# STAT270 Assignment 1

Session 2 2017

Justin Lam 45197083

Due: 13th September 2017 5:00pm

1(a)

```
dat = read.table(file = "heartbeats.txt", header = TRUE)

ns = with(dat,tapply(heartbeats,age_range, length))

ns means = with(dat,tapply(heartbeats, age_range, mean))

sds = with(dat,tapply(heartbeats, age_range, sd))

sds prod.data = data.frame(n=ns,means,sd = sds)

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

Observations:

From the data provided, it shows there are 4 different data sets, which are “10-19”, “20-39”, “30-49” and “60-69”. In order to understand the nature and features of the data, we need to comment on their sample sizes, means and standard deviations. Firstly, the sample size of age “20-39”, “40-59” and “60-69” are the same as the can be shown in the guidelines; on the other hand the sample size of age group “10-19” is 9 which is 1 less the other age groups. Secondly, the means of age group “20-39” and “60-69” as this can be shown via the summary of data and the boxplot provided later in the guidelines. Moreover, the standard deviation of age groups “10-19” and “60-69” are similar; this can be also shown through the summary of data and the boxplot.

1(b)

An appropriate statistical model to conduct a hypothesis test is the F-test, where the parameters are as follows. Firstly, we have a parameter ($Yij = \mu i + \varepsilon ij$), where Yij is the jth response; this can be (from 1, 2, 3….k levels) from the population i, μi is the3 mean response (or the mean of the treatment groups) and εij is the random error of the model. Usually, the mean response is fixed while the random error is usually a random response, as stated in the name itself. In addition, another parameter that we can use is the $\varepsilon ij \sim N(0, \sigma 2)$. Which represents a normal, constant variance of the data.

(c) ANOVA table.

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|---|---|---|---|---|---|
| age_range | 3 | 62.3 | 20.77 | 0.867 | 0.467 |
| Residuals | 35 | 838.6 | 23.96 | | |

Total degrees of freedom: 38; Total sum of squares: 900.9

(d)

(i) $H_0$: All the means are the same. $H_1$: At least one of the paired means are different.

(ii) The test-statistic of the test = 0.867

(iii) Null distribution. F ~ (3,35)

(iv) P-value = 0.467

(v) Statistical conclusion: Since the p-value is greater than 0.05, therefore we do not reject the null hypothesis. i.e. all the means are the same.

(vi) The contextual conclusion: Since we do not reject the null hypothesis, therefore the means of age on heartbeat rate in different groups (10-19, 20-39, 40-59, 60-69).

1(e) Pairwise comparison:

(i) The total number of pairwise comparisons = (4) (4-1)/2 = 6 pairs.

(ii) The significance level of each individual comparison = (the significance level/ number pairwise comparisons). ∴ 0.05/6 = 0.00833 (to the near nearest 5 decimal points)

(iii) The significance level for individual comparison (via Bonferroni approach) 2.796605

Question 2:

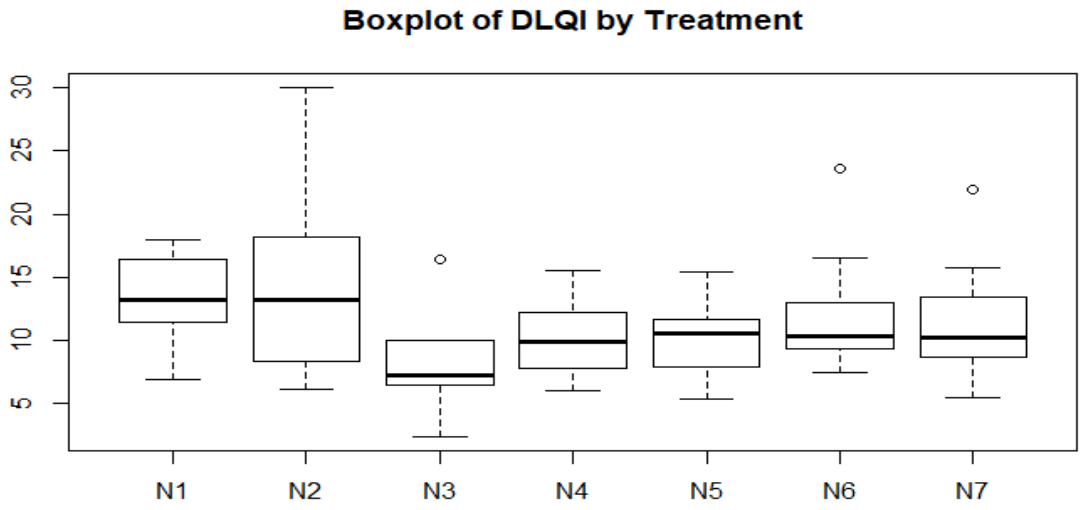
2(a) (i) The command below will help the user to generate the numerical summary of the data.

```
itch = read.table(file = "itch.txt", header = TRUE)

ns1 = with(itch,tapply(DLQI,Treatment, length))

ns1

means1 = with(itch,tapply(DLQI, Treatment, mean))

means1

sd1= with(itch,tapply(DLQI, Treatment, sd))

sd1
```

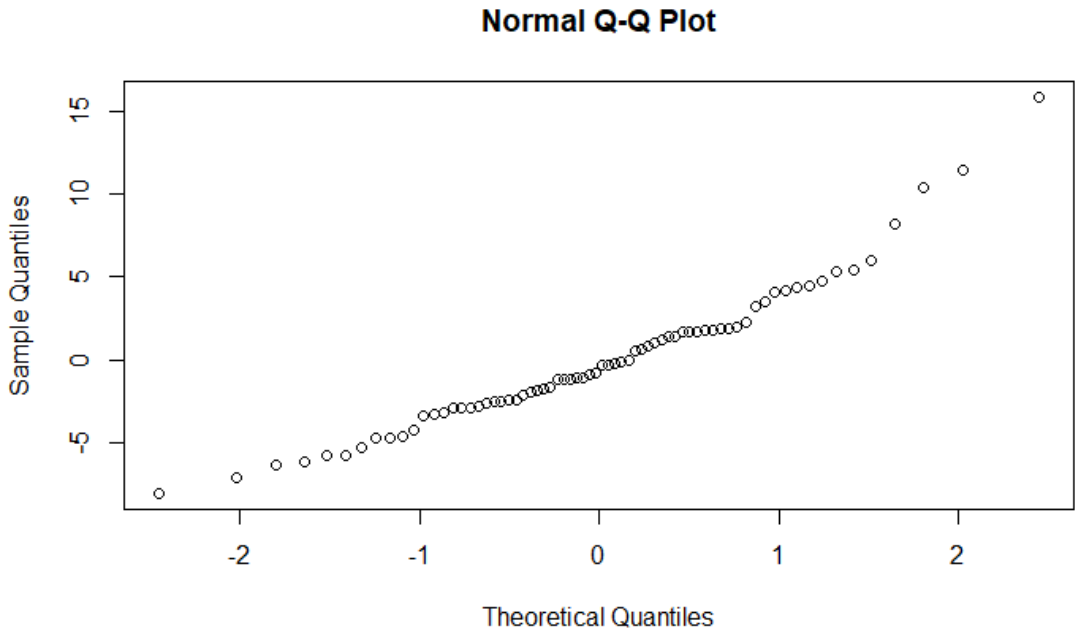
Observations:

From the data provided, there are 10 samples in each of the treatments groups and there are 7 treatment groups overall in the data set. In order to display and take note of the numerical summary of the data sets, as a result we need to interpret the key features of data. For instance, the means of each are as follows: 13.233256 (N1), 14.189376 (N2), 8.189376 (N3), 9.997691 (N4), 9.997691 (N5), 12.228637 (N6), 11.584296 (N7). In addition, the standard deviation of each treatments are as follows 7.32971 (N2), 3.658858 (N3), 3.099682 (N4), 2.915250 (N5) 4.770633 (N6), 4.676637(N7)

2(a)(ii) A graphical summary of data in the form of a boxplot spilt by the treatments.



2(a) (iii) A normal QQ plot for the residuals.




2(a)(iv) the features of data using part (i), (ii) and (iii)

- From the results in part (i), it shows that N4 and N5 have the same mean. While N1 and N2 have similar mean as this can be shown in the data itself and the boxplot (part ii). Lastly, N6 and N7 also have similar means. Besides means, we can use standard deviation to compare the features of the data. For instance, N6 and N7 have similar standard deviations as well as N and N3.
- From the boxplot above, it shows the N2 has an uneven spread which means the data (N2) may have an equal spread of variance. In addition, the median for N1 and N2 are similar as well as both N4 and N5.
- From the QQ plot generated in 2(a) (iii), the distribution of the data appears to be questionable as it does not form a straight line which represents the data may not draw from a normal distribution.

2(b) In order to test whether the assumptions of ANOVA, we can use a Bartlett's test to do so.

Code = bartlett.test (DLQI ~ Treatment, data = itch)

When we input the above code, we will get a p-value output of 0.07724. Since this value is greater than 0.05 therefore there's enough evidence to suggest that the assumption of ANOVA test is valid.

2(c) Perform an ANOVA test on the itch dataset.

i. $H_0$: All the means are equals; $H_1$: not all means are equal.
ii. Assumptions can be drawn from the Bartlett test.
iii. Null Distribution = F ~ (6, 63)
iv. The p-value is 0.07724.
v. Statistical conclusion: Since the p-value is less than 0.05, as a result there is not evidence to reject the hypothesis.
vi. Contextual conclusion: Since we do not reject the null hypothesis, as a result, there are significant evidence stated that the means of itch severity and the different drug treatments (N1, N2, N3, N4, 5, N6, N7) are the same.

2(d) We can the function “pairwise.t.test ()” in R to conduct a pairwise t-test (in terms of Bonferroni approach). The command of such test can be input as “ pairwise.t.test (itch$DLQI, itch$Treatment, p.adj = “bonf”). The output of this test will gives the user a series of p-values which determine all pairs of means are the same. From the output provided, all of the p-values are greater than 0.05; which suggests that all means are the same.