NYC DATA SCIENCE ACADEMY

Foundations of Statistics

Data Science with R: Machine Learning

Outline

- Part 1: All About Your Data
- Part 2: Statistical Inference
- Part 3: Introduction to Machine Learning
- Part 4: Review

PART 1

All About Your Data

Motivation

- Begin to explore the answers to questions like:
  - What is the distribution of MPG for various types of automobiles?
  - What is the efficacy of a new drug in a clinical trial?
    - Does gender impact the outcome?
    - Does age impact the outcome?
  - Is there a correlation between income and life expectancy?
  - Are you more likely to be imprisoned for a crime in different areas of the country?

Descriptive Statistics

- Before jumping into complex machine learning algorithms, it is important to take a step back and assess the nature of our variables:
  - What are the central tendencies?
  - What is the spread of the values?
  - How much do the values vary?
  - Are there any abnormalities that stand out?

- Having answers to these questions will help us make informed decisions about which machine learning algorithms will likely be appropriate.

Measures of Centrality

- The mean μ is defined for a numeric (quantitative) vector x as:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- The mean helps describe the average of the values; however, it is not robust to extreme values.
  - In the presence of extreme values, the perception of the mean may not be representative of the overall variable's central tendency.

- Along with the mean, it is helpful to examine the median of a vector as well. The median represents the 50th percentile, or "middle value", of a vector, and is robust to extreme values.

Measures of Centrality

- Consider the following example:

	Vector #1	Vector #2
x_1	1	1
x_2	2	2
x_3	3	15
Mean	2	6
Median	2	2

- For vectors that do not contain extreme values (like Vector #1), the mean and the median tend to be similar.
- For vectors that do contain extreme values (like Vector #2), the mean tends to be skewed towards the abnormalities.

Measures of Variability

- The variance σ^2 (or the standard deviation σ) is defined for a numeric vector x as:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)^2$$

- The variance helps describe the spread of the values.

- Note that the variance is calculated using the mean; thus, it is also not robust to extreme values. Is this an issue? No!
  - A small variance indicates that the data points tend to be close to the mean, and thus are similar in value.
  - A high variance indicates that the data points tend to be far from the mean, and thus are dissimilar in value.

Measures of Variability

- Consider the following example:

	Vector #1	Vector #2
x_1	1	1
x_2	2	2
x_3	3	15
Variance	0.667	**40.667**

- For vectors that do not contain extreme values (like Vector #1), the variance tends to be relatively small.
- For vectors that do contain extreme values (like Vector #2), the variance tends to be relatively large.

Frequency, Proportion, & Contingency Tables

- For categorical (qualitative) variables, it does not make sense to discuss the mean and variance because there is no sense of natural order:
  - What is the mean of *gender* (male, female)?
  - What is the variance of *state* (NY, CT, FL, etc.)?

- Instead, we can get an idea of these variables and investigate their relationships by inspecting frequency, proportion, and contingency tables.

Frequency, Proportion, & Contingency Tables

- A frequency table displays the number of times a data value occurs in a vector:

	Gender
Male	100
Female	200

	State
NY	50
CT	175
FL	75

- A proportion table displays the fraction of occurrence of a data value in a vector:

	Gender
Male	0.333
Female	0.667

	State
NY	0.167
CT	0.583
FL	0.250

Frequency, Proportion, & Contingency Tables

- A contingency table displays the frequencies/proportions among multiple categorical variables simultaneously:

	NY	CT	FL
Male	25	25	50
Female	25	150	25

- This helps us understand the conditional relationships among our categorical variables. For example, in our dataset:
  - "A New Yorker is equally likely to be male or female."
  - "There is a 75% chance that an arbitrary female is from Connecticut."

- **NB:** We can reconstruct individual variable frequencies/proportions by totaling the respective rows or columns; these are called margins.

Correlation

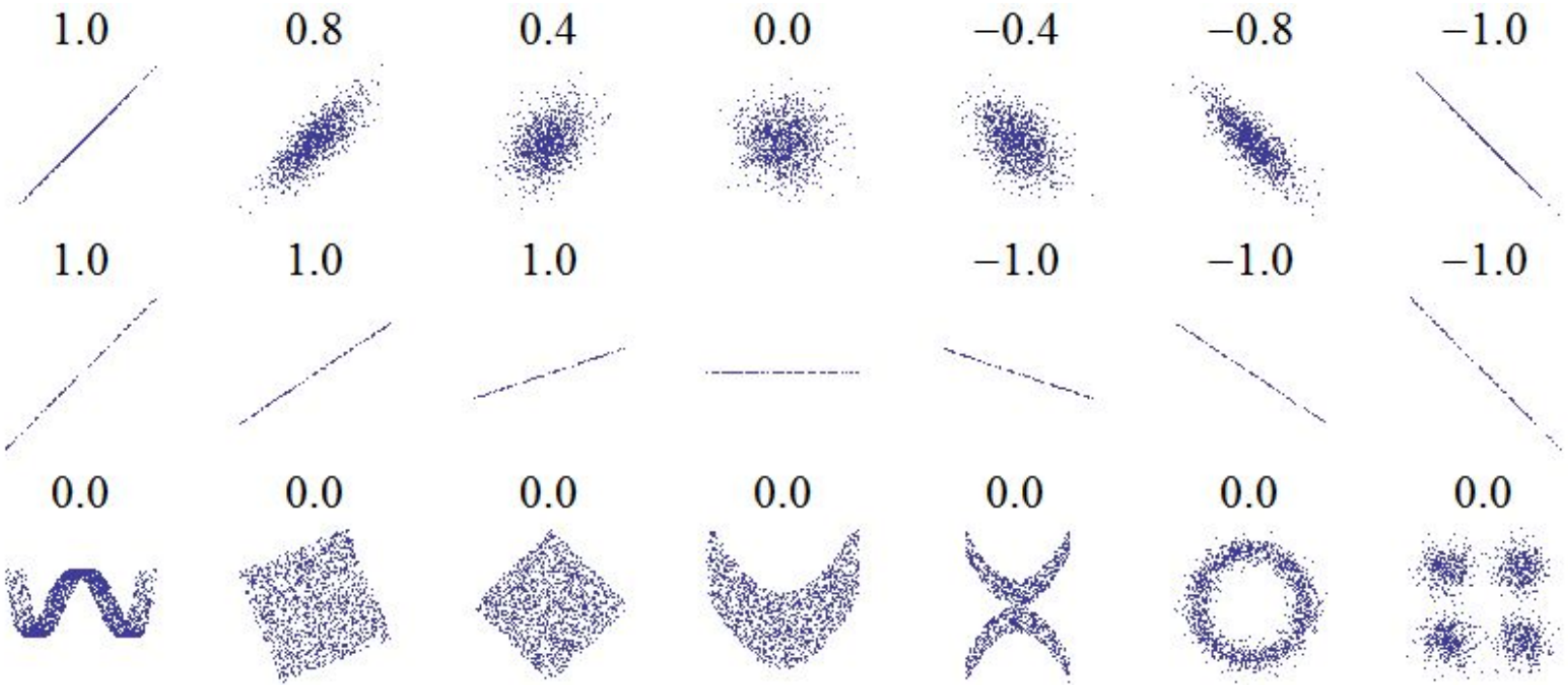
- The correlation ρ is defined for a numeric vectors x and y as:

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)}{n\sigma_x\sigma_y}$$

- The correlation helps quantify the linear dependence between two quantities (e.g., does knowing something about one variable inform anything about the other).

- Note the bounds of correlation: $-1 \leq \rho \leq 1$
  - A positive correlation means that the variables have a direct relationship (i.e., an increase in one variable usually implies an increase in the other).
  - A negative correlation means that the variables have an indirect relationship (i.e., an increase in one variable usually implies a decrease in the other).

Correlation



- **NB:** Independent variables necessarily have a correlation of 0; however, a correlation of 0 does not imply the variables are independent!

Correlation

- Examples of positive correlation:
  - The longer a person runs on a treadmill, the more calories they burn.
  - The taller a person gets, the heavier they will weigh.
  - The faster a roller coaster goes, the higher the g-forces a rider will experience.

- Examples of negative correlation:
  - The more classes a student misses, the lower their grade.
  - The more alcohol one consumes, the less judgment one has.
  - The faster a biker rides their bike, the sooner they will reach the finish line.

PART 2

Statistical Inference

What is Statistical Inference?

- Statistical inference is the process of deducing properties of an underlying distribution by analysis of data.
  - We aim to reach an educated conclusion about a population by inferring behavior from a sample.

- In order to infer conclusions about our population, we first have to pose a question that we wish to answer.
  - A statistical hypothesis is a scientific supposition that is testable on the basis of observing a process that is modeled by a set of random variables.

- Depending on the results of our hypothesis test, we can gain insight to the various behaviors existent in the population.

The Structure of a Hypothesis Test

1. State the null and alternative hypotheses:
   a. Null Hypothesis (H_0): The assumed default scenario in which nothing abnormal is observed (e.g., there is no difference among groups, etc.).
   b. Alternative Hypothesis (H_A): The scientific supposition we desire to test that contrasts H_0 (e.g, there is a difference among groups, etc.).

2. Assume that the null hypothesis is true. Calculate the probability (p-value) of observing results at least as extreme as what is present in your data sample.
   a. Because this part tends to be tedious, we most often use a table of theoretical quantities or a computer to help us with these calculations.

3. Based on the p-value, decide which hypothesis is more likely. Generally:
   a. If the p-value is > 0.05, retain the H_0.
   b. If the p-value is < 0.05, reject the H_0 in favor of H_A.

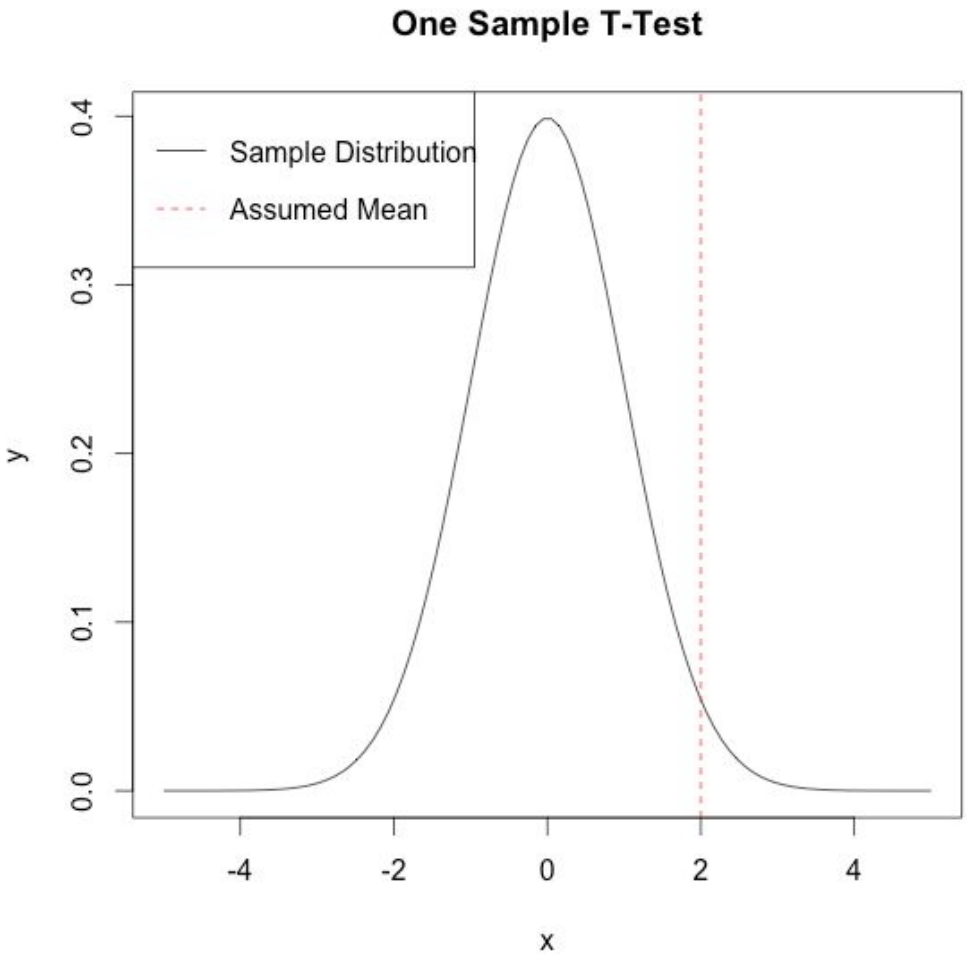
Basic Hypothesis Testing: One Sample T-Test

- When do we use the One Sample T-Test?
  - To examine the average difference between a sample and the known value of the population mean.
- Assumptions:
  - The population from which the sample is drawn is normally distributed.
  - Sample observations are randomly drawn and independent.
- P-value calculation:
  - Calculate the *t** test statistic, given by:

$$t^* = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim t_{n-1}$$

  - Compare the test statistic value with a standard table of t-values to determine whether the test statistic surpasses the threshold of statistical significance (yielding a significant p-value).

Basic Hypothesis Testing: One Sample T-Test



Basic Hypothesis Testing: One Sample T-Test

- Example:
  - You are told that the average height of a person in your building (μ) is 68 inches; however, you think the average person is actually much taller.

- For this scenario:
  - Null Hypothesis (H_0): μ = 68 inches
  - Alternative Hypothesis (H_A): μ > 68 inches

- Upon collecting a random sample of independent height measurements of people in your building, you can calculate the t-statistic.

Basic Hypothesis Testing: One Sample T-Test

- Suppose your data is as follows:
  - You collected 100 measurements (n = 100).
  - The average height, in inches, of your sample is 70 ($\bar{x}$ = 70).
  - The standard deviation of your sample is 1 (s = 1).

- Calculate the t-statistic as follows:

$$t^* = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{70 - 68}{1/\sqrt{100}} = 20$$

- Assuming the null hypothesis is true, we would expect to see a t-statistic at least as extreme as 20 less than 0.00001% of the time! (The p-value << 0.05)

- We have strong evidence to reject the null hypothesis in favor of the alternative that the average height is greater than 68 inches.

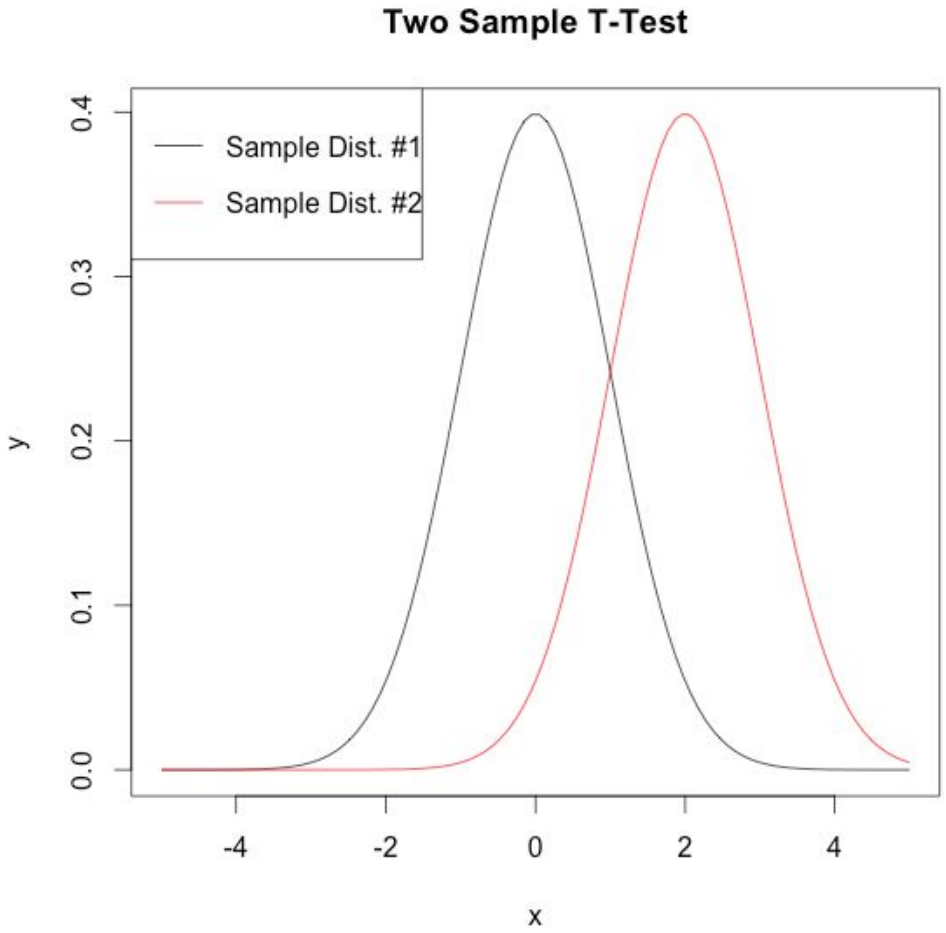
Basic Hypothesis Testing: Two Sample T-Test

- When do we use the Two Sample T-Test?
  - To examine the average difference between two samples drawn from two different populations.
- Assumptions:
  - The populations from which the samples are drawn are normally dist.
  - The standard deviations of the two populations are equal.
  - Sample observations are randomly drawn and independent.
- P-value calculation:
  - Calculate the *t** test statistic, given by:

$$t^* = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim t_{n_1+n_2-2}$$

  - Compare the test statistic value with a standard table of t-values to determine whether the test statistic surpasses the threshold of statistical significance (yielding a significant p-value).

Basic Hypothesis Testing: Two Sample T-Test



Basic Hypothesis Testing: Two Sample T-Test

- Example:
  - Although the difficulty of the SAT should not vary across its different administrations, you believe that timing is everything; you suppose that there is a difference in the average score from tests taken in spring and fall.

- For this scenario:
  - Null Hypothesis (H_0): $\mu_{Spring} = \mu_{Fall}$
  - Alternative Hypothesis (H_A): $\mu_{Spring} \neq \mu_{Fall}$

- Upon collecting two random samples of independent SAT scores, one from a spring administration and one from a fall administration, you can calculate the t-statistic.

Basic Hypothesis Testing: Two Sample T-Test

- Suppose your data is as follows:
  - You collected 180 scores ($n_{\text{Spring}} = 100$, $n_{\text{Fall}} = 80$).
  - The average score for spring was 1,550 and for fall was 1,500.
  - The standard deviation of your spring sample is 200; for fall, 210.

- Calculate the t-statistic as follows:

$$t^* = \frac{\bar{x}_{Spring} - \bar{x}_{Fall}}{\sqrt{\frac{s^2_{Spring}}{n_{Spring}} + \frac{s^2_{Fall}}{n_{Fall}}}} = \frac{1{,}550 - 1{,}500}{\sqrt{\frac{200^2}{100} + \frac{210^2}{80}}} \approx 1.62$$

- Assuming the null hypothesis is true, we would expect to see a t-statistic at least as extreme as 1.62 about 13.28% of the time. (The p-value > 0.05)

- We do not have strong evidence to reject the null hypothesis; the average spring and fall SAT scores are the same.

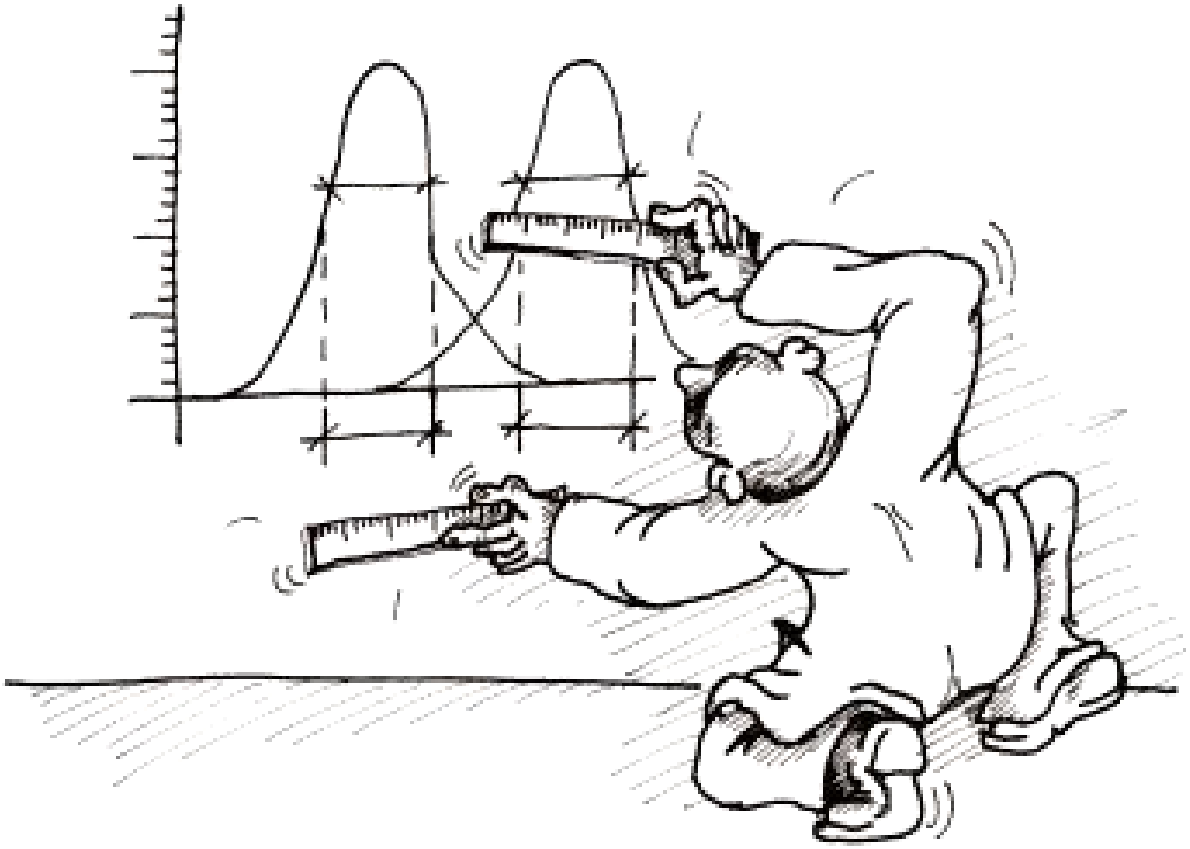
Basic Hypothesis Testing: F-Test

- When do we use the F-Test?
  - To assess whether the variances of two different populations are equal.
- Assumptions:
  - The populations from which the samples are drawn are normally dist.
  - Sample observations are randomly drawn and independent.
- P-value calculation:
  - Calculate the F^* test statistic, given by:

$$F^* = \frac{s_1^2}{s_2^2} \sim F_{n_1-1, n_2-1}$$

  - Compare the test statistic value with a standard table of F-values to determine whether the test statistic surpasses the threshold of statistical significance (yielding a significant p-value).

Basic Hypothesis Testing: F-Test



Image

Basic Hypothesis Testing: F-Test

- Example:
  - When we tested the difficulty of SAT exams in the previous example using the Two Sample T-Test, we should have had equal variances; however, the variances were slightly different. Were they significantly different?

- For this scenario:
  - Null Hypothesis (H_0): $\sigma^2_{Spring} = \sigma^2_{Fall}$
  - Alternative Hypothesis (H_A): $\sigma^2_{Spring} \neq \sigma^2_{Fall}$

- Upon collecting two random samples of independent SAT scores, one from a spring administration and one from a fall administration, you can calculate the F-statistic.

Basic Hypothesis Testing: F-Test

- Suppose your data is as follows:
  - The standard deviation of your spring sample is 200; for fall, 210.

- Calculate the F-statistic as follows:

$$F^* = \frac{s^2_{Fall}}{s^2_{Spring}} = \frac{210^2}{200^2} = 1.1025$$

- Assuming the null hypothesis is true, we would expect to see a F-statistic at least as extreme as 1.1025 about 65% of the time. (The p-value > 0.05)

- We do not have strong evidence to reject the null hypothesis; the variances of the spring and fall SAT score distributions are the same.

Basic Hypothesis Testing: One-Way ANOVA

- When do we use One-Way ANOVA?
  - To assess the equality of means of two or more groups. **NB:** When there are exactly two groups, this is equivalent to a Two Sample T-Test.
- Assumptions:
  - The populations from which the samples are drawn are normally dist.
  - The standard deviations of the populations are equal.
  - Sample observations are randomly drawn and independent.
- P-value calculation:
  - Calculate the F^* test statistic, given by:

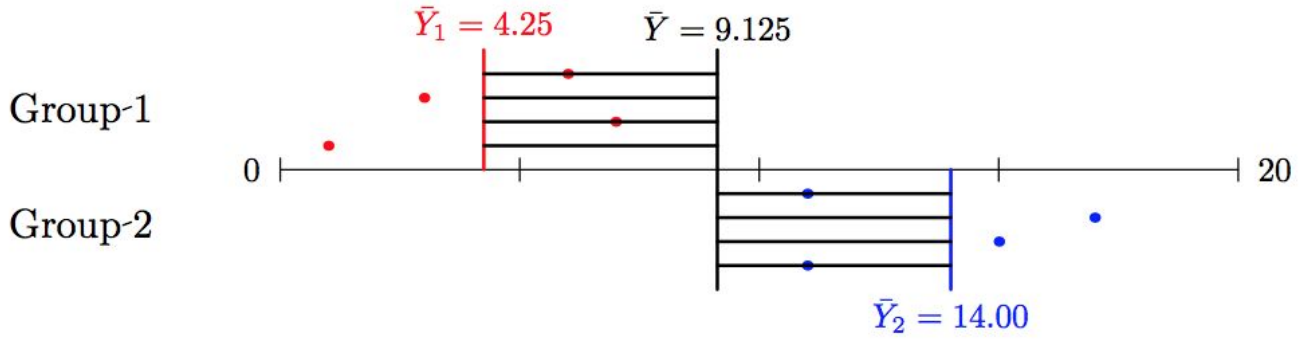
$$F^* = \frac{MS_{BetweenGroups}}{MS_{WithinGroups}} \sim F_{k-1,N-k}$$

  - Compare the test statistic value with a standard table of F-values to determine whether the test statistic surpasses the threshold of statistical significance (yielding a significant p-value).

Basic Hypothesis Testing: One-Way ANOVA

- Mean squares between groups:
  - A good estimate of the overall variance only when H_0 is true.
  - Quantifies the between-group deviations from the overall grand mean.
  - The formula is given below, where:
    - i is the index of the k groups.
    - n_i is the number of observations in the i^{th} group.

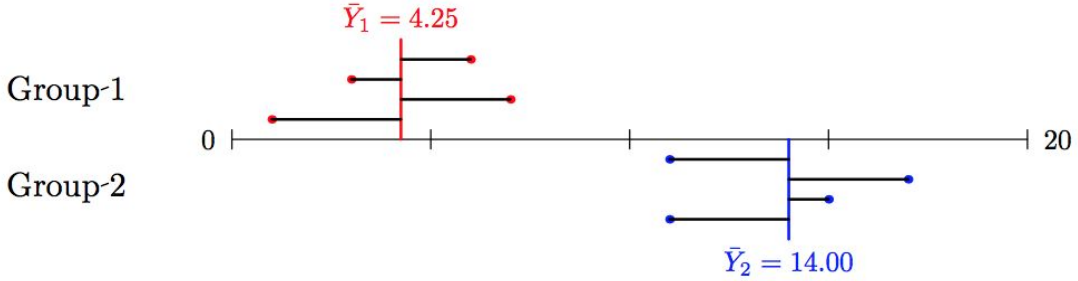
$$MS_{Between} = \frac{\sum_{i=1}^{k} n_i(\bar{Y}_i - \bar{Y})^2}{k-1}$$



Basic Hypothesis Testing: One-Way ANOVA

- Mean squares within groups:
  - A good estimate of the overall variance, unaffected by whether the null or alternative hypothesis is true.
  - Quantifies the within-group deviations from the respective group means.
  - The formula is given below, where:
    - i is the index of the k groups.
    - j is the index of the n_i observations of each group.
    - N is the overall number of observations.

$$MS_{Within} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2}{N - k}$$



Basic Hypothesis Testing: One-Way ANOVA

- Example:
  - You desire to test the efficacy of different types of diets on weight loss: low calorie, low carbohydrate, and low fat. You also have a control group for comparison purposes.

- For this scenario:
  - Null Hypothesis (H_0): $\mu_{\text{Low Calorie}} = \mu_{\text{Low Carbohydrate}} = \mu_{\text{Low Fat}} = \mu_{\text{Control}}$
  - Alternative Hypothesis (H_A): At least one of the average amounts of weight loss differs from the others.

- Upon collecting independent results from a clinical trial, you can calculate the F-statistic. (A manual calculation is omitted here for its tedious nature; we will observe how to automate this process in R.)

Basic Hypothesis Testing: χ^2 Test of Independence

- When do we use the χ^2 Test of Independence?
  - To test whether two categorical variables are independent.
- Assumptions:
  - Sample observations are randomly drawn and independent.
- P-value calculation:
  - Calculate the χ^2 test statistic, given by:

$$\chi^2 = \sum_{i=1}^{n_{rows}} \sum_{j=1}^{n_{cols}} \frac{(Observed_{ij} - Expected_{ij})^2}{Expected_{ij}} \sim \chi^2_{(n_{rows}-1)(n_{cols}-1)}$$

  - Compare the test statistic value with a standard table of χ^2-values to determine whether the test statistic surpasses the threshold of statistical significance (yielding a significant p-value).

Basic Hypothesis Testing: χ^2 Test of Independence

- Example:
  - A review session was held before a quiz was administered to a class of students. You would like to determine whether a student's grade on the quiz is dependent on whether or not they attended the review session.

- For this scenario:
  - Null Hypothesis (H_0): The two variables are independent of one another.
  - Alternative Hypothesis (H_A): The two variables are dependent on each other.

- Upon collecting independent samples on whether or not students went to the review session and whether or not they passed the quiz, you can calculate the χ^2-statistic.

Basic Hypothesis Testing: $\mathscr{X}^2$ Test of Independence

- Suppose your observed data is as follows:

Obs.	Pass	Fail	Total
Present	44	21	65
Absent	12	18	30
Total	56	39	95

- We then need to find the expected data under independence:

$$E_{ij} = \frac{n_{i.} \times n_{.j}}{N}$$

Exp.	Pass	Fail	Total
Present	38.32	26.68	65
Absent	17.68	12.32	30
Total	56	39	95

Basic Hypothesis Testing: $\mathscr{X}^2$ Test of Independence

- Calculate the $\mathscr{X}^2$ test statistic as follows:

$$\chi^2 = \frac{(44-38.32)^2}{38.32} + \frac{(21-26.68)^2}{26.68} + \frac{(12-17.68)^2}{17.68} + \frac{(18-12.32)^2}{12.32} \approx 6.5$$

- Assuming the null hypothesis is true, we would expect to see a $\mathscr{X}^2$-statistic at least as extreme as 6.5 about 1.08% of the time. (The p-value < 0.05)

- We have strong evidence to reject the null hypothesis in favor of the alternative that the attending the review session affected the likelihood of a student passing the pop quiz.

PART 3

Introduction to Machine Learning

What is Machine Learning?

- Machine learning developed from the combination of statistics and computer science; it aims to implement algorithms that allow computers to “learn” about the data it analyzes.

- Traditional algorithms require computers to follow a strict set of program instructions; machine learning algorithms instead assess the data at hand to make informed decisions.

- Machine learning algorithms have the ability to learn and adapt; traditional algorithms do not.

What is Machine Learning?

- A typical, explicitly hand-written computer algorithm often takes the following form:

```
if (...)
    then ...
    else if (...)
        if (...)
            then ...
            if (...)
                then ...
    else if ...
        then ...
else if ...
```

- Writing a program like this is not only tedious, but also can often not account for every scenario that might exist. If it did, the program could be endless!

What is Machine Learning?

- How would you write a program to tell a computer to directly recognize a tree? Why might this be difficult?
  - There are many different types of trees.
  - Different types of trees have different attributes.
  - Every tree doesn't necessarily share the same attributes.




- Instead, it would be easier to first observe a few examples of what is and is not a tree, and use that information to inform our future decision making.

What is Machine Learning?

- Which of the following problems are best suited for machine learning algorithms? Which are better for traditional hand-written algorithms?:
  a. Determining whether a number is prime.
  b. Determining whether credit card fraud is taking place.
  c. Determining when a falling object will hit the ground.
  d. Determining the optimal cycle for traffic lights at an intersection.
  e. Determining the age at which a particular medical test is recommended.

What is Machine Learning?

- Which of the following problems are best suited for machine learning algorithms? Which are better for traditional hand-written algorithms?:
  a. Determining whether a number is prime.
  b. Determining whether credit card fraud is taking place.
  c. Determining when a falling object will hit the ground.
  d. Determining the optimal cycle for traffic lights at an intersection.
  e. Determining the age at which a particular medical test is recommended.

- Answer:
  - Machine Learning: b, d, & e
  - Hand-written: a & c

Types of Machine Learning Algorithms

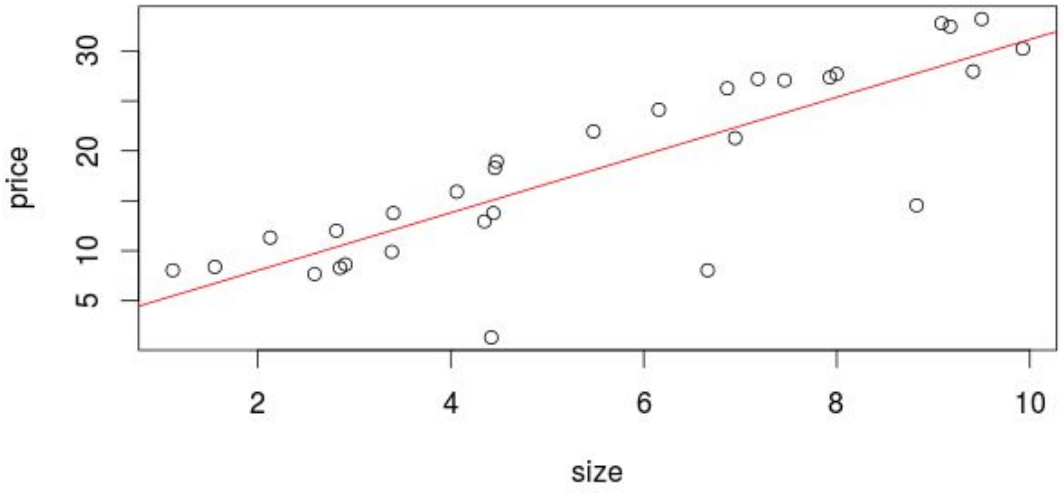
- Supervised Learning Algorithms:
  - Your data includes the "truth" that you wish to predict.
  - Use what you know about your observations to construct a model for future decision making.
    - Regression
    - Classification

- Unsupervised Learning Algorithms:
  - Your data does not include the "truth" that you wish to predict.
  - Use your data to find underlying structure to inform intrinsic behavior that is not already explicitly available.
    - Clustering
    - Dimension Reduction

Supervised Learning: Regression

- In regression, we aim to predict a continuous output given a slew of input variables. Our data contains the output that we wish to predict.

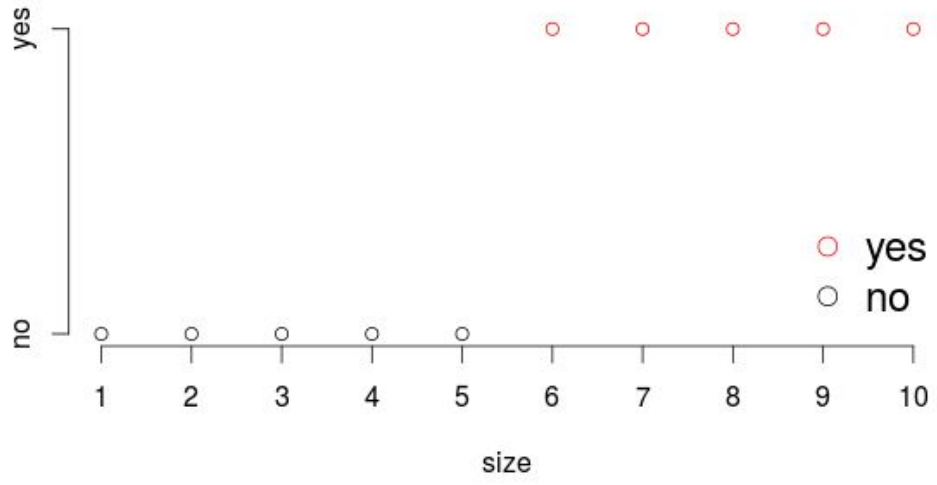
- Example: We are given a dataset of house sizes and prices. We want to examine the relationship between these two variables. Can we accurately predict the price of another house (not in our dataset) if we know its size?



Supervised Learning: Classification

- In classification, we aim to predict a categorical output given a slew of input variables. Our data contains the output that we wish to predict.

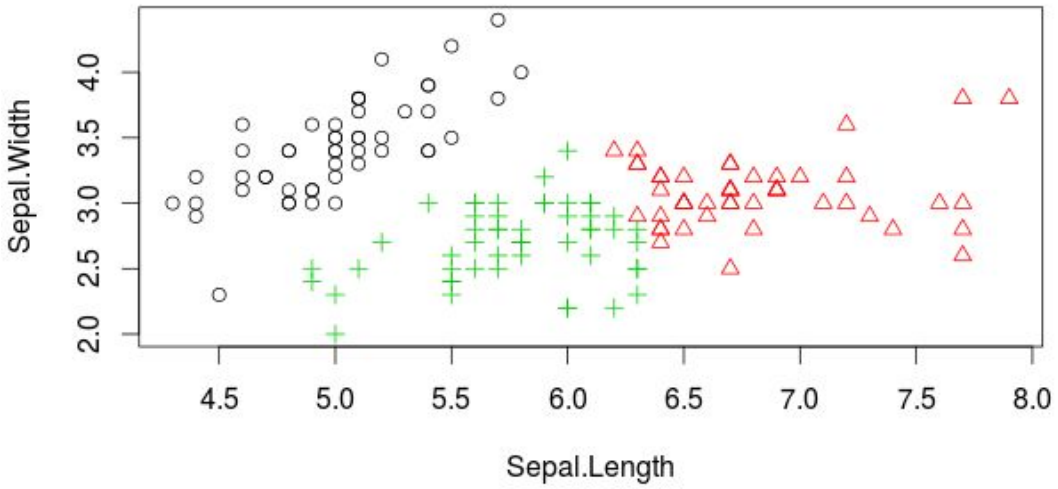
- Example: We are given a dataset of tumor sizes and malignancies. We want to examine the relationship between these two variables. Can we accurately predict the whether a tumor (not in our dataset) is malignant if we know its size?



Unsupervised Learning: Clustering

- In clustering, we aim to uncover commonalities in our data that help segment observations into different groups; within the groups, observations share some characteristics. Our data does not contain the group information that we seek.

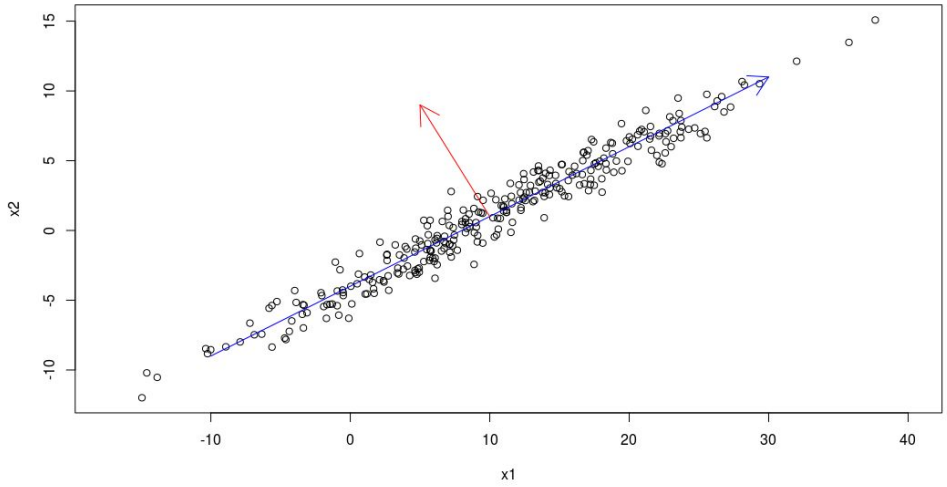
- Example: We are given a dataset of various flower measurements. Can we determine whether the flowers naturally fall into groups? Why might this be the case?



Unsupervised Learning: Dimension Reduction

- In dimension reduction, we aim to summarize massive amounts of data into smaller, more understandable components while retaining the structure of the original dataset. Our data does not tell us what the smaller components are.

- Example: We are given a dataset with highly correlated variables. Can we reduce the dimensionality of this dataset to eliminate structural redundancies without sacrificing information?



What is Machine Learning?

- Which machine learning technique would you use in the following scenarios?:
  a. You want to predict whether or not a student will be admitted to college based on their high school performance.
  b. You have a large dataset and believe there are some redundancies in your variables. You want to succinctly describe your data.
  c. You have measurements from different batches of rejected items from a manufacturing line. You want to discover similarities among defects.
  d. You want to understand the relationship between the age of a car and its estimated market value.

What is Machine Learning?

- Which machine learning technique would you use in the following scenarios?:
  a. You want to predict whether or not a student will be admitted to college based on their high school performance.
  b. You have a large dataset and believe there are some redundancies in your variables. You want to succinctly describe your data.
  c. You have measurements from different batches of rejected items from a manufacturing line. You want to discover similarities among defects.
  d. You want to understand the relationship between the age of a car and its estimated market value.
- Answer:
  a. Classification
  b. Dimension Reduction
  c. Clustering
  d. Regression

PART 4

Review

Review

- Part 1: All About Your Data
  - Descriptive Statistics
    - Measures of Centrality
    - Measures of Variability
    - Frequency, Proportion, & Contingency Tables
    - Correlation
- Part 2: Statistical Inference
  - Hypothesis Testing
    - One Sample T-Test
    - Two Sample T-Test
    - F-Test
    - One-Way ANOVA
    - $\mathcal{X}^2$ Test of Independence
- Part 3: Introduction to Machine Learning
  - Supervised Learning
    - Regression
    - Classification
  - Unsupervised Learning
    - Clustering
    - Dimension Reduction