Statistics for Computing

MA4413 Lecture 10A

Kevin O'Brien

Kevin.obrien@ul.ie

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

Autumn Semester 2013

Today's Class : Almost Completion of Inference Procedures

- One-tailed Hypothesis Testing Procedures
- The R Programming Language
- Inference Procedures with R
- Binary Classification Predictive Models

Next Class: A few more R procedures.
(Remark : Will not be covering *Sample Size estimation* to same extent as previous years.)

One Tailed Inference Procedures

- We will briefly look at one-tailed hypothesis.
- One tailed confidence intervals do exist but are rarely used in practice.
- Some procedures will apply corrections factors to their estimate, and are not symmetric. Although they are more or less two-tailed.

Number of Tails

Inference Procedures are either **One-tailed** or **Two-Tailed**.
Confidence Intervals are almost always two-tailed (in undergraduate statistics anyway). However Hypothesis tests can either be one-tailed or two-tailed. It is important to know how determine correctly the number of tails.

- The alternative hypothesis indicates the number of tails.
- A rule of thumb is to consider how many alternative to the H_0 is offered by H_1.
- When H_1 includes either of these relational operators;'$>$' ,'$<$' , only one alternative is offered.
- When H_1 includes the $\neq$ relational operators, two alternatives are offered (i.e.'$>$' or '$<$').

One Tailed Hypothesis test

- A one-sided test is a statistical hypothesis test in which the values for which we can reject the null hypothesis, H_0 are located entirely in one tail of the probability distribution (either the upper tail, or the lower tail, but not both).
- One tailed procedures are a more intuitive approach when determining if a certain values exceeds a certain threshold. (Recall the election question used in the midterm)
- Equivalently One-tailed tests are useful when determining if the population mean (or proportion) for one group is greater than that of another group.

One Tailed Hypothesis test

- In other words, the "critical region" for a one-sided test is the set of values beyond than the critical value for the test
- Again A rule of thumb is to consider the alternative hypothesis. If only one alternative is offered by H_1 (i.e. a ' $<$' or a ' $>$' is present, then it is a one tailed test.)
- (When computing quantiles from Murdoch Barnes table 7, we set $k = 1$)

Binary Classification

- A Binary classification procedure is a predictive approach used to classify cases as either one two possible outcomes.
- Almost always - there is some sort of hypothesis test taking place to make this prediction.

Binary Classification

Recall the possible outcomes of a hypothesis test procedure. In particular recall the two important types of error. Importantly the binary classification prediction procedure can yield wrong predictions.

	Null hypothesis (H_0) true	Null hypothesis (H_0) false
Reject null hypothesis	**Type I error** False positive	**Correct outcome** True positive
Fail to reject null hypothesis	**Correct outcome** True negative	**Type II error** False negative

Accuracy, Precision and Recall

Let us simplify the last table, and present it in the context of a binary prediction procedure.

	Predicted Negative	Predicted Positive
Observed Negative	True Negative	False Positive
Observed Positive	False Negative	True Positive

Accuracy, Precision and Recall

Important metrics for determining how usefulness of the prediction procedure are : Accuracy, Recall and Precision.
Accuracy, Precision and recall are defined as

$$\text{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn}$$

$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

Accuracy, Precision and Recall

Another measure is the F-measure. The F measure is computed as

$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Questions

	Predicted Negative	Predicted Positive
Negative Cases	TN: 9,700	FP: 165
Positive Cases	FN: 35	TP: 100

Accuracy, Precision and Recall

With reference to the table above, compute each of the following appraisal metrics. 2

a. Accuracy

b. Precision

c. Recall

d. F measure

Accuracy, Precision and Recall

- Why is the accuracy value so high?
- Why is the F-measure so low?
- This is the class-imbalance problem: more“negative” outcomes which skews the statistic, but these outcomes are the least relevant.
- F-measure disregards the irrelevant “true negatives, and concentrates on the more relevant potential outcomes.

R Programming Language

What is R?

- R is a suite of software facilities for
  - reading and manipulating data,
  - computation,
  - conducting statistical analyses,
  - displaying the results.
- open-source version (i.e. freely available version - no license fee) of the S programming language, a language for manipulating objects.
- a programming environment for data analysis and graphics
- a platform for development and implementation of new algorithms
- Software and packages can be downloaded from www.cran.r-project.org

p-values using R

- In every inference procedure performed using R, a p-value is presented to the screen for the user to interpret.
- If the p-value is larger than a specified threshold α/k then the appropriate conclusion is a failure to reject the null hypothesis.
- Conversely, if the p-value is less than threshold, the appropriate conclusion is to reject the null hypothesis.
- In this module, we will use a significance level$\alpha = 0.05$ and almost always the procedures will be two tailed ($k = 2$). Therefore the threshold α/k will be 0.025.

Using Confidence Limits

- Alternatively, we can use the confidence interval to make a decision on whether or not we should reject or fail to reject the null hypothesis.
- If the null value is within the range of the confidence limits, we fail to reject the null hypothesis.
- If the null value is outside the range of the confidence limits, we reject the null hypothesis.
- Occasionally a conclusion based on this approach may differ from a conclusion based on the p-value. In such a case, remark upon this discrepancy.

The paired t-test (a)

- 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 test statistic is computed as

$$TS = \frac{\bar{d} - \mu_d}{\frac{s_d}{\sqrt{n}}}$$

The Paired t-test (b)

- The calculation is dependent on the case-wise differences.
- Here the case-wise differences between paired measurements (e.g. “before” and “after”).
- Under the null hypothesis, the mean of case-wise differences is zero.
- As a quick example, the mean, standard deviation and sample size are presented in the next slide.

The paired t-test (c)

- Observed Mean of Case-wise differences $\bar{d} = 8.21$,
- Expected Mean of Case-wise differences under $H_0 : \mu_d = 0$,
- Standard Deviation of Case-wise differences $S_d = 7.90$,
- Sample Size $n = 14$.

$$TS = \frac{\bar{d} - \mu_d}{\frac{s_d}{\sqrt{n}}} = \frac{8.21 - 0}{\frac{7.90}{\sqrt{14}}} = 3.881$$

The paired t-test (e)

- Procedure is two-tailed, and you can assume a significance level of 5%.
- It is also a small sample procedure (n=14, hence df = 13).
- The Critical value is determined from statistical tables (2.1603).
- Decision Rule: Reject Null Hypothesis ($|TS| > CV$). Significant difference in measurements before and after.

The paired t-test (f)

Alternative Approach : using p-values.

- The p-values are determined from computer code. (We will use a software called `R`. Other types of software inlcude `SAS` and `SPSS`.)
- The null and alternative are as before.
- 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 a very arbitrary choice of threshold, suitable for some subject areas, not for others)

The paired t-test (g)

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 (h)

- 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 difference between ‘before’ and ‘after’.
- That is to say, we can expected a difference between two paired measurements.

The paired t-test (i)

- 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 (a)

- 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. This assumption is tested by comparing the variance of samples taken from both populations.
- We will not get into too much detail about that, but concentrate on how to assess equality of variance.
- 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 (b)

- 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(c)

You would be required to compute the test statistic for this procedure. The test statistic is the ratio of the variances for both data sets.

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

The standard deviations would be 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.

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

Variance Test (d)

```
> 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
```

Variance Test (e)

- 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 null values i.e. 1.

Shapiro-Wilk Test(a)

- 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()`.
- (Remark: You will not be required to compute the test statistic for this test.)

Shapiro Wilk Test(b)

For the data set used previously; x and y, we use the Shapiro-Wilk test to determine that both data sets are normally distributed.

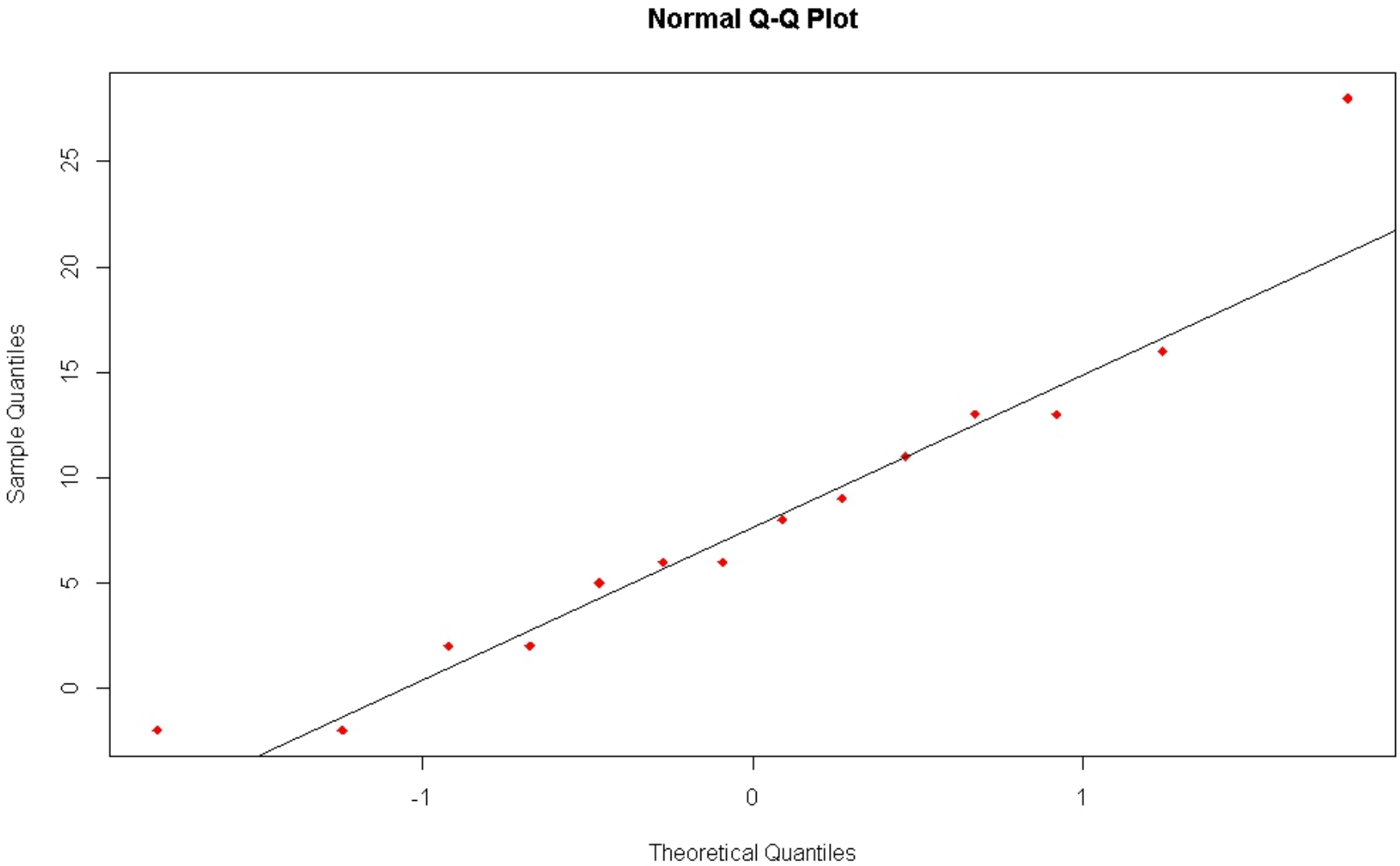
```
> shapiro.test(x)

        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`).
- 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 are any outliers in a data set.

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**). 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 are no outliers 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.