Problem Set 7

Michael Ghattas

Introduction

Collaboration

(1 point)

Other students who I worked with on this assignment (if any) : None.

Notes

These questions were rendered in R markdown through RStudio (https://www.rstudio.com/wp-content/uploads/2015/02/rmarkdown-cheatsheet.pdf, http://rmarkdown.rstudio.com).

Please complete the following tasks regarding the data in R. Please generate a solution document in R markdown and upload the .Rmd document and a rendered .pdf document. Your solution document should have your answers to the questions and should display the requested plots.

Question 1

The precipitation data in "precip.txt" are precipitation values for Boulder, CO from https://www.esrl.noaa.gov/psd/boulder/Boulder.mm.precip.html downloaded 2/17/2022.

Precipitation includes rain, snow, and hail. Snow/ice water amounts are either directly measured or a ratio of 1/10 applied for inches of snow to water equivalent.

The purpose of this analysis is to assess the null hypothesis that the total annual rainfalls in the early portion and the total annual rainfalls in the recent portion of the data are each independent identically distributed (i.i.d.) samples from Normally distributed populations with equal means, $Normal(\mu, \sigma^2_{early})$ and $Normal(\mu, \sigma^2_{recent})$.

Unlike in a class setting, in practice, data formatting is often a major component of a data analysis project. Some basic formatting of the data in "precip.txt" is included below for reference.

The symbol "Tr" represents a trace amount of precipitation. Observations marked by a "*" were made at a non-standard site. Some light-duty data formatting appears below that sets "Tr" values to 0 and drops years that include an observation made at a non-standard site.

The code provided below reads in the precipitation data. The values are tab-separated. Most columns are assigned the string class, "chr".

```
dat <- read.table("precip_2021.txt", sep = "\t",header = TRUE)
```

The following replaces all column names with lower case versions. For example, "TOTAL" becomes "total". The command "names(dat)" is used to verify that the replacement has succeeded.

```
# Change all characters in the variable names to lower case.
names(dat) <- str_to_lower(names(dat))
names(dat)
```

```
##  [1] "year"       "jan"        "feb"        "mar"        "apr"
##  [6] "may"        "jun"        "jul"        "aug"        "sep"
## [11] "oct"        "nov"        "dec"        "year.total"
```

Replace all occurrences of "Tr" with 0. Verify that this was successful.

```
# Replace "Tr".
dat <- mutate_all(dat, str_replace, "Tr", "0")
# Count all occurrences of "Tr".
sum(str_detect(unlist(dat), "Tr"))
```

```
## [1] 0
```

Drop all rows that include an asterisk indicating an observation at a non-standard location. The method for this is to write a function that takes a vector of strings as its argument and returns "TRUE" if none of the strings contains an asterisk, "FALSE" otherwise. Then apply this function to each row of the data to generate a Boolean vector. Finally,using this vector, reduce the data set to only those rows without asterisks.

Note that the asterisk has a special meaning in string manipulation so the backslashes are used to look for a literal asterisk.

https://cran.r-project.org/web/packages/stringr/vignettes/regular-expressions.html

```
# function to return TRUE if a string vector x contains no entries with an "*".
no_stars <- function(x){
  sum(str_detect(x, "\\*")) == 0
}
# Count asterisks in the data.
sum(str_detect(unlist(dat), "\\*"))
```

```
## [1] 8
```

```
# Identify the rows in the data with at least 1 "*".
all.standard <- apply(dat, 1, no_stars)
dat.trim <- dat[all.standard, ]
# Count asterisks in the trimmed data.
sum(str_detect(unlist(dat.trim), "\\*"))
```

```
## [1] 0
```

Set all precipitation columns in "dat.trim" to be of "numeric" class using the "as.numeric" function. Make the "year" column to be of class "integer". Verify the success of this by running "sapply(dat, class)" and displaying the results.

Verify that converting the strings to numeric values didn't produce any "NA"s.

```
dat.trim <- mutate_all(dat.trim, as.numeric)
dat.trim[, 1] <- as.integer(dat.trim[, 1])
sapply(dat.trim, class)
```

```
##       year        jan        feb        mar        apr        may        jun
##  "integer"  "numeric"  "numeric"  "numeric"  "numeric"  "numeric"  "numeric"
##        jul        aug        sep        oct        nov        dec year.total
##  "numeric"  "numeric"  "numeric"  "numeric"  "numeric"  "numeric"  "numeric"
```

```
sum(is.na(dat))
```

```
## [1] 0
```

```
which(is.na(dat), arr.ind = TRUE)
```

```
##      row col
```

Identify the omitted years in "dat.trim".

```
setdiff(min(dat.trim$year):max(dat.trim$year),dat.trim$year)
```

```
## [1] 1989 1990
```

Question 1.1

(5 points)

Since values in successive years may be related by persistent weather patterns, the data are thinned to every third entry in "dat.s"

For Welch's test to be a valid test of the null hypothesis of equality of population means, both populations should be (approximately) Normally distributed.

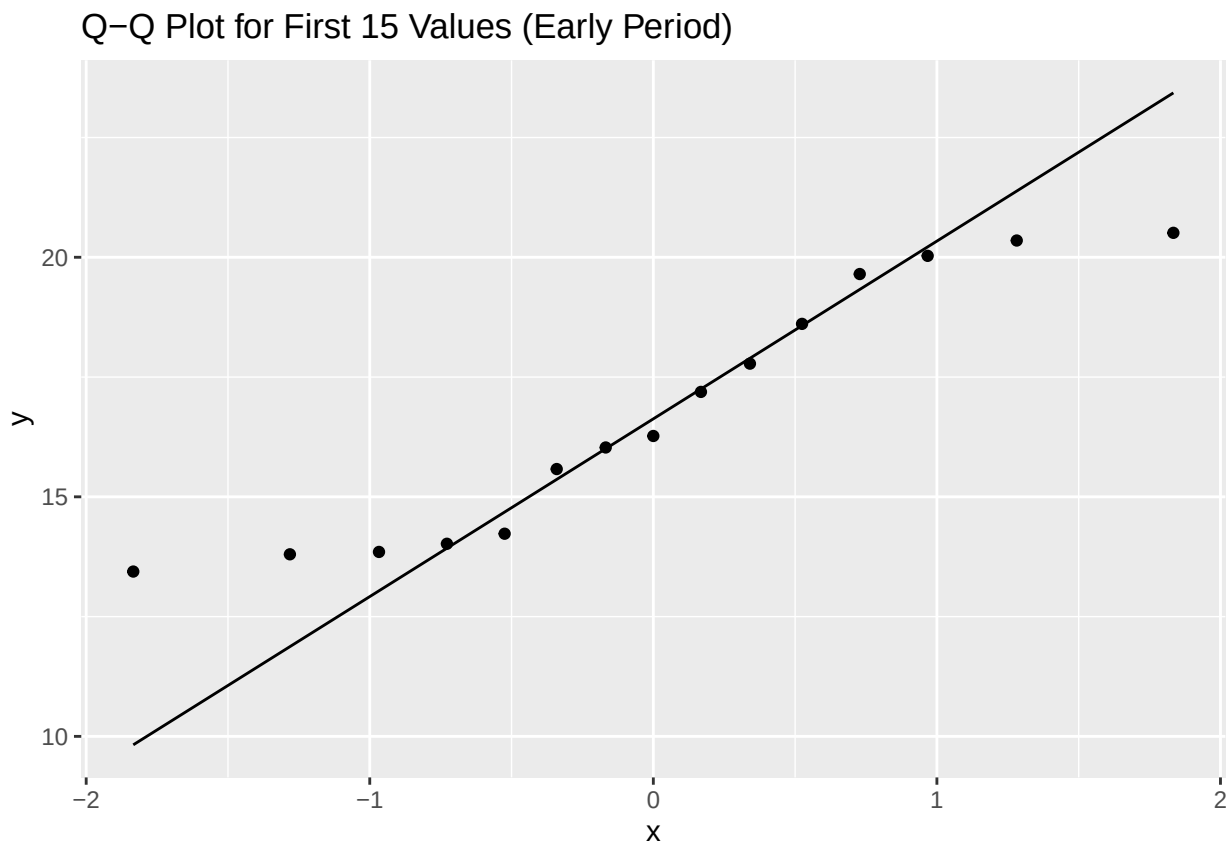
Please provide a visual assessment of the consistency with Normality of the first 15 values for "year.total" in "dat.s" and of the consistency with Normality of the last 15 values for "year.total" in "dat.s". Please give a verbal assessment based on the visualization. Within each period, are these data consistent with Normality?

Your answer and code here:

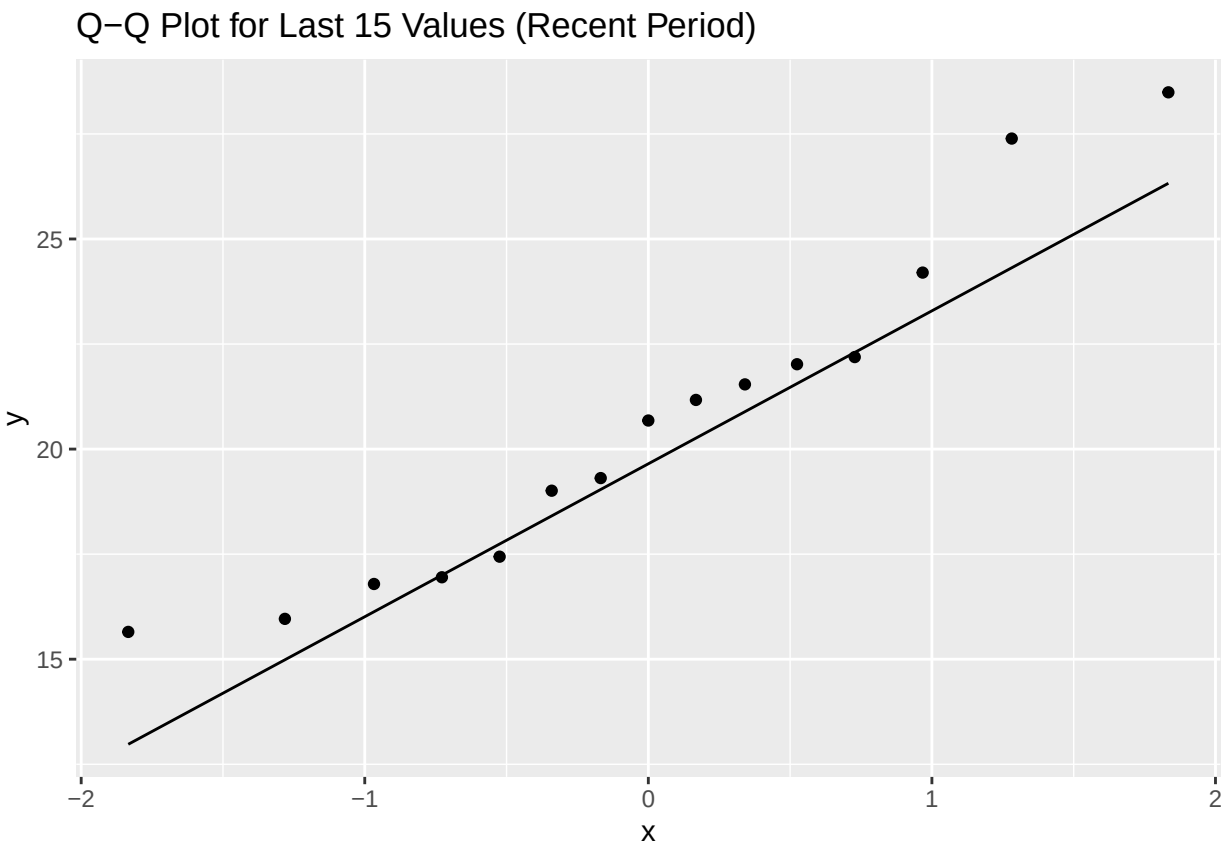
```
dat.s <- filter(dat.trim,year %% 3 == 2)
dat.sep <- dat.s[c(1:15, (nrow(dat.s) - 14):nrow(dat.s)), ]
dat.sep$era <- rep(c("early", "recent"), times = c(15, 15))
# your plotting code here

# Generate Q-Q plots
p1 <- ggplot(dat.sep %>% filter(era == "early"), aes(sample = year.total)) +
  geom_qq() +
  geom_qq_line() +
  ggtitle("Q-Q Plot for First 15 Values (Early Period)")
p2 <- ggplot(dat.sep %>% filter(era == "recent"), aes(sample = year.total)) +
  geom_qq() +
  geom_qq_line() +
```

```
  ggtitle("Q-Q Plot for Last 15 Values (Recent Period)")
# Display the plots
print(p1)
```




```
print(p2)
```



By visually examining the Q-Q plots, you can determine whether the data within each period are consistent with a normal distribution. If the points for either group closely follow the straight line, the data can be considered consistent with normality. Significant departures from the line would suggest that the assumption of normality may not hold for that group. While both samples show some deviations from a normal distribution, the data from the recent period (last 15 values) appear more normally distributed compared to the early period (first 15 values). Therefore, the assumption of normality holds better for the recent period data than for the early period data.

Question 1.2

(5 points)

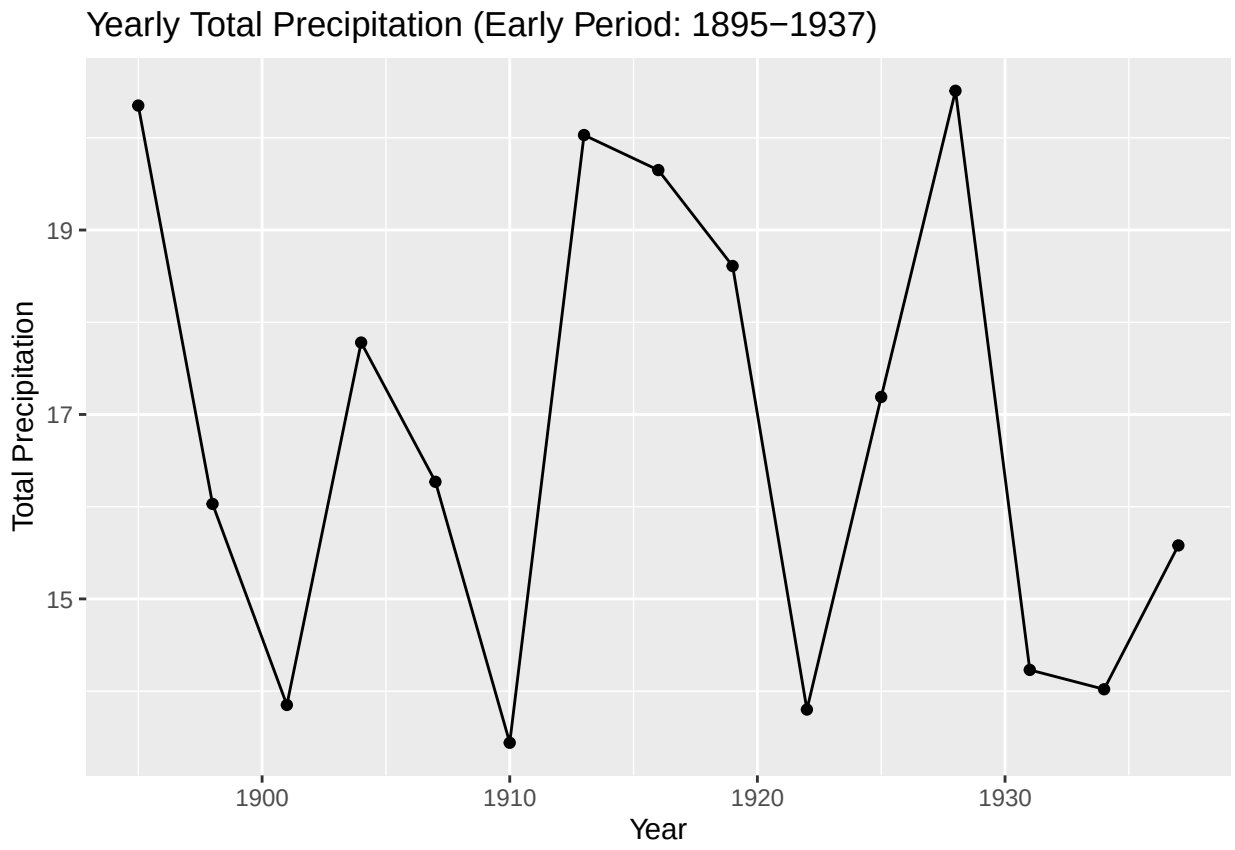
For Welch's test to be a valid test of the null hypothesis of equality of population means, the values in each group should be independent of one another.

Please provide a visualization to examine whether the "year.total" values show smooth variation over time, an indication of dependence, or whether the "year.total" values at consecutive time points in "dat.s" within the early (first 15 in dat.s, 1895-1937) period appear to be independent of one another and the "year.total" values at consecutive time points in "dat.s" within the recent(last 15 in dat.s, 1979-2021) periods appear to be independent. Please state your assessment.

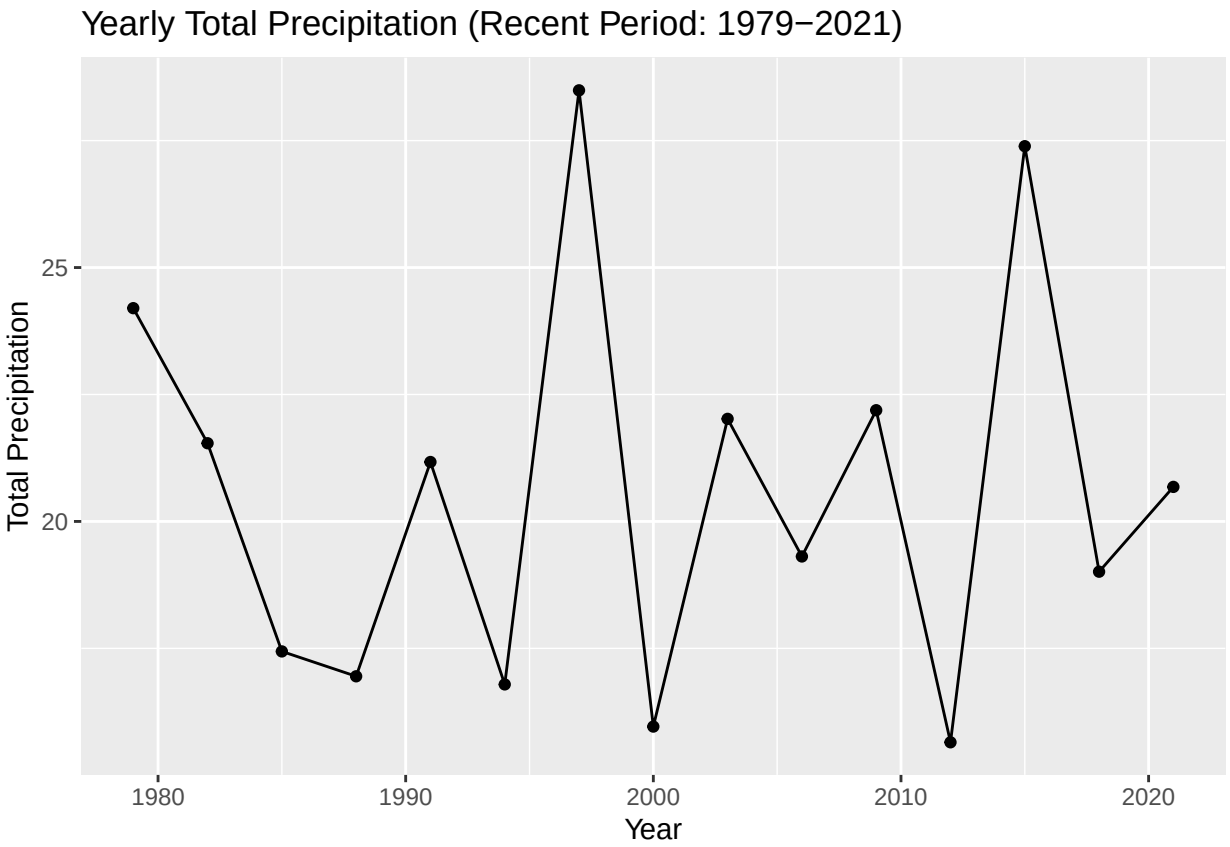
Your answer and code here:

```
# Plot for the early period
p1 <- ggplot((dat.sep %>% filter(era == "early")), aes(x = year, y = year.total)) +
  geom_line() +
  geom_point() +
```

```
  ggtitle("Yearly Total Precipitation (Early Period: 1895-1937)") +
  xlab("Year") +
  ylab("Total Precipitation")
# Plot for the recent period
p2 <- ggplot((dat.sep %>% filter(era == "recent")), aes(x = year, y = year.total)) +
  geom_line() +
  geom_point() +
  ggtitle("Yearly Total Precipitation (Recent Period: 1979-2021)") +
  xlab("Year") +
  ylab("Total Precipitation")
# Display the plots
print(p1)
```




```
print(p2)
```



Both plots show the "year.total" values fluctuating without a smooth, continuous trend, and the values appear scattered and do not follow a clear pattern. This suggests that the values at consecutive time points are likely independent for both periods.

Question 1.3

(5 points)

Please perform Welch's test of the null hypothesis that the total annual rainfalls in the early portion (first 15 values of dat.s) and the total annual rainfalls in the recent portion (last 15 values of dat.s) are each i.i.d. samples from Normally distributed populations with equal means, $Normal(\mu, \sigma^2_{early})$ and $Normal(\mu, \sigma^2_{recent})$. Please state your conclusion based on 1.a. and 1.b. regarding the null hypothesis that the means in the two populations are equal.

Your answer and code here:

```
# Extract the year.total values for the early and recent periods
early_values <- dat.sep %>% filter(era == "early") %>% pull(year.total)
recent_values <- dat.sep %>% filter(era == "recent") %>% pull(year.total)
# Conduct Welch's t-test
welch_test <- t.test(early_values, recent_values, alternative = "two.sided", var.equal = FALSE)
# Print the results of the test
print(welch_test)
```

```
##
##  Welch Two Sample t-test
```

```
##
## data:  early_values and recent_values
## t = -3.1518, df = 24.345, p-value = 0.004267
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -6.336094 -1.323906
## sample estimates:
## mean of x mean of y
##    16.756    20.586
```

Given the p-value of 0.004267, which is less than 0.05, we reject the null hypothesis. Additionally, the confidence interval for the difference in means does not include zero, further indicating a significant difference between the means. Thus, there is strong evidence to suggest that the means of the total annual rainfalls in the early portion (1895-1937) and the recent portion (1979-2021) are not equal. Therefore, we conclude that the total annual rainfalls between these two periods are significantly different, with the recent period having a higher mean annual rainfall compared to the early period.

Question 2

The goal in this analysis is to perform the strongest suitable test of whether the precipitation amount differs annually between October and November.

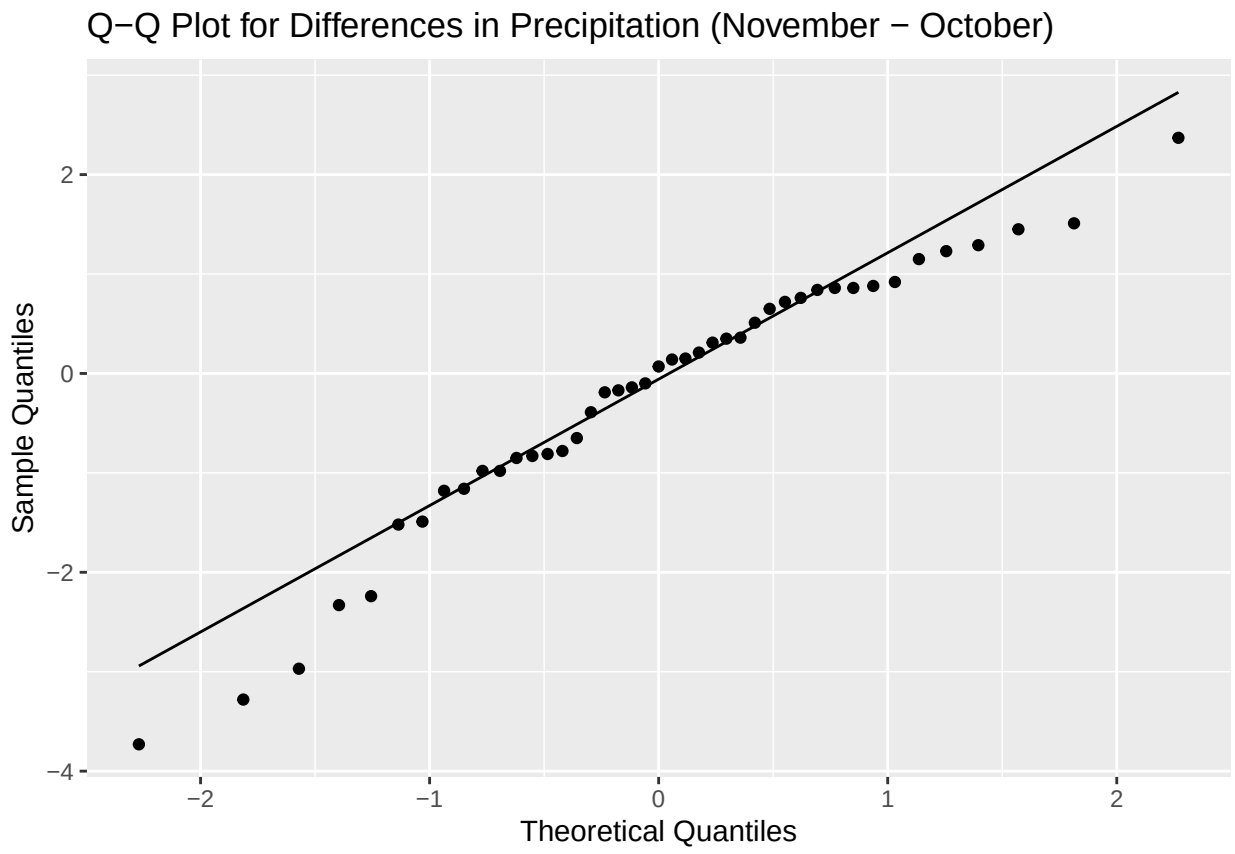
Question 2.1

(5 points)

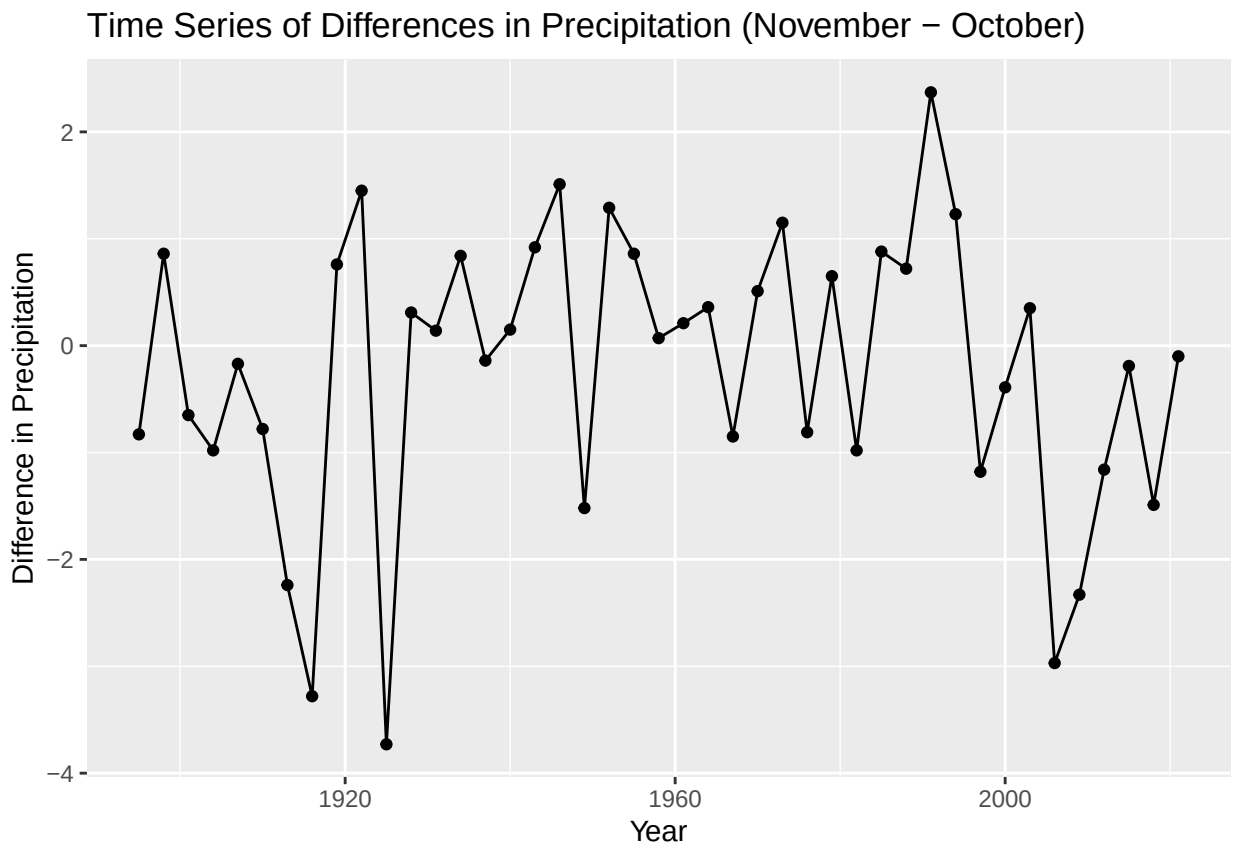
Please generate visualizations to address whether the differences between the precipitation in October and the following November in "dat.s" are consistent with being i.i.d. samples from a $Normal(\mu\sigma^2)$ distribution. Please address independence and Normality.

Your answer and code here:

```
# Calculate the differences between November and October precipitation
diff <- (dat.s$nov - dat.s$oct)
# Create a data frame for plotting
diff_data <- data.frame(year = dat.s$year, diff = diff)
# Q-Q Plot to check for normality
qq_plot <- ggplot(diff_data, aes(sample = diff)) +
  geom_qq() +
  geom_qq_line() +
  ggtitle("Q-Q Plot for Differences in Precipitation (November - October)") +
  xlab("Theoretical Quantiles") +
  ylab("Sample Quantiles")
# Time series plot to check for independence
time_series_plot <- ggplot(diff_data, aes(x = year, y = diff)) +
  geom_line() +
  geom_point() +
  ggtitle("Time Series of Differences in Precipitation (November - October)") +
  xlab("Year") +
  ylab("Difference in Precipitation")
# Display the plots
print(qq_plot)
```




```
print(time_series_plot)
```



The Q-Q plot shows some deviations from the diagonal line, especially in the tails. This indicates that the differences between October and November precipitation amounts may not be perfectly normally distributed. However, the deviations are not extreme, suggesting that the normality assumption weakly holds true. The time series plot shows values fluctuating without a smooth, continuous trend. The values appear scattered and do not follow a clear pattern, suggesting that the differences are likely independent. Given the visual assessment, we can proceed with a paired t-test, keeping in mind that while the normality assumption is not perfectly met, it is not severely violated.

Question 2.2

(5 points)

Please perform the strongest test of the null hypothesis that the difference in precipitation between October and November in each year has mean equal to 0. What do the results of this test imply?

Your answer and code here:

```
# Perform a paired t-test
paired_t_test <- t.test(dat.s$nov, dat.s$oct, paired = TRUE, alternative = "two.sided")
# Print the results of the test
print(paired_t_test)
```

```
##
##  Paired t-test
##
## data:  dat.s$nov and dat.s$oct
```

```
## t = -1.0474, df = 42, p-value = 0.3009
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  -0.6248408  0.1978641
## sample estimates:
## mean difference
##      -0.2134884
```

The p-value is 0.3009, which is greater than the significance level of 0.05. Therefore, we fail to reject the null hypothesis. The confidence interval for the difference in means is (-0.6248408, 0.1978641), which includes zero. The mean of the differences is -0.2134884. Based on the results of the paired t-test, there is not enough evidence to reject the null hypothesis that the mean difference in precipitation between October and November is zero. This suggests that there is no significant difference in the precipitation amounts between October and November. The differences between the precipitation amounts in these two months do not appear to be significantly different from zero.

Question 3

Assumptions can be dangerous! In this section you are asked to state the assumptions of the t-test, Wilcoxon signed rank test, and the sign test for a single population parameter.

Question 3.1

(2 points)

State the assumptions of the t-test for a single population mean.

Your Answer:

The data should be drawn randomly from a population with a distribution that is approximately normal, with observations independent of each other.

Question 3.2

(2 points)

State the assumptions of the Wilcoxon signed rank test for a single population parameter.

Your answer:

The data should be drawn randomly from a population with a symmetric distribution, with observations independent of each other.

Question 3.3

(2 points)

State the assumptions of the signed test for a single population parameter.

Your answer:

The data should be drawn randomly from a population, with observations independent of each other.

Question 3.4

(3 points)

Under what condition does the Sign and Wilcoxon signed rank test give a test of hypothesis on the population mean?

Answer: The Sign and Wilcoxon signed rank tests give a test of hypothesis on the population mean when the population distribution is symmetric. For a symmetric distribution, the median and the mean are equal, thus allowing these tests, which primarily test for differences in the median, to be used to infer about the mean as well.

Question 4

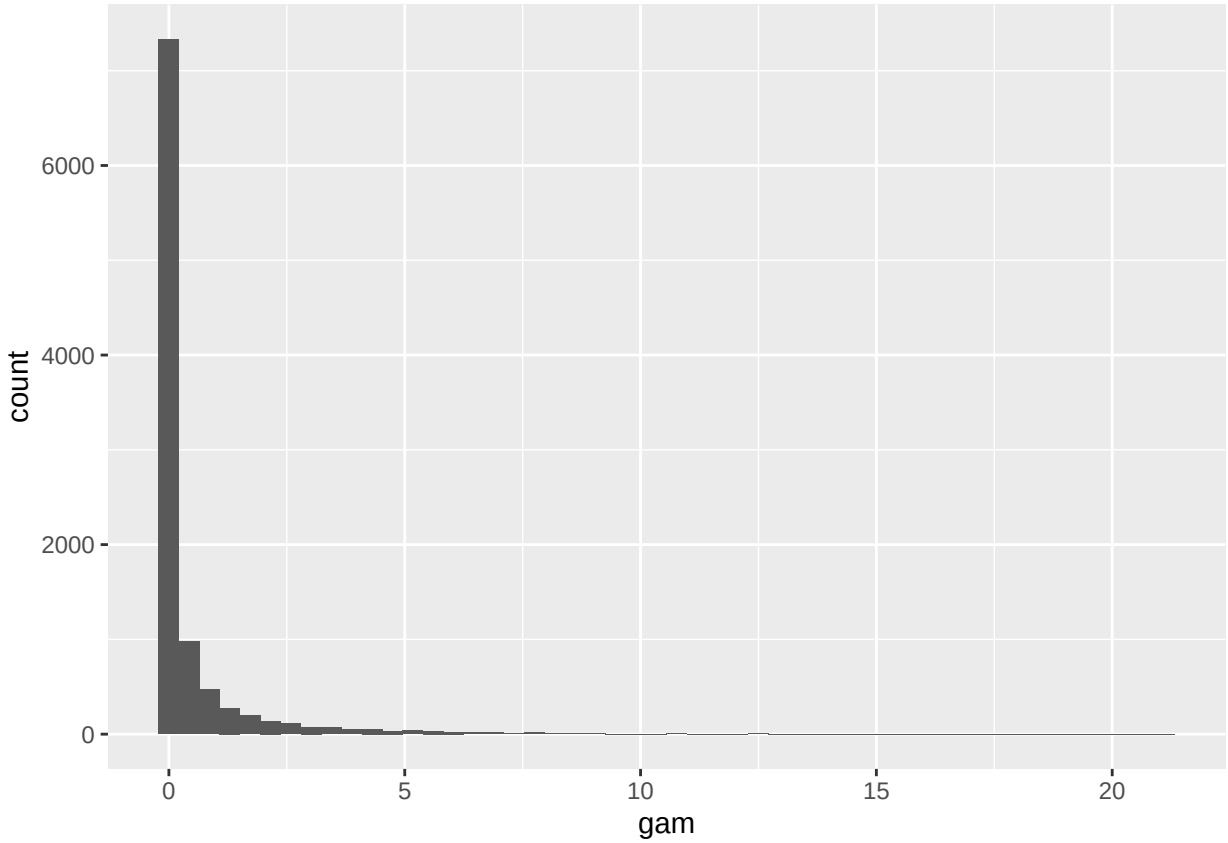
We now investigate the statistical power of the sign test, the Wilcoxon signed rank test, and the one sample t-test. To test the hypothesis that the mean is 1 when in fact the true mean is less. In the code chunk we calculate the probability that the test correctly rejects the null hypothesis thus calculating the power of each of the tests labeled 'tPower', 'wPower',and 'signPower'. In each question we vary the underlying distribution and parameters of the distribution.

Question 4.1

(5 points)

In the code chunk below we create functions to simulate drawing random samples of size 15 from a the Gamma distribution shown in the histogram, and calculate the power of each of the statistical tests. Each test is a test that the median/mean value is one against the alternative that the median/mean is less. Which statistical test would you select and why based on the power calculation? Additionally which statistical test best meets the assumptions of statistical test according to the histogram?

```
gam = rgamma(10000,shape = 1/8, scale = 4)
ggplot(data = data.frame(gam), aes(x=gam) ) + geom_histogram(bins = 50)
```




```
Rejection = function(data){
  n = nrow(data)
  tReject = t.test(data,mu = 1 , alternative = 'less')$p.value < .05
  wReject = wilcox.test(data,mu = 1 , alternative = 'less',exact = FALSE)$p.value < .05
# sign test
  k <- sum(data > 1)
  n <- length(data)
  p = pbinom(k, n, .5) # one sided sign test
  signReject = (p < .05)
 return(c(tReject, wReject, signReject))
}
PowerWrangle = function(Rpower){
# Some wrangling
Powers = data.frame(mean(Rpower[seq(1, to = length(Rpower), by = 3)]),
                            mean(Rpower[seq(2, to = length(Rpower), by = 3)]) ,
                            mean(Rpower[seq(3, to = length(Rpower), by = 3)]))
names(Powers) = c('tPower', 'wPower', 'signPower')
return(Powers) # final results
}
Rpower = replicate(1000, Rejection(rgamma(15, shape = 1/8, scale = 4))) # The data used is randomly gen
PowerWrangle(Rpower)
```

```
##   tPower wPower signPower
## 1  0.593  0.814     0.881
```

Your answer:

The sign test has the highest power (0.871), followed by the Wilcoxon signed rank test (0.830), and the t-test has the lowest power (0.588). The sign test is also recommended because it makes the fewest assumptions about the data distribution, making it most suitable for the highly skewed Gamma distribution in question. Thus, the sign test is the best choice due to its highest power and minimal assumptions, making it robust for the given skewed data distribution.

Question 4.2

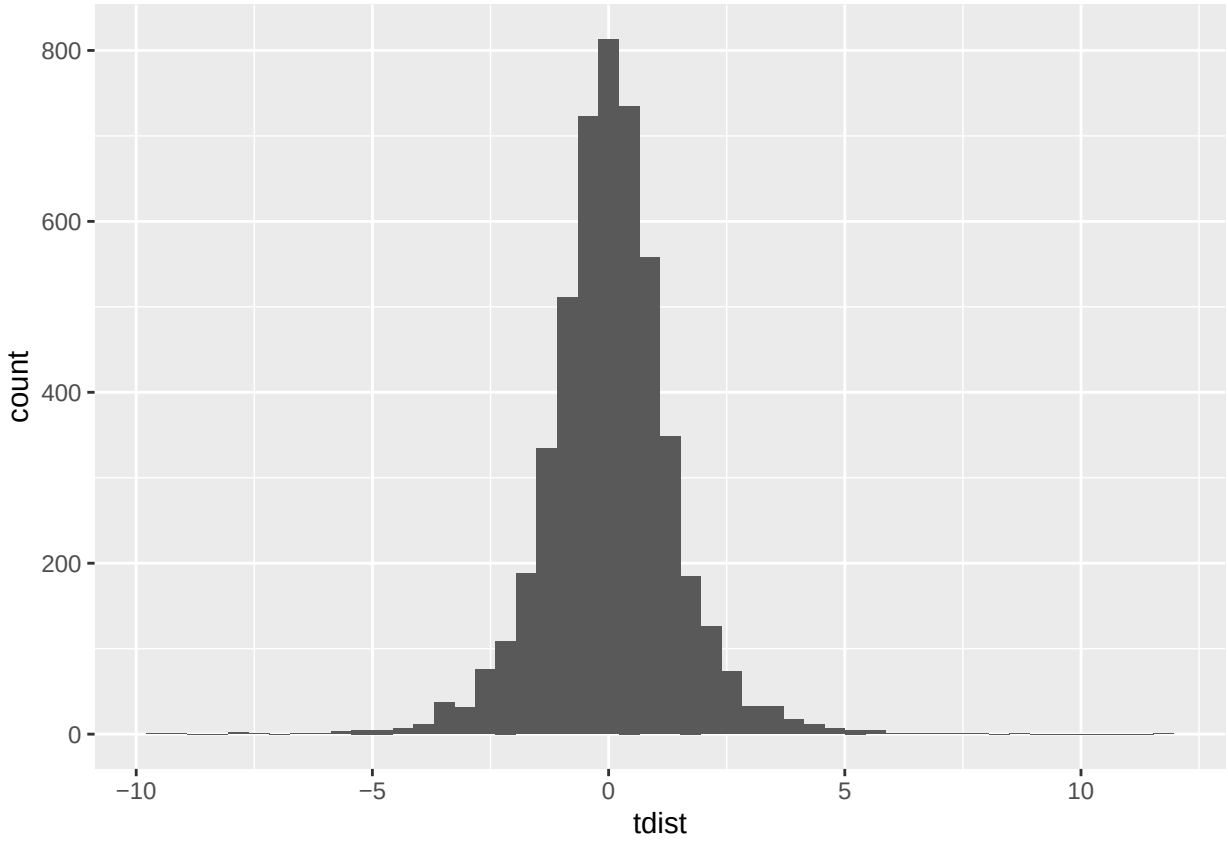
(5 points)

We simulate drawing random samples of size 15 from the student t-distribution with four degrees of freedom as shown in the histogram, and calculate the power of each of the statistical tests. Each test is a test that the median/mean value is one against the alternative that the median/mean is less. Which statistical test would you select and why based on the power calculation? Additionally, which statistical test best meets the assumptions of statistical test according to the histogram? Note the data is not normally distributed according to the shapiro test.

```
tdist = rt(5000, 4)
shapiro.test(tdist)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  tdist
## W = 0.95859, p-value < 2.2e-16
```

```
ggplot(data = data.frame(tdist), aes(x = tdist) ) + geom_histogram(bins = 50)
```




```
Rpower = replicate(1000, Rejection(rt(15, 4))) # The data used is randomly generated t random variables
PowerWrangle(Rpower)
```

```
##   tPower wPower signPower
## 1  0.836  0.882     0.691
```

Your answer:

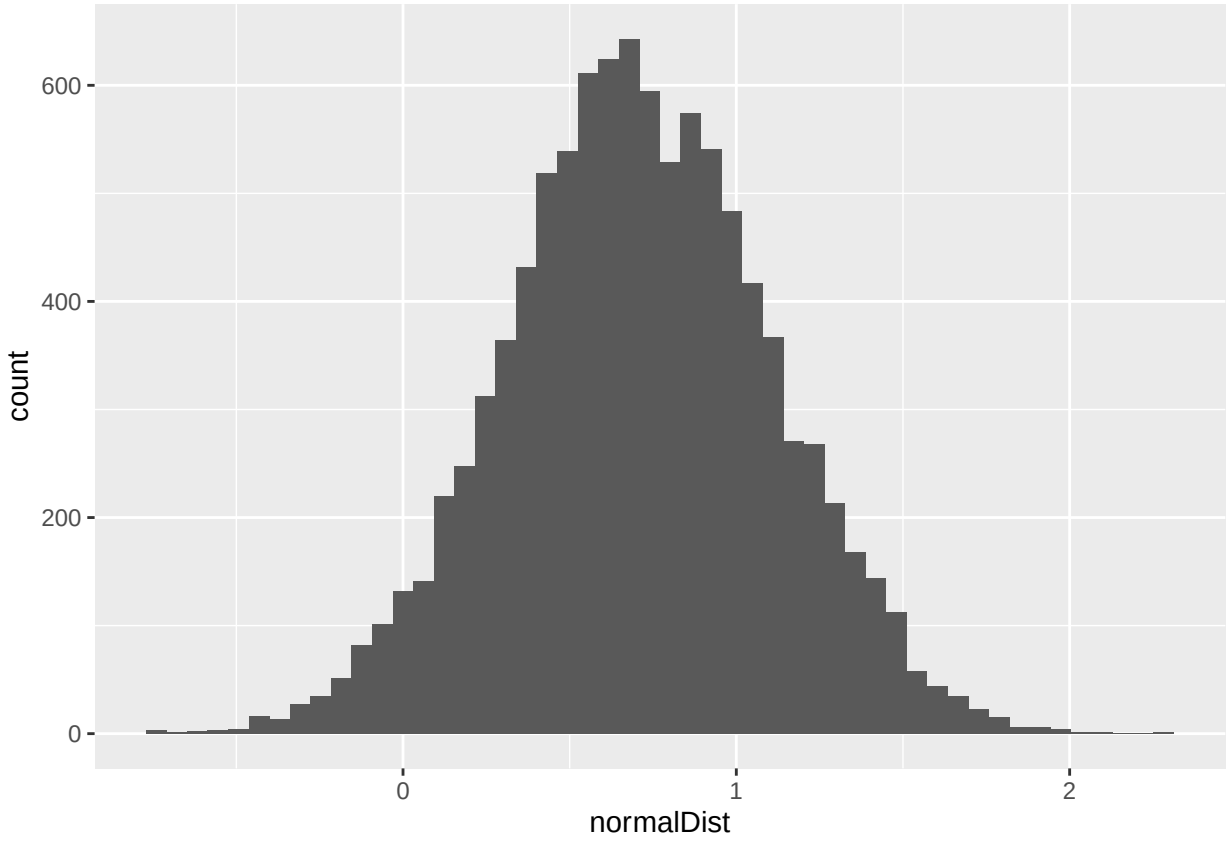
The Wilcoxon signed rank test has the highest power (0.869), followed closely by the t-test (0.834), with the sign test having the lowest power (0.697). The Wilcoxon signed rank test is recommended as it has the highest power (0.869). The Wilcoxon signed rank test is also recommended because it best meets the assumptions for the t-distribution with 4 degrees of freedom, being less stringent than the t-test and more powerful than the sign test. Thus, the Wilcoxon signed rank test is the best choice due to its highest power and its assumptions being most appropriate for the given t-distribution data.

Question 4.3

(5 points)

We simulate drawing random samples of size 15 from the normal distribution with a mean of .7 and variance of .4 as shown in the histogram. We calculate the power of each of the statistical tests. Each test is a test that the median/mean value is one against the alternative that the median/mean is less. Which statistical test would you select and why based on the power calculation? Additionally, which statistical test best meets the assumptions of statistical test according to the histogram?

```r
normalDist = rnorm(10000, .7, .4)
ggplot(data = data.frame(normalDist), aes(x = normalDist) ) + geom_histogram(bins = 50)
```




```r
Rpower = replicate(1000, Rejection(rnorm(15, .7, .4))) # The data used is randomly generated normal ran
PowerWrangle(Rpower)
```

```
##   tPower wPower signPower
## 1   0.88  0.854     0.559
```

Your answer:

The t-test has the highest power (0.872), followed closely by the Wilcoxon signed rank test (0.857), with the sign test having the lowest power (0.54). The t-test is recommended as it has the highest power (0.872). The t-test is also recommended because the data is normally distributed, perfectly meeting the assumptions of the t-test. Thus, the t-test is the best choice due to its highest power and its assumptions being perfectly met for the given normally distributed data.

End.