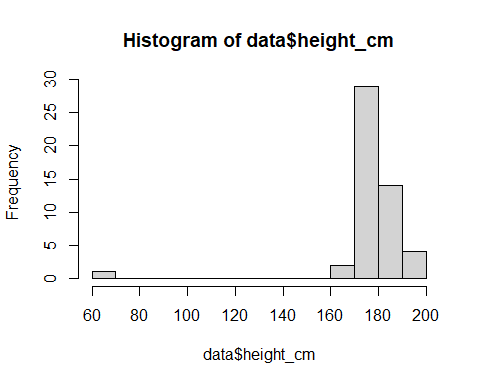
Biometry\_HW1

2025-01-12

## Question 1

#Read in the data and plot the distribution  
data <- read.csv("ABDLabs/DataForLabs/studentSampleData.csv")  
hist(data$height\_cm, breaks=10)



1. Plot the distribution of heights in the class. Describe the shape of the distribution.

*Ignoring the outlier of 69, the distribution of heights appears to be roughly symmetric and unimodal.*

1. Are there any large outliers that look as though a student used the wrong units for their height measurement?

*Yes, there is one value that is substantially smaller than the rest.*

#filter out any outliers  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

data.filter <- data %>% filter(height\_cm >= 140)

1. Use R to calculate the mean height of all students in the class, using the filtered data.

#Calculate mean of filtered data  
mean(data.filter$height\_cm)

## [1] 178.971

1. Use sd() to calculate the standard deviation of height, using the filter data.

#Calculate standard deviation of filtered data  
sd(data.filter$height\_cm)

## [1] 6.466048

## Question 2

caff <- read.csv("ABDLabs/DataForLabs/caffeine.csv")

1. What is the mean amount of caffeine in 16 oz. coffees?

mean(caff$caffeine\_mg\_16oz)

## [1] 188.0643

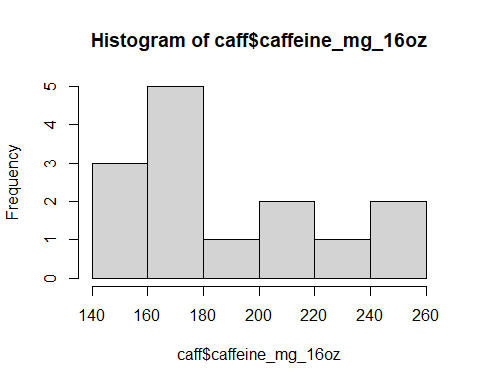
1. What is the 95% confidence interval for the mean?

t.test(caff$caffeine\_mg\_16oz)$conf.int

## [1] 167.5237 208.6049  
## attr(,"conf.level")  
## [1] 0.95

1. Plot the frequency distribution of caffeine levels for these data in a histogram.

hist(caff$caffeine\_mg\_16oz)



Is the amount of caffeine in a cup of coffee relatively consistent from one vendor to another?

*No, the amount of caffeine appears to vary widely between brands, ranging from approximately 140mg to almost 260mg.*

What is the standard deviation of caffeine level?

sd(caff$caffeine\_mg\_16oz)

## [1] 35.57535

What is the coefficient of variation?

sd(caff$caffeine\_mg\_16oz)/mean(caff$caffeine\_mg\_16oz)

## [1] 0.1891659

1. The file “caffeineStarbucks.csv” has data on six 16 oz. cups of Breakfast Blend coffee sampled on six different days. Calculate the mean (and the 95% confidence interval for the mean) for these data.

starbs <- read.csv("ABDLabs/DataForLabs/caffeineStarbucks.csv")  
mean(starbs$caffeine\_mg\_16oz)

## [1] 371.9667

t.test(starbs$caffeine\_mg\_16oz)$conf.int

## [1] 239.3527 504.5806  
## attr(,"conf.level")  
## [1] 0.95

Compare these results to the data taken on the broader sample of vendors in the first file. Describe the difference.

*The mean of the starbucks data is nearly double that of the broader sample and the 95% confidence interval is over 4 times as wide. Further, the 95% confidence intervals for the two datasets do not overlap.*

## Question 3

1. Calculate the 99% confidence interval for the mean caffeine level (for the caffeine data).

t.test(caff$caffeine\_mg\_16oz,conf.level=0.99)$conf.int

## [1] 159.4238 216.7047  
## attr(,"conf.level")  
## [1] 0.99

1. Compare this 99% confidence interval to the 95% confidence interval you calculate in question 2b. Which confidence interval is wider (i.e., spans a broader range)? Why should this one be wider?

*The 99% confidence interval encompasses approximately 8mg more in each direction than does the 95% confidence interval but both are centered around the mean of 188.06mg.* *The 99% confidence interval spans a broader range because it incorporates a greater number of values. Rather than including 95% of the potential true values, it includes 99%.*

1. Let’s compare the quantiles of the distribution of caffeine to this confidence interval. Approximately 95% of the data values should fall between the 2.5% and 97.5% quantiles of the distribution of caffeine. (Explain why this is true.)

*95% of the data values fall between the 2.5% and 97.5% quartiles because these values describe the data that was collected. These quantiles represent the middle 95% of data values, excluding the 2.5% that fall at the top and bottom of our distribution.*

quantile(caff$caffeine\_mg\_16oz, c(0.025, 0.975))

## 2.5% 97.5%   
## 144.765 254.685

Are these the same as the boundaries of the 95% confidence interval? If not, why not? Which should bound a smaller region, the quantile or the confidence interval of the mean?

*These quantiles are not the same as the boundaries of the 95% confidence interval. This is primarily because the quantiles describe the distribution of our sample while the confidence interval describes how our sample represents the true value within the population. Confidence intervals should generally bound a smaller region because they don’t need to incorporate extreme values that are found within the population. As we increase our sample size (and capture more of the population in our sample), the 95% confidence interval will typically get narrower since we are getting a better idea of what the true mean value is. However, assuming we started with a representative sample, the quantile values are less likely to change because they are simply a reflection of the range of values within our data.*

## Question 5

Return to the data on countries of the world, in “countries.csv”. Plot the distributions for ecological\_footprint\_2000, cell\_phone\_subscriptions\_per\_100\_people\_2012, and life\_expectancy\_at\_birth\_female.

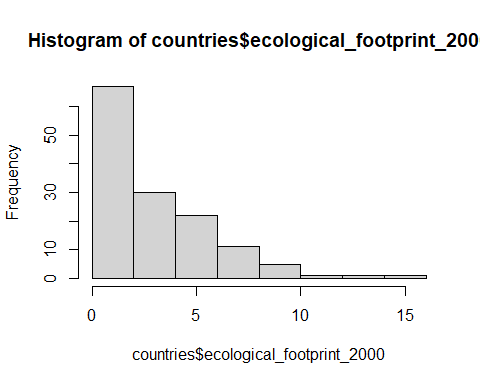
Match the difference in mean and median to the direction of skew on the histogram. Do you see a pattern?

*When the distribution is skewed right, the mean is greater than the median but when the distribution is skewed left, the median is greater than the mean. The mean is pulled by the tail of the distribution but the median is drawn to the highest point of the histogram.*

countries <- read.csv("ABDLabs/DataForLabs/countries.csv")

#Ecological footrpint

hist(countries$ecological\_footprint\_2000)



mean(countries$ecological\_footprint\_2000, na.rm=T)

## [1] 3.147391

median(countries$ecological\_footprint\_2000, na.rm=T)

## [1] 2.14

1. Is the variable skewed? If so, in which direction?

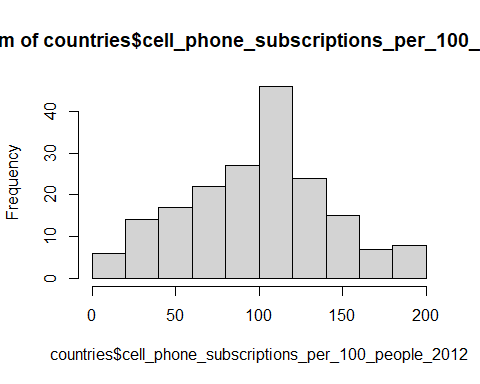
* *The variable appears to be skewed right since there is a long tail to the right. It is important to note that it is bounded on the left by 0 though.*

1. Calculate the mean and median. Are they similar?

* *The mean and median are not the same. The mean is greater than the median by a factor of approximately 1.47.*

#Cellphone subscriptions

hist(countries$cell\_phone\_subscriptions\_per\_100\_people\_2012)



mean(countries$cell\_phone\_subscriptions\_per\_100\_people\_2012, na.rm=T)

## [1] 99.90419

median(countries$cell\_phone\_subscriptions\_per\_100\_people\_2012, na.rm=T)

## [1] 103.25

1. Is the variable skewed? If so, in which direction?

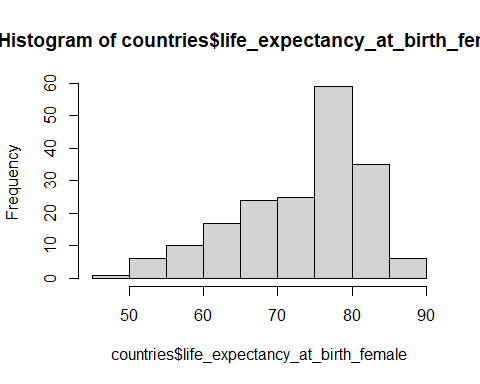
*The distribution of cellphone subscriptions appears very slightly skewed left.*

1. Calculate the mean and median. Are they similar?

*The mean and median are similar but the median is greater by a factor of 1.03.*

#Life expectancy

hist(countries$life\_expectancy\_at\_birth\_female, breaks=10)



mean(countries$life\_expectancy\_at\_birth\_female, na.rm=T)

## [1] 73.4153

median(countries$life\_expectancy\_at\_birth\_female, na.rm=T)

## [1] 75.9

1. Is the variable skewed? If so, in which direction?

*The distribution of life expectancy is skewed left.*

1. Calculate the mean and median. Are they similar?

*The mean and median are similar but the median is greater by a factor of 1.03.*