Social Sciences Intro to Statistics

Week 3.2 Distributions

Week 3: Learning goal - Articulate the descriptors of normal distribution and skewness.

Introduction

Load packages:

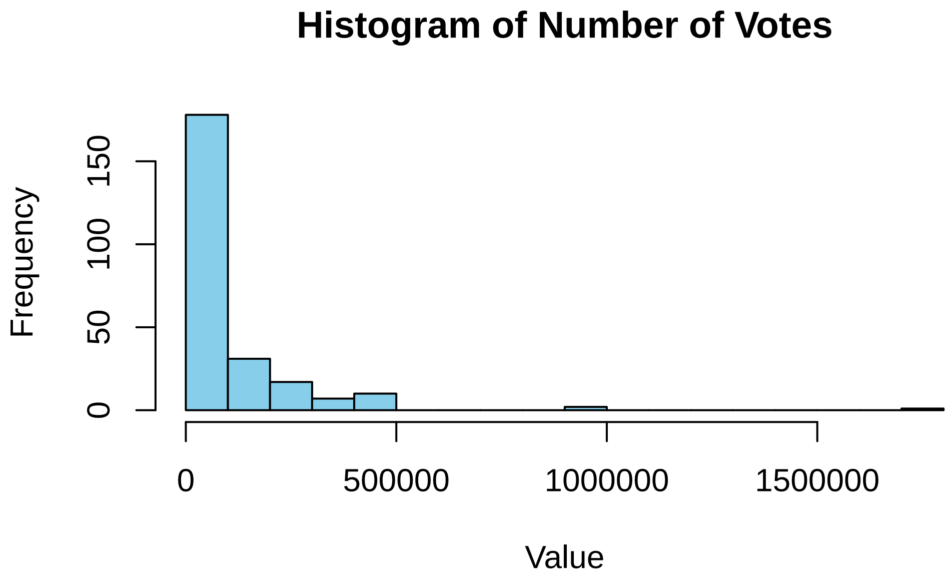
```
library(tidyverse)
library(ggplot2)
best_netflix <- read_csv("https://raw.githubusercontent.com/bcl96/Social-Sciences-Stats/ma
```

Resources used to create this lecture:

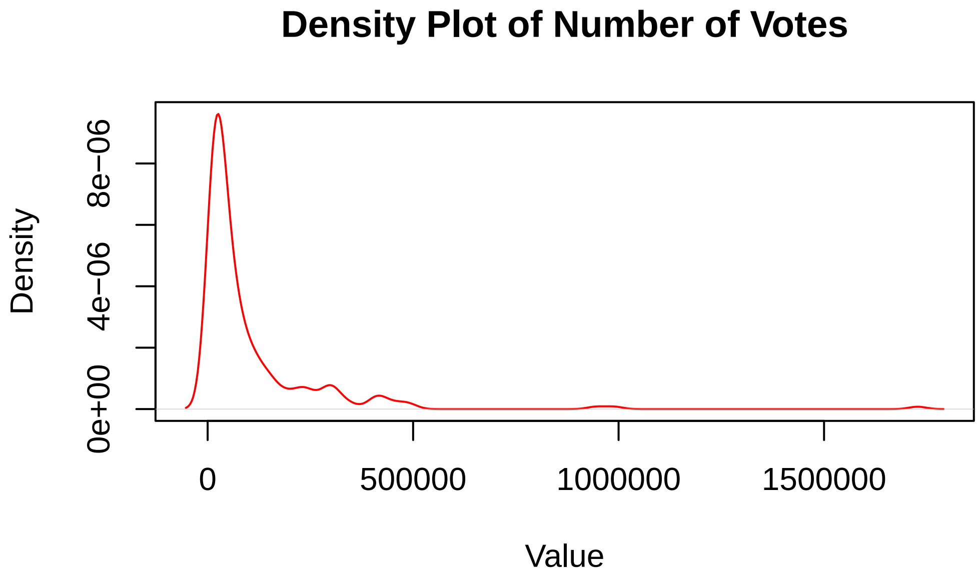
Distributions

Distributions help us further understand our data as it provides a snapshot of the data. Distribution shows us where the average value is (central tendency), the spread of the dataset (what the variability is), if the values are evenly spread out (normal) or if there is more values on one side (skewness).

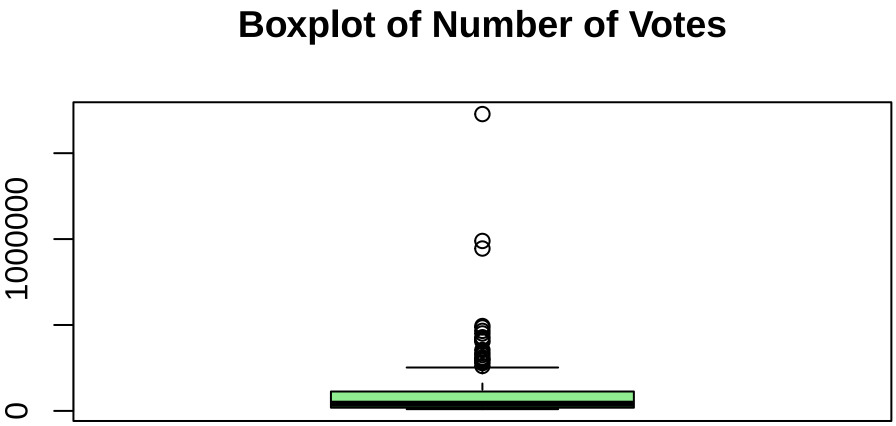
```
# Distribution with a histogram
hist(best_netflix$NUMBER_OF_VOTES, breaks = 20, col = "skyblue", main = "Histogram of Numb
```



```
# Distribution with a density plot
plot(density(best_netflix$NUMBER_OF_VOTES), main = "Density Plot of Number of Votes", xlab
```



```
# Distribution with a box plot
boxplot(best_netflix$NUMBER_OF_VOTES, main = "Boxplot of Number of Votes", col = "lightgre
```



Normal distribution

Normal distributions are continuous probability distributions that are symmetric around the mean. Normal distributions have a bell-shaped curve, where the mean, median, and mode of the distribution are all equal and located at the center of the distribution. The standard deviation of our normal distribution tells us the spread of the distribution. The larger the standard deviation, the wider the normal distribution. The smaller the standard deviation, the narrower the normal distribution. For datasets that have a normal distribution, about 68% of the data will fall within one standard deviation of the mean, and 95% of the data will fall within two standard deviations, and 99.7% of the data falls within three standard deviations.

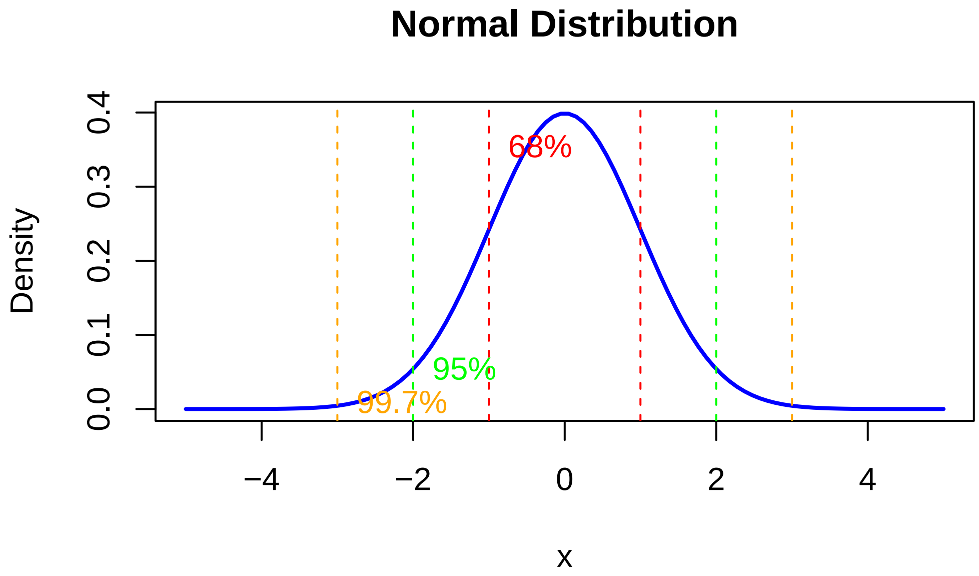
```
# Example of normal distribution
x <- seq(-5, 5, length.out = 100)  # Range of x values
y <- dnorm(x, mean = 0, sd = 1)      # PDF values for the normal distribution

# Plot the normal distribution
```

```
plot(x, y, type = "l", lwd = 2, col = "blue",
     xlab = "x", ylab = "Density",
     main = "Normal Distribution")

# Add vertical lines for one, two, and three standard deviations
abline(v = c(-1, 1), col = "red", lty = 2)  # One SD
abline(v = c(-2, 2), col = "green", lty = 2)  # Two SD
abline(v = c(-3, 3), col = "orange", lty = 2)  # Three SD

# Add text annotations for the percentages
text(-1, 0.35, "68%", col = "red", pos = 4)
text(-2, 0.05, "95%", col = "green", pos = 4)
text(-3, 0.005, "99.7%", col = "orange", pos = 4)
```



Skewness (normal, left-skewed, right-skewed)

Skewness measures the asymmetry of the distribution around its mean. In a normal distribution, the skewness is zero, which means that the distribution is symmetric. When the distribution leans towards the left side, it is left-skewed or negatively skewed. When the distribution leans towards the right side, it is right-skewed or positively skewed.

Left-skewed distribution has its mean less than its median, and its median less than its mode.

The tail of the distribution extends to the left side. Visually we will see that most of the data points are on the right side of the distribution. Since this is negatively skewed, the skewness will be less than zero.

Right -skewed distribution has its mean greater than its median, and its median greater than its mode. The tail of the distribution extends to the right side. Visually we will see that most of the data points are on the left side of the distribution. Since this is positively skewed, the skewness will be greater than zero.

```
- normal distributions and the empirical rule
- standard normal distribution
- z-scores
```