Fundamentals of mathematics and statistics - Assessment 1

This assessment is worth 50% of your final grade for the whole unit. It aims to test your understanding of univariable distributions, sampling distribution and hypothesis testing.

- All solutions should be clearly set out with any hypotheses carefully set up and described.
- You are encouraged to provide your R code to document your work. However, statistical outputs should be interpreted and described, with some demonstration of translating back to the problem domain. Relation to previous literature (etc) is not required.
- The open-ended nature of the assignment means that it is possible to make some choices during the analysis – mark scheme will be flexible to reflect this.

You have been provided with a subset of the data from the Second Manifestations of ARTerial disease (SMART) study. Here is a link to more information about the study.

The columns provide data on patient ID numbers ('ID'), sex ('SEX'), age ('AGE'), age group ('AGE_GRP'), diabetes ('DIABETES' (0=non-diabetic, 1=diabetic)), body mass index ('BMIO'), smoking status ('SMOKING'), systolic blood pressure ('SYSTBP') and whether they experienced a cardiovascular event during follow-up ('EVENT' (1=yes, 0=no)).

Questions

1. **Using the SMART ('FMS_assessment1.csv'') data. (11 marks)**

   a. Produce appropriate initial summaries of sex and smoking. (2 marks)

   b. Remove missing data (coded as NA) from systolic blood pressure and produce appropriate initial summaries of this variable. Discussion on whether systolic blood pressure follows a normal distribution. (6 marks)

   c. Calculate the 95% confidence interval for the true population mean blood pressure (using the sample mean, standard deviation, and the sample size *n*). The average blood pressure in the population is thought to be 145. Does the confidence interval contain 145? How would you interpret this? (3 marks)

Solutions:

Import data into R

```
> df <- read.csv("FMS_assessment1.csv", header=TRUE)
```

a.

```
> table(df$SEX)

Female   Male
   364   1146
```

- 0.5 mark

There are much more males (about 3 times) than females - 0.5 mark

```
> table(df$SMOKING)

Current  Former   Never
    156    1085     254
```

- 0.5 mark

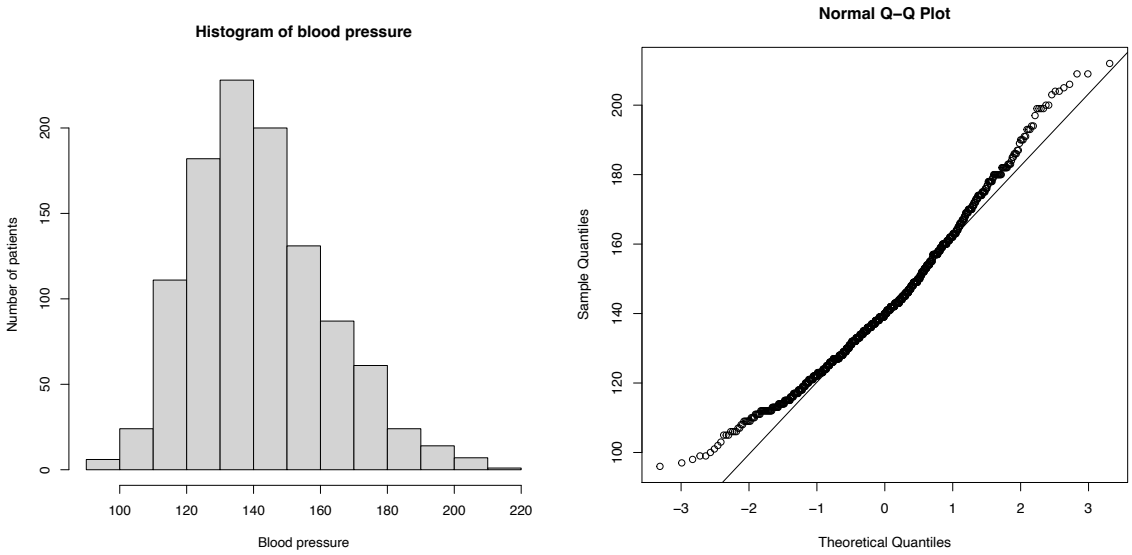
Most of patients are former smokers - 0.5 mark

b. bp.no.missing <- df$SYSTBP[!is.na(df$SYSTBP)] - 0.5 mark

> summary(bp.no.missing)

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
96.0	127.0	140.0	142.3	155.0	212.0 - 0.5 mark

Median is smaller than mean suggesting presence of skewness. 1 mark



Histogram: It is not symmetrical suggesting skewness - correct histogram (1 mark) and comment (1 mark)

QQ plot: Most data points are on or close to the straight line but some are not, showing deviation from normality. - correct QQ plot (1 mark) and comment (1 mark)

c. correct set up (mean/sd/n of data) - 1 mark

```
samplemean <-mean(bp.no.missing)
samplesd <- sd(bp.no.missing)
samplesize <- length(bp.no.missing)
z975 <- 1.96
```

correct upper and lower bounds of confidence interval - 1 mark

```
samplemean + z975*(samplesd/sqrt(samplesize))
samplemean - z975*(samplesd/sqrt(samplesize))
```

Presenting 95% confidence interval (141.08, 143.51)

95% CI calculated from *t*-statistic is also ok.

'95% confidence interval does not contain 145' and that the true population mean blood pressure is different from 145' or likewise. 1 mark

2. **There is a standard therapy used to treat breast cancer, with the cure rate 0.3 (i.e., 30% of patients with breast cancer are cured by the therapy). A new therapy has been developed by a pharmaceutical company. To test if it has a higher cure rate than the standard one, the company will allocate the new therapy to 1000 patients who are randomly drawn from those with breast cancer. The company will conclude that the new therapy is superior if at least 330 of the 1000 patients will cure from breast cancer. (14 marks)**

    a. Suppose X is the number of patients cured from the new therapy. What is the distribution of X? Please specify the null and alternative hypotheses. (4 marks)

    b. Use the central limit theorem to calculate the probability of type I error. Please interpret the result. (4 marks)

    c. If the true cure rate of the new treatment is 0.35, what is the power of the test? Please interpret the result. (2 marks)

    d. If the new treatment has resulted in 400 patients cured from breast cancer. What is the p-value of the test? What conclusion can you draw from the p-value? (4 marks)

Solutions:

a. $X \sim Bin(1000, p)$, where p is cure rate - 2 marks: 1 mark for distribution name, 1 mark for parameters)

   Null hypothesis H_0: the cure rate of the new therapy is the same as that of the standard therapy. $p = 0.3$ - 1 mark

   Alternative hypothesis H_1: the cure rate of the new therapy is higher than that of the standard therapy. $p > 0.3$ - 1 mark

b. P(type I error) = P(reject H_0| H_0)
   = P($X \geq 330 \mid p = 0.3$) - 1 mark
   $\approx$P $\left(Z \geq \frac{330-1000\times0.3}{\sqrt{1000\times0.3\times0.7}}\right)$, by central limit theorem - 2 marks
   = P $(Z \geq 2.0702)$
   = 0.0192 - 0.5 mark
   The probability of the company wrongly concluding that the new therapy is superior is 0.0192. - 0.5 mark

c. Power = P (reject H_0| H_1)
   = P($X \geq 330 \mid p = 0.35$) - 1 mark
   $\approx$P $\left(Z \geq \frac{330-1000\times0.35}{\sqrt{1000\times0.35\times0.65}}\right)$
   = P $(Z \geq -1.326)$
   = 0.9076 - 0.5 mark
   The probability of the company correctly concluding that the new therapy is superior is 0.9076. - 0.5 mark

d. P-value = P($X \geq x \mid H_0$)
   = P($X \geq 400 \mid p = 0.3$) - 1 mark
   $\approx$P $\left(Z \geq \frac{400-1000\times0.3}{\sqrt{1000\times0.3\times0.7}}\right)$

= P ($Z \geq 6.901$)
= 2.58×10^{-12} - 1 mark

The p-value is much smaller than 0.05. Calculation through binom.test() is also OK. Therefore, we reject the null hypothesis (1 mark) and conclude that the cure rate of the new therapy is higher than that of the standard therapy (1 mark).
For Questions c, d, it's OK to use either Binomial distribution or central limit theorem for calculations.

3. **Using the SMART ('FMS_assessment1.csv') data again. (11 marks)**

   a. Test if the true population proportion of male patients is 0.6. (4 marks)

   b. Calculate odds ratio of association between sex and cardiovascular event (EVENT). How would you interpret it? Is there statistical evidence of an association between sex and cardiovascular event? (7 marks)

Solutions:

a. Hypothesis are set up appropriately, with all terms defined: 1 mark. i.e. $H_0: p = 0.6; H_A: p \neq 0.6$, where p is the proportion of male patients.

Test carried out appropriately. Observed proportion of male patients = 1146/1510 1 mark.

Some discussion of assumptions – e.g. independence between patients, normality approximation from the central limit theorem (if using 'prop.test' or otherwise not conducting an exact test) – 1 mark.

```
prop.test(1146,1510, 0.60)

##
##  1-sample proportions test with continuity correction
##
## data:  1146 out of 1510, null probability 0.6
## X-squared = 158.28, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is not equal to 0.6
## 95 percent confidence interval:
##  0.7363866 0.7801599
## sample estimates:
##         p
## 0.7589404
```

Appropriate interpretation given: p<0.05 so there is evidence to reject H_0 i.e. evidence that proportion of males is different from 0.5 – also confidence interval for the proportion (0.736, 0.780) does not include 0.6. 1 mark.

b.

```
table(df$SEX, df$EVENT) 1 mark

##
##            0   1
```

```
##   Female 278  86
##   Male   772 374
```

Estimated odds ratio = (278*374)/(772*86) = 1.566 1 mark.
The odds of experiencing a cardiovascular event in male patients is 56.6% higher than (or 1.566 times) the odds experiencing the event in female patients. 1 mark

Hypotheses: 1 mark
H_0: There is no association between sex and cardiovascular event.
H_1: There is an association between sex and cardiovascular event.

```
chisq.test(df$SEX, df$EVENT)
```
1 mark (ok with or without correction)

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  df$SEX and df$EVENT
## X-squared = 10.163, df = 1, p-value = 0.001433
```

Interpret results: P-value is 0.0014 which is less than 0.05 so we reject the null hypothesis 1 mark. Therefore, there is statistical evidence of an association between sex and cardiovascular event. 1 mark

4. **Using the SMART ('FMS_assessment1.csv') data again, perform tests to investigate the following.** (**14 marks**)

   a. Is there statistical evidence of a difference in BMI between men and women? (6 marks)

   b. Is there statistical evidence of a difference in BMI between age groups? (8 marks)

Solutions:
a.

Set-up hypotheses and correctly determine that unpaired t-test is needed – 2 marks.

$$H_0: \mu_M = \mu_W$$
$$H_1: \mu_M \neq \mu_W$$

Where μ_M is mean BMI for men, and μ_W mean BMI for women
Carry out test correctly, including consideration of equal v unequal variances – 2 marks.
Discussion of other assumptions (e.g. independence, normality of sample statistics) – 1 marks.

```
var.test(BMIO~SEX,data=df)
```

```
##
##  F test to compare two variances
##
## data:  BMIO by SEX
## F = 1.7684, num df = 363, denom df = 1145, p-value = 2.358e-12
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  1.501684 2.097857
## sample estimates:
## ratio of variances
##           1.768411
```

Should NOT assume equal variances:

```
t.test(BMIO~SEX,data=df,var.equal=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  BMIO by SEX
## t = 0.47892, df = 499.99, p-value = 0.6322
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.4000739  0.6579872
## sample estimates:
## mean in group Female   mean in group Male
##             26.82360             26.69464
```

Interpret results: *p*=0.6322, which is greater than 0.05 so we DO NOT reject the null hypothesis H_0, i.e. there is no evidence to suggest that BMI differs between men and women among patients in this study. 1 mark.

b.

```
with(df, tapply(BMIO, AGE_GRP, median))
```

```
##   <56 yrs   >66 yrs 56-66 yrs
##    26.800    25.830    26.715
```

```
with(df, tapply(BMIO, AGE_GRP, mean))
```

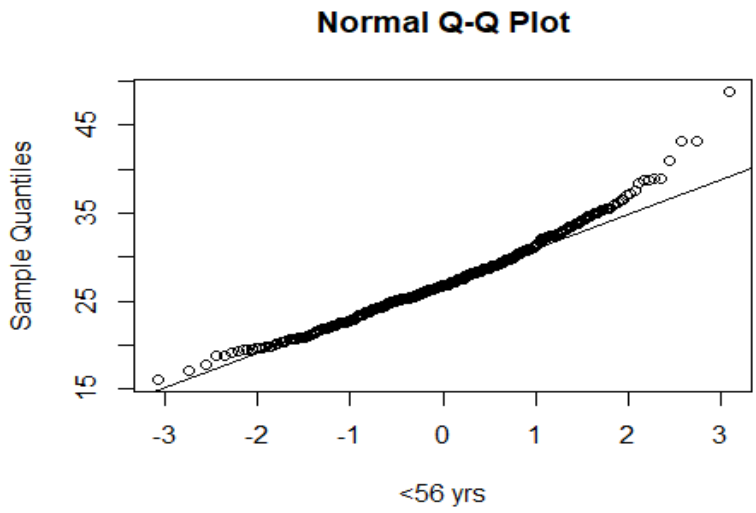
```
##   <56 yrs   >66 yrs 56-66 yrs
##  27.16039  26.27330  26.75831
```

Choice of analysis justified by consideration of assumptions (Normally distributed/central limit theorem applies with equal variance across groups). 3 marks (1 for normality check, 1 for variance check and 1 for appropriate choice of test)

E.g. QQ plot for youngest group suggests distribution may not be Normal therefore could choose Kruskal-Wallis test. Students may judge that deviation from Normality is only slight and/or that central limit theorem can be applied, and choose ANOVA analysis, marks given for justifying the decision either way.

```
with(df, qqnorm(BMIO[AGE_GRP=="<56 yrs"],xlab="<56 yrs"))

with(df, qqline(BMIO[AGE_GRP=="<56 yrs"]))
```



If ANOVA is chosen consider the assumption of equal variances in each group

```
with(df, tapply(BMIO, AGE_GRP, var))

##   <56 yrs   >66 yrs 56-66 yrs
##  19.61700  11.62018  13.58031

leveneTest(BMIO~AGE_GRP, data=df)

## Levene's Test for Homogeneity of Variance (center = median)
##         Df F value    Pr(>F)
## group    2   11.67 9.349e-06 ***
##       1507
```

Sample variances appear somewhat different across groups and from the Levene's test we can conclude that the is evidence the variances are not equal across groups. If ANOVA performed, Welch's ANOVA should be chosen and justified.

Set out null and alternative hypotheses for Kruskal-Wallis test (equality of medians) or Welch's ANOVA (equality of means) with terms defined. 1 mark

E.g.

$$H_0: \nu_1 = \nu_2 = \nu_3$$
$$H_1: \nu_1 \neq \nu_2 \ \ or \ \ \nu_1 \neq \nu_3 \ \ or \ \ \nu_2 \neq \nu_3$$

where ν_i denotes the median BMI for group i

Carry out the chosen test correctly. 1 mark

```
kruskal.test(BMIO~AGE_GRP, data=df)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  BMIO by AGE_GRP
## Kruskal-Wallis chi-squared = 11.84, df = 2, p-value = 0.002686
```

Or testing group means using ANOVA

```
oneway.test(BMIO~AGE_GRP, data=df)

##
##  One-way analysis of means (not assuming equal variances)
##
## data:  BMIO and AGE_GRP
## F = 6.5474, num df = 2.00, denom df = 987.84, p-value = 0.001497
```

Interpret the output. 3 marks

The *p*-value for the Kruskal-Wallis test [Welch's ANOVA] is less than 0.05, so the data provide evidence to reject the null hypothesis and conclude that median [mean] BMI is not equal across age groups.

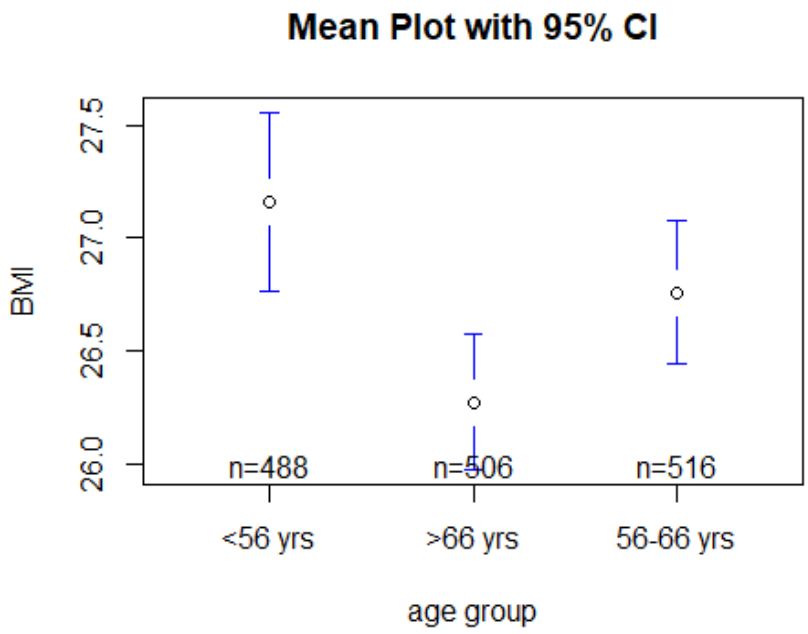
Plotting the means indicate <56 year olds and >66 year old have different BMI's. If ANOVA used then pairwise t.tests using Bonferroni correction and unequal variance (pool.sd = FALSE) can support that there is difference in BMI between the age group <56 and >66.

If assessing medians, visual inspection can conclude the same but differences between groups are less obvious. Pairwise Wilcox text using multiple correction should be used. Results show difference in median BMI between <56 age group and > 66 age group (as with the assessment of mean results) but also between the 56-66 age group and >66 age group.

```
pairwise.t.test(df$BMIO, df$AGE_GRP, p.adjust.method="bonferroni", pool.sd
 = FALSE)

##
##  Pairwise comparisons using t tests with non-pooled SD
##
## data:  df$BMIO and df$AGE_GRP
##
##           <56 yrs >66 yrs
## >66 yrs   0.0013  -
## 56-66 yrs 0.3580  0.0874
##
## P value adjustment method: bonferroni
```

```
with(df, plotmeans(BMIO~AGE_GRP,xlab="age group", connect=FALSE, ylab="BMI
", main="Mean Plot with 95% CI"))
```




```
pairwise.wilcox.test(df$BMIO, df$AGE_GRP, p.adjust.method="bonferroni")

##
##  Pairwise comparisons using Wilcoxon rank sum test
##
## data:  df$BMIO and df$AGE_GRP
##
##           <56 yrs >66 yrs
## >66 yrs   0.0039  -
## 56-66 yrs 1.0000  0.0300
##
## P value adjustment method: bonferroni
```