Social Sciences Intro to Statistics

Week 3.1 Descriptive Statistics

Week 3: Learning goal - Identify descriptive statistics in order to describe individual variables.

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:

How do we understand data?

Before we can interpret and run statistical tests, we need to first understand our data. One way to understand our data is using descriptive statistics to describe data. We can use summary functions to find out about the mean, median, percentiles, and standard deviation of our data.

Mean, Median, and Percentiles

Mean is the average of the dataset, the sum of all the values divided by the number of values. While the mean represents the typical value of a dataset, it is also sensitive to outliers. Outliers can drastically change the value of our mean.

```
mean(best_netflix$SCORE)
#> [1] 8.093496
```

Median on the other hand, is less sensitive to outliers. The median is the value where 50% of the dataset would be greater and 50% of the dataset would be lesser than it. To find the median, you arrange the dataset in ascending order (smallest to largest) and then find the value that is in the middle. If the dataset has an odd number of values, our median is the middle value. If the dataset has an even number of values, our median is the average of the two median values.

```
median(best_netflix$SCORE)
#> [1] 8
```

Percentiles help us visualize our dataset as 100 equal parts. The 25th percentile, also known as a the first quartile, is the value where 25% of the data would fall below. The 75th percentile, or third quartile, is the value where 75% of the data would fall below The median is the 50th percentile, it equally divides our dataset and 50% of the data would fall below. Similar to the median, because percentiles provide information on the spread of our dataset, it is less affected by outliers than compared to the mean.

```
# Find the 25th percentile (first quartile)
q1 <- quantile(best_netflix$SCORE, probs = 0.25)

# Find the median (50th percentile)
median <- quantile(best_netflix$SCORE, probs = 0.5)

# Find the 75th percentile (third quartile)
q3 <- quantile(best_netflix$SCORE, probs = 0.75)

# Print the results
print(q1)
#> 25%
#> 7.7
print(median)
#> 50%
#>   8
print(q3)
#> 75%
#> 8.4
```

Standard Deviation

Standard deviation is a measure of variability in the values of a dataset. Standard deviation measures how much the values deviate away from the mean. If all of the values in our dataset are the same, then the mean would be the same value and our standard deviation would be zero. A high standard deviation indicates that our values are spread out and there's more variability in our dataset. A low standard deviation indicates that our values are close to the mean.

Describing individual variables using R

Using R, we can also use summarise() to reduces multiple values down to a single summary output. This would give us the summary statistics such as mean, median, minimum, maximum, standard deviation for specific variables in the data frame. We can also use summarise() to create new variables that we build from existing variables.

```
#syntax of 'summarise()'
#summarise(data, new_variable = function(existing_variable))
```

Measures of central tendency (Mean, median, mode, etc.)

Measures of dispersion (Standard deviation)