CHAPTER 1: INTRODUCTION

1 Using statistics in the humanities

Were the people who emigrated from England to North America in the nineteenth century poor people fleeing poverty or prosperous skilled labourers seeking opportunities in a new country? Has the number of the people speaking Welsh in Wales increased in the past twenty years? When Jane Austin died in 1817 she left assets worth around £800. Was £800 a lot of money in 1817? [1]

Following the 1948 US presidential election why did the Chicago Tribune feel confident enough to run the headline "Dewey defeats Truman" before all the votes were counted only for it to be become evident that Truman had actually won? [2]

The average life expectancy in the early 1800s was about 40 years of age. Does this mean that there were no old people? The 2011 UK census reveals that 59% of people in the UK identify themselves as 'Christian',[3] but only 15% attend a Christian place of worship at least once a month. [4] What does this tell us about the relationship between identity, belief and practice?

2 Types of Quantitative Data

The answers to all of these questions rely on some sort of understanding of numbers and interpreting them. This book is a beginner's guide to statistics which uses examples from the humanities subjects. The case studies used are from archaeology, history, languages, linguistics, religious studies and area studies. I do not assume any prior knowledge of statistics other than that you know how to add, subtract, multiply and divide with the aid of a pocket calculator.

Some sources such as surveys and censuses are created with the express purpose that they will generate numerical data which will then be analysed. Whether it is a company researching the market for a new project, the government surveying the whole population in order to plan services in the longer term or educationalists undertaking questionnaires of primary school teachers about their opinions on learning a second language, these sources are intrinsically numerical. In a democratic society the principle of elections is that those with most votes are chosen. Again, the data is numerical by its very nature. Other sources are not designed to be quantitative, but quantitative data can be derived from them with ease. From Parish Registers it is possible to derive numerical data about length of life, age at marriage, family size, infant mortality and maternal death in childhood. Changes in these factors can be monitored over decades and centuries and different geographical areas can be compared. Financial data in the form of prices, taxation and public and private spending can provide insight into historical and contemporary conditions. These can be monitored over time or countries might be compared. We can get numerical information on the occupation, age, nationality and health of emigrants to the USA from immigration data and from ships' passenger lists. Again, trends can be identified and monitored over a period of time. Quantitative data can also be derived from non-quantitative sources such as newspapers. How many incidents of arson were there in Sussex in the 1830s? Although not every incident would necessarily be reported in the newspaper it is possible to count up the number of stories or column inches given to certain topics. It is possible to derive quantitative data from any source. The Oxford English Dictionary annually announces new words which have been invented or have come into more common use in the previous year. [5] Recent additions have included 'credit crunch', 'staycation' and 'jeggings'. OED staff monitor written language use to identify new trends. Researchers in linguistics might count incidents of an individual using certain phrases or metaphors or how many times they pause or say 'um', 'err' or 'ah'. These can be compared between individuals, sexes, languages or native and non-native speakers.

3 Where do statistics come from?

Statistics prove …", or do they? Some people think of statistics as a way of proving or disproving a

particular argument or relationship. We might begin an argument by saying "Statistics prove …" or "Statistics show …". In fact statistics are not morally, ethically and epistemologically neutral. There are inherent biases in the statistics we and others collect and why we collect them. These biases reflect the values both of those who collect data or statistics (including governments) and how we interpret them. We can only use statistics that are available to us whether we collected them ourselves or acquired them from other surveys or from other documents. We cannot make arguments on the basis of statistics we do not have. The UK census has taken place every ten years since 1801 (with the exception of 1941), but the questions have changed to reflect changes in society and changing views of what the Government needs to know about people. From 1951 to 1991 the census asked people if they had an inside toilet. In 1951 a lack of access to inside sanitation was seen as an important indicator of social deprivation. However, by 1991 there were very few houses without an inside toilet and the question was dropped. For the first time in 2011 the population was asked how well they could speak English. Our own reasons for being interested in a particular topic also impact on how we use statistics.

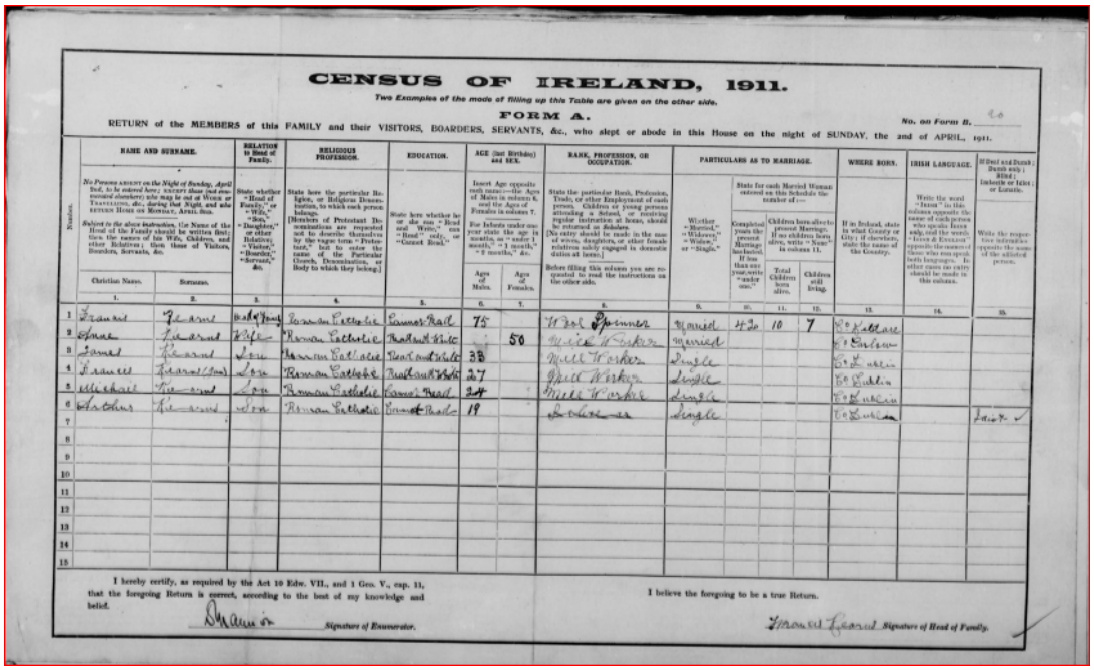
As researchers we all have our own values which are reflected in the things we are interested in and how we might use data relating to these topics. When we use the data produced by other people we are often examining them for a different purpose. The government of 1801 did not start collecting census data to make it easier for future generations to research their family history though many people use the census for this purpose. Taxation records were created for the purposes of collecting tax, not for producing maps of relative wealth in different parts of a city, though researchers have used these records to do just that. Today, schools count up how many pupils have free meals so that they know how many meals they need to cook and get reimbursement for. However, the proportion of children on free school meals is frequently used to measure social deprivation in any given school. A school will be considered 'deprived' if a high proportion of its pupils are eligible for free school meals. Schools do not ask parents to provide details of their income for the purpose of measuring deprivation. Statistics are also often used to support legal, moral or ethical arguments. Opinion polls are used by advocacy groups to demonstrate that the public is supportive, not supportive or doesn't care about alcohol regulation, abortion time limits, gay marriage or euthanasia. Such polls can be useful for governments unsure whether to proceed with a particular piece of legislation or policy change, but the amount of support for an opinion does not prove that one side is wrong and one side is right and researchers should use this sort of data with caution. In chapter 19 we will be exploring survey design and some of the implications this can have on the conclusions we come to about attitudes and beliefs.

4 References

Table 1: 1911 Census return for the Kearns family of Dublin. Note that the 19-year old son Arthur is described as an 'idiot'. Other people were listed as imbeciles or lunatics. These categories were considered scientific at the time.

Surname	Forename	Age	Sex	Relation to head	Religion	Specified Illnesses
Kearns	Francis	75	Male	Head of Family	Roman Catholic	0
Kearns	Anne	50	Female	Wife	Roman Catholic	0
Kearns	James	33	Male	Son	Roman Catholic	0
Kearns	Francis	27	Male	Son	Roman Catholic	0
Kearns	Michael	24	Male	Son	Roman Catholic	0
Kearns	Arthur	19	Male	Son	Roman Catholic	Idiot

Figure 1 Full original census return for the Kearns family. Unlike the UK Census, the 1911 Census of Ireland is available free online.



CENSUS OF IRELAND, 1911.

Two Examples of the mode of filling up this Table are given on the other side.

FORM A.

RETURN of the MEMBERS of this FAMILY and their VISITORS, BOARDERS, SERVANTS, &c., who slept or abode in this House on the night of SUNDAY, the 2nd of APRIL, 1911.

No. on Form B. 20

Number.	NAME AND SURNAME. Christian Name.	Surname.	RELATION to Head of Family.	RELIGIOUS PROFESSION.	EDUCATION.	AGE (last Birthday) and SEX. Ages of Males.	Ages of Females.	RANK, PROFESSION, OR OCCUPATION.	PARTICULARS AS TO MARRIAGE. Whether "Married," "Widower," "Widow," or "Single."	Completed years the present Marriage has lasted. If less than one year write "under one."	Children born alive to present Marriage. If no children born alive, write "None" in column 11. Total Children born alive.	Children still living.	WHERE BORN.	IRISH LANGUAGE.	If Deaf and Dumb; Dumb only; Blind; Imbecile or Idiot; or Lunatic.
	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.	15.
1	Edward	Kearns	Head	Roman Catholic	Can read	75		Retired Farmer	Married	42	10	7	Co. Kildare		
2	Anne	Kearns	Wife	Roman Catholic	Can read and write		50	Home Duties	Married				Co. Carlow		
3	Edward	Kearns	Son	Roman Catholic	Can read and write	33		Mill Worker	Single				Co. Kildare		
4	Patrick	Kearns	Son	Roman Catholic	Can read and write	27		Mill Worker	Single				Co. Kildare		
5	Michael	Kearns	Son	Roman Catholic	Can read	24		Mill Worker	Single				Co. Kildare		
6	Matthew	Kearns	Son	Roman Catholic	Can read	19		Labourer	Single				Co. Kildare		Deaf. ✓
7															
8															
9															
10															
11															
12															
13															
14															
15															

I hereby certify, as required by the Act 10 Edw. VII., and 1 Geo. V., cap. 11, that the foregoing Return is correct, according to the best of my knowledge and belief.

D. Allen Signature of Enumerator.

I believe the foregoing to be a true Return.

Edward Kearns Signature of Head of Family.

1].National Archives website. Famous wills: Jane Austin http://www.nationalarchives.gov.uk/museum/item.asp?itemid=33 http://www.ons.gov.uk/ons/rel/census/2011-census/key-statistics-for-local-authorities-in-england-and-wales/rpt-religion.html

[2]Chicago Tribune November 3, 1948

[3]Office for National Statistics (2012) Religion in England and Wales

[4] J. Ashworth et al (2007) Churchgoing in the UK: A research report from Tearfund on church attendance in the UK(London: Tearfund)

[

CHAPTER 2: HOW MANY AND HOW BIG?

1 How many and how big?

Counting is one of the first skills we learn as children and is an important milestone in a child's development. Counting is simply the question "How many are there? How many people were killed in the Holocaust? How many people were transported from Africa to the Americas as slaves? How many people speak Basque? How many mosques are there in Birmingham? How many copies of Harry Potter novels have been sold? How many Norse burial sites are there in the Orkney Islands? How many people lived in Liverpool at the time of the 1851 census? An accident which kills fifty people is more likely to get more time on the news than a similar accident which only claims one life. The news that more people are speaking Basque than before or that more people are unemployed then were six months ago may lead to calls for changes in policy. The statistic that six million Jews were killed in the Holocaust invokes a sense of moral outrage.

2 Why count?

Some linguists have suggested that "humans possess an innate number sense. Counting is essentially the first stage of working with numbers. Whether consciously or not we use counting to make analytical and moral judgements, often by using non-statistical language as we talk about numbers. Think of words like up, down, common, uncommon, important, major and minor. Although these words do not relate directly to any specific scale they invoke judgements about numerical scale. We think of success and historical events in terms of numbers, even if these numbers are unknown. Would Martin Luther King have come to prominence if he was one of only a small number of people supporting the Montgomery Bus Boycott? Would the Paris riots of 1968 still be talked about if only a handful of people participated over the course of a couple of hours? Even if we do not use statistics in our work we use words which employ ideas of size and scale (see Table 1).

Counting helps us to identify trends which are taking place or took place in the past;

For Example:

1. We can see if numbers of Welsh speakers are going up or down or how they went up or down in the past.

2. We can see how the population of Liverpool is going up or down or how it went up or down in the past.

3. Counting can challenge conventional wisdom about social phenomenon such as numbers of nineteenth century brides who were pregnant at the time of their marriage.

4. We can use numbers to make a qualitative judgement about the importance, scale, severity or impact of an event. For example, two earthquakes can have the same magnitude, but if one takes place in a city and another in an uninhabited area, the former is likely to impact on more people than the latter.

Table 1 Everyday words with an idea of numbers

Smaller number	,	Bigger number
A few	A lot	All
Uncommon	Common	Universal
Impoverished		Wealthy
Few	Many	Most

3 Problems in counting

Superficially, counting is a straightforward skill. However counting can become difficult in a number of situations.

3.1 Missing, unavailable or non-existent data

1. Data that never existed We are only able to count data that exists. We don't know what proportion of households in the UK had an inside toilet in 2001 and 2011 because this data was not collected.

2. Data that is missing or destroyed. Even if data was collected it can be lost or destroyed, deliberately or accidentally. Documents can fall victim to fires, floods, rodents as well as wear and tear damage.

3. Data that does exist, but will not be available until a future date. The original UK census returns will not be made public for 100 years after the census data. Therefore we cannot use data which depends on access to original census returns of 1921 and later. Some police, prison and legal records are also restricted. UK Government Papers are released after 30 years, but some information is restricted for reasons of national security

4. Data is missing because its subjects were deliberately or accidentally excluded, through their own actions or those of others. Even though censuses are an attempt to collect data on an entire population they are always an undercount. People may try to avoid censuses if they are in the country illegally or are concerned that the census is being used for taxation or military purposes. Many people oppose the census on the grounds that it invades their privacy

5. Data can be inaccurate. It may have been wrongly entered into a computer, mistakes may have been made in making calculations or someone might have lied when reporting data.

3.2 Non-comparable data

When we compare data taken at two time periods it may appear that we are comparing like with like. For example the question, 'how many people speak Welsh?'has been asked regularly over the years, but in different ways. There are lots of ways to asking questions to get a sense of someone's knowledge of the Welsh Language but they may lead to different answers, (See Table 2).

Comparisons across countries need to be undertaken with some degree of caution. The calculations and methods used for measuring inflation and unemployment, for example, sometimes differ between countries.

3.3 Changing geographical boundaries over time

1. Changes of borders within a country. What was the population of Oxfordshire in 1881? We need to clarify whether we are talking about Oxfordshire as it is now or Oxfordshire was it in 1881. The town of Abington was in Berkshire until 1974 when it was moved into Oxfordshire. There are numerous cases like this in the UK so they need to be checked. The exact administrative boundaries of towns and cities have also changed over time.

2. Movement of national boundaries can be a more difficult subject as useful data which was collected in the past might no longer be and vice versa. For example, the Alsace region of France has been part of both France and Germany over the past 400 years and was subject to the data collection regimes of both countries during different parts of the nineteenth and twentieth centuries.

3.4 Double (and triple) counting:

We will talk more about classifying data in the next section, but when we put data into different categories we can an end up counting the same data two or three times. Suppose you were counting the number of incidents of machine vandalism and arson which took place during the Swing Riots in Berkshire? What would you do if you came across a single incident in which protesters had vandalised a machine, then set fire to the barn in which it was housed? Is this one incident or two? Would you count it once or twice? What is being counted?

3.5 Differing definitions:

In October 2012 1.58 million people were registered as unemployed in the UK, a rate of 7.8 %. But who counts as unemployed? Possible answers include:

- Number of people who don't have jobs.
- Number of people claiming job seekers' allowance
- Number of people not working, but looking for a job.
- Number of people of working age who do not have a job.
- Number of people who could work, but are not working.
- Adults who do not have a job and are not studying

Table 2: Welsh language questions of the 1981, 1991 and 2001 censuses. [2]

1981 census For all persons aged 3 or over (born before 6 April 1978)	1991 census For all persons aged 3 or over (born before 22 April 1988) "	2001 census Can you understand, speak, read or write Welsh?"
"Does the person speak Welsh?	Speaks Welsh	Understand spoken Welsh "
If the person speaks Welsh, does he or she also:	Reads Welsh	Speak Welsh
"Speak English?	Writes Welsh	Read Welsh
Read Welsh?	Does not speak, read or write Welsh "	Write Welsh
"Write Welsh?		None of the Above

In order to make meaningful comparisons, definitions need to be agreed. The International Labour Organisation uses the following definition of unemployment:

> *"An unemployed person is defined by Eurostat, according to the guidelines of the International Labour Organization, as someone aged 15 to 74 without work during the reference week who is available to start work within the next two weeks and who has actively sought employment at some time during the last four weeks. The unemployment rate is the number of people unemployed as a percentage of the labour force."* [3]

4 Summary

The questions `How many?' and `How large?' are the starting point of any statistical analysis. Counting is not always as straight forward. We have to deal with changing questions, changing geographical boundaries, missing or inaccurate data, double counting and different definitions.

5 Exercises

Examine the Welsh language questions from the 1981, 1991 and 2001 UK censuses. How do they differ? How could they lead to different answers?

6 References

[1]C. Holden (2012), Life without numbers in the Amazon, Science 305 p.109e: An aggregate analysis, Area, 36(2), 187-201

[2]G. Higgs, C. Williams, and Dorling, D. (2004)Use of the Census of Population to discern trends in the Welsh language

[3]Eurostat: http://epp.eurostat.ec.europa.eu/statisticsexplained/index.php/Unemploym

CHAPTER 3: SUMMARISING DATA

1 Introduction

Lists of numbers and tables of data are useful, but a few statistical measures can usefully summarise a whole data set. We will read in the newspaper that the average income in the UK is £26,500. [1] The average weekly wage of an agricultural worker in 1850 was 9 shillings 3 12 pence. [2] We also use the word `average' in a qualitative sense. We might say that a student is of average academic ability or that we live in an average-sized house. When we talk about an average we are summarising a larger set of data in one figure. So when we say the average income is £26,000 it acknowledges that some people earn more than £26,000 and other people earn less, but someone on middle income earns around £26,000. We often associate the term `average' with `normal'. A person on an average income is not rich and not poor. A student of average ability is neither the one of the highest performers, nor one of the lowest performers. This chapter will show you how to calculate mode, median and mean, upper and lower quartiles and will address some of the issues surrounding the use of the mean, median and mode.

2 Mean

There are actually several types of average, but the most familiar average used is the mean average. This is calculated by adding the observations together then dividing by the number of observations. The following is a list of the heights in centimetres of 10 soldiers who enrolled in the French army in 1790. By adding all the heights together, then dividing our answer by the number of soldiers we can find the mean height: So:

$$\frac{168+165+165+168+168+165+165+173+175+165}{10}=\frac{1677}{10}=167.7$$

When you see the mean average reported academic papers you may see it written as $\bar{x}$ and spoken as `bar x' or `x bar'. If different averages are being compared you may see the different averages written as $\bar{y}$ or $\bar{z}$

We can present the sum above as an equation where x is one observation (in this case a soldier's height), $\bar{x}$ is the mean height of the soldiers (the mean of the x's) and n is the number of observations. As we have 10 observations we can write each of these as x_1 ,x_2, x_3 … etc. So our formula for calculating the mean is

$$\bar{x}=\frac{x_1,x_2,x_3,x_4...etc.x_n}{n}$$

$\bar{x}$ = Mean average.
n=Number of observations.
x_1 =Soldier 1's height
x_2 = Solider 2's height etc., up to x_{10} which is soldier 10' s height.

3 Median

The median average is simply the observation which comes in the middle. The following example comes from the register of burials in Accrington, Lancashire. Five people buried in succession died at the following ages:

8,20,32,17,82

All we need to do to find the median is to put the ages in order.

8,17,20,32,85

The median average is the one in the middle. In this case the median age at burial is 20 years of age. There are five observations of which two are below the median and two above. If we have an even number of observations we have a situation where there is no single middle number. In the example below we have six observations.

1,8,17,20,32,85

To find the median we must take the middle two values (17 and 20) and divide them by 2. This will give us our median.

$$\frac{17+20}{2}=18.5$$

The median of these six observations is 18.5. Notice that there are three observations which are less than 18.5 and three observations which are more than 18.5.

4 Mode

A third type of average is the mode. The mode is simply the value which occurs most frequently. Below we have added some more burial ages from Accrington to those we used in the example for the median.

8,20,32,17,82,0,0,22

In this example we can see that the most frequently occurring value is 0. Therefore the mode of this sample of burials is zero years of age.

5 Making sense of averages

An average summarises a set of data in one number. Each type of average has its own strength and weaknesses.

The mode is the least frequently used form of average. It only uses one number from the dataset. It is mostly used for describing nominal data (that is data with names or categories). For example if we did a questionnaire which asked people to name their religion and the most commonly occurring religion was Christian we would say that the mode or modal group was `Christian'. There is not a median or mean religion, sex, race or national identity.

The advantage of a mean average is that it takes account of all the observations. However, taking account of all values can be misleading. A few very high or very low values skew the data to give a misleading view of the data as a whole. This is very common in the case of income data where a small number of wealthy people drive the mean income up to a level which does not reflect anyone's income. For example consider the following five incomes

£18,000;£22,000;£20,000;£28,000;£100,000

The mean income is

$$\frac{18,000+22,000+20,000+28,000+100,000}{5}=37,600$$

As we can see the mean income is greater than four of these five incomes. The statement that the mean average income is £37,000 would be correct but it does not summarise the data very well. The median income, £20,000, is a much more realistic reflection of the income earned by four of these five people.

6 Case study: Life expectancy

A similar issue occurs when examining life expectancy. It is common to hear that the average person living in nineteenth century England had an average life span of around 30 years. This seems to suggest that most people were dead before the age of 40. Does this mean that there were no old people in the nineteenth century? Let's return to our example from Accrington in 1838. Table 1 records the age of death of 39 people buried that year.

The mode conveys that sad reality that the most common age to die was before the age of one. However, although it was the most common age at which to die it does not mean that most people died before the age of one. The mean age of death was 25.1 and the median 20 years. Both these averages indicate that people were able to live sufficiently long enough to have children themselves and deaths of people in their late teens and twenties were clearly common. All three averages fail to reflect the fact that people did live into their seventh, eighth and ninth decades. Not very many people lived this long, but it was not impossible. So there were old people living in nineteenth century England.

This pattern is also important in understanding the age profiles today in countries with low life expectancies. Societies with high levels of infant mortality reduce the mean, median and mode age of death far below that of societies where infant mortality is very low. A closer examination of the figures suggests that anyone who manages to live beyond the age of about four has a very good chance of reaching adulthood.

Life expectancy is particularly interesting in statistical terms. It is commonly expressed in terms of life expectancy at birth, but as you get older your life expectancy continues to increase.

As early as the seventeenth century an estimated 10% of the population were over 60. [3] Additionally, your life expectancy never stops increasing, so a 100 year-old's life expectancy is clearly over 100 years, much more than a five year-old or a 65 year-old.

To find the mean average we add together the 39 numbers then divide by 39. When we do this we find that the mean average age of death was 25.1 years old.

$$8+20+32+17+82+0+0+22+47+0+38+25+57+1+0+21+39$$

$$+15+54+48+1+3+78+1+29+22+63+41+73+1+2$$

$$\frac{+35+80+13+0+12+0+0+0}{39}=25.1 years$$

To find the mode average we look for the value which occurs the most frequently. In this case the most common age at which to die was 0. 8 out of the 39 people died before they were one year old.

To find the median average we put all the ages in order. The median is the value with the same number of values

Table 1 Age of death in Accrington, 1838

Age	Number	Age	Number	Age	Number
0	8	20	1	41	1
1	4	21	1	47	1
2	1	22	2	48	1
3	1	25	1	54	1
8	1	29	1	57	1
12	1	32	1	63	1
13	1	35	1	73	1
15	1	38	1	78	1
17	1	39	1	80	1

before it an after it. In this case our median is 20 years. 19 people died younger than this 20 year old and 19 people were older when they died.

To summarise, the mean average age of death was 25.1 years old. The mode average age of death was 0 years. The median average age of death was 20 years. You will notice that these three averages give us very different answers (see Figure 1). But, which is the most useful average?

7 The five figure summary

The above section has demonstrated some of the hazards of relying on the average value (whether mean, median or mode) alone. The five figure summary consists of

1. the median,
2. upper quartile
3. lower quartile
4. minimum observation
5. maximum observation.

We will calculate the five figure summary for a sample of farms in Chile. [4] The data below shows the size of 24 farms in Chile (in hectares).

It can be seen that farm sizes in Chile vary considerably.

0.5, 3.5, 15.1, 508.3, 0.7, 3.5, 13.1, 1701.7, 0.2, 7.1, 39, 10.1, 1.2, 8, 57, 19.5, 1.5, 9.9, 198.2, 4.9, 2.4, 6.2, 276.4, 3

8 Calculating the median

There are 24 farms altogether. We find which place the median is in by There are 24 farms altogether. We find which place the median is in by (

$$\frac{24+1}{2}=12.5$$

First we sort the farms out into order of size

510.2,0.5,0.7,1.2,1.5,2.4,3.0,3.5,3.5,4.9,6.2,7.1,8.0,9.9,10.1,13.1,15.1,19.5,39.0,57.0,198.2,276.4,508.5,1701.4

So if we were to place all the farms in order of size the median would be the farm in 12th and 13th place. As

Figure 1 Different types of average illustrated: Life expectancy in 1830s Accrington

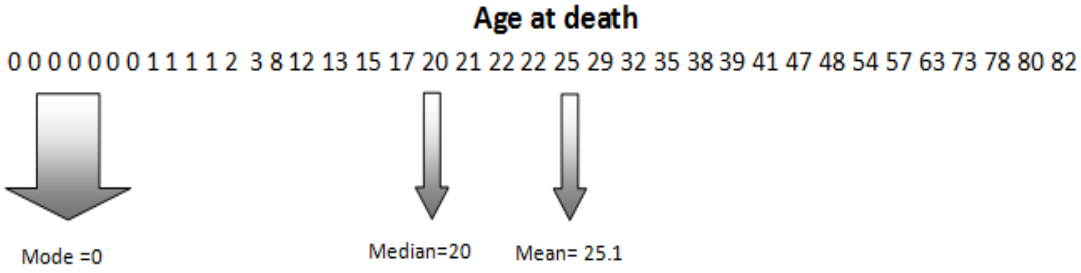


Figure 2 Boxplot: Farms in Chile

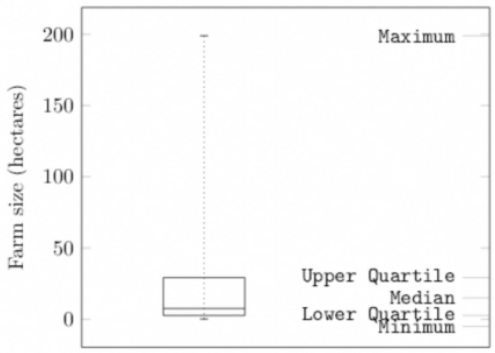
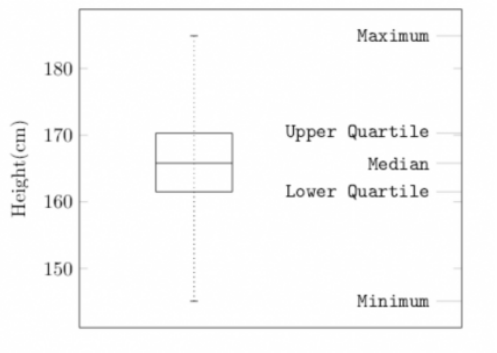


Figure 3Boxplot: Heights of Bavarian



there are an even number of farms and the median lies between two places we need to find:

$$\frac{7.1+8}{2}=7.55$$

It is instantly clear at this point that the median size farm at 7.55 hectares is considerably smaller than the mean at 120.5 hectares. This would seem to indicate that there are a high number of smaller farms and a small number of bigger farms. We can learn a bit more about the distribution of farm sizes by calculating the quartiles.

9 Upper and lower quartiles

The lower quartile lies halfway between the median and the lowest value. The upper quartile lies halfway between the median and the highest value. To find the lower quartile we need to find the median of all the values below the median of the whole dataset. There are 12 values below the median

0.2,0.5,0.7,1.2,1.5,2.4,3.0,3.5,3.5,4.9,6.2,7.1

To find where the median of these 12 values is located (12+1)2=6.5 The median lies between 6th and 7th place 0.2,0.5,0.7,1.2,1.5,2.4,3.0,3.5,3.5,4.9,6.2,7.1 (2.4+3.0)2=2.7 So our lower quartile is 2.7.

To find the upper quartile we need to find the median of all the values above the median. There are 12 values above the median: To find where the median of these 12 values is located

$$\frac{12+1}{2}=6.5$$

The median lies between 6th and 7th place

$$\frac{19.5+39.0}{2}=29.25$$

So our upper quartile is 29.25 So far we have three numbers for our five number summary:

Our median: 7.55

Our lower quartile: 2.7

Our upper quartile: 29.25

We need two more numbers for our summary, but these are easy to find. We need the smallest value and the largest value. The smallest number is 0.2 and the highest is 1701.4.

Therefore our five numbers are:

1. Minimum value: 0.2
2. Lower quartile: 2.7
3. Median 7.55
4. Upper quartile 29.25
5. Maximum value 1701.4

10 Presenting the five figure summary

This five figure summary gives us a better idea about the distribution of farm sizes. We can see that half of the farms are less than 7.55 acres, but a quarter are less than 2.7 hectares. This indicates that the vast majority of farms are small. The largest farm in the sample is huge in comparison, over 255 times the median and 58 times the size of the farm which forms the upper quartile. We can represent this graphically in the form of a box-plot . The ends of the box mark the two quartiles and the line inside the box is the median. The lines coming out of the box are known as whiskers. The whiskers go from the top of the box to the maximum value and from the bottom of the box to the minimum

Table 2: Number of booksellers registered for Value Added Tax(VAT) Turnover Size (thousands)

Turnover Size (thousands)	Number of booksellers
0-49	160
50-99	215
100-249	310
250-499	180
500-999	85
1000-4999	35
5000+	15
TOTAL	1000

Table 3 Number of booksellers registered for Value Added Tax(VAT) with midpoints calculated

Turnover Size (thousands)	Number of booksellers	Lower boundary	Upper boundary	Number of booksellers	Mid point
0-49	160	0	49	160	24.5
50-99	215	50	99	215	74.5
100-249	310	100	249	310	174.5
250-499	180	250	499	180	374.5
500-999	85	500	999	85	749.5
1000-4999	35	1000	4999	35	2999.5
5000+	15	5000	10000	15	7500
TOTAL	1000				

value. We can see from the boxplot (see Figure 2) the median and lower quartiles lines are very near the bottom of the box plot. We say the distribution is skewed.(Instructions on how to draw a boxplot can be found in Chapter 20. We can also get a sense of how skewed the data is by calculating the mean

(0.7+0.2+1.2+1.5+2.4+4.9+3.0+3.5+3.5+7.1+8.0+9.9+6.2+10.1+19.5+15.1+13.1+39+57+198.2+276.4+508.3+1707.7) ÷ 24=120

The mean farm size is 120 hectares, yet only four of the 24 farms are above the mean. A small number of large farms are so large that they are skewing the mean towards it. The boxplot in Figure 3 represents the heights of 20,000 Bavarian conscript soldiers in the early nineteenth century. It can be seen clearly that the soldier's heights are much more evenly distributed than the Chilean farms. The median height is 168cm and the mean height is 167.3 cm. If a soldier of mean height stood next to a soldier of median height it is unlikely that you would notice the difference. Although a small number of tall soldiers skew the mean slightly it is clear that most of soldiers have heights close to the mean. When most of the values are equally distributed around the mean, we might say that the values are normally distributed. We will be discussing the normal distribution in the next chapter. The five figure summary (minimum, lower quartile, median, upper quartile and maximum) is a useful way of describing our data which takes all the observations into account. Large

Table 4 Number of booksellers registered for Value Added Tax(VAT) with midpoints calculated and group turnover

Lower boundary	Upper boundary	Number of booksellers	Mid point	Group turnover
0	49	160	24.5	3920
50	99	215	74.5	16017.5
100	249	310	174.5	54095
250	499	180	374.5	67410
500	999	85	749.5	63707.5
1000	4999	35	2999.5	104982.5
5000	10000	15	7500	112500
		1000		422632.5

Table 5 Calculating the cumulative frequency Class

Class	Number of booksellers	To calculate cumulative frequency	Cumulative frequency
0-49	160	160	160
50-99	215	160+215	375
100-249	310	160+215+310	685
250-499	180	160+215+310+180	865
500-999	85	160+215+310+180+85	950
1000-4999	35	160+215+310+180+85+35	985
5000-10000	15	160+215+310+180+85+35+15	1000

Table 6 Mid points and cumulative frequency

Lower boundary	Upper boundary	Number of booksellers	Mid point	Cumulative frequency
0	49	160	24.5	160
50	99	215	74.5	375
100	249	310	174.5	685
250	499	180	374.5	865
500	999	85	749.5	950
1000	4999	35	2999.5	985
5000	10000	15	7500	1000

differences between the mean, median and mode indicate that our data is skewed. We will explore this further in the next chapter.

11 Dealing with data in classes

Sometimes we don't have access to all the data, but we are given a summary of the data classified into groups (sometimes referred to as `bins'.) The booksellers statistics are from: The Publishers Association (2011)[5] showing the turnover of 1000 booksellers is a good example of this. It tells us that that there were 160 booksellers with a turnover of between 0 and 49. (the numbers are in thousands here so 49 actually means 49,000). Although we know that 160 booksellers were making between 0 and 49,0000 we don't know the exact amount each one is making.

However, we can calculate a five figure summary for this grouped data group by calculating the midpoint for each bin. We add the lower boundary of the group to the upper boundary then divide by 2. For example to find the midpoint of 0-49 we add together 0 (the lower boundary) and 49 (the upper boundary) then divide the answer by 2.

$0+49=49$

$49 \div 2 = 24.5$

Table 3 shows the groups with all the midpoints calculated.

When we calculate the midpoint we are effectively assuming that all the booksellers with a with a turnover of between 0 and 49,000 had an actual turnover of 24,500. It is unlikely that this is actually the case, but it is the best estimate we can make with the data we have.

Mean

To find the mean average we need to estimate the amount of money that all the booksellers turn over. In order to do this we need to add another column to Table 3 to produce Table 4 which calculates the group turnover of each `bin'. We do this by multiplying the mid-point turnover by the number of booksellers.

Now to calculate the mean average we divide the total turnover of all the book sellers, but the number of booksellers

$442632.5 \div 1000 = 442.6325$

As the numbers are in thousands this makes our mean 442,632.50.

Median

To work out the median, lower quartile and upper quartile we need to calculate the cumulative frequency. This means starting with the number of booksellers in the lowest group then adding on each the number of book sellers in the next group (see Table 5).

As we don't know the exact turnover figures we cannot be sure of the exact median; however, we can find out what class it is in. As we have 1000 booksellers the median lies between the 500th and 501st booksellers. If we look at our cumulative frequency we can see that the 500th and 501st booksellers lies in the 100-249 group. As the mid point of this group is 175.5 we can say that the median group is 175.5, actually 175,500 .

The upper and lower quartiles

Finding the upper and lower quartiles is straight forward from here. The lower quartile is the median of the lower half of the book sellers meaning the bookseller in 250th place marks the lower quartile. $0+500\ 2 = 250$ The 250th bookseller is in the 50-100 group. As the midpoint of this group is 74.5 we can say that the lower quartile is 74.5. Similarly the upper quartile is the median of the upper half of book sellers meaning the bookseller in 750th place marks the upper quartile

$(500+1000) \div 2 = 750$

The bookseller in 750th is in the 250- 499 group. As the mid point of this group is 374.5 we can say the upper quartile is 374.5.

12 Exercises

1. Examine Table 7 . Calculate the mean, median and mode length of time each became Prime Minister. (Treat Winston Churchill's and Harold Wilson multiple times as Prime Minister as different entries).

2. Table 8 displays the turnover of UK booksellers by group. What do you think are the a) advantages and b)limitations of displaying the data in groups.

3. Table 9 shows the number and capacity (in tons) of freight wagons used on the Barbados Railway in the 1930s. [6]

a) Calculate the total capacity of the railway in tons.
b) Calculate the mean, median and modal wagon capacity.

References

1. ↑ BBC(2012) Average earnings rise by 1.4% to £26,500, says ONS. http://www.bbc.co.uk/news/business-

Table 7 British Prime Ministers since 1940

Prime Minister	Time as PM (years)
Gordon Brown	3
Tony Blair	10
John Major	7
Margaret Thatcher	11
James Callaghan	3
Harold Wilson (second term)	2
Edward Heath	4
Harold Wilson (first term)	6
Alec Douglas-Home	1
Harold Macmillan	6
Anthony Eden	2
Winston Churchill (second term)	4
Clement Atlee	6
Winston Churchill (first term)	5

Table 8 Number of booksellers registered for Value Added Tax(VAT) Market Research and Statistics Available from www.publishers.org.uk

Turnover Size (thousands)	Number of booksellers
0-49	160
50-99	215
100-249	310
250-499	180
500-999	85
1000-4999	35
5000+	15
TOTAL	1000

Table 9: Number and capacity (in tons) of freight wagons used on the Barbados Railway. Data from: Jim Horsfield (2001) From the Caribbean to the Atlantic: A Brief History of the Barbados Railway St. Austell: Paul Catchpole

Class	Number	Capacity (tons)
A	18	7
AA	31	8
B	2	6
C	10	6
D	4	8
E	2	15
AA converted	4	6
B converted	5	6
C converted	4	6
D converted	3	6
BK converted	2	6
BP converted	6	6