# 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.

- Useful free book if you are new to statistics (undergraduate level text): `https://www.openintro.org/stat/textbook.php?stat_book=os`

- Grades:
  - Every week or two I will set homework. This will be worth 10% of your final grade.
  - Midterm: 25% (30th October) in Blocker 150.
  - Assignment: 25% (set in the middle of November)
  - Final Exam: 40% (4th December 10:20-12:20) in Blocker 150.

# 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.

- **Always use common sense**

  – Example Researchers are looking for evidence that a stent (a medical device) reduces the risk of strokes in at risk patients. They conduct a study. 451 at risk patients volunteer for the study. They are *randomly* assigned to the treatment (given stents) group and control group (not given stents, but otherwise treated in 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.

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

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

Comparing the numbers, we see there is nothing in the data to suggest that a stent reduces the risk of a stroke. A correctly done statistical analysis will back this conclusion. Beware of interpreting statistical output incorrectly.

- 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 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 most of the 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/`. Full instructions are in the handout.

- I will sometimes use Statcrunch to illustrate certain ideas (such as distributions).

# What is statistics?

**The main idea**:

- Statistics is used in a multitude of disciplines. We use it to make medical diagnostics, on recommendation websites (Netflicks, Facebook, Amazon)....the list goes on.
- The applications are diverse, but the premise is the same.
- In statistics, we want to understand something about a population of individuals based on a sample from that population.
- Formally Our aim is make intelligent guesses about an unobserved population based on a substantially smaller sample (subset of the population) that is observed.

Is this possible, yes it is. But the important point is that we make intelligent guesses not definite statements about the unobserved population.

# OMG how random is that?

- We come across the word random all the time. You will often hear the word random in every day conversation.
  - **Example** Sam, went on holiday. Later Sam sees their best friend Jon at the same place. Sam thinks OMG, I saw Jon 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.
- Random here refers to the chance that two people independently (without knowledge of each others actions) choose the same holiday destination at the same time.
- The OMG means that they think this chance is small. The calculation of this chance can be tricky.

- Suppose the probability of a chance encounter is 0.1%. Using this chance we can come up with two possible explanations for the encounter.
  - It was by chance. Out of a 1000 holidaying 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 OMG random is that, what they really mean that the chance of this being a coincidence is small.

- Statistics is the interpretation of these chances. We give more precise examples below.

# Probability and its link to statistics

- We mentioned the word probability above.

  A probability is the "chance" of observing a particular outcome and is calculated under various assumptions.

- Suppose you toss a coin. There are two possible outcomes, heads or tails.

- What is the chance of getting a head? Usually we say 50% (or 0.5).

- But this is only true if the coin is *fair*. So the chance is 50% is only true under the assumption the coin is fair.

- The above is a probabilistic statement (it has nothing to do with statistics).

- A statistician may ask "Is this coin fair?". This we can check through experiments.

- We can toss the coin many times and record the outcome of each toss. This is an experiment, where we collect data.

- Suppose we toss the coin 100 hundred times. 55 are heads and 45 are tails. This is what we call a *sample*. Based on this sample do we think the coin is fair[1]?

- By simply tossing a coin repeatedly can we ever be absolutely sure that the coin is fair?

[1] To assess if the coin is fair we need to calculate the probability of obtaining 55 heads out of 100 under the assumption the coin is fair

# The difference between Population and Sample

Usually in statistics there is a population that cannot be observed and a sample (subsect) from the population, which can be observed.

Definitions

- Population The entire collection of individuals of interest Eg. (a) all students at A&M (b) the outcomes of coin toss tossed for ever (this is quite tentative).

- Sample A subset of the population that is measured.

- For example (a) a sample of A&M students. (b) The outcomes of 100 consecutive coin tosses.

# Returning to confident statements

In statistics we make confident statements about the population based on the sample.

- 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.

- By quantifying the uncertainty, we can make confident statements (intelligent guesses) about the populations. We use the word confident because it means we cannot be sure. Ie. I am confident I will get an A in this class, is different to I am certain I will get an A.

- If I toss a coin 1000 times, and 501 times out of 1000 I get a head. I can be confident that there is nothing in the data to say the coin is not fair. Or I can be confident that the coin is at least close to fair.

- In the next slide we briefly discuss bad data collection methods, where we cannot make an confident statements about the population.

# Example 1: Biased reporting

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 $0.78\%$.

- This is a relatively small probability, but it is not so small for it to be considered improbable (so unlikely the octopus was psychic).

  This type of argument will form the basis of hypothesis tests.

- But before we interpret the probability, it is important to understand the data is collected.

  This is an example of biased reporting.

  During the world cup there were probably thousands of animals trying to predicting the outcome of football matches. 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. Only the animals that get the correct prediction are reported (we not hear about the others). This is why it is biased.

# Example 2: 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 (collected from a series of the experiments), the students were asked if the data suggested that cell phone use increased reaction times in drived. The statistical analysis showed that the data did suggest that reaction time did 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".

- 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).
  - Further, it not clear whether the student made the comment, because they were a great driver. This is another example of biased reporting.

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

- We can only draw meaningful statistical conclusions from samples that are random sampled.

# Interesting examples

- We conclude this class by discussing two different studies from two different disciplines where both good data collection and sound statistical analysis is important.

# Example 1 - Sparkling water and weight

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

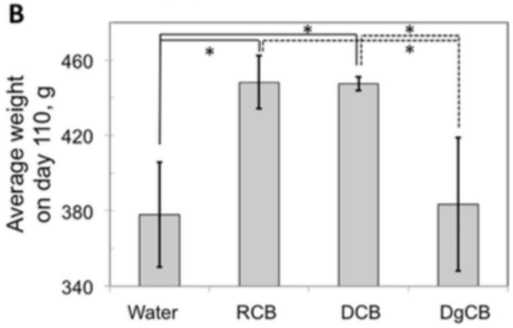
- It is widely believed that the sugar in the drink plays an important role. Recently, the role that artificial sweeteners in the soda have on obesity has been investigated.

- 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/CO2beverage.pdf`.

# Collecting the data

- 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 on the prescribed group.
- The rats had unlimited access to rodent food (it was not restricted).
- This is known as an experimental design.

- 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



- 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.

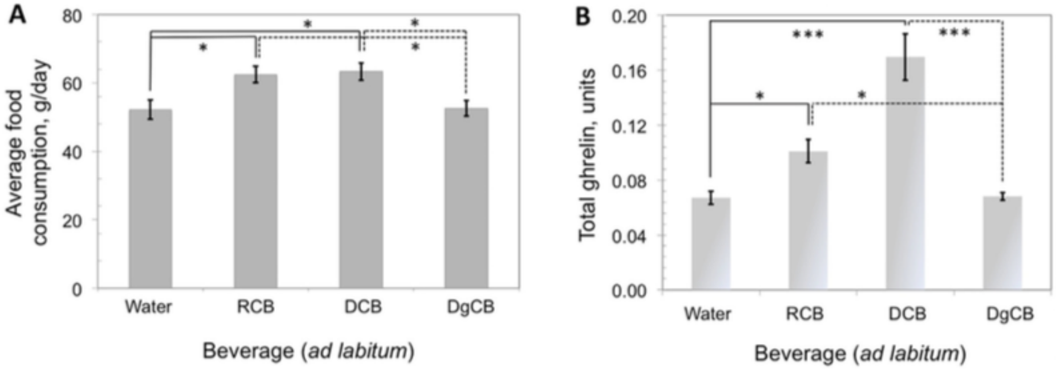
- The larger the difference the less believable it is due to chance.

- 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.



# Example 2: Looks and justice

Most psychological analysis involve the collection and analysis of data.

See `http://www.psychologicalscience.org/index.php/news/releases/the-trustworthiness-of-an-inmates-face-may-seal-his-fate.html` for interesting studies.

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.

# Collecting data

- 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).

- 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 in each sample.

- The photos of all prisoners were turned into black and white images.

- 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**.

# Summarising 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.