MT2508: Statistical Inference

Course Notes: 2015/2016

Prepared by Michail Papathomas
(Includes material by Ruth King and Steve Buckland)

University of St. Andrews

School of Mathematics and Statistics

MT2508 Structure - 2015-2016

Lecturers: Dr Michail Papathomas

Format:

- 2 lectures per week (odd weeks) on Tuesday and Thursday at 12.00 in Lecture Theatre D
- 3 lectures per week (even weeks) on Monday, Tuesday and Thursday at 12.00 in Lecture Theatre D
- 5 tutorials (one every odd week, starting in week 3) - see notice board and MMS - compulsory! Tutorial sheets to be handed out in even weeks (starting in week 2) and work to be handed in by 1pm the following Monday. The tutorial sheet will be composed of assessed questions, additional questions and past paper question. There will typically be 3-4 assessed questions and these will count towards the final mark for the course. You are encouraged to look at and/or hand in solutions to *additional* questions. This will improve understanding and aid learning in the future.
- 5 Example classes (one every even week, starting in week 2)
- 1 Microlab per week starting in week 2. See MMS and notice board for more details.

Assessment:

- 70% Examination (closed book) in May - compulsory questions (i.e. no choice - common to other Level 2000 modules).
- 30% Continuous Assessment:
  - Computer project performing statistical analysis using R: 15%
  - Fortnightly assessed tutorial questions: 5 x 3% = 15%

Academic Alerts:

- Tutorials are compulsory. If unable to attend e-mail tutor and explain why. Three failures to attend tutorials can result in a 0X mark awarded.
- Failure to hand in work constituting to at least 10% of the overall mark for the course may result in a 0X mark being awarded.

Note: checks will be made for plagiarism for work handed in. Students should not copy another's work and submit it as their own. If a case is detected then this is a very serious issue and those involved will be subject to the University's Academic Misconduct Policy. For example, a first offense often results in a 0 mark for the piece of work, a formal placement on a University register of offenders and a requirement to attend training relevant to plagiarism. Any further offense will be dealt with significantly harsher. Note that anyone giving their work to another to copy is just as guilty of academic misconduct and are likely to receive the same sanction. This does not mean that you cannot discuss important concepts and lecture material with colleagues, but all work must be written up independently!

Syllabus:

- Difference between population and sample;
- Unbiased and consistent estimators; Sample mean and variance as estimates of population mean and variance; sample covariance and correlation.
- Confidence intervals, interval estimation.
- Likelihood and maximum likelihood estimation: Discrete data and examples (sequence of binary trials, Poisson counts all with the same mean, Poisson with mean a function of a covariate); continuous data and example (n observations from $N(\mu, \sigma^2)$, m.l.e.s of mean and variance); invariance of m.l.e.s

- Basic properties of Normal distributions, Central Limit Theorem (statement and application to binomial and Poisson), assessing normality (normal scores).

- Hypothesis testing and interval estimation for normal distributions with known and unknown σ^2; χ^2, t and F distributions and their basic properties; one-sample t-test, paired t-test; two-sample t-tests and confidence intervals for means of normal distributions; F-tests for equality of variances of normal distributions; permutation tests: 2-sample permutation test; permutation test for matched pairs and one-sample test; randomization tests.

- Simple linear regression: least squares estimators, normal linear regression, regression in R, checking assumptions.

- A brief intorduction to multiple linear regression.

Reading list:

- John A. Rice, Mathematical Statistics and Data Analysis, Belmont, CA: Brooks/Cole CENGAGE, 2007.

- Richard D. De Veaux, Paul F. Velleman & David E. Bock, Stats: Data and Models, Pearson/Addison Wesley, 2005.

Statistics

Essentially, Statistics is the particular branch of Mathematics that involves the development of methodology for making inferences after observing data, whilst the associated uncertainty is evaluated. Through Statistics we answer questions using scientifically accepted methodologies. This allows for a consensus on the validity of the answers and inferences. Historically, the area of Statistics has had a "bad press". This is generally a result of misapplying or misunderstanding Statistics. Although sometimes this is done deliberately (for example when 'creative accounting' methods are used), it is usually accidental. It has been documented that our intuition when it comes to recognizing and handling uncertainty is often wrong. Thus, an understanding of the basic principles is essential to perform the appropriate data analysis and correctly interpret the results obtained.

Understanding and training in Statistics is a very desirable skill in the job market including:

- Government
- Pharmaceutical companies - clinical trials
- Banking and actuarial
- Public Health - Epidemiology
- Academia (Not only Maths/Stats departments; e.g. Comp. Science, Medical, Psychology, Biology etc.)
- Sports betting
- and many more...

0.1 Key Concepts

Statistics is the branch of Mathematics that deals with the analysis of data. However, the data collected (via some experiment or survey) typically varies if the experiment is repeated, e.g. length of time waiting for a bus; number of students obtaining a grade of 15+ on an exam; length of time a new Internet company takes to go bankrupt. Statistics is designed to deal with the variability that we may observe in the collected data and draw from it sensible conclusions. From an academic point of view, research on statistics can be either methodological, where the mathematical challenges can be as difficult as in pure mathematics, or it can be applied, where the aim is to solve difficult practical problems and apply in the best possible way the methodological developments.

Currently, there are two main scientific approaches for performing statistics. The classical (also called frequentist) approach is still the most popular, adopted and used by the majority of practitioners over the last century. Bayesian statistics (see MT4531 or MT5831) is becoming more and more popular over the last few decades and is now adopted by roughly 50% of academic statisticians. MT2508 will introduce the classical statistical approach. The main concepts we will deal with are:

0.1.1 Populations and samples

Often we observe data in the form of a sample, from which we wish to make inferences about the population. For example, for predicting results in general elections, take a sample of the voting intentions of 10000 citizens, representative of the population.

0.1.2 Variability and the replication principle

Suppose we are interested in elections as above and did a survey. If we repeat the survey and ask a different set of 10000 citizens we will, in general, obtain (slightly) different data that will lead to (slightly) different estimates. Can one estimate be regarded as more precise than the other? If we were to ask 1 million citizens then our estimate would be more accurate. Do we need, however, to consider taking a sample of size greater than 10000?

0.1.3 Confidence Intervals

Given observed data we estimate particular characteristics of interest (e.g. voting intentions). However, there will always be some error associated with the estimate, since we only observe a sample from the population. A confidence interval is an interval that includes the true population value with some (typically high) probability. *Note - the interval is dependent on the observed data and so will typically vary with different observed samples.*

0.1.4 Hypothesis Testing

With a statistical hypothesis we test a hypothesis about a population using only a sample. E.g. 10% of St. Andrews students are left-handed; more than 50% of St Andrews students know what is a random variable. To perform the test, we take a representative sample and decide whether the associated evidence would lead us to reject or not reject the hypothesis.

1 Statistical Inference

MT2504 focused on probability theory, i.e. how to measure our uncertainty on some quantity in a coherent manner, using random variables, and the properties of random variables (rv's). In MT2508 we extend this to statistical inference, i.e. how to make inferences and measure our uncertainty on some population characteristic after observing some relevant measurements (data) from a subset of the population (sample). Data are usually collected in two ways. The first is an observational study (survey), where the scientist just records measurements without any control over the manner in which the measurements were produced. Think, for example, of surveys via the telephone and internet, or the daily recording of stocks and shares prices. The second is a designed experiment, where the scientist controls the conditions that will create each measurement. Think, for example, of medical trials, where specific drugs are allocated in a specific manner to patients considering their characteristics (age, gender, etc.), and then the effect of the drugs is recorded.

The research study process usually consists of 6 stages.

- Define objectives and draft research question(s)
- Design of the study and plan of statistical analysis
- Collect the data (conduct the survey or experiment)
- Create the database
- Perform statistical analysis
- Report results in a comprehensive manner

This introductory module is mostly concerned with the last two stages of the research study process. This is because the first four stages can be better thought through when there is already knowledge of the basic statistical methodology and tools. Other modules are concerned with other stages; for example MT4614 'Design of Experiments', or MT4608 'Sampling Theory'.

1.1 Populations and samples

To make inferences on some population characteristic, we first consider a probability distribution that describes the variability in the measurements from the experiment or the observational study. Then, we make inferences on the distribution parameter that corresponds to the characteristic of interest. If the sample is representative of the whole population, then our inferences extend to the whole population.
Example: We want to estimate the average the average monthly accommodation expenditure for a St Andrews student. Let X_i represent the expenditure of student i. We assume that $X_i \sim N(\mu, 400)$ is a distribution that describes well the variability of expenditures in pounds. So, for every student i, before we measure their expenditure, we assume that $X_i \sim N(\mu, 400)$. We record the expenditure of 6 students, and observe $(x_1, x_2, x_3, x_4, x_5, x_6)$=(400,400,500,500,600,600). Based on these measurements we could infer that μ is close to the average of the observations, i.e. 500.

In the example above, if the measurements are only from PhD students, the sample is not representative of all students in St Andrews. In observational studies, a very good sampling method is 'Simple Random Sampling (SRS)', where every member of the population is equally likely to be selected. However, as this approach can be expensive and difficult to implement, other methods are often adopted. In 'Stratified Sampling', the population is divided in different groups and SRS takes place within each group. In 'Cluster Sampling', the population is divided into (geographical) clusters, and some clusters are chosen at random so that within cluster subjects are chosen with SRS. Finally, in 'Quota sampling' sample until enough subjects are measured to fulfill a certain plan.

Badly conducted sampling can introduce bias. Examples of dubious sampling are,

- 'Self-selecting Sampling' where the members of the population decide themselves if they will provide observations (e.g. internet surveys).
- 'Convenience Sampling' where observations are only provided by easily available members of the population (e.g. you only ask the accommodation expenditure of your flatmates).
- 'Judgemental Sampling' where observations are only provided by members of the population you think are trustworthy.