Introduction to Data Science
SKKU University, Summer 2015

Prof. Jevin D. West
University of Washington
Lecture 6 – July 6, 2015

Statistics Survey

1. What is the difference between mean, median and mode?
2. Define variance. How is this related to the standard deviation?
3. What is a random variable?
4. What is a probability distribution? Draw a normal distribution. Label the following: mean, median, mode, standard deviation, interquartile range.
5. What is a t-test? When would you use it and why? What are the assumptions of this statistical test?
6. What is a chi-square test? When would use it and why? What are the assumptions of this statistical test?
7. What is a null hypothesis?
8. What is regression? When should you use it?

Agenda

- 9:30 – 9:45 Statistics Survey
- 9:45 – 10:00 Week 2
- 10:00 – 10:30 Probability Distributions
- 10:30 – 10:40 Break
- 10:40 – 11:45 Probability Exercises
- 11:45 – 12:00 Questions

Week 2

- Tuesday, July 7: **Quiz #2**
- **DUE** Tuesday, July 7: Question Report
- **DUE** Thursday, July 9: Assignment #2
- **DUE** Friday, July 10: Preliminary Analysis

Quiz #2

- Tuesday, July 7: **Quiz #2** is based the lecture from last Friday and the following readings:
  - Thomas Davenport (2006). "Competing on Analytics", Harvard Business Review, Jan. 2006, Vol. 84 Isue 1, pp. 99-107
  - The Fourth Paradigm, *Jim Gray on eScience: A Transformed Scientific* Method, pgs xvii – xxxi
  - Provost & Fawcett (March 2013): "Data Science and its relationship to big data and data-driven decision making", Harvard Business Review

Group Project, Question

- **DUE** Tuesday, July 7: Group Project Question Identified including a 1-page description that answers the following:
  - Why are you investigating this question?
  - How are you going to try and answer your question?
  - What are the limitations of your question?
  - Who else has answered this question? How will you build upon other work done with this question?
  - Provide references
- Submit to your team repository on github

Assignment #2

- **DUE** Thursday, July 9: Assignment #2
  - This will be done mostly in class on Wednesday, July 8. The assignment will include answering questions about some example data sets. This is meant to help you prepare for your preliminary analysis.
- Submit to your team repository on github

Group Project, Preliminary Analysis

- **DUE** Friday, July 10: Preliminary Analysis and summary statistics (several tables and graphs). Conduct a first-pass descriptive analysis of your dataset. Produce tables and graphs that show exactly what data you have, and that contain summary statistics about the data. Questions to answer for each data source include:
  - How many unique observations to you have?
  - What information/features/characteristics do you have for each observation?
  - What are the min/max/mean/median/sd values for each of these features? What is the distribution of the core features (show a histogram)?
  - Are there obvious trends in the data (over time, across subgroups, etc.), and are the differences statistically significant?
  - What are the other salient aspects of the data (e.g. geospatial factors, text content, etc.)
  - Provide a bullet-list of the next 5-10 tasks you will perform in analyzing your dataset.
- Submit to your team repository on github

Random Variables, Probability Distributions and the Central Limit Theorem

Dice Simulation

- In R, roll two dice and sum their scores
- Plot your results for 10 experiments, 100 experiments and 10000 experiments
- What have you learned?

Height Percentile

- Gather height data for your group
- Plot your data
- What is the mean and standard deviation?
- What is *your* height percentile?