Estimating the Population Mean

The **population mean** is a measure of the center or "average" value in the whole population of a variable measured at the interval or ratio level.

The **sample mean** is a sample estimate of the population mean. It is a the same measure of center, obtained from a sample. The variable in your sample must be measured at the interval or ratio level.

Example: Current Population Survey from 2004 that includes data on average hourly earnings, marital status, gender, and age for thousands of people. A part of it is available for download from textbook website for Stock and Watson's *Introduction to Econometrics*.

1. Download the Dataset.

The code below downloads the data set and assigns the dataset to a variable we create and call `cps04`.

```
download.file(
  url="http://murraylax.org/datasets/cps04.csv",
  destfile="cps04.csv");
cps04 <- read.csv("cps04.csv");
```

The dataset `cps04` contains a variable called `ahe`, which stands for average hourly earnings.

2. Compute the Sample Mean.

```
mean(cps04$ahe)
```

```
## [1] 16.77115
```

The sample estimate for average hourly earnings for U.S. workers is 2004 is $16.77. This is not necessarily the population mean. Like every statistic, it includes a margin of error due to random sampling error.

3. Compute a 95% Confidence Interval

The confidence interval is a range of values for the population mean, based on our estimate of the sample mean, and an estimate for the margin of error due to random sampling.

The function `t.test` computes a number of statistics and statistical tests for a variable, including a confidence interval. In the code below, we use the function to compute our confidence interval and assign all the resulting output to a new variable we call `ahestats`.

```
ahestats <- t.test(cps04$ahe, conf.level = 0.95)
```

The output of `t.test` that we assigned to variable `ahestats` is a list which includes an item called `conf.int`.

Let's call this item to report our confidence interval:

```
ahestats$conf.int
```

```
## [1] 16.57902 16.96328
## attr(,"conf.level")
## [1] 0.95
```

The confidence interval for average hourly earnings for U.S. workers is 2004 is $16.58 - $16.96. We can say with 95% confidence that this interval estimate includes the true population mean.

4. One Sample T-Test (One-tailed):

Suppose a politician claimed that the average earnings of American workers was more than $16.50 per hour. We know that the sample estimate is larger from above, but let's test the hypothesis that the *population mean* is *more than* $16.50.

The appropriate statistical procedure is the **One-sample T-test for a Mean** which tests whether a single population mean is equal to or different than a particular value. Our null and alternative hypotheses for our one-sample t-test is given by the following:

Null hypothesis: $\mu = 16.50$
Alternative hypothesis: $\mu > 16.50$

The `t.test` function can also compute the one-sample t-test using the following code:

```
t.test(cps04$ahe, mu=16.50, alternative = "greater")
```

```
##
##  One Sample t-test
##
## data:  cps04$ahe
## t = 2.7665, df = 7985, p-value = 0.002839
## alternative hypothesis: true mean is greater than 16.5
## 95 percent confidence interval:
##  16.60992      Inf
## sample estimates:
## mean of x
##  16.77115
```

The output of the test reveals a p-value equal to 0.00028. Since this is below 5%, we reject the null hypothesis and conclude that we do have statistical evidence that the population mean is greater than $16.50.

5. Two-Tailed Test:

The previous example is a **one-tailed** test. That is, it involved an alternative hypothesis that looked for statistical evidence that the population parameter was in a particular direction away from the null hypothesized value (in the case above, *greater than* the null hypothesis).

A two tailed test instead tests an alternative hypothesis that simply says the population parameter is *different than* the null hypothesized value, leaving the possibility that it may be less than or may be greater than the value.

Let's test the following two-tailed hypotheses:

Null hypothesis: $\mu = 16.50$
Alternative hypothesis: $\mu \neq 16.50$

Notice the $\neq$ sign in the alternative hypothesis.

We use the `t.test` function again to compute the one-sample t-test using the following code:

```
t.test(cps04$ahe, mu=16.50, alternative="two.sided")
```

```
##
##  One Sample t-test
##
## data:  cps04$ahe
## t = 2.7665, df = 7985, p-value = 0.005679
## alternative hypothesis: true mean is not equal to 16.5
```

```
## 95 percent confidence interval:
##  16.57902 16.96328
## sample estimates:
## mean of x
##  16.77115
```

We can see from the output that the p-value is equal to 0.0057. Since this is below 5%, we reject the null hypothesis and conclude that we do have statistical evidence that the population mean *is different than* \$16.50.

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)
```

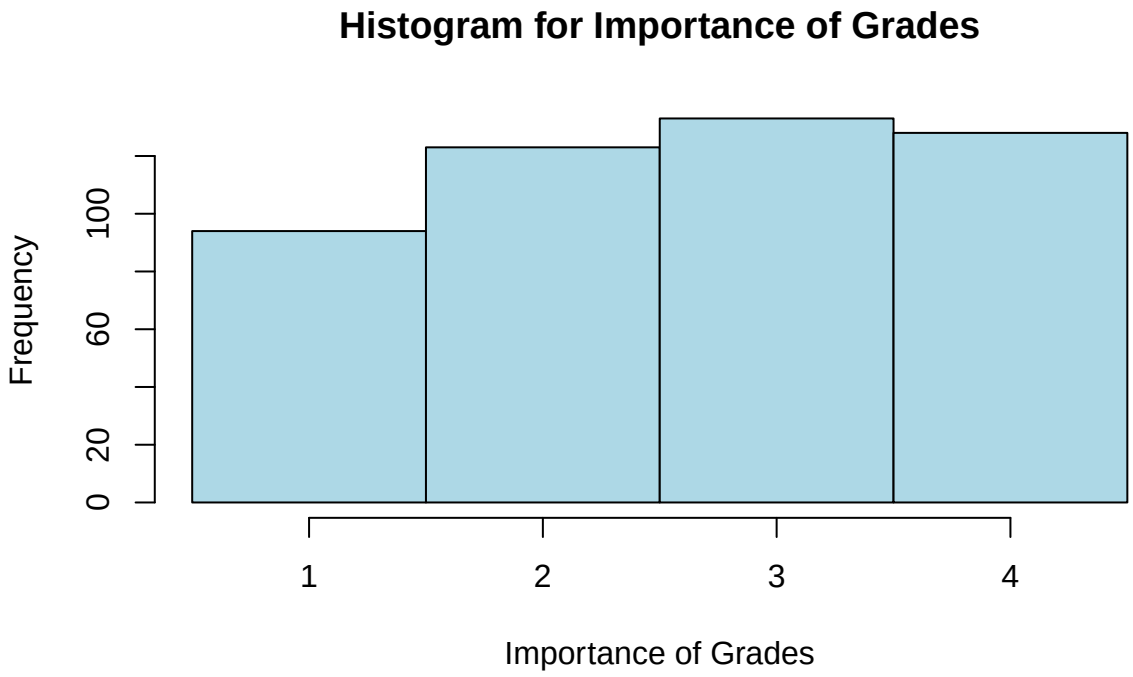
```
## [1] 3
```

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")
```

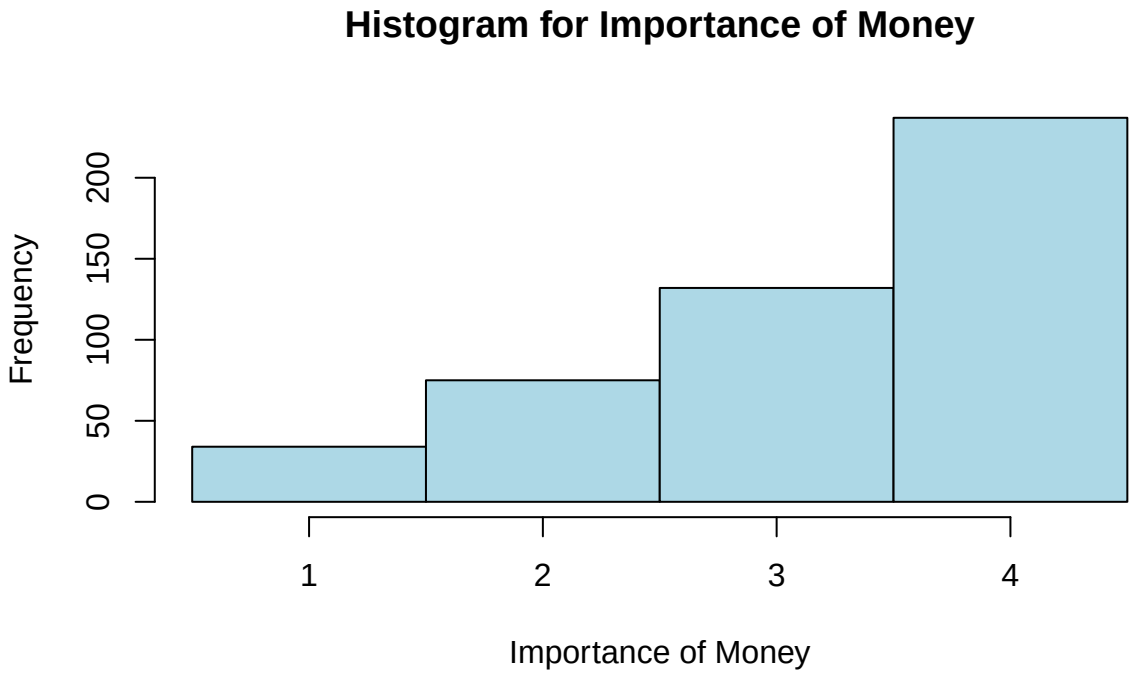


In the code above, the function `hist` displays a histogram. The first parameter, `kidsdata$Grades`, is the variable to display the histogram for. The parameter `breaks = c(0.5,1.5,2.5,3.5,4.5)` specifies the breakpoints for the intervals to use in the histogram. These breaks put the values 1, 2, 3, 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).

Estimating the Population 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.

An alternative measure for the median is the **interpolated median**. This is 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. For example, if the median of a variable is equal to 3, 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.

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 a variable called `Grades`, which is a ranking on a scale of 1-4 for how important students find good grades are for maintaining popularity. Let us compute the sample median and sample interpolated median for this variable:

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

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

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

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

The median ranking for grades is equal to 3, while the interpolated median equal to 2.67 indicates the distribution is more heavily weighted below 3. Therefore, even though the data is on a discrete scale for 1 to 4, the data is centered most closely to 2.67.

3. Confidence Intervals for the Median

It is most common to form confidence intervals for the median using bootstrapped samples. This procedure uses the data in the single sample to simulate thousands of possible samples, and for each simulation computes the median and/or interpolated median.

A short R script on my website http://www.murraylax.org contains bootstrapping procedures for calculating confidence intervals for the median and interpolated median. These procedures can be called into R with the following code:

```
source("http://www.murraylax.org/code/R/medianbs.r")
```

A confidence interval for both the median and interpolated median for the example data above can be computed with the following function call:

```
median.bs(kidsdata$Grades, conf.level=0.95, bootn=50000)
```

```
## $Confidence.Level
## [1] 0.95
##
## $Median.Confidence.Interval
##  2.5% 97.5%
##     3     3
##
## $Interpolated.Median.Confidence.Interval
##     2.5%    97.5%
## 2.507091 2.805195
##
## $Median
## [1] 3
##
## $Interpolated.Median
## [1] 2.665414
```

The function calculates the median and interpolated median for the variable `kidsdata$Grades` for `bootn=50000` simulated samples and reports the 2.5 and 97.5 percentiles (the middle 95% since `conf.level=0.95`)

Note that when you run the above procedure you may have found slightly different numbers for the ranges in the confidence interval. This is because these estimates are based on thousands of random simulations of samples, and your computer may have generated different random samples that resulted in different estimates for the confidence intervals. This is common for statistical procedures based on simulation methods.

4. Hypothesis Test on the Center of the Distribution

The **Wilcoxon Signed Rank test** considers the hypothesis that the distribution is centered around a particular value. Let us use the example above to illustrate the use of Wilcoxon Signed Rank test. We found

that median rank of importance that students assigned to getting good grades in our sample was equal to 3, and the sample interpolated median was 2.67. Let us test the hypothesis that in the population, the average rank for grades is less than 3. The null and alternative hypotheses are the following:

Null hypothesis: The population rank for grades is centered at 3.
Alternative hypothesis: The population rank for grades is centered ***below* 3.**

The following code calls the `wilcox.test()` function to test this hypothesis:

```
wilcox.test(kidsdata$Grades, alternative="less", mu=3)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  kidsdata$Grades
## V = 16128, p-value = 3.06e-15
## alternative hypothesis: true location is less than 3
```

The first parameter to the function call, `kidsdata$Grades`, specifies the variable for the hypothesis test, the second parameter, `alternative="less"` specifies that this is a one-tailed test with an alternative hypothesis that the distribution is centered below the given value, and the final parameter, `mu=3`, gives the value to test which is given in the null and alternative hypothesis.

With a p-value is significantly below 5%, we find statistical evidence that the population of students has ranks for grade importance centered below 3.

Estimating Differences in Means

Here we investigate estimating the difference in the means between two *independent samples*.

With **independent samples**, we have two groups of observations, distinguishable by some measured characteristic to divide into these groups, and no member of one group is also in the other group. The outcome of the observations in one group must be *independent* of the outcome of the observations in the other group.

Testing differences in *means* between two independent samples is appropriate when a variable measured from two independent samples are in the same units and at the interval or ratio scale.

Example: Current Population Survey from 2004 that includes data on average hourly earnings, marital status, gender, and age for thousands of people. A part of it is available for download from textbook website for Stock and Watson's *Introduction to Econometrics*.

Our goal with this example is to estimate the difference in average hourly earnings between men and women. The gender variable is our measured characteristic that will divide our sample into two independent samples.

1. Download the Dataset.

The code below downloads the data set and assigns the dataset to a variable we create and call `cps04`.

```
download.file(
  url="http://murraylax.org/datasets/cps04.csv",
  destfile="cps04.csv");
cps04 <- read.csv("cps04.csv");
```

The dataset `cps04` contains a variable called `ahe`, which stands for average hourly earnings, and a variable `female` which is equal to 1 if the observation is for a female and equal to 0 if for a male.

2. Calculate Means

The function `t.test` computes a number of statistics and statistical tests for a difference between two means, including sample estimates for each mean, a confidence interval, and a hypothesis test. In the code below, we call the function and assign all the resulting output to a new variable we call `ahestats`.

```
ahestats <- t.test(ahe ~ female, data=cps04, conf.level=0.95)
```

The first parameter, `ahe ~ female`, is a formula that says we are interested in the outcome variable `ahe` and how it is different for different values for `female`. The second parameter, `data=cps04`, tells the function in what dataset the variables `ahe` and `female` can be found. The last parameter `conf.level=0.95` will generate output that will be useful later for computing a 95% confidence interval.

The output of `t.test` that we assigned to variable `ahestats` is a list which includes an item called `estimate`. The `estimate` item includes the mean of each of the groups defined by `female`. Report this item with the following code:

```
ahestats$estimate
```

```
## mean in group 0 mean in group 1
##        17.77262        15.35857
```

We can see from above that men in our sample have average hourly earning equal to $17.77 and women have average hourly earnings equal to $15.36.

3. Calculate a 95% Confidence Interval

The confidence interval is a range of values for difference between the population means for our two independent groups, based on our estimates of the sample means and an estimate for the margin of error due to random sampling.

The output to the call to `t.test` above also included a confidence interval, in an item called `conf.int`. Let's call this item to report our confidence interval:

```
ahestats$conf.int
```

```
## [1] 2.039724 2.788384
## attr(,"conf.level")
## [1] 0.95
```

The confidence interval for the difference between average hourly earnings between men and women is between \$2.04 and \$2.79. We can say with 95% confidence that this interval estimate includes the true difference in population means.

4. Two-Tailed Independent Samples T-Test

An **independent samples t-test** lets us determine whether there is evidence that the mean of the first group is different than the mean of the second group in the population. The typical two-tailed test considers the following null and alternative hypotheses:

Null hypothesis: $\mu_0 - \mu_1 = 0$
Alternative hypothesis: $\mu_0 - \mu_1 \neq 0$

Notice that the alternative hypothesis includes a $\neq$ sign which implies that this is a two-tailed test. We are not explicitly testing which group has a larger average hourly earnings. We are only testing whether the population means are different from one another.

The output to the call to `t.test` above also included an independent samples t-test. If we call our return value, summary information from the test is output to the screen.

```
ahestats
```

```
##
##  Welch Two Sample t-test
##
## data:  ahe by female
## t = 12.642, df = 7792.6, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  2.039724 2.788384
## sample estimates:
## mean in group 0 mean in group 1
##        17.77262        15.35857
```

We can see from above that the p-value is 2.2e-16 which is much smaller than a significance level of 5%. We can say confidently that there is statistical evidence that the average hourly earnings is different for men and women.

5. One-Tailed Independent Samples T-Test

Suppose a politician claims that men earn on average more than \$2.00 per hour more than women. The null and alternative hypotheses for testing this claim are given by the following:

Null hypothesis: $\mu_0 - \mu_1 = 2.00$
Alternative hypothesis: $\mu_0 - \mu_1 > 2.00$

In the hypotheses above, μ_0 denotes the mean hourly earnings for men (female=0), and μ_1 denotes the mean hourly earnings for women (female=1). The alternative hypothesis has a *greater-than* symbol, because the claim we are testing suggested that men on average make more than \$2.00 per hour than women, when we subtract group 1 (women) from group 0 (men), the result should be *greater than* 2.00. This is therefore a one-tailed test.

The code that we ran above did conduct a hypothesis test, but we need to call the function again to give it the specifics that we want a one-tailed test and that we have a value of \$2.00 in the hypotheses. The relevant call to `t.test` is given by,

```
t.test(ahe ~ female, data=cps04, alternative="greater", mu=2.00)
```

```
##
##  Welch Two Sample t-test
##
## data:  ahe by female
## t = 2.1683, df = 7792.6, p-value = 0.01508
## alternative hypothesis: true difference in means is greater than 2
## 95 percent confidence interval:
##  2.099917      Inf
## sample estimates:
## mean in group 0 mean in group 1
##        17.77262        15.35857
```

The p-value is 0.015, which is less than 0.05, so we can say at the 5% significance level, we found sufficient statistical evidence that on average men earn more than \$2.00 per hour more than women.

Estimating Differences in Population Medians

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")
```

Here we investigate estimating the difference in the *medians* between two *independent samples*.

With **independent samples**, we have two groups of observations, distinguishable by some measured characteristic to divide into these groups, and no member of one group is also in the other group. The outcome of the observations in one group must be *independent* of the outcome of the observations in the other group.

Testing differences in *medians* between two independent samples is appropriate when a variable measured from two independent samples are in the same units and at the ordinal, interval, or ratio scale.

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.

An alternative measure for the median is the **interpolated median**. This is another measure of center which takes into account the percentage of the data that is strictly below versus strictly above the median.

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 a variable called `Grades`, which is a ranking on a scale of 1-4 for how important students find good grades are for maintaining popularity, and a variable called `Gender` which is equal to the text "boy" or "girl" for every observation. Let us compute the sample median importance of grades for boys and girls.

```
median( kidsdata$Grades[ kidsdata$Gender=="boy"] )
```

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

```
median( kidsdata$Grades[ kidsdata$Gender=="girl"] )
```

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

The code above calls the `median()` function and passes only a subset of the `Grades` observations in each call. The square brackets in `kidsdata$Grades[...]` are used to select specific rows of the `Grades` variable. For the first call that computes the median response for boys, the rows that are selected are the ones where `Gender` is equal to the text, "boy". Similarly, the second call computes the median for `Grades` but selecting only the rows where `Gender` is equal to "girl."

We see in the results above that the sample median response for the importance of grades is equal to 3 for both boys and girls.

Similarly, we can compute the interpolated median for each gender:

```
interp.median( kidsdata$Grades[ kidsdata$Gender=="boy"] )
```

```
## [1] 2.701493
```

```
interp.median( kidsdata$Grades[ kidsdata$Gender=="girl"] )
```

```
## [1] 2.628788
```

Here we can see that the distributions for how important grades are for each boys and girls are centered slightly below 3, and the distribution for boys is centered at a slightly higher value (2.70) than the distribution for girls (2.63), indicating boys on average put slightly *less* importance on grades than girls.

3. Hypothesis Test on Differences in the Center of the Distribution

The **Mann Whitney U-Test**, or sometimes refered to as the **Mann-Whitney-Wilcoxon** test, considers the hypothesis that the distributions for two independent samples are centered around the same value. In the example above, we found that median rank of importance earning good grades in our sample was equal to 3 for both boys and girls, but the sample interpolated medians differed slightly, with boys centered around 2.70 and girls centered around 2.63. Let us test the hypothesis that in the population, the distribution for importance of grades is centered around the same value for boys and girls

Null hypothesis: The center of the distribution for the importance of grades is *equal* for boys and girls
Alternative hypothesis: The center of the distribution for the importance of grades is *different* for boys and girls

The following code calls the `wilcox.test()` function to test this hypothesis:

```
wilcox.test(kidsdata$Grades ~ kidsdata$Gender, alternative="two.sided")
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  kidsdata$Grades by kidsdata$Gender
## W = 29388, p-value = 0.5373
## alternative hypothesis: true location shift is not equal to 0
```

The first parameter to the function call, `kidsdata$Grades ~ kidsdata$Gender`, is a formula that says we are interested in the outcome variable `Grades` and how it is different for different values for `Gender`. The second parameter, `alternative="two.sided"` specifies that this is a two-tailed test with an alternative hypothesis that the center of the two distributions are *different*.,

With a p-value is much above 5%, we *fail to find* statistical evidence that the median response for the importance for grades is different for boys versus girls.

Estimating Differences in Means in Paired Samples

Here we investigate estimating the difference in the means between two *paired samples*.

With **paired samples**, we have a single set of observations, but two measures or two variables for each observation. Obtaining two or more measures from each observation to pair can result from the following situations:

1. **Across-time measures:** The same dependent variable is measured for each individual but at two different time periods. For example, a sample of individuals may have their income measures once in 2013 and again in 2014, and a researcher could ask whether there was a change in average income from one year to the next.
2. **Different conditions measures:** The same dependent variable is measured for each individual, but under two different conditions, or before and after some treatment. For example, a sample of high school students may have test scores measured before introducing them to a new curriculum and afterwards. A researcher could ask whether the curriculum affected test scores.
3. **Related topics measures:** Two slightly different variables are measured for each individual. For example, foreign language students may take separate exams for writing proficiency and speaking proficiency. A researcher could ask whether students are more proficient with writing in a foreign language versus speaking.

Testing differences in *means* between paired samples is appropriate when the variables are measured at the interval or ratio scale.

Example: The Centers for Disease Control and Prevention (CDC) maintains data on motor vehicle fatalities by State, Age, and Gender. In our example, a dataset with 50 observations, one for each U.S. state, the motor vehicle occupant fatality rate per 100,000 members of the population. The dataset includes separate variables for the following age groups: 0-20, 21-34, 35-54, and 55+. The dataset also includes variables for the mortality rate for women as a whole and men as a whole.

1. Download the dataset

The code below downloads a csv file available on my personal website, http://murraylax.org/datasets/, then reads it into an R dataset that we name `fatalities`.

```
download.file(
  url="http://murraylax.org/datasets/vehiclefatalities.csv",
  dest="vehiclefatalities.csv")
fatalities <- read.csv("vehiclefatalities.csv");
```

2. Computing Means

The function `t.test` computes a number of statistics and statistical tests for a difference between two means, including sample estimates for the differences in the means, a confidence interval, and a hypothesis test. In the code below, we call the function to compare the means of variables `Age21.34` and `Age.35.54`, instruct the function that these are *paired* samples, and assign all the resulting output to a new variable we call `fatalstats`.

```
fatalstats <- t.test(x=fatalities$Age.21.34,
                     y=fatalities$Age.35.54,
                     paired=TRUE,
                     alternative="two.sided",
                     conf.level=0.95)
```

The first two parameters, `x` and `y`, into the function `t.test` are the two variables that we are comparing. As each variable is a member of the dataset named `fatalities`, we access each one by first naming the dataset, typing a dollar sign (`$`), then specifying which variable in the dataset we are referring to.

The third parameter specifies that we want a two tailed test. We do a two tailed test because our research question simply asked if there is a *difference* between the average fatality rate for the two groups, not whether a specific variable was *greater* than the other. Consistent with the research question and the two-tailed test is the not-equal sign in the alternative hypothesis. The last parameter `conf.level=0.95` will generate output that will be useful later for computing a 95% confidence interval.

The output of `t.test` that we assigned to variable `fatalstats` is a list which includes an item called `estimate`. The `estimate` item is equal to the mean of the `x` variable (the 21-34 age group) minus the mean of the `y` variable (the 35-54 age group). We report this item with the following code:

```
fatalstats$estimate
```

```
## mean of the differences
##                   4.625
```

Here we see a positive number, equal to `4.625`, which means the fatality rate for the 21-34 age group is higher than the 35-54 age group, by amount of 4.625 people per 100,000 in the population. If we wish to compute the mean for each age group, we can use the `mean()` function for each of the variables as follows:

```
mean(fatalities$Age.21.34, na.rm=TRUE)
```

```
## [1] 13.70435
```

```
mean(fatalities$Age.35.54, na.rm=TRUE)
```

```
## [1] 8.970455
```

The parameter `na.rm=TRUE` tells the function `mean` to ignore missing values, which are coded with NA (i.e. not available). We can see here that the mean fatality rate for the 21-34 age group is approximately 13.704 per 100,000 and the mean fatality rate for the 35-54 age group is 8.970 per 100,000. The difference is equal to the estimate found above, 4.625.

3. Calculate a 95% Confidence Interval

The confidence interval is a range of values for difference between the population means of the two variables, based on our samples estimates of the means and an estimate for the margin of error due to random sampling.

The output to the call to `t.test` above also included a confidence interval, in an item called `conf.int`. Let's call this item to report our confidence interval:

```
fatalstats$conf.int
```

```
## [1] 3.674189 5.575811
## attr(,"conf.level")
## [1] 0.95
```

The confidence interval places the mean difference between fatality rates of the 21-34 age group and the 35-54 age group in the range 3.674 and 5.576. We can say with 95% confidence that this interval estimate includes the true difference in population means.

4. Two-Tailed Paired Samples T-test

Let us test the hypothesis that the vehicle fatality rate 21-34 age group is different than the vehicle fatality rate for 35-54 age group. The null and alternative hypotheses are given as follows:

Null hypothesis: $\mu_{21/34} - \mu_{35/54} = 0$
Alternative hypothesis: $\mu_{21/34} - \mu_{35/54} \neq 0$

Notice that the alternative hypothesis includes a $\neq$ sign which implies that this is a two-tailed test. We are not specifying that a particular age group should be higher than the other. We are only testing whether the population means are *different* from one another.

The output to the call to `t.test` above also included a paired samples t-test. If we call our return value, summary information from the test is output to the screen.

```
fatalstats
```

```
##
##  Paired t-test
##
## data:  fatalities$Age.21.34 and fatalities$Age.35.54
## t = 9.8097, df = 43, p-value = 1.541e-12
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  3.674189 5.575811
## sample estimates:
## mean of the differences
##                   4.625
```

We can see from above that the p-value is 1.5e-12 which is much smaller than a significance level of 5%. We can say confidently that there is statistical evidence that the average fatality rate for the 21-34 age group is different than the 35-54 age group.

5. One-Tailed Independent Samples T-Test

Suppose our intuition tells us that the 21-34 age group may have a higher fatality rate than the 35-54 age group, because they have less experience driving and less maturity may lead to more dangerous decisions. To test this intuition, suppose a researcher is interested in instead testing the following one-tailed hypotheses:

Null hypothesis: $\mu_{21/34} - \mu_{35/54} = 0$
Alternative hypothesis: $\mu_{21/34} - \mu_{35/54} > 0$

The *greater-than* symbol implies that this is a one-tailed.

The code that we ran above did conduct a hypothesis test, but we need to call the function again to specify that this is a one-tailed test. The relevant call to `t.test` is given by,

```
t.test(x=fatalities$Age.21.34,
       y=fatalities$Age.35.54,
       paired=TRUE,
       alternative="greater",
       conf.level=0.95)
```

```
##
##  Paired t-test
##
## data:  fatalities$Age.21.34 and fatalities$Age.35.54
## t = 9.8097, df = 43, p-value = 7.707e-13
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  3.832424      Inf
## sample estimates:
## mean of the differences
```

```
##                   4.625
```

The p-value is 7.7e-13 which is much smaller than a significance level of 5%. Not coincidentally, this p-value is exactly half of the p-value we found in the two tailed test above. Given the low p-value, we can say confidently that there is statistical evidence that the average fatality rate for the 21-34 age group is *greater than* the 35-54 age group.

Estimating Differences in Medians with Paired Samples

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")
```

Here we investigate estimating the difference in the *medians* between two *paired samples*.

With **paired samples**, we have a single set of observations, but two measures or two variables for each observation. Obtaining two or more measures from each observation to pair can result from the following situations:

1. **Across-time measures:** The same dependent variable is measured for each individual but at two different time periods. For example, a sample of individuals may have their income measures once in 2013 and again in 2014, and a researcher could ask whether there was a change in average income from one year to the next.
2. **Different conditions measures:** The same dependent variable is measured for each individual, but under two different conditions, or before and after some treatment. For example, a sample of high school students may have test scores measured before introducing them to a new curriculum and afterwards. A researcher could ask whether the curriculum affected test scores.
3. **Related topics measures:** Two slightly different variables are measured for each individual. For example, foreign language students may take separate exams for writing proficiency and speaking proficiency. A researcher could ask whether students are more proficient with writing in a foreign language versus speaking.

Testing differences in *medians* is appropriate when the variables are in the same units and at the ordinal, interval, or ratio scale.

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.

An alternative measure for the median is the **interpolated median**. This is another measure of center which takes into account the percentage of the data that is strictly below versus strictly above the median.

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 a variables `Grades` and `Sports`, which are each rankings on a scale of 1-4, with lower values indicating higher performance, for how much importance students place on good grades and on participating in sports, respectively. Let us begin by computing the **sample median** for each variable.

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

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

```
median(kidsdata$Sports)
```

```
## [1] 2
```

We see in the results above that the sample median response for the importance of grades is equal to 3, while the median response for sports is equal to 2. This means that *in our sample*, on average children place more importance on sports than on grades.

Similarly, we can compute the **sample interpolated median** for each variable:

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

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

```
interp.median(kidsdata$Sports)
```

```
## [1] 1.980519
```

Here we can see that the difference in the center of the distributions is somewhat smaller than impled by the simple median. The response for grades is centered somewhat below 3 (interpolated median = 2.67) and the response for sports is centered very close to 2 (interpolated median=1.98).

3. Hypothesis Test on Differences in the Center of the Distribution

The **Wilcoxon Signed Rank Test** considers the hypothesis that the distributions for two paired samples are centered around the same value. Using the example above, let us test the hypothesis that in the population the distribution for *importance of grades* is centered around the same value as the distribution for the importance of sports.

Null hypothesis: The center of the distribution for grades is *equal* to the center of the distribution for sports
Alternative hypothesis: The center of the distribution for grades is *different* than the center of the distribution for sports

The following code calls the `wilcox.test()` function to test this hypothesis:

```
wilcox.test(kidsdata$Grades, kidsdata$Sports, alternative="two.sided", paired=TRUE)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  kidsdata$Grades and kidsdata$Sports
## V = 77586, p-value = 4.085e-12
## alternative hypothesis: true location shift is not equal to 0
```

The first and second parameters `kidsdata$Grades` and `kidsdata$Sports` specify the two variables to compare. The third parameter, `alternative="two.sided"` specifies that this is a two-tailed test with an alternative hypothesis that the center of the two distributions are *different*, and the final parameter `paired=TRUE` tells the function to run a paired-samples test (versus an independent-samples test).

With a p-value less than 5%, we do *find sufficient statistical evidence* that the median response for the importance for grades is different than the median response for the importance of sports.