Contents

1 Sampling and Confidence Intervals

1.1 Types of Sample

Often, it is impractical survey every subject in an entire population. Instead, we survey a much smaller sample, and see what we can infer about the population from that sample. However, for the inference to be reliable, the sample has to be carefully selected to ensure it is representative of the population.

1.1.1 Simple Random Sample

The simplest type of sample to draw inference from is the simple random sample. In this type of sampling, every individual has exactly the same probability of being selected for our sample. So if we have a population of 100,000, and want to sample 1,000 people from it, every person in the population should have a 1 in 100 chanve of being selected.

There are two big issues with simple random sampling. One is the fact that you need to have a list of everyone in your population. This may be hard to obtain. For example, GP lists are often used as a sampling frame, but not everyone is registered with a GP, so the sampling frame is incomplete. The same is true for electoral registers. This means that the population you are investigating may not be the population you wish to investigate: you are not sampling from "People living in this area", but "People registered with a G.P. in this area" or "People registered to vote in this area". Whether this is a problem or not depends on how different the population we are actually studying is from the population we intended to study.

The second problem is refusals. In a simple random sample, everyone has the same probability of being in the sample. However, some people have already assigned themselves a 0% probability of being in the sample by deciding they don't want to participate. Again, you can think of this as changing the population you are studying: "People from the population we wish to study who are happy to take part in our study". This may or may not be an issue, depending on the extend to which those who don't take part differ from those who do.

1.1.2 Stratified Sample

A stratified sample can be thought of as a number of simple random samples. The population to be studied is divided into strata, and a simple random sampling is performed within each stratum. The probability of being in the sample is the same for every individual within a given stratum, but can differ between strata.

A stratified sample can be useful if you wish to estimate something that varies widely between strata. For example, if you are interested in the prevalence of a condition that is very common in older subjects and rare in younger ones, you may want to recruit more younger subjects in order to get a equally precise estimate of the prevalence in the young as you would get in the older subjects.

1.1.3 Cluster Sample

Another commonly used type of sampling is cluster sampling: in which groups of subjects (clusters) are sampled rather than individuals. There are a number of situations where this can be useful.

1. There is no sampling frame for individuals, but there is for clusters. For example, you may have a list of houses, but not a list of the residents of those houses. It would then be impossible to sample individuals, but it is possible to sample houses.

2. It is cheaper and easier to recruit a number of people at the same time. Various addons to the ten-yearly census uses this method: they phone a landline and ask about everybody living in that house.

3. It may be that an intervention can not be allocated to individuals, only to clusters of individuals. For example, to assess the impact of posters in G.P.'s waiting rooms, you cannot randomly allocate individuals to look at or not look at the poster, but you can randomly allocate the posters to some waiting rooms and not to others

Cluster sampling requires special methods for analysing the data collected, which we will not mention in this course.

1.1.4 Other Types of Sample

There are other types of sample that you may hear mentioned, but I do not recommend using them. One is the quota sample, beloved of market researchers, where you have a fixed number of subjects to recruit to each subgroup. This can be thought of as a stratified sample, with the groups to which you are recruiting as the strata. However, you do not know what the sampling probability is for any stratum, so you don't know what proportion of the population each stratum represents. With a stratified sample, if you want to calculate a population statistic, you can reweight each stratum according to the size it should have in the population, rather than the size it has in the sample, and hence make the sample representative of the population. With a quota sample you can't do that.

In a systematic sample, you randomly select a starting point in your sampling frame, then take every n^{th} subject until you have your desired sample size. This can work, unless there is clustering or periodicity in your sampling frame. For example, if the sampling frame is in alphabetical order by surname, then you are unlikely ever to select two subjects from the same family, whilst in a simple random sample that would happen. That means that characteristics that are shared in a family, such as diet, will vary more in a systematic sample than in a simple random sample. You could fix this problem easily if you have an electronic sampling frame, by assigning each record a random number and sorting on that number, to put the sampling frame into a random order. Then a systematic sample would be a random sample.

One last type of sample to look out for is the convenience sample. This is where you simply recruit people who are easy to find. For example, when testing a questionnaire, you may ask your colleagues to complete it for you, and see how long it takes. However, your colleagues are likely to have more education than the population average, and so complete the questionnaire quicker. Provided that you are aware of the limited population you are recruiting from, this is not a problem, but it is not possible to make inference about the general population from a convenience sample.

1.2 From Sample to Population

The purpose of collecting sample data is to see what it can tell us about the population. We can be interested in point estimates ("What is our best guess at the population value ?"), interval estimates ("Can we be fairly certain the population value lies within a given range ?") and hypothesis tests ("Can we be fairly certain the population value is not one particular value ?"). In this chapter, we are only concerned with the first two questions, the next chapter covers the third one.

In order to help keep track of whether we are talking about population statistics or sample statistics, we generally use Greek letters for population values and Roman letters for sample values. The most commonly used notation is given in Table 1.1

Parameter	Population Value	Sample Value
Mean	μ	m
Standard deviation	σ	s
Proportion	π	p

Table 1.1: Notation for Population and Sample Parameters

1.2.1 Estimating a Mean

Suppose that we are interested in the mean value in the population (μ). The mean we calculate from the sample should be close to the population value, but will not be exactly equal to it. If we were to take a number of samples, we would expect the means to be different in each sample. However, we would expect the sample means to be close to the population mean, and to vary less in larger samples. This is illustrated in the example below.

Example of estimating a mean

For this example, the population consists of 10,000 individuals. The variable we are interested in, x, takes integer values 0, 1, ..., 9, and 1,000 individuals have each of the 10 possible values. The population distribution of x is shown in Figure 1.1.

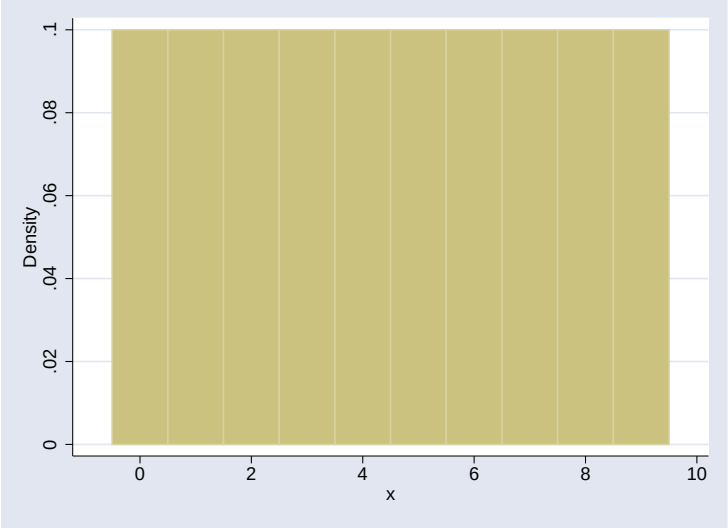
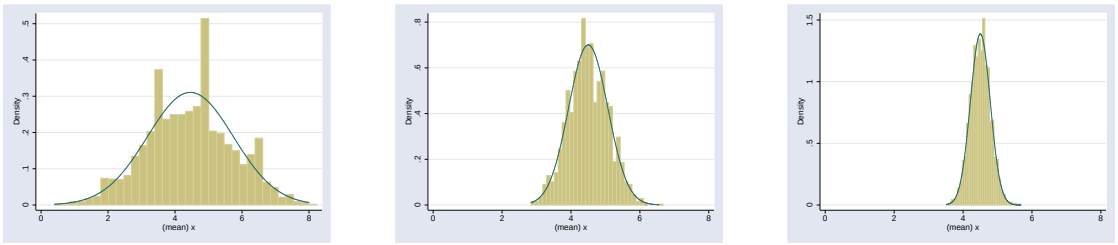


Figure 1.1: Population Distribution for Estimating Mean Example

From this population, samples of size 5, 25 and 100 were repeatedly taken, and the mean of each sample was calculated. Figure 1.2 shows the distribution of the sample means for each sample size, each histogram based on 2,000 samples.

It is clear from 1.2 that the distibution of the sample means is centred on the same value, 4.5, irrepective of the sample size. It is also clear that the distibution varies less as the sample size increases. Finally, the distribution of the sample mean appears to be normal, particularly for



(a) Sample size 5 (b) Sample size 25 (c) Sample size 100

Figure 1.2: Distributions of sample means of different sizes

the larger sample sizes, despite the fact that the distribution of the original variable (in Figure 1.1) is far from normal.

The Sampling Distribution of a mean

The distribution of the values taken by the sample mean ($\bar{x}$) as samples (of a given size) are repeatedly taken from a fixed population is called the *sampling distribution* of $\bar{x}$. All statistical inference (working out what our sample tells us about the population from which it was taken) depends on this distribution. There are three important properties of the sampling distribution of the mean:

$\boldsymbol{E(\bar{x}) = \mu}$ i.e. on average, the sample mean is the same as the population mean.

Standard Deviation of $\bar{x} = \frac{\sigma}{\sqrt{n}}$ i.e the uncertainty in $\bar{x}$ increases with σ, decreases with n. The standard deviation of the sampling distribution of the mean is also called the **Standard Error**

$\bar{x}$ is normally distributed This is true whether or not x is normally distributed, provided n is sufficiently large. Thanks to the *Central Limit Theorem*. This is clearly shown in Figure 1.2: the sampling distributions of the means are all normal, despite the fact that the distribution of x, shown in Figure 1.1, is clearly *not* normal.

The Standard Error

The standard error is the most important concept in statistics[a]. The standard error is the standard deviation of the sampling distribution. We will shortly see how the standard error can be used to calculate confidence intervals for population parameters, and in the next chapter how it can be used to perform hypothesis tests.

1.2.2 Estimating the Variance σ

The standard error of $\bar{x}$ is defined in terms of the population variance, σ^2. However, we do not know what this is, and need to estimate if from our sample.

In a population of size N, the variance of x is given by

$$\sigma^2 = \frac{\Sigma(x_i - \mu)^2}{N} \tag{1.1}$$

[a] At least in *frequentist* statistics, which is the only type of statistics this course is concerned with

with the summation being over every observation in the population. However, if we only have a sample from this population, we do not know the population mean μ. We can estimate it as the sample mean, $\bar{x}$, but this value will depend on the actual values we observe in the sample. Since $\bar{x}$ is the number that minimises $\Sigma(x_i - \bar{x})^2$, this sum will be smaller than the value we really want, $\Sigma(x_i - \mu)^2$. Hence, $s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n}$ will be smaller than σ^2. However, dividing by $n - 1$ rather than n is sufficient to correct this underestimate, and $s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n-1}$ is an unbiased estimate of the population variance σ^2, and can be used to calculate the standard error of the mean without problems.

1.2.3 Estimating a Proportion

Estimating a proportion works in exactly the same way as estimating a mean. In fact, estimating a proportion *is* estimating a mean: it is estimating the mean of a variable which takes the value 0 for all observations in which the characteristic of interest is absent, and 1 for all observations in which it is present. Adding all these values gives the number of observations with the characteristic, dividing by the sample size gives the proportion. Furthermore, we have seen that the sampling distribution of $\bar{x}$ is normal (in sufficiently large samples), even if x is not.

1.2.4 Confidence Intervals

We have seen that the sampling distribution of a mean is normal, with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$. The sample mean that we have can be thought of as a single observation drawn at random from this distribution. The $2.5th$ centile of this distribution is $\mu - 1.96\frac{\sigma}{\sqrt{n}}$, and the 97.5^{th} centile is $\mu + 1.96\frac{\sigma}{\sqrt{n}}$. Hence, there is a 95% chance that our sample value will lie in between these two values.

So, although we don't know the *exact* value of μ, we know that there is a 95% chance that it lies within $1.96 \times \frac{s}{\sqrt{n}}$, and we do know both s and n. Hence, we can calculate an interval, $\bar{x} - 1.96\frac{s}{\sqrt{n}}$ to $\bar{x} + 1.96\frac{s}{\sqrt{n}}$, in which we are 95% certain that the population will lie. This is known as a 95% confidence interval.

In theory, we can calculate a confidence interval with any chosen probability that the population mean lies within it. The less certain we need to be, the narrower the confidence interval can be. In practice, the 95% level is always used.

Confidence Interval Example

As an example of calculating a 95% confidence interval, suppose we measured serum albumin in a sample fo 216 patients with primary biliary cirrhosis. The mean value in the sample was 34.46 g/l, and the values had a standard deviation of 5.84 g/l.

To calculate a 95% confidence interval for the mean serum albumin level in primary biliary cirrhosis patients, we first have to calculate the standard error. This is

$$\begin{aligned} \text{Standard Error of } \bar{x} &= \frac{5.84}{\sqrt{216}} \\ &= 0.397 \\ \Rightarrow \quad 95\% \text{ Confidence Interval} &= 34.46 \pm 1.96 \times 0.397 \\ &= (33.68, 35.24) \end{aligned}$$

So we can be 95% certain that the population value is between 33.7 and 35.2 g/l.

Confidence Interval For a Proportion

If the sample size is large enough, the sampling distribution of a proportion p will be approximately normal, and its standard error in a sample of size n will be $\sqrt{\frac{p(1-p)}{n}}$. As a rule of thumb, if both $p \times n$ and $(1-p) \times n$ are bigger than 5, the sample size is large enough to assume normality.

Confidence Intervals in Stata

cmd:ci Stata has a command `ci` that can be used to calculate confidence intervals around a mean or proportion. If you want to calculate a confidence interval around a proportion in a small sample, where the normal approximation may not be very accurate, there is a `binomial` option which will use the exact binomial distribution to calculate the confidence interval.

1.2.5 Sample Size Calculations

In general, the aim of any study is to estimate a statistic: a mean, proportion, relative risk, hazard ratio or some other characteristic of a population. The larger our study is, the more precisely we can estimate this statistic. We can choose the size of study we need to provide adequate precision if we know how the sampling distribution depends on the sample size, and we have a definition of adequate, i.e. what is the widest confidence interval we will accept.

Given that the confidence interval is $\bar{x} \pm \frac{1.96\sigma}{\sqrt{n}}$, its width depends only on n and σ. If we have an estimate of σ, then we can choose n to give us a confidence interval of any width.

Suppose that we want the width of our confidence interval to be $2W$, so that the interval itself is $\bar{x} \pm W$ Then

$$\begin{aligned} W &= 1.96 \times \text{Standard Error} \\ &= 1.96 \times \frac{\sigma}{\sqrt{n}} \\ \Rightarrow \quad W^2 &= \frac{1.96^2\sigma^2}{n} \\ \Rightarrow \quad n &= \left(\frac{1.96\sigma}{W}\right)^2 \end{aligned}$$

So collecting a sample of at least $\left(\frac{1.96\sigma}{W}\right)^2$ subjects will provide a 95% confidence interval that is no wider than $2W$.

Sample Size Calculation Example

Suppose that using the primary biliary cirrhosis data in Section 1.2.4, we want to be know the population mean serum albumin level to within 0.5 g/l. How many patients would we need to study (assuming a standard deviation of 5.84 g/l) ?

$$\begin{aligned} W &= 0.5 \\ \sigma &= 5.84 \\ \Rightarrow \quad n &= \left(\frac{1.96\sigma}{W}\right)^2 \\ &= \left(\frac{1.96 \times 5.84}{0.5}\right)^2 \end{aligned}$$

$$\approx 524$$

So we would need a sample size of at least 524.

2 Sampling and Confidence Intervals Practical

2.0.1 Generating Random Samples

In this part of the practical, you are going to repeatedly generate random samples of varying size from a population with known mean and standard deviation. You can then see for yourselves how changing the sample size affects the variability of the sample mean. If you want to store your results in an Excel spreadsheet, double-click here to open a suitable one.

1. Ensure there is no data in stata's memory by entering the command `clear`
2. Set the sample size to 5 with the command `set obs 5`
3. Generate a variable `x` with a mean of 0 and a standard deviation of 1, using the command `generate x = invnorm(uniform())`
4. Obtain the mean of `x` in this sample from the command `summarize x`
5. Record the mean for this sample.
6. Repeat steps 1-5 10 times until the first column of the spreadsheet is full.
7. Now repeat the procedure a further ten times, but using `set obs 25` in step 2, to complete column 2 of the spreadsheet.
8. Complete column 3 of Table **??** using the command `set obs 100` in step 2.
9. Now calculate the mean and standard deviation of the values in each column.
   If you have used the Excel spreadsheet, it will do it for you. Otherwise, the easiest way to do this is to use the commands

   ```
   clear
   edit
   ```

   to get a spreadsheet view of an empty stata dataset, and type the values in as three columns. If you have stored them in a spreadsheet, you could cut and paste them, or use the command `import excel`: use the `help` command to find out how. Stata will call the three variables `var1`, `var2` and `var3` by default (unless you cut and paste the variable names from the spreadsheet), but you can rename them by double clicking on the name, and typing a new name in the dialog box that appears. When you have entered the data, click on the cross in the right hand top corner to close the spreadsheet view. (Once upon a time, stata would not carry out commands when a spreadsheet view was open. Current statas will, but the sheet will hide the results window).

   Now you can use the command

   ```
   summarize var1 var2 var3
   ```

   to get the mean and standard deviation of these variables.

2.0.2 Means

If the standard deviation of the original distribution is σ, then the standard error of the sample means is σ / $\sqrt{n}$, where n is the sample size.

0.1 If the standard deviation of measured heights is 9.31 cms, what will be the standard error of the mean in:

i a sample of size 49 ? .....

ii a sample of size 100 ? .....

0.2 Imagine we only had data on a sample of size 100, where the sample mean was 166.2cm and the sample standard deviation was 10.1cm.

i Calculate the standard error for this sample mean (using the sample standard deviation as an estimate of the population standard deviation).

.....

ii Calculate the interval ranging 1.96 standard errors either side of the sample mean.

.....

0.3 Imagine we only had data on a sample size of 36 where the sample mean height was 163.5 cm and the standard deviation was 10.5cm.

i Calculate the 95% confidence interval for the sample mean.

.....

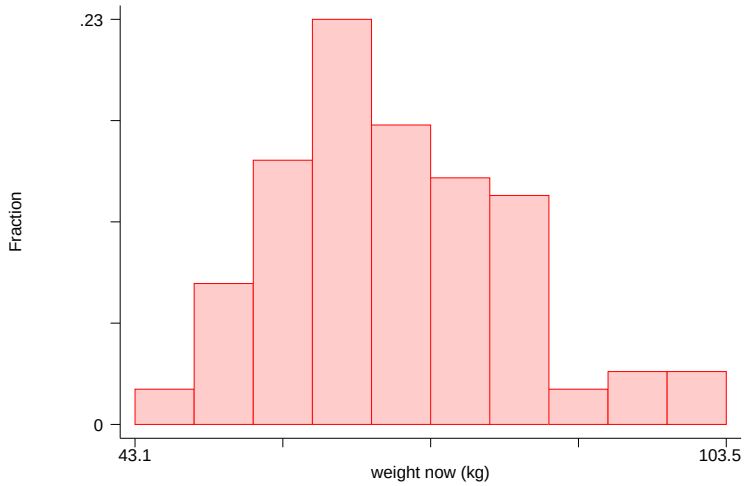


Figure 2.1: Weights in a random sample of 100 women

0.4 Figure 2.1 is a histogram of measured weight in a sample of 100 individuals.

i Would it be better to use the mean and standard deviation or the median and interquartile range to summarize this data ?

.....

ii If the mean of the data is 69.69kg with a standard deviation of 12.76kg, calculate a 95% confidence interval for the mean.

.....

2.0.3 Proportions

Again using our height and weight dataset of 412 individuals, 234 (56.8%) are women and 178 (43.2%) are men.

If we take a number of smaller samples from this population, the proportion of women will vary, although they will tend to be scattered around 57%. Figure 2.2 represents 50 samples, each of size 40.

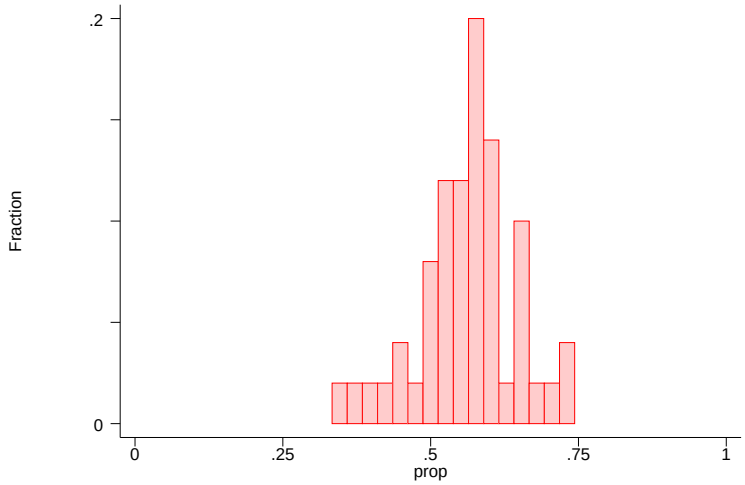


Figure 2.2: Proportion of Women in 50 samples of size 40

0.5 What would you expect to happen if the sample sizes were bigger, say n=100 ?

.....

.....

.....

0.6 In a sample of 40 individuals from a larger population, 25 are women. Calculate a 95% confidence interval for the proportion of women in the population.

.....

.....

Note: When sample sizes are small the use of standard errors and the normal distribution does not work well for proportions. This is only really a problem if p (or (1-p)) is less than 5/n (i.e. there are less than 5 subjects in one of the groups).

0.7 From a random sample of 80 women who attend a general practice, 18 report a previous history of asthma.

i Estimate the proportion of women in this population with a previous history of asthma, along with a 95% confidence interval for this proportion.

.....

ii Is the use of the normal distribution valid in this instance ?

.....

.....

0.8 In a random sample of 150 Manchester adults it was found that 58 received or needed to receive treatment for defective vision. Estimate the proportion of adults in Manchester who receive or need to receive treatment for defective vision, a 95% confidence interval for this proportion.

i Proportion

.....

ii 95% Confidence interval

.....

2.0.4 Confidence Intervals in Stata

Load the blood pressure data in its wide form into stata with the command

```
sysuse bpwide
```

This is fictional data concerning blood pressure before and after a particular intervention.

0.9 Use the command

```
histogram bp_before
```

to see if this variable is normally distributed. What do you think ?

.....

0.10 Create a new variable to measure the change in blood pressure and find its mean value with the commands

```
generate bp_diff = bp_after - bp_before
summarize bp_diff
```

What is the mean change in blood pressure ? .....

0.11 Create a confidence interval for the change in blood pressure with the command

```
ci bp_diff
```

Does the intervention reduce blood pressure in general ?

.....

0.12 Look at the histogram of changes in blood pressure using the command

```
histogram bp_diff
```

Does this confirm your answer to the previous question ?

.....

0.13 Create a new variable to measure whether blood pressure went up or down in a given subject using the command

```
generate down = bp_after < bp_before
```

Use the `tabulate` command to see how many subjects, and what proportion, showed a decrease in blood pressure.

.....

0.14 Create a confidence interval for the proportion of subjects showing a decrease in blood pressure with the command

```
ci down, binomial
```

Does this confirm the effect of the intervention on blood pressure ?

.....