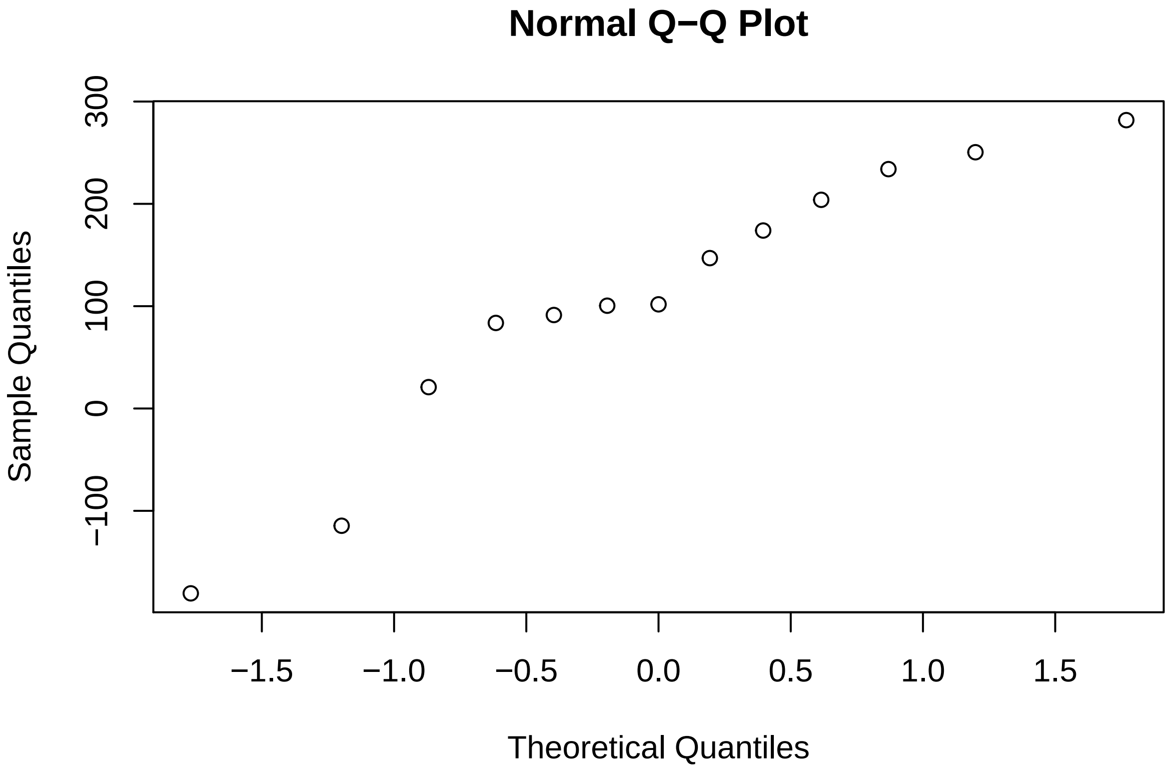
Homework3

Yanyu Zheng yz2690

February 14, 2016

Question1

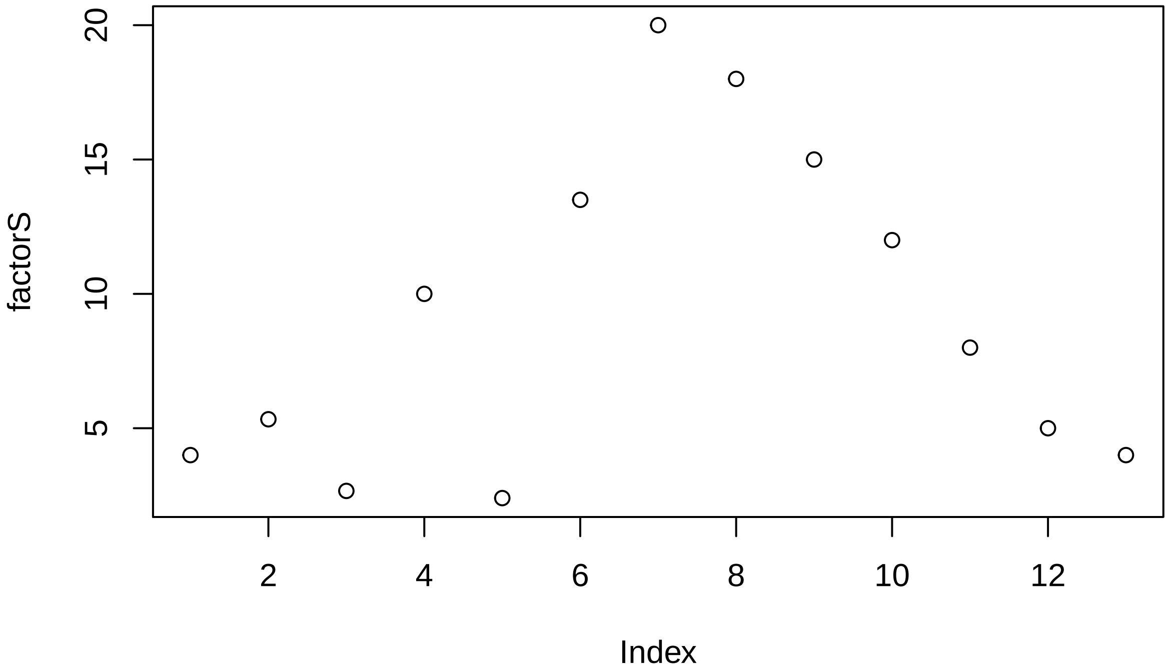
```r
library("Sleuth3")
data(ex0430)
## We do a linear regression without intercept to find the factor
model1 = lm(Sunscreen~0+PreTreatment, ex0430)
## Do a normality test on the residule to make sure the base assumption is satisfied to build a confiden
qqnorm(model1$residuals)
```



```r
## Looks okay. So we calculate the CI based on the model.
confint(model1)
```

```
##                  2.5 %   97.5 %
## PreTreatment 2.081547 4.524989
```

```r
## And we try to plot out the factors from the sample.
factorS = ex0430$Sunscreen/ex0430$PreTreatment
plot(factorS)
```



```
## The confidence interval does not cover much, obviously. So we try another approach. We use the sampl
test = t.test(log(ex0430$Sunscreen),log(ex0430$PreTreatment),paired=TRUE)
CI = exp(test$conf.int)
```

The 95% confidence interval is 4.7607549, 11.4262515. And from the R-square in the linear model we build, there's a high possibility that there exist donfounding variables. Besides, the study is not double-blind since the patient is not receiving placebo when without sunscreen. And "Can tolerent the sunlight" is highly subjective. So we have reason to believe the patients' awarness of "being protected of sunsreen" is an important confounding variable in the study.

Question2

First, we try a paired t test.

```
attach(ex0432)
t.test(ex0432$Marijuana, ex0432$Placebo, alternative="less", paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  ex0432$Marijuana and ex0432$Placebo
## t = -3.4397, df = 14, p-value = 0.001993
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -15.54936
## sample estimates:
## mean of the differences
##               -31.86667
```

P value is 0.001993, so marijuana treatment does reduce the frequency of episodes. And it reduces it by at least 15.54936 for a probability of 95%.

Then, we try a signed rank sum test.

```
wilcox.test(ex0432$Marijuana, ex0432$Placebo, alternative="less", paired = TRUE, conf.int=TRUE)
```

```
## Warning in wilcox.test.default(ex0432$Marijuana, ex0432$Placebo,
## alternative = "less", : cannot compute exact p-value with zeroes
```

```
## Warning in wilcox.test.default(ex0432$Marijuana, ex0432$Placebo,
## alternative = "less", : cannot compute exact confidence interval with
## zeroes
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  ex0432$Marijuana and ex0432$Placebo
## V = 0, p-value = 0.0008308
## alternative hypothesis: true location shift is less than 0
## 95 percent confidence interval:
##       -Inf -15.50001
## sample estimates:
## (pseudo)median
##            -28
```

P value is 0.0008308, so marijuana treatment does reduce the frequency of episodes. And it reduces it by at least 15.50001 for a probability of 95%.

The two test have very similar result in this case.

Question3

- a

```
## Type in the data
n = c(127,44,24,41,18,16,11,7,6)
mean = c(7.347, 7.368, 7.418, 7.487, 7.563, 7.568, 8.214, 8.272, 8.297)
sd = c(0.4979, 0.4235, 0.3955, 0.3183, 0.3111, 0.4649, 0.2963, 0.3242, 0.5842)
pooledEsd = sum((n-1)*(sd)^2)/sum(n-1)
```

The pooled estimate of variance is 0.1919322

- b

```
rownames = c("Between Groups", "Within Groups", "Total")
colnames = c("Sum of Squares", "df", "Mean Square", "F-Statistic", "p-value")
SST = 0.4962^2*(sum(n)-1)
SSW = sum((n-1)*(sd^2))
SSB = SST - SSW
SumofSquares = c(SSB, SSW, SST)
```

```r
df = c(8, 285, 293)
MeanSquare = SumofSquares/df
MeanSquare[3] = NA
FStatistic = MeanSquare[1]/MeanSquare[2]
pvalue = 1 - pf(FStatistic,8,285)
table = cbind(SumofSquares,df,MeanSquare,FStatistic,pvalue)
rownames(table) = rownames
colnames(table) = colnames
table[2:3,4:5] = NA
tableB = as.table(table)
tableB
```

```
##                 Sum of Squares           df  Mean Square  F-Statistic
## Between Groups   1.744015e+01 8.000000e+00 2.180019e+00 1.135828e+01
## Within Groups    5.470068e+01 2.850000e+02 1.919322e-01
## Total            7.214083e+01 2.930000e+02
##                      p-value
## Between Groups 5.662137e-14
## Within Groups
## Total
```

- c

```r
yBar = sum(n*mean)/sum(n)
BgSS = sum(n*mean^2)-sum(n)*yBar^2
```

The between group sum of square is 17.4540223, which is consistent with the result in b.

- d
- 1 Two big group VS separate-means

```r
rownames = c("Between Groups", "Two Big Groups", "Others")
colnames = c("Sum of Squares", "df", "Mean Square", "F-Statistic", "p-value")
BetweenGroups = table["Between Groups",]
n1 = n[1:6]
n2 = n[7:9]
mean1 = sum(mean[1:6]*n1)/sum(n1)
mean2 = sum(mean[7:9]*n2)/sum(n2)
meanT = sum(mean*n)/sum(n)
TwoBigGroups = (mean1-meanT)^2*sum(n1) + (mean2-meanT)^2*sum(n2)
TwoBigGroups = c(TwoBigGroups, 1, TwoBigGroups)
Others = BetweenGroups[1:2] - TwoBigGroups[1:2]
Others = c(Others, Others[1]/Others[2],NA,NA)
TwoBigGroups = c(TwoBigGroups, TwoBigGroups[3]/Others[3], 1-pf(TwoBigGroups[3]/Others[3],1,Others[2]))
table1 = rbind(BetweenGroups, TwoBigGroups, Others)
tableD1 = as.table(table1)
tableD1
```

```
##               Sum of Squares           df  Mean Square  F-Statistic
## BetweenGroups   1.744015e+01 8.000000e+00 2.180019e+00 1.135828e+01
## TwoBigGroups    1.578160e+01 1.000000e+00 1.578160e+01 6.660676e+01
## Others          1.658558e+00 7.000000e+00 2.369369e-01
##                     p-value
## BetweenGroups 5.662137e-14
## TwoBigGroups  8.022381e-05
## Others
```

P value is 8.022381e-05, which means the difference between the two big group is significant.

- Two Big Groups VS equal-means

```r
rownames = c("Between Groups", "Within Groups", "Total")
colnames = c("Sum of Squares", "df", "Mean Square", "F-Statistic", "p-value")
TwoBigGroups = (mean1-meanT)^2*sum(n1) + (mean2-meanT)^2*sum(n2)
TwoBigGroups = c(TwoBigGroups, 1)
Total = table["Total",]
WithinGroups = Total[1:2] - TwoBigGroups
WithinGroups = c(WithinGroups, WithinGroups[1]/WithinGroups[2],NA,NA)
TwoBigGroups = c(TwoBigGroups, TwoBigGroups[1]/TwoBigGroups[2])
TwoBigGroups = c(TwoBigGroups, TwoBigGroups[3]/WithinGroups[3],1-pf(TwoBigGroups[3]/WithinGroups[3],TwoB
table2 = rbind(TwoBigGroups, WithinGroups, Total)
rownames(table2) = rownames
colnames(table2) = colnames
tableD2 = as.table(table2)
tableD2
```

```
##                Sum of Squares          df Mean Square F-Statistic
## Between Groups     15.7815955   1.0000000  15.7815955  81.7652305
## Within Groups      56.3592354 292.0000000   0.1930111
## Total              72.1408309 293.0000000
##                    p-value
## Between Groups   0.0000000
## Within Groups
## Total
```

Question4

Power of the test = probability of rejecting the null when the alternative is true. Under the alternative hypothesis, we have $/frac/miu_x - /miu_y - 0.010.0214 \sim$ t(57). And we reject the null when $/frac/miu_x - /miu_y 0.0214$ is larger than 2.0024655 or less than $-\infty$.

```r
power = 1 - pt(qt(0.975, 57)-0.01/0.0214,57) + pt(qt(0.025, 57)-0.01/0.0214,57)
```

So the power of the test is 0.0734058.

Question 5

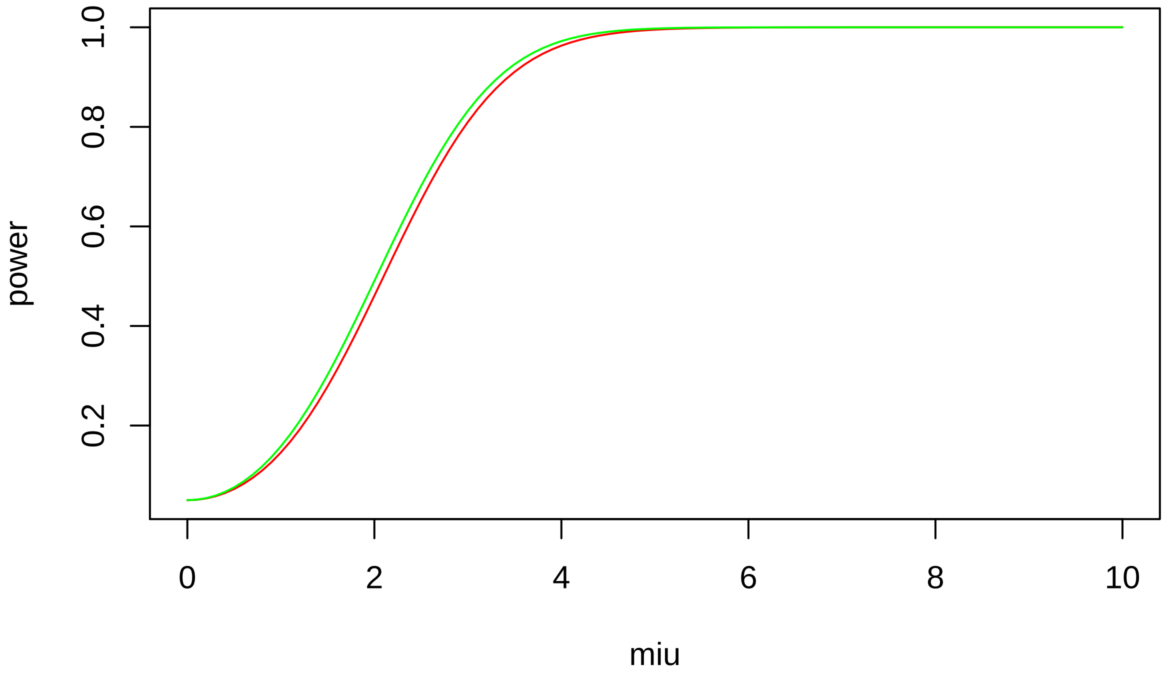
Question 6

Power of the test = probability of rejecting the null when the alternative is true. When $/sigma^2 = 1$, $/miu = 0.1$, Under the alternative hypothesis, we have $/frac/miu - 0.1/sigma \sim$ t(18). And we reject the null when $/frac/miu/sigma$ is larger than 2.100922 or less than $-\infty$.

```
powerP1 = 1 - pt(qt(0.975, 18)-0.1*1,18) + pt(qt(0.025, 18)-0.1*1,18)
powerP5 = 1 - pt(qt(0.975, 18)-0.5*1,18) + pt(qt(0.025, 18)-0.5*1,18)
power1 = 1 - pt(qt(0.975, 18)-1*1,18) + pt(qt(0.025, 18)-1*1,18)
```

The power for $/miu = 0.1$ is 0.0508734, for $/miu = 0.5$ is 0.0724313, for $/miu = 1$ is 0.1458012.

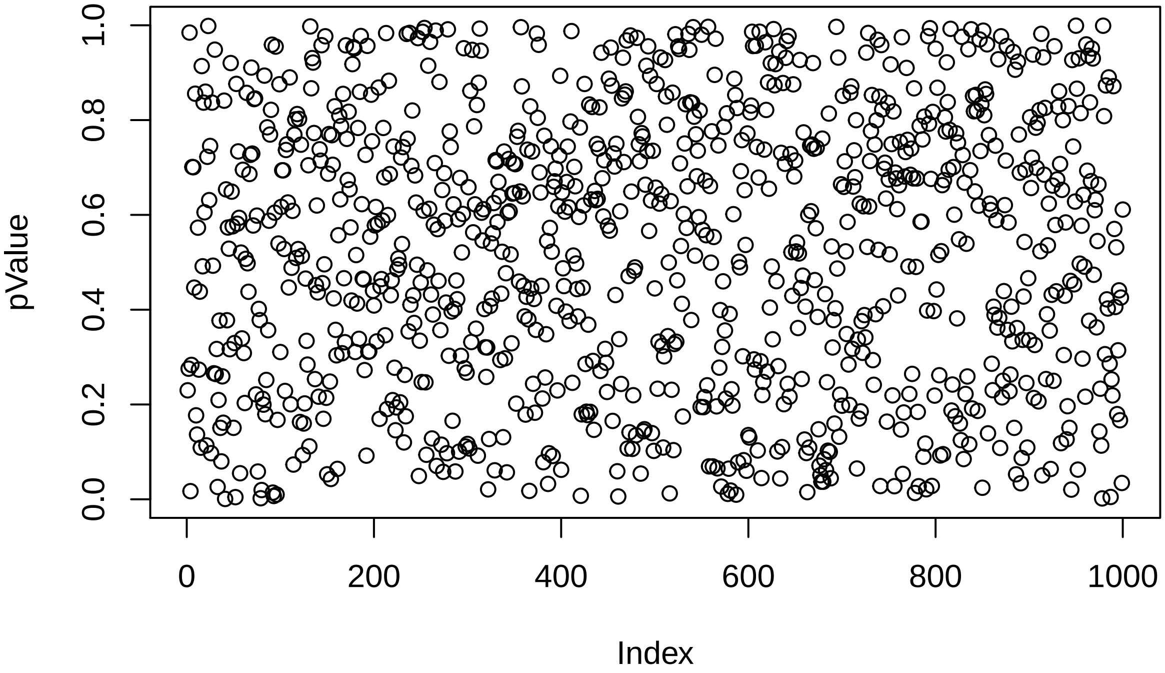
```
miu = seq(0,10,0.1)
power = 1 - pt(qt(0.975, 18)-miu,18) + pt(qt(0.025, 18)-miu,18)
plot(miu,power,type="l",col = "red")
power2 = 1 - pt(qt(0.975, 38)-miu,38) + pt(qt(0.025, 38)-miu,38)
lines(miu,power2,type="l", col = "green")
```



Question7

We do the simulation and do a plot of the pvalues we got and a qqplot with the uniform distribution.

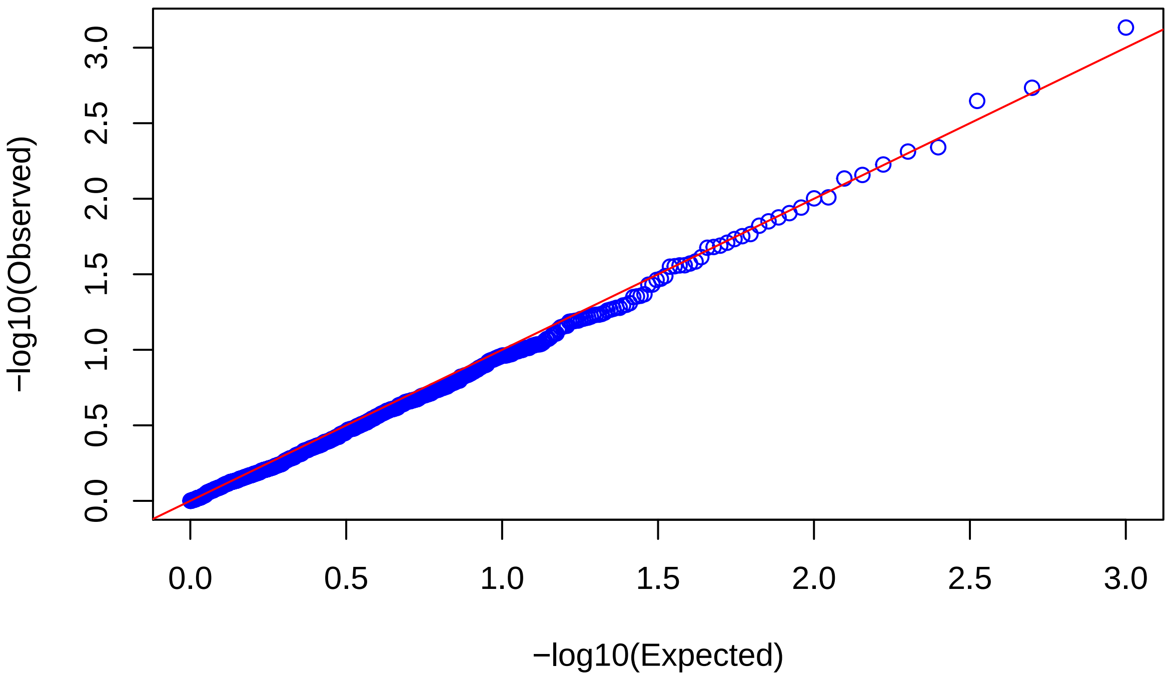
```
set.seed(7)
pValue = vector()
for(i in 1:1000){
  x = rnorm(10)
  y = rnorm(10)
  test = t.test(x,y)
  pValue = c(pValue,test$p.value)
}
plot(pValue)
```



```
library("gap")
```

```
## gap version 1.1-16
```

```
qqunif(pValue)
```



From the plot, we can see the plot is very much uniform in (0,1).