# Homework 3

*Alexander Van Roijen*

*January 25, 2019*

## Homework 3

## Authors: Alexander Van Roijen, Frank Chen, John Mahoney

### Problem 1(From excerise 3 problem 3)

We first tackled this problem by trying to understand the distribution of the mean differences that we do have available. Thus below, we examine the standard deviations of both populations and analyze what their combined variance would be in a new normal distribution.

```
processData = read.csv('process.csv')
lowtempvals = processData[processData$temp==50,]
hightempvals = processData[processData$temp==100,]

sd_diff_est_100_50 <- sqrt(var(hightempvals$output) + var(lowtempvals$output))
sd_diff_est_100_100 <- sqrt(var(hightempvals$output) + var(hightempvals$output))
print(paste("Standard Deviation of Difference with SD of Temp 120 as SD of Temp 50: ",
            sd_diff_est_100_50 , sep=""))
```

```
## [1] "Standard Deviation of Difference with SD of Temp 120 as SD of Temp 50: 220.213624777337"
```

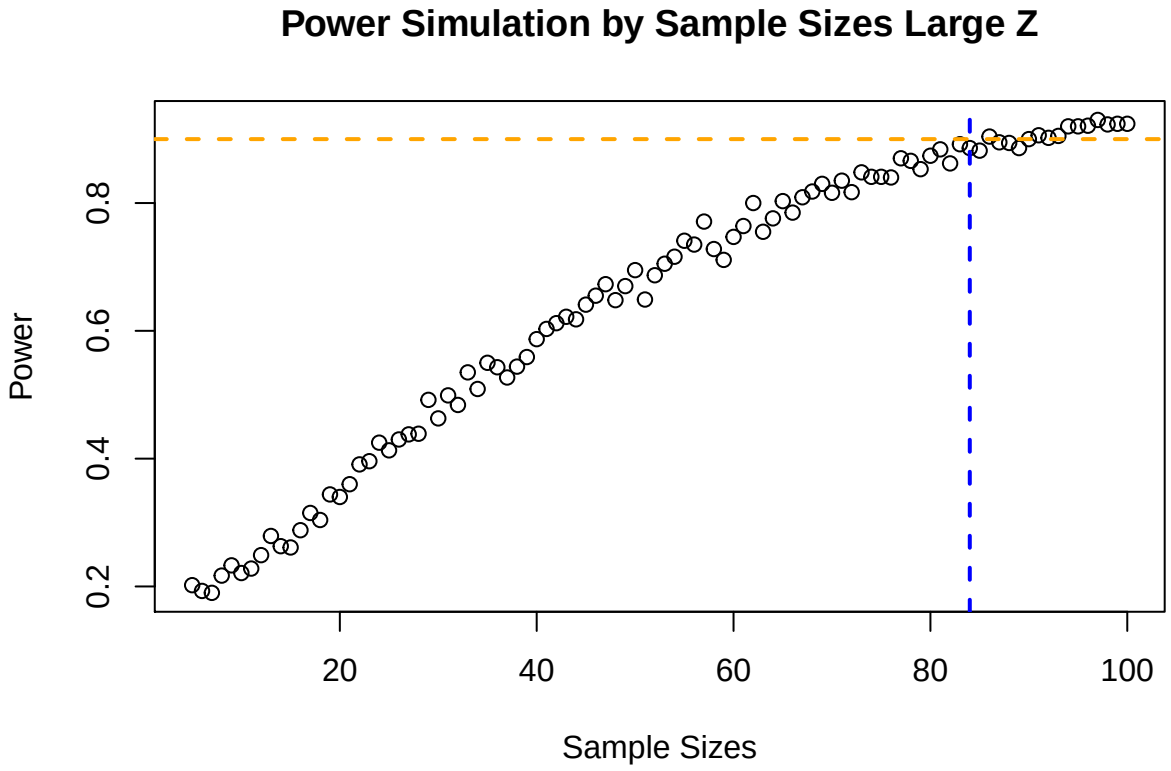
```
print(paste("Standard Deviation of Difference with SD of Temp 120 as SD of Temp 100: ",
            sd_diff_est_100_100 , sep=""))
```

```
## [1] "Standard Deviation of Difference with SD of Temp 120 as SD of Temp 100: 204.458338775336"
```

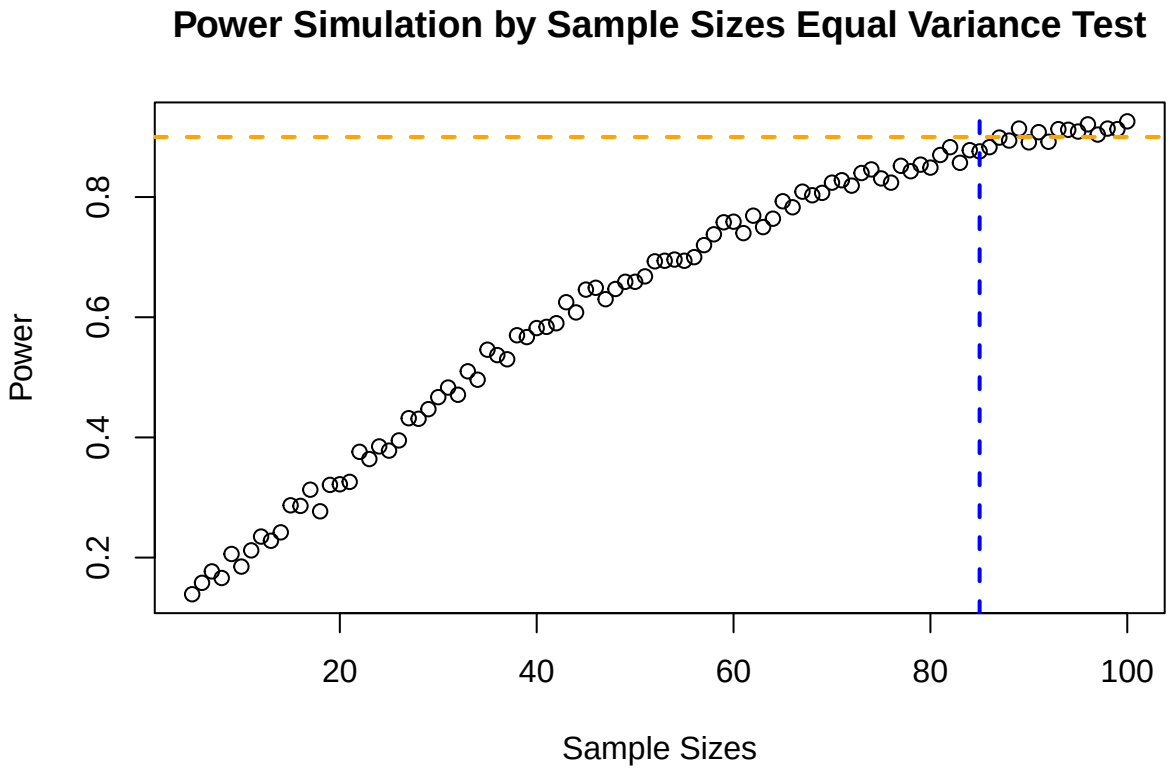
Using the variances above, we extrapolate that the normal distribution that we assume to be the case for the mean differences should have a standard deviation around the values listed above. Thus in the tests below, we will attempt various sample sizes on what this aggregated distribution would follow with rounded standard deviations shown above. We are taking a slightly different approach to standard procedure as we are simulating the Z distribution we believe to be followed by the mean differences between the 100 and 120 data, rather than simulating the two distributions independently. We apply this as we are more confident in what the aggregated hypothesis distribution will follow rather than what the 120 temperature distribution follows. We apply the simulated data in a large sample Z test, and equal variance t test. We believe that if the powers are achieved at similar sample sizes with the assumption of equal variance, then the Welch test will provide little additional information. Their powers over sample size are plotted below.

```
power_simulation <- function(n_list, sd_input) {
  N <- 1000
  power_list_z <- rep(NA, length(n_list))
  power_list_eq_var <- rep(NA, length(n_list))
  for (i in 1:length(n_list)) {
    sim_result <- replicate(N, {
      sim_data <- rnorm(n_list[i], mean=75, sd=sd_input)
      Z <- mean(sim_data)/(sd(sim_data)/sqrt(n_list[i]))
      abs(Z) > qnorm(0.975)
    })
    power_list_z[i] <- mean(sim_result)
  }
```

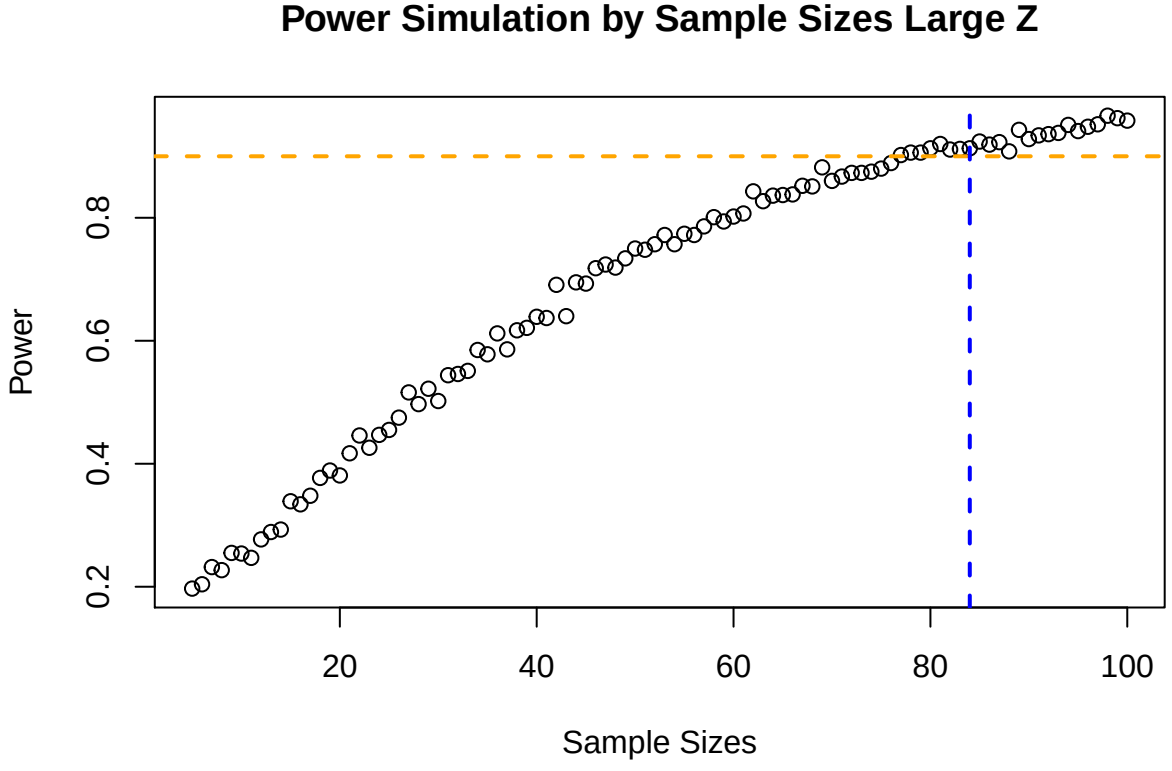
```r
  for (i in 1:length(n_list)) {
    sim_result <- replicate(N, {
      sim_data <- rnorm(n_list[i], mean=75, sd=sd_input)
      Z <- mean(sim_data)/(sd(sim_data)/sqrt(n_list[i]))
      abs(Z) > qt(0.975, n_list[i]*2-2)
    })
    power_list_eq_var[i] <- mean(sim_result)
  }
  return(list(power_list_z=power_list_z, power_list_eq_var=power_list_eq_var))
}
set.seed(1272019)
n_list <- c(5:100)
results <- power_simulation(n_list, 219)
plot(n_list, results$power_list_z, ylab="Power",
     xlab="Sample Sizes", main="Power Simulation by Sample Sizes Large Z")
abline(h=0.9, lty=2, col="orange", lwd=2)
abline(v=84, lty=2, col="blue", lwd=2)
```



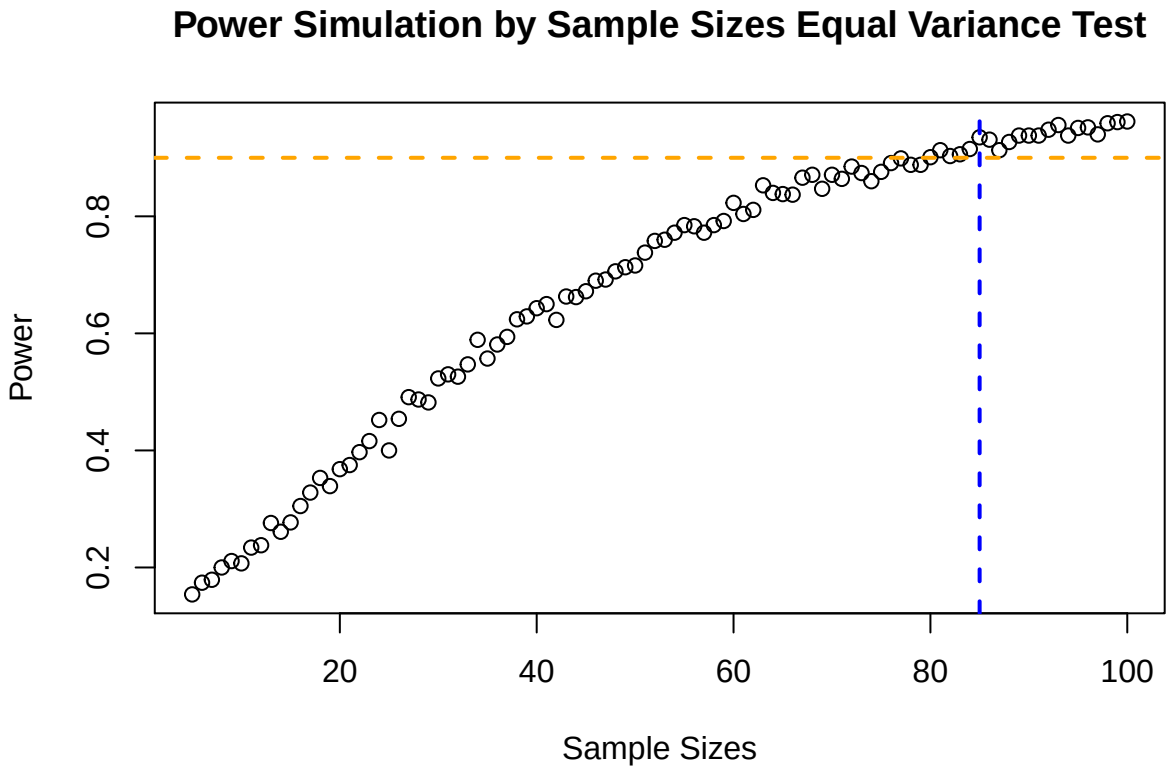

```r
plot(n_list, results$power_list_eq_var, ylab="Power",
     xlab="Sample Sizes", main="Power Simulation by Sample Sizes Equal Variance Test")
abline(h=0.9, lty=2, col="orange", lwd=2)
abline(v=85, lty=2, col="blue", lwd=2)
```




```
results <- power_simulation(n_list, 204)
plot(n_list, results$power_list_z, ylab="Power",
     xlab="Sample Sizes", main="Power Simulation by Sample Sizes Large Z")
abline(h=0.9, lty=2, col="orange", lwd=2)
abline(v=84, lty=2, col="blue", lwd=2)
```



```r
plot(n_list, results$power_list_eq_var, ylab="Power",
     xlab="Sample Sizes", main="Power Simulation by Sample Sizes Equal Variance Test")
abline(h=0.9, lty=2, col="orange", lwd=2)
abline(v=85, lty=2, col="blue", lwd=2)
```




As we can see, both the equal variance t test and large sample z test show an approximate sample size of size 83~84. This is also similar for the lower variance we deemed may be possible given the decline in standard deviation we see with increasing temperature. This is good to see as it shows a convergence we would expect with the larger sample size. On the err of caution, we will state that we believe the ideal sample size for a mean difference of seventy five with power of at least 90 can be achieved at a sample size of 85 assuming a significance level of .05. Looking back, it may have been better to simulate two different distributions and then run our tests rather that assume we can simulate what their sample distribution looks like after they have been combined. However this first approach seems somewhat logical considering lack of prior backrground.

## Problem 2

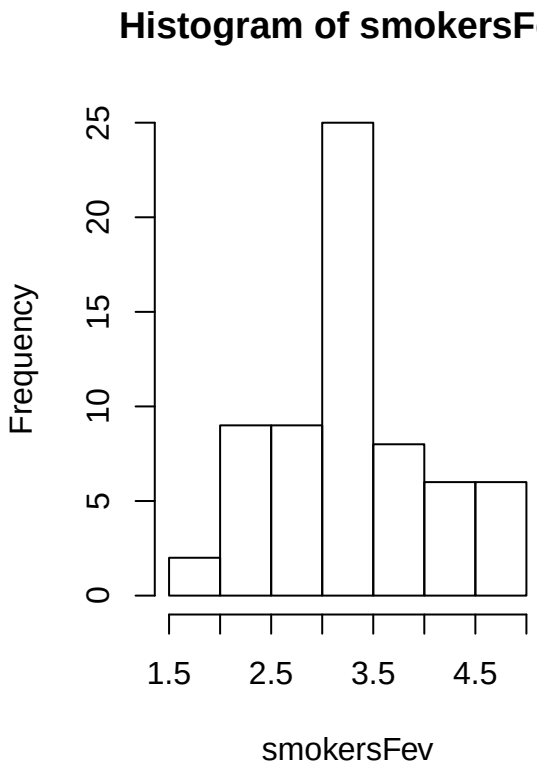
```r
fevData = read.csv('fev.csv')
smokers = fevData[fevData$smoke==1,]
nonSmokers = fevData[fevData$smoke==0,]
smokerSS = length(smokers[,1])
print(smokerSS)
```

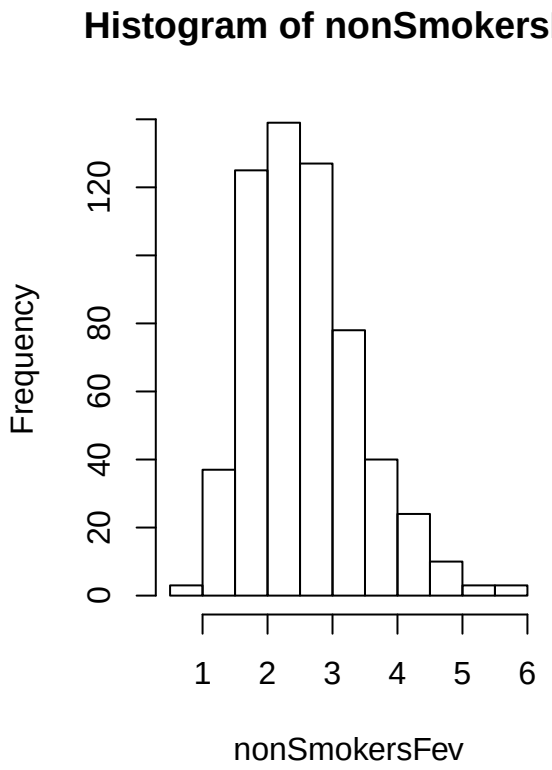
```
## [1] 65
```

```r
nonSmokerSS = length(nonSmokers[,1])
print(nonSmokerSS)
```

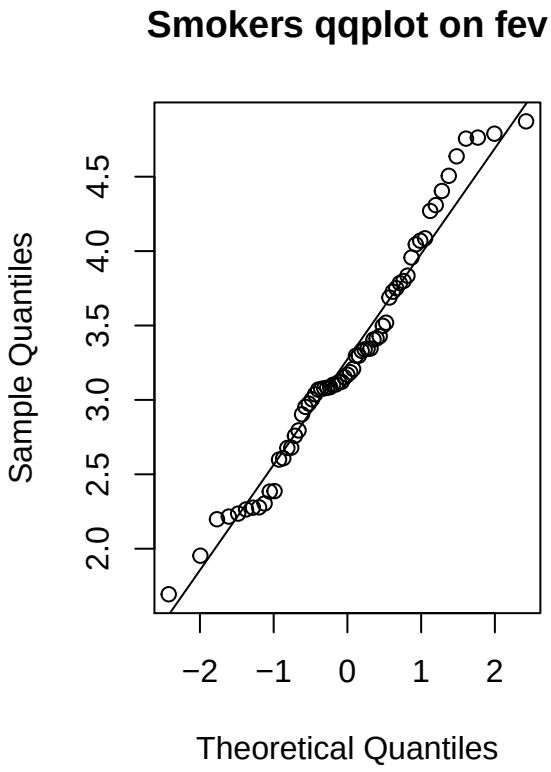
```
## [1] 589
```

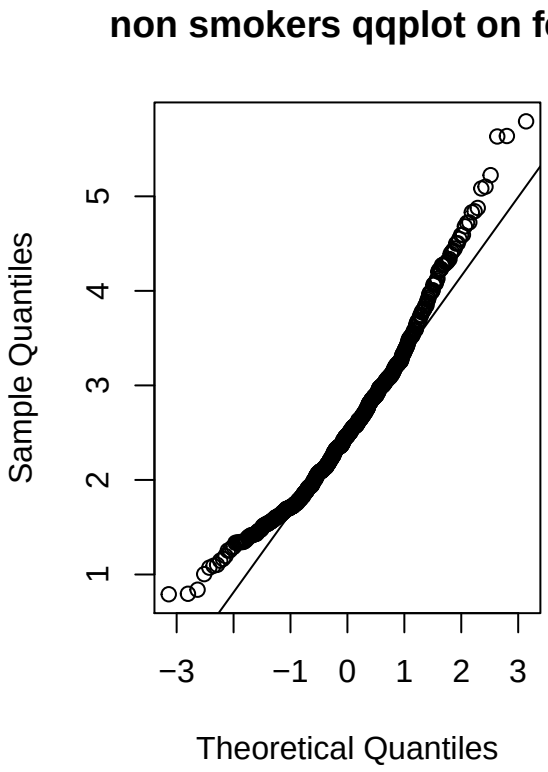
```r
par(mfrow = c(1,2))
smokersFev = smokers$fev
nonSmokersFev = nonSmokers$fev
hist(smokersFev)
hist(nonSmokersFev)
```




```r
qqnorm(smokersFev,main = "Smokers qqplot on fev")
qqline(smokersFev)
qqnorm(nonSmokersFev, main = "non smokers qqplot on fev")
qqline(nonSmokersFev)
```




```
smokerVar = var(smokersFev)
smokerMean = mean(smokersFev)
print(smokerVar)
```

```
## [1] 0.5624795
```

```
print(smokerMean)
```

```
## [1] 3.276862
```

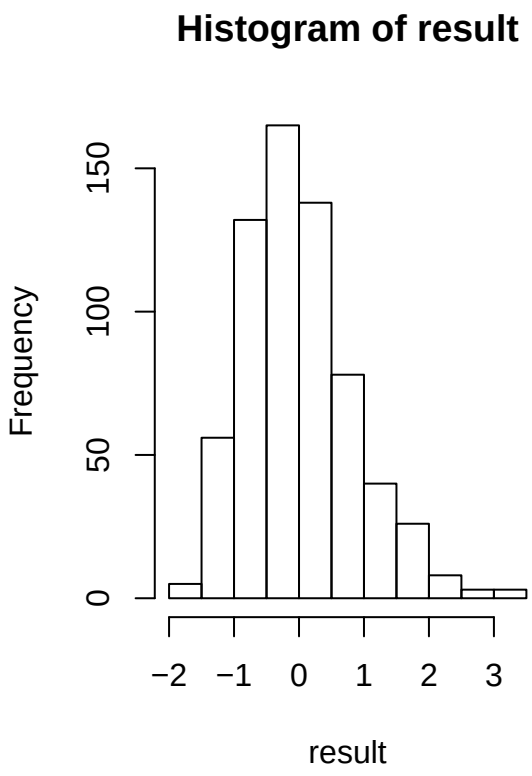
```
nonSmokerVar = var(nonSmokersFev)
nonSmokerMean = mean(nonSmokersFev)
print(nonSmokerVar)
```

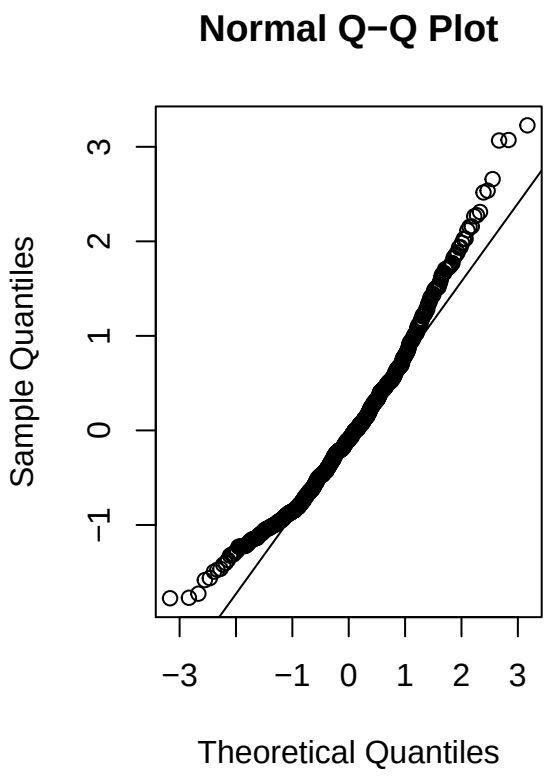
```
## [1] 0.7233869
```

```
print(nonSmokerMean)
```

```
## [1] 2.566143
```

```
smokerResid = smokersFev - smokerMean
nonSmokerResid = nonSmokersFev - nonSmokerMean
result = append(smokerResid,nonSmokerResid)
hist(result)
qqnorm(result)
qqline(result)
```




**Problem 2.1**

We are put into an interesting position. The data looks somewhat normal at first glance, particularly for the nonSmokers data except for its slight skew, but we can see that the qqplot indicates there is a rather heavy set of tails. Further, the plot for the smokers has a large central peak, but drops off quite rapidly. Looking at their variances, they seem somewhat far apart, however that may be attributed to the smaller amount of samples we have of smokers. We can see there are only 65 data points for our smokers and 589 for our non smokers. Finally, looking at their residuals, we see that it looks again somehwat normal with heavy-ish tails.Overall, I will assume a normal distribution for both data sets and simulate more samples with means and variances equivalent to the means and variances from the sample we have. I will also sample the same size population from each distribution as I do not want to assume equal sample sizes. Finally, I will attempt all tests and determine which test best fits the data.

```
numSimulations = 10000
zTestRes = 0
tTestRes = 0
wTestRes = 0
for( i in 1:numSimulations)
{
  generatedSampleSmokers = rnorm(smokerSS,mean=1.5,sd=sqrt(smokerVar))
  generatedSampleNonSmokers = rnorm(nonSmokerSS,mean=1.5,sd=sqrt(nonSmokerVar))
  smokersd = sd(generatedSampleSmokers)
  nonsmokersd = sd(generatedSampleNonSmokers)
  tms = mean(generatedSampleSmokers)
  tmns = mean(generatedSampleNonSmokers)
  se = sqrt((smokersd**2)/smokerSS + (nonsmokersd**2)/nonSmokerSS)
  zscore = abs(tms-tmns)/se
  if(zscore >1.96)
  {
```

```
    zTestRes = zTestRes+ 1
  }
  evTest = t.test(generatedSampleSmokers,generatedSampleNonSmokers,var.equal=T)
  nevTest = t.test(generatedSampleSmokers,generatedSampleNonSmokers,var.equal=F)
  if(evTest$p.value<0.05)
  {
    tTestRes = tTestRes + 1
  }
  if(nevTest$p.value<0.05)
  {
    wTestRes=wTestRes+1
  }
}
  print(zTestRes/numSimulations)
```

```
## [1] 0.0542
```

```
  print(tTestRes/numSimulations)
```

```
## [1] 0.03
```

```
  print(wTestRes/numSimulations)
```

```
## [1] 0.0497
```

Looking at the results of our simulation, we can see that the Welch test has the closest accuracy to the 0.05 significance level. Thus it appears that would be the best fit given the situation of our data, particularly dealing with the mixed sample size and difference in variance.

### Problem 2.2

```
smokeVnonSmokeTT =t.test(smokersFev,nonSmokersFev,var.equal=F)

print((smokeVnonSmokeTT))
```

```
##
##  Welch Two Sample t-test
##
## data:  smokersFev and nonSmokersFev
## t = 7.1496, df = 83.273, p-value = 3.074e-10
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.5130126 0.9084253
## sample estimates:
## mean of x mean of y
##  3.276862  2.566143
```

Judging by the results of our t-test above, we would reject the null hypothesis that the FEV between children who smoke and do not smoke is the same. Clearly, since our confidence interval does not contain 0 at this significance level, we can be confident that they are in all likelyhood not the same.

### Problem 2.3

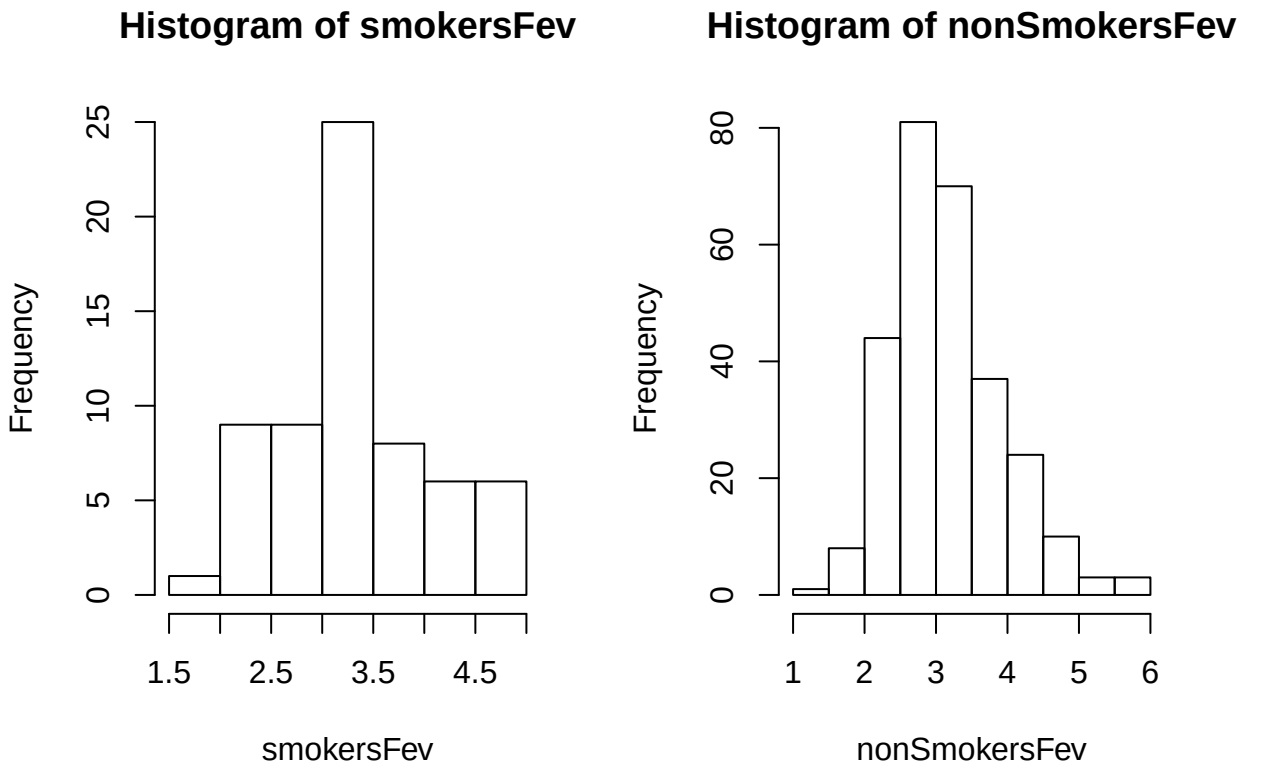
```
fevData = read.csv('fev.csv')
altFevData = fevData[fevData$age>=10,]
smokers = altFevData[altFevData$smoke==1,]
nonSmokers = altFevData[altFevData$smoke==0,]
print(length(smokers[,1]))
```

```
## [1] 64
```

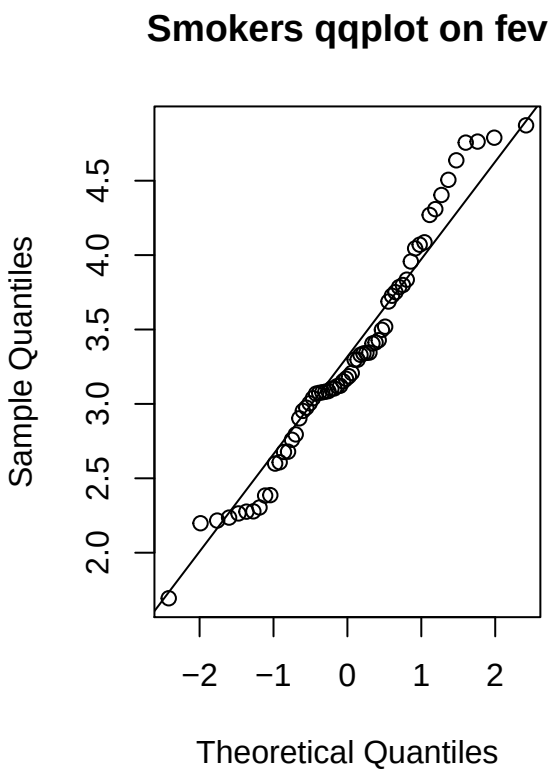
```
print(length(nonSmokers[,1]))
```

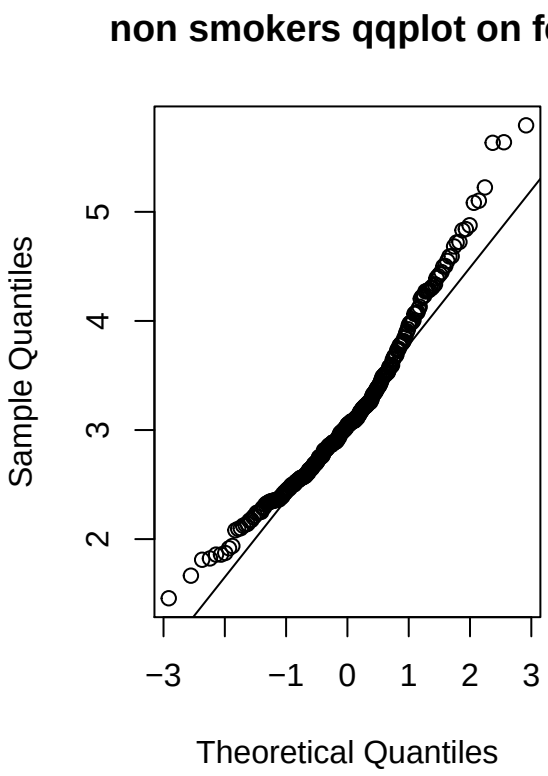
```
## [1] 281
```

```
par(mfrow = c(1,2))
smokersFev = smokers$fev
nonSmokersFev = nonSmokers$fev
hist(smokersFev)
hist(nonSmokersFev)
```




```
qqnorm(smokersFev,main = "Smokers qqplot on fev")
qqline(smokersFev)
qqnorm(nonSmokersFev, main = "non smokers qqplot on fev")
qqline(nonSmokersFev)
```



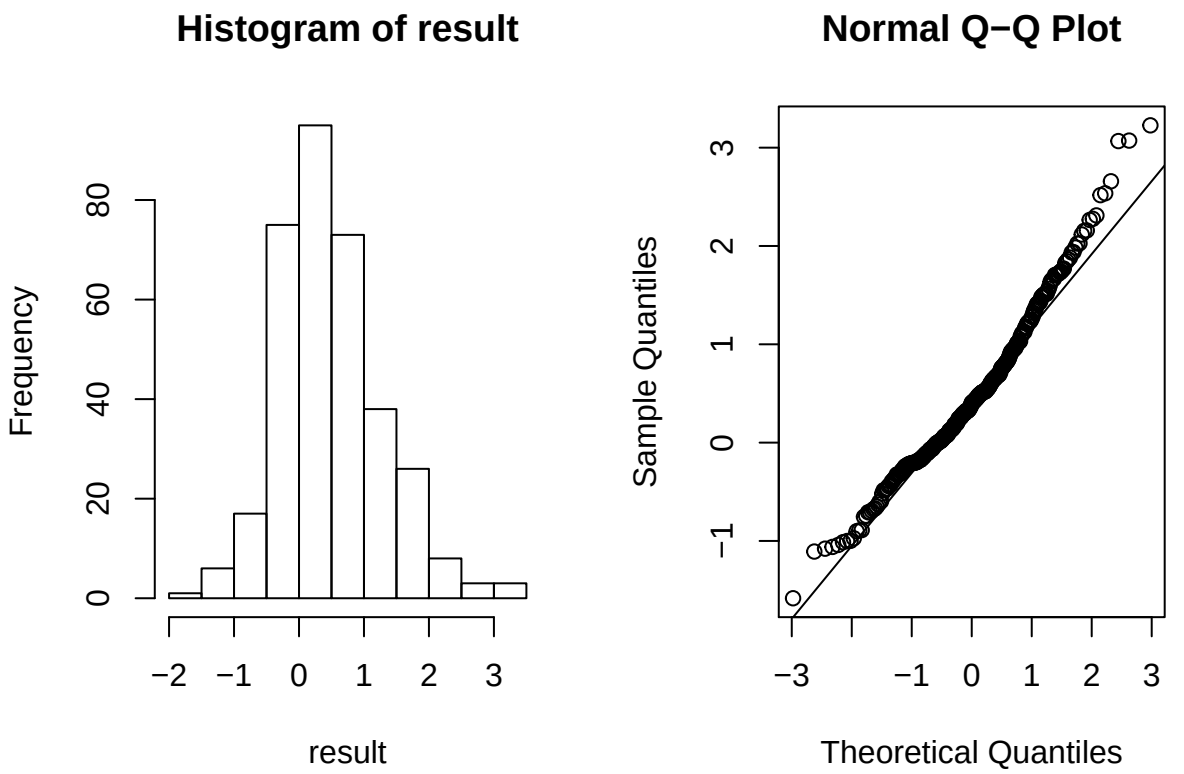

```
print(var(smokersFev))
```

```
## [1] 0.5431538
```

```
print(var(nonSmokersFev))
```

```
## [1] 0.57611
```

```
smokerResid = smokersFev - smokerMean
nonSmokerResid = nonSmokersFev - nonSmokerMean
result = append(smokerResid,nonSmokerResid)
hist(result)
qqnorm(result)
qqline(result)
```




Compared to our first analyses looking at all children, we are seeing similar distribution shape issues, but can see a smaller variance for the non smoker data set, and see that many of the children who were under 10 never smoke, in fact only one did in this study. Thus their sample sizes are much more similar. We can attribute the decrease in variance of non smokers likely to this fact. Further, looking again at the residuals, we see another normal looking distribution and am satisfied as such.

```
smokeVnonSmokeTT =t.test(smokersFev,nonSmokersFev,var.equal=F)

print((smokeVnonSmokeTT))
```

```
##
##  Welch Two Sample t-test
##
## data:  smokersFev and nonSmokersFev
## t = 1.4333, df = 95.857, p-value = 0.155
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.05663529  0.35088918
## sample estimates:
## mean of x mean of y
##  3.297547  3.150420
```

Unlike the previous instance, removing the children below the age of ten made a large difference. We see that it has been reflected in both our p value and our confidence interval. Thus, we will not reject the null hypothesis in this scenario. This large difference I think is indicative that we may want to investigate age as a more potent indicator of FEV in children as it appears removing younger children from the study has quite the impact.

**Problem 3**

I agree with the conclusions from John Ioannidis. Intuitively, it makes sense that bias and ulterior motives, especially nowadays, can alter results in favor of gaining prestige, funding, tenure, or other objectives rather than aiming to learn something new. I have been fortunate to have had previous professors and mentors who recognize this issue and take pride in careful articulation of results. In particular, fields with little to no ground truth are more likely than not to have results that seem useful, but in reality have no backing. I further appreciated the holistic view of the number of independent studies done on the same question and studying the probability a particular study has found something significant. I do not believe it was mentioned explicitly, but I would state that I think some of these studies can still be quite useful despite the lack of conclusive results. Data collection may still be useful, as well as methodology for future studies.