Contents

1 Summarizing Data **3**

1.1 Types of Data . . . 4
1.1.1 Quantitative data . . . 4
1.1.2 Qualitative data . . . 4
1.1.3 Caveats . . . 4
1.2 Summarising Qualitative Data . . . 5
1.2.1 Numerical Summaries . . . 5
1.2.2 Graphical Summaries . . . 6
1.3 Summarising Quantitative Data . . . 6
1.3.1 Numerical Summaries . . . 7
1.3.2 Graphical Summaries . . . 10
1.4 Table 1 . . . 12
1.5 Summarising Data Practical . . . 14
1.6 Hand Calculations . . . 14
1.7 Summarising Data in Stata . . . 16
1.8 `Summarise` . . . 17
1.9 Summarise by group. . . . 17
1.10 `tabstat` . . . 18
1.11 `table` . . . 18
1.12 Further exercises . . . 19

2 Summarizing Data Practical **21**

2.1 Hand Calculations . . . 22
2.2 Summarising Data in Stata . . . 24
2.3 `Summarise` . . . 25
2.4 Summarise by group. . . . 25
2.5 `tabstat` . . . 26
2.6 `table` . . . 26
2.7 Further exercises . . . 27

1 Summarizing Data

Having collected some data, you need to be able to describe it. Simply presenting all of the data to others is impractical, so there needs to be some way of summarising the data. This may be done graphically or numerically: both approaches have advantages and disadvantages. However, different types of data require different methods, so we must first consider the types of data we may be dealing with.

1.1 Types of Data

1.1.1 Quantitative data

Quantitative data is any kind of data that is meaured on an interval scale. It makes sense to do mathematics with the values that the data takes. For example, height and weight are both quantitative variables. In particular, it makes sense to use subtraction on these measures: for example if you weigh 75kg today and last week you weighed 76kg, you could say you have lost 1kg.

We further subdivide quantitative data into two types: discrete and continuous. Discrete data can only take a fixed number of distinct values, whereas continuous data can take any value within a given range. So height and weight are continuous, whereas number of children or number of visits to the G.P. this year are discrete.

1.1.2 Qualitative data

Qualitative, or categorical, variables are used when it is necessary to classify each observation into one of a number of groups. For example sex, ethnicity and marital status are all qualitative variables.

It may, in some cases, be possible to impose an ordering on a categorical variable. For example, a patient may be asked to describe their pain as none, mild, moderate or severe. Clearly, in this case, severe pain is always worse than any of the other categories. Such data is referred to as *ordinal*

1.1.3 Caveats

It is not always straightforward to decide what type of data you are dealing with. Indeed, it is often the case that the same variable may be appropriatedly treated as a different type in different circumstances.

Nominal vs ordinal

Whilst it may be possible to impose an ordering on a categorical variable, and thus treat it as ordinal, it may not necessarily be helpful. For example, hair colour is a nominal variable, but it could be ordered according to how dark it is, and this may be of use if the question being investigated relates to melanin.

Ordinal vs Discrete

If a person is asked to rate their pain on a scale of 1 to 10, that looks very like a discrete (or possibly even continuous, if people give fractional scores). However, treating it as discrete makes an assumption that a change in pain from 1 to 2 is somehow the same as a change in pain from 9 to 10. It is unlikely that anyone is capable of being that precise and rational about their own pain, so it would probably make more sense to treat it as ordinal.

Sometimes quantitative data may be grouped before you receive it, and thus become ordinal. For example, the number of years of education a person has received is a discrete variable. However, if education is simply classified as primary, secondary or university, so much information about the actual number of years has been lost that it would make more sense to treat that variable as ordinal.

Continuous vs Discrete

Technically, there is no such thing as a continuous variable in a computer analysis. There will always be a smallest possible increment in the variable that can be recorded by the computer, which makes the variable discrete. However, it makes sense to treat the variable as continuous since that smallest possible increment may be so small as to be negligible.

However, it may be that the smallest measurable increment is substantial. For example, I once worked with a study that involved measuring the diameter of the brachial artery on a computer screen. The resolution was not great, and the artery would be either 2, 3 or 4 pixels wide. Although the true diameter would be a continuous variable, the data we had to work with was discrete.

1.2 Summarising Qualitative Data

Qualitative data is easier to summarise simply because we cannot do maths with it. All we can do is say how many observations belong to each category. Those numbers may then be presented either numerically or graphically.

1.2.1 Numerical Summaries

When describing qualitative data numerically, it is best to give both the number and the proportion (or percentage) in each category. Proportions are useful since they are not affected by sample size: the proportion in each category will be roughly the same for a study of 1,000 people and a study of 100,000 people, but the numbers in each category will be completely different.

The stata command you would use to produce a numerical summary of a categorical variable is `tabulate`. The example below shows the results of typing `tabulate region` on a dataset taken from the

```
region |      Freq.     Percent        Cum.
------------+-----------------------------------
     Canada |        422       22.84       22.84
        USA |        541       29.27       52.11
     Mexico |        223       12.07       64.18
     Europe |        493       26.68       90.85
       Asia |        169        9.15      100.00
------------+-----------------------------------
      Total |      1,848      100.00
```

This shows the number of subjects recruited in each of 5 regions, in the column labelled "Freq.". The percentage from each region is in the column labelled "Percent". The column labelled "Cum." for "cumulative percentage" is completely meaningless for a nominal variable like region, but stata always produces it anyway. It only has any meaning for an ordinal variable: the categories of a nominal variable can be presented in any order, and changing the order would change the cumulative percentage columm.

1.2.2 Graphical Summaries

In general, a graphical summary of a qualitative variable will take up more space than a numeric summary, but for many people it will be easier to absorb the information in this form. When writing a paper, there would normally not be space to produce graphical summaries of every qualitative variable in the dataset (although they could always be included as supplementary material if necessary), but it may be worth including one or two for the most important variables, or if they clearly show an important feature of the data.

The best graphical summary of a qualitative variable is a bar chart. In this representation, there is a bar for each category, and the length of the bar is proportional to the number of observations in that category. The lengths of the bars may be labelled with either the number or the proportion in each category: this has no effect on the shape of the chart at all.

The vertical axis may be labelled with the number of observations that belong to that group, or the proportion. This does not change the shape of the graph in any way. Figure 1.1 is an example of a bar chart: it shows the number of patients recruited by region in the SLICC study.

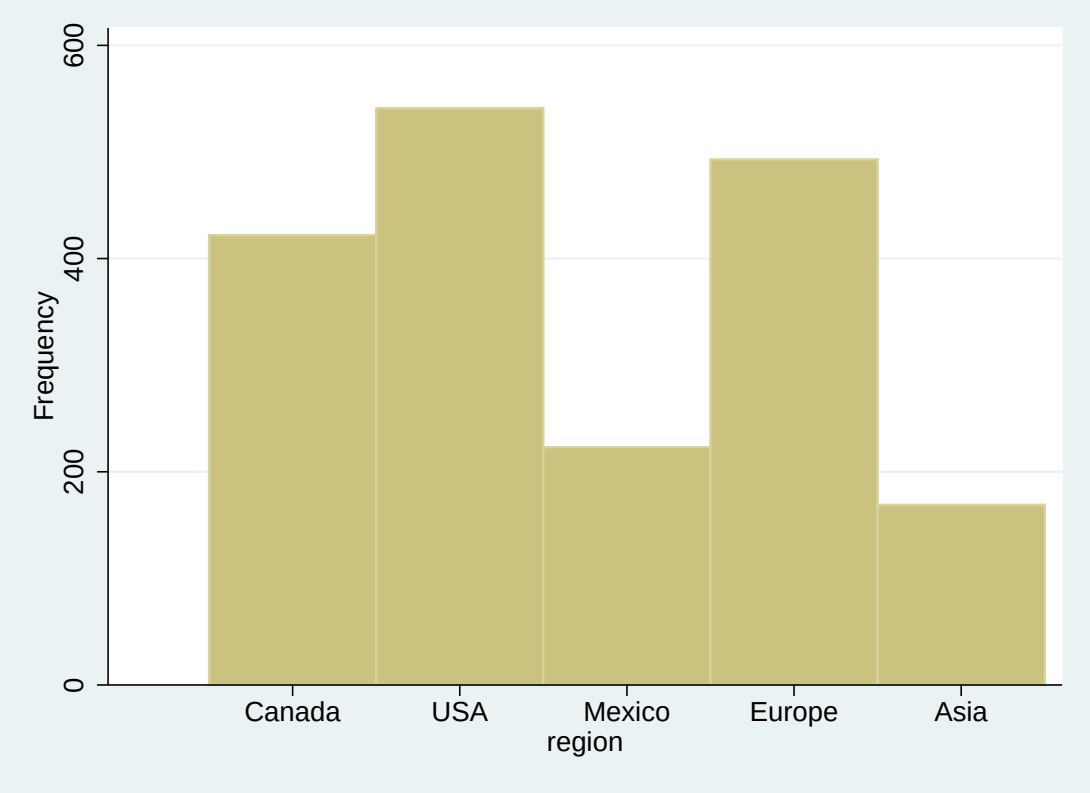


Figure 1.1: Bar Chart for Recruitment by Region

1.3 Summarising Quantitative Data

One simple approach to summarising numerical data is to divide the data into groups and treat the groups as a categorical variable. It may not even be necessary to create your own groups if the numerical variable is discrete with a reasonable small number of possible values. This approach is commonly used for age: the sample is divided into 5-year or 10-year age-bands,

depending on the range of ages in the sample, the number and proportion of people in each age band is given.

However, there a also a number of other ways to summarise numerical data. These summaries use the fact that it is possible to do mathemetics with the values of the variable, for example to calculate a mean.

1.3.1 Numerical Summaries

With numerical or quantiative data, there are a number of characteristics of the data that can be described numerically. The most important of these answering the questions "What is a typical value ?" (measure of location)and "How much do the values vary ?" (measure of scale). Some other descriptive statistics are discussed in section 1.3.1, but they are not widely used. This is partly because they are unnecessary for normally distributed data: given measure of location and a measure of scale, a normal distribution is completely determined.

Measures of Location

The most commonly used measures of location are the mean and the median. The mean, often written with a bar over the variable name ($\bar{x}$, pronounced "x bar"), is calculated by adding all of the values, then dividing by the number of values. In other words

$$\begin{aligned} \bar{x} &= \frac{x_1 + x_2 + \ldots + x_n}{n} \\ &= (\Sigma_{i=1}^{n} x_i)/n \end{aligned}$$

The median does not have a special notation, and is calculated by arranging all of the values in order, and picking out the middle one. If there are an even number of values, take the mean of the middle two.

Changing a single value will always change the mean (albeit very slightly in a large sample), but it is far less likely to change the median. The median can only change if the value that is changed goes from below the median to above it (or vice versa), or it is the middle observation (or one of the two middle observations if there is an even number. This makes the median far less sensitive to outliers, and it can be a better measure of location if the data is highly skewed.

For example, the list of numbers below are the number of days of absence from work due to sickness:

1,1,2,2,3,3,3,4,4,4,5,6,6,6,7,7,8,10,10,38,80

The mean value is 10, despite the fact that almost all of the observations are smaller than 10. This is not a good answer to the question "What is a typical value ?". The reason for this is the fact that there are two very large observations (38 and 80) which increase the mean considerably. The median for this data is 5, which gives a much better idea of a typical value: 9 of the 21 observations are within 1 day of this value, and 12 (more than half) are within 2 days. This is why the median should always be preferred to the mean when giving a measure of location for skewed data.

Measures of Scale

There are a number of statistics that can be used as a measure of how much a numerical variable varies.

Range You will occasionaly see the range of the data presented, that is the smallest value and the largest value. This is not a good idea, since it is determined completely by the two most extreme measurements. This means that as the sample size increases, the range can never decrease. If the new observation lies within the range of existing observations, the range is unchanged. If it lies outside this range, the range increases.

Inter-Quartile Range A better, yet still simple, measure of scale is the inter-quartile range. This is calculated by finding the quartiles of the data, those values that lie one quarter of the way through the list of observations in order and three quarters of the way through. For the sickness days data shown above, there are 21 observations, so the lower quartile would be at observation $\frac{21}{4} = 5.25$, i.e. between the 5th and 6th observation. Since both of these observations take the value 3, the lower quartile is 3. The upper quartile would be at observation $\frac{3\times21}{4} = 15.75$, i.e. between the 15th and the 16th observation. Since both of these observations are 7, the upper quartile is 7 and and the inter-quartile range is (3, 7). This means that the central half of the data lies between 3 and 7.

Standard Deviation A third possible measure of scale is the standard deviation. This can be thought of as the average distance of an observation from the mean. This is not quite true: thanks to the definition of the mean, the mean distance the observations in a sample from the sample mean is 0.

Since the average distance from the mean is not helpful, we sometimes use the average of the *squared* distance from the mean. This quantity, called the variance, can be calculated as

$$\text{Variance} \quad = \quad \Sigma(x_i - \bar{x})^2/n$$

The disadvantage of this statistic is that it is not in the same units as x, but in the square of that unit. For example, for the sickness days data, the variance is 316.2 days^2.

Variance is a particularly unintuitive measure of variability, as it is difficult to conceptualise what a squared day is. If we take the square root of the the variance, we get the standard deviation:

$$\text{Standard Deviation} \quad = \quad \sqrt{\Sigma(x_i - \bar{x})^2/n}$$

This statistic is in the same units as the original variable, which makes it easier to interpret. For example, the standard deviation of the sickness days data is 17.8 days. This seems quite high for a measure of how far the observations are from a typical value on average, but his is because it is strongly influenced by the outlying values of 38 and 80.

Other Descriptive Statistics

Quantiles and Percentiles We have seen that the median is a value such that half of the data is less than or equal to it, and half of the data is greater than or equal to it. However, there is nothing magical about half of the data: we could have a value such that one third of the data is less than or equal to it, and two thirds of the data is greater than or equal to it: this is the lower tertile.

Tertiles and quintiles are commonly used in data analysis: they can be used to divide your data up into a small number of equally sized groups. However, I would not recommend this common approach, since another study would have different tertiles and quintiles and generated

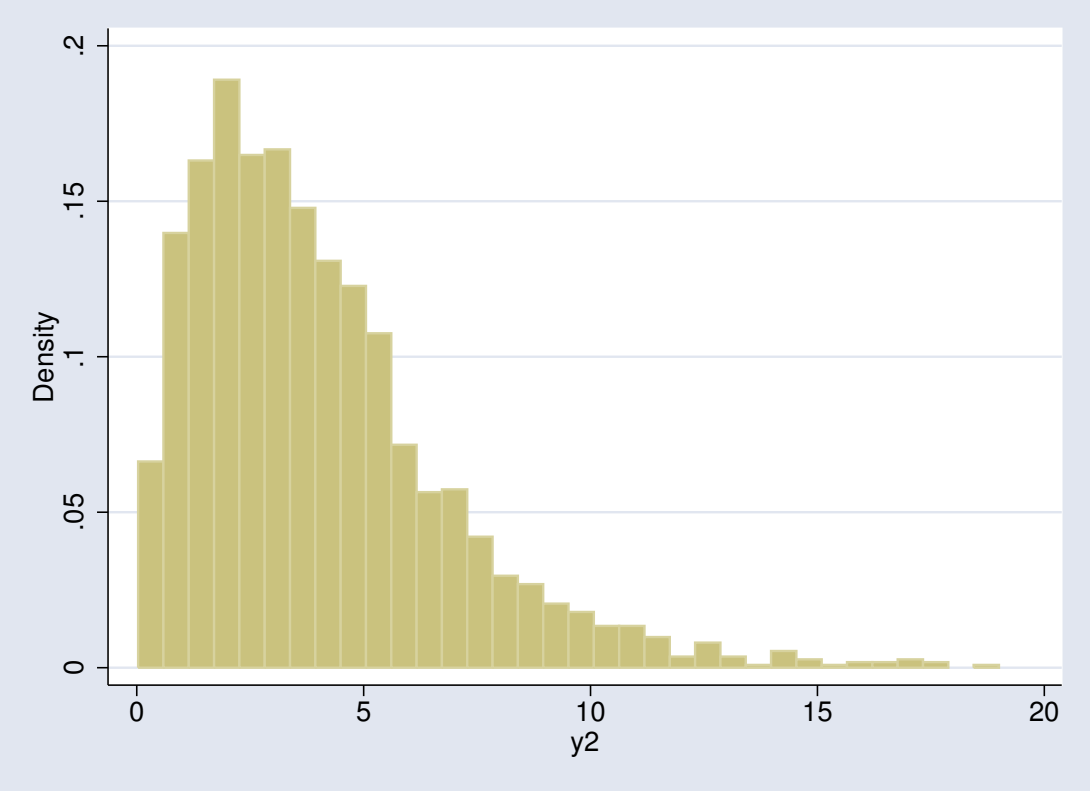


Figure 1.2: Positively Skewed Distributionn

different boundaries for their groups, thus making it harder to compare studies. Far better to use meaningful values for the variable in question: for example, if you wish to categorise BMI, use < 18; $18 - 25$; $25 - 30$; and > 30, since these are widely accepted thresholds, even if you do end up with groups of differing sizes.

It is possible to define any quantile (dividing the data into fractions, such as thirds or quarters) or percentile (selecting the bottom $x\%$ of the data, provided that there are enough observations.

Higher Moments If a variable follows a normal a normal distribution, the mean and the standard deviation are sufficient to describe it exactly. All quantiles and percentiles can be calculated if these two numbers are known. However, there are an infinite number of differently shaped distributions that have the same mean and standard deviation. For example, it may have some extremely high values (positively skewed, see Figure 1.2) or some extremely low values (negatively skewed). It may have more than one peak (bimodality, see Figure 1.3), usually due to two (or more) distinct populations being mixed and treated as one.

There are additional statistics that can be presented to describe the shapes of these distributions. For example, *skewness* is a measure of how lop-sided the distribution is, and *kurtosis* is a measure of how much of the weight of the distribution is the tails. Just as the variance is based on the squared differences between the observations and the means, skewness is based on cubed differences and kurtosis on differences raised to the power 4. The exact formulae are a bit more

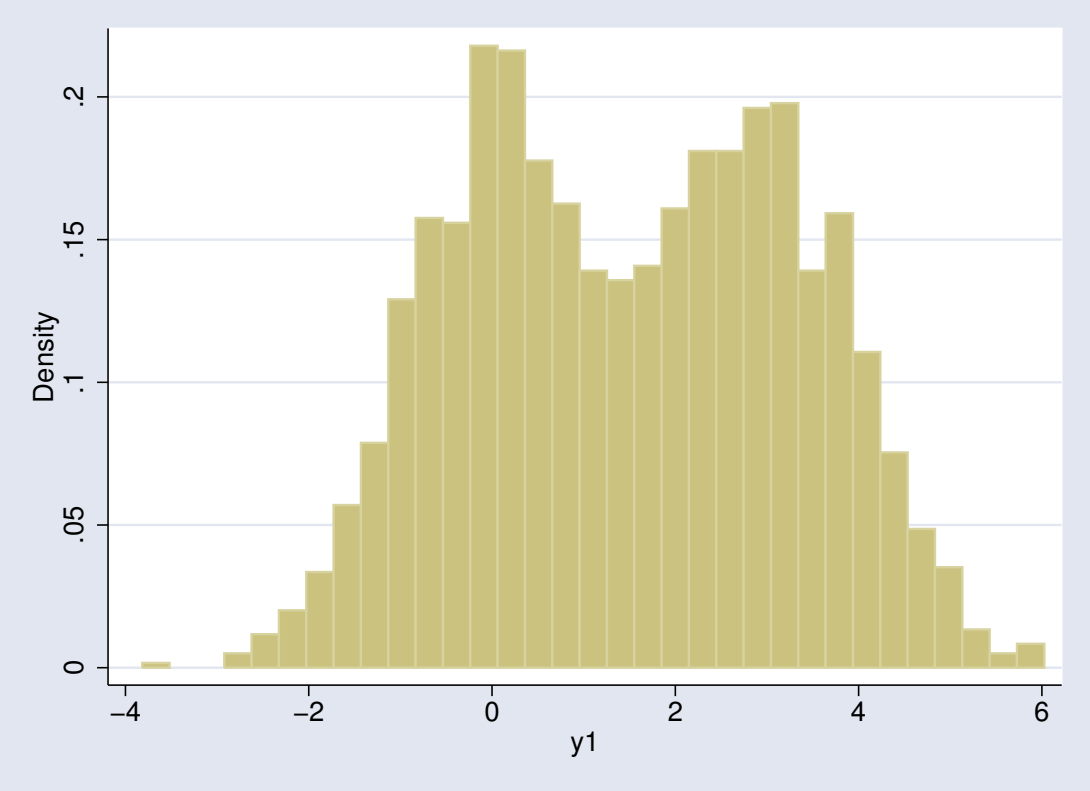


Figure 1.3: Bimodal Distribution

complex than the variance formula, and not of any interest, so look them up on wikipedia if you really want to see them. Skewness and kurtosis are rarely presented when giving a numerical summary of a quantitative variable, since they are not well understood. If you are presenting a distribution in which the non-normality matters, it is better and clearer to give a graphical summary.

1.3.2 Graphical Summaries

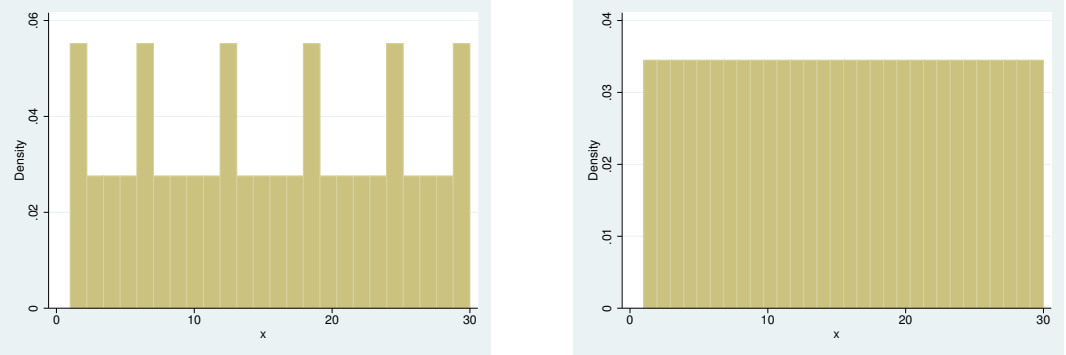
Histogram

A histogram is similar to a bar chart, but is used for quantitative, rather then qualitative variables. Rather than having pre-determined groups on the x-axis, the width of each bar can be determined individually. You can, in theory, have bars of different widths, although there is no way to achieve that in Stata. Which is a shame, since it could be useful: if you are recruiting subjects aged 18–50, it would make sense to have a group for those less than 20, then divied into 5 or 10 year age-bands. You would need the first bar to have a width of 2, the rest to have widths of 5 or 10.

The big difference between a bar chart and a histogram is that with a bar chart, the *height* of the bar is proportional to the size of te group, whereas with a histogram, it is the *area* of the bar. However, if the bars are all of equal width, those two statements are equivalent.

Setting the bar width correctly can be essential to produce a meaningful graph, and there is no single formula which tells you what it should be, although there are a number of recommendations based on the sample size and the range of the data. Particularly with discrete data,

it can be very important to set the bin width to an integer value. To see why, consider Figure 1.4. By default, Stata used 24 bins for 30 discrete values, meaning that some bins contained 2 values whilst most contained only 1. This gave the characteristic "comb" effect seen in Figure 1.4(a). Ensuring that each value had its own bin gave Figure 1.4(b), revealing that each value was equally likely to occur.



(a) 24 bins (default

(b) 30 bins (correct

Figure 1.4: Effet of Changing Number of Bins with a Discrete Variable

Bar charts and Histograms in Stata Both bar charts and histograms are produced with the command `histogram`. You can set the number of bars in the chart with the option `bin()`, or the width of each bar with the option `width()`. Alternatively, the option `discrete` tells stata to produce a bar for each value in the dataset. Stata also has the commands `graph bar` and `graph hbar` which produce graphs that look like bar charts, but they are really intended to show the association between a continuous variable and a categorical variable, rather than showing the distribution of a single continuous variable.

If you are concerned with the normality of a distribution, the option `normal` can be useful. This will overlay the histogram with a normal distribution having the same mean and standard deviation as the observed data. Any deviation from normality with then become clearer. An example is shown in Figure 1.5

Box and Whisker Plot

An alternative graphical summary of a quantitative variable is a box and whisker plot. The central box shows the median and the upper and lower quartiles, and the "whiskers" show the range of "normal" values, as well as any individual "outlying" values. The definitions of "normal" and "outlying" can vary, but Stata treats any observations more than 1.5 interquartile ranges away from the nearest quartile as outlying. See `help graph box` for details. A box and whisker plot does not show the "shape" of the distribution as well as a histogram would, but it can be very useful for comparing distributions in different subgroups.

For example, consider the two box plots in Figure 1.6. The left panel shows an normal distribution, the symmetry of this distribution is made clear: the median line is in the centre of the box, and the whiskers are of very similar lenghths. However, the right panel, of a skewed distribution, is markedly asymmetrical: the median line is is the lower part of the box, the upper whisker is much longer than the lower whisker, and the outlying values are all large values, there are no small outliers.

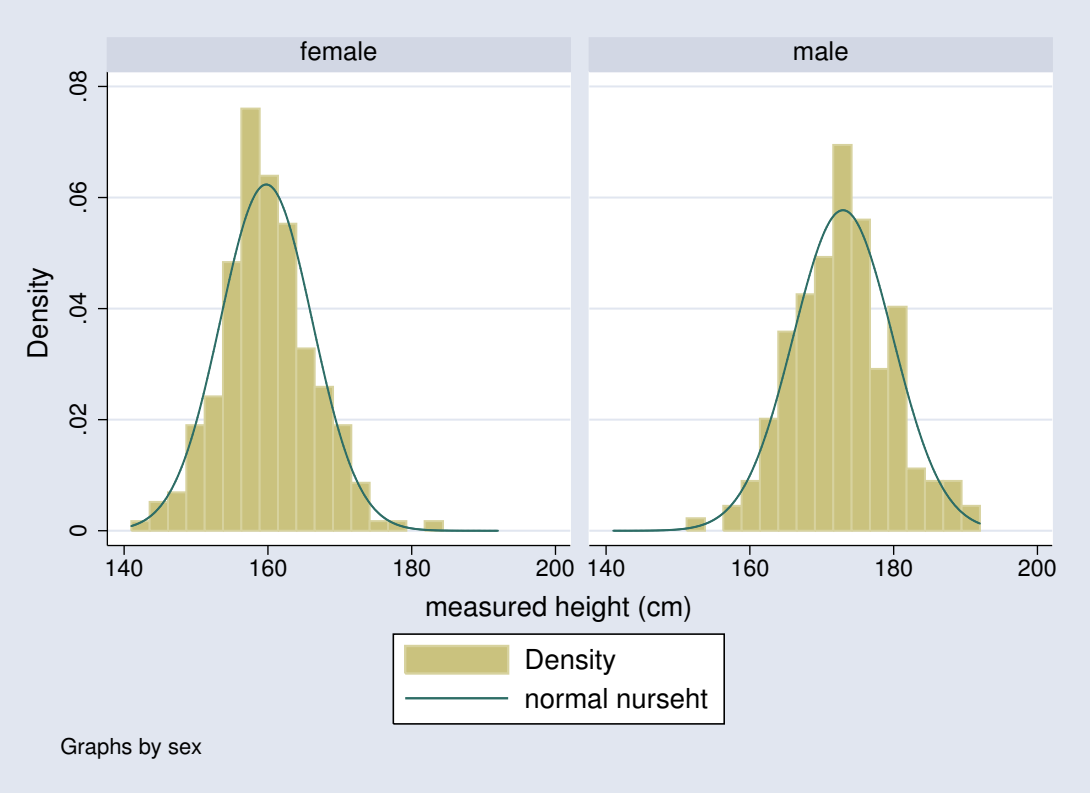


Figure 1.5: Histogram with superimposed normal distribution

The Stata command for producing a box and whisker plot is `graph box`.

1.4 Table 1

It is extremely common for the first table in a clinical or epidemiological research paper ("Table 1") to provide descriptive statistics for the main variables of interest in the study. When numerically summarising a quantitative variable, if the variable is normally distributed you can present the mean and standard deviation, otherwise present the median and the quartiles. It may be simpler to present the median and quartiles for all quantitative variables if some are normally distributed and others are not, but views differ on that. For qualitative variables, present the number and percentage in each category.

(a) Normal Distribution (b) Skewed Distribution

Figure 1.6: Boxplots of a normal and a skewed distribution

1.5 Summarising Data Practical

1.6 Hand Calculations

This section gives you the chance to do some calculations for yourself and see how the concepts we saw in the lecture work in real life. Once upon a time, these calculations would have been done by hand: I'm sure you could do them in your head, but getting stata to do them for you will be quicker. However, we are going to go through the steps that you would have to perform if you were calculating them by hand, so that you can see how it works. In practice, you would simply ask stata to churn the results out rather than calculating them this way.

To start Stata, click on Start Button ⇒ All Programs ⇒ Stata 14 ⇒ StataIC 14 (64 bit)

This should work for most of you, but depending on your faculty and whether you are staff or a student, the exact route may vary. Let me know if you have any difficulty.

Solutions for all practicals can be found at
`http://personalpages.manchester.ac.uk/staff/mark.lunt/stats_course.html`

Load in the PImax dataset with the commands

```
global datadir http://personalpages.manchester.ac.uk/staff/mark.lunt/stats
use $datadir/2_summarizing_data/data/pimax.dta
```

(Note that there are two separate commands above, and they need to be entered separately, each on its own line.)

This dataset contains a single variable, pimax, which is the maximal static inspiratory pressure, measured in cm H_2O, of 25 cystic fibrosis patients.

Sort the data into ascending order with the command

```
sort pimax
```

6.1 Now look at the data in the spreadsheet view using the command `browse` or the browse button, and look up the median value .....

6.2 From the same browse view, what are the 25th and 75th centiles (the values of the 7th and 18th observation respectively).

Lower quartile .....

Upper quartile .....

6.3 Now we will calculate the mean. First put the data back in order with the command

```
sort id
```

Then we can generate a running sum with the command

```
gen sum = sum(pimax)
```

If you look at the data with `browse`, you will see the variable `sum` contains a running sum, i.e. the sum of all observations from the top of the dataset to that observation inclusive. Hence, `sum` in the last observation contains the sum of `pimax` in all of the observations.
Finally, we have to divide by the number of observations. We can do this with two commands: first

```
gen n = _n
```

will put the number of the observation into a variable called `n`, then

```
gen mean = sum/n
```

will generate a running mean: the mean of all observations from the top of the dataset to the current observation. Again, you can browse the data.
What is the mean PImax in this sample ? .....

6.4 Finally, we are going to calculate the standard deviation. First, we need to create a variable containing the mean value for each observation:

```
drop mean
egen mean = mean(pimax)
```

If you `browse` the data, you will see that the variable `mean` now contains the same value for each observation, the overall mean. This is an important difference between `gen` and `egen`. (From now on, I will assume that you are `browse`ing the data each time you change it and will not explicitly mention it).
Next we calculate the difference between each observation and this mean (i.e. $x_i - \bar{x}$):

```
gen diff = pimax - mean
```

Square the difference to give $(x_i - \bar{x})^2$:

```
gen diff2 = diff * diff
```

and add them all up $(\Sigma\,(x_i - \bar{x})^2)$:

```
gen diff2_sum = sum(diff2)
```

Dividing by the observation number will give a "running variance", with the value of the variance in for the entire sample in the last observation $(\frac{\Sigma(x_i-\bar{x})^2}{n})$

```
gen variance = diff2_sum / n
```

Finally, take the square root of the variance to get the standard deviation $(\sqrt{\frac{\Sigma(x_i-\bar{x})^2}{n}})$:

```
generate sd = sqrt(variance)
```

Again, the value for the entire sample will be in the last observation.
What is the standard deviation of PImax in this sample ?[a] .....

1.7 Summarising Data in Stata

Read the stata file `htwt.dta` into stata by typing

```
global datadir http://personalpages.manchester.ac.uk/staff/mark.lunt/stats
use $datadir/2_summarizing_data/data/htwt.dta, clear
```

This file includes two BMI values: `bmi` which was based on measured data and `bmirep` which was based on reported data.

7.1 Examine the distribution of measured BMI scores by displaying a histogram of the data, using the command

```
histogram bmi
```

Is the data normally distributed, or does it show some skewness ?

.....

1.8 Summarise

8.1 Calculate summary measures of the measured BMI using the command

```
summarize bmi, det
```

8.2 Write down the mean BMI. .....

8.3 How does the mean compare to the median ?

.....

.....

.....

8.4 What are the lower and upper quartiles of the data ?

.....

.....

.....

1.9 Summarise by group.

9.1 Display the data separately for each sex, using the commands

```
sort sex
by sex:  summ bmi, det
```

9.2 The distributions for the two sexes can be neatly compared graphically using boxplots. The command to do this is `graph box bmi, by(sex)`. Write down a short description of what you see.

.....

.....

.....

.....

1.10 `tabstat`

This command can be used to produce tables of summary statistics: it is similar to `summarize`, but the output it produces is far more controllable. The basic syntax is `tabstat` *`varlist`*`, statistics(`*`statname [statname ...]`* The option `statistics` can be given one or more of the options in table 2.1.

For example,

```
tabstat bmi, statistics(mean sd)
```

would give the mean and SD of BMI. There is also a `by()` option, which enables you to obtain the statistics for different subgroups:

```
tabstat bmi, statistics(mean sd) by(sex)
```

will give the mean and SD of BMI for men and women separately.

10.1 Use `tabstat` to produce the mean and SD of height and weight, as measured by the nurse, for men and women separately (If you have forgotten the names of the variables to use for this, try typing `describe`.

.....

.....

1.11 `table`

An alternative to `tabstat` is `table`. This is more flexible in some ways and less flexible in others: in particular, it can only produce a maximum of 5 statistics per group. The syntax is

`table` *`groupvars`*`, contents(`*`contents_list`*`)`

where *`contents_list`* consists of pairs of statistic names and variable names. So, to get the mean and SD of BMI for men and women using `table`, you would type

statname	definition
mean	mean
count	count of nonmissing observations
n	same as count
sum	sum
max	maximum
min	minimum
range	range = max - min
sd	standard deviation
var	variance
cv	coefficient of variation (sd/mean)
semean	standard error of mean = sd/sqrt(n)
skewness	skewness
kurtosis	kurtosis
median	median (same as p50)
p1	1st percentile
p5	5th percentile
p10	10th percentile
p25	25th percentile
p50	50th percentile (same as median)
p75	75th percentile
p90	90th percentile
p95	95th percentile
p99	99th percentile
iqr	interquartile range = p75 - p25
q	equivalent to specifying "p25 p50 p75"

Table 1.1: Statistics available in `tabstat`

```
table sex, contents(mean bmi sd bmi)
```

11.1 Use `table` to produce the mean and SD of height and weight, as measured by the nurse, for men and women separately

.....

.....

1.12 Further exercises

12.1 What is the average age of the subjects ? .....

12.2 Draw a histogram of the ages, using the command `histogram age`. Do the ages follow a normal distribution ?

.....

.....

12.3 How old are the youngest and oldest males and females in the study ?

	Youngest	Oldest
Males	.....	.....
Females	.....	.....

12.4 What is the mean of the self-reported BMI. Is this greater or less than the mean of the BMI as measured by the nurse ?

.....

.....

12.5 Create a variable for the difference between measured BMI and self-reported BMI:

```
gen bmidiff = bmi - bmirep
```

Write down its mean value, standard deviation and the number of subjects for whom both BMI measures are available.

Mean .....

Standard Deviation .....

Both available on ..... subjects

12.6 Produce histograms of height in men and women, using the commands

```
histogram nurseht, by(sex)
```

and

```
histogram nursewt, by(sex)
```

Add a superimposed normal distribution to the graphs with the `normal` option to the `histogram` command.

12.7 Start a Word document. Select the Graph window in stata and use Ctrl+C to copy the graph. Switch to the Word document and use Ctrl+V to paste the graph.

2 Summarizing Data Practical

2.1 Hand Calculations

This section gives you the chance to do some calculations for yourself and see how the concepts we saw in the lecture work in real life. Once upon a time, these calculations would have been done by hand: I'm sure you could do them in your head, but getting stata to do them for you will be quicker. However, we are going to go through the steps that you would have to perform if you were calculating them by hand, so that you can see how it works. In practice, you would simply ask stata to churn the results out rather than calculating them this way.

To start Stata, click on Start Button ⇒ All Programs ⇒ Stata 14 ⇒ StataIC 14 (64 bit)

This should work for most of you, but depending on your faculty and whether you are staff or a student, the exact route may vary. Let me know if you have any difficulty.

Solutions for all practicals can be found at
`http://personalpages.manchester.ac.uk/staff/mark.lunt/stats_course.html`

Load in the PImax dataset with the commands

```
global datadir http://personalpages.manchester.ac.uk/staff/mark.lunt/stats
use $datadir/2_summarizing_data/data/pimax.dta
```

(Note that there are two separate commands above, and they need to be entered separately, each on its own line.)

This dataset contains a single variable, pimax, which is the maximal static inspiratory pressure, measured in cm H_2O, of 25 cystic fibrosis patients.

Sort the data into ascending order with the command

```
sort pimax
```

1.1 Now look at the data in the spreadsheet view using the command `browse` or the browse button, and look up the median value .....

1.2 From the same browse view, what are the 25th and 75th centiles (the values of the 7th and 18th observation respectively).

Lower quartile .....

Upper quartile .....

1.3 Now we will calculate the mean. First put the data back in order with the command

```
sort id
```

Then we can generate a running sum with the command

```
gen sum = sum(pimax)
```

If you look at the data with `browse`, you will see the variable `sum` contains a running sum, i.e. the sum of all observations from the top of the dataset to that observation inclusive. Hence, `sum` in the last observation contains the sum of `pimax` in all of the observations.
Finally, we have to divide by the number of observations. We can do this with two commands: first

```
gen n = _n
```

will put the number of the observation into a variable called `n`, then

```
gen mean = sum/n
```

will generate a running mean: the mean of all observations from the top of the dataset to the current observation. Again, you can browse the data.
What is the mean PImax in this sample ? .....

1.4 Finally, we are going to calculate the standard deviation. First, we need to create a variable containing the mean value for each observation:

```
drop mean
egen mean = mean(pimax)
```

If you `browse` the data, you will see that the variable `mean` now contains the same value for each observation, the overall mean. This is an important difference between `gen` and `egen`. (From now on, I will assume that you are `browse`ing the data each time you change it and will not explicitly mention it).
Next we calculate the difference between each observation and this mean (i.e. $x_i - \bar{x}$):

```
gen diff = pimax - mean
```

Square the difference to give $(x_i - \bar{x})^2$:

```
gen diff2 = diff * diff
```

and add them all up $(\Sigma\,(x_i - \bar{x})^2)$:

```
gen diff2_sum = sum(diff2)
```

Dividing by the observation number will give a "running variance", with the value of the variance in for the entire sample in the last observation $(\frac{\Sigma(x_i-\bar{x})^2}{n})$

```
gen variance = diff2_sum / n
```

Finally, take the square root of the variance to get the standard deviation $(\sqrt{\frac{\Sigma(x_i-\bar{x})^2}{n}})$:

```
generate sd = sqrt(variance)
```

Again, the value for the entire sample will be in the last observation.
What is the standard deviation of PImax in this sample ?[a] .....

2.2 Summarising Data in Stata

Read the stata file `htwt.dta` into stata by typing

```
global datadir http://personalpages.manchester.ac.uk/staff/mark.lunt/stats
use $datadir/2_summarizing_data/data/htwt.dta, clear
```

This file includes two BMI values: `bmi` which was based on measured data and `bmirep` which was based on reported data.

2.1 Examine the distribution of measured BMI scores by displaying a histogram of the data, using the command

```
histogram bmi
```

Is the data normally distributed, or does it show some skewness ?

.....

2.3 Summarise

3.1 Calculate summary measures of the measured BMI using the command

```
summarize bmi, det
```

3.2 Write down the mean BMI. .....

3.3 How does the mean compare to the median ?

.....

.....

.....

3.4 What are the lower and upper quartiles of the data ?

.....

.....

.....

2.4 Summarise by group.

4.1 Display the data separately for each sex, using the commands

```
sort sex
by sex:  summ bmi, det
```

4.2 The distributions for the two sexes can be neatly compared graphically using boxplots. The command to do this is `graph box bmi, by(sex)`. Write down a short description of what you see.

.....

.....

.....

.....

2.5 `tabstat`

This command can be used to produce tables of summary statistics: it is similar to `summarize`, but the output it produces is far more controllable. The basic syntax is `tabstat` *`varlist`*`, statistics(`*`statname [statname ...]`* The option `statistics` can be given one or more of the options in table 2.1.

For example,

```
tabstat bmi, statistics(mean sd)
```

would give the mean and SD of BMI. There is also a `by()` option, which enables you to obtain the statistics for different subgroups:

```
tabstat bmi, statistics(mean sd) by(sex)
```

will give the mean and SD of BMI for men and women separately.

5.1 Use `tabstat` to produce the mean and SD of height and weight, as measured by the nurse, for men and women separately (If you have forgotten the names of the variables to use for this, try typing `describe`.

.....

.....

2.6 `table`

An alternative to `tabstat` is `table`. This is more flexible in some ways and less flexible in others: in particular, it can only produce a maximum of 5 statistics per group. The syntax is

`table` *`groupvars`*`, contents(`*`contents_list`*`)`

where *`contents_list`* consists of pairs of statistic names and variable names. So, to get the mean and SD of BMI for men and women using `table`, you would type

statname	definition
mean	mean
count	count of nonmissing observations
n	same as count
sum	sum
max	maximum
min	minimum
range	range = max - min
sd	standard deviation
var	variance
cv	coefficient of variation (sd/mean)
semean	standard error of mean = sd/sqrt(n)
skewness	skewness
kurtosis	kurtosis
median	median (same as p50)
p1	1st percentile
p5	5th percentile
p10	10th percentile
p25	25th percentile
p50	50th percentile (same as median)
p75	75th percentile
p90	90th percentile
p95	95th percentile
p99	99th percentile
iqr	interquartile range = p75 - p25
q	equivalent to specifying "p25 p50 p75"

Table 2.1: Statistics available in `tabstat`

```
table sex, contents(mean bmi sd bmi)
```

6.1 Use `table` to produce the mean and SD of height and weight, as measured by the nurse, for men and women separately

.....

.....

2.7 Further exercises

7.1 What is the average age of the subjects ? .....

7.2 Draw a histogram of the ages, using the command `histogram age`. Do the ages follow a normal distribution ?

.....

.....

7.3 How old are the youngest and oldest males and females in the study ?

	Youngest	Oldest
Males	.....	.....
Females	.....	.....

7.4 What is the mean of the self-reported BMI. Is this greater or less than the mean of the BMI as measured by the nurse ?

.....

.....

7.5 Create a variable for the difference between measured BMI and self-reported BMI:

```
gen bmidiff = bmi - bmirep
```

Write down its mean value, standard deviation and the number of subjects for whom both BMI measures are available.

Mean .....

Standard Deviation .....

Both available on ..... subjects

7.6 Produce histograms of height in men and women, using the commands

```
histogram nurseht, by(sex)
```

and

```
histogram nursewt, by(sex)
```

Add a superimposed normal distribution to the graphs with the `normal` option to the `histogram` command.

7.7 Start a Word document. Select the Graph window in stata and use Ctrl+C to copy the graph. Switch to the Word document and use Ctrl+V to paste the graph.