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.