Contents

1 Non-Parametric Inference Procedures

Non-parametric inference procedures were developed to be used in cases when the distribution of the variable of interest in the population is known to be not-normal, or not known at all, and furthermore the distributional assumptions relevant to parametric statistical inference are undetermined (hence the name nonparametric).

Nonparametric tests are also referred to as ***distribution-free*** tests. These tests have the obvious advantage of not requiring the assumption of normality or the assumption of homogeneity of variance. They compare medians rather than means and, as a result, if the data have one or two outliers, their influence is negated.

Parametric tests are preferred because, in general, for the same number of observations, they are more likely to lead to the rejection of a false hull hypothesis. That is, they have more power. This greater power stems from the fact that if the data have been collected at an interval or ratio level, information is lost in the conversion to ranked data (i.e., merely ordering the data from the lowest to the highest value).

1.1 Non-Parametric Procedures

- Kolmogorov- Smirnov Test (`ks.test()`)
- Wilcoxon test (`wilcox.test()`)

1.2 Kolmogorov-Smirnov Test

For a single sample of data, the Kolmogorov-Smirnov test is used to test whether or not the sample of data is consistent with a specified distribution function. (Not part of this course) When there are two samples of data, it is used to test whether or not these two samples may reasonably be assumed to come from the same distribution. The null and alternative hypotheses are as follows:
H_0: *The two data sets are from the same distribution*
H_1: *The data sets are not from the same distribution*

Consider two sample data sets X and Y that are both normally distributed with similar means and variances.

```
X=rnorm(16,mean=20,sd=5)
Y=rnorm(18,mean=21,sd=4)
ks.test(X,Y)
```

```
> ks.test(X,Y)
        Two-sample Kolmogorov-Smirnov test

data:  X and Y
D = 0.2153, p-value = 0.7348
alternative hypothesis: two-sided
```

Remark: It doesnt not suffice that both datasets are from the same distribution. They must have the same value for the defining parameters. Consider the case of data sets; X and Z. Both are normally distributed, but with different mean values.

```
> X=rnorm(16,mean=20,sd=5)
> Z=rnorm(16,mean=14,sd=5)
> ks.test(X,Z)

        Two-sample Kolmogorov-Smirnov test

data:  X and Z
D = 0.5625, p-value = 0.0112
alternative hypothesis: two-sided
```

1.3 Wilcoxon Mann-Whitney Test

The Wilcoxon Mann-Whitney Test is one of the most powerful of the nonparametric tests for comparing two populations. It is used to test the null hypothesis that two populations have identical distribution functions against the alternative hypothesis that the two distribution functions differ only with respect to ***location*** (i.e. median), if at all.

The Wilcoxon Mann-Whitney test does not require the assumption that the differences between the two samples are normally distributed. In many applications, the Wilcoxon Mann-Whitney Test is used in place of the two sample t-test when the normality assumption is questionable. This test can also be applied when the observations in a sample of data are ranks, that is, ordinal data rather than direct measurements.

Remark - *Non-parametric procedures often give warning errors when tied values occur. In this module you may disregard these.*

1.4 Example

In the data frame column ***mpg*** of the ***mtcars*** data set , 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.

Exercise 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.

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.

```
> 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

..............
```

As the p-value turns out to be 0.001817, and is less than the .05 significance level, we reject the null hypothesis. At .05 significance level, we conclude that the gas mileage data of manual and automatic transmissions in mtcar are nonidentical populations.

1.4.1 Boxplot to accompany Wilcoxon Test Analysis

```
boxplot(mpg~am,data=mtcars,horizontal=TRUE,
    col=c("lightblue","lightpink"))
title("Miles Per Gallon by Transmission Type")
```