Exercise 1 - Solution

```
set.seed(557)

qqn <- function (x, ...) {
  qqnorm(x, ...)
  qqline(x)
}
```

Populations and Samples

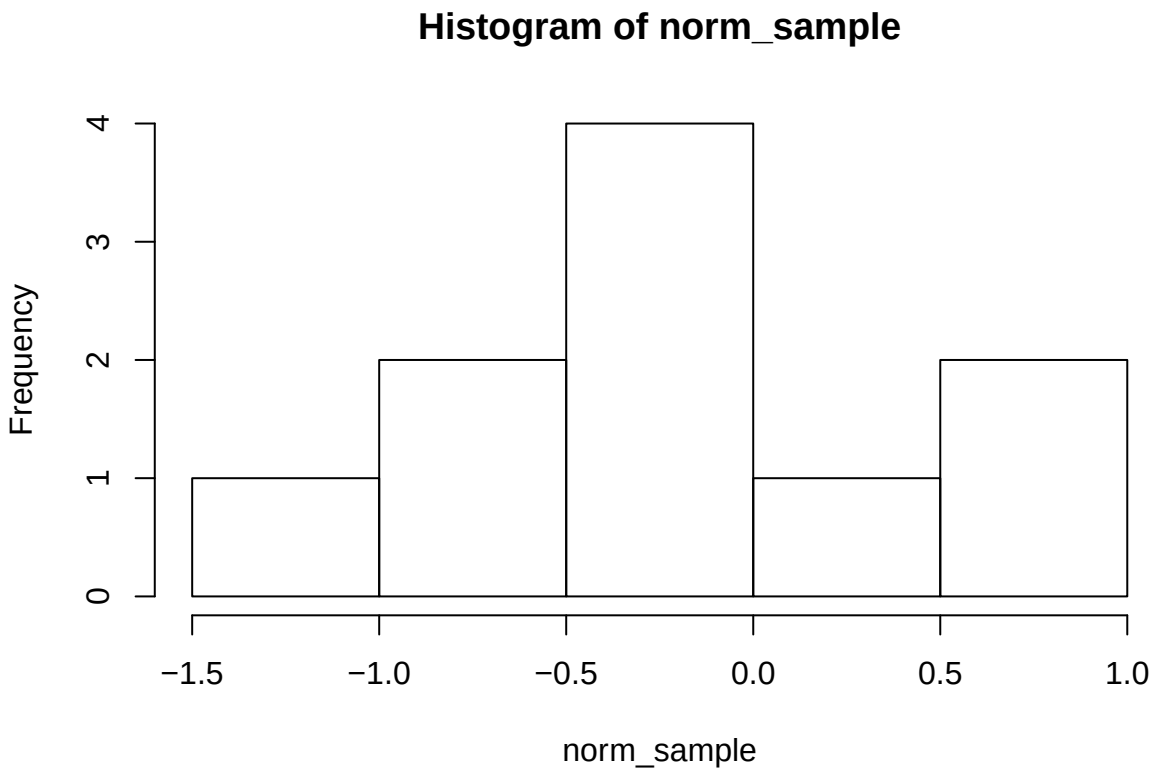
Problem 1

Take a random sample of size **10** from the standard normal distribution (mean 0, sd 1). Make a histogram and q-q plot of your sample. Does the distribution clearly look like a normal distribution? Does it clearly look non-normal? Repeat the sampling procedure at least 20 times. Do the samples consistently appear to be normal or non-normal?

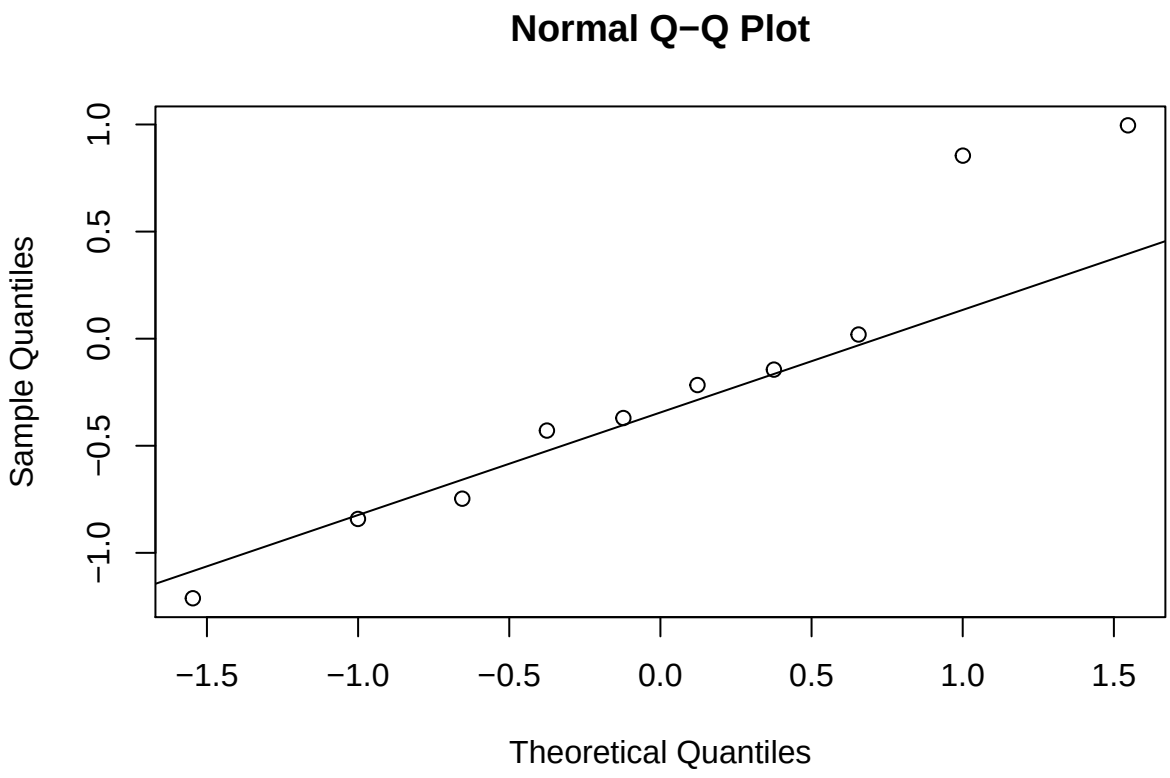
```
norm_sample <- rnorm(n = 10, mean = 0, sd = 1)
norm_sample
```

```
##  [1] -0.74716840  0.01930685 -1.21204429 -0.84174712 -0.21658881
##  [6]  0.85465529 -0.14454162  0.99601772 -0.42909576 -0.37066143
```

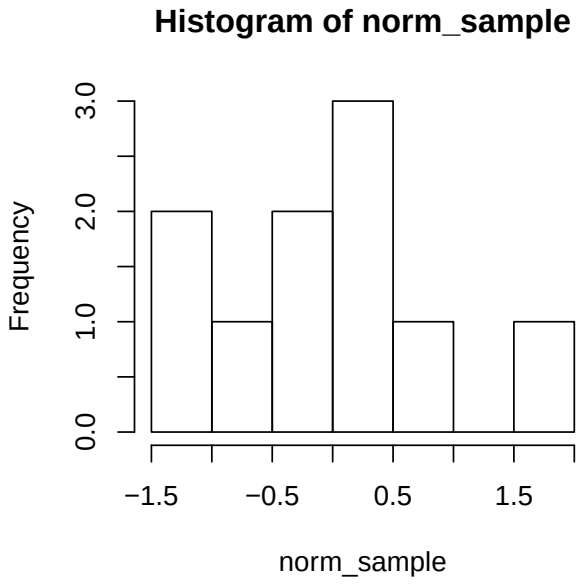
```
hist(norm_sample)
```

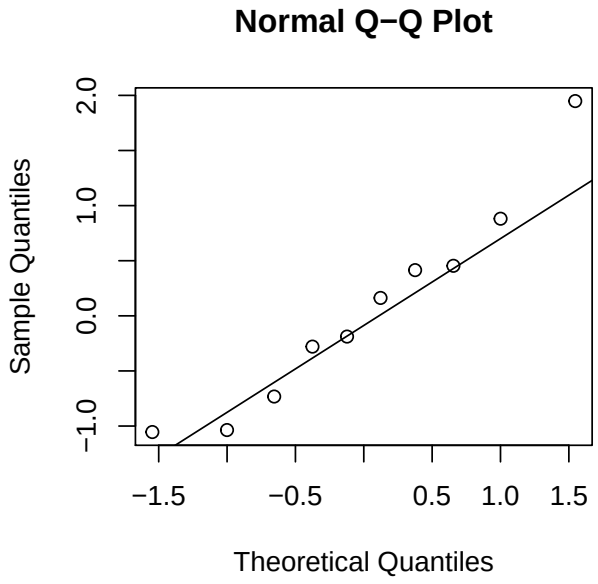


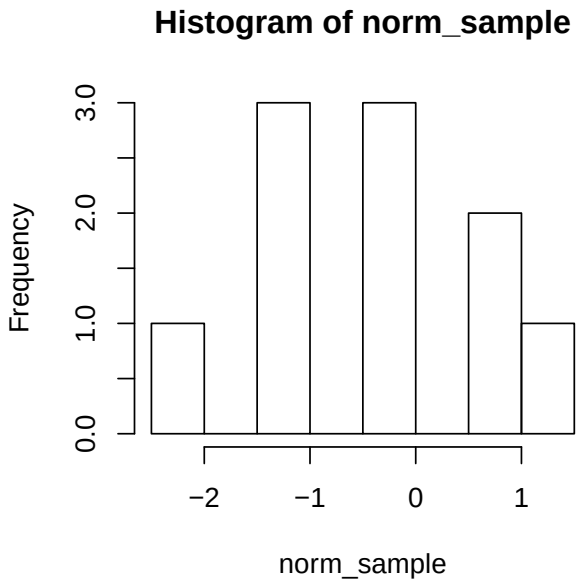

```
qqn(norm_sample)
```

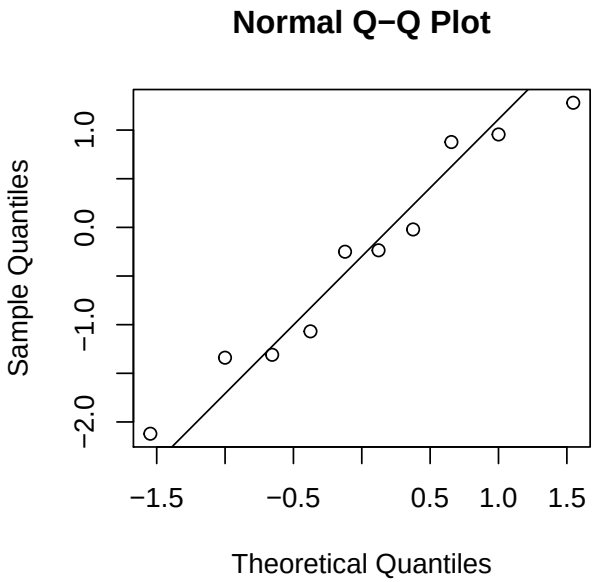


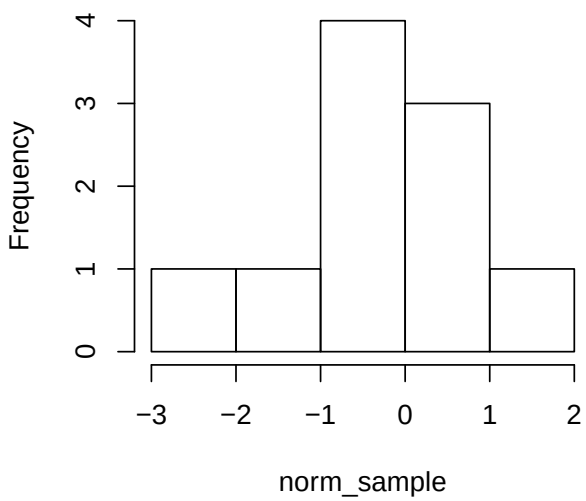

```r
par(mfrow=c(1,2))
for (i in 1:20) {
  norm_sample <- rnorm(n = 10, mean = 0, sd = 1)

  hist(norm_sample)

  qqn(norm_sample)
}
```

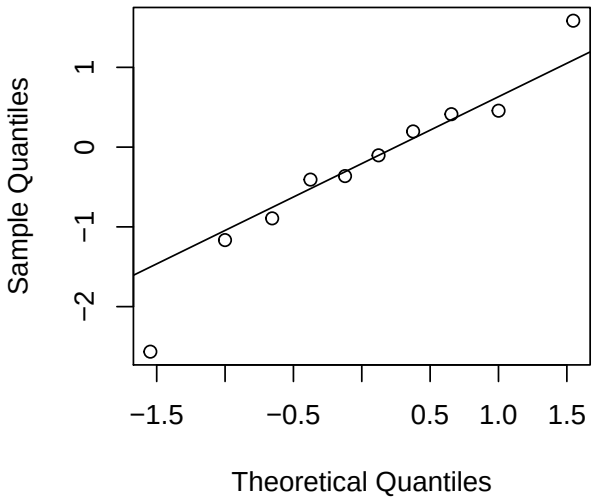


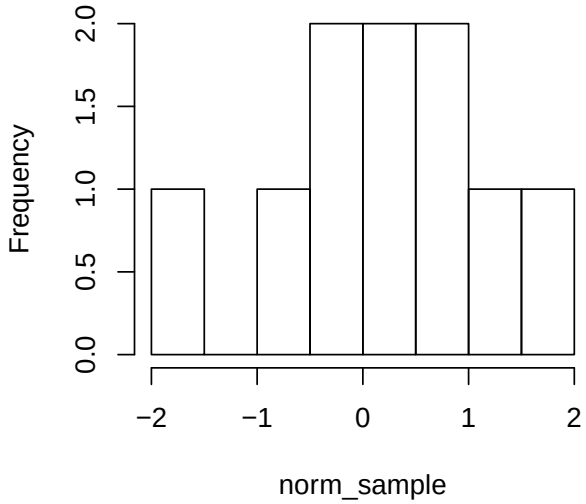


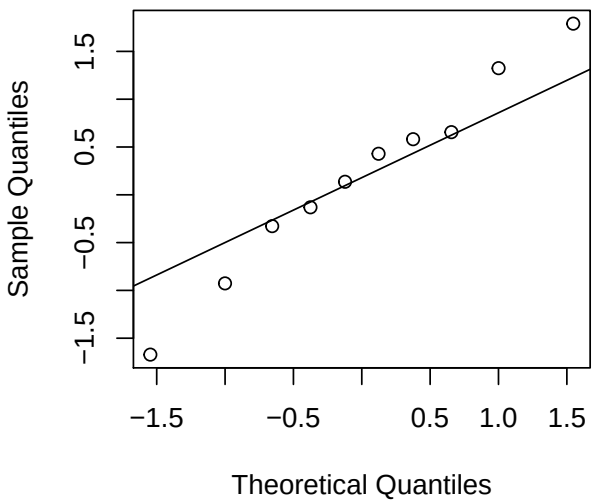

Histogram of norm_sample
Frequency
0.0
1.0
2.0
3.0
−2
−1
0
1
norm_sample

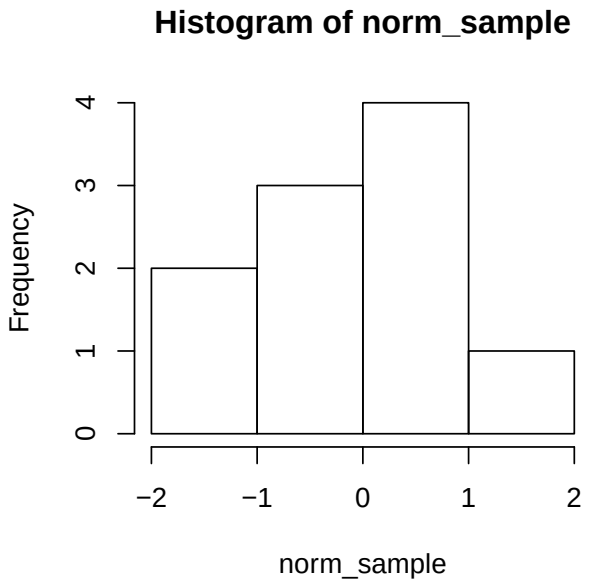

Normal Q−Q Plot
Sample Quantiles
−2.0
−1.0
0.0
1.0
−1.5
−0.5
0.5
1.0
1.5
Theoretical Quantiles

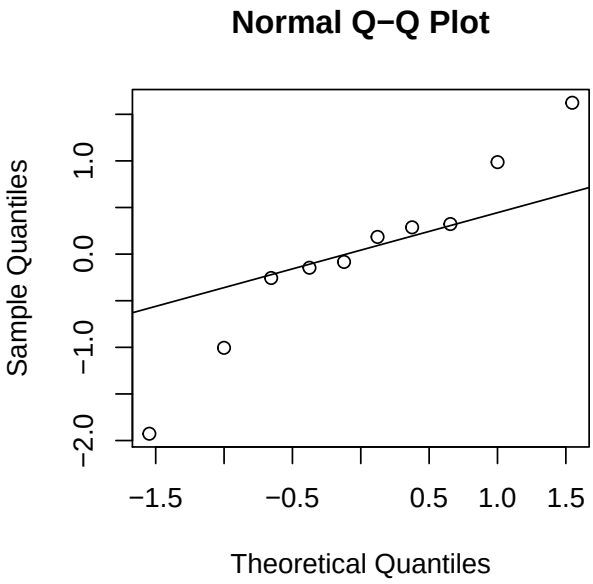

Histogram of norm_sample
Frequency
0
1
2
3
4
−3
−2
−1
0
1
2
norm_sample

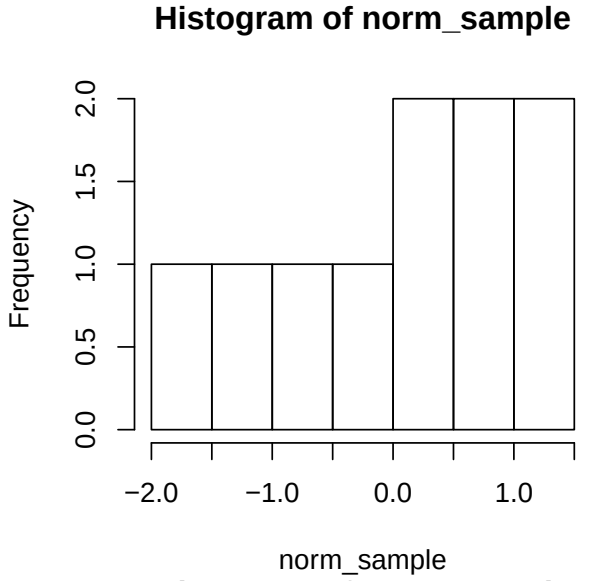

Normal Q−Q Plot
Sample Quantiles
−2
−1
0
1
−1.5
−0.5
0.5
1.0
1.5
Theoretical Quantiles

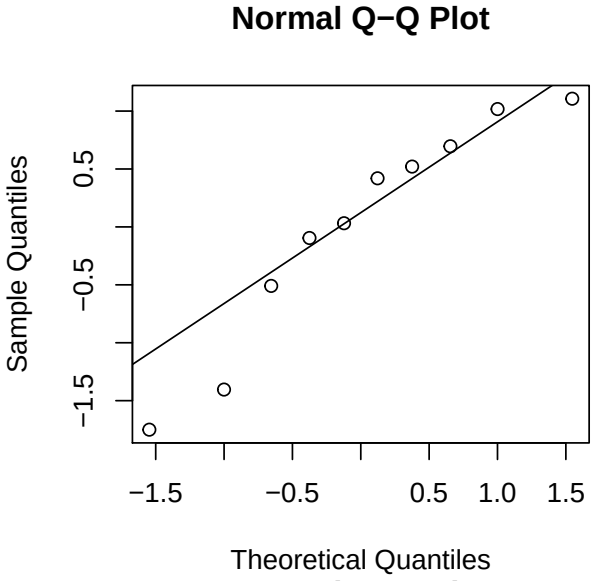

Histogram of norm_sample
Frequency
0.0
0.5
1.0
1.5
2.0
−2
−1
0
1
2
norm_sample

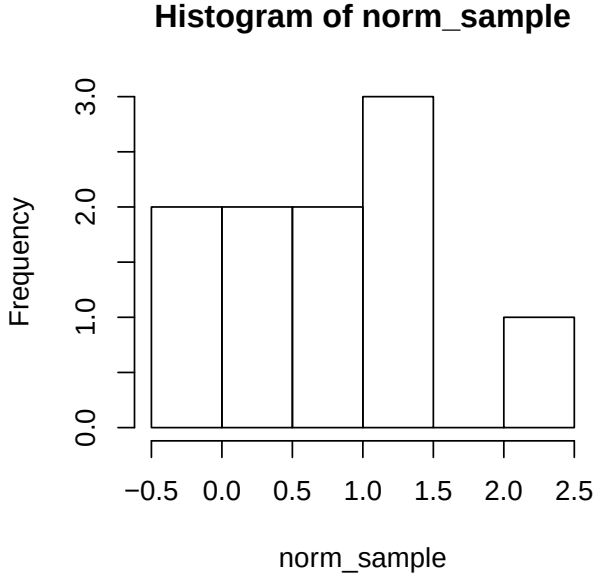

Normal Q−Q Plot
Sample Quantiles
−1.5
−0.5
0.5
1.5
−1.5
−0.5
0.5
1.0
1.5
Theoretical Quantiles

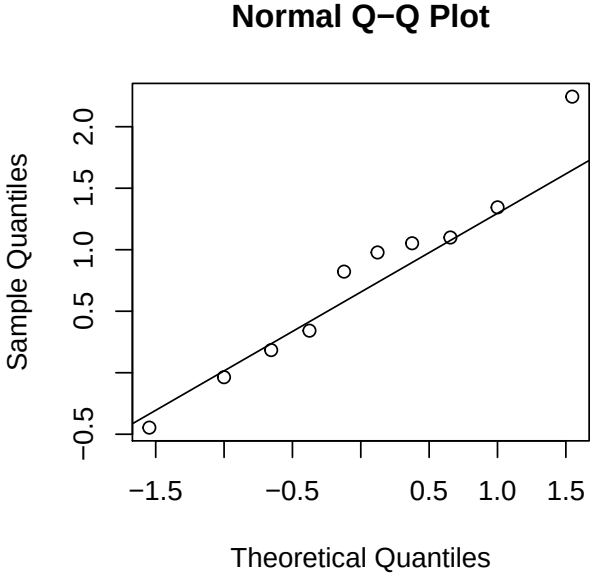

Histogram of norm_sample
Frequency
0
1
2
3
4
−2
−1
0
1
2
norm_sample

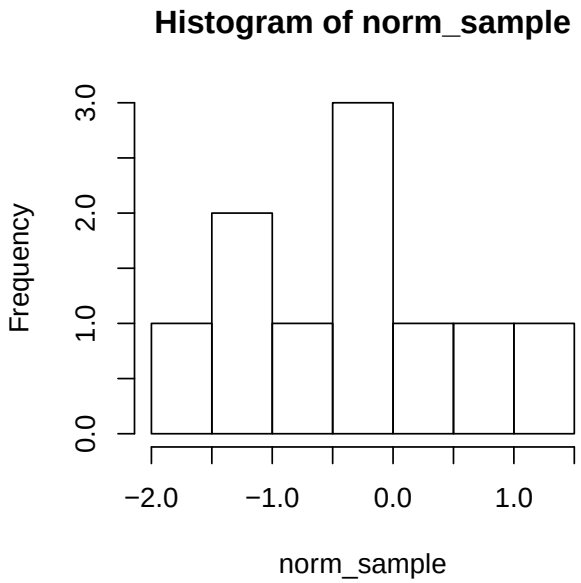

Normal Q−Q Plot
Sample Quantiles
−2.0
−1.0
0.0
1.0
−1.5
−0.5
0.5
1.0
1.5
Theoretical Quantiles

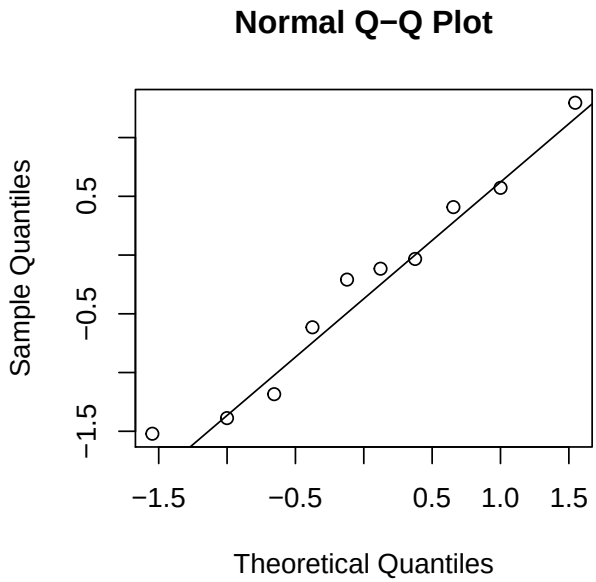

Histogram of norm_sample
Frequency
0.0
0.5
1.0
1.5
2.0
−2.0
−1.0
0.0
1.0
norm_sample

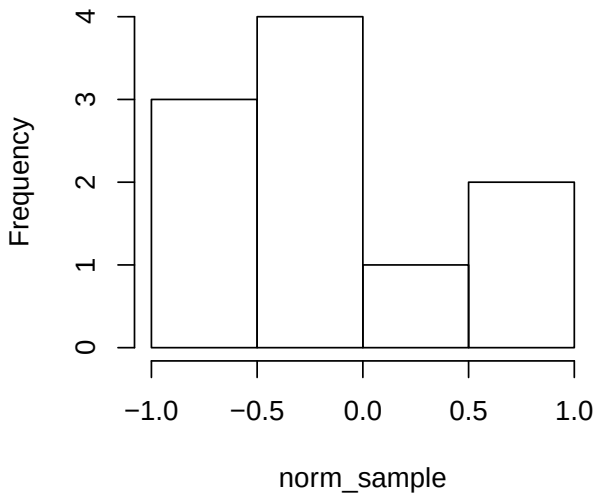

Normal Q−Q Plot
Sample Quantiles
−1.5
−0.5
0.5
1.0
1.5
−1.5
−0.5
0.5
1.0
1.5
Theoretical Quantiles

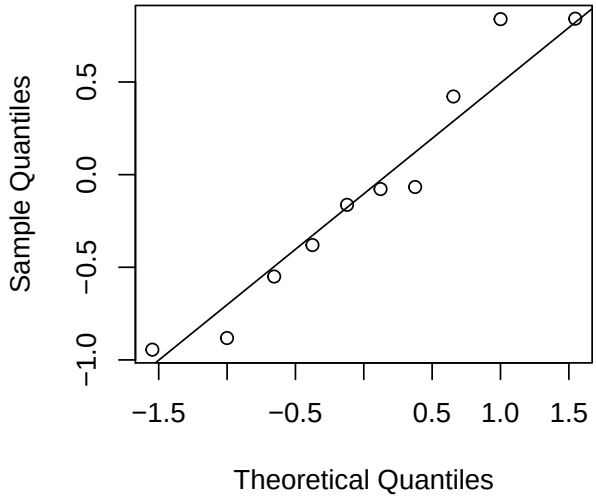

Histogram of norm_sample
Frequency
0.0
1.0
2.0
3.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
norm_sample

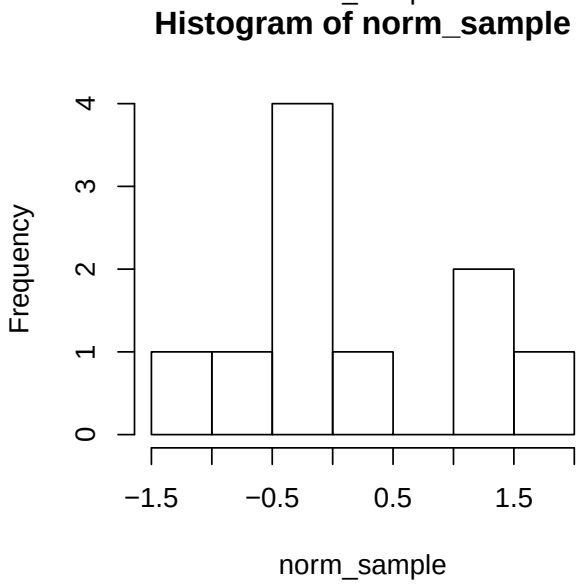

Normal Q−Q Plot
Sample Quantiles
−0.5
0.5
1.0
1.5
2.0
−1.5
−0.5
0.5
1.0
1.5
Theoretical Quantiles

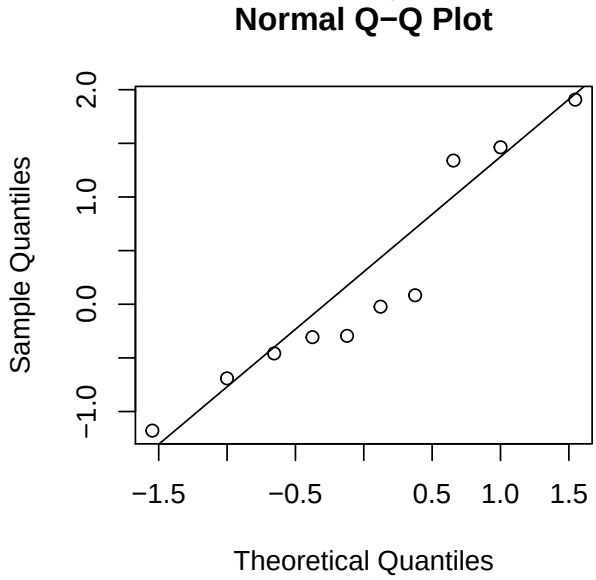

Histogram of norm_sample
Frequency
0.0
1.0
2.0
3.0
−2.0
−1.0
0.0
1.0
norm_sample

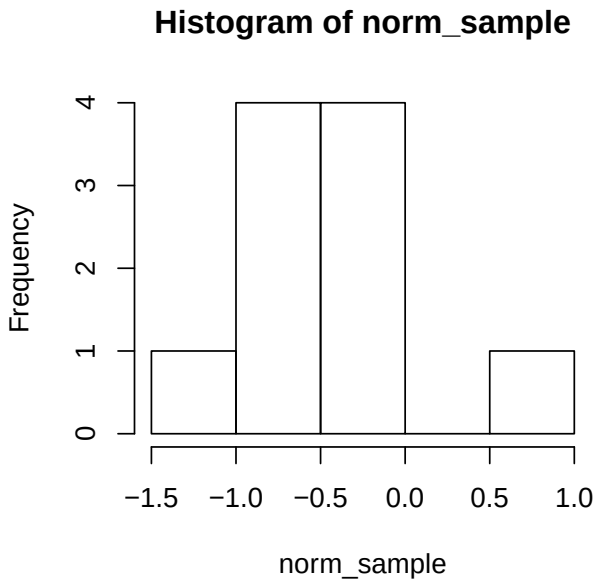

Normal Q−Q Plot
Sample Quantiles
−1.5
−0.5
0.5
−1.5
−0.5
0.5
1.0
1.5
Theoretical Quantiles

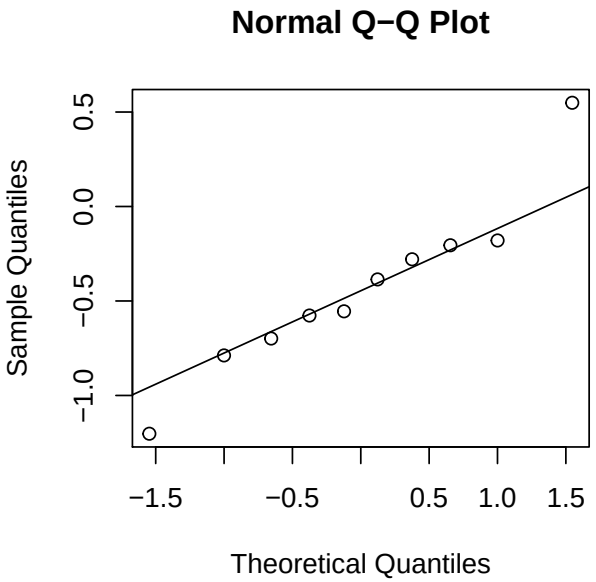

Histogram of norm_sample
Frequency
0
1
2
3
4
−1.0
−0.5
0.0
0.5
1.0
norm_sample

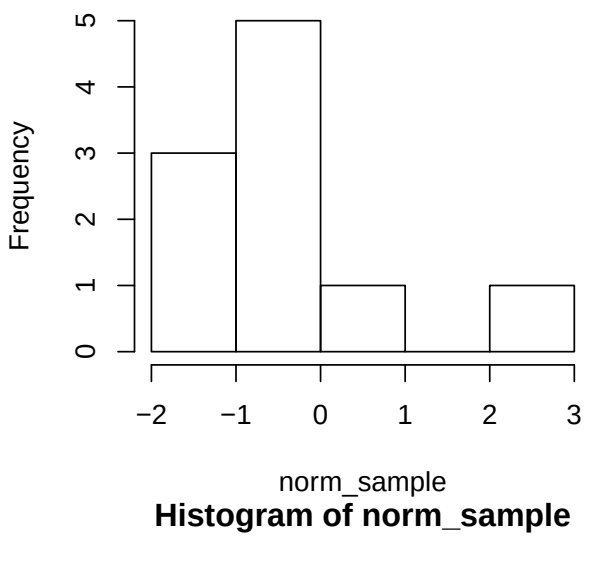

Normal Q−Q Plot
Sample Quantiles
−1.0
−0.5
0.0
0.5
−1.5
−0.5
0.5
1.0
1.5
Theoretical Quantiles

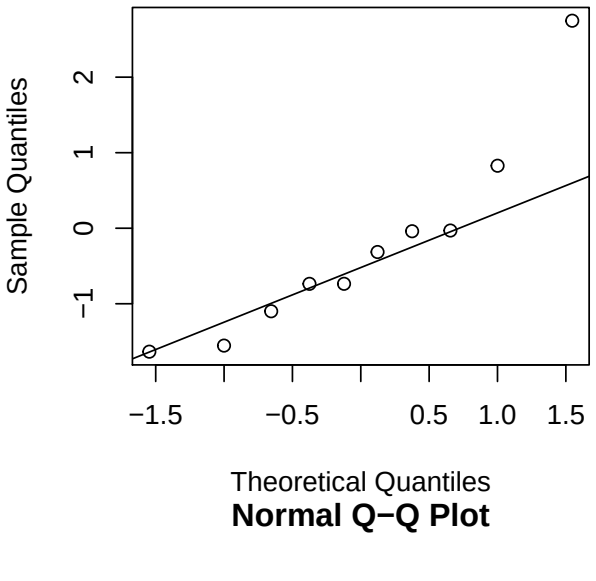

Histogram of norm_sample
Frequency
0
1
2
3
4
−1.5
−0.5
0.5
1.5
norm_sample

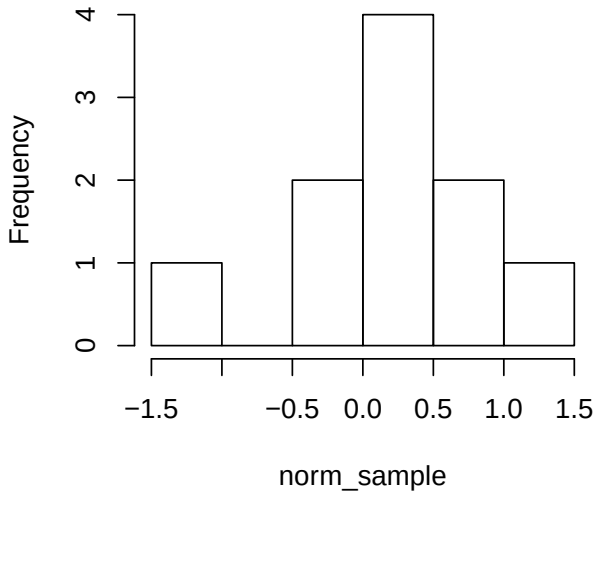

Normal Q−Q Plot
Sample Quantiles
−1.0
0.0
1.0
2.0
−1.5
−0.5
0.5
1.0
1.5
Theoretical Quantiles

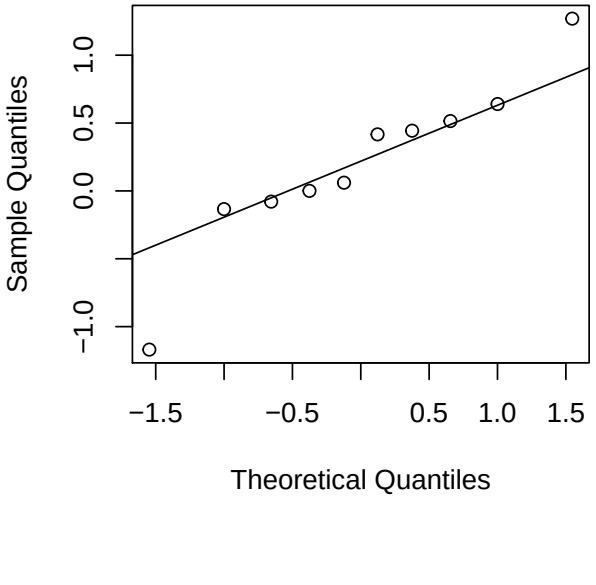

Histogram of norm_sample
Frequency
0
1
2
3
4
−1.5
−1.0
−0.5
0.0
0.5
1.0
norm_sample

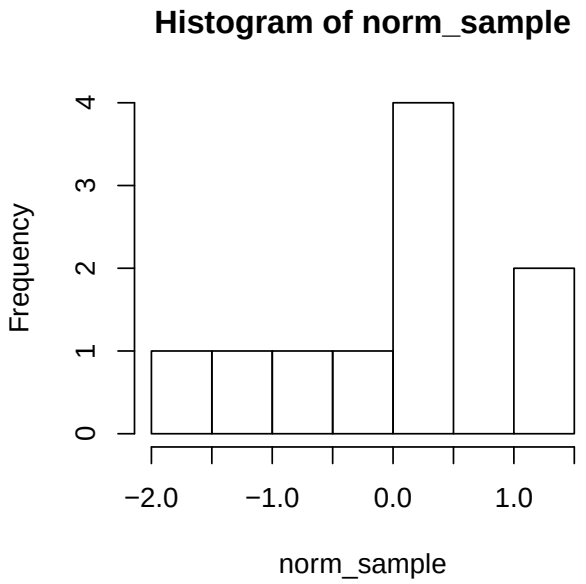

Normal Q−Q Plot
Sample Quantiles
−1.0
−0.5
0.0
0.5
−1.5
−0.5
0.5
1.0
1.5
Theoretical Quantiles

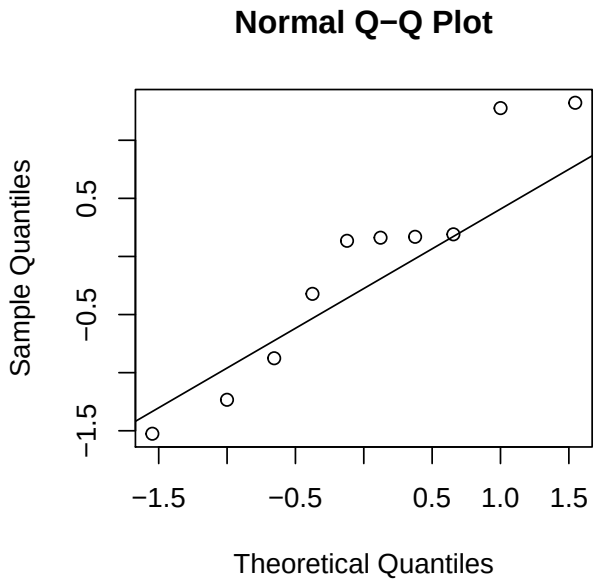

Histogram of norm_sample
Frequency
0
1
2
3
4
5
−2
−1
0
1
2
3
norm_sample

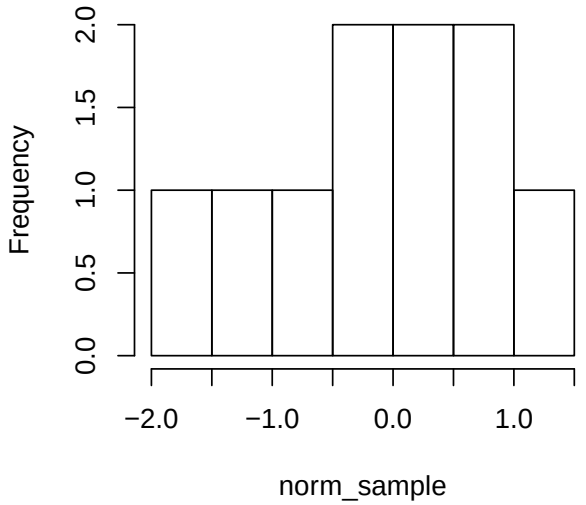

Normal Q−Q Plot
Sample Quantiles
−1
0
1
2
−1.5
−0.5
0.5
1.0
1.5
Theoretical Quantiles

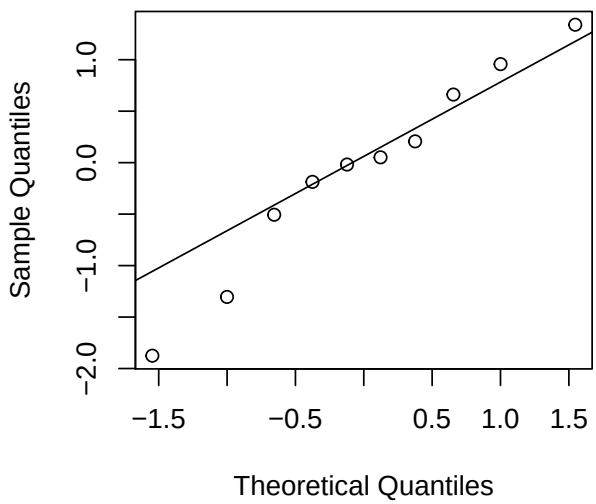

Histogram of norm_sample
Frequency
0
1
2
3
4
−1.5
−0.5
0.0
0.5
1.0
1.5
norm_sample

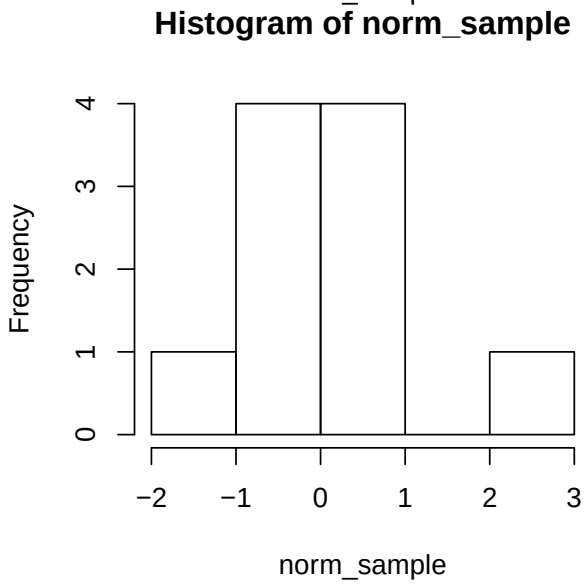

Normal Q−Q Plot
Sample Quantiles
−1.0
0.0
0.5
1.0
−1.5
−0.5
0.5
1.0
1.5
Theoretical Quantiles

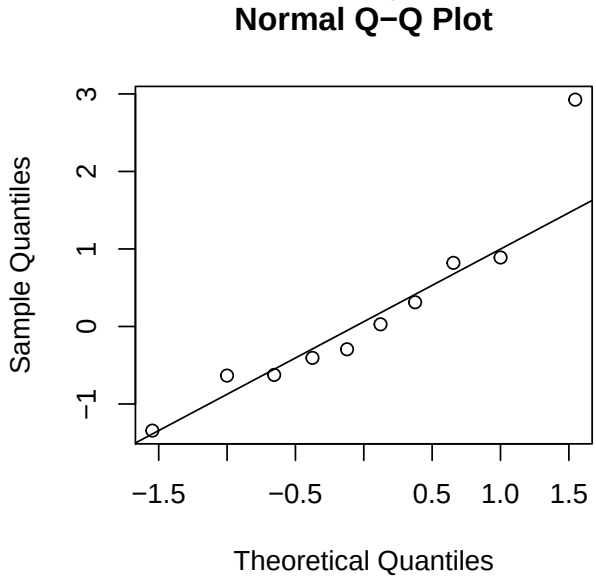

Histogram of norm_sample
Frequency
0
1
2
3
4
−2.0
−1.0
0.0
1.0
norm_sample

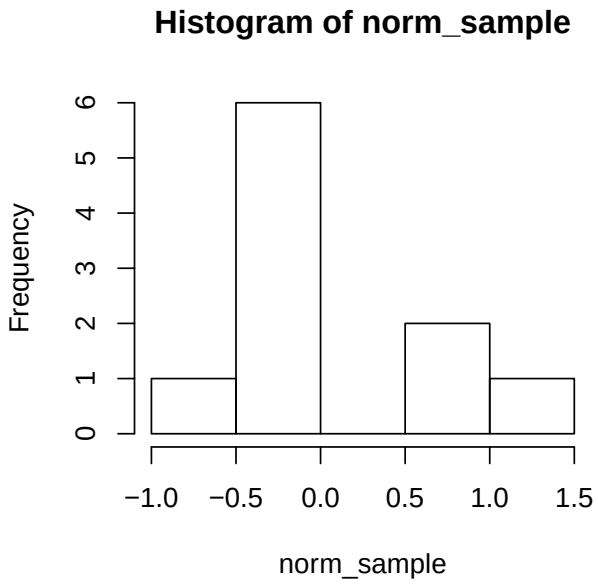

Normal Q−Q Plot
Sample Quantiles
−1.5
−0.5
0.5
−1.5
−0.5
0.5
1.0
1.5
Theoretical Quantiles

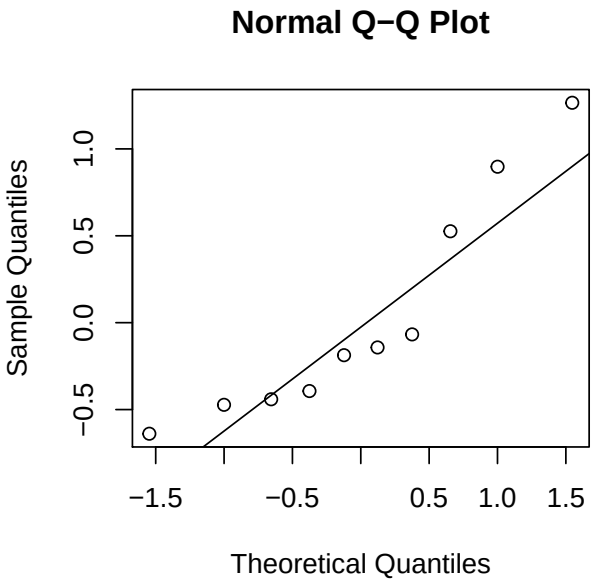

Histogram of norm_sample
Frequency
0.0
0.5
1.0
1.5
2.0
−2.0
−1.0
0.0
1.0
norm_sample

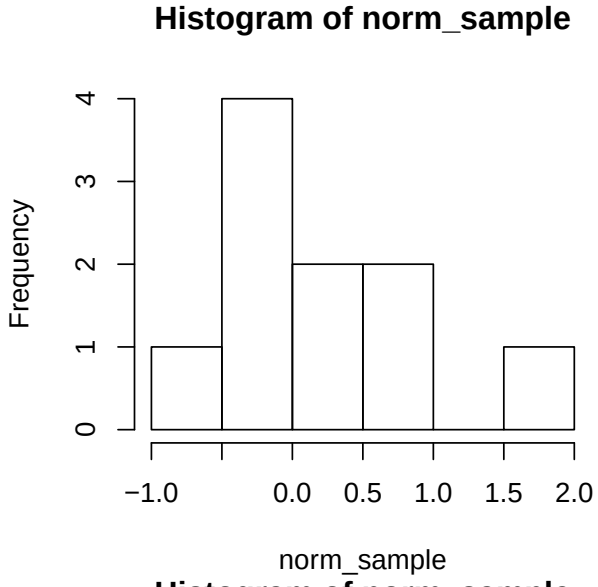

Normal Q−Q Plot
Sample Quantiles
−2.0
−1.0
0.0
1.0
−1.5
−0.5
0.5
1.0
1.5
Theoretical Quantiles

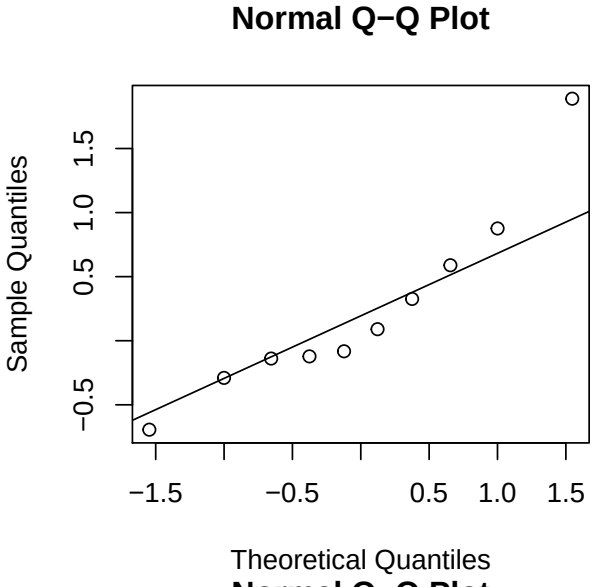

Histogram of norm_sample
Frequency
0
1
2
3
4
−2
−1
0
1
2
3
norm_sample

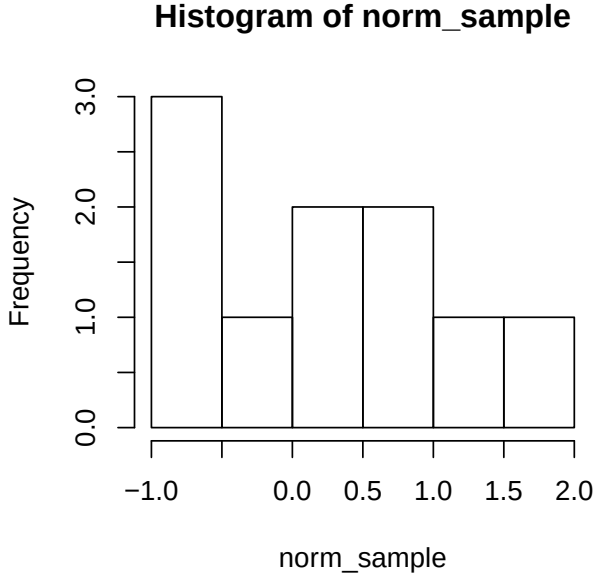

Normal Q−Q Plot
Sample Quantiles
−1
0
1
2
3
−1.5
−0.5
0.5
1.0
1.5
Theoretical Quantiles

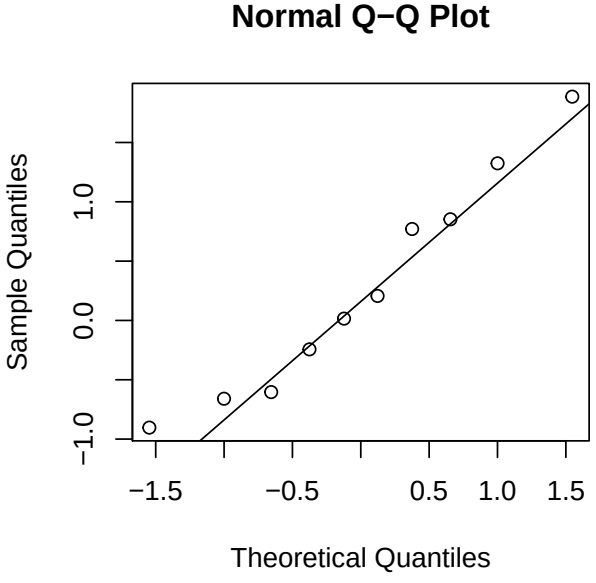

Histogram of norm_sample
Frequency
0
1
2
3
4
5
6
−1.0
−0.5
0.0
0.5
1.0
1.5
norm_sample

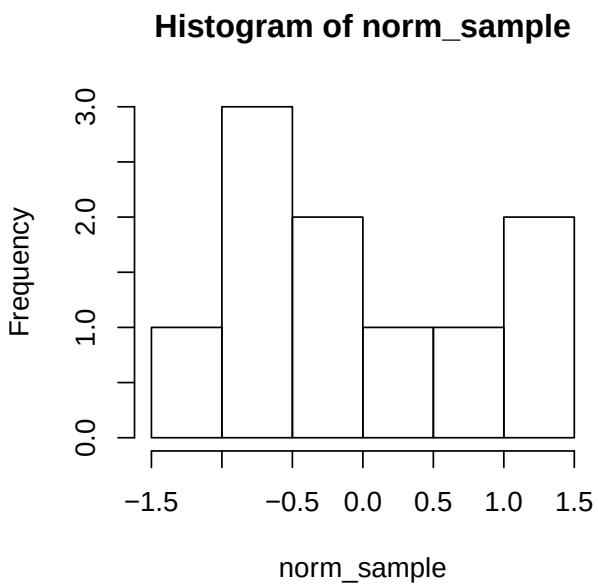

Normal Q−Q Plot
Sample Quantiles
−0.5
0.0
0.5
1.0
−1.5
−0.5
0.5
1.0
1.5
Theoretical Quantiles

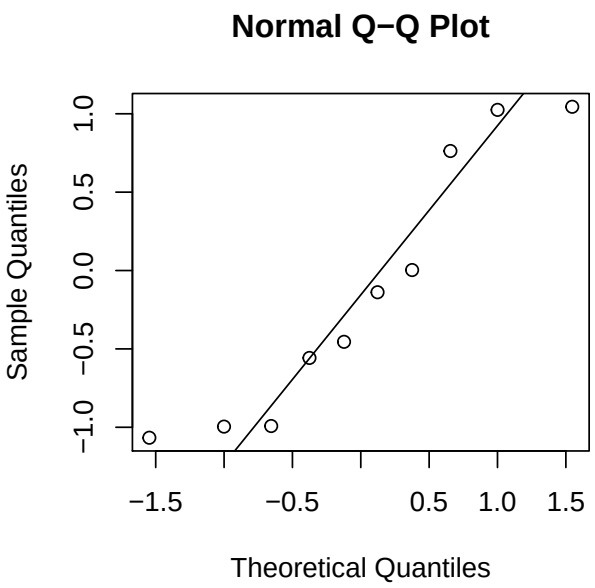

Histogram of norm_sample
Frequency
0
1
2
3
4
−1.0
0.0
0.5
1.0
1.5
2.0
norm_sample


Normal Q−Q Plot
Sample Quantiles
−0.5
0.5
1.0
1.5
−1.5
−0.5
0.5
1.0
1.5
Theoretical Quantiles


Histogram of norm_sample
Frequency
0.0
1.0
2.0
3.0
−1.0
0.0
0.5
1.0
1.5
2.0
norm_sample


Normal Q−Q Plot
Sample Quantiles
−1.0
0.0
1.0
−1.5
−0.5
0.5
1.0
1.5
Theoretical Quantiles

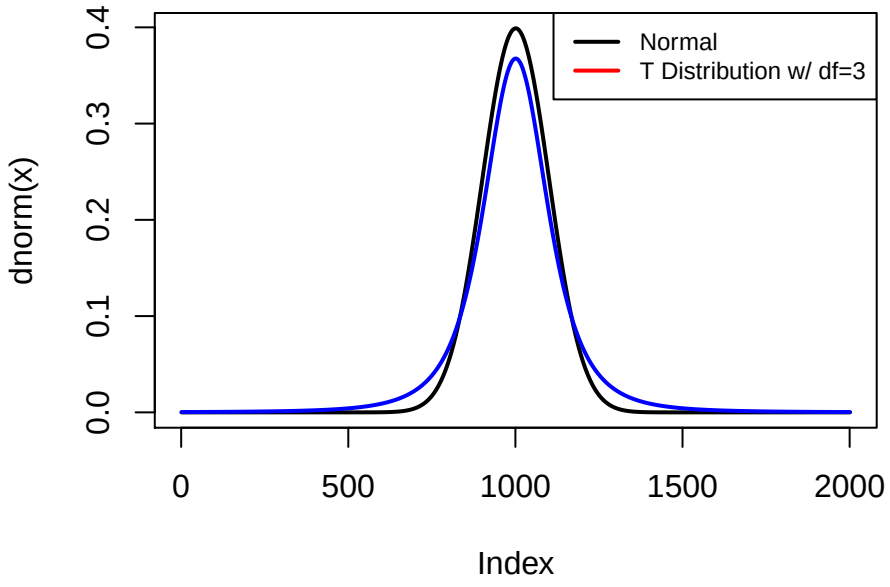




Problem 2

One way to model data from a heavy-tailed distribution is to use a t distribution. The family of t distributions is indexed by a parameter called the degrees of freedom. For this question we will use the t distribution with 3 degrees of freedom. We will denote this distribution as t(3). Make a plot of the pdf of the t distribution with 3 degrees of freedom. Add the standard normal distribution to the plot. Describe the differences between the two distributions.

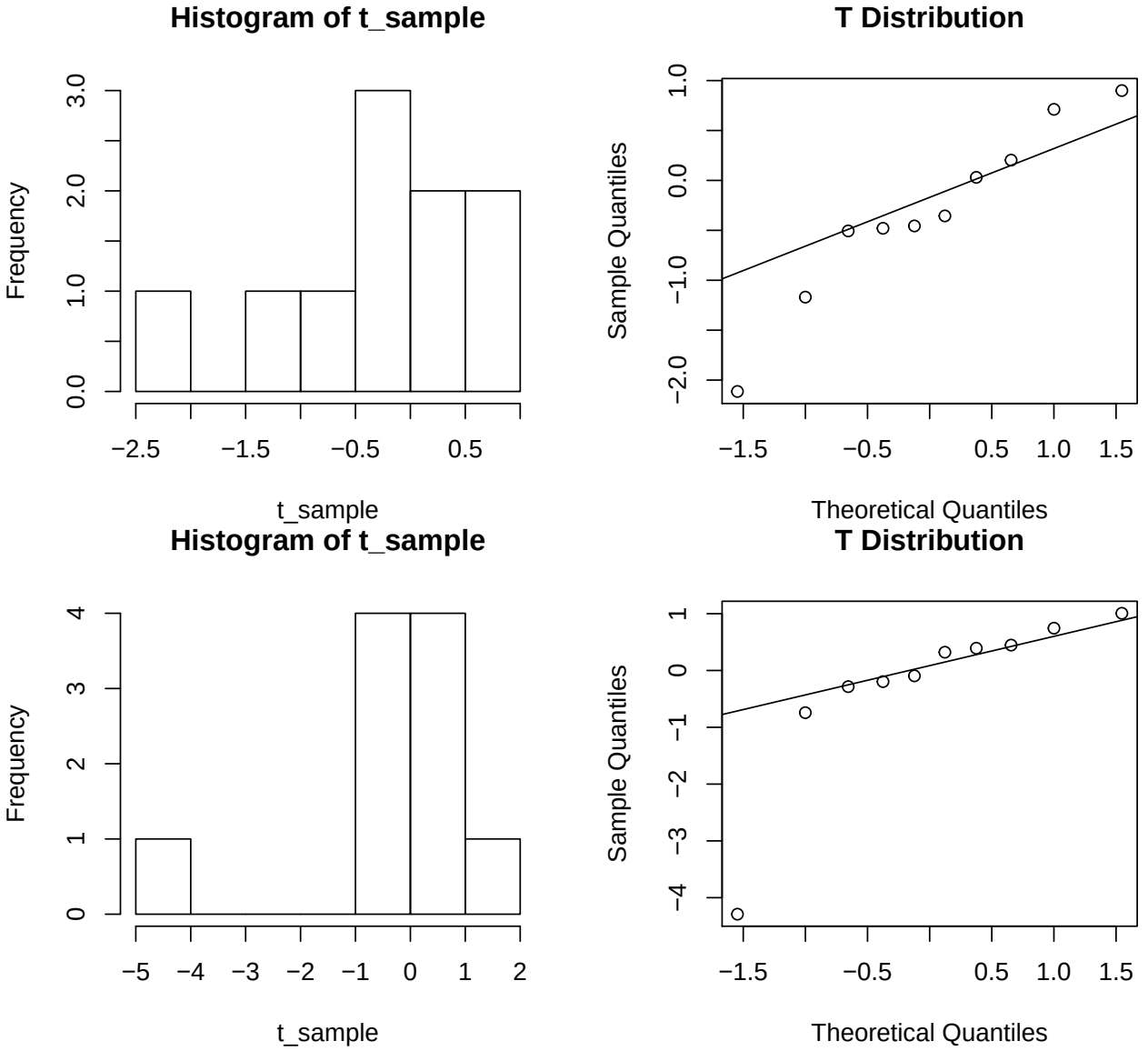
```
x <- seq(-10, 10, 0.01)
plot(dnorm(x), lwd=2, type='l')
lines(dt(x, df=3), col='blue',lwd=2)
legend("topright",
       col=c('black', 'red'),
       legend = c("Normal","T Distribution w/ df=3"),
       lty=1 , lwd=2, cex=0.75)
```

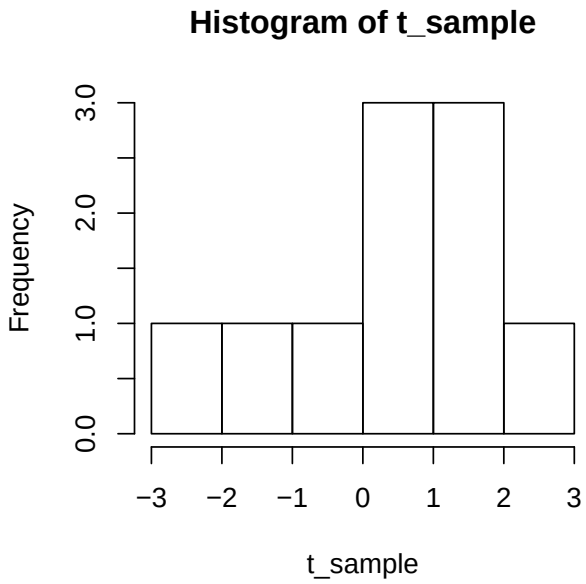


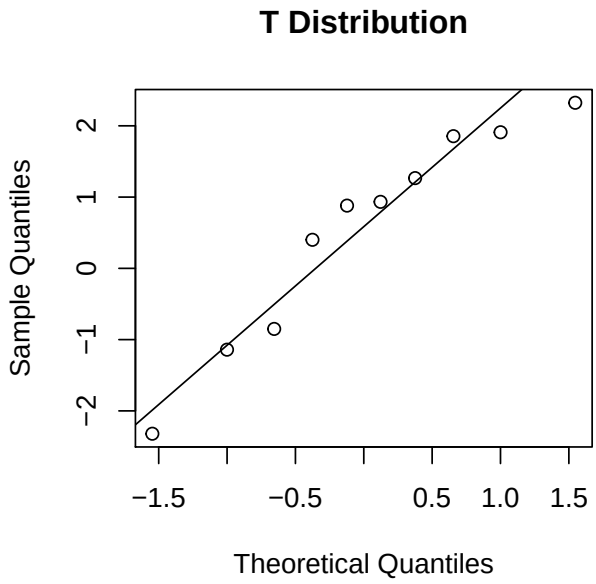
Problem 3

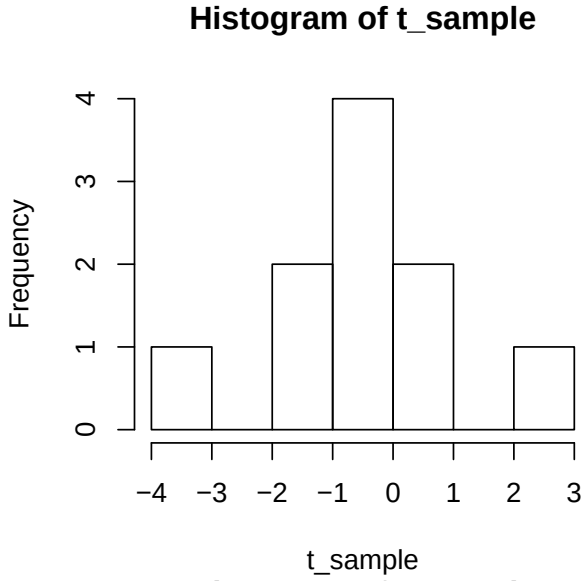
Repeat question 1 using samples from the t(3) distribution. Compare results with the results of question 1.

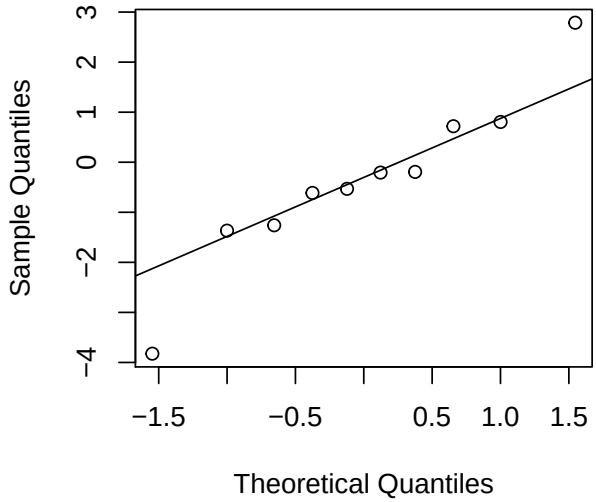
```
par(mfrow=c(1, 2))
for (i in 1:20) {
  t_sample <- rt(n = 10, df=3)

  hist(t_sample)

  qqn(t_sample, main = "T Distribution")
}
```

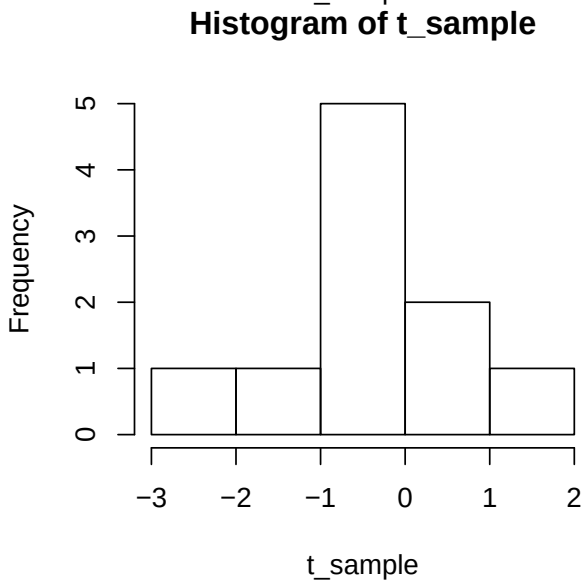


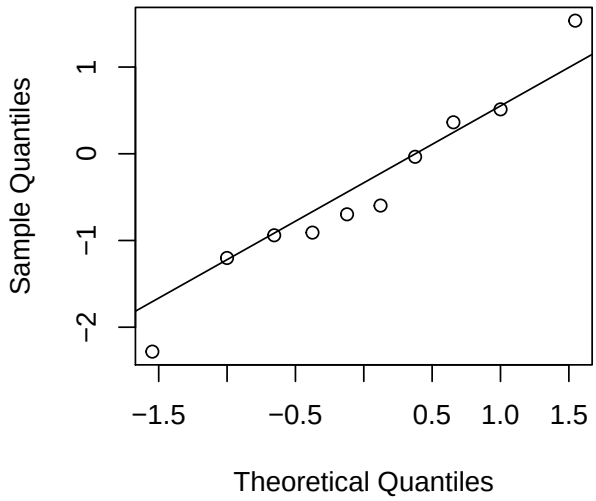

Histogram of t_sample
Frequency
0.0
1.0
2.0
3.0
−3
−2
−1
0
1
2
3
t_sample

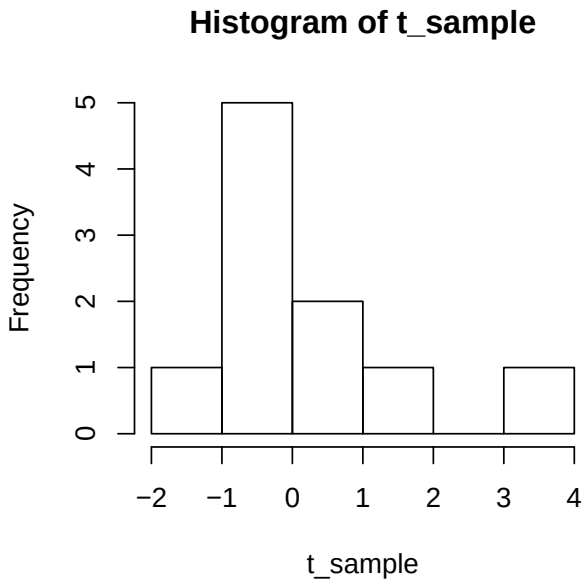

T Distribution
Sample Quantiles
−2
−1
0
1
2
−1.5
−0.5
0.5
1.0
1.5
Theoretical Quantiles

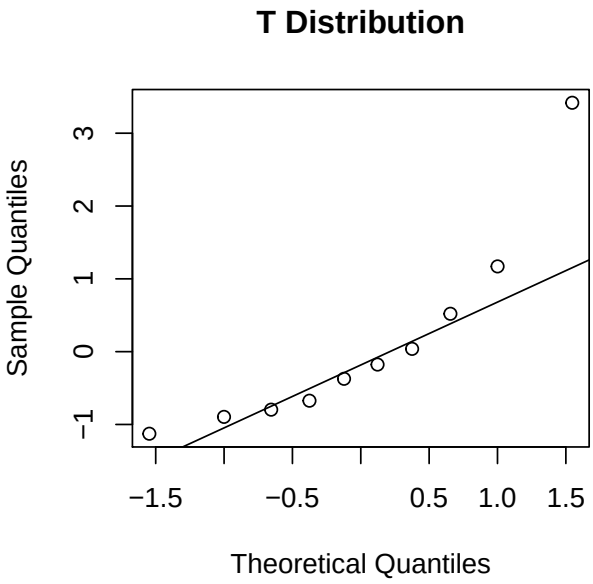

Histogram of t_sample
Frequency
0
1
2
3
4
−4
−3
−2
−1
0
1
2
3
t_sample

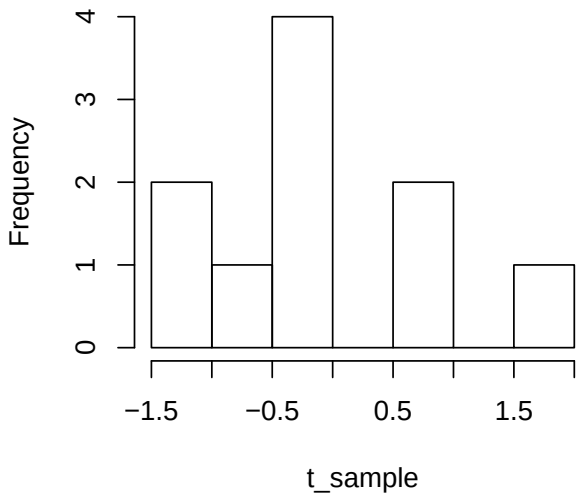

T Distribution
Sample Quantiles
−4
−2
0
1
2
3
−1.5
−0.5
0.5
1.0
1.5
Theoretical Quantiles

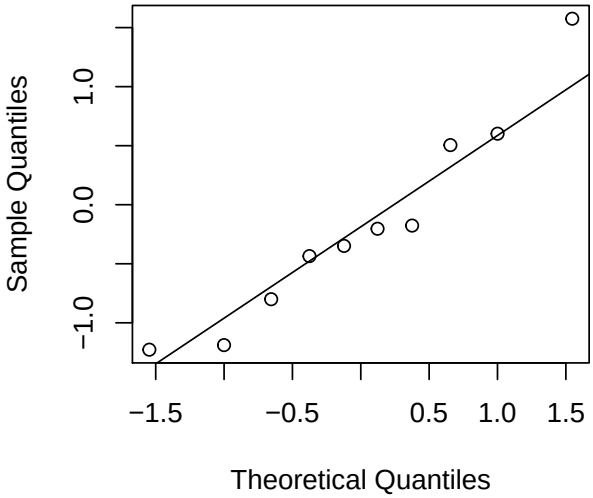

Histogram of t_sample
Frequency
0
1
2
3
4
5
−3
−2
−1
0
1
2
t_sample

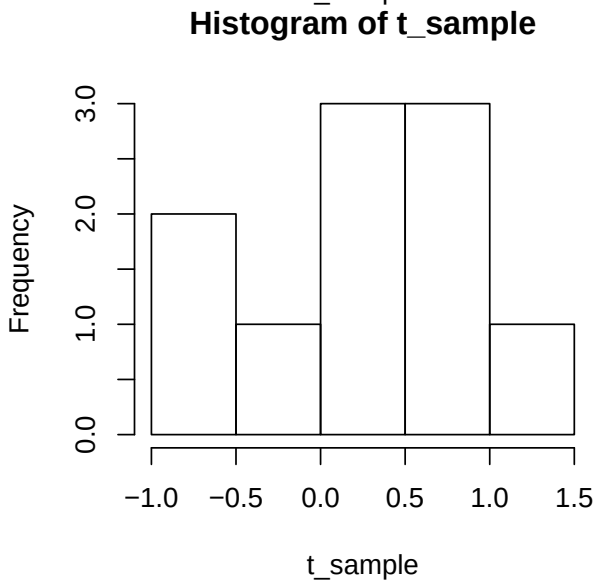

T Distribution
Sample Quantiles
−2
−1
0
1
−1.5
−0.5
0.5
1.0
1.5
Theoretical Quantiles

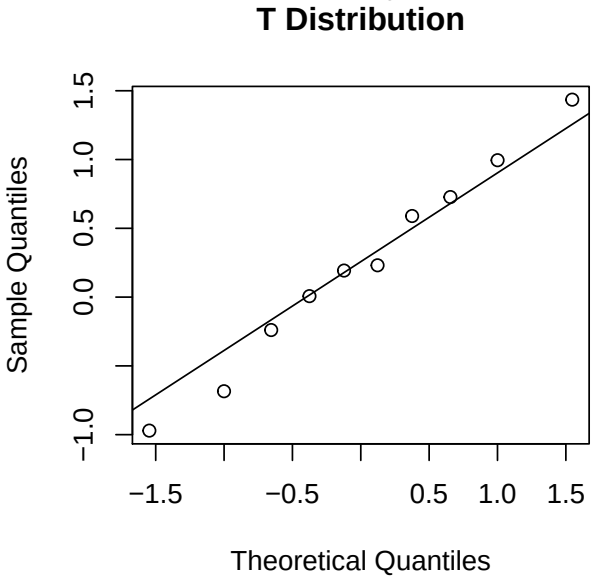

Histogram of t_sample
Frequency
t_sample
0
1
2
3
4
5
−2
−1
0
1
2
3
4

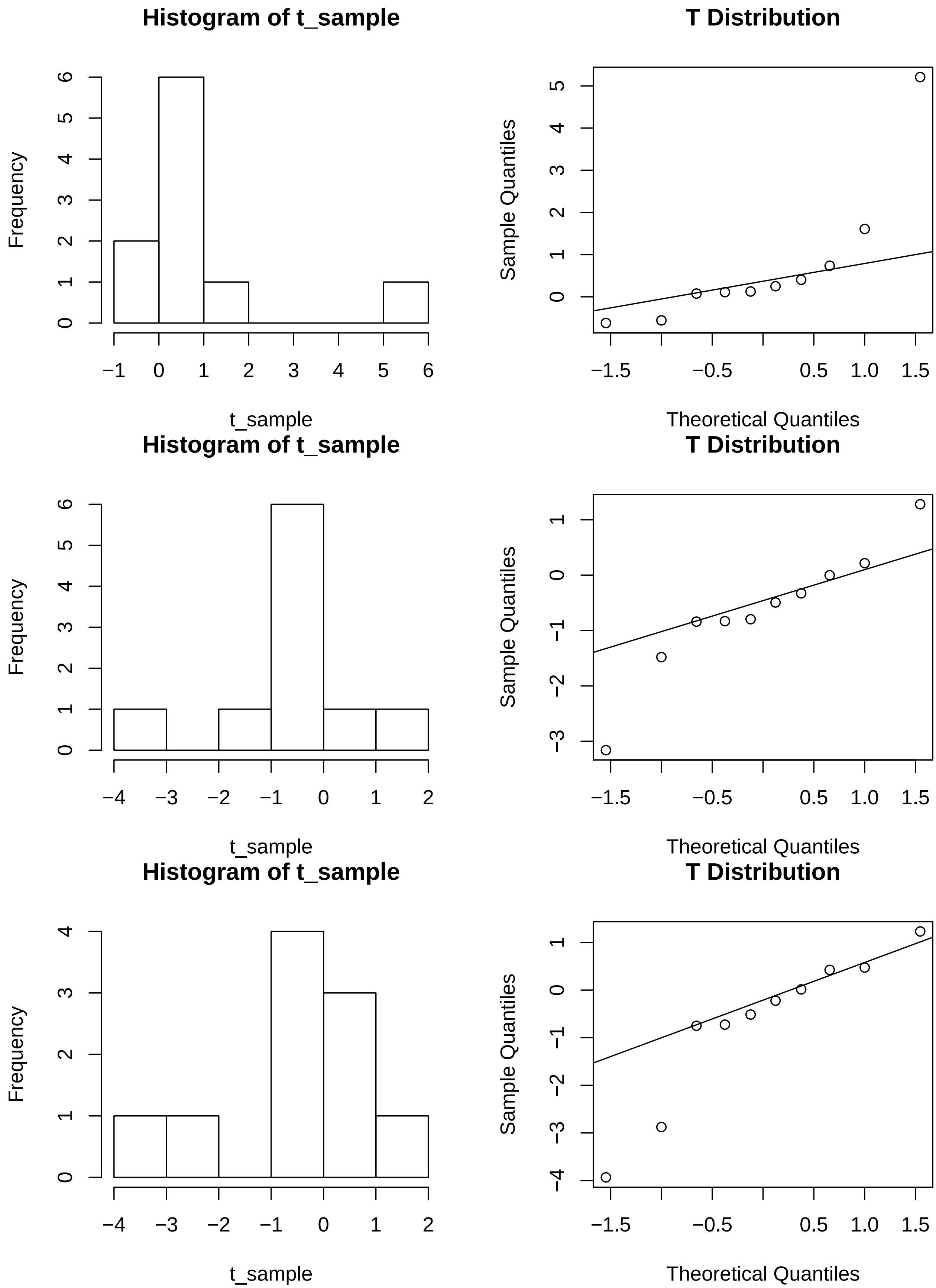

T Distribution
Sample Quantiles
Theoretical Quantiles
−1
0
1
2
3
−1.5
−0.5
0.5
1.0
1.5

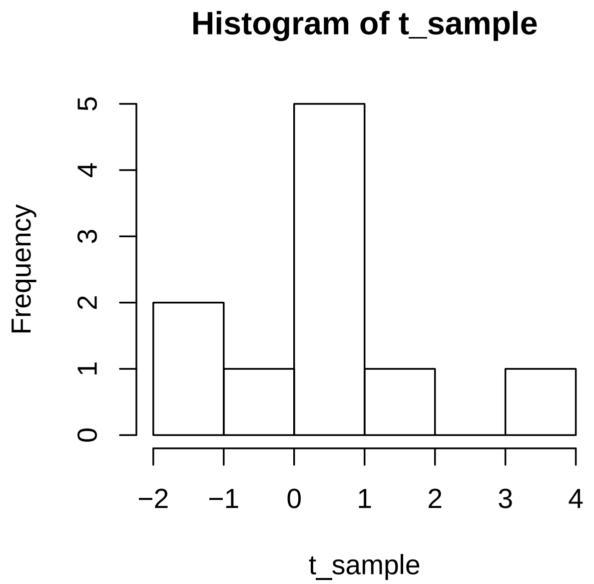

Histogram of t_sample
Frequency
t_sample
0
1
2
3
4
−1.5
−0.5
0.5
1.5

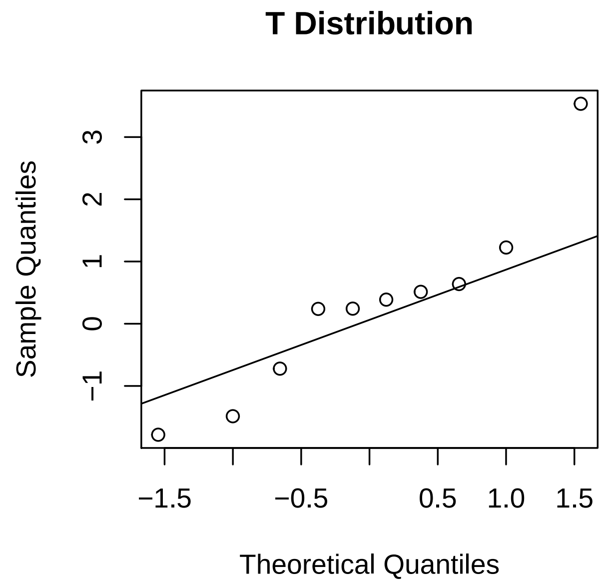

T Distribution
Sample Quantiles
Theoretical Quantiles
−1.0
0.0
1.0
−1.5
−0.5
0.5
1.0
1.5

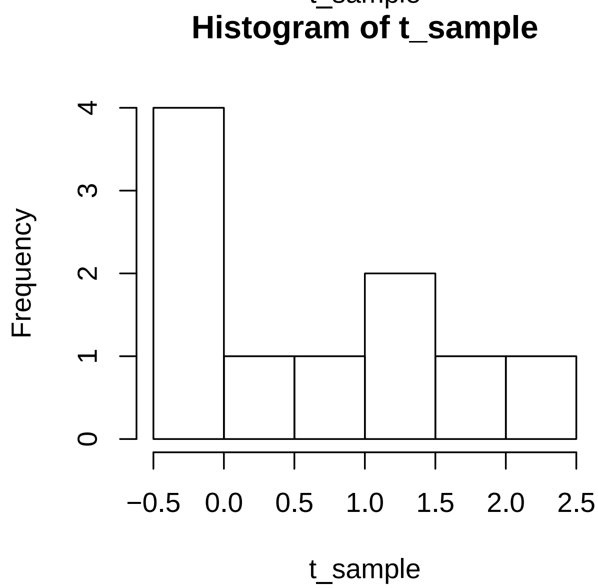

Histogram of t_sample
Frequency
t_sample
0.0
1.0
2.0
3.0
−1.0
−0.5
0.0
0.5
1.0
1.5

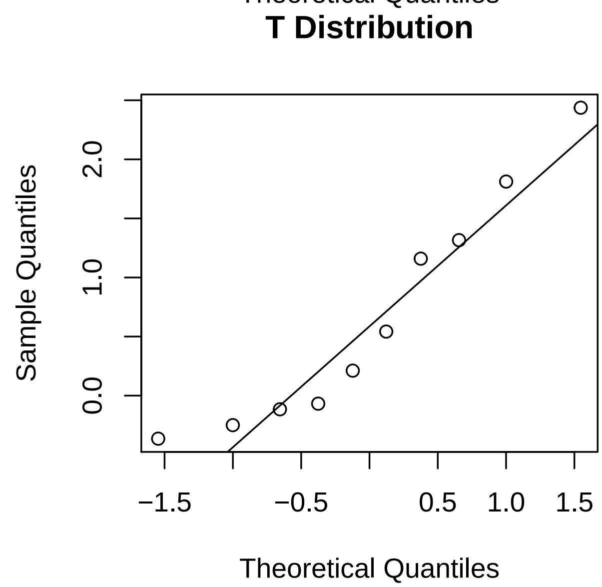

T Distribution
Sample Quantiles
Theoretical Quantiles
−1.0
0.0
0.5
1.0
1.5
−1.5
−0.5
0.5
1.0
1.5

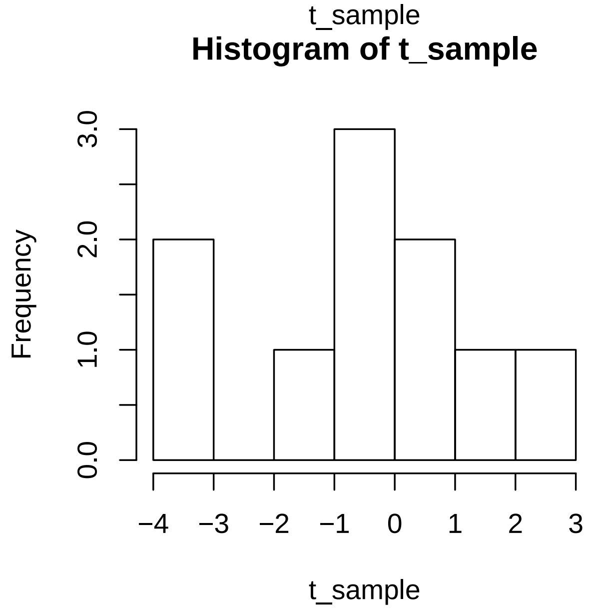

Histogram of t_sample
Frequency
t_sample
T Distribution
Sample Quantiles
Theoretical Quantiles
Histogram of t_sample
Frequency
t_sample
T Distribution
Sample Quantiles
Theoretical Quantiles
Histogram of t_sample
Frequency
t_sample
T Distribution
Sample Quantiles
Theoretical Quantiles

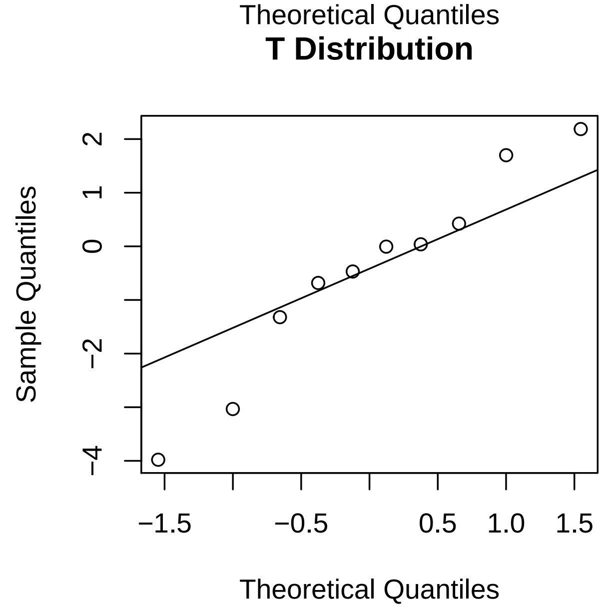

Histogram of t_sample
Frequency
0 1 2 3 4 5
−2 −1 0 1 2 3 4
t_sample

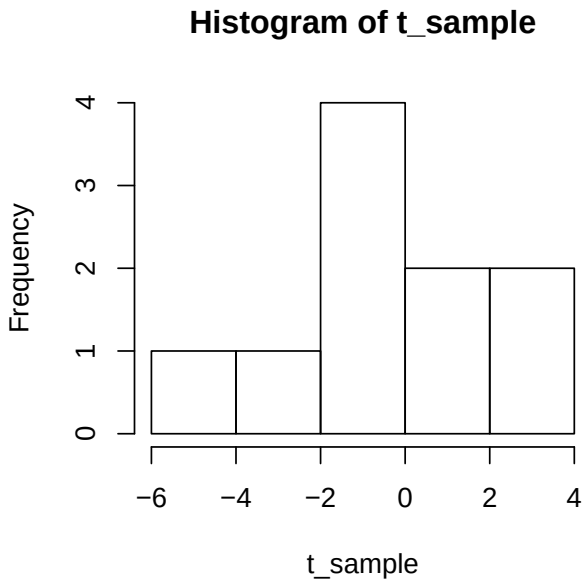

T Distribution
Sample Quantiles
−1 0 1 2 3
−1.5 −0.5 0.5 1.0 1.5
Theoretical Quantiles

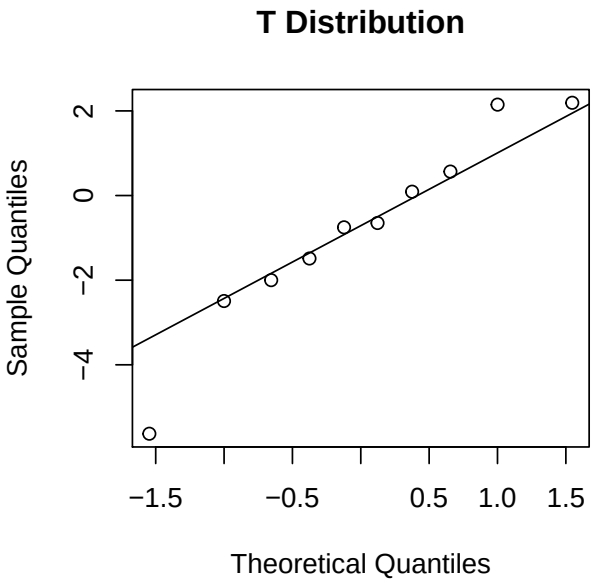

Histogram of t_sample
Frequency
0 1 2 3 4
−0.5 0.0 0.5 1.0 1.5 2.0 2.5
t_sample

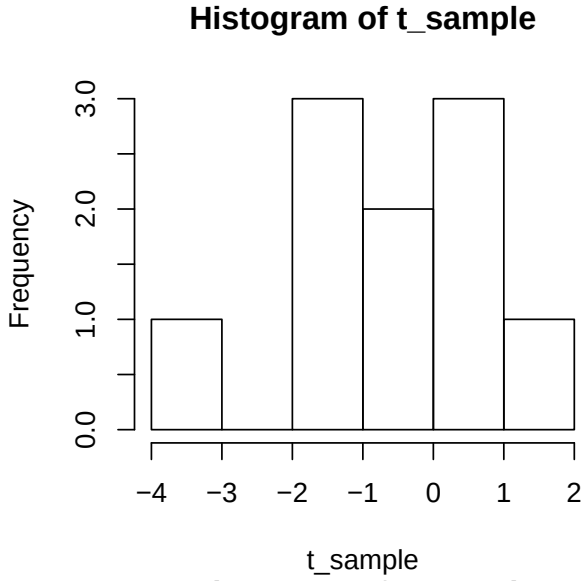

T Distribution
Sample Quantiles
0.0 1.0 2.0
−1.5 −0.5 0.5 1.0 1.5
Theoretical Quantiles

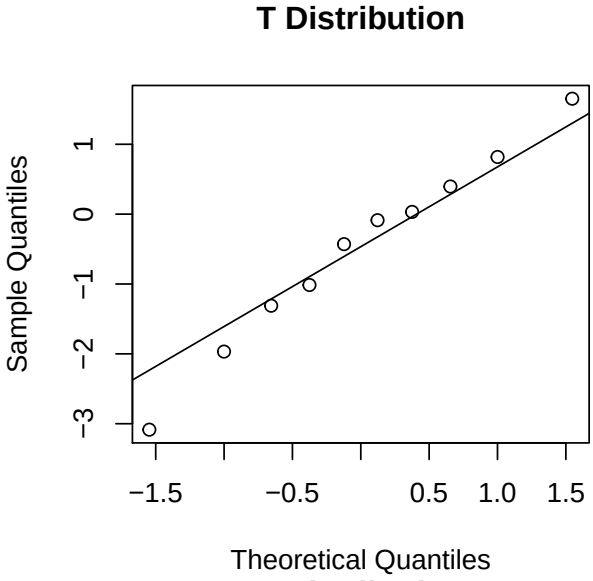

Histogram of t_sample
Frequency
0.0 1.0 2.0 3.0
−4 −3 −2 −1 0 1 2 3
t_sample

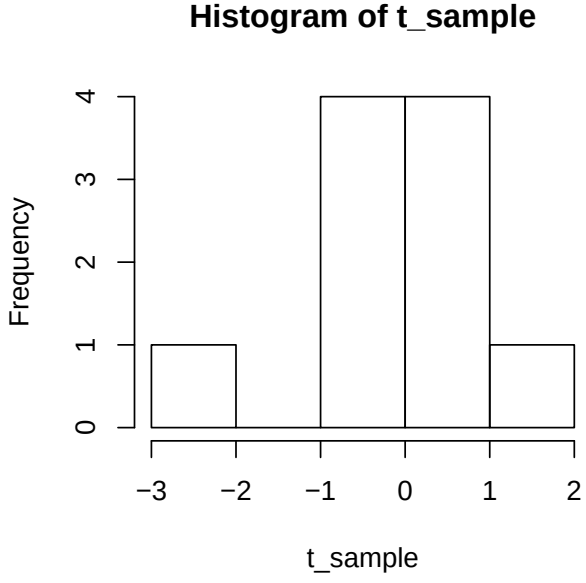

T Distribution
Sample Quantiles
−4 −2 0 1 2
−1.5 −0.5 0.5 1.0 1.5
Theoretical Quantiles

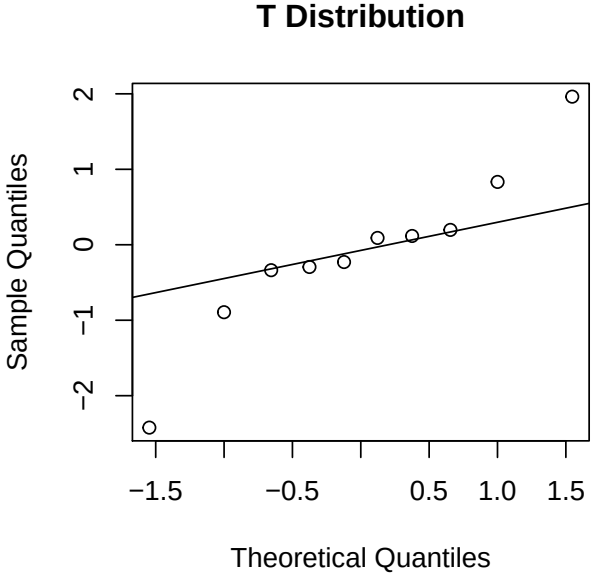

Histogram of t_sample
Frequency
t_sample
0
1
2
3
4
−6
−4
−2
0
2
4

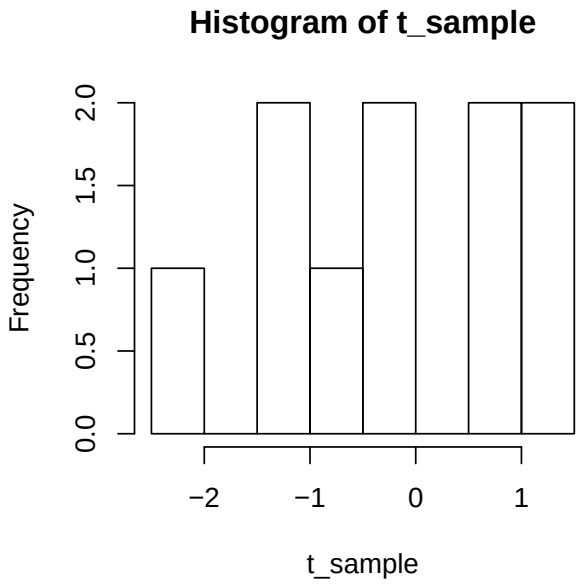

T Distribution
Sample Quantiles
Theoretical Quantiles
−4
−2
0
2
−1.5
−0.5
0.5
1.0
1.5

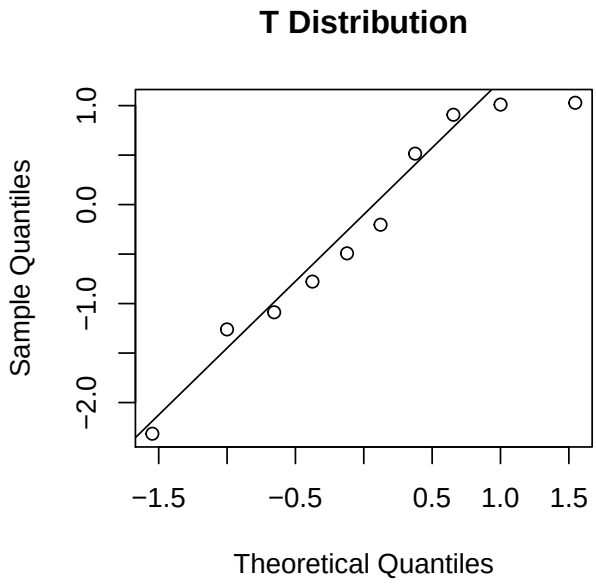

Histogram of t_sample
Frequency
t_sample
0.0
1.0
2.0
3.0
−4
−3
−2
−1
0
1
2

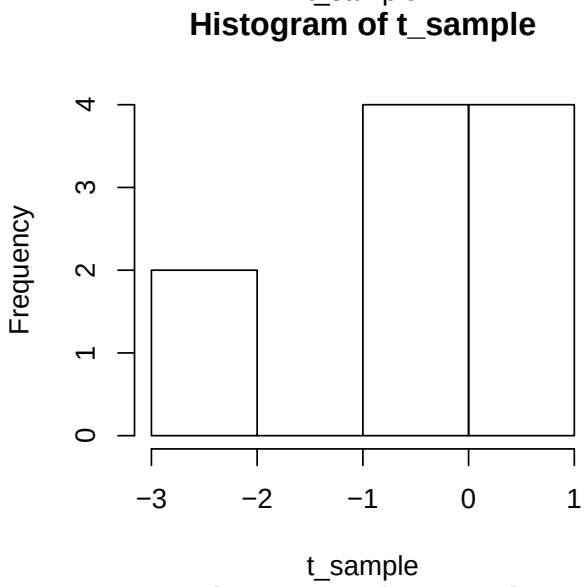

T Distribution
Sample Quantiles
Theoretical Quantiles
−3
−2
−1
0
1
−1.5
−0.5
0.5
1.0
1.5

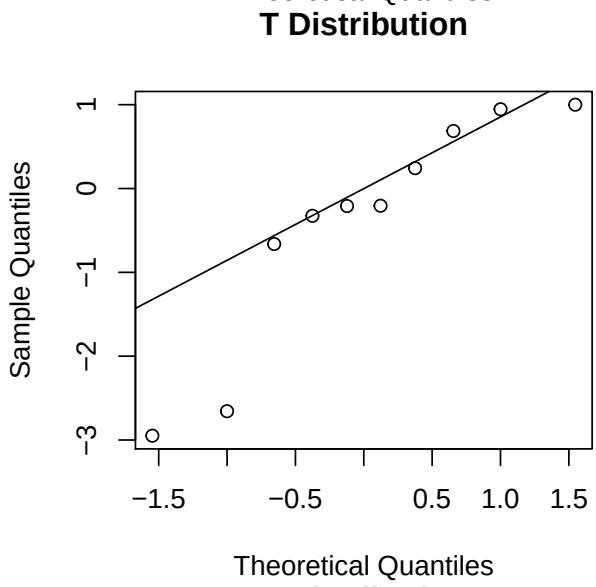

Histogram of t_sample
Frequency
t_sample
0
1
2
3
4
−3
−2
−1
0
1
2

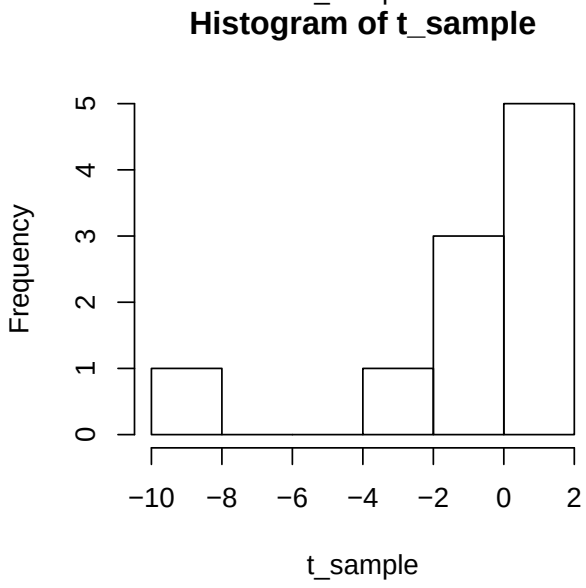

T Distribution
Sample Quantiles
Theoretical Quantiles
−2
−1
0
1
2
−1.5
−0.5
0.5
1.0
1.5

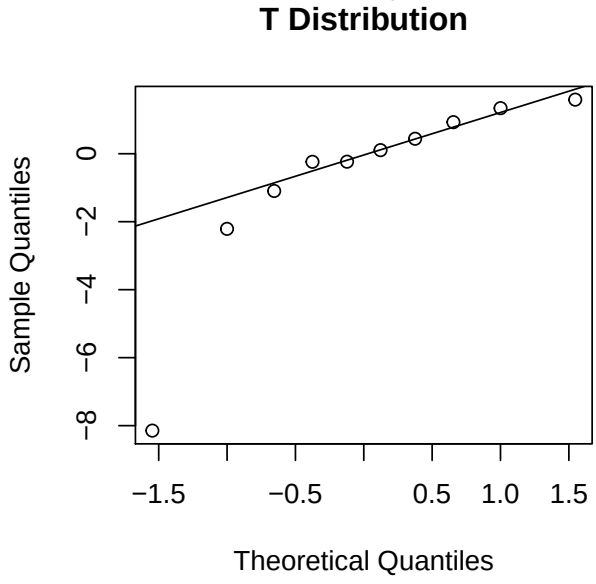

Histogram of t_sample
Frequency
0.0
0.5
1.0
1.5
2.0
−2
−1
0
1
t_sample


T Distribution
Sample Quantiles
−2.0
−1.0
0.0
1.0
−1.5
−0.5
0.5
1.0
1.5
Theoretical Quantiles


Histogram of t_sample
Frequency
0
1
2
3
4
−3
−2
−1
0
1
t_sample


T Distribution
Sample Quantiles
−3
−2
−1
0
1
−1.5
−0.5
0.5
1.0
1.5
Theoretical Quantiles


Histogram of t_sample
Frequency
0
1
2
3
4
5
−10
−8
−6
−4
−2
0
2
t_sample

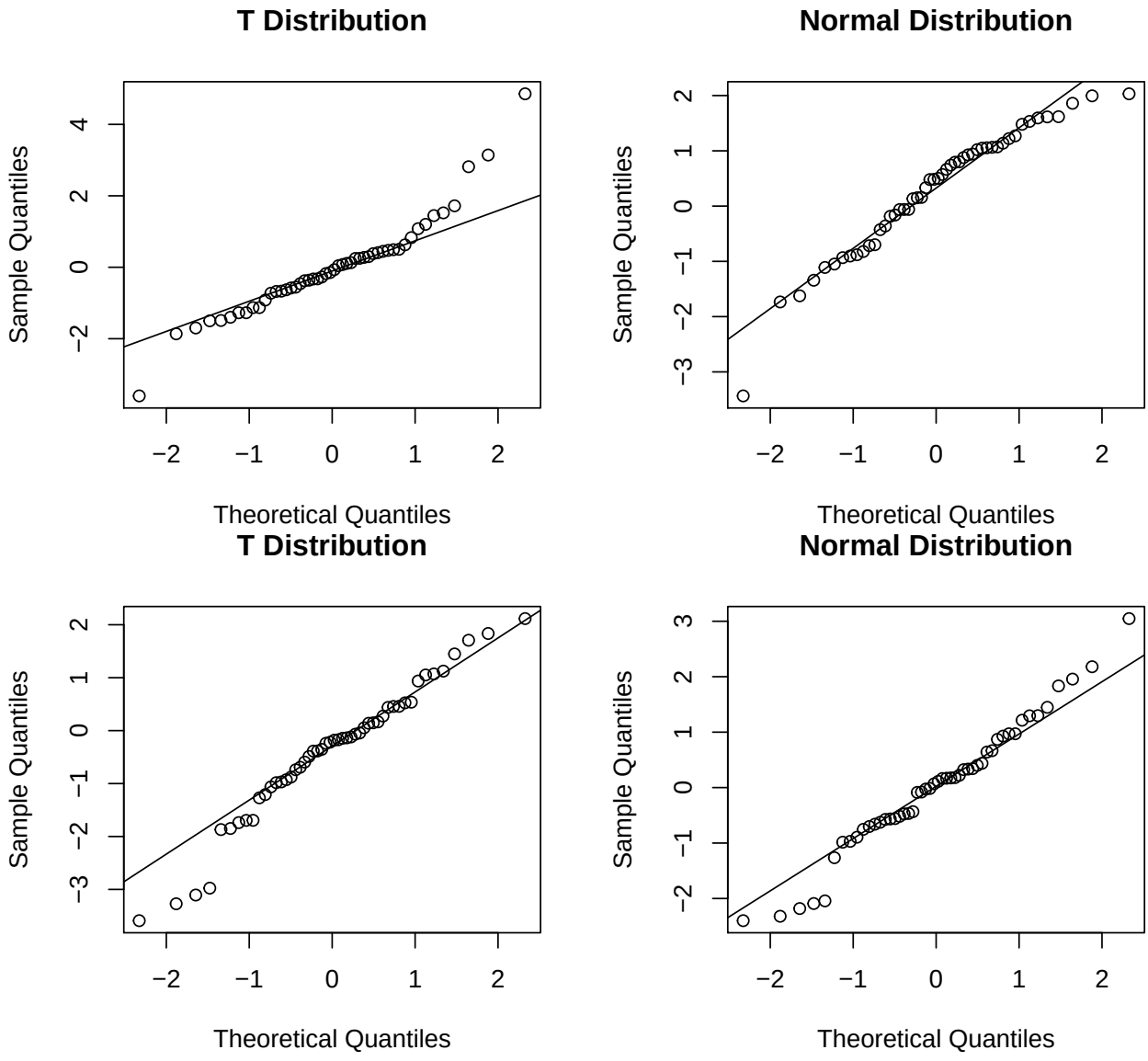

T Distribution
Sample Quantiles
−8
−6
−4
−2
0
−1.5
−0.5
0.5
1.0
1.5
Theoretical Quantiles

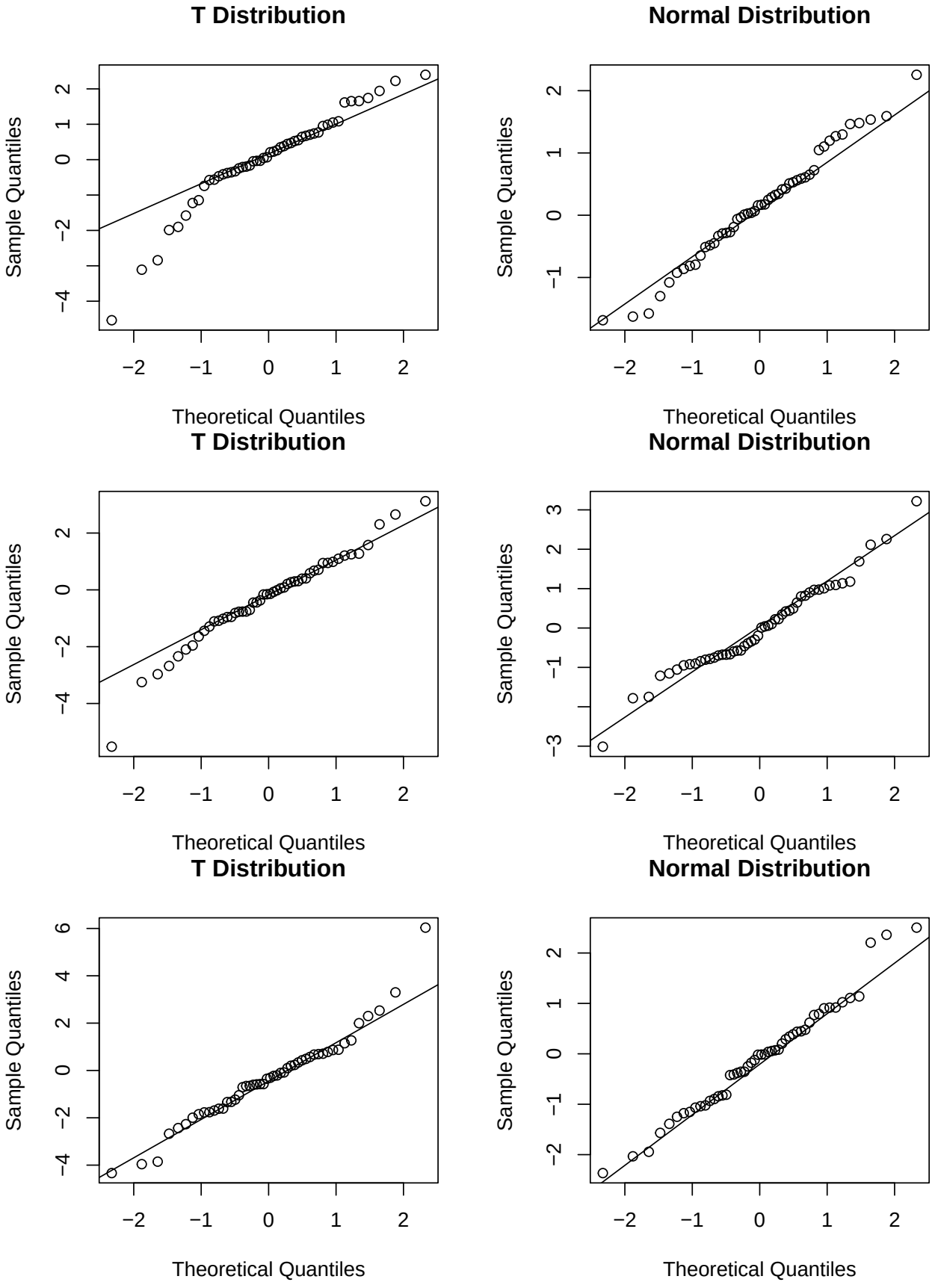
The plots are pretty similar to the ones obtained in question 1, i.e. they aren't distinguishable from a Normal distribution.

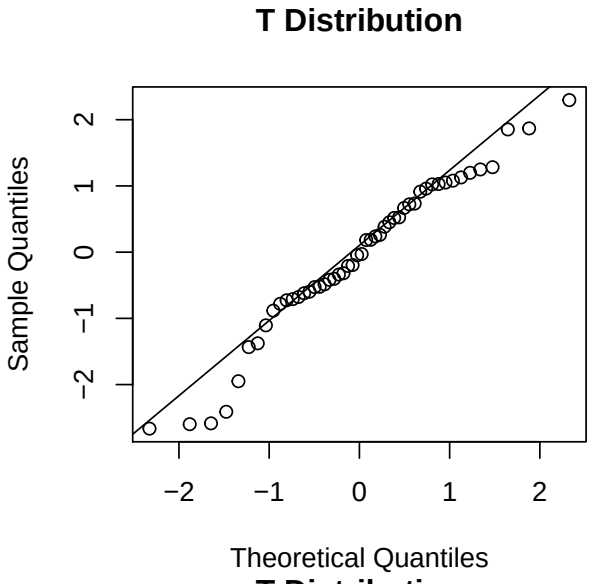
Problem 4

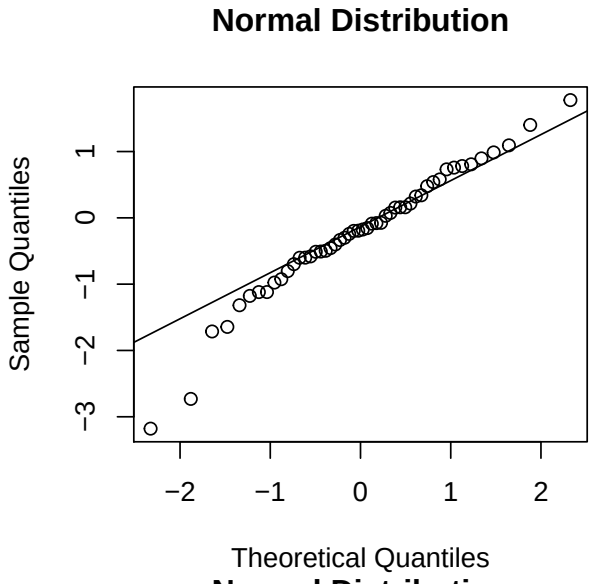
Repeat questions 1 and 3 with a sample size of 50. How do the results differ? How does the ability to detect non-normality depend on sample size?

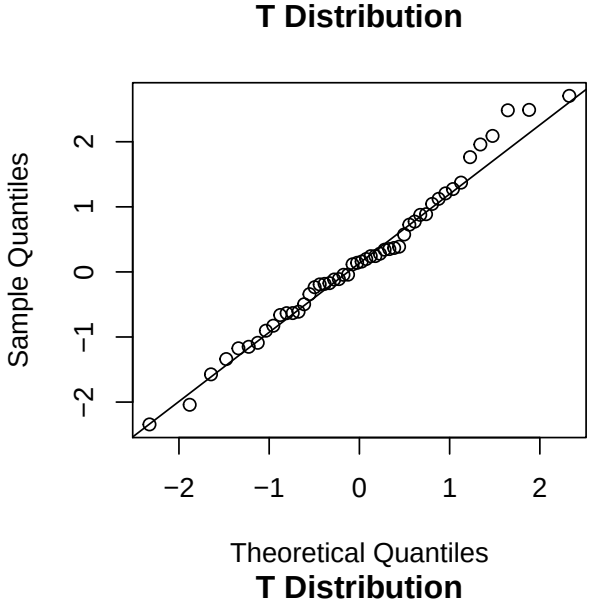
```
# leaving out histograms to save space
par(mfrow=c(1,2))
for (i in 1:20) {
  t_sample <- rt(n = 50, df=3)
  norm_sample <- rnorm(n = 50, mean = 0, sd = 1)

  qqn(t_sample, main="T Distribution")
  qqn(norm_sample, main="Normal Distribution")
}
```

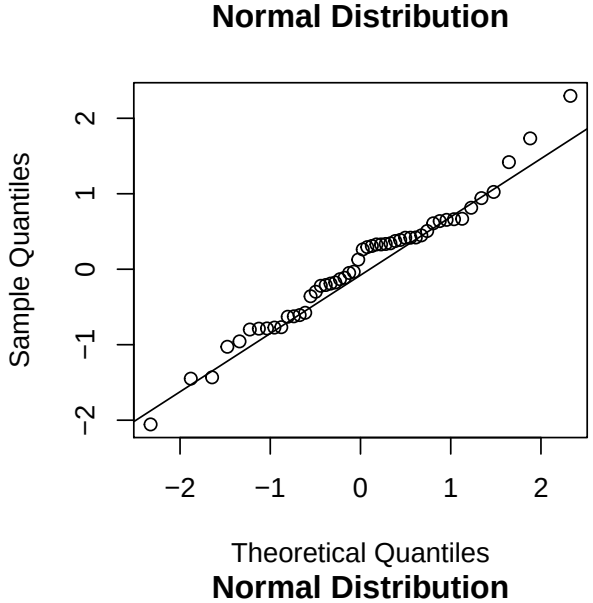


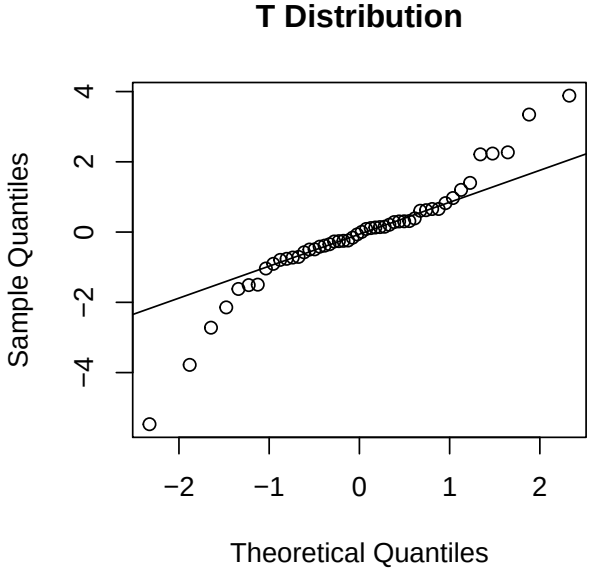

T Distribution
Normal Distribution
T Distribution
Normal Distribution
T Distribution
Normal Distribution
Sample Quantiles
Theoretical Quantiles

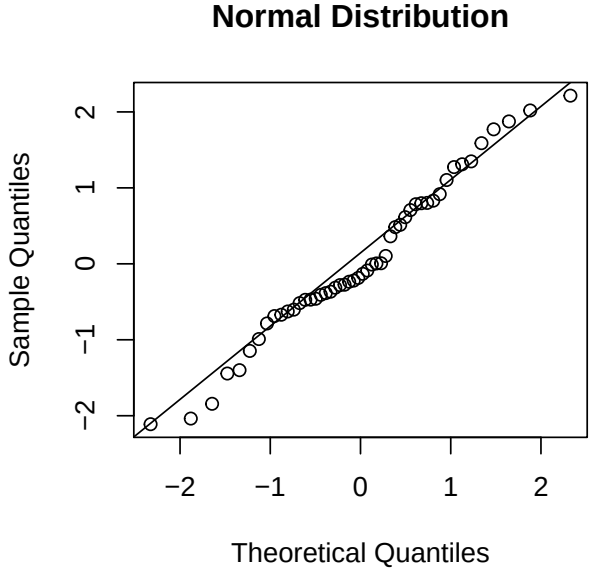

T Distribution
Sample Quantiles
Theoretical Quantiles
−2
−1
0
1
2

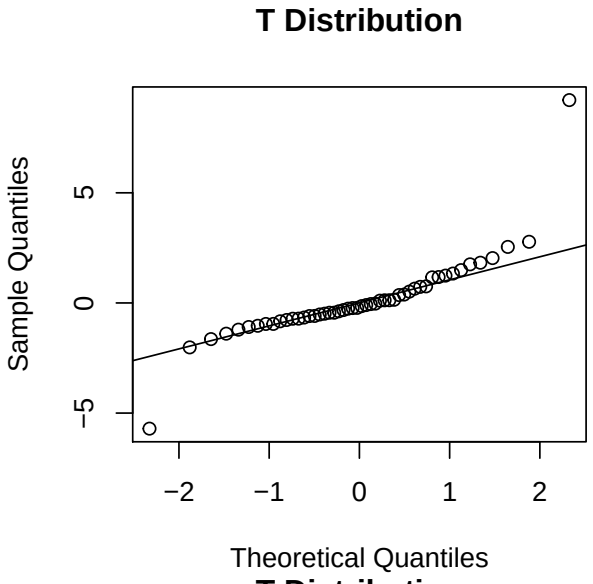

Normal Distribution
Sample Quantiles
Theoretical Quantiles
−3
−2
−1
0
1
2

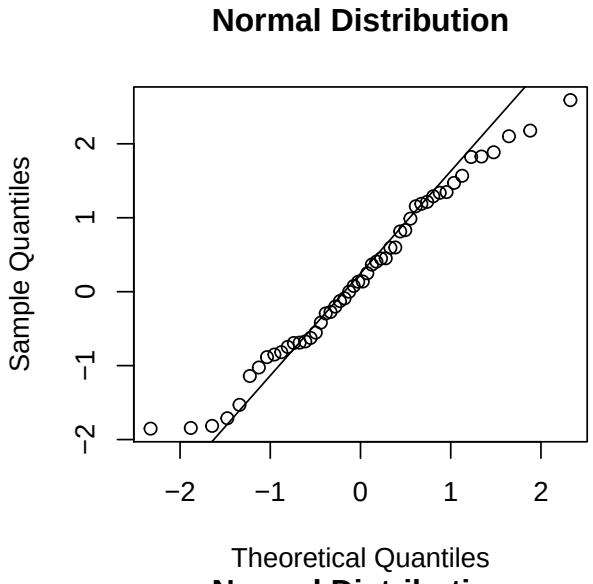

T Distribution
Sample Quantiles
Theoretical Quantiles
−2
−1
0
1
2

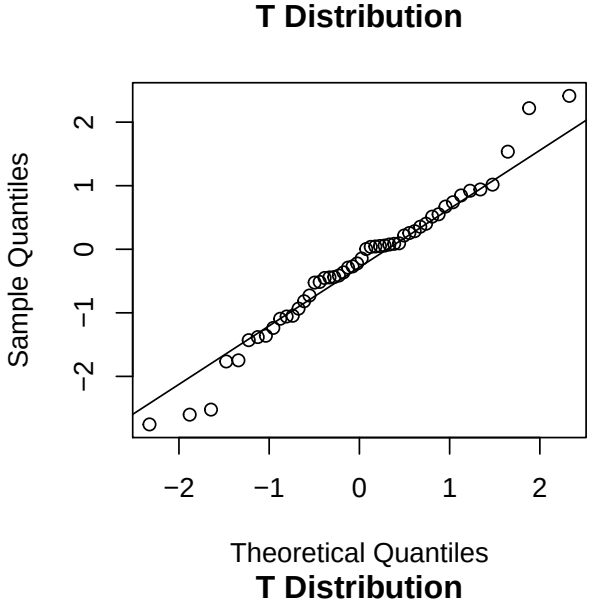

Normal Distribution
Sample Quantiles
Theoretical Quantiles
−2
−1
0
1
2

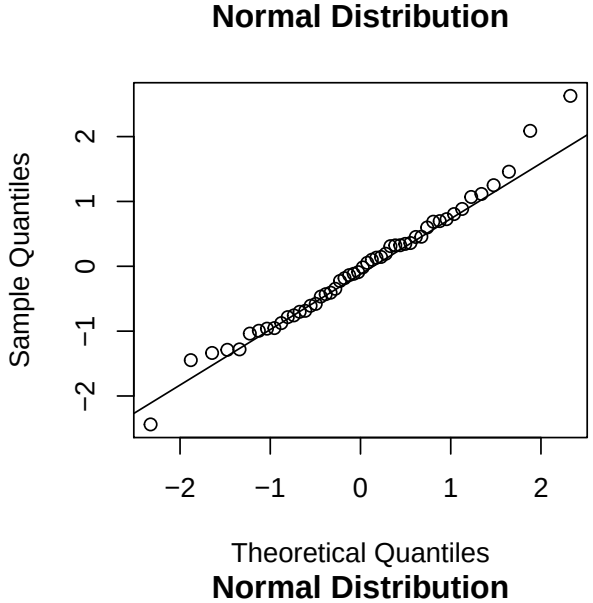

T Distribution
Sample Quantiles
Theoretical Quantiles
−4
−2
0
2
4

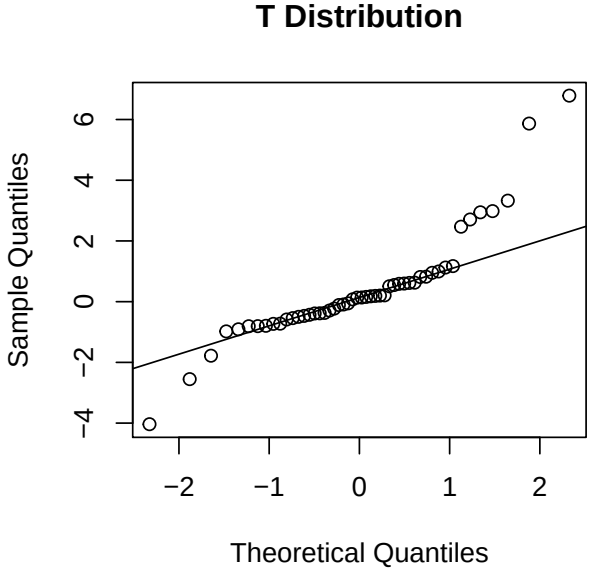

Normal Distribution
Sample Quantiles
Theoretical Quantiles
−2
−1
0
1
2

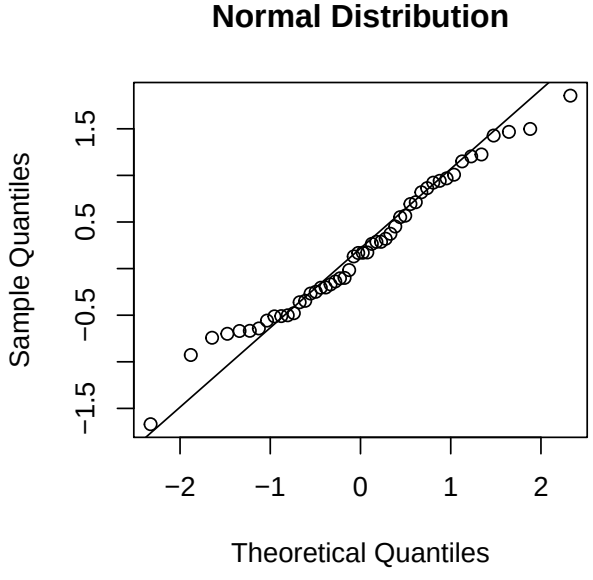

T Distribution
Sample Quantiles
Theoretical Quantiles
−5
0
5
−2
−1
0
1
2

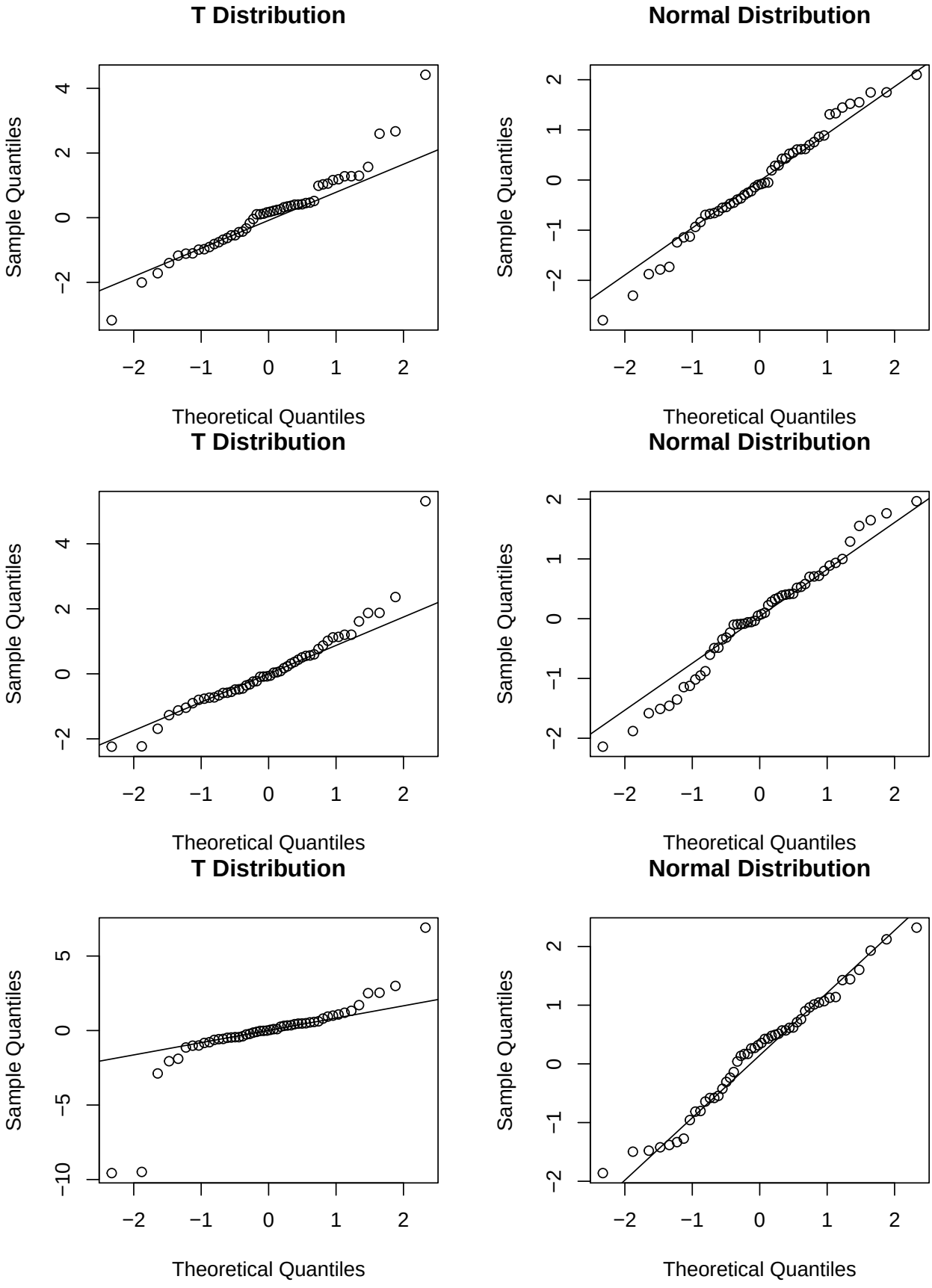

Normal Distribution
Sample Quantiles
Theoretical Quantiles
−2
−1
0
1
2
−2
−1
0
1
2

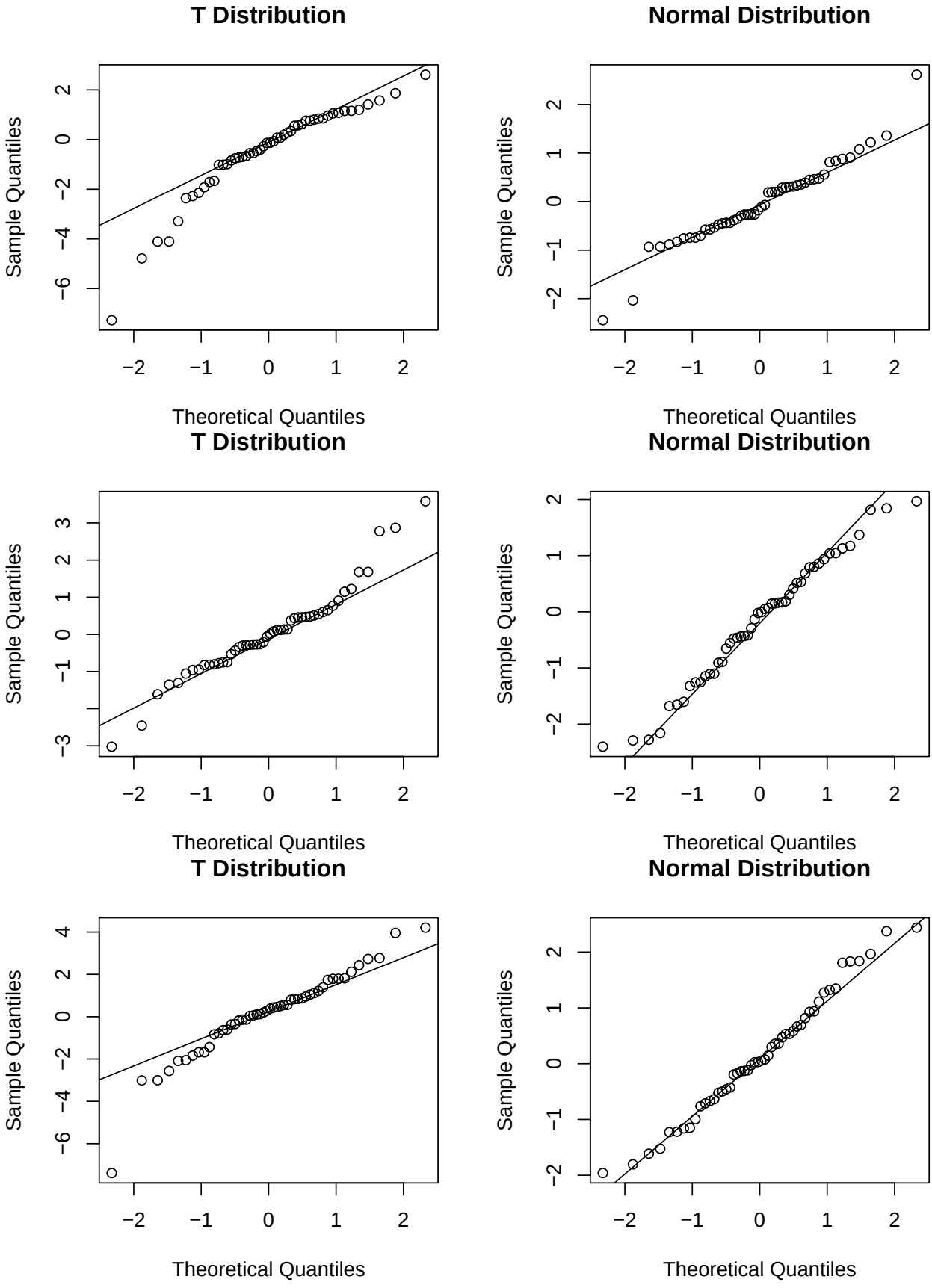

T Distribution
Sample Quantiles
Theoretical Quantiles
−2
−1
0
1
2
−2
−1
0
1
2

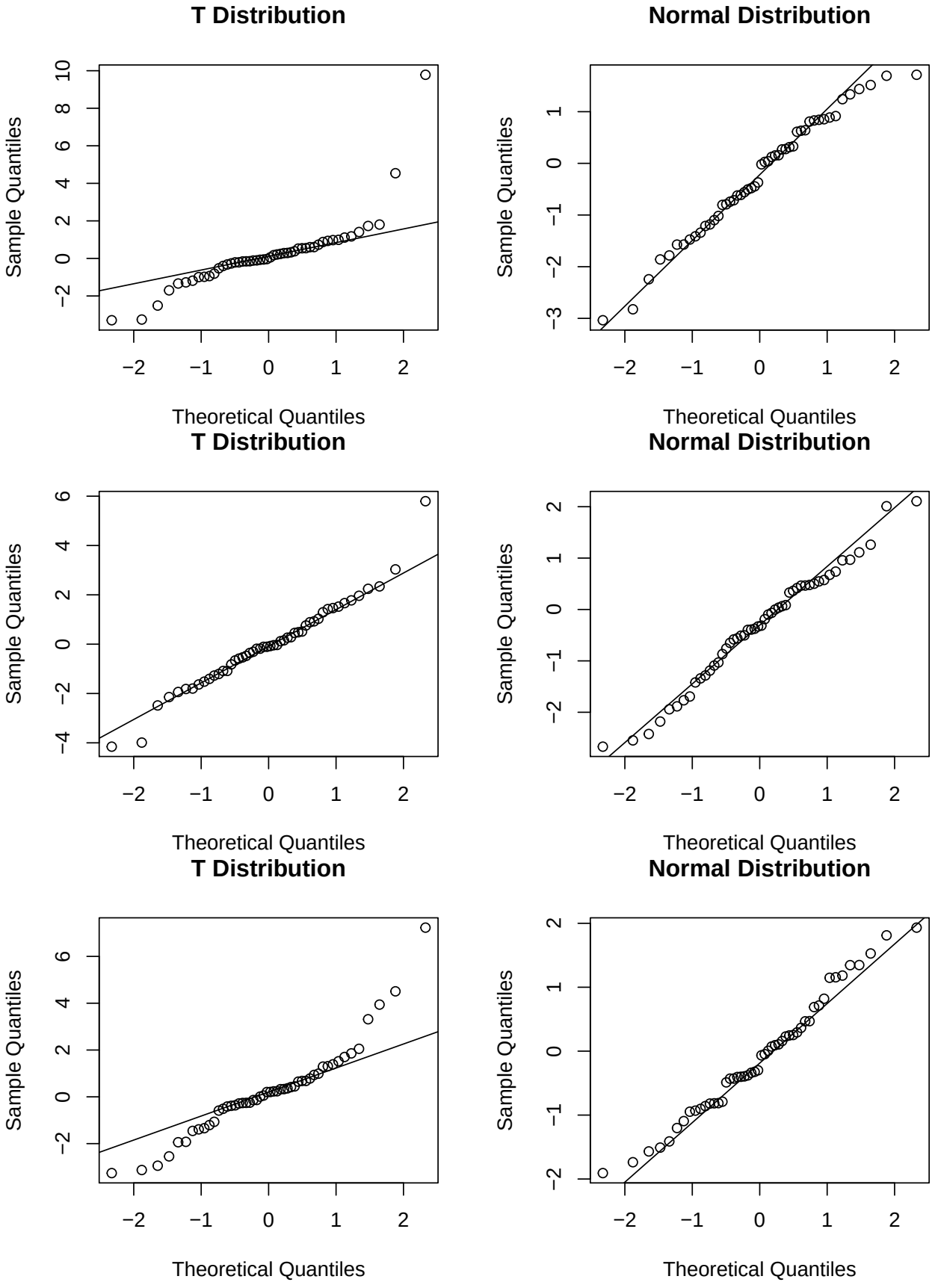

Normal Distribution
Sample Quantiles
Theoretical Quantiles
−2
−1
0
1
2
−2
−1
0
1
2


T Distribution
Sample Quantiles
Theoretical Quantiles
−4
−2
0
2
4
6
−2
−1
0
1
2


Normal Distribution
Sample Quantiles
Theoretical Quantiles
−1.5
−0.5
0.5
1.5
−2
−1
0
1
2


T Distribution
Normal Distribution
T Distribution
Normal Distribution
T Distribution
Normal Distribution
Sample Quantiles
Theoretical Quantiles


T Distribution
Normal Distribution
Sample Quantiles
Theoretical Quantiles
T Distribution
Normal Distribution
Sample Quantiles
Theoretical Quantiles
T Distribution
Normal Distribution
Sample Quantiles
Theoretical Quantiles


T Distribution
Sample Quantiles
Theoretical Quantiles
Normal Distribution
Sample Quantiles
Theoretical Quantiles
T Distribution
Sample Quantiles
Theoretical Quantiles
Normal Distribution
Sample Quantiles
Theoretical Quantiles
T Distribution
Sample Quantiles
Theoretical Quantiles
Normal Distribution
Sample Quantiles
Theoretical Quantiles

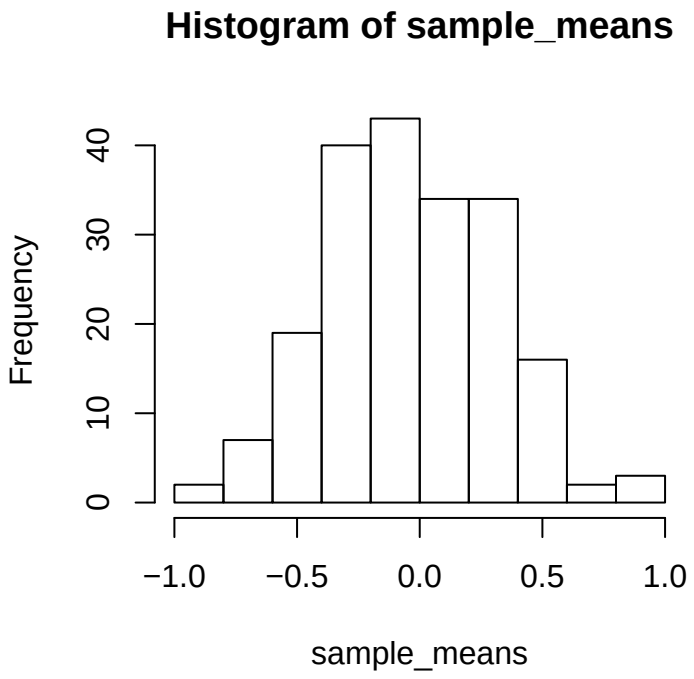
The results are not too different, but with a larger sample size, it is more often possible to distinguish the t-distribution from the normal distribution.

Standard Error of the Sample Mean

Problem 5

Perform a simulation study to assess the sampling distribution of the sample mean for sampling from the standard normal distribution with sample size 10.

```
simulation_study_sample_mean_norm <- function(n, mu=0, sigm=1) {
  # n is the sample size

  # Generate 200 independent random samples from the standard normal distribution
  # for each of the 200 samples calculate the sample mean
  sample_means <- replicate(200, mean(rnorm(n, mean = mu, sd = sigm)))

  # Make a histogram of the 200 sample means;
  hist(sample_means)

  # Calculate the mean, variance and SD of the sample means.
  experimental <- c(mean(sample_means), var(sample_means), sd(sample_means))
  theoretical <- c(0, sigm^2/n, sigm/sqrt(n))

  # How well do the mean, variance and SD of the sample means compare
  # to the values you would expect from the theoretical SE formula?
  res <- data.frame(experimental, theoretical)
  rownames(res) <- c("Mean", "Variance", "Standard Deviation/Error")
  res
}

simulation_study_sample_mean_norm(n = 10)
```

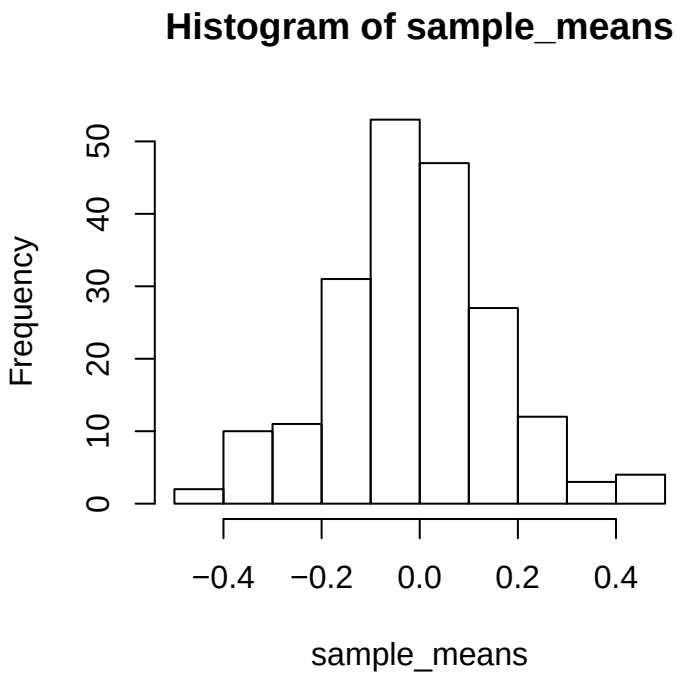


```
##                           experimental theoretical
## Mean                       -0.03493236   0.0000000
## Variance                    0.12130671   0.1000000
## Standard Deviation/Error    0.34829113   0.3162278
```

Problem 6

Repeat the simulation study using a sample of size 40.

```
simulation_study_sample_mean_norm(40)
```




```
##                           experimental theoretical
## Mean                       -0.01316545   0.0000000
## Variance                    0.02861348   0.0250000
## Standard Deviation/Error    0.16915521   0.1581139
```

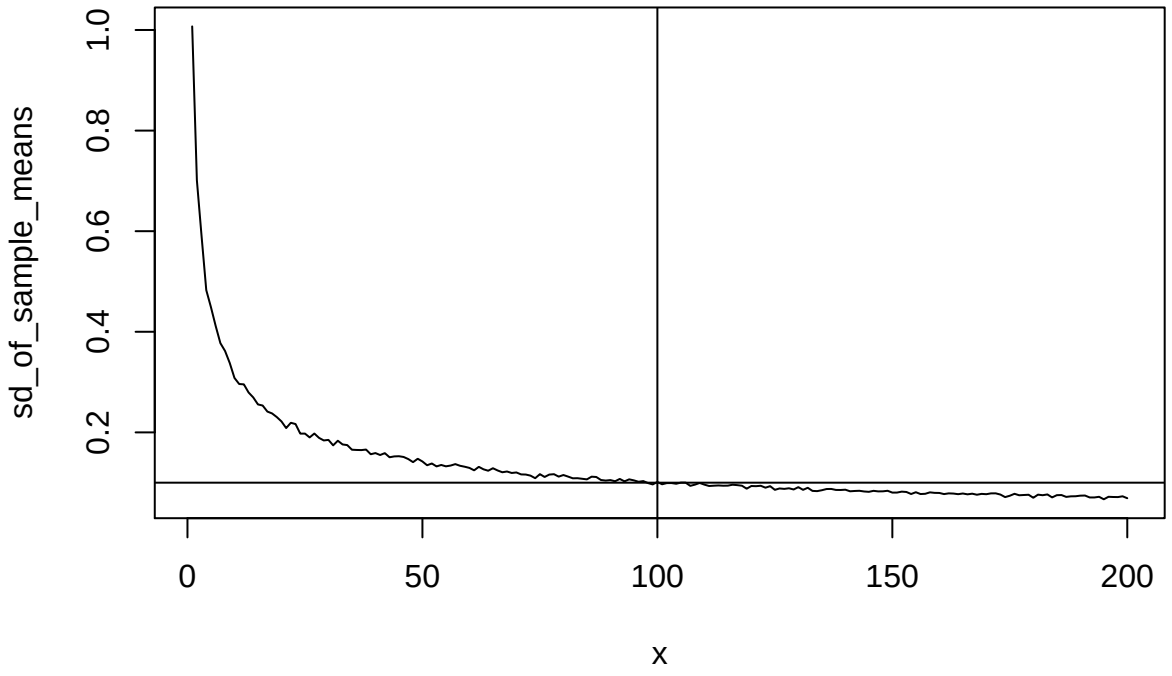
We can see that by increasing the sample size, the results are a closer match to the theoretical values.

Problem 7

Using the SE formula, what is the minimum sample size required in order for the sample mean to have SD of 0.1 or less when sampling from the standard normal distribution? Perform a simulation study to verify your answer.

```
# for 1/sqrt(n) = 0.1, n = 100

sd_of_sample_mean_norm_from_n_samples <- function(n) {
  sd(replicate(1000, mean(rnorm(n))))
}

x <- seq(1, 200)
sd_of_sample_means <- lapply(X = x, FUN = sd_of_sample_mean_norm_from_n_samples)
plot(x, sd_of_sample_means, type='l')
```

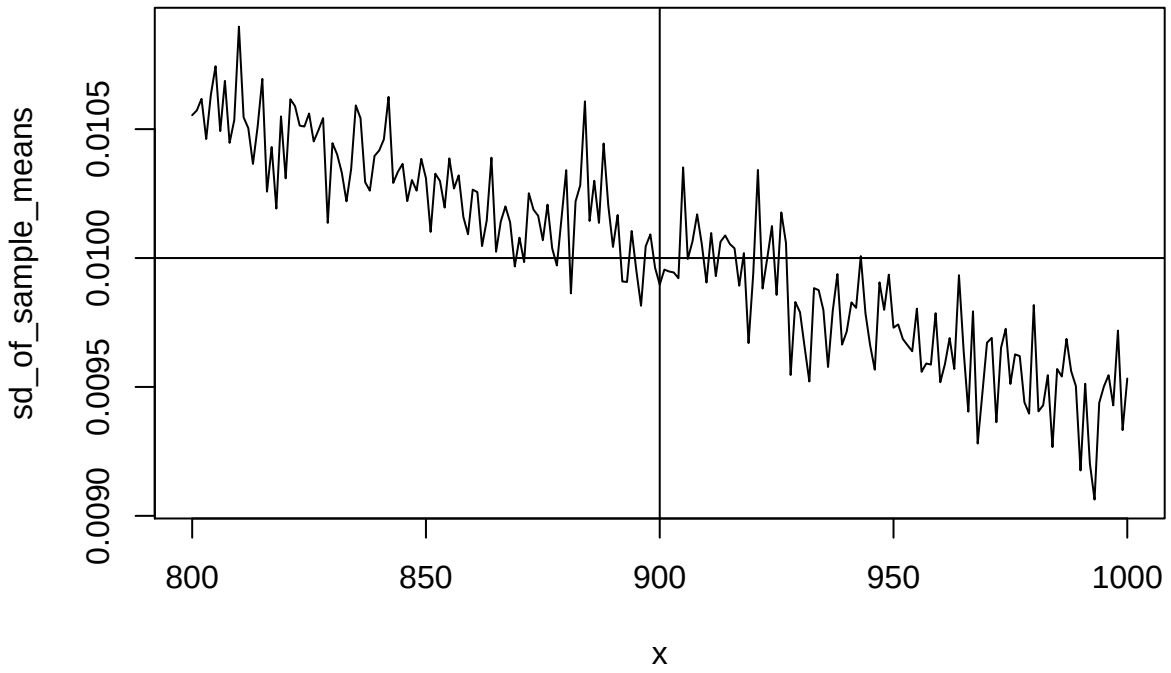
```
abline(v=100)
abline(h=0.1)
```



Problem 8

Consider sampling from a Bernoulli distribution with success probability p=0.1. The sample means are now sample proportions. What is the minimum sample size required to have a sample proportion with SD of 0.01 or less? Verify using a simulation study.

```
# Minimum sample size needed for SD = 0.01
# SD = sqrt(p(1-p)/n)
# 0.0001 = p(1-p)/n
# 0.0001 = 0.1(0.9)/n
# n = 0.1 * 0.9 / 0.0001 = 900

rbern <- function(n, p=0.1) {
  rbinom(n = n, size = 1, prob = p)
}

sd_of_sample_mean_from_bern_n_samples <- function(n) {
  sd(replicate(2000, mean(rbern(n))))
}

x <- seq(800, 1000)
sd_of_sample_means <- lapply(x, sd_of_sample_mean_from_bern_n_samples)
plot(x, sd_of_sample_means, type='l')
abline(v=900)
abline(h=0.01)
```



The Bernoulli experiment has much more variability, but our estimate of 900 seems about right.

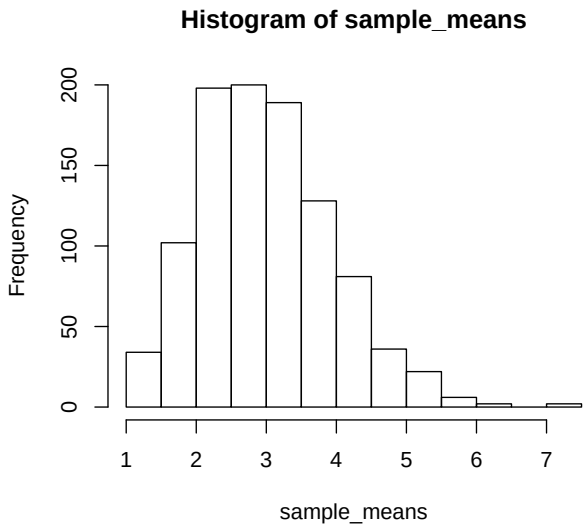
Central Limit Theorem

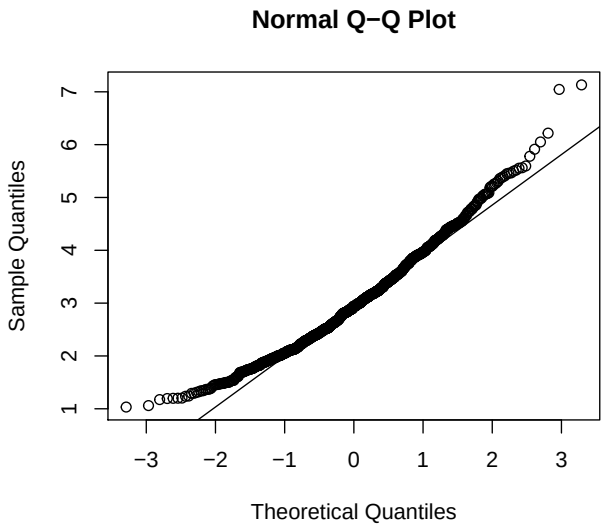
Problem 9

Perform a simulation study to assess the sampling distribution of the sample mean for sampling from the exponential distribution with **mean equal to 3**, with **sample size 10**. Does the distribution of the sample mean appear to be closely approximated by a normal distribution? Use a histogram and q-q plot.

```
simulation_study_expo <- function(n) {
  lambd <- 3
  sample_means <- replicate(1000, mean(rexp(n = n, rate = 1/lambd)))

  par(mfrow=c(1, 2))

  hist(sample_means)

  qqn(sample_means)

  # Calculate the mean, variance and SD of the sample means.
  experimental <- c(mean(sample_means), var(sample_means), sd(sample_means))
  theoretical <- c(lambd, lambd^2 / n, lambd / sqrt(n))
  abs_error <- abs(theoretical - experimental)

  # How well do the mean, variance and SD of the sample means compare
  # to the values you would expect from the theoretical SE formula?
  res <- data.frame(experimental, theoretical, abs_error)
  rownames(res) <- c("Mean", "Variance", "Standard Deviation/Error")
  res
}
```

```
res <- simulation_study_expo(10)
```





It looks not too much different than a Normal distribution, but is somewhat skewed.

Problem 10

How well do the mean, variance and SD of the sample means compare to the values you would expect from theory?

```
res
```

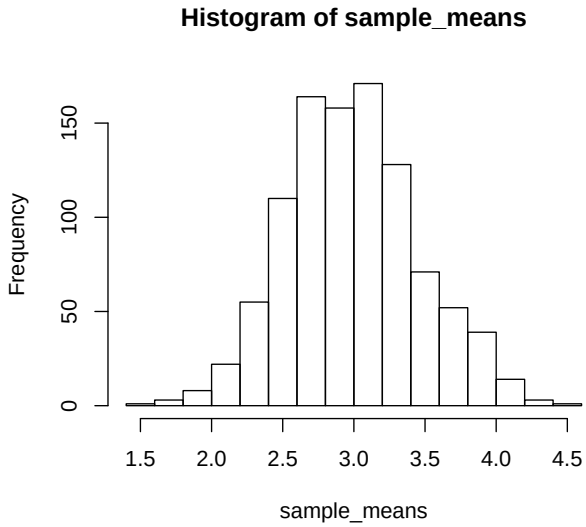
```
##                           experimental theoretical   abs_error
## Mean                         3.0109371   3.0000000 0.010937052
## Variance                     0.9058002   0.9000000 0.005800157
## Standard Deviation/Error     0.9517353   0.9486833 0.003052042
```

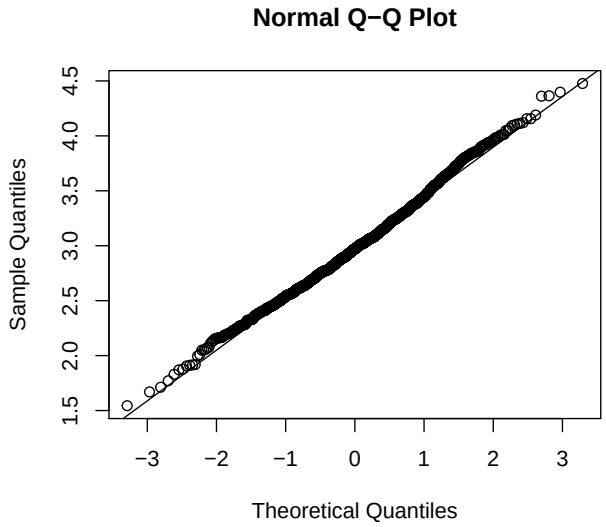
The experimental values are pretty close to the theoretical values.

Problem 11

Repeat questions 9 and 10 using a sample of size 40.

```
res <- simulation_study_expo(40)
```






```
res
```

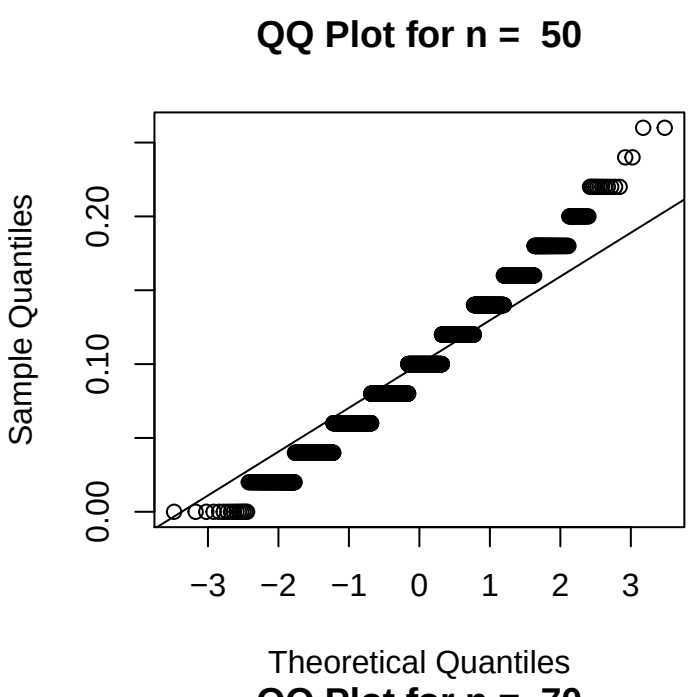
```
##                           experimental theoretical  abs_error
## Mean                         2.9892377   3.0000000 0.01076230
## Variance                     0.2144486   0.2250000 0.01055137
## Standard Deviation/Error     0.4630860   0.4743416 0.01125567
```

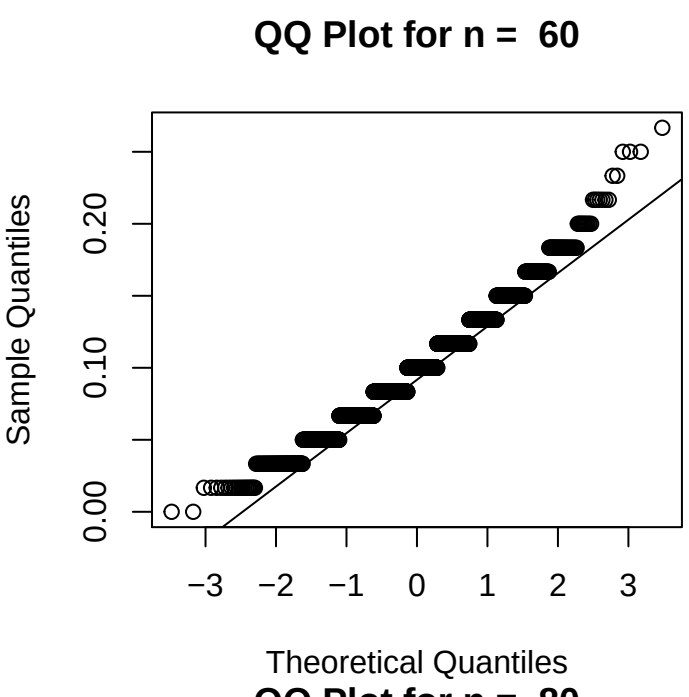
The distribution looks closer to the Normal than it did before, as we'd expect.

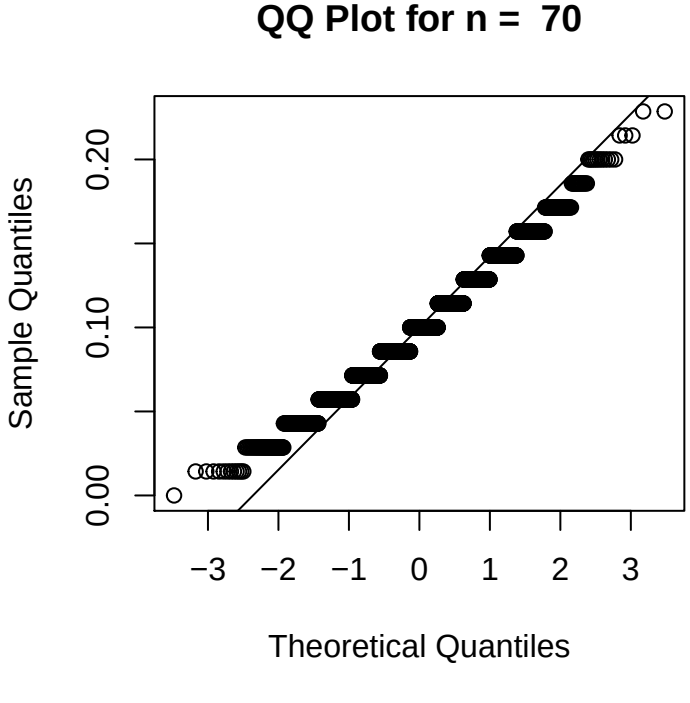
Problem 12

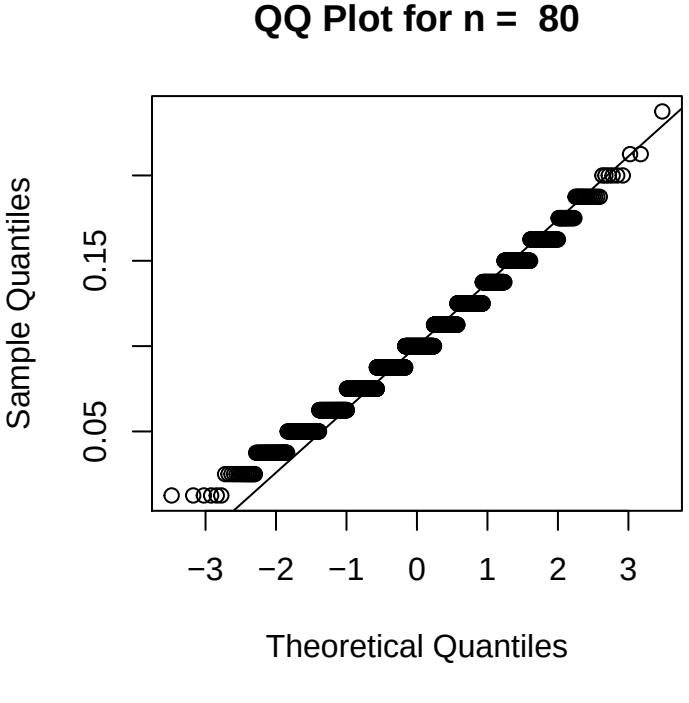
Now consider sampling from the Bernoulli distribution with success probability p=0.1. What is the minimum sample size required in order that the sample proportion has an approximately normal distribution?

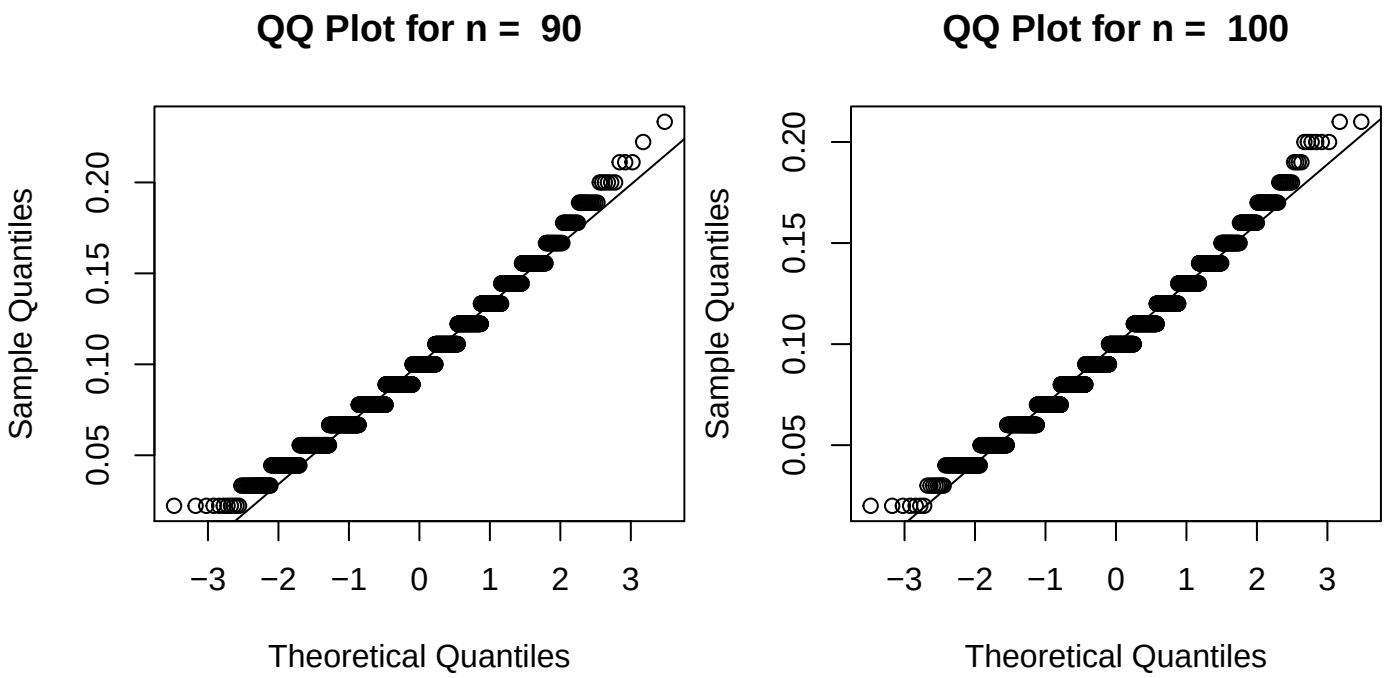
```
sample_means_for_bern <- function(n) {
  replicate(2000, mean(rbern(n, p = 0.1)))
}

# The values in the for loop were repeatedly increased
# to get plots that looked progressively more Normal
# This was subjective, and you may have a different standard for deciding
# how close the QQPlot is to being Normal.
for (i in 10*(5:10)) {
  x <- sample_means_for_bern(i)
  qqn(x, main=paste("QQ Plot for n = ", i))
}
```


QQ Plot for n = 50
Sample Quantiles
Theoretical Quantiles
0.00
0.10
0.20
−3
−2
−1
0
1
2
3


QQ Plot for n = 60
Sample Quantiles
Theoretical Quantiles
0.00
0.10
0.20
−3
−2
−1
0
1
2
3


QQ Plot for n = 70
Sample Quantiles
Theoretical Quantiles
0.00
0.10
0.20
−3
−2
−1
0
1
2
3


QQ Plot for n = 80
Sample Quantiles
Theoretical Quantiles
0.05
0.15
−3
−2
−1
0
1
2
3

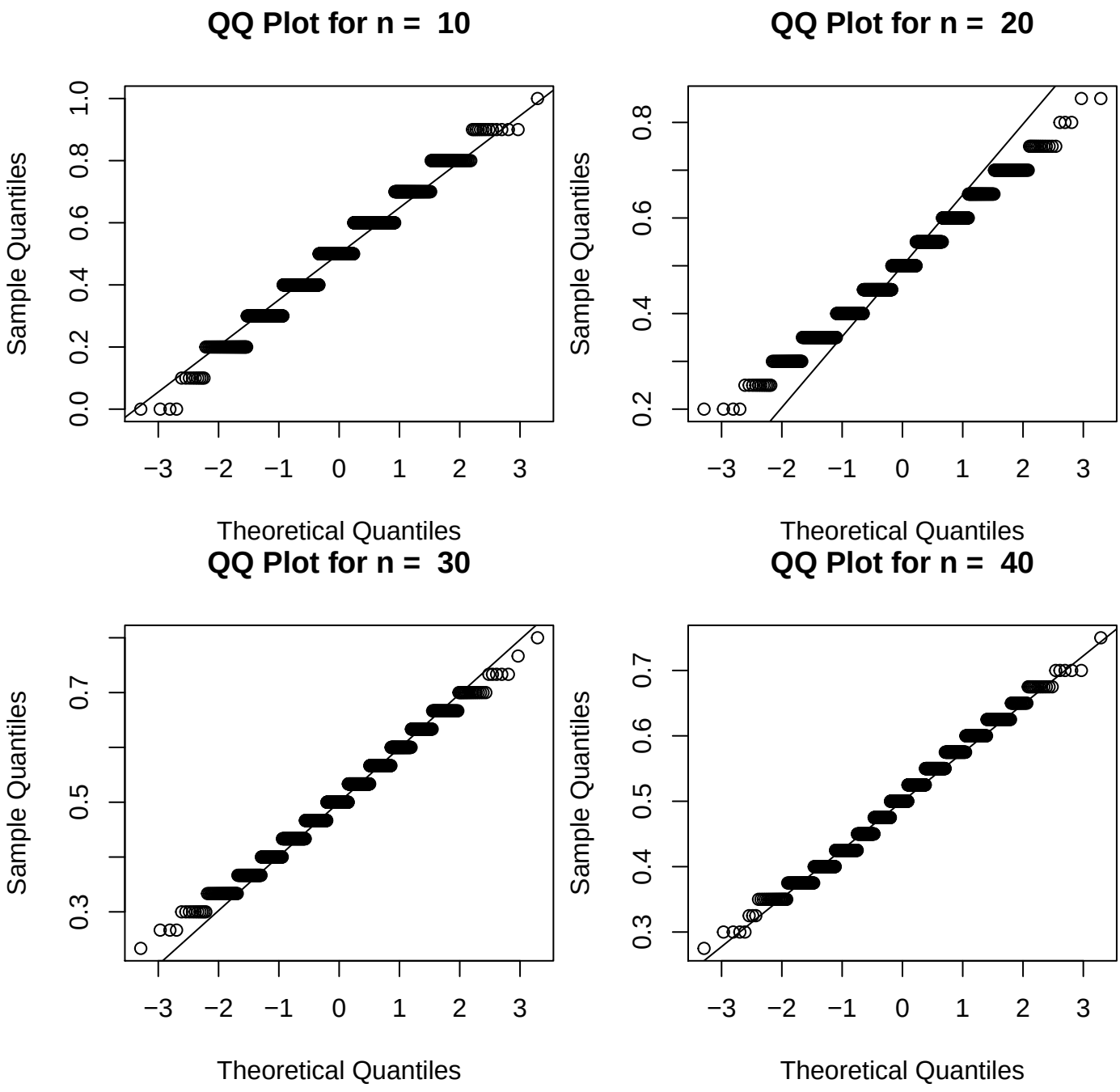


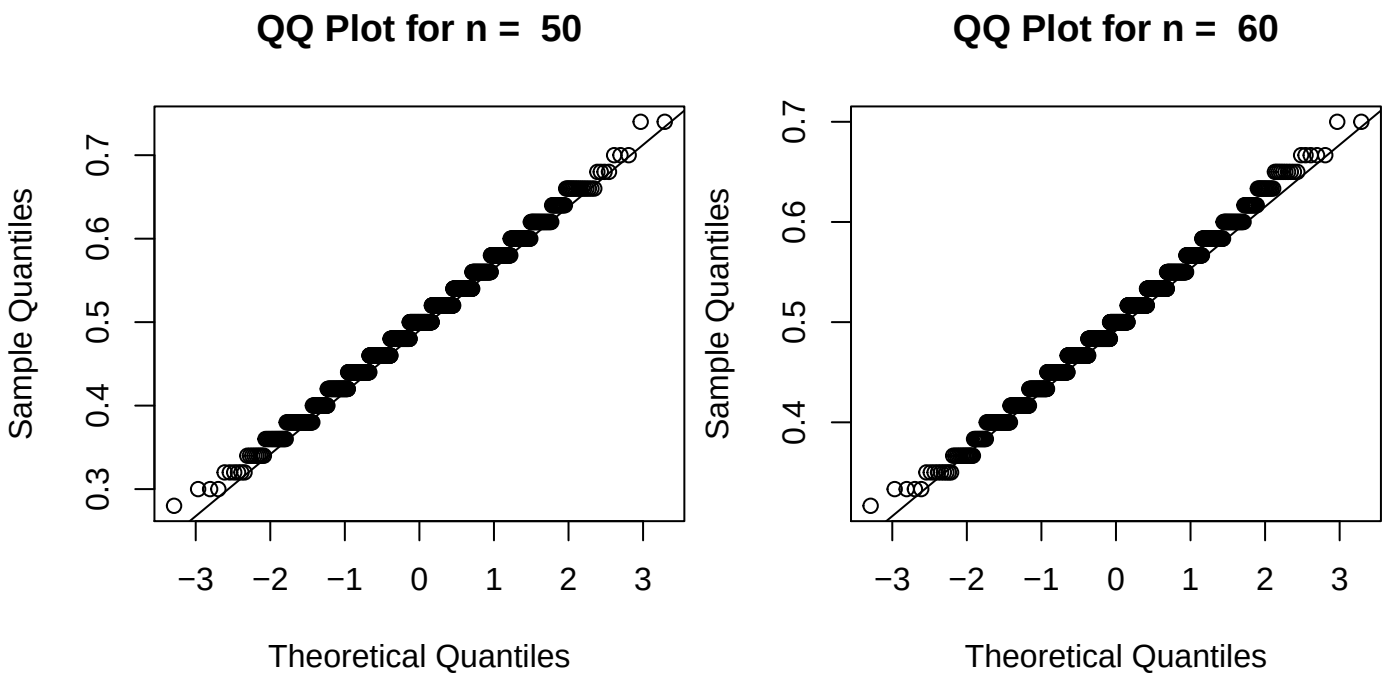
Approximately 100 is sufficient.

Problem 13

How would the answer to question 12 change if the success probability was p=0.5 instead of 0.1?

```
sample_means_for_bern <- function(n) {
  replicate(1000, mean(rbern(n = n, p = 0.5)))
}

# The values in the for loop were repeatedly increased
# to get plots that looked progressively more Normal
# This was subjective, and you may have a different standard for deciding
# how close the QQPlot is to being Normal.
for (i in 10*(1:6)) {
  x <- sample_means_for_bern(i)
  qqn(x, main=paste("QQ Plot for n = ", i))
}
```


QQ Plot for n = 10
Sample Quantiles
Theoretical Quantiles
QQ Plot for n = 20
Sample Quantiles
Theoretical Quantiles
QQ Plot for n = 30
Sample Quantiles
Theoretical Quantiles
QQ Plot for n = 40
Sample Quantiles
Theoretical Quantiles



A smaller sample size is needed for p=0.5. In this case about 40 is sufficient.

Confidence Intervals

Questions 14-16 will explore the calculation and properties of confidence intervals. We will use the built-in R data set 'morley', which contains measurements of the speed of light (in km/sec, with 299000 subtracted) performed in 5 experiments with 20 runs each. We will only use the 20 measurements from Experiment 1 for this assignment. We will consider these 20 measurements to be a random sample from a population with population mean equal to the true value of 792 (the true speed of light is 299,792 km/sec).

Problem 14

Calculate a 95% confidence interval for the speed of light using the 20 measurements from Experiment 1.

```
measurements <- morley[morley$Expt == 1, "Speed"]
true_mean <- 792
n <- length(measurements)

se <- sd(measurements)/sqrt(n)
mu_est <- mean(measurements)

upper <- mu_est + 1.96*se
lower <- mu_est - 1.96*se

c(lower, upper)
```

```
## [1] 863.0141 954.9859
```

Problem 15

Is the true value for the speed of light in the 95% confidence interval? What does this say about the bias or lack of bias of the experimental measurement technique.

The true value for the speed of light does not fall in the confidence interval. It is actually very far from the confidence interval, so the experimental measurement technique is probably biased.

Problem 16

Perform a simulation study to assess the performance of the confidence interval for the speed of light based on 20 measurements. For the simulations, treat the Experiment 1 sample as if it were the population and sample with replacement from this "population". Use a sample size of 20, i.e., the same as the actual sample size. Generate 1000 samples of size 20 from the "population" and calculate the confidence interval for each sample. What percentage of the confidence intervals contain the true speed of light? What percentage contains the "population" mean? Explain.

```
pop_mean <- mean(measurements)
true_mean <- 792

get_conf_interval <- function(sample_measurements) {
  se <- sd(sample_measurements)/sqrt(length(sample_measurements))
  mu_est <- mean(sample_measurements)

  upper <- mu_est + 1.96*se
  lower <- mu_est - 1.96*se

  list(lower=lower, upper=upper)
}

interval_contains_true_mean <- function(interval) {
  true_mean >= interval$lower && true_mean <= interval$upper
}

interval_contains_population_mean <- function(interval) {
  pop_mean >= interval$lower && pop_mean <= interval$upper
}

samples <- replicate(1000, sample(measurements, size = 20, replace = T))
intervals <- apply(samples, FUN=get_conf_interval, MARGIN=2)

# sapply will return a vector of booleans,
# specifying whether each interval contains true mean or not.
# So percent of intervals containing true mean is simply the proportion of 1's (TRUE = 1 in R)
percent_intervals_with_true_mean <- mean(sapply(intervals, FUN=interval_contains_true_mean)) * 100
cat("% Intervals Containing True Mean: ", percent_intervals_with_true_mean, "%\n")
```

```
## % Intervals Containing True Mean:  0.4 %
```

```
percent_intervals_with_pop_mean <- mean(sapply(intervals, FUN=interval_contains_population_mean)) * 100
cat("% Intervals Containing Population Mean: ", percent_intervals_with_pop_mean, "%")
```

```
## % Intervals Containing Population Mean:  91.4 %
```

The percentage of confidence interval (CI) covering the true mean is 0.4%, much lower than the nomial 95%. This is reasonable because the experiment is biased, and the CI is not a valid CI for the true mean. The

percentage of CI covering the "population" mean is 91.4%. It is lower than the nominal 95%, but not too bad. One reason for undercoverage is that the sample size is only 20, which might be too small for the central limit theorem, and we used the normal approximation. If we use a t-interval, the coverage could be improved (we will see this later in class).

Visualizing a random selection of the confidence intervals:

```
get_lower <- function(inter) inter$lower
get_upper <- function(inter) inter$upper

n <- 100

lowers <- sapply(intervals[1:n], FUN = get_lower)
uppers <- sapply(intervals[1:n], FUN = get_upper)

matplot(rbind(1:n,1:n),
        t(cbind(lowers, uppers)),
        type="l",lty=1,lwd=1,col=4,xlab="Sample #",ylab="Mean")

abline(h=true_mean, lty=2,col=2, lwd=2)
abline(h=pop_mean, lty=2, col=3, lwd=2)

legend("topright", legend = c("True Mean", "Population Mean"), col = c(2, 3), lty=2)
```