Median and Interpolated Median

R Tutorials for Applied Statistics

Note on required packages: The following code required the packages `tidyverse` and `psych`. The `tidyverse` actually contains many packages that allow you to organize, summarize, and plot data. The package `psych` is used to perform statistics related to the median. If you have not already done so, download, install, and load the library with the following code:

```
# This only needs to be executed once for your machine
install.packages("psych")

# This only needs to be executed once for your machine
install.packages("tidyverse")

# This needs to be executed every time you load R
library("psych")

# This needs to be executed every time you load R
library("tidyverse")
```

1 Introduction

The **population median** is the value of the 50th percentile of some variable for all the members of the population. When members of the population are sorted by this value, the median is the middle value.

The **sample median** is the sample estimate of the population median.

The median can be measured on ordinal, interval, or ratio data. Because ordinal data is categorical data, the mean is not an appropriate measure of center. However, since ordinal data can be sorted or ranked, it is possible to calculate the median.

While one can also measure the mean of interval or ratio data, it is often desirable to compute the median for populations that have a skewed distribution. That is, an asymmetric distribution where one end of the distribution extends farther from the median than another end. The extreme values of the long end of the distribution cause the mean to move towards that tail, away from the middle of the distribution.

For example, the distribution of income (and most economic and financial variables related to income) is skewed to the right. For example, the median household income in the United States was $55,775 in 2015. What separates a median income household from the poorest person in poverty is approximately $55,000 per year. What separates a median income household from the richest household is millions of dollars per year. The right end of the income distribution stretches much farther beyond the median than the left. When very high income households end up in a sample, this pushes the mean higher, but not the median.

2 Example Dataset

In this dataset, students in fourth through sixth from three school districts in Michigan ranked their how important each of the following were for achieving popularity: achieving good grades, athletic ability, having popularity, and having money. A rank of 1 indicates highest importance and a rank of 4 indicates lowest importance. The data set comes from Chase, M. A., and Dummer, G. M. (1992), “The Role of Sports as a Social Determinant for Children,” *Research Quarterly for Exercise and Sport*, 63, 418-424.

The code below downloads and loads the dataset.

```r
load(url("http://www.murraylax.org/datasets/gradeschool.RData"))
```

3 Compute Medians

The dataset includes variables called `Grades` and `Money`, among others. Compute the median importance for each of these variables with the following code:

```r
median(df$Grades)
```

```
## [1] 3
```

```r
median(df$Money)
```

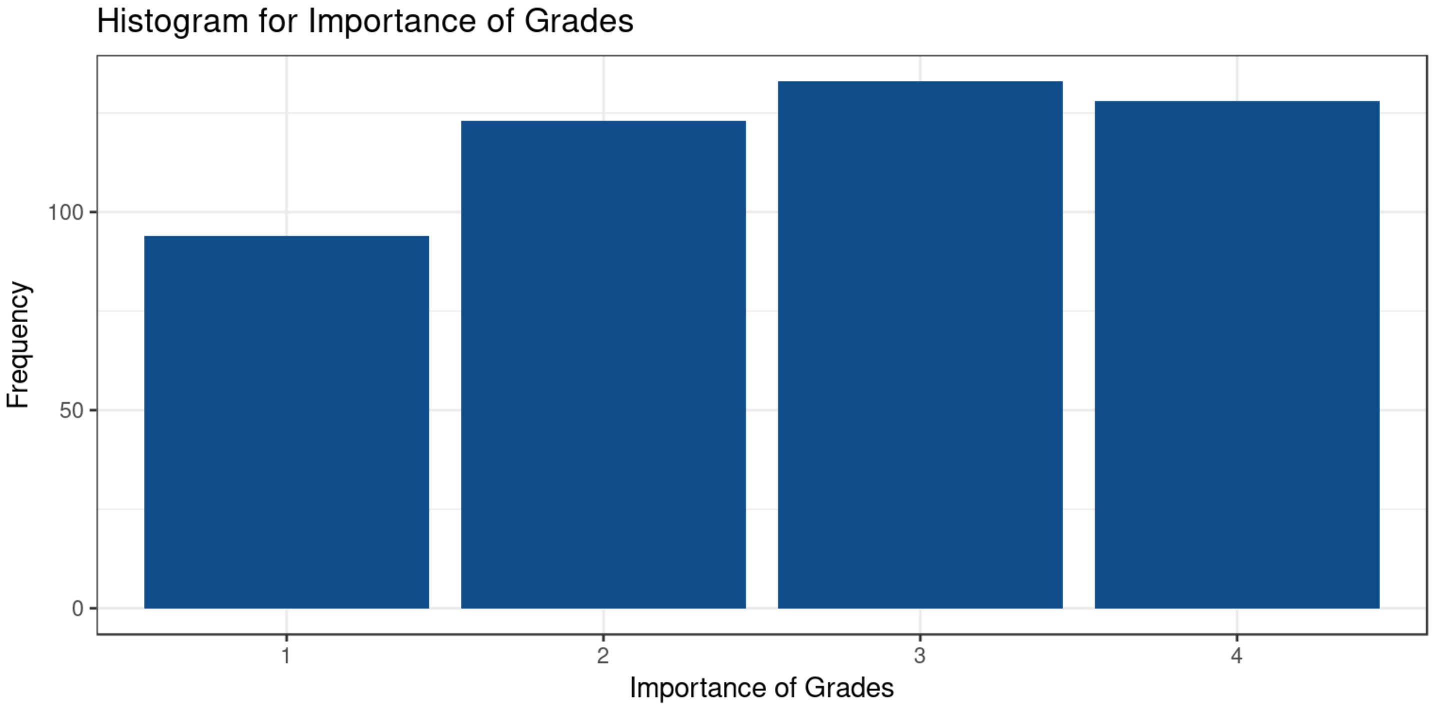
```
## [1] 3
```

The median value for both of these variables is equal to 3.

4 Display histograms

A **histogram** is a bar graph illustrating the number of observations in a sample fall in several intervals. Let’s display a histogram of the importance students place on grades in terms of being popular with the following code:

```
ggplot(data=df, mapping=aes(x=as.factor(Grades))) +
  geom_bar(fill="dodgerblue4") +
  labs(title="Histogram for Importance of Grades",
       x="Importance of Grades",
       y="Frequency") +
  theme_bw()
```



The code above uses the `ggplot2` package to create a histogram plot. The first segment,

`ggplot(data=df, mapping=aes(x=as.factor(Grades))) +`

sets up a plot, establishes the data frame to be used (`df`), and maps the `Grades` variable to the x-axis. The `as.factor(Grades)` tells R to treat the `Grades` variable as a `factor`, i.e. a categorical variable.

The next segment of the code, `geom_bar(fill="dodgerblue4")` sets up the *geometry* layer of the plot. The geometry layer specifies what kind of shape will be plotted. In this case, we plot bars that are shaded blue.

The segment of code creates the labels,

`labs(title="Histogram for Importance of Grades",`

`x="Importance of Grades",`

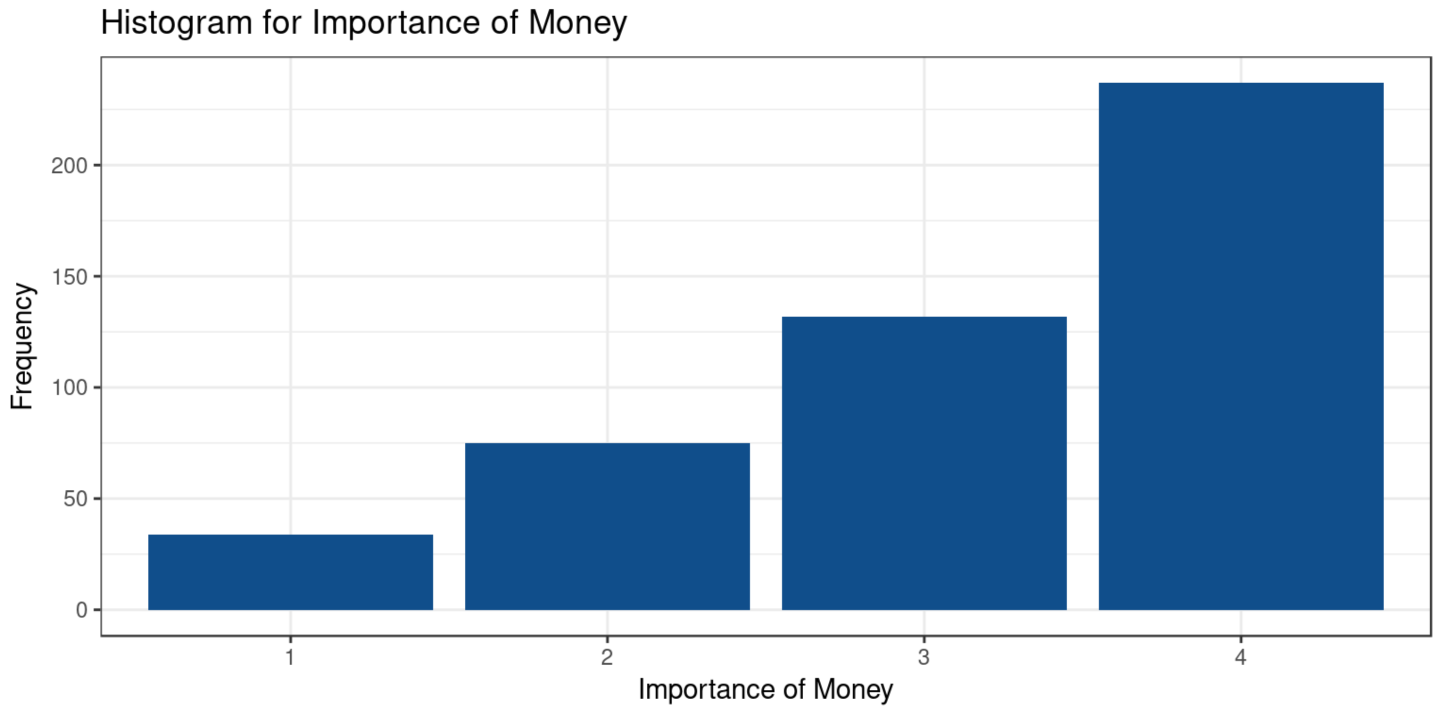
`y="Frequency")`

set up the title of the plot and the labels for the x-axis and y-axis.

Finally, the segment `theme_bw()` sets up a simple black and white theme for the background.

Let’s display a histogram for the importance of money:

```
ggplot(data=df, mapping=aes(x=as.factor(Money))) +
  geom_bar(fill="dodgerblue4") +
  labs(title="Histogram for Importance of Money",
       x="Importance of Money",
       y="Frequency") +
  theme_bw()
```



While the money had the same median importance (3) as grades, we can see from the histogram that a much smaller portion of students ranked money below the median (at 1 or 2) than above the median (at 4).

5 Interpolated Median

The situation above often occurs when comparing medians of ordinal data with a limited number of responses. While the medians may be equal, it may be clear from the histograms that one distribution is more heavily weighted above or below the median than the other distribution.

The **interpolated median** provides another measure of center which takes into account the percentage of the data that is strictly below versus strictly above the median.

The interpolated median gives a measure within the upper bound and lower bound of the median, in the direction that the data is more heavily weighted. Using the example above, the median of each variable is equal to 3, but the interpolated median can take any value between 2.5 and 3.5, depending on whether the distribution is more heavily weighted above or below 3.

While the interpolated median returns a value on a continuous scale (i.e. fractional numbers above and below the median), it is appropriate to use on ordinal data, as well as interval and ratio data.

Let’s calculate the interpolated median for `Grades` and `Money`:

```
interp.median(df$Grades)
```

```
## [1] 2.665414
```

```
interp.median(df$Money)
```

```
## [1] 3.484848
```

We can see from these measures of interpolated medians that the center of the sample distribution for the level of importance for grades (2.67) is less than the center of the sample distribution for money (3.48). Therefore, while both of the samples had an equal median equal to 3, we can say that in our sample the centers of the samples imply the students put a higher level of importance for grades than money (lower numbers were used to indicate more important rank).