Homework 6 KEY

Point and Interval Estimation

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.

The method of moments estimator of δ_0 is the value that solves the equation $E[X] = \bar{X}$.

Since we have that $E[X] = 4.84 + \delta_0$ using our method of moments estimator method, we have $\bar{x} = 4.84 + \hat{\delta_0}$. Thus we have the MoM estimator:

$$\hat{\delta_0}^{MoM} = \bar{X} - 4.84$$

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

We have $E[\hat{\delta_0}^{MoM}] = E[\bar{X} - 4.84]$.

Since we know that $E[\bar{X}] = E[X] = 4.84 + \delta_0$, we have $E[\bar{X} - 4.84] = 4.84 + \delta_0 - 4.84 = \delta_0$ where we have used the linearity of expectation to simplify.

Thus the estimator is unbiased.

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

We have that $Var(\hat{\delta_0}^{MoM}) = Var(\bar{X} - 4.84) = Var(\bar{X}) = \frac{\sigma_0^2}{n}$ by Theorem 18.1. We see that as n goes to ∞, $Var(\hat{\delta_0}^{MoM})$ goes to zero. Since the estimator is also unbiased, it is thus 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)

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

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

Let X denote the time until a randomly selected component fails. Then we are given that

$$X \sim Gamma(k_0 = 5, \lambda_0 = \frac{1}{10})$$

where k_0 is the shape parameter and λ_0 the rate parameter. We know from STAT 340 that

$$\begin{aligned}
\mu_0 &= E\left[X\right] \\
&= \frac{k_0}{\lambda_0}, \\
&= 50. \\
\sigma_0^2 &= Var\left[X\right] \\
&= \frac{k_0}{\lambda_0^2}, \\
&= 500.
\end{aligned}$$

Let $S = X_1 + X_2 + \cdots + X_{40}$. We are being asked to calculate $P(S \geq 2191.5)$. From the Central Limit Theorem (Theorem 18.2), assuming n is large:

$$S \quad \approx \quad N(n\mu_0 = 2000, \quad \sqrt{n}\sigma_0 = 141.421).$$

Therefore

$$\begin{aligned}
P(S \geq 2191.5) &= \text{pnorm(2191.5, mean = 2000, sd = 141.421, lower.tail=F)} \\
&= 0.088.
\end{aligned}$$

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.

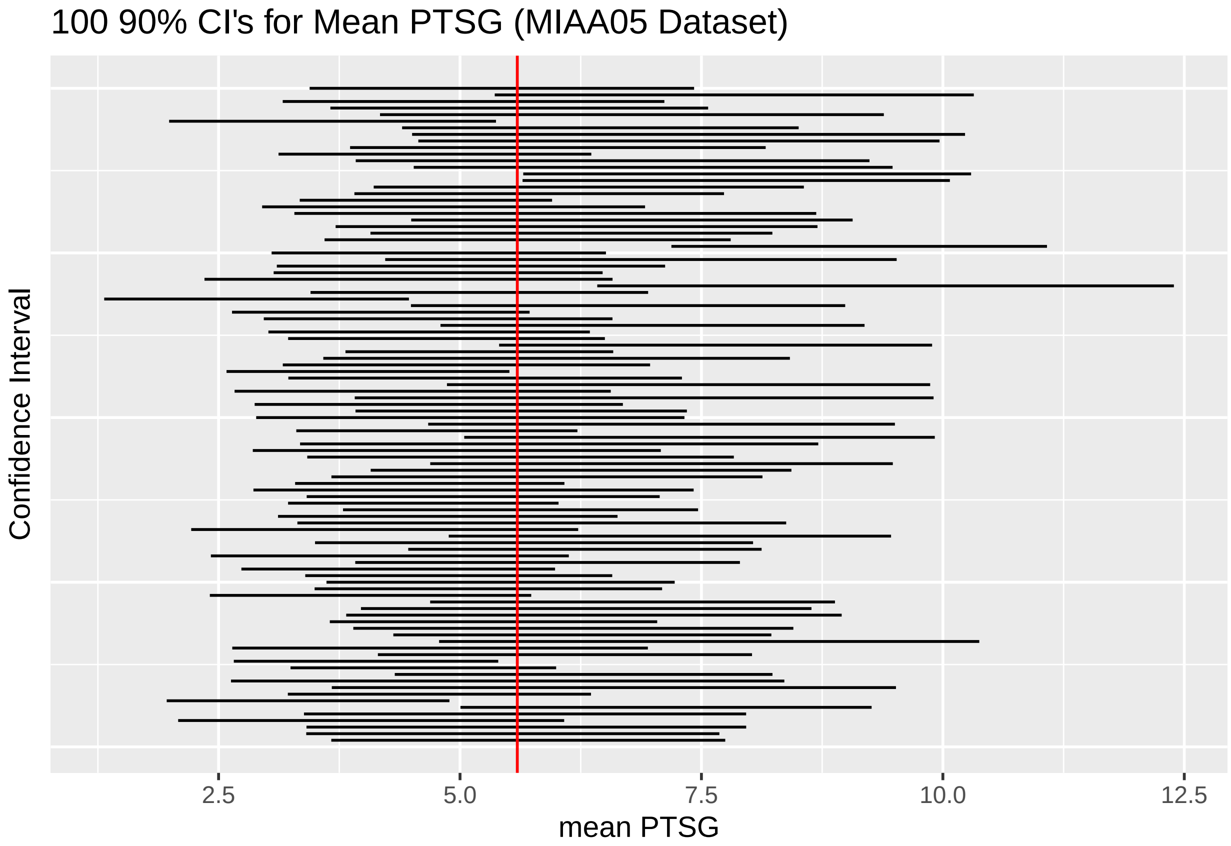
a. From each sample, calculate a 90% confidence interval for the mean PTSG (points per game) of MIAA players. Add the lower and upper limits of the confidence interval as additional columns called `lower` and `upper` in the `sample_summary` dataframe. Then print the first 10 rows of the dataframe. (Show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

```
# I am using the sample SD since we are evaluating how the method does in practice when we don't know t

sample_summary <- sample_summary %>%
  mutate( lower = sample_mean - qnorm(p=0.95)*sample_sd/sqrt(sampsize),
          upper = sample_mean +qnorm(p=0.95)*sample_sd/sqrt(sampsize))

sample_summary %>% slice_head(n = 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.)

```
ggplot(data=sample_summary)+
  geom_segment(mapping = aes(x = lower,
                             xend = upper,
                             y = 1:nsamp,
                             yend = 1:nsamp)) +
  labs(x = "mean PTSG",
       y = "Confidence Interval",
       title = "100 90% CI's for Mean PTSG (MIAA05 Dataset)") +
  theme(axis.text.y=element_blank(),
        axis.ticks.y=element_blank())+
   geom_vline(xintercept = mean(MIAA05$PTSG),col='red')
```



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.

```
sample_summary <- sample_summary %>%
  mutate(
  cover = ifelse(lower < mean(MIAA05$PTSG) & mean(MIAA05$PTSG) < upper, 1,0) )
sample_summary %>% summarize("Coverage Proportion" = mean(cover))
```

```
##   Coverage Proportion
## 1                0.91
```

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?

We would expect the same proportion of coverage (90%), since we should expect 90% of intervals to cover the true mean regardless of sample size. The confidence intervals are constructed in such a way that on average 90% of the intervals will cover the true mean, so even if the intervals will shrink if n increases to 25, the coverage proportion should remain constant.

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?

Based on the formula relating the sample size n to the margin of error for a 95% confidence interval:

$$n \approx= \frac{1.96^2\, \sigma_0^2}{\text{margin of error}^2}$$

we can see that if we use a value of σ_0 which is too high, then we end up with a sample size that is larger than needed for the confidence level we desire.

The reverse holds true when we lowball σ_0. The sample size is then calibrated to have 95% confidence for a smaller value of σ_0. So if the actual σ_0 is larger, then our interval will not have the right confidence level.

Assuming the cost incurred by taking a larger sample is not prohibitive, we should take a larger sample. This way, we achieve our desired confidence level even though it is at the expense of intervals which are much more precise (narrow) than is necessary.

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

Since men and women likely have different shoe size distributions, it is best to consider each separately. It is likely that the distribution of shoe size within each sex is fairly symmetric already, so we would also require a smaller sample size for the Central Limit Theorem to kick in.

However, if for some reason, we really want an interval for **all adults** then we have no choice but to pool all the data with the understanding that we will need a bigger sample to invoke the use of the Central Limit Theorem since the population distribution will likely not be symmetric in that case.

   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?

Using the formula from part a, we need:

$$\begin{aligned} n &\geq \frac{1.96^2 \times 4}{0.25^2} \\ &\geq 245.86. \end{aligned}$$

Therefore in order to have a margin of error of 0.25 we should sample 246 individuals.