# Basic Statistical Inference

## Graphical Exploratory Data Analysis

### Introduction to exploratory data analysis

#### Exploratory to exploratory data analysis

* The process of organizing, plotting, and summarizing a data set

#### Plotting a histogram

* Always label your axes
* Sns.set() sets seaborn to default settings
* Square root rule
  + Choose the number of bins to be the square root of the number of samples
* Binning bias
  + The same data may be interpreted differently depending on choice of bins
    - Use bee swam plot to avoid this

#### Plotting all of your data: Bee swam plots

* Organization of the data frame
  + It must be where every column is a feature and every observation is a row
* They avoid the binning bias of a histogram

#### Empirical Cumulative Distribution Functions



## Quantitative Exploratory Data Analysis

### Percentiles, outliers, and box plots

### Variance and Std Dev

### Covariance and Pearson correlation coefficient

## Thinking probabilistically – Discrete Variables

### Probabilistic logic and statistical inference

### Random number generators and hacker statistics

### The Binomial distribution

* Probability mass function
  + The set of probabilities of discrete outcomes
* Binomial distribution stories
  + The number of r of successes in n Bernoulli trials with probability p of success

### Poisson processes and the Poisson distribution

#### Poisson process

* The timing of the next event is completely independent of when the previous event happened

#### Examples

* Natural births in a given hospital
* Hit on a website during a given hour

#### Poisson Distribution

* The number r of arrivals of a Poisson process in a given time interval with average rate of lambda arrivals per interval

#### Relationship between Poisson and Binomial

* When we have rare events (low p, high n) the Binomial distribution is Poisson.

## Continuous Variables

### Probability Density Functions

* Mathematical description of the relative likelihood of observing a value of a continuous variable

#### Interpreting PDFs

#### Interpreting CDFs

### The Normal Distribution

#### The Normal PDF

#### The Normal CDF

### The Normal Distribution: Properties and warnings

### The Exponential Distribution

* The waiting time between arrivals of a Poisson process is Exponentially distributed

## Parameter Estimation by Optimization

### Optimal Parameters

* Parameter values that bring the model in closest agreement with the data

### Linear Regression by least squares

### The importance of EDA

## Bootstrap Confidence Intervals

### Generating bootstrap replicates

* Bootstrapping is the use of resampled data to perform statistical inference
* Use the resampling engine: np.random.choice()

### Bootstrap confidence intervals

### Pairs bootstrap

#### Nonparametric inference

* Make no assumptions about the model or probability distribution underlying the data

#### Pairs bootstrap for linear regression

* Resample data in pairs
* Compute slope and intercept from resampled data
* Each slope and intercept is a bootstrap replicate
* Compute confidence intervals from percentiles of bootstrap replicates

## Introduction to hypothesis testing

### Formulating and simulating a hypothesis

* We permutated the samples

### Test statistics and p-values

### Bootstrap hypothesis tests

#### Pipeline for hypothesis testing

* Clearly state the null hypothesis
* Define your test statistic
* Generate many sets of simulated data assuming the null hypothesis is true
* Compute the test statistic for each simulated data set
* The p-value is the fraction of your simulated data sets for which the test statistic is at least as extreme as for the real data

#### One Sample test

* Compare one set of data to a single number

### Permutation Tests

* Permutations tests are exact because it uses all the data and elimantes any correlation between the variables
* It is used to see if samples come from the same distribution.

## Hypothesis Test Examples

### A/B Testing

* Used by organizations to see if a strategy change gives a better result
* The null hypothesis of an A/B test is that the test statistic is impervious to the change

### Test of correlation

# A/B Testing

## What is A/B Testing?

* Method for measuring two versions of one element so you can determine which one is more successful
* Things commonly tested
  + Layout
  + Design
  + Style
  + Pricing
  + Promotions
  + Headlines
  + Copywriting
  + Images
* You must run your own tests to see what works for your site.

## A/B Split Testing vs. Multivariate Testing

* Multivariate is the same but you have more variables to measure than the outcomes you desire
* Cons
  + Increased complexity and possibly duration of testing period

## A/B Testing Tools

* Google analytics content research
  + Ideal starting point
  + Free
* Optimizely
  + Let’s you perform simple and advanced A/B test via a point and click WYSIWYG editor
  + Paid
* Visual website optimizer
* Unbounce
* Webtrend’s conversion optimizer
  + Geared for testing enterprise clients