UW PCE-Data Science 350
"Methods for Data Analysis"
Instructor: Stephen Elston

Overall Class Objectives:

- Learn methods to explore and understand data.
- Understand the core concepts of statistics.
- Understand and implement various statistical procedures in R.
- Describe and interpret analytical results from common statistical methods.
- Expand R programming skills to be able to write/test/log code from scratch.
- Work with structured and unstructured data.

Core Topics

Accessing and Storing Data Topics:

- Introduction to SQL, SQLite in R.

Data Visualization and Exploration Topics:

- Explore data with multiple views, using the techniques of histograms, distributions, box/chart plots, chart aesthetics, and conditioned plots.

Statistical Topics:

- Probability, distributions, hypothesis testing, bias vs variance, confidence intervals, law of large numbers, central limit theorem, linear regression, spatial statistics, time series, Bayesian statistics, Bayesian inference, and computational statistics.

Programming Topics:

- Functional programming for analytics, vectorizing code, reading and writing data, and data munging techniques, testing statistical code.

Code, data and slides for this course can be found at: https://github.com/StephenElston/DataScience350

	Topic	Subjects & Concepts	Learning Objectives
Week 1	Introduction, Data Exploration, R Overview	Logging, data loading, using functions, apply operators and functional programming, base statistics functions, graphing in R, summarizing data frames	• Explore and develop understanding of data sets • Understand techniques to visualize data • Understand summary statistics • Use functions in R • Understand the concepts of functional programming • Testing statistical code
Week 2	Probability,	Probability distributions, 3	• Recognize differences between

	Introduction to Conditional Probability, Missing Data, Getting/Storing Data	axioms of probability, counting, permutations, combinations, factorials, mutually exclusive, conditional, independent events, introduction to simulation, using SQL in R, imputation of missing values in R	combinations and permutations • Be able to setup equations to count outcomes • Distinguish conditional, mutually exclusive, and independent events • Understand the 3 axioms of Probability • Understand the principles of simulation • Understand how to access and store data with R • Use different imputation methods for missing data
Week 3	Applying conditional probability, detecting, verifying and treating outliers, advanced techniques for missing values introduction to hypothesis testing	Conditional probability trees, multiple imputation, sample vs population, sampling procedures in R, Law of Large numbers and the Central Limit Theorem, standard deviation/standard error, z scores, students t-test, welch's t-test, Chi-squared test, Fisher's Exact test, testing for outliers	• Understand and apply multiple imputation methods • Recognize different sampling procedures and know the benefits and uses of each • Understand Law of Large Numbers and the Central Limit Theorem • Understand the differences between sample and population • Describe the difference between standard error and standard deviation • Understand the principles of hypothesis testing • Interpret p-values
Week 4	Hypothesis Testing, Bootstrap resampling, and simulation	Kolmogorov-Smirnov test, Shapiro-Wilk test, ANOVA, Bonferonni Correction, confidence intervals, introduction to resampling methods, permutation tests with resampling methods, hierarchical simulation	• Apply various hypothesis tests • Account for testing multiple hypotheses • Use Central Limit Theorem • Understand the concepts of bootstrap resampling • Know how to apply resampling methods • Know how to apply simulation
Week 5	Introduction to Linear Regression,	Regression, least squares, homoscedasticity, leverage and cook's distance, prediction vs confidence intervals, predictor/feature selection, variable	• Identify linear vs non linear regression. • Understand the method of least squares • Be able to identify outliers in regression • Transform independent

		transformations, introduction to multiple linear regression	variables • Be able to understand the extension to multiple regression
Week 6	More on Linear Regression	Matrices, basic linear algebra operations, SVD, SVD regression, clustering and storing data via SVD, Ridge regression, Lasso regression, Logistic regression	• Understand how to work with underdetermined problems and regularization terms • Implement various regression techniques in R • Interpret linear regression outcomes • Understanding SVD and how it is used in feature reduction • Understand loss functions
Week 7	Time Series and Spatial Statistics	Dependent data representations, moving averages, auto regressive models, seasonality, Fourier transform, ARIMA methods, spatial median polish, point estimation, global estimation, and variograms.	• Understand how a series can be dependent on prior values • Be able to explain Random noise vs. Random walks • Identify seasonality in time series • Be able to quantify the spatial dependence in a data set • Predict points and global means • Understand and apply numerical methods to measure clustering
Week 8	Bayesian Inference and Computational Statistics	Bayes rule, Bayesian inference, prior, likelihood, and posterior distributions, MCMC, computational p-values via simulation, cross validation	• Understand Bayesian Inference • Be able to apply Bayesian Inference iteratively for data observations • Use MCMC and bootstrapping to estimate distributions • Generate/Interpret p-values computationally. • Use k-fold cross validation and understand the trade-off for high/low k values.
Week 9	Unstructured Data Part 1: Introduction to text analytics and NLP	Text normalization, string distance, stop words, dictionaries and corpus, Naive Bayes, distance metrics, word frequencies, Latent Dirichlet Allocation,	• Understand the text normalization process • Perform text normalization in R • Be able to explain TF-IDF • Apply and interpret Naive Bayesian • Understand the current state of NLP and what it can be used for.
Week 10	Unstructured Data Part 2:	Nature of image data, loading image data,	• Understand the nature of image data

	Introduction to image processing and understanding	normalization of image data, common image operations, feature extraction from images, Analytics for images	• Know how to manipulate and normalize image data • Understand image feature extraction • Apply analytics to image data