MATH524 Assignment 1

Question 1

The effectiveness of vitamin C in orange juice and in synthetic ascorbic acid was compared in 12 guinea pigs (divided at random into 2 groups of 6) in terms of the lengths of the odontoblasts (type of tooth precusor) after 6 weeks, with the following results.

	(8)	(11)	(5)	(12)	(7)	(6)
Orange juice	9.7	15.2	8.2	17.6	9.6	9.4
Ascobic acid	11.3	11.2	4.2	7.3	5.8	6.4
	(10)	(9)	(1)	(4)	(2)	(3)

(a) If it is suspected a priority that orange juice would produce longer odontoblasts, test appropriate hypotheses. State your hypotheses, your test statistic, define your p-value, calculate you p-value and state your conclusion.

hypotheses :

H_0: Tr and C are equivalent (i.e. odontoblasts of orange juice and ascorbic acid are equivalent)

H_a: Tr (Orange juice) would produce longer odontoblasts than ascorbic acid.

Test-statistic

$$T = W_{xy_{obs}} = W_s - \frac{1}{2}n(n+1) = (8+11+5+12+7+6) - \frac{1}{2} \times 6 \times 7$$

$$= 28$$

P-value

$$P = P_{H_0}(W_{xy} \geq W_{xy_{obs}})$$

$$= P_{H_0}(W_{xy} \geq 28)$$

$$= 0.066$$

conclusion

The p-value is small ($0.066 < 0.1$). There is evidence that in this group of 12 Orange juice would produce longer odontoblasts than ascorbic acid.

(b) Repeat part (a) using a 2-sided alternative

hypotheses :

H_0: Tr and C are equivalent (i.e. odontoblasts of orange juice and ascorbic acid are equivalent)

H_a: Tr (Orange juice) would produce longer or shorter odontoblasts than orange juice.

Test statistics :

$$W_{xy_{obs}} = W_s - \frac{1}{2}n(n+1) = 28$$

P-value :

$$P = 2 \min\{P_{H_0}(W_{xy} \geq 28, W_{xy} \leq 28)\} = 0.132$$

conclusion :

The p-value = 0.132 > 0.1 is large. Therefore, there does not appear to be evidence that in this group of 12 subjects orange juice would produce longer or shorter odonblasts than orange juice.

Question 2

A group of 22 people with allergies was randomly divided into two treatment groups of size 9 and 22 respectively. All 22 were exposed to a certain allergy and their sputum Histamine levels were recorded. They are shown below.

TABLE 4.3. Sputum Histamine Levels ($\mu g/g$ Dry Weight Sputum)

Tr 1	Tr 2
1651.0 22	48.1 15
1112.0 21	48.0 14
102.4 20	45.5 13
100.0 19	41.7 12
67.6 18	35.4 10
65.9 17	34.3 9
64.7 16	32.4 8
39.6 11	29.1 6
31.0 7	27.3 5
	18.9 4
	6.6 3
	5.2 2
	4.7 1

Source: H. V. Thomas and E. Simmons (1969).

It was a-priori suspected that Treatment 2 was more effective in lowering histamine levels. A-priori it was not known which treatment would produce more variable histamine levels.

(a) Test the hypothesis that Sputum Histamine levels are ..... Fill in the appropriate words yourself.

H_0 : Treatment 1 & Treatment 2 are equivalent

H_a : Treatment 2 was more effective in lowering histamine levels

$$W_{xy} = W_s - \frac{1}{2}n(n+1)$$

$$= (1+2+3+4+5+6+8+9+10+12+13+14+15) - \frac{1}{2} \times 13 \times 14$$

$$= 11$$

$$P = P(W_{xy} \leq W_{xy,obs})$$

$$= P(W_{xy} \leq 11)$$

$$= 0.000386$$

..... will be filled by: " less likely to be equivalent for treatment 1 and treatment 2. Since the p-value $0.000386 < 0.05$, there is evidence that treatment 2 was more effective in lowering histamine level

(b)Test for ..... in variation ....., again filling in the appropriate words yourself.

Test for treatment 2 is less than treatment 1 in variation.

hypotheses:

H_0: treatment 1 and treatment 2 have same level of variation

H_a: treatment 2 is less variable than treatment 1

According to Siegel - Turkey Test:

4.7	5.2	6.6	18.9	27.3	29.1	31.0	32.4	34.3	35.4	39.6	41.7	46.5	48.0	48.1	64.7
1	4	5	8	9	12	13	16	17	20	21	22	19	18	15	14
S_1	S_2	S_3	S_4	S_5	S_6		S_8	S_9	S_2		S_{13}	S_{11}	S_{10}	S_7	

65.9	67.6	100.0	102.4	1112.0	1651.0
11	10	7	6	3	2

Test statistic

$$W_s = 1+4+5+8+9+12+17+20+22+19+18+15+16$$
$$= 166$$

$$P = P_{H_0}(W_s \geq W_{s\,obs})$$
$$= P_{H_0}(W_s \geq 150)$$
$$= 0.146$$

Conclusion: As P-value > 0.1, there is no evidence that treatment2 is less variable than treatment 1.

Question 5

Find $P_{H0}(S_r = k)$ for any $r \leq n$

Since $P_{H0}(S_r = k) = \dfrac{\#\text{ of cases when } S_r = k}{\#\text{ of all possible cases}}$

$$= \frac{\#\text{of cases when } (S_1 \cdots S_{r-1} \leq k-1) \text{ and } (S_{r+1}, \cdots S_n \geq k+1)}{\#\text{ of all possible cases}}$$

$$\textcircled{*} = \frac{\binom{k-1}{r-1} \cdot \binom{N-k}{n-r}}{\binom{N}{n}}$$

Also, # of all possible cases $= \sum_{k=r}^{N-(n-r)} \#\text{ of cases when } S_r = k$

$$= \sum_{k=r}^{N-(n-r)} \binom{k-1}{r-1}\binom{N-k}{n-r}$$

$$\textcircled{*} = \frac{(k-1)!}{(r-1)!\,(k-r)!} \cdot \frac{(N-k)!}{(n-r)!\,(N-k-n+r)!} \div \frac{n!\,(N-n)!}{N!}$$

$$= \frac{(k-1)!\,(N-k)}{(N-1)!} \cdot \frac{1}{N} \cdot \frac{n(n-1)!}{(r-1)!\,(n-r)!} \cdot \frac{(N-n)!}{(k-r)!\,(N-k-n+r)!}$$

$$= \frac{n}{N} \cdot \frac{\binom{n-1}{r-1}\binom{N-n}{k-r}}{\binom{N-1}{k-1}}$$

Question 6

Prove that the Smirnov test is a rank test. i.e that the test statistic is a function of only of the ranks

$$D_{m,n} = \sup_{-\infty<x<\infty} |F_m(x) - G_n(x)| = \max_{z_{(k)} \in S} |F_m(z_{(k)}) - G_n(z_{(k)})|$$

Let the Rank of treatment group be : $S_1 < S_2 \cdots < S_n$

The Rank of control group be : $R_1 < R_2 \cdots < R_m$

$S = z_{(1)} < z_{(2)} \cdots < z_{(m+n)}$. Each jump is either treatment group $(X_{(i)})$ or control group $(Y_{(j)})$, so $z_{(k)}$ is either $X_{(i)}$ or $Y_{(j)}$ for each k in $1, \ldots, m+n$

Let

$$D_{m,n} = \max_{z_{(i)} \in S} |F_m(z_{(i)}) - G_n(z_{(i)})|$$

$$= \max \left\{ \max_{X_{(i)}} |F_m(X_{(i)}) - G_n(X_{(i)})|, \max_{Y_{(j)}} |F_m(Y_{(j)}) - G_n(Y_{(j)})| \right\}$$

For each $X_{(i)}$: $F_m(X_{(i)}) = \frac{i}{n}$ and $G_n(X_{(i)}) = \frac{S_i - i}{m}$

For each $Y_{(j)}$: $G_n(Y_{(j)}) = \frac{j}{m}$ and $F_m(Y_{(j)}) = \frac{R_j - j}{n}$

We then have :

$$D_{m,n} = \max \left\{ \max_{S_i} \left|\frac{i}{n} - \frac{S_i - i}{m}\right|, \max_{R_j} \left| \frac{R_j - j}{n} - \frac{j}{m} \right| \right\}$$

$\Rightarrow$ $D_{m,n}$ is a function of ranks

Question 7

Prove that the Mann-Whitney statistic is symmetrically distributed about $\frac{mn}{2}$, under H_0

For any C,

$$\begin{aligned} & P_{H_0}(W_{xy}(m,n) = C) \\ = \; & P_{H_0}(W_s - \tfrac{1}{2}n(n+1) = C) \\ = \; & P_{H_0}(W_s - \tfrac{1}{2}n(N-m+1) = C) \\ = \; & P_{H_0}(W_s - \tfrac{1}{2}n(N+1) = C - \tfrac{mn}{2}) \end{aligned}$$

Also,

$$\begin{aligned} & P_{H_0}(W_{xy}(m,n) = C) \\ = \; & P_{H_0}(W_s - \tfrac{1}{2}n(n+1) = C) \\ = \; & P_{H_0}(-W_s + \tfrac{1}{2}n(N-m+1) = -C) \\ = \; & P_{H_0}(\tfrac{1}{2}n(N+1) - W_s = \tfrac{mn}{2} - C) \end{aligned}$$

We can conclude that

$$P_{H_0}(W_s - \tfrac{1}{2}n(N+1) = C) = P_{H_0}(\tfrac{1}{2}n(N+1) - W_s = C) \quad \text{for any } G$$

(Symmetry of the null distribution of W_s)

- Let m be the size of control group, We have that :

$$P_{H_0}(W_s - \tfrac{1}{2}n(m+n+1) = C) = P_{H_0}(\tfrac{1}{2}n(n+m+1) - W_s = C)$$

$$P_{H_0}(W_s - \tfrac{1}{2}n(n+1) = C + \tfrac{mn}{2}) = P_{H_0}(\tfrac{1}{2}n(n+1) - W_s = C - \tfrac{mn}{2}) :$$

$$P_{H_0}(W_s - \tfrac{1}{2}n(n+1) = \tfrac{mn}{2} + C) = P_{H_0}(W_s - \tfrac{1}{2}n(n+1) = \tfrac{mn}{2} - C)$$

$$P_{H_0}(W_{xy} = \tfrac{mn}{2} + C) = P_{H_0}(W_{xy} = \tfrac{mn}{2} - C) \quad \text{for } \forall C$$

$\Rightarrow$ Mann-Whitney statistic is symmetrically distributed about $\frac{mn}{2}$

Question 8

A student once asked me whether it is true that under $H_0 W_s$ is symmetrically distributed about $\frac{1}{2}n(N+1)$, implies immediately that $E_{H_0}(W_s) = \frac{1}{2}n(N+1)$, I replied that this may not be so, since a symmetric distribution may have a point of symmetry that is not the mean.
(a) Find an example to support my assertion.

For example, the cauchy distribution with its pdf:

$$f(x) = \frac{1}{b\pi\left[1 + \left(\frac{x-a}{b}\right)^2\right]}$$

Which is symmetric about a as $f(a-x) = f(a+x)$ for any x. However, cauchy distribution have a long tails on both ends, the means jump considerably as number of random points changing. The mean of Cauchy distribution is not defined.

(b) However, show that under an appropriate (simple) assumption that (applies to W_s) the symmetry of a distribution about a point, implies that this point must be the mean.

The simple assumption: $m, n \to \infty$

According to the CLT:

$$\lim_{m,n\to\infty} P_{H_0}\left(\frac{W_s - E_{H_0}(W_s)}{\sqrt{Var_{H_0}(W_s)}} \le x\right) = P(Z \le x), \quad \text{where } Z \sim N(0,1)$$

$$\Rightarrow \lim_{m,n\to\infty} P_{H_0}\left(W_s - E_{H_0}(W_s) \le x\right) = P(Z \le x), \text{ where } Z \sim N(0, Var_{H_0}(W_s))$$

$$\Rightarrow \lim_{m,n\to\infty} P_{H_0}\left(W_s \le x\right) = P(Z \le x), \text{ where } Z \sim N(E_{H_0}(W_s), Var_{H_0}(W_s))$$

It implies that as $m, n \to \infty$, we have that the distribution of W_s tends

to be normal distribution at any point x. As the symmetric point of normal distribution must be the mean. We can conclude that as $m, n \to \infty$, the symmetric point of W_S is exactly the mean $\frac{1}{2}n(N+1)$.