MA4605 - Lecture 2B part i- Nonparametric Tests and Outliers

Occasionally, the assumptions of the t-tests are seriously violated. In particular, if the type of data you have is ordinal in nature. On such occasions an alternative approach is to use non-parametric tests. Nonparametric tests may be, and often are, more powerful in detecting population differences when certain assumptions are not satisfied.

All tests involving ranked data, i.e. data that can be put in order, are nonparametric.

We are not going to place much emphasis on them in this unit as they are only occasionally used. But you should be aware of them and have some familiarity with them.

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).

Generally, running nonparametric procedures is very similar to running parametric procedures, because the same design principle is being assessed in each case. So, the process of identifying variables, selecting options, and running the procedure are very similar. The final p-value is what determines significance or not in the same way as the parametric tests.

The following table gives the non-parametric analogue for the paired sample t-test and the independent samples t-test. There is no obvious comparison for the one sample t-test

Parametric test	**Non-parametric analogue**
One-sample t-test	Nothing quite comparable
Paired sample t-test	Wilcoxon T Test
Independent two samples t-test	Wilcoxon-Matt-Whitney test
Pearson's correlation	Spearman's correlation

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 (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.

The ***Wilcoxon-Matt-Whitney test*** is also known as the ***Wilcoxon rank sum test*** and the ***Mann-Whitney U-test***. (There are some slight differences between the two, but we will progress any further with such matters)

To implement this test in R, we use the command `wilcox.test()`, specifying the namesof the relevant data sets.

```
> mean(Y);median(Y)
[1] 34.33333
[1] 37
> mean(Z);median(Z)
[1] 30.61538
[1] 30
> wilcox.test(Y,Z)

        Wilcoxon rank sum test with continuity correction

data:  Y and Z
W = 82.5, p-value = 0.8276
alternative hypothesis: true location shift is not equal to 0
```

Here we fail to reject the null hypothesis. We are able to assume that the population for both sample data sets have very similar measures of centrality.

Wilcoxon Signed Ranks Test

The Wilcoxon Signed Ranks test is designed to test a hypothesis about the location (***median***) of a population distribution. It often involves the use of paired data, for example, "*before and after*" data, in which case it tests for a median difference of zero.

In many applications, this test is used in place of the one sample t-test when the normality assumption is questionable. It is a more powerful alternative to the ***sign test***, but does assume that the population probability distribution is symmetric.

This test can also be applied when the observations in a sample of data are ranks, that is, ordinal data rather than direct measurements.

Sign Test

The sign test is designed to test a hypothesis about the location of a population distribution. It is most often used to test the hypothesis about a ***population median***, and often involves the use of matched pairs, for example, before and after data, in which case it tests for a median difference of zero.

The Sign test does not require the assumption that the population is normally distributed.

In many applications, this test is used in place of the one sample t-test when the normality assumption is questionable. It is a less powerful alternative to the Wilcoxon signed ranks test, but does not assume that the population probability distribution is symmetric.

This test can also be applied when the observations in a sample of data are ranks, that is, ordinal data rather than direct measurements.

Runs Test

In studies where measurements are made according to some well defined ordering, either in time or space, a frequent question is whether or not the average value of the measurement is different at different points in the sequence. The runs test provides a means of testing this.

Example

Suppose that, as part of a screening programme for heart disease, men aged 45-65 years have their blood cholesterol level measured on entry to the study. After many months it is noticed that cholesterol levels in this population appear somewhat higher in the Winter than in the Summer. This could be tested formally using a Runs test on the recorded data, first arranging the measurements in the date order in which they were collected.

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

The R command used to perform the Kolmogorov-Smirnov Test is `ks.test()`, with the names of the data sets specified as arguments. Consider the case of the data sets X,Y and Z.

X and Y are normally distributed with similar means and variances. The fact that both are normally distributed is merely incidental.

```
> ks.test(X,Y)

        Two-sample Kolmogorov-Smirnov test

data:  X and Y
D = 0.2797, p-value = 0.6437
alternative hypothesis: two-sided
```

We fail to reject the null hypothesis. Both data sets can be assumed to be from the same distribution.

Remark: It doesn't not suffice that both datasets are from the same type of distribution. They must have the same value for the defining parameters. In the case of the normal distribution that would be the normal mean μ and normal variance σ^2.

Consider the case of data sets; X and Z. Both are normally distributed, but with different mean values.

```
> ks.test(X,Z)

        Two-sample Kolmogorov-Smirnov test

data:  X and Z
D = 0.7692, p-value = 5e-04
alternative hypothesis: two-sided
```

Here, we reject the null hypothesis. Both data sets have different distributions.

The Kolmogorov-Smirnov test is a very useful tool in assessing comparability of an experimental group and a control group prior to a study being carried out.

Kruskal-Wallis Test

(For future reference. We have not seen the parametric equivalent yet.)

The Kruskal-Wallis test is a nonparametric test used to compare three or more samples. It is used to test the null hypothesis that all populations have identical distribution functions against the alternative hypothesis that at least two of the samples differ only with respect to location (median), if at all.

It is the analogue to the F-test used in analysis of variance. While analysis of variance tests depend on the assumption that all populations under comparison are normally distributed, the Kruskal-Wallis test places no such restriction on the comparison.

Outliers

An outlier is an observation that appears to deviate markedly from other observations in the sample.

Identification of potential outliers is important for the following reasons.

1. An outlier may indicate bad data. For example, the data may have been coded incorrectly or an experiment may not have been run correctly. If it can be determined that an outlying point is in fact erroneous, then the outlying value should be deleted from the analysis (or corrected if possible).
2. In some cases, it may not be possible to determine if an outlying point is bad data. Outliers may be due to random variation or may indicate something scientifically interesting. In any event, we typically do not want to simply delete the outlying observation.

However, if the data contains significant outliers, we may need to consider the use of ***robust statistical techniques***.

Grubbs' Test for Outliers

Grubbs' test is used to detect a single outlier in a univariate data set that follows an approximately normal distribution.
If you suspect more than one outlier may be present, it is recommended that you use other procedures (beyond the scope of this course). Grubbs' test is also known as the maximum normed residual test.

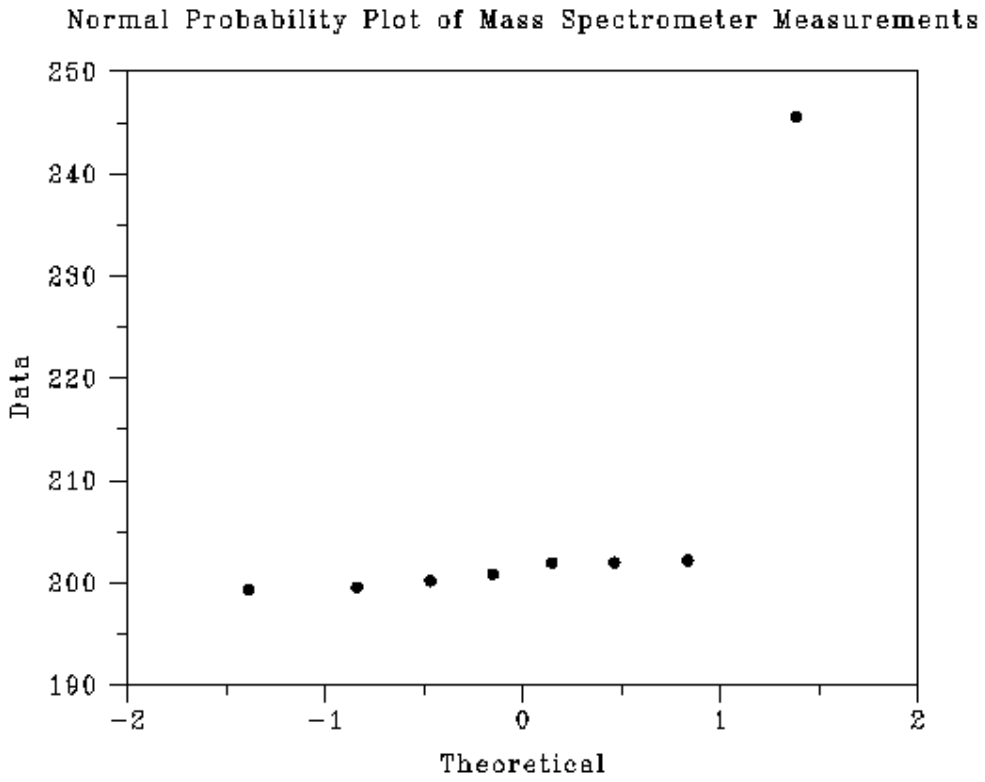
Grubbs' test is defined for the hypothesis:

H_0: There are no outliers in the data set
H_1: There is exactly one outlier in the data set

The Tietjen and Moore paper gives the following set of 8 mass spectrometer measurements on a uranium isotope:

199.31 199.53 200.19 200.82 201.92 201.95 202.18 245.57

As a first step, a normal probability plot was generated



This plot indicates that the normality assumption is reasonable with the exception of the maximum value. We therefore compute Grubbs' test for the case that the maximum value, 245.57, is an outlier.

```
> W=scan()
1: 199.31 199.53 200.19 200.82 201.92 201.95 202.18 245.57
9:
Read 8 items
>
> #Install Package "Outliers" to run Grubb's Test
> library(outliers)
> grubbs.test(W)

      Grubbs test for one outlier

data:  W
G = 2.4688, U = 0.0049, p-value = 1.501e-07
alternative hypothesis: highest value 245.57 is an outlier
```

For this data set, we reject the null hypothesis and conclude that the maximum value is in fact an outlier .

Remark: To generate the Q-Q plot

```
qqnorm(W,main="Normal Probability Plot of Mass Spectrometer Measurements")
```

Relevant to the end of year exam:

- Be able to explain the necessity for non-parametric methods.
- Be familiar with the Wilcox Mann Whitney Test.
- Be familiar with the Kolmogorov- Smirnov Two Sample Test.
- Be familiar with the Grubb's Outlier Test.