

Statistics:

Concepts of statistics for researchers

How to Use This Course Book

This course book accompanies the face-to-face session taught at IT Services. It contains a copy of the slideshow and the worksheets.

Software Used

We might use Excel to capture your data, but no other software is required. Since this is a Concepts course, we will concentrate on exploring ideas and underlying concepts that researchers will find helpful in undertaking data collection and interpretation.

Revision Information

Version	Date	Author	Changes made
1.0	January 2014	John Fresen	Course book version 1
2.0	October 2014	John Fresen	Updates to slides
3.0	January 2015	John Fresen	Updates to slides
4.0	February 2015	John Fresen	Updates to slides and worksheets
5.0	May 2015	John Fresen	Updates to slides and worksheets
6.0	June 2015	John Fresen	Updates to slides and worksheets
7.0	October 2015	John Fresen	Updates to slides and worksheets
8.0	December 2015	John Fresen	Updates to slides and worksheets
9.0	February 2016	John Fresen	Updates to slides and worksheets
10	March 2016	John Fresen	Updates to slides and worksheets

March 2016

Copyright

The copyright of this document lies with Oxford University IT Services.

Contents

1 Introduction	1
1.1. What You Should Already Know	1
1.2. What You Will Learn	1
2 Your Resources for These Exercises	2
2.1. Help and Support Resources	2
3 What Next?	3
3.1. Statistics Courses	3
3.2. IT Services Help Centre	3

1 Introduction

Welcome to the course **Statistics: Concepts**.

This is a statistical concepts course, an ideas course, a think-in-pictures course. What are the basic notions and constructs of statistics? Why do we differentiate between a population and a sample? How do we summarize and describe sample information? Why, and how, do we compare data with expectations? How do hypotheses arise and how do we set about testing them? With inherent uncertainty in any sample, how can one extrapolate from a sample to the population? And then, how strong are our conclusions?

This course is designed to prepare you to get the most from the statistical applications that we teach. It involves discussion of real-life examples and interpretation of data. We strive to avoid mathematical symbols, notation and formulae.

1.1. What You Should Already Know

We assume that you are familiar with entering and editing text, rearranging and formatting text - drag and drop, copy and paste, printing and previewing, and managing files and folders.

The computer network in IT Services may differ slightly from that which you are used to in your College or Department; if you are confused by the differences, ask for help from the teacher.

1.2. What You Will Learn

In this course we will cover the following topics:

- Descriptive statistics and graphics
- Population and sample
- Probability and probability distributions
- Comparing conditional distributions
- Confidence intervals
- Linear regressions
- Hypothesis testing
- From problem – to data – to conclusions

Where to get help....

Topics covered in related *Statistics* courses, should you be interested, are given in Section 3.1.

2 Your Resources for These Exercises

The exercises in this handbook will introduce you to some of the tasks you will need to carry out when working with *WebLearn*. Some sample files and documents are provided for you; if you are on a course held at IT Services, they will be on your network drive **H:** (Find it under **My Computer**).

During a taught course at IT Services, there may not be time to complete all the exercises. You will need to be selective, and choose your own priorities among the variety of activities offered here. However, those exercises marked with a star * should not be skipped.

Please complete the remaining exercises later in your own time, or book for a Computer8 session at IT Services for classroom assistance (See section 8.2).

2.1. Help and Support Resources

You can find support information for the exercises on this course and your future use of *WebLearn*, as follows:

- *WebLearn* Guidance <https://weblearn.ox.ac.uk/info> (This should be your first port of call)

If at any time you are not clear about any aspect of this course, please make sure you ask John for help. If you are away from the class, you can get help and advice by emailing the central address weblearn@it.ox.ac.uk.

The website for this course including reading material and other material can be found at <https://weblearn.ox.ac.uk/x/Mvkigl>

You are welcome to contact John about statistical issues and questions at john.fresen@gmail.com

3 What Next?

3.1. Statistics Courses

Now that you have a grasp of some basic concepts in Statistics, you may want to develop your skills further. IT Services offers further Statistics courses and details are available at <http://courses.it.ox.ac.uk>.

In particular, you might like to attend the course

Statistics: Introduction: this is a four-session module which covers the basics of statistics and aims to provide a platform for learning more advanced tools and techniques.

Courses on particular discipline areas or data analysis packages include:

R: An introduction

R: Multiple Regression using R

Statistics: Designing clinical research and biostatistics

SPSS: An introduction

SPSS: An introduction to using syntax

STATA: An introduction to data access and management

STATA: Data manipulation and analysis

STATA: Statistical, survey and graphical analyses

3.2. IT Services Help Centre

The IT Services Help Centre at 13 Banbury Road is open by appointment during working hours, and on a drop-in basis from 6:00 pm to 8:30 pm, Monday to Friday.

The Help Centre is also a good place to get advice about any aspect of using computer software or hardware. You can contact the Help Centre on (2)73200 or by email on help@it.ox.ac.uk

Statistical Concepts for Researchers

John Fresen
March 2016



IT Learning Programme

Your safety is important

Where is the fire exit?

Beware of hazards:

Tripping over bags and coats

Please report any equipment faults to us

Let us know if you have any other concerns

Your comfort is important

The toilets are along the corridor outside the lecture rooms

The rest area is where you registered;
it has vending machines and a water cooler

The seats at the computers are adjustable

You can adjust the monitors for height, tilt and brightness



3

Session 1: Setting the scene

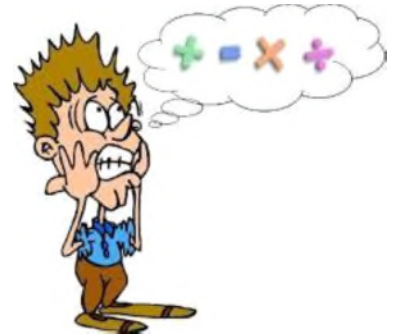
*We are drowning in information but starving for knowledge
– Rutherford D. Roger*

Thanks to:
Dave Baker, IT Services
Steven Albury, IT Services
Jill Fresen, IT Services
Jim Hanley, McGill University
Margaret Glendining, Rothamsted Experimental Station
Ian Sinclair, REES Group Oxford

4

- This course is designed for people with little no or previous exposure to statistics

-- even those suffering some 'statistical anxiety'
 -- but who need to use statistics in their research



- This a pre-computing course
 discussing real-life examples
 and interpretation of data

- We strive to avoid mathematical symbols, notation and formulae
 but have put some formulae into the text
 because of requests for them
 but you can ignore them if you wish



Research question

– particular problem

- collect data

- draw conclusions



steps in research

We don't do statistics for statistics sake

- but to answer questions

Fundamental assumptions of statistics

variation/diversity/noise/error is everywhere/ubiquitous

Statistical models: observe = truth + error

observe = model + error

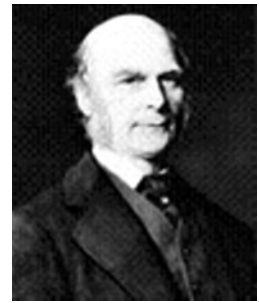
observe = signal + noise

Sir Francis Galton

(16 February 1822 – 17 January 1911)

http://en.wikipedia.org/wiki/Francis_Galton

Sir Francis Galton was an incredible polymath
Cousin of Charles Darwin.



General: Genetics – What do we inherit from our ancestors?

Particular: Do tall parents have tall children and short parents, short children?
i.e. Does the height of children depend on the height of parents?

Data: Famous 1885 study: 205 sets of parents 928 offspring
mph = average height of parents; ch = child height

Galton Peas Experiment: Selected 700 pea pods of selected sizes
average diam of parent peas ; average diam of child peas



7

Figure 1. Photograph of the entries for the first 12 families listed in Galton's notebook.
Published with the permission of the Director of Library Services of University College London.

FAMILY HEIGHTS. from R.F.F.				
(add 60 inches to every entry in the Table)				
	Father	Mother	Sons in order of height	Daughters in order of height.
1	18.5	7.0	13.2	9.2, 9.0, 9.0
2	15.5	6.5	13.5, 12.5	5.5, 5.5
3	15.0	about 4.0	11.0	8.0
4	15.0	4.0	10.5, 8.5	7.0, 4.5, 3.0
5	15.0	1.5	12.0, 9.0, 8.0	6.5, 2.5, 2.5
6	14.0	8.0		9.5
7	14.0	8.0	16.5, 14.0, 13.0, 13.0	10.5, 4.0
8	14.0	6.5		10.5, 8.0, 6.0
9	14.5	6.0		6.0
10	14.0	5.5		5.5
11	14.0	2.0	14.0, 10.0	8.0, 7.0, 7.0, 6.0, 3.5, 3.0
12	14.0	1.0		5.0

Sir Ronald Fisher - The grandfather of statistics (17 February 1890 – 29 July 1962)

http://en.wikipedia.org/wiki/Ronald_Fisher



Stained glass window in the dining hall of Gonville and Caius College, Cambridge, commemorating Ronald A. Fisher, geneticist and statistician, who was a fellow and president of the college. The window represents a 7 by 7 Latin square, an experimental design used in statistics; the text on the windows reads:
R.A. FISHER; FELLOW 1920-1926 1943-1962;
PRESIDENT 1956-1959

We'll use his potato data from Rothamsted, that I've taken from his book *Statistical Methods for Research Workers*.

9

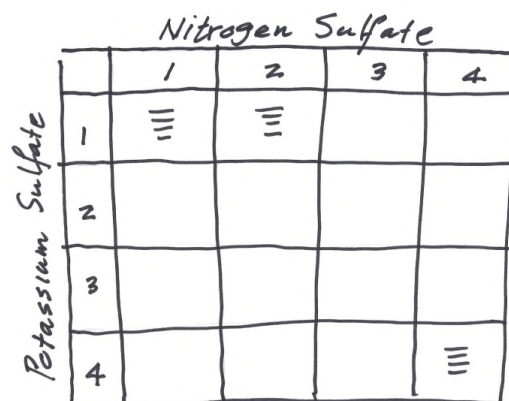
Fisher's Potato Data

T. Eden and R. A. Fisher (1929) *Studies in Crop Variation. VI. Experiments on the Response of the Potato to Potash and Nitrogen. J. Agricultural Science* 19, 201–213.

nitrogen	potash	yield
0	0	317.5
0	1	363
0	2	368
0	4	381.5
1	0	314
1	1	383
1	2	434.5
1	4	447.5
2	0	302.5
2	1	444.5
2	2	471.5
2	4	449

General question: Can we improve crop production?

Particular question: This experiment



In every experiment there is always the assumption that the variables cover a big enough range to be effective

H. V. Roberts *Harris Trust and Savings Bank:
An analysis of employee compensation.* (1979)
Report 7946, Center for Mathematical Studies in Business and Economics,
University of Chicago Graduate School of Business

Starting salaries (\$U.S.) for 32 male and 61 female clerical hires at a bank

Males			Females					
4,620	5,700	6,000	3,900	4,500	4,800	5,220	5,400	5,640
5,040	6,000	6,000	4,020	4,620	4,800	5,220	5,400	5,700
5,100	6,000	6,000	4,290	4,800	4,980	5,280	5,400	5,700
5,100	6,000	6,300	4,380	4,800	5,100	5,280	5,400	5,700
5,220	6,000	6,600	4,380	4,800	5,100	5,280	5,400	5,700
5,400	6,000	6,600	4,380	4,800	5,100	5,400	5,400	5,700
5,400	6,000	6,600	4,380	4,800	5,100	5,400	5,400	6,000
5,400	6,000	6,840	4,380	4,800	5,100	5,400	5,520	6,000
5,400	6,000	6,900	4,440	4,800	5,100	5,400	5,520	6,120
5,400	6,000	6,900	4,500	4,800	5,160	5,400	5,580	6,300
	6,000	8,100						6,300

General question: Do women earn less than men?

Particular question: These data from an American bank

11

Data sets summary:

1. Galton parent-child height data: *Do tall parents have tall children?*
2. Galton Peas data: *Do big peas produce big peas?*
3. Fisher potato data: *Response of the Potatoes to Potash and Nitrogen*
4. Roberts salary data: *Do woman earn less than men?*

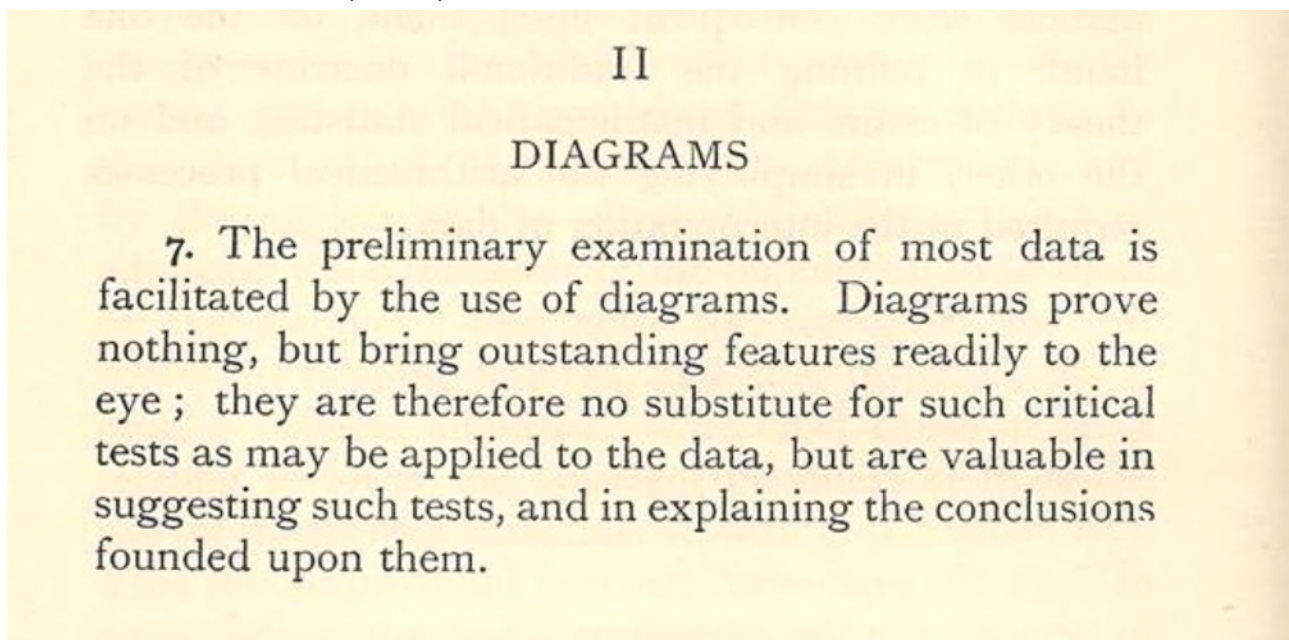
12

Session 2: Descriptive statistics

*The quiet statisticians have changed our world;
not by discovering new facts or technical developments,
but by changing the ways that we reason, experiment and form our opinions*
–Ian Hacking

13

From Ch 2 of Sir Ronald Fisher's book *Statistical Methods for Research Workers* (1925):



I wonder what he would have produced and said if he had the graphical power we have nowadays?

I strongly recommend the work by Paul J Lewi *Speaking of Graphics* that can be found at <http://www.datascope.be/sog.htm>

14

observation/perception is interpretive
... describe your data
... tell the story of your data
... what is your data saying?

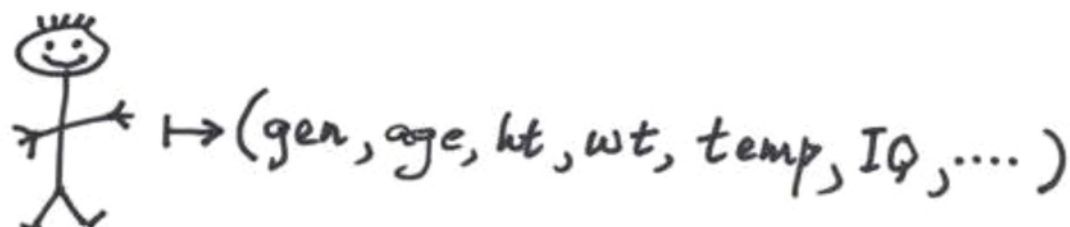
narration depends on many things
... amount of information/extent of knowledge
... purpose of description

e.g. describe your research



*There is no true interpretation of anything;
interpretation is a vehicle in the service of human comprehension.
The value of interpretation is in enabling others to fruitfully think about an idea*
—Andreas Buja

15



We consider one (univariate) eg ht
or two (bi-variate) eg ht, wt

Any real analysis is usually
multivariate

One consequence is that we can never collect all the information
associated with an experiment or observational study

16

Source and location of data

- How was data obtained?
- Where is it stored?
- What processing has been done on the data
- Who has access to data?

Numerical descriptors of a data set

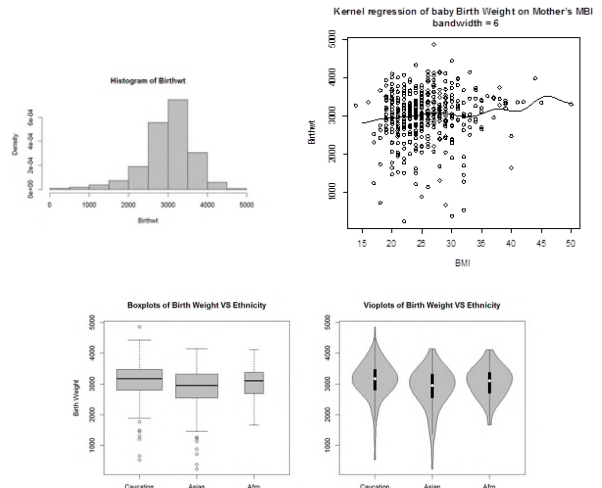
(Usually most uninformative)

- Order statistics – smallest to biggest
- Mean/average
- Variance and standard deviation
- Quartiles, percentiles
- Correlation coefficient
- Prevalence of HIV
- ... Many more

Graphical descriptors of a data set:

(A picture says a thousand words)

- Dot plot
- Box and whisker plot
- Histogram
- Pie chart
- Scatterplot
- ... Many more



17

Average, Mean, Centre of Gravity

Consider data: 3, 2, 5, 10, 5, 2, 6, 7, 2, 5

Sort (order stats): 2, 2, 2, 3, 5, 5, 5, 6, 7, 10
most basic statistics.

dot plot / stripchart

Sum = 47

average = 4.7

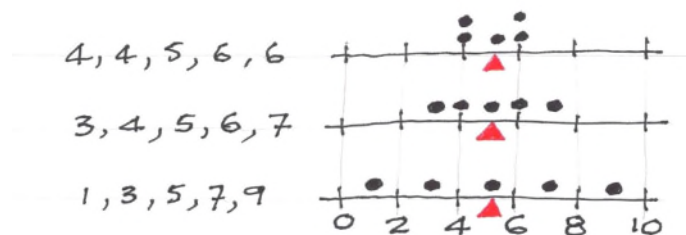
Average is balance point or C.O.G.

So we can get a pretty good idea of average by guessing C.O.G on a dot plot.

The average/mean/COG is the **balance point** of a **distribution**

18

degrees of freedom,
total variation, variance, standard deviation



degrees of freedom (df) :

starts off as the number of observations, (n).

each time we do a computation we use up a df.

(total) variation = sum of squared deviations about the mean (df = n-1)

variance = average squared deviation about the mean = variation/df

standard deviation = square root of variance

Task: Compute the variation, variance and sd for the three data sets

As a rough approximation the sd is about range/6

It conveys the “average” or typical deviation of the data about the mean

Median, Quartiles, Percentiles, Box and Whisker plot

Median: 50% of data are less or equal to the median.

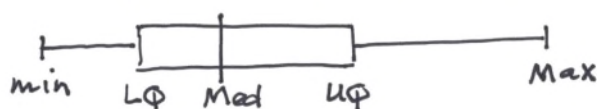
Sort data smallest → largest

Quartiles: 25% data ≤ Lower quartile
50% " " Middle "
75% " " Upper "

Percentiles: 90% data ≤ 90th percentile
50% " ≤ 50th "
etc.

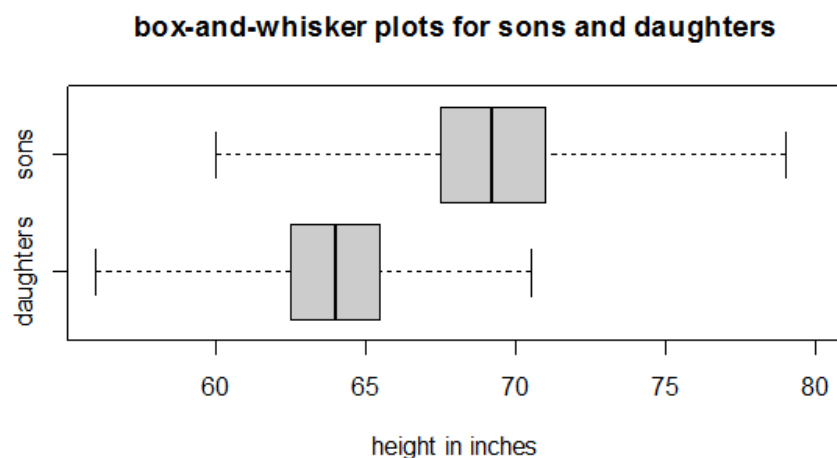
Box & Whisker plot:

Graphical display of (min, LQ, Med, UQ, max)



The dot diagram is only useful for a small grainy data set.
For a large data set we may use a box-and-whisker plot.

Consider the Galton family data.

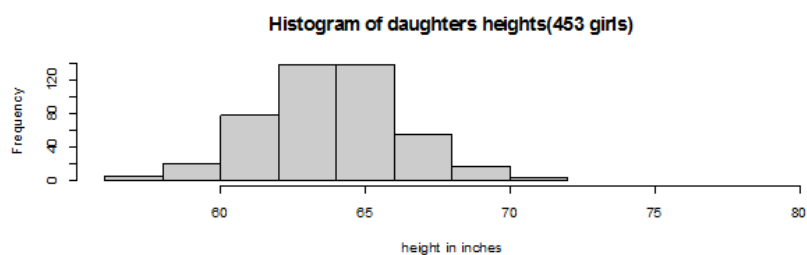
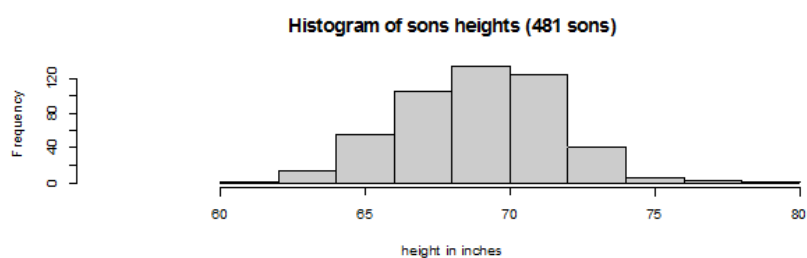


The box-and-whisker plot (due to John Tukey) is a graphical representation of a five number summary of the data:

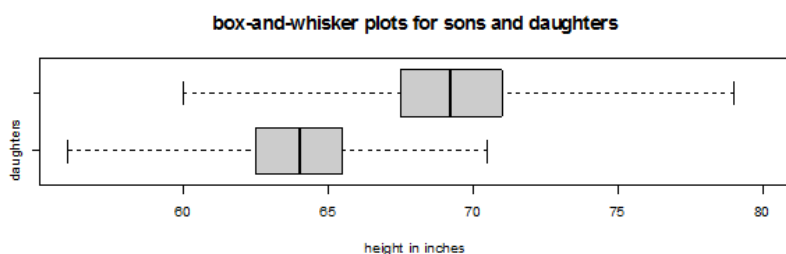
minimum, lower quartile, median, upper quartile, maximum

21

Frequency histograms and Box-and-whisker plots for the Galton data



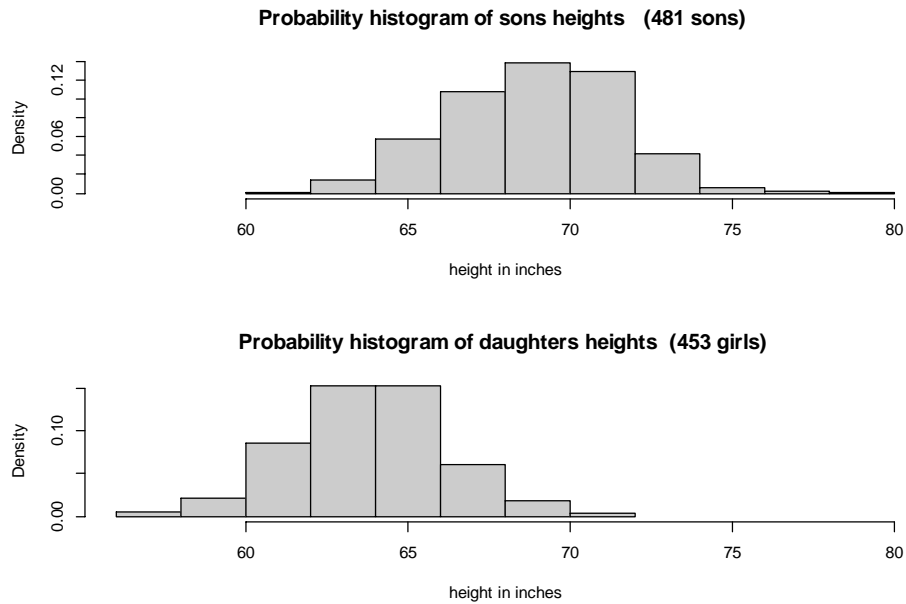
*Guess means and sd's
for these two distributions*



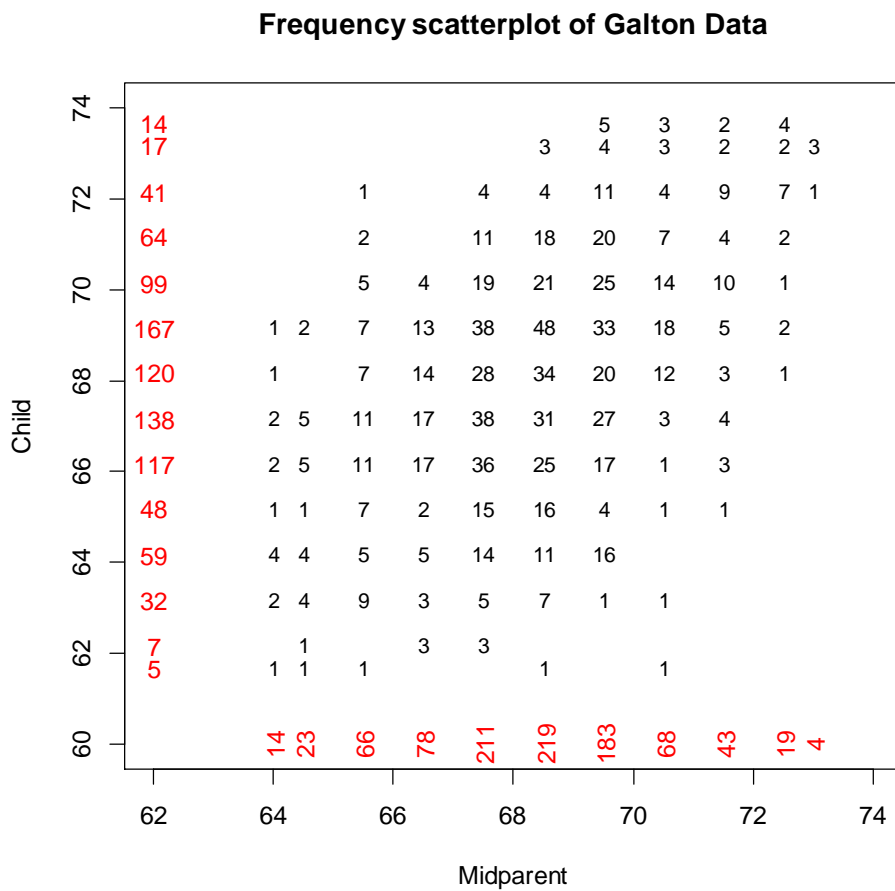
22

Probability density histogram

mass density = mass per unit volume
probability density = probability per unit length



23

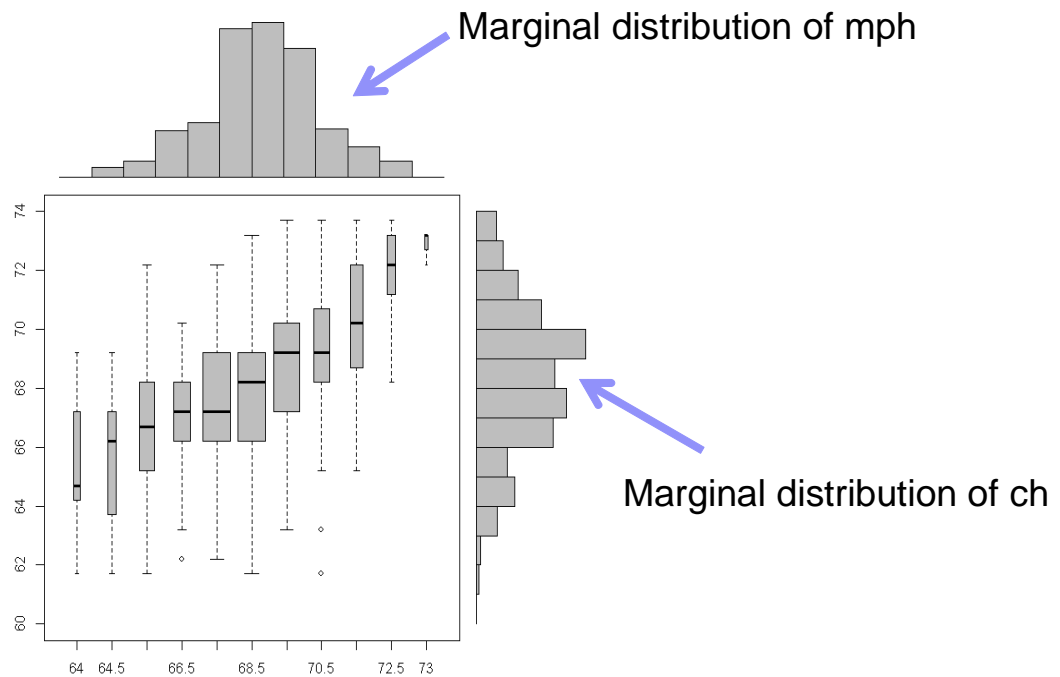


24

Galton data:

Boxplots of conditional distributions

histograms of the marginal distributions



Width of boxplots proportional to sqrt of sample size

25

Do worksheet 1 at the back of the book

Work in pairs or groups if you like

26

Session 3: Probability and probability distributions

27

What is an experiment? Give an example.

Essential feature is that the outcome is unknown . . .

. . . leads to a probability distribution of the possible outcomes

At least three notions of probability

Classical: *Assumes equally likely outcomes (games of chance)*

toss a coin

roll a die

spin a roulette wheel

winning a cake on a church raffle

Empirical: *empirical probability is a percentage*

probability of smokers developing lung cancer (Richard Doll)

probability of an motor insurance claim

Subjective: probability of a business venture being successful

probability of a heart replacement surgery being successful

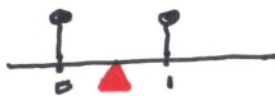
probability of Oxford winning the boat race

28

Discrete probability distributions (can only take on certain values)

Toss Coin

outcome	H=1	T=0
prob	$\frac{1}{2}$	$\frac{1}{2}$



Roll die

outcome	1	2	3	4	5	6
prob	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$



Roulette wheel



outcome	1	2	3	4	5
prob	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$



mean of prob dist =
sum of
outcome \times prob

outcome	1	2	3	4
prob	0.1	0.2	0.3	0.4



variance of prob dist =
sum of
sqrd deviation from mean \times prob

Three things that make a probability distribution

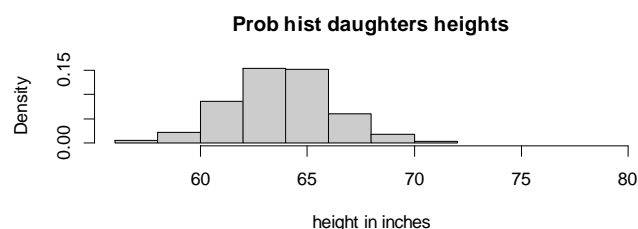
1. List all outcomes
2. List all the associated probabilities
3. Probabilities add to unity

Task: compute means and
variances of the distributions of the
last two distributions **29**

Continuous probability distributions:

can take on any value in a certain range, e.g. height, weight, . .
can't list the possible values or their probabilities
use a probability density function

A probability histogram is a crude example of a probability density function

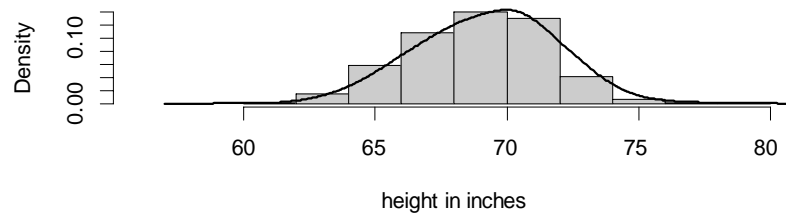


Properties

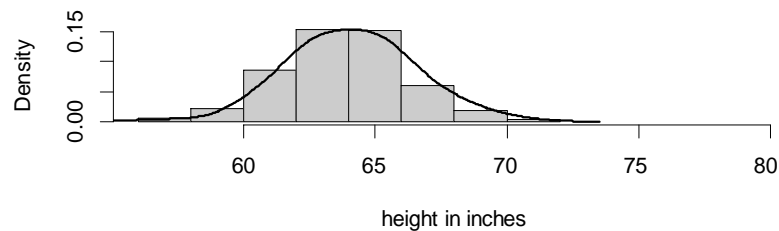
1. Possible values are displayed on horizontal axis
2. Total area under curve is unity
3. Probabilities are represented by areas under the curve

Probability density estimate of heights using a moving bin

Prob hist sons heights with density estimate



Prob hist daughters heights with density estimate



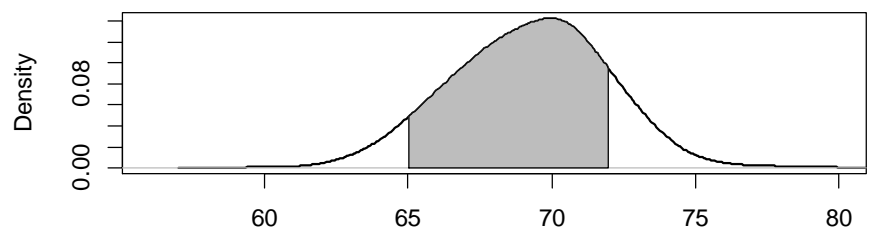
Properties of a probability density function

1. Possible values are displayed on horizontal axis
2. Total area under density curve is unity
3. Probabilities are represented by areas under the curve

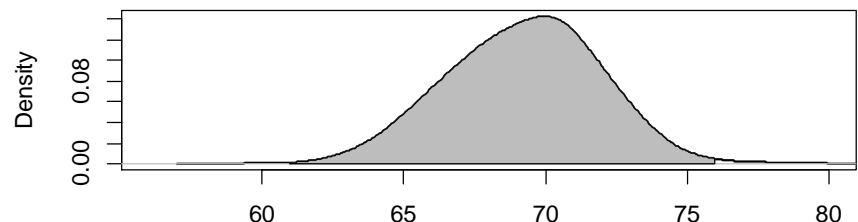
31

Examples:

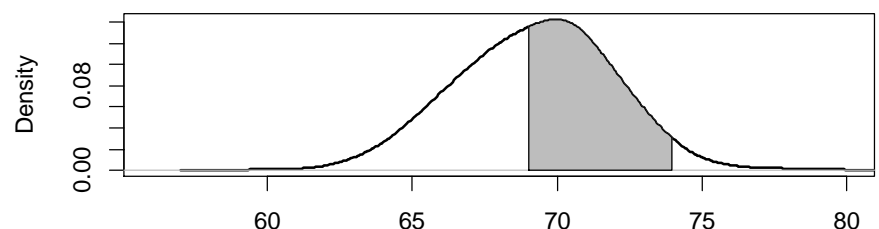
Probability of selecting a son
between 65 and 72 inches
area = 0.78



Probability of selecting a son
between 61 and 76 inches
area = 0.99



Probability of selecting a son
between 69 and 74 inches
area = 0.51



Properties of a probability density function

1. Possible values are displayed on horizontal axis
2. Total area under density curve is unity
3. Probabilities are represented by areas under the curve

32

The mean and variance of a continuous distribution generalise the ideas from a discrete distribution

Draw probability density function

Partition the x-axis into cuts of width “dx”

Compute area of slices under the curve above each cut – to give probabilities

mean = sum of cut centre points X area of slice above cut (prob)

variance = sum of sqrd distance of cut centre points from mean X area of slice above (prob)

$$\text{mean} = \mu = \int xf(x)dx$$

$$\text{variance} = \sigma^2 = \int (x - \mu)^2 f(x)dx$$

33

Normal or Gaussian distribution – *Affectionately called the bell-curve*

developed mathematically by DeMoivre (1733) as an approximation to the binomial distribution. His paper was not discovered until 1924 by Karl Pearson.

Laplace used the normal curve in 1783 to describe the distribution of errors.

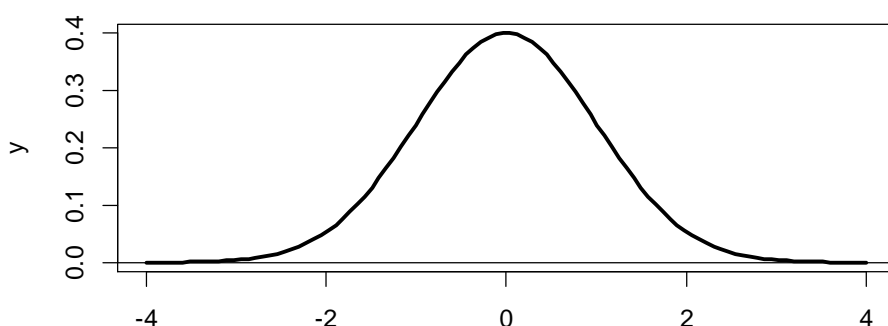
Gauss (1809) used the normal curve to analyse astronomical data

Assumptions about the noise/error:

- small errors are more likely than large errors

- negative errors just as likely as positive errors

This led him to the symmetrical bell-shaped curve known as the normal distribution



34



Do worksheet 2

35



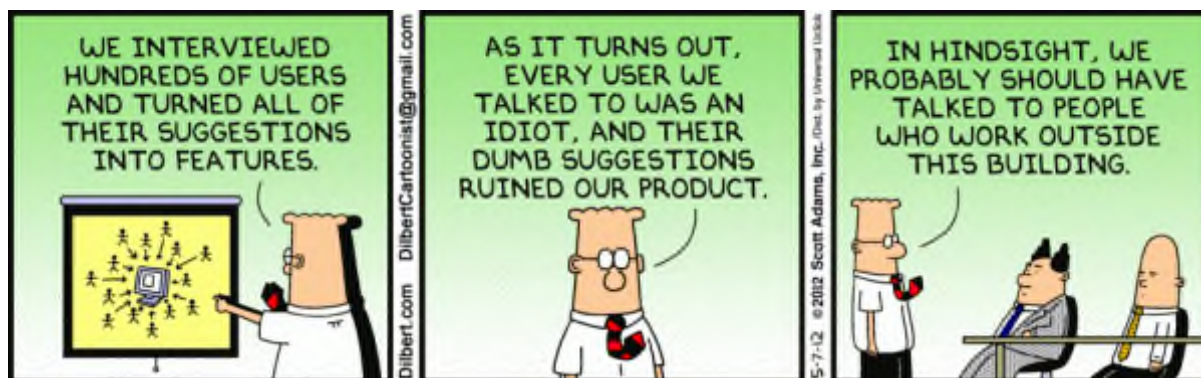
Session 4: Population and Sample

36

The fundamental **notions** of statistics are:

1. Population
2. Variation between the members of a population
3. Sample
4. Description of variation by a probability distribution

(There is hardly any point to our research if we can't generalize from our sample to some broader population)



37

37

The fundamental **strategy** of statistics:
compare observations with expectations

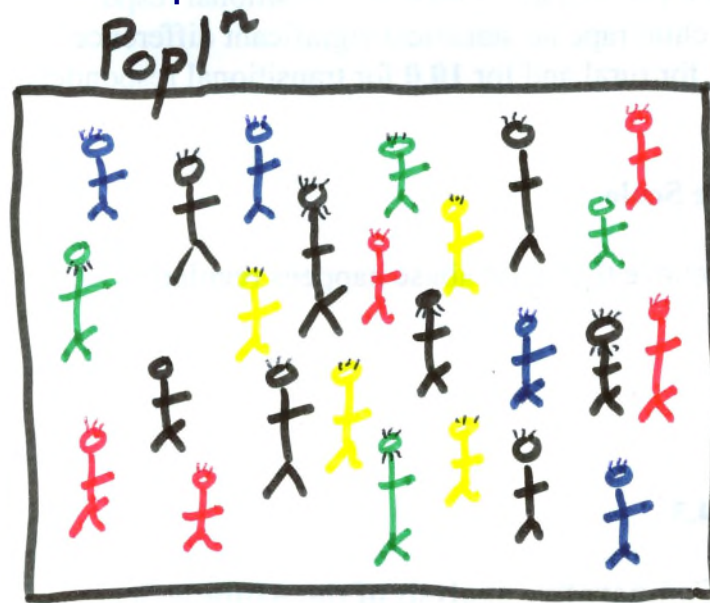
Eg: do men and women earn similar salaries?
is yield under fertilizer A same as yield under fertilizer B?
is the generic alternative as good as the brand name drug?
how do ART children compare with normal children educationally?

The fundamental **method** of statistics:
compare conditional distributions

38

38

Population and sample



What are the pro's and con's of these constructs?

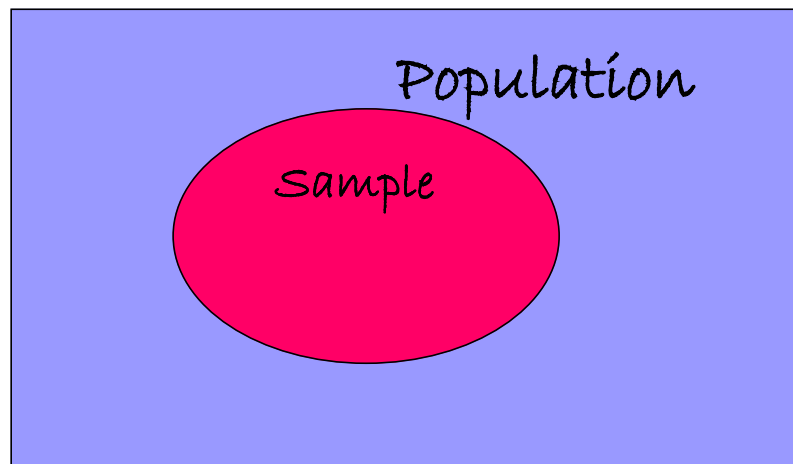
- Population difficult to define

- Changing over time

- Many potential populations from which our sample may have come

We have a definite sample but our population is more an intellectual construct

39

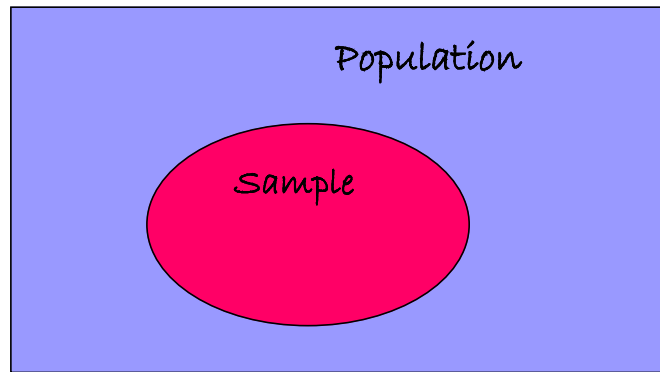


Problem: We wish to know something about the population

Solution: Take a sample to get an idea/estimate

Question: How reliable is our estimate?

Partial answer: The larger the sample the better



Anything **calculated** from a **sample** is called a **statistic**

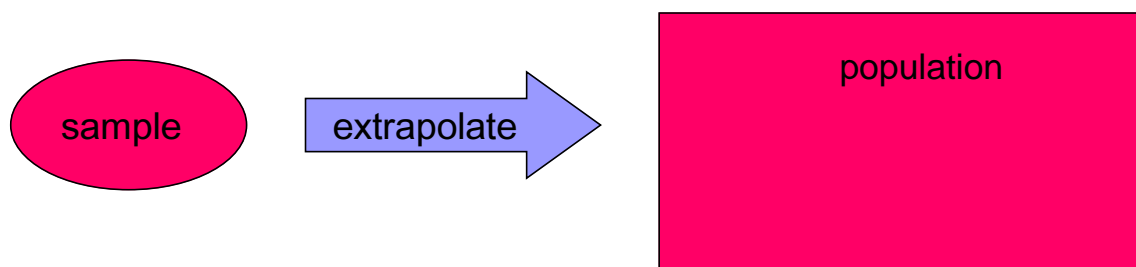
i.e. statistic = something calculated from a sample

e.g. average, maximum, range, proportion having HIV/Aids
or a combination of these

41

41

Statistical Inference: *extrapolating from sample to population*



One can describe the statistical aspects of any sample –

but can only reliably extrapolate from a **random sample**

Why a random sample?

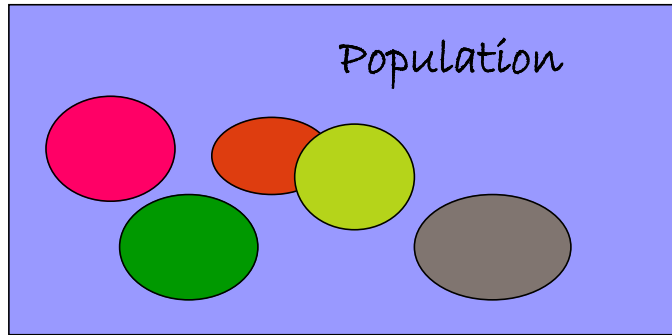
How would we achieve a random sample?

What's bad about a random sample?

Going from particular to general – inductive inference.
Controversial issue. Hume 1777.

42

42



Problem with taking a random sample: *We get a random answer!*

Each time we take another random sample, we get a different answer

the resulting distribution is called the sampling distribution

standard deviation of the sampling distribution is called the *standard error*

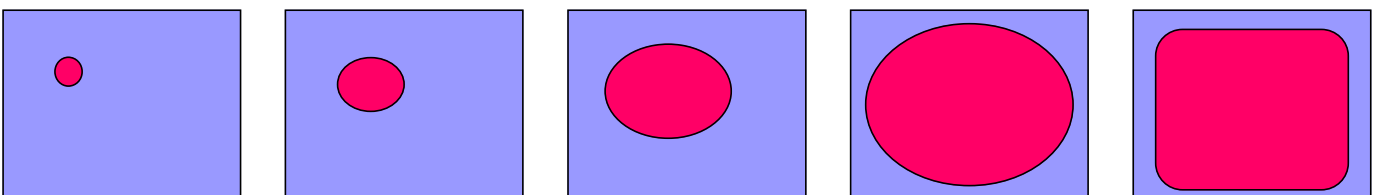
Nearly all statistical theory assumes a random sample

If the sample is *not random* – we can't rely on statistical theory

43

What happens as sample gets large?

Two central pillars of statistics: LLN and CLT



sample average **converges** to population average
(*Law of Large Numbers LLN*)

distribution of sample average **converges** to the normal distribution
(*Central Limit Theorem CLT*)

any random quantity that arises as the sum of many independent contributions is distributed very much like a Gaussian random variable.

Same applies to most other statistics:

sample quantity **converges** to population quantity *LLN*

sampling distribution **converges** to the normal distribution *CLT*

Example of a sampling distribution

At lunch time we'll work in pairs.

Each pair take a random sample of size 10 from the population of 600 buttons.

Compute five statistics:

- average ht, wt, BMI

- proportion having hypertension (bp) and diabetes (db)

 - (proportion = average of the indicator data)

approximately 1.54×10^{21} possible random samples of size 10 from the 600 buttons. If we were able to compute the five statistics for each of these random samples we would have their sampling distributions – we could draw histograms and compute their means and standard deviations.

The standard error (SE) of a statistics is the standard deviation of the sampling distribution

We can't do this. But we can get an approximations

using our few random samples

using the central limit theorem: $SE(\text{approx.}) = \text{sample sd} / \sqrt{\text{sample size}}$

Sometimes it's difficult to decide what is a population and what is a sample
e.g. RAU – MEDUNSA M-STUDENTS - Sept 2003

What question would make this a population?

What would make it a sample?





Moure open cast coal mine in Queensland Australia:

How would we obtain a random sample to assess coal quality???

What is the population to which we wish to extrapolate results???

Are units distinct?

What functions are we interested in to measure the quality of the coal?

47

To estimate the density of stomata on undersides of loblolly pine needles it is hoped to obtain a random sample of needles from loblolly pine trees. Describe how one would attempt to define the population and then how to obtain a random sample from a forest of loblolly pine trees.

TABLE I: Number of stomata per centimetre on each of ten loblolly pine needles.

needle	1	2	3	4	5	6	7	8	9	10
	149	136	143	121	148	129	127	134	117	129
	143	139	142	133	121	134	130	137	128	132
	138	129	124	126	124	127	123	119	119	131
	131	143	134	130	128	113	125	130	118	137



48

Closing Remarks:

Our conclusions will be no stronger than the degree to which the assumptions and mathematical models correlate with the real world:

Population (Can we define this?)

Random sample (How do we achieve this?)

Limited information (What variables or factors are important?)

Simplified models (Linear regression)

Probability of an error

Quasi-Modus Tollens Argument (to come later)

These might seem like great limitations –
but statistical arguments are the best we have
for assessing empirical evidence

49

Session 5: Confidence intervals

50

When *random sampling* from a population to obtain an estimate of a population parameter (mean, prevalence of HIV) the sample estimate is a *random quantity*.

A good question is: *What is the likely error in our estimate?*

Answer: *We go back to the sampling distribution*

*we could use the lower and upper quartile of the sampling distribution
(about a 50% Confidence Interval)*

or the 5th and 95th percentiles (about a 90% Confidence Interval)

use the CLT the sampling distribution will be approx. normal

90% CI: *sample estimate $\pm 1.64 \times$ standard error of estimate*

95% CI: *sample estimate $\pm 1.96 \times$ standard error of estimate*

99% CI: *sample estimate $\pm 2.58 \times$ standard error of estimate*

(The more confidence you want, the wider is the CI) **51**

For our sampling exercise of the buttons data, a 95% CI

for the mean (ht or wt or BMI) of the population is

95% CI: *sample average $\pm 1.96 \times$ standard deviation / $\sqrt{\text{sample size}}$*

for the prevalence (%diabetic or % hypertensive)

95% CI: *sample prevalence $\pm 1.96 \times \sqrt{\text{prevalence} \times (1 - \text{prevalence}) / \text{sample size}}$*

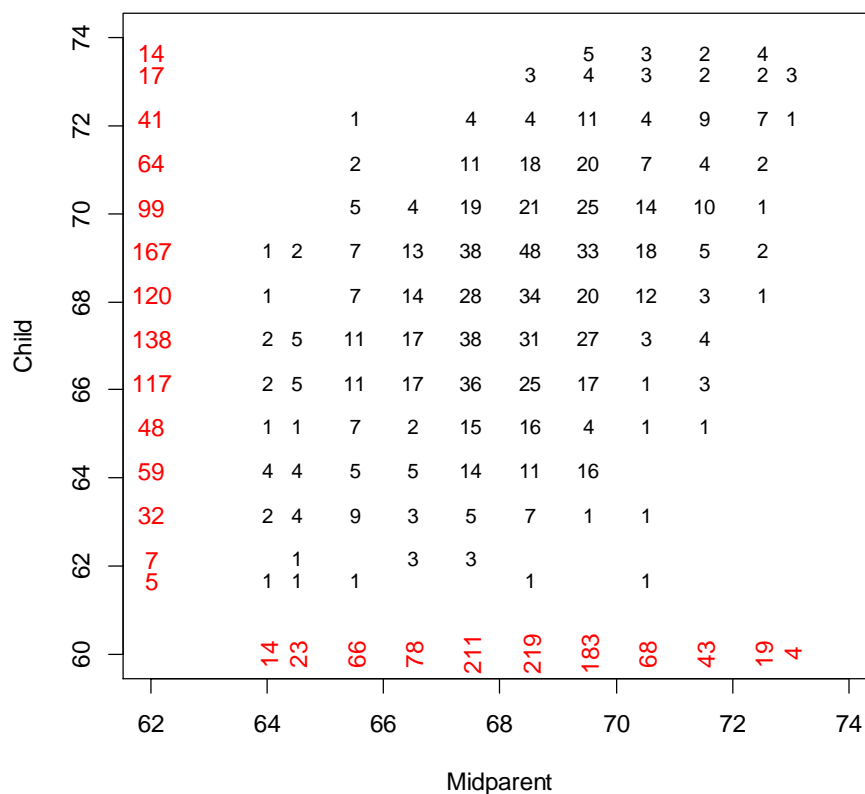
Do worksheet 3

Session 6: Regression

53

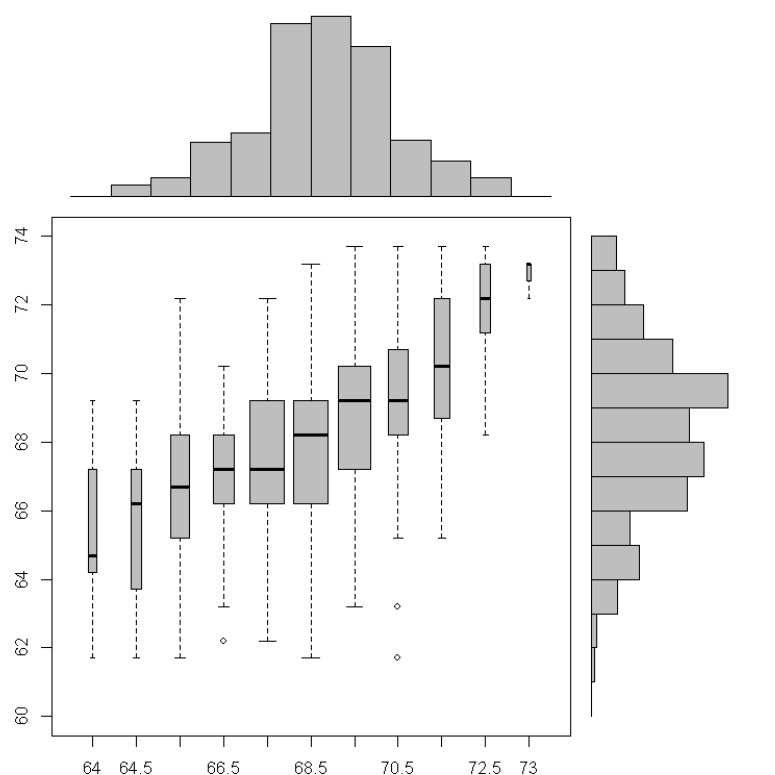
Francis Galton: Do tall parents have tall children, short parents short children?
Does height of child depend on height of parents?

Frequency scatterplot of Galton Data



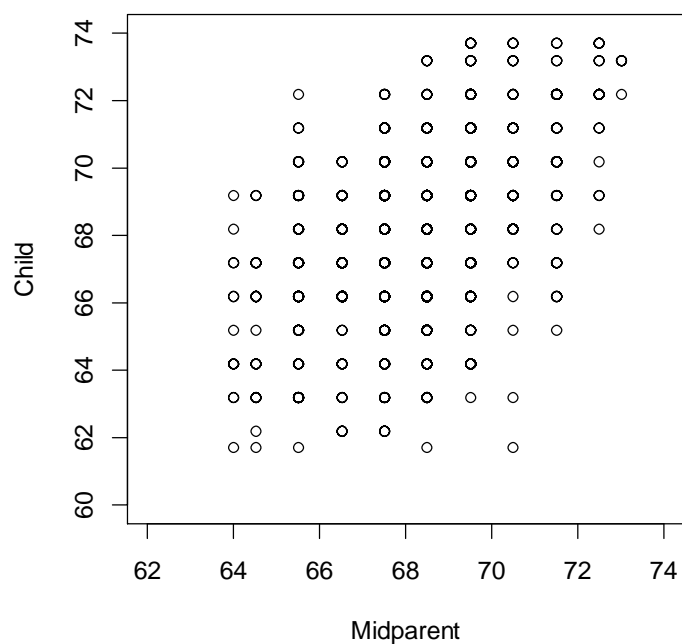
54

Galton data:
Boxplots of conditional distributions
histograms of the marginal distributions

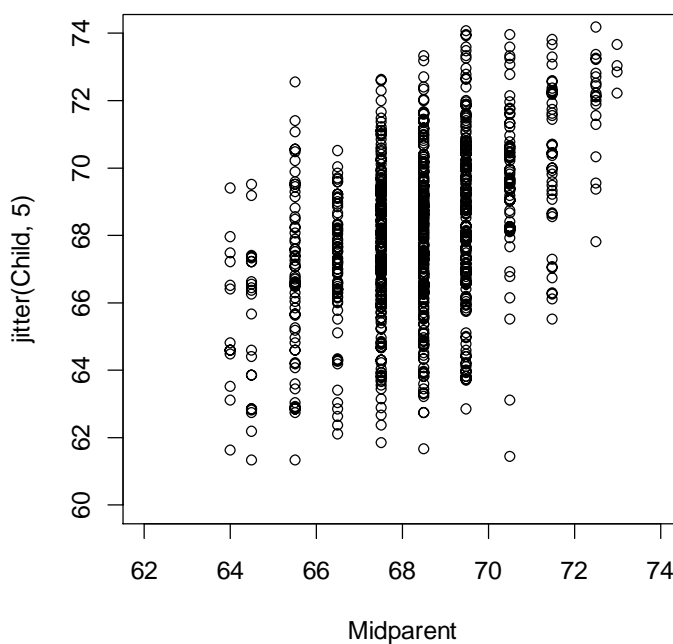


55

Scatterplot of data



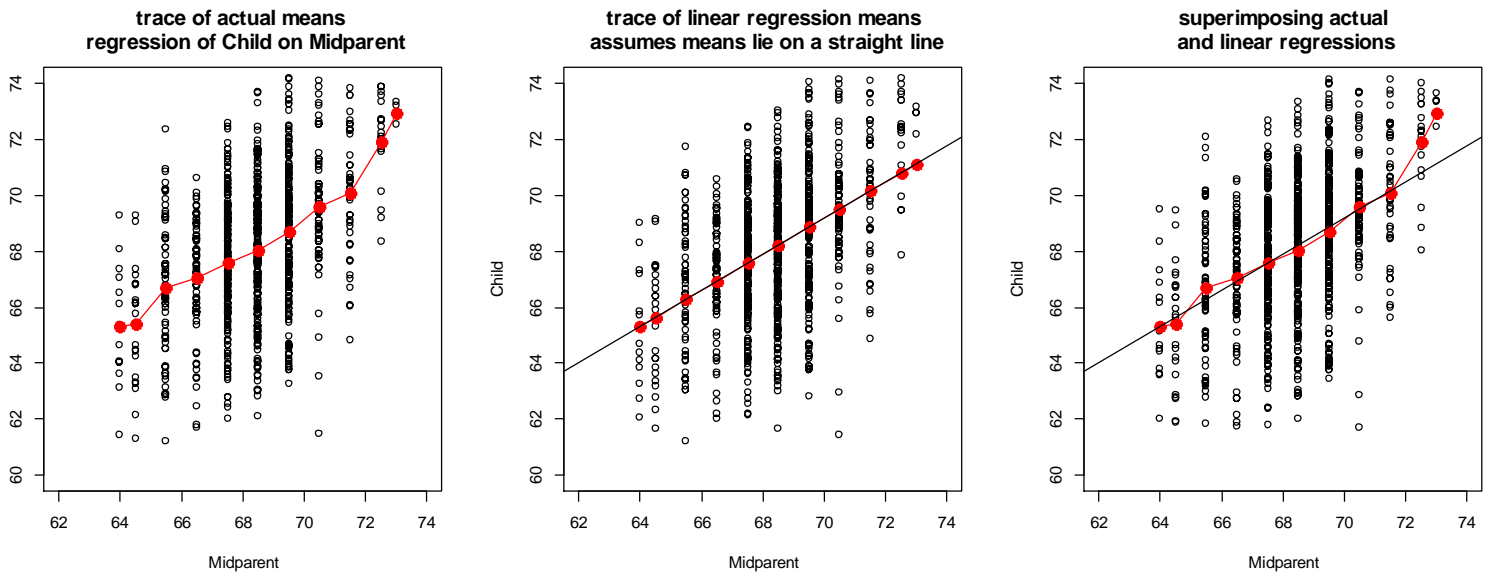
Scatterplot with Child jittered



The distribution of points is not evident in this plot
because many points land on one spot

56

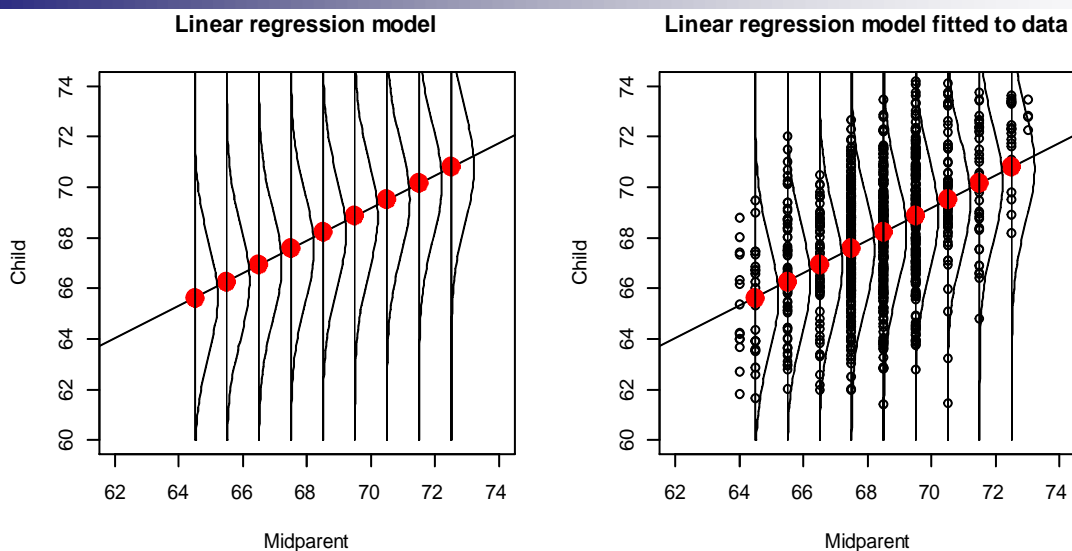
Regression is a plot/trace of the means of the conditional distributions



The trace of actual means has no assumptions in it – but end distributions have a lot of sampling variation because of the small number of observations in those distributions

Linear regression stabilises that – but let's take a deeper look at the linear regression model

57



Linear regression model assumes:

1. Conditional distributions are normal
2. Conditional means lie on a straight line
3. Conditional distributions all have same spread

In words: the distribution of child height, conditional on a given midparent height, is normal, with means lying on the straight line, and constant spread

In mathematics: $Y|x \sim \text{Normal}(\mu = a + bx, \sigma^2)$

58

What did Galton get from his linear regression?

mid(ph)	64.0	64.5	65.5	66.5	67.5	68.5	69.5	70.5	71.5	72.5
ave(ch)	65.3	65.6	66.3	66.9	67.6	68.2	68.9	69.5	70.2	70.8

He concluded that:

YES, tall parents do tend to have tall children

but their children regress down to the population average

YES, short parents tend to have short children

but their children tend to regress up to the population average

How fortunate. Imagine if this were not so.

He drew similar conclusions about other hereditary factors, such as intelligence for example. Intelligence testing began to take a concrete form with Sir Francis Galton, considered to be the father of mental tests.

What is the purpose of doing regression?

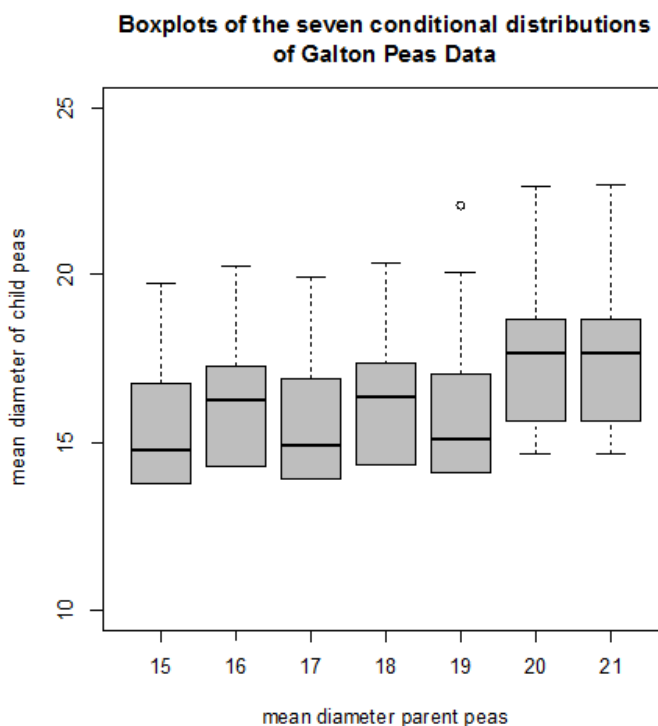
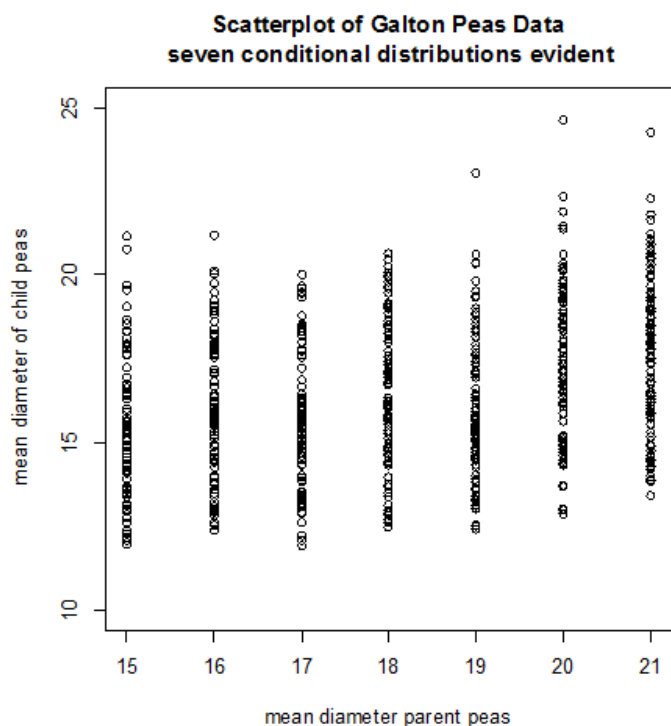
Prediction? e.g. Credit scoring - predict the probability of bad debt from various characteristics of a client

Explanation? e.g. which type of advertising, radio, TV, bill boards, is most effective for improving TESCO sales


Other?

59

Does the average diam of child peas depend on average diam of parent peas?



60



Do worksheet 4 on regression

61



Session 7: Hypothesis formulation and testing

62

H. V. Roberts

Harris Trust and Savings Bank: An analysis of employee compensation. (1979)
Report 7946, Center for Mathematical Studies in Business and Economics,
University of Chicago Graduate School of Business

Starting salaries (\$U.S.) for 32 male and 61 female clerical hires at a bank

Males			Females					
4,620	5,700	6,000	3,900	4,500	4,800	5,220	5,400	5,640
5,040	6,000	6,000	4,020	4,620	4,800	5,220	5,400	5,700
5,100	6,000	6,000	4,290	4,800	4,980	5,280	5,400	5,700
5,100	6,000	6,300	4,380	4,800	5,100	5,280	5,400	5,700
5,220	6,000	6,600	4,380	4,800	5,100	5,280	5,400	5,700
5,400	6,000	6,600	4,380	4,800	5,100	5,400	5,400	5,700
5,400	6,000	6,600	4,380	4,800	5,100	5,400	5,400	6,000
5,400	6,000	6,840	4,380	4,800	5,100	5,400	5,520	6,000
5,400	6,000	6,900	4,440	4,800	5,100	5,400	5,520	6,120
5,400	6,000	6,900	4,500	4,800	5,160	5,400	5,580	6,300
	6,000	8,100						6,300

Why did Roberts collect these data?

63

The purpose of the research was to demonstrate that female starting salaries are less than male starting salaries. *(Actually the conditional distribution of female salaries was substantially lower than the conditional distribution of male salaries.)*

The Research Hypothesis (*often called the alternate hypothesis*) is what we are trying to prove, i.e.

HA: female starting salaries < male starting salaries
(should have expressed this in terms of conditional distributions)

The Null Hypothesis is the negation of the research hypothesis
i.e. the research hypothesis is null and void.

HO: female starting salaries = male starting salaries
(should have expressed this in terms of conditional distributions)

We gather evidence, data, and reject the Null Hypothesis when the evidence against it is overwhelming.



This is similar to the strategy used in a court of law, where a suspect is brought in with the intention of convicting him of a certain crime. But he is assumed innocent until the weight of evidence against his innocence is overwhelming.

Of course, we can always make an error in our decision



64

The thinking behind a test of a hypothesis can be likened to that in a court of law. The theory behind these tests was developed by Neyman and Pearson and published in *Biometrika* in 1928.

In a law court

	<i>True State of Nature</i>	
<i>Our Decision:</i>	Innocent	Guilty
Aquit		Type 2 error prob = β
Convict	Type 1 error prob = α	

In a statistical test

	<i>True State of Nature</i>	
<i>Our Decision:</i>	Null Hypothesis true	Alternate Hypothesis true
Accept Null Hypothesis		Type 2 error prob = β
Do Not Accept Null Hypothesis	Type 1 error prob = α	

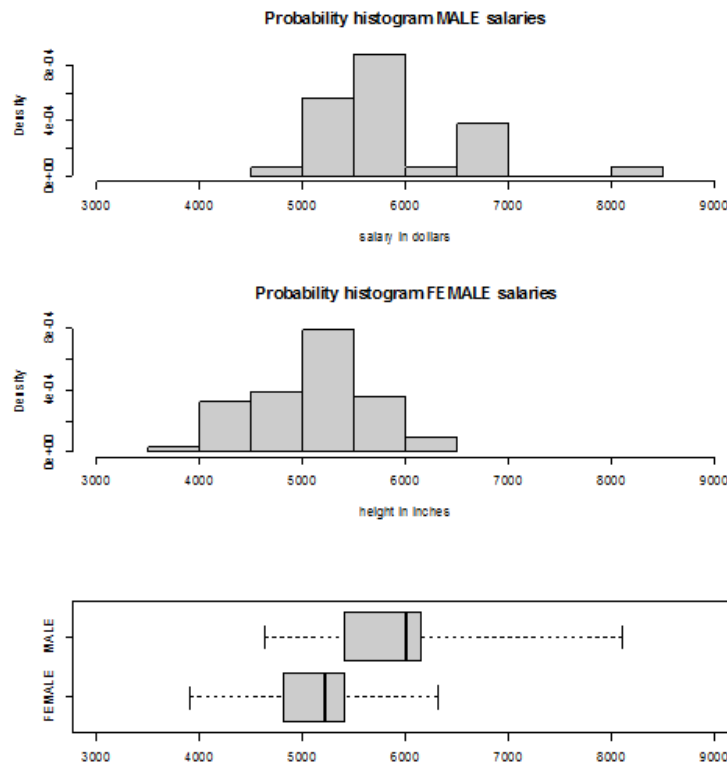
65

Elements of a Test of Hypothesis:

1. *Alternative Hypothesis (Research Hypothesis)*
What we are trying to prove or establish.
2. *Null Hypothesis:*
Usually the nullification of the research hypothesis.
3. *Test statistic or procedure:* How to compare your data with your null hypothesis. A formula or test statistic.
4. *Rejection region:* The values of the test statistic that will lead one to reject the null hypothesis. The rejection region is chosen so that the probability of a type 1 error is small, usually 0.05 or 0.01 that is called the level of significance.
5. *Conclusion:*
If calculated test statistic falls into the rejection region we reject the null hypothesis at the level of significance stated.
If the calculated test statistic does not fall into the rejection region, we do not reject the null hypothesis.

66

Distributions of starting salaries, conditional on gender
using histograms and box-and-whisker plots,
tell a powerful story: distributions have a different shape and location



67

I found the delightful description of a significance test in John Polkinghorn's book *Science and Creation* that he attributes to Batholomew

Let us set up the hypothesis that there is no directing purpose behind the universe so that all change and development is the product of "blind chance". We then proceed to calculate the probability that the world (or that aspect under consideration) would turn out as we find it. If that probability turns out to be extremely small we argue that the occurrence of something so rare is totally implausible and hence that the hypothesis on which it is calculated is almost certainly false. The only reasonable alternative open to us is to postulate a grand intelligence to account for what has occurred. This procedure is based on the logical disjunction *either* an extremely rare event has occurred *or* the hypothesis on which the probability is based is calculated as false. Faced with this choice the rational thing to do is to prefer the latter alternative.

68

Strength of a Statistical Argument

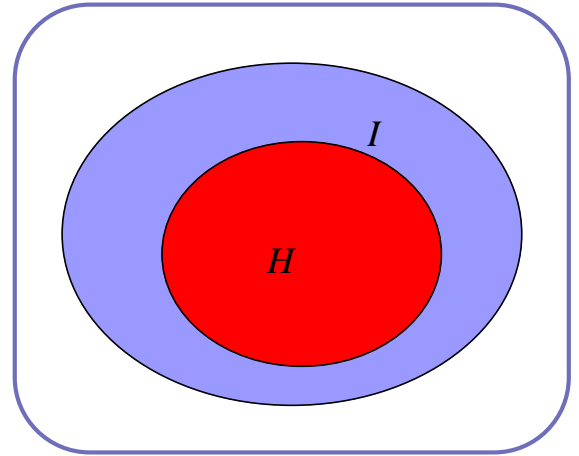
Modus Tollens Argument
(very strong)

Have some hypothesis H

Has some implication I

But evidence shows not I

Therefore conclude not H



Example: Hypothesis: high school results have predictive value for subsequent performance at university
Implication: strong positive correlation coefficient between university performance and high school results.
Evidence: correlation coefficient is found to be small
Conclusion: the hypothesis cannot be true
i.e. high school results not predictive of university results

69

Real life implementation of Modus Tollens

Quasi Modus Tollens Argument
(not so strong)

Have some hypothesis H

together with Assumptions A_1, A_2, \dots, A_k

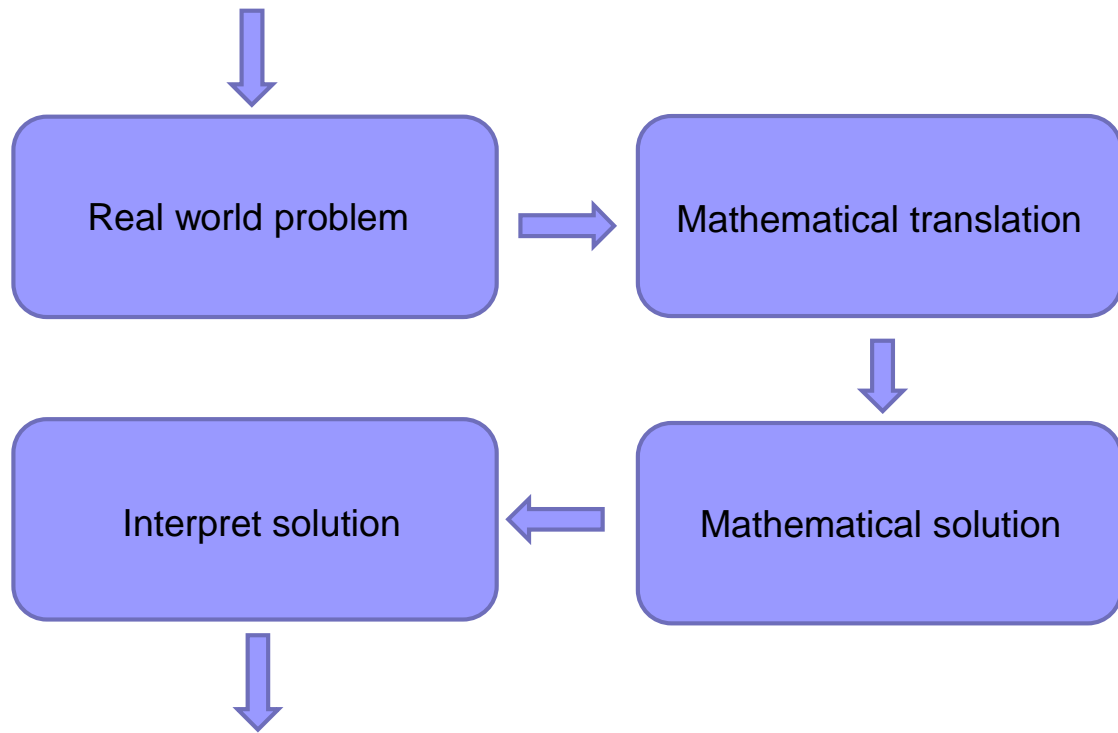
Have some implication I

But evidence shows that I is unlikely

Therefore conclude not H

70

The typical mathematical modeling setting



*we translate our real world problem into mathematics
solve the mathematics problem
translate the mathematical solution back real life . . .
. . . how good are our translations???*

71

Please do worksheet 5

Session 8: ANOVA

73

Variation = sum of squared deviations from the mean

Variance = average variation = variation/degrees of freedom

e.g. the data 1, 2, 3, 4, 5 has mean = 3 and df = 4

variation $(1 - 3)^2 + (2 - 3)^2 + (3 - 3)^2 + (4 - 3)^2 + (5 - 3)^2 = 10$

variance $10/4 = 2.5$

Notice that

As a data set gets larger the variation grows without bound

But the variance converges to the variance of the population

The Analysis of Variance Equation (ANOVA) (Due to Fisher)

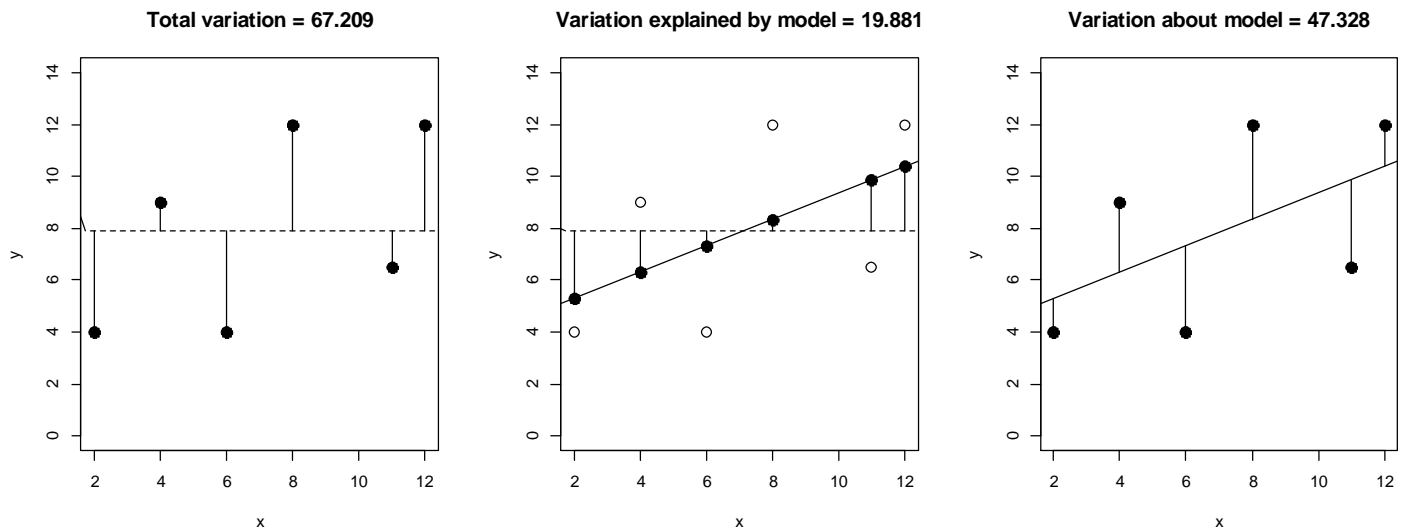
Total variation = variation explained by the model + variation due to noise
= variation explained by the model + variation about the model
= explained variation + unexplained variation

Total variation for a given set of data is fixed

Total df = df for model + df due to noise

74

ANOVA in pictures - for a simple regression model



Analysis of Variance Table for these data

		variation	variance	variance ratio
Source	Df	Sum Sq	Mean Sq	F-ratio
x	1	19.881	19.881	1.6803
Residuals	4	47.328	11.832	
Total	5	67.209		

Percentage variation explained by model = $19.881/67.209 = 29.58\%$
 Variance ratio = 1.68

75

The ANOVA for Fisher's potato data is as follows

Source	Df	Sum-Sq	Mean-Sq	F-ratio	Pr(>F)
nitrogen	3	209646	69882	31.220	4.53e-12
potash	3	32926	10975	4.903	0.00423
Residuals	57	127589	2238		
TOTAL	63	370161			

Conclusions:

Model explains about 66% of the total variation in the data.

The variance ratios suggest : Exceptionally strong evidence against the null hypothesis that that nitrogen is not effective
 Strong evidence against the null hypothesis that potash is not effective

Go back to graphs on slide 4 – see how that adds to the interpretation.
 But, recall the inadequacy of the mathematical model.

76

Do worksheet 6



Bibliography

1. Cole, T. J. (2000), "Galton's Midparent Height Revisited," *Annals of Human Biology*, 27, 401–405.
2. Daly, C. (1964) Statistical Games *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, Vol. 13, No. 2, pp. 74-83
3. Friendly, M. (2008) The Golden Age of Statistical Graphics. *Statistical Science*, Vol. 23, No. 4, pp. 502-535
4. Friendly, M. and Denis, D. (2005) The early origins and development of the scatterplot. *Journal of the History of the Behavioral Sciences*, Vol. 41(2), 103–130
5. Friendly, M. (2004) The Past, Present and Future of Statistical Graphics. (An Ideo-Graphic and Idiosyncratic View). <http://www.math.yorku.ca/SCS/friendly.html>
6. Eden, T. and Fisher, R.A. (1929) Experiments in the response of potato to potash and nitrogen. *Studies in Crop Variation*, Vol XIX, pp 201 - 213.
7. Fisher, R.A. (1921) An examination of the yield of dressed grain. *Studies in Crop Variation*. Vol. XI, pp107 – 135.
8. Fisher, R.A. (1934) The Contributions of Rothamsted to the Development of Statistics Rothamsted Experimental Station Report For 1933 pp 43 – 50
9. Galton, F. (1869). Hereditary Genius: An Inquiry into its Laws and Consequences. London: Macmillan.
10. Galton, F. (1886). Regression towards mediocrity in hereditary stature. *Journal of the Anthropological Institute of Great Britain and Ireland*, 15, 246-263.
11. Galton, F. (1877), "Typical Laws of Heredity," in *Proceedings of the Royal Institution of Great Britain*, 8, pp. 282–301.
12. (1886), "Regression Towards Mediocrity in Hereditary Stature," *Journal of the Anthropological Institute of Great Britain and Ireland*, 15, 246–263.
(1889), *Natural Inheritance*, London: Macmillan.
(1901), "Biometry," *Biometrika*, 1, 7–10.
(1908), *Memories of My Life* (2nd ed.) London: Methuen.
13. Hadley Wickham, Dianne Cook, Heike Hofmann, and Andreas Buja, (2010)
14. Hanley, J. (2004) "Transmuting *Graphical Inference for Inforis*" Women into Men: Galton's Family Data on Human Stature. *The American Statistician*, Vol. 58, No. 3 1
15. Hanley, J. A. (2004), Digital photographs of data in Galton's notebooks, and related material, available online at <http://www.epi.mcgill.ca/hanley/galton>.
16. Hanley, J and Turner, E. (2010) Age in medieval plagues and pandemics: Dances of Death or Pearson's bridge of life? *Significance*, June 2010, 85-87
17. Handley, J., Julien, M., and Moodie, E.E.M. (2008) Student's z, t, and s: What if Gosset had R? *The American Statistician*, February 2008, Vol. 62, No. 1
18. Jacques, J.A. and Jacques, G.M. (2002) Fisher's randomization test and Darwin's data – A footnote to the history of statistics. *Mathematical Biosciences*, Vol 180, 23–28
19. Jaggard, K.W., Qi, A. and Ober, E.S. Possible changes to arable crop yields by 2050. *Phil. Trans. R. Soc. B*, 365, 2835–2851
20. Nievergelt, Y. (2000). A tutorial history of least squares with applications to astronomy and geodesy. *Journal of Computational and Applied Mathematics*, 121, 37-72.
21. Pagano, M. and Anoke, S. (2013) Mommy's Baby, Daddy's Maybe: A Closer Look at Regression to the Mean. *CHANCE*, 26:3, 4-9
22. Pearson, K. (1930). The Life, Letters and Labours of Francis Galton, Vol.III: Correlation, Personal Identification and Eugenics . Cambridge University Press.
23. Stigler, S. M. (1986). The History of Statistics: The Measurement of Uncertainty before 1900. Harvard University Press.

24. Stigler, S. M. (1999). *Statistics on the Table: The History of Statistical Concepts and Methods*. Harvard University Press.
25. Sung Sug Yoon, R.N., Vicki Burt, R.N., Tatiana, L., Carroll, M.D. (2012). Hypertension Among Adults in the United States, 2009–2010. *NCHS Data Brief*, No. 107, October 2012.
26. Wachsmuth, A. and Wilkinson, L. (2003) Galton's Bend: An Undiscovered Nonlinearity in Galton's Family Stature Regression Data and a Likely Explanation Based on Pearson and Lee's Stature Data. Publication details unknown.
27. Wright, K (2013) Revisiting Immer's Barley Data. *The American Statistician*, 67:3, 129-133
28. A review of basic statistical concepts. Author unknown.
29. Diabetes in the UK 2012.
30. Pearson, K. (1896), "Mathematical Contributions to the Theory of Evolution. III Regression, Heredity and Panmixia," *Philosophical Transactions of the Royal Society of London, Series A*, 187, 253–318.
31. (1930), *The Life, Letters and Labours of Francis Galton*, (Vol. IIIA), London: Cambridge University Press.
32. Pearson, K., and Lee, A. (1903), "On the Laws of Inheritance in Man: I. Inheritance of Physical Characters," *Biometrika*, 2, 357–462.
33. Stigler, S. (1986), "The English Breakthrough: Galton," in *The History of Statistics*:
34. *The Measurement of Uncertainty before 1900*, Cambridge, MA: The Belknap Press of Harvard University Press, chap. 8.
35. Tredoux, G. (2004), Web site <http://www.galton.org>.
36. Wachsmuth, A., Wilkinson, L., and Dallal, G. E. (2003), "Galton's Bend: A Previously Undiscovered Nonlinearity in Galton's Family Stature Regression Data," *The American Statistician*, 57, 190–192.
37. Paul J Lewi *Speaking of Graphics* that can be found at <http://www.datascope.be/sog.htm>

"The Power to See: A New Graphical Test of Normality," Aldor-Noiman, S., Brown, L. D., Buja, A., Rolke, W., Stine, R.A., *The American Statistician*, 67 (4), 249 {260 (2013).

"Valid Post-Selection Inference," Berk, R., Brown, L., Buja, A., Zhang, K., Zhao, L., *The Annals of Statistics*, 41 (2), 802 {837 (2013).

"Statistical Inference for Exploratory Data Analysis and Model Diagnostics," Buja, A., Cook, D., Hofmann, H., Lawrence, M., Lee, E.-K., Swayne, D.F., and Wickham, H., *Philosophical Transactions of the Royal Society A*, 367, 4361 {4383 (2009).

"The Plumbing of Interactive Graphics," Wickham, H., Lawrence, M., Cook, D., Buja, A., Hofmann, H., and Swayne, D.F., *Computational Statistics*, (April 2008).

"Visual Comparison of Datasets Using Mixture Distributions," Gous, A., and Buja, A., *Journal of Computational and Graphical Statistics*, 13 (1) 1 {19 (2004).

"Exploratory Visual Analysis of Graphs in GGobi," Swayne, D.F., Buja, A., and Temple-Lang, D., refereed proceedings of the *Third Annual Workshop on Distributed Statistical Computing* (DSC 2003), Vienna.

"GGobi: Evolving from XGobi into an Extensible Framework for Interactive Data Visualization," Buja, A., Lang, D.T., and Swayne, D.F., *Journal of Computational Statistics and Data Analysis*, 43 (4), 423-444 (2003).

March 2016

XGobi: Interactive Dynamic Data Visualization in the X Window System," Swayne, D.F., Cook, D., and Buja, A., *Journal of Computational and Graphical Statistics*, 7, 113{130 (1998).

March 2016

Worksheet 1: Descriptive statistics:

Question 1:

Plot the data 2, 3, 5, 8, 12 on a hand drawn dot diagram below

Compute the mean for these data ($Sum = 30$)

Compute the variation for these data ($sum\ of\ squares\ of\ data\ about\ their\ mean$)

Compute the variance of these data ($variation / df$)

Compute the sd of these data ($square\ root\ of\ variance$)

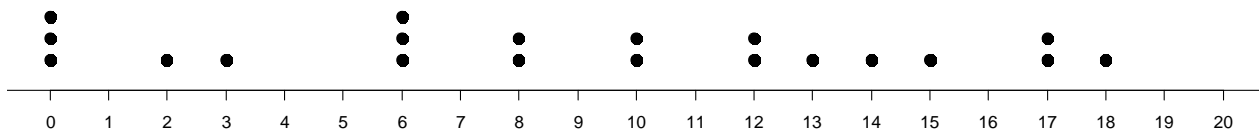
Answers:

Question 2

The number of people in the lunch queue at noon at the Mathematical Institute on 20 working days during January was:

15, 8, 10, 0, 17, 12, 18, 8, 13, 14, 17, 0, 10, 12, 3, 6, 0, 2, 6, 6

order statistics: 0, 0, 0, 2, 3, 6, 6, 6, 8, 8, 10, 10, 12, 12, 13, 14, 15, 17, 17, 18



Dotplot of data

Compute the three quartiles.

Create a hand drawn Box-and-whisker plot above the dotplot.

Visually estimate the mean and sd.

LQ = any number between 3 and 6. Some would use 4.5

Median = any number between 8 and 10. Some would use 9.

UQ = any number between 13 and 14. Some would use 13.5.

(about 9 and 4)

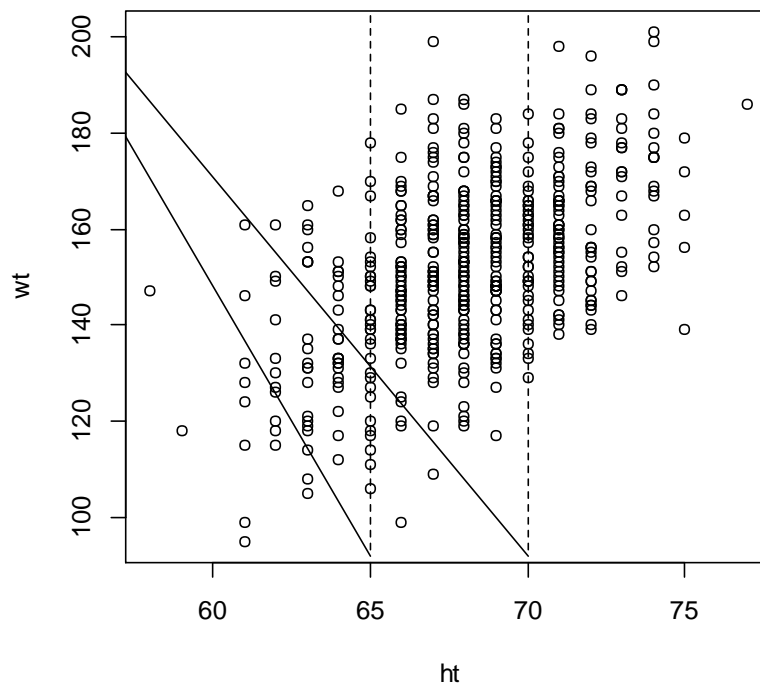
Answers:

March 2016

Question 3:

The following scatterplot shows wt (the weight in lbs of an individual) plotted against ht (height in inches of an individual) for the button data (to come later). Estimate visually the means and sd's for the conditional distributions of weights at heights of 65 and 70 inches respectively.

Plot of weight vs height for button data



Answers:

R code for dotplot

```
x <- c(0, 0, 0, 2, 3, 6, 6, 6, 8, 8, 10, 10, 12, 12, 13, 14, 15, 17, 17, 18)
y <- c(1, 2, 3, 1, 1, 1, 2, 3, 1, 2, 1, 2, 1, 2, 1, 1, 1, 1, 2, 1)
points <- 0:20
plot(x,y,pch=19, cex=1.5,xlab="Dotplot of data", xlim=c(0,20), ylim=c(0,10),
     at=points,cex.lab=2)
```

R code for plot of weight vs height

```
bd <- read.table("E:/buttondata.txt",header=T)
attach(bd)
plot(ht,wt,main="Plot of weight vs height for button data")
abline(v=65,lty=2);abline(v=70,lty=2)
```

Worksheet 2: Probability: 10 Min

Task 1: The standard statistical model is: what you observe = truth + error/noise

What contributes to the noise in the reduced Galton data on midparent height and child height?

Answers:

Task 2: What meanings of probability are invoked in the following statements:

- a Smokers are 23 times more likely to get lung cancer than are non-smokers
How would one compute this?
How might this statement be re-phrased?
- b Insurance costs are usually based on risk.
Women get a 40% discount on motor insurance in South Africa.
Does this mean that women are better drivers? Discuss.
- c You bought four tickets in a lottery. What are your chances of winning?

Answers:

March 2016

Worksheet 3: Sampling and Confidence intervals:

Do this during lunch break

1. Record your data in excel.
2. Use Excel to compute means, sd's, and the prevalence of diabetics and hypertensives i.e. %(diabetics), %(hypertensives). Compute the quantiles by sorting data smallest to largest.
3. Compute 95% confidence intervals for each of the parameters measured.

case	ht	wt	BMI	bp	db	
1						
2						
3						
4						
5						
6						
7						
8						
9						
10						
mean						average
sd						stdev.s
min						min
LQ						
MED						
UQ						
max						max
LCL						
UCL						

LCL = Lower Confidence Limit

UCL = Upper Confidence Limit

a 95% CI for the mean (ht or wt or BMI) of the population is:

$$\text{sample average} \pm 1.96 \times \text{standard deviation} / \sqrt{\text{sample size}}$$

a 95% CI for the prevalence (%diabetic or % hypertensive):

$$\text{sample prevalence} \pm 1.96 \times \sqrt{\text{prevalence} \times (1 - \text{prevalence}) / \text{sample size}}$$

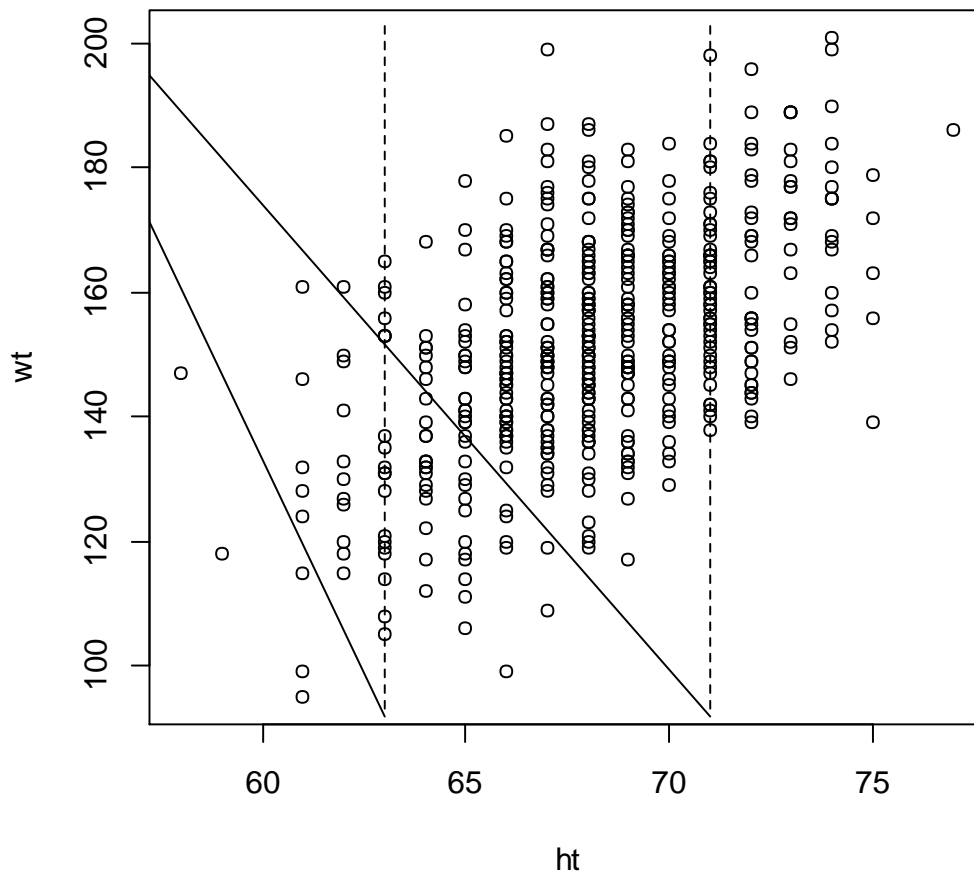
Use the following functions in Excel:

average, stdev.s, min, max

Worksheet 4: Regression:

A plot of weight vs. height for the button data is given in Figure 1 below.

Plot of weight vs height for button data



1	Assume the linear regression model holds for these data. What are the assumptions of the linear regression model?
2	Locate the means of the conditional distributions at $ht = 63$ and 71 visually and fit a linear regression line through these data by hand.
3	Estimate the slope of your fitted line. NB: Slope = rise/run

Continued overleaf

March 2016

4	Estimate the standard deviation of the data about the regression line by visually.
5	Describe the conditional distributions of weights at heights of 63 and 71 inched.
6	Can one predict weight from height? What can one predict? What are the uncertainties?
7	Can you describe or interpret what the regression model is telling us?

R code for plot of weight vs height

```
bd <- read.table("E:/buttondata.txt",header=T); attach(bd)
plot(ht,wt,main="weight vs height for button data");
abline(v=63,lty=2);abline(v=71,lty=2)
```

Worksheet 5: Hypothesis formulation and testing:

Question 1: Discuss possible the type I and type II errors in the following settings. One might also want to consider the problem from the perspective of different individuals or groups.

1	<p>A new drug for treating HIV/Aids is proposed and a clinical trial is designed to compare it with the existing drug.</p> <p>Possible individuals groups: The researchers who propose the drug, the person with HIV/Aids and their families, the drug companies who produce drugs for treating HIV/Aids,</p>
2	<p>A teenager gets caught up in gang violence and is given a life sentence, at age 18, for the conviction of second degree murder of an opposing gang member.</p> <p>Possible individuals groups: The teenager who is sentenced and her family, the family of the person who was killed, the legal or justice community given the burden of justice, the community at large.</p>
3	<p>A student is accused of plagiarism, her dissertation is rejected and she is excluded from the university because of it.</p> <p>Possible groups: Student. University. Community.</p>
4	<p>For the salary data, the null hypothesis is that the average salaries paid to men and women are equal, and that the observed differences were due to other circumstantial factors.</p> <p>Possible groups: Males. Females. The bank. Community.</p>

Question 2: In a Galton regression setting, where one is comparing the conditional distributions of child height at given values of Midparent heights, what is the null hypothesis of interest for this research problem. What would be considered sufficient or convincing evidence for rejecting the null hypothesis.

Hint: Whatever is computed from a random sample, is itself a random variable, whose sampling distribution we can compute.

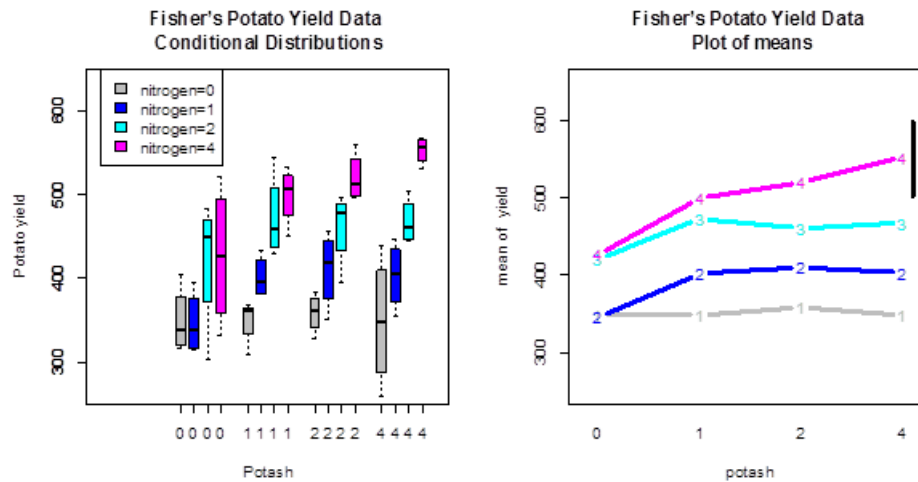
Question 3: In a general research setting, how would one statistically test any null hypothesis?

Hint: Whatever is computed from a random sample, is itself a random variable, whose sampling distribution we can compute.

March 2016

Worksheet 6: Fisher Potato Data ANOVA : 10 min

Question 1: For the Fisher data there are a number of hypotheses one might wish to consider about the effect of potash, nitrogen and the interaction of potash and nitrogen.



	Formulate conceptually and in words the mathematical model for these data.
	What would be sensible null hypotheses for this experiment?
	How convincing is the evidence against the null hypotheses judging from the graphs? Does one in fact need a statistical test against the null hypotheses? Or is the evidence against the null hypotheses simply overwhelming?

ANOVA for Potato Data

Source	Df	Sum Sq	Mean Sq	F value	p-value
nitrogen	3	209646	69882	32.299	2.72e-11
potash	3	32926	10975	5.073	0.00413
nitrogen:potash	9	18925	2103	0.972	0.47556
Residuals	45	97361	2164		
TOTAL	60	358858			

p-value = measure of type I error probability.

Interpret the ANOVA for these data keeping in mind the assumptions of the model and the possible effect on the computed statistics of the model assumptions being violated.