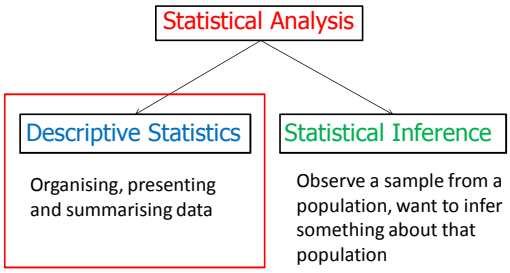
Descriptive Statistics I

Trinity College Dublin
The University of Dublin

What do we mean by Descriptive Statistics?



2

Outline

- Population and Sample
- Types of data (numerical, categorical)
- Graphical presentation (tables and plots)
- Measures of the centre of a set of observations
- Measures of variability
- Probability distributions

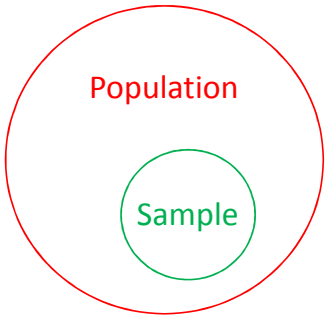
3

Population and Sample

- A *population* is the set of all individuals that are of interest to the investigator in a particular study
- A *sample* is usually a relatively small number of individuals taken from a relatively large population. The sample is only part of the available data
- It is very important to understand the distinction between what the population is and what the sample is – especially when carrying out inference

4

Population and Sample



5

Population and Sample: Examples

- Want to examine the blood pressure of all adult males with a schizophrenia diagnosis in Ireland
- Population is all adult males with a schizophrenia diagnosis in Ireland
- Take a random sample of 100 adult males with a schizophrenia diagnosis and measure their blood pressure

6

Population and Sample: Examples

- A medical scientist wants to estimate the average length of time until the recurrence of a certain disease
- Population is all *times* until recurrence for all individuals who have had a particular disease
- Take a sample of 20 individuals with the particular disease and record for each individual their time to recurrence

7

Defining the Population

- Sometimes it's not so easy to exactly define the population
- A clinician is studying the effects of two alternative treatments:
  - How old are the patients?
  - Are they male/female, male and female?
  - How severe is, or at what stage is, their disease?
  - Where do they live, what genetic/ethnic background do they have?
  - Do they have additional complications/conditions?
  - and so on…
- When writing up research findings precise information on the specific important details that characterise the population are necessary in order to draw valid inferences from the sample, about the population

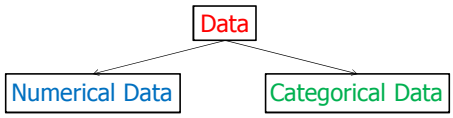
8

Data

- Data are what we collect/measure/record
- There are many different types of data
- It is vital to be able to distinguish the type(s) of data that we have in order to decide how best to both describe and analyse this data

9

Types of Data

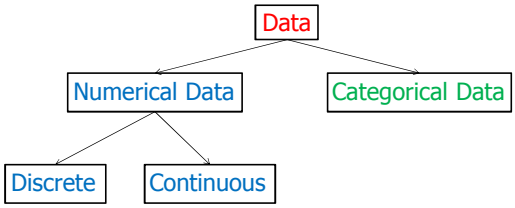


- Data can take numerical values
- Also known as *quantitative* or *metric*

- Typically not numerical but can be coded as numerical values
- Also known as *qualitative*

10

Types of Data



- Values vary by finite steps

- Can take any numerical value

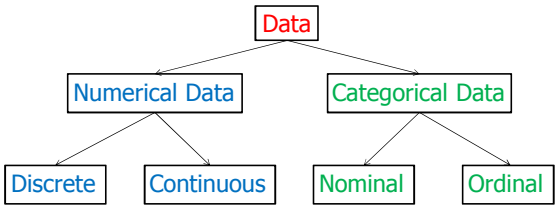
11

Numerical Data

- Discrete Numerical Data: values vary by finite steps
  - Number of siblings
  - Number of doses
  - Number of children
- Continuous Numerical Data : can take any numerical value
  - Birth weight
  - Body temperature
  - Proportion of individuals responding to a treatment

12

Types of Data



- Values vary by finite steps
- Can take any numerical value
- Categories have no particular order
- There is a natural ordering to the categories

13

Categorical Data

- Nominal Categorical Data (non-ordered) : categories have no particular order
  - Male/female
  - Eye colour (blue, green, brown etc.)
- Ordinal Categorical Data (ordered): there is a natural ordering to the categories
  - Disagree/neutral/agree
  - Poor/fair/good

14

Other Types of Data

- Ranks: Relative positions of the members of a group in some respect
  - Order that an individual comes in a competition or examination
  - Individuals asked to rank their preference for a treatment type
- Rates: ratio between two measurements (sometimes with different units)
  - Birth rate: e.g.: number of births per 1,000 people per year
  - Mortality rate: e.g.: number of deaths in a population, scaled to the size of that population, per unit of time

15

Plotting Data

- As well as sometimes being necessary, it is always good practice to display data, using plots, graphs or tables, instead of just having a long list of values for each variable for each individual
- It is always a good idea to plot the data in as many ways as possible, because one can learn a lot just by looking at the resulting plots

16

Plotting Data

- How to display data?
- Choice of how to display data depends on the type of data
- Here are a few of the most common ways of presenting data

17

Tables

- Objective of a table is to organise the data in a compact and readily comprehensible form
- Categorical data can be presented in a table
- One way, count the number of observations in each category of the variable and present the numbers and percentages in a table
- Need to be careful not to attempt to show too much in a table – in general a table should be self-explanatory

18

Tables

- Three groups of 10 patients each received one of 3 treatments (A, B, C)
- For each treatment a certain number of patients responded positively (positive = 1, negative = 0)
- A subset of the total data is shown here

Patient No.	Treat A	Treat B	Treat C	Result
1	1	0	0	1
2	1	0	0	1
3	1	0	0	0
4	1	0	0	0
5	1	0	0	0
6	1	0	0	1
7	1	0	0	1
8	1	0	0	0
9	1	0	0	0
10	1	0	0	1
11	0	1	0	0
12	0	1	0	0
13	0	1	0	0
14	0	1	0	0
15	0	1	0	0
16	0	1	0	1
17	0	1	0	1

19

Tables

- Summarising the data in a table allows easier understanding of the data, here is one way the data could be presented:

Treatment	No. of Positive Outcomes	% of Total
A	5	16.7
B	4	13.3
C	7	23.3

20

Tables

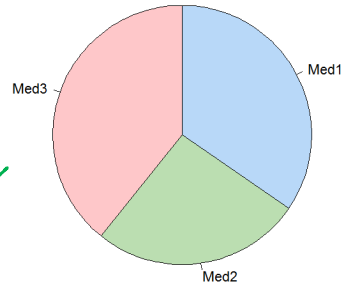
- Here is another way the same data could be presented:

Treatment	No. of Positive Outcomes	% of Total Receiving that Treatment
A	5	50
B	4	40
C	7	70

21

Pie Charts (Categorical Data)

- Pie charts are a popular way of presenting categorical data

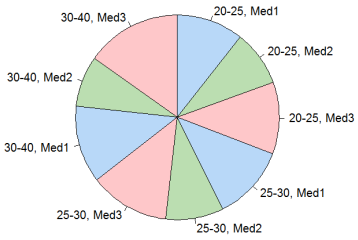


22

Pie Charts (Categorical Data)

- Be careful not to divide the circle into too many categories as this can be confusing and misleading as the human eye is not good with angles! (rough guide: 6 max)

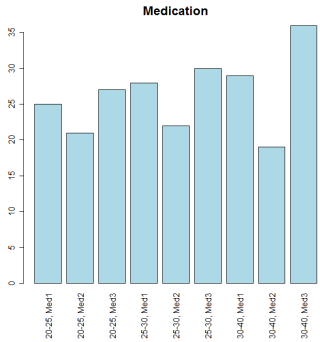
X



23

Bar Charts

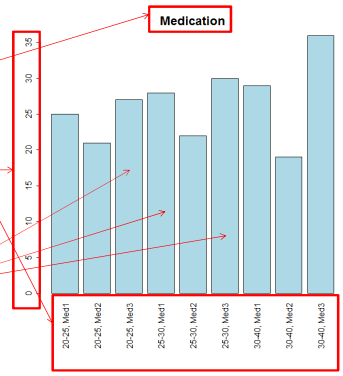
- Use a bar chart for discrete numerical data or categorical data
- Usually the bars are of equal width and there is space between them



24

Bar Charts

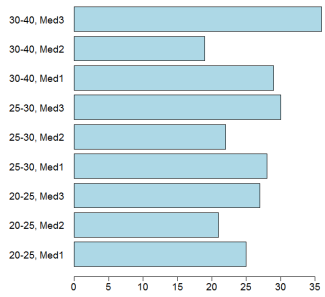
- Title and axis labels
- Axis scale
- Data bars



25

Bar Charts

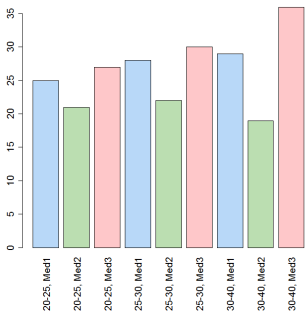
- Another bar chart for the same data



26

Bar Charts

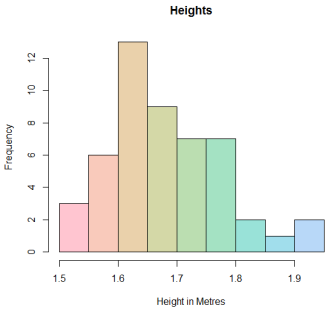
- Another bar chart for the same data



27

Histogram

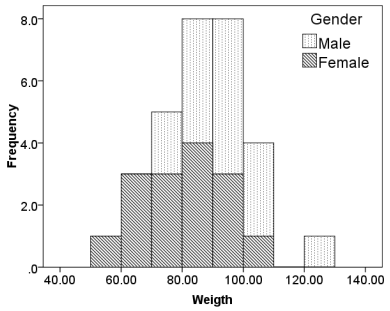
- Histogram is used to display continuous numerical data
- The total area of all the bars is proportional to the total frequency
- The width of the bars does not always have to be the same



28

Histogram

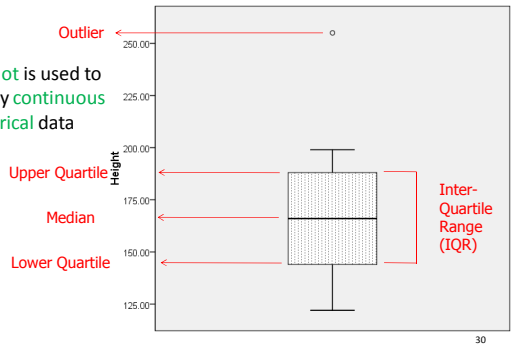
- Another example of a histogram



29

Box-and-Whisker Plot or Box Plot

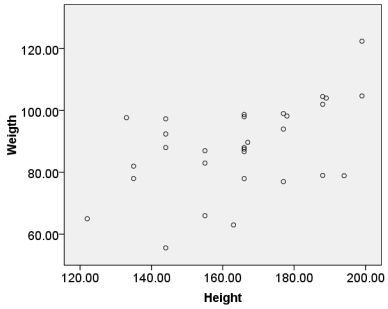
- Box plot is used to display continuous numerical data



30

Scatter Plot

- Scatter plots can be used to investigate correlations or relationships between two sets of measurements



31

Descriptive Statistics

- After presenting and plotting the data, the next step in descriptive statistics is to obtain some measurements of the centre and spread of the data

32

Measuring the Centre of the Observations

- Suppose we have a set of numerical observations and we want to choose a single value that will represent this set of observations
- How do we choose such a value?
- What is meant by the average of a set of observations?
- We will look at 3 measures of the centre of the observations:
  - Median
  - Mean
  - Mode

33

Median

Individual ID	IQ Score
1	75
2	81
3	79
4	69
5	85
6	98
7	100
8	102
9	76
10	84

- Table contains data for IQ scores for 10 individuals
- Rank the observations, i.e., write them down in order of size beginning with the smallest

  69, 75, 76, 79, 81, 84, 85, 98, 100, 102
- Median is the observation that has as many observations above it as below it in the ranked order

34

Median

Individual ID	IQ Score
1	75
2	81
3	79
4	69
5	85
6	98
7	100
8	102
9	76
10	84

- When n (total number of observations) is odd:

  Median = $((n + 1)/2)^{th}$ observation
- When n is even:

  Median = half way between the $(n/2)^{th}$ observation and the $((n/2) + 1)^{th}$ observation
- Here n is even:

  69, 75, 76, 79, 81, 84, 85, 98, 100, 102

  median = (81 + 84)/2 = 82.5

35

Mean

- The arithmetic mean, often just simply called "the mean" or the average, is defined to be the sum of all the observations divided by the number of observations:

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

- The mean is calculated using the actual values of all the observations (unlike the median) and is therefore particularly useful in detecting small differences between sets of observations
- x_i refers to each of the individual observations, there are n of these

36

Mean

- $\overline{x}$ is the mean of the observations in the sample, it is not necessarily equal to the mean of the population, which we term μ
- $\overline{x}$ is used as an estimate of μ, the mean of the population
- For the IQ data above, the mean IQ is $\overline{x} = 84.9$

  (75 + 81 + 79 + 69 + 85 + 98 + 100 + 102 + 76 + 84)/10 = 84.9

37

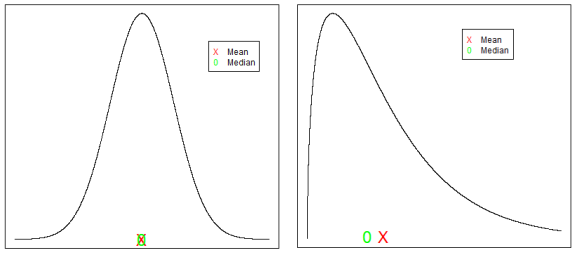
Mean or Median

Individual ID	IQ Score
1	75
2	81
3	79
4	69
5	85
6	98
7	100
8	102 500
9	76
10	84

- Median is unaffected by outliers
  Here median = 82.5
- The mean, because it takes all values into account, is affected
  Here mean = 124.7
- Mean has better mathematical properties as it takes all data into account
- Median is usually used for descriptive statistics

38

Mean or Median



- For symmetric data, the median and the mean are the same
- The median can be a better measure than the mean when the data are skewed

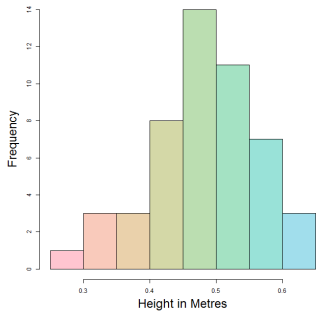
39

Mode

- The mode is that value of the variable which occurs most frequently
- As a measure of the central value of a set of observations, the mode is less commonly used than either the mean or median
- Some sets of observations may have no mode and some may have more than one mode (unimodal = 1 peak, bimodal = 2 peaks)
- The mode can be used for categorical measurements

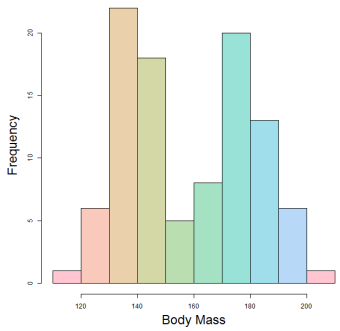
40

Mode - Unimodal



41

Mode - Bimodal



42

Summary I

- Need to be able to define what the population is and what the sample is in the study you are carrying out and in the data you are analysing
- Need to be able to determine the type of data that you have
  - First in order to be able to describe, plot or put the data into tabular form
  - Later on to choose the best, most appropriate way to analyse your data

43

Summary II

- Descriptive statistics:
  - Measures of centrality for your data, choosing the most appropriate for your data
    - Mode
    - Median
    - Mean

44

Descriptive Statistics II

Trinity College Dublin
The University of Dublin

Overview

- Measures of variability
  - Range
  - Interquartile Range
  - Variance
  - Standard Deviation
- Probability distributions:
  - Binomial
  - Normal
  - Standard Normal
  - Student's t

2

Variability

- Statistics may be defined as the study of variability
- If there was no variability there would be no need for statistics
- How do we measure the variability in the data?

3

Range

Individual ID	IQ Score
1	75
2	81
3	79
4	69
5	85
6	98
7	100
8	102
9	76
10	84

- The *range* of a set of observations is the difference between the largest and smallest observations
- The range for the IQ data is

  102 – 69 = 33

4

Range

- In small sets of observations, the range can be a useful measure of variability
- As the range only uses two observations, the highest and the lowest, and ignores the pattern of distribution of the observations in between, it can be relatively uninformative in larger data sets

5

Inter-Quartile Range and Box Plot

- A box plot shows the distribution of the data based on various percentile values
  - A rectangular box shows where most of the data lie
  - A line in the box marks the centre of the data
  - Whiskers, which encompass all or nearly all of the remaining data, extend from either end of the box
  - Outliers are represented as far out dots or circles, etc.

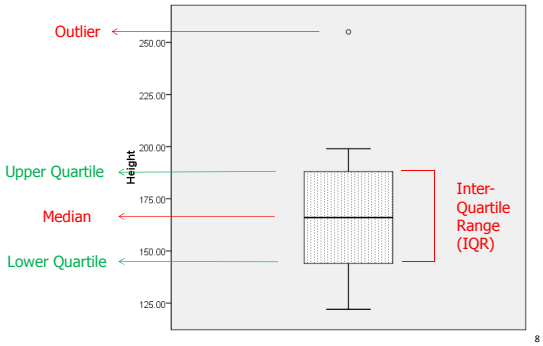
6

Inter-Quartile Range and Box Plot

- 25th percentile = lower quartile = median of the lower half of the data
- 50th percentile = median of the data
- 75th percentile = upper quartile = median of the upper half of the data
- Difference between the upper and lower quartiles is called *the inter-quartile range (IQR)*

7

Inter-Quartile Range and Box Plot



8

Variance and Standard Deviation

- Suppose we have calculated the mean
- We would like to measure the variability of the observations by seeing how closely the individual observations cluster around the mean
- The sample variance is defined as:

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

- s^2 is the sample estimate of the population variance σ^2

9

Variance and Standard Deviation

- The sample standard deviation is given by the square root of the variance

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

- Small standard deviation says the observations cluster closely around the mean, larger standard deviation says the observations are more scattered
- Standard deviation is often used as it has the same units as the mean

10

Probability Distributions

- Consider an experiment: toss a coin
  - Coin comes up either a head or a tail
- Another experiment: throw a dice
  - Either a 1, 2, 3, 4, 5, 6 will come up
- With each of the outcomes there is a probability associated
  - Coin Toss: probability of 0.5 for either a head or a tail
  - Throw of Dice: probability of 1/6 for each of 1, 2, 3, 4, 5, 6

11

Random Variables

- A Probability Distribution assigns a probability to each of the possible outcomes of a random experiment
- Constant: the value does not change
- Variable: the value can change
- Random variable: a variable whose value depends on chance, it is random (stochastic variable)

12

Discrete Probability Distributions

- A Probability Distribution assigns a probability to each of the possible outcomes of a random experiment
- Experiment: Treatment Effectiveness
  - do patients respond to the treatment or not?
  - binary outcome (yes or no) as to whether they respond or not
- Discrete probability distribution: can easily assign a probability to each of the possible outcomes

13

Binomial Distribution

- Binomial Distribution is a discrete probability distribution
- Gives the probability for the number of successes in a sequence of n independent yes/no experiments
- Each of the individual experiments has a probability p of success
- Only two possible outcomes: success and failure
- n and p are referred to as the parameters of the distribution

14

Parameters

- The parameters of a distribution define the distribution – determine its shape
- Change the values of the parameters and the distribution changes
- Distributions are defined by a number of parameters

15

Binomial Distribution

- Blood groups: B, O, A, AB
- Probability of an individual having blood group B = 0.08
- Probability of an individual not having blood group B, being one of O, A, AB = 1- 0.08 = 0.92
- Two random, unrelated individuals
  - What is the probability neither have blood group B?
  - What is the probability one has blood group B?
  - What is the probability both have blood group B?

D. Altman, Practical Statistics for Medical Research

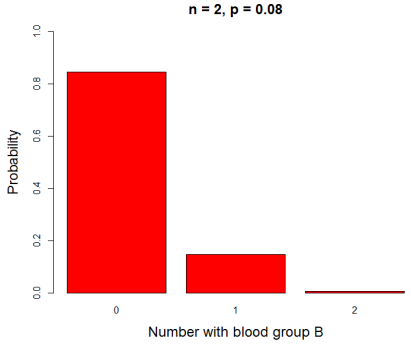
16

Binomial Distribution

- Are the assumptions of the binomial distribution satisfied?
- Only two possible outcomes:
  - Blood group B
  - Not blood group B (O, A, AB)
- The individuals are unrelated – independence
- The probability of each person having blood group B does not change from person to person(p = 0.08)

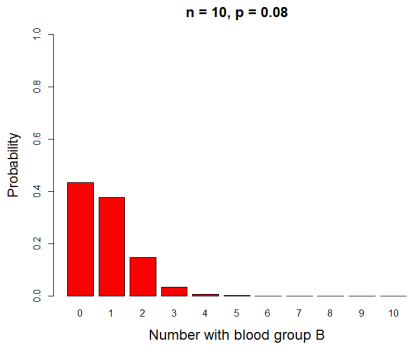
17

Binomial Distribution



18

Binomial Distribution



19

Discrete Probability Distributions

- Many other discrete probability distributions
  - Multinomial – more than two possible outcomes
  - Poisson – count data
  - Hypergeometric – sampling without replacement
  - etc.

20

Continuous Probability Distributions

- When the random variable can take values from a continuum, we need to consider continuous probability distributions
- For example
  - Height
  - Weight
- With continuous probability distributions (densities) the probability of the random variable taking on a particular value is zero
- Can only think about the probability for an interval of values

21

Normal Distribution

- The Normal, Gaussian or bell-shaped distribution is a very important continuous probability distribution
- Many statistical tests are based on the assumption that the data are Normally distributed
- The distribution is described/defined by two parameters - the mean μ and σ the standard deviation

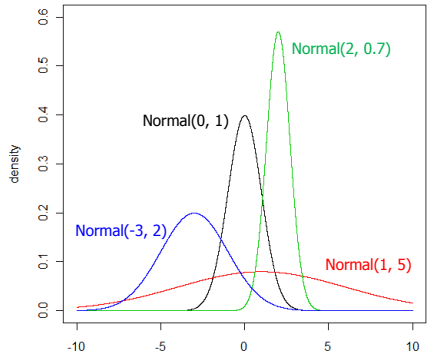
22

Normal Distribution

- The curve of the Normal distribution is
  - bell-shaped
  - symmetric about the mean
  - the shape of the curve depends on the standard deviation, the larger the standard deviation the more spread out the distribution

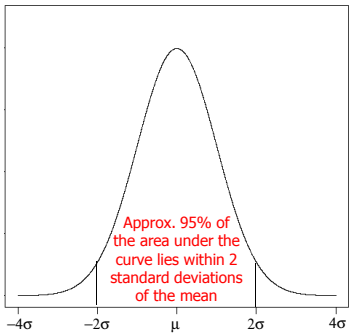
23

Normal Distributions



24

Normal Distribution



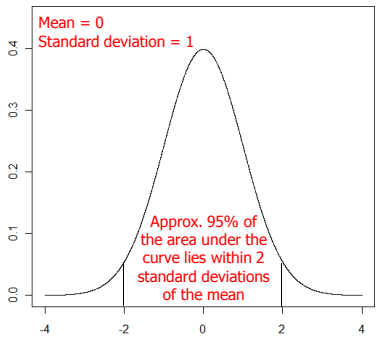
25

Standard Normal Distribution

- The area under a part of the curve gives a particular probability
- To find out the area/probability we use the *Standard Normal Distribution*
  (mean = 0, standard deviation = 1)
  and look up the area in tables or use a computer
- z has a standard Normal distribution: $z \sim N(0,1)$

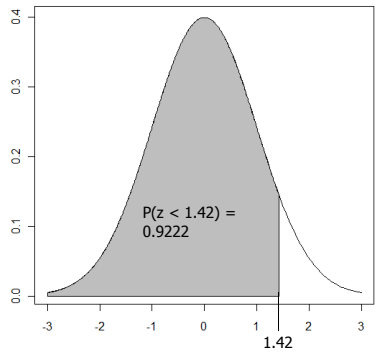
26

Standard Normal



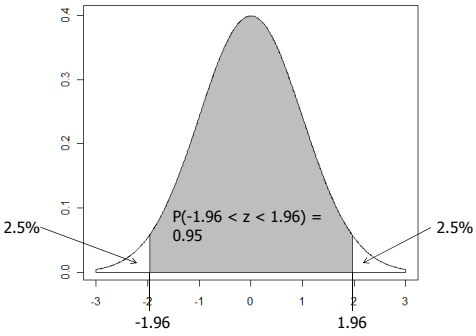
27

Area Under Standard Normal



28

Area Under Standard Normal



29

Normal Distribution Example

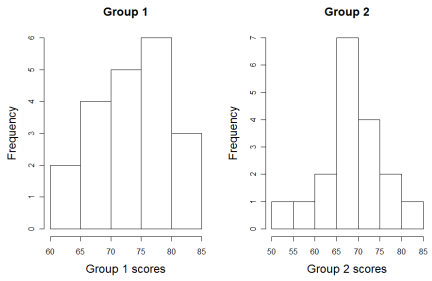
- Here are some data from psychological test scores

Patient ID	Group	Score
01	1	71.2
02	2	68.0
03	1	73.6
04	2	75.6
05	1	62.3
06	2	74.5
04	1	75.4
05	2	65.9
06	1	74.9
.	.	.
.	.	.
.	.	.

30

Normal Distribution Example

- And the distribution of these scores for each group



31

The Standard Normal distribution

- Any Normal distribution can be converted to a standard Normal by subtracting the mean and dividing by the standard deviation

$$X \sim N(\mu, \sigma)$$

$$z = \frac{X - \mu}{\sigma}$$

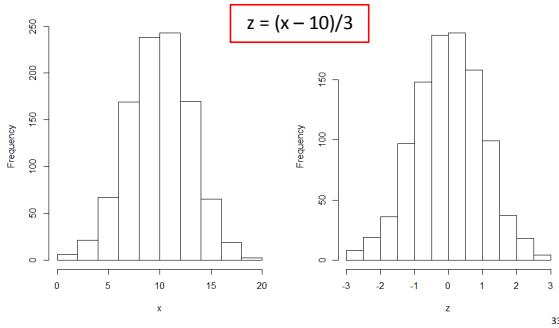
32

The Standard Normal distribution

X ~ Normal(10, 3)

Z ~ Normal(0, 1)

z = (x – 10)/3



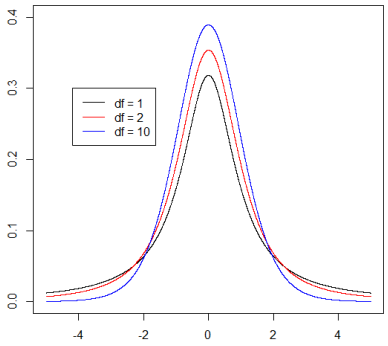
33

Student's t-distribution

- The Student's t-distribution is another symmetric continuous probability distribution
- This distribution is very similar to the Normal Distribution but has heavier tails
- Appear in many statistical tests when the sample size is relatively small
- Has one parameter: degrees of freedom(df)

34

Student's t-distribution



35

Continuous Probability Distributions

- Other continuous probability distributions:
  - Chi- square distribution: describes the sum of a number of squares of standard Normally distributed random variables
  - Uniform distribution: all intervals of the same length are equally probable

36

Summary I

- Descriptive statistics:
  - Measures of spread:
    - range
    - IQR
    - Variance
    - standard deviation
- Random variables
- Probability distributions
  - Discrete distributions: Binomial Distribution
  - Continuous distributions: Normal distribution, Standard Normal Distribution

37

Study Design

Trinity College Dublin
The University of Dublin

Outline

- Types of Study
- Sampling and Experimental Strategies
- Errors
- Hypotheses
- Results of a Hypothesis Test
- Statistical Significance
- Outcome Measures
- Effect Size
- Power

2

Scientific Studies and Experiments

- Exploratory
  - To collect *data* about the natural world
  - To identify *associations* and *dependencies* amongst the variables of interest
- Investigative
  - To test *hypotheses*
  - To investigate *causality*

3

Observational Studies

Methodical observation of a system without intervention

- Examples:
  - Epidemiology:
    relationship between smoking and lung cancer
  - Astronomy:
    relationship between the mass of a star and its brightness

4

Controlled Experiments

Manipulate one or more variables in order to determine the effect of the intervention

- Examples:
  - Medicine:
    clinical drug trials
  - Physics:
    relationship between electrical current, voltage and resistance

5

Case – Control Studies

- Compares group of patients with group of unaffected controls
- Relatively quick and cheap
- Difficult to select an appropriate group of controls
- Can detect correlations but not cause and effect

6

Cohort Studies

- Observes a fixed group over a period of time
- Can be *retrospective* or *prospective*
- Retrospective studies are cheap and quick, but affected by confounding variables
- Prospective studies can be controlled for confounding variables but are expensive and time consuming

7

Randomised Controlled Trials

- Subjects are assigned randomly to different groups
- Possible to control for confounding variables
- Difficult to generalise to background population
- Difficult to investigate variation over time
- Expensive

8

Example

- Background
  - It is conjectured that patients with bipolar disorder tend to have a cognitive deficit (as measured by IQ) compared with unaffected people
- Objective
  - To determine whether this is in fact the case
- This is an observational study

9

Methodology

In our example we will use a case control design

- Select a group of affected people and a group of unaffected people and see whether those that are affected have a lower than average IQ

10

Sampling and Experimental Strategy

- Randomisation
  - Assign subjects to intervention and control groups randomly to minimise the effect of confounding variables
  - This does not apply to our observational study
- Blinding of subjects
  - Subjects do not know which groups they are assigned to
  - This does not apply to our observational study

11

Sampling and Experimental Strategy

- Blinding of experimenters
  - Experimenters do not know which subject is assigned to each group
  - We can and should implement this in our study
- Matching
  - Match individual cases and controls with similar characteristics
  - We will not apply this in our study

12

Sampling and Experimental Strategy

- Stratification
  - Divide groups into sub-groups by particular characteristics, eg. age, sex
  - In our example we should stratify (at least) by age and sex

13

Stochastic Errors

- Caused by intrinsic variability in the data
  - In our study this arises because of natural differences in IQ between individuals
- These should be:
  - estimated in advance of the experiment
  - accounted for in the statistical analysis

14

Measurement Errors

- Caused by limitations in the measurement procedures
- In our study this will depend on:
  - uncertainties in the BPD diagnosis
  - the precision with which IQ can be measured
  - the care with which the measurements are taken

15

Systematic Errors

- Caused by defective experimental procedures
- In our study these may arise from:
  - differences in the calibration of different IQ scales
  - differences in diagnostic procedures between different clinicians
- Systematic errors are also known as bias

16

Hypotheses

- A hypothesis is a specific conjecture about a system
- A hypothesis should:
  - address a question of scientific interest
  - relate to the *system* and not to the experiment
  - be *specific*
  - be *testable*

17

Hypothesis Testing Procedure

1. Define the *research* hypothesis, H_1
2. Define the *null* hypothesis, H_0
3. Define the *significance threshold,* α

Conduct the Experiment

18

Example

- In our example:
  - Research Hypothesis:
    There is a difference in the mean IQ between affected and unaffected people
  - Null Hypothesis:
    There is no difference in the mean IQ between affected and unaffected people
  - Significance threshold:
    We will set this later

19

Testing the Hypothesis

- Given our data, how likely is it that our hypothesis is true?

  We cannot answer this question!

- Given that an hypothesis is true, how likely is our data?

  We can answer this question

20

Possible Outcomes of the Hypothesis Test

	H_0 True	H_0 False
Reject	✗	✓
Don't Reject	✓	✗

21

False Positives and False Negatives

- H_0 is true but is rejected
- This is also called a *false positive* or Type I Error

- H_0 is false but is not rejected
- This is also called a *false negative* or Type II Error

22

True Positives and True Negatives

- H_0 is true and is not rejected
- This is also called a *true negative*

- H_0 is false and is rejected
- This is also called a *true positive*

23

Possible Outcomes of the Hypothesis Test

	H_0 True	H_0 False
Reject	False Positive (Type I Error)	True Positive
Don't Reject	True Negative	False Negative (Type II Error)

24

Statistical Significance I

- Statistical significance, *p*:
  - The probability of rejecting the null hypothesis when it is in fact true (Type I error)
- Significance threshold, α:
  - The critical value of *p* below which we reject the null hypothesis

25

Statistical Significance II

- What threshold should we choose for our experiment?
- In *theory* this should depend on the experiment:
  - How do we want to balance Type I and Type II errors?
  - What prior evidence is there for our hypothesis?
  - How important is it that we get the answer right?

26

Statistical Significance III

- In *practice*:
  - Everyone chooses 0.05
- The critical value should be decided before the experiment is performed
- We will choose 0.05 as our significance threshold

27

Outcomes if H_0 is True

- The *probability* of a false positive equals the significance threshold, α
- The *probability* of a true negative equals $1-\alpha$
- This is also called the specificity

28

Outcomes if H_0 is False

- The *probability* of a false negative is denoted β
- The probability of a true positive equals $1-\beta$
- This is also called the sensitivity or power

29

Outcome Probabilities

	H_0 True	H_0 False
Reject	α Type I Error Rate	$1-\beta$ Power
Don't Reject	$1-\alpha$ Specificity	β Type II Error Rate
	1	1

30

Outcome Measures

- An outcome measure is the effect that we hope to observe and should be clearly defined at the design stage
- An effect size is the size of the outcome measure that we observe
- In general, it represents the strength of a relationship between two variables
- Outcome measures and effect sizes should always be clearly reported

31

Effect Size

- Some examples:
  - Differences in means
  - Differences in proportions
  - Correlation coefficient
  - Odds ratios
  - Relative risks

32

Odds Ratio and Relative Risk

Probability of occurrence in Group 1 = p
Probability of non-occurrence in Group 1 = $1 - p$

Probability of occurrence in Group 2 = q
Probability of non-occurrence in Group 2 = $1 - q$

33

Odds Ratio

Odds:

$$O(p) = p / (1 - p)$$
$$O(q) = q / (1 - q)$$

Odds Ratio:

$$OR = Odds(p) / Odds(q)$$
$$= p(1 - q) / q(1 - p)$$

34

Relative Risk

$$RR = p / q$$
$$RR = (1 - p) / (1 - q) \times OR$$

When p and q are almost equal or p and q are small:

$$RR \approx OR$$

When p is much larger than q:

$$1 << RR << OR$$

When p is much smaller than q:

$$1 >> RR >> OR$$

35

Example

$p = 0.05 \quad 1 - p = 0.95$
$q = 0.04 \quad 1 - q = 0.96$

RR = **1.25**; OR = **1.26**

$p = 0.95 \quad 1 - p = 0.05$
$q = 0.80 \quad 1 - q = 0.20$

RR = 1.19 OR = 4.75

36

Relative Risk and Odds Ratio

Relative risk is easier to understand intuitively

but can be deceptive

eg:

RR can be close to 1 or far from 1 depending on how we define the "event"

37

Example

$p = 0.050 \quad 1 - p = 0.950$

$q = 0.025 \quad 1 - q = 0.975$

RR = 2.00 OR = 2.05

BUT

$p = 0.950 \quad 1 - p = 0.050$

$q = 0.975 \quad 1 - q = 0.025$

RR = 0.97 OR = 0.49

38

Relative Risk and Odds Ratio

- RR is usually used in randomised controlled trials and cohort studies
- OR is usually used in case-control studies

39

Example

- In our example:
- The outcome measure is the difference in mean IQs between the two groups
- The effect size is the numerical value of this difference

40

Effect Size and Statistical Significance

- Statistical significance does not imply scientific significance
- Effect size may imply scientific significance
- Effect size does not determine the significance
- Significance does not determine the effect size
- Effect size tells you something about nature
- Significance tells you something about your experiment

41

Power

- Power depends on (amongst other things):
  - Effect size
  - Significance threshold required
  - Stochastic variability in the data (noise)
  - and finally.... sample size

42

Power

- What power should we choose for our experiment?
- In *theory* this should depend on the experiment:
  - How do we want to balance Type I and Type II errors?
  - What are the practical considerations regarding sample size?

43

Power

- In *practice*:
  - Everyone chooses 0.8
- The power should be decided before the experiment is performed
- We will choose 0.8 for our power

44

Power versus Type I Errors

- Example: Cheap, simple test for a medical condition

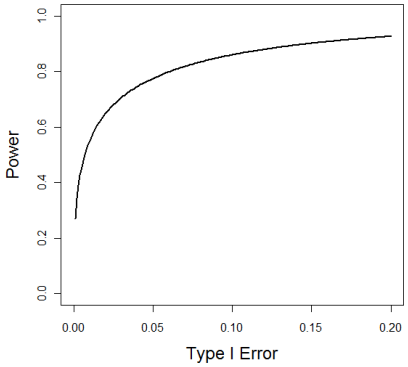
  Procedure: toss a coin

  Specify outcomes:
  Heads → positive result
  Tails → positive result
  Side → positive result

- This test has 100% power to detect any medical condition
- And 100% Type I error rate

45

Variation of Power with Type I Error



Sample Size = 60
Variability = 10
Effect Size = 5

46

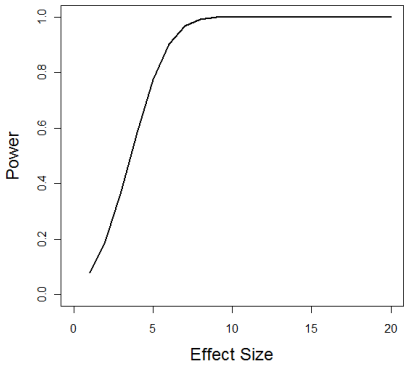
Type I Error Rate in Our Example

- We will assume a Type I error rate (significance threshold) of 0.05...

  ... based on tradition

47

Variation of Power with Effect Size



Sample Size = 60
Variability = 10
Significance= 0.05

48

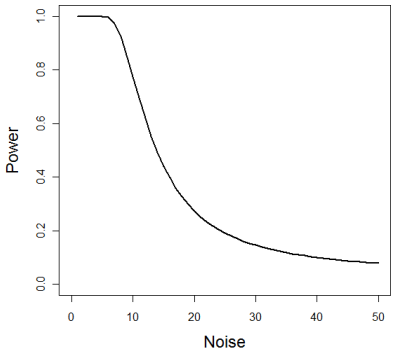
Effect Size in Our IQ Example

- We will assume a difference in the means of 5 IQ points...

  ... based on expert opinion and experience

49

Variation of Power with Stochastic Variability



Sample Size = 60
Effect Size = 5
Significance= 0.05

50

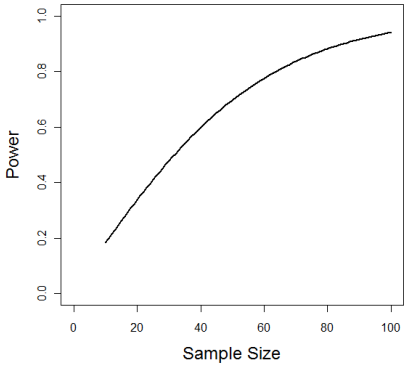
Data Variability in Our Example

- We will assume a standard deviation of 10 IQ points...

  ... based on experience and preliminary testing

51

Variation of Power with Sample Size



Variability = 10
Effect Size = 5
Significance= 0.05

52

Sample Size in Our Example

- We will use a sample size of 60 individuals per group...

  ... based on our other assumptions of a power requirement of 0.8

53

Summary of Our Study I

- Scientific Question: Do people with BPD tend to have lower IQs than unaffected people?
- Methodology: Case control study
- Experimental Strategy:
  - Blinding of experimenters
  - Stratification by age and sex
- Null Hypothesis: The difference in the mean IQ between case and control groups is zero

54

Summary of Our Study II

- Outcome Measure: Difference in the Means
- Estimated Effect Size: 5 IQ points
- Estimated Variability: 10 IQ points
- Significance Threshold: 0.05
- Sample Size: 60 per group
- Power: 0.8

55

Summary I

- Types of Study:
  Exploratory, investigative, observational studies, controlled experiments
- Methodologies:
  Prospective, retrospective, case–control
- Sampling and Experimental Strategy:
  Randomisation, blinding, matching, stratification
- Errors:
  Stochastic errors, measurement errors, systematic errors (bias)

56

Summary II

- Hypotheses:
  Good and bad hypotheses, null and alternative hypothesis
- Hypothesis Testing:
  Likelihood of data, rather than likelihood of hypothesis, false positives and false negatives, true positives and true negatives, Type I and Type II errors
- Statistical Significance:
  Significance threshold, specificity, sensitivity (power)

57

Summary III

- Outcome Measure:
  Effect size, relationship between effect size and statistical significance
- Power:
  Relationship between power, sample size, data variability and effect size

58

Take Home Message

"To propose that poor design can be corrected by subtle analysis techniques is contrary to good scientific thinking"

Stuart Pocock ("Controlled Clinical Trials", pg. 58) regarding the use of retrospective adjustment for trials with historical controls

59

Hypothesis Testing I

Parametric Hypothesis Testing

- A *statistical hypothesis* is a statement of belief regarding the value of one or more *population* characteristics
- Note: About a population, not a sample
- A hypothesis test is a test of that belief
- *Parametric hypothesis test* makes assumptions about the distribution of the population, typically a Normal distribution assumption

2

Hypothesis Test

- Hypothesis testing typically involves four steps:
  1. Formulation of the hypothesis
  2. Select and collect sample data from the population of interest
  3. Application of an appropriate test
  4. Interpretation of the test results

3

Hypothesis Test: Example

- The average height of males in the population is believed to be approximately 175cm
- We want to know if male patients attending particular out-patient clinics are also this tall on average or are they smaller or taller?

4

Null and Alternative Hypotheses

- The *null hypothesis*, denoted H_0, is a claim about a population characteristic
- Initially we assume the null hypothesis is true
- The opposite hypothesis is termed the *alternative hypothesis* and is denoted by H_1
- Need to turn the research/clinical question into a statistical hypothesis that we can test

5

Hypotheses: Example

- For our example data set, the research question could be: *"Are male patients who attend out-patient clinics of average height?"*
- Null hypothesis: *the mean height of male patients is the same as the average height of males:*

$$H_0: \mu = 175\text{cm}$$

- Alternative hypothesis: *the mean height of male patients is not the same as the average height of males:*

$$H_1: \mu \neq 175\text{cm}$$

- μ = the population mean height of male patients attending the particular type of out-patient clinics

6

Hypotheses: One-Sample z-Test

- To test the null hypothesis we will use a one-sample z-test
- Assumptions of the one-sample z-test:
  - Independent random sampling
  - Large sample size (rough guide at least 30)
  - Normally distributed population
  - Standard deviation of the population known

7

Hypothesis Testing: Significance Level

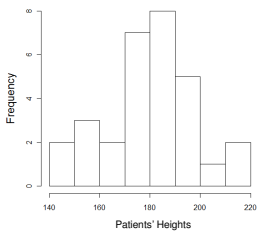
- The significance level is the probability of wrongly rejecting the null hypothesis H_0, if it is in fact true
- Usually $\alpha = 0.05$, this is just a convention, sometimes $\alpha = 0.01$ is used. The level is based on the importance of the decision being made and the consequences of falsely accepting or rejecting H_0
- We will use a significance level: $\alpha = 0.05$

8

Hypothesis Test: Example

- We collect data on the heights of 30 male patients from out-patient clinics
- Here is a subset of the data and a plot of all the data

Patient ID	Height (cm)
01	148
02	197
03	173
04	192
05	174
.	.
.	.
.	.



9

Hypotheses: Example

- For our example data set, the sample mean:

$$\overline{x} = 180.1$$

- Is this just by chance? Did we pick a sample that just happens to be taller than the general male population? Or are male patients taller than the average male population?
- To answer these questions we test our null hypothesis

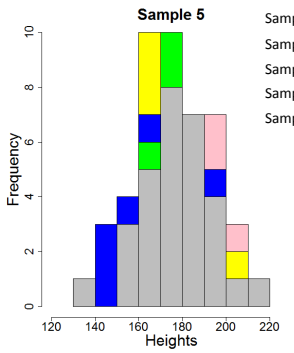
10

Sampling Distribution of the Mean

- In order to test the hypothesis we first need to understand what we mean by the sampling distribution of the mean
- If we take repeated samples of size *n* from a population, we would expect the means of each of these samples to vary
- These means will have their own mean and standard deviation

11

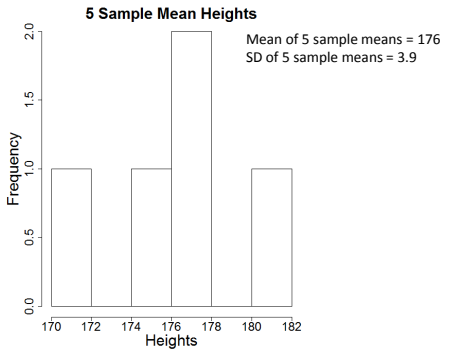
Sampling Distribution of the Mean



Sample 1: Mean = 181.4 SD = 14.1
Sample 2: Mean = 175.4 SD = 12.9
Sample 3: Mean = 170.4 SD = 16
Sample 4: Mean = 176.3 SD = 12.9
Sample 5: Mean = 176.5 SD = 17.2

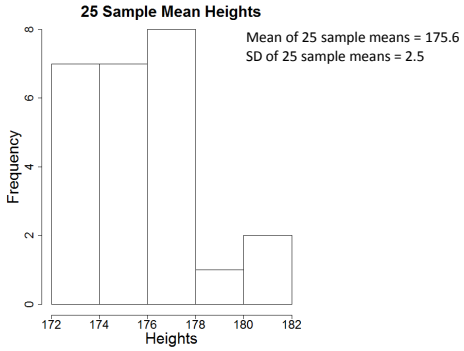
12

Sampling Distribution of the Mean



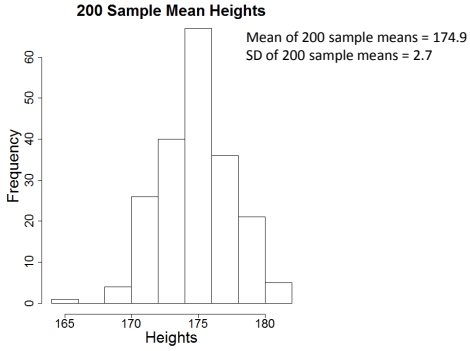
13

Sampling Distribution of the Mean



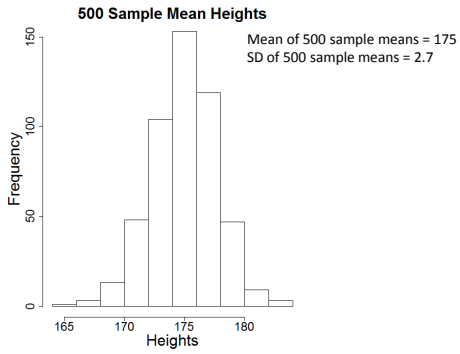
14

Sampling Distribution of the Mean



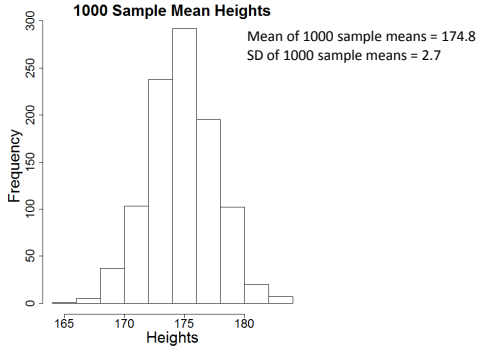
15

Sampling Distribution of the Mean



16

Sampling Distribution of the Mean



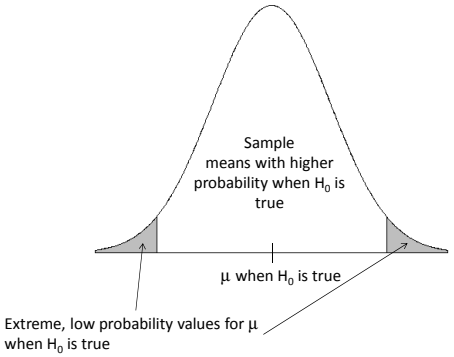
17

Sampling Distribution of the Mean

- If the true population mean and standard deviation are μ and σ respectively, then the sample means will have a mean of μ and a standard deviation of $\frac{\sigma}{\sqrt{n}}$ also called the standard error of the mean
- For large samples the distribution of the sample means will be Normal

18

Sample Means When H_0 is True



19

Hypotheses: One-sample z-Test

- Start with a normal variable that has a given mean and standard deviation
- Transform this normal variable so that it has a mean of 0 and standard deviation of 1
- The transformed variable has a standard normal distribution: Normal(0, 1)

20

Hypothesis Testing: P-Value

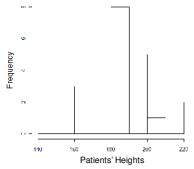
- We have obtained sample data from the population
  - Sample of male out-patients' heights
- We now evaluate the probability that we could have observed this data if the null hypothesis were true
- This probability is given by the P-value
- The smaller the P-value the more unlikely this is
- We evaluate this probability using a test statistic

21

Hypothesis Testing: Test Statistic

- For the male patients' heights:

Sample mean: $\overline{x}$ = 180.1



Standard error of the sample mean = $\frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{30}} = 2.7$

22

Hypothesis Testing: Test Statistic

Test Statistic = $\frac{\textit{Observed Value} - \textit{Hypothesized Value}}{\textit{Standard Error of the Observed Value}}$

Sample Mean — Hypothesized Mean

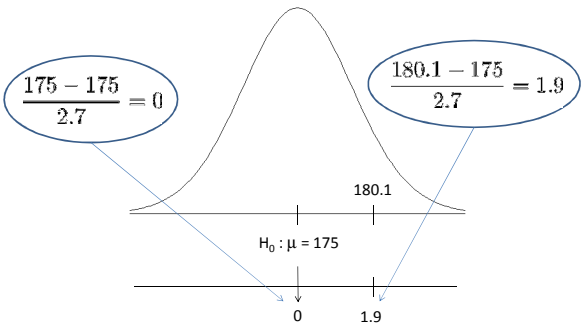
$$\frac{180.1 - 175}{2.7} = 1.9$$

Test Statistic

Standard Error of the Mean

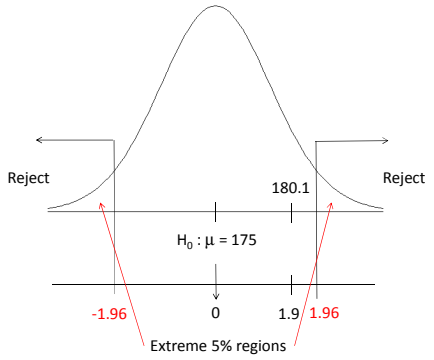
23

Hypothesis Testing: Test Statistic



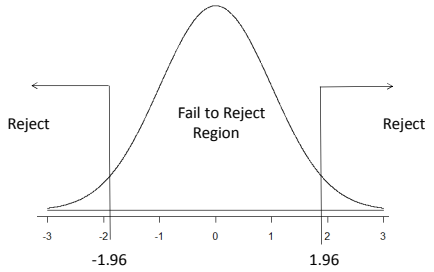
24

Hypothesis Testing: Test Statistic



25

Hypothesis Testing: Normal(0,1)



- Standard Normal Distribution: Normal(0, 1)
  - The rejection region for H_0 and the fail to reject region for H_0 for a z-test at a two-sided significance level of 5%

26

Hypothesis Testing: Test Statistic

- A test statistic is calculated from the sample data
- It is used to decide whether or not the null hypothesis should be rejected
- The general form for the test statistic is the following:

$$\textit{Test Statistic} = \frac{\textit{Observed Value} - \textit{Hypothesized Value}}{\textit{Standard Error of the Observed Value}}$$

- The test statistic expresses the distance between the observed value and the hypothesized value as a number of standard errors

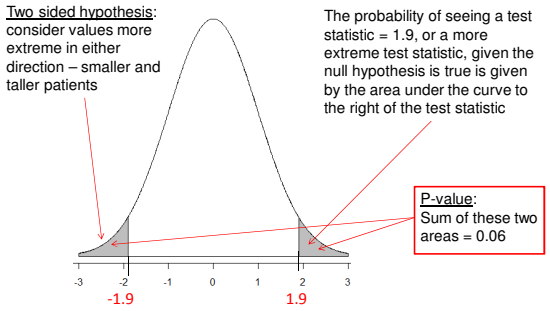
27

Hypothesis Testing: Significance Level

- What's the probability of observing the test statistic 1.9, or a more extreme test statistic, given the null hypothesis is true?
- This is the P-value
- Use the z tables to compute this probability

28

Hypothesis Testing: Normal(0,1)



29

Hypothesis Testing: Significance Level

- Is this P-value large? Do we reject H_0?
- The answer to these questions depends on the significance level: $\alpha = 0.05$
- H_0 should be rejected if the P-value < α
- H_0 should not be rejected if the P-value >= α

30

Hypothesis Testing: Interpreting P-value

- The P-value is 0.06 for the analysis carried out on the heights of the male patients
- Thus at a significance level (α) $= 0.05$ we fail to reject the null hypothesis that the mean height of the male patients attending the out-patient clinics is equal to 175cm

31

Failing to Reject the Null Hypothesis

- The null hypothesis is never accepted
- We either reject or fail to reject the null hypothesis
- Failing to reject means that no difference is one of the possible explanations but we haven't shown that there is no difference
- The data may still be consistent with differences of practical importance

32

Hypothesis Testing: Errors

- Associated with every hypothesis test are errors:
- Type I Error: (false positive)
  is the error of rejecting H_0 when it is actually true
- Type II Error: (false negative)
  is the error of failing to reject H_0 when it is false

	***H_0* True**	***H_0* False**
Reject	False Positive (Type I Error)	True Positive
Don't Reject	True Negative	False Negative (Type II Error)

33

Hypothesis Testing: Errors

- The probability of a Type I Error is predetermined by the significance level α
- The probability of a Type II Error is denoted β
- The power of a statistical test is defined as $1-\beta$ and is the probability of rejecting H_0 when H_0 is false
- A good test is one which minimises α and β

34

Confidence Intervals

- Remember we are interested in some aspect of a population
- We take a random representative sample from this population and collect some data from this sample
- Suppose we consider the mean of the data
- The mean of the sample ($\overline{x}$) is a point estimate of the population mean (μ)

35

Confidence Intervals

- If we took another random sample from the population and collected data for this second sample we may get a different sample mean
- We would like to consider the range within which the true population mean would be expected to lie, not just the point estimate
- We can use confidence intervals to do this

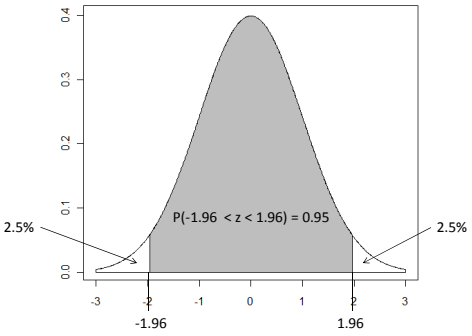
36

Confidence Intervals

- A confidence interval for a population characteristic (doesn't have to be the mean) is an interval of plausible values for that characteristic of interest
- Associated with each confidence interval is a confidence level
- If we took repeated samples and calculated confidence intervals, the confidence level says what proportion of those would be expected to contain the true population parameter
- Usual choices are 95%, 99% etc.

37

Area Under Standard Normal



38

Confidence Intervals

Sample mean

Population mean

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

Standard deviation of the sample mean, also known as the standard error of the mean

39

Confidence Intervals

- Looking up the z-tables we can write down the following:

$$P(-1.96 \leq z \leq 1.96) = 0.95$$

- Replace z:

$$P(-1.96 \leq \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq 1.96) = 0.95$$

- Re-arranging:

$$P(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}) = 0.95$$

40

Confidence Intervals

- Which gives us our 95% confidence interval:

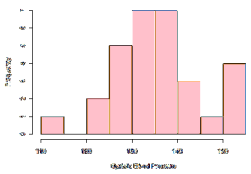
$$(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96\frac{\sigma}{\sqrt{n}})$$

- All we need to know is the sample mean and the standard deviation to obtain the confidence interval

41

Confidence Intervals

- Example: we want to estimate an interval of possible values for the mean systolic blood pressure of patients
- We take a random sample of 30 patients and record their systolic blood pressure
- Mean systolic blood pressure = 135.5
- Standard deviation of the systolic blood pressure = 9



42

Confidence Intervals

- General formula:

$$(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96\frac{\sigma}{\sqrt{n}})$$

- For the blood pressure data:

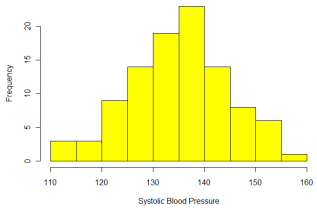
$$(135.5 - 1.96\frac{9}{\sqrt{30}}, 135.5 + 1.96\frac{9}{\sqrt{30}})$$

- A 95% confidence interval for the population mean systolic blood pressure is: (132.3, 138.7)

43

Confidence Intervals

- Suppose we increase the sample size to 100 and measure the systolic blood pressure on this random sample of size 100
- What do we expect to happen to the confidence interval? Should it become narrower or wider?



44

Confidence Intervals

- Interval should become narrower
- The 95% confidence interval for the systolic blood pressure based on 100 samples is: (133.5, 137)

45

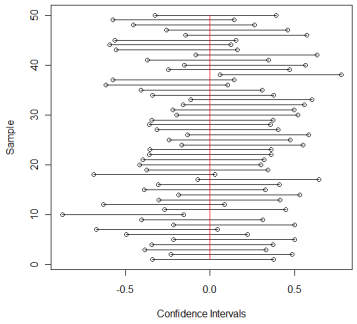
Interpreting Confidence Intervals

- A 95% confidence interval:
  - if samples were repeatedly taken from the population of interest
  - calculate confidence intervals for each sample
  - 95% of the time, these intervals would contain the true population value of the parameter of interest

46

Interpreting Confidence Intervals

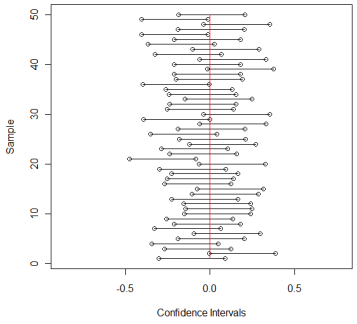
- 50 samples each of size 30, true population mean = 0, 95% CI



47

Interpreting Confidence Intervals

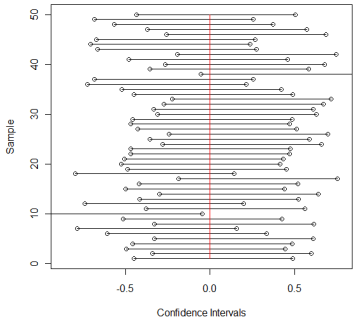
- 50 samples each of size 100, true population mean = 0, 95% CI



48

Interpreting Confidence Intervals

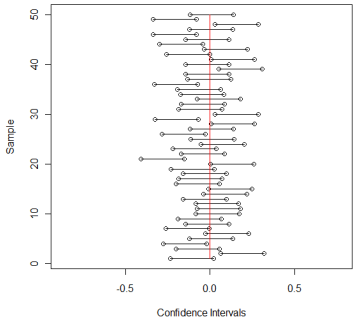
- 50 samples each of size 100, true population mean = 0, 99% CI



49

Interpreting Confidence Intervals

- 50 samples each of size 100, true population mean = 0, 80% CI



50

Interpreting Confidence Intervals

- Confidence intervals and hypothesis tests are related and provide complementary information
- For every hypothesis test, we can also consider an equivalent statement about whether or not the hypothesized value is contained in the confidence interval

51

Two Sample Hypothesis Test

- Group 1: Students received extra tuition before a test
- Group 2: Students did not receive extra tuition before a test

- Research Question:
  Does extra tuition help students to achieve better test scores or do they perform similarly to those who don't receive extra tuition?

52

Hypothesis Generation

- Null hypothesis: *the population mean test score is the same in both groups :*

  $H_0: \mu_1 = \mu_2$

  or equivalently

  $H_0: \mu_1 - \mu_2 = 0$

- Alternative hypothesis: *the population mean test score is not the same in both groups :*

  $H_1: \mu_1 \neq \mu_2$

  or equivalently

  $H_0: \mu_1 - \mu_2 \neq 0$

- μ_1, μ_2 = the population mean test score for those receiving extra tuition and those not receiving extra tuition, respectively

53

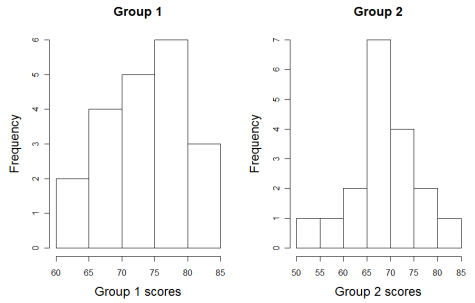
Two Sample Hypothesis Test

- Two groups of students' test scores were collected
- Here is some of the data

Student ID	Group	Test Score
01	1	71.2
02	2	68.0
03	1	73.6
04	2	75.6
05	1	62.3
06	2	74.5
07	1	75.4
08	2	65.9
09	1	74.9
.	.	.
.	.	.
.	.	.

54

Two Sample Hypothesis Test



55

Independent Two Sample *t*-test

- We carry out an independent two sample *t*-test for means
  - Two samples must be independent and random
  - The underlying populations must not be skewed
  - The standard deviation in the two samples must be the same

56

Independent Two Sample *t*-test

- Test statistic:

Sample difference in means

Hypothesized value

$$t = \frac{\bar{x}_1 - \bar{x}_2 - 0}{se(\bar{x}_1 - \bar{x}_2)}$$

Standard error of the sample difference in means

$$2.2 = \frac{73.2 - 68.7 - 0}{2.04}$$

57

Two Sample *t*-test

- t = 2.2, P-value = 0.03
- P-value < 0.05, therefore we reject the null hypothesis and conclude that the extra tuition does have an impact on the test scores of the students
- 95% confidence interval: (0.24, 8.6)
- Confidence interval also leads to the same conclusion as it does not cover 0

58

Summary I

- Hypothesis test is a statement of belief regarding the value of one or more population characteristics
- Parametric hypothesis test – makes assumptions about the population
- Setting up the hypothesis:
  - Null hypothesis
  - Alternative hypothesis
  - Significance level
- Sampling distribution of the mean
  - standard error of the mean

59

Summary II

- One sample Z-test
  - Assumptions
  - Test statistic
  - P-value
  - Rejecting or failing to reject the null hypothesis
  - Type I, Type II errors and power
- Confidence Interval
  - Relationship between confidence interval and hypothesis testing
- Independent 2 sample t-test

60

Hypothesis Testing II

Overview

- Hypothesis tests for
  - Comparing Proportions: Chi squared test
  - Paired Data
- When the Assumptions don't hold
  - Transforming Data
  - Non-parametric tests
- Exact Tests
  - Fisher's Exact Test
  - Permutation Test

2

Comparing Proportions

- We have examined how to compare means: t tests
- One of the next most common comparisons we might want to make is between proportions

3

Comparing Proportions

- Suppose we have two groups of individuals and some event happening or not in the group (e.g. responding to a treatment), a binary outcome
- How do we examine whether the proportion of individuals responding is the same in each group?

4

Comparing Proportions

- Categorical data are very common: when we can categorize individuals/objects/cells etc. into two or more mutually exclusive groups
- The number of individuals that fall into a particular group is called the frequency
- The data can be displayed in frequency tables/contingency tables or cross tabulated
- When there are only two categories for one of the variables, we can consider proportions

5

Comparing Proportions

- Suppose we have two groups of individuals
  - The individuals in Group 1 have received a treatment
  - The individuals in Group 2 have received a placebo
  - The trial was set up to be a blind trial
- After a period of time the individuals will either have responded to the treatment/placebo or not
- We want to examine whether the proportion of individuals that respond is the same in Group 1 and Group 2

6

2x2 Contingency table

	Respond	Don't Respond	
Group 1	20	40	60
Group 2	35	35	70
	55	75	130

7

2x2 Contingency table

	Respond	Don't Respond	
Group 1	a	b	60
Group 2	c	d	70
	55	75	130

Cells of the Table

8

2x2 Contingency table

	Respond	Don't Respond	
Group 1	a	b	M_1
Group 2	c	d	M_2
	M_3	M_4	130

Marginal Totals

9

2x2 Contingency table

	Respond	Don't Respond	
Group 1	a	b	a + b
Group 2	c	d	c + d
	a + c	b + d	N = a + b + c+ d

Overall Total

10

Comparing Proportions

- Want to know if the proportion of individuals responding is the same in each of the groups
- Another way of asking the same question is: whether the row and column variables are independent or not
- The null hypothesis is that responding to the treatment is independent of whether the treatment was received or the placebo

11

Chi Squared Test

- This hypothesis is tested using a Chi squared test

$$\chi^2 = \sum_{a,b,c,d} \frac{(observed - expected)^2}{expected}$$

12

Chi Squared Test

- The *observed* is the count that we have observed in our data
- The *expected* is what count we would expect to observe:

$$\textit{Expected cell frequency} = \frac{\text{row total} \times \text{column total}}{N}$$

13

Chi Squared Test

	Respond	Don't Respond	
Group 1	20	40	60
Group 2	35	35	70
	55	75	130

- Cell *a*:
- Observed = 20
- Expected = (60 X 55)/130 = 25.38

14

Chi Squared Test

	Respond	Don't Respond	
Group 1	20	40	60
Group 2	35	35	70
	55	75	130

- Cell *b*:
- Observed = 40
- Expected = (60 X 75)/130 = 34.62

15

Chi Squared Test

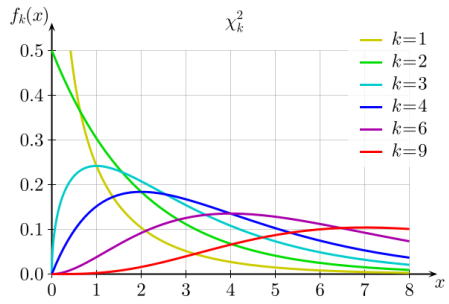
	Respond	Don't Respond	
Group 1	20	40	60
Group 2	35	35	70
	55	75	130

$$\chi^2 = \sum_{a,b,c,d} \frac{(observed - expected)^2}{expected}$$

$$\chi^2 = \frac{(20-25.38)^2}{25.38} + \frac{(40-34.62)^2}{34.62} + \frac{(35-29.62)^2}{29.62} + \frac{(35-40.38)^2}{40.38} = 3.67$$

16

Chi Squared Distribution



17

Chi Squared Test

- The test statistic is compared with a Chi Square distribution having a particular number of degrees of freedom and a p-value is obtained
- For the example data the test statistic is = 3.67 and the corresponding p-value is = 0.055
- Thus we would fail to reject the null hypothesis

18

Chi Squared Test Assumptions

- Random sample
- Independent observations in the cells
- Expected cell counts need to be >= 5

19

Paired Samples

- Sometimes we may have two groups of data which are not independent samples
- One of the most common scenarios is when measurements are taken before and after some intervention on the same individuals
- For example individuals treated with a new drug
  - Blood metabolite measurements are taken before and after the drug has been taken
  - Test to see if the drug has changed the mean blood metabolite measurement

20

Paired Samples

- Cannot treat the groups as independent as the same individuals are in the *before* and *after* groups
- The data are paired
- H_0: There is no difference between the means

  H_0: $\mu_1 - \mu_2 = 0$
- H_1: There is some difference between the means

  H_1: $\mu_1 - \mu_2 \neq 0$

21

Paired sample example

- Measurements of a blood metabolite taken before and after a treatment
- The measurements are not independent

Individual	1	2	3	4	…	…	n
Before	40.1	19.1	21.2	18.4	…	…	26.3
After	45.1	22.4	29.1	19.0	…	…	33.3
Difference	5	2.7	7.9	0.6	…	…	7.0

- Can now carry out a 1 sample t test for the hypothesis:

H_0: $\mu = 0$

μ is the population mean difference

22

Paired sample example

- We have reduced the data to one sample by calculating the difference of each pair
- Some statistical programs will allow you to specify that the data are paired so the difference won't need to be calculated beforehand
- NOTE: we are still assuming that the people are independent, as in the usual t test
- There are other paired data tests e.g. McNemar Test

23

Parametric Tests

- So far we have used probability distributions, and **assumed** that if the sample size is large enough, then the data will match some underlying distribution, e.g. the t distribution, the chi-squared distribution, etc.
- These tests are referred to as parametric tests
- If possible we want to use parametric tests as they are often the most powerful tests for a given data set
- But sometimes the test assumptions will not be satisfied, particularly the assumption of Normality

24

Test Assumptions Not Satisfied

- What to do if your data is not roughly Normally distributed?
- For example:
  - Censored – e.g. cut off at zero
  - Bimodal or multimodal
  - Asymmetrical – skewed to the left or right
  - Non-numeric data, such as ordered categorical
  - Groups with different variance

25

Alternative: Transforming your data

- Sometimes the data don't look Normally distributed but we can transform the data so that it looks Normal
- A number of different transformations are possible:
  - Take the square root of each of the data points
  - Take the square of each of the data points
  - Take the logarithm of each of the data points

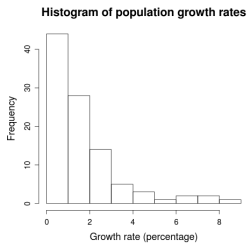
26

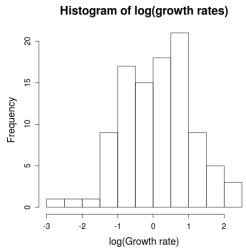
Alternative: Transforming your data

- The choice will depend on the shape of the original data
- Remember that all tests carried out on the transformed data relate to the transformed data and not the original data

27

Alternative: Transforming your data





28

Non-Parametric Tests

- If we cannot find a transformation that allows us to use a parametric test, then the next alternative is to use a non-parametric test

29

Non-Parametric Tests

- The z test, etc. depend on being able to define the data using the parameters: the mean and the variance
- We can avoid this by using **ranks**, and instead of comparing means we compare **medians**
- There are non-parametric versions of several of the most popular parametric tests

30

Non-Parametric Tests: Guidelines

- Ranking the data results in a loss of information: we now have no information about how spread out the data is, just about the ordering of the data points
- Non-parametric tests will have less power
- They can be computationally easier for simple cases (small samples)
- They are not completely assumption-free

31

Mann-Whitney U test

- If you want to compare two samples, but they are not normally distributed
- The Mann-Whitney U test is the non-parametric alternative
- H_0: median of group 1 = median of group 2
- H_1: median of group 1 ≠ median of group 2

32

Mann-Whitney U test

- Procedure:
  - Pool the two groups, and rank all the data points
  - In each of the two groups, sum the ranks
  - Test statistic: U is then calculated from these sums and the sample sizes in the groups

33

Non-Parametric Tests

Parametric Test	Non-Parametric Version
t test: Two sample	Mann-Whitney U test
	Wilcoxon rank sum test
t test: Paired	Sign test
	Wilcoxon signed rank test
ANOVA	Kruskal-Wallis test
Pearson correlation	Spearman correlation

34

Exact Tests

- Another type of test is an exact test
- An exact test is a test where the distribution of the test statistic is exactly calculable, either by complete enumeration or by simulation
- Using an exact test, we get an exact p-value

35

Fisher's Exact test

- Suppose we have a contingency table in which we are comparing the side effect of a drug, with a placebo
- When at least one expected cell count is low (<5), Fisher's exact test is usually employed
- Given that we have a certain number of people in each category (marginal totals), this table can be seen as one possible instance of all possible tables

36

Fisher's Exact test

- It is clearly unbalanced – it has 2:4 in the first row, and 7:3 in the second row
- How many tables **more unbalanced** than that are possible, while keeping the marginal totals the same?

	Drug	Placebo	Total
Side Effects	2	4	6
No Side Effects	7	3	10
Total	9	7	16

37

Fisher's Exact test

- More extreme:
  - One with 1:5 and 8:2

	Drug	Placebo	Total
Side Effects	1	5	6
No Side Effects	8	2	10
Total	9	7	16

38

Fisher's Exact test

- More extreme:
  - One with 1:5 and 8:2
  - One with 0:6 and 9:1

	Drug	Placebo	Total
Side Effects	0	6	6
No Side Effects	9	1	10
Total	9	7	16

39

Fisher's Exact test

- Here we've made a table to show the permutations for each of the four cells
- Calculate a probability for each table – this is done with a formula based on the hypergeometric distribution
- Add up the probabilities of **as extreme, or more extreme** tables to obtain the exact total probability

a	b	c	d	probability
. . .	. . .	. . .	. . .	. . .
3	3	6	4	…
2	4	7	3	p_{obs}
1	5	8	2	p_1
0	6	9	1	p_2

Exact p-value = $p_{obs} + p_1 + p_2$

40

Permutation tests

- These are a type of Exact test
- They follow the same principle of *shuffling* the data
- With a permutation test, we obtain the distribution of the test statistic under the null hypothesis by calculating all possible values for the test statistic by rearranging the labels of the observed data points

41

Permutation test: example

- Two sample t test:
- Remember our hypotheses:

H_0: $\mu_1 = \mu_2$

H_1: $\mu_1 \neq \mu_2$

Patient ID	Group 1	Patient ID	Group 2
01	71.2	02	68.0
03	73.6	04	75.6
05	62.3	06	74.5
07	75.4	08	65.9
09	74.9	10	67.5
11	68.3	12	67.4
. . .	. . .	. . .	. . .

42

Permutation test: example

- Two sample t test:
- H_0: the mean is the same in both groups
- So, if we swapped people from one group to the other, this should not affect the mean

Patient ID	Group 1	Patient ID	Group 2
01	71.2	02	68.0
03	73.6	04	75.6
05	62.3	06	74.5
07	75.4	08	65.9
09	74.9	10	67.5
11	68.3	12	67.4
.	.	.	.
.	.	.	.
.	.	.	.

43

Permutation test: example

- Procedure:

Permute the data points between groups, that is each point has a 50% chance to swap group

Keeping the sample sizes in each group the same

Perform t test, and see if the test statistic is larger

Patient ID	Group 1	Patient ID	Group 2
01	71.2	02	68.0
03	73.6	04	75.6
05	62.3	06	74.5
07	75.4	08	65.9
09	74.9	10	67.5
11	68.3	12	67.4
.	.	.	.
.	.	.	.
.	.	.	.

44

Permutation test: example

- Original t = 2.15

Patient ID	Group 1	Patient ID	Group 2
01	71.2	02	68.0
03	73.6	04	75.6
05	62.3	06	74.5
07	75.4	08	65.9
09	74.9	10	67.5
11	68.3	12	67.4
.	.	.	.
.	.	.	.
.	.	.	.

45

Permutation test: example

- Original t = 2.15
- t_{p1} = 1.94

Patient ID	Group 1	Patient ID	Group 2
01	65.9	02	68.0
03	74.5	04	75.6
05	55.1	06	73.6
07	67.4	08	71.2
09	74.9	10	67.5
11	68.3	12	75.4
.	.	.	.
.	.	.	.
.	.	.	.

46

Permutation test: example

- Original t = 2.15
- t_{p1} = 1.94
- t_{p2} = 0.98
- etc...
- 1000 permutations and 39 gave a larger t statistic

Patient ID	Group 1	Patient ID	Group 2
01	66.6	02	68.0
03	71.2	04	75.6
05	55.1	06	73.6
07	67.5	08	74.5
09	74.9	10	67.4
11	75.4	12	68.3
.	.	.	.
.	.	.	.
.	.	.	.

47

Permutation tests

- The **empirical p-value** is calculated by counting the proportion of times that your permuted data sets show a larger test statistic than the one you saw in the original data

$$\text{empirical p-value} = \frac{R}{N}$$

- R = the number of permutations where your test statistic was exceeded
- N = the total number of permutations

48

Permutation tests

- The **empirical p-value** is calculated by counting the proportion of times that your permuted data sets show a larger test statistic than the one you saw in the original data

$$\text{empirical p-value} = \frac{R}{N}$$

$$= \frac{39}{1000} = 0.039$$

- Permutation tests are similar to the exact test, except we don't run through ***all*** permutations, just a representative subset of them

49

Summary I

- Comparing Proportions: Chi squared test
  - What hypothesis is being tested
  - The assumptions underlying the test
- Paired Data
  - Sometimes can reduce a paired dataset to a single sample, by taking differences

50

Summary II

- When the assumptions don't hold
  - Transforming Data
  - Non-parametric tests
- Exact Tests
  - Fisher's Exact Test (expected cell counts low)
  - Permutation Test (swapping labels)

51

One-Way ANOVA

Trinity College Dublin
The University of Dublin

Introduction

- Analysis of variance (ANOVA) is a method for testing the hypothesis that there is no difference between three or more population means
- Often used for testing the hypothesis that there is no difference between a number of treatments

2

Independent Two Sample t-test

- Recall the independent two sample t-test which is used to test the null hypothesis that the population means of two groups are the same
- Let $\bar{x}_1$ and $\bar{x}_2$ be the sample means of the two groups, then the test statistic for the independent t-test is given by:

$$t = \frac{\bar{x}_1 - \bar{x}_2 - 0}{se(\bar{x}_1 - \bar{x}_2)}$$

- The test statistic is compared with the t-distribution with $(n_1 + n_2 - 2)$ degrees of freedom (df)

3

Why Not Use t-test Repeatedly?

- The t-test, which is based on the standard error of the difference between two means, can only be used to test differences between *two* means
- With more than two means, could compare each mean with each other mean using t-tests
- Conducting multiple t-tests can lead to severe inflation of the Type I error rate (false positives) and is NOT RECOMMENDED

4

Why Not Use t-test Repeatedly?

- ANOVA is used to test for differences among several means without increasing the Type I error rate
- The ANOVA uses data from all groups to estimate standard errors, which can increase the power of the analysis

5

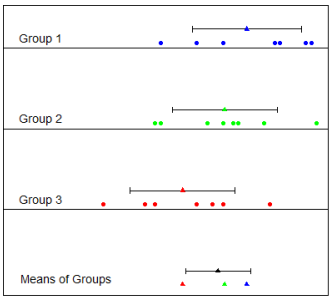
Why Look at Variance When Interested in Means?

- Basic Idea
  - Calculate the mean of the observations within each group
  - Compare the variance of these means to the average variance within each group
  - As the means become more different, the variance among the means increases

6

Why Look at Variance When Interested in Means?

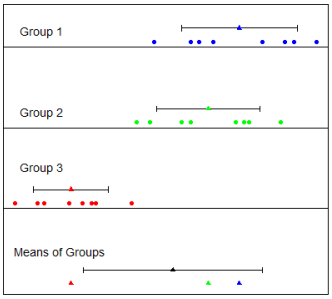
- The variability within each sample is approximately the same
- The variability in the mean values of the samples is consistent with the variability within the individual samples



7

Why Look at Variance When Interested in Means?

- The variability in the sample means is much larger than would be expected given the variability within each of the samples



8

Why Look at Variance When Interested in Means?

- To distinguish between the groups, the variability between (or among) the groups must be greater than the variability of, or within, the groups
- If the within-groups variability is large compared with the between-groups variability, any difference between the groups is difficult to detect
- To determine whether or not the group means are significantly different, the variability between groups and the variability within groups are compared

9

One-Way ANOVA and Assumptions

- One-Way ANOVA
  - When there is only one categorical variable which denotes the groups and only one measurement variable (numerical), a one-way ANOVA is carried out
  - For a one-way ANOVA the observations are divided into I mutually exclusive categories, giving the one-way classification

10

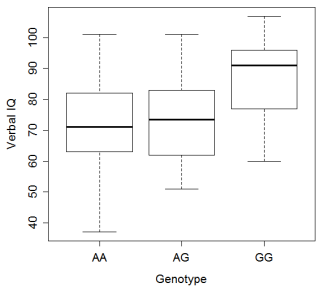
One-Way ANOVA and Assumptions

- ASSUMPTIONS
  - Each of the populations is Normally distributed with the same variance (homogeneity of variance)
  - The observations are sampled independently, the groups under consideration are independent

ANOVA is robust to moderate violations of its assumptions, meaning that the probability values (P-values) computed in an ANOVA are sufficiently accurate even if the assumptions are moderately violated

11

Simulated Data Example



- 54 observations
- 18 AA observations
  mean IQ for AA = 71.6
- 18 AG observations
  mean IQ for AG = 72.7
- 18 GG observations
  mean IQ for GG = 87.1

12

Introduction of Notation

- Consider I groups, whose means we want to compare
- Let $n_i, i = 1, 2, \ldots, I$ be the sample size of group i
- For the simulated verbal IQ and genotype data, ($I = 3$), representing the three possible genotypes at the particular locus of interest. Each person in this data set, as well as having a genotype, also has a verbal IQ score

13

Null Hypothesis for ANOVA

- Want to examine if the mean verbal IQ score is the same across the 3 genotype groups
  - Null hypothesis is that the mean verbal IQ is the same in the three genotype groups

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

14

Within-Groups Variance

- Remember assumption that the population variances of the three groups is the same
- Under this assumption, the three variances of the three groups all estimate this common value
  - True population variance = σ^2
  - Within-groups variance
    = within-groups mean square
    = error mean square
    = s_w^2

15

Within-Groups Variance

- For groups with equal sample size this is defined as the average of the variances of the groups

$$s_w^2 = \frac{1}{I}\sum_{i=1}^{I} s_i^2 = \frac{1}{I}\sum_{i=1}^{I}\sum_{j=1}^{n_i}\left(\frac{(x_{ij}-\bar{x}_i)^2}{n_i-1}\right)$$

x_{ij} = observation *j* in group *i*

$\bar{x}_1, \bar{x}_2, \bar{x}_3$ = sample means of the genotype groups AA, AG, GG

16

Within-Groups Variance

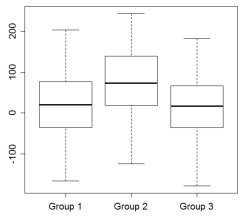
- In our example data

$$\begin{aligned} s_w^2 &= \frac{(s_{AA}^2 + s_{AG}^2 + s_{GG}^2)}{3} \\ &= \frac{(247.3 + 169.2 + 229.3)}{3} \\ &= 215.3 \end{aligned}$$

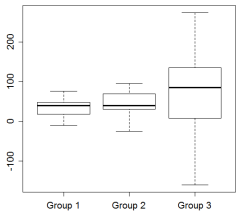
17

Within-Groups Variance

- Since the <u>population variances are assumed to be equal</u> the estimate of the population variance, derived from the separate within-group estimates, is valid whether or not the null hypothesis is true



EQUAL



NOT EQUAL

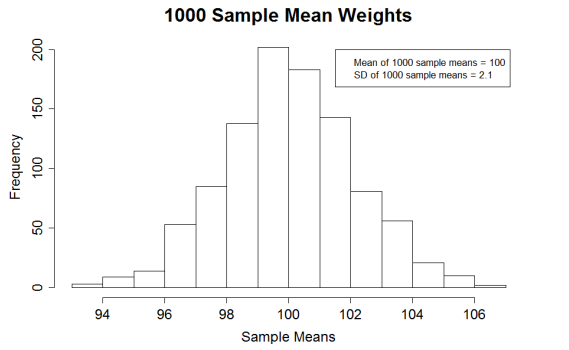
18

Between-Groups Variance

- If the null hypothesis is true:
  - the three groups can be considered as random samples from the same population
  - assumed equal variances, and because the null hypothesis is true, then the population means are equal
  - The three means are three observations from the same sampling distribution of the mean

19

Between-Groups Variance



20

Between-Groups Variance

- The sampling distribution of the mean has variance

$$\sigma^2/n$$

- This gives a second method of obtaining an estimate of the population variance (n = number of observations in each group)
- The observed variance of the treatment means is an estimate of σ^2/n and is given by:

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

21

Between-Groups Variance

- For equal sample sizes, the between-groups variance is then given by:

$$s_b^2 = n \sum_{i=1}^{I} \frac{(\bar{x}_i - \bar{x})^2}{I-1}$$

$$= 18 \times \frac{(71.6-77.1)^2 + (72.7-77.1)^2 + (87.1-77.1)^2}{2}$$

$$= 1346$$

Means of the 3 groups

Mean of all the observations

22

Unequal Sample Sizes

- There are adjustments to these formulae when the sample sizes are not all equal in the groups:
  - Within Groups variance:

$$s_w^2 = \sum_{i=1}^{I} \frac{(n_i - 1)s_i^2}{N - I}$$

  - Between Groups Variance:

$$s_b^2 = \sum_{i=1}^{I} n_i \frac{(\bar{x}_i - \bar{x})^2}{I-1}$$

23

Testing the Null Hypothesis, F-test

- If the null hypothesis is true then
  - the between-groups variance s_b^2
  - and the within-groups variance s_w^2
  - are both estimates of the population variance σ^2

24

Testing the Null Hypothesis, F-test

- If the null hypothesis is *NOT* true then
  - the population means are not all equal
  - then s_b^2 will be greater then the population variance, σ^2
  - it will be increased by the treatment differences

25

Testing the Null Hypothesis, F-test

- To test the null hypothesis we compare the ratio of s_b^2 and s_w^2 using an F-test
- F statistic is given by:

$$F = \frac{s_b^2}{s_w^2}$$

with *I-1* and *I(n-1)* degrees of freedom

26

Testing the Null Hypothesis, F-test

- Another way of thinking about this ratio: $F = \frac{s_b^2}{s_w^2}$

$$F = \frac{\text{variability due to treatment effect and variability due to chance}}{\text{variability due to chance}}$$

27

F Distribution and F-test

- The F distribution is the continuous distribution of the ratio of two estimates of variance
- The F distribution has two parameters: degrees of freedom numerator (top) and degrees of freedom denominator (bottom)
- The F-test is used to test the hypothesis that two variances are equal

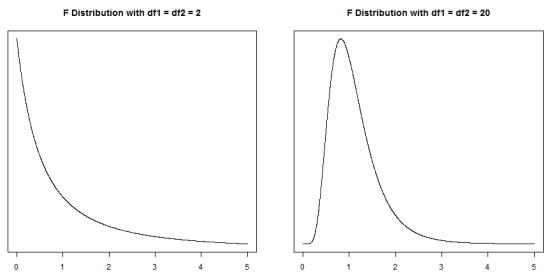
28

F Distribution and F-test

- The validity of the F-test is based on the requirement that the populations from which the variances were taken are Normal
- In the ANOVA, a one-sided F-test is used

29

F Distribution



30

Output: ANOVA Table

- For the genotype, verbal IQ data:

One-way ANOVA: IQ versus Genotype

Source	DF	SS	MS	F	P
Genotype	2	2691	1346	6.25	0.004
Error	51	10979	215		
Total	53	13671			

S = 14.67 R-Sq = 19.69% R-Sq(adj) = 16.54%

31

Output: ANOVA Table

- For the genotype, verbal IQ data:

Between-Groups
Between treatments

One-way ANOVA: IQ versus Genotype

Source	DF	SS	MS	F	P
Genotype	2	2691	1346	6.25	0.004
Error	51	10979	215		
Total	53	13671			

P-Value

F Statistic

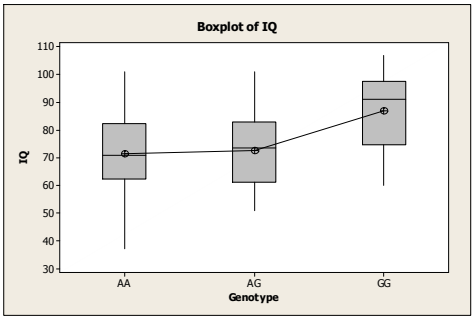
Within-Groups
Residual Variation

I-1 = 3-1 = 2, since 3 genotype groups, AA, AG, GG

I(n-1)= 3(18-1) = 51

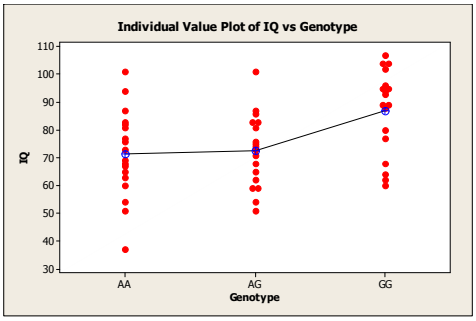
32

Output: Plots



33

Output: Plots



34

Assumption Checking

- Homogeneity of variance = homoscedasticity
  - The dependent variable (quantitative measurement) should have the same variance in each category of the independent variable (qualitative variable)
  - Needed since the denominator of the F-ratio is the within-group mean square, which is the average of the group variances

35

Assumption Checking

- ANOVA is robust for small to moderate departures from homogeneity of variance, especially with equal sample sizes for the groups
- Rule of thumb: the ratio of the largest to the smallest group variance should be 3:1 or less, but be careful, the more unequal the sample sizes the smaller the differences in variances which are acceptable

36

Assumption Checking

- Testing for homogeneity of variance
  - Levene's test of homogeneity of variance
  - Bartlett's test of homogeneity of variance (Chi-square test)
  - Examine boxplots of the data by group, will highlight visually if there is a large difference in variability between the groups
  - Plot residuals versus fitted values and examine scatter around zero,
    residuals = observations – group mean
    group mean = fitted value

37

Assumption Checking

Normality Assumption

- The dependent variable (measurement, quantitative variable) should be Normally distributed in each category of the independent variable (qualitative variable)
- Again ANOVA is robust to moderate departures from Normality

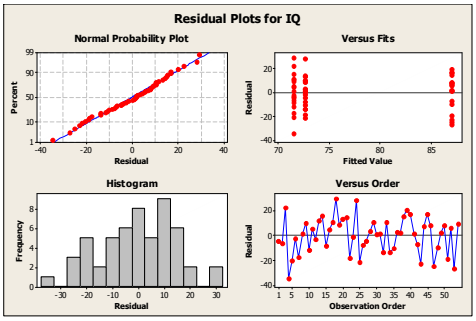
38

Assumption Checking

- Checking the Normality assumption
  - Boxplots of the data by group allows for the detection of skewed distributions
  - Quantile-Quantile plots (QQ plots) of the residuals, which should give a 45-degree line on a plot of observed versus expected values,
  - Usual tests for Normality may not be adequate with small sample sizes, insufficient power to detect deviations from Normality

39

Assumption Checking: Plots



40

What to do with a Significant ANOVA Result (F-test)

- If the ANOVA is significant and the null hypothesis is rejected, the only valid inference that can be made is that at least one population mean is different from at least one other population mean
- The ANOVA does not reveal which population means differ from which others

41

What to do with a Significant ANOVA Result (F-test)

- Only think about investigating differences between individual groups when the overall comparison of groups (ANOVA) is significant, or that you had intended particular comparisons at the outset
- Need to consider whether the groups are ordered or not

42

What to do with a Significant ANOVA Result (F-test)

- Modified t-test:
  - based on the pooled estimate of variance from all the groups (within groups, residual variance in the ANOVA table), not just the pair being considered
- Least significant difference:
  - The least difference between two means which is significant is used
  - Arrange the treatment means in order of magnitude and the difference between any pair of means can be compared with the least significant difference

43

What to do with a Significant ANOVA Result (F-test)

- Tukey's honest significance test
  - The test compares the means of every group to the means of every other group and corrects for the multiple comparisons that are made
- Linear Trend
  - When the groups are ordered we don't want to compare each pair of groups, but rather investigate if there is a trend across the groups (linear trend)
- There are many other tests available...

44

Assumption Checking

- If data are not Normal or don't have equal variances, can consider
  - transforming the data, (eg. log, square root)
  - non-parametric alternatives

45

Non Parametric ANOVA – Kruskal-Wallis Test

- ANOVA is the more general form of the t-test
- Kruskal-Wallis test is the more general form of the Mann-Whitney test
- Kruskal-Wallis test:
  - doesn't assume normality, compares medians
  - based on ranking the data (some information is lost, less powerful)

46

Summary I

- The hypothesis that the means of at least three groups are the same is tested using a one-way ANOVA
- An ANOVA assumes
  - the observations are independent and random
  - the data is Normally distributed in each group
  - the variances are the same across the groups
- The ANOVA works by comparing estimates of variance which should be the same if the null hypothesis of equal means is true

47

Summary II

- Within-Groups Variance
  - This is the average of the variances in each group
  - This estimates the population variance regardless of whether or not the null hypothesis is true
  - This is also called the error mean square, or the within-groups mean square

48

Summary III

- Between-Groups Variance
  - This is calculated from the variance between groups
  - This estimates the population variance, if the null hypothesis is true
  - This is also called the residual variance or the between-groups mean square

49

Summary IV

- Use an F-test to compare these two estimates of variance
- Relationship with t-test
  - If the one-way ANOVA is used for the comparison of two groups only, the analysis is exactly and mathematically equivalent to the use of the independent t-test, (the F statistic is exactly the square of the corresponding t statistic)

50

Correlation

Outline

- What is correlation?
- Pitfalls
- Looking at the data
- The correlation coefficient
- Assessing significance
- Non-parametric tests

2

Correlation

- Measures the degree to which two (or more) variables change together
- Linear (straight line) relationship → *correlation*
- Nonlinear relationship → *association*
- Captured by a single number
- Correlation/association does *not* imply causation!

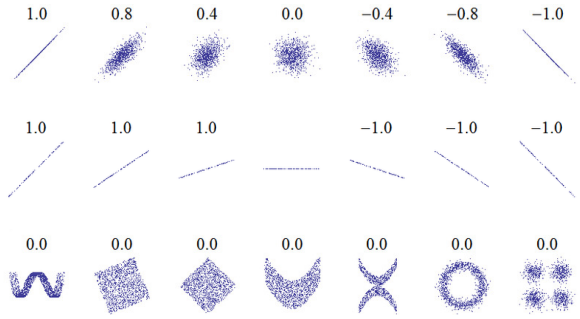
3

Correlation and Dependence

Purpose	Y Random X Controlled	Y Random X Controlled
Investigate dependence	Model 1 Regression	Model 2 Regression
Investigate relationship	Meaningless	Correlation

4

Examples of Correlation and Non-Linear Relationships



5

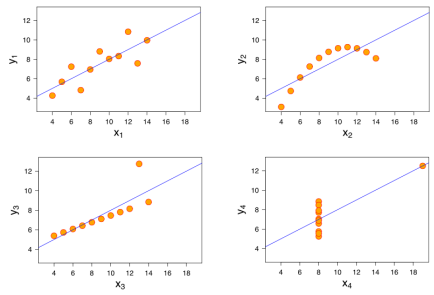
Data Quality

- The measure of correlation is highly sensitive to anomalies in the data:
  - Outliers
  - Clustered data points
  - Nonlinearity
  - Spurious correlations/associations

6

Look at the Data!

- All these datasets have the same mean, variance, correlation coefficient and regression line:

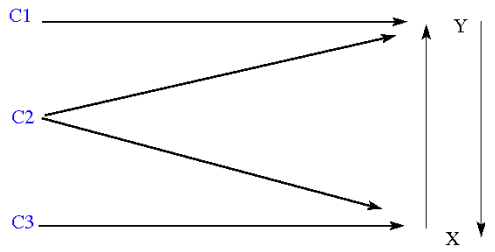


- Anscombe, Francis J. (1973) Graphs in statistical analysis. American Statistician, 27

7

Dependency Structure

- Correlation is not causation...



8

Pearson Correlation

- Measures the degree of *linear relationship* between two sets of *n* measurements,
  eg. weights W_i and heights H_i
- Varies between -1 and 1
- A value of 1 means perfect linear correlation
- A value of 0 means no correlation
- A value of -1 means perfect linear anticorrelation
- The Pearson sample correlation coefficient, r is an **estimator** of the population coefficient, ρ *(rho)*

9

Toy Example

- Measure the heights and weights of 3 people:

ID	Height (m)	Weight(kg)
P1	1.65	66
P2	1.75	70
P3	1.85	74

10

Computing the Correlation Value I

- Start with two sets of measurements, (heights and weights) H_i and W_i:
- Subtract the means to get data centred with mean 0:

$$x_i = H_i - \bar{H} \qquad y_i = W_i - \bar{W}$$

11

Toy Example

- First calculate the means:

$$\bar{H} = 1.75m \quad \bar{W} = 70kg$$

- Subtract the means to centre the data on zero:

$$x_1 = -0.1; \quad x_2 = 0; \quad x_3 = 0.1$$
$$y_1 = -4; \quad y_2 = 0; \quad y_3 = 4$$

12

Computing the Correlation Value II

- Divide by sample standard deviations to get the standardised heights and weights:

$$u_i = \frac{x_i}{s_x} \qquad v_i = \frac{y_i}{s_y}$$

13

Toy Example

- Work out the standard deviations

$$s_x = \sqrt{\frac{1}{(3-1)}(-0.1^2 + 0^2 + 0.1^2)} = 0.1$$

$$s_y = \sqrt{\frac{1}{(3-1)}(-4^2 + 0^2 + 4^2)} = 4$$

- Divide by the standard deviations to get our standardised variables:

$$u_1 = -1; \quad u_2 = 0; \quad u_3 = 1$$

$$v_1 = -1; \quad v_2 = 0; \quad v_3 = 1$$

14

Computing the Correlation Value III

- Consider the product of standardised height and weight for a single person:

$$c_3 = u_3 v_3$$

- Positive Correlation:

$$u_3 > 0; \quad v_3 > 0 \quad \rightarrow \quad c_3 > 0$$
$$u_3 < 0; \quad v_3 < 0 \quad \rightarrow \quad c_3 > 0$$

- Negative Correlation:

$$u_3 > 0; \quad v_3 < 0 \quad \rightarrow \quad c_3 < 0$$
$$u_3 < 0; \quad v_3 > 0 \quad \rightarrow \quad c_3 < 0$$

15

The Correlation Coefficient: Version I

- The overall correlation is the mean (almost!) of the n products c_i:

$$r = \frac{1}{(n-1)} \sum_{i=1}^{n} c_i$$

- Question: Why do we divide by n - 1 instead of by n?

16

Toy Example

- Work out the c values:

$$c_1 = u_1 v_1 = -1 \times -1 = 1$$
$$c_2 = u_2 v_2 = 0 \times 0 = 0$$
$$c_3 = u_3 v_3 = 1 \times 1 = 1$$

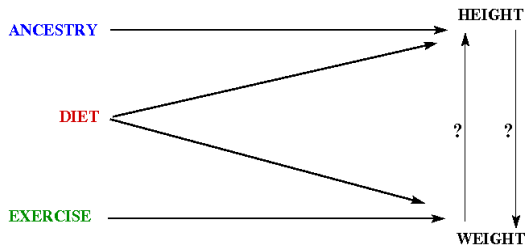
- Work out the correlation coefficient:

$$r = \frac{1}{(3-1)}(1+0+1) = 1$$

17

Dependency Structure

- Correlation is not causation...



18

The Correlation Coefficient: Version II

- You may also have seen the Pearson correlation written in terms of the original measurements
- In our case, these were the weights, W_i and the heights, H_i:

$$r = \frac{\sum_{i=0}^{n}(W_i - \bar{W})(H_i - \bar{H})}{\sqrt{\sum_{i=0}^{n}(W_i - \bar{W})^2}\sqrt{\sum_{i=0}^{n}(H_i - \bar{H})^2}}$$

19

Assessing Significance I

- The correlation coefficient r is an estimator of the population coefficient ρ
- Null Hypothesis: $\rho = 0$
- Assumption: the variables X and Y are normally distributed

20

Assessing Significance II

- Consider the distribution of the quantity:

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

- If $\rho = 0$, this is distributed as Student's t, with two degrees of freedom

$\rightarrow$ we can obtain the critical value of t and hence the critical value of r

21

The Spearman Coefficient

- The Spearman coefficient measures correlation between *rank ordered* data
- Can handle non–linear (monotonic) data
- The Spearman coefficient is *not* an estimator of any simple population parameter

22

Computing the Spearman Coefficient

- Replace the data values by their ranks in the expression for the Pearson coefficient:

$$s = \frac{\sum_{i=0}^{n}(R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum_{i=0}^{n}(R_i - \bar{R})^2}\sqrt{\sum_{i=0}^{n}(S_i - \bar{S})^2}}$$

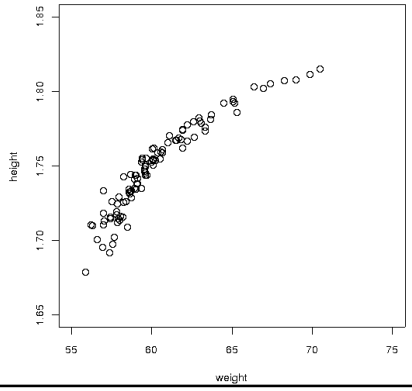
23

Toy Example

ID	Height(m)	Rank(R)	Weight(kg)	Rank(S)
P1	1.65	1	66	1
P2	1.75	2	70	2
P3	1.85	3	74	3
		$\bar{R} = 2$		$\bar{S} = 2$

24

Example: Non Linear Data



25

R Output for Spearman's Rank Correlation

Spearman's rank correlation rho

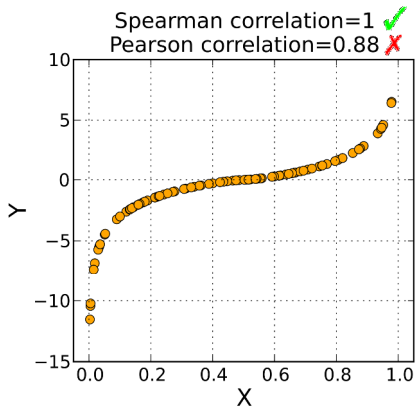
data: height and weight

S = 4328, p-value < 2.2e-16

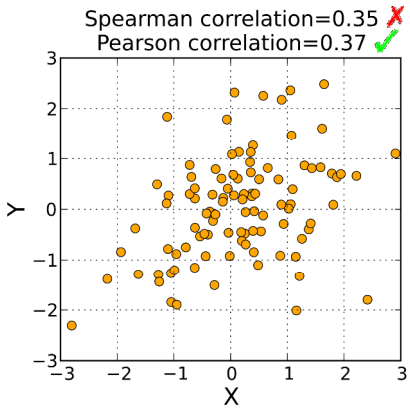
alternative hypothesis: true rho is not equal to 0

sample estimates:
rho
0.9740294

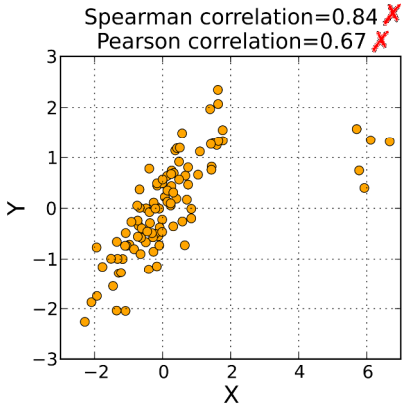
26



27



28



29

Summary I

- Correlation measures how two variables change together
- It does *not* imply causation
- It is sensitive to anomalies in the data
- Data should always be examined visually before doing a correlation analysis

30

Summary II

- The Pearson coefficient r measures linear relationships and varies between -1 and +1
- If both variables are normally distributed we can determine the statistical significance
- The Spearman coefficient measures non-linear monotonic relationships and varies between -1 and +1

31

Regression

Trinity College Dublin
The University of Dublin

Introduction

- Correlation and regression – for quantitative/numeric variables
  - Correlation: assessing the association between quantitative variables
  - Simple linear regression: description and prediction of one quantitative variable from another

2

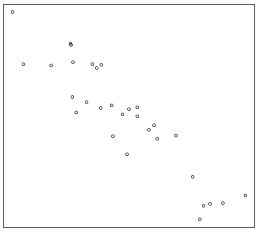
Introduction

- Simple linear regression: only considering linear (straight-line) relationships
- When considering correlation or carrying out a regression analysis between two variables always plot the data on a scatter plot first

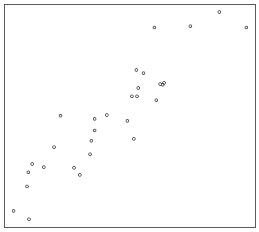
3

Scatter Plots

Linear



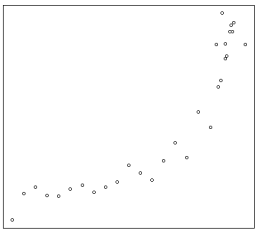
Linear



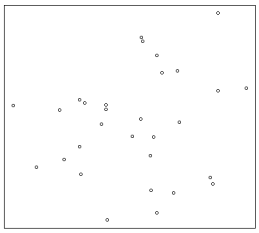
4

Scatter Plots

Non-Linear



No Relationship



5

Simple Linear Regression

- Data on two quantitative variables
- Aim is to ***describe*** the relationship between the two variables and/or to ***predict*** the value of one variable when we only know the other variable
- Interested in a linear relationship between the two variables X and Y

Y	Predicted Variable	Dependent Variable	Response Variable	Outcome Variable
X	Predictor Variable	Independent Variable	Carrier Variable	Input Variable

6

Simple Linear Regression

- Simple linear regression - when there is only one predictor variable, which we will consider here
- Multiple or multivariate regression - when there is more than one predictor variable

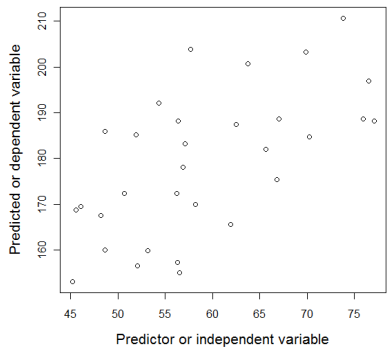
7

Simple Linear Regression

- The aim is to fit a straight line to the data that predicts the mean value of the dependent variable (*Y*) for a given value of the independent variable (*X*)
- Intuitively this will be a line that minimizes the distance between the data and the fitted line
- Standard method is *least squares regression*
- Notation: n pairs of data points, (x_i, y_i)

8

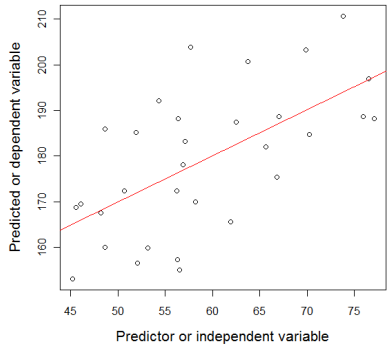
Two Quantitative Variables



9

Two Quantitative Variables,

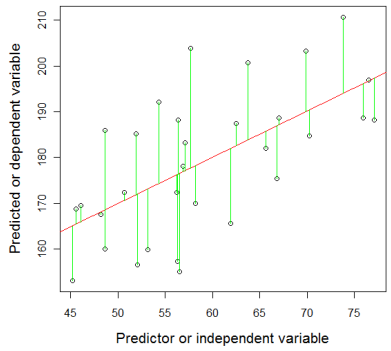
Regression Line



10

Two Quantitative Variables,

Regression Line



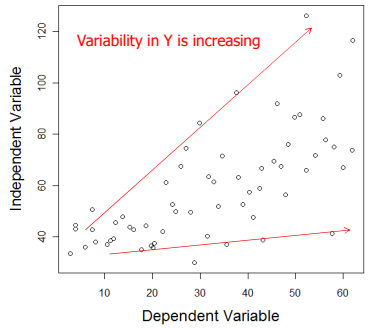
11

Linear Regression Assumptions

- The values of the dependent variable *Y* should be Normally distributed for each value of the independent variable *X* (needed for hypothesis testing and confidence intervals)
- The variability of *Y* should be the same for each value of *X* (homoscedasticity)

12

Linear Regression Assumptions



13

Linear Regression Assumptions

- The relationship between the two variables should be linear
- The observations should be independent
- Values of *X* do not have to be random
- Values of *X* don't have to be Normally distributed

14

Linear Regression Assumptions

- It is easier to check many of these assumptions after the regression has been carried out
- Use residuals to do this and we will return to these later

15

Linear Regression Assumptions

- The straight line or linear relationship is described by the equation for a straight line

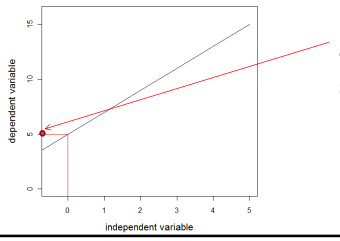
$$y = a + bx$$

Dependent variable

Independent variable

Intercept

Slope



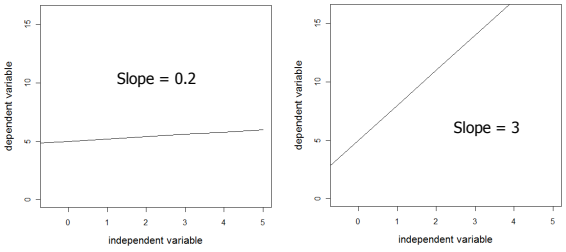
Intercept, value of y when x = 0, a = 5, y = 5 + b*0

Slope of the line, b = 2 here

y = 5 + 2x

16

Slopes



Same intercept = 5

17

Least Squares Regression

- No line could pass through all the data points in our example
- We want the best "average" equation (*regression equation*) that would represent a line through the middle of the data, this is the *regression line:*

$$\hat{y}_i = \hat{a} + \hat{b}x_i$$

- The constants a, the intercept and b, the slope or regression coefficient are computed using the method of least squares

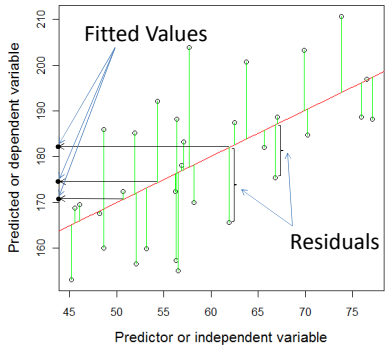
18

Least Squares Regression

- Fitted value = value of Y given by the line for any value of the variable X
  (remember what a fitted value was in ANOVA)
- Residual = difference between the observed value of Y and the fitted value
  (again, remember what a residual was in ANOVA)
- Least squares aim: to minimize the sum of squares of the residuals

19

Fitted Values and Residuals



20

Least Squares Regression

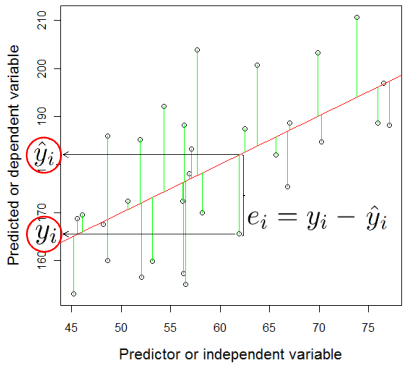
- At any point x_i, the corresponding point on the line is given by: $\hat{a} + \hat{b}x_i$

Regression equation: $\hat{y}_i = \hat{a} + \hat{b}x_i$

Residuals (errors): $e_i = y_i - \hat{y}_i$

21

Fitted Values and Residuals



22

Least Squares Regression

- Linear model:

$$y_i = \hat{a} + \hat{b}x_i + e_i, \quad e_i \sim \text{Normal}(0, \sigma^2)$$

- Note: if the errors/residuals are correlated or have unequal variances then least squares is not the best way to estimate the regression coefficient

23

Least Squares Regression

- Minimize the sum of squares *(S)* of the vertical distances of the observations from the fitted line (residuals)

$$S = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} [y_i - \hat{y}_i]^2$$

- In order to find the intercept and regression coefficient that minimize S the mathematical technique of differentiation is employed

24

Least Squares Regression

- The solution for these two equations results in the following two formulae for the estimates of the intercept and regression coefficients respectively:

$$\hat{a} = \bar{y} - \hat{b}\bar{x} \qquad \hat{b} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})}$$

- For the systolic blood pressure and age data in the previous plots:

$\hat{a}$ = 118.7

$\hat{b}$ = 1.0

x_i	y_i	$\hat{y}_i$	$e_i = y_i - \hat{y}_i$
49	186	168.7	17.3
46	169	165.7	3.3
58	170	177.9	-7.9
53	160	172.8	-12.8
:	:	:	:

25

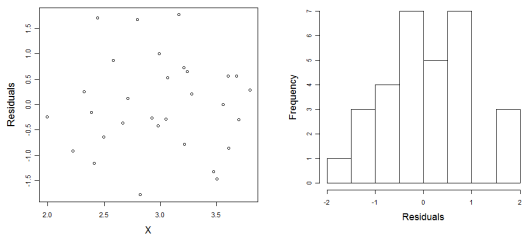
Residuals

- Checking assumptions:
  - Residuals should have a Normal distribution with zero mean
  - Plot *X* against residuals, looking for even scatter at all *X* values
  - Consider transformations of the data if these are not satisfied (eg., log, square root)

26

Residuals

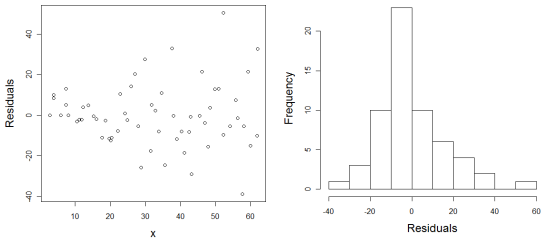
- These residuals appear reasonable



27

Residuals

- These residuals show increasing variability



28

Regression Coefficient b

- Regression coefficient:
  - this is the slope of the regression line
  - indicates the strength of the relationship between the two variables
  - interpreted as the expected change in y for a one-unit change in x

29

Regression Coefficient b

- Regression coefficient:
  - can calculate a standard error for the regression coefficient
  - can calculate a confidence interval for the coefficient
  - can test the hypothesis that $b = 0$, i.e., that there is no relationship between the two variables

30

Regression Coefficient b

- To test the hypothesis that $b = 0$, testing the hypothesis that there is no relationship between the *X* and *Y* variables, the test statistic is given by:

$$t = \frac{b}{se(b)}$$

comparing this ratio with a *t* distribution with $n-2$ degrees of freedom

- Can also calculate a confidence interval for b:

$$b \pm t_{0.975} se(b)$$

31

Intercept a

- Intercept:
  - the estimated intercept a gives the value of y that is expected when $x = 0$
  - often not very useful as in many situations it may not be realistic or relevant to consider $x = 0$
  - it is possible to get a confidence interval and to test the null hypothesis that the intercept is zero and most statistical packages will report these

32

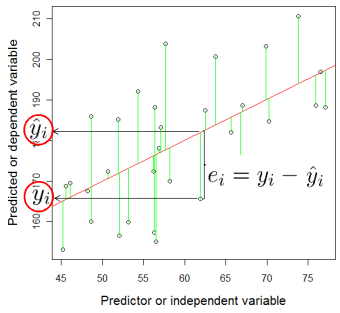
Coefficient of Determination, R-Squared

- The coefficient of determination or R-squared is the amount of variability in the data set that is explained by the statistical model
- Used as a measure of how good predictions from the model will be
- In linear regression R-squared is the square of the correlation coefficient

33

Residual Sum of Squares

$$\sum_i (y_i - \hat{y}_i)^2$$

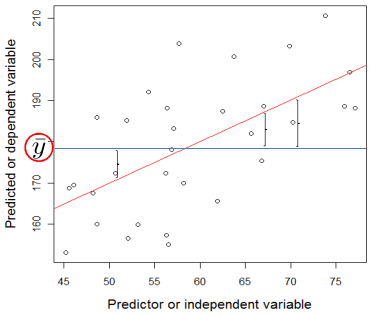


34

Regression Sum of Squares

$$\sum_i (\hat{y}_i - \bar{y})^2$$

Here $\bar{y}$ = 179

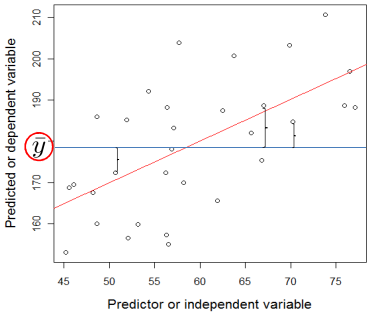


35

Total Sum of Squares

$$\sum_i (y_i - \bar{y})^2$$

Here $\bar{y}$ = 179



36

Sum of Squares

$$\sum_i (y_i - \bar{y})^2 = \sum_i (y_i - \hat{y}_i)^2 + \sum_i (\hat{y}_i - \bar{y})^2$$

Total sum of squares	=	Residual sum of squares	+	Regression sum of squares

Total Variation	=	Unexplained Variation	+	Explained Variation

37

Coefficient of Determination

Total sum of squares	=	Residual sum of squares	+	Regression sum of squares
Total Variation	=	Unexplained Variation	+	Explained Variation

$$\text{Coefficient of Determination} = \frac{\text{Explained Variation}}{\text{Total Variation}}$$

$$= \frac{\text{Regression SS}}{\text{Total SS}}$$

38

Coefficient of Determination
R-Squared

- Coefficient of determination
  = R-Squared
  = R^2
  = R-Sq
- The regression analysis can be displayed as an ANOVA table, many statistical packages present the regression analysis in this format

39

Coefficient of Determination
R-Squared

- R-Sq must lie between 0 and 1
- If it is equal to one then all the observed points must lie exactly on a straight line – no residual variability
- Often expressed as a percentage
- High R-squared says that the majority of the variability in the data is explained by the model (good!)

40

Adjusted R-Squared

- Sometimes an adjusted R-squared will be presented in the output as well as the R-squared
- Adjusted R-squared is a modification to the R-squared adjusting for the sample size and for the number of explanatory or predictor variables in the model (more relevant when considering multiple regression)
- The adjusted R-squared will only increase if the addition of the new predictor improves the model more than would be expected by chance

41

Residual Standard Deviation

- Remember the linear model formulation:

$$y_i = \hat{a} + \hat{b}x_i + e_i, \quad e_i \sim \text{Normal}(0, \sigma^2)$$

- The residual standard deviation is an estimate of the standard deviation of the residuals
- Measures the spread of the y values about the regression line

42

Residual Standard Deviation

- The residual standard deviation is a goodness-of-fit measure
- The smaller the residual standard deviation the closer the fit to the data

43

Output

Regression Analysis: Systolic BP versus Age

```
The regression equation is
Systolic BP = 119 + 1.02 Age

Predictor     Coef  SE Coef     T      P
Constant    118.73    14.29  8.31  0.000
Age         1.0205   0.2386  4.28  0.000

S = 12.5103   R-Sq = 38.7%   R-Sq(adj) = 36.6%

Analysis of Variance
Source           DF      SS      MS      F      P
Regression        1  2863.2  2863.2  18.29  0.000
Residual Error   29  4538.7   156.5
Total            30  7401.9

Unusual Observations
            Systolic
Obs    Age        BP     Fit  SE Fit  Residual  St Resid
 16  58.0    204.00  177.91    2.26     26.09      2.12R
R denotes an observation with a large standardized residual.
```

44

Output

Regression Analysis: Systolic BP versus Age

```
The regression equation is
Systolic BP = 119 + 1.02 Age

Predictor     Coef  SE Coef     T      P
Constant    118.73    14.29  8.31  0.000
Age         1.0205   0.2386  4.28  0.000

S = 12.5103   R-Sq = 38.7%   R-Sq(adj) = 36.6%
```

Estimate of the intercept, *a*

Estimate of the regression coefficient, *b*

S is the standard deviation of the residuals

R-Sq says that 38.7% of the variability in the data is explained by the model

Adjusted R-Sq, slightly lower as it has been adjusted for the number of predictor variables and the sample size

45

Output

Regression sum of squares

F-statistic = square of t-statistic for regression coefficient, $(4.28)^2$

```
Analysis of Variance
Source           DF      SS      MS      F      P
Regression        1  2863.2  2863.2  18.29  0.000
Residual Error   29  4538.7   156.5
Total            30  7401.9
```

Residual sum of squares

Total sum of squares

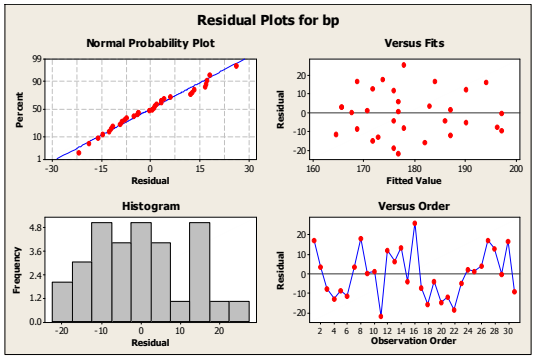
R^2 = Regression sum of squares / Total sum of squares = 2863.2/ 7401.9 = 0.387

Output

Statistical packages sometimes also highlight any unusual observations

```
Unusual Observations
          Systolic
Obs   Age        BP     Fit  SE Fit  Residual  St Resid
 16  58.0    204.00  177.91    2.26     26.09      2.12R
R denotes an observation with a large standardized residual.
```


Assumption Checking Output

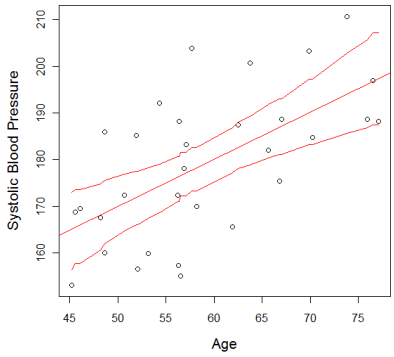


Confidence Interval on Fitted Values

- Can calculate a confidence interval on the fitted value: $\hat{y}_i$
- This is a confidence interval for the mean value of y, given a value of x
- The width of the confidence interval depends on the value of x_i and will be a minimum at $x_i = \bar{x}$ and will widen as $|x_i - \bar{x}|$ increases

49

95% Confidence Interval



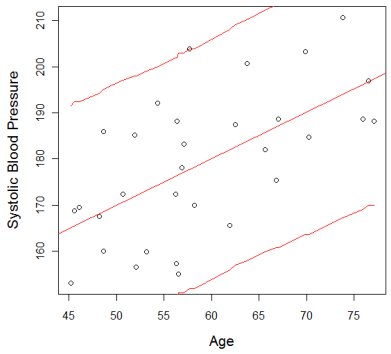
50

Prediction Interval for Future Values

- Can predict the range of possible values of y for a new independent value of x not used in the regression model
- The prediction interval describes the spread of the observations around the mean value: $\hat{y}_i$
- The prediction interval is wider than the confidence interval
- The interval widens with distance from the mean value of x, but is not so obvious to see

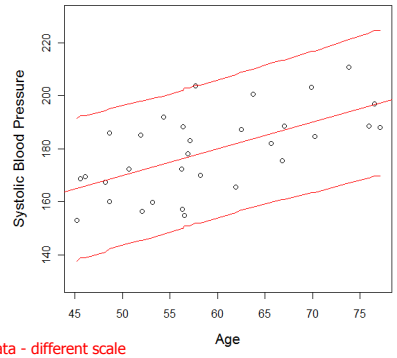
51

95% Prediction Interval



52

95% Prediction Interval



Same data - different scale

53

Interpolation and Extrapolation

- *Interpolation*
  - Making a prediction for *Y* within the range of values of the predictor *X* in the sample used in the analysis
  - Generally this is fine
- *Extrapolation*
  - Making a prediction for *Y* outside the range of values of the predictor *X* in the sample used in the analysis
  - No way to check linearity outside the range of values sampled, not a good idea to predict outside this range

54

Correlation and Regression

- Correlation only indicates the strength of the relationship between two variables, it does not give a description of the relationship or allow for prediction
- The t-test of the null hypothesis that the correlation is zero is exactly equivalent to that for the hypothesis of zero slope in the regression analysis

55

Correlation and Regression

- For correlation both variables must be random, for regression *X* does not have to be random
- Correlation is often over used
- One role for correlation is in generating hypotheses, remember correlation is based on one number, limit to what can be inferred with one number

56

Summary I

- Simple linear regression- describe and predict linear relationship
- Least squares regression
- Assumptions:
  - Dependent variable *Y* Normally distributed for each value of the independent variable
  - Variability of *Y* same for each value of *X*
  - Linear relationship
  - Independent observations
  - *X* doesn't have to be random or Normally distributed

57

Summary II

- Need to be familiar with:
  - Regression coefficient (slope)
  - Intercept
  - Residuals – Normal(0, σ^2)
  - Fitted Value
  - R^2 (coefficient of determination)
  - Residual standard deviation
- Confidence and Prediction Intervals
- Interpolation and Extrapolation

58