Homework 2

Spring 2023

KEY

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.)

Let $q_1 = qgamma(\alpha/2, n, n)$ and $q_2 = qgamma(1-\alpha/2, n, n)$.

Thus we know that:

$$P(q_1 \leq \lambda_0 \, \bar{X} \leq q_2) = 1 - \alpha$$

Doing some algebra we have:

$$\begin{aligned} P(q_1 \leq \lambda_0 \, \bar{X} \leq q_2) = & \\ & = P(\frac{q_1}{\bar{X}} \leq \lambda_0 \leq \frac{q_2}{\bar{X}}) \\ & = P(\frac{\bar{X}}{q_2} \leq \frac{1}{\lambda_0} \leq \frac{\bar{X}}{q_1}) \end{aligned}$$

Thus a $100(1-\alpha)\%$ CI will be of the form:

$$\left[\frac{\bar{X}}{q_2}, \frac{\bar{X}}{q_1}\right]$$

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

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?

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)

Students: We only want to see the interpretative portion of your answer to part b in the main document: this means answers to the questions asked in part iii. and iv written as a brief paragraph with the simulated coverage rates and P-value clearly indicated. Your code needs to be shown in an Appendix. Please type your code in the chunk below as given.

iii)

We see from our results that the t interval gave us a coverage rate of 0.8885 and the exponential interval we derived gave us a coverage rate of 0.9541. We also see that in general the t-intervals underestimated the true means. We saw this the proportion of times this occurred was 0.1075. For the exponential interval we derived we saw that about 2.5% of the time the interval overestimated/underestimated. This means the derived interval was not necessarily biased in favor of underestimating or overestimating the mean like we saw with the t interval.

iv)

We can run a test where we have a null hypothesis $H_0 : \pi_0 = .95$ and $H_1 : \pi_0 \neq .95$.

Across our 10,000 samples we can derive a null sampling distribution, using the CLT:

$$\bar{X} \sim N(\mu = .95, \sigma = \sqrt{\frac{.95(1-.95)}{10000}})$$

Where each X_i represents whether the interval covered the true mean or not in each of the 10,000 simulations.

Doing this test we have the following p-values:

```
c("p_value" = p_val_t)
```

```
## p_value.t_coverage
##      3.503646e-175
```

```
c("p_value" = p_val_exp)
```

```
## p_value.exp_coverage
##           0.05994349
```

Thus at a $\alpha = .01$ level, we would reject the null that coverage rate is 95% for the t-interval and fail to reject the null for the derived exponential interval. This makes sense as the assumptions of the t interval are violated when sampling from an exponential distribution, but the exponential interval should have a true coverage rate of .95 .

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

$$\frac{\Gamma(m)\,\Gamma(n)}{\Gamma(m+n)} = \int_0^1 x^{m-1}\,(1-x)^{n-1}\,dx$$

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

$$f(s) = \int_0^s f_1(x)\, f_2(s-x)\, dx$$

Plug in for f_1 and f_2 and simplify.

$$
\begin{aligned}
f(s) &= \int_0^s f_1(x)\, f_2(s-x)\, dx \\
&= \int_0^s \frac{(1/2)^{m/2}}{\Gamma(m/2)} x^{\frac{m}{2}-1} e^{-x/2} \frac{(1/2)^{n/2}}{\Gamma(n/2)} (s-x)^{\frac{n}{2}-1} e^{-(s-x)/2} dx \\
&= \frac{(1/2)^{m/2}}{\Gamma(m/2)} \frac{(1/2)^{n/2}}{\Gamma(n/2)} \int_0^s x^{\frac{m}{2}-1} e^{-x/2} (s-x)^{\frac{n}{2}-1} e^{-(s-x)/2} dx \\
&= \frac{(1/2)^{m/2}}{\Gamma(m/2)} \frac{(1/2)^{n/2}}{\Gamma(n/2)} \int_0^s x^{\frac{m}{2}-1} (s-x)^{\frac{n}{2}-1} e^{-s/2} dx \\
&= \frac{(1/2)^{m/2}}{\Gamma(m/2)} \frac{(1/2)^{n/2}}{\Gamma(n/2)} e^{-s/2} \int_0^s x^{\frac{m}{2}-1} (s-x)^{\frac{n}{2}-1} dx
\end{aligned}
$$

Now let u=x/s, so du = 1/s dx

$$
\begin{aligned}
&= \frac{(1/2)^{m/2}}{\Gamma(m/2)} \frac{(1/2)^{n/2}}{\Gamma(n/2)} e^{-s/2} s \int_0^1 s u^{\frac{m}{2}-1} (s-su)^{\frac{n}{2}-1} du \\
&= \frac{(1/2)^{m/2}}{\Gamma(m/2)} \frac{(1/2)^{n/2}}{\Gamma(n/2)} e^{-s/2} s^{\frac{m}{2}-1} s^{\frac{n}{2}-1} s \int_0^1 u^{\frac{m}{2}-1} (1-u)^{\frac{n}{2}-1} du \\
&= \frac{(1/2)^{m/2}}{\Gamma(m/2)} \frac{(1/2)^{n/2}}{\Gamma(n/2)} e^{-s/2} s^{\frac{m}{2}-1} s^{\frac{n}{2}-1} s \frac{\Gamma(m/2)\,\Gamma(n/2)}{\Gamma(m/2+n/2)} \\
&= \frac{(1/2)^{m/2+n/2}}{\Gamma(m/2+n/2)} e^{-s/2} s^{m/2+n/2-1}
\end{aligned}
$$

Since s>0 clearly, we are done.

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.

The null hypothesis H_0 for both researchers is that dark matter does not exist. When H_0 is true, then we are told $-2\ln(\text{P-value}) \sim \chi^2_2$. What direction is predicted for a P-value when H_0 is not true? Well, small p-values give credence to H_1, therefore let's keep this in mind as we proceed.

If we let P_1 denote the P-value obtained by researcher 1 and P_2 be the P-value for researcher 2. We can suggest that the scientists multiply the two P-values together. Since the tests are independent and share a null hypothesis, the product of the P-values gives a combined assessment of the evidence against the null

hypothesis. For our test statistic, we will use $W = -2\ln(P_1\ P_2)$ instead of just the product since we know the sampling distribution of W under the null.

Specifically, when the null hypothesis is true

$$W = -2\ln(P_1\ P_2) = -2\ln(P_1) - 2\ln(P_2) \sim \chi^2_4$$

We see that if we calculate the value for this test stat we have:

```
w = -2*log(.06*.08)
```

We can calculate the P-value using this combined test statistic by examining the right tail of the χ^2_4 distribution. We only need to examine the right tail since small values of P_1 and P_2 are predicted when the alternative is true. This implies that large values of W are predicted when H_1 is true.

Therefore the combined P-value is

```
round( pchisq(q = w, df = 4, lower.tail = F), 4)
```

```
## [1] 0.0304
```

This gives a combined P-value which is smaller than the individual ones. The act of combining the P-values gives us a greater ability to detect departures from the null hypothesis.

An alternate method is to compare the calculated value of w with the 95th percentile of the null distribution. Since this value is larger than the 95th percentile of a χ^2_4 distribution, this would imply we would reject the null hypothesis that no dark matter exists under the null hypothesis at a 5% level since we have a more extreme value.

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 $\sigma^2_0 = 35$.

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^2_0, then it is of primary interest to make inference about σ^2_0.

The data is in the file `systolic.csv`. Calculate and interpret (in context) a 95% confidence interval for σ^2_0. (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.)

```
systolic <- read.csv("systolic.csv")
systolic <- systolic %>% mutate(diff = observer.1 - observer2)
```

In this case the data represents readings of BP by two observers across 10 patients. We are testing to see whether the differences have a lower variance than a standard blood pressure cuff which has been show to be 35 through various studies.

To conduct this analysis, we will create a 95% CI for the variance of the differences of the readings for these 10 patients. This means that if 35 is included in the interval, we would fail to reject the null hypothesis that Arterisonde machines have the same variability as a standard cuff. If 35 is not in the interval then we would reject the null at a 5% level in favor of the alternative that the machines have different variability than BP cuffs.

In doing this test we are making the key assumption that the BP readings for each observer are normally distributed, so that the differences will be independent and identically distributed according to a normal distribution.

We see that we have a 95% interval for the variance of the differences as:

```
n <- 10
s_2 <- var(systolic$diff)
lower <- (n-1)*s_2/qchisq(1-.05/2,df=n-1)
upper <- (n-1)*s_2/qchisq(.05/2,df=n-1)
c(lower,upper)
```

```
## [1]  3.869048 27.255327
```

Since 35 is not in this CI we reject the null hypothesis that the two methods have the same variability at a 5% level.

Appendix

Problem 1 code

```
set.seed(544)

## code to generate samples of size 6 from Exp(lambda=2)
## and calculate 95% t-confidence interval and also exact confidence
## interval from part a.
B <- 10000
interval_coverage <- function(i){

  x <- rexp(6,2)
  mu <- 1/2

  lower_t <- mean(x) - qt(p = 0.975, df = length(x) - 1) * sd(x) / sqrt(length(x))
  upper_t <- mean(x) + qt(p = 0.975, df = length(x) - 1) * sd(x) / sqrt(length(x))
  under_t <- mu<=lower_t
  over_t <- mu>=upper_t
  cover_t <- (mu<=upper_t)*(mu>=lower_t)

  lower_exp <- (mean(x))/qgamma(1-.05/2,length(x),length(x))
  upper_exp <- (mean(x))/qgamma(.05/2,length(x),length(x))
  under_exp <- mu<=lower_exp
  over_exp <- mu>=upper_exp
  cover_exp <- (mu<=upper_exp)*(mu>=lower_exp)

  return(data.frame("t_coverage" = cover_t,
                    "t_below_interval" = under_t,
                    "t_over_interval" = over_t,
                    "exp_coverage" = cover_exp,
                    "exp_below_interval" = under_exp,
                    "exp_over_interval" = over_exp))
}

sim_coverage <- do.call(rbind, lapply(1:B,interval_coverage)) %>% colMeans()
## code to calculate coverage rates of the two intervals for part iii.

sim_coverage
```

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

p_val_t <- 2*pnorm(sim_coverage[1],mean=.95, sd = sqrt(.95*.05/B))
p_val_exp <-  2*pnorm(sim_coverage[4],mean=.95, sd = sqrt(.95*.05/B),lower.tail=F)
c("p_value" = p_val_t)
c("p_value" = p_val_exp)
```