Inferential Statistics with Python - Rounak Banik

About the speaker

I am a final year undergraduate at IIT Roorkee. Although currently pursuing Electronics and Communication Engineering, my professional interests lie in Web Development and Data Science. I have previously interned as a Software Engineer at Parceed, a New York based startup and Springboard, a Data Science EdTech startup based in San Francisco and Bangalore. I also worked as a Backend Development Instructor with Acadview, teaching Python and Django to around 35 college students from Delhi and Dehradun. I am currently working directly under the Director of IIT Roorkee and Dr. Durga Toshniwal for my B.Tech Project on Fake News and Review Detection. I am also a student of Springboard's Data Science Career Track, being mentored by Baran Toppare, former Lead Data Scientist at Getir.

Abstract

Link to talk material: https://github.com/rounakbanik/inferential_stats_pycon

Inferential Statistics is the art of making conclusions and predicting outcomes from data. It is an incredibly important component of exploratory data analysis and A/B testing.

In this talk, I will be giving a brief overview of the major theories underlying inferential statistics, its many tools and techniques and its implementation using Python. Through the course of the talk, I will also be walking the audience through six real world datasets and giving them a taste of how to proceed with gaining insights from your data through hypothesis testing and data visualisation.

My talk has the following contents:
1. What is Statistics? The Difference between Descriptive and Inferential Statistics
2. A brief primer on Descriptive Statistics: Central Tendencies, Binomial and Normal Distributions, Z-Scores.
3. The importance of Sampling. Various kinds of sampling bias. Quality and quantity of sampled data.
4. Estimation of a population proportion and mean. Sampling error, confidence intervals. Central Limit Theorem.
5. Basics of Hypothesis Testing
6. One Sample and Two Sample Significance Tests, Chi Square Significance Test
7. Correlation, Scatter plots and Linear Regression
8. Which Statistical Test to use on what kind of data
9. Statistical and Practical Significance of test results

To demonstrate the above concepts, we will be implementing the methods in Python and working through synthetic data as well as real world datasets.
1. Exploring Literacy Rates in Punjab and Delhi: From data retrieved from Kaggle, we will try to determine if there is a significant difference in literacy rates of Punjab and Delhi.
2. NBA Player Heights: From a sample of NBA Players, we will try to find out if the mean height is actually 6'7" as reported by most publications.
3. Suicide Rates in India: From the suicide statistics between 2001-11, we will try to determine if men are as likely as women to commit suicide.
4. Do Men and Women prefer certain countries to book Airbnbs in: We will use Airbnb's data to deduce if there is a relationship between sex and country preference for booking Airbnbs.

5. Olympian Weights: We will try to estimate the average weight of Olympians given a small sample.
6. Credit Card Fraud: We will try and estimate the fraction of fraudulent transactions given a small subset of the data.