# HW1

Haojie Liu

2025-04-09

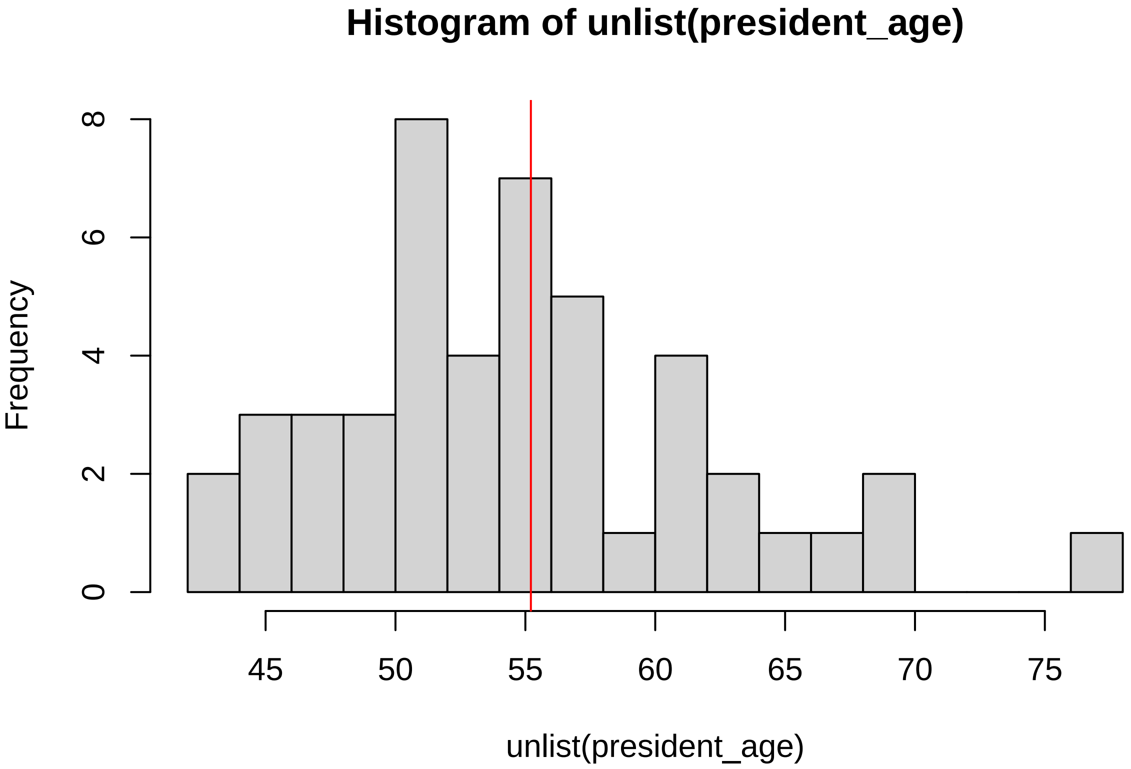
```
library(tidyverse)
```

**1. The ages of presidents at inauguration are shown in the file "president_age.csv". This file can be found under the Data folder.**

```
president_age <- read.csv("president_age.csv")
```

Draw a relative frequency histogram of presidential ages at inauguration. (Hint: Read in the data file in Module 1, and plot a histogram). What does this histogram tell you?

```
hist(unlist(president_age),breaks = 20)
abline(v = mean(unlist(president_age)), col = "red")
```



This data is quite normal distributed across the axis.

What is the median age of presidents at inauguration?

```
median(unlist(president_age))
```

```
## [1] 55
```

What is the mean age of presidents at inauguration?

```r
mean(unlist(president_age))
```

```
## [1] 55.21277
```

Is the median or mean a more useful measure of central tendency in this data set? Why?

median, since it is mode close to the mode.

Was Joe Biden, at 78, unusually old to become president? (Hint – think about outliers – you will have to define what they are.)

```r
summary(president_age)
```

```
##        x        
##  Min.   :42.00  
##  1st Qu.:51.00  
##  Median :55.00  
##  Mean   :55.21  
##  3rd Qu.:59.00  
##  Max.   :78.00  
```

We can see that the 3rd quantile is around 59, therefore, we can count Joe Biden as a outlier since 78 is beyond that point.

### 2. Retrieve the Los Angeles precipitation data from the class website under Module 1 ("LArain.csv").

```r
LArain <- read.csv("LArain.csv")
```

Find the mean and standard deviation of these data. Assume these values represent population parameters of a normally distributed random variable.

```r
mean(LArain$y)
```

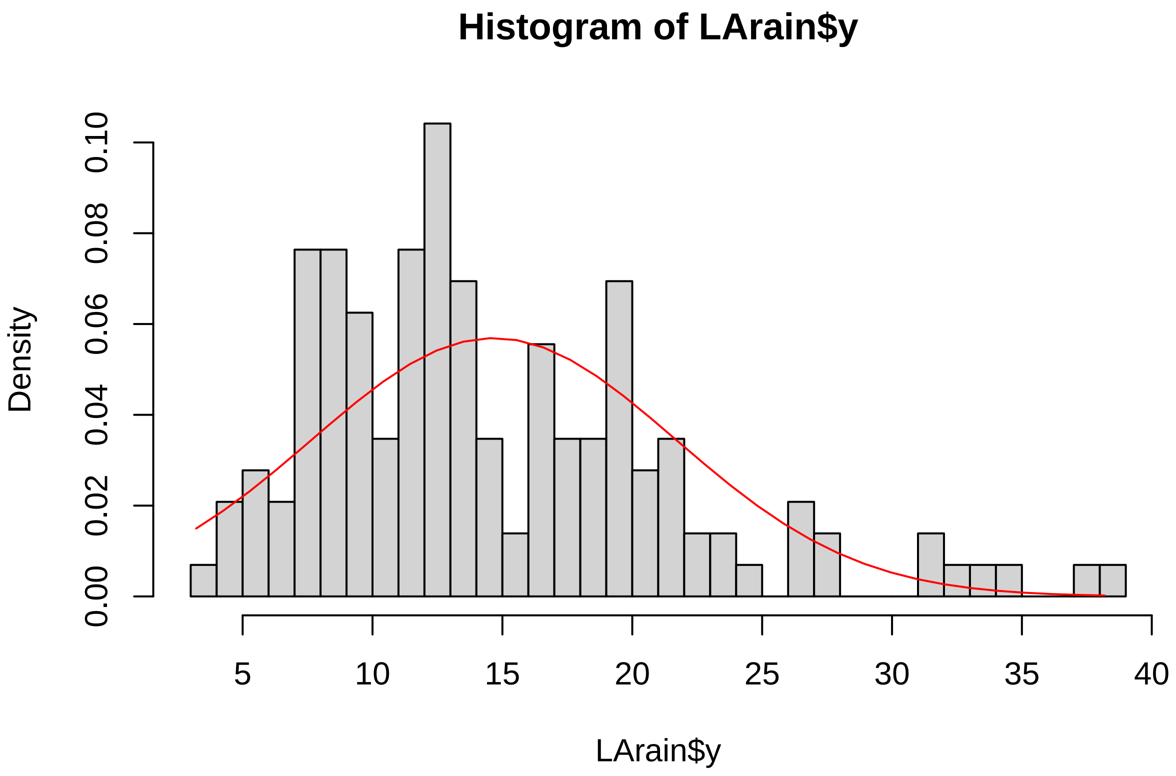
```
## [1] 14.67097
```

```r
sd(LArain$y)
```

```
## [1] 7.010294
```

Generate a histogram of the precipitation data and overlay the normal density curve for these data. Briefly interpret what the histogram and density curve tell you.

```r
hist(LArain$y, freq = FALSE, breaks = 40)
x <- seq(min(LArain$y), max(LArain$y), length.out = max(LArain$y) - min(LArain$y))
m <- dnorm(x, mean(LArain$y), sd(LArain$y))
lines(x, m, col = "red")
```



This data is kind right skewed.

Now find the following probabilities: The probability that next year annual rainfall will be less than 26 inches.

```
mean_val <- mean(LArain$y)
sd_val <- sd(LArain$y)

pnorm(26, mean = mean_val, sd = sd_val)
```

```
## [1] 0.9469589
```

The probability that next year annual rainfall will be between 10 and 20 inches.

```
pnorm(20, mean = mean_val, sd = sd_val) - pnorm(10, mean = mean_val, sd = sd_val)
```

```
## [1] 0.523815
```

The probability that next year annual rainfall will be at least twice the normal average.

```
pnorm(2 * mean_val, mean = mean_val, sd = sd_val, lower.tail = FALSE)
```

```
## [1] 0.0181846
```

**3. A biogeographer collects species counts from a sample of 40 plots in a study area. In those plots the sample mean is equal to 21.5. Prior work establishes that the population standard deviation for species counts across plots in the study area is 4.3. Find the 99% confidence interval for the unknown population mean number of species across similar sized plots.**

```
xbar <- 21.5
sigma <- 4.3
n <- 40
z <- 2.576
```

```r
error <- z * (sigma / sqrt(n))
lower <- xbar - error
upper <- xbar + error
c(lower, upper)
```

```
## [1] 19.7486 23.2514
```

**4. An historical geographer is interested in the average number of children of households in a certain city in 1800. Rather than spend the time analyzing each entry in the city directory, she decides to sample randomly from the directory and estimate the average number of children per household in the city from this sample. In a sample of n = 60 households, she finds the average number of children to be 4.56 with a standard deviation of 2.03. Use the Student's t-distribution to find the 95% confidence interval for the average number of children per household in the city as a whole.**

```r
xbar <- 4.56
s <- 2.03
n <- 60
t <- qt(0.975, df = n - 1)

error <- t * (s / sqrt(n))
lower <- xbar - error
upper <- xbar + error
c(lower, upper)
```

```
## [1] 4.035595 5.084405
```

**5. In a sample of 30 US cities, average male income was found to be $92,000. Assume that the population standard deviation is known and equal to $16,000. Write out your hypothesis statements, identify the appropriate test statistic, specify your decision rules and conclusions for the following 2 questions:**

Test the hypothesis that this sample could have been drawn from a population with an average male income equal to $86,000, using a significance level of 0.05.

```r
mu <- 86000
xbar <- 92000
sigma <- 16000
n <- 30
z <- (xbar - mu) / (sigma / sqrt(n))
abs(z) > 1.96
```

```
## [1] TRUE
```

Reject the null hypothesis since z score is greater than 1.96

Test the hypothesis that this sample could have been drawn from a population with an average male income greater than $86000, again using a significance level of 0.05.

```r
mu <- 86000
xbar <- 92000
sigma <- 16000
n <- 30
z <- (xbar - mu) / (sigma / sqrt(n))
z > 1.645
```

```
## [1] TRUE
```

Reject the null hypothesis since z score is greater than 1.645