STAT22000 Summer 2020 Homework 11 Solutions

Problems to Turn In: due **midnight on Monday, July 20, on Canvas**.

1. If the 95% confidence interval for a parameter μ is $(3.1, 5.1)$ and the 99% confidence interval for μ is $(2.8, 5.4)$. One wants to test the hypotheses H_0: $\mu = 3$ v.s. H_a: $\mu \neq 3$.

   (a) Will H_0: $\mu = 3$ be rejected at significance level $\alpha = 0.05$? Why or why not?

   (b) Will H_0: $\mu = 3$ be rejected at significance level $\alpha = 0.01$? Why or why not?

Answer: [*Give 0pt if the reason is wrong*] A $100(1-\alpha)\%$ CI is equivalent to a two-sided test with significance level α in the sense that μ_0 is in the $100(1-\alpha)\%$ CI if and only if H_0: $\mu = \mu_0$ is NOT rejected at level α, or if and only if the corresponding two-sided P-value exceeds α.

(a) [*1pt*] No. As $\mu = 3$ is not in the 95% CI $(3.1, 5.1)$, we will reject H_0: $\mu = 3$ at level 0.05.

(b) [*1pt*] No. As $\mu = 3$ is in the 99% CI $(2.8, 5.4)$, we will not reject H_0: $\mu = 3$ at level 0.01.

---

2. The P-value for a two-sided test of the null hypothesis H_0: $\mu = 10$ is 0.03.

   (a) Does the 95% confidence interval for μ include 10? Why or why not?

   (b) Does the 99% confidence interval for μ include 10? Why or why not?

Answer: [*Give 0pt if the reason is wrong*]

(a) [*1pt*] No. To see if 10 is in the 95% CI, we need to check if H_0: $\mu = 10$ is rejected at level 0.05. As the two-sided P-value $0.03 < 0.05$, the H_0 is rejected, we know 10 is NOT in the 95% CI.

(b) [*1pt*] No. To see if 10 is in the 99% CI, we need to check if H_0: $\mu = 10$ is rejected at level 0.01. As the two-sided P-value $0.03 > 0.01$, the H_0 is NOT rejected, we know 10 is in the 99% CI.

---

3. It has been hypothesized that allergies result from a lack of early childhood exposure to antigens. If this hypothesis were true, then we would expect allergies to be more common in very hygienic households with low levels of bacteria and other infectious agents. To test this theory, researchers at the University of Colorado sampled the houses of 61 children 9-24 months old and recorded two variables: (1) whether the child tested positive for allergies and (2) the concentration of bacterial endotoxin in the house dust (endotoxin units per ml, EU/ml)[1]. The following are the endotoxin levels at the homes of the 51 children tested negative for allergies.

```
 708.23  911.60  976.81 1316.63  262.74 9772.08  370.76  229.16 2570.51
 891.19 3163.20 1777.65 1288.57  436.23 2631.63 1173.52  911.67 7942.42
 740.32  356.92 1175.48 1480.55 2754.61  575.62  573.89  468.26 1000.71
 364.22 1025.26 1022.04  645.41  363.57  977.47 1022.75 1860.63  371.13
 174.73  399.68 1479.77 2882.96  601.99 1697.32 2291.00  646.49 1176.27
1995.43  955.54 1480.05  456.71 1174.70 5494.22
```

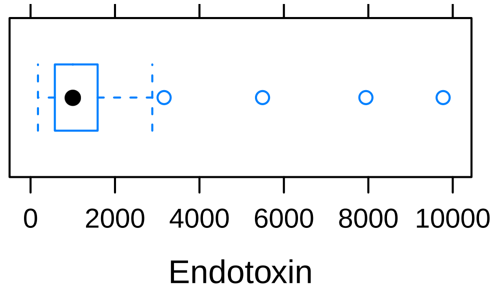
[1] Gereda JE, Leung DYM, Thatayatikom A, Streib JE, Price MR, Klinnert MD, and Liu AH. (2000). Relation between house-dust endotoxin exposure, type 1 T-cell development, and allergen sensitisation in infants at high risk of asthma. *The Lancet*, **355**: 1680-1683.

(a) Make a boxplot for the endotoxin levels at the homes of the 51 children without allergy. Comment on whether it is appropriate to construct a t-confidence interval for the mean endotoxin level at the homes of children without allergy.

```
Endotoxin =
c(708.23,  911.60,  976.81, 1316.63,  262.74, 9772.08,  370.76,  229.16, 2570.51,
 891.19, 3163.20, 1777.65, 1288.57,  436.23, 2631.63, 1173.52,  911.67, 7942.42,
 740.32,  356.92, 1175.48, 1480.55, 2754.61,  575.62,  573.89,  468.26, 1000.71,
 364.22, 1025.26, 1022.04,  645.41,  363.57,  977.47, 1022.75, 1860.63,  371.13,
 174.73,  399.68, 1479.77, 2882.96,  601.99, 1697.32, 2291.00,  646.49, 1176.27,
1995.43,  955.54, 1480.05,  456.71, 1174.70, 5494.22)
library(mosaic)
bwplot(Endotoxin, horizontal=T)
```

Answer:

[2pts = 1pt for the plot + 1 pt for the comment] The distribution of endotoxin levels is severely right-skewed. There are outliers more than 3 or 4 IQRs above Q3 in the normal group. Even when the sample size 51 is moderately large, the one-sample t-CI might not be reliable with such extreme outliers.



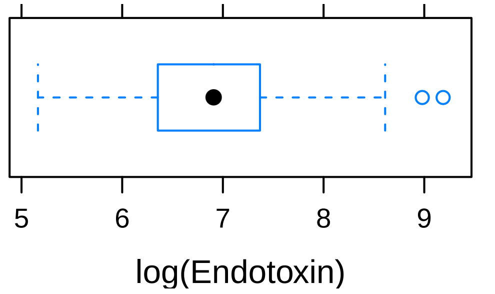
(b) Make a boxplot of the log of the endotoxin levels at the homes of the 51 children without allergy.

```
bwplot(log(Endotoxin), horizontal=T)
```

Comment on whether it is more appropriate to construct a t confidence interval for the mean of the log endotoxin level at the homes of children without allergy.

Answer:

[2pts = 1pt for the plot + 1 pt for the comment] After log-transformation, the distribution of endotoxin levels becomes a lot more symmetric. The two outliers tagged by the 1.5 IQR rule are not as extreme as in the untransformed data. It's more appropriate to use use the one-sample t-CI on log endotoxin levels than on endotoxin levels.



(c) Construct a 95% t-confidence interval for the mean of the log endotoxin level at the homes of children without allergy.

Important Note: *You may find the sample mean and the sample SD of the log endotoxin levels using R, but you are NOT suppose to find the confidence interval using the R function* `t.test`. *Please write down the formula for the confidence interval and show your calculation.*

Answer: *[5pts in total = 2pts for the mean and the SD + 1pt for t^* + 2pts for the CI]* The mean and SD for the log endotoxin levels are 6.9166 and 0.8584 respectively.

```
> mean(log(Endotoxin))
[1] 6.916581
> sd(log(Endotoxin))
[1] 0.8584405
```

With df $= n - 1 = 51 - 1 = 50$, the critical value for 95% CI is $t^* \approx 2.008559$), found in R as follows

```
> qt(0.05/2, df=50, lower.tail=F)
[1] 2.008559
```

The 95% CI is

$$\overline{x} \pm t^* s/\sqrt{n} = 6.9166 \pm 2.008559 \times 0.8584/\sqrt{51} \approx 6.9166 \pm 0.2414 = (6.675, 7.158).$$

---

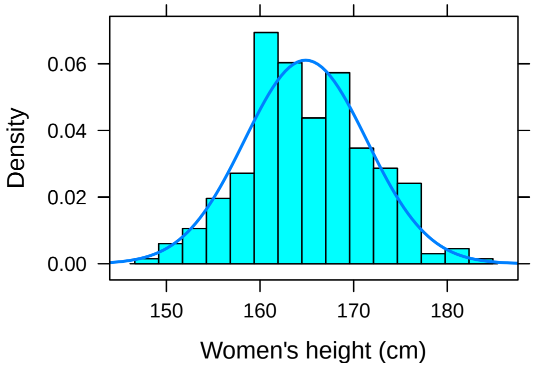
4. Please complete the "On Your Own" part of Lab 7:

`http://www.stat.uchicago.edu/~yibi/s220/labs/lab07.html`

(a) Now use the 260 women in the data set as the population. Make a histogram for the height of the 260 women. Do you find any outlier or clear skewness? Find the population mean and SD of the height of the 260 women.

Answer: [*2pts = 1 pt for the histogram and the comment + 1 pt for the mean and SD*] The histogram of the 260 women's height is roughly symmetric and bell-shaped (at least not clearly skewed), and has no outlier. The population mean is about 164.87 cm and the SD is about 6.54 cm.

```
download.file("http://www.openintro.org/stat/data/bdims.RData", destfile = "bdims.RData")
load("bdims.RData")
library(mosaic)
fdims = subset(bdims, sex == 0)
population = fdims$hgt
histogram(population, fit="normal", nint=15, xlab="Women's height (cm)")
```



```
> mu = mean(population); mu
[1] 164.8723
> sigma = sd(population); sigma
[1] 6.544602
```

(b) Take 100 samples of size 5 from the population of the 260 women. For each of the sample, construct the 3 kinds of confidence intervals ($\overline{x} \pm z^* \sigma/\sqrt{n}$, $\overline{x} \pm z^* s/\sqrt{n}$, and $\overline{x} \pm t^* s/\sqrt{n}$) for the mean height of the 260 women at 90% level. Use the `plot_ci` function in Lab 7 to plot the 3 kinds of confidence intervals. For each of the 3 kinds of intervals, calculate the proportion of intervals that include the true population mean. Are those percentages close to the nominal level 90%?

Answer: *[6pts = 1 pt for 1.645 (or 1.65)+ 1 pt for* $t^* = 2.1318$ *+ 3pts for the plots + 1pt for the comparison.]* R codes for computing and plotting the 100 90% z-intervals $\overline{x} \pm 1.645\,\sigma/\sqrt{n}$, where σ is the population SD:

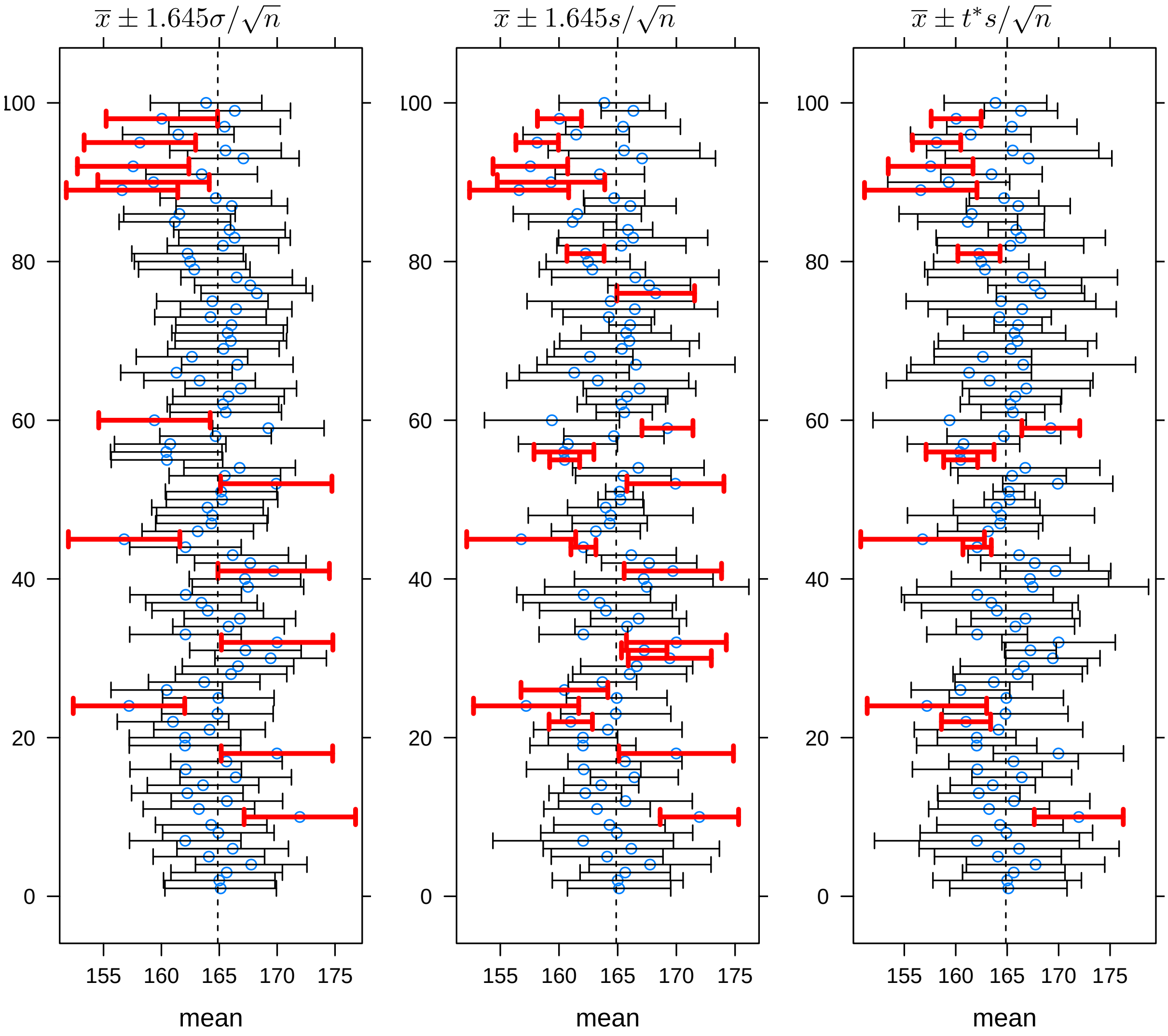
```
samp = do(100) * favstats(sample(population, size = 5))
sigma = sd(population)
samp = transform(samp, lower = mean - 1.645 * sigma/sqrt(5))
samp = transform(samp, upper = mean + 1.645 * sigma/sqrt(5))
plot_ci(samp, mu = mean(population))
```

R codes for computing and plotting the 100 90% z-intervals $\overline{x} \pm 1.645\,s/\sqrt{n}$, where s is the sample SD:

```
samp = transform(samp, lower = mean - 1.645 * sd/sqrt(5))
samp = transform(samp, upper = mean + 1.645 * sd/sqrt(5))
plot_ci(samp, mu = mean(population))
```

R codes for computing and plotting the 100 90% t-intervals $\overline{x} \pm t^* s/\sqrt{n}$, where $t^* \approx 2.1318$ (or 2.13) for a 90% CI is obtained in R as `qt(0.1/2, df=4, lower.tail=F)`

```
samp = transform(samp, lower = mean - 2.1318 * sd/sqrt(5))
samp = transform(samp, upper = mean + 2.1318 * sd/sqrt(5))
plot_ci(samp, mu = mean(population))
```



13 CIs miss μ 87% cover μ

22 CIs miss μ 78% cover μ

13 CIs miss μ 87% cover μ

[The plots and percentages covering μ may vary.]

Conclusion: From the simulation above, we see the z-interval with known population SD σ $\overline{x} \pm 1.645\sigma/\sqrt{n}$ and the t-interval $\overline{x} \pm t^* s/\sqrt{n}$ have coverage probabilities close to the nominal 90% level, but the coverage probability of the z-interval with unknown population SD $\overline{x} \pm 1.645s/\sqrt{n}$ is substantially lower than the nominal 90% level.