NYU Center for Data Science

# DS-UA 112
# Introduction to Data Science

## Week 9: Lecture 2

## Models – Working with Random Variables

NYU Center for Data Science

How can we generalize from a sample to a population?

# DS-UA 112
# Introduction to Data Science

## Week 9: Lecture 2
## Models – Working with Random Variables

*Adapted from Dudoit, Nolan, Gonzalez, Lau*

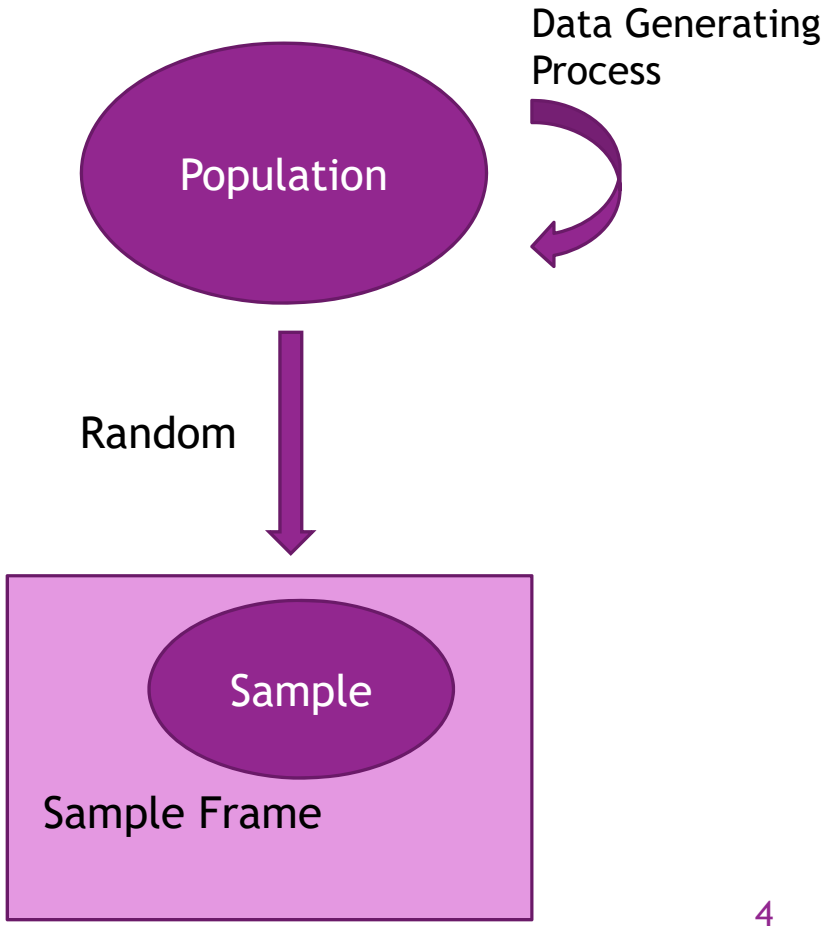
# Announcements

- Please check Week 9 agenda on NYU Classes
  - Lab 7
    - Due on Friday March 27 at 12PM
  - Project 1
    - Due on Monday April 6 at 12PM
  - Survey


https://nyu.qualtrics.com/jfe/form/SV_3DCWUa4yc08L0wt

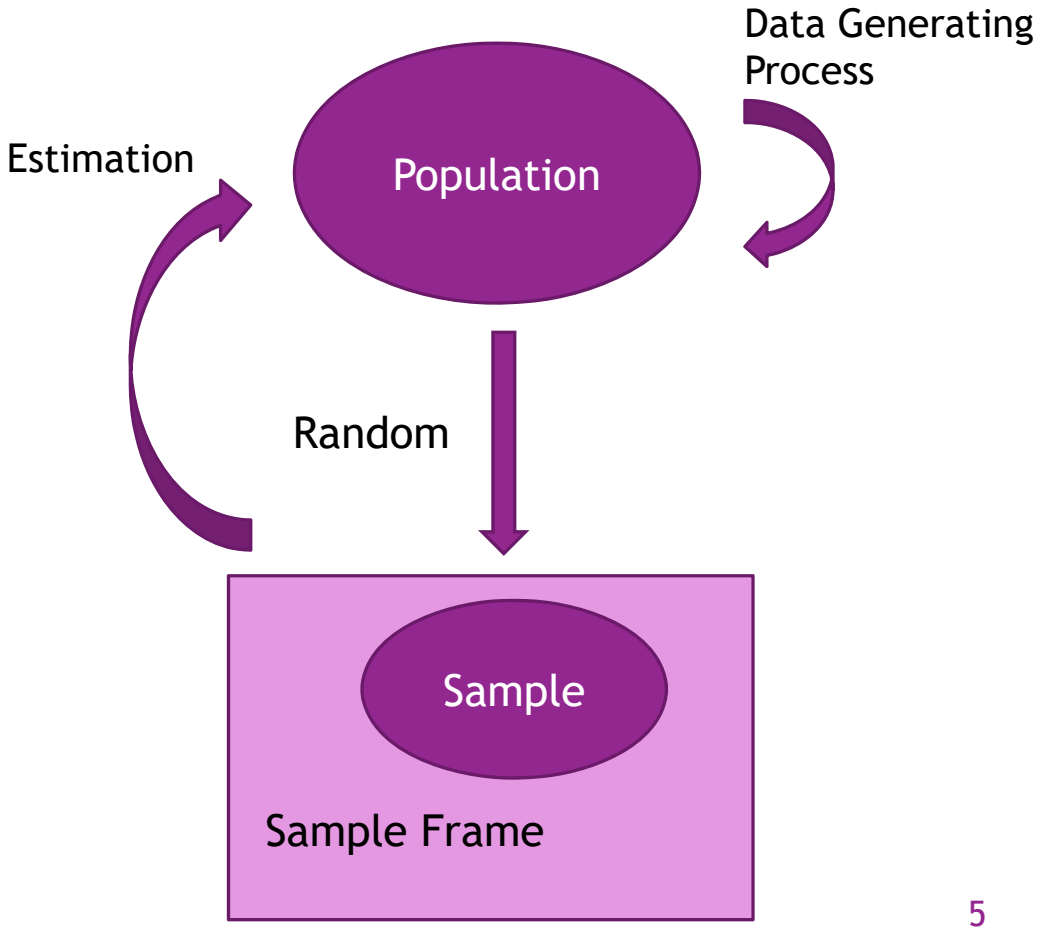
# Review

- Model
  - We want to understand a population through random samples
  - If we have guesses about the processes generating the data, then we can propose a model.
  - The data helps us to validate the assumptions behind the model



# Review

- Model
  - We can associate numbers called parameters to the population
  - A statistic is an estimate for a parameter obtained from a summary of data
  - How can we determine appropriate estimators of the parameters in the population?

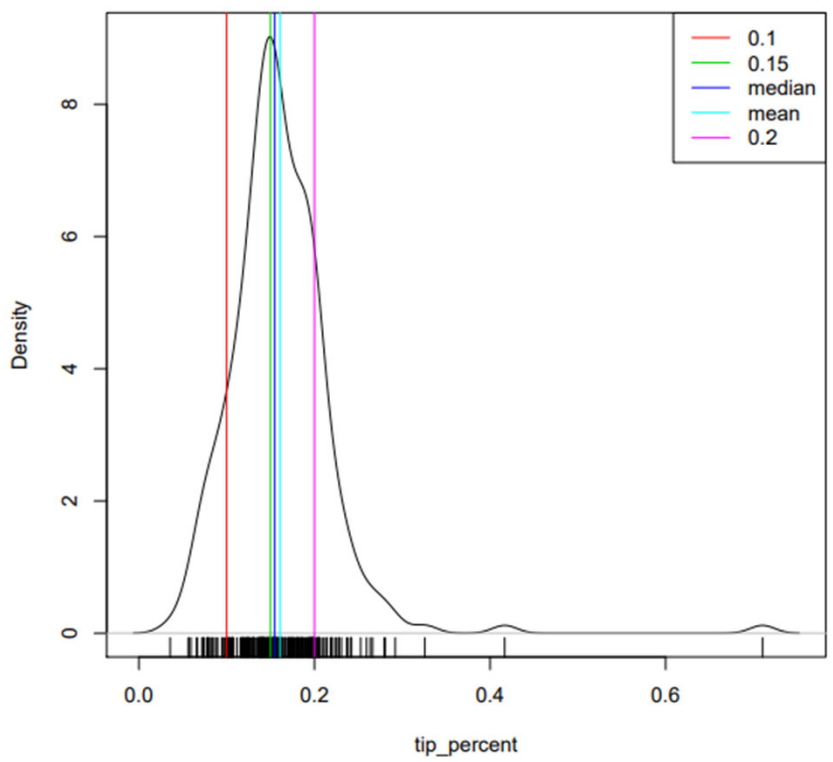


# Review

| | total_bill | tip | sex | smoker | day | time | size |
|---|---|---|---|---|---|---|---|
| **0** | 16.99 | 1.01 | Female | No | Sun | Dinner | 2 |
| **1** | 10.34 | 1.66 | Male | No | Sun | Dinner | 3 |
| **2** | 21.01 | 3.50 | Male | No | Sun | Dinner | 3 |
| **...** | ... | ... | ... | ... | ... | ... | ... |
| **241** | 22.67 | 2.00 | Male | Yes | Sat | Dinner | 2 |
| **242** | 17.82 | 1.75 | Male | No | Sat | Dinner | 2 |
| **243** | 18.78 | 3.00 | Female | No | Thur | Dinner | 2 |

- For example, the population could be tips at restaurants in the United States
- The sample could consist of 244 observations collected at a restaurant
- The customary amount for a tip is 15%. So our model could assert that tips are 15% on average.
- How could we estimate the average?

# Review




- For sample size n = 244 we can label the observations as $x_1,\ldots,x_n$
- We could summarize with the mean

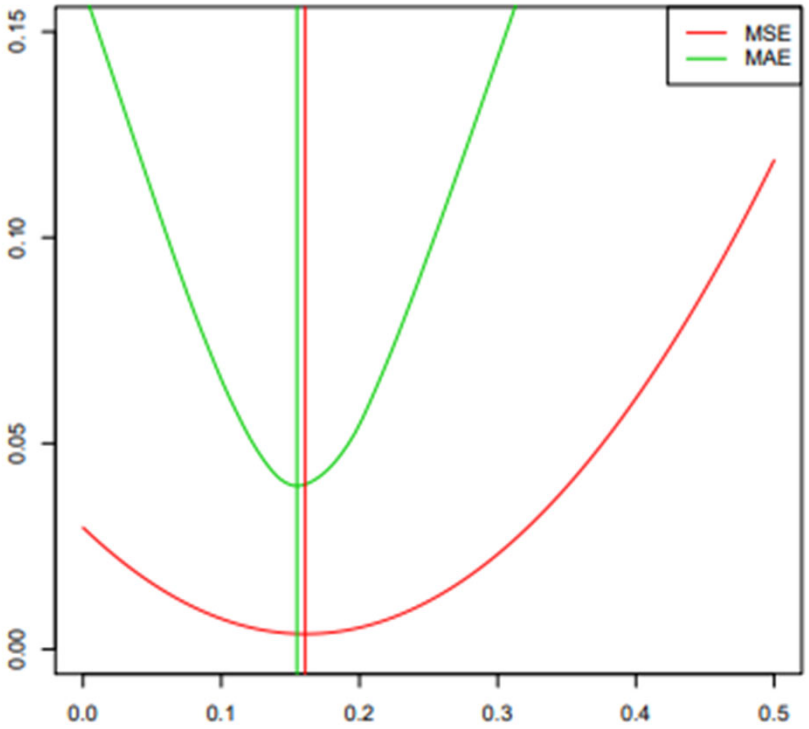
$$x_{mean} = \frac{1}{n}(x_1 + \ldots + x_n)$$

or the median $x_{median}$

$$\sum_{x_i < x_{median}} (1) = \sum_{x_i > x_{median}} (1)$$

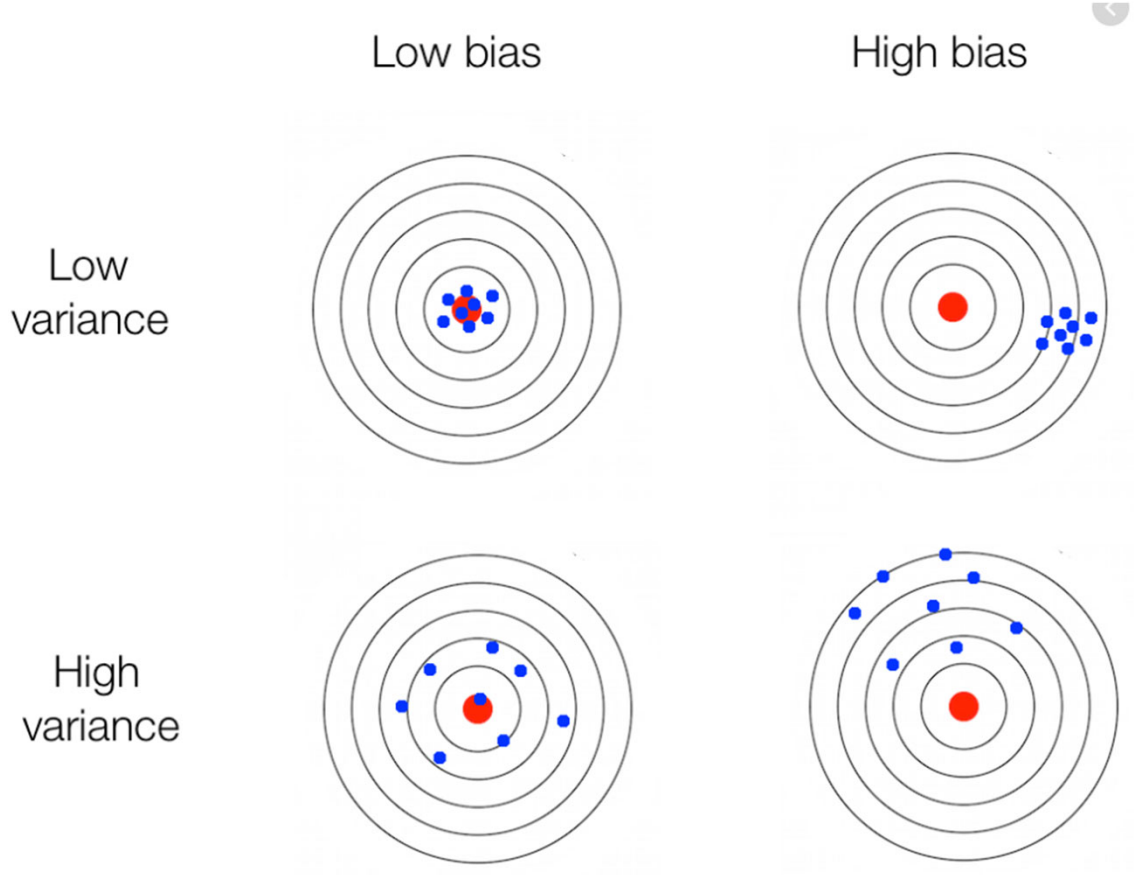
- How would one summary determine a better estimator than another summary?

# Review




- Loss Functions
  - Functions depending on $x_1,\ldots,x_n$ along variables. These variables are the unknown quantities corresponding to different choices for summaries.
  - We choose between the different summaries by the value of the loss function.
    - Low value is good
    - High value is bad

# Review

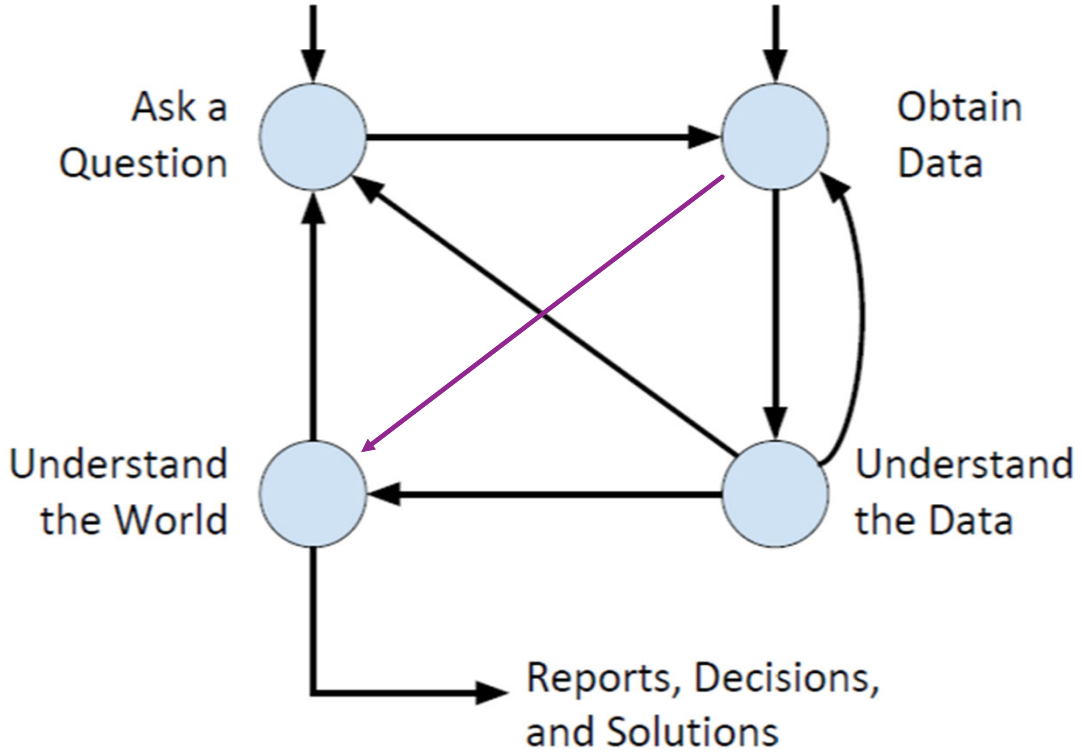



- Loss Functions
  - Find the values of the unknown quantities that minimizes the loss function gives a systematic way to choose between statistics
  - Loss functions should allow us to assess the
    - Bias
    - Variance

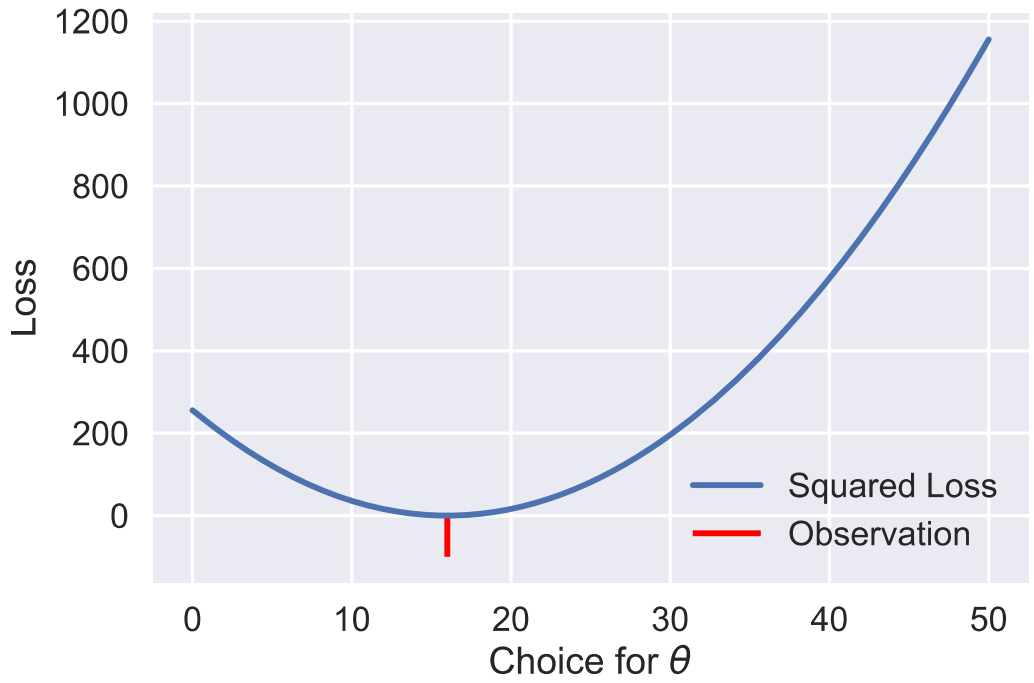
  of the choices

# Agenda

- Loss Functions
- Probability Distributions
  - Bernoulli
  - Binomial
- Random Variables
  - Expectation
  - Variance



# Loss Functions




- Mean Square Error has minimum value at the mean of the data
- The derivative tells us the rate of change of a function. We calculate the change in output divided by the change in input.
- If the derivative is zero, then mean square error has obtained it minimum value

# Loss Functions

$$L(\theta, x_1, \ldots, x_n) = \frac{1}{n}\sum(x_i - \theta)^2$$

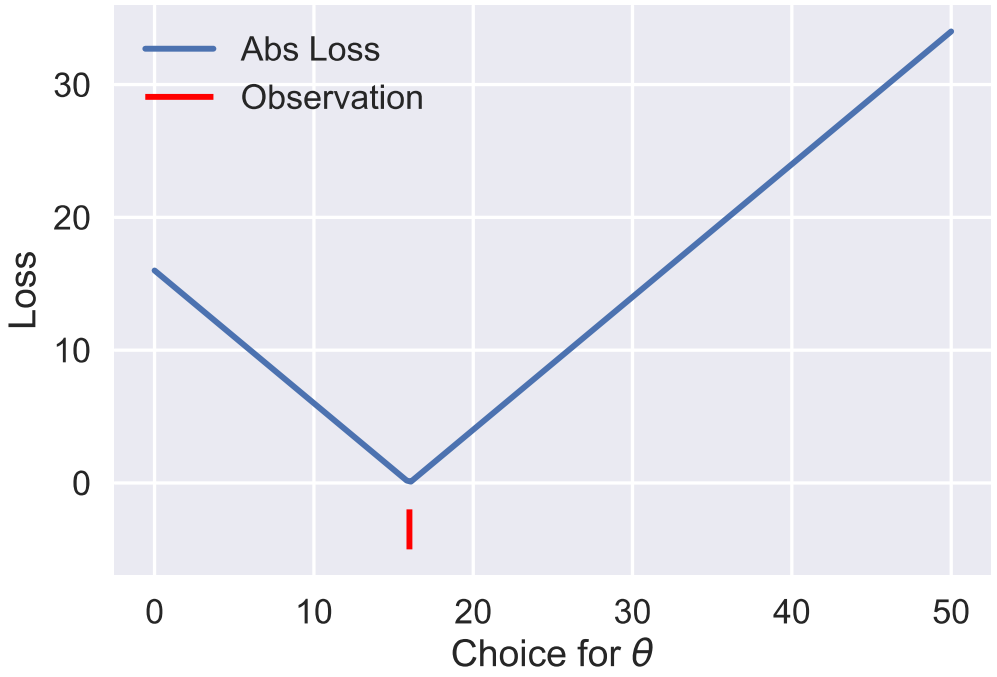
$$\frac{\partial}{\partial\theta}L(\theta, x_1, \ldots, x_n) = \frac{1}{n}\sum(2)(x_i - \theta)(-1)$$

$$= -\frac{2}{n}\left(\sum(x_i) - n\theta\right)$$

$$\sum(x_i) - n\hat{\theta} = 0$$

- Mean Square Error has minimum value at the mean of the data
- The derivative tells us the rate of change of a function. We calculate the change in output divided by the change in input.
- If the derivative is zero, then mean square error has obtained it minimum value

# Loss Functions




- Mean Absolute Error has minimum value at the median of the data
- The derivative of absolute value function gets tricky around the value 0 where it jumps from -1 to 1
- However we can split up the summation in the loss function to make the calculation easier

# Loss Functions

$$L(\theta, \mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}|x_i - \theta|$$

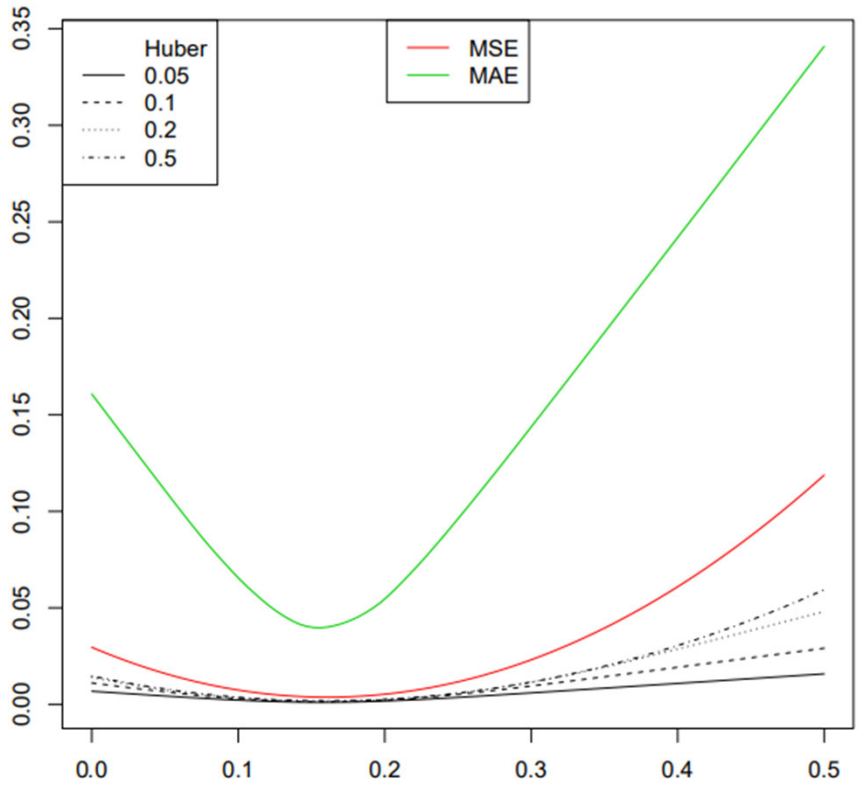
$$= \frac{1}{n}\left(\sum_{x_i<\theta}|x_i - \theta| + \sum_{x_i=\theta}|x_i - \theta| + \sum_{x_i>\theta}|x_i - \theta|\right)$$

$$\frac{1}{n}\left(\sum_{x_i<\theta}(-1) + \sum_{x_i=\theta}(0) + \sum_{x_i>\theta}(1)\right) = 0$$

$$\sum_{x_i<\theta}(1) = \sum_{x_i>\theta}(1)$$

- Mean Absolute Error has minimum value at the median of the data
- The derivative of absolute value function gets tricky around the value 0 where it jumps from -1 to 1
- However we can split up the summation in the loss function to make the calculation easier

# Loss Functions




- The mean square error has derivatives at all values of the function. Derivatives are helpful for finding minimum values.
- However, the mean square error has large output for large input. The function is not robust to outliers
- While the mean absolute error has a tricky derivative, the function does not have the same problem with outlier
- Huber Loss combines benefits of both loss functions

# Random Variables

- Represents a numeric value related to a random event.
- Random variables can be discrete (integer values) or continuous (any floating point number).
- We tend to use capital letters for the variables with P denoting probability.
  - For example, X could be the tip as a percentage of the bill
  - P(X = 2) is the probability that X has the value 2.
  - P(X < 10) is the probability that X is less than 10

# Random Variables

- Represents a numeric value related to a random event.
- Random variables can be discrete (integer values) or continuous (any floating point number).
- We tend to use capital letters for the variables with P denoting probability.
  - For example, X could be the tip as a percentage of the bill
  - P(X = 2) is the probability that X has the value 2.
  - P(X < 10) is the probability that X is less than 10

# Distributions

- The probability distribution of a random variable consists of possible values along with their frequency of occurrence
- We use stem-plots to visualize distributions

$$P(X = 1) = \frac{1}{6}$$
$$P(X = 2) = \frac{1}{6}$$
$$\dots$$
$$P(X = 6) = \frac{1}{6}$$

| X | P(X) |
|---|---|
| 1 | 1/6 |
| 2 | 1/6 |
| ... | ... |
| 6 | 1/6 |

# Bernoulli Distribution

- Random variable that takes the value 0 or 1 is called a Bernoulli random variable
- We need to specify the probability that the value is 1

$$X \sim \text{Bern}(p)$$

$$P(X = 1) = p \quad P(X = 0) = 1 - p$$

- Remember by the complement rule that the probability of 0 is 1-0

# Binomial Distribution

- Sum of Bernoulli random variables is a Binomial random variable
- Here each random variable is independent
- We need to specify the number of Bernoulli random variables and the probability that each is p

$$X \sim \text{Binom}(n, p)$$

$$P(X = k) = P(k) = \binom{n}{k} p^k (1-p)^{n-k} = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$$

# Binomial Distribution

$$X \sim \text{Binom}(5, 1/3)$$

$$P(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

$$P(3) = \binom{5}{3} \left(\frac{1}{3}\right)^3 \left(\frac{2}{3}\right)^2 = 0.165$$

- We need to specify the number of Bernoulli random variables and the probability that each is p

$$X \sim \text{Binom}(n, p)$$

$$P(X = k) = P(k) = \binom{n}{k} p^k (1-p)^{n-k} = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$$

# Expectation

- Expectation is weighted of the values in the distribution.
- We can think of expectation like the mean expect different values have different weights depending on their frequency.

$$X \sim \text{Bern}(p) \qquad E(X) = \sum_{x \in \mathbb{X}} x \cdot P(X = x)$$
$$= (1)(p) + (0)(1-p)$$
$$= p$$

$$E(X) = \sum_{x \in \mathbb{X}} x \cdot P(X = x)$$

- For example, with Bernoulli random variables we obtain the probability of value 1.

# Variance

- Variance is expectation of square difference between random variable and expectation
- We can think of variance like the mean square error. It measures the spread of values of the random variable away from the expectation

$$Var(X) = E\left((X - E(X))^2\right)$$

$$X \sim \text{Bern}(p)$$

$$E(X) = p$$

$$E(X^2) = (1^2)(p) - (0^2)(1-p) = p$$

$$Var(X) = p - p^2 = p(1-p)$$

- For example, with Bernoulli random variables we obtain the probability of value 1 times the probability of value 0

# Variance

- Variance is expectation of square difference between random variable and expectation
- We can think of variance like the mean square error. It measures the spread of values of the random variable away from the expectation

$$Var(X) = E(X^2) - E(X)^2$$

$$X \sim \text{Bern}(p)$$

$$E(X) = p$$

$$E(X^2) = (1^2)(p) - (0^2)(1-p) = p$$

$$Var(X) = p - p^2 = p(1-p)$$

- For example, with Bernoulli random variables we obtain the probability of value 1 times the probability of value 0

# Summary

- Loss Functions
- Probability Distributions
  - Bernoulli
  - Binomial
- Random Variables
  - Expectation
  - Variance

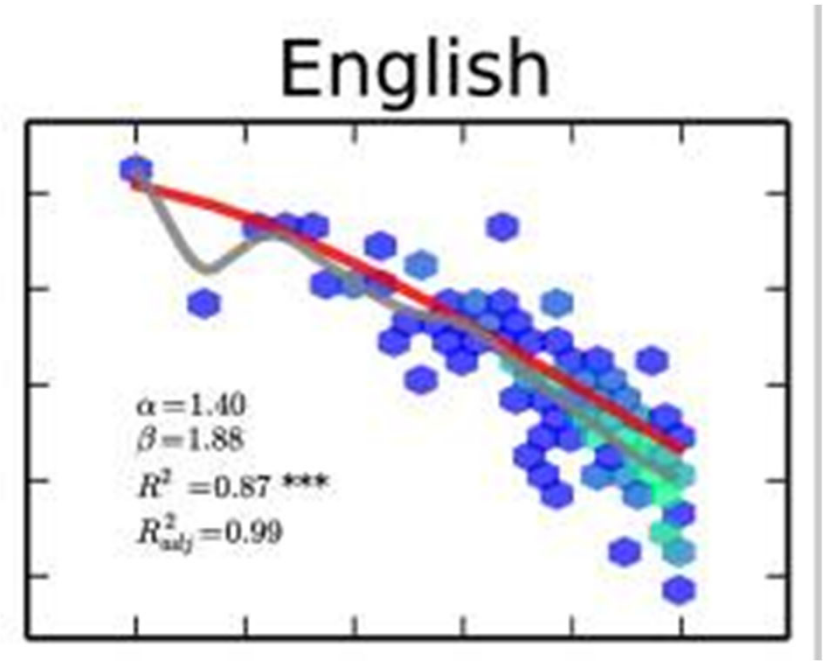
**Goals**

- How are MSE and MAE different?
- Why would bias and variance help us with estimation of parameters?

# Questions

- Questions on Piazza?
  - Please provide your feedback along with questions
- Question for You!

Where else can you find the Zipf Rule?



# Questions

- Questions on Piazza?
  - Please provide your feedback along with questions
- Question for You!

Where else can you find the Zipf Rule?