Andrew Moore, 10/11/2021

MATH-471, Homework 4

1a.

We have a sample of data $\underline{\text{x}}$ = 6, 4, 8, 4, 4, 4, 2, 5, 5, 7. The sample size is 10.

We are told that the data represent the nightly count of animals caught in Alex's trap. Based on the description, we will assume the data comes from a *Poisson* distribution. The probability mass function for the Poisson distribution is

$$f(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}, \text{ for } x = 0, 1, 2, ... \text{ and } \lambda > 0.$$

1b.

For the Poisson distribution, both the mean and variance are controlled by the same parameter, λ. To estimate the mean and variance, we will maximize the likelihood function, defined as

$$L(\lambda|x_1, x_2, ..., x_{10}) = \prod_{i=1}^{10} f(x_i|\lambda) = \prod_{i=1}^{10} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}.$$

Taking the natural log of L, we have

$$\begin{aligned} lnL(\lambda|x_1, x_2, ..., x_{10}) &= \sum_{i=1}^{10} ln\frac{e^{-\lambda}\lambda^{x_i}}{x_i!} \\ &= \sum_{i=1}^{10} \big[-\lambda + x_i ln\lambda - ln(x_i!) \big] \\ &= -\sum_{i=1}^{10} \lambda + \sum_{i=1}^{10} x_i ln\lambda - ln(x_i!) \\ &= -10\lambda + \sum_{i=1}^{10} x_i ln\lambda - ln(x_i!). \end{aligned}$$

Differentiating lnL results in

$$\frac{\partial}{\partial\lambda}\big[lnL(\lambda|x_1, x_2, ..., x_{10})\big] = \frac{\partial}{\partial\lambda}\big[-10\lambda + \sum_{i=1}^{10} x_i ln\lambda - ln(x_i!)\big] = -10 + \frac{\sum_{i=1}^{10} x_i}{\lambda}.$$

We can test whether this point is the global maximum by using the 2nd-derivative test. Taking the second derivative of lnL we have

$$\frac{\partial^2}{\partial^2\lambda}\big[ln(L(\lambda|x_1, x_2, ..., x_{10}))\big] = \frac{\partial}{\partial\lambda}\big[-10 + \frac{\sum_{i=1}^{10} x_i}{\lambda}\big] = \frac{-\sum_{i=1}^{10} x_i}{\lambda^2}$$

This result is negative, and will be negative for all λ (because, by definition, $\lambda > 0$). This means that we have found the maximum.

Setting the equation to 0, and solving for $\hat{\lambda}$ we have

$$\begin{aligned}
0 &= -10 + \frac{\sum_{i=1}^{10} x_i}{\hat{\lambda}} \\
10 &= \frac{\sum_{i=1}^{10} x_i}{\hat{\lambda}} \\
10\hat{\lambda} &= \sum_{i=1}^{10} x_i \\
\hat{\lambda} &= \frac{\sum_{i=1}^{10} x_i}{10} \\
\hat{\lambda} &= \frac{49}{10} = 4.9.
\end{aligned}$$

Thus, our estimate for the parameter is $\hat{\lambda} = 4.9$, i.e., $\mu = 4.9$ and $\sigma^2 = 4.9$.

5.1

The county government in a city that is dominated by a large state university is concerned that a small subset of its population has been overutilized in the selection of residents to serve on county court juries. The county decides to determine the mean number of times that an adult resident of the county has been selected for jury duty during the past 5 years. They will then compare the mean jury participation for full-time students to that of nonstudents.

a. Identify the populations of interest to the county officials.

Full-time students versus non-students.

b. How might you select a sample of voters to gather this information?

The county government could request lists of students who were enrolled full-time over the past 5 years. It could then compile lists of all residents recorded as living in the county over the past 5 years. After removing duplicates from each list, the county could then compare its list of county-residents against the list of students. Once the students had been removed from the list of county residents, the government would have their two populations available for sampling. They could then draw a random sample of an appropriate size from each.

5.5

A company that manufacturers coffee for use in commercial machines monitors the caffeine content in its coffee. The company selects 50 samples of coffee every hour from its production line and determines the caffeine content. From historical data, the caffeine content (in milligrams, mg) is known to have a normal distribution with $\hat{\sigma} = 7.1$ mg. During a 1-hour time period, the 50 samples yielded a mean caffeine content of $\bar{y} = 110$ mg.

a. Identify the population about which inferences can be made from the sample data.

The population is the universe of possible n = 50 samples drawn from the production line.

b. Calculate a 95% confidence interval for the mean caffeine content μ of the coffee produced during the hour in which the 50 samples were selected.

```r
n <- 50
s <- 7.1
ybar <- 110

# Margin of Error at 95% confidence
moe <- 1.96 * s / sqrt(n)

(ci <- ybar + c(c(-1, 1) * moe))
```

```
## [1] 108.032 111.968
```

c. Explain to the CEO of the company in nonstatistical language the interpretation of the constructed confidence interval.

Based on the sample collected, we would estimate that the average level of caffeine is between 108.03 and 111.97. If we repeated this experiment 100 times, and computed an interval for each in the same fashion, approx. 95 of the intervals would contain the true average.

5.6

Refer to Exercise 5.5. The engineer in charge of the coffee manufacturing process examines the confidence intervals for the mean caffeine content calculated over the past several weeks and is concerned that the intervals are too wide to be of any practical use. That is, they are not providing a very precise estimate of μ.

a. What would happen to the width of the confidence intervals if the level of confidence of each interval is increased from 95% to 99%?

The interval width would increase.

b. What would happen to the width of the confidence intervals if the number of samples per hour was increased from 50 to 100?

The interval width would decrease.

5.8

As part of the recruitment of new businesses, the city's economic development department wants to estimate the gross profit margin of small businesses (under $1 million in sales) currently residing in the city. A random sample of the previous years annual reports of 15 small businesses shows the mean net profit margin to be 7.2% (of sales) with a standard deviation of 12.5%.

a. Construct a 99% confidence interval for the mean gross profit margin of μ of all small businesses in the city.

We are asked to compute a 99% confidence interval for the sample of 15 businesses, but are not given indication as to what distribution the data should be modeled by. This means that the Central Limit Theorem for the sample mean does not apply.

For the sake of the exercise, we will assume the data does come from a normal distribution, and use a T distribution as the distribution for our pivot.

```
n <- 15
s <- 0.125
ybar <- 0.072

tval <- qt(1 - 0.01 / 2, n - 1)

(ci <- ybar + c(c(-1, 1) * (tval * s / sqrt(n)))) * 100
```

```
## [1] -2.407719 16.807719
```

b. The city manager reads the report and states that the confidence interval for μ constructed in part (a) is not valid because the data are obviously not normally distributed and thus the sample size is too small. Based on just knowing the mean and standard deviation of the sample of 15 businesses, do you think the city manager is valid in his conclusion about the data? Explain your answer.

I would agree with the city manager, and would recommend a sample size of at least 30 businesses if we we want to rely on the Central Limit Theorem.

5.10

The susceptibility of the root stocks of a variety of orange tree to a specific larva is investigated by a group of researchers. Forty orange trees are exposed to the larva and then examined by the researchers 6 months after exposure. The number of larvae per gram is recorded on each root stock. The mean and standard deviation of the logarithm of the counts are recorded to be 9.02 and 1.12, respectively.

a. Use the sample information to construct a 90% confidence interval on the mean of the logarithm of the larvae counts.

```
n <- 40
s <- 1.12
ybar <- 9.02
```

```
zval <- -qnorm((1 - 0.9) / 2)

(ci <- ybar + c(c(-1, 1) * (zval * s / sqrt(n))))
```

```
## [1] 8.728717 9.311283
```

b. Identify the population for which this confidence interval could be used to assess the susceptibility of the orange trees to the larva.

This could be used to assess a Normal/LogNormal distribution.

5.14

The housing department in a large city monitors the rent for rent-controlled apartments in the city. The mayor wants an estimate of the average rent. The housing department must determine the number of apartments to include in a survey in order to be able to estimate the average rent to within $100 using a 95% confidence interval. From past surveys, the monthly charge for rent-controlled apartments ranged from $1,000 to $3,500. How many renters must be included in the survey to meet the requirements?

```
s <- (3500 - 1000) / 4
moe <- 100
zval <- 1.96

# calculating N
zval^2 * s^2 / moe^2
```

```
## [1] 150.0625
```

150 renters should be included in the survey.

5.15

Refer to Exercise 5.14. Suppose the mayor's staff reviews the proposed survey and decides that in order for the survey to be taken seriously the requirements need to be increased.

a. If the level of confidence is increased to 99% with the average rent estimated within $50, how many apartments need to be included in the survey?

```
s <- (3500 - 1000) / 4
moe <- 50
zval <- 2.575

# calculating N
zval^2 * s^2 / moe^2
```

```
## [1] 1036.035
```

1036 renters should be included.

b. Suppose the budget for the survey will not support increasing the level of confidence to 99%. Provide an explanation to the mayor, who has never taken a statistics course, of the impact on the accuracy of the estimate of the average rent of not raising the level of confidence from 95% to 99%.

Increasing the level of confidence to 99% requires a sample size 10 times as large as that required for 95%, which is unfeasible based on the project budget. However, even if the resources were available, a 95% confidence level should provide an adequate level of accuracy. If we were to repeat our data collection and estimation procedures 100 times, 95 of the collections/estimations would provide an interval that contains the true average.