Final Exercises

C. Durso

## Instructions

Please work these problems on your own. You may use web searches, but not interactive methods such as asking others online or in person.

Please complete the questions on this template and upload your solutions in a single knitted Word or pdf document. Please also upload your completed template.

In light of the exam context, the data sets for the questions have been generated clearly to satisfy or obviously to violate the requirements of the statistical procedures. If reasonable exploratory analysis is done, there should be little ambiguity as to whether the given data satisfy the requirements. This is unrealistic, but less stressful for students and graders alike.

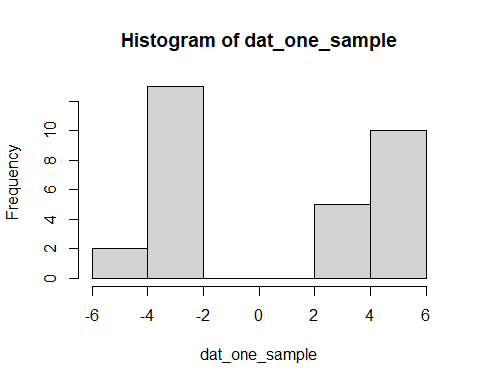
## Questions

1. Student’s t, one sample

load("dat\_one\_sample.RData")

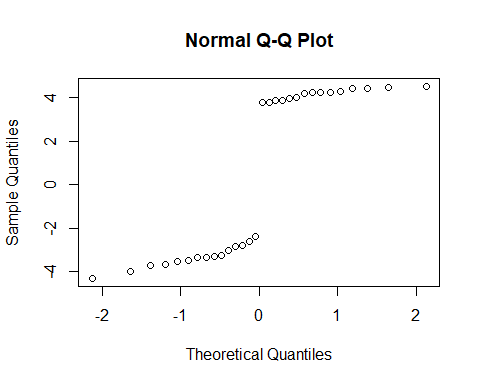
1.a. Please make a histogram of dat\_one\_sample with informative bin widths. (2 points)

hist(dat\_one\_sample)



1.b. Please generate a Normal qq plot for dat\_one\_sample.(2 points)

qqnorm(dat\_one\_sample)



1.c. Please perform a Student’s t-test of the null hypothesis that dat\_one\_sample is drawn from a Normal population with mean equal to 1. Report the 90% (not 95%) confidence interval for the mean. Please do this whether or not your work in 1.a and 1.b indicates that the hypotheses making the one sample Student’s test a test of location of the mean are satisfied. (3 points)

t.test(dat\_one\_sample,  
 mu = 1,  
 conf.level = 0.9)

##   
## One Sample t-test  
##   
## data: dat\_one\_sample  
## t = -0.82763, df = 29, p-value = 0.4146  
## alternative hypothesis: true mean is not equal to 1  
## 90 percent confidence interval:  
## -0.7595529 1.6068837  
## sample estimates:  
## mean of x   
## 0.4236654

1.d. Considering your work in 1.a and 1.b, how do you interpret the results (p-value and confidence interval) in 1.c? (3 points)

While the sample size of 30 is sufficiently large, the normality assumption is flagrantly violated. The histogram is not unimodal, but two clusters with a gap in between. Naturally the test can still be run and the confidence interval generated, but the results are not especially meaningful. If the normality had not been violated, the p-value of .4146 would not be sufficiently low enough to reject the null hypothesis. It would be quite possible that 1 is the true mean. As further evidence, 1 is inside the confidence interval.

1. Wilcoxon signed rank

2.a. Please perform a Wilcoxon signed rank test of the null hypothesis that dat\_one\_sample is drawn from a population symmetric around its mean with mean equal to 1. (5 points)

wilcox.test(dat\_one\_sample,mu=1)

##   
## Wilcoxon signed rank exact test  
##   
## data: dat\_one\_sample  
## V = 124, p-value = 0.02479  
## alternative hypothesis: true location is not equal to 1

2.b. Considering your work in 1.a and 1.b, how do you interpret the results in 2.a? (5 points)

Since the data set is more or less symmetric, the assumptions for the Wilcoxon signed rank test are met, as there is no normality presumption. Given the p-value of .02479, we would reject the null hypothesis at the alpha=.05 level. There is statistical evidence that the true location of mu is not = 1.

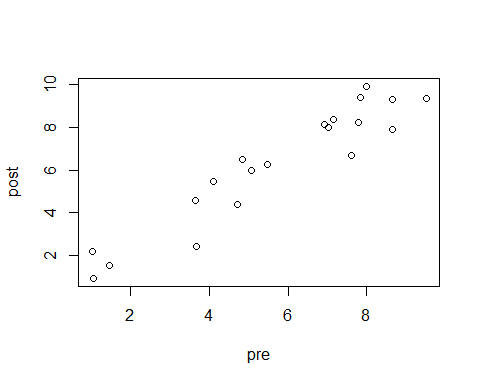
1. The data set dat\_pre\_post simulates pre-intervention measurements for 20 individuals together with their post-intervention measurements. Carry out the most powerful test we have learned of the null hypothesis that the intervention is not associated with any systematic increase or decrease in the measurement. Please justify your choice of test. (Note that to have evidence that any change was caused by the intervention, a controlled experiment would be required.) To help you decide which test to use, please generate a scatter plot of the pre values against the post values. (10 points)

Looking at the scatterplot as well as the normal qqplots and histograms for the pre and post samples, it looks as though the paired samples Wilcoxon Test would be the most appropriate, as well as the most powerful. We will conduct a two sided test to maximize its power. Given the p-value of .01923, we would reject the null hypothesis at the alpha=.05 level. There is statistical evidence to suggest a statistically significant difference between the pre and post observations.

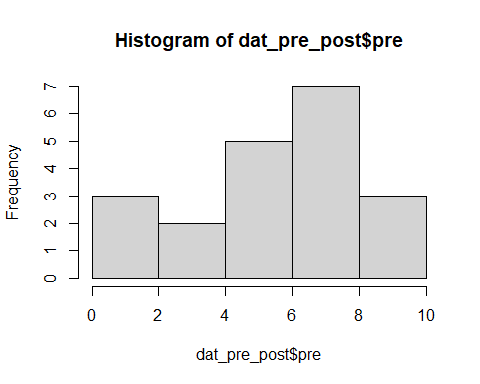
load("dat\_pre\_post.RData")  
dat\_pre\_post

## pre post  
## 1 7.785807 8.2474490  
## 2 7.618600 6.6725355  
## 3 8.653174 9.3144843  
## 4 3.653029 4.5814522  
## 5 5.071998 5.9621342  
## 6 7.980290 9.9253405  
## 7 8.654312 7.9064772  
## 8 3.673627 2.3893749  
## 9 1.020867 2.1852386  
## 10 7.024605 8.0020798  
## 11 4.721275 4.3643713  
## 12 5.463265 6.2380824  
## 13 1.453562 1.5214667  
## 14 4.109938 5.4593081  
## 15 9.506002 9.3546408  
## 16 6.911345 8.1579171  
## 17 7.143192 8.3701818  
## 18 4.842457 6.4780037  
## 19 1.065420 0.9123392  
## 20 7.834550 9.3966180

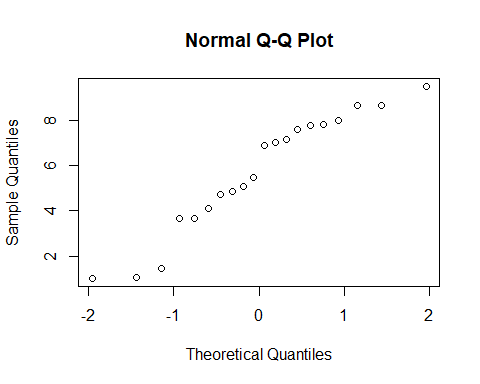
plot(dat\_pre\_post)



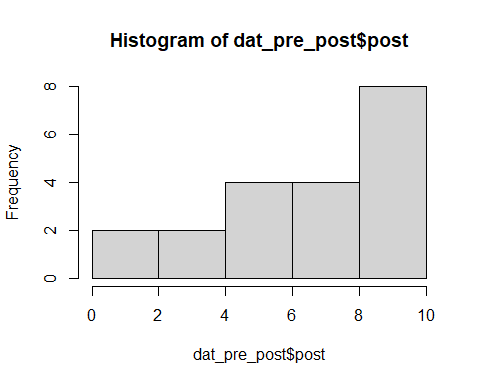
hist(dat\_pre\_post$pre)



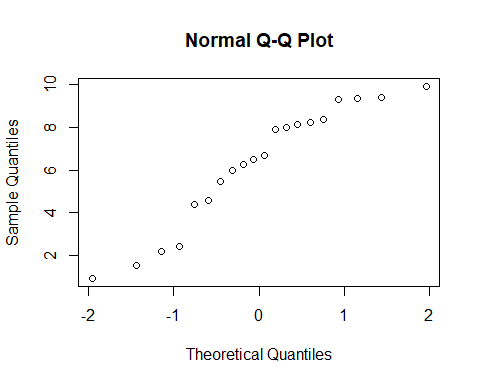
qqnorm(dat\_pre\_post$pre)



hist(dat\_pre\_post$post)



qqnorm(dat\_pre\_post$post)



wilcox.test(dat\_pre\_post$pre, dat\_pre\_post$post, paired = TRUE, alternative = "two.sided")

##   
## Wilcoxon signed rank exact test  
##   
## data: dat\_pre\_post$pre and dat\_pre\_post$post  
## V = 43, p-value = 0.01923  
## alternative hypothesis: true location shift is not equal to 0

1. The data dat\_two\_sample simulate independent, identically distributed samples from a population with the samples from in the “val” column, labeled with “gp”=“x” and independent, identically distributed samples from a population with the distribution in the “val” column, labeled with “gp”=“y”

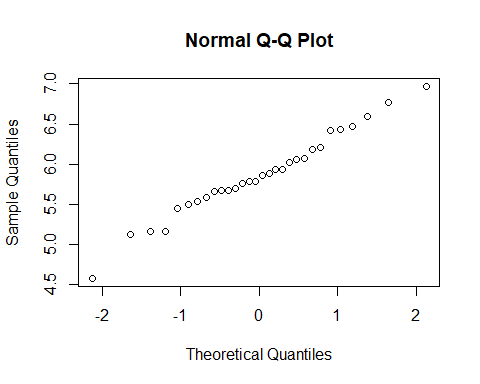
load("dat\_two\_sample.RData")  
dat\_two\_sample

## gp val  
## 1 x 5.670434  
## 2 x 5.587114  
## 3 x 6.596941  
## 4 x 5.667562  
## 5 x 6.066487  
## 6 x 4.580501  
## 7 x 5.940727  
## 8 x 5.657193  
## 9 x 6.182293  
## 10 x 5.757841  
## 11 x 6.968209  
## 12 x 5.787352  
## 13 x 5.165130  
## 14 x 5.124298  
## 15 x 6.420755  
## 16 x 5.698060  
## 17 x 5.449311  
## 18 x 6.017764  
## 19 x 6.765347  
## 20 x 5.936359  
## 21 x 5.542808  
## 22 x 5.497695  
## 23 x 5.882798  
## 24 x 5.865071  
## 25 x 6.431611  
## 26 x 5.783280  
## 27 x 6.056209  
## 28 x 6.470350  
## 29 x 6.212378  
## 30 x 5.166367  
## 31 y 8.237196  
## 32 y 5.997303  
## 33 y 5.547691  
## 34 y 5.301229  
## 35 y 4.798721  
## 36 y 2.525102  
## 37 y 7.699789  
## 38 y 8.510127  
## 39 y 6.833164  
## 40 y 8.225090  
## 41 y 6.208186  
## 42 y 5.933597  
## 43 y 5.357341  
## 44 y 4.069184  
## 45 y 3.883922  
## 46 y 7.756473  
## 47 y 3.710417  
## 48 y 7.041952  
## 49 y 7.877799  
## 50 y 4.689452  
## 51 y 6.682377  
## 52 y 5.970157  
## 53 y 9.670289  
## 54 y 8.104578  
## 55 y 8.882708  
## 56 y 8.293125

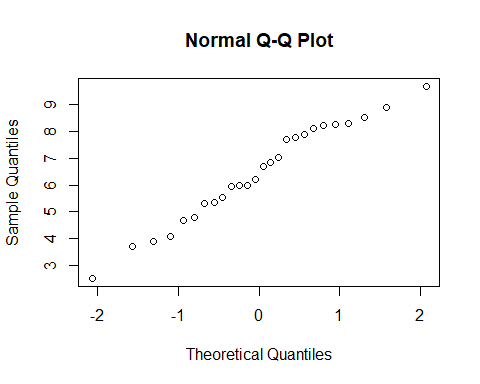
4.a.Please visually assess the Normality of the x’s and the y’s. (5 points)

Separately looking at the normal qq plots of the x and y values, the x values theoretical and sample quantiles line up nicely. The y value qq plot is not as straight, as there are a cluster of values to the right of the theoretical quantile of 0. We could feel more comfortable with the presumption of normality of the x values than the y values.

xvals <- dat\_two\_sample$val[1:30]  
yvals <- dat\_two\_sample$val[31:56]  
qqnorm(xvals)

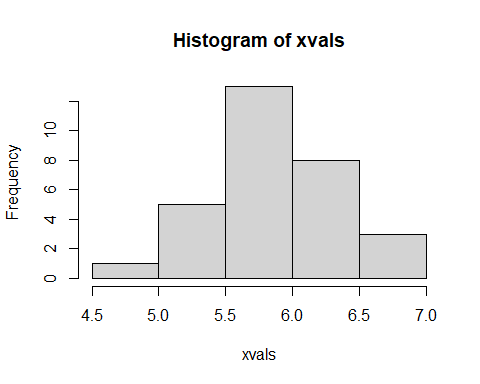


qqnorm(yvals)

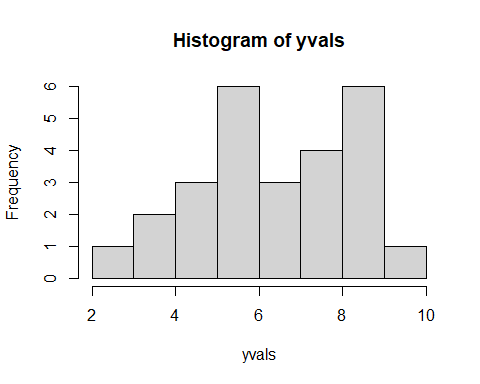


4.b. Please display histograms or density plots of the x’s and the y’s. (5 points)

hist(xvals)



hist(yvals)



4.c. Please carry out an F test of the equality of the variance of the x’s and y’s.(5 points)

var.test(xvals, yvals, ratio = 1)

##   
## F test to compare two variances  
##   
## data: xvals and yvals  
## F = 0.079381, num df = 29, denom df = 25, p-value = 1.828e-09  
## alternative hypothesis: true ratio of variances is not equal to 1  
## 95 percent confidence interval:  
## 0.03624478 0.17002503  
## sample estimates:  
## ratio of variances   
## 0.07938087

4.d. Please carry out Welch’s test of the null hypothesis that the means of x and y are equal. Please interpret the result using the work in 4.a-4.c. (5 points)

Given the results of the F-test of equality of variance in 4c, there is overwhelming statistical evidence that the two samples come from distributions with different variances. Because of this, Welch’s t-test is the appropriate test for the difference of means of the two samples. Given the p-value of .1238, we cannot reject the null hypothesis, that the difference in means is = 0. There is not enough statistical evidence to suggest that the two samples come from distributions with different means.

t.test(xvals,yvals)

##   
## Welch Two Sample t-test  
##   
## data: xvals and yvals  
## t = -1.5858, df = 28.442, p-value = 0.1238  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -1.349688 0.171343  
## sample estimates:  
## mean of x mean of y   
## 5.864942 6.454114

1. Please carry the Mann Whitney U test on x and y. Please interpret the result using the work in 4.a-4.c.(10 points)

Most of the assumptions for the Mann Whitney U test are met, specifically independent groups, and independence of observations. There is a major assumption violation however, with the difference in variances as well as the observable difference in shape of the samples. Therefore we would defer to the results of the Welch two sample t-test in 4d. That said, the tests generate similar p-values, which would lead to similar conclusions. There would not be rejection of the null hypothesis that the true difference of means=0.

wilcox.test(xvals,yvals)

##   
## Wilcoxon rank sum exact test  
##   
## data: xvals and yvals  
## W = 298, p-value = 0.1336  
## alternative hypothesis: true location shift is not equal to 0

1. Please carry out a test of independence on the contingency table in “mat”. Please interpret the results.(10 points)

Given the p-value of .01459, we reject the null hypothesis at the alpha = .05 level, that the categorical variables are independent. There is statistical evidence to suggest that the variables are not independent.

load("mat.Rdata")  
chisq.test(mat)

##   
## Pearson's Chi-squared test  
##   
## data: mat  
## X-squared = 8.4554, df = 2, p-value = 0.01459

1. Please carry out Fisher’s exact test on “mat” and interpret the results. (10 points)

Given the p-value of .009642, we would reject the null hypothesis at the .05 or .01 levels, that the row and column variables are independent. There is strong statistical evidence that the row and column variables are not independent.

fisher.test(mat)

##   
## Fisher's Exact Test for Count Data  
##   
## data: mat  
## p-value = 0.009642  
## alternative hypothesis: two.sided

8.a. Please fit a linear regression model giving post as a linear function of pre from the dat\_pre\_post data set. Display the coefficients with their p-values. (5 points)

prepostmod <- lm(post ~ pre, data=dat\_pre\_post)  
summary(prepostmod)

##   
## Call:  
## lm(formula = post ~ pre, data = dat\_pre\_post)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.8109 -0.6704 0.2777 0.6682 1.3423   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.46163 0.51651 0.894 0.383   
## pre 1.01769 0.08264 12.315 3.32e-10 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.94 on 18 degrees of freedom  
## Multiple R-squared: 0.8939, Adjusted R-squared: 0.888   
## F-statistic: 151.7 on 1 and 18 DF, p-value: 3.321e-10

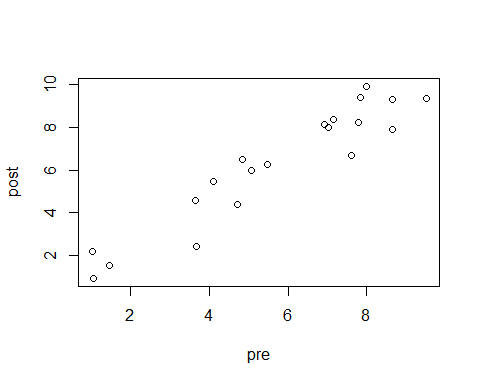
8.b. Please use your work in question 3 and any additional plots you find informative to address the validity of the model, the coefficients, and the p-values. Please address the linearity of the relation between pre and post, the Normality of the residuals, and the issue of influential observations.(15 points)

There are several factors which call into question the linear model that is generated for the pre and post observation data. One is the clear lack of normality of the residuals. The normal qq plot does not follow a straight line, and a histogram of the residuals shows a lack of symmetry and unimodality. There are also several points from the data set which are exerting a disproportionate influence on the model. For example, the point at approximately (3.67,2.39) as well as the point located at (7.62,6.67). These are helping cause a high standard error and consequently high p-value for the y- intercept, speaking to a lack of certainty of it’s true location.

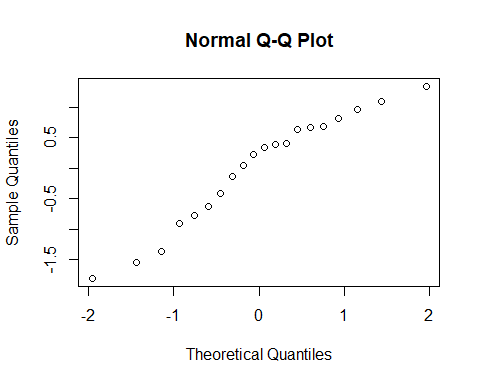
dat\_pre\_post

## pre post  
## 1 7.785807 8.2474490  
## 2 7.618600 6.6725355  
## 3 8.653174 9.3144843  
## 4 3.653029 4.5814522  
## 5 5.071998 5.9621342  
## 6 7.980290 9.9253405  
## 7 8.654312 7.9064772  
## 8 3.673627 2.3893749  
## 9 1.020867 2.1852386  
## 10 7.024605 8.0020798  
## 11 4.721275 4.3643713  
## 12 5.463265 6.2380824  
## 13 1.453562 1.5214667  
## 14 4.109938 5.4593081  
## 15 9.506002 9.3546408  
## 16 6.911345 8.1579171  
## 17 7.143192 8.3701818  
## 18 4.842457 6.4780037  
## 19 1.065420 0.9123392  
## 20 7.834550 9.3966180

plot(dat\_pre\_post)



prepostres <- resid(prepostmod)  
qqnorm(prepostres)



hist(prepostres)

