project 1

1

1.a

$$H0 : There\ is\ no\ difference\ of\ finishing\ times\ between\ the\ tortoise\ and\ hare\ groups.$$

$$H\alpha : There\ is\ difference\ of\ finishing\ times\ between\ hare\ and\ tortoise\ groups$$

We are only interested in whether there is difference in this two groups now, instead of which one is longer or shorter.

1.b

```
data<-read.csv('race.csv')
dif_mean<-function(data){
  mean(data[,1])-mean(data[,2])
}
raw_df_mean<-dif_mean(data)
raw_df_mean
```

```
## [1] -5.045642
```

1.c when caculate the variance of difference of sample mean, we have a total df=N1+N2.

Since the variance is additive

$$Var(\bar{X} - \bar{Y}) = Var(\bar{X}) + Var(\bar{Y})$$

$$recall\ that\ (N_1 - 1)S_1^2/\sigma_1^2\ follows\ chisq(N_1 - 1)$$

$$(N_2 - 1)S_2^2/\sigma_2^2\ follows\ chisq(N_2 - 1)\ with\ \sigma_1^2 = \sigma_2^2 = \sigma^2$$

$$So\ we\ have$$

$$S_p^2/\sigma^2 = (N_1 - 1)S_1^2/((N_1 + N_2 - 2) * \sigma^2) + (N_2 - 1)S_2^2/((N_1 + N_2 - 2) * \sigma^2)$$

$$=$$

$$also\ S_p^2/\sigma^2 = 1/(1/N_1 + 1/N_2)$$

$$thus\ \sigma^2 = ((N_1 - 1)S_1^2 + (N_2 - 1)S_2^2)/(((N_1 + N_2 - 2)) * (1/N_1 + 1/N_2))$$

1.d

```
var_sam_mean<-function(data){
  ((var(data[,1])+var(data[,2]))*9/((10+10-2))*(1/10+1/10))
}
raw_var_sam_mean<-var_sam_mean(data)
raw_var_sam_mean
```

```
## [1] 82.62257
```

1.e.i

```
t_stat<-function(data){
  t_s<-dif_mean(data)/sqrt(var_sam_mean(data))
  p_val<-1-2*(abs(0.5-pt(t_s,df=18)))
  list(t_statistic=t_s,p_value=p_val)
}
raw_t_stat<-unlist(t_stat(data))
raw_t_stat
```

```
## t_statistic     p_value
##  -0.5550947   0.5856628
```

1.e.ii

```
left<-qt(0.025,18)
right<-qt(0.975,18)
null_region<-c(left,right)
null_region
```

```
## [1] -2.100922  2.100922
```

the reject region is |t|>2.101. Given that the statistic fall out of the reject region, we cannot reject the null.

1.e.iii

No. we assume that the two sample is normally distributed and with equal variance. We test on these two assumptions.

```
####normally distributed test####
shapiro.test(data[,1])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data[, 1]
## W = 0.52331, p-value = 6.83e-06
```

```
shapiro.test(data[,2])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data[, 2]
## W = 0.7704, p-value = 0.006324
```

```
####equality of variance test####
library(car)
```

```
## Loading required package: carData
```

```
leveneTest(c(data[,1],data[,2]),c(rep(1,10),rep(2,10)))
```

```
## Warning in leveneTest.default(c(data[, 1], data[, 2]), c(rep(1, 10),
## rep(2, : c(rep(1, 10), rep(2, 10)) coerced to factor.
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  1   0.613 0.4438
##       18
```

The results show that in the sample distribution test, both p-value<0.01, which means that the two sample is not normally distributed.In the equality test, p-value=0.44>0.05, which means that the variance of two groups have no significant differences. t-test may not fit this data well.

2.a

```
####U statistics####

u_stat<-function(a,b){
  score=0
  for (i in 1:length(a)){
    result<-a[i]<b
    score=score+sum(result)
  }
  score
}
raw_u_hare<-u_stat(data$Hare,data$Tortoise)
raw_u_tortoise<-u_stat(data$Tortoise,data$Hare)
raw_u_hare
```

```
## [1] 81
```

```
raw_u_tortoise
```

```
## [1] 19
```

2.b

The U-statistic for each team should both be 50. Note that sum of both U-statistics for these two teams should be 100, in each comparison, for each i-th hare and j-th tortoise, given that all the numbers in the 20 finishing time not exactly the same

$$X_{hare,i} < X_{tortoise,j}$$

$$X_{tortoise,j} < X_{hare,i}$$

####one of these two boolean must be true and the other be false. Thus

$$I(X_{hare,i} < X_{tortoise,j}) = 0 and I(X_{tortoise,j} < X_{hare,i}) = 1$$

$$I(X_{hare,i} < X_{tortoise,j}) = 1 and I(X_{tortoise,j} < X_{hare,i}) = 0$$

$$U_{hare} + U_{tortoise} = (\sum_{i=1}^{n1})(\sum_{i=1}^{n2})(1+0) = n1 * n2 = 100$$

if the null hypothesis is true, u_statistics should be equal, both equals to 100/2=50

2.c.i

```
sd_mu0<-sqrt(10*10*21/12)
z_stat<-function(a,b){
  z<-(u_stat(a,b)-50)/sd_mu0
  p_val<-1-2*abs(pnorm(z)-0.5)
  list(z_statistic=z,p_value=p_val)

}
raw_z_stat<-unlist(z_stat(data$Hare,data$Tortoise))
raw_z_stat
```

```
## z_statistic     p_value
##  2.34337973  0.01910992
```

p-value is 0.019

2.c.ii

p<0.05,we reject the null hypothesis.

2.c.iii

```
####wilcoxon test####
wilcox.test(data$Hare,data$Tortoise,exact=F,correct=F)
```

```
##
##  Wilcoxon rank sum test
##
## data:  data$Hare and data$Tortoise
## W = 19, p-value = 0.01911
## alternative hypothesis: true location shift is not equal to 0
```

```
wilcox.test(data$Tortoise,data$Hare,exact=F,correct=F)
```

```
##
##  Wilcoxon rank sum test
##
## data:  data$Tortoise and data$Hare
## W = 81, p-value = 0.01911
## alternative hypothesis: true location shift is not equal to 0
```

the result is the same.

3

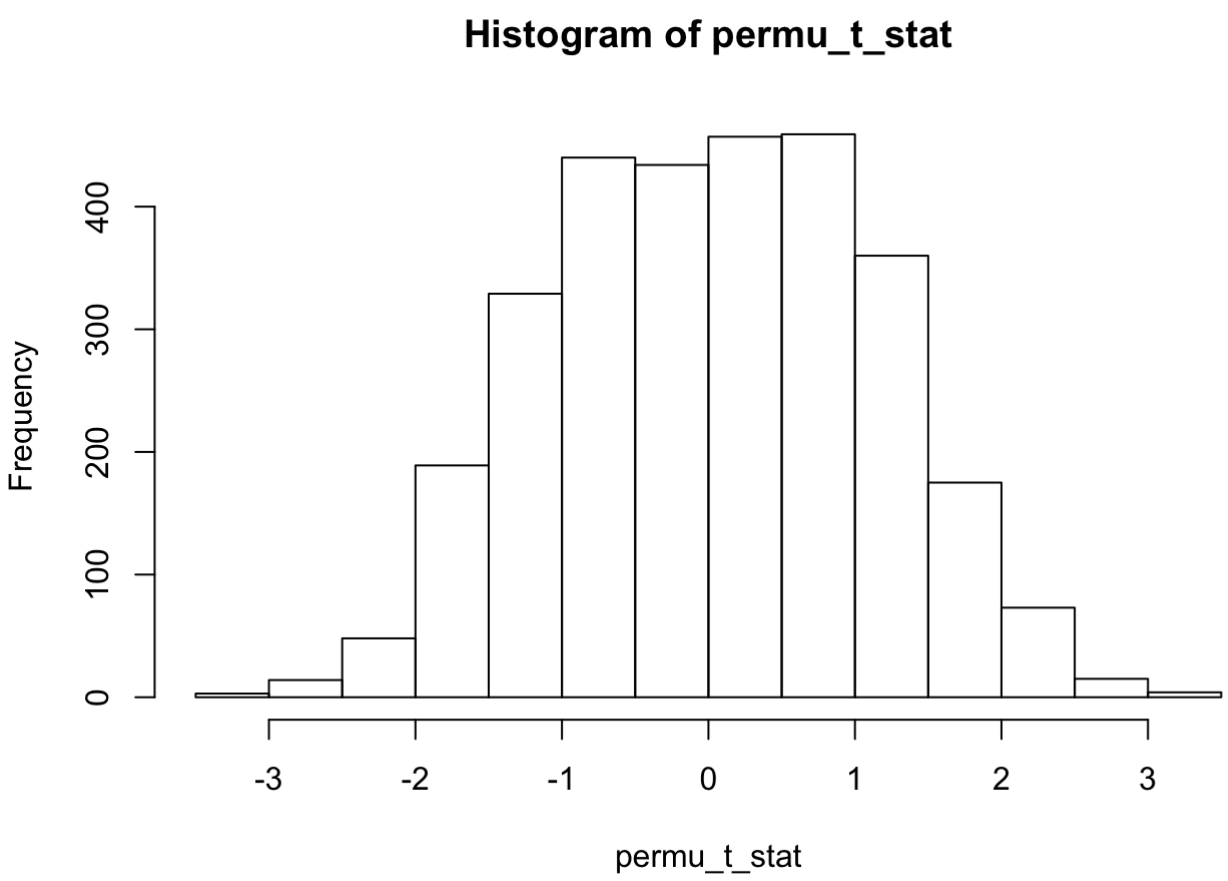
3.a

```
####generate permuted datasets####
set.seed(1)
permu_data<-matrix(0,nrow=30000,ncol=2)
for (i in 1:10){
  index<-seq(i,29990+i,10)
  for (j in index){
  permu_data[j,]=sample(c(data[i,1],data[i,2]),size=2,replace=F)}
}
```

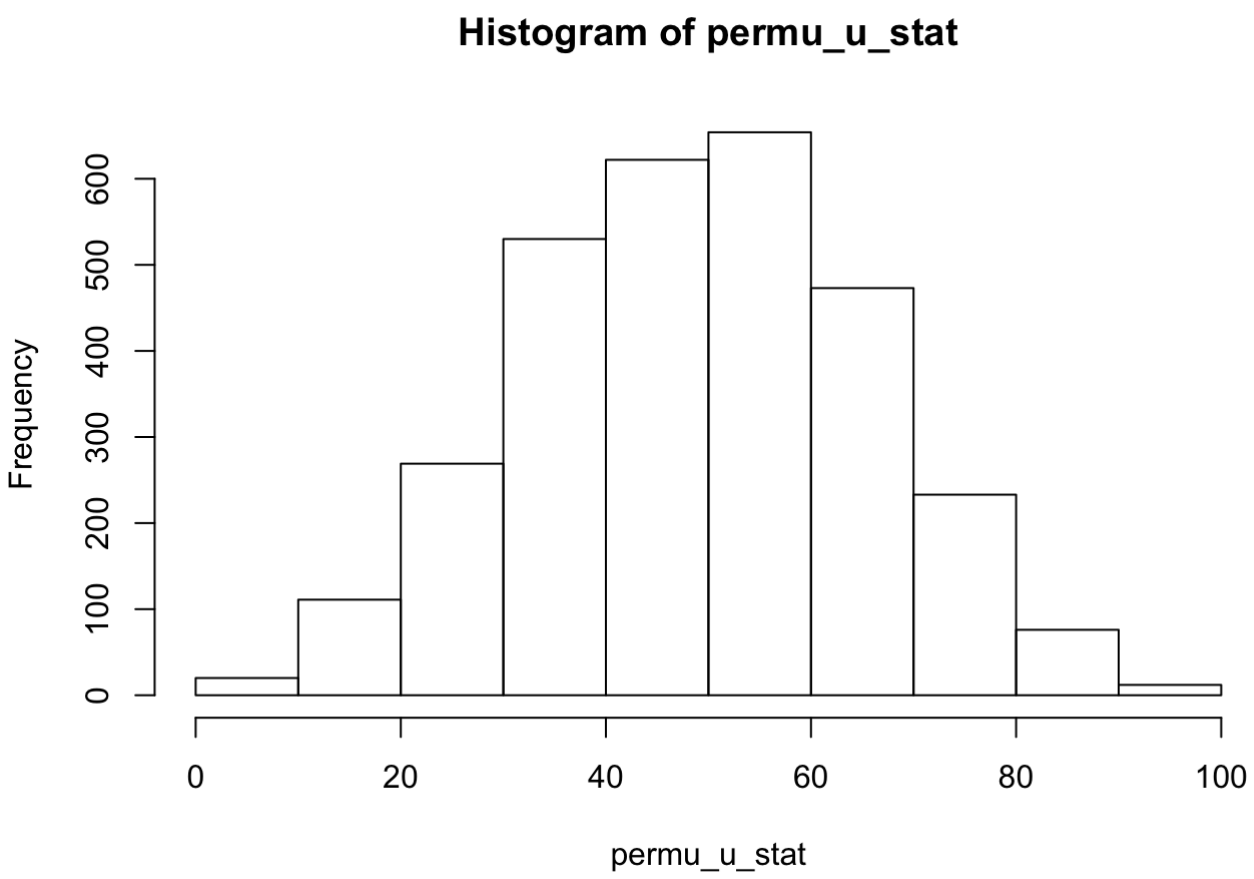
3.b

```
permu_mean_dif<-vector()
permu_t_stat<-vector()
permu_u_stat<-vector()
permu_z_stat<-vector()
permu_wilc_stat<-vector()

for (i in 1:3000){
  new_data<-permu_data[(10*i-9):(10*i),]
  permu_mean_dif<-c(permu_mean_dif,dif_mean(new_data))
  permu_t_stat<-c(permu_t_stat,unlist(t_stat(new_data)$t_statistic))
  permu_u_stat<-c(permu_u_stat,u_stat(new_data[,1],new_data[,2])[1])
  permu_z_stat<-c(permu_z_stat,unlist(z_stat(new_data[,1],new_data[,2])$   z_statistic))
  permu_wilc_stat<-c(permu_wilc_stat,wilcox.test(new_data[,2],new_data[,1],exact=F,corre
ct=F)$statistic)
}

hist(permu_mean_dif)
```

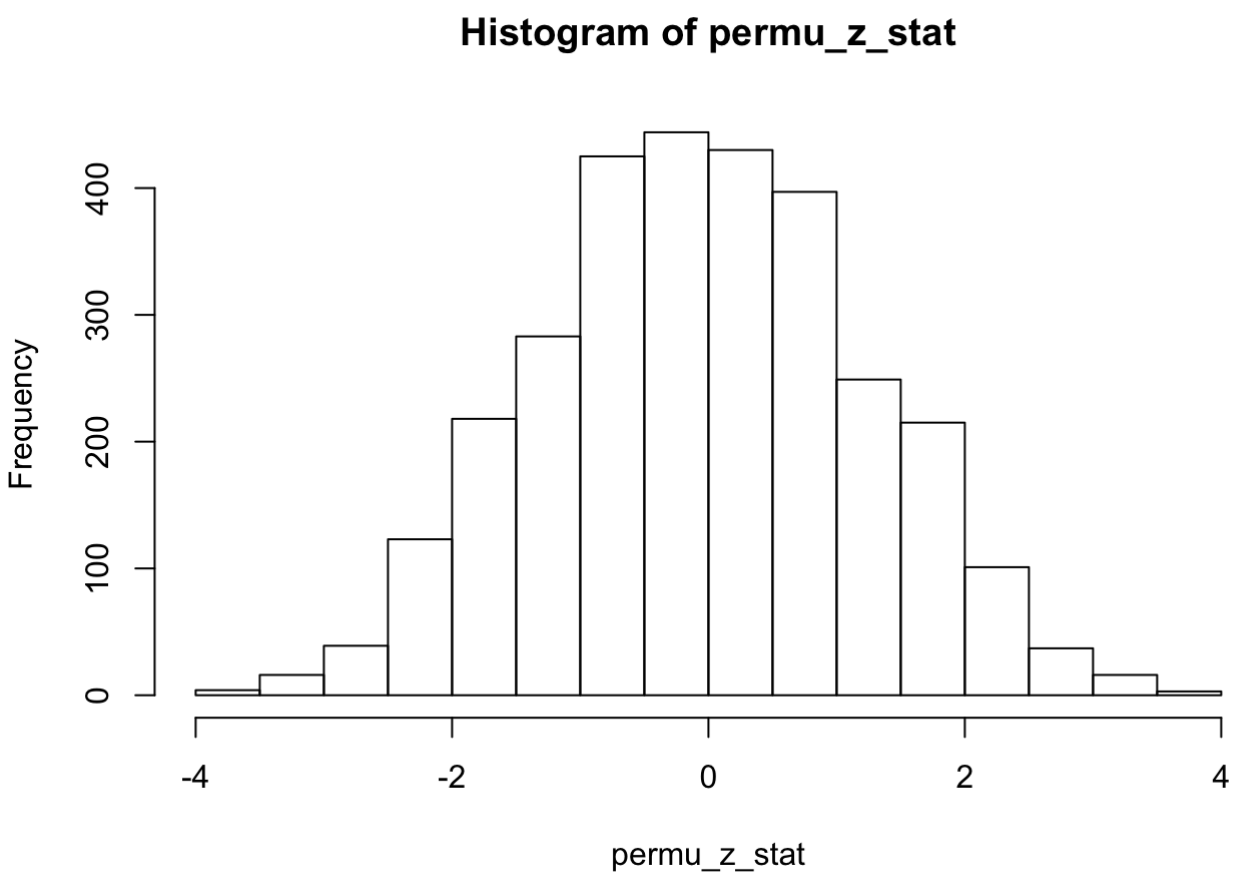

```
hist(permu_t_stat)
```



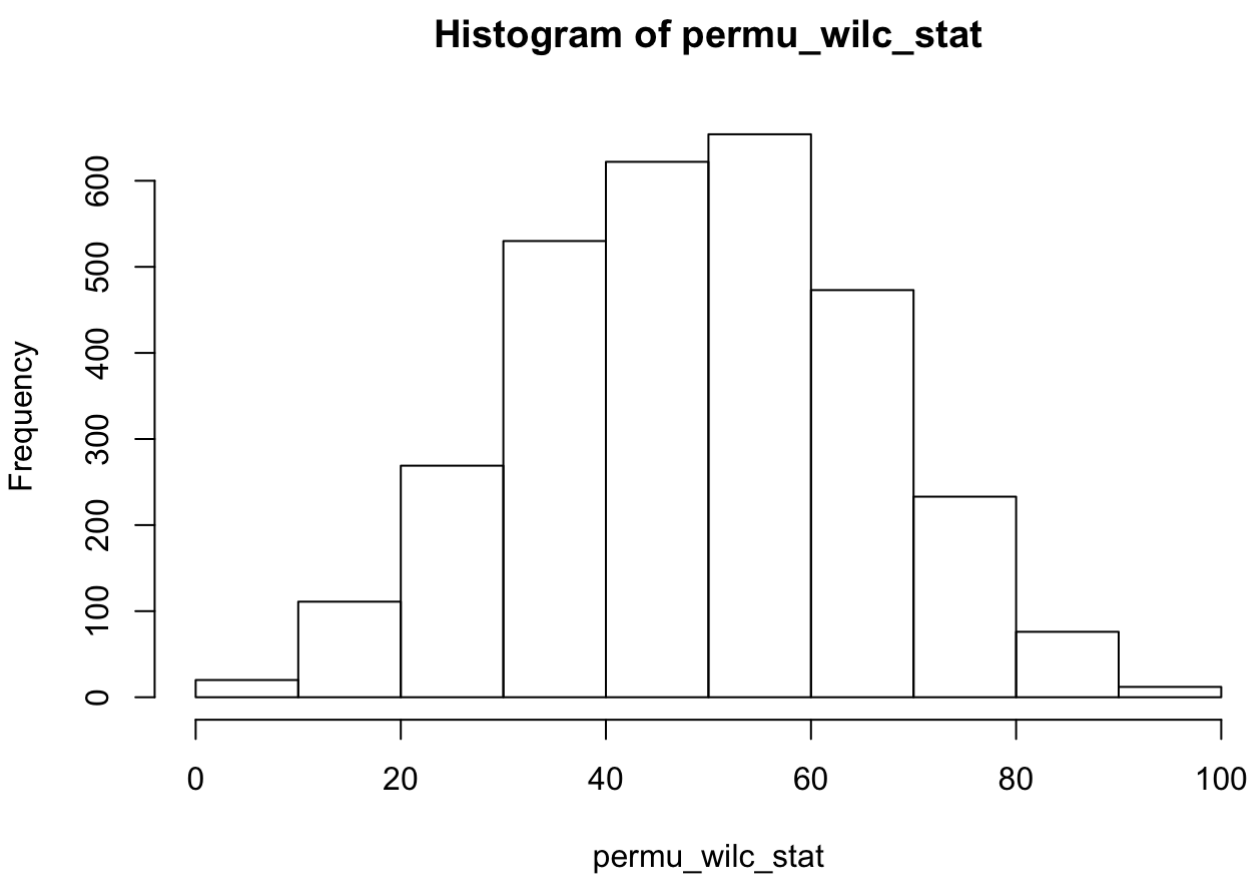

```
hist(permu_u_stat)
```




```
hist(permu_z_stat)
```

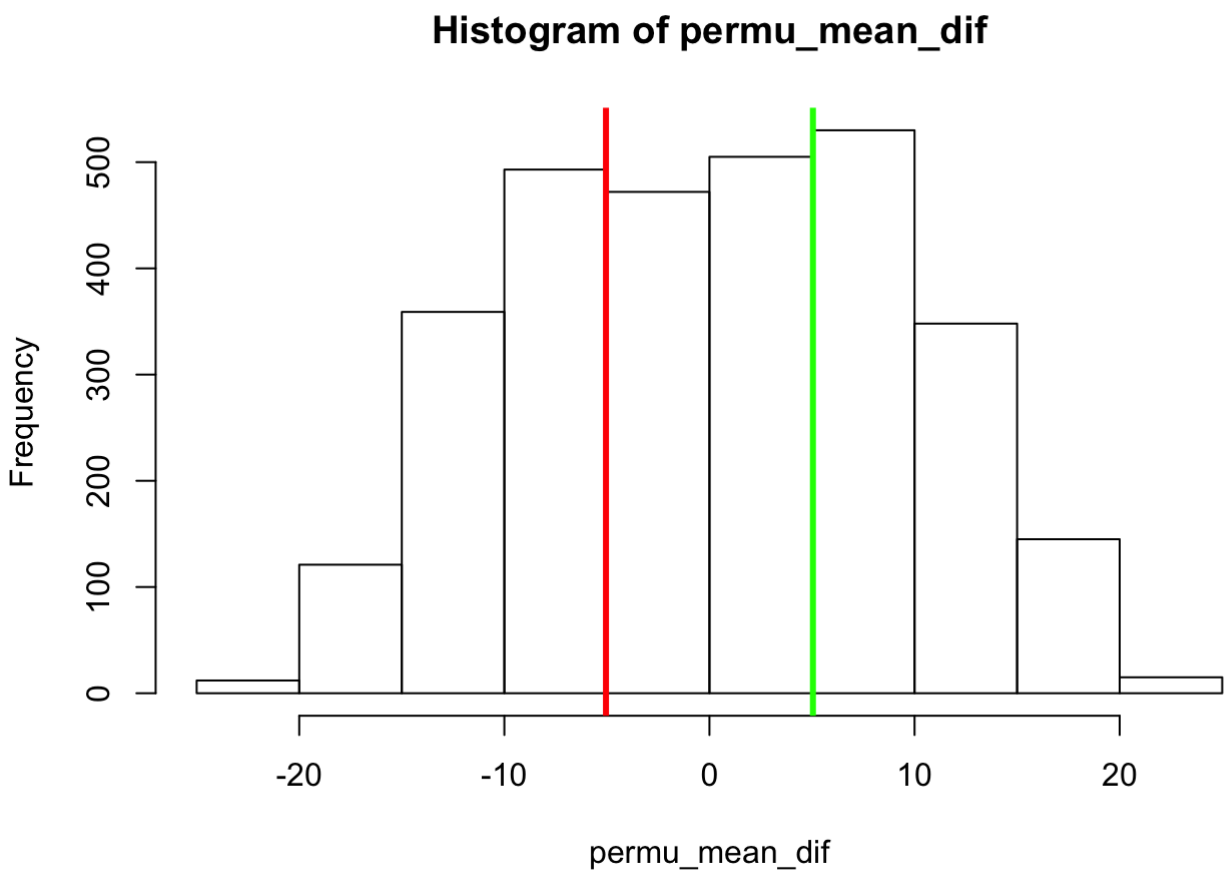



```
hist(permu_wilc_stat)
```

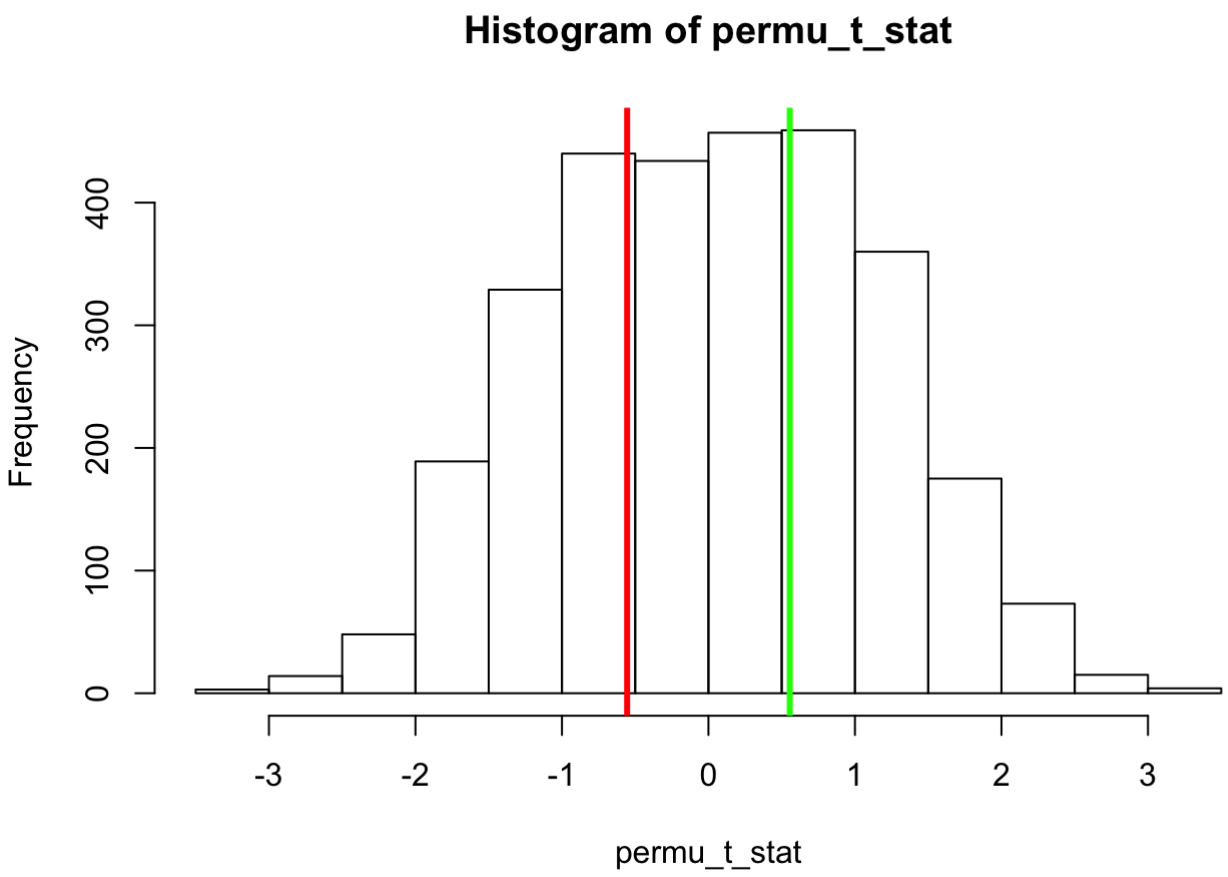


####3.c

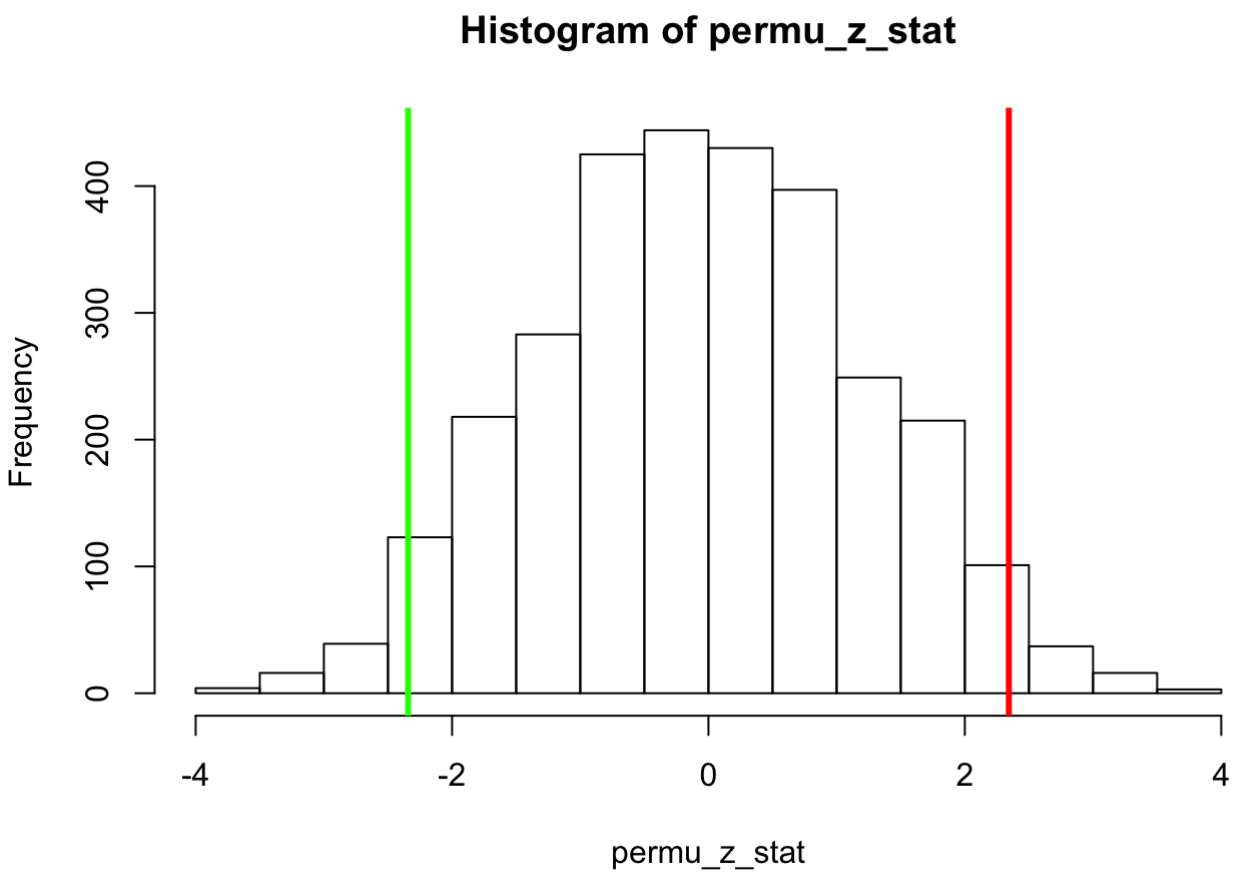
```
p_val<-function(x,data){
  tile<-sum(x>=data)/length(data)
  p_value<-1-2*abs(tile-0.5)
  p_value
}

hist(permu_mean_dif)
abline(v=raw_df_mean,col='red',lwd=3)
abline(v=-raw_df_mean,col='green',lwd=3)
```



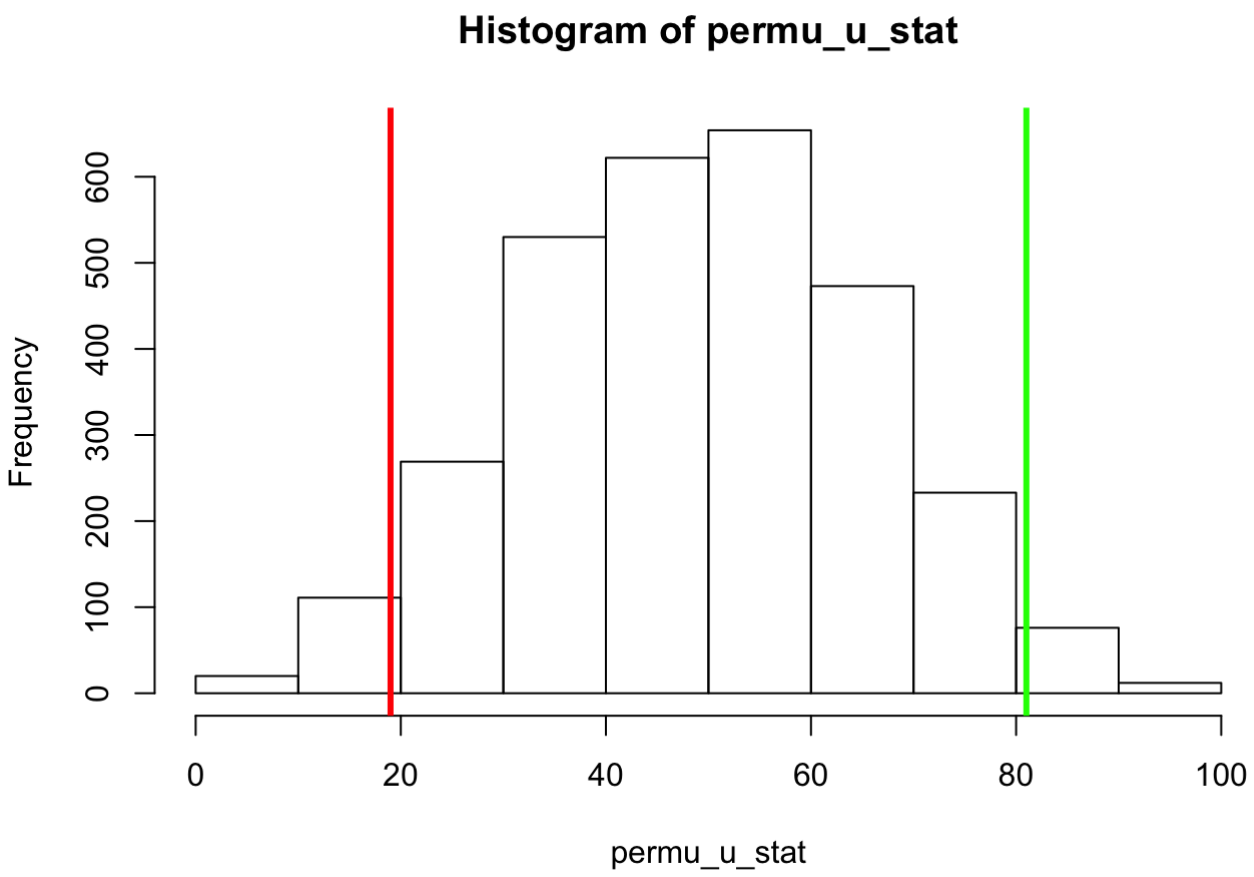

```
hist(permu_t_stat)
abline(v=raw_t_stat[1],col='red',lwd=3)
abline(v=-raw_t_stat[1],col='green',lwd=3)
```



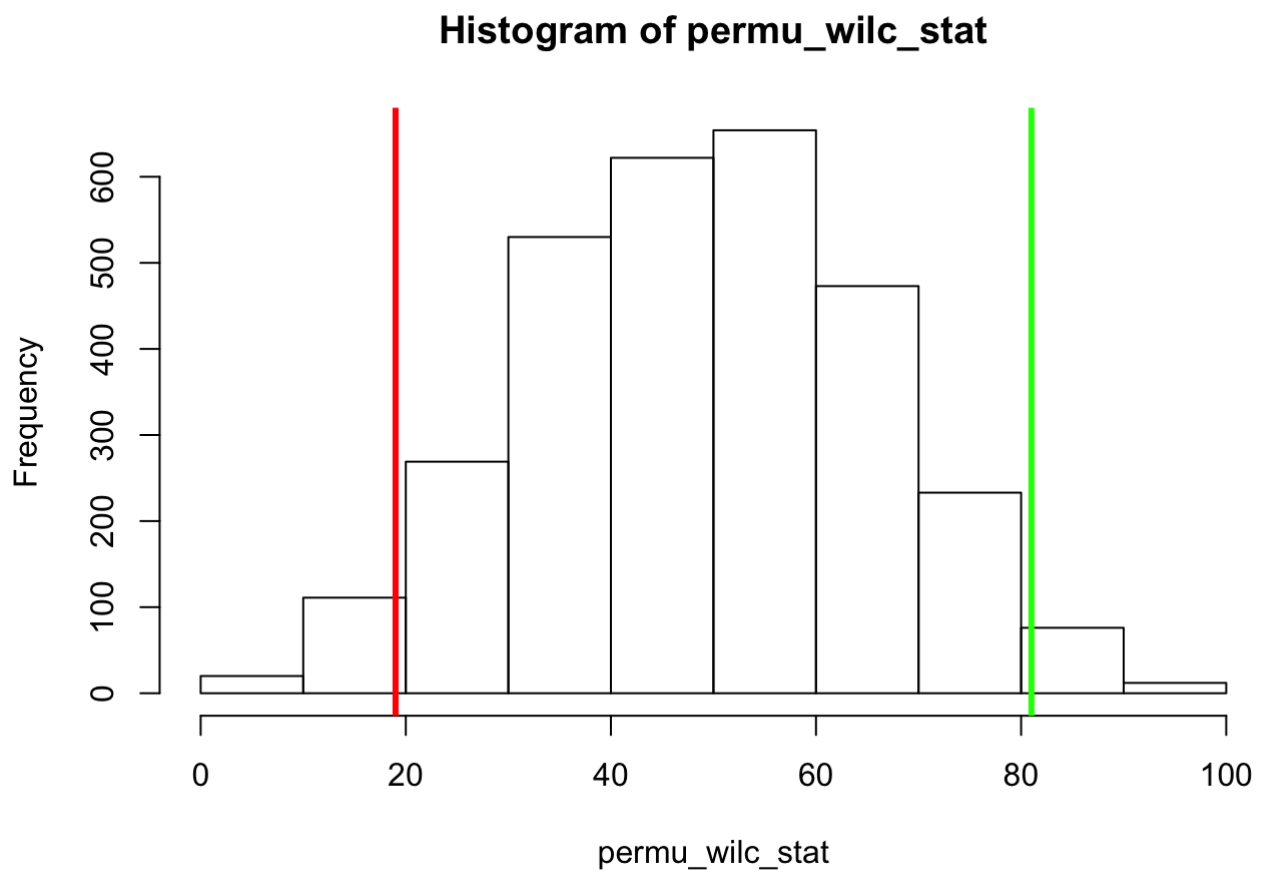

```
hist(permu_z_stat)
abline(v=raw_z_stat[1],col='red',lwd=3)
abline(v=-raw_z_stat[1],col='green',lwd=3)
```




```
hist(permu_u_stat)
abline(v=raw_u_tortoise,col='red',lwd=3)
abline(v=100-raw_u_tortoise,col='green',lwd=3)
```




```
hist(permu_wilc_stat)
abline(v=19,col='red',lwd=3)
abline(v=81,col='green',lwd=3)
```



####3.c.i ####these distributions are approximately normally distributed.while after centralization and standardization, t_statistic and z_statistic have smaller variance.sample mean difference tend to be thin tail distribution with smaller scale of the sample, while only with 10 number, the number range of u_statistic aka. wilcoxon statistics scale are much bigger and could clarify the statistic more efficiently.

3.c.ii

mean value of mean difference of two sample should be 0, of t statistics should be 0,u statistics should be 50, z statistics should be 0, wilcoxon statistic should be 50.

3.c.iii

refer to the position of raw statistics in the statistics distribution on permuted dataset.in a two-sided test, the region space out of the region between raw statistic and the symmetric position on the other half part of the distribution(shown in green lines) should be p-value

```
p_dif_mean<-1-2*abs(sum(raw_df_mean<permu_mean_dif)-1500)/3000
p_dif_mean
```

```
## [1] 0.6526667
```

```
p_t_stat<-1-2*abs(sum(raw_t_stat[1]<permu_t_stat)-1500)/3000
p_t_stat
```

```
## [1] 0.6526667
```

```
p_u_stat<-1-2*abs(sum(raw_u_tortoise<permu_u_stat)-1500)/3000
p_u_stat
```

```
## [1] 0.07933333
```

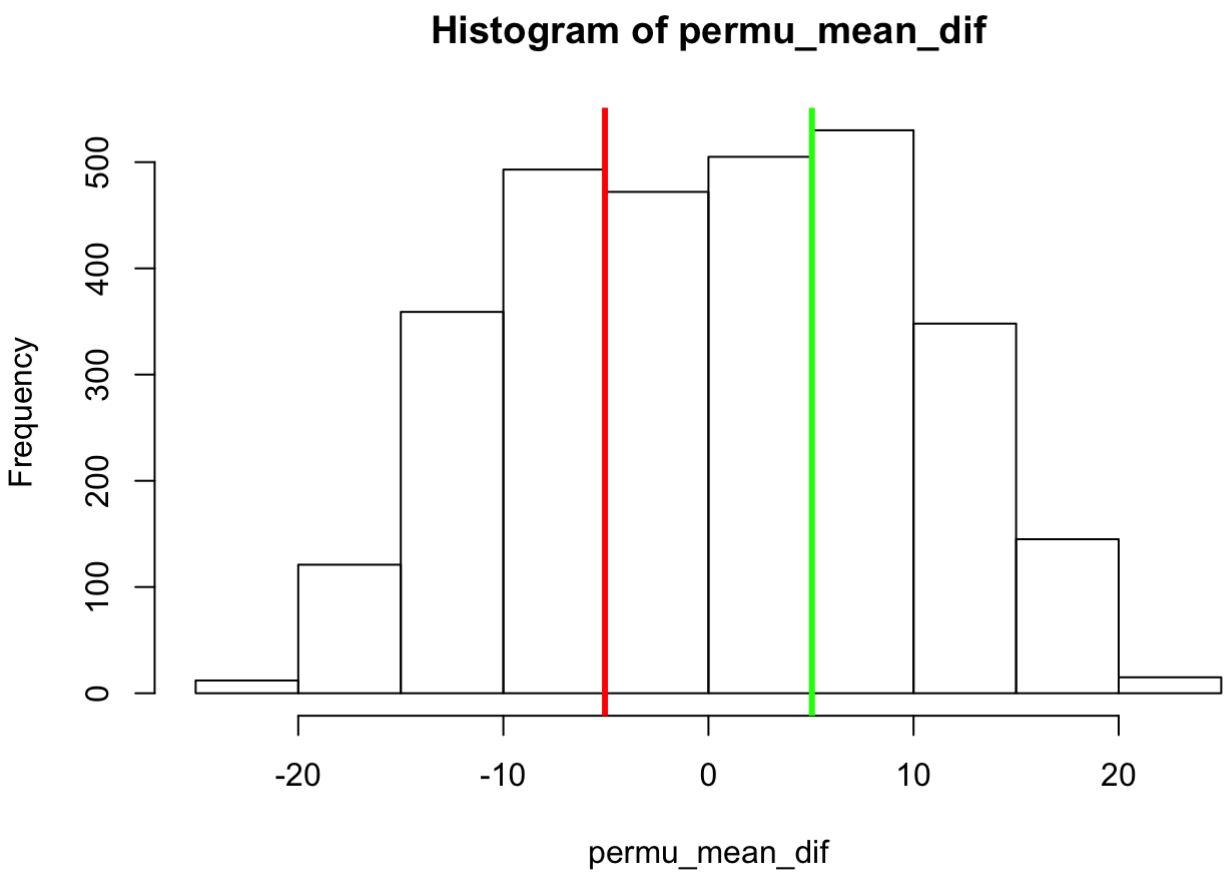
```
p_z_stat<-1-2*abs(sum(raw_z_stat[1]<permu_z_stat)-1500)/3000
p_z_stat
```

```
## [1] 0.052
```

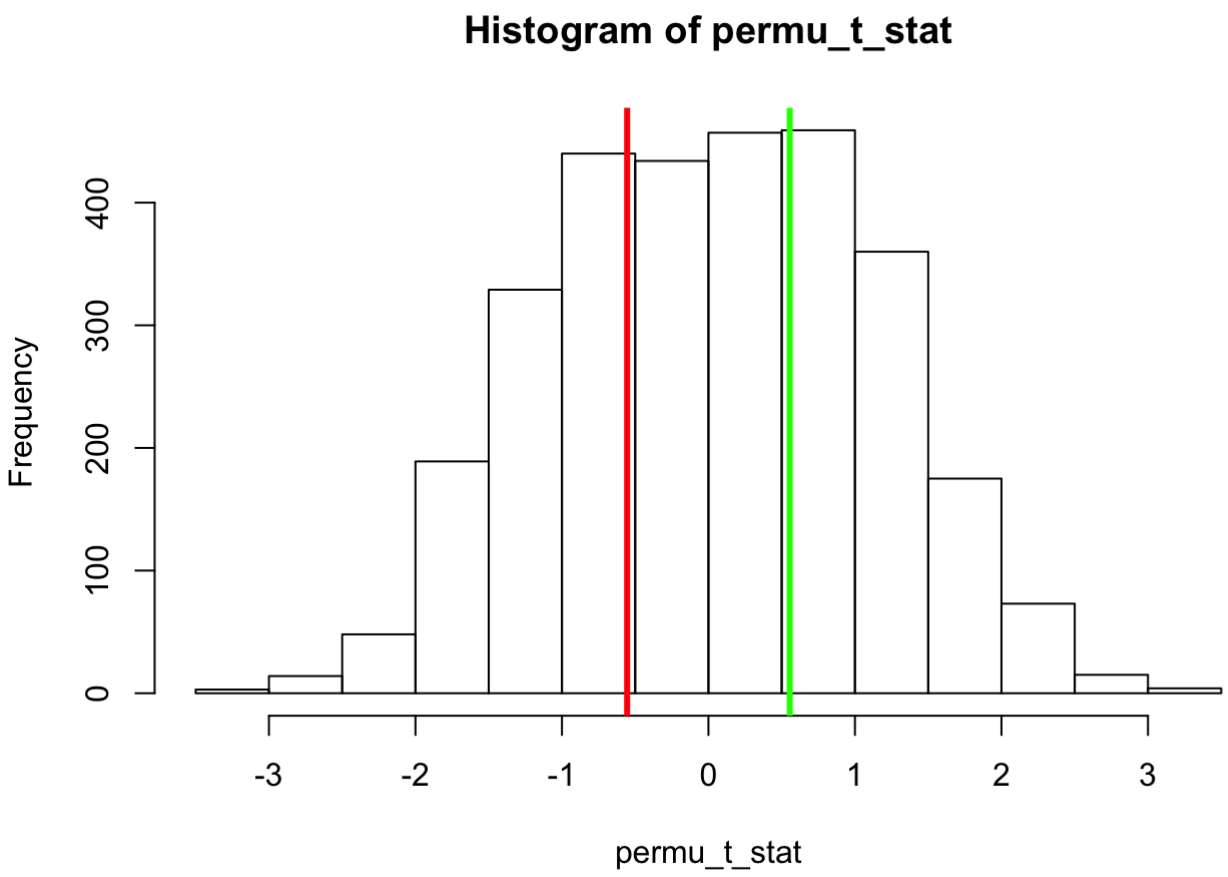
we cannot reject the null based on the sample mean difference distribution and t-statistics, however, we could not reject the null with z and u statistics either, but the p-value is much smaller.

3.c.iv

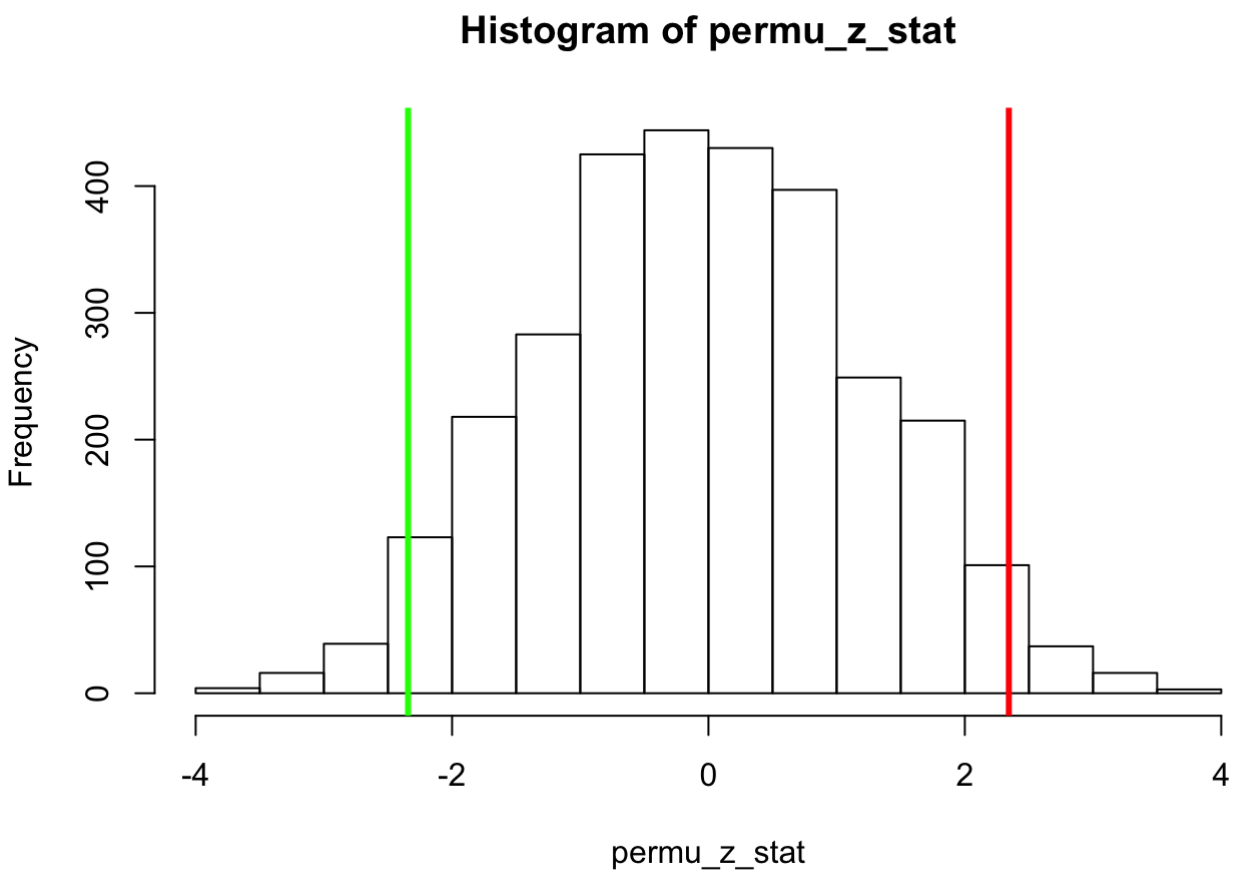
```
p_val<-function(x,data){
  tile<-sum(x>=data)/length(data)
  p_value<-1-2*abs(tile-0.5)
  p_value
}

hist(permu_mean_dif)
abline(v=raw_df_mean,col='red',lwd=3)
abline(v=-raw_df_mean,col='green',lwd=3)
```



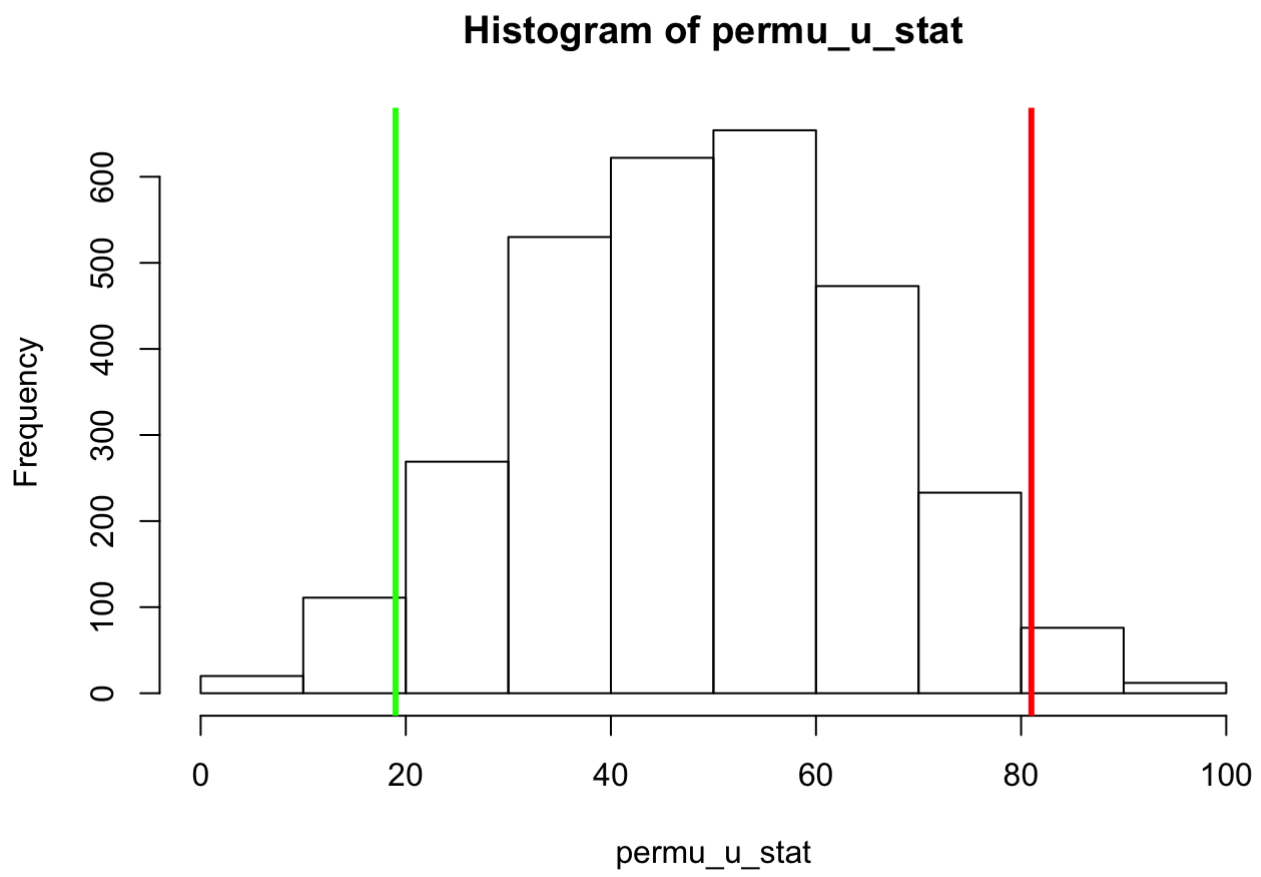

```
hist(permu_t_stat)
abline(v=raw_t_stat[1],col='red',lwd=3)
abline(v=-raw_t_stat[1],col='green',lwd=3)
```



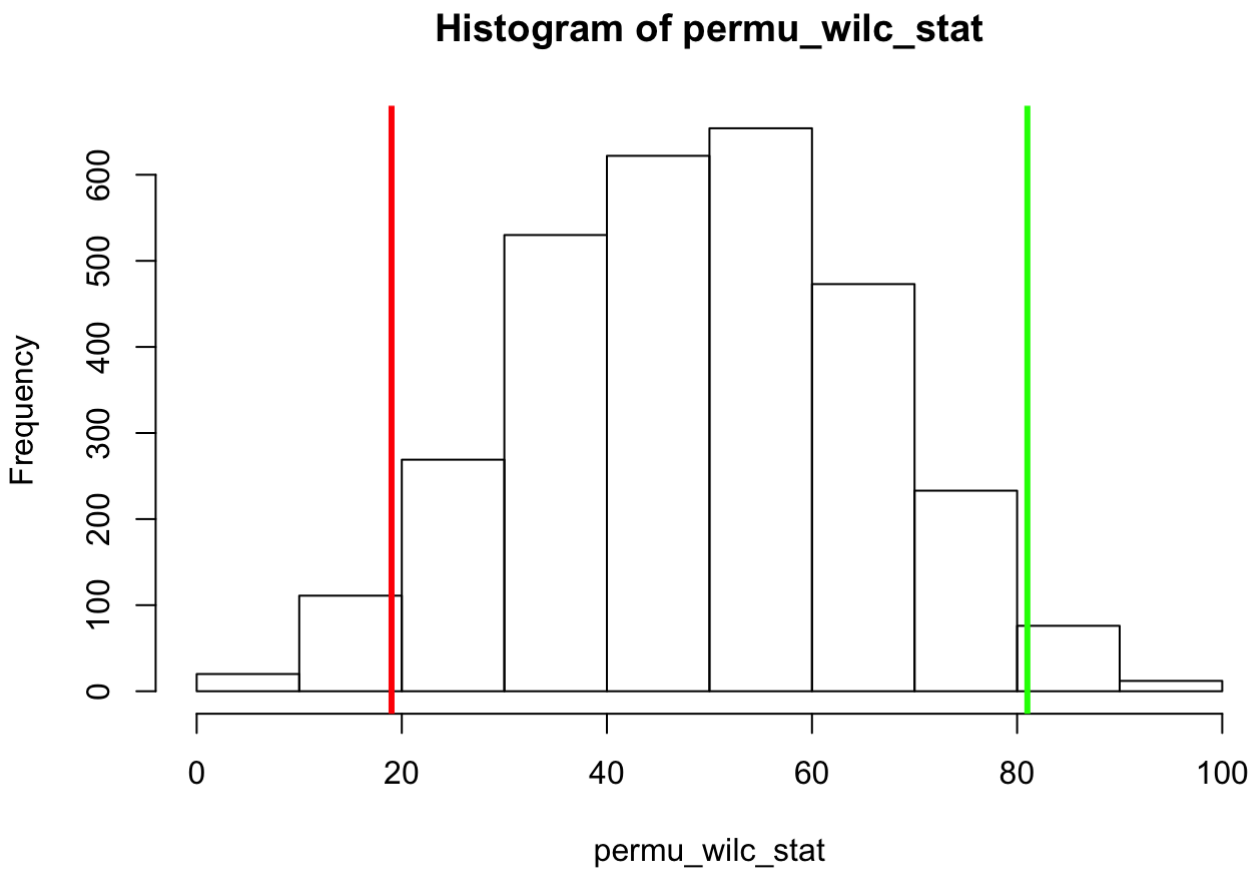

```
hist(permu_z_stat)
abline(v=raw_z_stat[1],col='red',lwd=3)
abline(v=-raw_z_stat[1],col='green',lwd=3)
```




```
hist(permu_u_stat)
abline(v=raw_u_hare,col='red',lwd=3)
abline(v=100-raw_u_hare,col='green',lwd=3)
```




```
hist(permu_wilc_stat)
abline(v=19,col='red',lwd=3)
abline(v=81,col='green',lwd=3)
```



4

(1)Wilcoxon rank sum test aka. Mann-Whitney U test and test based on z statistics:

generally these methods are non-parametric methods with no assumptions on the sample distribution.

pros:when the sample is not normally dirtributed or with distributions unkown,we apply Wilcoxon rank sum test. for the distribution of data. These models are robot especially against outliers.

Cons: when we know the data is normally distributed, Mann-Whitney test does not perform as well as t-test.

(2)When the sample are normally distributed with same variance, t_test and the test based on sample mean difference perform better. With more information taken into consideration including the sample mean and variance, the results would be more believable.While when the model assumption about the sample distribution and equality of variance don`t meet, there are not preferable methods.