Homework #5

Derek Sonderegger

1

An experiment is conducted to examine the susceptibility of root stocks of a variety of lemon trees to a specific larva. Forty of the plants are subjected to the larvae and examined after a fixed period of time. The response of interest is the logarithm of the number of larvae per gram of root stock. For these $n = 40$ plants, the sample mean is $\bar{x} = 11.2$ and the sample standard deviation is $s = 1.3$. Use these data to construct a 90% confidence interval for μ, the mean susceptibility of lemon tree root stocks from which the sample was taken.

$$\bar{x} \pm t_{n-1}^{1-\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$

$$11.2 \pm t_{39}^{0.95}\left(\frac{1.3}{\sqrt{40}}\right)$$

```
qt(0.95, df=39)
```

```
## [1] 1.684875
```

$$11.2 \pm 1.68(0.206)$$

$$11.2 \pm 0.345$$

2

A social worker is interested in estimating the average length of time spent outside of prison for first offenders who later commit a second crime and are sent to prison again. A random sample of $n = 100$ prison records in the count courthouse indicates that the average length of prison-free life between first and second offenses is 4.2 years, with a standard deviation of 1.1 years. Use this information to construct a 95% confidence interval for μ, the average time between first and second offenses for all prisoners on record in the county courthouse.

$$\bar{x} \pm t_{n-1}^{1-\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$

$$4.2 \pm t_{99}^{0.975}\left(\frac{1.1}{\sqrt{100}}\right)$$

```
qt(0.975, df=99)
```

```
## [1] 1.984217
```

$$4.2 \pm 1.984\,(0.11)$$

$$4.2 \pm 0.21$$

3.

A biologist wishes to estimate the effect of an antibiotic on the growth of a particular bacterium by examining the number of colony forming units (CFUs) per plate of culture when a fixed amount of antibiotic is applied. Previous experimentation with the antibiotic on this type of bacteria indicates that the standard deviation of CFUs is approximately 4. Using this information, determine the number of observations (i.e. cultures developed) necessary to calculate a 99% confidence interval with a half-width of 1.

```
qnorm(0.995)
```

```
## [1] 2.575829
```

$$
\begin{aligned}
n &= \left[z_{1-\alpha/2}\left(\frac{\sigma}{ME}\right)\right]^2 \\
&= \left[z_{0.995}\left(\frac{4}{1}\right)\right]^2 \\
&= [2.58(4)]^2 \\
&= [10.32]^2 \\
&= 106.5 \qquad \text{so consider } n = 107
\end{aligned}
$$

4.

In the R package Lock5Data, the dataset `FloridaLakes` contains information about the mercury content of fish in $n = 53$ Florida lakes. For this question, we'll be concerned with the average ppm of mercury in fish from those lakes which is encoded in the column `AvgMercury`.

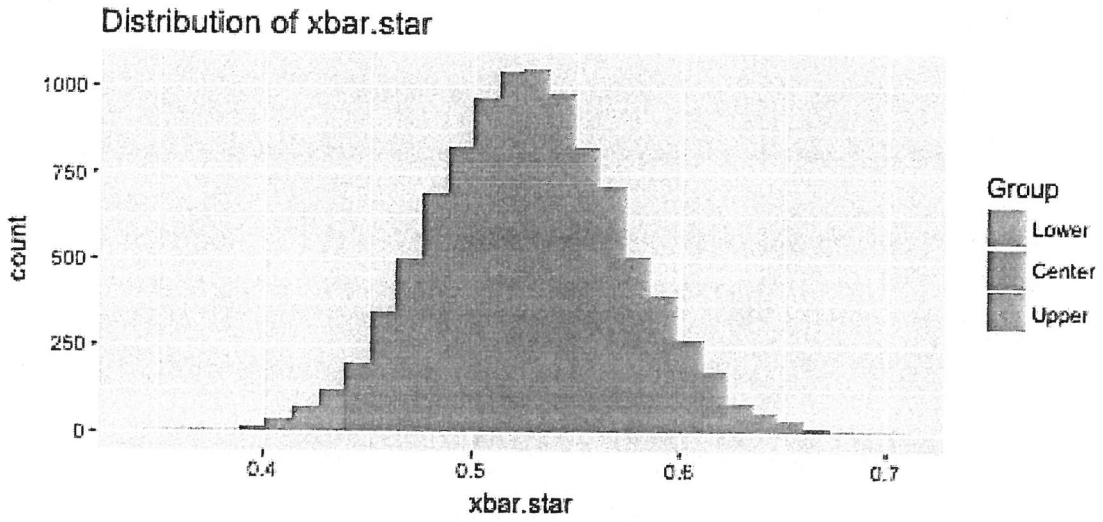
(a) Using the bootstrapping method, calculate a 95% confidence interval for μ, the average ppm of mercury in fish in all Florida lakes.

```
data(FloridaLakes)
bootDist <- do(10000)*{
  resample(FloridaLakes) %>% summarise(xbar.star=mean(AvgMercury))
}
bootDist %>% summarise( lwr=quantile(xbar.star, probs=c(0.025)),
                        upr=quantile(xbar.star, probs=c(0.975)))
```

```
##         lwr       upr
## 1 0.4369811 0.6192453
```

```
bootDist <- bootDist %>%
  mutate(Group = cut(xbar.star,
                     breaks=c(-Inf,quantile(xbar.star,c(0.025, 0.975)),Inf),
                     labels=c('Lower','Center','Upper')))
ggplot(bootDist, aes(x=xbar.star, fill=Group)) +
  geom_histogram() +
  ggtitle('Distribution of xbar.star')
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



(b) Using the asymptotic approximations discussed in this chapter, calculate a 95% confidence interval for μ, the average ppm of mercury in fish in all Florida lakes.

```
FloridaLakes %>% summarise( xbar = mean(AvgMercury),
                            s    = sd(AvgMercury))
```

```
##        xbar         s
## 1 0.5271698 0.3410356
```

$$\bar{x} \pm t_{n-1}^{1-\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$

$$\bar{x} \pm t_{52}^{0.975}\left(\frac{s}{\sqrt{53}}\right)$$

$$0.527 \pm 2.007\left(\frac{0.34}{\sqrt{(53)}}\right)$$

$$0.527 \pm 0.0937$$

$$(0.433,\ 0.621)$$

(c) Comment on the similarity of these two intervals.

These two intervals are very close... both about 0.43 to 0.62. This shouldn't be too surprising because the bootstrap distribution of $\bar{x}^*$ looks pretty normal so either method is appropriate.

5.

In the R package `Lock5Data`, the dataset `Cereal` contains nutrition information about a random sample of $n = 30$ cereals taken from an on-line nutrition information website (see the help file for the dataset to get the link). For this problem, we'll consider the column `Sugars` which records the grams of sugar per cup.

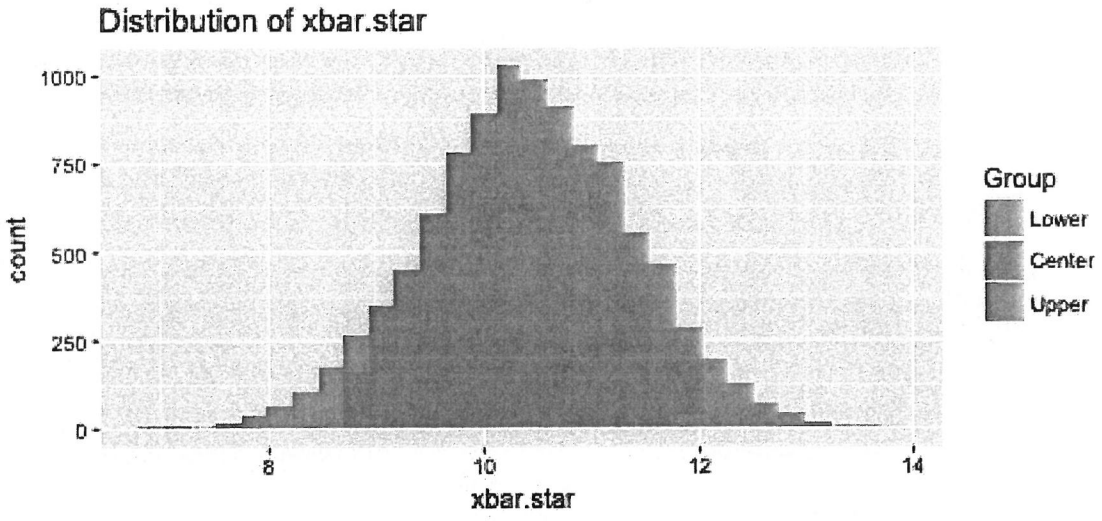
(a) Using the bootstrapping method, calculate a 90% confidence interval for μ, the average grams of sugar per cup of all cereals listed on this website.

```
data(Cereal)
bootDist <- do(10000)*{
  resample(Cereal) %>% summarise(xbar.star=mean(Sugars))
}
bootDist %>% summarise( lwr=quantile(xbar.star, probs=c(0.05)),
                        upr=quantile(xbar.star, probs=c(0.95)))
```

```
##        lwr      upr
## 1 8.786667 11.97667
```

```
bootDist <- bootDist %>%
  mutate(Group = cut(xbar.star,
                     breaks=c(-Inf,quantile(xbar.star,c(0.05, 0.95)),Inf),
                     labels=c('Lower','Center','Upper')))
ggplot(bootDist, aes(x=xbar.star, fill=Group)) +
  geom_histogram() +
  ggtitle('Distribution of xbar.star')
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



(b) Using the asymptotic approximations discussed in this chapter, calculate a 90% confidence interval for μ the average grams of sugar per cup of all cereals listed on this website.

```
Cereal %>% summarise( xbar = mean(Sugars),
                      s    = sd(Sugars))
```

```
##    xbar        s
## 1 10.42 5.331843
```

$$\bar{x} \pm t_{n-1}^{1-\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$

$$\bar{x} \pm t_{29}^{0.95}\left(\frac{s}{\sqrt{30}}\right)$$

$$10.42 \pm 1.70\left(\frac{5.33}{\sqrt{(30)}}\right)$$

$$10.42 \pm 1.65$$

(8.77, 12.07)

(c) Comment on the similarity of these two intervals.

These two intervals are very close... both about 8.8 to 12.0. This shouldn't be too surprising because the bootstrap distribution of $\bar{x}^*$ looks pretty normal so either method is appropriate.

(d) We could easily write a little program (or pay an undergrad) to obtain the nutritional information about all the cereals on the website so the random sampling of 30 cereals is unnecessary. However, a bigger concern is that the website cereals aren't representative of cereals Americans eat. Why? For example, consider what would happen if we added 30 new cereals that were very nutritious but were never sold.

The issue is that because the cereals are not representative of what is actually sold, we don't know if Americans actually only eat the most sugary (Raisin Bran!) or if they eat a mix of the other ones. Furthermore, there are a bunch of cereals that are less popular and not on this web site and and we don't know where they fall on the sugar scale.