Data Analysis and Statistical Methods
Statistics 651

http://www.stat.tamu.edu/~suhasini/teaching651/

Lecture 1 (MWF)

Suhasini Subba Rao
Summary of Handout

• Strongly recommended book: ‘Statistical Methods and Data Analysis’, by Ott and Longnecker. An online version is available, contact me for more information.


• Grades:
  – Every week or two I will set homework. This will be worth 10% of your final grade.
  – Midterm: 25% (after Spring Break, we will decide closer to the time)
  – Assignment: 25% (set in the middle of April)
  – Final Exam: 40% (7th May, Room 150, between 15:30-17:30).
Overview of class

• Understand how to interpret probabilities - what they are and how the are calculated. In a couple of weeks, we will do some probability calculations, but the our goal will be interpretation.

• Understand how meaningful information about the data and what it tells us about the population.

• Understand the correct procedure to apply and how to interpret results, and whether they are meaningful or not.

• If statistics is done properly you cannot draw crazy conclusions:
  – Example: I want to see if there is any evidence that there is any evidence that stents (a medical device) reduces the risk of strokes in
at risk patients. 451 at risk patients volunteer for the study. They are randomly assigned to a treatment (given stents) and control group (not given stents, but treated the same as the treatment group). Of the 224 in the treatment group 45 went on to get a stroke within the year. Of the 227 in the control group 28 went on to get a stroke within the year.

\[
\text{proportion in treatment group who had a stroke} = \frac{45}{224} = 0.2
\]

\[
\text{proportion in control group who had a stroke} = \frac{28}{227} = 0.12
\]

Clearly, there is nothing in the data to suggest that stents reduces the risk. And any statistical analysis would back this conclusion.

- Statistics is not about putting numbers into formulas (computers can do this). But on understanding and interpreting data.
• I will spend the first half of the course on the properties of the sample mean (average).

The focus will be on the sample mean, because the more advanced material simply builds on this. For this part of the course we will do a lot of calculations by hand. This is to facilitate our understanding of the underlying concepts.

• In the second half of the course we will cover several different methods in more complex situations. We will rely on JMP statistical software to do the calculations. We will use the ideas we have learnt in the first half of the course to understand and interpret the statistical output.

• JMP is free to you, to install it go to [http://www.stat.tamu.edu/jmpinstall/jmp13/](http://www.stat.tamu.edu/jmpinstall/jmp13/). Full instructions are in the handout.

• I will sometimes use Statcrunch to illustrate certain ideas (such as
distributions). It is not compulsory, however it may be useful. If you want to use it, it will cost 13.75$ and can be obtained at http://www.statcrunch.com/
What is statistics?

The main idea:

- Statistics is used in a multitude of disciplines. We use it to make medical diagnostics, on recommendation websites (Netflix, Facebook, Amazon)....the list goes on.

- The applications are diverse, but the premise is the same. We do not observe the whole picture and only have partial information. Is it possible to say anything about the population based on very limited information?

- The answer is yes, so long as the sample has been collected in the correct way. However, we cannot get definitive answers about the population but only make ‘confident statements’ (intelligent guesses).

- In statistics, we want to understand something about a population of individuals. However, we are unable to observe the entire population (it
is too costly, too large, plain impossible) and instead we only observe a subset or a sample from that population.

• One of the main aims in statistics is to draw conclusions on an unobserved population based on a substantially smaller sample (subset of the population) that is observed.

• A good method for learning in this course is to keep in mind examples which are meaningful to you.
Interpreting probabilities: Paul the psychic octopus

Paul the octopus lived in a tank at a Sea Life Centre in Oberhausen, Germany (sadly he passed away). During the FIFA (football) world cup in summer 2010 he was used to predict the outcomes of all matches that Germany played.

He correctly predicted the outcome of 7 consecutive football matches. There are various scientific explanations on how he did this (related to the experiment and the colour of flags). But let us consider this statistically:

- Suppose that Paul is not psychic and his predictions were purely by chance.

- Under the assumption that Paul is not psychic, the probability of his correctly predicting the outcome of any Germany’s football matches is 0.5.
Therefore the probability of his correctly predicting the outcome of 7 matches in a row is 
\[ 0.5 \times 0.5 \times \ldots \times 0.5 = (0.5)^7 = 0.0078 = 0.78\% \] (we will learn how to do this calculation later).

- This is a relatively small probability, but it is not so small for it to be improbable and to suggest that the octopus was psychic.
- Further, during the world cup there were probably thousands of psychic animals predicting the outcome of football matches. Suppose none of these animals were psychic. We would expect that on average out of 1000 non-psychic predicting animals, about \( 1000 \times 0.0078 = 7.8 \) to correctly predict the outcome, without any psychic ability. Of course, only the animals that get the correct prediction are reported. This is an example if biased reporting.

These type of arguments will form the basis of hypothesis testing which we will cover later in the course.
Don’t be fooled by anecdotal evidence

- Sometime ago I gave an exam and one of the questions asked whether cell phone use in a car increased average reaction times. Given a data set (sample from a population), the students if the data suggested that cell phone use increased reaction times in drivers. Based on the given data, Student A showed that there was evidence to suggest that reaction time did indeed increase with cell phone use.

- Student A answered the question correctly. However, at the end of the solution Student A made the comment “I don’t believe the conclusion of the test, using a cell phone does not effect my driving”.

- Whether the student’s driving is good or bad is subjective, but despite doing the question correctly Student A had not understood the basic ideas behind statistics.
There are two problems with this Student A’s conclusion
– The data is based on just one example (too small to come up with any meaningful conclusion).
– More importantly, it not clear whether the data is representative of the population. It may be note worthy simple because it is not representative.

• Data collected in such a way is called anecdotal evidence.

• We can only draw meaningful statistical conclusions from samples that are random sampled (and to be reliable they cannot be too small).
Example 1 - Sparkling water

• There has been an intense discussion about the role that soda has on obesity.

• It is widely believed that the sugar in the drink plays an important role. Recently, the role that artificial sweeteners in the soda have on obesity has been investigated.

• A recent paper published in “Obesity Research and Clinical Practice” suggests that the CO2 in the water may also play a role. [https://www.stat.tamu.edu/~suhasini/C02beverage.pdf](https://www.stat.tamu.edu/~suhasini/C02beverage.pdf).

• The aim is to understand the statistical analysis behind this paper by the end of this course. But we now summarize their experiment.
• 16 rats (born on the same day), were randomly assigned to one of four groups. Each group was given one of four drinks
  – Regular water (Water).
  – Regular carbonated beverage made with sugar (CB)
  – Diet carbonated beverage (DGB)
  – Degassed carbonated beverage made with sugar (DgCB).

• They were not given any other drink besides the one in the prescribed group.

• The rats had unlimited access to rodent food (it was not restricted).
• Below is a plot of the average weight in each group after 110 days. The height of the bars is the average weight in each group.

![Graph showing average weight comparison between different groups]

• We see there are differences. As expected the weight of the rats consuming regular soda is higher than those on water. But those consuming Diet soda is also higher. Surprisingly those consuming the flat regular soda was less.

• But remember these are the averages of just 4 rats. Are these differences real?
– Are the differences there because of differences in the drink.
– Or can the differences be explained by chance.

• The larger the difference the less believable it is due to chance.

• But it is not just the difference that matters, it is the spread of the data too.

• Are the differences we see so large that we can reject the notion that it was due to chance?

• To answer this question we require statistical tools which we will develop in this course.
You may be curious as to why there was a weight gain. The following two plots may offer some clues.

![Graph A: Average food consumption, g/day](image)

![Graph B: Total ghrelin, units](image)
Example II - Climate change

- There has been a lot of speculation about ‘dramatic changes’ in the climate. To see if there is any merit to these conjectures, climatological data is collected (depending on the source, these can go back hundreds and even thousands of years). Statistical models are fitted to the data and ‘hypotheses’ about changes in the climate are tested.

- Data has been collected over the past 100 hundred years. From this we see that the average temperature this decade is greater than the average temperature 60 years ago. The question is, whether this change in the average is due to random variation (simply chance/coincidence), or whether there is a real change which cannot be explained well by random variability.

- Statistical analysis strongly suggests that there evidence to support the view that the change seen is more than just random variability.
What is random chance?

• We come across the word random all the time. You will often hear the word random in every day conversation.

  – Example Sam, a student at A&M, went on holiday. Later they tells a friend OMG, I saw Jon walk past on holiday, how random is that!.

• Random has a precise meaning in statistics, but it is related to the how it is used in the example above.

  Sam means the chance that two students independently (without knowledge of each others actions) choose the same holiday location. The OMG means that they thinks this chance is small. Suppose the probability of a chance encounter is 0.1%. There are two possible explanations for this encounter:

  – It was by chance. Out of a 1000 students there is likely to be about 1 chance encounter.
– It was not by chance. That is Jon’s choice of holiday location was not independent of Sam’s choice. The smaller the this probability the more likely you are to reject the idea that the meeting was by chance.

• In other words, when people say how random is that, they really mean that the chance of this happening by mere coincidence is small.
Example III - Heights

• Suppose I want to know the average height of a student at A&M.

• Of course I will never know the heights of all A&M students. However I can draw a sample from this population, for example this class.

• Question How large a sample should I draw? Suppose I have a choice of drawing a sample of 5 people, 8 people or 40 people, which sample is more reliable. Why?

  – It is incorrect to say that samples of size 40 will be closer to the total population mean/average than a sample size of 8. A sample is random we have no idea whether it will be or not.
  – However, in general (on average) it will be. How do we precisely quantify this?
The difference between Population and Sample

Underlying these examples is that we will never know (observe) the population (population does not refer to just people, it can mean all the vehicles on a road, all the patients with a particular illness, all trees in the world etc). If the population is known, then the statistics becomes redundant. However it is very rare that we observe the entire population.

Definitions

- **Population** The entire collection of individuals of interest Eg. (a) all students at A&M (b) all oak trees in Texas this year.

- **Sample** A subset of the population that is measured.

- For example (a) a sample of A&M students, i.e. students in a class (b) A sample of all oak trees in Texas could be all oaks in a square kilometer of a forest.
Confident statements

• This brings us back to the start of the lecture.

• Since we do not observe the population and only a sample from it, Statistical analysis can not be used to ‘prove’ a result. However, if the sample has been collected in a way that is ‘representative’ of the population, then we can use statistics to quantify the amount of uncertainty in the sample. We can use this to make confident statements about the populations. We use the word confident because it means we cannot be sure. I.e. I am confident I will get an A in this class, is different to I am certain I will get an A.

• In the next example, we give an example of a statistical analysis that has recently been in the media.
Example IV: Looks can Kill

Most psychological analysis involve the collection and analysis of data.


The American justice system is built on the idea that it is blind to all but the objective facts, as exemplified by the great lengths we go to make sure the jurors enter the courts unbiased and protected from outside influence during their service. Of course this idea does not always match reality say psychological sciences John Paul Wilson and Nicholas Rule, co-authors in the study.
Example IV: Setting up the study

• The researchers compared deathrow and life sentencers in Florida. They chose Florida because it has many people on deathrow and it also keeps a database of photos of all convicts.

• The researchers obtained the photos of 371 convicts on deathrow convicted of first degree murder (226 white and 145 black). This is a sample. In this case the population is a little abstract. But roughly speaking, it is all deathrowers convicted of first degree murder (in Florida) - really it is more than this (essentially 1st degree murders all people ever, who if put on trial in Florida, would be given a death sentence).

• To make their comparison, they obtained the photos of 371 convicts convicted of first degree murder who were given a life sentence (not on death row), again to ensure that race was not an issue in this sample.
there were 226 white convicts and 145 black convicts. This is another sample.

- The photos of all prisoners were turned into black and white photos and placed on sheets of papers.

- 208 Adult Americans (who did not know that any of the men were convicts) were asked to rate the trust-worthiness of each convict using just their photo. The scale was from 1 to 8, 8 being very trust-worthy and 1 being not at all trust worthy. The researchers found that the average trust-worthy score given to convicts on deathrow was 2.76 compared with those given a life sentence which was 2.87.
Example IV: The Statistical analysis

• The difference is not huge, but it is statistically significant as the p-value is less than 1% (technical jargon). What this means, is that a difference of $2.87 - 2.76 = 0.11$ or larger happening between the two group by just random chance is less than 1%. In other words, suppose the photos of two groups of 371 life-sentencers were compared 100 times, on average less than one out of a 100 times would we see a difference of $2.87 - 2.76 = 0.11$ in their honesty scores.

• This suggests that untrustworthy looking people are more likely to face the death penalty than trustworthy people (of convicted).

• Of course, it could be that people who look more dishonest commit more terrible crimes than more honest looking people.
The researchers took this into account by making an independent study and compared the honesty ratings of death sentence convicts who were acquitted (usually on DNA evidence) with life sentencers. Again a statistically significant difference in honesty ratings was seen. Suggesting that looking dishonest does not mean you are more dishonest.