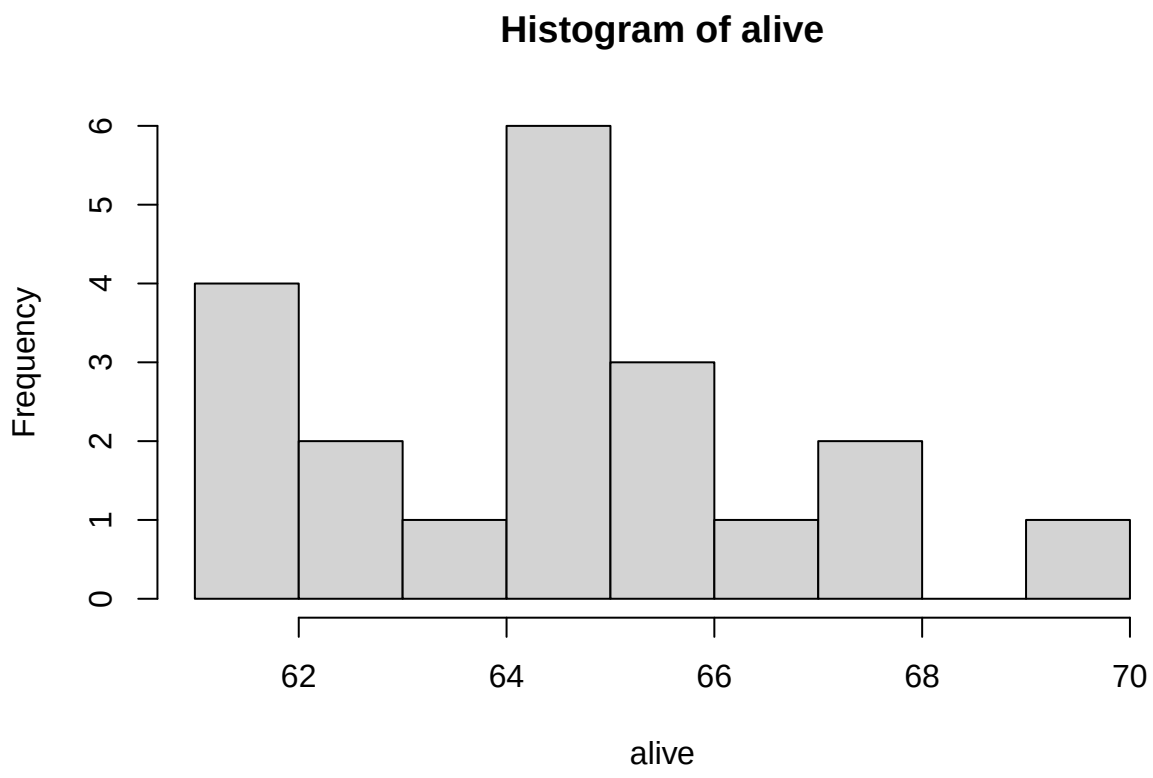
hypothesis testing Raza fall 2020

This dataset contains simulated data. It is based on the following covid dataset: https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset Import the data

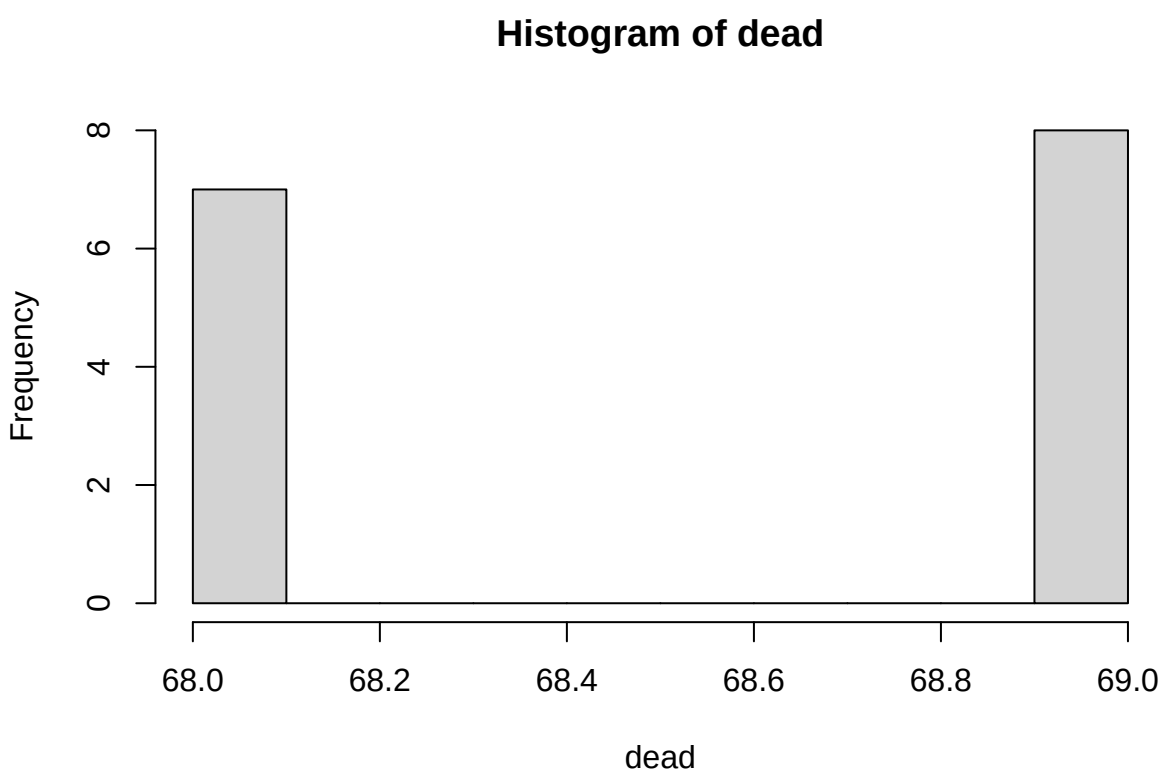
```
covid=read.csv("covid_simulated.csv")
```

Is there a difference between the mean age of people who survive and those who dont?

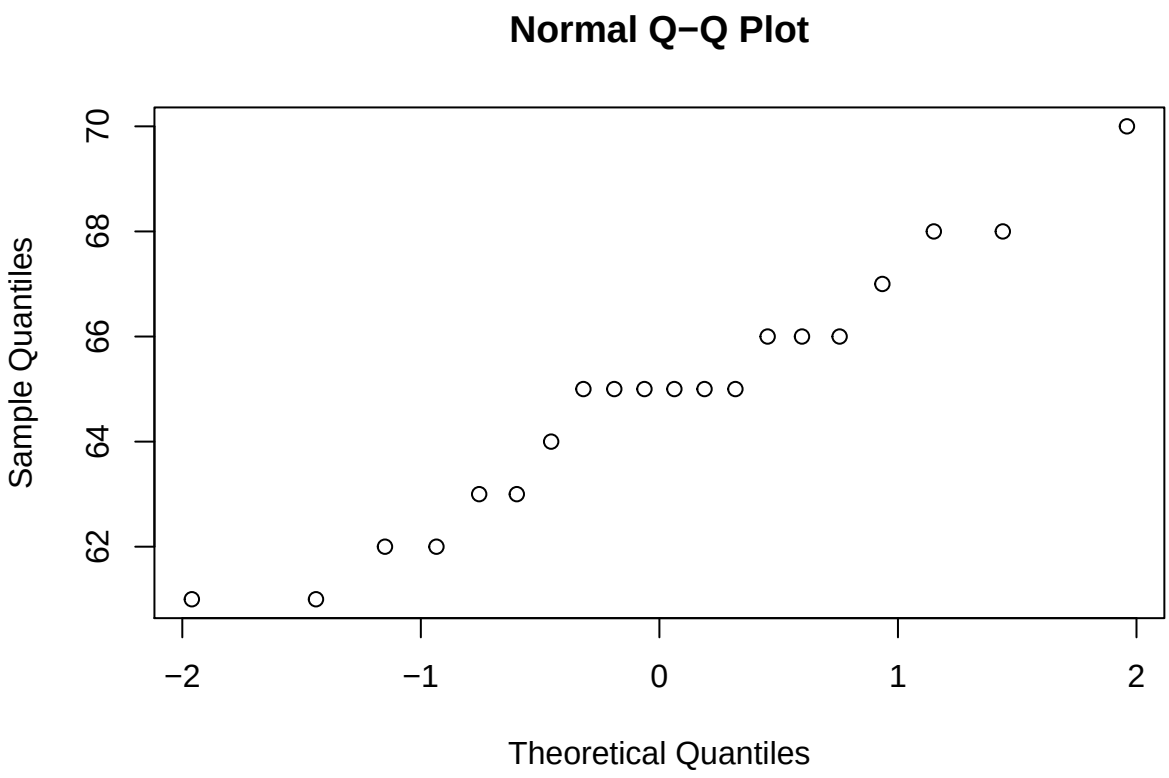
```
alive=covid[covid$Survival=="Yes","Age"]
dead=covid[covid$Survival=="No","Age"]

#check if the two groups are normal with histogram, qqnorm and shapiro.tets
hist(alive,breaks = 10)
```



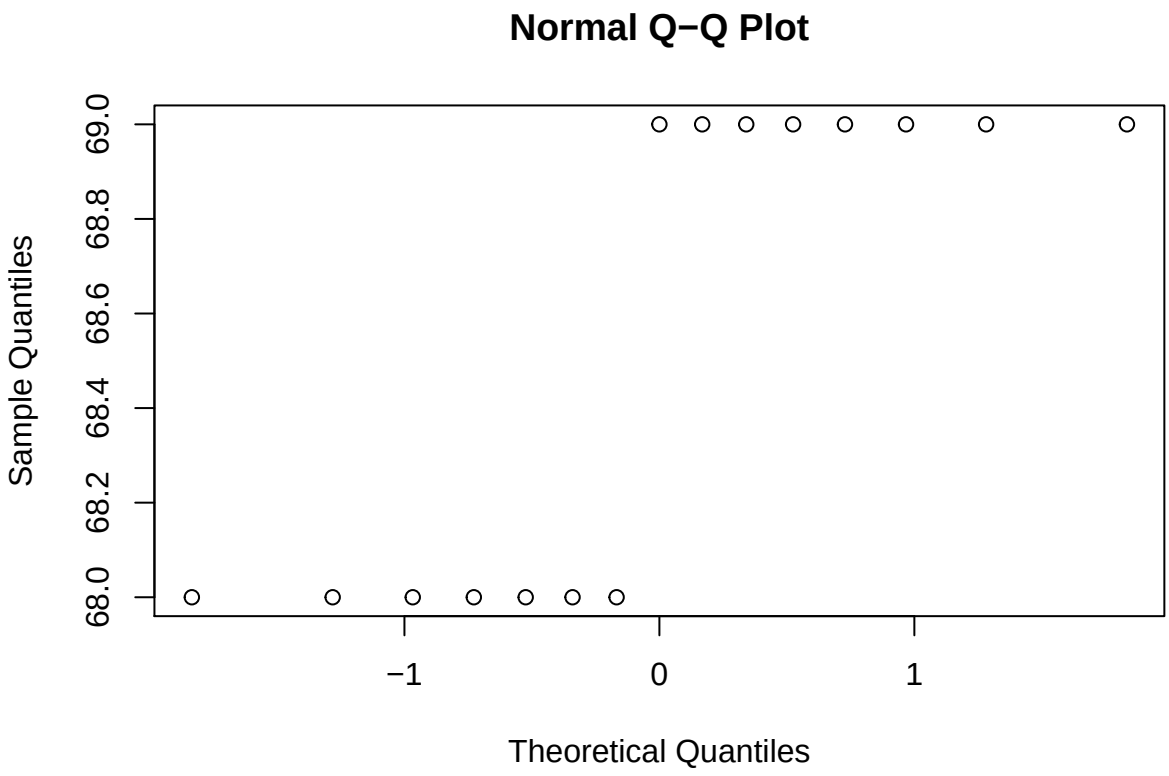

```
hist(dead,breaks = 10)
```




```
qqnorm(alive)
```


Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
62
64
66
68
70
−2
−1
0
1
2

```
qqnorm(dead)
```




```
shapiro.test(alive)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  alive
## W = 0.958, p-value = 0.5048
```

```
shapiro.test(dead)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dead
## W = 0.64341, p-value = 6.562e-05
```

```
#Age of dead patients is not normal. Use wilcox.test
wilcox.test(alive,dead,paired = F)
```

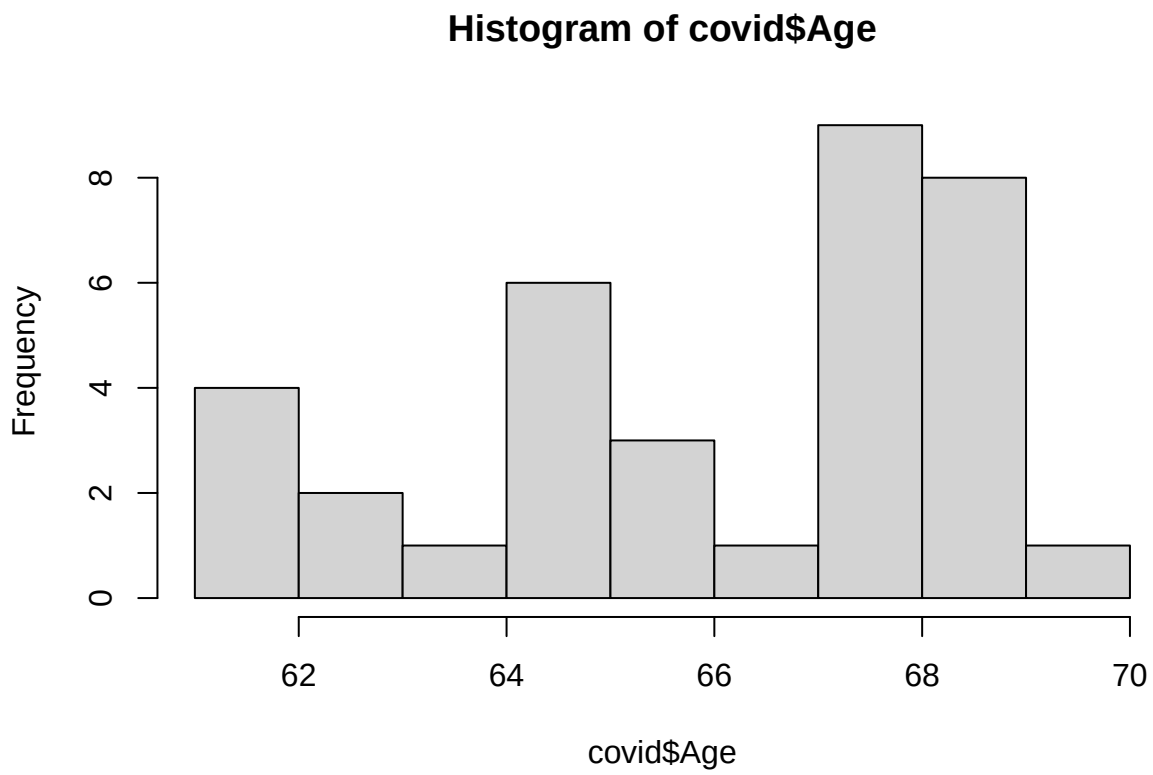
```
## Warning in wilcox.test.default(alive, dead, paired = F): cannot compute exact p-
## value with ties
```

```
##
```

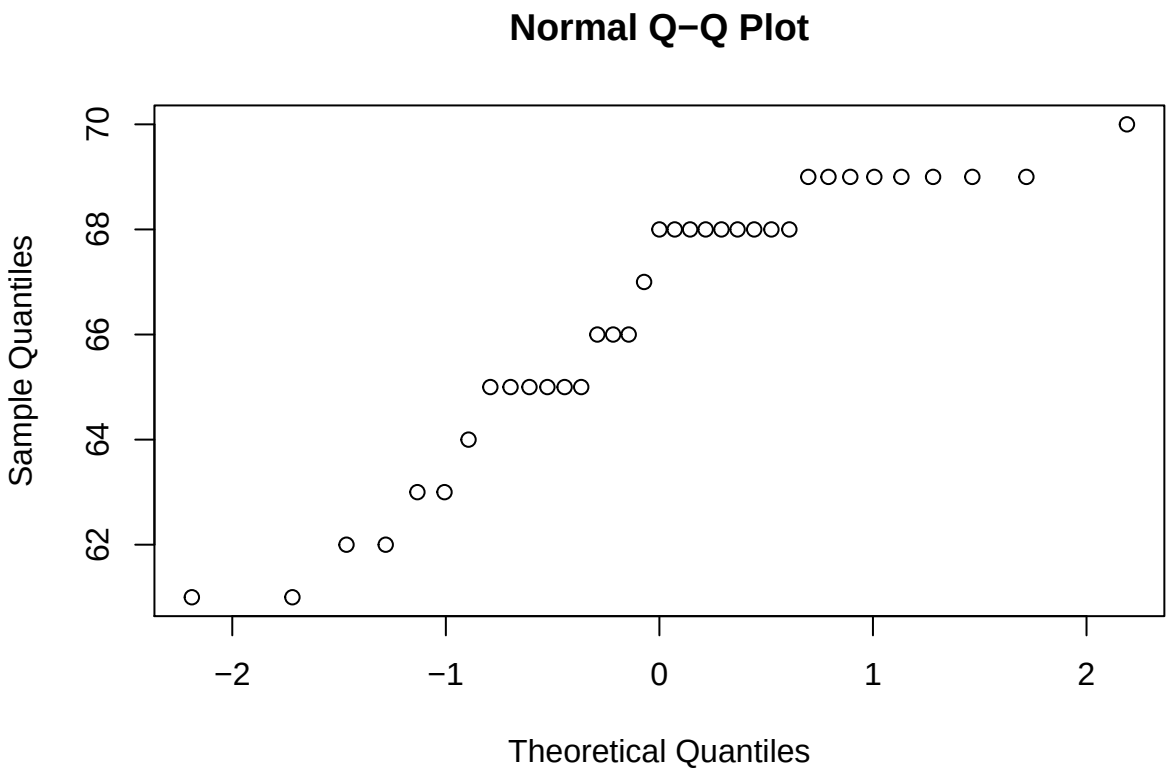
```
##  Wilcoxon rank sum test with continuity correction
##
## data:  alive and dead
## W = 22, p-value = 1.524e-05
## alternative hypothesis: true location shift is not equal to 0
```

Is the mean age of all covid patients greater than 60?

```
hist(covid$Age,breaks = 10)
```




```
qqnorm(covid$Age)
```




```
shapiro.test(log(covid$Age))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  log(covid$Age)
## W = 0.88424, p-value = 0.001531
```

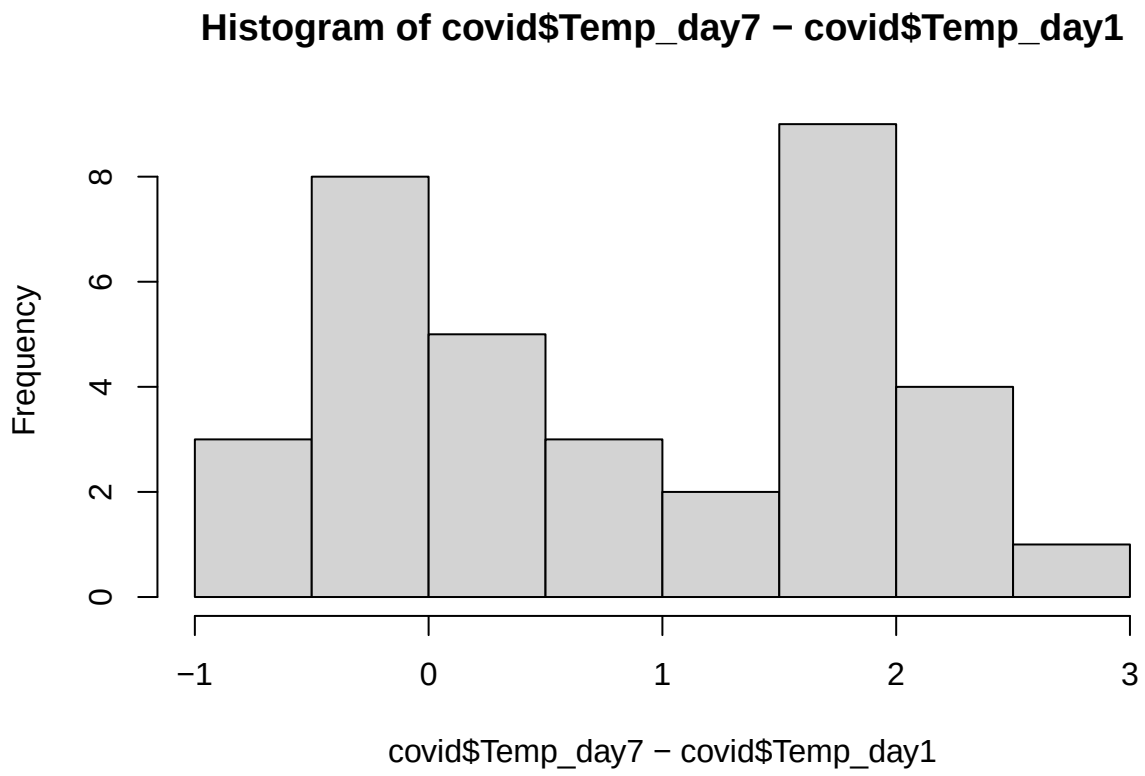
```
#data is not normal, use non-paramtric test
#but since sample size is large t.test might also work
wilcox.test(covid$Age,mu = 60,alternative = "g")
```

```
## Warning in wilcox.test.default(covid$Age, mu = 60, alternative = "g"): cannot
## compute exact p-value with ties
```

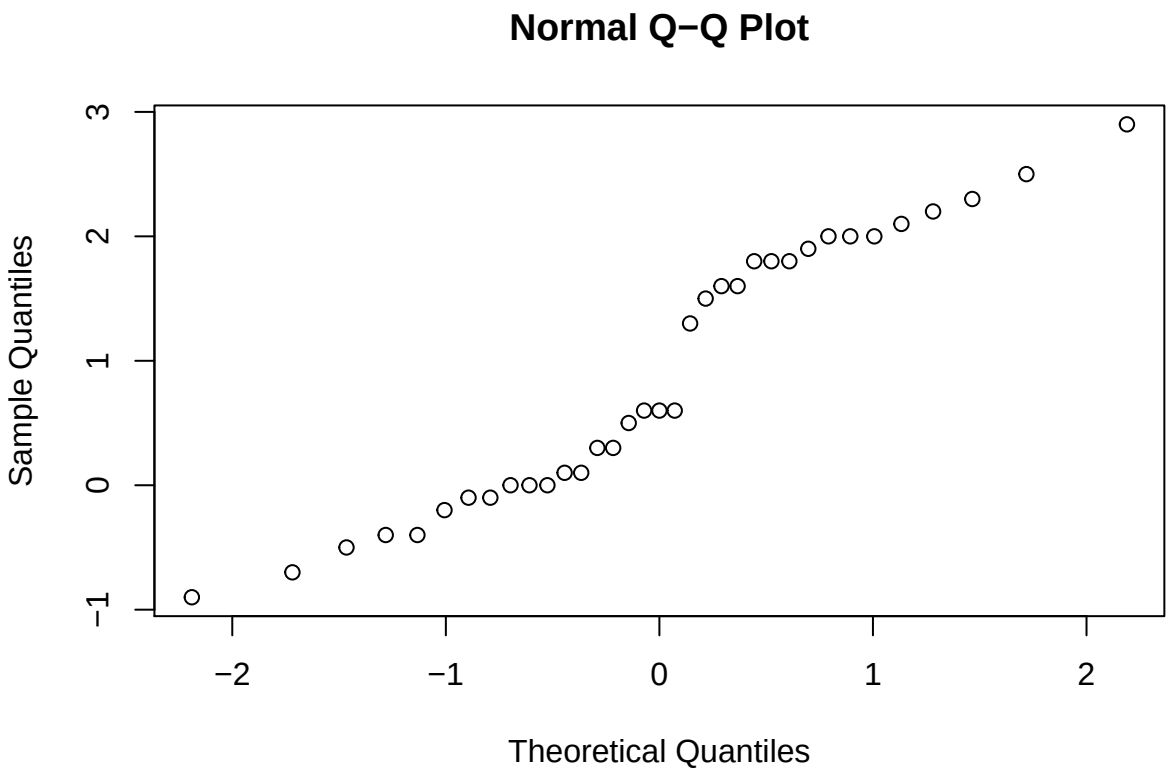
```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  covid$Age
## V = 630, p-value = 1.154e-07
## alternative hypothesis: true location is greater than 60
```

Is their a difference between the mean body temperature of all patients on day1 and day7 after infection

```
hist(covid$Temp_day7-covid$Temp_day1)
```




```
qqnorm(covid$Temp_day7-covid$Temp_day1)
```




```
shapiro.test(covid$Temp_day7-covid$Temp_day1)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  covid$Temp_day7 - covid$Temp_day1
## W = 0.9273, p-value = 0.02328
```

```
#data is not normal. Use non-parametric test with paired=T
#but since sample size is large paired t.test might also work

wilcox.test(covid$Temp_day1,covid$Temp_day7,paired = T)
```

```
## Warning in wilcox.test.default(covid$Temp_day1, covid$Temp_day7, paired = T):
## cannot compute exact p-value with ties

## Warning in wilcox.test.default(covid$Temp_day1, covid$Temp_day7, paired = T):
## cannot compute exact p-value with zeroes

##
##  Wilcoxon signed rank test with continuity correction
##
## data:  covid$Temp_day1 and covid$Temp_day7
## V = 67.5, p-value = 0.0002461
## alternative hypothesis: true location shift is not equal to 0
```

```
wilcox.test((covid$Temp_day1-covid$Temp_day7),mu=0)
```

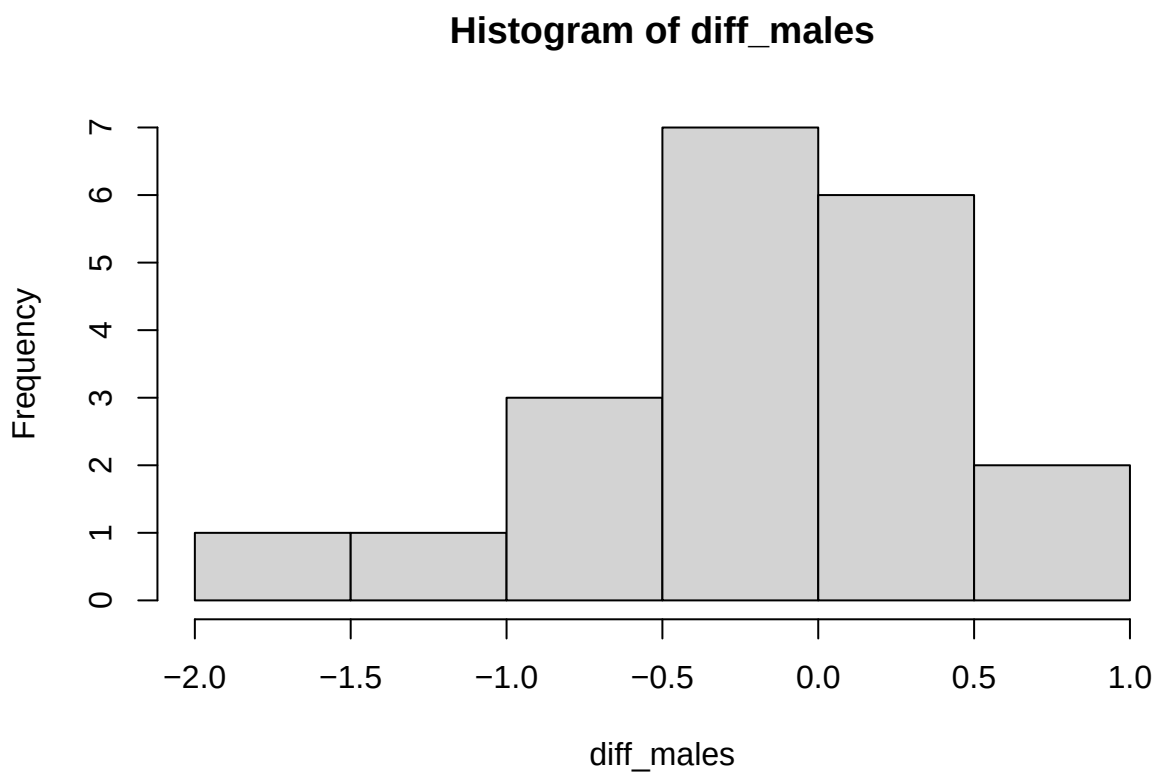
```
## Warning in wilcox.test.default((covid$Temp_day1 - covid$Temp_day7), mu = 0):
## cannot compute exact p-value with ties
```

```
## Warning in wilcox.test.default((covid$Temp_day1 - covid$Temp_day7), mu = 0):
## cannot compute exact p-value with zeroes
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  (covid$Temp_day1 - covid$Temp_day7)
## V = 67.5, p-value = 0.0002461
## alternative hypothesis: true location is not equal to 0
```

Is their a difference between the mean body temperature of male patients on day1 and day7 after infection

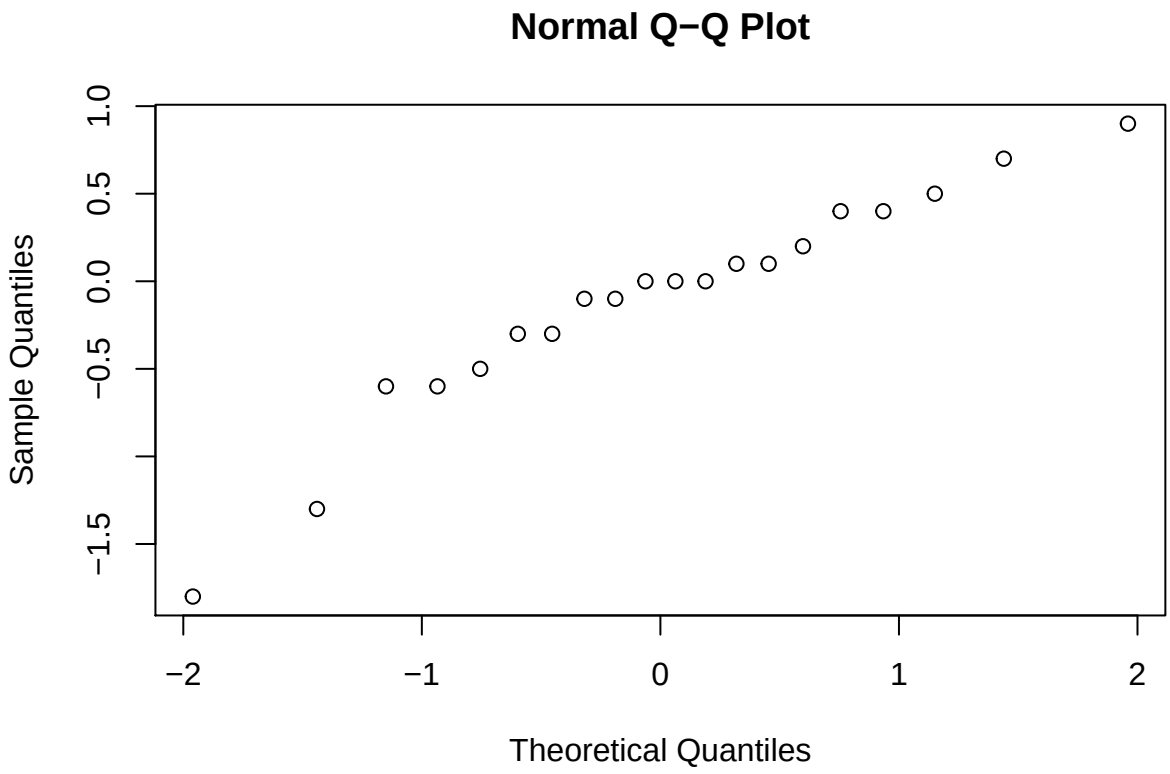
```
males_day1=covid[covid$Gender=="Male","Temp_day1"]
males_day7=covid[covid$Gender=="Male","Temp_day7"]
diff_males=males_day1 - males_day7
hist(diff_males)
```



```
shapiro.test(diff_males)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  diff_males
## W = 0.93097, p-value = 0.1612
```

```
qqnorm(diff_males)
```




```
#data is normal, use paired t.test
t.test(diff_males,mu=0)
```

```
##
##  One Sample t-test
##
## data:  diff_males
## t = -0.80446, df = 19, p-value = 0.4311
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.4142038  0.1842038
## sample estimates:
## mean of x
##    -0.115
```

```r
#or
t.test(males_day1,males_day7,paired = T)
```

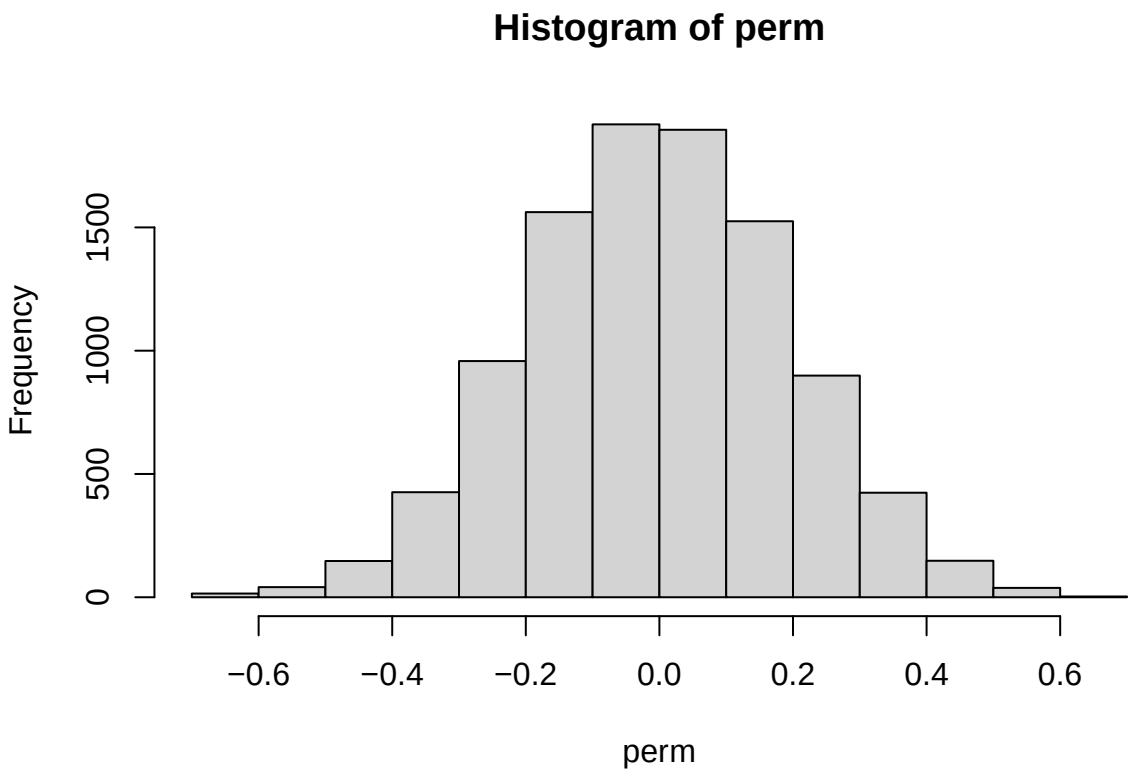
```
##
##  Paired t-test
##
## data:  males_day1 and males_day7
## t = -0.80446, df = 19, p-value = 0.4311
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.4142038  0.1842038
## sample estimates:
## mean of the differences
##                  -0.115
```

```r
#manual
t=mean(diff_males)/(sd(diff_males)/sqrt(length(diff_males)))
2*pt(t,df=19,lower.tail = T)
```

```
## [1] 0.4310823
```

Use a permutation test to answer the previous question and calculate the p-value

```r
diff=mean(males_day1)-mean(males_day7)

p=c(males_day1,males_day7)
fac=c(rep("one",10),rep("seven",10))

perm=as.numeric()

for (i in 1:10000){
perm[i]=diff(tapply(sample(p,20,replace = F), fac, mean))
  }

hist(perm)
```




```
sum(abs(perm)>=abs(diff))/10000
```

```
## [1] 0.5655
```

Fever is defined as body temp of >100.4 F. Is there any relationship between the patients gender and having fever on day 7 after infection?

```
covid$Fever=covid$Temp_day7>100.4
tab=table(covid$Gender,covid$Fever)

#Use fisher exact test as assumptions of chisq are not met
chisq.test(tab)
```

```
## Warning in chisq.test(tab): Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tab
## X-squared = 7.0444, df = 1, p-value = 0.007951
```

```
fisher.test(tab)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  tab
## p-value = 0.003083
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.0000000 0.4959655
## sample estimates:
## odds ratio
##          0
```