![Macquarie University Sydney-Australia logo]

STAT270/680: Applied Statistics - Assignment 1

Semester 2 – 2017. Due : Wednesday 13th September 5:00pm

You are expected to write the assignment using MS Word or similar word processor (or in PDF format). You are required to write your name, student ID and Unit Code written on the first page. Please submit your assignment via the provided submission link on iLearn.

You may discuss the assignment in the early stages with your fellow students. However, the assignment submitted should be your own individual work.

You are encouraged to use R Markdown if you are familiar but this is not required. A 'Cheatsheet' from the RStudio team is given here.

In your answers to the questions below, produce the appropriate `R` output and/or explanation of the steps and results. Don't include any more `R` output than necessary and include only concise explanations.

Question 1

An experiment was conducted to examine the effect of age on heartbeat rate when a subject is subjected to a specific amount of exercise. Male subjects were selected at random from four age groups. Each subject walked a treadmill at a fixed grade for 12 minutes and the increase in heartbeats per minute was recorded.
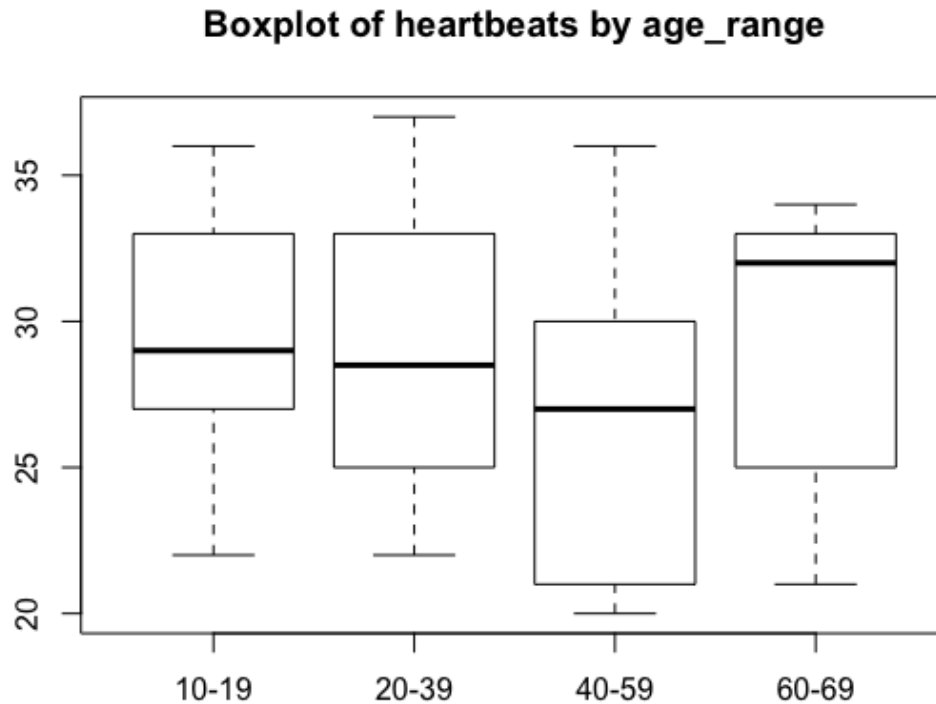
heartbeats	Increase in heartbeats per minute
age_range	Four age groups ("10–19", "20–39", "40-59", "60–69")

a. Comment on the features of the data using the summary statistics and the boxplot shown below.

```
> dat = read.table(file="heartbeats.txt",header=TRUE)
> ns = with(dat,tapply(heartbeats,age_range,length))
> means = with(dat,tapply(heartbeats,age_range,mean))
> sds = with(dat,tapply(heartbeats,age_range,sd))
> prod.data = data.frame(n=ns,mean = means,sd = sds)

##        n mean       sd
## 10-19  9 30.0 4.609772
## 20-39 10 29.0 4.760952
## 40-59 10 26.7 5.355164
## 60-69 10 29.5 4.790036

> boxplot(heartbeats~age_range,data=dat,main="Boxplot of heartbeats by age_range")
```

Boxplot of heartbeats by age_range

b. Define an appropriate statistical model to conduct a hypothesis test that the mean increase in heartbeats are the same between all age groups. In your answer:

   i. Define all appropriate parameters and fully state the distribution required to conduct the test.

   ii. Using the information provided, state whether the requirements for the hypothesis test are satisfied for this dataset. If any requirements cannot be checked using the information provided, state what additional information is needed. **Note**: In this part of the question, you are **not** required to conduct any hypothesis test or state any hypothesis, only write down the mathematical model that is appropriate for this situation.

c. An ANOVA table for this dataset is shown below

Source	d.f.	Sum Sq	Mean Sq	F
age_range	?	62.32	?	?
Residuals	?	?	23.960	
Total	?	?		

Determine the missing parts of the table using the properties of ANOVA tables and/or the information provided earlier.

d. Assuming an ANOVA is valid for this dataset, use the completed ANOVA table in part c. to answer the research question "Is there a difference in the increase of heartbeats due to the age groups?". In your answer:

   i. State the appropriate Hypotheses;

   ii. Compute an appropriate test statistic;

iii. Describe the null distribution of that statistic;
iv. Compute a rejection region **or** P-value;
v. State the statistical conclusion;
vi. State the contextual conclusion.

e. Suppose all multiple pairwise comparisons of all the four age groups ("10–19", "20–39", "40-59", "60–69") at an overall 5% significance level are required.

  i. What is the total number of pairwise comparisons?
  ii. What is the significance level for each individual comparison (test) if using the Bonferroni approach at an overall 5% significance level?
  iii. What is the t-critical value that should be used for each of those pairwise comparisons if using the Beonferroni approach? State also the degrees of freedom for this t-distribution.

Question 2

Let us consider an experimental study of drugs to relieve itching. Five drugs (N3-N7) were compared to a placebo (N2) and no drugs (N1) with 10 volunteer male subjects aged 20-30. Itch severity is measured in terms of the Dermatology Life Quality Index (DLQI) and the scores range from 0 (no impairment of life quality) to 30 (maximum impairment). The data is available in the file `itch.txt` on `iLearn`.

DLQI	Itch severity
Treatment	Five drugs (N3–N7), a placebo (N2) and no drugs (N1)

a. Do a preliminary analysis to summarise the data.

  i. Compute numerical summaries including the number of observations, the mean and standard deviation of itch severity for each treatment;
  ii. Show a graphical summary of the data in form of a boxplot split by the treatments;
  iii. Show a normal QQ plot for the residuals;
  iv. Comment on the features of the data using the results of i., ii. and iii..

b. Comment on whether or not the data appears to satisfy the assumptions for an analysis of variance.

c. Assuming an ANOVA is valid for this dataset, perform an ANOVA to check whether the itch severity is the same across all treatments or not. In your answer:

  i. State the appropriate Hypotheses;
  ii. Compute an appropriate test statistic;
  iii. Describe the null distribution of that statistic;
  iv. Compute a rejection region **or** P-value;
  v. State the statistical conclusion;
  vi. State the contextual conclusion.

d. Suppose all multiple pairwise comparisons of all the treatments at an overall 5% significance level are required by using the Bonferroni approach. Find pairs for which the means are significantly different at an overall 5% significance level.