Stat 3202 Homework 4

Nathan Johnson

2023-03-01

Due on Carmen Saturday, March 04 before 11:59 pm. All uploads must be .pdf. Submissions will be accepted for 24 hours past this deadline, with a deduction of 1% per hour. Absolutely no submissions will be accepted after this grace period.

Problems will be graded for a combination of correctness and completion.

Academic Integrity

Acceptable for this assignment:

- Working together in a small group
- Working together in a Zoom chat
- Discussing or explaining your approach with students who have also already attempted a problem
- Discussing the assignment with your TA or using MSLC tutoring resources
- Asking me for help in class, office hours, or through Zoom

Not acceptable for this assignment:

- Copying answers from another student or letting another student copy your answers
- Posting solutions to online or communal forums such as a group chat, Chegg, or Stack Exchange
- Using solutions from any previous course or section of Stat 3202

Instructions

- Replace "FirstName LastName (name.n)" above with your information.
- Provide your solutions below in the spaces marked "Solution:".
- Include any R code that you use to answer the questions; if a numeric answer is required, show how you calculated it in R.
- Knit this document to **pdf** and upload both the **pfd** file and your completed **Rmd** file to Carmen
- Make sure your solutions are clean and easy-to-read by
  - formatting all plots to be appropriately sized, with appropriate axis labels.
  - only including R code that is necessary to answer the questions below.
  - only including R output that is necessary to answer the questions below (avoiding lengthy output).

– providing short written answers explaining your work, and writing in complete sentences.

Total: 20 points

Question 1

The Columbus State Department of Ecology provides information on the mean lead contamination in parts per million (ppm) in trout in the Olentangy River basin. The data file `Trout.csv` contains a random sample of $n = 100$ trout with their lead contamination in parts per million.

Please download the `Trout.csv` from Carmen and do the analysis for the following questions.

a)

Supposing the population standard deviation is known to be $\sigma = 0.5$ ppm, and with a large sample size of $n = 100$, please construct a 95% confidence interval for μ, the population mean lead contamination in all trout in the Olentangy River basin. You may use ZTest() or t.test() to verify your answer, but replicate the values mathematically yourself and show your work as well.

b)

Please interpret the confidence interval found in part (a). Explain in both statistical and scientific terms.

c)

Suppose the population standard deviation is unknown and the sample of $n = 100$ trout observations comes from a normal distribution. Please construct a 95% confidence interval for μ, the population mean lead contamination in all trout in the Olentangy River basin. You may use ZTest() or t.test() to verify your answer, but replicate the values mathematically yourself and show your work as well.

d)

Suppose the population standard deviation is unknown and the sample of $n = 100$ trout observations DO NOT come from a normal distribution, would you be able to construct a 95% confidence interval for μ, the population mean lead contamination in all trout in the Olentangy River basin using what you have learnt about constructing confidence intervals. Hint: think of what can you say about the distribution of the statistic $\frac{\bar{x}-\mu}{s/\sqrt{n}}$, where s is the sample standard deviation.

Your answers go here for question 1

Part a:

```r
trout = read.csv("C:\\Users\\natha\\Desktop\\trout.csv")

t.test(trout$lead)
```

```
##
##  One Sample t-test
##
## data:  trout$lead
## t = 20.473, df = 99, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.8855898 1.0756698
## sample estimates:
## mean of x
## 0.9806298
```

```r
mean = mean(trout$lead)
var = 0.5

LowerB = mean - (1.9842 * var / sqrt(100))
UpperB = mean + (1.9842 * var / sqrt(100))

print(LowerB)
```

```
## [1] 0.8814198
```

```r
print(UpperB)
```

```
## [1] 1.07984
```

Since $\alpha = 0.05$ and $n = 100$, the critical t value is 1.9842. So the standard deviation from the sample mean is $1.9842 * 0.5/\sqrt{100}$ where $\sigma = 0.5$ is given to us and $n = 100$ based on the sample size given. So the lower bound is 0.8814198 and the upper bound is 1.07984.

Part b: The confidence interval found in part a shows the true mean of trout with their lead contamination in ppm is between 0.8814198 and 1.07984 with 95% confidence.

Part c:

```r
library(BSDA)
```

```
## Loading required package: lattice

##
## Attaching package: 'BSDA'
```

```
## The following object is masked from 'package:datasets':
##
##     Orange
```

```
trout = read.csv("C:\\Users\\natha\\Desktop\\trout.csv")
mean = mean(trout$lead)
var = var(trout$lead)

z.test(trout$lead, sigma.x = var)
```

```
##
##  One-sample z-Test
##
## data:  trout$lead
## z = 42.744, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.935664 1.025596
## sample estimates:
## mean of x
## 0.9806298
```

```
LowerB = mean - (1.9842 * var / sqrt(100))
UpperB = mean + (1.9842 * var / sqrt(100))

print(LowerB)
```

```
## [1] 0.935108
```

```
print(UpperB)
```

```
## [1] 1.026152
```

Do all the same in part a except $\sigma = 0.5$ is replaced with $s = 0.23$ and the result is a a confidence interval from 0.935108 to 1.026152 of the true mean of trout with their lead contamination in ppm with 95% confidence.

Part d: No, it will not be able to be constructed. The first requirement for a confidence interval is that we are working with a normal distribution. If it were sampling distributions that were not normal, the distribution of the samples would be normal and we could use that. But since it is one sampling distribution that is not normal, it would not apply.

Question 2

a)

Show (or argue technically) that the $100(1-\alpha)\%$ confidence interval for population mean μ

$$\bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

is shorter than

$$\bar{x} - z_{1-\alpha/3}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{1-\alpha/3}\frac{\sigma}{\sqrt{n}}$$

Here, $z_{1-\alpha/2}$ is the $(1-\alpha/2)$ percentile of a standard normal distribution and α is significant level.

b)

Suppose we have a sample of size n from a population with known standard deviation σ. We build a $(1-\alpha)\%$ confidence interval for μ. How would this interval chance under the following circumstances? Explain how or why the change would take place.

- We collect more data (n increases), but everything else stays the same.
- We increase the significance level α, but everything else stays the same.
- The population standard deviation σ were larger, but everything else stays the same.
- The population mean μ were larger, but everything else stays the same.

Your answers go here for question 2 Part a: Change the formulas so that they look like this: $-z_{1-\alpha/2} \leq \frac{\mu-\bar{x}}{\sigma/\sqrt{n}} \leq z_{1-\alpha/2}$ and $-z_{1-\alpha/3} \leq \frac{\mu-\bar{x}}{\sigma/\sqrt{n}} \leq z_{1-\alpha/3}$.

This way, we can look at the alpha level and determine which is smaller or larger.

At $\alpha = 0.5$, $z_{1-\alpha/2} = 1.1503$ and $z_{1-\alpha/2} = 1.5011$. At $\alpha = 0.1$, $z_{1-\alpha/2} = 1.96$ and $z_{1-\alpha/2} = 2.128$.

We can apply this to all values of α because there is no point where $\alpha/3$ will be smaller than $\alpha/2$, so $-z_{1-\alpha/2} \leq \frac{\mu-\bar{x}}{\sigma/\sqrt{n}} \leq z_{1-\alpha/2}$ will be shorter than $-z_{1-\alpha/3} \leq \frac{\mu-\bar{x}}{\sigma/\sqrt{n}} \leq z_{1-\alpha/3}$ since $z_{1-\alpha/2}$ will always be smaller.

Part b: If n increases, the confidence interval would get smaller since the variance would decrease due to $\sigma/\sqrt{n}$ being smaller.

If significance level was increased, the z-score value would get smaller, thus the confidence interval would be smaller.

If the population standard deviation increased, the confidence interval would get larger since the variance is larger.

If the population mean increased, the confidence interval would be the same difference between the lower and upper bound, but it would shift however much the population mean shifted.