DATA557 Homework 2

Will Wright

January 17, 2019

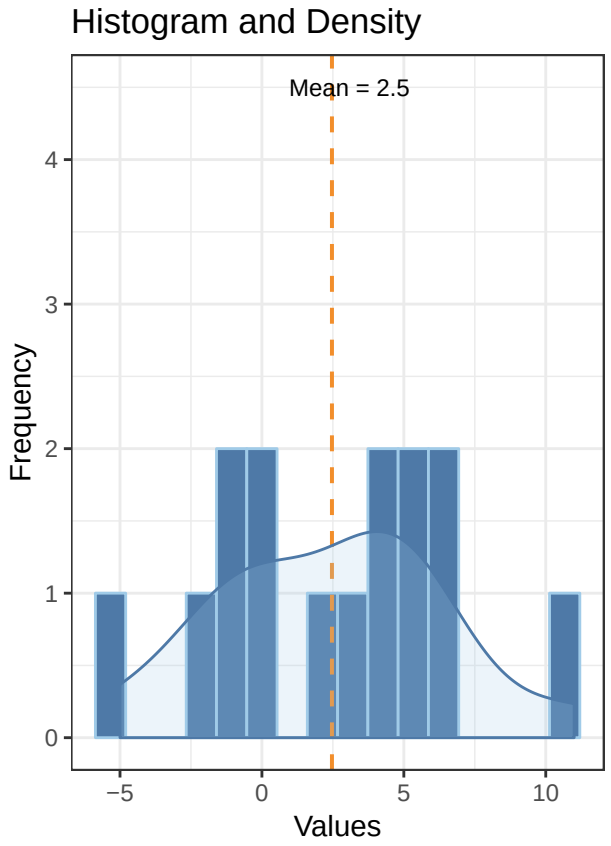
Problems

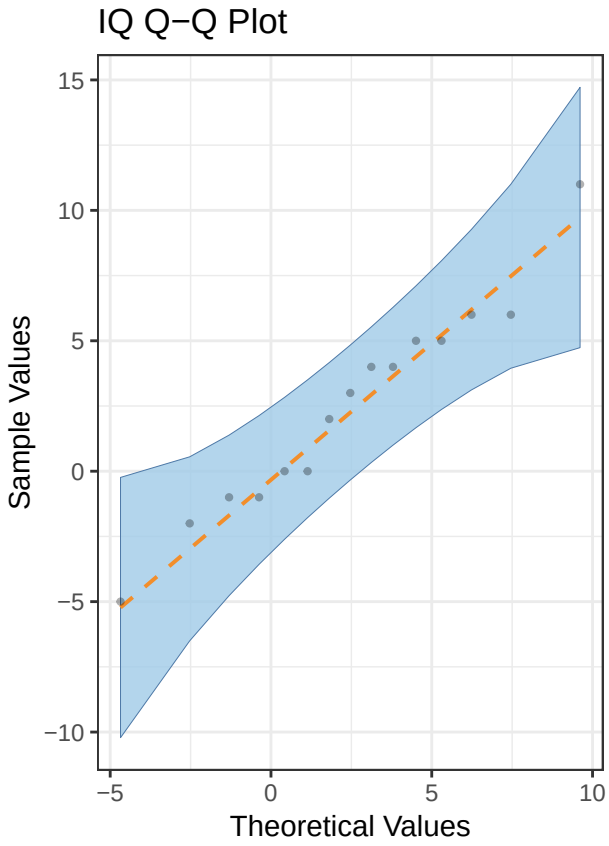
Question 1. (This is taken from Q3 of Exercise 2.)

A researcher is interested in measurements of a pollutant in water samples. In particular, there is a question about whether the value changes if the sample is tested when it is older compared with being tested right after it is collected. The researcher does not know whether aging could increase or decrease the pollutant concentration. To test the hypothesis 15 samples of water were taken from a lake. Each sample was divided into 2 aliquots, one to be analysed right away and the other to be analysed 1 month later. The difference between pollutant concentrations was recorded for each of the samples. The values obtained for the differences (fresh sample - aged sample), arranged from smallest to largest, were as follows: -5, -2, -1, -1, 0, 0, 2, 3, 4, 4, 5, 5, 6, 6, 11.

1.1 Calculate a 95% confidence interval for the mean difference in concentration between aged and fresh samples. Also perform a test of the null hypothesis that the mean difference is equal to 0. Do not assume a known variance. What is the name of the test? Compare results with the results from Exercise 2, Question 2. Explain how and why they differ.

```
waterSamples <- c(-5,-2,-1,-1,0,0,2,3,4,4,5,5,6,6,11)
n <- length(waterSamples)
xbar <- mean(waterSamples)
s <- sd(waterSamples)
se <- s/sqrt(n)
t <- xbar/se
# ci <- c(xbar - qnorm(0.975)*se, xbar + qnorm(0.975)*se) #used with known population variance
critical_value <- round(qt(0.975, df = n-1),2)
ci_lower <- round(xbar-critical_value*se,2)
ci_upper <- round(xbar+critical_value*se,2)
# p_val <- 2*(1-pnorm(t)) # not asked in this question
distribution_visualizer(waterSamples)
```






```
t
```

```
## [1] 2.368682
```

```
critical_value
```

```
## [1] 2.14
```

The 95% confidence interval for the mean difference in concentration between aged and fresh samples is:

$0.24 \leq \mu \leq 4.7$.

The **null hypothesis** is that there is no difference in in pollution concentration between aged and fresh samples: $H_0 : \mu = 0$.

The **alternate hypothesis** is that there is a difference: $H_1 : \mu \neq 0$

Given that the sample size is low at n = 15, we'll use a T-test to determine C and whether or not to reject:

$$\text{Reject } H_0 : \mu = 0 \text{ if } T = \frac{|\bar{X}|}{s/\sqrt{n}} > C,$$

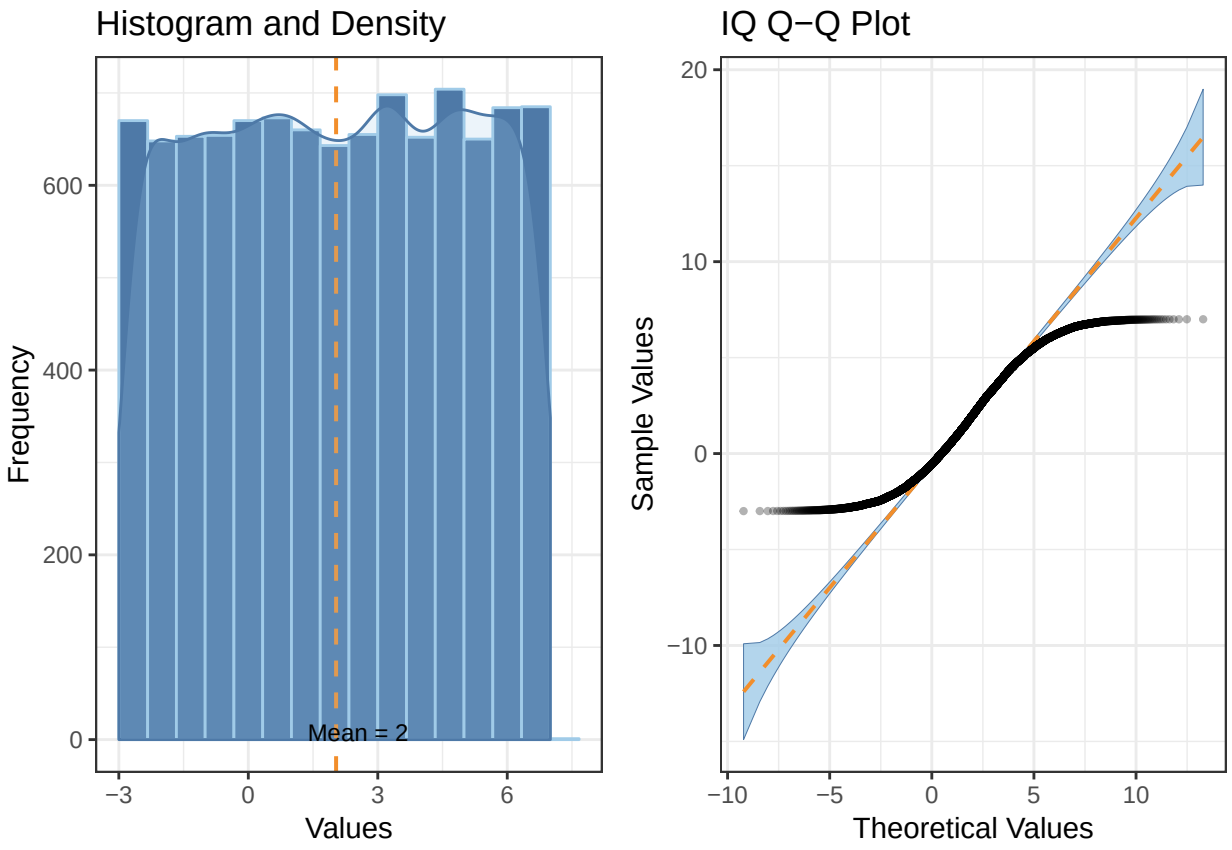
Because $C = 2.14 < T = 2.37$, we reject the null hypothesis. While we reject the null hypotheses in both this question and question 2, using the sample variance of 16.27 instead of the population variance of 15 and the t_{n-1} distribution instead of the normal distribution led to a higher critical value for this question.

1.2. Conduct a simulation study to assess the validity of the test used in 1.1 for this experiment. Choose a non-normal distribution for the population that is a plausible distribution for the differences in pollutant measurements (it is not necessary to require that the variance is equal to 15). Estimate the type I error of

the test. Is the test valid? Compare to the results for Exercise 2, Question 2.4. Explain the differences that you observe.

Given the shape of the sample data's distribution, we'll use a uniform from -3 to 7.

```
set.seed(4432)
# see what a high-n single sample looks like
uniformData <- runif(10000,-3,7)
distribution_visualizer(uniformData)
```




```
n=1000
reps=200
z=rep(NA,reps)
for(i in 1:reps){
  x=runif(n,-3,7)
  # shifting down 2 since we're assuming the mean is closer to 2, like the sample data
  z[i]=(mean(x)-2)/(sd(x)/sqrt(n))
}

summary(z)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -3.60421 -0.70517 -0.04280 -0.04654  0.64267  3.00363
```

```
unif_xbar <- mean(z)
critical_value = qnorm(0.975)
a <- mean(abs(z)>critical_value)
```

```
a
```

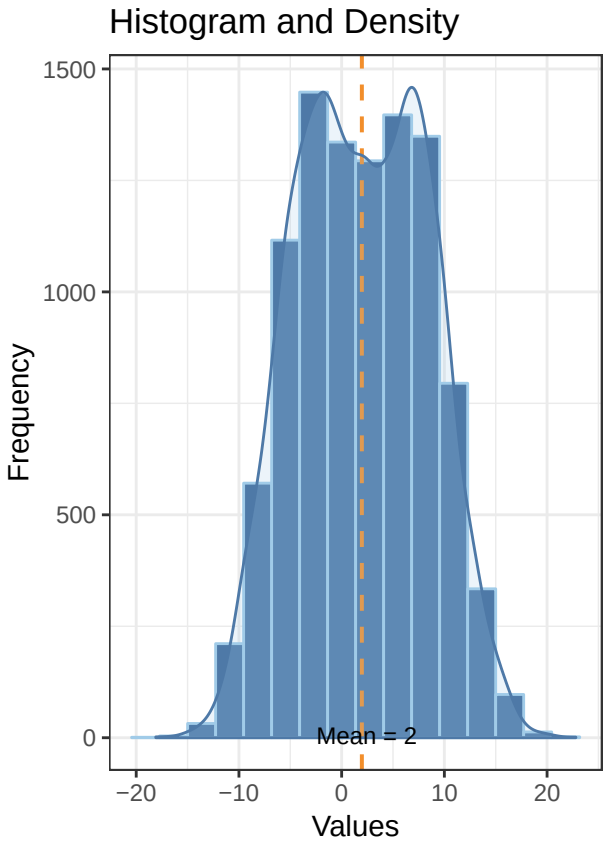
```
## [1] 0.05
```

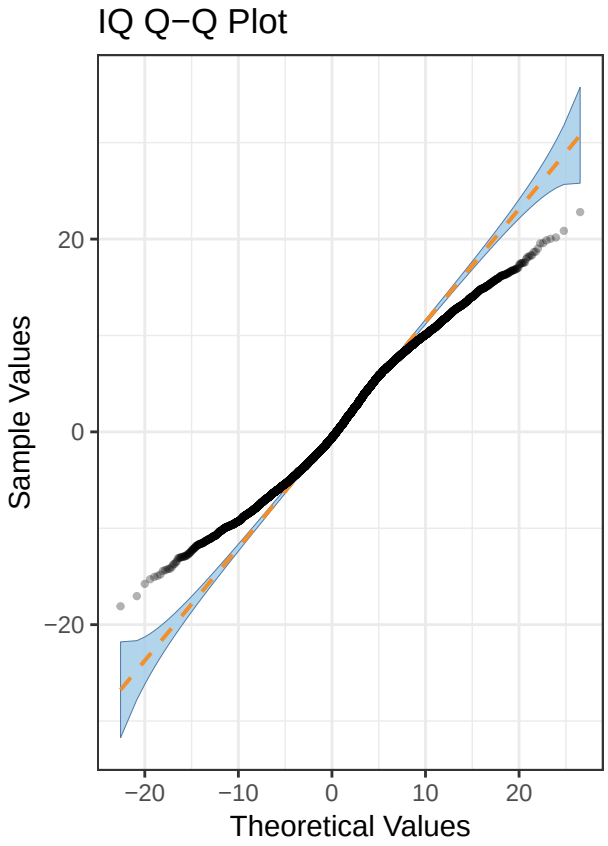
The type I error rate is $\alpha = 0.05$, which approximately meets the expectation of 0.05 and seems valid. Question 2.4 resulted in a mean of 0.49 and $\alpha =$ of 0.05 using a poisson distribution, which is very similar to the -0.05 we achieved. Any differences would be the result of the differences in the shape of the uniform to the poisson distribution, the sample variances, the sample sizes, number of repetitions, and the random nature of draws.

1.3. Propose a second plausible distribution for the population that you think might give different results for the type I error. Estimate the type I error probability of the test under this distribution and compare with results of the previous simulation. Explain any differences.

For this question, let's assume a bimodal bivariate normal distribution with 2 modes centered at -3 and 7:

```
set.seed(4432)
mu1 <- -3
mu2 <- 7
sig1 <- sqrt(15)
sig2 <- sqrt(15)
cpct <- 0.5

bimodalDistFunc <- function (n,cpct, mu1, mu2, sig1, sig2) {
  y0 <- rnorm(n,mean=mu1, sd = sig1)
  y1 <- rnorm(n,mean=mu2, sd = sig2)

  flag <- rbinom(n,size=1,prob=cpct)
  y <- y0*(1 - flag) + y1*flag
}

bimodalData <- bimodalDistFunc(n=10000,cpct,mu1,mu2, sig1,sig2)
# see what a high-n single sample looks like
distribution_visualizer(bimodalData)
```






```
# perform simulation
n=1000
reps=200
z=rep(NA,reps)
for(i in 1:reps){
  x <- bimodalDistFunc(n, cpct, mu1, mu2, sig1, sig2)
  # shifting down 2 since we're assuming the mean is closer to 2, like the sample data
  z[i]=(mean(x)-2)/(sd(x)/sqrt(n))
}

summary(z)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -3.26526 -0.53122  0.13310  0.08075  0.76357  3.21698
```

```
bimodal_xbar<-mean(z)
critical_value = qnorm(0.975)
a <- mean(abs(z)>critical_value)
a
```

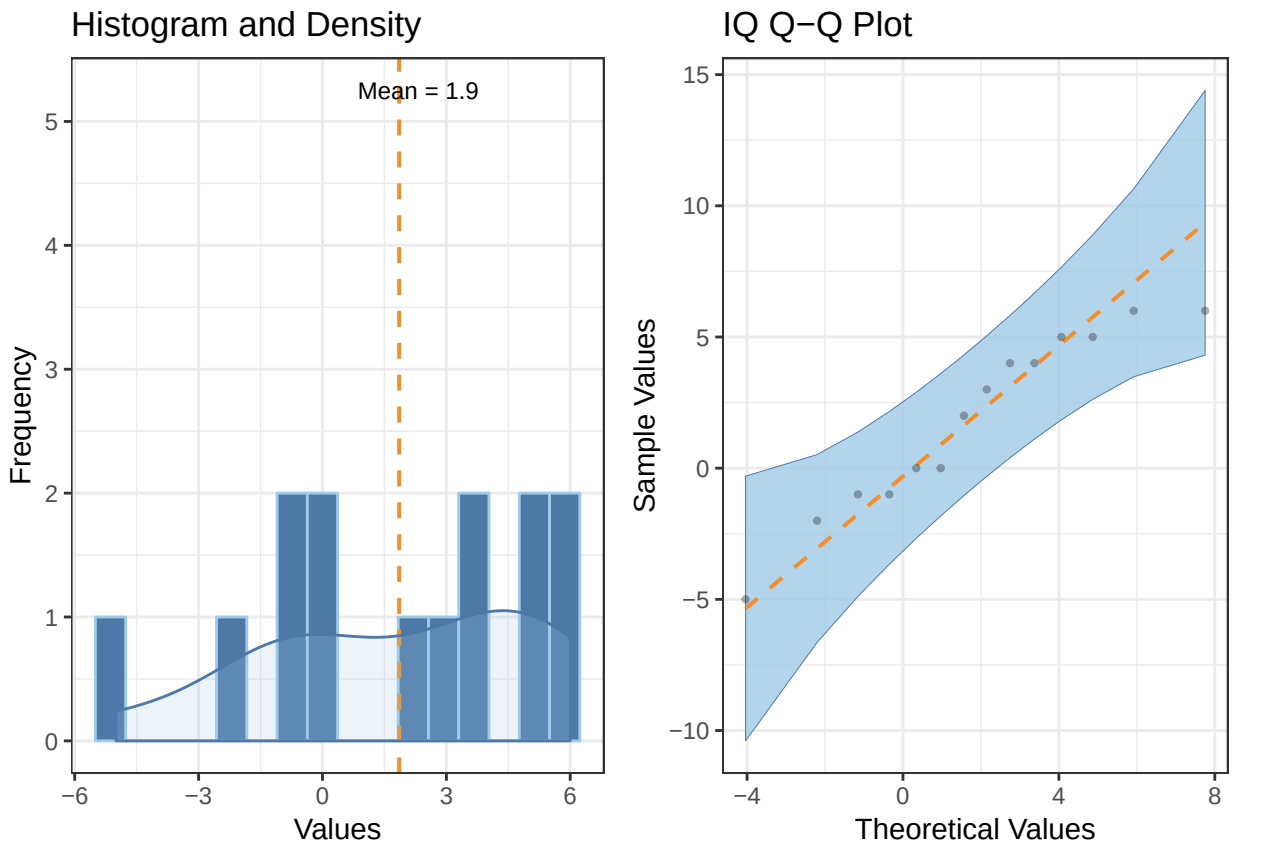
```
## [1] 0.055
```

In this case, we're seeing a mean of 0.08 instead of -0.05, but they are still very similar. The type I error rate here is $\alpha = 0.06$. Again, very similar. The shape of this distribution better matches that of the original sample, but both give similar results.

1.4. Suppose that it was determined that the last data value (11) was an error due to failure of the measuring equipment. Re-run the test and confidence interval for the mean with this value excluded. How did the

results change?

```
waterSamples <- c(-5,-2,-1,-1,0,0,2,3,4,4,5,5,6,6) # remove 11
n <- length(waterSamples)
xbar <- mean(waterSamples)
s <- sd(waterSamples)
se <- s/sqrt(n)
t <- xbar/se
critical_value <- round(qt(0.975, df = n-1),2)
ci_lower <- round(xbar-critical_value*se,2)
ci_upper <- round(xbar+critical_value*se,2)
distribution_visualizer(waterSamples)
```




```
t
```

```
## [1] 2.04762
```

```
critical_value
```

```
## [1] 2.16
```

The 95% confidence interval for the mean difference in concentration between aged and fresh samples is:

-0.1 $\leq \mu \leq$ 3.82.

The **null hypothesis** is that there is no difference in in pollution concentration between aged and fresh samples: $H_0 : \mu = 0$.

The **alternate hypothesis** is that there is a difference: $H_1 : \mu \neq 0$

Rejection rule:

$$\text{Reject } H_0 : \mu = 0 \text{ if } T = \frac{|\bar{X}|}{s/\sqrt{n}} > C,$$

Because $C = 2.16 > T = 2.05$, we fail to reject the null hypothesis. While the 11 value added higher variance to the data, it was a strong pulling force for a higher mean. Without it, the mean of the sample is closer to 0 and, in fact, the 95% confidence interval included it.

Question 2

A researcher wants to evaluate the sensitivity of their assay for measuring urinary mercury. The standard is to have a 99% probability of detecting a sample with mercury concentration 1 ppm. Therefore, they wish to test the null hypothesis H_0: p=0.99, where p is the probability of a positive test when the concentration is 1 ppm. The alternative hypothesis is that p is not equal to 0.99. The researcher created 500 samples with a 1ppm mercury concentration and tested them. The number of samples that tested positive was recorded.

2.1. Suppose that they decide to reject H_0 if the number of positive samples is 494 or less. What is the type I error probability of this rejection rule?

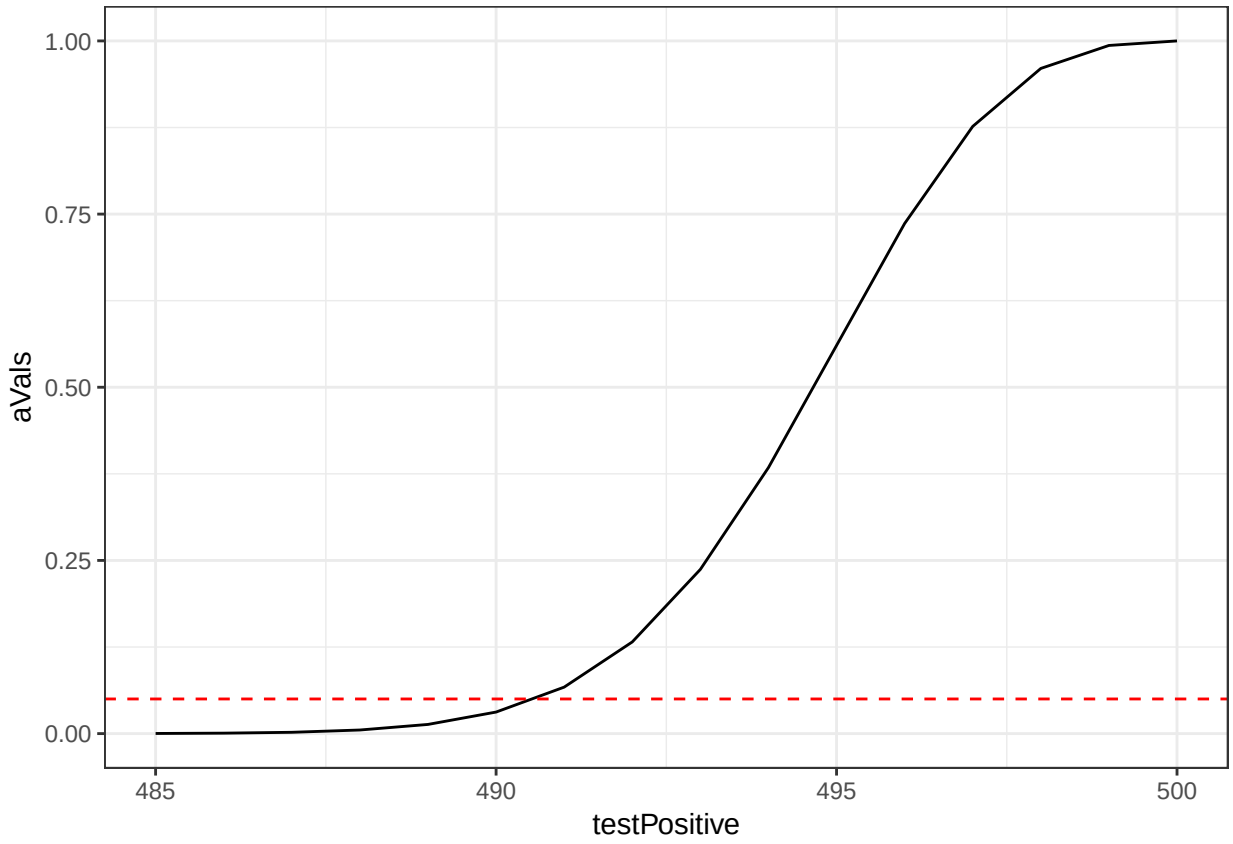
```
a <-sum(dbinom(c(0:494), size=500, prob=0.99))
```

The type I error rate is 0.38.

2.2. Find a different rejection rule for the test that gives a type I error probability that is less than or equal to 0.05 and as close to 0.05 as possible.

```
aVals <- rep(NA,500)
for(i in 1:length(aVals)){
  aVals[i] <- sum(dbinom(c(0:i), size=500, prob=0.99))
}
aValsDf <- data.frame(testPositive = 1:500, aVals = aVals)
g <- ggplot(aValsDf, aes(x = testPositive, y = aVals))
g + geom_line() +
  scale_x_continuous(limits = c(485,500)) +
  scale_y_continuous(limits = c(0,1)) +
  theme_bw() +
  geom_hline(yintercept = 0.05, col = "red", linetype = "dashed")
```

```
## Warning: Removed 484 rows containing missing values (geom_path).
```




```
max(aValsDf$testPositive[which(aValsDf$aVals<=0.05)])
```

```
## [1] 490
```

Because >490 testing positive results in a type I error rate ≤ 0.05 and positive tests ≤ 490 results in a type I error rate > 0.05, the new rejection rule is:

$$\text{Reject } H_0 : \text{Positive Tests} \leq 490$$

2.3. Suppose that the number of positive samples was 490, i.e., 98% of 500. Using your rejection rule from Q2, would you reject the null hypothesis?

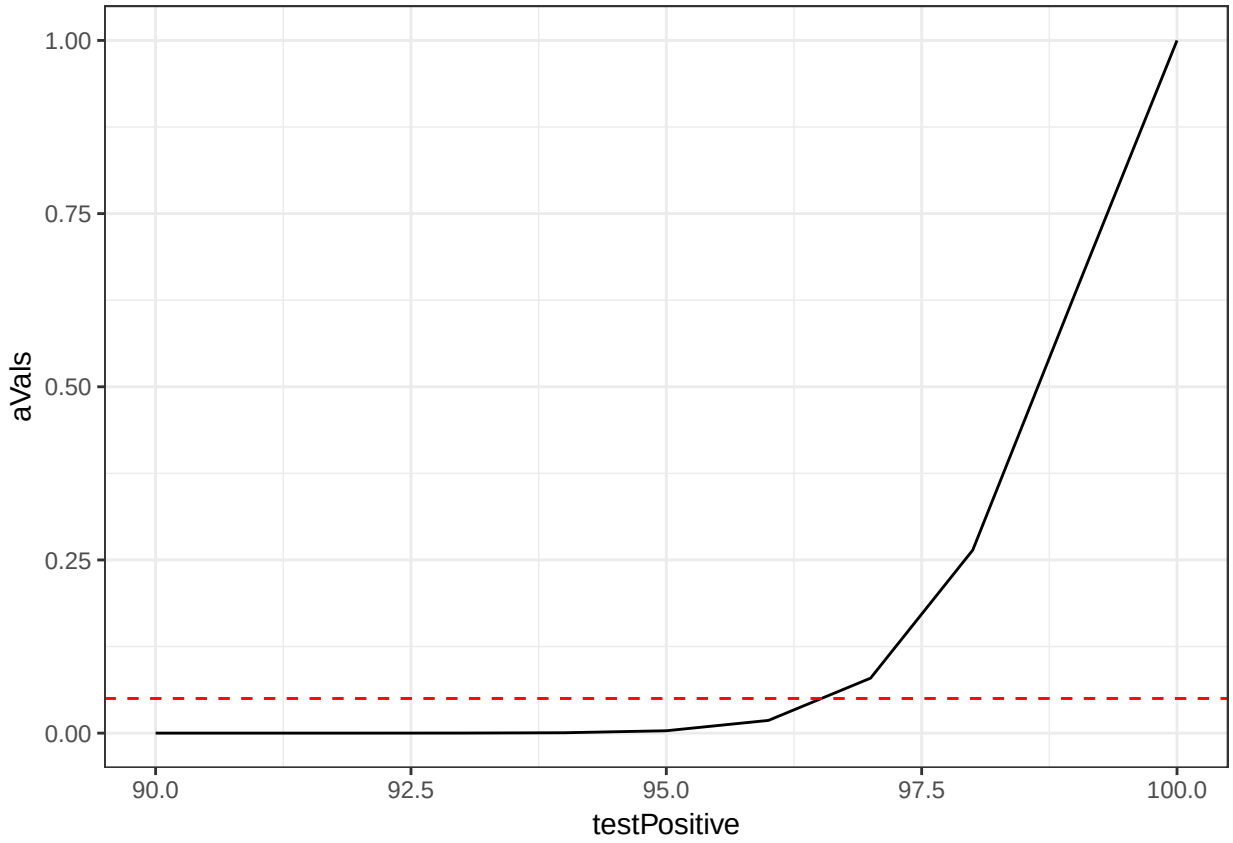
Yes.

2.4. Now suppose that the lab had performed only 100 samples and had found the same proportion of positive tests, i.e., 98 out of 100 positive tests? What is the rejection rule for the test in this situation and what would be the outcome of the test? Explain any differences in the outcomes between this test and the test based on 500 samples.

```
aVals <- rep(NA,100)
for(i in 1:length(aVals)){
  aVals[i] <- sum(dbinom(c(0:i), size=100, prob=0.99))
}
```

```
aValsDf <- data.frame(testPositive = 1:100, aVals = aVals)
g <- ggplot(aValsDf, aes(x = testPositive, y = aVals))
g + geom_line() +
  scale_x_continuous(limits = c(90,100)) +
  scale_y_continuous(limits = c(0,1)) +
  theme_bw() +
  geom_hline(yintercept = 0.05, col = "red", linetype = "dashed")
```

```
## Warning: Removed 89 rows containing missing values (geom_path).
```




```
max(aValsDf$testPositive[which(aValsDf$aVals<=0.05)])
```

```
## [1] 96
```

In this case, our new rejection rule is to reject if the number of positive tests is ≤ 96 so we fail to reject the null hypothesis with 98 positive tests. The reason for this difference is that with a smaller sample size (100 instead of 500) our SE is higher (since sqrt(n) is in the denominator). When SE is higher, our rejection region gets smaller and, hence, it becomes more difficult to reject the null hypothesis.

2.5. Calculate p-values for the tests based on 500 samples and 100 samples. Explain the difference between the p-values.

```
p_val_smallSample <-sum(dbinom(c(0:98), size=100, p=0.99))
p_val_largeSample <-sum(dbinom(c(0:494), size=500, p=0.99))
```

The 100-sample test with 98 positives has a p-value of 0.264 while the 500-sample test with 494 positives has

a p-value of 0.38.

Question 3 (Data set: 'iq.csv'. See HW1 for background).

3.1. Test the null hypothesis that the mean IQ score in the community is equal to 100 with a 2-sided significance level of 0.05. Use the 2-sided 1-sample t-test. State whether or not you reject the null hypothesis at significance level 0.05 and also give the p-value for your test.

```
# Read data
iqData <- read.csv("../WEEK01/iq.csv")

# Perform calculations
mu <- 100
s <- sd(iqData$IQ)
n <-nrow(iqData)
se <- s/sqrt(n)
xbar <- mean(iqData$IQ)
z <-abs(xbar-mu)/se
t_val <-qt(0.975, df=n-1)
p_val <- 2*(1-pt(z, df=(n-1)))
```

We reject the null hypothesis that the mean IQ score in the community is equal to 100 because the test's p-value of ≈ 0 is $\leq \alpha = 0.05$.

3.2. Compute a 95% confidence interval for the mean IQ. Do the confidence interval and hypothesis test give results that agree or conflict with each other? Explain. (2 points)

```
conf_upper <- xbar + t_val*se
conf_lower <- xbar - t_val*se
```

The 95% confidence interval for the mean IQ is 88.52 to 93.64. Because the range does not intersect 100, the results agree with the hypothesis test.

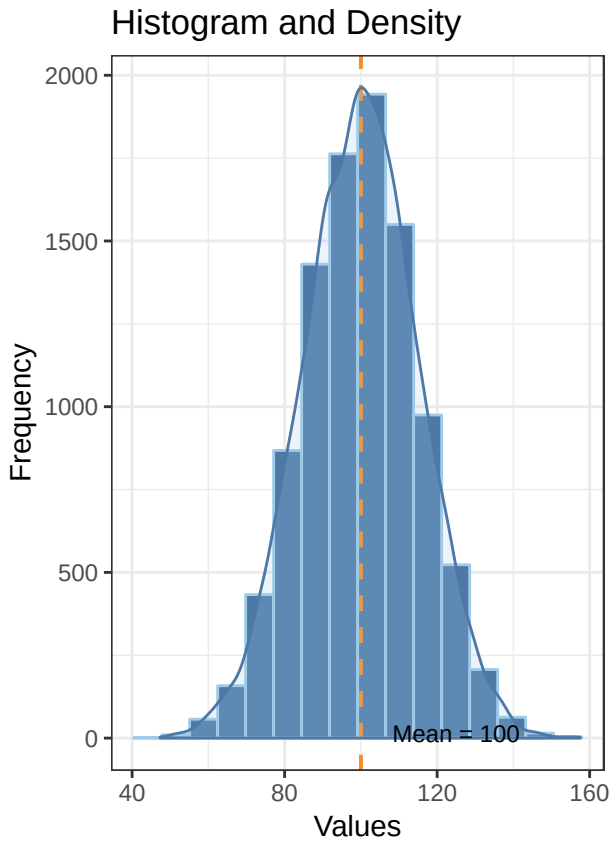
3.3. Repeat Q3.1 and Q3.2 using a significance level of 0.01 and a 99% confidence interval.

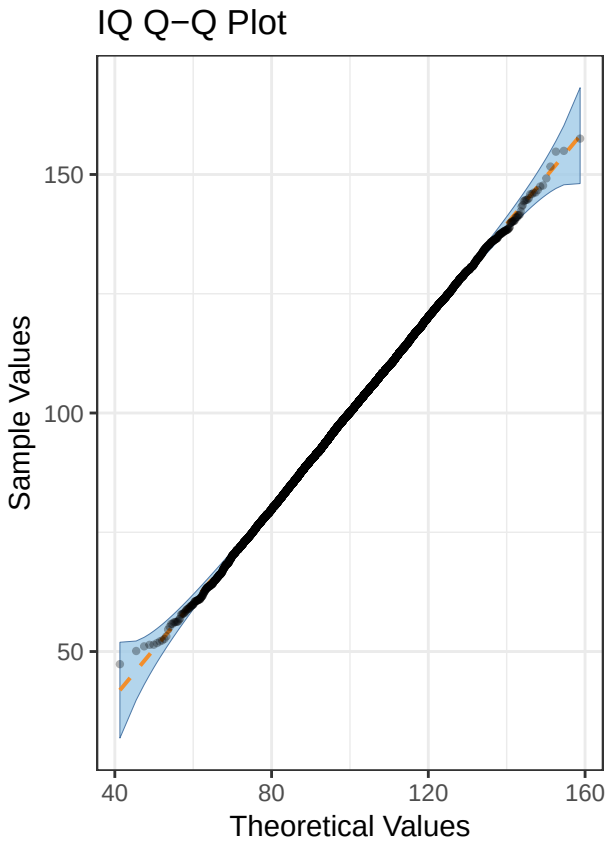
```
t_val <-qt(0.995, df=n-1)
p_val <- 2*(1-pt(z, df=(n-1)))

conf_upper <- xbar + t_val*se
conf_lower <- xbar - t_val*se
```

In this case, our p_val is still ≈ 0, which is $\leq \alpha = 0.01$ so we reject the null hypothesis. The 99% confidence interval is 87.7 to 94.46, which results in the same conclusion of rejection.

3.4. Perform a simulation study to assess the type I error probability of the test. For the simulation, generate samples of IQ scores using the normal distribution with mean 100 and SD 15. The sample size should be the same as for the data set. Report the observed type I error based on your simulation and comment on how well it agrees with theory.

```r
set.seed(121)
# see what a high-n single sample looks like
uniformData <- rnorm(10000,100,15)
distribution_visualizer(uniformData)
```






```r
n=nrow(iqData)
reps=200
z=rep(NA,reps)
for(i in 1:reps){
  x=rnorm(n,100,15)
  z[i]=(mean(x)-100)/(sd(x)/sqrt(n))
}

summary(z)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -3.69909 -0.77203 -0.06823 -0.02658  0.83123  2.30804
```

```r
critical_value = qnorm(0.975)
a <- mean(abs(z)>critical_value)
a
```

```
## [1] 0.05
```

The results agree with the theory since 5% of the results are outside the 95% confidence interval.

3.5. Perform a simulation study to estimate the power of the test to detect an alternative mean value for the mean IQ equal to 95. Generate samples (of same size as the data set) from a normal distribution with SD equal to 15.

```
set.seed(999)

mu <- 95
xbar <- mean(iqData$IQ)
n <- nrow(iqData)
n=nrow(iqData)

reps=200
z <- rep(NA,reps)
for(i in 1:reps){
  x <- rnorm(n,xbar,15)
  se <- sd(x)/sqrt(n)
  mean <- mean(x)
  z[i] <- (95-mean)/se
}

power <- mean(abs(z)>qnorm(0.975))
```

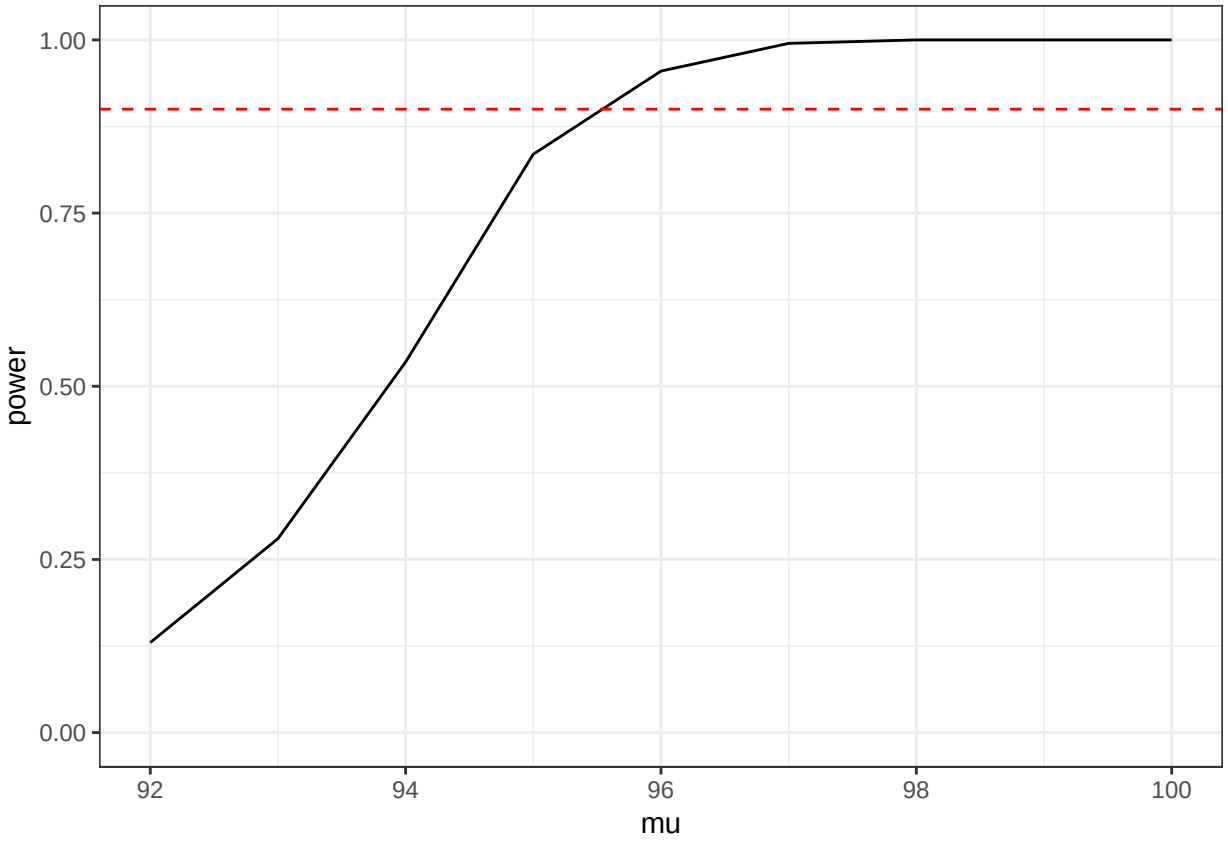
The estimated power from the simulation is 0.89.

3.6. Find the largest value of the alternative hypothesis mean that would be rejected with power of approximately 0.9. (Consider only values less than 100 for the alternative hypothesis mean.)

```
set.seed(999)

mu_vals <- 90:100
xbar <- mean(iqData$IQ)
n <- nrow(iqData)
n=nrow(iqData)
reps=200

powers <- rep(NA,length(mu_vals))
for(j in 1:length(mu_vals)){
  z <- rep(NA,reps)
  for(i in 1:reps){
    x <- rnorm(n,xbar,15)
    se <- sd(x)/sqrt(n)
    mean <- mean(x)
    z[i] <- (mu_vals[j]-mean)/se
  }
  powers[j] <- mean(abs(z)>qnorm(0.975))
}
results <- data.frame(mu = mu_vals, power = powers)

g <- ggplot(results, aes(x = mu, y = power))
g + geom_line() +
  scale_x_continuous(limits = c(92,100)) +
```

```
  scale_y_continuous(limits = c(0,1)) +
  theme_bw() +
  geom_hline(yintercept = 0.9, col = "red", linetype = "dashed")
```

```
## Warning: Removed 2 rows containing missing values (geom_path).
```




```
max(results$mu[which(results$power<=0.9)])
```

```
## [1] 95
```

The largest value of the alternative hypothesis mean that would be rejected with a power of 0.9 is 95.