MA-250 Probability and Statistics

Nazar Khan

PUCIT

Introduction

Administrative

- Course webpage
  - http://faculty.pucit.edu.pk/nazarkhan/teaching/Spring2018/MA250/MA250.html
- Course material also available on \\printsrv
- Office Hours: MW 1:30-2:30 pm
- Lots of quizzes
- No scaling
- No such thing as a stupid question.

What is Statistics and Probability

- **Statistics** is the ‘art’ of
  - understanding the “real” world as it is and not “how we think” it is,
  - intelligently summarizing large amounts of data,
  - making numerical guesses for puzzling questions.
- **Probability** is the ‘tool’
  - to work with statistics,
  - to make conclusions/predictions from statistics,
  - to assign numeric value to uncertainty.

Statistics

- Even a lifeless calculator can give you statistics by plugging numbers into formulae.
- But the true meaning of those statistics requires careful thinking.
- One aim of this course is to make you think like a statistician, not like a calculator!

Probability

- One of the more important branches of mathematics.
- Can be a bit unintuitive.
- Has its own terminology.
- Every probability problem requires **thinking**.
  - Fortunately, there are some tricks.
- One aim of this course is to make you develop thinking skills that help solve probability problems!

Applications of Probability and Statistics

- Computer Networks
- Machine Learning
- Computer Vision, Image Processing, Graphics
- Algorithms
- Data Mining

Applications of Probability and Statistics

- Politics
- Economics
- Social Sciences
- Medicine
- Physics
- Everything involves probability and statistics!

Applications of Probability and Statistics

- Every two days we create as much data as we did from the beginning of mankind till 2003.
- The **only** way to deal with such large amounts of data is to summarize it.
- Statistics is the method of summarization.

The Scientific Method

1. Define the question
2. Background research, observation
   - Have others tried to answer this earlier?
3. Formulate a hypothesis
   - If we do X, then Y will happen.
4. Design and run an experiment
5. Analyze the results
6. Communicate the results

- Experimental measurements are noisy (randomness).
- Statistics is critical in steps 4 and 5!

Design is more important than the experiment itself

DESIGN OF EXPERIMENTS

Is Polio vaccine effective?

- Some one makes a vaccine for Polio.
- You need to find if it is effective or not.
- How will you go about finding an answer to this question?
  - Follow the scientific method

Is Polio vaccine effective?

- Follow the scientific method
  1. Define the question
  2. Background research, observation
     - Have others tried to answer this earlier?
  3. Formulate a hypothesis
     - If we do X, then Y will happen.
  4. Design and run an experiment
     - Do X
  5. Analyze the results
     - Did Y happen?
     - So what do we conclude?
  6. Communicate the results

Is Polio vaccine effective?

- Which **hypothesis** is better?
  1. Children that get vaccinated will have lesser polio cases
  2. Children that get vaccinated will have lesser polio cases **compared to** children that don’t get vaccinated.
- What will each hypothesis prove?

Is Polio vaccine effective?

- **Compare**
  - those that use the vaccine – **treatment group**
  - those that don’t – **control group**
- If the treatment group has lesser percentage of polio, then the treatment is effective
- Otherwise the treatment is useless.

	Treatment	Control	Conclusion
Outcome	Less Polio	More Polio	Effective

How to lie with statistics

- What if the treatment and control groups are different?
  - Treatment group is from people that are more immune to Polio – poor children
  - Control group is from people that are less immune to Polio – rich children
  - Statistics will falsely show that the vaccine is effective.

	Treatment	Control	Conclusion
Outcome	Less Polio	More Polio	Effective

How not to lie with statistics

- The treatment and control groups **must be similar**.
- How to ensure that?
  - Pick randomly
- Statistical studies can mix up hidden factors.
  - Polio is a disease of hygiene. This factor must be accounted for in the treatment and control groups.

Questions that statistics can answer

- Is Homeopathy effective?
  - Bad study: ask those that use Homeopathy.
  - Good study: perform a **controlled** experiment with **randomly selected** treatment and control groups.

Statistics is also an ‘art’

- Blind statistics: Plugging numbers into formulae without thinking.
- Proper statistics: Controlling for the hidden factors.
- Blind application of statistics can be disastrous.
  - Terming a medical treatment effective when, in fact, it is harmful.
  - Terming a medical treatment ineffective when, in fact, it is effective.

Confounding Factor

- The treatment and control groups should differ from each other **only in terms of the treatment**.
- If they differ with respect to some other factor, then this is a **confounding factor**.
  - Are the results due to treatment or due to the confounding factor?

The Randomized Controlled Experiment

- **Random** selection of treatment and control groups
- Elimination of **confounding factors**
  - **Placebo effect** – some people are cured by the idea of treatment. So give both groups the impression that they are being treated.
  - **Double blinding** – neither the patient nor the doctors know which group the patient is in. So the doctors don’t give different treatment.

The Observational Study

- Is smoking harmful?
  - Can we do a randomized controlled experiment to answer this?
- We can **observe** smokers and non-smokers over time.

Is Smoking Harmful?

- If smokers are less healthy compared to non-smokers, then “yes, smoking is harmful”.
- **Wrong!!**
- Association is not causation!
- What about confounding factors?
  - For example, a gene that causes lung cancer and also causes people to smoke.
- Careful studies have concluded that there are no confounding factors. **Smoking really is harmful.**

Association is not Causation

- Many people die in hospitals.
- Do hospitals cause death?
- Hospitals and deaths are **associated** with each other.
- But hospitals don’t **cause** deaths (in general).
- Blind statistics will tell you that hospitals lead to deaths.

Randomized Controlled Experiment vs. Observational Study

- Observational studies prove association but not causation.
- Confounding factors can be at work.
- Randomized controlled experiments try to minimize the effects of confounding factors.
- So, wherever possible, a randomized controlled experiment should be performed to understand the real world.

Next Lecture

- Descriptive statistics
  - Histograms
  - Mean, Standard Deviation
  - The Normal Curve
- Read Ch. 3-5.