Homework 2

Spring 2023

Jaiden Atterbury

Instructions

- This homework is due in Gradescope on Wednesday April 19 by midnight PST.
- Please answer the following questions in the order in which they are posed.
- Don't forget to knit the document frequently to make sure there are no compilation errors.
- When you are done, download the PDF file as instructed in section and submit it in Gradescope.

Exercises

1. (Simulation) From problem session 1, we saw from our robustness studies that the ratio

$$T = \frac{\sqrt{n}(\bar{X} - \mu_0)}{S}$$

is not well approximated by a t-distribution for small samples taken from $Exp(\lambda_0 = 2)$. Therefore, use of the t confidence interval formula $\bar{X} \pm t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}$ in small samples where the data may have come from an exponential will likely not work very well.

In this problem, you will develop an alternate confidence interval formula for μ_0 which works for small samples. Specifically, suppose $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} Exp(\lambda_0)$. Then it can be shown that (you do not have to show this):

$$\lambda_0\, \bar{X} \sim Gamma(n, n).$$

a. Use the distribution of $\bar{X}$ to construct an *equal tailed* $100(1-\alpha)\%$ confidence interval for the mean $\mu_0 = \frac{1}{\lambda_0}$. Show your work and the formula explicitly. (You may leave q_1 and q_2 in terms of the R function that will be used to calculate them.)

An equal tailed $100(1-\alpha)\%$ confidence interval for the mean $\mu_0 = \frac{1}{\lambda_0}$ can be constructed by manipulating the probability statement $P\left(q_{1-\frac{\alpha}{2}} \leq \lambda_0\bar{X} \leq q_{\frac{\alpha}{2}}\right) = 1-\alpha$, where q_p is the $1-p$th quantile of a $Gamma(n, n)$ distribution. Since $q_{1-\frac{\alpha}{2}} \leq \lambda_0\bar{X} \implies \frac{1}{\lambda_0} \leq \frac{\bar{X}}{q_{1-\frac{\alpha}{2}}}$ and $\lambda_0\bar{X} \leq q_{\frac{\alpha}{2}} \implies \frac{1}{\lambda_0} \geq \frac{\bar{X}}{q_{\frac{\alpha}{2}}}$, we can rewrite the above probability statement as $P\left(\frac{\bar{X}}{\alpha_{\alpha/2}} \leq \frac{1}{\lambda_0} \leq \frac{\bar{X}}{\alpha_{1-\frac{\alpha}{2}}}\right) = 1-\alpha$. Thus it follows that $\left[\frac{\bar{X}}{\alpha_{\alpha/2}}, \frac{\bar{X}}{\alpha_{1-\frac{\alpha}{2}}}\right]$ is a $100(1-\alpha)\%$ confidence interval for the mean $\mu_0 = \frac{1}{\lambda_0}$

b. Use simulations to verify the coverage probability of your interval from part a by following the steps below.

   i. Generate $B = 10,000$ samples each of size 6 from an $Exp(\lambda_0 = 2)$ distribution. Please set the seed at the top of the code chunk to 544.

   ii. Calculate the 95% t confidence interval for each sample. Also calculate your confidence interval from part a.

iii. Calculate and report the coverage rates for the two confidence intervals across the 10,000 samples. Also contrast the intervals in terms of their pattern of non-coverage. This means, when they miss the true mean, is it because the mean lies below the interval? Or above? Or both above and below?

t confidence interval:
As computed through the use of simulation, the coverage rate of the t confidence interval when 10000 resamples of size 6 from an $Exp(\lambda = 2)$ were taken was 0.8885. There was an apparent trend in terms of its pattern of non-coverage, as 1075 of the intervals that didn't cover the true mean were too low, while only 40 of the intervals that didn't cover the true mean were too high.

Gamma confidence interval:
As computed through the use of simulation, the coverage rate of the Gamma confidence interval that we computed in part a. when 10000 resamples of size 6 from an $Exp(\lambda = 2)$ were taken was 0.9541. Unlike the t confidence interval, there was not one side of the interval that the Gamma confidence interval tended to miss the true value of the mean at a higher rate. Instead, the Gamma confidence interval missed the true mean value almost equally on both sides of the interval. In particular, the Gamma confidence interval missed the true mean value 238 times on the high end of the interval, while missing the true mean value 221 times on the low end of the interval.

iv. Perform a large sample significance test to evaluate whether the intervals from part iii. have a nominal coverage rate of 95%. Use $\alpha = 0.01$ to make a decision. (Hint: this is related to problem 1 on HW 1)

Setup:
If we let X_i denote whether the ith confidence interval covers the true value, then the simulated coverage rate is $\bar{X} = \frac{1}{B}\sum_{i=1}^{B} X_i$ where $B = 10000$. Hence it follows that $X_i \sim Binom(n = 1, \pi_0)$, where π_0 is the true coverage rate of the confidence interval. Since both high and low values lead to evidence that the true coverage rate is not 95%, it follows that we are looking at a two sided alternative. Thus our competing hypotheses are: $H_0 : \pi_0 = 0.95$ versus $H_1 : \pi_0 \neq 0.95$. Since $X_i \sim Bernoulli(\pi_0)$ it follows that $E[X_i] = \pi_0$, thus we are concerned about the mean of the distribution. Since $B = 10000$ is large enough for $\bar{X}$ to be approximately normal, we will run a Z test to assess these hypotheses. In particular, 10000 is sufficiently large for the central limit theorem to hold, thus $\bar{X} \approx Norm\left(\pi_0, \sqrt{\frac{\pi_0(1-\pi_0)}{10000}}\right)$.

Furthermore, under the assumption that the null hypothesis is true, $\bar{X} \approx Norm\left(0.95, \sqrt{\frac{0.95 \cdot 0.05}{10000}}\right)$. To find whether the intervals from part iii. have a nominal coverage rate of 95%, we will calculate P-values and use $\alpha = 0.01$ to make a decision.

Test on t interval:
The observed coverage rate of the t confidence interval as computed through simulation and reported above was 0.8885, the P-value associated with this was 3.504×10^{-175}. Thus we have a highly significant result. At the $\alpha = 0.01$ level, we have significant evidence that the coverage rate is not 95% for the t confidence interval when samples of size 6 are drawn from and $Exp(\lambda = 2)$ distribution. In particular, we have evidence that the coverage rate of the t interval under these conditions is less than 95%.

Test on Gamma interval:
The observed coverage rate of the Gamma confidence interval that we found in part a. as computed through simulation and reported above was 0.9541, the P-value associated with this was 0.0599. At the $\alpha = 0.01$ level. we fail to reject the null hypothesis, thus we don't have significant evidence that the true coverage rate of the Gamma confidence interval is anything other than 95%.

Appendix 1

Problem 1 code

```
set.seed(544)

B = 10000

# Code to generate samples of size 6 from Exp(lambda=2) and to calculate 95%
# t-confidence interval and also exact confidence interval from part a.

exp_sim_df <- lapply(X=1:B, FUN=function(X) {
                samp = rexp(n = 6, rate = 2)
                data.frame(t_lower = mean(samp) - qt(p=0.975, df=5) * sd(samp) / sqrt(6),
                           t_upper = mean(samp) + qt(p=0.975, df=5) * sd(samp) / sqrt(6),
                           g_lower = mean(samp) / qgamma(p = 0.975, shape = 6, rate = 6),
                           g_upper = mean(samp) / qgamma(p = 0.025, shape = 6, rate = 6)
)})

exp_sim_stats <- do.call(rbind, exp_sim_df)

# Code to calculate coverage rates of the two intervals for part iii.
true_mean <- 1/2

# Function to find coverage rate and high or low intervals:
find_int_stats <- function(lower, upper) {
  df = data.frame(covg_rate = length(which(lower <= true_mean & upper >= true_mean)) / B,
                  low = length(which(upper < true_mean)),
                  high = length(which(lower > true_mean)))
  as.vector(do.call(rbind, df))
}

# t confidence interval stats:
t_upper <- exp_sim_stats$t_upper
t_lower <- exp_sim_stats$t_lower

t_stats <- find_int_stats(t_lower, t_upper)

t_cov_rate <- t_stats[1]
t_low <- t_stats[2]
t_high <- t_stats[3]

# gamma confidence interval stats:
g_upper <- exp_sim_stats$g_upper
g_lower <- exp_sim_stats$g_lower

g_stats <- find_int_stats(g_lower, g_upper)

g_cov_rate <- g_stats[1]
g_low <- g_stats[2]
g_high <- g_stats[3]

# iv) code to calculate P-value for part iv.
```

```r
# Find P-value for t distribution:
t_pval <- 2 * pnorm(q = t_cov_rate, mean = 0.95, sd = sqrt(0.0475 / 10000))

# Find P-value for Gamma(n, n):
q_pval <- 2 * pnorm(q = g_cov_rate, mean = 0.95, sd = sqrt(0.0475 / 10000), lower.tail = F)
```

2. (Reproducing chi squares) Suppose $X \sim \chi^2_m$ independently of $Y \sim \chi^2_n$. That is, X has PDF

$$f_1(x) = \frac{(1/2)^{m/2}}{\Gamma(m/2)} x^{\frac{m}{2}-1} e^{-x/2}, \quad x > 0$$

and Y has PDF

$$f_2(y) = \frac{(1/2)^{n/2}}{\Gamma(n/2)} y^{\frac{n}{2}-1} e^{-y/2}, \quad x > 0$$

Let $S = X + Y$ be their sum. Show, using the method of convolution, that

$$S = X + Y \sim \chi^2_{m+n}.$$

That is, show that S has PDF

$$f(s) = \frac{(1/2)^{(m+n)/2}}{\Gamma((m+n)/2)} s^{(m+n)/2-1} e^{-s/2} \quad s > 0.$$

You may use without proof that

$$\frac{\Gamma(m)\,\Gamma(n)}{\Gamma(m+n)} = \int\limits_0^1 x^{m-1}\,(1-x)^{n-1}\,dx$$

Hint: Denote the PDF of S by f Recall from chapter 16 that

$$f(s) = \int\limits_0^s f_1(x)\, f_2(s-x)\, dx$$

Plug in for f_1 and f_2 and simplify.

Solution:
Suppose $X \sim \chi^2_m$ independently of $Y \sim \chi^2_n$. As noted above, the PDF of X is $f_1(x) = \frac{(1/2)^{m/2}}{\Gamma(m/2)} x^{\frac{m}{2}-1} e^{-x/2}$, $x > 0$. While the PDF of Y is $f_2(y) = \frac{(1/2)^{n/2}}{\Gamma(n/2)} y^{\frac{n}{2}-1} e^{-y/2}$, $x > 0$. In order to show that their sum $S = X + Y \sim \chi^2_{m+n}$ we will use the convolution formula which states $f(s) = \int\limits_0^s f_1(x)\, f_2(s-x)\, dx$.

Plugging in f_1 and f_2 into the convolution formula gives us the result

$$f(s) = \int\limits_0^s \frac{(1/2)^{m/2}}{\Gamma(m/2)} x^{\frac{m}{2}-1} e^{-x/2} \frac{(1/2)^{n/2}}{\Gamma(n/2)} (s-x)^{\frac{n}{2}-1} e^{-(s-x)/2} dx$$

Since $(1/2)^{m/2}$ and $(1/2)^{n/2}$ have the same base we will add the exponents and obtain $\left(\frac{1}{2}\right)^{\frac{m+n}{2}}$. Furthermore, we will take $\Gamma(m/2)$ and $\Gamma(n/2)$ and put them in the same denominator, this leads to the updated integral

$$f(s) = \frac{\frac{1}{2}^{\frac{m+n}{2}}}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} \int\limits_0^s x^{\frac{m}{2}-1} e^{-x/2} (s-x)^{\frac{n}{2}-1} e^{-(s-x)/2} dx$$

The next part of our simplification is noticing that the term $e^{-(s-x)/2}$ can be simplified to $e^{-s/2+x/2}$, which can further be simplified to $e^{-s/2}e^{x/2}$. This is significant, because the $e^{-x/2}$ and $e^{x/2}$ terms cancel in the integral which can now be rewritten as

$$f(s) = \frac{\frac{1}{2}^{\frac{m+n}{2}}}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} e^{\frac{-s}{2}} \int\limits_0^s x^{\frac{m}{2}-1}(s-x)^{\frac{n}{2}-1}dx$$

The second to last step in our simplification involves a substitution to remove the s in the limits of our integral, this substitution is letting $u = \frac{x}{s} \implies du = \frac{dx}{s} \implies dx = s \cdot du$. Hence the integral becomes $\int\limits_0^1 (su)^{m/2-1}(s-su)^{n/2-1}s\,du$. Factoring out an s from each of the terms simplifies the integral to $s^{(m+n)/2-1}\int\limits_0^1 u^{m/2-1}(1-u)^{n/2-1}du$. The entire equation for $f(s)$ hence becomes

$$f(s) = \frac{\frac{1}{2}^{\frac{m+n}{2}}}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} s^{\frac{m+n}{2}-1} e^{\frac{-s}{2}} \int\limits_0^1 u^{m/2-1}(1-u)^{n/2-1}du$$

The last step in our simplification lies on the fact from above that $\frac{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})}{\Gamma(\frac{m+n}{2})} = \int\limits_0^1 x^{\frac{m}{2}-1}(1-x)^{\frac{n}{2}-1}dx$. Multiplying both sides by the denominator leaves us with $\Gamma(\frac{m}{2})\Gamma(\frac{n}{2}) = \Gamma(\frac{m+n}{2})\int\limits_0^1 x^{\frac{m}{2}-1}(1-x)^{\frac{n}{2}-1}dx$ and hence our equation for $f(s)$ becomes

$$f(s) = \frac{\frac{1}{2}^{\frac{m+n}{2}}}{\Gamma(\frac{m+n}{2})} s^{\frac{m+n}{2}-1} e^{\frac{-s}{2}} \frac{\int\limits_0^1 u^{m/2-1}(1-u)^{n/2-1}du}{\int\limits_0^1 x^{\frac{m}{2}-1}(1-x)^{\frac{n}{2}-1}dx}$$

Notice however that the integrals in each fraction are the same except for the variable of integration. However, this variable of integration is just a "dummy" variable and we could change it to whatever variable we want. With that said, the definite integrals evaluate to the same numerical constant and thus the ratio of the two is just equal to 1. With this taken into account we now have our formula for the PDF of S.

$$f(s) = \frac{\frac{1}{2}^{\frac{m+n}{2}}}{\Gamma(\frac{m+n}{2})} s^{\frac{m+n}{2}-1} e^{\frac{-s}{2}}, \quad s > 0$$

Hence we have shown that the sum of two chi square random variables is a chi square random variable with the degrees of freedom of the two independent chi square random variables added up.

3. (Dark Matter) Two independent research teams claim to have discovered the elusive dark matter. They have used completely independent methods, and completely different statistical tests (although in both cases, rejecting the null hypothesis implies the discovery of dark matter). However, neither group has obtained a significant P-value, achieving 0.06 and 0.08,
respectively. They want to combine their results somehow. Here are two facts:

   - When a null hypothesis is true, P-values follow a $Unif(0,1)$ distribution.
   - If $U \sim Unif(0,1)$ then $-2\ln(U) \sim \chi^2_2$.

   Knowing that the 95th percentile of a χ^2_4 distribution is 9.49, how would you suggest they combine their results?

   Hint: you will need to also use what you learned from problem 2 to define a combined test statistic whose distribution you know under the null hypothesis. Then see whether the observed value of the test statistic provides a contradiction to this distribution.

Setup and recommendation:
In this problem, two independent research teams claim to have discovered dark matter. The researchers have used completely independent methods, and completely different statistical tests. However, both tests regard around the same competing hypotheses: H_0 : Dark matter wasn't discovered versus H_1 : Dark matter was discovered. In both cases rejecting the null hypothesis implies the discovery of dark matter. However, neither research team has obtained a significant P-value, achieving 0.06 and 0.08, respectively.

Since both tests follow the same overarching hypotheses, and both teams found low but insignificant P-values, we need to find a way to combine their results to see if two independent teams finding rare results is significant enough to claim that they found dark matter. However, since they ran two independent and different statistical tests we can't simply add the P-values together. In order to decide what method to use we rely on two key facts. The first fact is, when a null hypothesis is true, P-values follow a $Unif(0,1)$ distribution, and the second fact is if $U \sim Unif(0,1)$ then $-2\ln(U) \sim \chi^2_2$.

With these facts in mind, the research teams should use Fisher's method to combine their P-values. Fisher's method relies on the test statistic $S = -2\sum_{i=1}^{k} \ln(P_i)$ where k is the number of independent P-values and P_i is the ith P-value. In this test, If the p-values from the independent tests tend to be small, the test statistic χ^2 will be large, which would suggests that the null hypotheses are not true for every test, leading us to accept the alternative. Using this test statistic and the results from Problem 2, our test statistic will follow a chi squared distribution. Then using the researchers observed P-values, we can compare our observed value under the null with the $95th$ percentile of a χ^2_4 distribution to see if we contradict the null hypothesis.

Calculation and conclusion:
Let X denote the P-value observed by the first research team, and let Y denote the P-value observed by the second research team, where X is independent of Y. In this case the observed values of our random variables would be $x = 0.06$ and $y = 0.08$. Under the hypotheses described above, and assuming the null hypothesis is true, it follows that $X \sim Unif(0,1)$ and $Y =\sim Unif(0,1)$. Using Fisher's method we obtain the test statistic $S = -2\sum_{i=1}^{2} \ln(P_i) \implies S = -2\ln(\prod_{i=1}^{2} P_i)$ where the P_i's in this case represent the observed P-values of the two researchers.

Using the fact that if $U \sim Unif(0,1)$ then $-2\ln(U) \sim \chi^2_2$, and the fact that the sum of two independent chi square random variables is a chi square random variable with degrees of freedom being the sum of the degrees of freedom of the two independent chi square random variables, we see that $S \sim \chi^2_4$. Our observed value of S is thus $-2\ln(0.06) + -2\ln(0.08) = -2\ln(0.06 \cdot 0.08) = 10.678$. However, the 95th percentile of a χ^2_4 distribution is 9.49. Thus the probability that two independent researchers see something as extreme as these researchers did under the null is 0.0304. Therefore, at the $\alpha = 0.05$ level we contradict our null hypothesis, and conclude that the researchers did in fact find dark matter.

Appendix 2

Problem 3 code

```
fisher_pval <- pchisq(q = -2*log(0.06) + -2*log(0.08), df = 4, lower.tail = F)
```

4. (Blood pressure) An *Arterisonde machine* prints blood-pressure readings on a tape so that the measurement can be read rather than heard. A major argument for using such a machine is that the variability of measurements obtained by different observers on the same person will be lower than the variability with a standard blood-pressure cuff. From previously published work, the variance with a standard blood pressure cuff is $\sigma_0^2 = 35$.

Suppose we have data consisting of systolic blood pressure (SBP) measurements obtained on 10 people and read by two observers. We use the difference between the first and second observers to assess inter-observer variability. In particular, if we assume the underlying distribution of these **differences** is normal with mean μ_0 and variance σ_0^2, then it is of primary interest to make inference about σ_0^2.

The data is in the file `systolic.csv`. Calculate and interpret (in context) a 95% confidence interval for σ_0^2. (Even though investigators think the variability of the new method will be lower, we calculate a two sided confidence interval as the observers are less experienced in using it and this might result in an increase in the variability.)

Create a brief report (of sorts) where you include a description of the data and scientific problem, model/assumptions, the confidence interval, and a conclusion. (Put R code in the appendix.)

Description of the scientific problem:
An Arterisonde machine is a modern tool for obtaining blood-pressure readings, it does so by printing these readings on a tape so that the measurements can be read by observers rather than heard. Another, more standard, tool for obtaining blood pressure is through the use of a blood-pressure cuff. The main scientific problem at hand is that one major argument in support of the Arterisonde machine is that the variability of measurements obtained by different observers on the same person will be lower than the variability of measurements taken with a standard blood-pressure cuff.

Description of the data:
In order to analyze this scientific problem/claim from proponents of the Arterisonde machine, we will use the dataset provided in `systolic.csv`. In this dataset, there are three main variables: patients, measurements from observer 1, and measurements from observer 2. The measurements, in the form of systolic blood pressure (SBP) readings, were obtained on 10 people and read by two observers. The main way in which we can use this data to analyze the claims about the two methods of obtaining blood pressure readings is by using the difference between the measurements taken by the first and second observers to assess the variability between the two observers, then using that data to compare to previously published work which states that the variance with a standard blood pressure cuff is $\sigma_0^2 = 35$.

Description of the model/assumptions:
If we assume that the underlying distribution of the differences between the measurements taken by observer 1 and observer 2 is normal with mean μ_0 and variance σ_0^2, then our main goal is to make inference about σ_0^2. In terms of random variables, if we let X_i denote the ith blood pressure reading from observer 2 and Y_i be the ith blood pressure reading from observer 1, then it follows that $D_i = X_i - Y_i \sim Norm(\mu_0, \sigma_0)$. Under this assumption of the normality of the D_i's, we can use the fact that $\frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2_{n-1}$ where S^2 is the sample variance, to asses the claims made about the Arterisonde machine.

In order to find if there is any significant difference in the variability of the two blood pressure measurement methods, we will use the fact that previously published work has found that the variance with a standard blood pressure cuff is $\sigma_0^2 = 35$. Thus by creating a 95% confidence interval for σ_0^2, the inter-observer variability of the Arterisonde machine, we can test the hypotheses: $H_0 : \sigma_0^2 = 35$ versus $H_1 : \sigma_0^2 \neq 35$ at the $\alpha = 0.05$ significance level. It is important to note that even though investigators think the variability of the Arterisonde machine will be lower, we will calculate a two sided confidence interval as the observers are less experienced in using this newer method which might result in an increase in the variability.

Results and conclusion:
The confidence interval for σ_0^2, the inter-observer variability of the Arterisonde machine, which takes the form $\left[\frac{(n-1)S^2}{\chi^2_{\alpha/2}}, \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}}\right]$, is $[3.870, 27.255]$. Due to the fact that 35 is not in this interval, we reject the null hypothesis that the variability between these two methods is the same. In particular, we have significant evidence at the $\alpha = 0.05$ level that the Arterisonde machine has less inter-observer variability than the traditional blood pressure cuff method.

Appendix 3

Problem 4 code

```r
# Read in the data:
blood_data <- read_csv("systolic.csv")

# Create the difference column:
blood_diff <- blood_data %>%
```

```
                mutate(diff = observer2 - `observer 1`)

# Number of observations:
n <- length(blood_diff$diff)

# Obtain confidence interval:
diff_ci <- tibble(
  lower = (n - 1) * var(blood_diff$diff) / qchisq(p = 0.975, df = n - 1),
  upper = (n - 1) * var(blood_diff$diff) / qchisq(p = 0.025, df = n - 1)
)
```