Non-parametric Methods

A statistical method is called non-parametric if it makes no assumption on the population distribution or sample size.

This is in contrast with most parametric methods in elementary statistics that assume the data is quantitative, the population has a normal distribution and the sample size is sufficiently large.

In general, conclusions drawn from non-parametric methods are not as powerful as the parametric ones.

However, as non-parametric methods make fewer assumptions, they are more flexible, more robust, and applicable to non-quantitative data.

Sign Test

A sign test is used to decide whether a binomial distribution has the equal chance of success and failure.

Example

A soft drink company has invented a new drink, and would like to find out if it will be as popular as the existing favorite drink. For this purpose, its research department arranges 18 participants for taste testing. Each participant tries both drinks in random order before giving his or her opinion.

Problem

It turns out that 5 of the participants like the new drink better, and the rest prefer the old one. At .05 significance level, can we reject the notion that the two drinks are equally popular?

Solution

The null hypothesis is that the drinks are equally popular. Here we apply the binom.test function. As the p-value turns out to be 0.096525, and is greater than the .05 significance level, we do not reject the null hypothesis.

```
> binom.test(5, 18)

        Exact binomial test

data:  5 and 18
number of successes = 5, number of trials = 18,
p-value = 0.09625
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.09695 0.53480
sample estimates:
probability of success
               0.27778
```

Answer

At .05 significance level, we do not reject the notion that the two drinks are equally popular.

Wilcoxon Signed-Rank Test

Two data samples are matched if they come from repeated observations of the same subject.

Using the Wilcoxon Signed-Rank Test, we can decide whether the corresponding data population distributions are identical without assuming them to follow the normal distribution.

Example

In the built-in data set named immer, the barley yield in years 1931 and 1932 of the same field are recorded. The yield data are presented in the data frame columns Y1 and Y2.

```
> library(MASS)        # load the MASS package
> head(immer)

  Loc Var    Y1    Y2
1  UF   M  81.0  80.7
2  UF   S 105.4  82.3
    .....
```

Problem

Without assuming the data to have normal distribution, test at .05 significance level if the barley yields of 1931 and 1932 in data set immer have identical data distributions.

Solution

The null hypothesis is that the barley yields of the two sample years are identical populations.

To test the hypothesis, we apply the wilcox.test function to compare the matched samples.

For the paired test, we set the "paired" argument as TRUE. As the p-value turns out to be 0.005318, and is less than the .05 significance level, we reject the null hypothesis.

```
> wilcox.test(immer$Y1, immer$Y2, paired=TRUE)

        Wilcoxon signed rank test with continuity correction

data:  immer$Y1 and immer$Y2
V = 368.5, p-value = 0.005318
alternative hypothesis: true location shift is not equal to 0

Warning message:
In wilcox.test.default(immer$Y1, immer$Y2, paired = TRUE) :
  cannot compute exact p-value with ties
```

Answer

At .05 significance level, we conclude that the barley yields of 1931 and 1932 from the data set immer are nonidentical populations.

Exercise

One-sample nonparametric tests

First, load the Hipparcos dataset and recall the variable names using the names function. By using attach, we can automatically create temporary variables with these names (these variables are not saved as part of the R session, and they are superseded by any other R objects of the same names).

```
#> hip <- read.table("http://astrostatistics.psu.edu/datasets/HIP_star.dat",
#+ header=T, fill=T)
> hip <- read.table("HIP_star.dat", header=T, fill=T)
> names(hip)
> attach(hip)
```

Let's take a look at the declination values using a boxplot:

```
> boxplot(DE, notch=T)
```

Recall that the notch=TRUE argument tells R to add "notches" to the box to indicate (by default) an approximate 95% confidence interval for the population median (actually, this CI is supposed to be used to test a difference of two medians, but it is still revealing in this one-sample case). Let's add a horizontal line at 0:

```
> abline(h=0, lty=2)
```

It appears that 0 is not in the CI, so let's see if we can reject the null hypothesis that the median is zero using a simple sign test. We need to count the number of declination measurements (out of 2719) that are greater than zero:

```
> sum(DE>0)
```

We see that there are 1419 such measurements. Because this is a two-sided test, the p-value will be two

times the probability that a fair coin flipped 2719 times will come up 1419 or more as "heads". This may be found using the binomial distribution function:

```
> 2*(1-pbinom(1418, 2719, .5)) # Do you see why we used 1418?
```

So the p-value is smaller than .05 and we can reject the null hypothesis that the median equals zero.

Next, we will perform a Wilcoxon signed rank test of the same null hypothesis:

```
> wilcox.test(DE)
```

What happened here? We rejected the null hypothesis using a more simplistic test that is known to be less powerful (the sign test) and we failed to reject using a more powerful test (the Wilcoxon signed rank test). What has happened here is the fact that the Wilcoxon, although it makes no parametric assumption, does assume that the distribution is symmetric around its median. This is a bizarre case in which the observations less than zero, though less numerous, tend to be slightly larger in absolute value. Thus, when we add the absolute ranks of the positive values, there are more of them but they tend to be smaller than they would for a symmetric distribution, and these two competing forces almost exactly balance out. Weird!

```
> hist(DE)                  # See the non-symmetry?
> sum(rank(abs(DE))[DE>0])  # Same as Wilcox test stat
> 2719*2720/4               # Expected value of stat under H0
```

Mann-Whitney-Wilcoxon Test

Two data samples are independent if they come from distinct populations and the samples do not affect each other.

Using the Mann-Whitney-Wilcoxon Test, we can decide whether the population distributions are identical without assuming them to follow the normal distribution.

Example

In the data frame column mpg of the data set mtcars, there are gas mileage data of various 1974 U.S. Automobiles.

```
> mtcars$mpg
[1] 21.0 21.0 22.8 21.4 18.7 ...
```

Meanwhile, another data column in mtcars, named am, indicates the transmission type of the automobile model (0 = automatic, 1 = manual). In other words, it is the differentiating factor of the transmission type.

```
> mtcars$am
 [1] 1 1 1 0 0 0 0 0 ...
```

In particular, the gas mileage data for manual and automatic transmissions are independent.

Problem

Without assuming the data to have normal distribution, decide at .05 significance level if the gas mileage data of manual and automatic transmissions in mtcars have identical data distribution.

Solution

The null hypothesis is that the gas mileage data of manual and automatic transmissions are identical populations. To test the hypothesis, we apply the wilcox.test function to compare the independent samples. As the p-value turns out to be 0.001817, and is less than the .05 significance level, we reject the null hypothesis.

```
> wilcox.test(mpg ~ am, data=mtcars)

        Wilcoxon rank sum test with continuity correction

data:  mpg by am
W = 42, p-value = 0.001871
alternative hypothesis: true location shift is not equal to 0

Warning message:
In wilcox.test.default(x = c(21.4, 18.7, 18.1, 14.3, 24.4, 22.8,  :
  cannot compute exact p-value with ties
```

Answer

At .05 significance level, we conclude that the gas mileage data of manual and automatic transmissions in mtcar are nonidentical populations.

Exercise

Two-sample nonparametric tests

In the exploratory data analysis and regression tutorial, we used exploratory techniques to identify 92 stars from the Hipparcos data set that are associated with the Hyades. We did this based on the values of right ascension, declination, principal motion of right ascension, and principal motion of declination. We then excluded one additional star with a large error of parallax measurement:

```
> filter1 <-  (RA>50 & RA<100 & DE>0 & DE<25)
> filter2 <- (pmRA>90 & pmRA<130 & pmDE>-60 & pmDE< -10)
> filter  <-  filter1 & filter2 & (e_Plx<5)
> sum(filter)
```

In this section of the tutorial, we will compare these Hyades stars with the remaining stars in the Hipparcos dataset on the basis of the color (B minus V) variable. That is, we are comparing the groups in the boxplot below:

```
> color <- B.V
> boxplot(color~filter,notch=T)
```

For ease of notation, we define vectors H and nH (for "Hyades" and "not Hyades") that contain the data values for the two groups.

```
> H <- color[filter]
> nH <- color[!filter & !is.na(color)]
```

In the definition of nH above, we needed to exclude the NA values.

Let us run a two-sample Wilcoxon rank sum test to compare the medians of H and nH. The assumption

of this test is that the *shapes* of the two distributions are the same, but their medians may be different. The null hypothesis is that the medians are equal.

```
> wilcox.test(H, nH, conf.int=T)
```

The confidence limits are obtained using the pairwise differences between H and nH. For details, see the Bauer (1972) reference in the R help file for wilcox.test. Of course we reject H0 with such a small p-value, but note that the estimated difference in location is not equal to the difference of the sample medians:

```
> median(H) - median(nH)
```

Instead, the estimator is the *Hodges-Lehmann* estimator of location, which is the median of all pairwise differences:

```
> median(outer(H, nH, "-"))
```

Notice how we used the outer "product" of two vectors but substituted the subtraction operation for the multiplication operation.

```
> sum(outer(H, nH, "-")>0)
```

Here is a quotation from the help file for wilcox.test in R:

> *The literature is not unanimous about the definitions of the Wilcoxon rank sum and Mann-Whitney tests. The two most common definitions correspond to the sum of the ranks of the first sample with the minimum value subtracted or not: R subtracts and S-PLUS does not, giving a value which is larger by m(m+1)/2 for a first sample of size m. (It seems Wilcoxon's original paper used the unadjusted sum of the ranks but subsequent tables subtracted the minimum.)*

So the statistic calculated above, 84531, is the sum of the ranks of the Hyades minus m(m+1)/2. We can check this easily:

```
> sum(rank(c(H, nH))[1:92]) - 92*93/2
```

Kruskal-Wallis Test

A collection of data samples are independent if they come from unrelated populations and the samples do not affect each other.

Using the Kruskal-Wallis Test, we can decide whether the population distributions are identical without assuming them to follow the normal distribution.

Example

In the built-in data set named airquality, the daily air quality measurements in New York, May to September 1973, are recorded.

The ozone density are presented in the data frame column Ozone.

```
> head(airquality)
  Ozone Solar.R Wind Temp Month Day
1    41     190  7.4   67     5   1
2    36     118  8.0   72     5   2
   .....
```

Problem

Without assuming the data to have normal distribution, test at .05 significance level if the monthly ozone density in New York has identical data distributions from May to September 1973.

Solution

The null hypothesis is that the monthly ozone density are identical populations.

To test the hypothesis, we apply the kruskal.test function to compare the independent monthly data.

The p-value turns out to be nearly zero (6.901e-06). Hence we reject the null hypothesis.

```
> kruskal.test(Ozone ~ Month, data = airquality)

        Kruskal-Wallis rank sum test

data:  Ozone by Month
Kruskal-Wallis chi-squared = 29.267, df = 4, p-value = 6.901e-06
```

Answer

At .05 significance level, we conclude that the monthly ozone density in New York from May to September 1973 are nonidentical populations.

Exercise

From Nieppola et al. 2011 (2011A&A..535..A69N) work, titled “Correlation between Fermi/LAT gamma-ray and 37 GHz radio properties of northern AGN averaged over 11 months”. Find if exist an identical data distributions into several AGNs class.