# Motivating Example

# Roasting Coffee: Introduction

# Roasting Coffee: Field Trip

# Statistics, Parameters, and Estimators

# Introduction to Properties of Estimators

# Estimating Characteristics of Populations

- Data scientists are not given the joint density distribution.
- One of our principle tasks is to produce *estimates* about the *universe* given only the limited information that we get from a sample of data.

# Estimating Characteristics of Populations (cont.)

## Parameters, Statistics, and Estimators

- A **paramater**, $\theta$, is some summary of the joint distribution.
- A **statistic** is some function that maps data from $\mathbb{R}^n \to \mathbb{R}$.
- An **estimator**, $\hat{\theta}$, is some statistic of the data that we use to produce a guess for $\theta$.

# $\overline{X}$ is an Estimator

We refer to statistics that are “good” guesses for a population parameter as **estimators**.

- $\theta = \mathrm{E}[X]$ is a parameter.
- $T_{(n)} =$ choose the $30^{th}$ percentile is a statistic
- $\hat{\theta} = \overline{X}$ is an estimator for $\theta$.

# Introducing Estimators

In this section, you will learn about the properties that make estimators good guesses for a population parameter.

**Properties of estimators**

- **Consistency:** As sample size grows, does the estimator converge in probability to the true value?
- **Bias** and **unbiasedness**: Is the estimator systematically too low or too high?
- **Efficiency:** Does an estimator have relatively *large or small* sampling variance, standard error, and MSE?

# Reading: Core Estimation Theory

# Reading: Core Estimation Theory

**Note: This is a READING CALL, just placing it here for organization.** Read pages 102–105, stopping at 3.2.3, Variance Estimators.

# Desirable Properties of Estimators

# Desirable Properties of Estimators

Throughout, we refer to $\theta$ as a population *feature* or *parameter*.

- $\theta$ has a fixed, true value, but this value is not directly observable.
- Without ground truth, how can we know if our estimator is doing its job?

## Desirable Properties of Estimators: Unbiasedness

**Bias of an estimator**

The bias of an estimator is the expected difference between the *true* population value and the estimator.

- The *bias* of $\hat{\theta}$ is $\mathrm{E}[\hat{\theta}] - \theta$.
- If $\mathrm{E}[\hat{\theta}] = \theta$, then there is no bias, and the estimator is *unbiased*.

# Desirable Properties of Estimators: Unbiasedness (cont.)

If $\hat{\theta}$ is unbiased, does that mean that, for any sample taken, the estimate produced by $\hat{\theta} = \theta$?

- $\hat{\theta}$ as an *estimator* is a random variable.
- The value that it takes on given a sample is the *estimate*.

# Desirable Properties of Estimators: Unbiasedness (cont.)

# Desirable Properties of Estimators: Efficiency

## Mean Squared Error of an Estimator

- The MSE of an estimator $\hat{\theta}$ is

$$E[(\hat{\theta} - \theta)^2] = V[\hat{\theta}] + (E[\hat{\theta}] - \theta)^2$$

# Desirable Properties of Estimators: Efficiency

# Desirable Properties of Estimators: Consistency

**Consistency**

An estimator $\hat{\theta}$ is consistent for $\theta$ if $\hat{\theta} \xrightarrow{p} \theta$.

# Desirable Properties of Estimators: Consistency

# Applying Properties of Estimators

# Evaluating Estimators

Which do you prefer?

# Evaluating Estimators (cont.)

# Desirable Properties of Estimators: Review

- An *inconsistent* estimator is of little use.
- A *more efficient* is estimator is preferable to a *less efficient* estimator, all else equal.
- An *unbiased* estimator is preferable to a *biased* estimator, all else equal.

# Random Sampling

# Reading: Independent and Identically Distributed

## Reading: Independent and Identically Distributed

- Read Section 3.0 and 3.1 in *Foundations of Agnostic Statistics* (pages 91–96 in our copy of the book.)
    - Work to place the formal math definition into plan understanding. We will discuss this when you come back.
    - Notice how $\mu$ and $\sigma^2$ are concepts that are immediately familiar, but they now represent population parameters.

# Independent and Identically Distributed

# Independent and Identically Distributed

**Definition: independent and identically distributed (IID)**

- Undertake some process an arbitrary number of times—call it *sampling*—to create a value from a phenomenon.
- If each instance of sampling draws from the same probability distribution, then we say the collection of values are **identically distributed**.
- If none of the instances of sampling provide information about other instances of sampling, then we say the collection of values are **independent**.

# Independent and Identically Distributed (cont.)

# Making Assumptions

- What is a statistical assumption?
- When are they important?
- When can I violate them?
- What do I do if they are violated?

# Making Predictions

# Making Predictions

- When data are independent and identically distributed (IID), it is possible to learn desirable things from a sample:
    - Accurate characterizations of the probability distribution function and values
    - Reliable characterizations of certainty and uncertainty
- We can learn about unseen *population parameters* from a sample and then use our domain knowledge to evaluate whether these population parameters apply to some circumstance.

# What Is a Population?

**U.S. Senator fundraising**

There are 100 U.S. Senators.

- One could reason about a finite population with 100 elements.
- Or, one could assume a continuous distribution for fundraising.

# Learnosity: Is This IID?

# Is This IID?

**Note: This is a learnosity activity, just placing it here for organization.**

For each of these, would you say that this sample is distributed IID? What do you understand about the process that brings you to your conclusion?

- Coffee
  - **Goal:** Understand how roasted a batch of coffee is.
  - **Sampling process:** Dip into a drum to pull a set of 30 beans.
- Conduct a draft
  - **Goal:** Randomly select people for military service.
  - **Sampling process:** Draw balls from an urn with a day of the year. Everybody born on that day goes to war.

# Is This IID? (cont.)

- Teach a computer to *really* read
  - **Goal:** Teach a computer to understand theme and plot (the work of School of Information professor David Bamman).
  - **Sampling process:** Feed a neural network each word (in each sentence [in each paragraph (in each chapter)]) of the English language literary canon.
- Any others?
  - **Goal:**
  - **Sampling process:**

# Reading: Sample Statistics

## Reading: Sample Statistics

- Read pages 96–98.
- Stop before theorem 3.2.5.

# Sample Statistics

# Sample Statistics

Sample statistics are:

- Functions that are applied to samples of random variables
- *Themselves* random variables

# Conduct Sampling

# Conduct Sampling

**Note: This is a learnosity activity, just placing it here for organization.**

The goal is that this will solidify the understanding that students have from the reading and will move us forward into a conversation about expected value and the sample variance of the sample average.

- Students will be provided with data and starter code that produces a population.
- From this population, they will have sliders they can pull that will increase or decrease the number of samples that they take, the number of draws per sample, and the underlying population variation.

# The Sample Mean as an Estimator

# The Sample Mean

**Definition: Sample Mean**

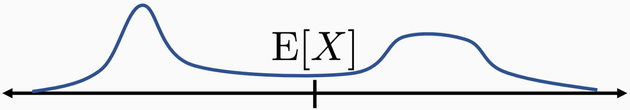
For IID random variables, $X_1, ..X_n$, the *sample mean* is

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- The sample mean is an *estimator* for $\mathrm{E}[X]$.

# The Sample Mean is an Estimator

1. Population Distribution




2. Sample

3. Sampling Distribution of the Statistic

# Evaluating the Sample Mean as an Estimator

First Questions:

1. Is the sample mean biased?
2. How efficient is the sample mean?

# Unbiasedness of the Sample Mean

**Theorem: The Sample Mean is Unbiased**

For IID random variables, $X_1, ..X_n$, the sample mean $\overline{X}$ is an unbiased estimator for the population mean $\mathrm{E}[X]$.

# Efficiency of the Sample Mean

**Theorem: Sampling Variance of the Sample Mean**

For IID random variables, $X_1, ..X_n$, with population variance, $\mathbb{V}[X]$, the variance of the sample mean is

$$\mathbb{V}\left[\overline{X}\right] = \frac{\mathbb{V}[X]}{n}$$

# Learnosity: Sampling Variance of the Sample Mean

**Note: This is a learnosity activity, just placing it here for organization.**

- If you increase the sample size, does the sampling variance of the sample mean increase, decrease, or stay the same?
- At what rate does this change? (Options include faster than the data, at the same rate at the data, slower than the data.)
- If you increase the sample size, does the population variance, $\mathbb{V}[X]$, change? (Hint: It is a population parameter.)

# Consistency and the Continuous Mapping Theorem

# Consistency

*"If you can't get it right as n goes to infinity, you shouldn't be in this business."*

- Clive W.J. Granger

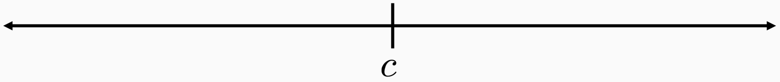
How can we formalize "right as $n$ goes to infinity?"

- Estimators may converge at different rates.
- Deterministic guarantees are not possible.

# Intuition for Convergence in Probability

- Let $T_{(1)}$ be the statistic with 1 datapoint.
- Let $T_{(2)}$ be the statistic with 2 datapoints.
- Let $T_{(3)}$ be the statistic with 3 datapoints.

$$\vdots$$



**Definition: Convergence in Probability**

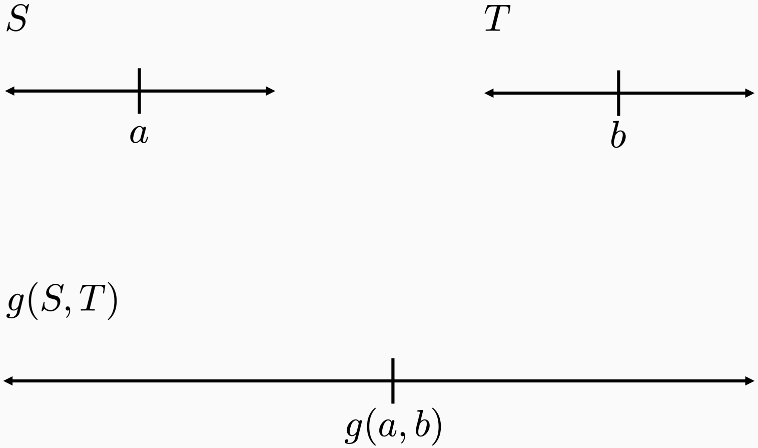
Let $(T_{(1)}, T_{(2)}, T_{(3)}, \ldots)$ be a sequence of random variables and let $c \in \mathbb{R}$. $T_{(n)}$ *converges in probability* to $c$ if for all $\epsilon > 0$,

$$\lim_{n \to \infty} P\Big[T_{(n)} \in (c - \epsilon, c + \epsilon)\Big] = 1$$

We write this as $T_{(n)} \xrightarrow{p} c$.

- An estimator $\hat{\theta}$ is *consistent* for $\theta$, if $\hat{\theta} \xrightarrow{p} \theta$.

# Continuous Mapping Intuition



## The Continuous Mapping Theorem

Let $(S_{(1)}, S_{(2)}, S_{(3)}, \ldots)$ and $(T_{(1)}, T_{(2)}, T_{(3)}, \ldots)$ be two sequences of random variables. Let $g : \mathbb{R}^2 \to \mathbb{R}$ be a continuous function. If $S_{(n)} \xrightarrow{p} a \in \mathbb{R}$ and $T_{(n)} \xrightarrow{p} b \in \mathbb{R}$, then $g(S_{(n)}, T_{(n)}) \xrightarrow{p} g(a, b)$

# Reading: Weak Law of Large Numbers

# Reading: Weak Law of Large Numbers

Read page 100, beginning at theorem 3.2.8, through the end of page 102.

# Weak Law of Large Numbers

# The Weak Law of Large Numbers

**Theorem: The Weak Law of Large Numbers**

Let $(X_1, X_2, X_3, ...)$ be a sequence of i.i.d. random variables with finite variance. Let $\overline{X}_{(n)} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then

$$\overline{X}_{(n)} \xrightarrow{p} \mathrm{E}[X]$$

- **Equivalently:** The sample mean is *consistent* for the population mean.

# Consequences of WLLN

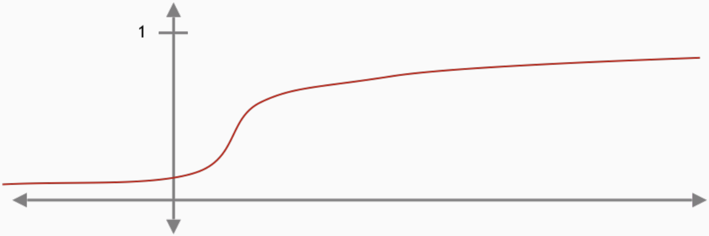
**The sample mean is consistent.**

- The error converges in probability to zero.
- The sample mean is accurate, as long as we can make $n$ large enough.

**As a building block to study other estimators.**

- Combine WLLN with Continuous Mapping Theorem.

# Applying the WLLN




**Objective:** Estimate the cdf $F$ at a point $x$.

**Idea:** Use empirical cdf

# Proof of the Weak Law of Large Numbers

# Proof of the Weak Law of Large Numbers

Given random variable $X$ and $\epsilon > 0$.

# Proof of the Weak Law of Large Numbers

Given random variable $X$ and $\epsilon > 0$.

Let $D = |\bar{X} - \mathrm{E}[X]|$

$\mathrm{V}[\bar{X}] = \mathrm{E}[D^2] = \mathrm{V}[X]/n$

$\mathrm{V}[\bar{X}] = \mathrm{E}[D^2] = \mathrm{E}[D^2|D < \epsilon]P(D < \epsilon) + \mathrm{E}[D^2|D \geq \epsilon]P(D \geq \epsilon)$

$\geq 0 + \epsilon^2 P(D \geq \epsilon)$

$P(D \geq \epsilon) \leq \frac{\mathrm{V}[X]}{\epsilon^2 n} \xrightarrow{p} 0$

# Simulating the WLLN

**Note: This is a learnosity activity, we're just including it here for organization.** Students will work through the notebook called `WLLN.Rmd`.

# Reading: Estimating Population Variance

# Reading: Estimating Population Variance

- Read pages 105–108, stopping before the beginning of the next section.
- You are going to use *the plug-in principle*, which is a general approach for designing estimators in a sample.

# Estimating Population Variance

# Why Estimate Variance?

- How much coffee do students drink on average?
- How much does coffee intake differ from student to student?

Population Variance: $\mathrm{V}[X] = \mathrm{E}[X^2] - \mathrm{E}[X]^2$

**The Plug-In Variance Estimator**

Given a sample $\boldsymbol{X} = (X_1, X_2, ..., X_n)$, the *plug-in variance estimator* is,

$$\tilde{\mathrm{V}}(\boldsymbol{X}) = \overline{X^2} - \overline{X}^2$$

# Consistency of the Plug-In Estimator

**Theorem: Consistency of the Plug-In Variance Estimator**

Let $(X_1, X_2, X_3, \ldots)$ be a sequence of i.i.d. random variables with finite variance $\mathbb{V}[X]$. Then $\tilde{\mathbb{V}}(\boldsymbol{X}) = \overline{X^2} - \overline{X}^2$ is consistent for $\mathbb{V}[X]$.

# Bias of the Plug-In Variance Estimator

Let $(X_1, X_2, X_3, \dots)$ be a sequence of i.i.d. random variables with finite variance $\mathrm{V}[X]$. Then

$$\mathrm{E}[\tilde{\mathrm{V}}(\boldsymbol{X})] = \frac{n-1}{n}\mathrm{V}[X]$$

# A Better Variance Estimator

**The Unbiased Variance Estimator**

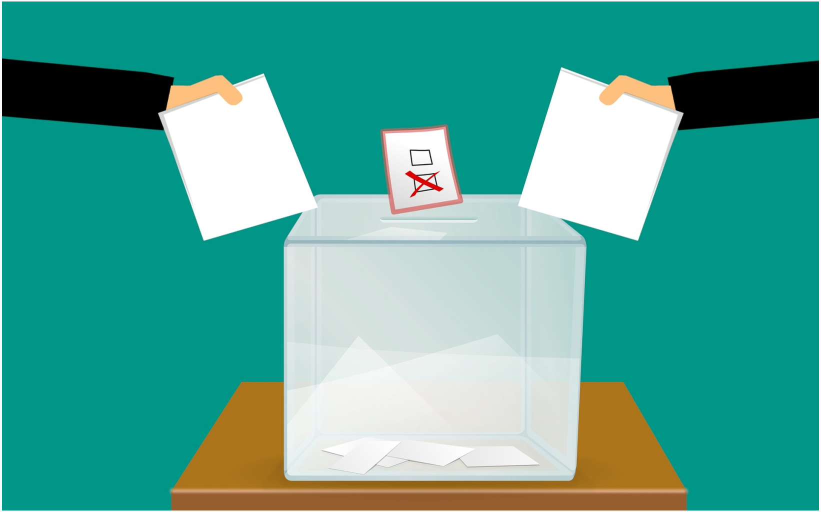
Given a sample $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$, the *unbiased sample variance estimator* is,

$$\hat{\mathrm{V}}(\boldsymbol{X}) = \frac{n}{n-1}\left(\overline{X^2} - \overline{X}^2\right)$$

# Standard Errors

# Point Estimate $\Leftrightarrow$ Uncertainty

# Importance of Uncertainty



**Estimate:** Our candidate has 52% support.

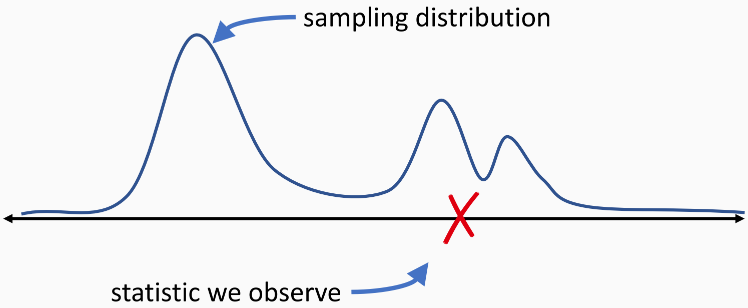
# Importance of Uncertainty

**Estimate:** The maximum dose of the medication is 26mg.

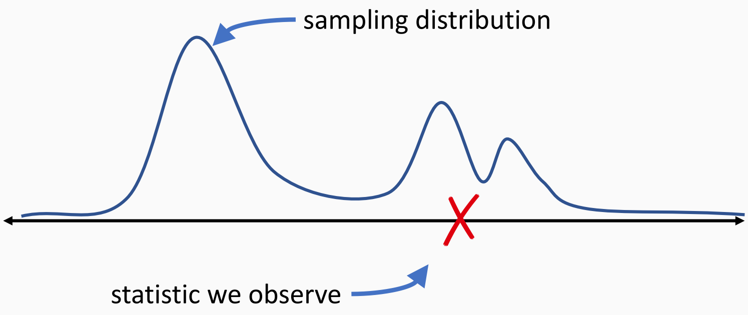
# Importance of Uncertainty

**Estimate:** The angle of the tower is 90 degrees.

# The Sampling Distribution of the Statistic



# The Sampling Distribution of the Statistic




**Standard Error:** (Estimated) standard deviation of the sampling distribution

# Reporting standard errors

1. The mean number of mushrooms per pizza was 13.2 (SE 3.6).
2. The time for mice to navigate the maze was 35 $\pm 2$ seconds.
3. 

| Vitamin W | Vitamin X |
|---|---|
| 2.3 | 3.4 |
| (0.3) | (0.9) |

# Standard Errors, The Sample Variance, and Standard Deviation

# Many Measures of Dispersion

**Sample Variance**

**Sampling Variance of the Sample Mean**

**Sample Standard Deviation**

**Standard Error of the Sample Mean**

# Motivating The Central Limit Theorem

# Capturing Uncertainty

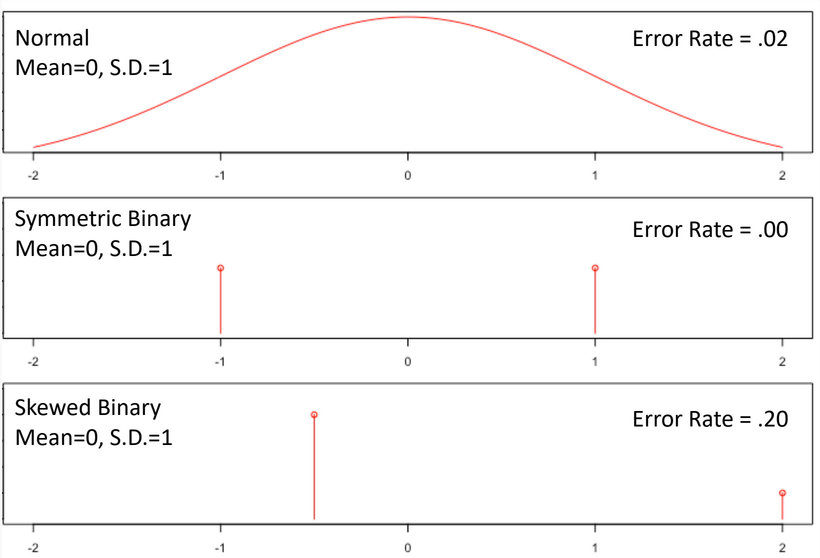
Need tools to capture how far off our estimate may be.

- Standard Error: the (estimated) standard deviation of the sampling distribution of the estimator.
- But a standard error is just one number...

# Uncertainty Example

- Hidden fact: Population mean = 0
- Idea: Market if estimate is 2 standard errors above 0.

# The Importance of Shape



**Central limit theorem (CLT): idea**

For a **very broad** class of population distributions, the sampling distribution of the mean becomes approximately normal as the sample size grows large.

# Reading: The Central Limit Theorem

# Reading: The Central Limit Theorem

**Note: This is a reading call, we're just placing it here for organization.** Read pages 108 and 109 of section 3.2.4. There's no *need* to read through the demonstration of Slutsky's theorem, but you can if you would like.

- Rather than proving the CLT, we are going to ask you to work through a short demonstration against data that we hope will convince you of the CLT's effectiveness.
- The book is terse in its presentation of when and how the CLT applies. We will fill that out when we come back together.

# Apply the Central Limit Theorem

**Note: This is a learnosity activity, just placing it here for organization.**

# Central Limit Theorem

# Reminder of Context

- The WLLN tells us what happens to the sample mean as $n \to \infty$:
$$\overline{X} \xrightarrow{p} \mathrm{E}[X]$$
- We also know that, as $n \to \infty$, we can generate an increasingly good estimate for $\mathrm{V}[X]$ because
$$\overline{X^2} - \overline{X}^2 \xrightarrow{p} \mathrm{V}[X]$$
- For more precise statements about uncertainty, we need the sampling distribution of the statistic.

# Convergence in Distribution

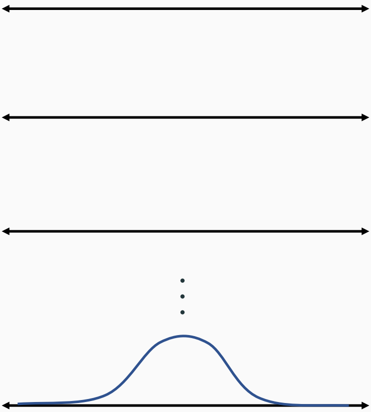
Statistic for $n$=1 $T_{(1)}$

Statistic for $n$=2 $T_{(2)}$

Statistic for $n$=3 $T_{(3)}$

⋮ ⋮

Target Random Variable $T$
(Usually Normal)



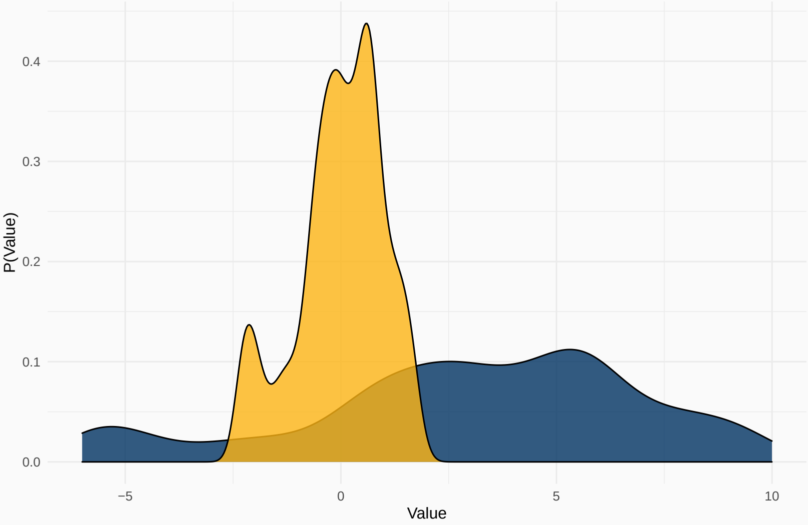
**Definition: Convergence in distribution**

Let $(T_{(1)}, T_{(2)}, T_{(3)}, \ldots)$ be a sequence of random variables, with cdfs $(F_{(1)}, F_{(2)}, F_{(3)}, \ldots)$, and let $T$ be a random variable with cdf $F$. Then $T_{(n)}$ *converges in distribution* to $T$ if, for all $t \in \mathbb{R}$ at which $F$ is continuous,

$$\lim_{n\to\infty} F_{(n)}(t) = F(t).$$

We denote this as $T_{(n)} \xrightarrow{d} T$.

# The need to Standardize



# Standardizing the Sample Mean

**Definition: Standardized sample mean**

For IID random variables $(X_1, X_2, ..., X_n)$ with finite $\mathrm{E}[x] = \mu$ and finite $\mathrm{V}[X] = \sigma^2$, then **the standardized sample mean** is:

$$Z = \frac{\left(\overline{X} - \mathrm{E}\left[\overline{X}\right]\right)}{\sigma\left[\overline{X}\right]} = \frac{\sqrt{n}\left(\overline{X} - \mu\right)}{\sigma}.$$

- The standardized sample mean always has mean 0 and standard deviation 1.

**The Central Limit Theorem**

Let $(X_1, X_2, X_3, \ldots)$ be a sequence of i.i.d. random variables with finite mean $\mathrm{E}[X] = \mu$ and finite variance $\mathrm{V}[X] = \sigma^2$,

$$Z = \frac{\sqrt{n}\left(\overline{X} - \mu\right)}{\sigma} \xrightarrow{d} N(0, 1)$$

# Generality of the CLT

The CLT applies to:

- Continuous Random Variables
- Discrete Random Variables
- Symmetric Random Variables
- Asymmetric Random Variables

Only Requirements:

- Data points are IID.
- Population has finite variance.

Versions of the CLT exist for many other statistics.

# What Sample Size is Enough?

The CLT works in the limit as $n \to \infty$. What can we say for a finite $n < \infty$?

- Rule of Thumb: $n = 30$ for CLT to "kick in."
- Reality: Convergence depends on how non-normal population is.
    - A normal population requires $n = 1$.
    - Highly skewed distributions may require $n = 100$, $n = 1000$, or more.

# Learnosity: When Does the CLT Apply?

# Where and When Does the CLT Apply? Part III

**Which of these random variables has an approximately normal distribution because of the CLT?**

- $X =$ hours spent by one individual on a 203 homework assignment
- $X =$ average hours spent by randomly assigned study groups of size 4 on the same 203 homework
- $X =$ number of barks by a dog named Rex for a one-hour period at night

# Where and When Does the CLT Apply? Part III

**Which of these random variables has an approximately normal distribution because of the CLT?**

- $X =$ total number of barks by a neighborhood of dogs for a one-hour period at night
- $X =$ age of a randomly selected MIDS student
- $X =$ average of sample of 10 randomly selected MIDS students
- $X =$ VADER sentiment of a sample of 100 SMS messages

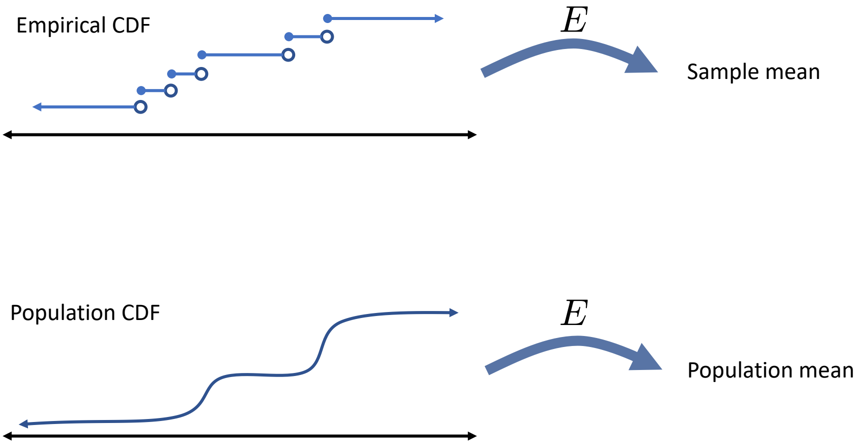
# The Plug-In Principle

# The Plug-In Principle

- Want mean of population $\rightarrow$ Mean of sample.
- Want variance of population $\rightarrow$ Variance of sample.
- Want $f(E[X], V[X])$? $\rightarrow f(\bar{X}, \hat{V}(X))$

See the Pattern?

# The Plug-In Principle



# Reading Assignment

# Reading Assignment

**Only if you're interested,** read section 3.3 - 3.3.1 to learn more about the Plug-In Principle.

# Asymptotic Theory

# Asymptotic Theory

## Asymptotics Rescue Data Science

**Across a range of applications, small samples are problems.**

- If we've got a *lot* of data, though, we can rely on weaker requirements
- Asymptotic properties of estimators ask the question, **What happens when we have a lot of data?**
- In general, as $n \to \infty$, does $\hat{\theta}$ converge in distribution or probability to something that is useful or desirable?

# Asymptotic Theory

## Reading: Asymptotics

# Reading: Asymptotics

**Note: This is a READING CALL. We're placing it here for organization.**

- The interested student can read pages 111-114.
- But, we might recommend skipping it if you're constrained.
- The take home is that there is an asymptotic statement of:
    - Being Normally Distributed
    - SE, MSE and Efficiency
    - Sampling Variance and Sampling Standard Error

# Asymptotic Theory

## Reading: The Plug-In Principle

# Reading: The Plug-In Principle

- $F(x)$ is the cumulative distribution function, the *CDF*
- $f(x)$ is the probability distribution function, the *PDF*
- $\hat{F}(x) \neq F(x)$ and $\hat{f}(x) \neq F(x)$
- As $n \to \infty$ we hope that $\hat{F}(x) \xrightarrow{d} F(x)$ and $\hat{f}(x) \xrightarrow{d} f(x)$.

# Asymptotic Theory

## The Plug-In Principle

**Expectation and Variance Functionals**

You *know* what the processes for calculating an expectation and variance are.

$$E[X] = T_E(F) = \int x \cdot dF(x)$$

$$V[E] = T_V(F) = \int (x - E[X])^2 \cdot dF(x)$$

These are the *statistical functionals* for expectation and variance.

# The Plug-In Principle (cont.)

## Plug-In Estimators

For i.i.d. random variables $X_1, X_2, \ldots, X_n$, with a common CDF $F$, the plug in estimator for $\theta = T(F)$ is just $\hat{\theta} = T(\hat{F})$.

$$\begin{aligned}
\hat{E}(X) &= T_E(\hat{F}) \\
&= \sum x \cdot \hat{f}(x) \\
&= \sum x \cdot \frac{I(X_i = x)}{n} \\
&= \frac{1}{n} \sum x \\
&= \overline{X}
\end{aligned}$$

# Example of the Plug-In Principle

## Example with Discrete Data

Suppose you there is some discrete process that places values into the random variable $X$.

- $F_X(1) = 1/3$
- $F_X(2) = 1/6$
- $F_X(3) = 0$
- $F_X(4) = 1/2$

But, you do get to produce i.i.d. random draws from the process. Of 1000 draws:

| 1 | 2 | 3 | 4 |
|---|---|---|---|
| 331 | 156 | 0 | 513 |

# Example of the Plug-In Principle, cont'd

## Example with Discrete Data

$$\hat{E}(\boldsymbol{X}) = T_E(\hat{F})$$

# Asymptotic Theory

## Reading: Kernel Methods Estimate the PDF

# Reading: Kernel Methods Esttimate the PDF

Read pages 121 - 124.

# Asymptotic Theory

## Kernel Methods Estimate the PDF

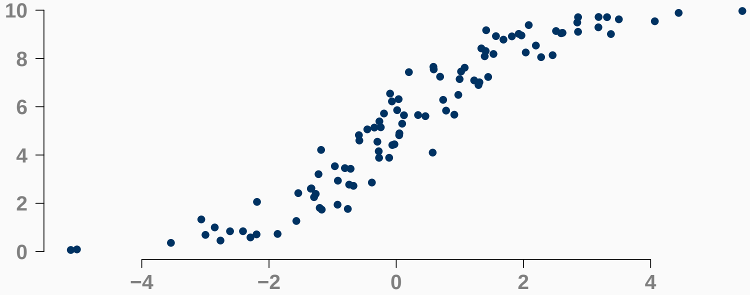
## Kernel Density Estimator of PDF

- Cannot *directly* observe the joint PDF.
- For discrete RV, approximations come through frequency tables.
- For continuous RV, approximations come through kernel density estimates:

$$\hat{f}_K(x) = \frac{1}{n}\sum_{i=1}^{n} K_b(x - X_i), \forall x \in \mathbb{R}$$

# Kernel Methods, Part II

- Kernel methods are smoothing methods
- The *kernel* is some weighting function



As smoothing function, these permit plug in estimates that are directly analogous to the estimating functionals of the CDF.

**Kernel Plug-In Estimator**

The *Kernel plug-in estimator* of $\theta = T(F)$ is

$$\hat{\theta}_k = T\left(\hat{F}_K\right),$$

where, $\hat{F}_K = \int_{-\infty}^{x} \hat{f}(u) du$.

## Kernel Plug-In Estimator for Expected Value

We know the functional for the expected value has a specific "shape"

$$\mathrm{E}[Y] = \int y \cdot df(y).$$

So, the feasible kernel method is

$$\hat{\mathrm{E}}_k(\boldsymbol{Y}) = \int_{-\infty}^{\infty} y \cdot d\hat{f}_K(y)$$

# Kernel Methods, part V

- E[$Y$] and E[$Y|X$] are lowest MSE estimates of $Y$

$$\hat{\mathrm{E}}_K(\boldsymbol{Y}) = \int y\hat{f}(y)$$

$$\hat{\mathrm{E}}_K(\boldsymbol{Y}|X = x]) = \int y\hat{f}_{Y|X}(y|x)$$

- If you can produce i.i.d. samples from $f_{Y|X}$, you can produce estimates, $\hat{f}_K(y|x)$ that get ever closer to the true value

# Asymptotic Theory

## Apply Kernel Methods

# Apply Kernel Methods

**Note: This is a LEARNOSITY activity. We're just placing it here for organization.**