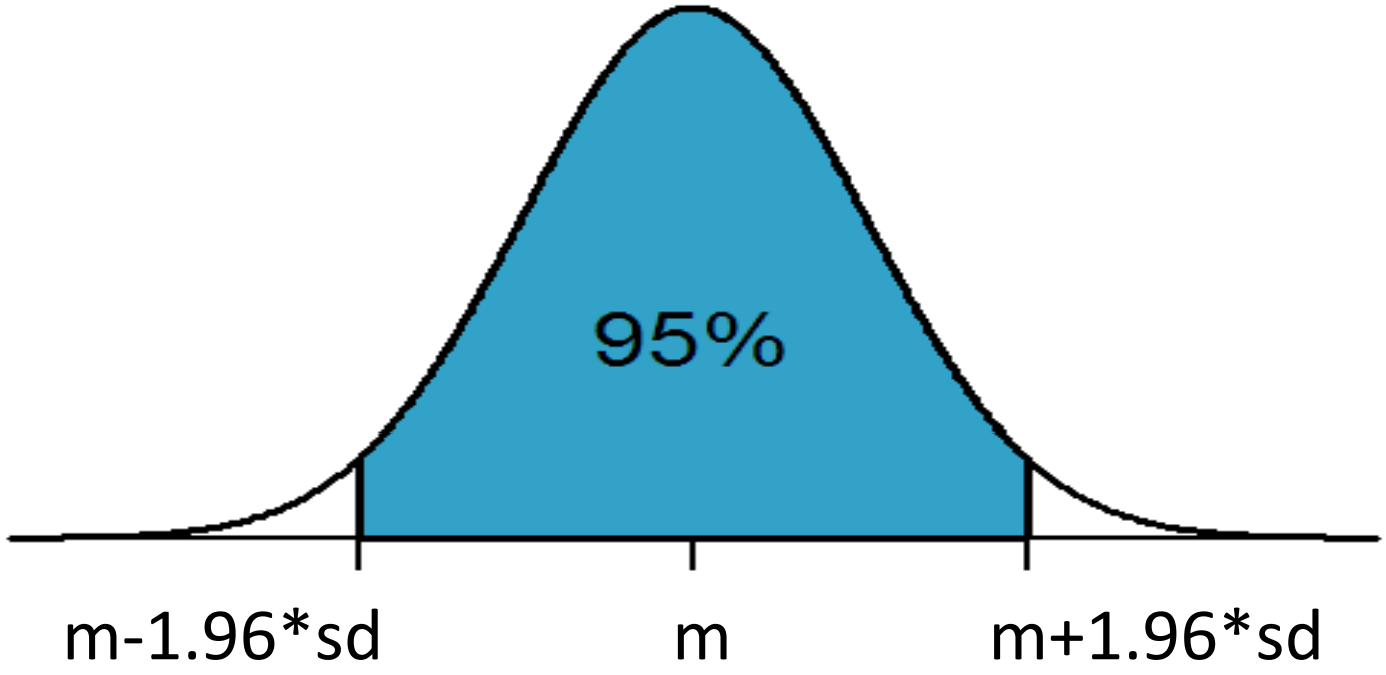
- The adjacent values are the furthest away from the median but still within 1.5 times the interquartile range

Standard Deviation

- The standard deviation is the measure of dispersion, or scatter, in the data.
- Take the following 2 sets of measurements:
  - s_1: {6, 7, 8, 7, 7, 9, 8, 9, 8, 7}
  - s_2: {2, 18, 10, 7, 5, 10, 12, 1, 3, 8}
- Both sets have a mean of 7.6. The second set however is much more disperse than the first.

- $sd = \sqrt{(\sum(y_i - y)^2/N)}$

Normally Distributed Data



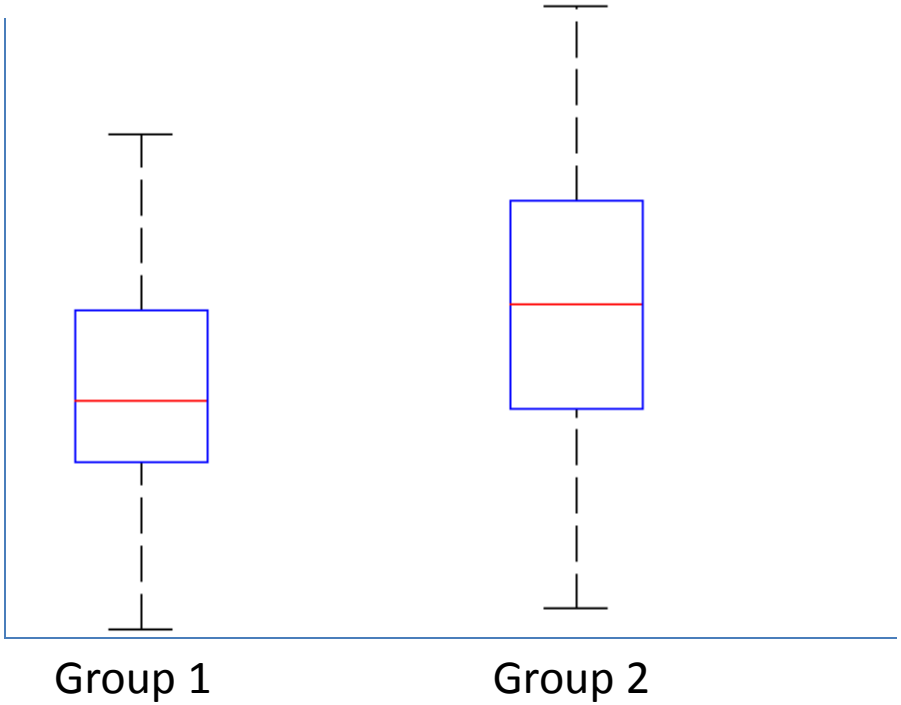
Approximately, 95% of the data lie within 1.96 of the standard deviation

Comparing Means

- Student's t-test can be used to compare means of 2 independent samples that can be reasonably modelled by a normal distribution.

Comparing Means

- For example, we want to compare birth weight of 2 groups of infants diagnosed with SIRDS, Group 1 died and Group 2 survived



Setting the hypothesis

- Often, we want to show differences in means between 2 groups. In statistics, we hypothesize that the difference is zero (the null hypothesis) and see if we can find evidence against this hypothesis.

Significance level

- If the null hypothesis H_0 is true then, in repeated experiments, H_0 will be rejected in some of the experiments, even though it is true. The significance level gives the proportion of the repeated experiments in which H_0 will be rejected falsely.
- Note that this statement refers to repeated experiments in which *the null hypothesis is true*

Type I and Type II errors

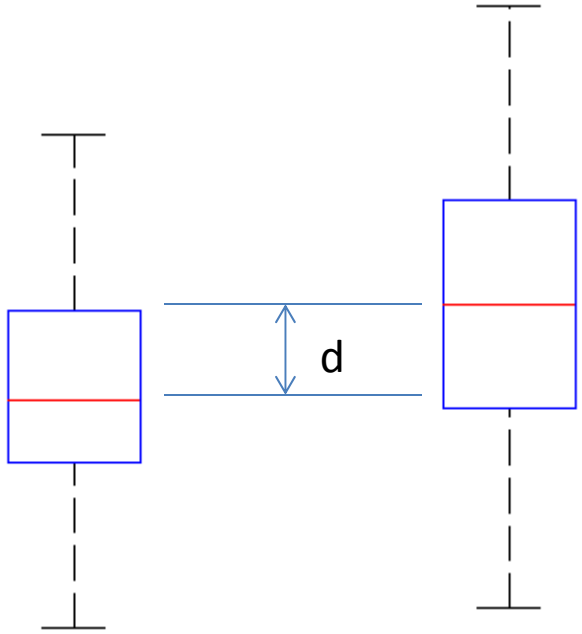
- **Type I error:** rejecting the null hypothesis when it is true. This is often referred to as α and is usually set to 0.05.
- **Type II error:** not rejecting the null hypothesis when it is false. Probability of avoiding it is the power of the test. This is often referred to as γ and is usually set to 0.8.
  - Note: α and γ are related. Setting 1-α too high might result in low γ and vice versa.

Comparing Means

- In the SIRDS example:

  Student's t-test:

  t = 2.25
  p = 0.014



Given a standard deviation *s* in the data, if the two groups were similar, there is a 1.4% probability of observing a difference of magnitude *d* or higher.

The P Value
Definition

The probability of having observed our data (or more extreme) given the null hypothesis is true

Interpreting the P value

The probability of obtaining the data given the null hypothesis is true

P value	Rough interpretation
$p > 0.1$	little evidence against H_0
$0.1 \geq p > 0.05$	weak evidence against H_0
$0.05 \geq p > 0.01$	moderate evidence against H_0
$p \leq 0.01$	strong evidence against H_0

Problems with P Values

- **Misinterpretation:** The P value is sometimes misinterpreted as the probability of the null hypothesis being correct or the probability that the observed effect is not real
- **Publication Bias:** Research findings with $p > 0.05$ sometimes do not get published
- **Over-Reliance:** Researchers sometimes change their conclusion radically depending on which side of 0.05 the P value is

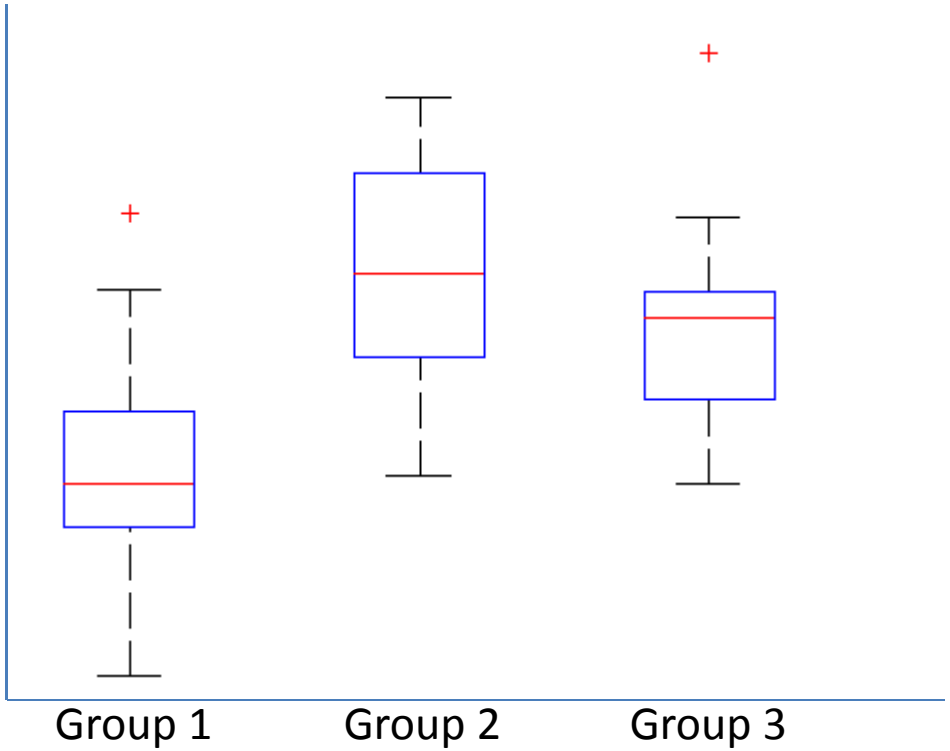
ANOVA

- If you have more than 2 groups and you wish to compare their means, you use a test called 1-way ANOVA (Analysis of Variance)

ANOVA

- For example, we wish to compare CO_2 measurements in 3 groups of children: Group 1 – Normals, Groups 2 – Hypoxia, Group 3 – Down Syndrome

f = 26.91
p < 0.001



Calculating Sample Size (Power Calculations)

- If you are collecting continuous scalar data (e.g. blood pressure, weight, height, etc.) and you wish to compare the difference in means, you need the following:
  - Underlying variation (standard deviation): from literature, pilot studies, models
  - Difference considered to be significant: from clinical experience
  - Power of the study: usually 0.8 or 0.9
  - Level of significance: usually 0.05 or 0.01

Calculating Sample Size (Power Calculations)

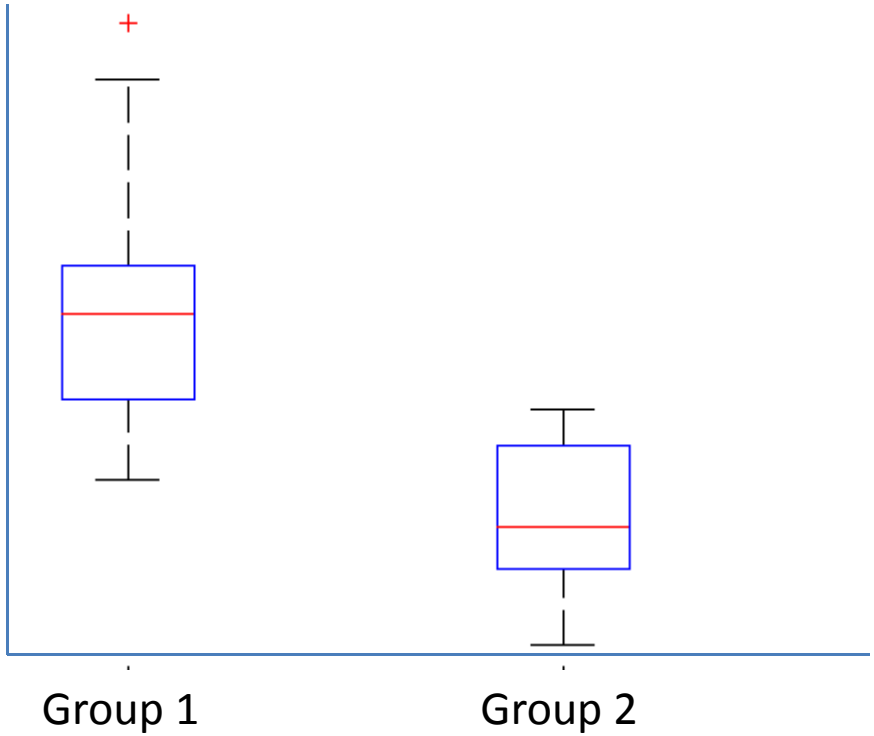
- If data represent proportions (e.g. proportion of males affected, proportion of patients developing metastases, etc.) and you wish to compare the difference in proportions, you need the following:
  - Proportion in the control group: from literature, pilot studies, models
  - Change in proportion that is considered to be significant: from clinical experience
  - Power of the study: usually 0.8 or 0.9
  - Level of significance: usually 0.05 or 0.01

Comparing Medians

- If normal distribution cannot be assumed for the data, we use a family of tests called non-parametric tests to compare medians rather than means.

Comparing Medians

- For example, we would like to compare level of arousal score using auditory-evoked response in two groups: Group 1 – awake subjects and Group 2 – anaesthetised subjects

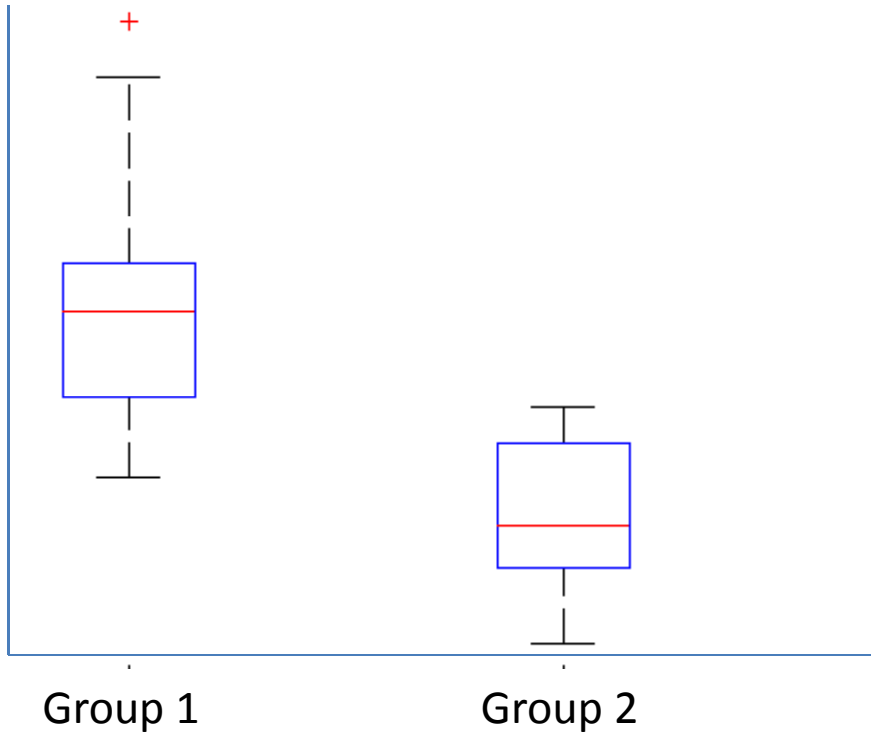


Comparing Medians

- Mann-Whitney Test

$z = 4.578$

$p < 0.001$



Differences in Categories

- For categorical variables, data is presented in cross tabulation format. For example, for a study looking at association between smoking and lung cancer, we might have the following data:

Exposure	Cancer	No Cancer	Total
Smoking	31	35	66
Non-smoking	8	62	70

$$RR = \frac{31/66}{8/70} = 4.11$$

Differences in Categories

- Chi-squared analysis are used to test for association between exposure to risk and outcome:

- For the smoking data:

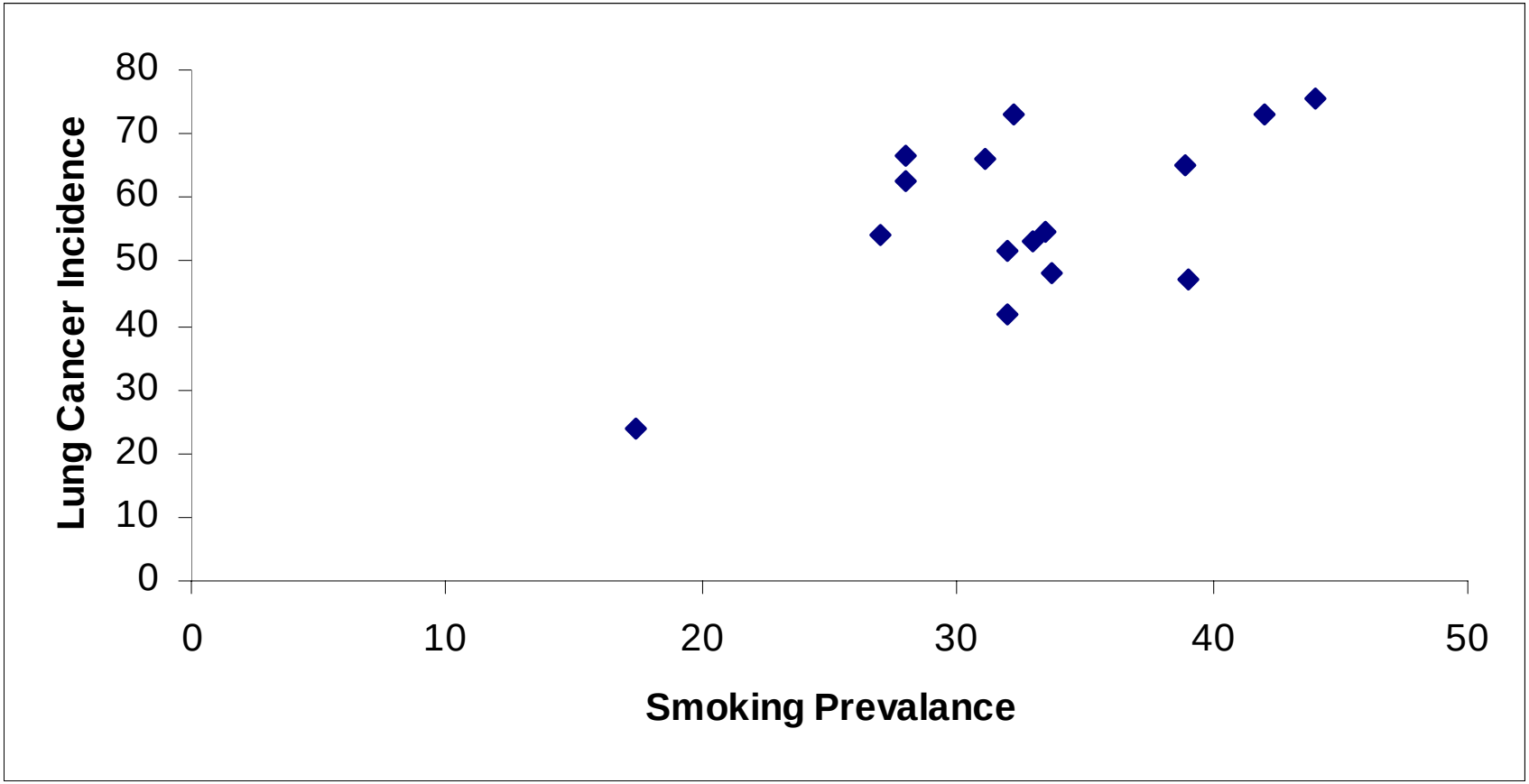
- $\chi^2 = 20.98$
- $p < 0.01$

- Therefore strong evidence against the null hypothesis for no association

Correlation and Regression

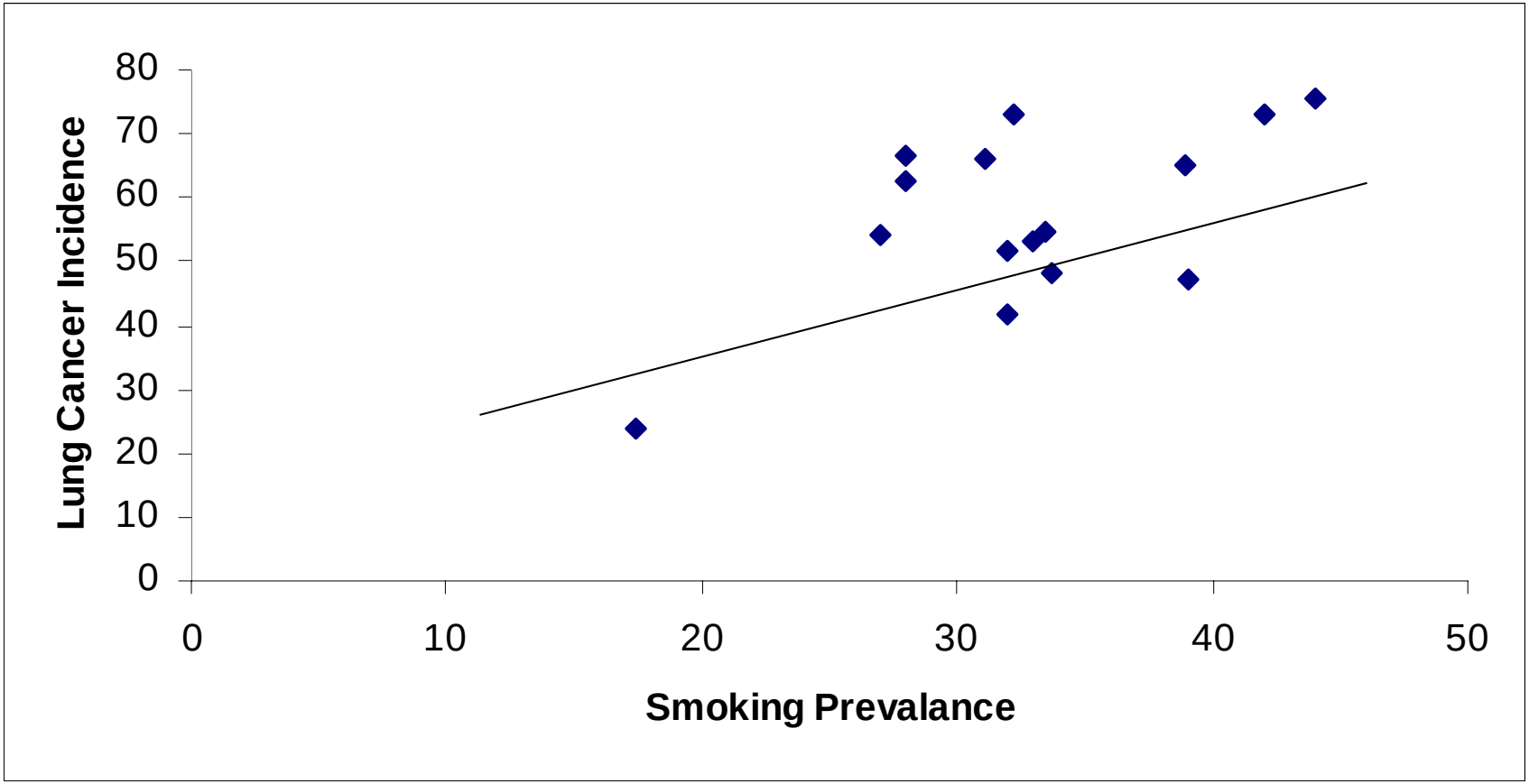
- Two variables are said to be correlated if knowing the value of one of the variables tells you something about the value of the other.
- Regression is the process of modelling how a certain random variable (response) is correlated to an associated variable (explanatory).

Example
Smoking and Lung Cancer in Europe



* Figures taken from Cancer Research and International Agency for Research in Cancer (IARC) websites

Regression



$y = 15.26 + 1.28 *x$

$R^2 = 0.368$

From Correlation to Causation

- Strength of the association
- Dose-response relationship
- Consistency of the association (with other studies)
- Temporality (exposure precede disease)
- Biological plausibility
- Lack of conflict

Hill, A. B. (1965) The environment and disease: Association or causation? Proceedings of the Royal Society of Medicine, **58,** 295-300

Summary

- For differences in means between two continuous variables, use Student's t-test if the data is normally distributed. Otherwise use non-parametric tests such as the Mann-Whitney
- For more than 2 variables, use ANOVA for normally distributed data. Otherwise, use a non-parametric version such as Kruskall-Wallis
- For looking at association between categorical variables, use chi-squared analysis