STAT 422 Final Project

Cassie Noble, Kendra Philipsek, Lisa Bowersock

The dataset we have chosen for this project was collected by the Montana Early Childhood Project Practitioner Registry. The Practitioner Registry is a voluntary database of childcare providers and individuals working in childcare settings in the state of Montana. Information collected by the Early Childhood Project, including this dataset, is publicly available upon request. Our sample consists of 461 early childhood teachers who currently work in childcare programs with children ages infant to elementary school. This sample represents about one quarter of the total population of early childhood teachers in the state of Montana. The random variable of interest is the hourly wage (in U.S. dollars) for early childhood teachers in Montana. Wage information was collected through voluntary self-report and subsequently verified by employers.

Looking at the hourly wages of early childhood care teachers, we could reasonably assume the data follows a normal distribution because hourly wage is a continuous random variable. However, due to the observance of the federal minimum wage and the high proportion of childcare providers who earn a salary close to the minimum wage, the distribution of our dataset shows heavy right skew versus a normal distribution. Therefore, the normal distribution does not seem appropriate for this data. The histogram of the data shows a shape that could resemble a chi-squared, gamma, or beta distribution. A beta distribution does not seem appropriate because we are not dealing with proportions, that is the wages do not fall between 0 and 1. A chi-squared or gamma distribution also does not make sense in this situation because we are not looking at times until an event occurs or time between events. Therefore, this dataset does not seem to reasonably follow any established distribution.

While investigating this dataset, we noticed that the hourly wage for childcare providers in Montana seemed quite low. The U.S. Department of Health and Human Services lists the 2018 poverty

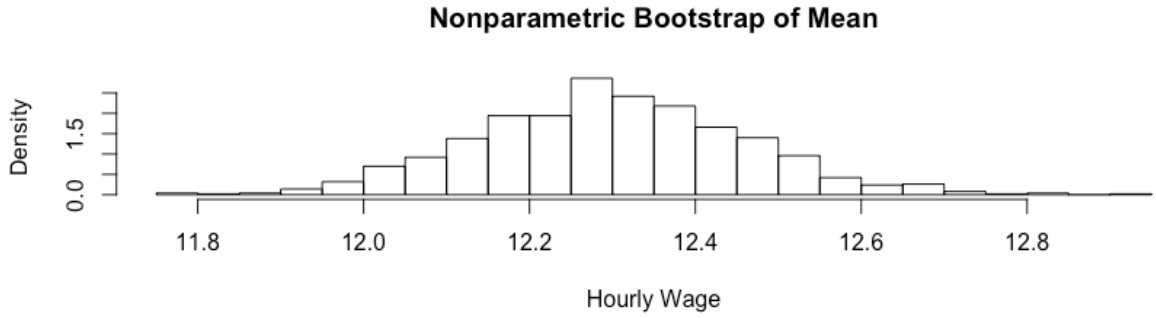
guideline for a family of four as an annual wage of $25,100, which translates to an hourly wage of $12.07 for a full-time employee. Based off this guideline, the research question we have chosen to explore is if the average Montana child care provider makes enough money to be considered above the national poverty guideline. This question references the underlying characteristic of the average hourly wage of the population of child care providers in Montana. The answer to this question would be of great interest to anyone who is considering a career as a child care provider in Montana. These individuals would want to know if this line of work would be monetarily sufficient to support their family. Also, this could be an important piece of information related to the quality of childcare providers in Montana; if the average child care provider in the state is living below the national poverty line, then there is less incentive for quality employees to choose this profession in Montana.

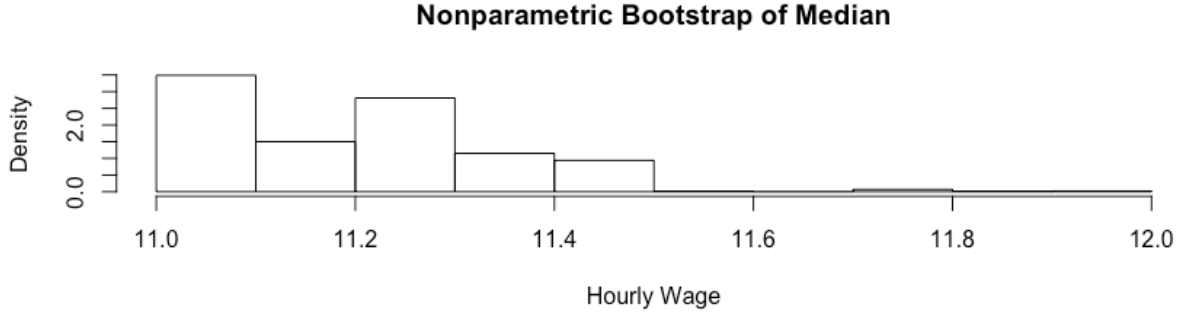
To help us answer this question, there are two estimators which could be used to estimate the average hourly wage of the population of childcare providers in Montana; the mean and the median of our dataset. The mean hourly wage of our dataset is a clear contender to estimate the average hourly wage of the population of all childcare providers in Montana as our dataset contains a fairly large portion of the wages of child care providers in the state. However, sometimes the median value is more appropriate for getting a sense of the average of a dataset if there is heavy skew in the data; a small number of extreme outliers are not always appropriate to include in estimates of averages. Since the distribution of our data set does show right skew, the median might be a better estimator to use. We explored both of these potential estimators for the population average hourly wage and decided which one should be used to answer our research question using the bootstrap method.

The calculated mean and median hourly wage from our data set are $12.30 and $11.15 per hour, respectively. We performed a non-parametric bootstrap of both the mean and median hourly wage. The figures below show the distribution of the mean and median bootstrap. The bootstrap distribution of the sample mean hourly wage appears normally distributed with a mean of $12.30 and a

standard deviation of 0.16. We are 95% confident that the true mean is between $11.99 and $12.64. The median of the bootstrap distribution of the sample mean hourly wage is also $12.30, showing that the bootstrap distribution of the sample mean has the same center as our sample dataset. The interquartile range of the bootstrap distribution of the sample mean is 0.21 and the estimated bias of the sample mean is 0.

The bootstrap distribution of the sample median hourly wage is skewed to the right, with a mean of $11.19, median of $11.25, and standard deviation of 0.17. We are 95% confident that the true median is between $11.00 and $11.50. The interquartile range of the bootstrap distribution of the sample median is 0.25. The estimated bias of the sample median is 11.19-11.15 = 0.04.





When choosing a good estimator, we need to compare the bias and the variance of the possible estimators. In our analysis we found that the bootstrap distribution for the sample mean has a smaller

variance than the bootstrap distribution for the sample median (0.026 compared to 0.029). The bootstrap distribution for the sample mean was also unbiased compared to a bias of 0.04 for the median. Since the sample mean is unbiased and has a smaller variance, it is clearly the better estimator.

Using our simulated distribution of the sample mean, we found the probability of observing a value greater than or equal to the null hypothesis value of $12.07. This probability came out to 0.916, meaning that there is a 91.6% chance that the mean hourly wage is greater than or equal to $12.07. As this is a very high probability of success, we would fail to reject the null hypothesis that the average Montana child care provider makes at least enough money to be considered above the national poverty guideline.

References:

Montana Early Childhood Project Practitioner Registry. Retrieved April 4, 2018.

https://www.mtecp.org/practitioner.html

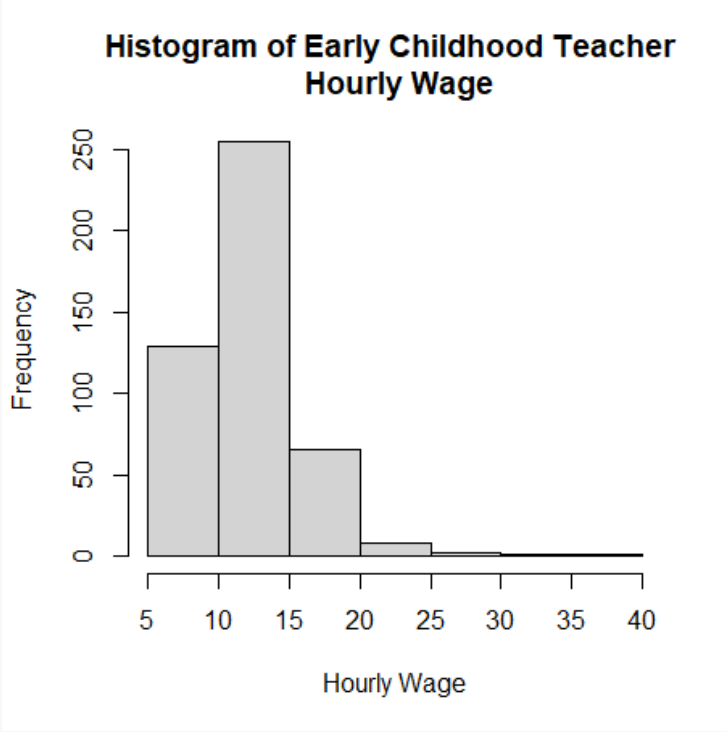
U.S. Federal Poverty Guidelines Used to Determine Financial Eligibility for Certain Federal Programs. *U.S. Department of Health and Human Services*. Published January 13, 2018.

https://aspe.hhs.gov/poverty-guidelines.

Appendix A: Code and output

```
# read in data
teaching <- read.csv("teacherWages.csv", stringsAsFactors = FALSE)

# inspect data
hist(teaching$Hourly.Wage,
     xlab = "Hourly Wage",
     main = "Histogram of Early Childhood Teacher \n Hourly Wage")
```

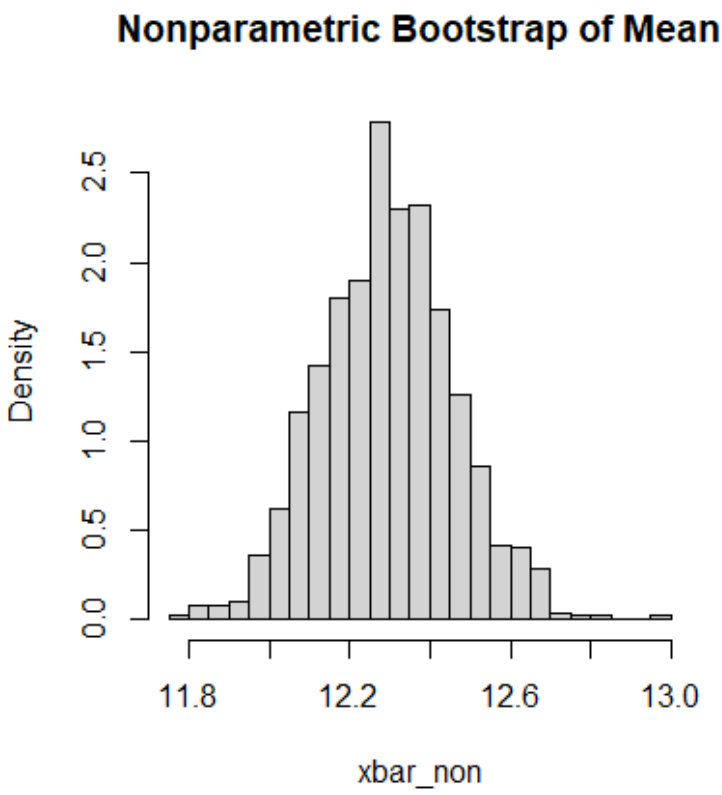



```
summary(teaching$Hourly.Wage)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.30   10.00   11.15   12.30   14.00   38.78

# nonparametric bootstrap of mean
set.seed(10)
m = 999
xbar_non = mean(sample(teaching$Hourly.Wage, replace = T))

for(i in 1:m){
  xsamp_mean = sample(teaching$Hourly.Wage, replace = T)
  xbar_non = c(xbar_non, mean(xsamp_mean))
}

hist(xbar_non, freq = F, breaks = 20,
     main = "Nonparametric Bootstrap of Mean")
```

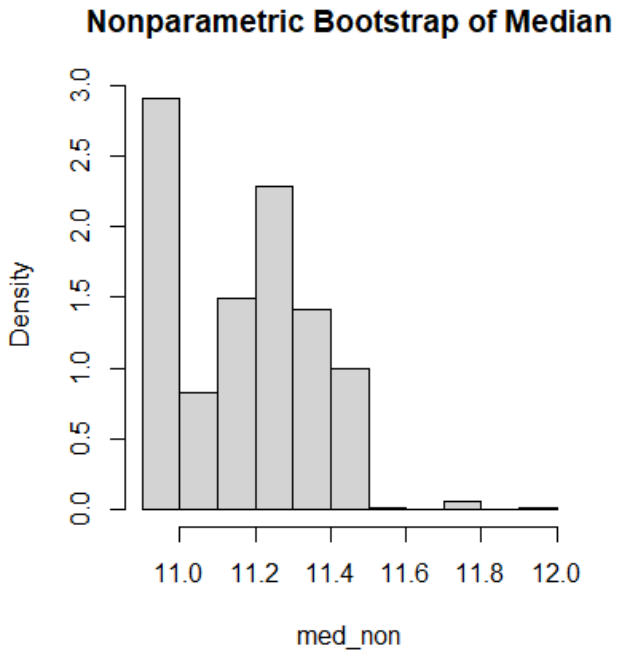



```
summary(xbar_non)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   11.76   12.19   12.30   12.30   12.40   12.94

# nonparametric bootstrap of median
set.seed(10)
m = 999
med_non = median(sample(teaching$Hourly.Wage, replace = T))

for(i in 1:m){
  xsamp_med = sample(teaching$Hourly.Wage, replace = T)
  med_non = c(med_non, median(xsamp_med))
}

hist(med_non,freq = F, breaks = 10,
     main = "Nonparametric Bootstrap of Median")
```




```
summary(med_non)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   11.00   11.00   11.25   11.19   11.25   12.00

length(xbar_non[12.07 <= xbar_non]) / length(xbar_non)
## [1] 0.916

length(med_non[12.07 <= med_non]) / length(med_non)
## [1] 0

# confidence intervals
CI.boot.mean = quantile(xbar_non,probs = c(0.025, 0.975))
CI.boot.mean
##     2.5%    97.5%
## 11.99197 12.63603

CI.boot.med = quantile(med_non,probs = c(0.025, 0.975))
CI.boot.med
##  2.5% 97.5%
##  11.0  11.5

# standard deviations
sd(xbar_non)

## [1] 0.1604704

sd(med_non)

## [1] 0.170449
```

```
# variances
var(xbar_non)

## [1] 0.02575076

var(med_non)

## [1] 0.02905285


# mean bias
12.30-12.30

## [1] 0

# median bias
11.19-11.15

## [1] 0.04
```