Probability & Sampling Distributions

Dr Austin R Brown

Kennesaw State University

Why Should I Care?

- Whether conscious of it or not, you already do care about probability, or you would likely not be in college.
  - People with college degrees have higher rates of employment than those who don't (Kurtzleben, 2014).
  - People with college degrees tend to make more money than those who don't (Kurtzleben, 2014).

- As researchers, we care about quantifying the likelihood that our sample data provide statistical support for our research hypothesis.
  - We call this likelihood, *probability*

Sample Spaces & Probability

- Let's start off with some fundamental definitions and concepts:

- A **probability experiment** is a chance process that leads to well defined outcomes
  - Flipping a coin is a probability experiment because it is a chance process that leads to one of two well defined outcomes: heads or tails

- An **outcome** is the result of a single trial of a probability experiment.

- A **sample space** is the set of all possible outcomes of a probability experiment.

Sample Spaces & Probability

- For us, this practically means that for whatever it is that we're measuring/studying, there exists some total set of possible observations.

- For example, if we're studying the systolic blood pressure of chronic marijuana users, we will for sure observe a number greater than 0 and less than 200.
  - We can consider that interval our sample space!

- The sample space for resting heart rate among adults living in Georgia is also likely an interval such as 1 - 200 beats per minute.
  - There's 100% chance we will observe something in that interval.

Sample Spaces & Probability

- We often hear about probability in our daily lives and typically think of it as a measuring stick of uncertainty.
  - 70% chance of rain? The closer that number gets to 100%, the more likely we think the event is to happen.

- In statistics, though, probability has a slightly different interpretation. It's actually the proportion of outcomes of interest in a sample space.

- Let's say we have 10 students in a class. 3 are freshmen, 0 are sophomores, 2 are juniors, and 5 are seniors. What's the probability a student from this group is a junior?
  - We have 2 in the junior category and 10 total students so $P(\text{Junior}) = 2/10 = 0.20$.

Sample Spaces & Probability

- More specifically, probability is thought of as the "long run" proportion of outcomes called the **law of large numbers**.
  - While we could flip a coin and get heads 1000 times in a row, if we performed the same experiment a huge number of times, we would expect the proportion of times we observed heads to be 0.50.
  - This is the right way to interpret probability

Sampling Distributions

- Obviously, there are lots of different measures we can calculate using data collected from a sample.
  - Mean resting heart rate, for example!

- In the resting heart rate example, we wouldn't sample every single adult in Georgia in order to determine the exact mean resting heart rate for the whole population.
  - We would take a smaller, more manageable/practical sample to come up with an estimate.

- If we were to take another sample, we obviously wouldn't have exactly the same data and thus our sample mean resting heart rate would be different than what we obtained in the first sample.
  - And if we did it again, we'd get something different!

Sampling Distributions

- Now thinking about this in terms of sample spaces, we know that adult resting heart rate also has some sample space likely ranging from 1 to some practical upper limit (I'm not a medical doctor haha).

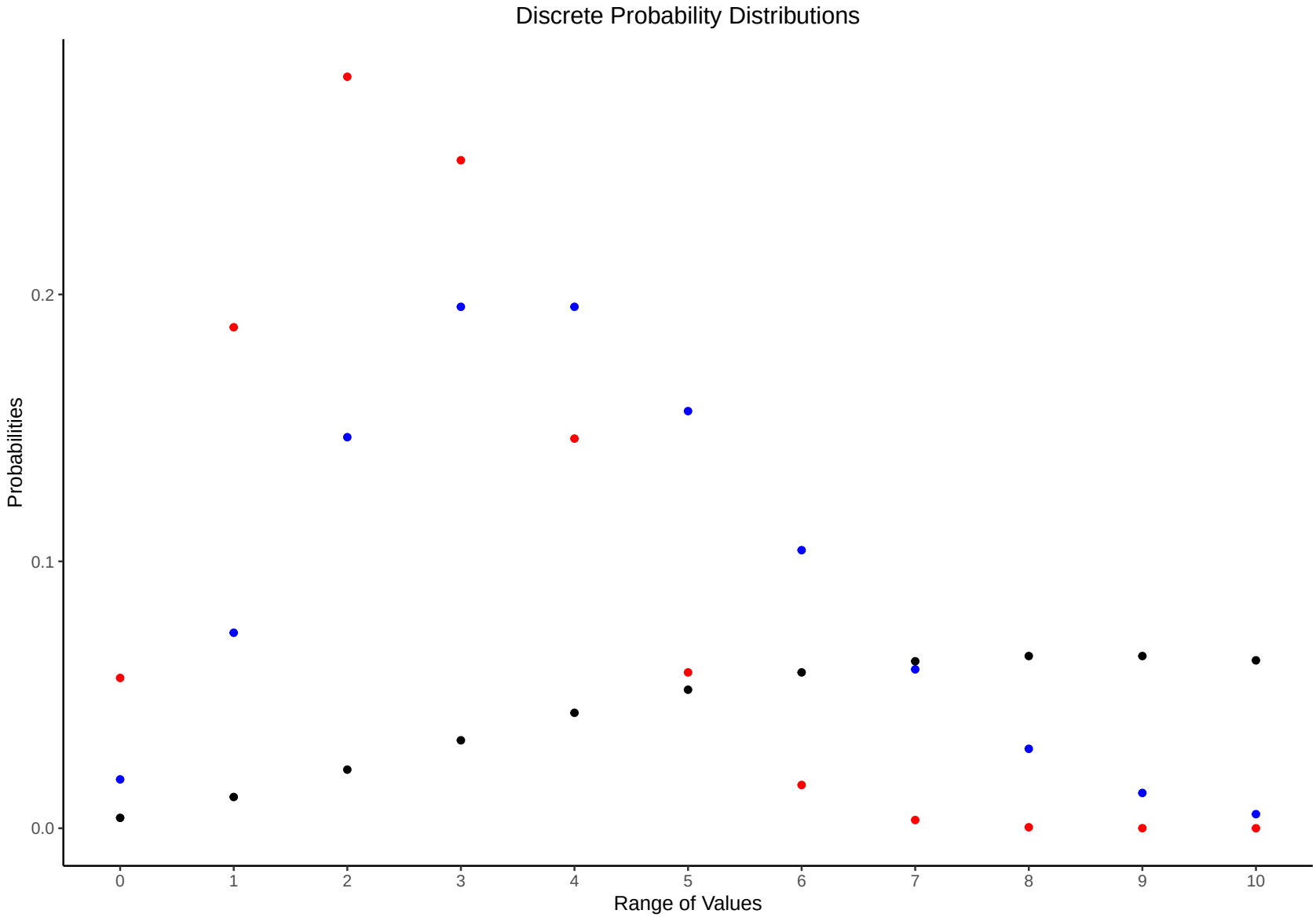
- The same is true of the *sample mean* of resting heart rate.

- However, it is unlikely that each of the possible outcomes in a sample space are equally likely.
  - It's more likely for a human male to be between 5 and 6 feet tall than it is over 7 feet.
  - We probably aren't going to run into many people who have a resting heart rate of 1 beat per minute.

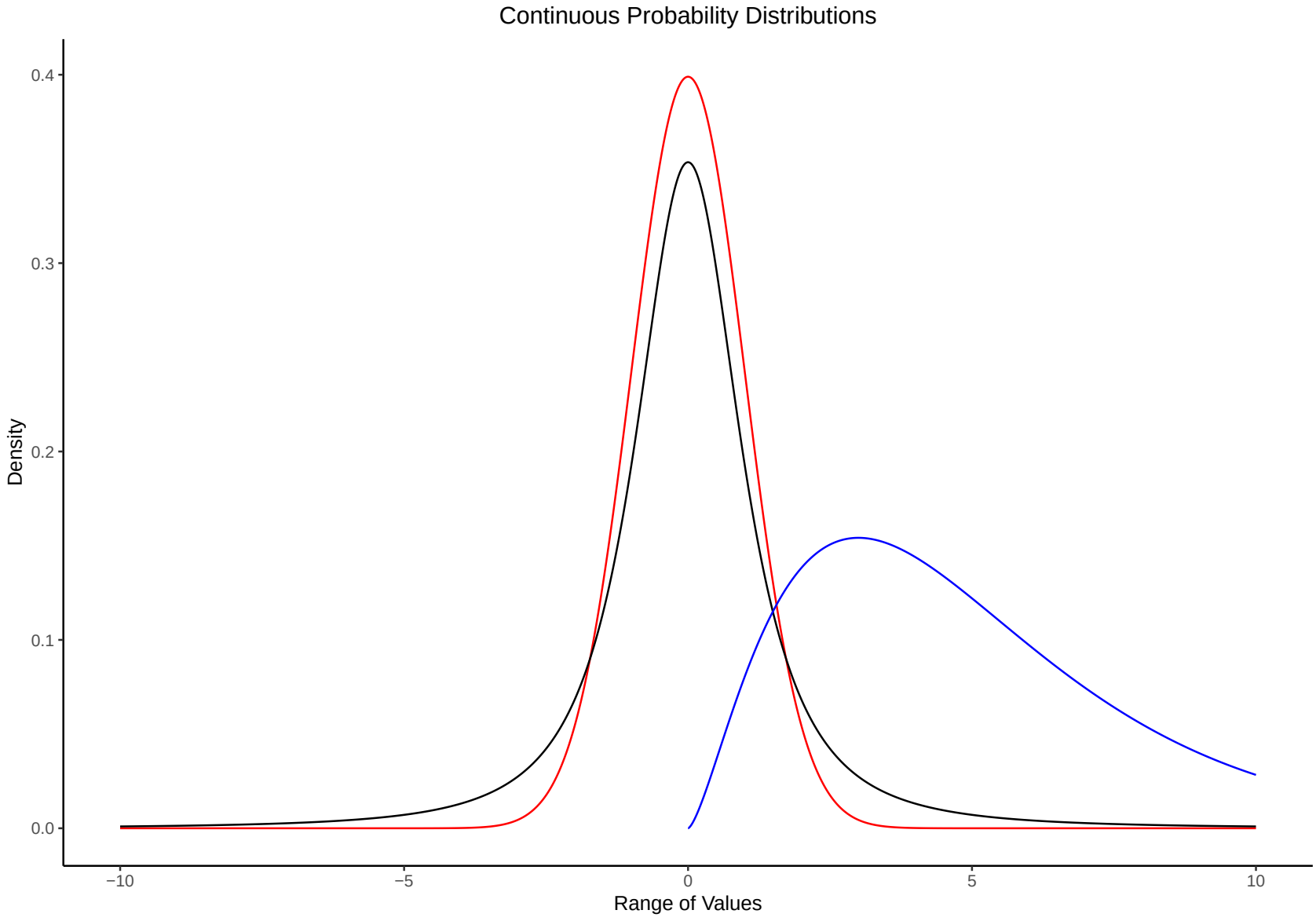
Sampling Distributions

- Each outcome, or range of outcomes, has some associated probability of occurrence, which is akin to a function in mathematics (remember the horizontal line rule in high school algebra?).
- In probability/statistics, we refer to the function which describes the probability of some outcome, or range of outcomes, in a sample space as a **probability distribution function** or PDF for short.
- In theory, anything that is measurable or observable has an associated PDF.

Sampling Distributions



Sampling Distributions



Sampling Distributions

- Distributions model the probabilistic behavior of an entire population.
  - What proportion of the human population is taller than 7 foot?
- Like we will find with sample data, distributions can have characteristics like a mean and variance/standard deviation (ways of quantifying the similarity of quantitative values), too.

Sampling Distributions

- This is well and good, but how can we know what the distribution of human height, resting heart rate of adults living in Georgia, or systolic blood pressure of chronic marijuana users is without measuring all of the members of the population?

- In general, we can't know. So what good are they to us then??

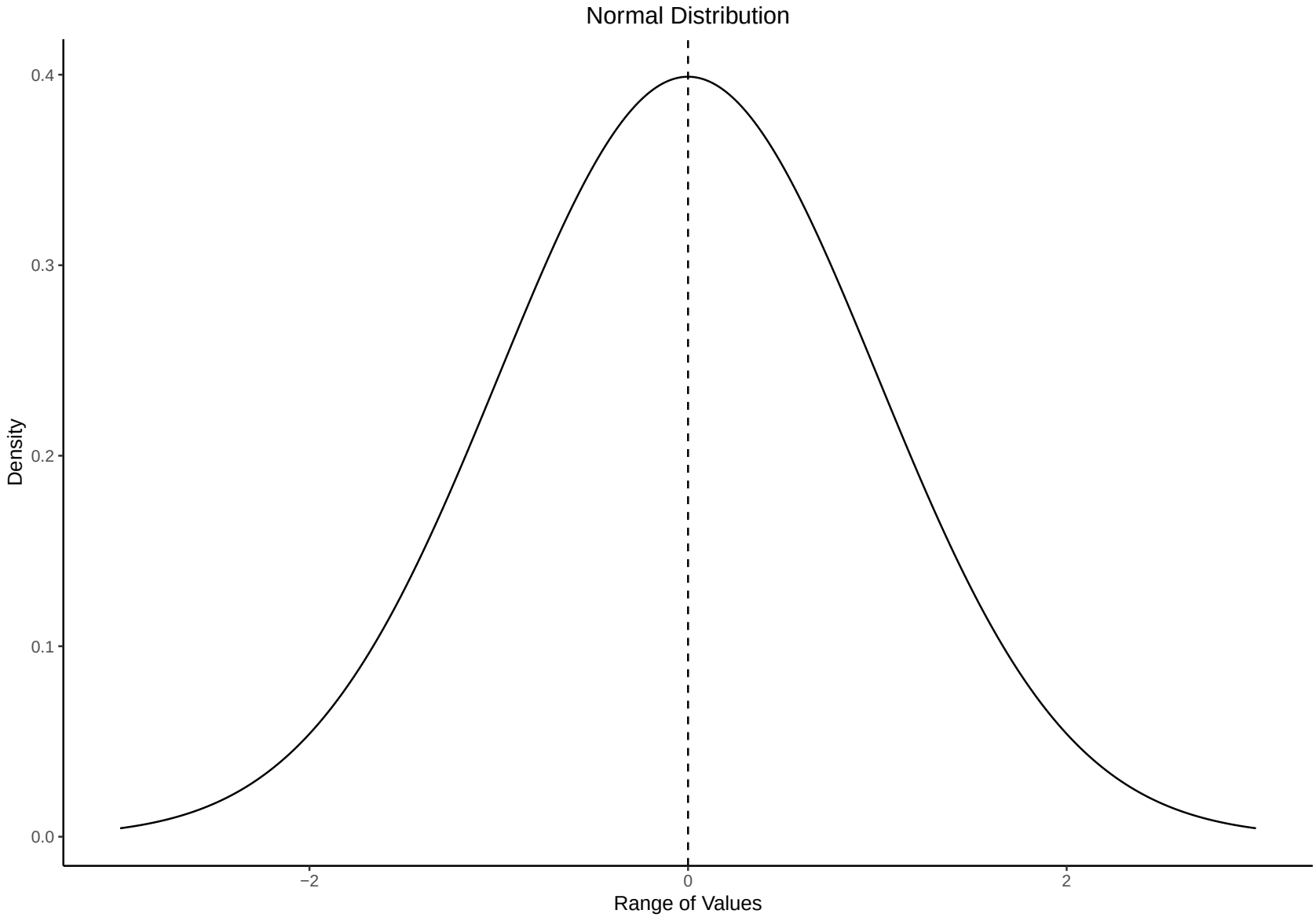
- Fortunately, through a theorem called the "Central Limit Theorem," we can know what the distribution of various **sample statistics** are.
  - A distribution for a sample statistic is called a **sampling distribution**.

Sampling Distributions

- Basically, the CLT says that for any sample statistic meeting certain conditions (which nearly all do), the sample statistic will have an approximately *Normal Distribution* as the sample size used to calculate said sample statistic is reasonably large ($n > 25$)!

- This is cool because, regardless of the PDF of the population (whic again, is typically unknowable to us), we can at least know what the PDF of a sample statistic, like the mean or proportion is!

Sampling Distributions



Sampling Distributions

- Here are some characteristics of the Normal Distribution
  - Has two parameters, a mean, μ and a variance, σ^2
  - Bell-shaped
  - Mean, median, and mode are at the center
  - Unimodal
  - Symmetric about the mean
  - The curve is continuous (meaning for any value of the variable, there is a corresponding probability)
  - The curve approaches, but never touches the x-axis (all values of the variable have some non-zero probability of being observed)
  - The area under the curve equals 1 (same concept as sample spaces!)
  - The area within $\pm 1\sigma$ of the mean is about 68% of the observations, $\pm 2\sigma$ of the mean is about 95% of the observations, and $\pm 3\sigma$ of the mean is about 99.73% of the observations (this is called the "Empirical Rule")

Sampling Distributions

- In the case of the sample mean, we know it has an approximately Normal distribution, but what is its mean and variance?

- Its mean is the mean of the population, which means that it is an unbiased estimator
  - Practically this means we are not systematically overestimating or underestimating the true population parameter being estimated with sample data.

- Its variance is the population variance, σ^2, divided by the size of our sample, n:

$$\bar{x} \dot{\sim} N\left(\mu, \frac{\sigma^2}{n}\right)$$

- Practically, this means as our sample size gets big, the value of our sample mean will approach the value of the population mean

Sampling Distributions

- One of the other main sample statistics whose sampling distribution we will use is the sample proportion.
  - We may want to know, for example, the proportion of pine trees in a national forest afflicted with some disease. Or, what is the proportion of people in Georgia who got the flu vaccine last fall?

- The population proportion is typically denoted with a p and the sample proportion is typically denoted as $\widehat{p}$.

Sampling Distributions

- The sampling distribution of $\hat{p}$ is:

$$\hat{p} \dot{\sim} N\left(p, \frac{p(1-p)}{n}\right)$$

- So again, $\hat{p}$ is an unbiased estimator as its mean is the populatin proportion. And while perhaps not as evident, its variance is the population variance divided by the sample size.

Takeaways

- We need to ensure that we have the appropriate understanding of how to interpret probability as this concept will start coming up a lot as we get into confidence intervals and hypothesis testing
  - Long-run proportion of outcomes of interest
- While we may not now the distribution of our population of interest, we can know the sampling distribution for a sample statistic we calculate from a sample collected from that population!
  - This is fundamental to statistical inference!