Statistics for Computing

MA4413 Lecture 10B

Kevin O'Brien

Kevin.obrien@ul.ie

Dept. of Mathematics & Statistics,
University *of* Limerick

Autumn Semester 2013

Inference Procedures with R

The paired t-test with R

- Previously we have seen the paired $t-$test. It is used to determine whether or not there is a significant difference between paired measurements.
- Equivalently whether or not the case-wise differences are zero.
- The mean and standard deviation of the case-wise differences are used to determine the test statistic.
- Under the null hypothesis, the expected value of the case-wise differences is zero (i.e $H_0 : \mu_d = 0$).

The paired t-test with R

- Conclusions about all inference procedures can be made based on the $p-$values.
- The p-values can be determined from computer code. (We will use a software called R. Other types of software inlcude SAS and SPSS.)
- The computer software automatically generates the appropriate test statistic, and hence the corresponding $p-$value.
- The user then interprets the $p-$values. If $p-$value is small, reject the null hypothesis. If the p-value is large, the appropriate conclusion is a failure to reject H_0.
- The threshold for being considered small is less than α/k, usually 0.0250.
  (This is actually a very arbitrary choice of threshold, suitable for some subject areas, not for others.)

The paired t-test with R

- In the following procedure (next slide), there are two sets of values: the `Before` values and the `After` values.
- The R command is `t.test()`, with the additional specification "`paired=`".
- The alternative hypothesis is specified in the output. (Another way of expressing it: True mean of case-wise differences is not zero)
- Also included in the output is a 95% confidence interval for the sample mean of case-wise differences.

The paired t-test with R

Implementing the paired t-test using R for the example previously discussed.

```
> t.test(Before,After,paired=TRUE)

        Paired t-test

data:  Before and After
t = 3.8881, df = 13, p-value = 0.001868
alternative hypothesis: true difference in means is not 0
95 percent confidence interval:
  3.650075 12.778496
sample estimates:
mean of the differences
                8.214286
```

The paired t-test with R

- The p-value (0.001868) is less than the threshold is less than the threshold 0.0250.
- We reject the null hypothesis (mean of case-wise differences being zero, i.e. expect no difference between “before” and “after”).
- We conclude that there is a likely to be a difference between ‘before’ and ‘after’ measurements.
- That is to say, we can expected a difference between two paired measurements.

The paired t-test with R

- We could also consider the confidence interval. We are 95% confident that the expected value of the case-wise difference is at least 3.65.
- Here the null value (i.e. 0) is not within the range of the confidence limits.
- Therefore we reject the null hypothesis.

```
> t.test(Before,After,paired=TRUE)
...
...
95 percent confidence interval:
  3.650075 12.778496
...
```

Test for Equality of Variance

- In this procedure, we determine whether or not two populations have the same variance.
- The assumption of equal variance of two populations underpins several inference procedures.
- The null and alternative hypotheses are as follows:

$$H_0 : \sigma_1^2 = \sigma_2^2$$

$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

Test for Equality of Variance

- When using R it would be convenient to consider the null and alternative in terms of variance ratios.
- Two data sets have equal variance if the variance ratio is 1.

$$H_0 : \sigma_1^2/\sigma_2^2 = 1$$
$$H_1 : \sigma_1^2/\sigma_2^2 \neq 1$$

Test for Equality of Variance

The test statistic is the ratio of the variances for both data sets.

$$TS = \frac{s_x^2}{s_y^2}$$

The standard deviations are provided in the R code.

- Sample standard deviation for data set $x = 3.40$
- Sample standard deviation for data set $y = 4.63$

To compute the test statistic.

$$TS = \frac{3.40^2}{4.63^2} = \frac{11.56}{21.43} = 0.5394$$

Test for Equality of Variance

```
> var.test(x,y)

        F test to compare two variances

data:  x and y
F = 0.5394, num df = 9, denom df = 8, p-value = 0.3764
alternative hypothesis:
 true ratio of variances is not equal to 1
95 percent confidence interval:
 0.1237892 2.2125056
sample estimates:
ratio of variances
         0.5393782
```

Test for Equality of Variance

- The $p-$value is 0.3764 (top right), above the threshold level of 0.0250.
- We fail to reject the null hypothesis.
- We can assume that there is no significant difference in sample variances. Therefore we can assume that both populations have equal variance.
- Additionally the 95% confidence interval (0.1237, 2.2125) contains the expected value under the assumption of equal variance i.e. 1.

Shapiro-Wilk Test for Normality

- We will often be required to determine whether or not a data set is normally distributed.
- Again, this assumption underpins many statistical models.
- The null hypothesis is that the data set is normally distributed.
- The alternative hypothesis is that the data set is not normally distributed.
- One procedure for testing these hypotheses is the Shapiro-Wilk test, implemented in R using the command `shapiro.test()`.

Shapiro-Wilk Test for Normality

For the data set used previously; x and y, we use the Shapiro-Wilk test to determine that both data sets are normally distributed. (Both data sets can be assumed to be normally distributed, based on these values.)

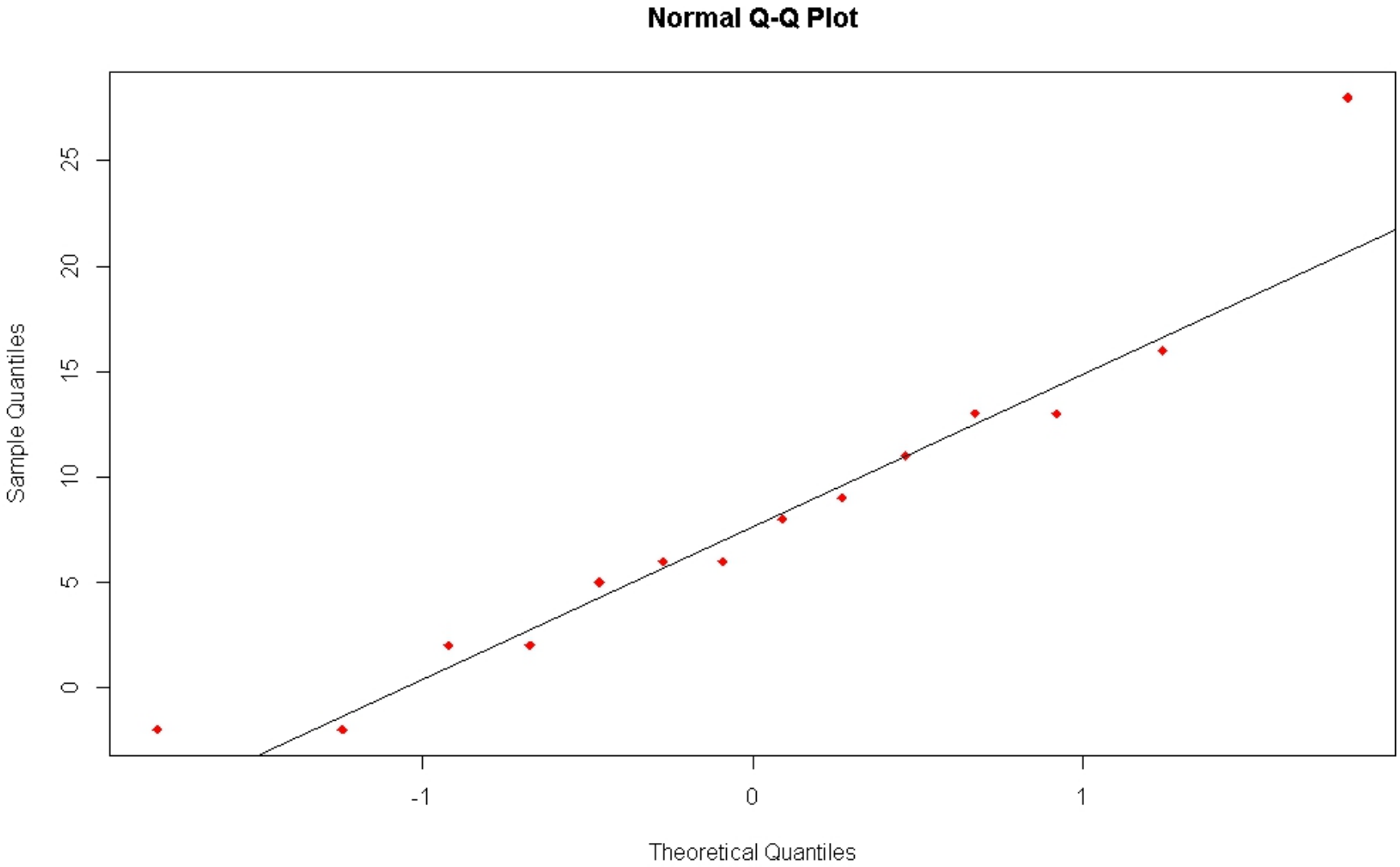
```
> shapiro.test(x)
        Shapiro-Wilk normality test

data:  x
W = 0.9474, p-value = 0.6378
>
> shapiro.test(y)
        Shapiro-Wilk normality test

data:  y
W = 0.9347, p-value = 0.5273
```

Graphical Procedures for assessing Normality

- The normal probability (Q-Q) plot is a very useful tool for determining whether or not a data set is normally distributed.
- Interpretation is simple. If the points follow the trendline (provided by the second line of `R` code `qqline`), then the data set can be assumed to be normally distributed.
- One should expect minor deviations. Numerous major deviations would lead the analyst to conclude that the data set is not normally distributed.
- The Q-Q plot is best used in conjunction with a formal procedure such as the Shapiro-Wilk test.

```
>qqnorm(CWdiff)
>qqline(CWdiff)
```

Graphical Procedures for Assessing Normality



Grubbs Test for Determining an Outlier

- The Grubbs test is used to determine if there is exactly one outlier in a complete data set.
- (This is actually the most common of several variants of the Grubbs' test).
- Importantly, the test requires that the data is normally distributed to be a valid approach.
- There is no agreed formal definition for an outlier. The definition of outlier used for this procedure is a value that unusually distance from the rest of the values.
- (For the sake of clarity , we shall call this type of outlier a **Grubbs Outlier**).

Grubbs Test for Determining an Outlier

Consider the following data set: is the lowest value (4.01) an outlier?

```
6.98 8.49 7.97 6.64
8.80 8.48 5.94 6.94
6.89 7.47 7.32 4.01
```

Under the null hypothesis, there is no outlier present in the data set. We reject this hypothesis if the $p-$value is sufficiently small.

Grubbs Test for Determining an Outlier

```
> grubbs.test(x, two.sided=T)
Grubbs test for one outlier
data: x
G = 2.4093, U = 0.4243, p-value = 0.05069
alternative hypothesis: lowest value 4.01 is an outlier
```

We conclude that while small by comparison to the other values, the lowest value 4.01 is not an outlier.