Lecture 12
Statistics

Ryan McWay†

† *Applied Economics,*
University of Minnesota

Mathematics Review Course, Summer 2023
University of Minnesota
August 22nd, 2023

Last Lecture Review

- Probability:
  - Probability Limits
  - Independence
  - Law of Total Probability
  - Conditional Probability
  - Cumulative Distribution Function
  - Probability Distribution Function
  - Joint & Marginal Distributions
  - Gaussian (Normal) Distribution
  - Bayes Rules
  - Moments of a Distribution
  - Covariance
  - Correlation

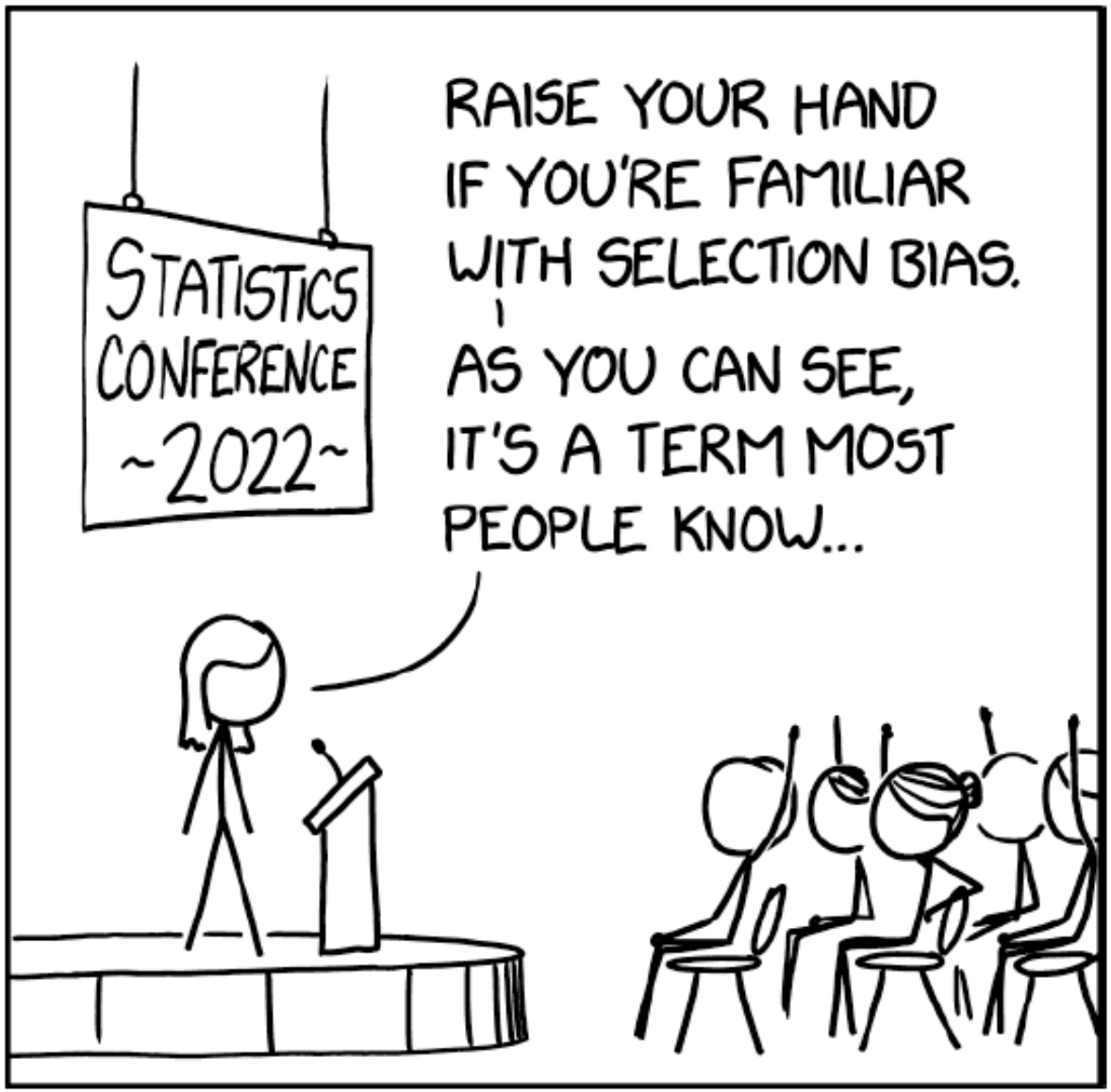
Review Assignment

1. Problem Set 11 solutions are available on Github.
2. Any issues or problems **You** would like to discuss?

Daily Icebreaker

- Attendance via prompt:
  - Name
  - Program and track
  - Daily icebreaker subject...

Selection Bias...



Statistics

Motivation

- General background
  - This is how we map real world data into the world of probabilities.
- Why do economists' care?
  - This is one of the primary tools of economists.
  - We use statistics to determine what information we choose to believe is 'true'.
- Application in this career
  - Throughout applied research.
  - A core component of econometrics.

Motivation

- General background
  - This is how we map real world data into the world of probabilities.
- Why do economists' care?
  - This is one of the primary tools of economists.
  - We use statistics to determine what information we choose to believe is 'true'.
- Application in this career
  - Throughout applied research.
  - A core component of econometrics.

Motivation

- General background
  - This is how we map real world data into the world of probabilities.
- Why do economists' care?
  - This is one of the primary tools of economists.
  - We use statistics to determine what information we choose to believe is 'true'.
- Application in this career
  - Throughout applied research.
  - A core component of econometrics.

Overview

1. Population, Parameters, and Distributions
2. Random Variables
3. Discrete & Continuous Variables
4. Law of Iterated Expectations
5. Sampling
6. Data Types
7. Data Structure
8. Randomization
9. Estimate, Estimator, & Estimand
10. Parametric vs. Non-parametric
11. Expectations
12. Conditional Expectation Function
13. Law of Large Numbers
14. Central Limit Theorem
15. Continuous Mapping Theorem
16. Delta Method
17. Standardizing Units
18. Hypothesis Testing
19. Causation
20. Prediction
21. Descriptive Analysis
22. Inferential Analysis

1. Population, Parameters, and Distributions

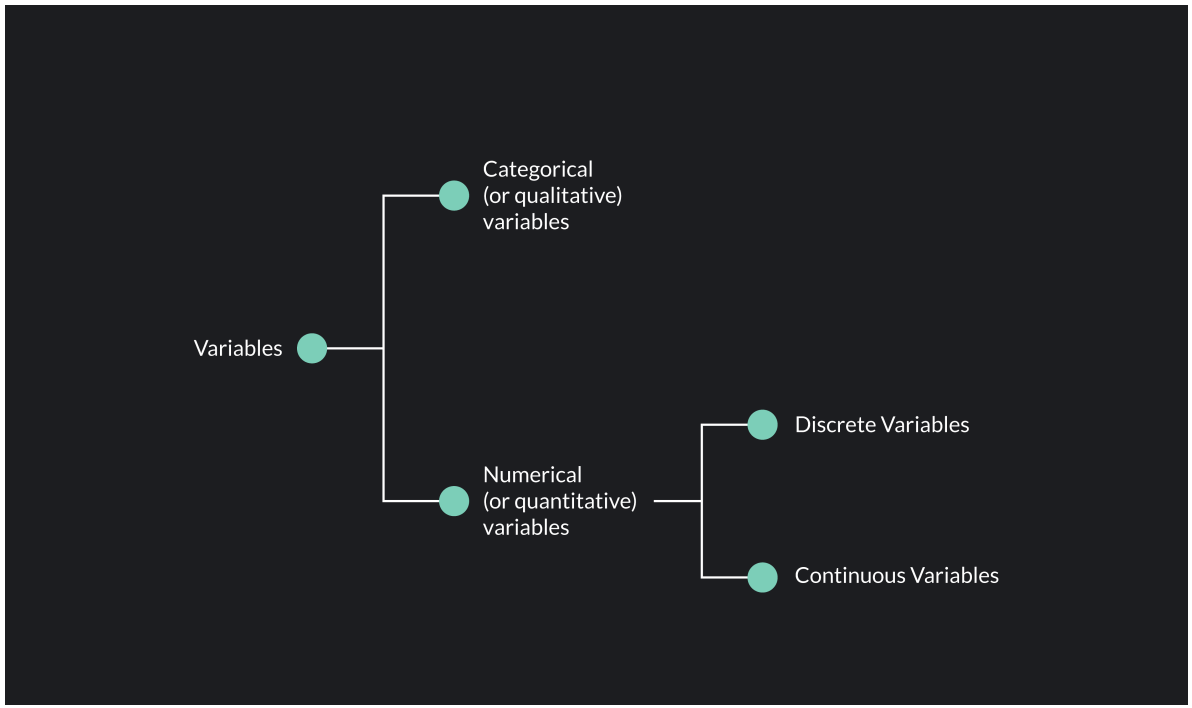
- Population
  - A well-defined group of subjects of interest.
  - Theoretically, an infinite sized group of all members fitting the definition.
- Parameters
  - A constant, unknown value describing the true relationship of variables in the population.
  - Parameters are defined, but never known. You estimate to approximate their true value.
- Distributions
  - The frequency of values (outcomes) about the range of the sample space for a population.

2. Random Variables

- Experiment: Procedure that can be infinitely repeated and has well-defined set of outcomes. But, the actual outcome may be of unknown certainty.
- Random variable: $x \in X$ is a numerical value outcome in a set of possible outcomes where the outcome is determined by an experiment.
- Binary (Bernoulli) random variables: Takes values 0 or 1 such that $Pr(X = 1) = \theta : 0 \leq \theta \leq 1$.

3. Discrete & Continuous Variables

- Discrete Variables: Random variable that has a finite or countably infinite domain of values.
- Continuous Variables: Random variable with infinite domain of values.
  - Can take on any real value with $Pr(X) \geq 0$.



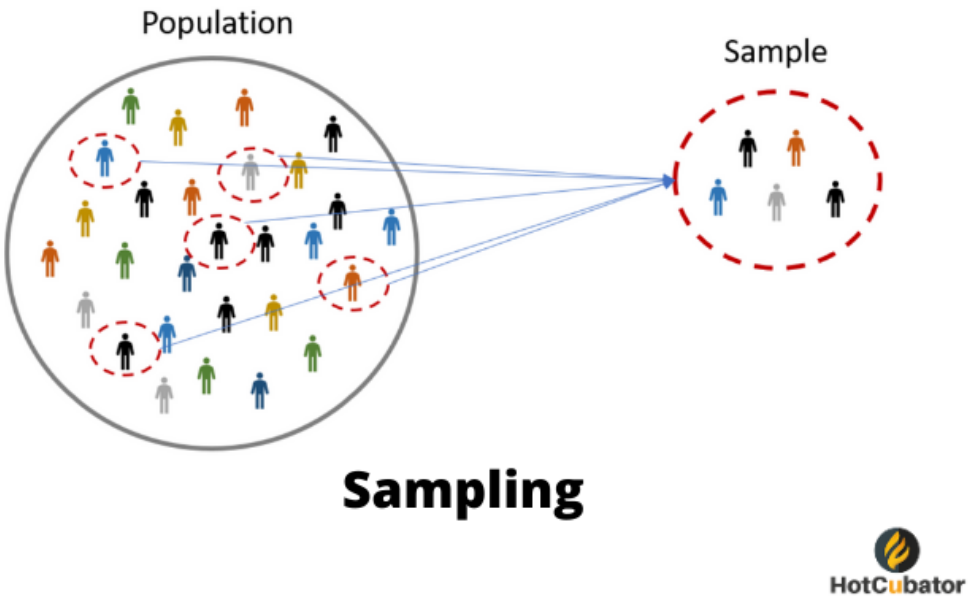
5. Law of Iterated Expectations

- The expected value of the conditional expected value of X given Y is the same as the expected value of X.
- The average of the group averages is the average of the whole distribution.
- E.g., you can recover the population average from aggregated (group) averages.

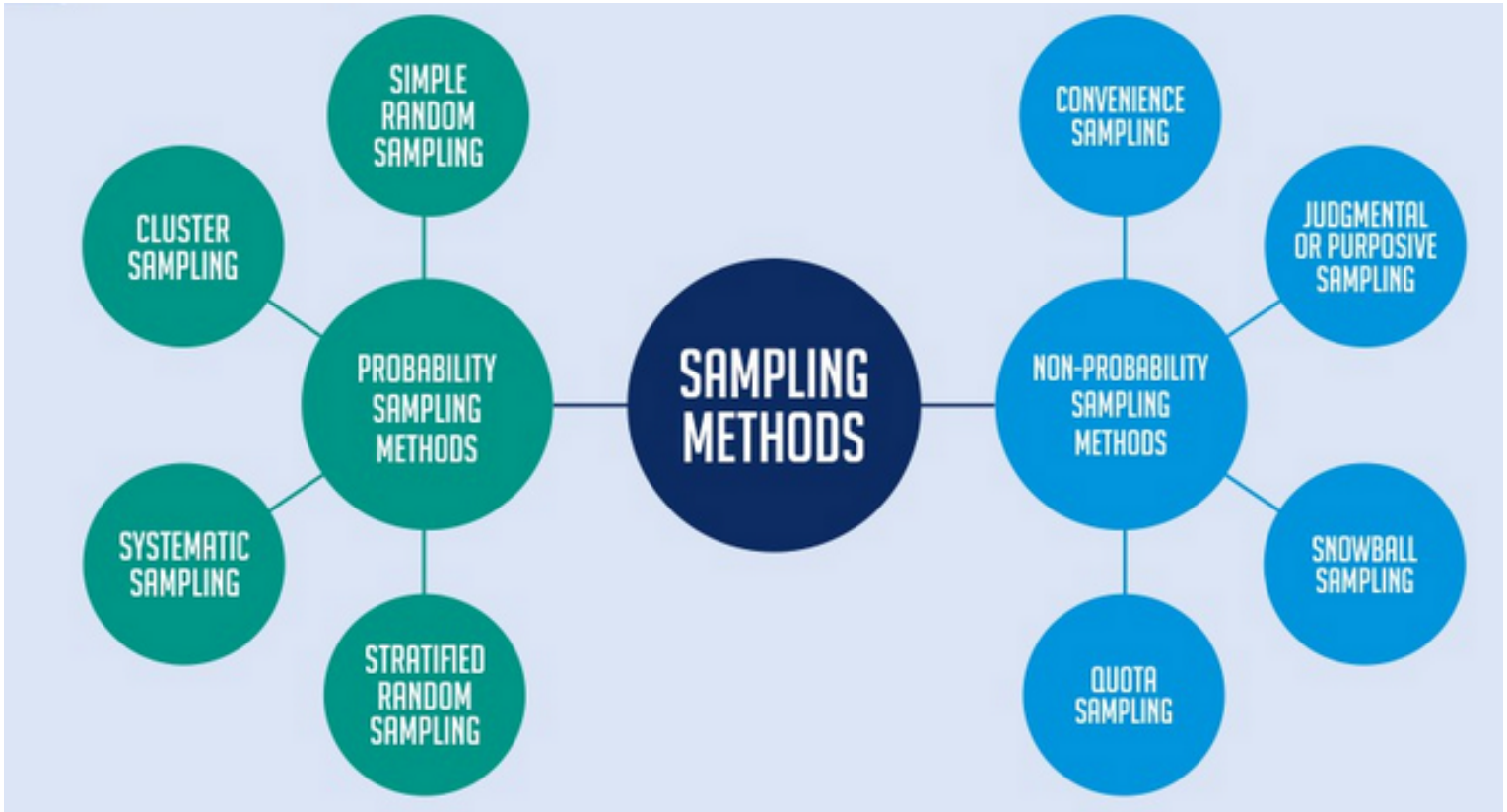
$$\mathbb{E}(X) = \mathbb{E}(\mathbb{E}(X|Y))$$

6. Sampling

- Sampling pulls units from the population distribution as the data used to estimate population parameters.
- Random Sampling
  - A sample of independently drawn units.
  - Independence means that no individual is more likely to be sampled than another from the population.



6. Sampling

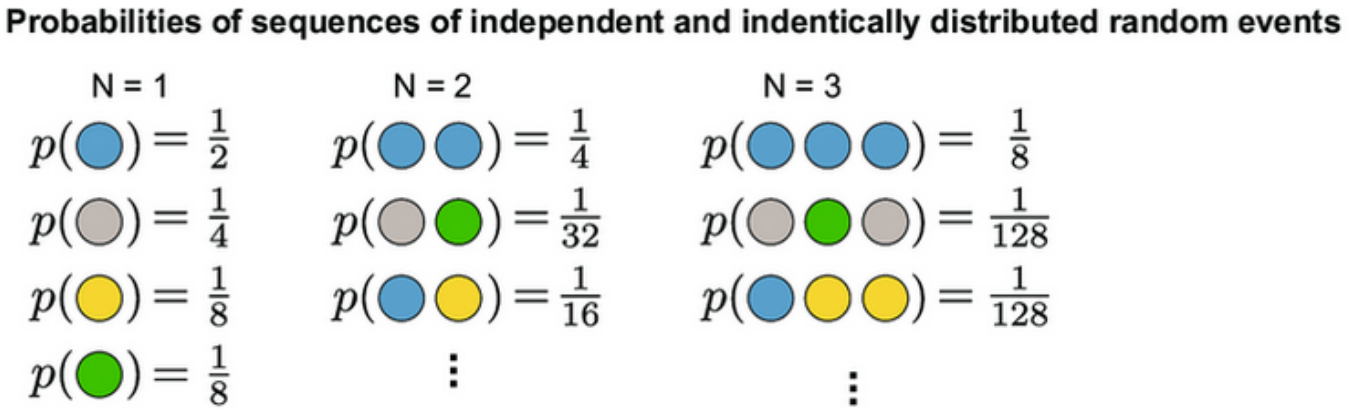

SAMPLING METHODS
PROBABILITY SAMPLING METHODS
SIMPLE RANDOM SAMPLING
CLUSTER SAMPLING
SYSTEMATIC SAMPLING
STRATIFIED RANDOM SAMPLING
NON-PROBABILITY SAMPLING METHODS
CONVENIENCE SAMPLING
JUDGMENTAL OR PURPOSIVE SAMPLING
SNOWBALL SAMPLING
QUOTA SAMPLING

7. Data Types

- The data you gather may come in many forms – qualitative and quantitative.
- Most of the data you will work with are samples formatted into tabular arrays (viz., matrix).
- Types:
  - Float: Fractions
  - Boolean: True/False
  - Integer: ...
  - Categorical: Cardinally ranked integers
  - Character: Letter or special character
  - String: Combination of characters
  - Date: Formatted string or integer
  - Null: Empty Set
  - N/A: Missing value

8. Data Structure

- The data generating process (DGP)
  - The way that the data is constructed in the real world.
  - Typically unknown – often an assumption.
- Independently Identically Distributed (IID)
  - Each random variable has the same probability distribution as the others and are mutually independent of one another.
  - The sample is 'representative' of the population.
  - Identically distributed: Distribution does not fluctuate by sample.
  - Independent: Sample items (outcomes) are independent events



8. Data Structure

- Cross-sectional
  - A sample of individuals (units) at a given point in time (period) with many variables (characteristics).
- Time Series
  - Potentially several units on a single variable over many periods.
- Pool Cross-section
  - A series of cross-sections with some overlapping units at different time periods.
- Panel Datasets
  - A time series of many periods for many variables across many units.

9. Randomization

- A random process is a random sequence that does not follow a deterministic process.
- Random Sample: Individuals have a known probability of sampling from the population.
- Ensure the data is I.I.D. and that the estimates are representative for the parameters in the population.

10. Estimate, Estimator, & Estimand

- Estimate: Approximation of the population parameter.
- Estimator: The function or algorithm doing the estimation.
- Estimand: The parameter in the population you aim to estimate.

ESTIMAND
What you seek

E.g. The true difference in Y due to exposure

ESTIMATOR
How you will get there

Method

1. Preheat your oven to 190°C /170°F / Gas Mark 5. Grease and line the base of 2 cake tins, one 8 inch/20cm and one 6 inch/15cm
2. Cream together the butter and caster sugar until light and fluffy.
3. Add the eggs one at a time with a spoonful of flour and blend in well.
4. Sift in the flour and baking powder and gently fold in. Finally add the milk and mix until you have a smooth batter.
5. Pour 1/3 of the batter into the small tin and 2/3 into the large tin.
6. Bake on the same shelf in the preheated oven, the smaller tin at the front.
7. Check the smaller cake after 20 minutes. When it is cooked remove from the oven, leaving the larger one still baking. The large cake should be done by 30 minutes.
8. Leave the cakes for 5 minutes in the tins, then turn out onto a rack to cool completely.
9. To make the icing, beat together the butter and icing sugar, add the vanilla and then the milk. Whisk the icing hard using an electric stand mixer if you can. Whisk it for 5 minutes and it will become really pale and light.

E.g. Your regression model

ESTIMATE
What you get

E.g. the estimated difference in Y from model coefficient

11. Parametric vs. Non-parametric

- Parametric: Imposes some distribution on the underlying population (e.g., the distribution of the parameters)
- Non-parametric: Is agnostic to how the distribution of the population might look
- Parametric estimators rely on asymptotic properties of the population (e.g., Assume to know the underlying distribution in large samples).
- Non-parametric estimators are driven by distributional properties of the sample.
- Can often use non-parametric estimators when the moments of the distribution do not converge to a known distribution.

12. Expectations

- Expected Values (Mean): A weighted average of all possible values of X, with weights equal to the probability of each outcome occurring.
- Discrete:

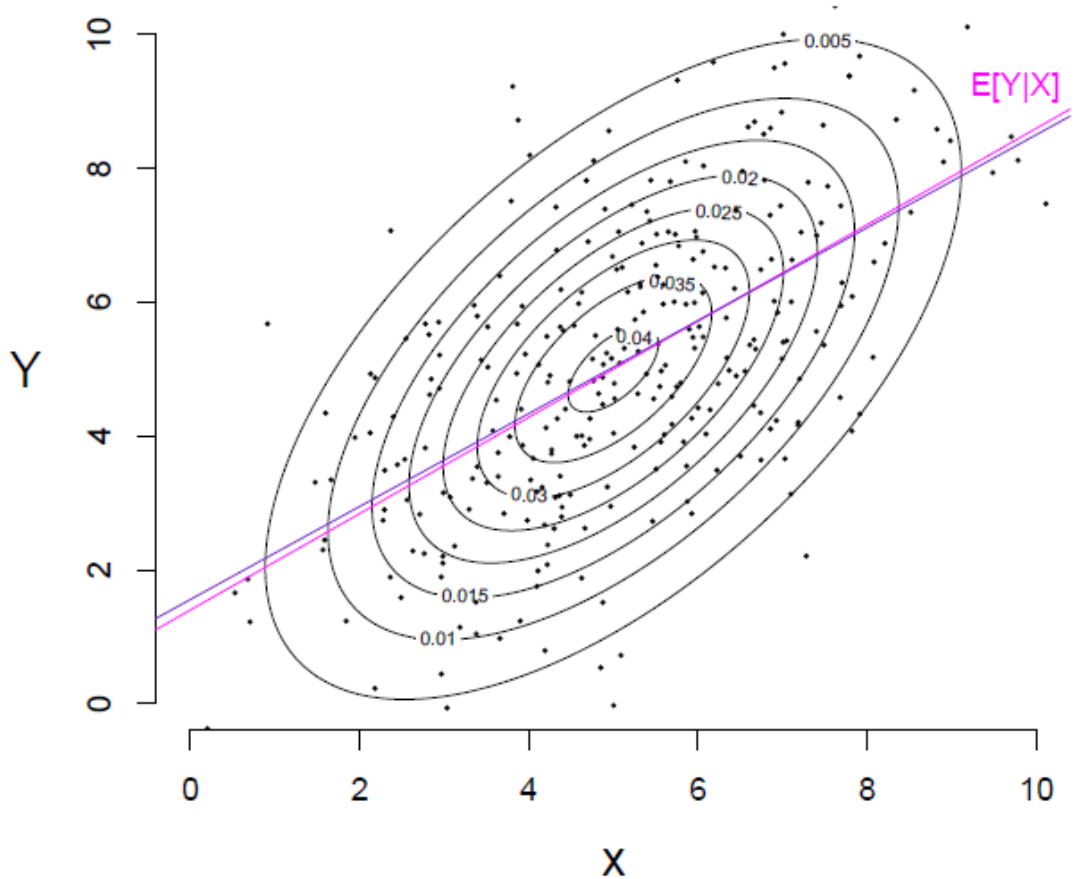
$$\mathbb{E}(X) = \sum_{j}^{k} x_j f(x_j) = \mu_x$$

- Continuous:

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f(x) dx$$

13. Conditional Expectation Function

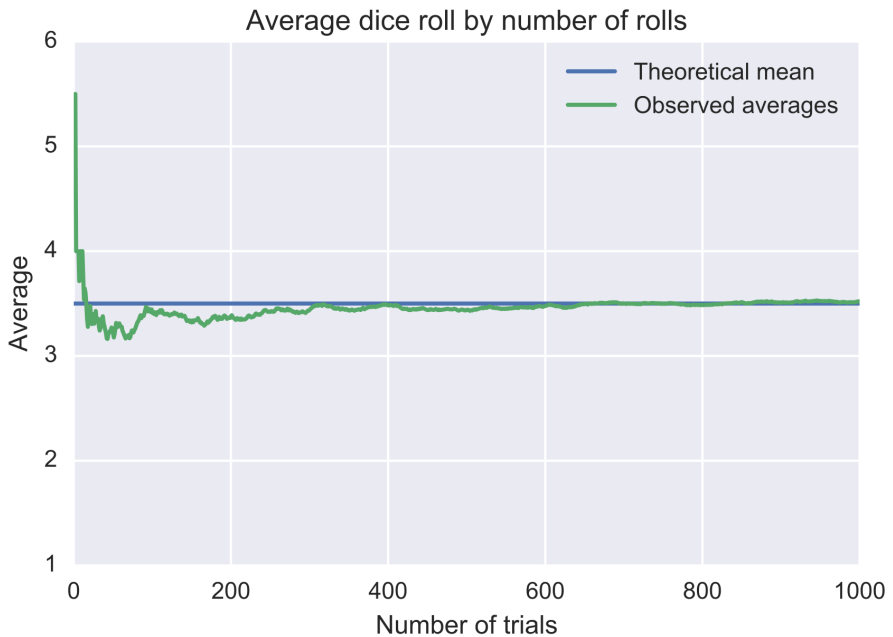
$$\mathbb{E}[Y|X = x] = \int_{-\infty}^{\infty} y f_{Y|X}(y|x) dy$$



14. Law of Large Numbers

- The expected value of the distribution converges to the population parameter as the sample becomes larger.
- If X_i are I.I.D. and $\mathbb{E}[X] < \infty$, then as $n \to \infty$:

$$\bar{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p} \mathbb{E}[X]$$

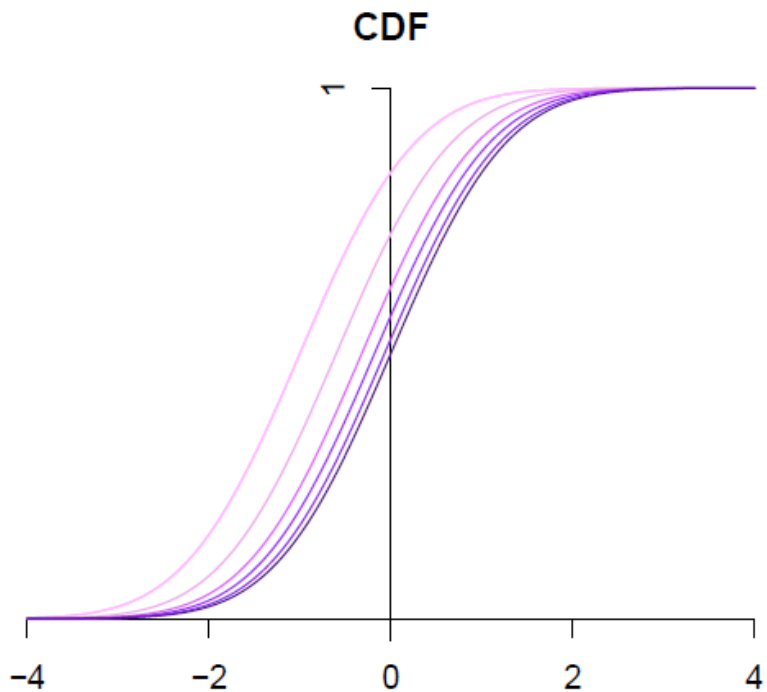


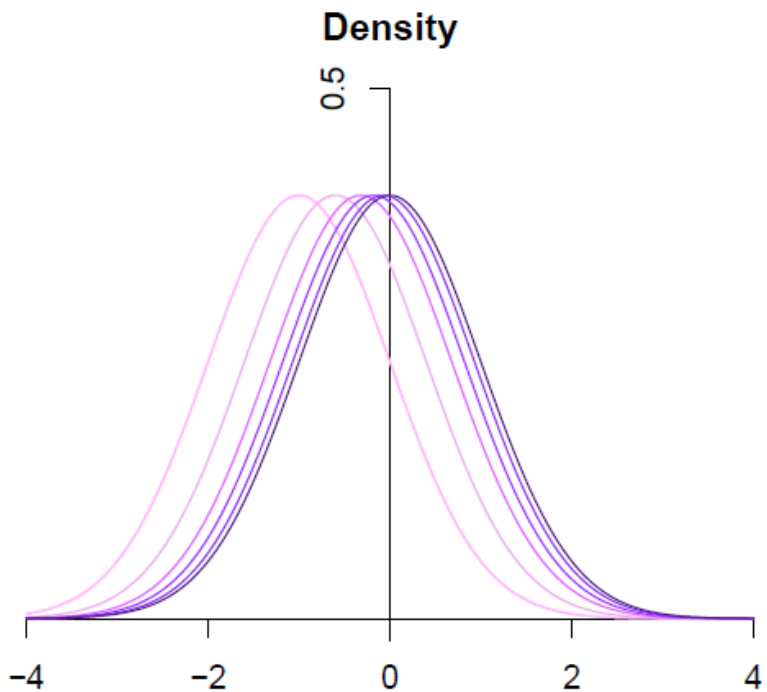
15. Central Limit Theorem

- The distribution of the sample approximates a normal distribution as the sample becomes larger – **regardless of the population's underlying distribution**.
- Let $X_i \in \mathbb{R}^k$ be I.I.D. with $\mathbb{E}[||X_i||^2] < \infty$.
- Define $\mu = \mathbb{E}[X]$.
- Define $\Sigma = \mathbb{E}[(X - \mu)(X - \mu)^T]$
- Then, as $n \to \infty$:

$$\sqrt{n}(\bar{X} - \mu) \xrightarrow{d} N(0, \Sigma)$$

15. Central Limit Theorem





16. Continuous Mapping Theorem

- Idea: $g(Z_n) = g(plim Z_n)$
- So: $\hat{\beta} \xrightarrow{p} \beta$
- If $Z_n \xrightarrow{p} c$ as $n \to \infty$, and $h(\cdot)$ is continuous at c, then:

$$h(Z_n) \xrightarrow{p} h(c) \text{ as } n \to \infty$$

17. Delta Method

- Variance will be approximately normal.
- Let $\theta \in \mathbb{R}^k$ and $h : \mathbb{R}^k \to \mathbb{R}$ be C^1 in the neighborhood of θ.
- Let $h = \nabla h(\theta)$.
- If $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \eta$, then as $n \to \infty$,

$$\sqrt{n}(h(\hat{\theta}) - h(\theta)) \xrightarrow{d} h^T \cdot \eta$$

- If $\eta \sim N(0, V)$ (by CLT), then

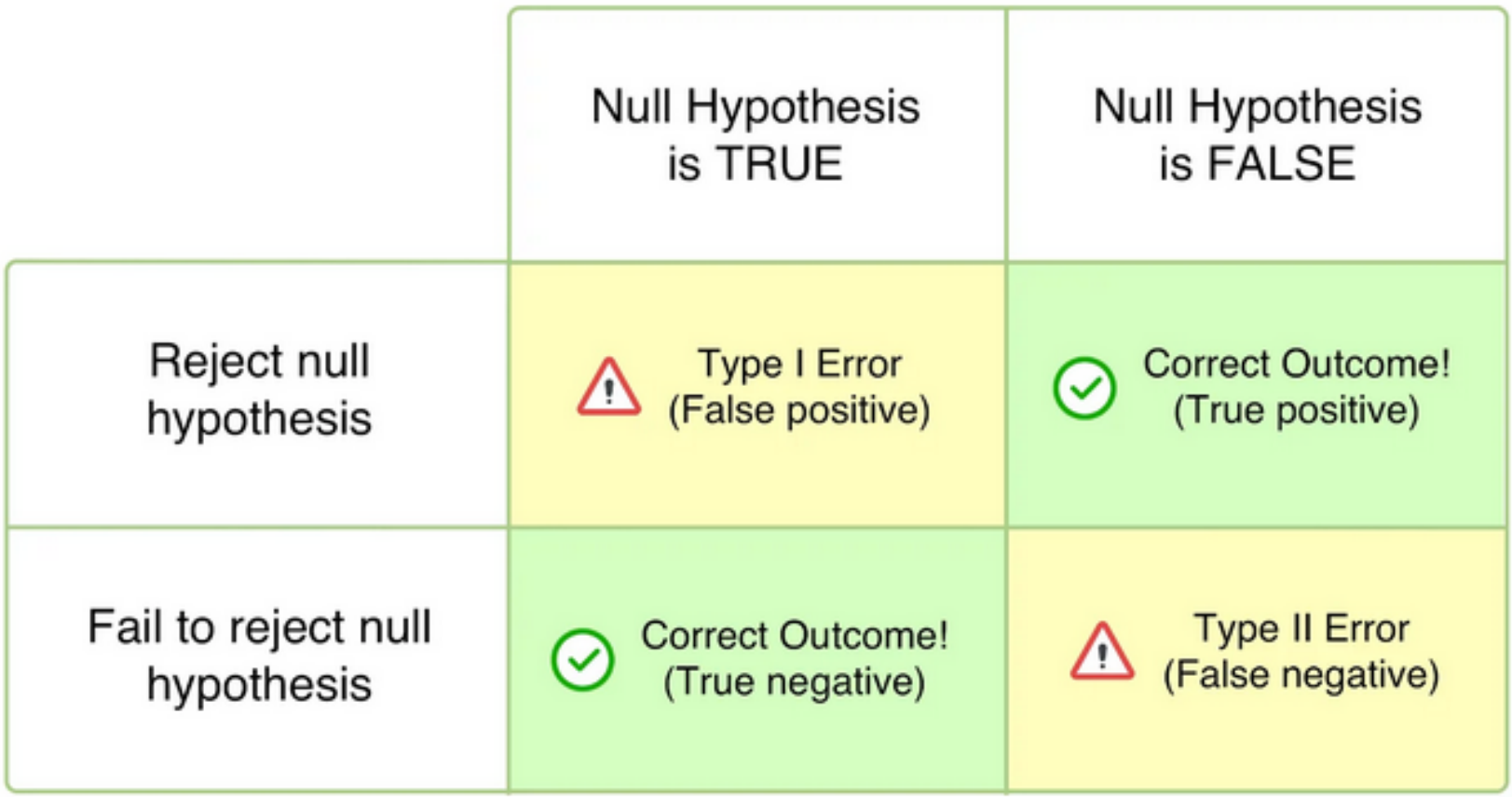
$$\sqrt{n}(h(\hat{\theta}) - h(\theta)) \xrightarrow{d} N(0, h^T V h)$$

18. Standardizing Units

- Commonly referred to as 'taking a z-score'.
- De-meaning the random variable and weighting it by the standard error to create a new random variable measured in standard deviations from the mean.
- When the units are standard deviations, it is now easy to compare many types of outcomes with various ranges and distributions.
- $\mathbb{E}[Z] = 0$
- $Var(Z) = 1$

$$Z = \frac{x - \mu_x}{\sigma_x}$$

19. Hypothesis Testing



19. Hypothesis Testing

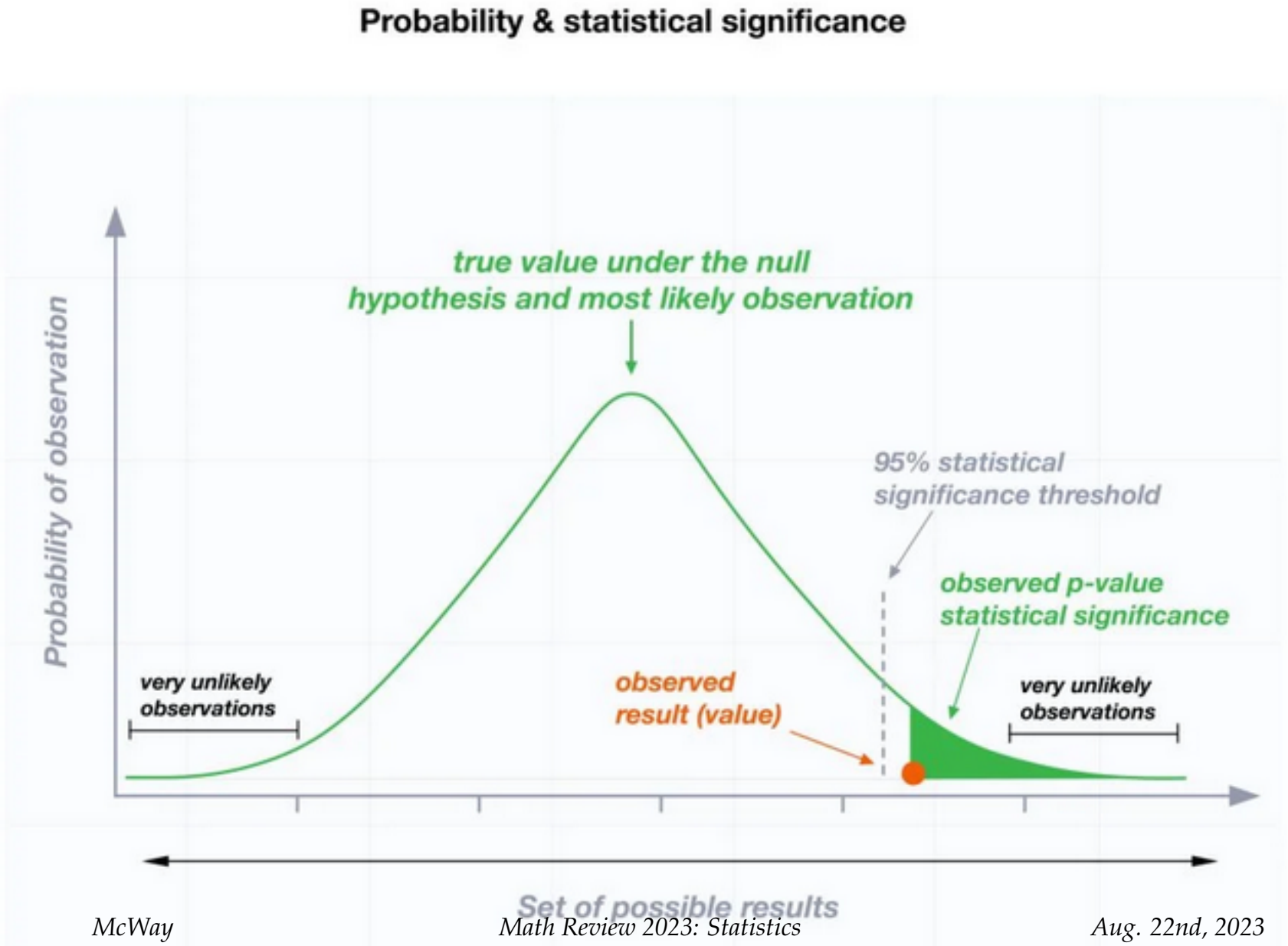
- Student's T-Test: Creates a measure of the distance a standardized value is from the mean weighted by the size of the standard error.

$$t = \frac{Z}{s} = \frac{\bar{X} - \mu}{\hat{\sigma}/\sqrt{n}}$$

- P-Value: The probability of obtaining the outcome as compared to the null hypothesis of a given distribution (i.e., standard normal).

$$p = Pr(T \geq t|H_0)$$

19. Hypothesis Testing



20. Causation

21. Prediction

22. Descriptive Analysis

23. Inferential Analysis

Practice: Statistics

1.

Review: Statistics

1. Population, Parameters, and Distributions
2. Random Variables
3. Discrete & Continuous Variables
4. Law of Iterated Expectations
5. Sampling
6. Data Types
7. Data Structure
8. Randomization
9. Estimate, Estimator, & Estimand
10. Parametric vs. Non-parametric
11. Expectations
12. Conditional Expectation Function
13. Law of Large Numbers
14. Central Limit Theorem
15. Continuous Mapping Theorem
16. Delta Method
17. Standardizing Units
18. Hypothesis Testing
19. Causation
20. Prediction
21. Descriptive Analysis
22. Inferential Analysis

Assignment

- Readings on Dynamic Programming before Lecture 13:
  -
- Assignment:
  - Problem Set 12 (PS12)
  - Solution set will be available following end of Lecture 12
- Struggling?
  1. Read the 'Encouraged Reading'
  2. Review 'Supplementary material'
  3. Reach out directly