NYU Center for Data Science

# DS-UA 112
# Introduction to Data Science

## Week 10: Lecture 1
## Testing Hypotheses

How can we validate the assumptions in a model with data?

# DS-UA 112
# Introduction to Data Science

## Week 10: Lecture 1
## Testing Hypotheses

*Adapted from Adhikari, DeNero, Wagner, Milner*

# Announcements

- Please check Week 10 agenda on NYU Classes
  - Homework 3/4
  - Lab 6
- Please check the Calendar linked to NYU Classes

# Review

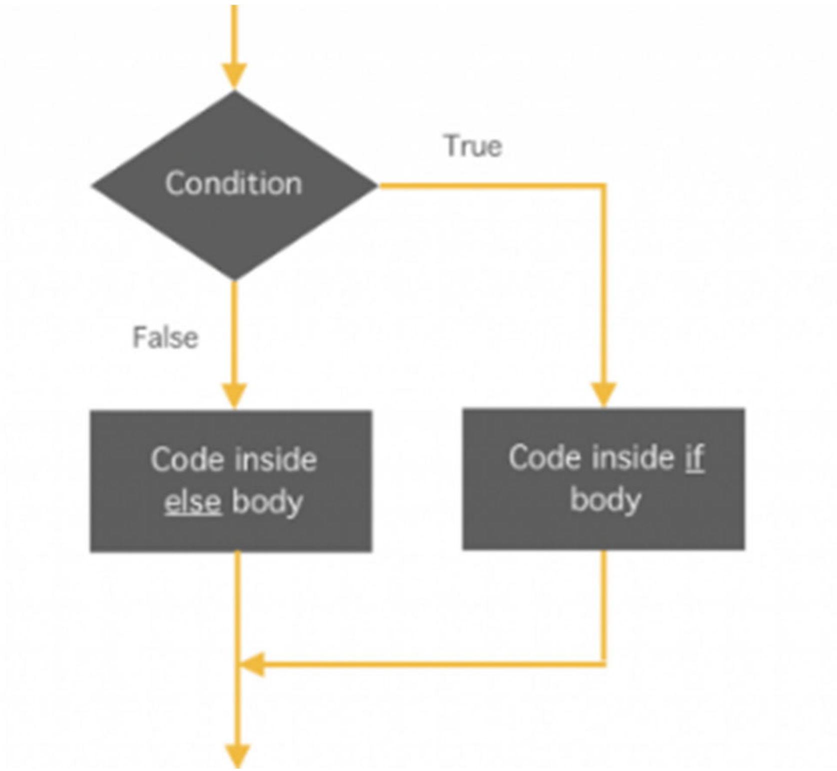
- Simulation
  - Conditional Statements
  - Loops
  - Random Selection
- Distributions
  - Probability Distribution
  - Empirical Distribution
  - Parameters

References

- Simulation:
  - Chapters 9.1,9.2,9.3, 10.2,10.3

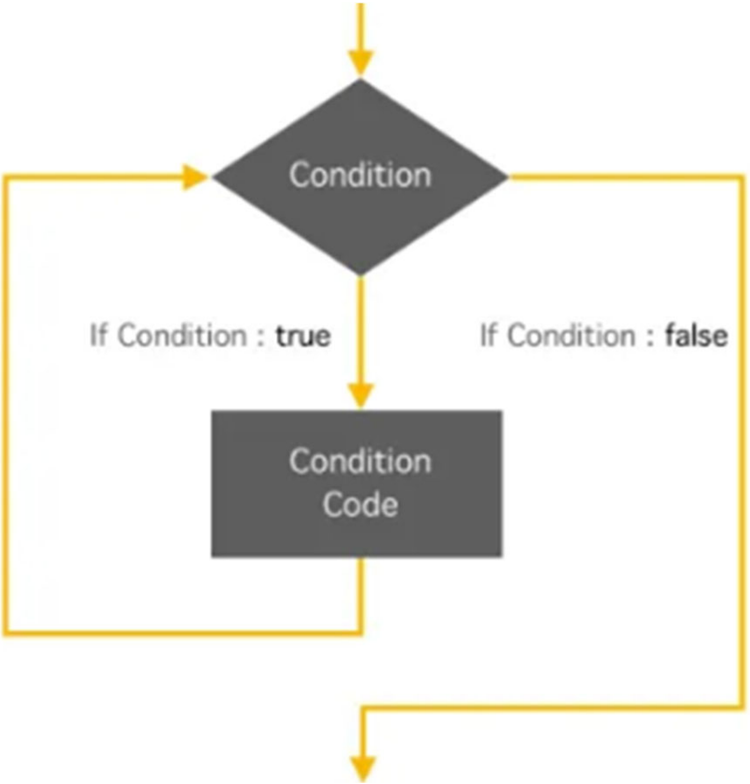
# Review

- Conditional Statements
  - We use a special computational data type called Boolean for True and False in Python
  - Think of True/False as
    - Yes/No
    - 1/0
    - Not Empty/Empty…



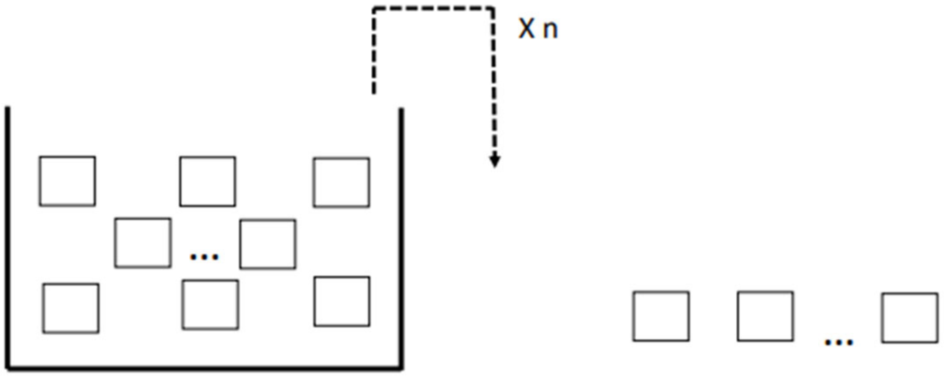
# Review

- Loops
  - We can repeatedly run a block of code in Python using a loop
  - for loop
    - Runs the block of code for specified number of iterations
  - while loop
    - Combines conditional statement and for loops
    - Runs block of code while the logical expression is True



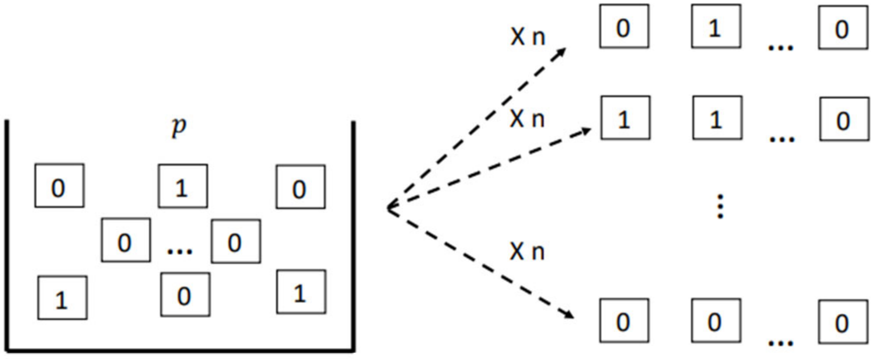
# Review

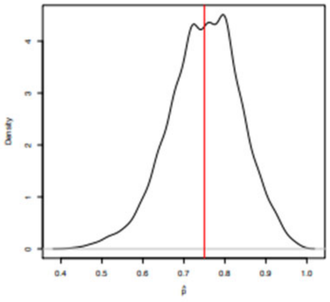
- Random Selection
  - We have random and deterministic approaches to gathering observations
  - Simple Random Sample (SRS) means randomly picking from the population with equal probability for each observation
  - With replacement means we put the observations back. Without replacement means we don't put the observations back.



# Review

- Distributions
  - Random quantity with different possible values
  - Probability Distribution
    - Chance of any possible values in population
  - Empirical Distribution
    - Observed values in a random sample
    - We compute chance of value in the random sample by proportion of occurrences

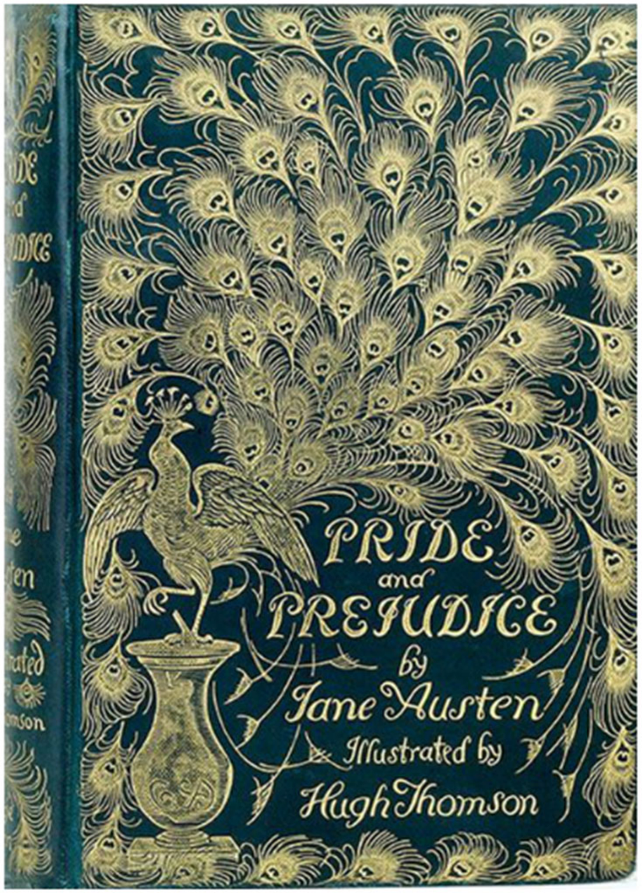




# Exercise

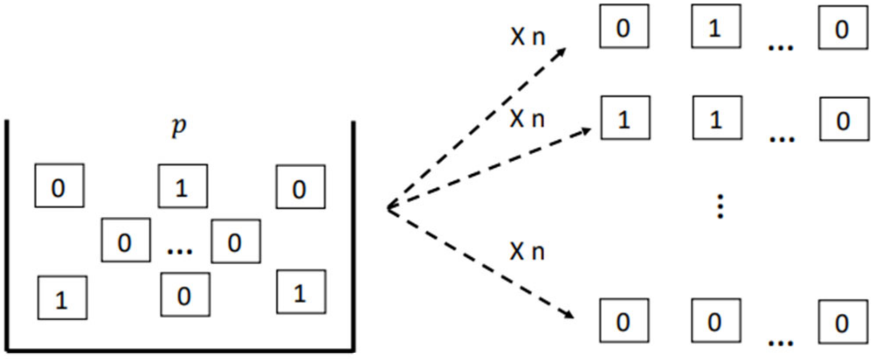
Pride and Prejudice by Jane Austen

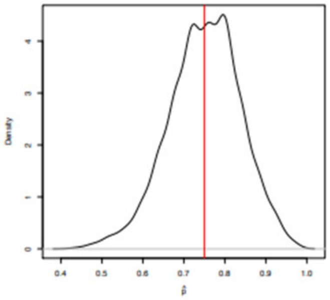
- Can we make guesses about the length of words in the novel through sampling instead of counting?



# Review

- Simulation
  - Often we try to determine numerical attributes of the probability distribution nicknamed parameters
  - If we compute a statistic to estimate the parameter across many random samples, then we expect these estimates to converge on average to the parameter





# Agenda

- Testing a Hypothesis
  - Null hypothesis
  - Alternative hypothesis
- Comparing Distributions
  - Statistics for goodness of fit

References

- Hypothesis Testing
  - Chapters 11.1,11.2

# Goodness of Fit

- Suppose we have a sample that might come from randomly sampling a population.
- Assuming we have a guess about the probability distribution for the population, then we can simulate random sampling

- Step 1: Hypotheses
- Step 2: Statistic
- Step 3: Probability Distribution

# Goodness of Fit

- If we can associate a statistic to the distributions, then we can compare the statistic of the sample to the empirical distribution of the statistics simulated from the population

- Step 1: Hypotheses
- Step 2: Statistic
- Step 3: Probability Distribution

# Goodness of Fit

- Step 1
  - Test chooses between two possible possibilities
  - Null hypothesis assumes the model captures the process behind the population generating the samples
  - Alternative hypothesis assumes the model captures the process behind the population generating the samples

- Step 1: Hypotheses
- Step 2: Statistic
- Step 3: Probability Distribution

# Goodness of Fit

- Step 2
  - Compute a statistic that helps us to choose between hypotheses
  - Statistic should estimate the parameters in the population
- Step 3
  - Under the null hypothesis we simulate random sample from the population to generate an empirical distribution of the statistic

- Step 1: Hypotheses
- Step 2: Statistic
- Step 3: Probability Distribution

# Goodness of Fit

- Step 2
  - Compute a statistic that helps us to choose between hypotheses
  - Statistic should estimate the parameters in the population
- Step 3
  - Under the null hypothesis we simulate random sample from the population to generate an empirical distribution of the statistic

Accept:

If the observed statistic is consistent with the empirical distribution

# Goodness of Fit

- Step 2
  - Compute a statistic that helps us to choose between hypotheses
  - Statistic should estimate the parameters in the population
- Step 3
  - Under the null hypothesis we simulate random sample from the population to generate an empirical distribution of the statistic

Reject:

If the observed statistic is not consistent with the empirical distribution

# Example

Gregor Mendel

- Botanist studying the genetics of pea plants.
- Validated assumptions in model for expression of features like color

# Total Variation Distance

- Suppose we have two distributions whose values correspond to categories. In other words, the statistical data type of the values is qualitative.
- How can we determine a statistic to compare them?

- Step 1: Differences
- Step 2: Absolute Value
- Step 3: Divide by Two

# Total Variation Distance

- Step 1
  - Take the difference between the proportions corresponding to each category
- Step 2
  - Apply absolute value transformation to obtain positive numbers
- Step 3
  - Add the transformed numbers. Divide the summation by 2.

- Step 1: Differences
- Step 2: Absolute Value
- Step 3: Summation

# Example

Juries

- Courts need to have jurors for trials
  - Eligible members of community
  - Chosen by identification
  - Selected from a panel to sit on jury

# Summary

- Testing a Hypothesis
  - Null hypothesis
  - Alternative hypothesis
- Comparing Distributions
  - Statistics for goodness of fit

Goals

- Compare a sample and simulated samples to accept / reject hypotheses
- Compute the total variation distance as statistic to compare two distributions with multiple categories