Carson Slater STA 5380 Homework #5

---

1. An EPA researcher wants to design a study to estimate the mean lead level of fish in a lake located near an industrial area. Based on past sample data, the researcher estimates that σ for the lead level in the fish population is approximately 0.016 mg/g. He wants to use a 95% confidence interval having a margin of error no greater than 0.005 mg/g.

   (a) How many fish does he need to catch?
   We know that in order to obtain a 95% confidence interval, the researcher must catch n fish, such that n satisfies the following equation,

$$\begin{aligned} Z_{0.025} \cdot \frac{\sigma}{\sqrt{n}} &= 0.005 \\ \implies n &= \left(Z_{0.025} \cdot \frac{\sigma}{0.05} \right)^2 \\ &= \left(1.96 \cdot \frac{0.016}{0.005} \right)^2 \\ &= 39.336. \end{aligned}$$

   So the researcher would need to catch at least 40 fish to meet the desired margin of error.

   (b) The researcher collects a sample of $n = 50$ fish and computes a 95% confidence interval for the mean lead level to be (0.9932, 1.0045). Interpret this particular interval in the context of the problem.
   Assuming this confidence interval was obtained correctly, using the correct methodology, we are 95% confident that mean lead level of fish in a lake located near this industrial area is between 0.9932 mg/g and 1.0045 mg/g.

   (c) In your own words, explain the concept of 95% confidence in general for this problem (interpretation #1 from page 120 of the notes).
   In repeated samples of size 40 or more, we believe the methodology that is used to obtain the confidence interval is 95% likely to contain the true mean lead level in fish, and it is 5% likely to not contain the true mean lead level of fish in this particular lake.

2. For a semester project, a student was interested in determining the average weight gain of college students during their freshman year. She took a SRS of 20 freshman and found a sample mean weight gain of $\bar{x} = 5.25$ pounds with a standard deviation of 10 pounds. Suppose that it is known that weight gain during the freshman year is normally distributed. Use this to answer the following questions:

   (a) Calculate a 90% CI for μ, the average weight gain of college students during their freshman year.
   Given size $n = 20$, we can find the 90% confidence interval for μ using the following formula,

   $$\bar{x} \pm t_{0.05,\,19} \cdot \frac{s}{\sqrt{n}} = 5.25 \pm 1.729 \cdot \frac{10}{\sqrt{20}}$$

   So then the 90% confidence interval for the mean weight gained during freshman year is (1.383, 9.116).

   (b) Which of the following is/are correct interpretation(s) for the confidence interval from the previous part? Explain the reasoning for your choice(s).

      i. We are 90% confident that the average weight gain of college students during their freshman year is in the interval.
      ii. We are 90% confident that 5.25 is in the interval.
      iii. There is a 90% probability that 5.25 is in the interval.
      iv. Out of 100 intervals, approximately 90% will contain $\bar{X}$, and 10 of them will not.

   The correct interpretation would be the first option, as the sample has already been selected. The true parameter, μ is the true average weight gain of college students during their freshman year, and so a confidence interval of the 90% will yield an upper and lower bound with 90% confidence that the set contained between upper and lower bound also contains the true population parameter.

3. A gasoline company tested 19 samples of gasoline produced during a day to check whether the day's production meets the nominal octane rating of 87. The results are as follows:

```
octane <- c(87.5, 86.9, 87.3, 87.9, 88.0, 86.7, 87.5, 87.2, 87.0, 88.1,
            87.5, 86.5, 87.7, 88.0, 87.1, 87.0, 87.6, 87.5, 88.3)
```

Find a one-sided lower 95% confidence limit on the mean octane rating. Use this confidence limit to determine if it appears that the mean octane rating exceeds 87.

```
t.test(octane, mu = 87, alternative = "greater", conf.level = 0.95)$conf.int
```

```
## [1] 87.23961      Inf
## attr(,"conf.level")
## [1] 0.95
```

The 95% one-side confidence interval for the mean octane rating is $(87.239, \infty)$. Using only this confidence interval, we can be 95% confident that the mean octane rating is above 87.

4. Tell in each of the following instances whether the study uses an independent samples or a matched pairs design. In each, identify and define the parameter of interest.

   (a) Two computing algorithms are compared in terms of the CPU times required to do the same six test problems.
   This scenario would require a match pairs design for the mean difference in computing time between the two algorithms. Each sample would be matched by the computing time for each of the six problems.

   (b) An agronomist compares the yields of two varieties of soybean by planting each variety in 10 separate plots of land (for a total of 20 plots).
   Assuming the plots of land are geographically alike and similar in climate and soil composition, this experiment uses independent samples where they ought not to be paired together to test the mean yield between two soybean varieties.

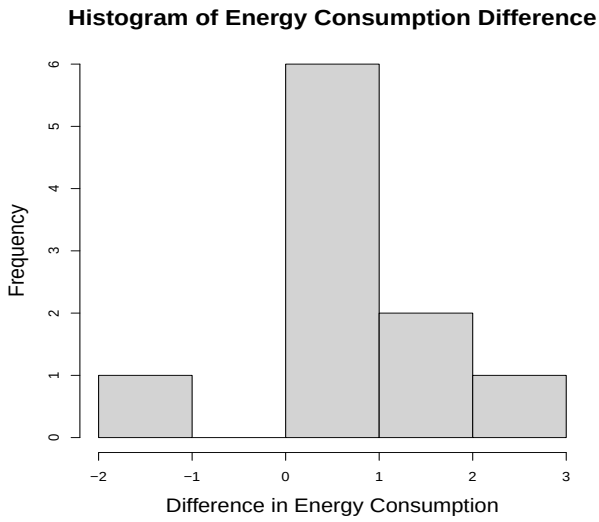
   (c) An advertising agency has come up with two different commercials for a household detergent. To determine which one is more effective, a test is conducted in which a sample of 100 adults is randomly divided into two groups. Each group is shown a different commercial, and the people in the group are asked to score the commercial.
   Knowing each group only sees one commercial and rates it, each group is then considered independent, and the parameter of interest would be the mean rating that each group gives their respective commercial.

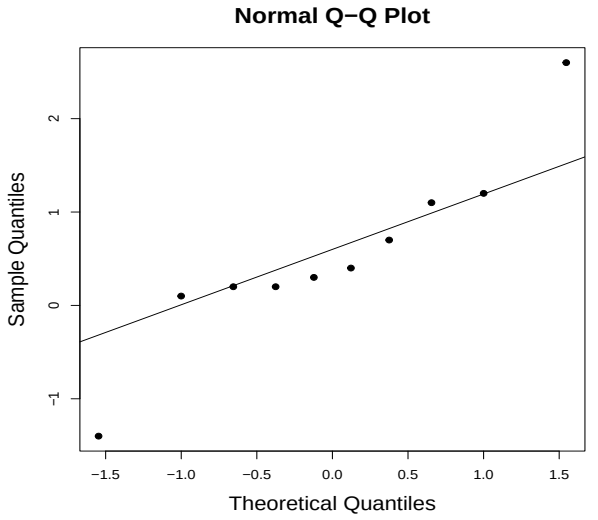
   (d) Military test pilots who had at least one accident are matched by length of experience to pilots without accidents. The two groups are then surveyed about the number of childhood accidents to determine if the former group of pilots is more "accident prone."
   This is a match pairs design to determine if the mean number of childhood accidents for pilots with an accident on their record is greater than those who have not had an accident.

5. To study the effectiveness of wall insulation in saving energy for home heating, the energy consumption (in MWh) for 10 houses in Bristol, England, was recorded for two winters; the first winter was before insulation, and the second winter was after insulation.

```
before <- c(12.1, 11.0, 14.1, 13.8, 15.5, 12.2, 12.8, 9.9, 10.8, 12.7)
after <- c(12.0, 10.6, 13.4, 11.2, 15.3, 13.6, 12.6, 8.8, 9.6, 12.4)
```

Find a 90% confidence interval for the true mean difference in energy consumption. Does it appear that the wall insulation has reduced the mean energy consumption?






```
t.test(before - after, conf.level = 0.9)$conf.int

## [1] -0.04875625  1.12875625
## attr(,"conf.level")
## [1] 0.9
```

The 90% confidence interval for the difference in energy consumption is (-0.0487, 1.1287), which indicates there is not sufficient evidence that the wall insulation has reduced energy consumption (Zero is contained in the confidence interval).

6. Two brands of water filters are to be compared in terms of the mean reduction in impurities measured in parts per million (ppm). Twenty-one water samples were tested with each filter and reduction in impurity level was measured, resulting in the following data:

   Filter 1: $n_1 = 21 \quad \bar{x}_1 = 8.0 \quad s_1^2 = 4.5$
   Filter 2: $n_2 = 21 \quad \bar{x}_2 = 6.5 \quad s_2^2 = 2.0$

   (a) Calculate a 95% confidence interval for the mean difference $\mu_1 - \mu_2$ between the two filters assuming that $\sigma_1^2 = \sigma_2^2$. Is there a statistically significant difference between the mean reduction in impurity levels between the two filters?

   Assuming $\sigma_1^2 = \sigma_2^2$, we must also assume that all observations are independent, that both $\bar{x}_1$ and $\bar{x}_2$ are normally distributed, and that there are no outliers skewing the estimated means. Finding a confidence interval assuming equal variances, we can pool s_1^2 and s_2^2 into S_p^2 using the following formula,

$$\begin{aligned} S_p^2 &= \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)} \\ &= \frac{20(4.5) + 20(2.0)}{20 + 20} \\ &= 3.25. \end{aligned}$$

   Then the confidence interval for the reduction of impurity levels between the two filters is,

$$\bar{x}_1 - \bar{x}_2 \pm t_{0.025,\,40} \cdot S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}},$$

   which in this case yields a 95% confidence interval of (0.375, 2.624). This confidence interval indicates that there does not appear to be a statistically significant between the mean reduction in impurity levels between the two filters, as zero is contained in the confidence interval.

   (b) Repeat (a) without assuming $\sigma_1^2 = \sigma_2^2$. Compare the results.

   Assuming unequal variances, we must use the following confidence interval for the difference of means, using the Satterthwaite degrees of freedom, $\nu \approx 34.845$,

$$\bar{x}_1 - \bar{x}_2 \pm t_{0.025,\,34.845} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

   Using the given sample means and variances, the following confidence interval yields the interval (0.370, 2.629). This confidence interval does not contain zero and would be indicative that there exists a significant difference between the mean reduction in impurity levels between the two filters.

7. In 1993 *Time* magazine reported a telephone poll survey of 800 adults in the U.S., of whom 45% stated that they had guns in their homes.

   (a) Check that the assumptions are met to compute a 95% confidence interval for the true proportion of adults in the U.S. who have a gun in their home.

   We can check if we have a sufficiently large sample by checking if $n\hat{p} > 10$ and $n(1-\hat{p}) > 10$.

   $$n\hat{p} = 800 \cdot 0.45 = 360 \quad \text{and} \quad n(1-\hat{p}) = 800 \cdot 0.55 = 400.$$

   Both are greater than 10, so we may proceed assuming a sufficiently large sample to suppose that $\hat{p} \sim \mathcal{N}(0.45, 0.018)$.

   (b) Find the 95% confidence interval for the proportion stated in part (a) using the score interval and the large-sample interval.

```
ptilde <- (.45 + qnorm(.975)^2/(2*800))/(1+ qnorm(.975)^2/800)
me <- qnorm(.975) *(sqrt((.45*.55/800)+(qnorm(.975)^2)/(4*(800^2))))/
      (1+((qnorm(.975)^2)/800))

cat("(", ptilde - me, "," , ptilde + me, ")")

## ( 0.4158467 , 0.4846312 )
```

   First using the score interval,

   $$\frac{\hat{p} + \frac{z^2_{0.025}}{2n}}{1 + \frac{z^2_{0.025}}{n}} \pm z_{0.025} \frac{\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2_{0.025}}{4n^2}}}{1 + \frac{z^2_{0.025}}{n}},$$

   we obtain the interval (0.3827, 0.5175). Then using the large sample interval,

   $$\hat{p} \pm z_{0.025} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}},$$

   we obtain the interval (0.4155, 0.4845).

   (c) Is it guaranteed that the true proportion lies within the bounds that you computed in part (b)? Explain why or why not.

   It is not guaranteed that the true proportion lies within the bounds that were computed above. We can expect the true parameter to be contained in the bounds we computed 19 times for every twenty random intervals created from the sample.

8. The data in the table below are from a study by chemist and Nobel Laureate Linus Pauling, and it shows the incidence of colds among 279 French skiers who were randomized to the Vitamin C and Placebo Groups.

	Cold		
Group	Yes	No	Row Total
Vitamin C	17	122	139
Placebo	31	109	140
Column Total	48	231	279

(a) Compute a 95% CI for the difference in proportions of skiers who contracted a cold while taking the placebo versus those who took Vitamin C.

We can see that if p_1 is the proportion of skiers who contracted a cold while taking the placebo, $\hat{p}_1 = \frac{31}{140} \approx 0.2214$. Likewise if p_2 is the proportion of skiers who contracted a cold while taking the Vitamin C, $\hat{p}_2 = \frac{17}{139} \approx 0.1223$. We can check if each proportion has a sufficiently large sample to be approximated a large sample distribtion.

$$n\hat{p}_1 = 140 \cdot \frac{31}{140} = 31 \quad \text{and} \quad n(1-\hat{p}_1) = 140 \cdot \frac{140-31}{140} = 109.$$
$$n\hat{p}_2 = 139 \cdot \frac{17}{139} = 17 \quad \text{and} \quad n(1-\hat{p}_2) = 139 \cdot \frac{139-17}{139} = 122.$$

All of these values are greater than 10, so we can assume $\hat{p}_1$ and $\hat{p}_2$ are distributed normally. This implies their difference is normally distributed, as any linear combination of normally distributed variables is also normally distributed. So then

$$\hat{p}_1 - \hat{p}_2 \sim \mathcal{N}\left(p_1 - p_2, \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}^2\right).$$

We can create a 95% confidence interval for the difference in proportions of skiers who contracted a cold while taking the placebo versus those who took Vitamin C using the following interval,

$$\hat{p}_1 - \hat{p}_2 \pm z_{0.025} \cdot \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}},$$

which yields the interval (0.0114, 0.1869).

(b) Interpret the interval you computed in part (b) in the context of this problem.

We are 95% confident that the difference in proportions of skiers who contracted a cold while taking the placebo versus those who took Vitamin C is between (0.0114, 0.1869).

9. In your own words, describe the difference between a confidence interval for the mean, μ, a prediction interval, and a tolerance interval.

   A **confidence interval** for a population *parameter* is an interval that is expected to contain a population parameter (say, μ) with a certain confidence (not probability). As for a **prediction interval**, rather than being expected to contain a population parameter, this interval is expected to contain a new independent, identically distributed *observation* (say, random variable X_{n+1}) with a certain confidence (again, not probability). Lastly, a **tolerance interval** is expected to contain at least a specified *proportion* of all future observations given a population with some confidence. Hence, if you were to repeatedly sample from the population and construct tolerance intervals, you would expect a certain arbitrary percentage (say, 95%) of those intervals to contain the specified proportion of the population.

10. The durability of an upholstery fabric is measured in double rubs (DR), which simulates a person getting in and out of a chair. The manufacturing label on one fabric gives its durability range as 68,000-82,000 DR. If the durability can be modeled by a normal distribution, this would mean that two standard deviations is roughly 7,000 DR, and one standard deviation is 3,500 DR. The company's quality control department independently evaluated the fabric by testing 40 one-yard samples. The data is given in the file "durability.csv" on the class website.

    (a) Write an R function that will automatically compute a CI for the standard deviation. This function should accept arguments for the data and the confidence level and should return the upper and lower bounds of the interval.

```r
# data must be coerced into a single vector for ci_sd()
ci_sd <- function(data, conf.level = 0.95) {
  stopifnot(conf.level > 0 & conf.level < 1)

  stopifnot(is.vector(data))

  # obtain necessary parameters
  nu <- length(data)
  alpha <- (1 - conf.level)

  # calculate bounds (square roots of variance CI)
  lower <- sqrt(((nu - 1)*sd(data)^2)/qchisq(p = 1 - alpha/2, df = nu - 1))
  upper <- sqrt(((nu - 1)*sd(data)^2)/qchisq(p = alpha/2, df = nu - 1))

  # print CI
  cat(paste(conf.level*100, "% CI for SD is ", "(",round(lower,3),", ",round(upp
  invisible(list(upper, lower))
}
```

    (b) The mean is acceptable, but is the standard deviation consistent with the labeled range? Answer by constructing a 95% confidence interval for σ.

```r
ci_sd(durability)

## 95% CI for SD is (3393.815, 5319.808)
```

    Since 3,500 is contained in our 95% confidence interval for the standard deviation, we have insufficient evidence to reject that one standard deviation would be 3,500 DR.

    (c) Find a 95% prediction interval for the durability of this fabric. If an office is buying this fabric to cover furniture in a waiting room and requires a fabric with durability of at least 70,000 DR, would this be a good purchase?

    The 95% prediction interval for the durability of this fabric would be between 65,774 DR and 82,743 DR. If the bare minimum for fabric durability is 70,000 DR, then this would not be a good purchase.

    (d) Find a 95% tolerance interval that would include the DR values for 99% of all fabric made by the same manufacturing process. Does this tolerance interval fall within the

manufacturing specifications of 68,000-82,000 DR? (See your textbook, p. 685, to find the correct value of K.)

```
xbar <- mean(durability)
s <- sd(durability)
K <- 3.213
xbar - K*s; xbar + K*s
```

```
## [1] 60946.84
## [1] 87570.01
```

For our particular tolerance interval of interest, we would use $K = 3.213$. Using this K, we find that we are 95% confident that 99% of all future fabrics' durability will be between 60,947 DR and 87,570 DR.

11. The measurements on the compressive modulus of silicone rubber used in high voltage transformers are given below. Ten sample sheets from each batch of rubber were tested.

    These values can be entered into R with the following code:

```
b1 <- c(997.5, 972.4, 1064.2, 972.0, 994.4, 1044.5, 982.6, 956.3, 991.3, 1001.5)
b2 <- c(870.1, 1064.1, 925.9, 906.2, 969.8, 982.4, 936.0, 930.0, 927.2, 924.5)
b3 <- c(1018.7, 993.6, 939.6, 1006.3, 1027.0, 877.1, 926.4, 980.2, 966.6, 925.0)
b4 <- c(973.0, 1048.7, 1058.3, 961.9, 1002.2, 973.3, 967.6, 992.3, 956.1, 973.5)
```

    We are interested in finding confidence intervals for the parameters μ_{b1}, μ_{b2}, μ_{b3}, and μ_{b4}.

    (a) Find a 95% confidence interval for each individual parameter.

Parameter	95% Confidence Interval
μ_{b1}	(973.91, 1021.42)
μ_{b2}	(906.14, 981.09)
μ_{b3}	(931.54, 1000.55)
μ_{b4}	(965.07, 1016.30)

    (b) Find 95% simultaneous confidence intervals for all 4 parameters using the Bonferonni method.

Parameter	Bonferonni Simultaneous Confidence Interval
μ_{b1}	(965.01, 1030.33)
μ_{b2}	(892.08, 995.16)
μ_{b3}	(918.60, 1013.50)
μ_{b4}	(955.46, 1025.92)

    (c) Compare the average width of the intervals computed in parts (a) and (b).

```
mean(diff(t.test(b1)$conf.int),
     diff(t.test(b2)$conf.int),
     diff(t.test(b3)$conf.int),
     diff(t.test(b4)$conf.int))

## [1] 47.50257

new_conf <- 1 - 0.05/4

mean(diff(t.test(b1, conf.level = new_conf)$conf.int),
     diff(t.test(b2, conf.level = new_conf)$conf.int),
     diff(t.test(b3, conf.level = new_conf)$conf.int),
     diff(t.test(b4, conf.level = new_conf)$conf.int))

## [1] 65.32588
```

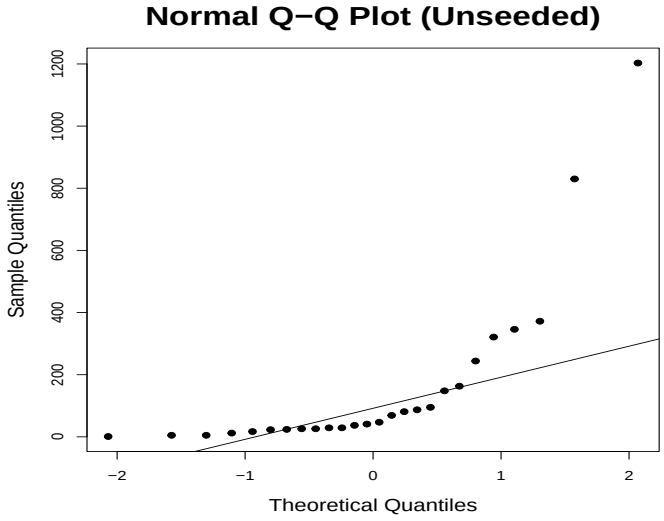
    The average width of the Bonferroni confidence intervals was 37.5% bigger than the individual confidence intervals. The mean width of the individual confidence intervals was 47.5 and the mean width of the simultaneous confidence intervals was 65.3.

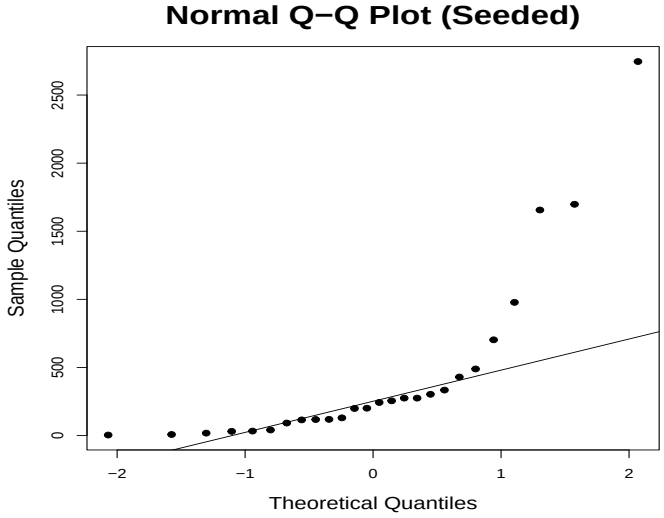
12. In Example 7.15 of the class notes, we computed a confidence interval for the true ratio of variances of the log rainfall of seeded versus unseeded clouds.

    (a) Why is it important when computing confidence intervals for the variance or for a ratio of variances that the data comes from a normal distribution?

    The normality assumption of the population which the data is drawn is critical in order that the distribution of the sample variance follows a chi-square distribution. This enables the ratio of the variances to be the ratio of two Chi-Square distributions with their respective degrees of freedom. This ratio yields an F-distribution, a nice result.

    (b) Take the log rainfall data and apply an exponential transform. For example, if the variable seeded contains the log rainfall data, then orig.seed $<-$ exp(seeded) will contain the original rainfall data. Does it appear to be normally distributed?





These data do not look normally distributed.

(c) Regardless of your answer to part (b), compute a 95% confidence interval (based on the F-distribution) for a ratio of the variances of the rainfall of the seeded versus unseeded clouds.

```
type <- c(rep("seed", 26), rep("unseed", 26))
data <- c(seed, unseed)
var.test(data ~ type)$conf.int

## [1]  2.449628 12.185054
## attr(,"conf.level")
## [1] 0.95
```

(d) Now, compute a 95% bootstrap confidence interval for the ratio of the variances of the rainfall of the seeded versus unseeded clouds.

```
set.seed(613)
unseed_pm <- array()
seed_pm <- array()
B <- 10000
n1 <- length(seed)
n2 <- length(unseed)
# nonparametric bootstrap
unseed_npm <- array()
seed_npm <- array()
for(i in 1:B){
  seed_npm[i] <- var(sample(seed, n1, replace = TRUE))
  unseed_npm[i]  <- var(sample(unseed, n2, replace = TRUE))
}

npm <- seed_npm/unseed_npm

quantile(npm, c(0.025, 0.975))

##       2.5%      97.5%
##  0.8005954 52.4332571
```

Using a nonparametric bootstrap (because normality appears to be violated), we obtained a 95% confidence interval for the ratio of the variances of the rainfall of seeded versus unseeded clouds. We are 95% confident that the true ratio of the variances of the rainfall of seeded versus unseeded clouds is contained in the interval (0.801, 52.4333).

(e) How do the intervals in parts (c) and (d) compare? Which one is more trustworthy?

Both intervals are vastly different. Unfortunately, the most reliable would probably be the nonparametric bootstrap interval. It is very wide and very unhelpful. This is because the normality assumption appears to be violated when observing QQ-plots. The one alternative approach that might be reasonable would be to use a parametric approach where the bootstrap was drawn from exponentials distribution.

13. When the stated assumptions for a particular confidence interval are not met, then the stated confidence level does not match the actual confidence level. A C.I. for the variance of a population based on the χ^2 distribution assumes that the sample is drawn from a normal population regardless of the sample size.

    (a) Set the seed to 100. Generate 10,000 samples of size 30 from a normal population with mean 0 and variance 1. For each sample, compute the 95% C.I. for the variance based on the χ^2 distribution,

    $$\left(\frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha/2}}, \frac{(n-1)s^2}{\chi^2_{n-1,\alpha/2}} \right).$$

    What proportion of these intervals should capture the true variance of a N(0,1), and what proportion of these intervals do capture the true variance in the simulation? In other words, what is the nominal coverage, and what is the coverage probability?

```
set.seed(100)

n <- 10000
x <- replicate(n, rnorm(30)) |> as.data.frame()
num <- 0

for (i in 1:n) {
  ci <- ci_sd(x[[i]])^2
  ci <- as.list(ci)
  var <- var(x[[i]])
  if ((ci[[1]] < 1) & (ci[[2]] > 1)) num <- num + 1
}

(coverage_prob1 <- num/n)

## [1] 0.9492
```

    Here we have a 95% nominal coverage, but a 94.92% coverage probability.

    (b) Repeat part (a) but now generate 10,000 samples of size 30 from an exponential population with rate 1.

```
set.seed(100)
x <- replicate(n, rexp(30, rate = 1)) |> as.data.frame()
num <- 0

for (i in 1:n) {
  ci <- ci_sd(x[[i]])^2
  ci <- as.list(ci)
  var <- var(x[[i]])
  if ((ci[[1]] < 1) & (ci[[2]] > 1)) num <- num + 1
}

(coverage_prob2 <- num/n)
```

```
## [1] 0.7178
```

Here we have a 95% nominal coverage, but a 71.78% coverage probability.

(c) Repeat part (b) but now generate 10,000 samples of size 100 from an exponential population with rate 1.

```
set.seed(100)
x <- replicate(n, rexp(100, rate = 1)) |> as.data.frame()
num <- 0

for (i in 1:n) {
  ci <- ci_sd(x[[i]])^2
  ci <- as.list(ci)
  var <- var(x[[i]])
  if ((ci[[1]] < 1) & (ci[[2]] > 1)) num <- num + 1
}

(coverage_prob3 <- num/n)

## [1] 0.6886
```

Here we have a 95% nominal coverage, but a 68.86% coverage probability.

(d) How do the results in parts (a), (b), and (c) compare? In particular, does increasing the sample size when sampling from an exponential distribution help improve the observed coverage?

The interval with the highest observed coverage was with the samples drawn from a normal distribution. The observed coverage dropped substantially when normality was violated, and increasing the sample size in our case actually make the coverage probability worse.