Median and Interpolated Median

Note on required packages: The following code required the package `psych` to perform statistics related to the median. If you have not already done so, download, install, and load the library with the following code:

`install.packages("psych")`

`library("psych")`

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.

Example: 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.

1. Download the dataset

The code below downloads the dataset and assigns the dataset to a variable we call `kidsdata`.

```
kidsdata <- read.csv("http://www.murraylax.org/datasets/gradeschool.csv")
```

2. 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:

```
median(kidsdata$Grades)
```

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

```
median(kidsdata$Money)
```

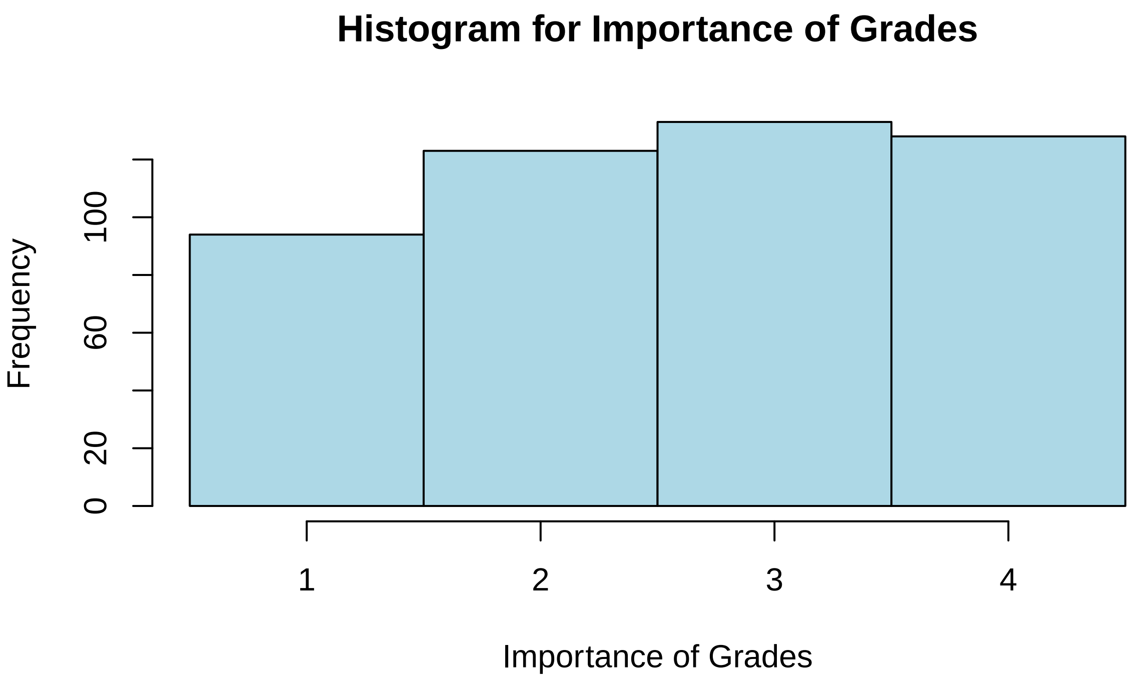
```
## [1] 3
```

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

3. 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:

```
hist(kidsdata$Grades,
     breaks = c(0.5,1.5,2.5,3.5,4.5),
     xlab = "Importance of Grades",
     main = "Histogram for Importance of Grades",
     col="lightblue")
```

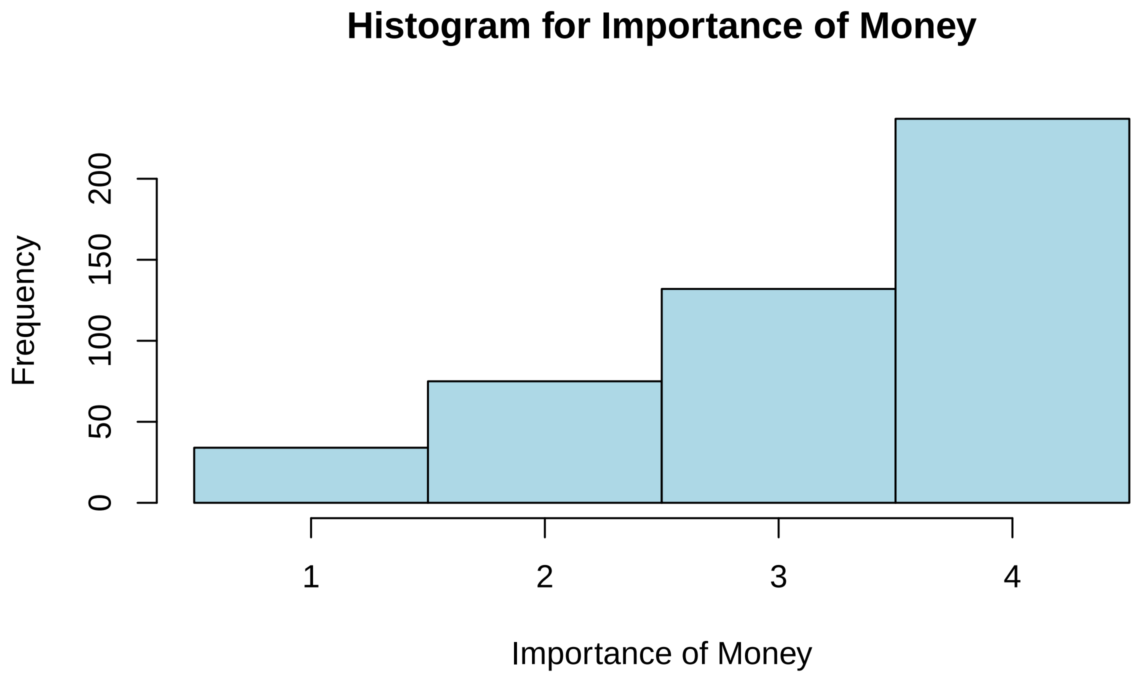


In the code above, the function `hist` displays a histogram. The first parameter, `kidsdata$Grades`, is the variable to display the histogram for. The parameter `breaks = c(0.5,1.5,2.5,3.5,4.5)` specifies the breakpoints for the intervals to use in the histogram. These breaks put the values 1, 2, 3, and 4 in the center of each interval. Finally, to make the histogram more visually attractive, we set a label for the horizontal axis using the parameter, `xlab = "Importance of Grades"`, set the title text for the histogram using `main = "Histogram for Importance of Grades"`, and made the bars light blue with the parameter `col="lightblue"`.

We can see from the histogram above that while the median importance is equal to 3, nearly half of the students ranked grades at 1 and 2.

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

```
hist(kidsdata$Money,
     breaks = c(0.5,1.5,2.5,3.5,4.5),
     xlab = "Importance of Money",
     main = "Histogram for Importance of Money",
     col="lightblue")
```



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).

4. 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(kidsdata$Grades)
```

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

```
interp.median(kidsdata$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).