Lab 2

Revisit the NormTemp dataset from Lab 1, where we examined the observed mean body temperature (temperature) in comparison to the well-known "average" of 98.6.

```r
library(UsingR)
```

```
## Warning: package 'UsingR' was built under R version 4.2.3

## Warning: package 'HistData' was built under R version 4.2.3
```

```r
library(tidyverse)
```

```
## Warning: package 'ggplot2' was built under R version 4.2.3

## Warning: package 'tibble' was built under R version 4.2.3

## Warning: package 'tidyr' was built under R version 4.2.3

## Warning: package 'purrr' was built under R version 4.2.3

## Warning: package 'dplyr' was built under R version 4.2.3

## Warning: package 'stringr' was built under R version 4.2.3
```

Perform a statistical test (alpha = 0.05) to determine whether this well-known number is actually the mean body temperature. What is your p-value? Explain in words what this p-value means. What is your conclusion?

The p value of the one sample t-test is 2.411e-07, lower than our significance level, meaning that we reject the Null Hypothesis and average temperature is significantly different than 98.6 degrees

```r
library(UsingR)
t.test(normtemp$temperature, mu=98.6, alternative = 'two.sided', p.value=0.05)
```

```
##
##  One Sample t-test
##
## data:  normtemp$temperature
## t = -5.4548, df = 129, p-value = 2.411e-07
## alternative hypothesis: true mean is not equal to 98.6
## 95 percent confidence interval:
##  98.12200 98.37646
## sample estimates:
## mean of x
##  98.24923
```

Give the 95% Confidence Interval for temperature. Explain in words what a 95% confidence interval represents.

A 95% confidnece interval is a range of values that we are 95% confident contains the true population parameter.

```
t.test(normtemp$temperature, mu=98.6, conf.level = 0.95)
```

```
##
##  One Sample t-test
##
## data:  normtemp$temperature
## t = -5.4548, df = 129, p-value = 2.411e-07
## alternative hypothesis: true mean is not equal to 98.6
## 95 percent confidence interval:
##  98.12200 98.37646
## sample estimates:
## mean of x
##  98.24923
```

If we restrict our analysis to only the females in this dataset, would our conclusion change?

The confidence interval is closer to the 98.6 value, but the value is significantly different than the expected average temperature.

```
fem_temp <- normtemp %>% filter(gender == 2)
t.test(fem_temp$temperature, mu=98.6, conf.level = 0.95)
```

```
##
##  One Sample t-test
##
## data:  fem_temp$temperature
## t = -2.2355, df = 64, p-value = 0.02888
## alternative hypothesis: true mean is not equal to 98.6
## 95 percent confidence interval:
##  98.20962 98.57807
## sample estimates:
## mean of x
##  98.39385
```

Is there any difference (alpha=0.05) in temperature between the two genders recorded in this dataset (be sure to look at assumptions and perform the correct test)?

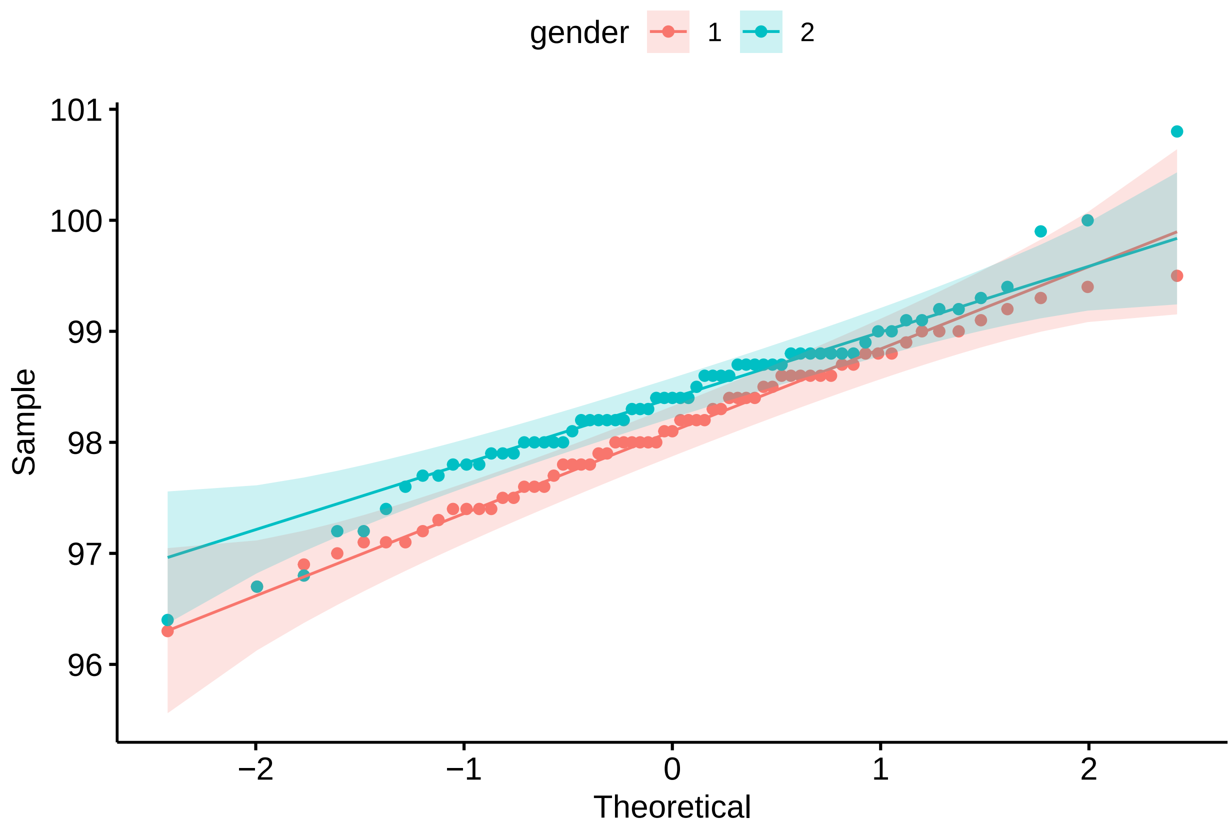
Assumptions: - Normality of both groups: True, QQ-plot and Shapiro-Wilks show normality in both groups - Independence between the groups: True, in this dataset there are no observations that are both male and female. - Equal variance between groups

There is a significant difference in the average of temperature of the genders.

```r
library(ggpubr)
```

```
## Warning: package 'ggpubr' was built under R version 4.2.3
```

```r
normtemp <- normtemp %>% mutate(gender = as.factor(gender))
# Test normality
ggqqplot(normtemp, x="temperature", color = "gender")
```




```r
shapiro.test(normtemp$temperature[normtemp$gender == 1])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  normtemp$temperature[normtemp$gender == 1]
## W = 0.98941, p-value = 0.8545
```

```r
shapiro.test(normtemp$temperature[normtemp$gender == 2])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  normtemp$temperature[normtemp$gender == 2]
## W = 0.96797, p-value = 0.09017
```

```r
# Test variance
var.test(temperature~gender,data=normtemp)
```

```
##
##  F test to compare two variances
##
## data:  temperature by gender
## F = 0.88329, num df = 64, denom df = 64, p-value = 0.6211
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.5387604 1.4481404
## sample estimates:
## ratio of variances
##          0.8832897
```

```r
t.test(temperature ~ gender, data=normtemp)
```

```
##
##  Welch Two Sample t-test
##
## data:  temperature by gender
## t = -2.2854, df = 127.51, p-value = 0.02394
## alternative hypothesis: true difference in means between group 1 and group 2 is not equal to 0
## 95 percent confidence interval:
##  -0.53964856 -0.03881298
## sample estimates:
## mean in group 1 mean in group 2
##        98.10462        98.39385
```

The Airline dataset contains information regarding the number of international airline travelers (variable air) across different months of the year from 1949-1960. To obtain this data set, you will need to:

```r
data("AirPassengers")
```

We are interested in knowing if during this time period there was a significant difference between air travel in the Summer months of June, July, and August vs. the remainder of the year? Use a statistical hypothesis test (alpha=0.05) to support your answer. In order to get month information, you will need to:

There is a significant difference in average air travel between the summer months and non-summer months

```r
# install.packages('tseries')
# install.packages('forecast')

library(tseries)
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```r
library(forecast)
```

```
##
## Attaching package: 'forecast'

## The following object is masked from 'package:ggpubr':
##
##     gghistogram
```

```r
cycle(AirPassengers)
```

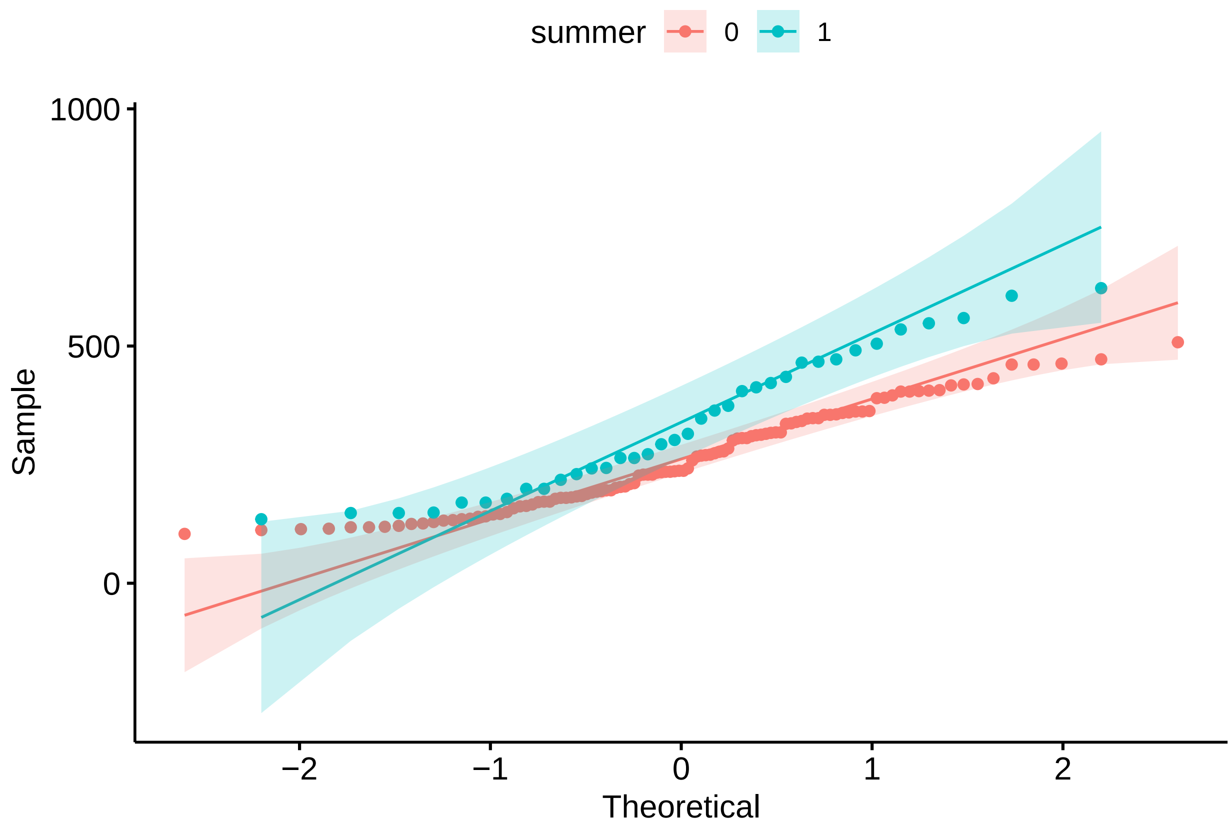
```
##      Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
## 1949   1   2   3   4   5   6   7   8   9  10  11  12
## 1950   1   2   3   4   5   6   7   8   9  10  11  12
## 1951   1   2   3   4   5   6   7   8   9  10  11  12
## 1952   1   2   3   4   5   6   7   8   9  10  11  12
## 1953   1   2   3   4   5   6   7   8   9  10  11  12
## 1954   1   2   3   4   5   6   7   8   9  10  11  12
## 1955   1   2   3   4   5   6   7   8   9  10  11  12
## 1956   1   2   3   4   5   6   7   8   9  10  11  12
## 1957   1   2   3   4   5   6   7   8   9  10  11  12
## 1958   1   2   3   4   5   6   7   8   9  10  11  12
## 1959   1   2   3   4   5   6   7   8   9  10  11  12
## 1960   1   2   3   4   5   6   7   8   9  10  11  12
```

```r
air1 = data.frame(AirPassengers)
air2 = air1 %>% mutate(summer=ifelse(cycle(AirPassengers) %in% 6:8,1,0))
air2$summer <- as.factor(air2$summer)
```

Normality

```r
ggqqplot(air2, x = 'AirPassengers', color='summer')
```

```
## Don't know how to automatically pick scale for object of type <ts>. Defaulting
## to continuous.
## Don't know how to automatically pick scale for object of type <ts>. Defaulting
## to continuous.
## Don't know how to automatically pick scale for object of type <ts>. Defaulting
## to continuous.
```



Equivariant: False

```
var.test(AirPassengers~summer, data=air2)
```

```
##
##  F test to compare two variances
##
## data:  AirPassengers by summer
## F = 0.50274, num df = 107, denom df = 35, p-value = 0.007707
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.2801247 0.8360404
## sample estimates:
## ratio of variances
##          0.5027447
```

```
t.test(AirPassengers~summer, data= air2, var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  AirPassengers by summer
## t = -2.9233, df = 47.279, p-value = 0.0053
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
```

```
## -129.93398 -24.01046
## sample estimates:
## mean in group 0 mean in group 1
##        261.0556        338.0278
```