Computer practical 11

Statistics 1, Term 2

Aim

- To learn how to implement statistical methods in R
  - To generate confidence intervals for a population mean
  - To validate assumptions underlying their use through inspecting the normality of the data
  - To employ transformations to make the data look more normal.

Statistical techniques

Suppose *data* is a vector of data.

Normal Quantile plots

Type:

```
> qqnorm(data)
```

Overlaying the normal shape on a histogram

This is a bit more difficult. The following are general instructions. Read them and then try and use them later on. Type:

```
> hist(data, freq=FALSE)
> curve(dnorm(x, mean=mean(data), sd=sd(data)), add=T)
```

For more information, read the R help for the individual functions (hist, dnorm, curve).

Obtaining two-sided confidence intervals

If you are not yet familiar with qnorm and qt, read their documentation.

Suppose alpha is the level of significance.

- If σ is known, and n is large or we're sampling from normal, we must calculate

$$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}.$$

  Suppose sigma contains the value of σ. Type:

```
> mean(data)+qnorm(1-alpha/2)*sigma/sqrt(length(data))
> mean(data)-qnorm(1-alpha/2)*sigma/sqrt(length(data))
```

to get the confidence interval. The p-value of

$$H_0 : \mu = m$$

$$H_a : \mu \neq m$$

is calculated via

```
> 2*(1-pnorm(abs(mean(data)-m)/(sigma/sqrt(length(data)))))
```

- If σ is unknown, but n is large, then simply replace σ by the sample standard deviation s (sd(data)).
- If σ is unknown, n is small, and we are sampling from a normal distribution (use a normal quantile plot to check this!), then in addition to replacing σ by sd(data), also use the t distribution with n-1 degrees of freedom, instead of the normal distribution:

  ```
  > mean(data)+qt(1-alpha/2, length(data)-1)*sd(data)/sqrt(length(data))
  > mean(data)-qt(1-alpha/2, length(data)-1)*sd(data)/sqrt(length(data))
  > 2*(1-pt(abs(mean(data)-m)/(sd(data)/sqrt(length(data))), + length(data)-1))
  ```

  Another way of obtaining a 95% confidence interval for a population mean based on a vector data is as follows:

  ```
  > confint(lm(data~1))
  ```

  What's actually happening is we are exploiting the regression facilities to fit a model which only has an intercept, i.e. the population mean. We can vary the confidence level by adding the optional level argument, as in

  ```
  > confint(lm(data~1), level=.99)
  ```

Applications

Dorsal lengths data

Dataset: octopod {durham}

Load the octopod data. The data are the dorsal lengths (in millimetres) of taxonomically distinct octopods. These data were also explored during computer practical 4.

Find a 95% confidence interval for the mean dorsal length. Write down the mean and standard error.

```
# INSERT YOUR CODE HERE ...
```

- WRITE DOWN YOUR COMMENTS HERE...

Assess the Normality of the distribution. Write down your interpretations.

```
# INSERT YOUR CODE HERE ...
```

- WRITE DOWN YOUR COMMENTS HERE...

Clearly, the octopod data are heavily skewed with a long tail to the right. What is the implication of this for the confidence interval you calculated above?

```
# INSERT YOUR CODE HERE ...
```

- WRITE DOWN YOUR COMMENTS HERE...

For heavily skewed positive data, a logarithmic transformation often works well. Assess the normality for the logarithm (base 10, use log10) of the octopod data.

Has the transformation made the distribution more Normal in shape?

```
# INSERT YOUR CODE HERE ...
```

- WRITE DOWN YOUR COMMENTS HERE...

Find a 95% confidence interval for the mean of the logarithm to base 10 of dorsal length.

```
# INSERT YOUR CODE HERE ...
```

- WRITE DOWN YOUR COMMENTS HERE...

What is the advantage of base 10 logarithms over natural logarithms when analysing data?

```
# INSERT YOUR CODE HERE ...
```

- WRITE DOWN YOUR COMMENTS HERE...

Cherry tree volumes data

Dataset: *trees* {datasets}

The trees data consists of heights, diameters and volumes of cherry trees. We will ignore the heights and diameters.

Assess the Normality of the volumes. Write down your interpretations.

```
# INSERT YOUR CODE HERE ...
```

- WRITE DOWN YOUR COMMENTS HERE...

Find a 95% confidence interval for the mean cherry tree volume.

```
# INSERT YOUR CODE HERE ...
```

- WRITE DOWN YOUR COMMENTS HERE...

Sampling distributions

It's easy to take samples from distributions in R. For a finite list of numbers, one can use the sample function. For standard distributions, R has special functions which you can use. For the normal distribution, function is called rnorm and to take a sample of size n from a normal distribution with mean mu and standard deviation sigma, one types

```
> rnorm(n, mu, sigma)
```

Try taking a sample of size 30 from the standard normal. Repeat the command to see that the answer changes each time.

To study sampling distributions, one needs to take multiple samples. An easy way to do this is to take a single large sample and then arrange it as a matrix. For example, to take 10000 samples of size 15 from the normal distribution with 5 and standard deviation 2, one sample in each row of the matrix, type

```
> samples = matrix(rnorm(10000*15, 5, 2), nrow=10000)
```

Now one can study the sampling distribution. One can obtain the mean of each sample by typing:

```
> means = apply(samples, 1, mean)
```

Produce a histogram of the means. What distribution should it have?

```
# INSERT YOUR CODE HERE ...
```

- WRITE DOWN YOUR COMMENTS HERE...

Produce a histogram of the standard deviations. Describe the distribution. How much variation is there relative to the original population standard deviation? How does the mean of these standard deviations compare to the population standard deviation?

```
# INSERT YOUR CODE HERE ...
```

- WRITE DOWN YOUR COMMENTS HERE...

How can you obtain all 10000 confidence intervals?

```
# INSERT YOUR CODE HERE ...
```

- WRITE DOWN YOUR COMMENTS HERE...

What proportion of the confidence intervals contain the population mean? How does it compare to the theoretical expected proportion?

```
# INSERT YOUR CODE HERE ...
```

- WRITE DOWN YOUR COMMENTS HERE...

Now repeat all of this exercise using a uniform distribution (see the R help for the *runif* function). To what extent do you expect the same outcomes and to what extent do you expect them to differ?

```
# INSERT YOUR CODE HERE ...
```

- WRITE DOWN YOUR COMMENTS HERE...

Save me

Generate the document as a Notebook, PDF, Word, or HTML by choosing the relevant option (from the pop-up menu next to the Preview button). Then save your Markdown code by choosing the relevant option (from the task bar menu).

Save the *.Rmd script, so that you can edit it later.