Assignment 2

Alexander Williams

2024-10-15

1.

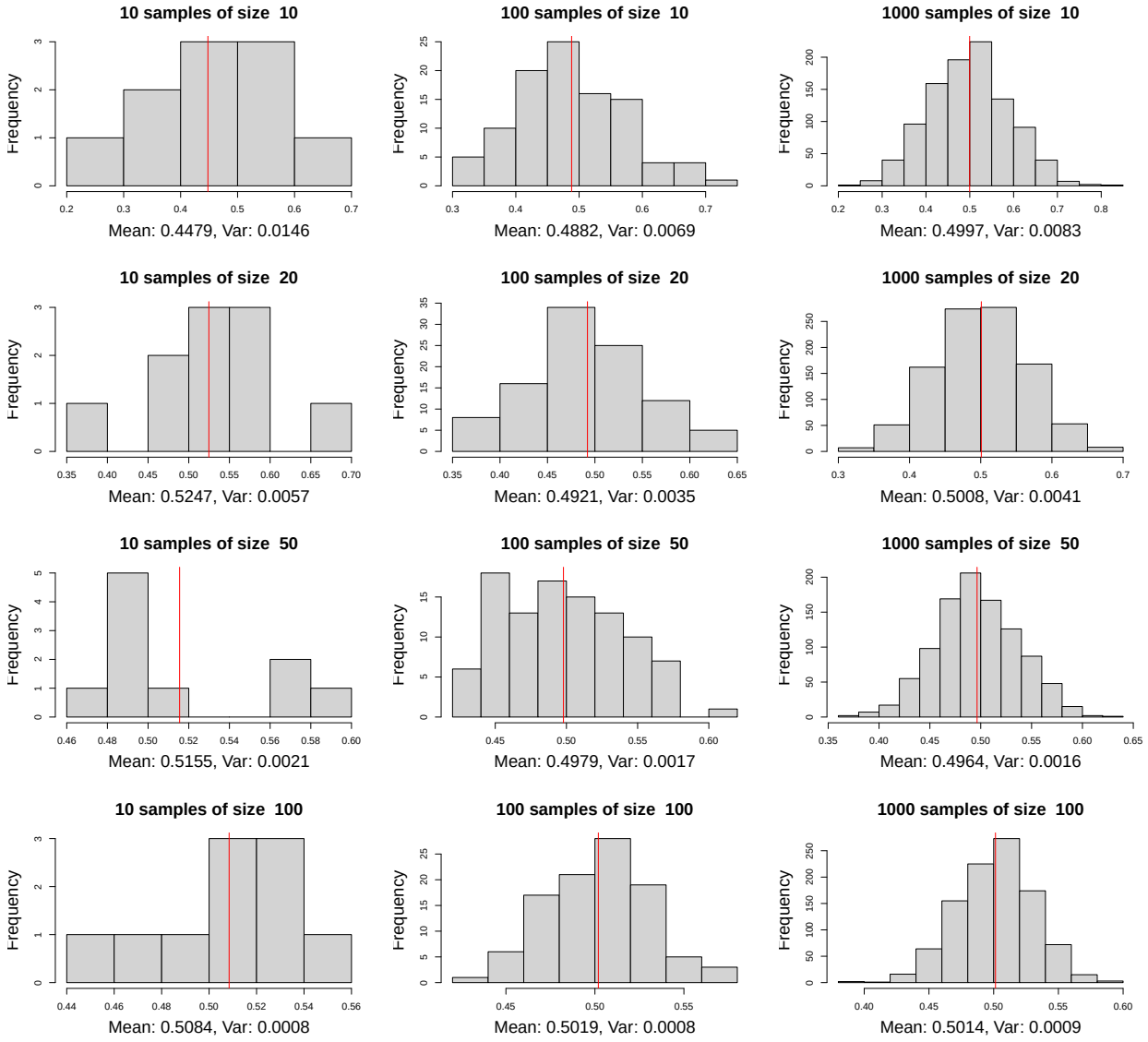
To make this process of drawing so many histograms easier, I'm using a function to automate most of the process, given a function to sample from the distribution and the desired sample size

```
triple.hist <- function(sample.size, FUN, arg1, arg2)
{
x10 <- matrix(data=FUN(10*sample.size,arg1,arg2), nrow=sample.size, ncol=10)
x10.mean <- apply(x10, 2, mean)
hist(x10.mean, main=paste('10 samples of size ', sample.size),
     cex.lab=1.75, cex.main=1.75,
     xlab=sprintf("Mean: %.4f, Var: %.4f", mean(x10.mean), var(x10.mean)))
abline(v=mean(x10.mean), col='red')

x100 <- matrix(data=FUN(100*sample.size,arg1,arg2), nrow=sample.size,ncol=100)
x100.mean <- apply(x100, 2, mean)
hist(x100.mean, main=paste('100 samples of size ', sample.size),
     cex.lab=1.75, cex.main=1.75,
     xlab=sprintf("Mean: %.4f, Var: %.4f", mean(x100.mean), var(x100.mean)))
abline(v=mean(x100.mean), col='red')

x1000 <- matrix(data=FUN(1000*sample.size,arg1,arg2), nrow=sample.size,ncol=1000)
x1000.mean <- apply(x1000, 2, mean)
hist(x1000.mean, main=paste('1000 samples of size ', sample.size),
     cex.lab=1.75, cex.main=1.75,
     xlab=sprintf("Mean: %.4f, Var: %.4f", mean(x1000.mean), var(x1000.mean)))
abline(v=mean(x1000.mean), col='red')
}
```

i) Uniform Distribution

{r, figures-side, fig.show="hold", out.width="33%"}

```
triple.hist(sample.size=10, runif, 0, 1)
triple.hist(sample.size=20, runif, 0, 1)
triple.hist(sample.size=50, runif, 0, 1)
triple.hist(sample.size=100, runif, 0, 1)
```



As the sample size increases, the mean values do tend to approach the mean of the distribution, 0.5, and the variance decreases, meaning that on average the mean of each sample is closer to 0.5. This shows that in accordance with the central limit theorem, as the sample size increases, the sample mean does approach the distribution mean.

as the number of samples increases, the sample means do approach the distribution mean, and the overall distribution appears more normal, but it does not appear that variance has a significant increase. This could be tested further to confirm

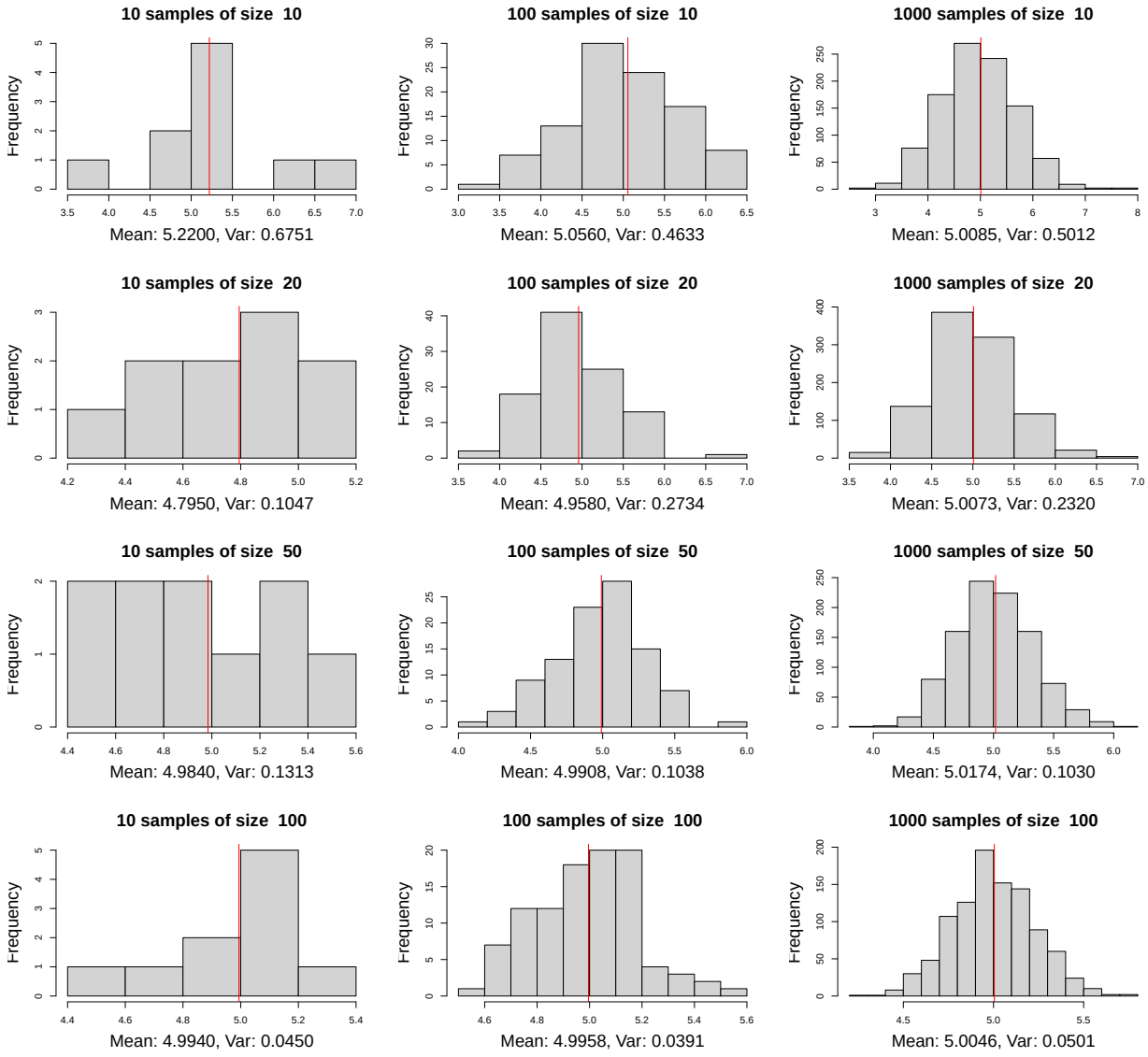
ii) Poisson Distribution

The function I wrote to plot everything for me requires 2 arguments for the distribution, while the Poisson distribution only requires one. To get around this, I'll make a small function that takes 2 arguments but only uses one, and pass that to the histogram function.

```
mpois <- function(n, real, fake)
{
  rpois(n, real)
}
```

The histograms will be on the next page

```
triple.hist(sample.size=10, mpois, 5, 999)
triple.hist(sample.size=20, mpois, 5, 999)
triple.hist(sample.size=50, mpois, 5, 999)
triple.hist(sample.size=100, mpois, 5, 999)
```

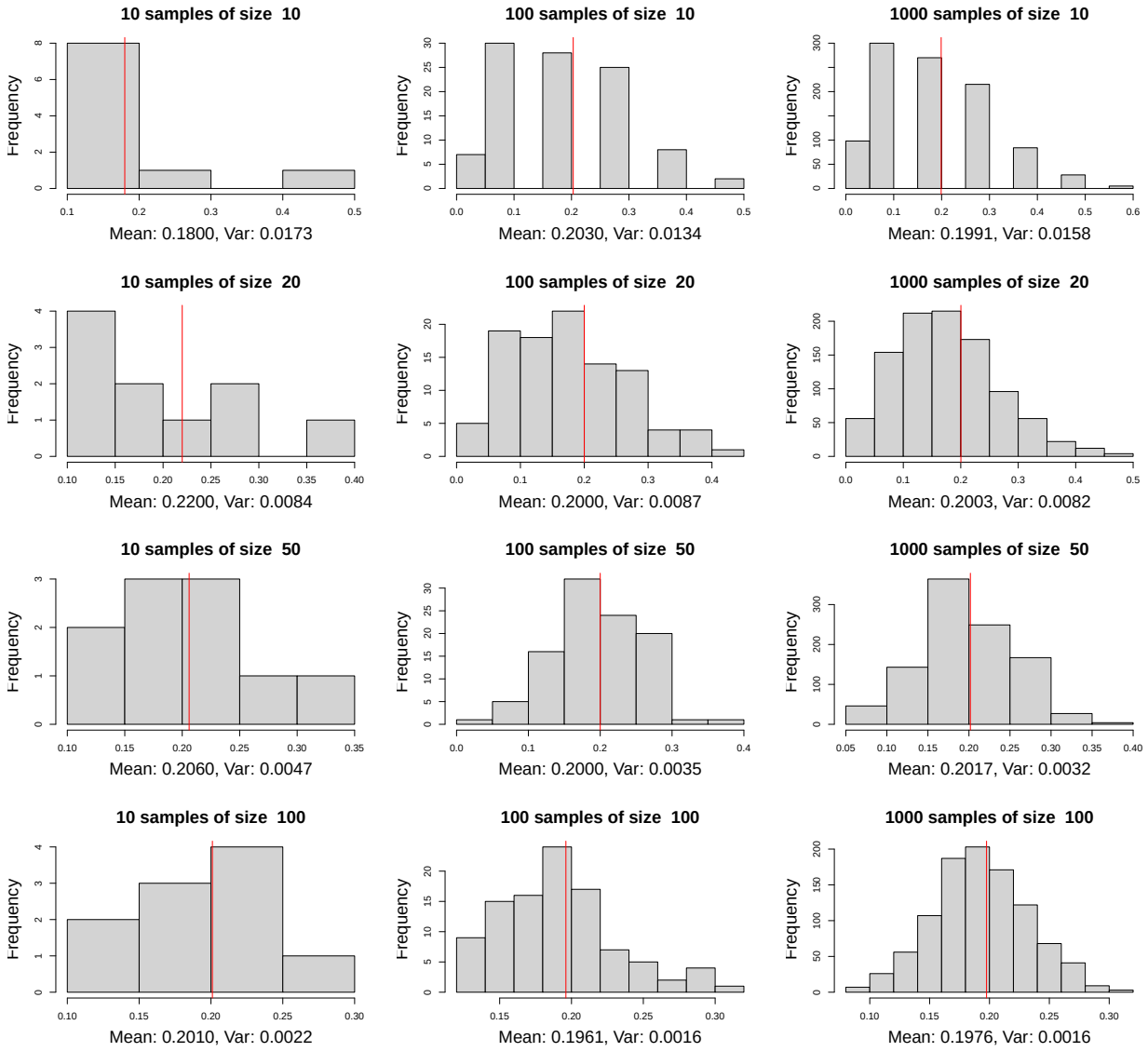


Just like the normal distribution, as the sample size increases, the sample mean approaches the distribution mean, and the sample variance decreases. As the number of samples increases, The sample means appear appear to approach the distribution mean, but there's no significant decrease in variance.

Compared to the uniform distribution, the variance for this t-distribution is significantly larger.

iii) Bernoulli Distribution

```
triple.hist(sample.size=10, rbinom, 1, 0.20)
triple.hist(sample.size=20, rbinom, 1, 0.20)
triple.hist(sample.size=50, rbinom, 1, 0.20)
triple.hist(sample.size=100, rbinom, 1, 0.20)
```



Once again, as sample size increases, the sample mean approaches the distribution mean, and variance decreases. As the number of samples increase, the sample means approach the distribution mean, and variance does not appear to significantly change.

Compared to the previous two distributions, the Bernoulli distribution visually appears the least normal, even as sample size and sample count increases. While it isn't apparent in all histograms, some appear to be skewed to the right. This does make intuitive sense, since the mean is 0.2 and the distribution can only return a value from 0-1. This means that, even if the values are unlikely, there's more possible values above the mean compared to below. Further testing would be required to accurately determine if this distribution is skewed.

In all examples, as the number of samples increases, the sample mean does approach the distribution mean. This is evidence that supports the central limit theorem.

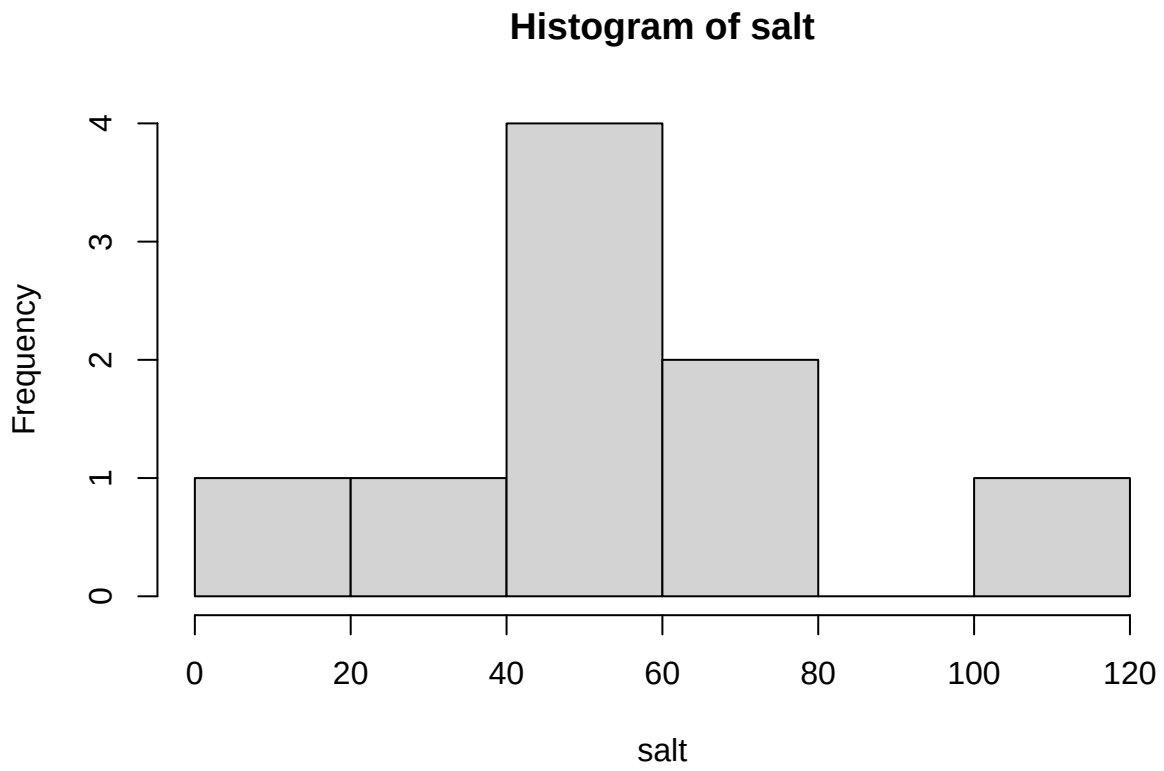
Something I find interesting is that when you compare `sample.size=10,sample.count=100` to `sample.size=100,sample.count=10` in any of these distributions, you can see that the mean is similar, but variance is much lower in the histograms with the larger sample size.

2.

```
Salt <- read.table(file='~/Documents/uni/STAT359/data/salt.txt',
                   header=TRUE,
                   sep="")
attach(Salt)
```

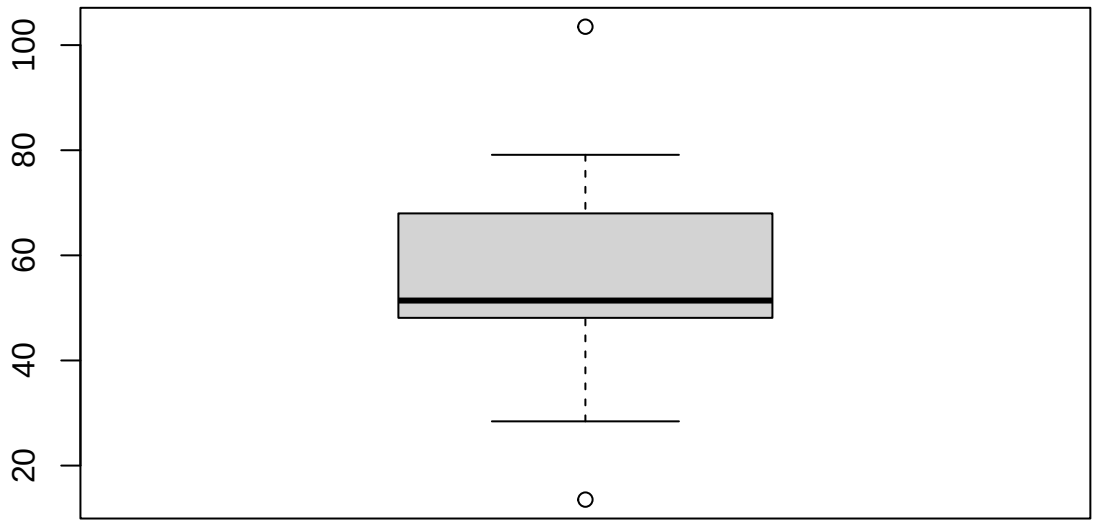
a)

```
hist(salt)
```



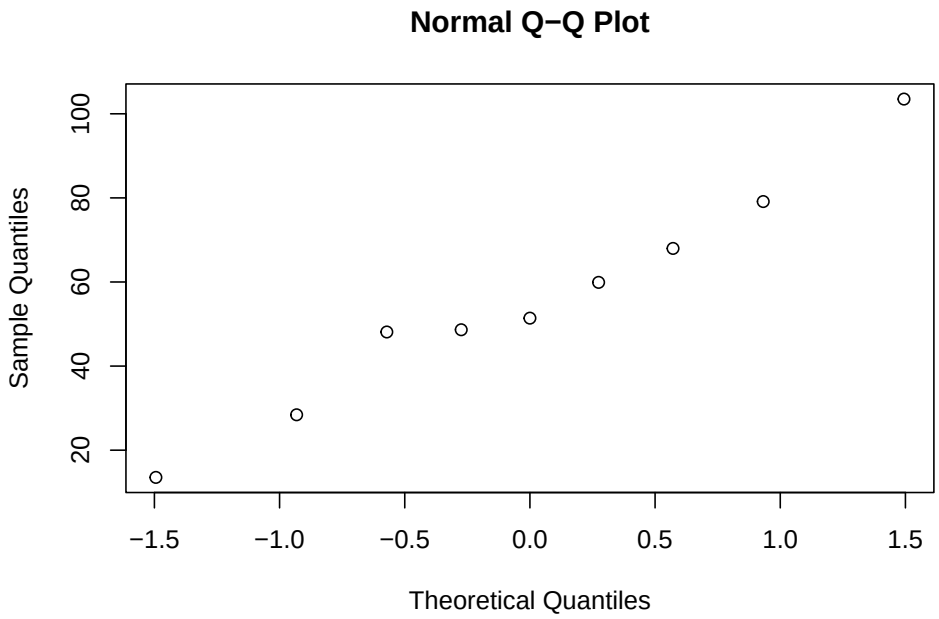
Hard to say if data is symmetric from the histogram, maybe a boxplot will be clearer

```
boxplot(salt)
```



While the mean is low, the distribution does appear somewhat symmetrical

```
qqnorm(salt)
```



The left of the plot might be curved down while the right is curved up, implying there could be more weight than a normal distribution in both tails, but it does appear to be symmetric.

b)

```
#calculate skew
skew <- function(x)
  {
  m3 <- sum((x-mean(x))^3) / length(x)
  s3 <- sqrt(var(x))^3
  skew<- m3/s3
  skew
}
skew(salt)
```

```
## [1] 0.1723753
```

```
# Calculate kurtosis
kurtosis <- function(x)
  {
  m4 <- sum((x-mean(salt))^4) / length(x)
  s4 <- var(x)^2
  kurtosis <- m4/s4 - 3
  kurtosis
  }
kurtosis(salt)
```

```
## [1] -0.9342198
```

c)

```r
B <- 1000

bootstrap <- matrix(data=sample(salt, length(salt)*B,replace=TRUE),
                    nrow=length(salt),
                    ncol=B)
boot.skew <- apply(bootstrap,2,skew)
skew.interval<-quantile(boot.skew,
                        probs=c(0.025,0.975))
skew.interval
```

```
##       2.5%      97.5%
## -0.9314516  1.0927449
```

Since the confidence interval contains 0, there's no evidence that the data is skewed.

d)

```r
B <- 1000

bootstrap <- matrix(data=sample(salt, length(salt)*B,replace=TRUE),
                    nrow=length(salt),
                    ncol=B)
boot.kurt <- apply(bootstrap,2,kurtosis)
kurt.interval<-quantile(boot.kurt,
                        probs=c(0.025,0.975))
kurt.interval
```

```
##      2.5%     97.5%
## -1.672355  4.215875
```

Again, the confidence interval does contain 0, so there is no evidence against the null hypothesis that this data has no kurtosis.

e)

Based on the very small values for both skew and kurtosis in the sample, and how the confidence values for both skew and kurtosis contained zero, it's fairly likely that the data is normal. However, the lack of datapoints makes it difficult to say for sure.

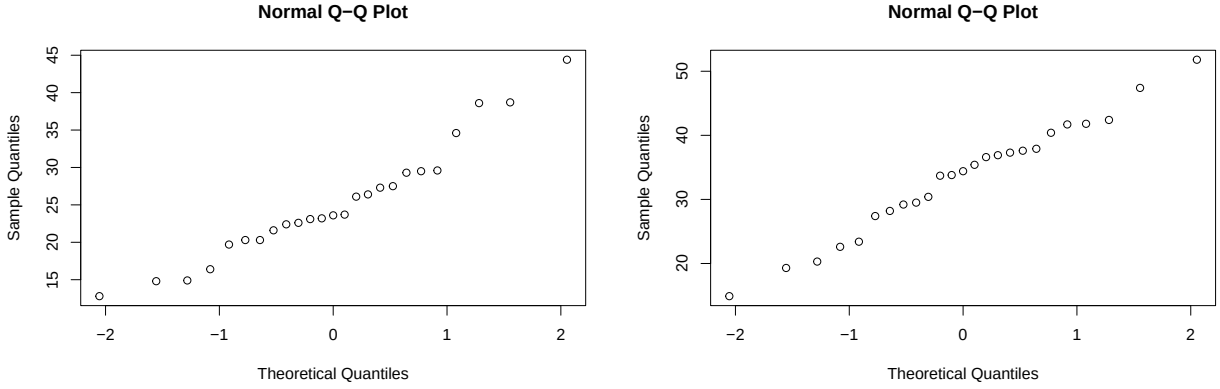
3.

```
fecund <- read.table(file='~/Documents/uni/STAT359/data/fecundity.txt',
                     sep="", header=TRUE)
sprintf("%f, %f", var(fecund$RS), var(fecund$NS))
```

```
## [1] "60.410067, 79.959600"
```

```
qqnorm(fecund$RS)
qqnorm(fecund$NS)
var.test(fecund$RS, fecund$NS)
```

```
##
##  F test to compare two variances
##
## data:  fecund$RS and fecund$NS
## F = 0.75551, num df = 24, denom df = 24, p-value = 0.4974
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.3329286 1.7144557
## sample estimates:
## ratio of variances
##          0.7555074
```



Given the high p-value of 0.50, there's no evidence that the two samples don't share the same variance. We can move forward under the assumption that they share the same variance. The data also appears normal in the qqplots, and we have no reason to believe the data is paired, so the data can be compared using a pooled t-test.

```
t.test(fecund$RS, fecund$NS, alternative="two.sided", var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  fecund$RS and fecund$NS
## t = -3.4251, df = 48, p-value = 0.001268
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -12.880308  -3.351692
## sample estimates:
## mean of x mean of y
##    25.256    33.372
```

The p-value for this test is quite low at 0.001, there's very strong evidence that the two samples don't share the same mean. the resistant bred fruit flies on average have a lower fecundity than the non-selectively bred fruit flies.

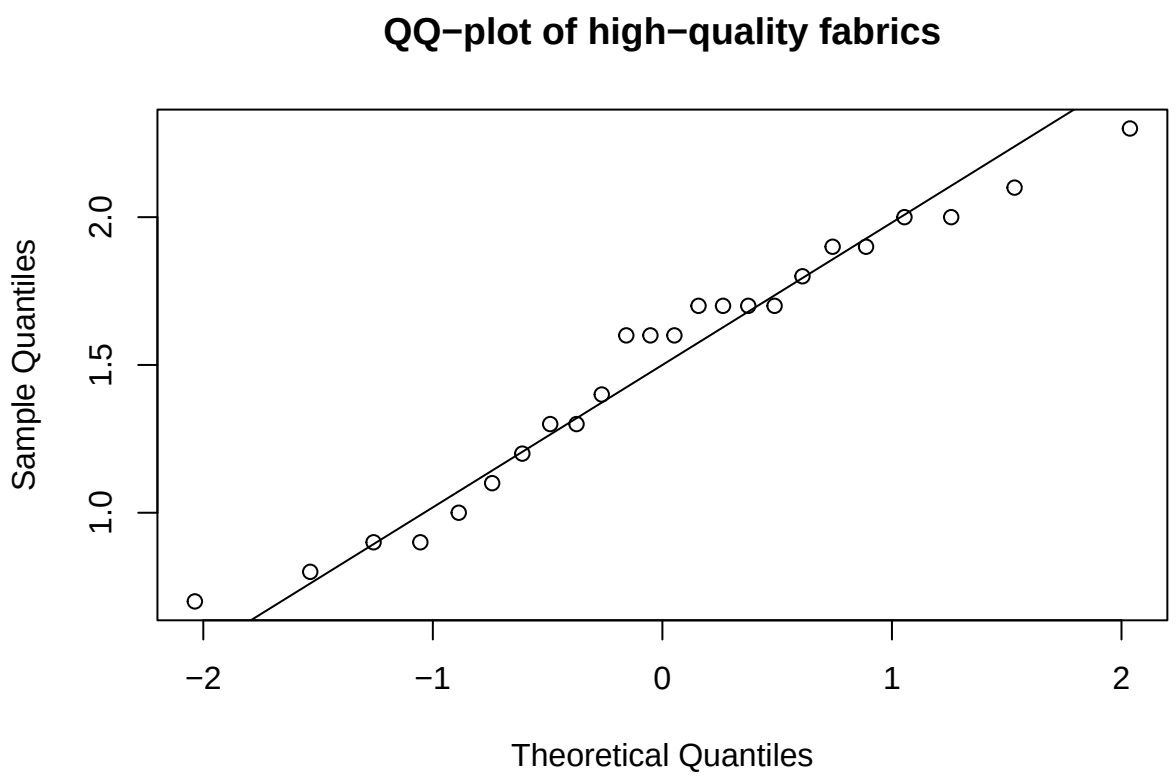
In summary, the two samples share the same variance, but the selectively bred fruit flies on average lay less eggs than the non-selectively bred fruit flies.

4.

```
fabric.flip <- read.delim(file='~/Documents/uni/STAT359/data/fabric.txt',
                      sep="",
                      header=FALSE)
fabric <- setNames(as.data.frame(t(fabric.flip[-1])), fabric.flip[,1])
```

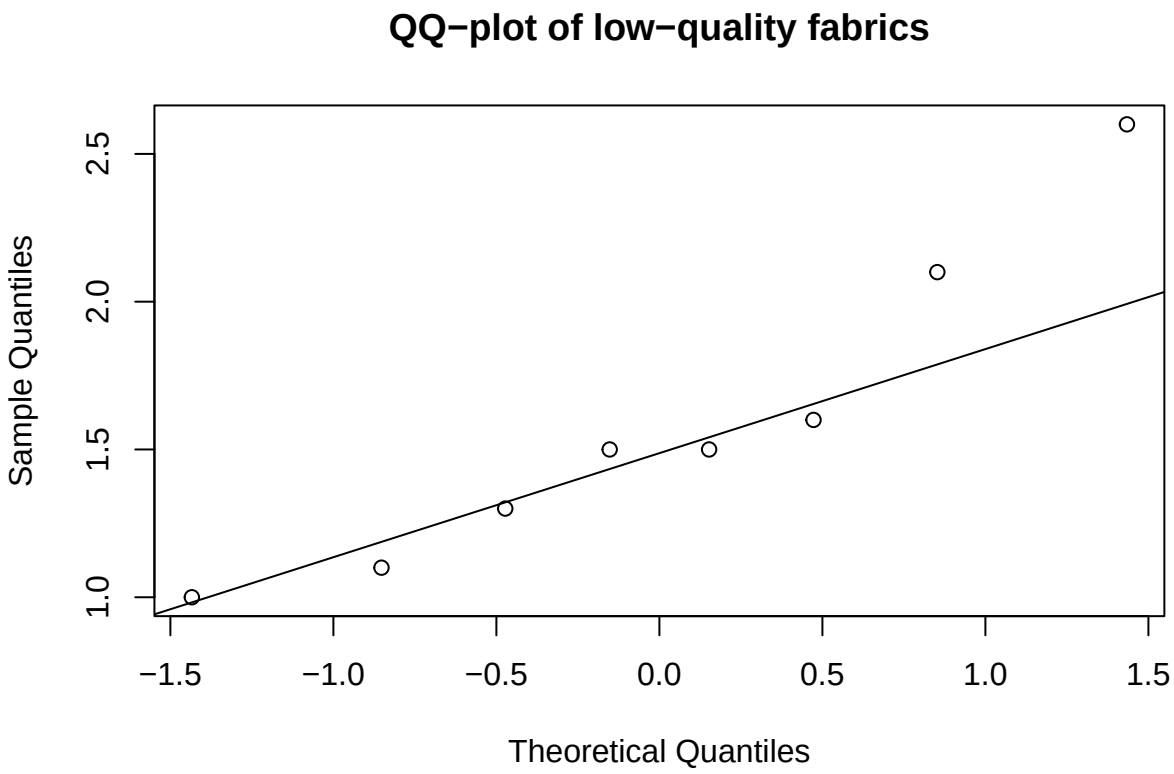
a)

```
qqnorm(fabric$H, main="QQ-plot of high-quality fabrics")
qqline(fabric$H)
```



The high-quality fabric's extensibility appears normal

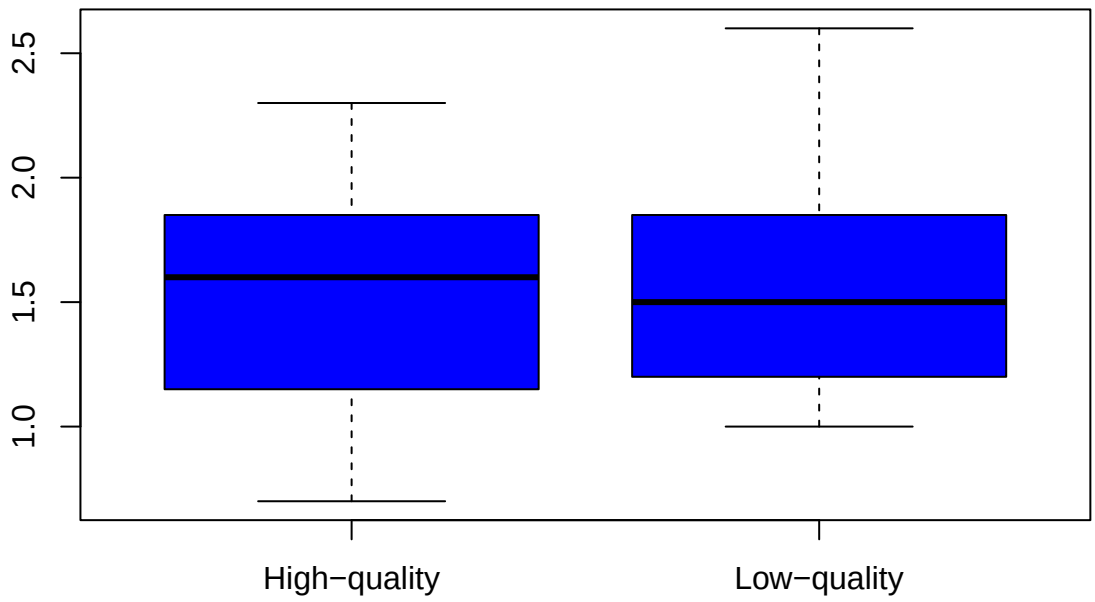
```
qqnorm(fabric$P, main="QQ-plot of low-quality fabrics")
qqline(fabric$P)
```



It's much harder to say if the low quality fabric's extensibility is normal due to the small number of data points, but the qqplot may imply a positive skew.

b)

```
boxplot(fabric$H, fabric$P, col='blue', names = c("High-quality", "Low-quality"))
```



The means of both populations are very close, though the extensibility of the low-quality fabric does appear to have some positive skew. It does not suggest that the true mean extensibility is different for the two populations.

c) Due to the small number of samples, a t-test should be used. The samples are not paired, and we need to check if the variance is close enough for a pooled t-test

```
var.test(fabric$H, fabric$P)
```

```
##
##  F test to compare two variances
##
## data:  fabric$H and fabric$P
## F = 0.70158, num df = 23, denom df = 7, p-value = 0.4862
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.1585015 2.0362234
## sample estimates:
## ratio of variances
##          0.7015781
```

With such a high p-value, an unpooled t-test is ideal

```
t.test(fabric$H, fabric$P, alternative = "two.sided", var.equal =FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  fabric$H and fabric$P
## t = -0.38011, df = 10.482, p-value = 0.7115
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.5403506  0.3820172
## sample estimates:
## mean of x mean of y
##  1.508333  1.587500
```

Since the p-value is greater than 0.05, there is no evidence that extensibility differs between low and high quality fabric.