Scalable Data Analysis (CIS 602-02)

Machine Learning

Dr. David Koop
Descriptive Statistics

- Mean, median, mode, range
- Standard deviation (variance) measures how far-flung data is (difference from mean)

\[ s_x = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}}. \]

- Interquartile Range: Inner 25% to 75% of values (by positions)
Probability Distributions

- Distribution of a random variable is the possible outcomes and their probabilities
- Types: Discrete, continuous, and mixed

![Diagram of probability distribution]

Event \( x_1 < X < x_2 \)

\[ P(x_1 < X < x_2) \]

[J. Isotalo., Basics of Statistics]
Frequentists and Bayesians

DID THE SUN JUST EXPLODE?
(ITS NIGHT, SO WE'RE NOT SURE)

THIS NEUTRINO DETECTOR MEASURES WHETHER THE SUN HAS GONE NOVA.

THEN, IT ROLLS TWO DICE. IF THEY BOTH COME UP SIX, IT TELLS TO US.
OTHERWISE, IT TELLS THE TRUTH.

LET'S TRY.
DETECTOR! HAS THE SUN GONE NOVA?

ROLL

YES.

FREQUENTIST STATISTICIAN:
THE PROBABILITY OF THIS RESULT HAPPENING BY CHANCE IS 1/36 = 0.027.
SINCE P<0.05, I CONCLUDE THAT THE SUN HAS EXPLODED.

BAYESIAN STATISTICIAN:
BET YOU $50 IT HAasn't.

[https://xkcd.com/1132/]
Bayesian Methods

• Inferential statistics
• Difference between Frequentist and Bayesian perspectives: both are useful
• Law of large numbers
• Prior and posterior
• Distributions
• Statistical models
Assignment 2

• Bikes and Weather
• Bike sharing data for Washington DC
• Weather data for Washington DC
• Hypothesis might be that people are more likely to use the program when the weather is "nicer"
• Integrate data, visualize it, do regression
Statistics for Hackers

J. VanderPlas
Statistics

• p-value: Test for a statistical hypothesis
  - Need a significance level (often 0.05)
  - Tells when the null hypothesis is true or not

• t-test: test whether the mean matches a specific value (or another mean)
  - Assume normal distribution
  - Paired versus unpaired

• Bootstrapping: resample the sample data (assume the sample is the population)
  - Use for small samples and when the distribution is unknown

• Overfitting: simpler can be better in the long run
Statistical Modeling: The Two Cultures

L. Breiman
Statistical Modeling and Machine Learning

- Differences
- Multiplicity of Good Models
- Simplicity vs. Accuracy
- Dimensionality
- "Data looking for a question" - Cox
- "Algorithms often appear in the form of black boxes with enormous numbers of adjustable parameters, …sometimes more knobs than data points" - Efron
- Understanding vs. Prediction
Next Class

• Unsupervised Learning: Clustering
• Reading Response
• Assignment 2 Details