STAT 500 Midterm Exam

Vu Thi-Hong-Ha; NetID: 851924086

October 7, 2020

1 Question 1

$\sigma = 2.5$; $n_{drug} = n_1 = 6$; $\alpha = 0.05$; $\delta = 5$.

$\delta = 2t_{n_1+n_2-2;1-\alpha/2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

$\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = \frac{\delta}{2t_{n_1+n_2-2;1-\alpha/2} S_p}$

$\frac{n_1+n_2}{n_1 n_2} = \frac{\delta^2}{4(t_{n_1+n_2-2;1-\alpha/2})^2 S_p^2}$

First iteration using initial value $z_{1-\alpha/2} = 1.96$:

$\frac{6+n_2}{6n_2} = \frac{5^2}{4(1.96)^2(2.5)^2} \implies n_2 = 11$

Second iteration: $t_{n_1+n_2-2;1-\alpha/2} = t_{15,0.975} = 2.13$

$\frac{6+n_2}{6n_2} = \frac{5^2}{4(2.13)^2(2.5)^2} \implies n_2 = 19$

Third iteration: $t_{n_1+n_2-2;1-\alpha/2} = t_{23,0.975} = 2.07$

$\frac{6+n_2}{6n_2} = \frac{5^2}{4(2.07)^2(2.5)^2} \implies n_2 = 15$

Fourth iteration: $t_{n_1+n_2-2;1-\alpha/2} = t_{19,0.975} = 2.093$

$\frac{6+n_2}{6n_2} = \frac{5^2}{4(2.093)^2(2.5)^2} \implies n_2 = 17$

Fifth iteration: $t_{n_1+n_2-2;1-\alpha/2} = t_{21,0.975} = 2.08$

$\frac{6+n_2}{6n_2} = \frac{5^2}{4(2.08)^2(2.5)^2} \implies n_2 = 16$

Sixth iteration: $t_{n_1+n_2-2;1-\alpha/2} = t_{20,0.975} = 2.086$

$\frac{6+n_2}{6n_2} = \frac{5^2}{4(2.086)^2(2.5)^2} \implies n_2 = 16$

Number of mice that should be used in the control group is 16.

2 Question 2

(a)

- Treatments: 4 types of paints.

- Experimental units: 200-meter road segments.

- Response variable: the visibility of the paints.

- Replications: yes. Each type pf paints is observed on 5 road segments at 5 locations.

- Randomization: yes. Paint A and B are randomly assigned to 2 road segments at 5 locations, which are also randomly selected. The same procedure goes for C and D.

- Blocking: Blocking in locations of 5 locations used for paint A and B. The same blocking is used for paint C and D.

(b)

In the above design, paint A and B are grouped together and observed at the same locations, and paint C and D are grouped together. However, there is no information about A and B share any common characteristics (neither do C and D). Therefore, this design will still potentially have unwanted variations between locations while comparing the 4 paints together. Moreover, each location is 1600-meter long, which allows us to make more observations than this design does.

In my opinion, we can design an experiment that utilizes blocking with respect to locations. Within each location, divide the location into 8 segments, each of which is 200-meter long, and is adjacent to each other. Within these 8 segments, randomly assign 2 segments for paint A, 2 segments for paint B, 2 segments for paint C, and 2 segments for paint D. This way, we reduce the variations between different locations, and also have larger sample size (20 observations for each paint).

3 Question 3

Number of treatments $r = 4$, number of samples per treatment $n_1 = n_2 = n_3 = n_4 = 4$.

(a)

$$\text{Pooled estimate of variance} = \frac{\sum_{i=1}^{r=4}(n_i - 1)S_i^2}{\sum_{i=1}^{r=4} n_i - r} = \frac{3 * 2.3 + 3 * 2.4 + 3 * 1.8 + 3 * 2.3}{16 - 4} = 2.2$$

$$\bar{Y_{..}} = \frac{\sum_{i=1}^{r=4} n_i \bar{Y_{i.}}}{\sum_{i=1}^{r=4} n_i} = \frac{4 * (31 + 29 + 25 + 27)}{16} = 28$$

$$SS_{model} = \sum_{i=1}^{r=4} n_i(\bar{Y_i}. - \bar{Y_{..}})^2 = 4 * (31 - 28)^2 + 4 * (29 - 28)^2 + 4 * (25 - 28)^2 + 4 * (27 - 28)^2 = 80$$

$$SS_{error} = \sum_{i=1}^{r=4}(n_i - 1)S_i^2 = 3 * 2.3 + 3 * 2.4 + 3 * 1.8 + 3 * 2.3 = 26.4$$

ANOVA table:

Source of variation	Degrees of freedom	Sums of squares	Mean square	F Value	Pr > F
Model	$r-1=3$	$SS_{model}=80$	$\frac{SS_{model}}{r-1}=26.6667$	$\frac{MS_{model}}{MS_{error}}=12.12$	0.00061
Error	$N-r=12$	$SS_{error}=26.4$	$\frac{SS_{error}}{N-r}=2.2$		
Total	$N-1=15$	$SS_{total}=106.4$			

(b)

Denote μ_A to be the true mean of bean pod yields under treatment A, μ_B to be the true mean of bean pod yields under treatment B, μ_C to be the true mean of bean pod yields under treatment C, μ_D to be the true mean of bean pod yields under treatment D.

$H_0 : \mu_A = \mu_B = \mu_C = \mu_D$

H_a : At least one group mean is different from the rest.

(c)

Test statistic: $F = \frac{MS_{model}}{MS_{error}} = 12.12$

The F statistic has a central F-distribution with 3 numerator and 12 denominator degrees of freedom.

$P - value = P(F_{3;12} > F) = 0.00061$

At a significance level of 0.05, we reject the null that four group means are equal.

(d)

Assumptions: Data are normally distributed and samples are independent. Under these assumptions and $H_0 : \mu_A = \mu_B = \mu_C = \mu_D$:

$(N-r)MS_{error}/\sigma^2 \sim \chi^2_{N-r}$

$(r-1)MS_{model}/\sigma^2 \sim \chi^2_{r-1}$

MS_{error} and MS_{model} are independent.

$F = \frac{((r-1)MS_{model}/\sigma^2)/(r-1)}{((N-r)MS_{error}/\sigma^2)/(N-r)} = \frac{MS_{model}}{MS_{error}}$ has a central F-distribution with $r-1$ numerator and $N-r$ denominator degrees of freedom.

(e)

Consider the contrast $\gamma = \frac{\mu_B + \mu_C + \mu_D}{3} - \mu_A$.

$H_0 : \gamma = 0$

$H_a : \gamma \neq 0$

$\hat{\gamma} = \sum_{i=1} c_i \bar{Y_{i.}} = -31 + \frac{1}{3} * 29 + \frac{1}{3} * 25 + \frac{1}{3} * 27 = -4$

$MS_{error} = 2.2$

Then $S_{\hat{\gamma}} = \sqrt{2.2 * (\frac{1}{4} + \frac{1/9}{4} + \frac{1/9}{4} + \frac{1/9}{4})} = 0.8563$

Also, $N - r = 12$, $1 - \alpha/2 = 0.975$ if we choose $\alpha = 0.05$. Then, $t_{12,0.975} = 2.1788$.

100(1-α)% confidence intervals: $\hat{\gamma} \pm t_{N-r,1-\alpha/2} S_{\hat{\gamma}} = (-4 - 2.1788 * 0.8563; -4 + 2.1788 * 0.8563) = (-5.8657; -2.1343)$

0 is not in $(-5.8657; -2.1343)$, so at a significance level of 0.05, we reject the null that this contrast is 0, or there

is no difference in the mean of different treatments.

If we carry out one sided test with $H_0 : \gamma = 0$ against $H_a : \gamma < 0$, we can obtain $p - value = 0.00088$, which also leads us to reject the null at a significance level of 0.05, and we can conclude that mean yields under treatment B, C and D are smaller that yields without pollution.

(f)

$HSD = \frac{1}{\sqrt{2}} q_{(4,12,0.95)} \sqrt{MS_{error}(\frac{2}{n})} = \frac{1}{\sqrt{2}}(4.20)\sqrt{2.2(\frac{2}{4})} = 3.115$

Order sample means in increasing order:

C	D	B	A
25	27	29	31

For an experiment-wise error rate of 5%, the pairs of means (C, B), (C, A), (D, A) are significantly different.

4 Question 4

(a)

The experiment is a randomized complete block experiment because units within each block are similar; within each block, the treatments are randomly assigned to the units so that one unit is for one treatment. Also, the number of units in each block is the same as the number of treatments.

Treatments: two timing schedule. Number of treatments: $J = 2$.

Response variable: the stem tissue nitrate amount.

Blocks: parts of an irrigated field. Number of blocks $n = 4$.

(b)

Experimental units: land plots. Number of units per block $J = 2$.

(c)

(i) Using sign test.

(ii)

H_0 : The median difference is equal to zero.

H_a : The median difference is different from zero.

Block	|Difference (Schedule 1 - Schedule 2)|	Sign
1	|-2.2|	−
2	|-4.63|	−
3	|-3.29|	−
4	|0.77|	+

$S = 1$ positive difference out of $n = 4$ pairs.

p-value is obtained from the binomial distribution. $p - value = 2 * [(\binom{4}{0} + \binom{4}{1}) * 0.5^4] = 0.625$.

(iii) At a significance level of 0.05, we fail to reject the null that the median difference is equal to zero.

5 Question 5

(a) This study is an experiment because some patients received a form of treatment (therapy), and some patients did not.

(b)

We are interested in finding whether group therapy treatment is effective or not in increasing survival time.

To decide on which analysis to use, we need to check on assumptions first:

Independence: We have the information that the patients were randomly assigned to each treatment. Moreover, it can be safe to assume that each individual is independent as there is no information if the patients are biologically related, and no person received both treatments.

Homogeneous variance: We may test if the two populations have the same variance by using graphical methods, ratio of sample standard deviations, folded F-test for equality of variances, or Brown-Forsythe test. As Brown-Forsythe test is more objective (it does not depend on our perception like graphs or previous study-based thresholds like ratio test), and it is not sensitive to departures from normal distributions, I will use Brown-Forsythe test here.

GLM results for Difference in Survival Time

The GLM Procedure

Brown and Forsythe's Test for Homogeneity of Months Variance
ANOVA of Absolute Deviations from Group Medians

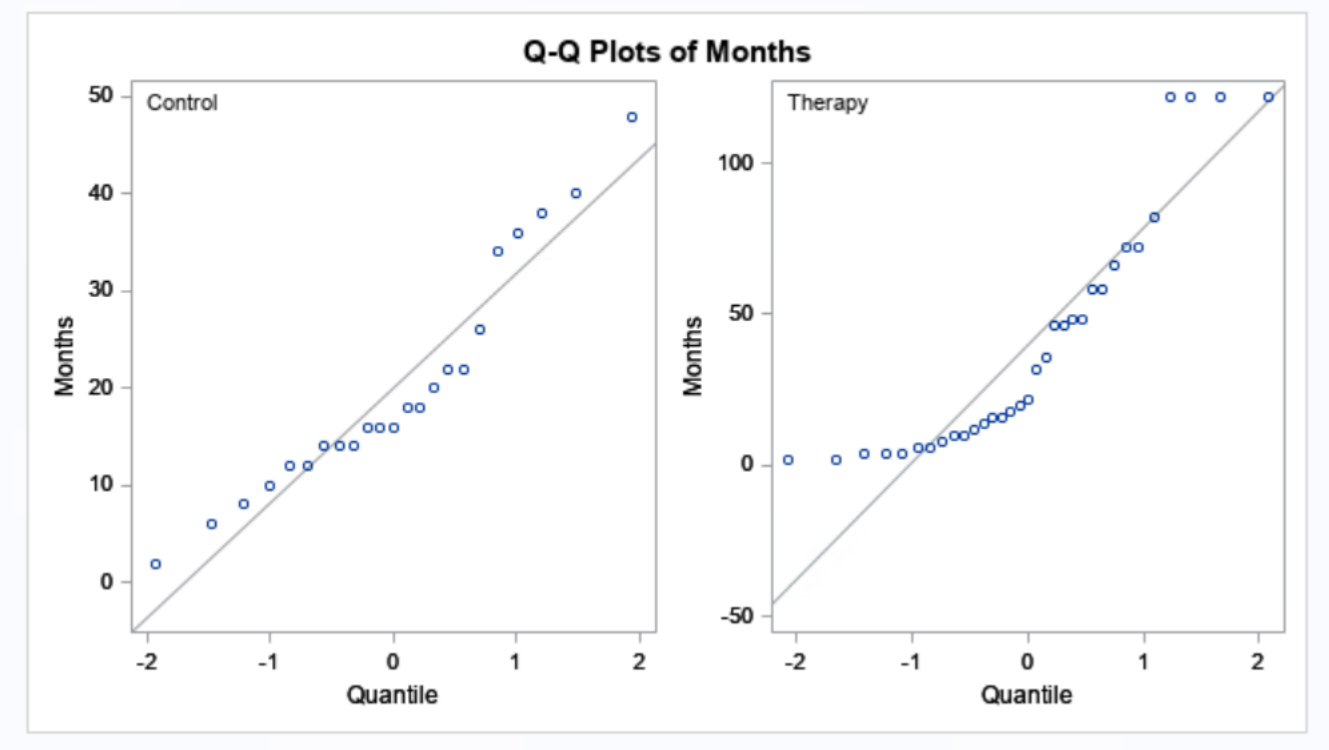
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Group	1	6378.9	6378.9	11.34	0.0014
Error	54	30388.4	562.7		

At a significance level of 0.05, we reject the null that the two populations have homogeneous variance.

Normality: We can use graphical methods or tests like Shapiro-Wilk test to check this assumption:

Treatment	Shapiro-Wilk Statistic	p-value
Control	0.919472	0.0649
Therapy	0.837274	0.0002

For group Control, the S-W test statistic is small, but the p-value is greater than 0.05, so at a significance level of 0.05, we fail to reject the null that the Control group follows normal distribution. However, this p-value indicates a weak evidence of the null. For Therapy group, at a significance level of 0.05, we reject the null that this group follows normal distribution as the p-value is smaller than 0.05.



Looking at the Q-Q plot above, we can see in Control group, the data appeared to be heavy tailed and right skewed; while in Therapy group, the plot clearly shows the data does not follow normal distribution.

As only the assumption of independence holds, I will carry out non-parametric test to study the research question. In this case, I will use Wilcoxon rank sum test.

(c)

H_0 : The two populations have the same distribution.

H_a : The two populations have the different distribution.

Assumption: Samples are random and independent.

Test statistics:

Treatment	Wilcoxon Rank Statistic
Control	585.0
Therapy	1011.0

Let W be the sum of ranks for Control group. Then $W = 585.0$.

$$E_0(W) = \frac{n_1(n_1+n_2+1)}{2} = \frac{23(23+33+1)}{2} = 655.5$$

$$V_0(W) = \frac{n_1 n_2(n_1+n_2+1)}{12} = \frac{23*33(23+33+1)}{12} = 3605.25$$

$$z = \frac{|W - E_0(W)| - 0.5}{\sqrt{V_0(W)}} = \frac{|585 - 655.5| - 0.5}{\sqrt{3605.25}} = 1.1658$$

Wilcoxon Two-Sample Test					
				t Approximation	
Statistic (S)	Z	Pr < Z	Pr > |Z|	Pr < Z	Pr > |Z|
585.0000	-1.1674	0.1215	0.2431	0.1240	0.2481
Z includes a continuity correction of 0.5.					

p-value is 0.2431 $>$ 0.05, so at the significance level of 0.05, we fail to reject the null that the survival time distributions are the same between two treatment groups.

(d)

We observe the survival times of breast cancer patients who were randomly assigned to either a control group or a treatment group. The data of the two treatments are random and independent. However, these two populations do no have homogeneous variance. We have weak evidence that the control group follows normal distribution (p-value = 0.0649, significance level = 0.05), but we reject the null that the therapy group follows normal distribution (p-value = 0.0002, significance level = 0.05). To study whether group therapy treatment is effective or not in increasing survival time, we test the null hypothesis that the two populations have the same distribution, against the alternative hypothesis that the two populations do not have the same distribution. We carry out a non-parametric test, Wilcoxon rank sum test, which only assumes independence between samples. By Wilcoxon rank sum test, we fail to reject the null that the two populations have the same distribution (p-value = 0.2431, significance level = 0.05).