BEE552 Biometry Week 4

Maria Feiler

2/16/2022

My Learning Journey

Over the last week, I participated in Biometry in the following ways:

- I asked / answered **5** questions posed in class.
- I asked **0** questions in Slack.
- I answered **1** questions posed by other students on Slack.
- I came to Heather's office hours: **No**
- I came to Jose's office hours: **No**
- I met with Heather or Jose separately from office hours: **No**

Anything not falling into one of the above categories?

> **No**

On a scale of 1 (no knowledge) to 10 (complete expert), how would I rate my comfort with R programming after this week?

> **6**

Any topics from last week that you are still confused about?

> **Doing fine for now**

Problem Set

Part I

We will be using a recently published dataset on ancient rings and ribs that may have been used as early forms of money. These data come from the paper "The origins of money: Calculation of similarity indexes demonstrates the earliest development of commodity money in prehistoric Central Europe" by M.H.G. Kuijpers and C. Popa (PLoS ONE, January 20, 2021).

Download the data in the file "rings_and_ribs.csv".

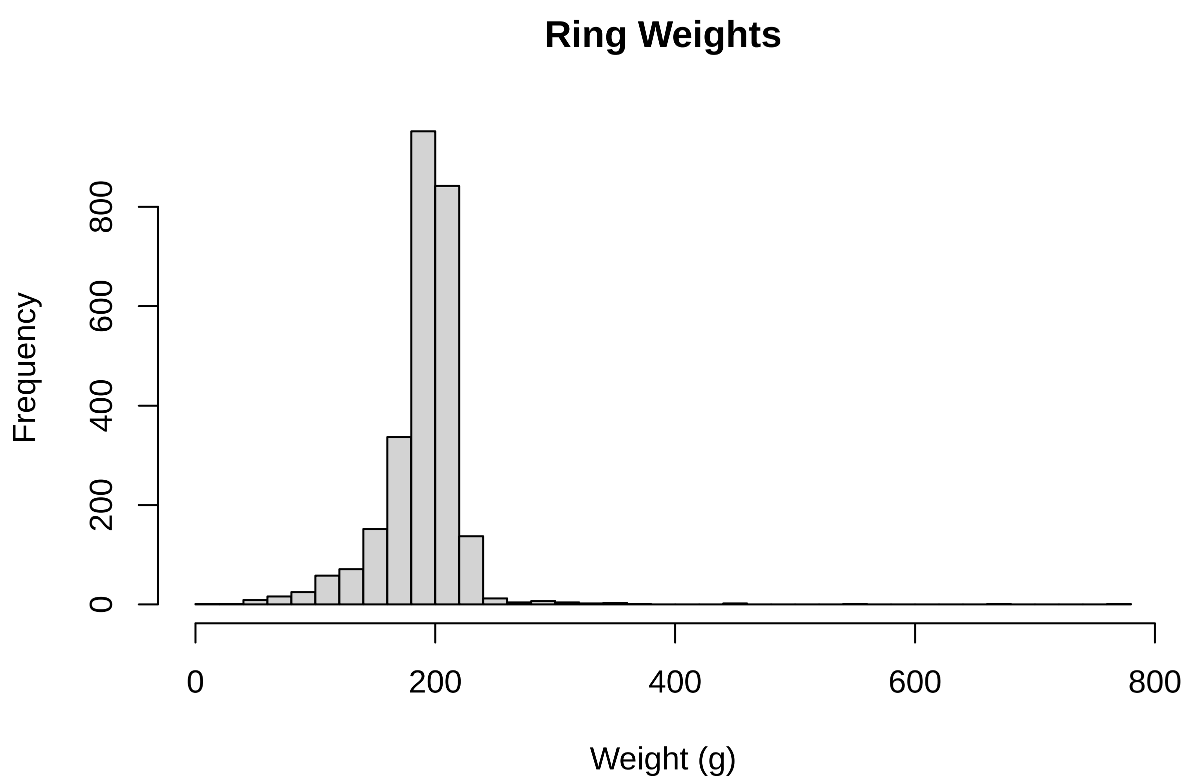
```r
money <- read_excel("journal.pone.0240462.s002.xlsx",
                    sheet = 1
                    )

# Rename the weight column to make it tidy
money <- rename(money, Weight = `Weight (g)`)
```

Table 1: A selection of data from Kuijpers and Popa 2021.

Location	Country	Zone	Date	Weight	Type	Museum	Source
Obereching	Austria	1	EBA	186.5	Rib	Salzburg	Moolsteiner and Maoesta 1988
Mürfelndorf (Pöggstall)	Austria	1	EBA	190.0	Ring	Vienna	Lenerz- de Wilde documentation
München-Luitpoldpark	Germany	1	EBA	198.0	Rib	Munich	Lenerz- de Wilde documentation
Obereching	Austria	1	EBA	202.5	Rib	Salzburg	Moolsteiner and Maoesta 1988
Traisenmündung	Austria	1	EBA	149.0	Ring	Vienna	Lenerz- de Wilde documentation
Radostice	Czech Republic	1	EBA	201.0	Ring	Ceské Budejovice	Moucha 2005
Schleching	Germany	1	EBA	186.3	Rib	Munich	Lenerz- de Wilde documentation
München-Luitpoldpark	Germany	1	EBA	167.0	Rib	Munich	Lenerz- de Wilde documentation
Köschinger Forst	Germany	1	EBA	72.0	Rib	Private Property	Lenerz- de Wilde documentation
Luštenice	Czech Republic	1	EBA	218.0	Ring	Mladá Boleslav//Dobrovice	Moucha 2005
Osterfeld	Germany	2	EBA	149.0	Ring	Halle	Lenerz- de Wilde documentation
München-Luitpoldpark	Germany	1	EBA	184.0	Rib	Munich	Lenerz- de Wilde documentation
Uttenweiler	Germany	1	EBA	93.0	Rib	Stuttgart	Lenerz- de Wilde documentation
Kobylí	Czech Republic	1	EBA	196.0	Ring	Brno	Lenerz- de Wilde documentation
Mauthausen	Germany	1	EBA	237.0	Ring	Bad Reichenhall	Lenerz- de Wilde documentation; Menke 1987/1979
Mauthausen	Germany	1	EBA	198.0	Ring	Bad Reichenhall	Lenerz- de Wilde documentation; Menke 1987/1979
Waging am See	Germany	1	EBA	127.0	Rib	Munich	Lenerz- de Wilde documentation
Mauthausen	Germany	1	EBA	207.0	Ring	Bad Reichenhall	Lenerz- de Wilde documentation; Menke 1987/1979
München-Luitpoldpark	Germany	1	EBA	179.0	Rib	Munich	Lenerz- de Wilde documentation
Unknown (Brno)	Czech Republic	1	EBA	191.0	Ring	Brno	Lenerz- de Wilde documentation
Ebersdorf an den Zaya	Austria	1	EBA	205.0	Ring	Vienna	Lenerz- de Wilde documentation
Mauthausen	Germany	1	EBA	216.0	Ring	Bad Reichenhall	Lenerz- de Wilde documentation; Menke 1987/1979
Wildendürnbach	Austria	1	EBA	203.0	Ring	Asparn	Lenerz- de Wilde documentation
Blucina	Czech Republic	1	EBA	211.0	Ring	Brno	Lenerz- de Wilde documentation
München-Luitpoldpark	Germany	1	EBA	195.0	Rib	Munich	Lenerz- de Wilde documentation
Obereching	Austria	1	EBA	205.0	Rib	Salzburg	Moolsteiner and Maoesta 1988
Bermatingen	Germany	1	EBA	74.0	Rib	Konstanz	Lenerz- de Wilde documentation
Waging am See	Germany	1	EBA	141.0	Rib	Munich	Lenerz- de Wilde documentation
Amselfing	Germany	1	EBA	125.0	Ring	Straubing	Lenerz- de Wilde documentation
Obereching	Austria	1	EBA	187.0	Rib	Salzburg	Moolsteiner and Maoesta 1988

Question 1 *For now, let's group all the locations together. Make a histogram of ring weight.*



Question 2 *What is the FORMULA for the sample mean and standard deviation (in other words, what is the formula you would want to use if you wanted to estimate the population mean [the μ parameter assuming a normal distribution] and the population variance [σ^2 if we assume a normal distribution] from a sample that represented a random subset of the entire population)?*

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

$$S = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}}$$

Question 3 *What is the population about which we are trying to make inference?*

The population is the total collection of rings and ribs ever used for currency in the context of Germany, Austria, the Czech Republic, and Poland during this time period.

Question 4 *Using R, what are the mean and the standard deviation of ring weight?*

The mean of the ring weights is 172.633. The standard deviation of ring weights is 48.296.

Question 5 *What is the formula for the standard error of the mean (heretofore s.e.)?*

$$SEM = \sqrt{\frac{S^2}{n}}$$

Question 6 *Describe the difference between the s.e. (of the mean) and the s.d.*

The standard deviation is a measure of the variance of your sample The standard error of the mean describes how close your sample mean approximates the population mean.")

Question 7 *Finish the sentence: The standard error is the standard deviation of...*

...the mean.

Question 8 *Use the MASS package's 'fitdistr' function to fit a normal distribution to the ring weight data. Do you get the roughly same answer as above?*

```
ringWeightFit <- fitdistr(money["Type" = "Ring",]$Weight,
                          "normal")

# Calculate difference in calculated and fitdistr() means
meandif <- unname(ringWeightFit$estimate[1]) - ringMean

# Calculate difference in calculated and fitdistr() standard deviations
sddif <- unname(ringWeightFit$estimate[2]) - ringSD
```

The difference between the calculated mean and the mean from fitdistr() is 0 The difference between the calculated standard deviation and the standard deviation from fitdistr() is -0.00546.

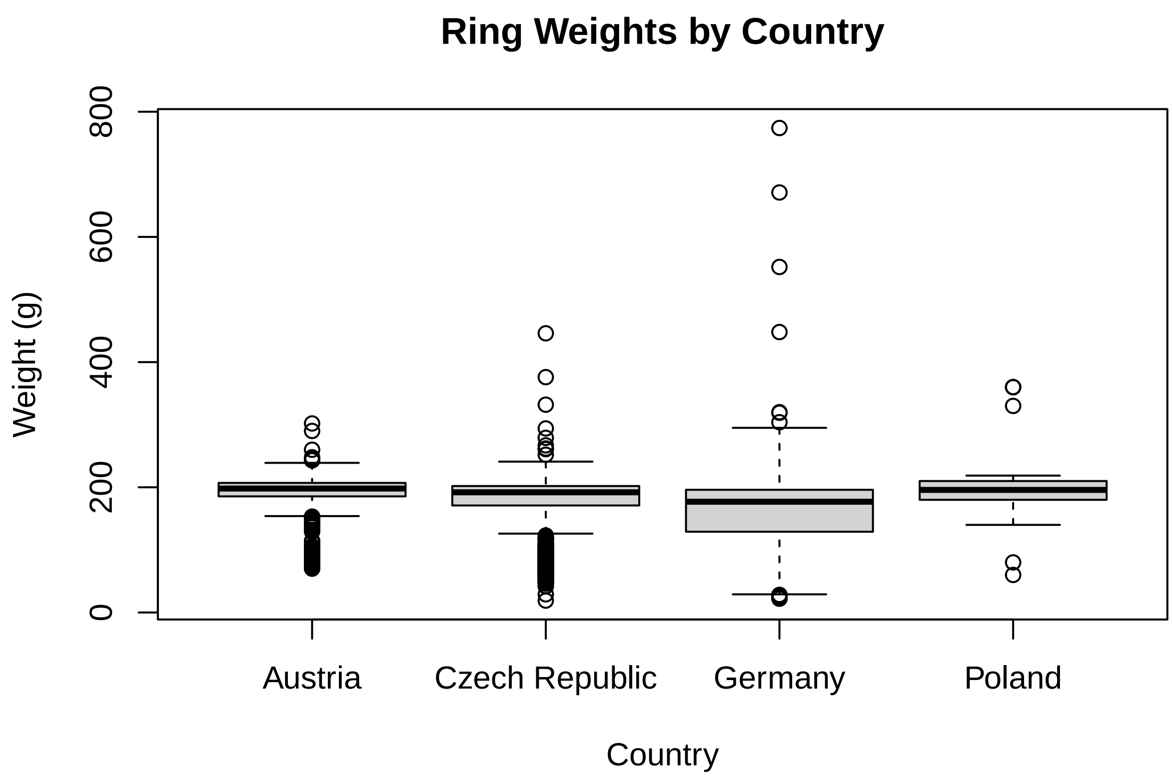
Question 9

Why might it not be valid to group the different locations when summarizing the data?

Though there might be differences between the rings used as currency in each country, all were considered to have monetary value and might have changed hands outside of country/political entity lines. Since they are all from the same time period, if you want to characterize how the earliest money worked, you must use all the material that was used as money.

Question 10 *Using the R command 'boxplot', create a boxplot to compare ring weight across different countries.*

```
ringsByCountry <- melt(data = as.data.table(money["Type" = "Ring",]),
                       id = "Country",
                       measure = "Weight"
                       )
```



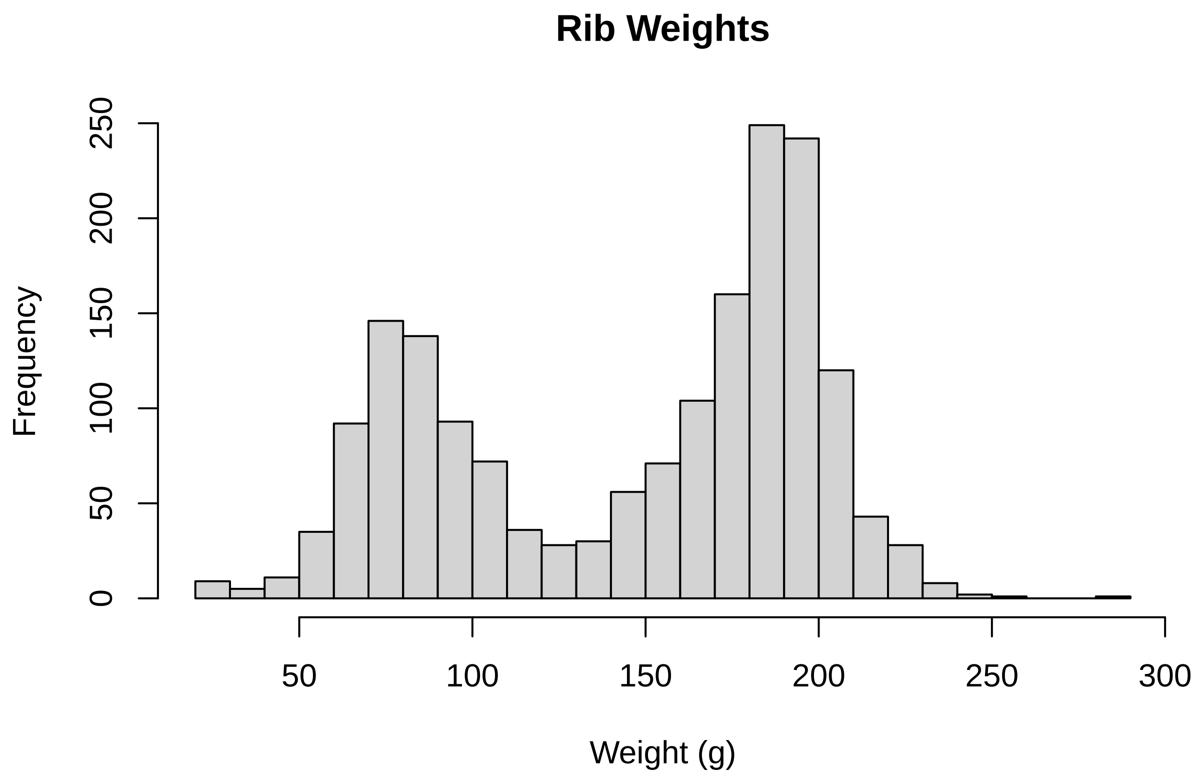
Question 11 *Make the case (in words and/or mathematically) that average ring weight in Germany is or is not statistically different from the average ring weight in Austria.*

I will test whether or not the mean weight of Austrian and German rings are statistically different using a critical value of $\alpha_c = 0.05$. The null hypothesis ($H_0 : \overline{x}_{German} = \overline{x}_{Austrian}$) will be tested using a student's t-test and will be rejected if $p > 0.05$.

```
GermanVSAustrian <- t.test(ringsByCountry$value[ringsByCountry$Country == "Germany"],
                           ringsByCountry$value[ringsByCountry$Country == "Austria"]
                           )
```

The average weight of Austrian rings (192.01 grams) is statistically greater than the average weight of German rings (161.91 grams) ($p < 0.001$).

Question 12 *So far, we've only been focused on ring weight. Let's go back and make a histogram of rib weight. Why would testing a hypothesis about average rib weight be harder?*



Making predictions or testing hypotheses about rib weight would be difficult because they appear to follow a bimodal distribution.

Part II

Assume an experiment in which the number of plants in 16 experimental plots is counted as:

```
plants <- c(8, 0, 3, 3, 4, 3, 2, 3, 3, 2, 4, 2, 3, 1, 4, 1)
```

We want to model the number of plants in each plot as being distributed according to a Poisson distribution.

Question 1 *Starting with the probability density function for the Poisson, manually derive the maximum likelihood estimate for λ, the parameter for the Poisson distribution.*

$$
\begin{aligned}
f(x \mid \lambda) &= \frac{e^{-\lambda}\lambda^x}{x!} \\
f(X_1, X_2..., X_n \mid \lambda) &= \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{X_i}}{X_i!} \\
L(\lambda \mid X_1, X_2, ..., X_n) &= \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{X_i}}{X_i!} \\
LL &= \sum_{i=1}^{n} \left(ln \frac{e^{-\lambda}\lambda^{X_i}}{X_i!} \right) \\
LL &= \sum_{i=1}^{n} (ln\ e^{-\lambda} + ln\ \lambda^{X_i} - ln\ X_i!) \\
LL &= \sum_{i=1}^{n} (-\lambda + X_i\ ln\ \lambda - ln\ X_i!) \\
NLL &= \sum_{i=1}^{n} (\lambda - X_i\ ln\ \lambda + ln\ X_i!) \\
\frac{\partial NLL}{\partial \lambda} &= \sum_{i=1}^{n} \left(1 - \frac{X_i}{\lambda}\right) \\
0 &= \sum_{i=1}^{n} \left(1 - \frac{X_i}{\hat{\lambda}}\right) \\
0 &= n - \sum_{i=1}^{n} \left(\frac{X_i}{\hat{\lambda}}\right) \\
n &= \sum_{i=1}^{n} \left(\frac{X_i}{\hat{\lambda}}\right) \\
n &= \frac{1}{\hat{\lambda}} \sum_{i=1}^{n} X_i \\
\hat{\lambda} &= \frac{\sum_{i=1}^{n} X_i}{n}
\end{aligned}
$$

Question 2 *Write an R function to calculate the negative log-likelihood for this data as described by the Poisson distribution.*

From question 1, we know the negative log-likelihood of a Poisson distribution is $NLL = \sum_{i=1}^{n}(\lambda - X_i\ ln\ \lambda) = n\lambda - \sum_{i=1}^{n} X_i\ ln\ \lambda$

```
# Function inspired by https://www.ime.unicamp.br/~cnaber/optim_1.pdf
neg.ll <- function(x, lambda){
        result <- length(x)*lambda - sum(x)*log(lambda)
        return(result)
}
```

Question 3 *Using your function for the negative log-likelihood, calculate the MLE for λ and the 95th percentile confidence interval. What would the 99th percentile confidence interval be?*

```
# Define test lambda values
lambdaVals <- seq(min(plants),
                  max(plants),
                  by = 0.05
                  )

# Create vector to catch log likelihood values of test lambdas
plantLogLiks <- rep(0, length(lambdaVals))

# Run neg.ll() over all lambdaVals
for (i in 1:length(lambdaVals)){
        plantLogLiks[i] <- neg.ll(plants, lambdaVals[i])
}

# Use optimize() to determine the minimum value
plantMinLogLik <- optimize(f = neg.ll, x = plants, interval = lambdaVals)

# 95th percentile confidence interval
# Collect the positions of the lambda values whose negative log-likelihoods are
# within 1.92 of the minimum
vals95 <- which(plantLogLiks < plantMinLogLik$objective + 1.92)

# Use the first and last to select the corresponding lambda value
LL95 <- lambdaVals[vals95[1]]
UL95 <- lambdaVals[vals95[length(vals95)]]

# 99th percentile confidence interval
# Collect the positions of the lambda values whose negative log-likelihoods are
# within 3.32 of the minimum
vals99 <- which(plantLogLiks < plantMinLogLik$objective + 3.32)

# Use the first and last to select the corresponding lambda value
LL99 <- lambdaVals[vals99[1]]
UL99 <- lambdaVals[vals99[length(vals99)]]
```

The calculated lambda for the plants per plot is 2.875. The optimized lambda value with the lowest log-likelihood is 2.875, NLL = -2.578.

95$^{\text{th}}$ percentile confidence interval: (2.15, 3.75)

99$^{\text{th}}$ percentile confidence interval: (1.95, 4.1)

Question 4 *Plot the likelihood over a range of parameter values and plot the boundaries of the 95^{th} and 99^{th} percentile confidence interval. Remember: The confidence interval is a range of parameter values, it is NOT the likelihood values itself.*