Estimating Differences in Medians - Independent Samples

James M. Murray, Ph.D.
University of Wisconsin - La Crosse

Updated: September 13, 2017

PDF file location: http://www.murraylax.org/rtutorials/two_medians.pdf

HTML file location: http://www.murraylax.org/rtutorials/two_medians.html

---

Note on required packages: The following code requires the `psych` package 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") # Only need to execute once per machine

library("psych") # Need to execute every time you load R
```

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.

1. Data

In this data set, students in fourth through sixth from three school districts in Michigan ranked their how important each of the following were to them: achieving good grades, having 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. Available at: http://shapeamerica.tandfonline.com/doi/abs/10.1080/02701367.1992.10608764

The code below downloads the data and assigns the it to a data frame that we call `kidsdata`.

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

2. Compute Medians

The data frame 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 Distributions

The **Mann Whitney U-Test**, or sometimes referred 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(Grades ~ Gender, data=kidsdata, alternative="two.sided")
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Grades by Gender
## W = 29388, p-value = 0.5373
```

```
## alternative hypothesis: true location shift is not equal to 0
```

The first parameter to the function call, `Grades ~ Gender`, is a *formula* that says we are interested in the outcome variable `Grades` and how it is different for different values of the explanatory variable, `Gender`. The second parameter, `data=kidsdata`, tells the function in what data frame it can find these variables. The final 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.*,

The p-value is equal to 0.5373 which is much above 0.05 or 5%. Therefore we *fail to find* statistical evidence that the median response for the importance for grades is different for boys versus girls.