Lab 3: Part 2

The aim of this part of the lab exercise is to give you practical experience in Hypothesis testing using R Studio.

1   Hypothesis testing

1. (a) A company produces synthetic diamonds that have an mean weight of 0.5 carat.

    (b) An experiment has been conducted to evaluate a new process for producing synthetic diamonds. Six diamonds have been generated by the new process, with recorded weights of $0.46, 0.61, 0.52, 0.48, 0.57$ and 0.54 carat.

    (c) It is essential that the new process produces diamonds with a weight in excess of 0.50 carat.

    (d) *Do the six diamond weight measurements present sufficient evidence to indicate that the mean weight of the diamonds produced by the process is in excess of 0.5 carat?* Use R studio to test at the 1% significance level.

    (e) Make sure you open a new markdown file and that you comment and explain your steps such as what hypothesis you are testing.

2. Using the data on breast cancer diagnosis from the University of Wisconsin. This data set contains 4 variables: Diagnosis (M = malignant, B = benign), radius.mean, radius.standard deviation and radius max. These are based on measurements taken from tumour cells in each patient.

    (a) Read in the data set *breastCancer_Wissonsin.csv*. This is a sample of data on Breast Cancer.

    (b) Explore the data and the variables. You can do this through the use of numerical summaries and graphs.

    (c) Compute the proportion of tumours classified as benign (b) in the sample data you read in in step (a)

    (d) The proportion of benign tumours in the population from which this sample data is drawn from is 0.6. Is the proportion in this sample data supportive of this proportion in the population? Write out the hypothesis to test and run this hypothesis test in R. Is H_0 rejected when $\alpha = 0.05$?

    (e) Explore the `radius.mean` attribute (column) in the data. Do so numerically and graphically.

    (f) Is there a significant difference between the radius mean for benign vs malignant? (Hint: use both numerical and visual clues)

    (g) in step (c) you computed the proportion of tumours classified as benign. What is the 95% confidence interval for this proportion?

2 R commands and syntax

```r
#reading in a csv file
df<-read.csv("name-of-csv-file")

#the cumulative distribution function of the t values
pt(test.statistic, df)

#the cumulative distribution function for the standardised Normal distribution Z
pnorm(test statistic)

#Testing if a variable is normally distributed
shapiro.wilks(variable)

#If there are two samples - test for equal variances
var.test(variable1, variable2)

#If there are two samples - testing for equal means
# R will handle non-equal variances
t.test(variable1, variable2)
```