ST3300 Data Analysis 1 : Assignment 1

Student No: 120224791

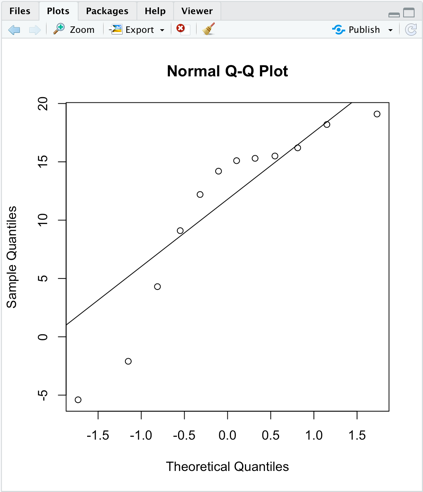
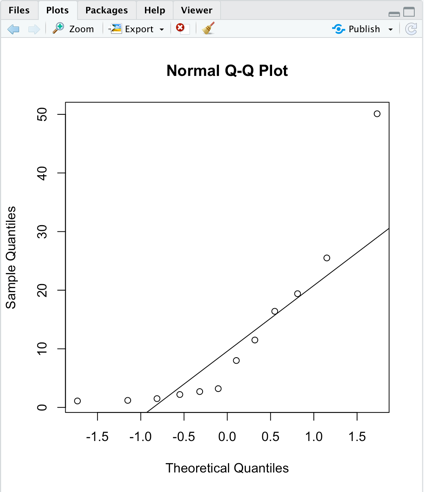
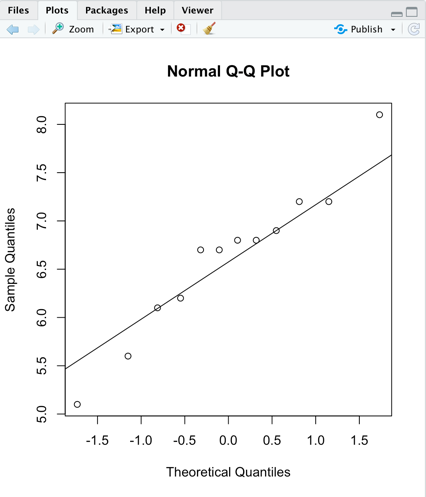
Name: James Henry Hehir

**Question 1:**

**A)**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Mean** | **Median** | **Variances** | **Standard Deviation** | **Interquartile Range** |
| **sample1** | 6.616667 | 6.75 | 0.619697 | 0.7872083 | 0.8 |
| **sample2** | 11.9 | 5.6 | 211.7982 | 14.55329 | 15.125 |
| **sample3** | 10.975 | 14.65 | 63.7075 | 7.981698 | 7.775 |

**B) sample1 sample2 sample3**



**> shapiro.test(sample1)**

**Shapiro-Wilk normality test**

**data: sample1**

**W = 0.95833, p-value = 0.7598**

**> shapiro.test(sample2)**

**Shapiro-Wilk normality test**

**data: sample2**

**W = 0.7687, p-value = 0.004219**

**> shapiro.test(sample3)**

**Shapiro-Wilk normality test**

**data: sample3**

**W = 0.84158, p-value = 0.02894**

The graphs above are calculated using qqnorm() which plots the sample quantiles (percentiles) against those of a standard normal distribution with qqline() which fits a best fit line. You can see that sample1 was clearly normally distributed while sample 2 and 3 gave a few suspicions that they may not be normally distributed. To check numerically we perform a shapiro test on each sample and the answers are printed above. In this test if the p value is above 5% we can say that there is no evidence that the sample is not normally distributed. The low p-values for sample 2 and 3 would suggest to me that they are not normally distributed.

**C) > t.test(sample2, mu = 5, alternative = "greater", conf.level=0.90)**

**One Sample t-test**

**data: sample2**

**t = 1.6424, df = 11, p-value = 0.06438**

**alternative hypothesis: true mean is greater than 5**

**90 percent confidence interval:**

**6.171994 Inf**

**sample estimates:**

**mean of x**

**11.9**

The alternative or research hypothesis is H1: mu>5, where with mu we denote the mean of sampe2.#(also, Ho: mu<=5) so we want an upper rejection region i.e. set alternative="greater”. Since our p value is less than 10% (0.06438) we reject the null hypothesis.

**D)** > **t.test(sample1, sample2, paired=F)**

**Welch Two Sample t-test**

**data: sample1 and sample2**

**t = -1.2557, df = 11.064, p-value = 0.2351**

**alternative hypothesis: true difference in means is not equal to 0**

**95 percent confidence interval:**

**-14.537000 3.970333**

**sample estimates:**

**mean of x mean of y**

**6.616667 11.900000**

H1: sample1 = sample2

H0: sample1> sample2

P-value approach suggests that we would accept the null hypothesis as the p-value(0.2351) is larger than 5%. The two sample t test suggests that two means are not the same.

**E) > t.test(sample2, sample3, conf.level=0.90, paired=T)**

**Paired t-test**

**data: sample2 and sample3**

**t = 0.21383, df = 11, p-value = 0.8346**

**alternative hypothesis: true difference in means is not equal to 0**

**90 percent confidence interval:**

**-6.843775 8.693775**

**sample estimates:**

**mean of the differences**

**0.925**

The null hypothesis suggests that the true mean difference is equal to zero. And using the p-value for the test above we can accept the null hypothesis as the p-value is greater than 10%.

**F)** In testing (c),(d) and (e) they were causes for concern as the t-test that were carried out on sample2 and sample3 would suggest there is a problem with the relationship of these data samples. This was a suspicion in (b) that the data from sample2 and sample3 was not normally distributed and in the further testing in the rest of the questions it would suggest these data sets do not fit the model.

**Question 2**

**A) > wilcox.test(sampleA, mu=42, correct=F)**

**Wilcoxon signed rank test**

**data: sampleA**

**V = 25.5, p-value = 0.7209**

**alternative hypothesis: true location is not equal to 42**

**> wilcox.test(sampleA, mu = 42, conf.level=0.10,correct=F)**

**Wilcoxon signed rank test**

**data: sampleA**

**V = 25.5, p-value = 0.7209**

**alternative hypothesis: true location is not equal to 42**

H0: sampleA>42 for both cases. Since in both cases the p-value is greater than 5% and 10% respectively, we accept the null hypothesis.

**B) > wilcox.test(sampleA,sampleB, correct = F)**

**Wilcoxon rank sum test**

**data: sampleA and sampleB**

**W = 71, p-value = 0.1113**

**alternative hypothesis: true location shift is not equal to 0**

The null hypothesis is that sampleA’s median is equal to that of the population underlying sample. Using the p-value approach we accept the null hypothesis because the p-value (0.1113) is greater than the level of significance which is 5%.

**C) > wilcox.test(sampleA,sampleB, m=0, conf.level=0.95, correct=F,paired = T)**

**Wilcoxon signed rank test**

**data: sampleA and sampleB**

**V = 48.5, p-value = 0.03153**

**alternative hypothesis: true location shift is not equal to 0**

The null hypothesis suggests that the median distance to wear-out for BrandA is equal to that of BrandB. Using the p value approach, we would reject the null hypothesis as the p-value is less than 5% level of significance. The test statistic is shown by V which is 48.5.

For the critical value approach

**D)** The effect of changing the significance level in part a had no overall effect on either output. In changing from independent to dependent samples in part b and c we saw a decrease in the p-value making the second test a better fit of the data to the model.

**Question 3**

**A)**

|  |  |
| --- | --- |
| **One way Parametric ANOVA** | **One way Nonparametric ANOVA** |
| All populations are normally distributed | Assume all populations of interest have a broadly similar shape |
| We sample randomly from each population | Assume we sample randomly from each population of interest |
| All populations have the same variance | Assume all populations of interest have the same variance |
| **Advantages** |  |
| Robust design | Assess median which can be better in certain studies |
| Increases Statistical power | Good for small sample sizes |
| Trustworthy results | Can analyze ordinal data, ranked data and outliers |
| **Disadvantages** |  |
| If the median is a better measure, nonparametric will be a better fit | Have other firm assumptions that can be harder to meet |

**B) > av1**

**Call:**

**aov(formula = response ~ sample, data = st\_eg1)**

**Terms:**

**sample Residuals**

**Sum of Squares 93.56933 130.10000**

**Deg. of Freedom 2 12**

**Residual standard error: 3.292669**

**Estimated effects may be unbalanced**

**> summary(av1)**

**Df Sum Sq Mean Sq F value Pr(>F)**

**sample 2 93.57 46.78 4.315 0.0387 \***

**Residuals 12 130.10 10.84**

**---**

**Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1**

The p-value 0.0387 is less than 5% so there is some effect on the means.

I perform a Bonferroni confidence interval on the data and the result are below:

**> PostHocTest(av1, method = "bonferroni");**

**Posthoc multiple comparisons of means : Bonferroni**

**95% family-wise confidence level**

**$sample**

**diff lwr.ci upr.ci pval**

**sampleB-sampleA 5.46 -0.3281596 11.24816 0.0669 .**

**sampleC-sampleA 5.12 -0.6681596 10.90816 0.0903 .**

**sampleC-sampleB -0.34 -6.1281596 5.44816 1.0000**

**---**

**Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1**

**C)**

**i)** Overall there is 3 factors which are Delivery, Regular and Block**.** Delivery has 2 factors, Regular has 2 factors and Block has 2 factors. There is 8 treatment groups present in the study design. I would say that the study has a well-balanced design as each treatment group has 5 points.

**ii)** There are two factors each with two levels. Type and Delivery and their interaction is a major source of variance.

**iii) > summary(anova2, conf.level=0.10)**

**Df Sum Sq Mean Sq F value Pr(>F)**

**delivery 1 10.0 10.0 0.588 0.4480**

**type 1 14.4 14.4 0.847 0.3634**

**delivery:type 1 324.9 324.9 19.118 0.0001 \*\*\***

**Residuals 36 611.8 17.0**

**---**

**Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1**

H0,1: Factors Delivery and Form(Regular/Block) don􏰝t interact –

H1,1: Factors Delivery and Form interact

H0,2: no main effect due to Delivery - H1,2: main effect due to Delivery

H0,3: no main effect due to Form - H1,3: main effect due to Form

There seems to be an effect because the p value is quiet low in delivery:form.

**> tk\_anova2 <-TukeyHSD(anova2)**

**> tk\_anova2**

**Tukey multiple comparisons of means**

**95% family-wise confidence level**

**Fit: aov(formula = output ~ delivery \* type, data = uniResults)**

**$delivery**

**diff lwr upr p adj**

**r-b -1 -3.643879 1.643879 0.4480296**

**$type**

**diff lwr upr p adj**

**t-d 1.2 -1.443879 3.843879 0.3634347**

**$`delivery:type`**

**diff lwr upr p adj**

**r:d-b:d 4.7 -0.265254 9.665254 0.0690763**

**b:t-b:d 6.9 1.934746 11.865254 0.0033990**

**r:t-b:d 0.2 -4.765254 5.165254 0.9995342**

**b:t-r:d 2.2 -2.765254 7.165254 0.6349427**

**r:t-r:d -4.5 -9.465254 0.465254 0.0873357**

**r:t-b:t -6.7 -11.665254 -1.734746 0.0045861**

**Question 4**

**A)** N= -29, N=25

N= total points, N=4n

**n=7**

**B)**

H0=m1=m2=m3=m4

H=12/28\*29 (60^2/7 + 60^2/7 + 122^2/7 + 164^2/7) – 3(29)

H= 153.422

**> kruskal.test(list(SA,SB,SC,SD))**

**Kruskal-Wallis rank sum test**

**data: list(SA, SB, SC, SD)**

**Kruskal-Wallis chi-squared = 3, df = 3, p-value = 0.3916**

**> qchisq(0.95,3)**

**[1] 7.814728**

H>7.814728 meaning we reject the null hypothesis.