Homework 6

Point and Interval Estimation

Jaiden Atterbury

Instructions

Please answer the following questions in the order in which they are posed. Add a few empty lines below each and write your answers there. **Focus on answering in complete sentences and show work whether we ask for it or not**. You will also need scratch paper/pen to work out the answers before typing it.

For help with formatting documents in RMarkdown, please consult R Markdown: The Definitive Guide. Another option is to search using Google.

Exercises

1. (Measurement error) A ph-meter is known to have systematic error[1] of size δ_0. In order to estimate δ_0, six measurements are made from a solution with pH **known** to be 4.84.

   The measurement model is that $X_1, X_2, \ldots, X_6$ are independently drawn from a probability distribution with mean $\mu_0 = 4.84 + \delta_0$ and some standard deviation σ_0. In other words, you are being told that

$$E[X] = 4.84 + \delta_0, \quad Var[X] = \sigma_0^2.$$

a. Find the method of moments **estimator** of δ_0. Show your work.

To use the method of moments to find an estimator of δ_0, we must start with the equation $E[X] = \bar{X}$. In our case, this equations equates to $4.84 + \delta_0 = \bar{X}$. Thus, our method of moments estimator for δ_0 is $\hat{\delta}_0^{mom} = \bar{X} - 4.84$ where $\bar{X} = \frac{X_1+X_2+X_3+X_4+X_5+X_6}{6}$.

b. Is your estimator from part a. unbiased for δ_0? Show your work.

In order to see if our estimator from part a. is unbiased for δ_0, we must show that $E[\hat{\delta}_0^{mom}] = \delta_0$.

$$\begin{aligned}
E[\hat{\delta}_0^{mom}] &= E[\bar{X} - 4.84] \quad (\text{Where } \bar{X} = \frac{X_1 + X_2 + X_3 + X_4 + X_5 + X_6}{6}) \\
&= E[\frac{1}{6}\sum_{i=1}^{6} X_i - 4.84] \\
&= \frac{1}{6}\sum_{i=1}^{6} E[X_i] - 4.84 \quad (\text{By linearity of expectation}) \\
&= \frac{1}{6} \cdot 6 \cdot (4.84 + \delta_0) - 4.84 \quad (\text{Since } E[X] = 4.84 + \delta_0) \\
&= \delta_0 \\
&\therefore \hat{\delta}_0^{mom} \text{ is an unbiased estimator of } \delta_0
\end{aligned}$$

[1] this means it gives readings that are systematically higher or lower than what they should be

c. Is your estimator from part a. consistent? Show your work.

In order to see if our estimator from part a. is consistent, we must show that $\lim_{n\to\infty} Var[\hat{\delta}_0^{mom}] = 0$.

$$\begin{aligned}
Var[\hat{\delta}_0^{mom}] &= Var[\bar{X} - 4.84] \quad (\text{Where } \bar{X} = \frac{X_1 + X_2 + X_3 + X_4 + X_5 + X_6}{6}) \\
&= Var[\frac{1}{6}\sum_{i=1}^{6} X_i - 4.84] \\
&= \frac{1}{36} Var[\sum_{i=1}^{6} X_i] \quad (\text{By nonlinearity of variance}) \\
&= \frac{1}{36} \sum_{i=1}^{6} Var[X_i] \quad (\text{By independence of the } X_i\text{'s}) \\
&= \frac{1}{36} \cdot 6 \cdot \sigma_0^2 \quad (\text{Since } Var[X_i] = \sigma_0^2) \\
&= \frac{\sigma_0^2}{6}
\end{aligned}$$

Notice however, that the 6 in $\frac{\sigma_0^2}{6}$ just represents the number of random variables in our sum. If we let this number of random variables in our sum be represented by n, and we take the limit as n goed to infinity we see $\lim_{n\to\infty} Var[\hat{\delta}_0^{mom}] = \lim_{n\to\infty} \frac{\sigma_0^2}{n} = 0$. Thus, our estimator from part a. is consistent.

2. (CLT) Suppose that the time (in days) until a component fails has a gamma distribution with shape $k = 5$ and rate $\lambda = \frac{1}{10}$. When a component fails, it is immediately replaced by a new component. Use the Central Limit Theorem to estimate the probability that 40 components will together be sufficient to last for at least 6 years. (You may assume a year has exactly 365.25 days)

   You may use R to perform the calculations but be sure to set up the problem mathematically first, and show your work and code.

Setup:
Let X denote the time (in days) until a component fails. It follows that $X \sim Gamma(k = 5, \lambda = \frac{1}{10})$. Furthermore, since when a component fails it is immediately replaced by a new component, we can let Y denote the total amount of time (in days) for 40 components to fail. Thus, it follows that $Y = X_1 + X_2 + \cdots + X_{40}$. Since the amount of random variables we are summing is large ($n = 40$), and $40 > 30$, we can use the Central Limit Theorem to approximate the distriubtion of Y. From the Central Limit Theorem we know that $Y \approx Norm(\mu = n\mu_0, \sigma = \sigma_0\sqrt{n})$. In our case, using Theorem 13.2 from the STAT 340 notes which tells us the formula for the expected value and variance of a gamma random variable, we know $E[X] = \frac{k}{\lambda} = \frac{5}{\frac{1}{10}} = 50$ and $Var[X] = \frac{k}{\lambda^2} = \frac{5}{\frac{1}{100}} = 500$. Thus we can plug the values of $E[X]$, $Var[X]$,and n into Y and see that $Y \sim Norm(\mu = 2000, \sigma = \sqrt{20000})$.

In this problem, we want to estimate the probability that 40 components will together be sufficient to last for at least 6 years. However, since we know that a year has exactly 365.25 days, this is exactly the same as calculating the probability that 40 components will together be sufficient to last for at least 2191.5 days. In terms of a probability statement this is the same as trying to find $P(Y \geq 2191.5)$.

Finding P($Y \geq 2191.5$):
Since we know $Y \sim Norm(\mu = 2000, \sigma = \sqrt{20000})$, we can simply calculate $P(Y \geq 2191.5)$ using pnorm in R.

```
prob_of_event <- pnorm(q = 2191.5, mean = 2000, sd = sqrt(20000), lower.tail = F)
```

Thus, the probability that 40 components will together be sufficient to last for at least 2191.5 days, is approximately 0.0878507.

3. The `MIAA05` basketball data contains statistics on 134 players in the MIAA 2005 Men's Basketball season. The following code chooses 100 different samples of size 15 from the dataset. From each sample, the mean and standard deviation of `PTSG` is calculated.

a. From each sample, calculate a 90% confidence interval for the mean PTSG (points per game) of MIAA players. Add the lower and upper limits of the confidence interval as additional columns called `lower` and `upper` in the `sample_summary` dataframe. Then print the first 10 rows of the dataframe. (Show your code and output)

Calculating Confidence Intervals for MIAA PTSG Variable:

In order to find a $100(1-\alpha)\%$ confidence interval we must find the critical z-value that corresponds with it, this equates to finding the $1-\frac{\alpha}{2}$ quantile of a standard normal distribution. Thus in our case, in order to find 90% confidence intervals, we must find the 0.95 quantile of the standard normal, which turns out to be around $z_{\frac{\alpha}{2}} = 1.65$. Our resulting confidence interval will thus be $(\bar{x} - 1.65\frac{s}{\sqrt{15}}, \bar{x} + 1.65\frac{s}{\sqrt{15}})$, where $\bar{x}$ represents the sample mean, and s represents the sample standard deviation.

```
z_crit <- qnorm(p = 0.95)
sample_summary$lower <- sample_summary$sample_mean - z_crit * sample_summary$sample_sd / sqrt(sampsize)
sample_summary$upper <- sample_summary$sample_mean + z_crit * sample_summary$sample_sd / sqrt(sampsize)
```

First 10 Rows of the Dataset:

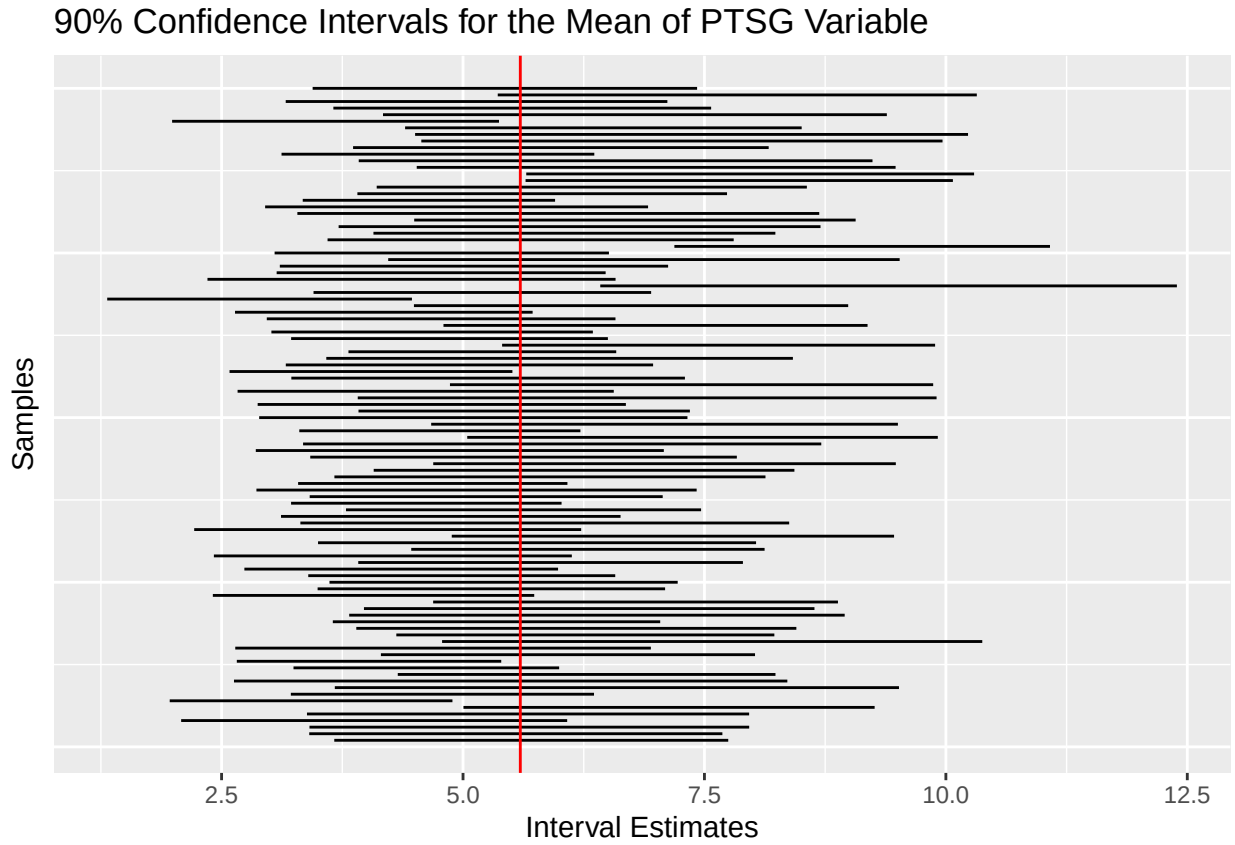
```
head(sample_summary, 10)
```

```
##    sample_mean sample_sd    lower    upper
## 1     5.706667  4.802896 3.666880 7.746454
## 2     5.546667  5.036845 3.407522 7.685812
## 3     5.686667  5.360686 3.409987 7.963347
## 4     4.080000  4.706105 2.081320 6.078680
## 5     5.673333  5.390662 3.383922 7.962744
## 6     7.133333  5.012793 5.004403 9.262263
## 7     3.426667  3.446627 1.962886 4.890447
## 8     4.786667  3.695918 3.217013 6.356321
## 9     6.593333  6.877963 3.672267 9.514400
## 10    5.493333  6.746371 2.628154 8.358513
```

b. The following code represents the confidence intervals you calculated in part a.as horizontal line segments. Fill in the `labs` layer. Also add a vertical line corresponding to the true mean `PTSG` in red. (You will need to calculate this from the `MIAA05` data.)

```
bball_data <- MIAA05
ptsg_mean <- bball_data %>%
  summarise(mean = mean(PTSG))

ggplot(data=sample_summary)+
  geom_segment(mapping = aes(x = lower,
                             xend = upper,
                             y = 1:nsamp,
                             yend = 1:nsamp)) +
  geom_vline(xintercept = ptsg_mean$mean, color = "red") +
  labs(x = "Interval Estimates",
       y = "Samples",
       title = "90% Confidence Intervals for the Mean of PTSG Variable") +
  theme(axis.text.y=element_blank(),
        axis.ticks.y=element_blank())
```



c. Of the confidence intervals you calculated in part a., how many contain the true mean `PTSG`? Write code to calculate this and show your code and answer below.

```
num_intervals <- length(which(sample_summary$lower <= ptsg_mean$mean &
                              ptsg_mean$mean <= sample_summary$upper))
```

The number of confidence intervals created in part a that contain the true mean of the PTSG variable is 91.

d. Suppose you bump up the sample size from 15 to 25. Would you *expect* more intervals to cover the true mean `PTSG`? Why or why not?

Hypothetically, if we were to bump up the sample size from 15 to 25, we would not expect more intervals to cover the true mean of PTSG. The reason we wouldn't expect more intervals to contain the true mean value of the PTSG variable is because the sample size taken doesn't control the amount of intervals we expect to contain the true mean. Instead, this is controlled by the confidence level. In particular, no matter what sample size we use, when we create a 90% confidence interval, over many hypothetical trials/samples we expect 90% of these intervals to contain the true value of the parameter we are estimating. Instead, a larger sample size will result in a tighter confidence interval with a smaller margin of error.

4. Suppose you want to estimate the mean shoe size of adults in a city. You would like to have a 95% confidence interval that is no wider than 0.5 shoe sizes (the margin of error would be at most 0.25). How large a sample must you get?

a. This calculation will require that you make a guess about what approximately the standard deviation will be. What are the implications of guessing too high or too low? Should you guess on the low side or the high side?

What are the implications of guessing too high or too low:
When making a guess of the standard deviation, if we don't guess the value exactly right, there will be

implications that impact our confidence level and width of our interval. In particular, if we guess too high we will get a wider interval and we will have a lower bound of 95% confidence that our interval contains the true mean value. Since our interval got wider than it would've been had we guessed the standard deviation correctly, in order to get our bounds on the margin of error/width of the interval, we would thus need to take a larger sample size to compensate for this extra width in our interval.

On the other hand, if we were to make an incorrect guess of the standard deviation that was too low, we would see the opposite of what we saw in the previous case. In particular, we would get a narrower interval at the cost of a lower confidence level, in this case we will have an upper bound of 95% confidence that our interval contains the true mean value. Thus, we will never reach our intended level of confidence if we underestimate the standard deviation. However, unlike the above case, we won't have to take as large of a sample to get our intended bounds for the margin of error/width of the interval.

Should you guess on the low side or the high side:
After analyzing the implications of guessing too high or too low, if we want to create at least a 95% confidence interval for the mean shoe size of adults in a city with a margin of error of at most 0.25, we would want to guess on the high side. The reason we would want to guess on the high side comes down to what allows us to have the most things in our control. If we were to make a guess that is too low, we would have to no chance that our confidence level would be 95% however, if we were to guess too high, we would know our confidence level is **at least** 95%. Thus, unlike guessing too low, if we guess too high we could still meet all of the criteria of our interval by taking a large enough sample.

b. Should you include men and women in your sample or just one or the other? Why?

Since we are trying to estimate the mean shoe size of adults in a city, and adults consist of both men and women, it only make sense that we contain both in our sample if we truly want to model this phenomenon at the population level. However, if we only sampled men or women exclusively, although we would still be able to construct a 95% confidence interval of the mean with an interval that is no wider than 0.5 shoe sizes, we would not be creating an interval estimate for the mean shoe size of adults in a city. Instead, we would be creating an interval estimate for the mean shoe size for adult males or females in a city, depending on who we decided to include in our sample.

c. Suppose you guess that the standard deviation of the population will be approximately 2. How large must your sample be to get the desired confidence interval?

If we make a guess that the standard deviation of the shoe size of adults in a city will be approximately 2, in order to find how large the sample must be to get a 95% confidence interval that is no wider than 0.5 shoe sizes, we must use the formula from slide 33 of the STAT 341 Interval Estimation notes that states for 95% confidence intervals: $n = \frac{1.96^2\sigma_0^2}{\text{margin of error}^2}$. Thus in order to get a 95% with the desired properties, we must take a sample of size $n = \frac{1.96^2 \cdot 2^2}{0.25^2} = 245.862$ adults.