# Solution for Session 3
# Sampling & Confidence Intervals

### *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.

### *0.2 Means*

If the standard deviation of the original distribution is $\sigma$, then the standard error of the sample means is $\sigma$ / $\sqrt{n}$, 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 ? $9.31/\sqrt{49} = 1.33$

ii a sample of size 100 ? $9.31/\sqrt{100} = 0.931$

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).
$Standard\ Error = 10.1/\sqrt{100} = 1.01cm$

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

$$\begin{aligned} 95\%\ Confidence\ Interval &= Mean \pm 1.96 \times Standard\ Error \\ &= 166.2cm \pm 1.96 \times 1.01cm \\ &= 164.2cm, 168.2cm \end{aligned}$$

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.

$$\begin{aligned} Standard\ Error &= 10.5cm/\sqrt{36} \\ &= 1.75 \\ \Rightarrow Confidence\ Interval &= 163.5cm \pm 1.96 \times 1.75cm \\ &= (160.1cm, 166.9cm) \end{aligned}$$

0.4 Figure 1.1 is a histogram of measured weight in a sample of 100 individuals.

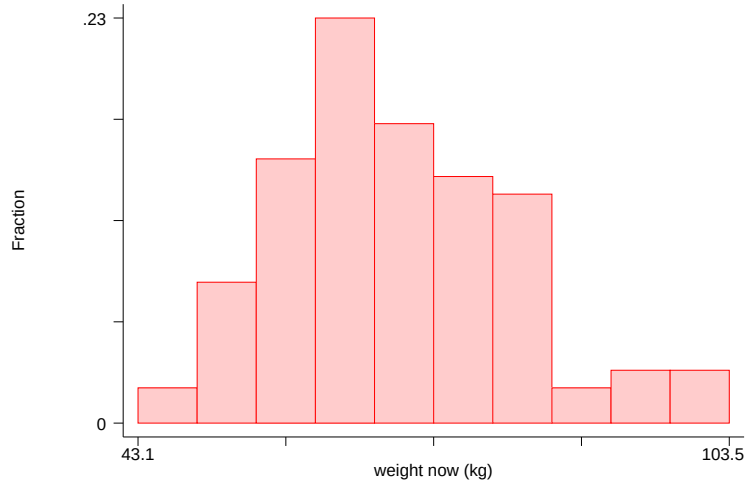


Figure 1.1: Weights in a random sample of 100 women

i Would it be better to use the mean and standard deviation or the median and interquartile range to summarize this data ?
*The data appears to be positively skewed, so the median and inter-quartile range would provide a better summary than the mean and SD.*

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.

$$\begin{aligned} \textit{Standard Error} &= 12.76kg/\sqrt{100} \\ &= 1.276kg \\ \Rightarrow \textit{Confidence Interval} &= 69.69kg \pm 1.96 \times 1.276 \\ &= (67.2kg, 72.2kg) \end{aligned}$$

### *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 1.2 represents 50 samples, each of size 40.

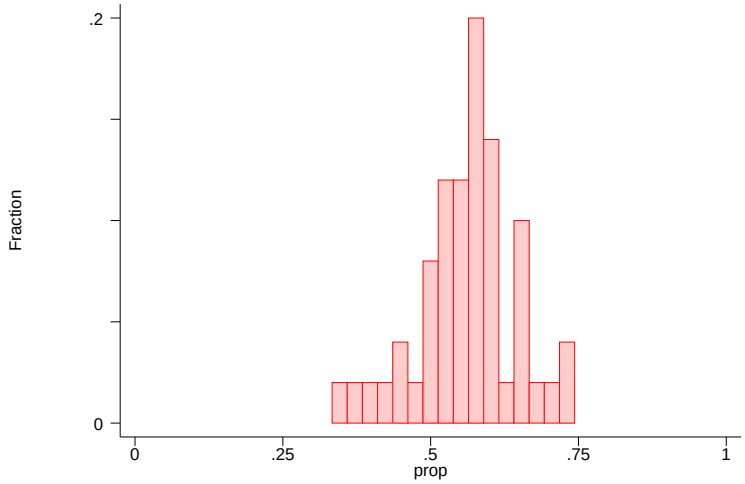


Figure 1.2: Proportion of Women in 50 samples of size 40

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

. . . . . .

. . . . . .

*Larger samples would lead to less variation in the sample means*

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.

. . . . . .

$$
\begin{aligned}
\textit{Proportion } p &= \frac{25}{40} \\
&= 0.625 \\
\Rightarrow \textit{Standard Error} &= \sqrt{\frac{p(1-p)}{n}} \\
&= \sqrt{\frac{0.625 \times 0.375}{40}} \\
&= 0.07655 \\
\Rightarrow \textit{Confidence Interval} &= 0.625 \pm 1.96 \times 0.07655 \\
&= 0.475, 0.775
\end{aligned}
$$

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.

$$\begin{aligned} \textit{Proportion } p &= \frac{18}{80} \\ &= 0.225 \\ \Rightarrow \textit{Standard Error} &= \sqrt{\frac{0.225 \times 0.775}{80}} \\ &= 0.04669 \\ \Rightarrow \textit{Confidence Interval} &= 0.225 \pm 1.96 \times 0.04669 \\ &= (0.133, 0.317) \end{aligned}$$

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

.....

*Yes, since there are more than 5 subjects in each of the two groups (with and without asthma).*

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

$\textit{Proportion } p = \frac{58}{150} = 0.387$

ii 95% Confidence interval

$$\begin{aligned} \textit{Proportion } p &= \frac{58}{150} \\ &= 0.387 \\ \Rightarrow \textit{Standard Error} &= \sqrt{\frac{0.387 \times 0.613}{150}} \\ &= 0.03977 \\ \Rightarrow \textit{95\% Confidence Interval} &= 0.387 \pm 1.96 \times 0.03977 \\ &= (0.309, 0.465) \end{aligned}$$

### *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 ?
*No: the data is positively skewed*

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 ? *-5.09*

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 ?
*Yes, the confidence interval for the mean change in the population is (-8.1, -2.1), suggesting that there has been a reduction in blood pressure.*

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 ?
*The bulk of the histogram is to the left of 0, suggesting there has been an overall reduction in blood pressure*

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.
*73 subjects, 60.8%*

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 ?
*The confidence interval is (51.5%, 69.6%), suggesting that in the population more than half and possibly as many as two thirds of subjects will have a reduction in blood pressure.*