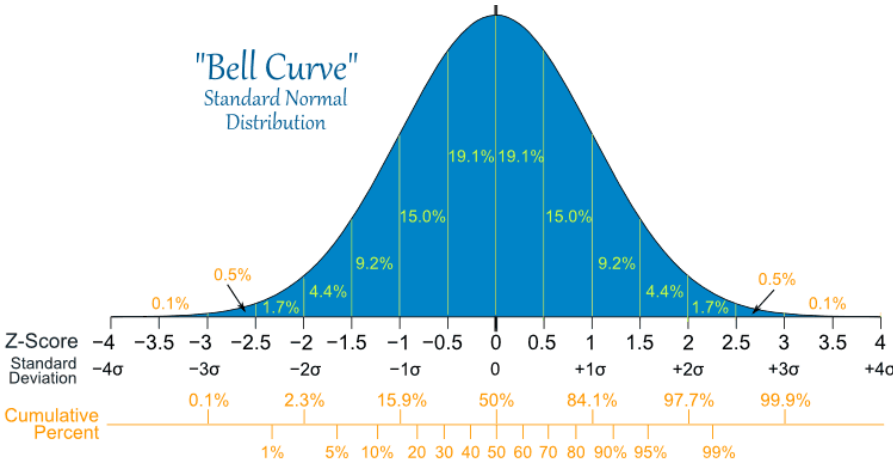
UNIVERSITY of WASHINGTON

Data Science UW Methods for Data Analysis

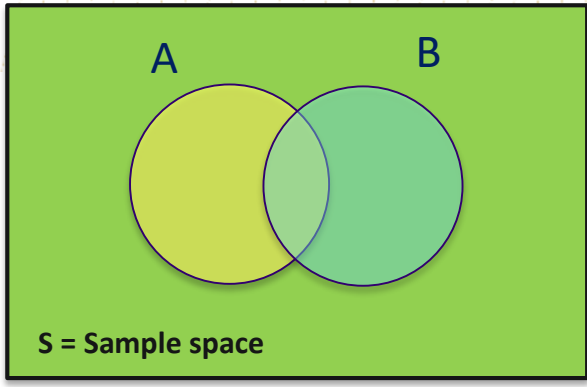
Class Review
Lecture 10
Nick McClure

Lecture 1

- R review
- Discrete Distributions
  - Bernoulli, Binomial, Poisson
- Continuous Distributions
  - Uniform, Normal, Students-T, Beta
- Covariance: Expected value of the differences between x,y and their corresponding means.
- Correlation: Normalized Covariance.
- Variable Transformations (must be monotonic)



Lecture 2

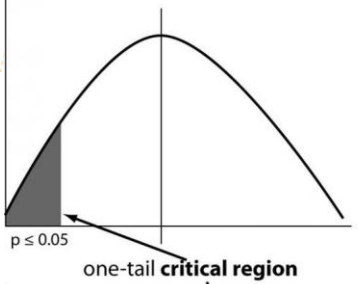


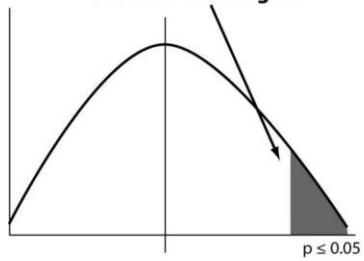
- Counting
  - Multiplication Principle, Factorial, Combinations, Permutations, expand.grid()
- Probability
  - 3 axioms: $0 \leq P(A) \leq 1$

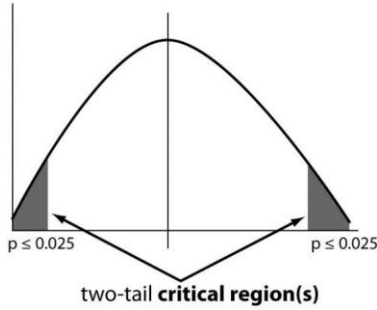
    $P(S) = 1$

    $P(A \cup B) = P(A) + P(B)$ If A and B are M.E.
  - Venn Diagrams
- Conditional Probability
- Mutually Exclusive $P(A \cap B) = 0$
- Independence $P(A|B) = P(A)$
- Simulations in R
- Imputation
  - Multiple Imputation: Amelia package

Lecture 3

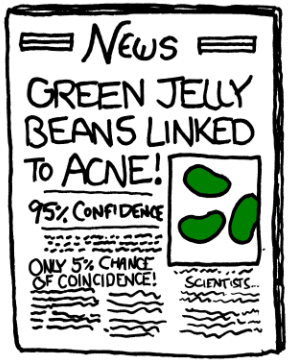






- Conditional Probability Trees
  - Rare disease testing
- Sampling Data
- Law of Large Numbers
- Standard Deviation: Measure of variability in a sample or population.
- Standard Error: Measure of variability in the statistics of the sample.
- Hypothesis Testing
  - Normal curve, one tailed vs two tailed, interpreting the p-value
- Student's T-test: Test differences of means of two populations with known variance.
- Welch's T-test: Test differences of means of two populations with *unknown* variance.
- Chi-Squared Test: Test difference in Counts, needs larger sample.
- Fisher's Exact Test: Same as above, but exact (Stricter).
- Testing for outliers

Lecture 4



- K-S Statistic
- Shapiro-Wilk test for normality
- ANOVA: analysis of variants, i.e., is at least one mean of the groups different?
- Bonferroni correction: If you test n hypotheses, significance level should be alpha/n.
- Central Limit Theorem: The distribution of summary statistics is normally distributed:

$$\bar{X} \sim N(mean, \frac{variance}{\sqrt{n}})$$

- Confidence Intervals
- Introduction to Regression:

$$y_i = mx_i + b + \varepsilon_i$$

$$\varepsilon_i \sim N(0, \sigma)$$

Lecture 5

DayOfWeek
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
…

→

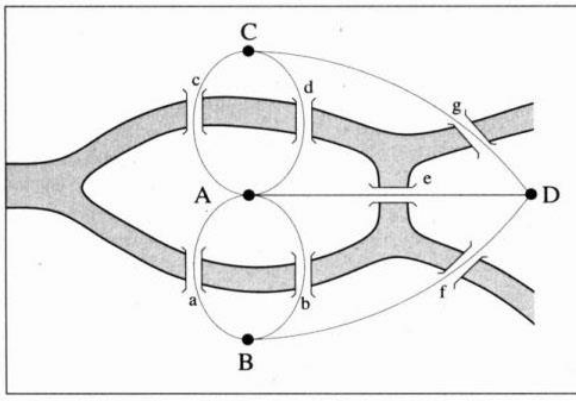
DayOfWeek
1
2
3
4
5
6
7
…

Eye Color
Brown
Brown
Blue
Green
Green
Blue
Brown
…

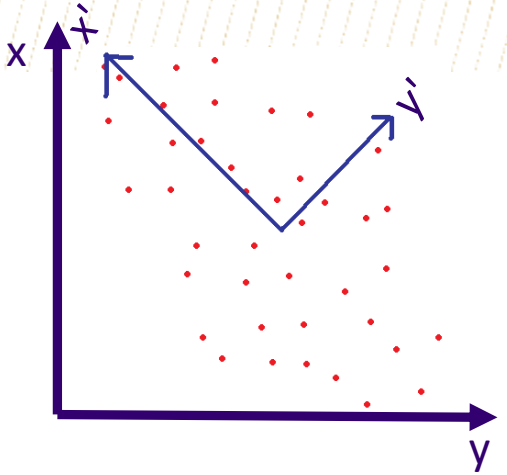
→

Brown	Blue
1	0
1	0
0	1
0	0
0	0
0	1
1	0
…	…

- More on Regression:
  - MSE, R^2, Least Squares Fitting.
- Homoscedasticity
  - Errors are random, heteroscedastic otherwise.
- Leverage and Cook's Distance
- Prediction and Confidence bands
- Encoding categorical variables
- Multiple linear regression
- Introduction to graph theory with python:
  - Triangle completions
  - Centrality
  - Graph labeling
  - Clustering
- Gephi
- Testing for Degree Distributions



Lecture 6



- Matrix operations/Linear algebra
- Singular Value Decomposition (SVD)
- SVD as regression (Deming regression or Total least squares)
- Using SVD to compress information.
- SVD as a way of clustering data.
- Ridge Regression
  - Regularize partial slopes with a squared term in the loss function:

$$\min \sum (y - y_i)^2 + \alpha \sum \beta^2$$

- Lasso Regression
  - Regularize partial slopes to have total sum less than a value:

$$\min \sum (y - y_j)^2 \quad \text{Such that} \quad \sum |\beta_i| < \lambda$$

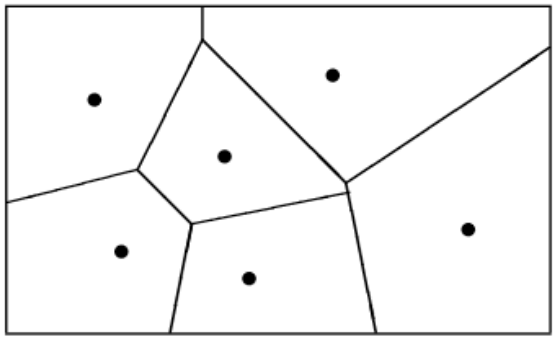
- Logistic Regression

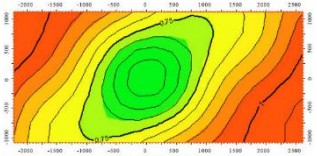
Lecture 7

$$X_k = \frac{1}{N}\sum_{n=0}^{N-1} x_n e^{i2\pi k \frac{n}{N}}$$

To find the energy at a particular frequency, spin your signal around a circle at that frequency, and average a bunch of points along that path.

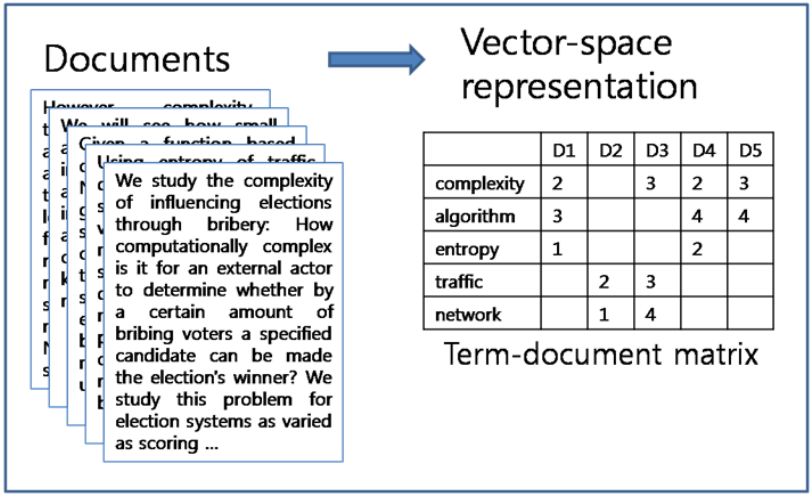
- Time Series
- Moving Averages
- Seasonality
  - Fourier Transform
- ARIMA models
  - Auto-regressive Integrated Moving Average
- Spatial Statistics
- Moving Windows
- Median Polish
  - Removes spatial trends
- Point estimate
  - Weighted Averages: weighted by voronoi polygon area
- Global estimation
  - Kriging: weight prediction at any spot by spatial dependence or variance.
- Clustering
  - Ripley's K



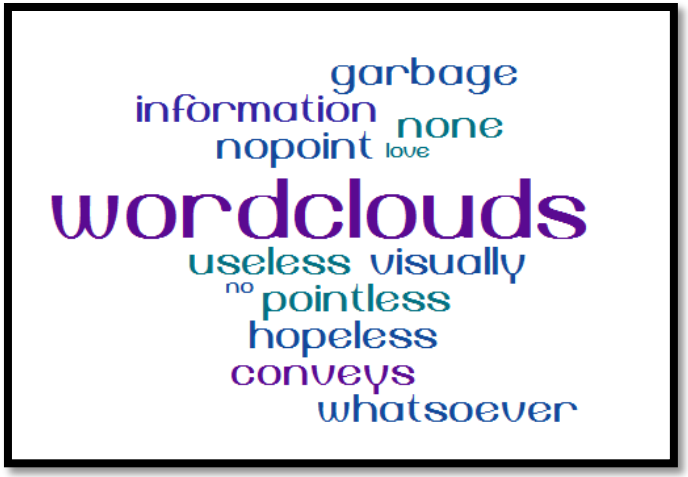


Lecture 8

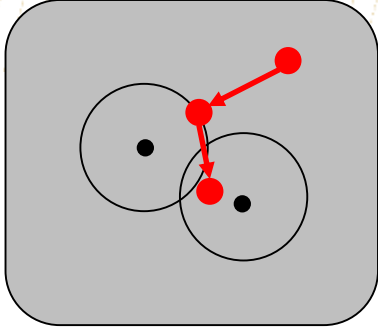
- Guest lecture on Bandit Algorithms
  - Used primarily as a substitute for A/B testing.
- Natural Language Processing
- Text Normalization
- Word Clouds ☹
- Text Distances
- Corpus/Dictionaries
- Naïve Bayes
- Word Frequencies (TF-IDF)



	D1	D2	D3	D4	D5
complexity	2		3	2	3
algorithm	3			4	4
entropy	1			2	
traffic		2	3		
network		1	4		



Lecture 9



- Bayesian Statistics
  - Prior, Likelihood, and Posterior
- Bayesian Inference
  - Estimating p(heads)
  - Estimating linear regression parameters
- Monty Carlo Markov Chain Estimation
  - Accepting/Rejecting points to estimate a distribution.
- Computational Statistics
  - Simulate the Null Hypothesis, and find p-value.
- Bootstrapping
  - Bootstrapping for small samples and getting errors on linear regression.

$$Posterior = Likelihood\frac{Prior}{P(data)}$$

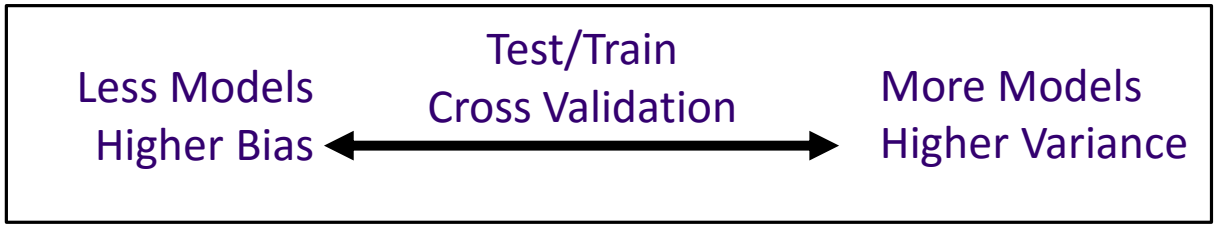
Class Overview: Important Themes

> Hypothesis Testing

> Linear Regression

– Ordinary Linear Regression, Multiple Linear Regression, Logistic Regression, Ridge Regression, Lasso Regression, SVD

> Bayesian Statistics and Computational Statistics



Class Overview

- Remember this class is an overview of many methods.
- Hopefully you will know what and where to lookup subjects that you may need for work, projects, dinner party jokes, etc…
- This certification class is a great step in the right direction.
  - It shows employers and colleagues that you are serious about the analytical field and have had formal training.
- You are now (and have been) a resource for others.
- Last piece of advice:
  - Don’t ever stop learning. The day we stop learning for/at our jobs is the day we should be looking for a new job.