STAT3005 Assignment 2
Lam Hoi Chun (1155192755)

Exercise 2.1

Ex 2.1 1(a)

Since we have to randomly selected one record if the same animal was observed more than one time, with the fact that the dataset contains 25 observations with 21 distinct animals. We first extract the observations with the distinct animals.

Notice that we are testing median$(X_1, \dots, X_n) = \theta$. We should perform sign test in this case. So. we can first compute T, where

$$\begin{aligned} T &= \sum_{i=1}^{n} \psi_i \\ &= \sum_{i=1}^{n} \operatorname{sign}(X_i - 21) \\ &= -11 \end{aligned}$$

$$\begin{aligned} \mathrm{E}(T) &= 0 \\ \mathrm{Var}(T) &= n \end{aligned}$$

Then, we can find:

$$\begin{aligned} t &= \frac{T - \mathrm{E}(T)}{\sqrt{\mathrm{Var}(T)}} \\ &= -2.400397 \end{aligned}$$

And the corresponding p-value = 0.01637731.

We can also use the binom.test() function in R: binom.test((T_value+n)/2, n)$p.value will yield the result of 0.0266037. Although the p-values from the two methods are slightly different, they are both smaller than 0.05.

$\because$ The exact test and the binom.test() function in R both yield the results of $\hat{p} < 0.05$

$\therefore$ We reject H_0 in 0.05 significance level.

As we apply the sign test, we assume that the data $X_i = \mu + \varepsilon_i$, where ε_i are continuous, independent and of median 0.

Ex 2.1 1(b)

We get a p-value ≈ 0.02 means that if the actual median $= 21$, there is only 2% chance of getting such observed values. We can also say that if we keep doing such experiment for 100 times, there is only 1 time that this case, or even more extreme cases happen, which means there is less than or equal to 5 observations that is greater than 21.

Ex 2.1 2(a)

$$H_0 : \theta_1 - \theta_2 = 0 \qquad H_1 : \theta_1 - \theta_2 < 0$$

Ex 2.1 2(b)

For 2 sample t-test, we construct:

$$\begin{aligned} t &= \frac{\text{median(sample1) - median(sample2)}}{s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \\ &= -0.3933845 \end{aligned}$$

where

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

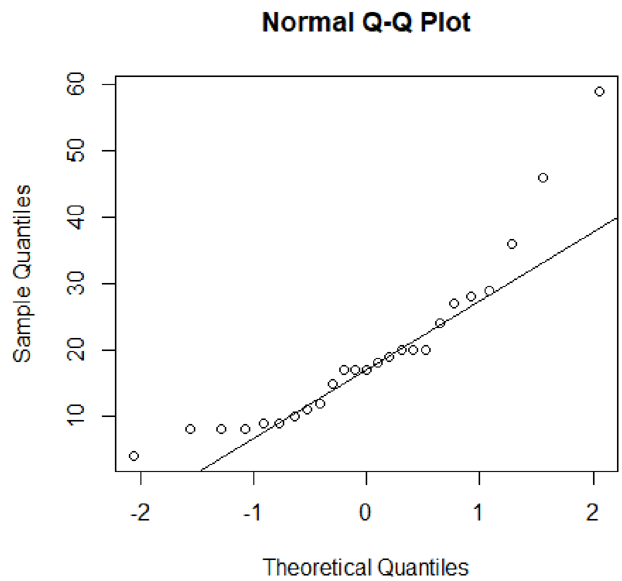
The p-value found using 2-sample t-test is 0.3488.

For non-parametric test, we can perform Wilcoxon rank sum test using the wilcox.test() function in R, which gives the p-value of 0.05929. As both p-values < 0.05, both tests cannot reject H_0 at the significance level of 0.05.

In the 2-sample t-test. we assume that the sample is normally distributed, since mean equals median in normal distribution. Besides, in the non-parametric test. we assume the data follow: $X_i^{(j)} = \mu_j + \varepsilon_i^{(j)}$, where $\varepsilon_i^{(j)}$, where $j = 1, 2$, are CIID.

Ex 2.1 2(b)(Cont'd)

As we assume that the sample is normally distributed in the 2-sample t-test, we can verifty this using QQ Plot. From the QQ Plot, it can be seen that the data does not form a straight line



passing through origin. So, the normality of the sample does not holds, and so I would trust the t-test, which is more reliable because it makes less and weaker assumptions about the data.

Ex 2.1 2(c)

To compare the results of the 2-sample t-test and the Wilcoxon rank sum test, it seems that they can give the same results, as both p-values are smaller than 0.05. The p-value of 0.3488 in the 2-sample t-test will lead to a result that of not rejecting H_0 since it is much larger than the common significance level (0.05). However, the p-value of 0.05929 in the Wilcoxon rank sum test is much closer to 0.05. Although we still cannot reject H_0 at the significance level of 0.05, the result shows a stronger indication that heavier mammals may have a longer urinate time, than the 2-sample t-test.

Exercise 2.2

Ex 2.2 (1)

The null and alternative hypotheses are

$$H_0 : \mu_1 = \mu_2 = \mu_3 \quad \text{and} \quad H_1 : \mu_1 < \mu_2 < \mu_3$$

Ex 2.2 (2)

We can rewrite the test statistic T as

$$T = \sum_{i=1}^{n} \frac{R_i}{n+1} \cdot \left[\mathbb{1}(i \le n_1) - \mathbb{1}(i \ge n_1 + n_2 + 1)\right]$$

Note that T is a test statistic with

$$c(i) = \mathbb{1}(i \le n_1) - \mathbb{1}(i \ge n_1 + n_2 + 1) \quad \text{and} \quad s(i) = \frac{i}{n+1}$$

Also note that

$$\begin{aligned}
\bar{c} &= \frac{1}{n} \sum_{i=1}^{n} c(i) = \frac{n_1 - n_3}{n} = p_1 - p_3 \\
\bar{s} &= \frac{1}{n} \sum_{i=1}^{n} \frac{i}{n+1} = \frac{1}{n(n+1)} \cdot \frac{n(n+1)}{2} = \frac{1}{2}
\end{aligned}$$

By Theorem 3.1, Under H_0, we have

$$\therefore \mathrm{E}(T) = n\bar{c}\bar{s} = n \cdot (p_1 - p_3) \cdot \frac{1}{2} = \frac{n(p_1 - p_3)}{2}$$

To find $\mathrm{Var}(T)$, we have to find σ_c^2 and σ_s^2

$$\begin{aligned}
\sigma_c^2 &= \bar{c^2} - \bar{c}^2 \\
&= \frac{1}{n} \sum_{i=1}^{n} [c(i)]^2 - \bar{c}^2 \\
&= \frac{1}{n} \sum_{i=1}^{n} \left[\mathbb{1}(i \le n_1) - \mathbb{1}(i \ge n_1 + n_2 + 1)\right]^2 - \bar{c}^2 \\
&= \frac{n_1 + n_3}{n} - (p_1 - p_3)^2 \\
&= p_1 + p_3 - (p_1 - p_3)^2
\end{aligned}$$

Ex 2.2 (2) (Cont'd)

$$\begin{aligned}
\sigma_s^2 &= \bar{s^2} - \bar{s}^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}[s(i)]^2 - \bar{s}^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{i}{n+1}\right)^2 - \bar{s}^2 \\
&= \frac{1}{n(n+1)^2}\cdot\sum_{i=1}^{n} i^2 - \bar{s}^2 \\
&= \frac{1}{n(n+1)^2}\cdot\frac{n(n+1)(2n+1)}{6} - \frac{1}{4} \\
&= \frac{n-1}{12(n+1)}
\end{aligned}$$

By Theorem 3.1, Under H_0, we have

$$\begin{aligned}
\mathrm{Var}(T) &= \frac{n^2}{n-1}\cdot\sigma_s^2\cdot\sigma_c^2 \\
&= \frac{n^2}{n-1}\cdot\left[\frac{n-1}{12(n+1)}\right]\cdot\left[p_1+p_3-(p_1-p_3)^2\right] \\
&= \frac{n^2\left[p_1+p_3-(p_1-p_3)^2\right]}{12(n+1)}
\end{aligned}$$

Ex 2.2 (3)

We have to check the following conditions in Theorem 3.1.

- (Data Condition) The data $X_1, \ldots, X_n$ are CIID under H_0.
- (Score Condition) The score generating function is $S(u) = u$ is not a constant and satisfies

$$\int_0^1 S^2(u)du = \int_0^1 u^2 du = \frac{1}{3}\left[u^3\right]_0^1 = \frac{1}{3} < \infty$$

- (Coefficient Condition) Note that $\bar{c} = \mu_c = p_1 - p_3$. As $n \to \infty$, we have

$$\begin{aligned}
\max\nolimits_{1\le i\le n}\left[c(i)-\bar{c}\right]^2 &= \max\left[(p_1-p_3)^2, (p_1-p_3-1)^2, (p_1-p_3+1)^2\right] \le 4 \\
\sum_{i=1}^{n}\left[c(i)-\bar{c}\right]^2 &= n\sigma_c^2 = \frac{n^3\left[p_1+p_3-(p_1-p_3)^2\right]}{12(n+1)} \xrightarrow{n\to\infty} \infty \\
\therefore\ \mathrm{UAN}_n &= \frac{\max_{1\le i\le n}\left[c(i)-\bar{c}\right]^2}{\sum_{i=1}^{n}\left[c(i)-\bar{c}\right]^2} \xrightarrow{n\to\infty} \frac{4}{\infty} = 0
\end{aligned}$$

Ex 2.2 (3) (Cont'd)

Since all the conditions are satisfied, by Theorem 3.1,

$$t := \frac{T - \mu_T}{\sigma_T} \xrightarrow{d} \text{N(0,1)}$$

Ex 2.2 (4)

$$\begin{aligned} \text{Under } H_1 &\Rightarrow R_1, \ldots, R_{n_1} \text{ are likely to be smaller} \\ &\Rightarrow T \text{ is likely to be small relative to the null distribution} \\ &\Rightarrow t \text{ is likely to be small relative to the null distribution} \\ \text{By critical value} &\Rightarrow \text{We reject } H_0 \text{ iff } t \text{ is small} \quad \iff \quad t < \text{qnorm}(1-\alpha_0) \end{aligned}$$

Consequently, the asymptotic p-value is

$$\hat{p} = \text{pnorm}(t)$$

and we will reject H_0 iff $\hat{p} < \alpha_0$

Ex 2.2 (5)

Using random seed = 123, we can get the result below:

	80%	90%	95%	99%	99.9%
$n_1 = 10$	0.8529217	1.2793825	1.6165841	2.3207404	2.9753578
$n_1 = 20$	0.8274101	1.2604800	1.6336948	2.3097768	3.1934649
$n_1 = 30$	0.8463572	1.2935255	1.6428157	2.3183661	2.8346555
$n_1 \to \infty$	0.8416212	1.2815516	1.6448536	2.3263479	3.0902323

Exercise 2.3

Ex 2.3 (1)

$V(\theta)$ is the total number of Walsh averages that are greater than or equal to θ.

Ex 2.3 (2)

Let $Z_i = X_i - \theta$ for $i = 1, \ldots, n$, and arrange them to $Z_{[1]}$, ..., $Z_{[n]}$ according to their absolute values: $|Z_{[1]}|$, ..., $|Z_{[n]}|$. Then, we can rewrite

$$W_{i,j} = \frac{X_i + X_j}{2} = \frac{Z_i + Z_j}{2} + \theta$$

Recall that we want to find the total number of Walsh averages that are greater than or equal to θ, i.e. $W_{i,j} > \theta \iff Z_i + Z_j \geq 0$. So we can rewrite

$$\begin{aligned}
V(\theta) &= \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbb{1}(W_{i,j} \geq \theta) \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbb{1}(Z_i + Z_j \geq 0) \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbb{1}(Z_{[i]} + Z_{[j]} \geq 0)
\end{aligned}$$

Note that $Z_{[i]} + Z_{[j]} \geq 0$ iff $Z_{[i]} > 0$. Also, we know that if $Z_{[i]} > 0$, there are i values for j such that $Z_{[i]} + Z_{[j]} \geq 0$. So, we can rewrite

$$V_i(\theta) = i\mathbb{1}(Z_{[i]} \geq 0)$$

From the last part, we find that there are i values for j such that $Z_{[i]} + Z_{[j]} \geq 0$. In fact, rank($|Z_i|$) also equals to the number of values. So, we can rewrite

$$\begin{aligned}
V(\theta) &= \sum_{i=1}^{n} i\mathbb{1}(Z_{[i]} \geq 0) \\
&= \sum_{i=1}^{n} \text{rank}(|Z_i|) \times \mathbb{1}(Z_i \geq 0)
\end{aligned}$$

Ex 2.3 (2)(Cont'd)

Using the identity $\text{sign}(Z_i) = 2\mathbb{1}(Z_i \geq 0) - 1$, we can rewrite

$$\begin{aligned}
T(\theta) &= \frac{1}{n+1}\sum_{i=1}^{n} \text{sign}(Z_i) \times \text{rank}(|Z_i|) \\
&= \frac{1}{n+1}\sum_{i=1}^{n} \left[2 \cdot \mathbb{1}(Z_i \geq 0) - 1\right] \times \text{rank}(|Z_i|) \\
&= \frac{2}{n+1}\sum_{i=1}^{n} \text{rank}(|Z_i|) \cdot \mathbb{1}(Z_i \geq 0) - \frac{1}{n+1}\sum_{i=1}^{n} \text{rank}(|Z_i|) \\
&= \frac{2}{n+1}\sum_{i=1}^{n} \text{rank}(|Z_i|) \cdot \mathbb{1}(Z_i \geq 0) - \frac{1}{n+1} \cdot \frac{n(n+1)}{2} \\
&= \frac{2}{n+1}\sum_{i=1}^{n} \text{rank}(|Z_i|) \cdot \mathbb{1}(Z_i \geq 0) - \frac{n}{2} \\
&= \frac{2}{n+1} \cdot V(\theta) - \frac{n}{2} \\
V(\theta) &= \frac{n+1}{2}\left[F(\theta) + \frac{n}{2}\right]
\end{aligned}$$

Ex 2.3 (3)

$$\begin{aligned}
\hat{\theta} &= \text{arg min}_\theta |T(\theta) - 0| \\
&= \text{arg min}_\theta |T(\theta)| \\
&= \text{arg min}_\theta \left|\frac{2}{n+1} \cdot V(\theta) - \frac{n}{2}\right| \\
&= \text{arg min}_\theta \left|V(\theta) - \left(\frac{n+1}{2}\right)\left(\frac{n}{2}\right)\right| \\
&= \text{arg min}_\theta \left|V(\theta) - \frac{N}{2}\right|
\end{aligned}$$

To minimise the absolute difference when N is odd, we can rewrite

$$\frac{N}{2} = m + \frac{1}{2}$$

for some m, where we can choose between m or $m+1$. But it is sensible to choose

$$\hat{\theta} = W_{(m+1)} = W_{(\frac{N+1}{2})}$$

as it is the median of Walsh averages.

Ex 2.3 (3)(Cont'd)

When N is even, we can use the same idea from last part. We have

$$\frac{N}{2} = m$$

for some m. To minimise the absolute difference we can use the median of Walsh averages

$$\hat{\theta} = \frac{W_{(\frac{N}{2})} + W_{(\frac{N+1}{2})}}{2}$$

So we can conclude that one possible solution is

$$\hat{\theta} = \begin{cases} W_{((N+1)/2)} & \text{if } N \text{ is odd;} \\ \{W_{(N/2)} + W_{(N/2+1)}\}/2 & \text{if } N \text{ is even} \end{cases}$$

Ex 2.4 (1)

Recall from *Ex 2.3 (2)*

$$T(\theta) = \frac{2}{n+1} \cdot V(\theta) - \frac{n}{2}$$

So we can rewrite

$$\begin{aligned} T^2(\theta) &\leq c_{1-\alpha} \\ \left[\frac{2}{n+1} \cdot V(\theta) - \frac{n}{2}\right]^2 &\leq c_{1-\alpha} \\ \left|\frac{2}{n+1} \cdot V(\theta) - \frac{n}{2}\right| &\leq \sqrt{c_{1-\alpha}} \end{aligned}$$

Then we can get

$$-\sqrt{c_{1-\alpha}} \leq \frac{2}{n+1} \cdot V(\theta) - \frac{n}{2} \leq \sqrt{c_{1-\alpha}}$$

Consider the upper bound:

$$\begin{aligned} \frac{2}{n+1} \cdot V(\theta) - \frac{n}{2} \leq \sqrt{c_{1-\alpha}} & \\ V(\theta) &\leq \frac{n+1}{2}\left(\sqrt{c_{1-\alpha}} + \frac{n}{2}\right) \\ &= \frac{n(n+1)}{4} + \frac{(n+1)\sqrt{c_{1-\alpha}}}{2} \\ &= N_\alpha^+ \end{aligned}$$

Then consider the lower bound:

$$\begin{aligned} -\sqrt{c_{1-\alpha}} &\leq \frac{2}{n+1} \cdot V(\theta) - \frac{n}{2} \\ V(\theta) &\geq \frac{n+1}{2}\left(-\sqrt{c_{1-\alpha}} + \frac{n}{2}\right) \\ &= \frac{n(n+1)}{4} - \frac{(n+1)\sqrt{c_{1-\alpha}}}{2} \\ &= N_\alpha^- \end{aligned}$$

So we proved

$$\{\theta : T^2(\theta) \leq c_{1-\alpha}\} = \{\theta : N_\alpha^- \leq V(\theta) \leq N_\alpha^+\}$$

Ex 2.4 (2)

Recall that $V(\theta)$ counts the number of Walsh averages $W_{i,j}$ that are greater than or equal to θ. Note that N is the total number of Walsh averages. We can rewrite

$$V(\theta) = N - i \quad \text{for } \theta \in (W_{(i)}, W_{(i+1)}]$$

Using the relationship above, we can rewrite

$$\begin{aligned} V(\theta) \le N_\alpha^+ &\iff \theta > W_{N-N_\alpha^+} \\ V(\theta) \ge N_\alpha^- &\iff \theta \le W_{N-N_\alpha^-+1} \end{aligned}$$

So we proved

$$\{\theta : N_\alpha^- \le V(\theta) \le N_\alpha^+\} = \{\theta : W_{N-N_\alpha^+} < \theta \le W_{N-N_\alpha^-+1}\}$$

$$\begin{aligned} \therefore 1 - \alpha &= P(\hat{\theta} \in \hat{I}) \\ &= P\left\{\hat{\theta} \in \left(W_{N-N_\alpha^+}, W_{N-N_\alpha^-+1}\right]\right\} \end{aligned}$$