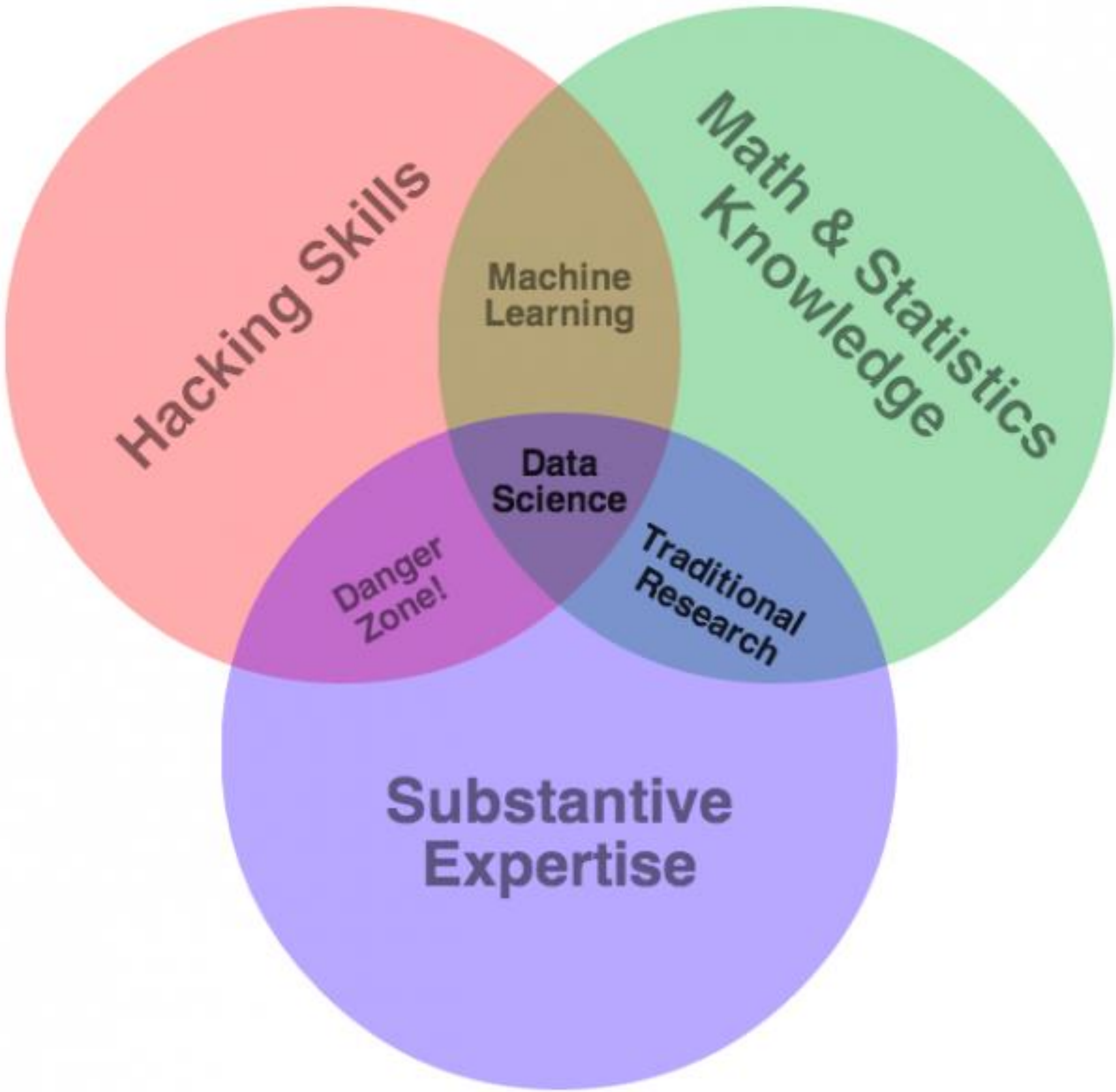
UNIVERSITY *of* WASHINGTON

Data Science UW Methods for Data Analysis

Introduction and Data Exploration
Lecture 1
Nick McClure


Hacking Skills
Math & Statistics Knowledge
Machine Learning
Data Science
Danger Zone!
Traditional Research
Substantive Expertise

Course Purpose

- This course isn't designed to make you an expert
- This course is designed to point you in the right direction
- Course Objectives:
  - Statistical tools for data exploration
  - The use of R to apply these tools to real data
  - Using inferential statistics to interrogate data
  - Testing and experimental design
  - Bayesian and classical statistics
- See syllabus for more information

Course Outline

	Date	Topic	Chapter	Assignment
Week 1	2015-06-22	Introduction; Overview of Modeling; Data Exploration; R overview	Intro DS: Ch.3,9 StatThink: Ch.2	Data Exploration; Start thinking about project; Vote on preferred extra topics.
Week 2	2015-06-29	Probability Distributions; Conditional Probability; Missing Data; Getting/Storing Data	Intro DS: Ch.7,10 StatThink: Ch.4	Conditional Probability and Outliers
Week 3	2015-07-06	Outliers and Missing Data; Introduction to Hypothesis Testing	Intro DS: Ch.6	Conditional Probability and Topic Chosen
Week 4	2015-07-13	Hypothesis Testing Continued; Modeling Exercise; The Central Limit Theorem	StatThink: Ch.6,7	Hypothesis Testing and Modeling Mini Project
Week 5	2015-07-20	Hypothesis Testing Continued; Confidence Intervals; Graph Algorithms	StatThink: Pg.93-97	Hypothesis Testing
Week 6	2015-07-27	Regression; Feature Selection	Intro DS: Ch.16	Regression
Week 7	2015-08-03	Feature Selection Continued; Simpson's Paradox; Intro To Bayes		Data Analysis and Feature Selection
Week 8	2015-08-10	Bayesian Statistics	StatThink Pg. 97-101	Bayesian Analysis
Week 9	2015-08-17	Bayesian Inference with R; Computational Statistics; Model Selection		Finish Project up this week!
Week 10	2015-08-24	Review and Possible Extra Topics (Graph Databases, Time Series, Spatial Statistics, NLP, Regex, Basket Analysis, ...)		

Course Requirements and Grading

This course will be graded by attendance, homework, and an individual project.

- Attendance: You MUST attend at least 8 out of 10 classes. This is non-negotiable, a UW requirement.
- Homework must be completed by the start of the next class. (Assigned weeks 1-8).
  - Returned as a 0,1, or 2.
    - 0 = Not done or a major part wrong/missing.
    - 1 = Completed, but missing or got wrong 1 or 2 parts.
    - 2 = Completed with at most minor issues. Demonstrates full understanding of subject.
- Individual Project: Due at the start of the last class.
  - Counts as 8 points.

Course Requirements and Grading

There is a total of 24 possible points. (16 pts for hmk + 8 project)

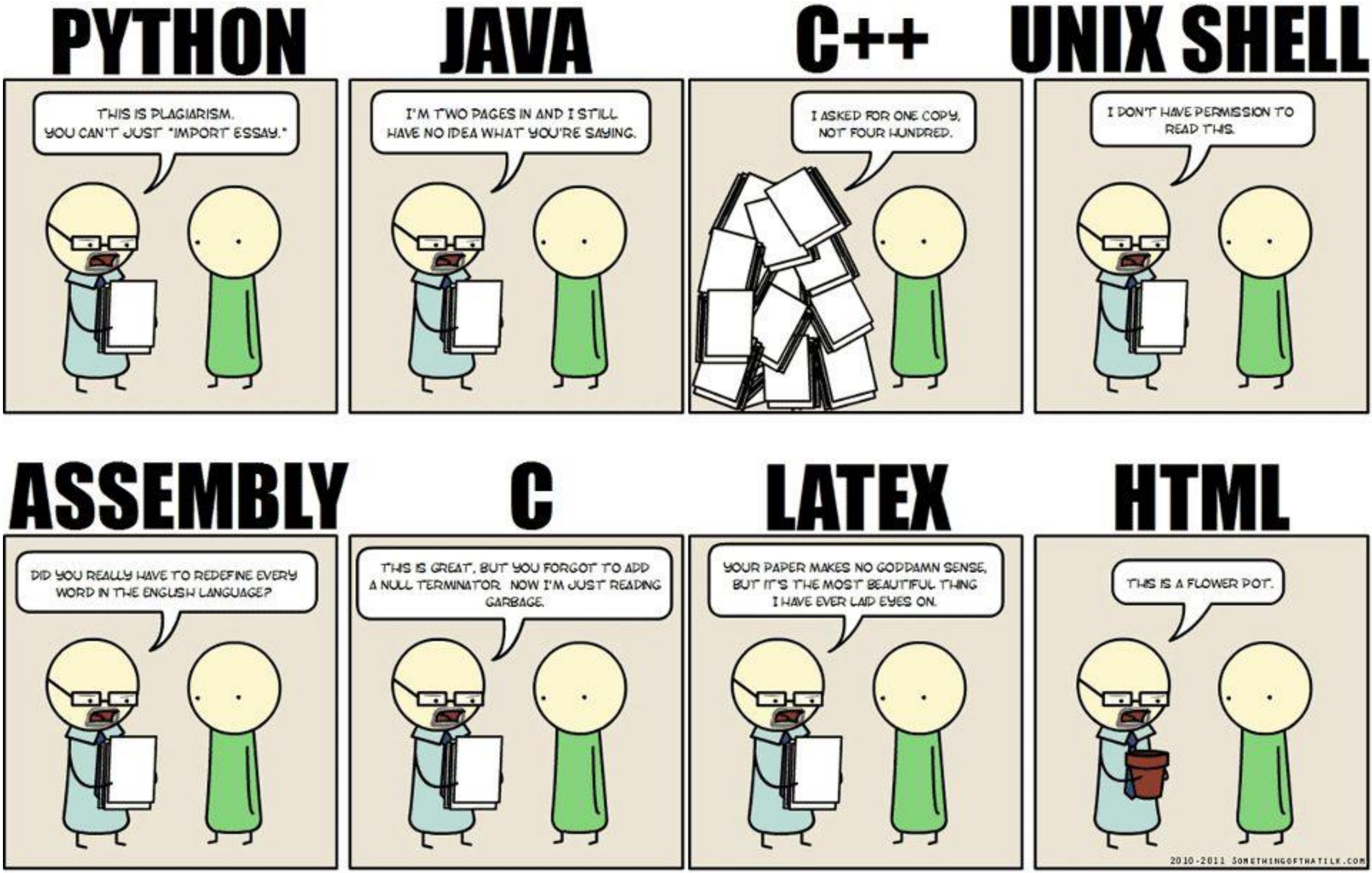
- Must get 18 total points to pass.
- 4 homework assignments must be made in a production level script (every other one = 1,3,5,7).
- 4 homework assignments are regular script writing (every other one = 2,4,6,8).
- The individual project must be production level code.

Office Hours and Contact Information

- List of ways to contact me:
  - nickmc@uw.edu
  - Linkedin group
- When I’m *usually* available:
  - Off/on for simple things during work. (M-F 8am-5pm PST)
  - Tuesday-Thursday 7pm-10pm.
  - Sunday various afternoon/evening times.

Emergency contact: 402-980-3192

Review


PYTHON
THIS IS PLAGIARISM. YOU CAN'T JUST "IMPORT ESSAY."
JAVA
I'M TWO PAGES IN AND I STILL HAVE NO IDEA WHAT YOU'RE SAYING.
C++
I ASKED FOR ONE COPY, NOT FOUR HUNDRED.
UNIX SHELL
I DON'T HAVE PERMISSION TO READ THIS.
ASSEMBLY
DID YOU REALLY HAVE TO REDEFINE EVERY WORD IN THE ENGLISH LANGUAGE?
C
THIS IS GREAT, BUT YOU FORGOT TO ADD A NULL TERMINATOR. NOW I'M JUST READING GARBAGE.
LATEX
YOUR PAPER MAKES NO GODDAMN SENSE, BUT IT'S THE MOST BEAUTIFUL THING I HAVE EVER LAID EYES ON.
HTML
THIS IS A FLOWER POT.
2010-2011 SOMETHINGOFTHATILK.COM

R Review

- R resources:
  - R page:
    - http://www.r-project.org/other-docs.html
  - Stackoverflow:
    - http://www.stackoverflow.com
  - ‘Little’ R intro:
    - http://cran.r-project.org/doc/contrib/Rossiter-RIntro-ITC.pdf
  - Quick R:
    - http://statmethods.net/
  - There are many tutorials available online, e.g.,
    - http://cyclismo.org/tutorial/R/
  - Notes from a two day course at UW:
    - http://faculty.washington.edu/tlumley/Rcourse/
  - Google’s Style Guide:
    - http://google-styleguide.googlecode.com/svn/trunk/google-r
style.html

Statistics Review

> Familiar Concepts:
  - Discrete vs. Continuous Distributions
  - Probability
  - y = mx + b vs $\bar{Y} = \mathbf{M} \cdot \bar{X} + \mathbf{B}$

> This area is the emphasis of the course.

SQL Review

> SQL (to know):
  - Create tables
  - Drop tables
  - Joins (Inner, outer, right, left)
  - Temp tables
  - Coalesce, Cast, Case

> (Accessing SQLite in R Demo)
> Access to SQLite Files:
  - Windows: http://www.yunqa.de/delphi/doku.php/products/sqlitespy/index
  - Mac OS X: https://www.sqlitepro.com/

Counting Review

- Factorials
  - Count # ways to order N things = N!
- Permutations
  - Count # of ways to **order** R things from N things = N!/(N-R)!
  - Ordering matters
  - P(N,R)
- Combinations
  - Count # of ways to **group** R things from N things = N!/(R!(N-R!))
  - Ordering doesn't matter
  - C(N,R) or $\binom{N}{R}$
- We will talk about this in depth next class.

Data Distributions (Discrete)

> Discrete Distribution Properties
  - Sum of all events must equal 1.
  - Probability of event equal to value of distribution at point.
  - No Negative values or values greater than 1.

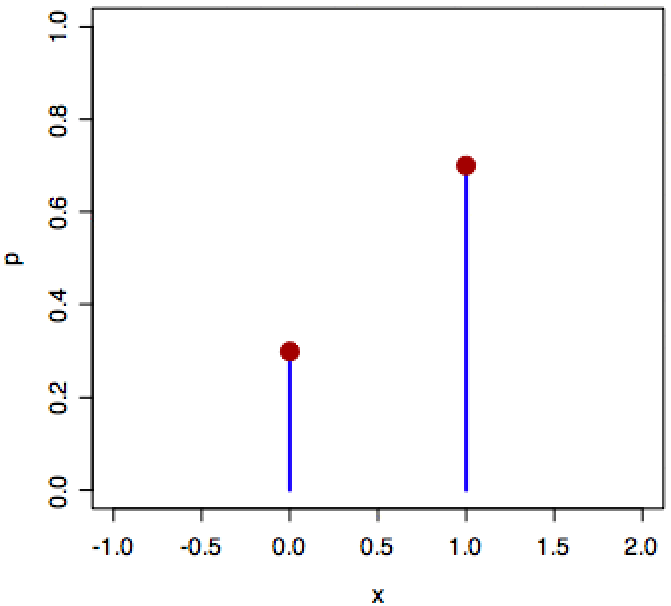
Data Distributions (Discrete)

> Bernoulli (1 event, e.g.: coin flip)

$$P(x) = \begin{cases} p \; if \; x = 1 \\ (1 - p) \; if \; x = 0 \end{cases}$$

$$P(x) = p^x (1 - p)^{(1-x)} \quad x \in \{0,1\}$$

- Mean = p
- Variance = p(1-p)



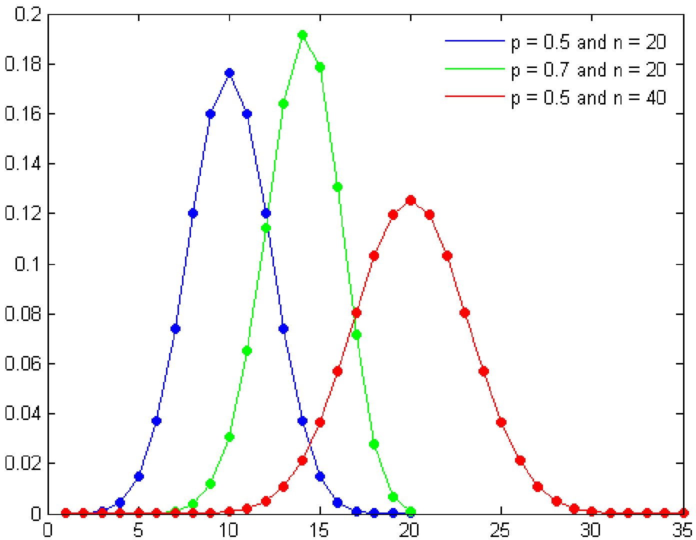
Data Distributions (Discrete)

> Binomial (Multiple Bernoulli's Events)

- Multiple Independent events = Product of Bernoulli Probabilities

$$P(x|N,p) = \binom{N}{x} p^x (1-p)^{(N-x)}$$

- Mean = np
- Variance = np(1-p)



Note: for larger n, we approximate this by a normal distribution.

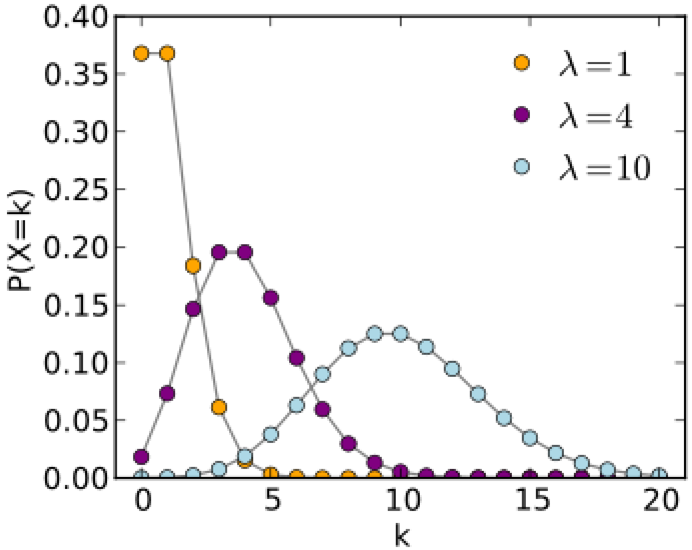
Data Distributions (Discrete)

> Poisson (Count of number of events in a time span)

$$P(x|\lambda) = \frac{\lambda^x}{x!} e^{-\lambda}$$

– Mean = λ
– Variance = λ

Interpret as the rate of occurrence of an event is equal to lambda in a finite period of time.



Data Distributions (Continuous)

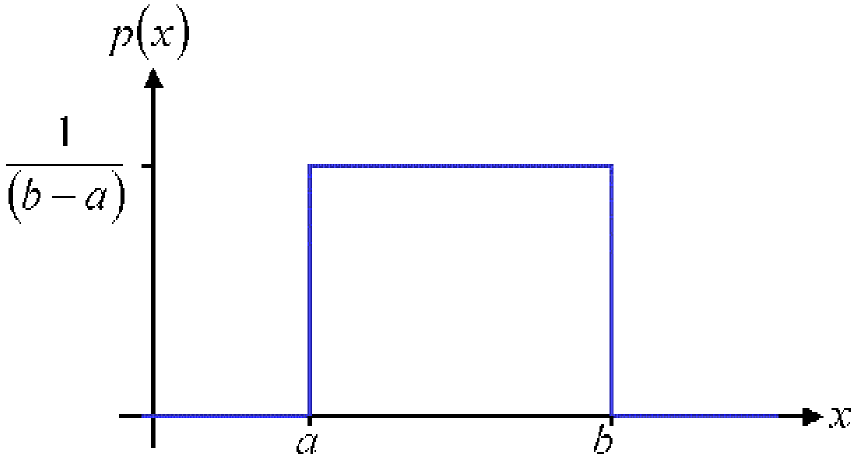
- Continuous Distribution Properties
  - Area under the curve must be equal to 1.
  - Probability of event equal to AREA under curve.
  - No negative values.
  - Probability of a single, exact value is 0.

Data Distributions (Continuous)

- Uniform (flat, bounded)

$$P(x) = \begin{cases} \frac{1}{(b-a)} \; if \; a \leq x \leq b \\ 0 \; if \; x < a \; or \; x > b \end{cases}$$

- Very useful for parameter priors. (future discussion)
  - Mean=(a+b)/2
  - Variance=(1/12)(b-a)^2

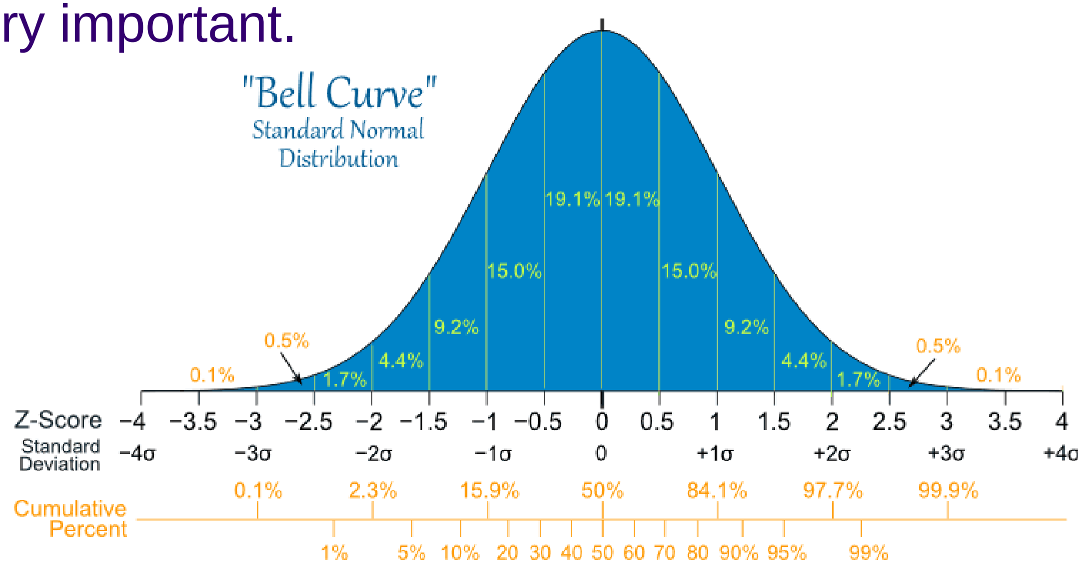


Data Distributions (Continuous)

> Normal (Gaussian) distribution
  - Most common and occurs naturally.
  - Defined by a mean and variance only. (standard = N(0,1))

$$P(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

  - Has very nice properties.
  - Tests for normality are very important.



Data Distributions (Continuous)

- Student’s T (normal for small samples)
  - Important for hypothesis testing smaller sample sizes.
  - Used for:
    - Testing of mean value when st. dev. is unknown.
    - Testing difference between two distribution means.
  - Looks very similar to the normal distribution.

Data Exploration (Descriptive Statistics)

> Purpose: To gain a clear understanding of your data.
  - How large is it?
  - What columns are of interest?
  - Missing data?
  - Outliers?

Numerical Exploration

> str(): structure of the data frame

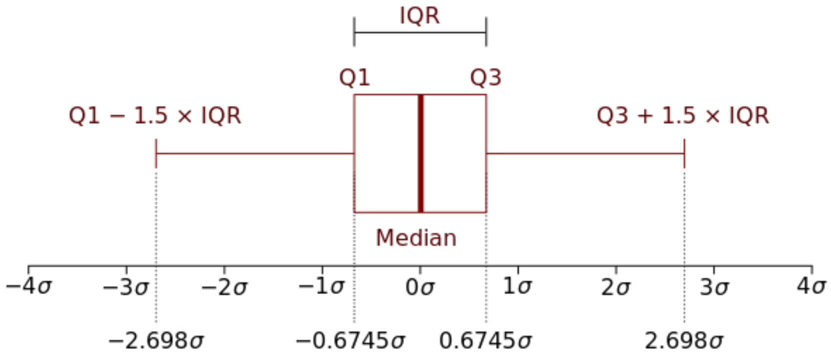
> summary(): summary of each of the columns

> head() / tail(): top / bottom of data frame

> table(): frequency table

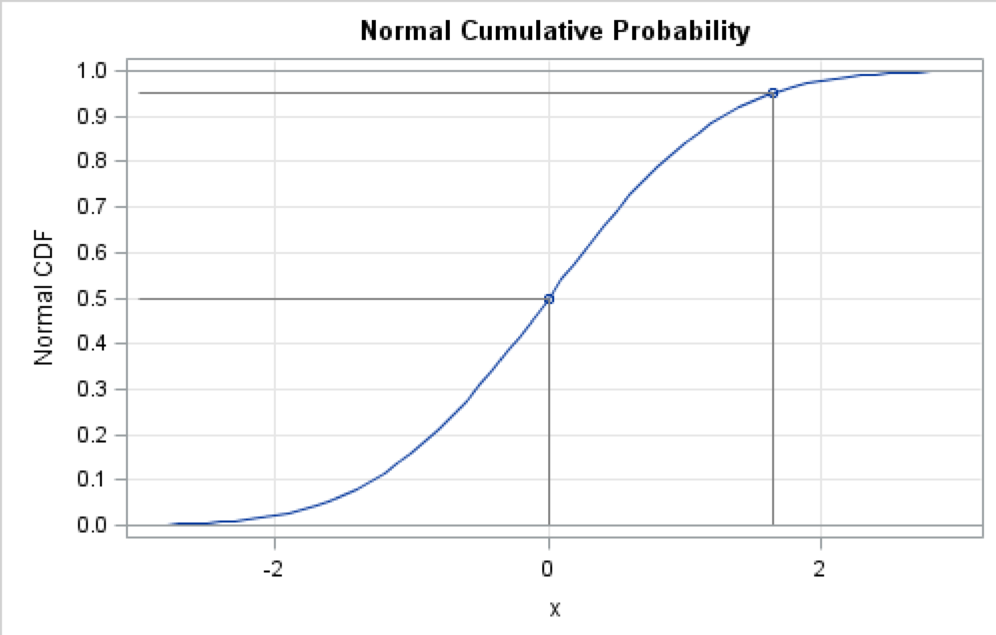
Numerical Exploration

> IQR(): inner quartile range (Q3 – Q1)



Numerical Exploration

- quantile(): quantiles of numerical vectors
  - Quantiles are inverse values of the CDF (cumulative distribution function).
  - Standard Normal: (shown in figure)
    - Quantile(0.5) = 0, means at x=0, 50% of the distribution lies to the left. (This is also the median)
    - Quantile(0.95) = 1.65



Numerical Exploration

- Relationships:
  - cov(): covariances

$$cov(x, y) = E\big((x - \mu_x)(y - \mu_y)\big)$$

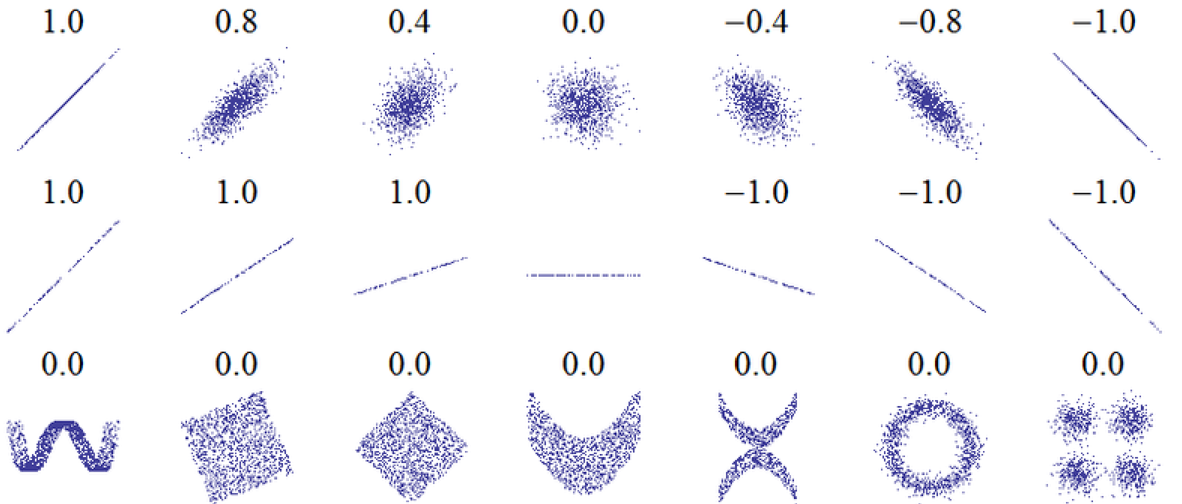
  - Interpretation: Expected value of the differences between x and y and their corresponding mean.
  - E.g. if x is above it's mean when y is also above it's mean, then they will have a high covariance.
  - Highly interpretable, but not bounded.

Numerical Exploration

> Relationships:
  - cor(): correlations (pearsons)

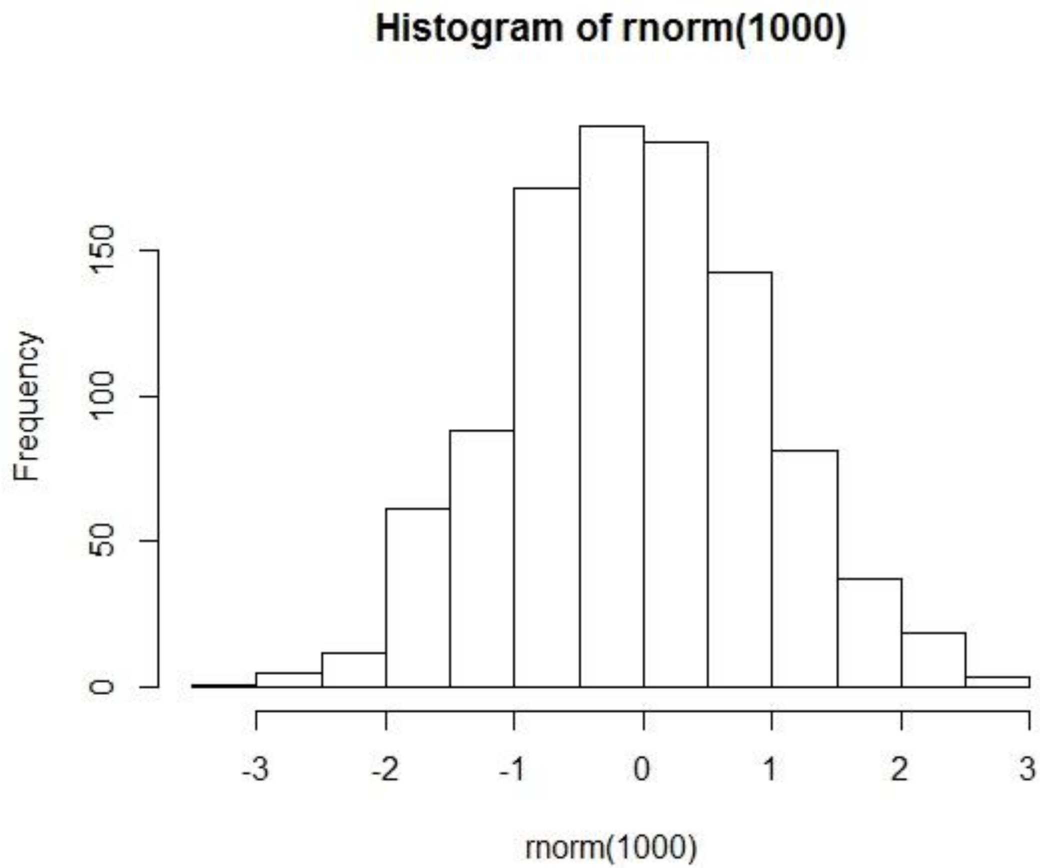
$$cor(x,y) = \frac{E\big((x-\mu_x)(y-\mu_y)\big)}{\sigma_x \sigma_y}$$

  - Bounded between 0 and 1.
  - Not as interpretable.



Visual Exploration

> Histograms:

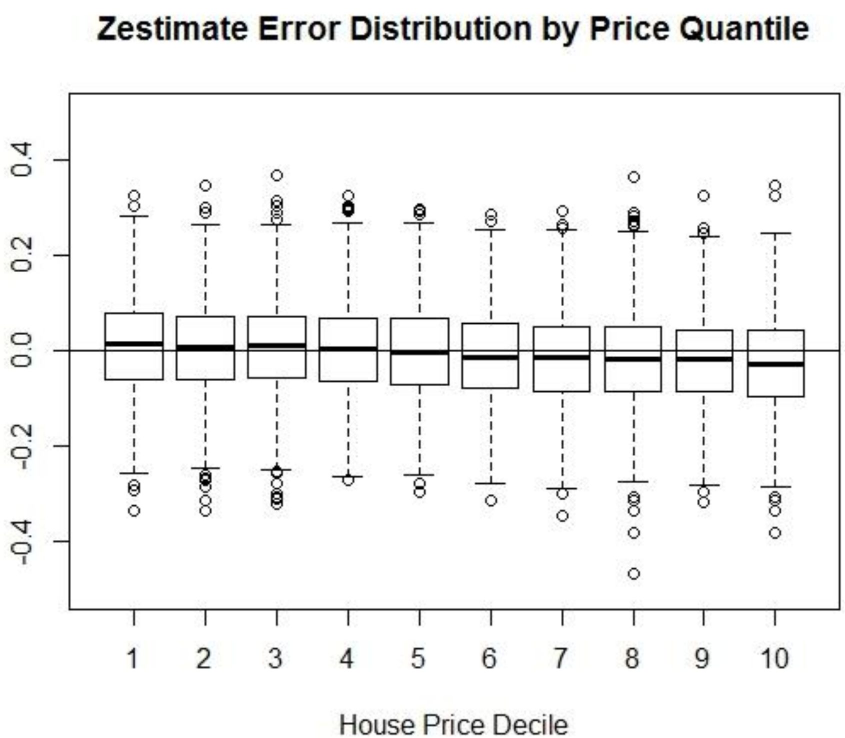


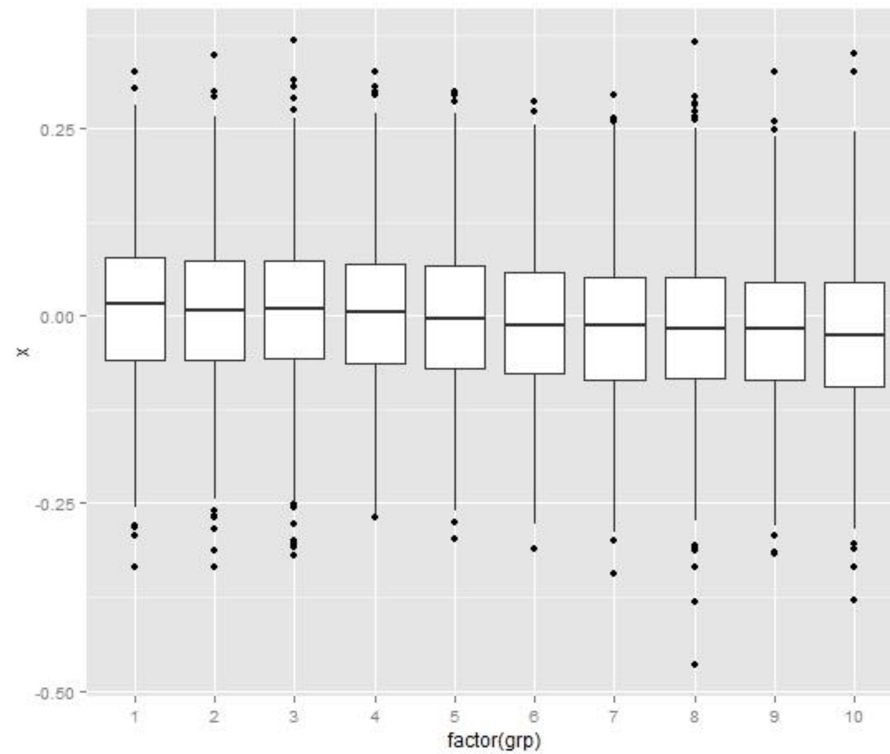
Base:
hist()

ggplot2:
+ geom_histogram()

Visual Exploration

> Boxplots:



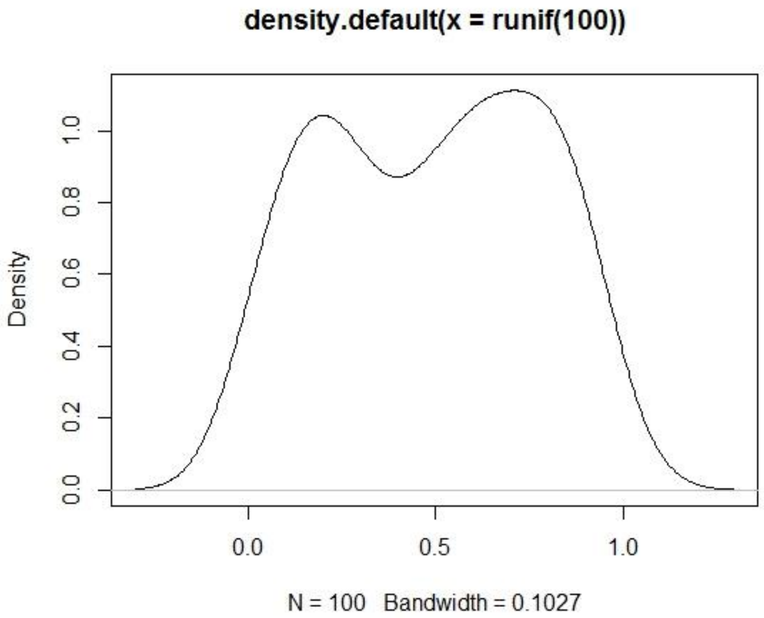


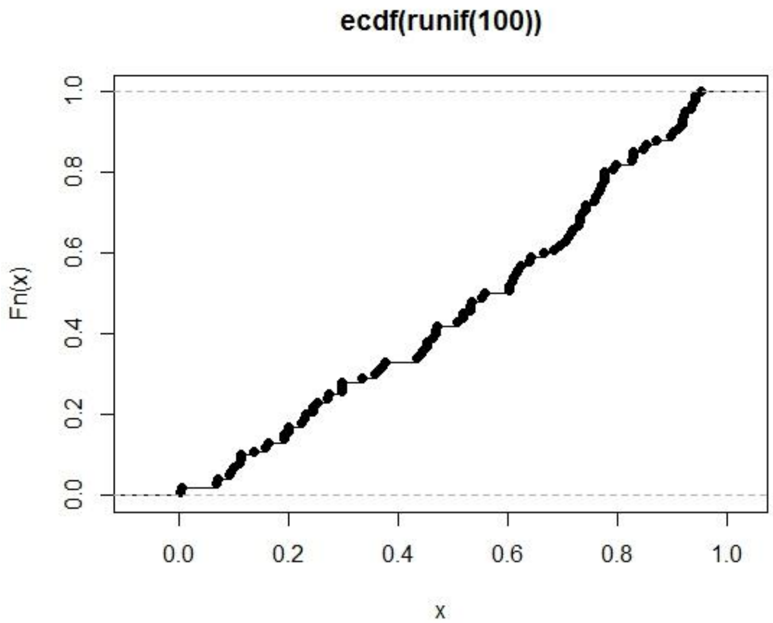
Base:
boxplot()

ggplot2:
+ geom_boxplot()

Visual Exploration

> Densities/CDFs:



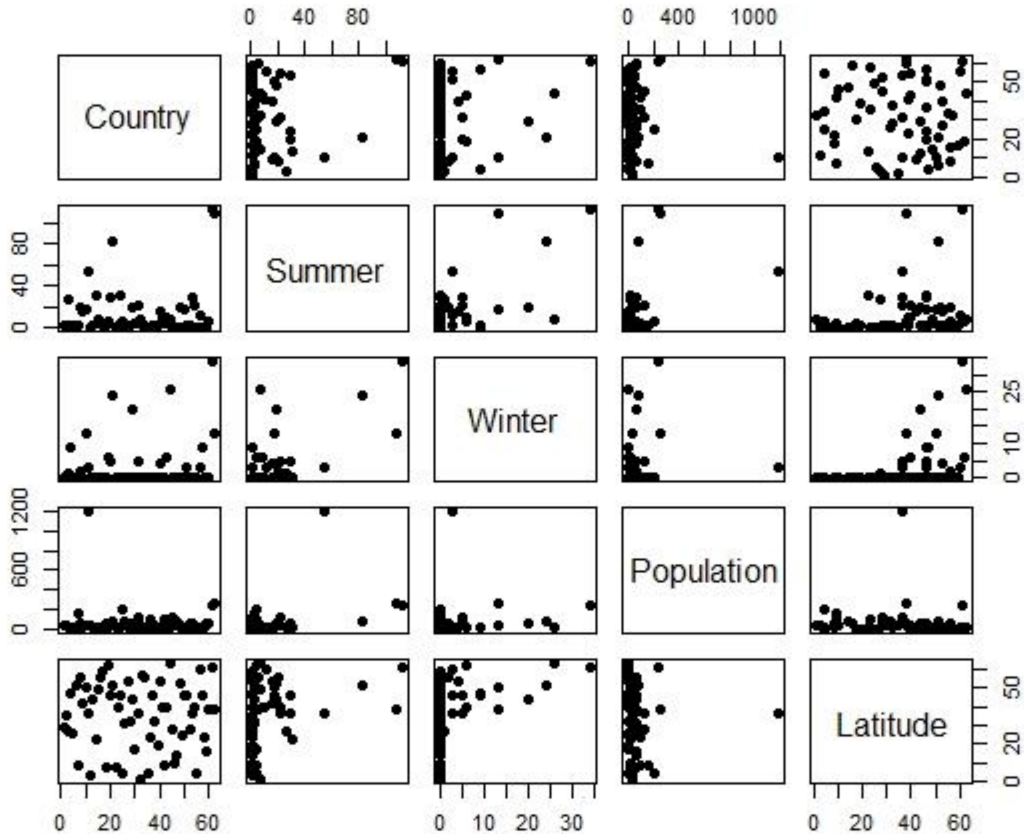


Base:
plot(density())
plot(ecdf())

ggplot2:
+ geom_density()
+ stat_ecdf()

Visual Exploration

> Scatterplots



Base:
pairs()

ggplot2:
ggpairs()

Distribution Transformations

- The purpose of transforming a variable is to make it easier to distinguish between values.
  - Most commonly we are looking to transform a distribution to be normal.
- Common Transformations
  - Log-based:
    - Log(x), log(x+1), log(x-min(x) + 1)
  - N-th Root based:
    - X^(1/n)
  - Any combination you can think of (remembering math rules).
- We will cover normality tests in a later class.

Simpsons Paradox

> Slicing up data in different ways can create different results.

Department	#male applicants	#female applicants	%male admit	%female admit
A	825	108	62	82
B	560	25	63	68
C	325	593	37	34
D	417	375	33	54

The explanation is that women applied in larger numbers to departments that had lower admission rates.

> http://vudlab.com/simpsons/

Production Level Scripts

- Logging
- Functionalize everything possible
- interactive()
- R-example: Weather Scraping R script

Assignment

- Go to:
  - Vote for extra topics (time permitting)
- Complete Homework 1:
  - Explore 'JitteredHeadCount.csv', a data set from Caesar's Entertainment that has falsified/jittered table headcounts.
  - Write ***production level*** script that shows/illustrates 3 key takeaways of your choosing from exploring the data.
  - You should submit:
    - **ONE R-script.**
    - **One word document with 5 key points.** (example next page).

Example Takeaway

> The aggregate table headcounts on the weekends are X% higher than non-weekends (figure 1). In fact, the game that has the highest difference between average highs and average low days is Gamecode AA with a difference of 5.71 heads/table.