Lancaster University

# Quantitative Research Methods

**Matthew Ivory**

matthew.ivory@lancaster.ac.uk

# 1. Introduction to the course

- Session 1: Introduction to quantitative research methods using R
- Session 2: Data management and data wrangling
- Session 3: Exploratory data analysis
- Session 4: Data visualization
- Session 5: Mid-term Assignment
- **Session 6: Probability and distributions**
- Session 7: Tests for discrete variables: Analysing contingency tables
- Session 8: Correlations and t-tests.
- Session 9: ANOVA and linear regression
- Session 10: Multiple regression, introduction to generalised linear regression

# Our plan for today

Lancaster University

**1. Important concepts**

- Populations vs samples
- Descriptive vs inferential statistics
- Random samples and sample bias
- Why is random sampling so important?

**2. Distributions**

- Uniform
- Exponential
- Bimodal
- Normal
- Sampling distributions

**3. Probability theory**

- Law of Large Numbers
- Central Limit Theorem

# Important concepts

- Populations and samples
- Descriptive statistics vs inferential statistics
- Random samples and sample bias
- Why is random sampling so important?

# Populations

- A defined set of individuals, events, objects that we want to know something about.
- The population includes all members of a defined group.

# Populations do **not** need to be large…

- The distinguishing characteristic is not size!
- It is that all those who meet the definition for membership are included.
- Example: The population of the UK, students at Lancaster University, postgraduates at FASS, enrolled students in FASS512…

# Samples

- A subset of the population that we are interested in.
- To determine characteristics of the population, we take random samples of the population.

# Populations and samples

- Usually, we cannot count and observe everything because it would be too time-consuming or expensive.
- Instead, we often rely on examining parts of a population (samples) to learn something about the population.

# Conventions

When referring to populations:

- Descriptive measures of a population are called *parameters*.
- When referring to population parameters, we use Greek letters (or uppercase letters).
- Example Greek letters:
  - population mean $\mu$,
  - population standard deviation $\sigma$

# Conventions

When referring to samples:

- Descriptive measures of a sample are called (sample) *statistics*.
- When referring to sample statistics, we use Roman letters.
- Example Roman letters:
  - sample mean $\bar{x}$,
  - sample standard deviation $s$

# Descriptive statistics and inferential statistics

Descriptive statistics:

- Procedures for classifying or summarizing numerical data.

Inferential statistics:

- Procedures for using our sample data to estimate parameters and to test hypotheses about parameters.

# Random samples

- Subsets of the population in which every member of the population had an equal chance of being selected.

- When we sample randomly, each member of the population must have an equal probability of being drawn.

# Sample bias

- Occurs when the sample is not drawn randomly from the entire population.

- That is, every member of the population did not have an equal chance of being drawn.

# Why is random sampling so important?

Drawing randomly eliminates the possibility of sample bias.

This is because all confounding factors that could bias the results are equally likely to be present in our random sample.

# Examples of Sampling Bias

- **Self-selecting** – participants volunteering or responding to adverts (a study on exercise interventions for health will get you motivated participants who want to achieve better health
- **Attrition** – People who drop out of a study may significantly differ from those who stay (a study on burnout in workplace – who's likely to not have the time or energy to participate?)
- **Under-coverage** – some population members are not represented (an online study precludes those without internet access, ~6-10%)
- **Advertising** – (advertising on campus will get you students)

# Distributions

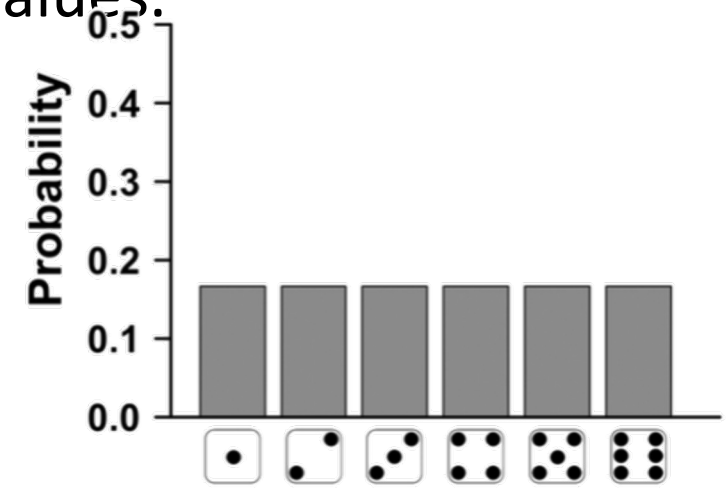
# Distributions

- A probability distribution is the mathematical function that gives the probabilities of occurrence of different possible outcomes for an experiment.

- It is a mathematical description of a random phenomenon in terms of its sample space and the probabilities of events (subsets of the sample space).

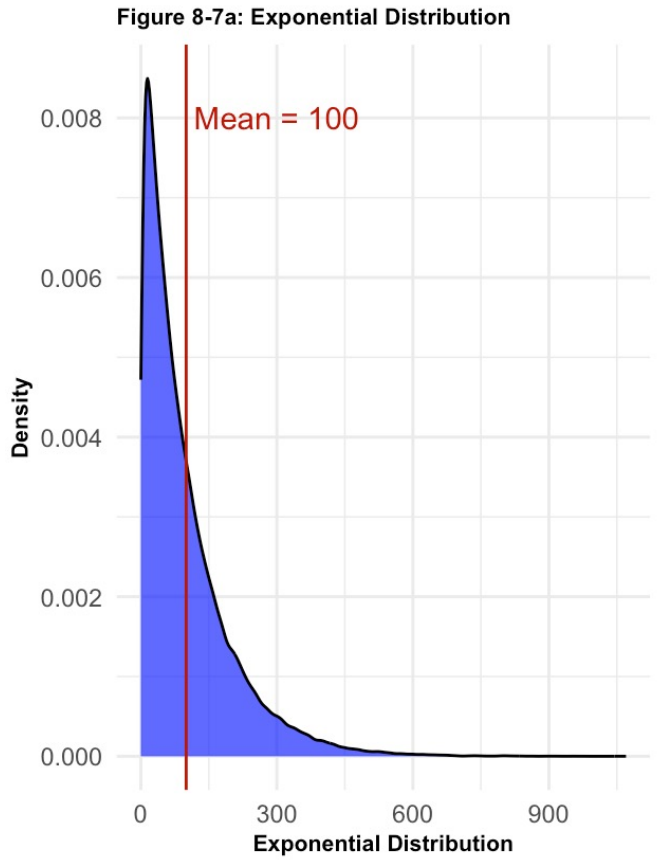
# Uniform distribution

- A type of distribution in which all outcomes are equally likely.
- The shape resembles a square block: the frequency of cases does not change over the range of possible values.



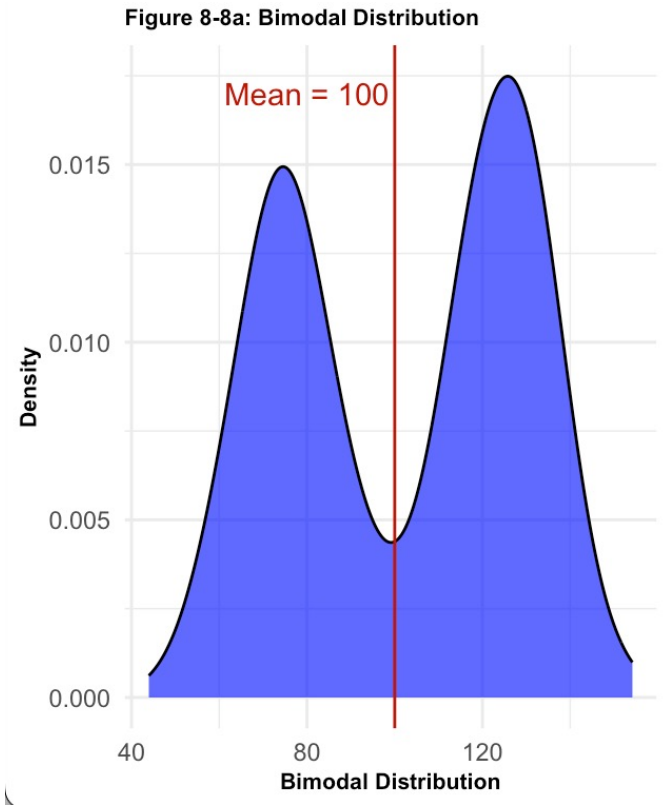
# Exponential distribution

- A type of distribution that is very skewed either in a negative or positive direction.
- A large majority of cases might be grouped together at the low end (positive skew) or the high end (negative skew), with very few cases on the opposing end.



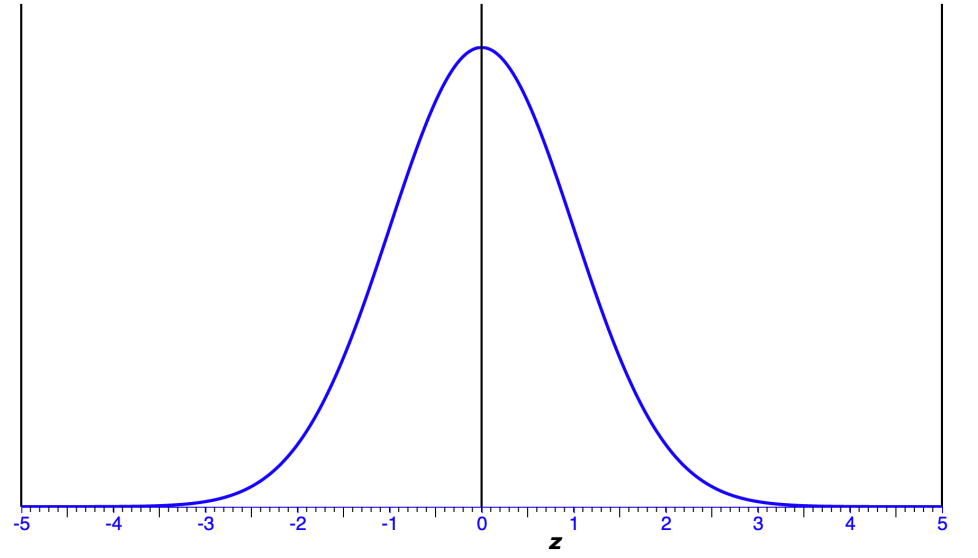
# Bimodal distribution

- A type of distribution with two modes rather than one.
- The shape of a bimodal distribution is characterized by two humps.



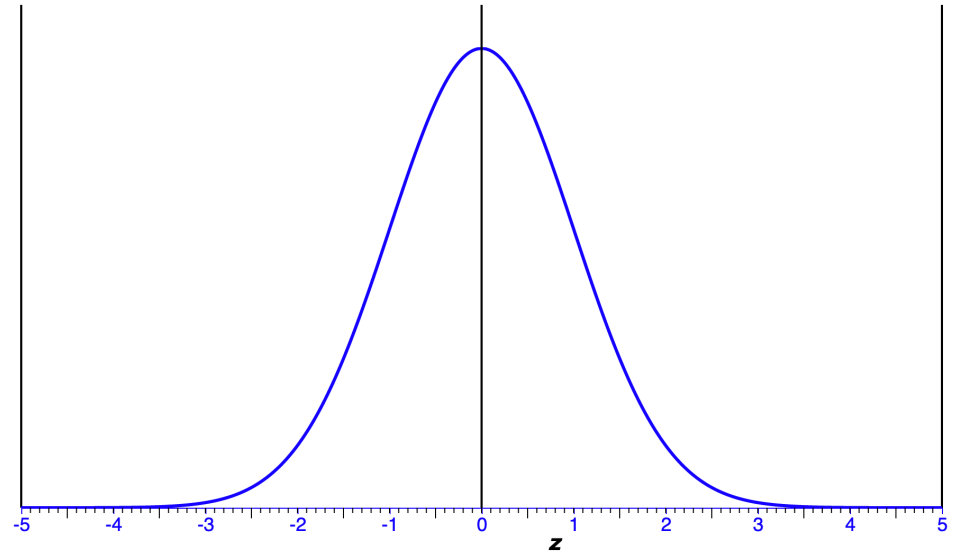
# Normal distribution

- A type of distribution in which the mean, median and mode are exactly the same.
- The distribution is symmetric: half the values fall below the mean and half above the mean.



# Normal distribution

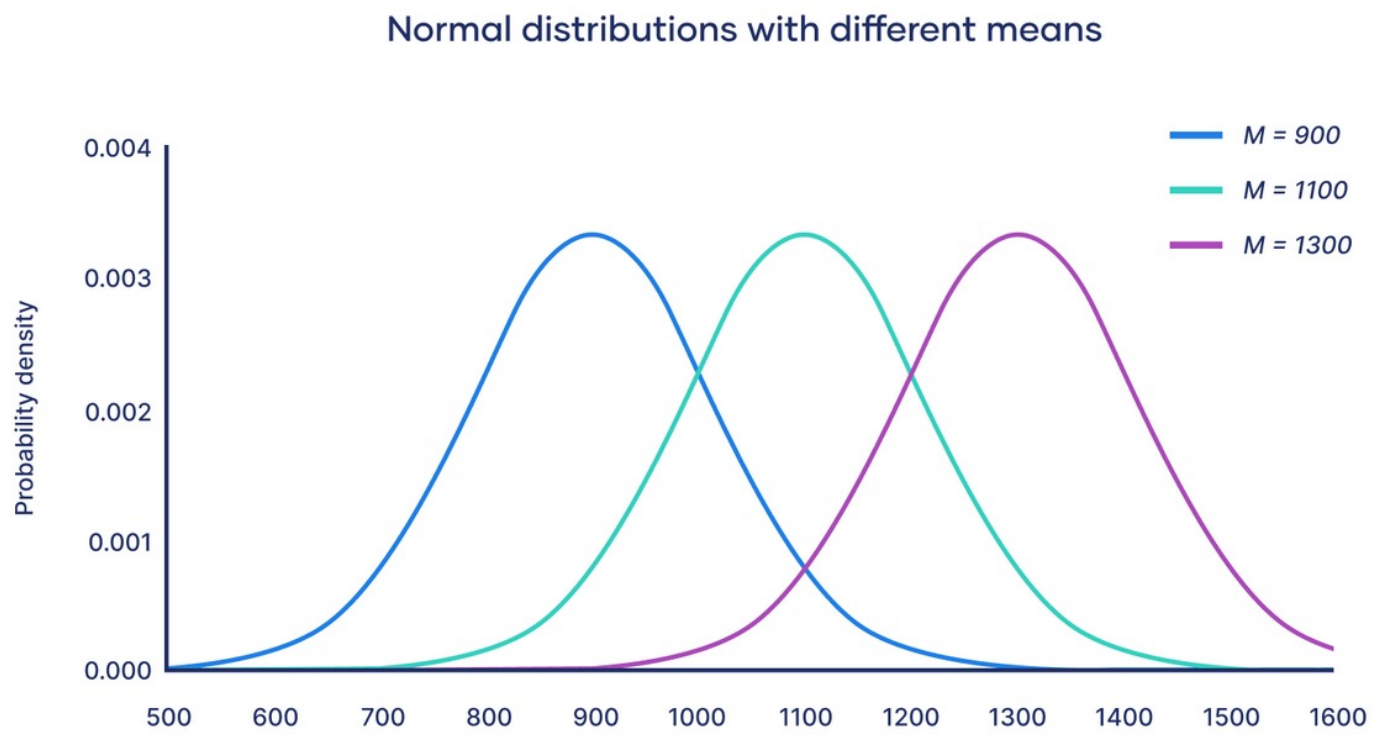
- Normal distributions resemble a bell shape, with most values clustering the central region (mean), tapering off as the scores move further away from the centre.



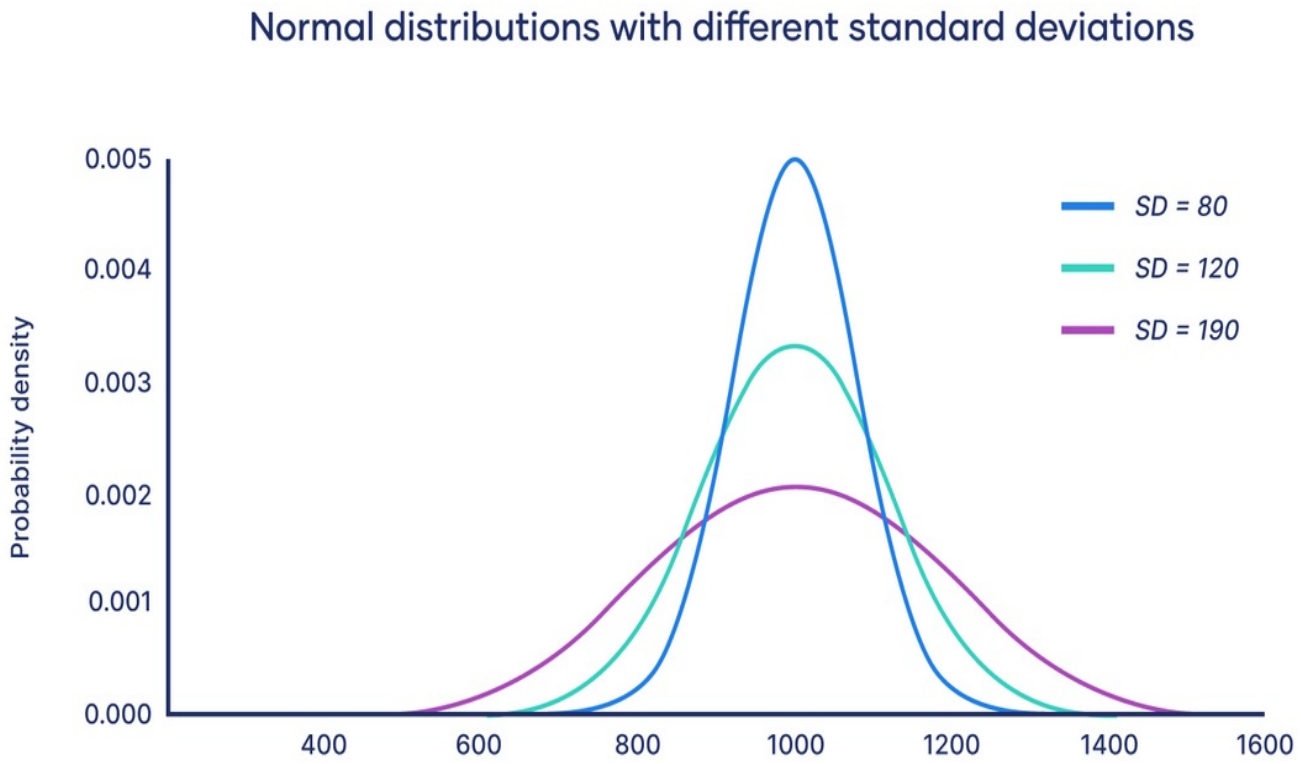
# Normal distribution

The distribution can be described by two values: the mean and the standard deviation.

# Normal distributions with different means



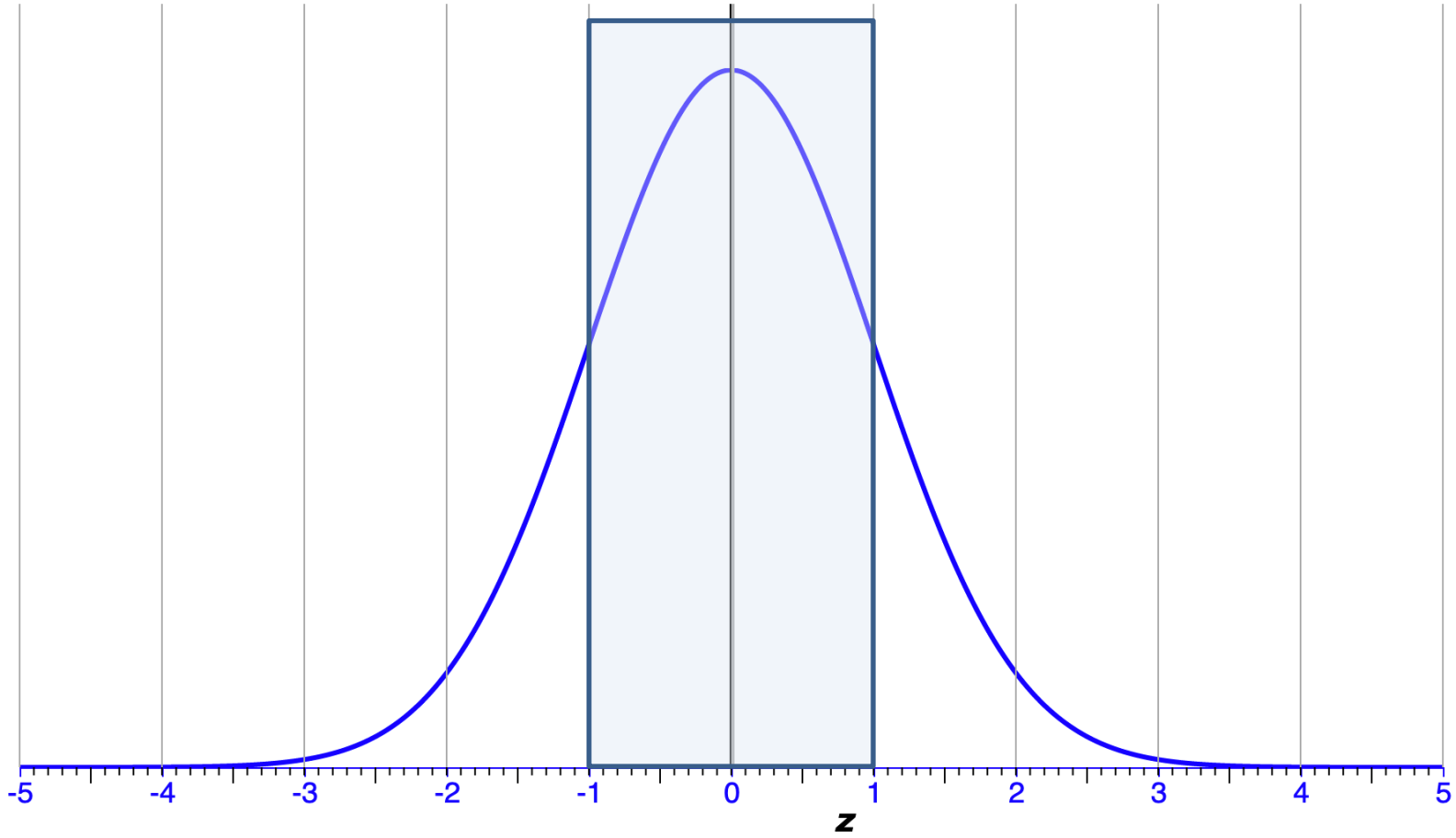
# Normal distributions with different standard deviations



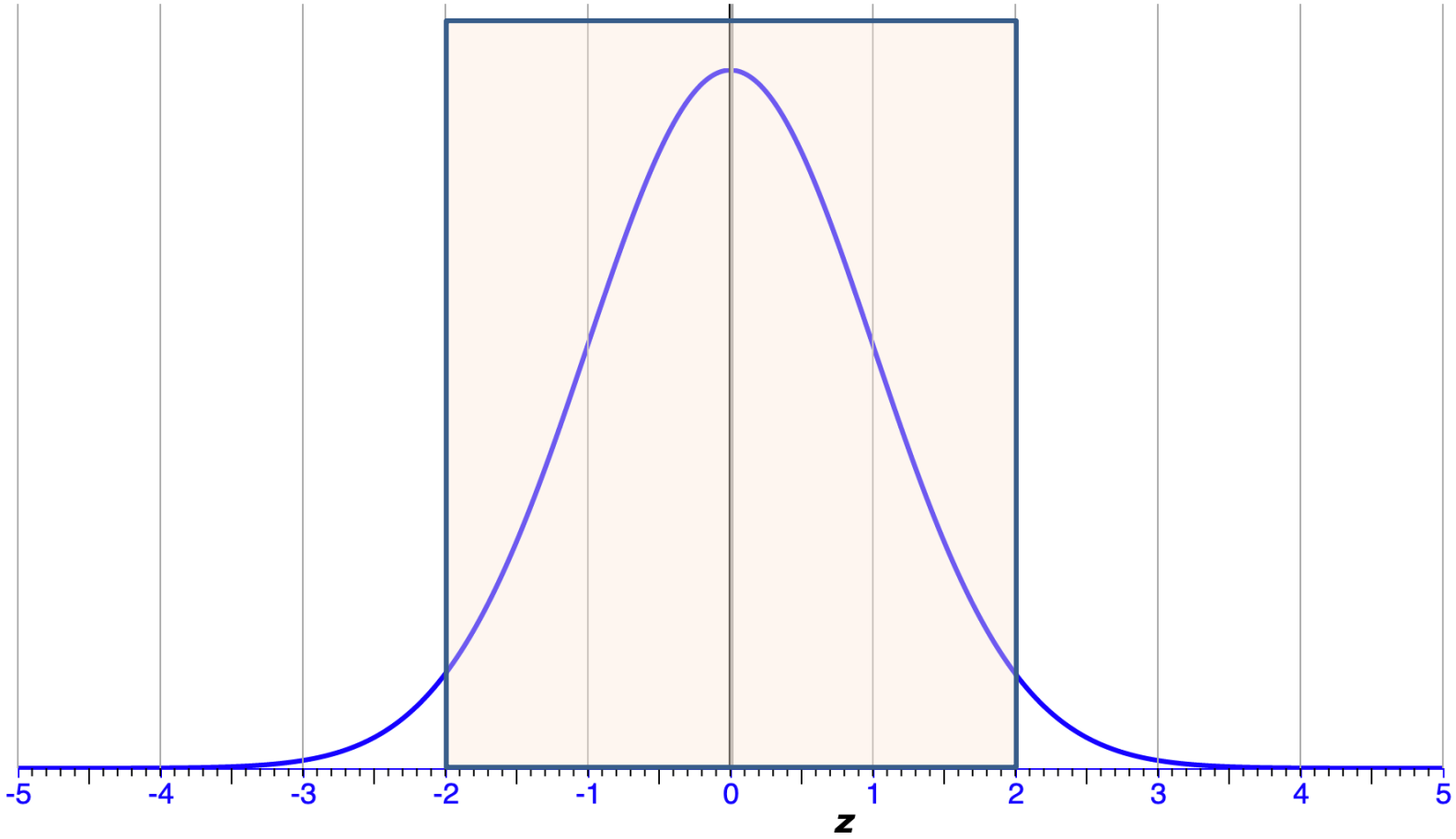
# Empirical rule
# (also known as 68-95-99.7 rule)

This rule tells us where most of your values lie in a normal distribution.

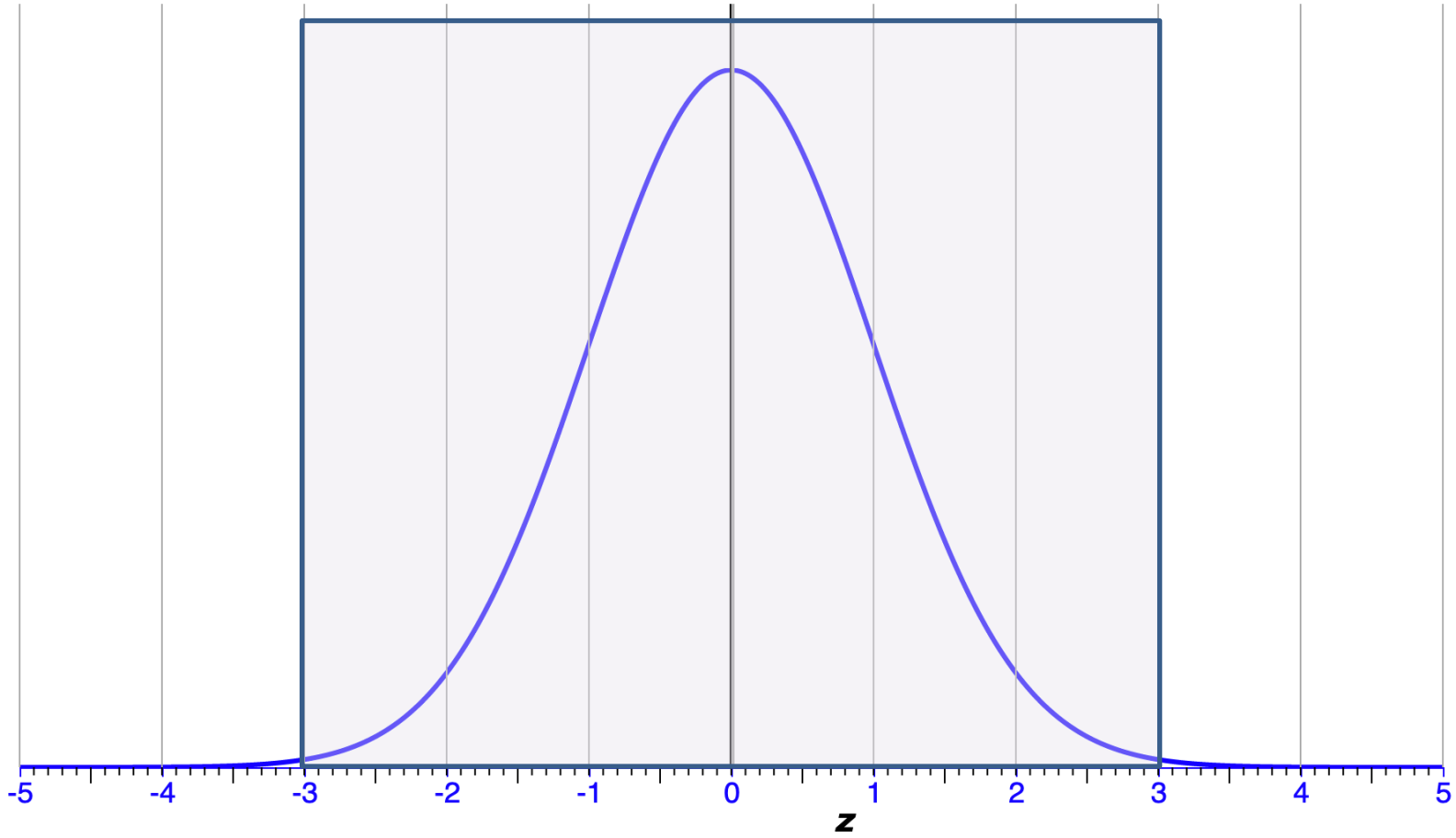
# Around 68% of values are within 1 SD from the mean.



# Around 95% of values are within 2 SDs from the mean.

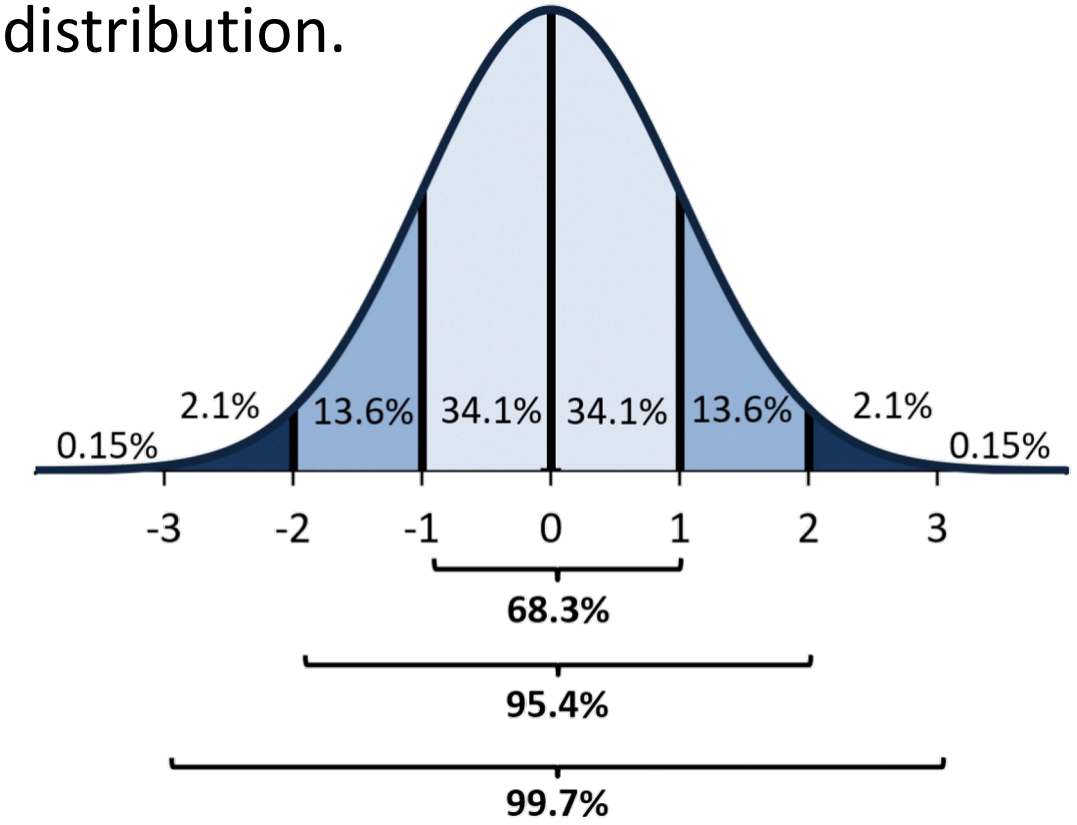


# Around 99.7% of values are within 3 SDs from the mean.



# Empirical rule (also known as 68-95-99.7 rule)

This rule tells us where most of your values lie in a normal distribution.



# What is a sampling distribution?

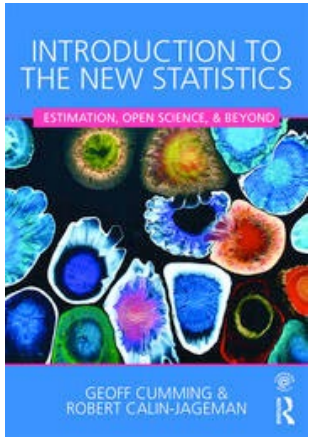
- A sampling distribution is the distribution of sample statistics (e.g., means) collected from many independent samples.
- To build a sampling distribution, we draw a sample from the population, calculate a statistic (e.g., mean) and record it.
- We then repeat this again and again.

# Handout task 1:
# What is a sampling distribution?

- No datasets to download this week, we will be generating data using R

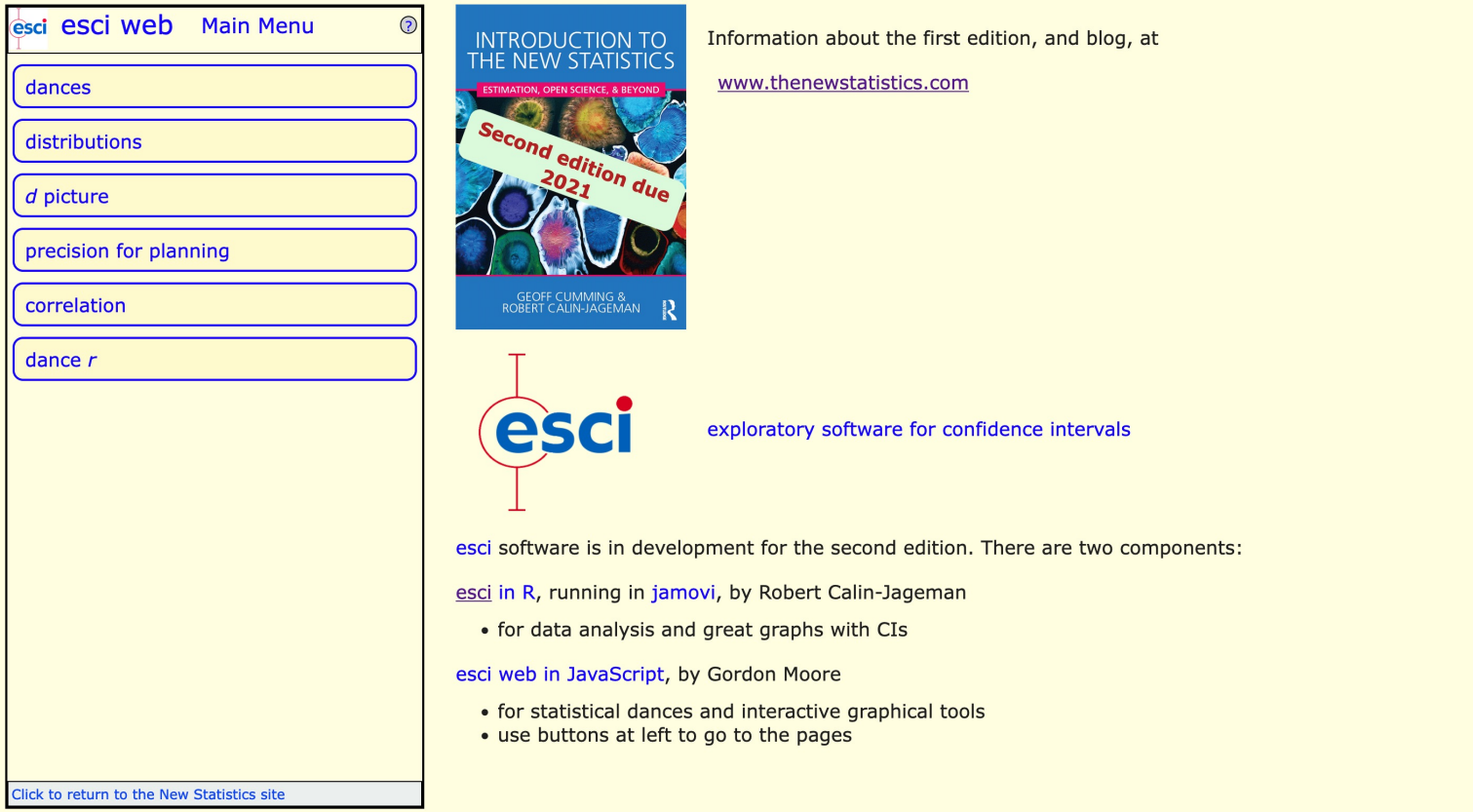
# Handout task 1:
# What is a sampling distribution?

- The psychologist Geoff Cumming wrote an excellent textbook on The New Statistics.
- Cumming advocates a significant reform of the way we use statistics in the social sciences.
- See also: Cumming, G. (2014). The New Statistics: Why and How. *Psychological Science, 25*(1), 7–29.

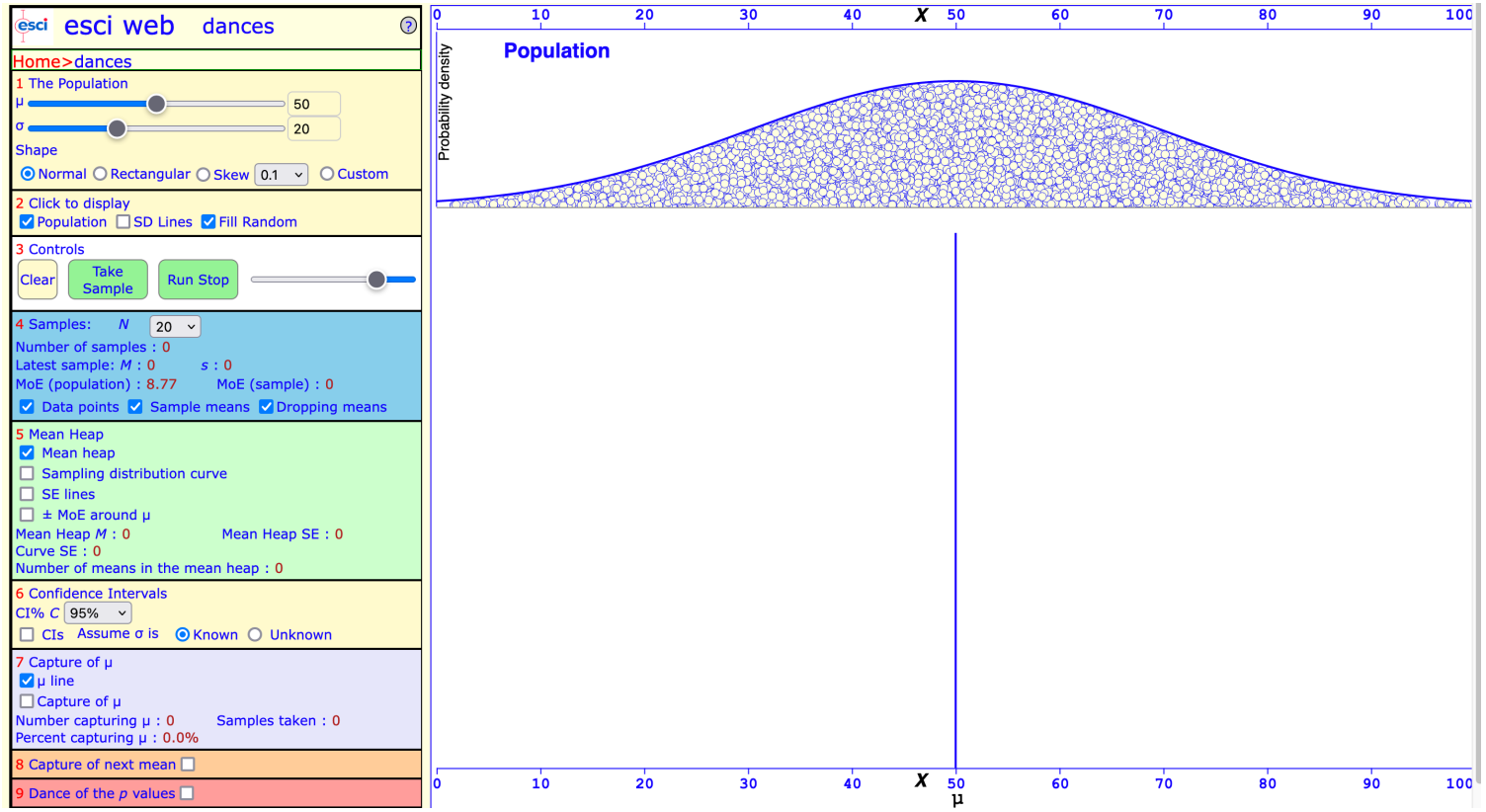


# ESCI: Exploratory Software for Confidence Intervals

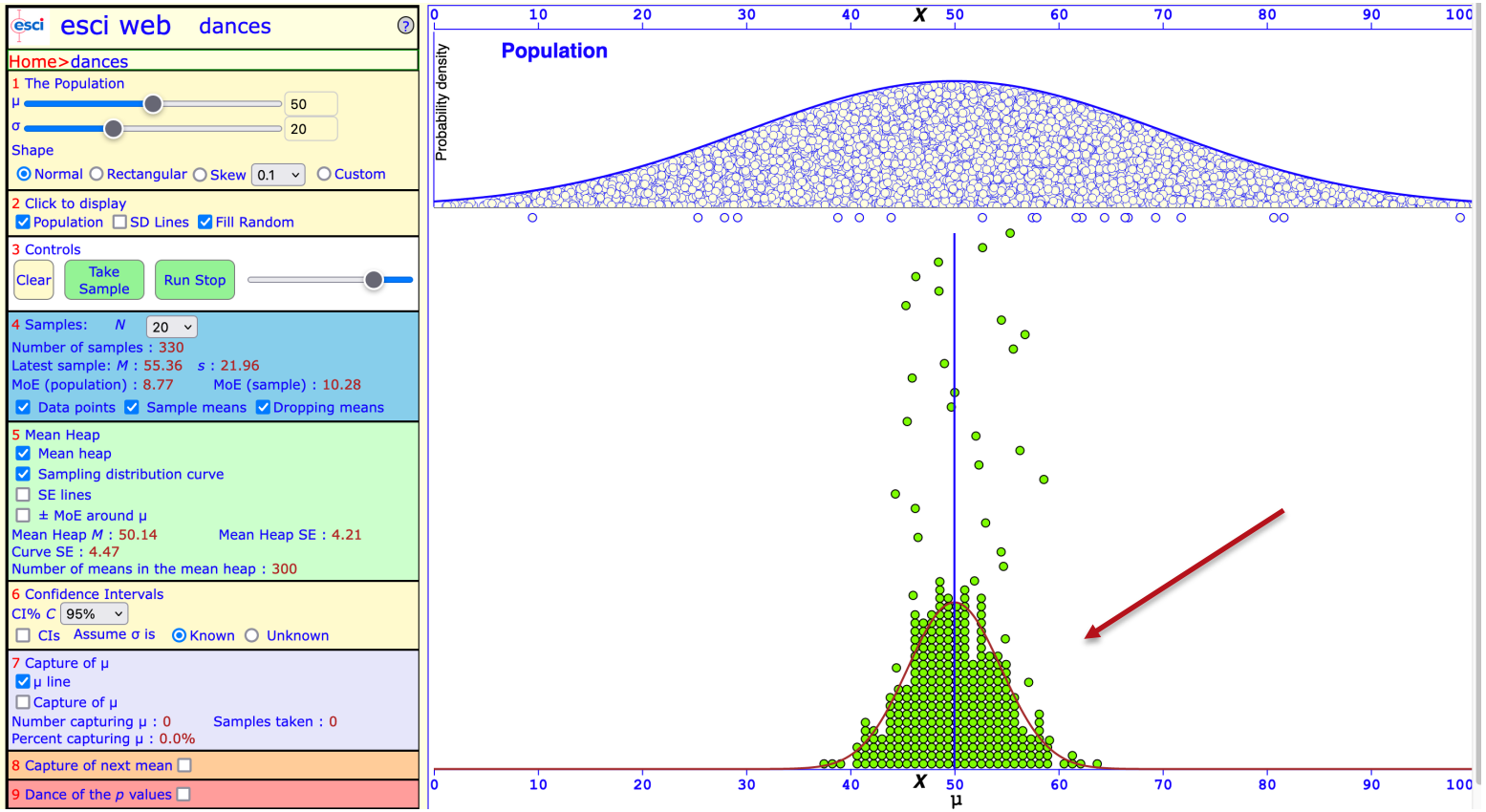
https://www.esci.thenewstatistics.com/



# Handout task 1:
# What is a sampling distribution?



# Handout task 1:
# What is a sampling distribution?



# Questions?

# Why do normal distributions matter?

- Normal curve approximates the distribution of many naturally occurring and human-made phenomena (Tijms, 2004).

- Adolphe Quetelet (1835): Showed that human traits were distributed according to a normal curve (height, birth weight, etc.)

# Why do normal distributions matter?

Many other phenomena (shoe size, exam scores, IQ score, stock market prices, etc.) also follow a normal distribution.

Because normally distributed variables are so common, many statistical tests are designed for normally distributed populations.

# Why do normal distributions matter?

Understanding the properties of normal distributions (e.g., the Empirical Rule) means we can use probability theory to compare different groups and make estimates about populations using samples.

# Empirical and theoretical distributions

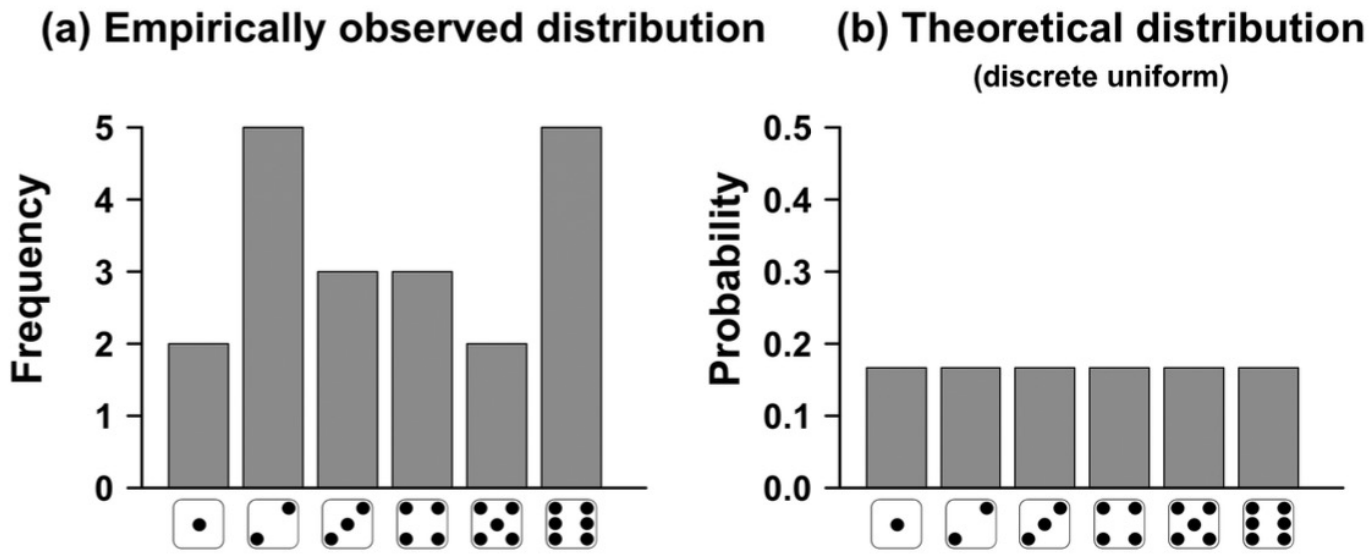



*Figure 3.1.* (a) An empirically observed distribution based on 20 throws of a die; (b) A theoretical distribution displaying the expected probabilities for an infinite number of throws

# Handout task 2

# Handout task 2: Playing around with the normal distribution

What happens when we run the following command five times? Observe the shape of the histogram

```
hist(rnorm(n = 20)) # Five times
```

How does sample size affect the shape of the distribution?

```
hist(rnorm(n = 5, 10, …, 1000))
```

Lancaster University

# Questions?

# Probability theory

# Probability theory

- Remember that, ultimately, we are interested in populations, not the samples drawn from them.
- We take a random sample and calculate a statistic from the sample (e.g., the sample mean).
- We then use probability theory to calculate precisely how closely our sample statistic approximates the true population parameter (e.g., the population mean).

# Probability theory

Two fundamental concepts of probability theory underpin this:

- Law of Large Numbers
- Central Limit Theorem

# The Law of Large Numbers

- A theorem that describes the result of performing the same experiment a large number of times.
- This could mean rolling dice many times, drawing many independent samples from our population of interest, and so forth.

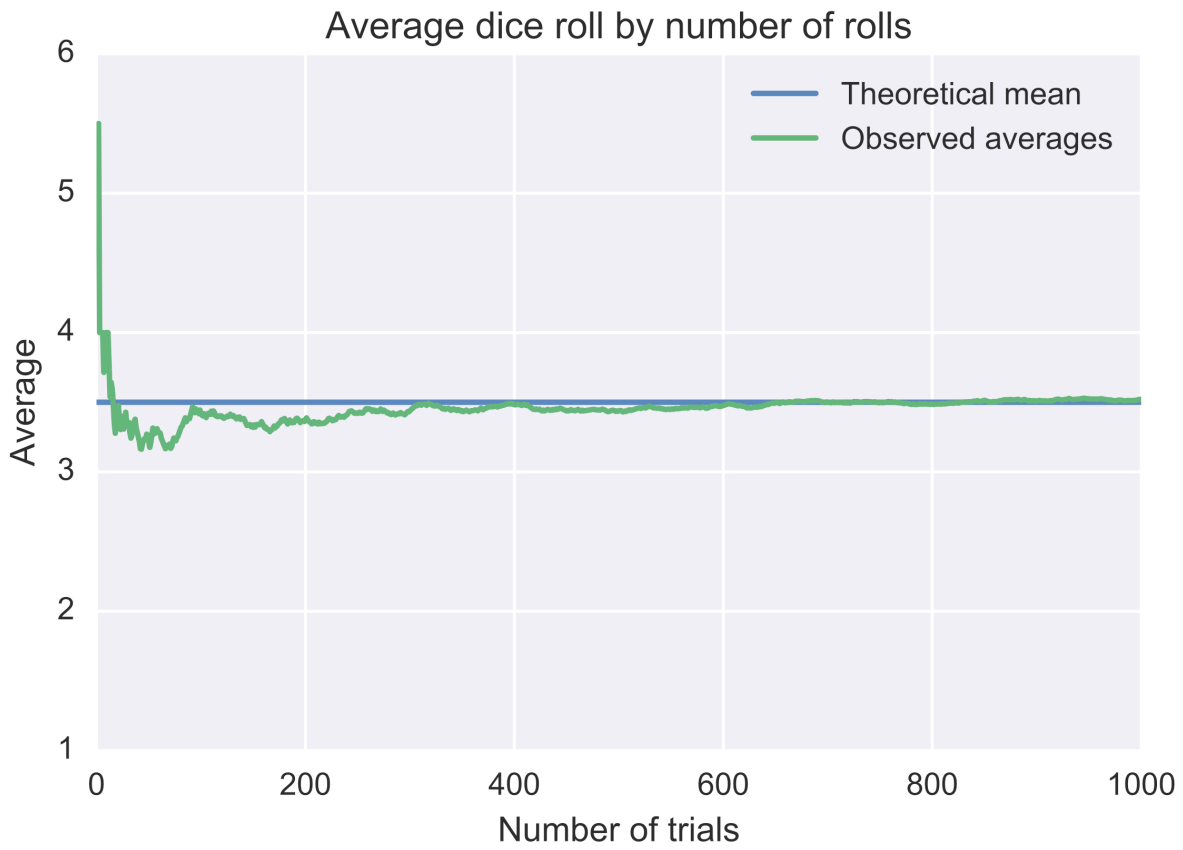

First formal proof by Jakob Bernoulli (1713)

# The LLN states….

We can determine characteristics of a population as long as we draw enough samples.

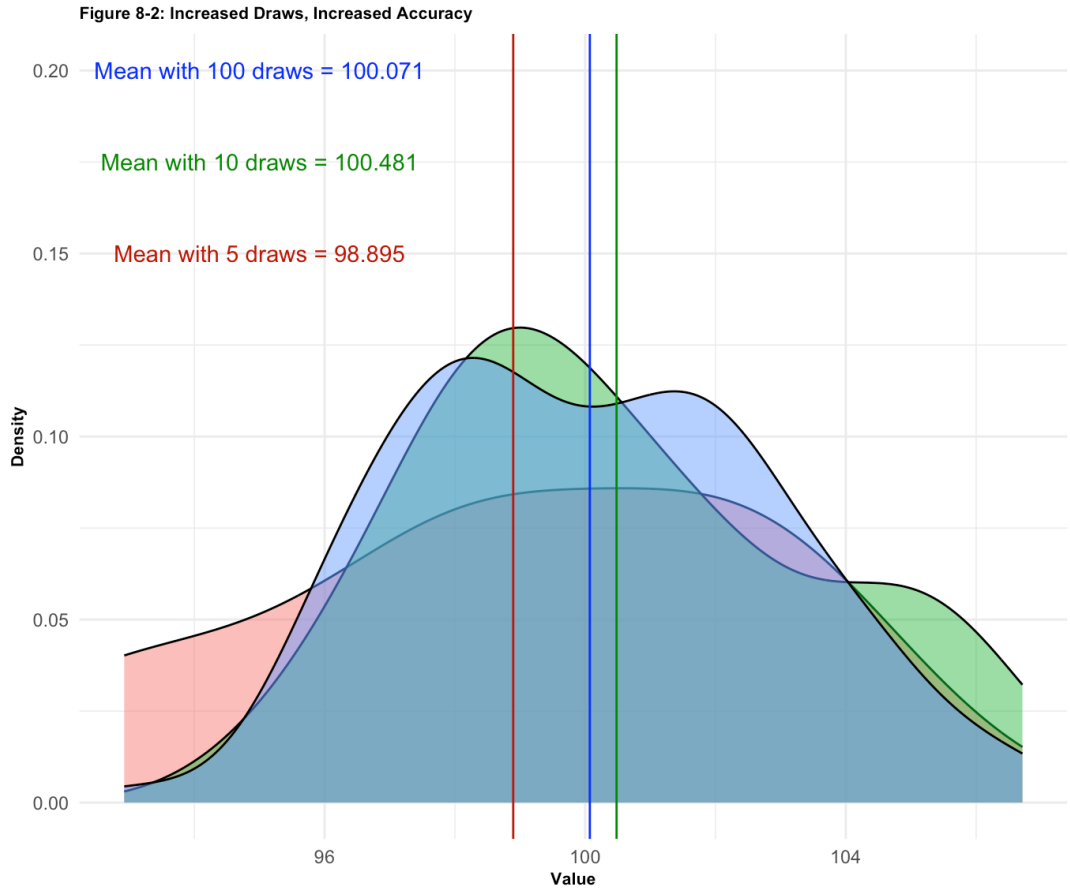
The more random samples we draw from the population, the closer we get to the population parameter.

# What happens to the mean scores when you keep rolling a fair die



# Handout task 3: The LLN in action

How does the number of randomly-drawn samples (5, 10, 100 draws) affect the three sampling distributions and the mean of each?



Lancaster University

# Questions?

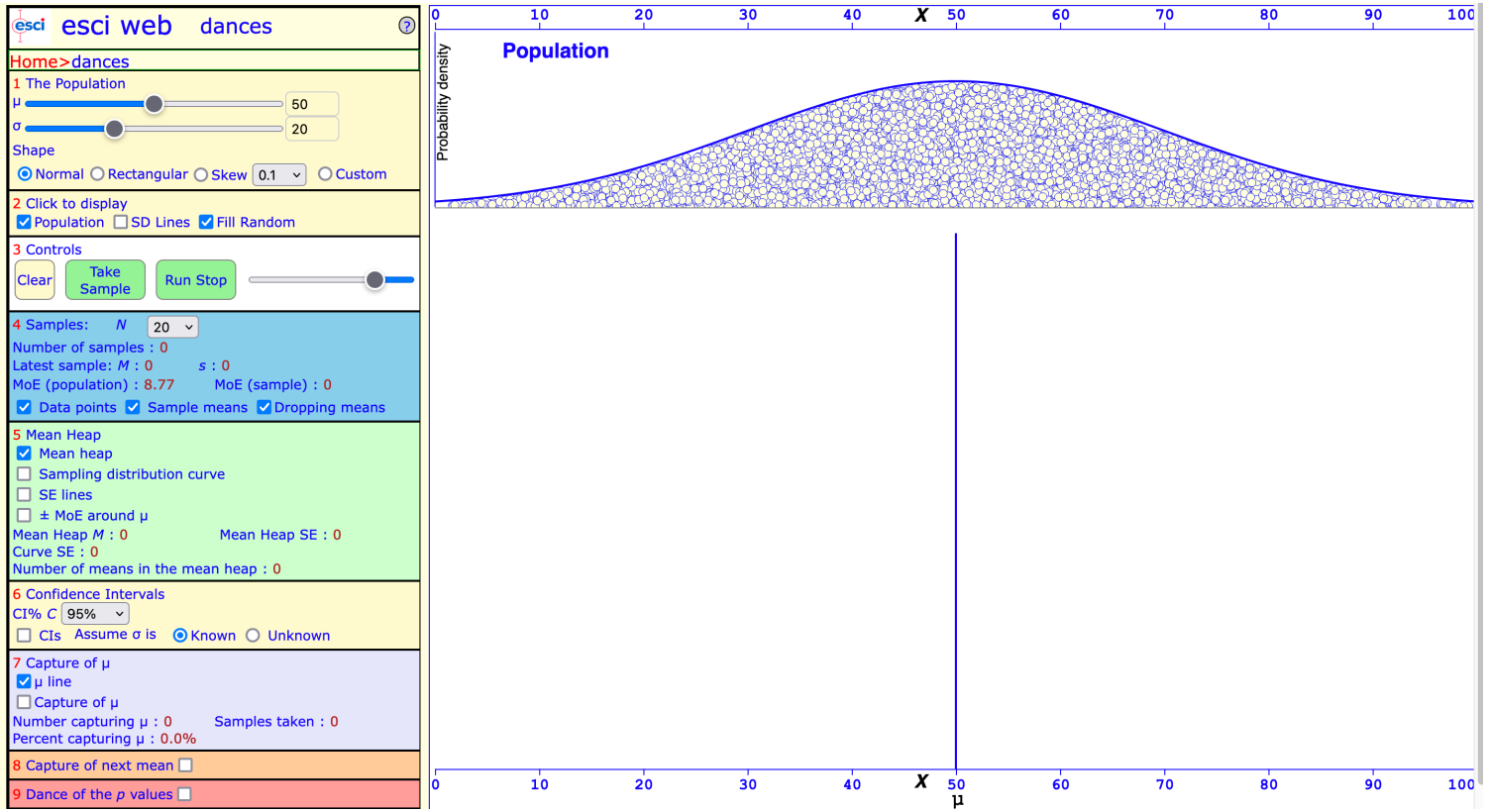
# The Central Limit Theorem

- LLN states: The more you sample, the truer your sample mean is to the average mean of the population.
- But: We cannot draw an infinite number of samples as time and money are finite! Luckily, we don’t have to, and the CLT explains why.

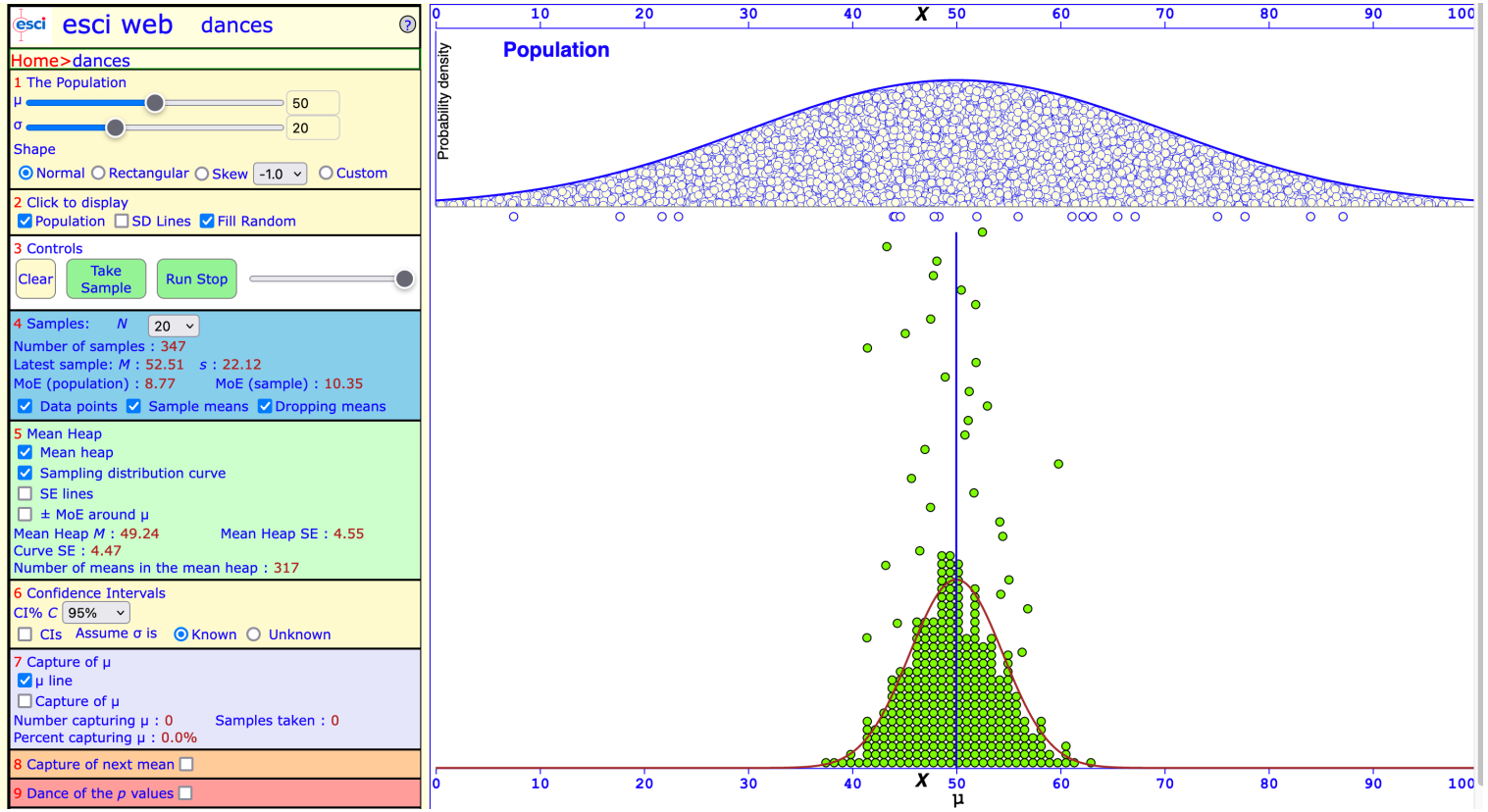
# The CLT states…

- As the number of observations increases in each sample, the shape of the sampling distribution will approximate a normal distribution.
- The sampling distribution of the mean will always be normally distributed, as long as the sample size is large enough.
- This holds true even if the population is not normally distributed!

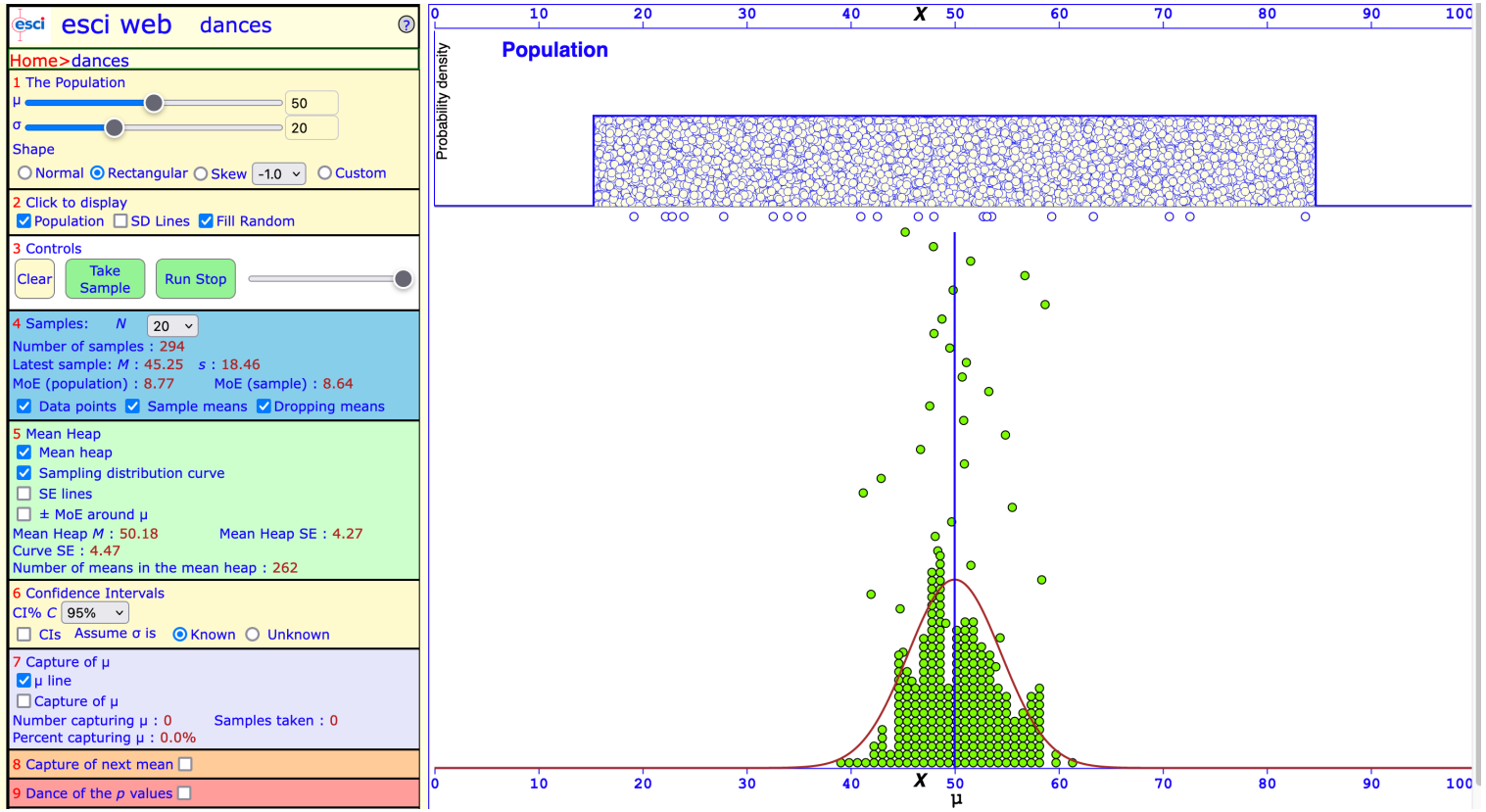
# Handout task 4:
# Let's confirm this!



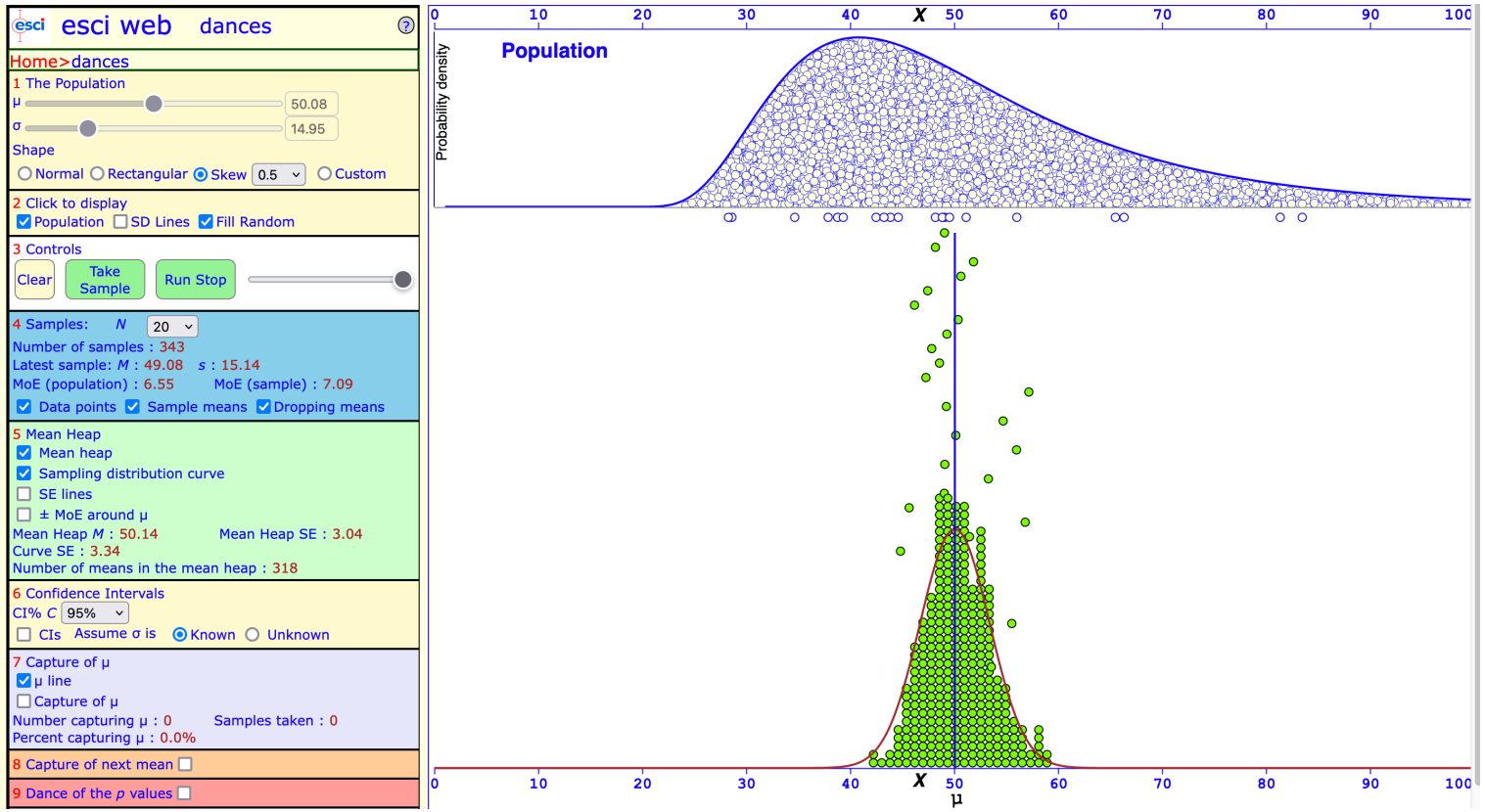
# Population distribution: Normal



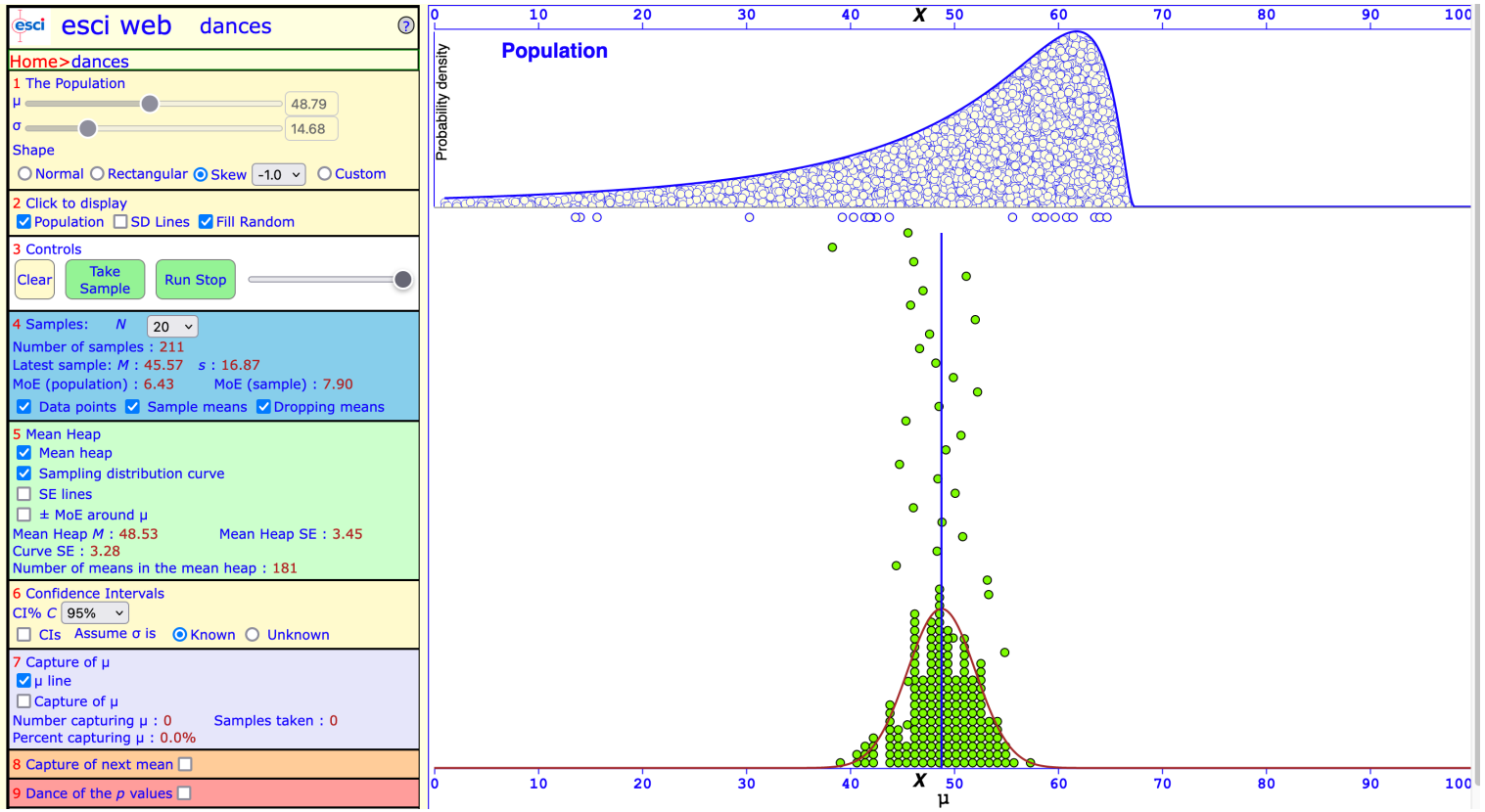
# Population distribution: Uniform (rectangular)



# Population distribution: Positive skew



# Population distribution: Negative skew



Lancaster University

# Questions?

# The shape of sampling distributions approaches the standard normal curve

- This is really useful!
- We can use what we know about normal distributions (e.g., the 68-95 rule) to calculate how far our sample statistic might be from the population parameter.
- We can also obtain a sense of how confident we should be in our estimate.

# Sample size matters

- If we know that the population is normally distributed, we don't need a lot of observations in our samples.
- In this case, the sampling distribution will approach normality even when the number of observations in our sample is small.
- This is helpful as we need fewer observations in our sample in this case.

# Sample size matters

- However, if the population distribution is not normally distributed (or if we simply don't know the distribution of the population), we need to have more observations.

# Sample size matters

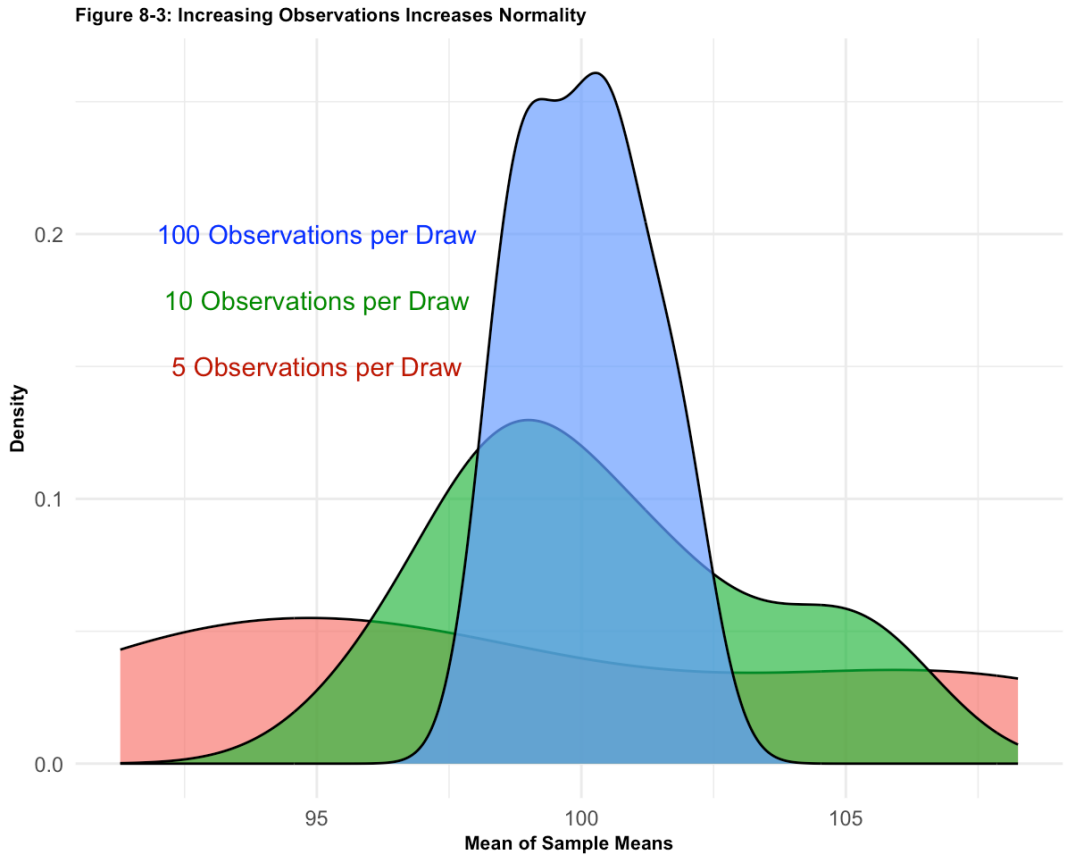
Increasing the number of observations in each sample increases the accuracy of our estimate.

The more information we have, the more accurate our predictions!

We can determine the necessary sample using statistical power - "If I run my experiment 100 times, how many times will I find a significant result?"

# Handout task 5: The CLT in action

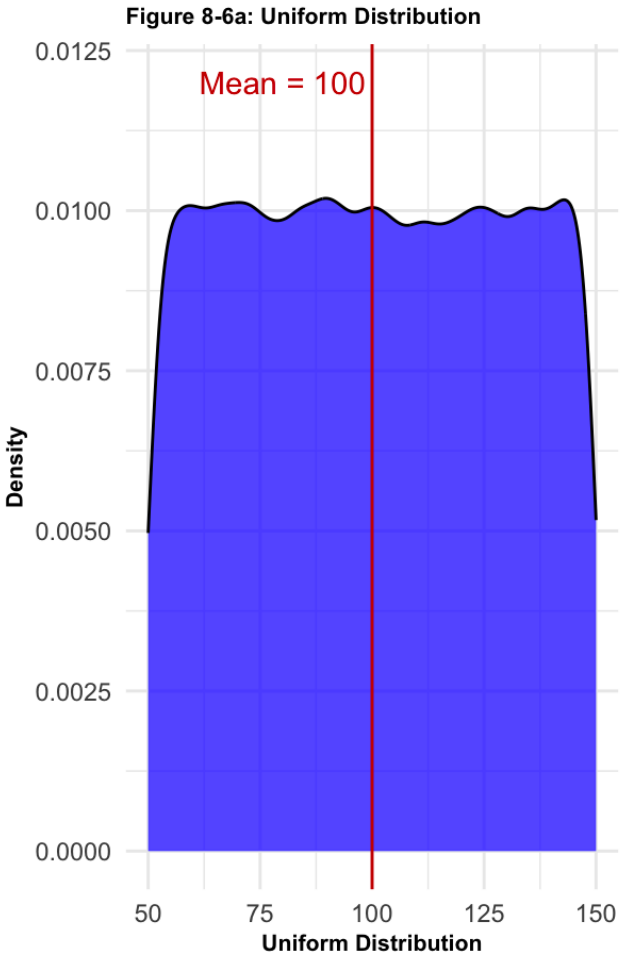
How does changing the number of observations (n = 5, 10, 100) affect the three sampling distributions?

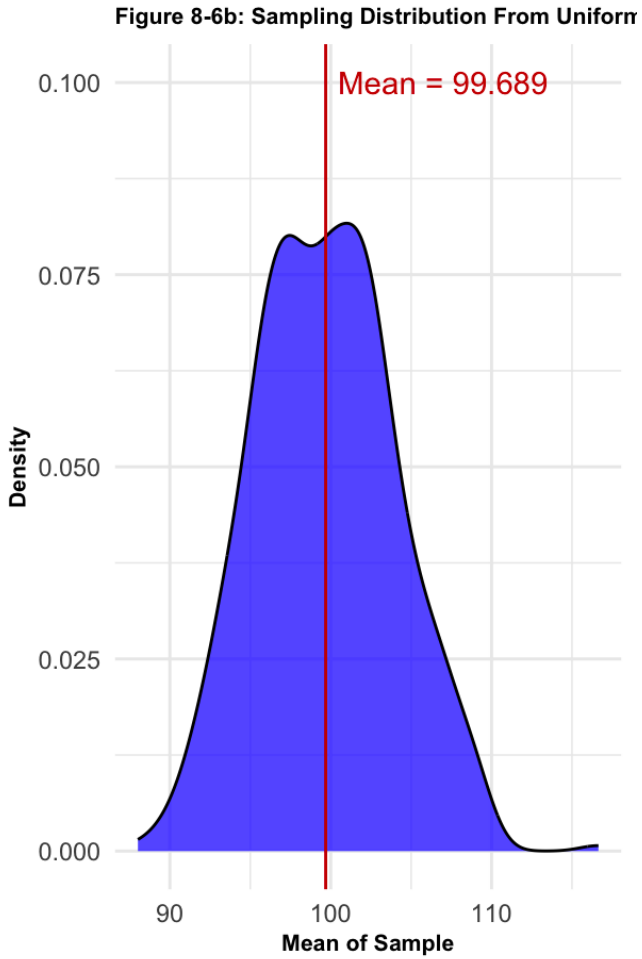


# Sampling from different distributions

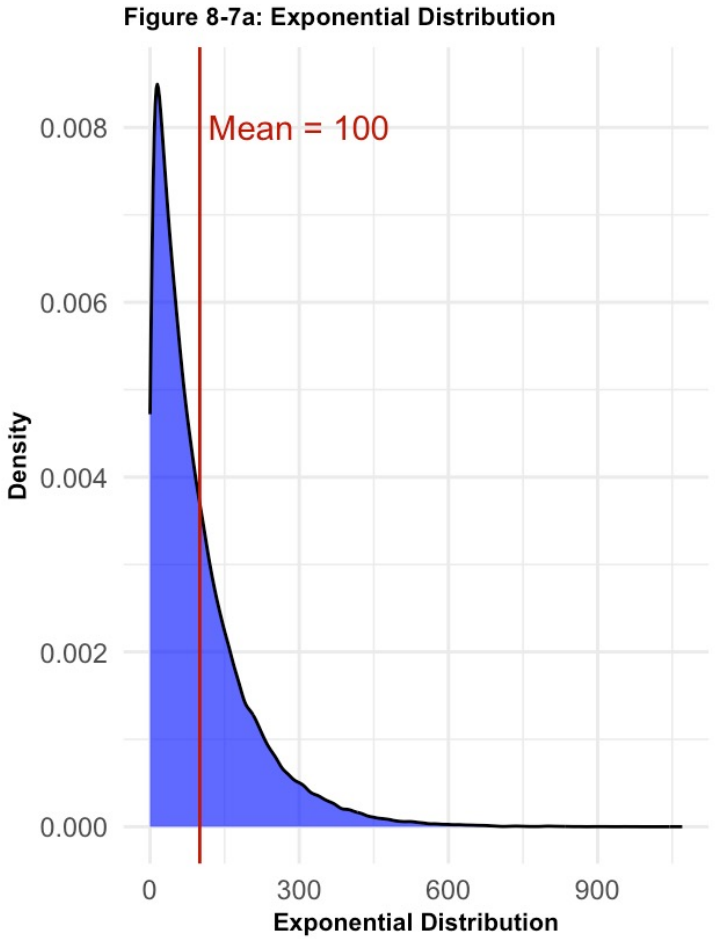
- No matter the shape of our population's distribution, the sampling distribution we construct from it will approximate a symmetrical, normal-looking distribution.

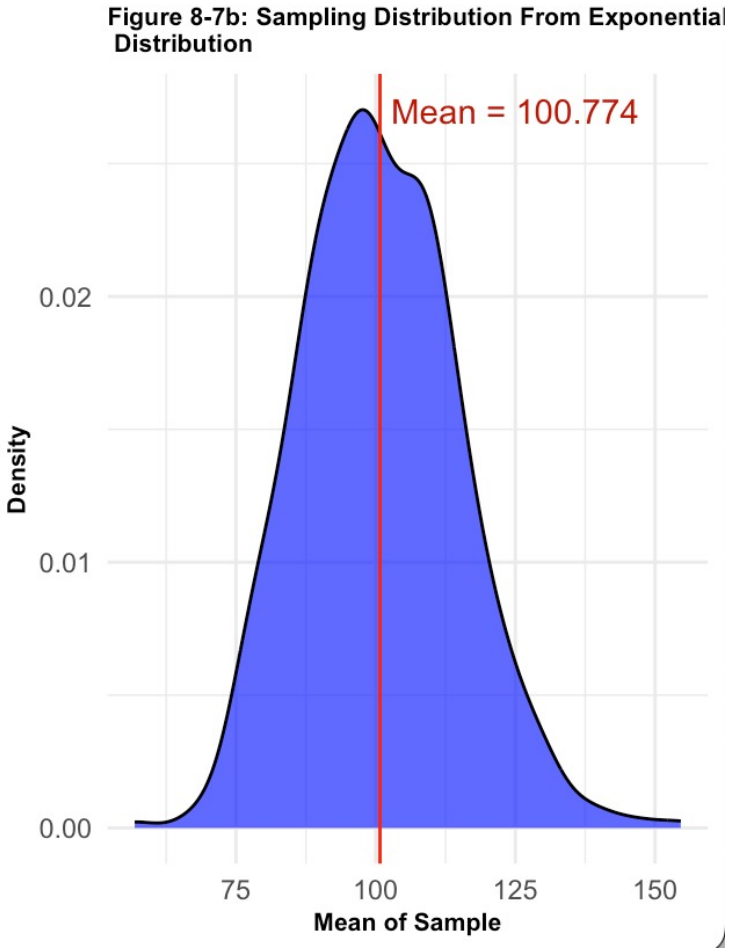
# Sampling from different distributions



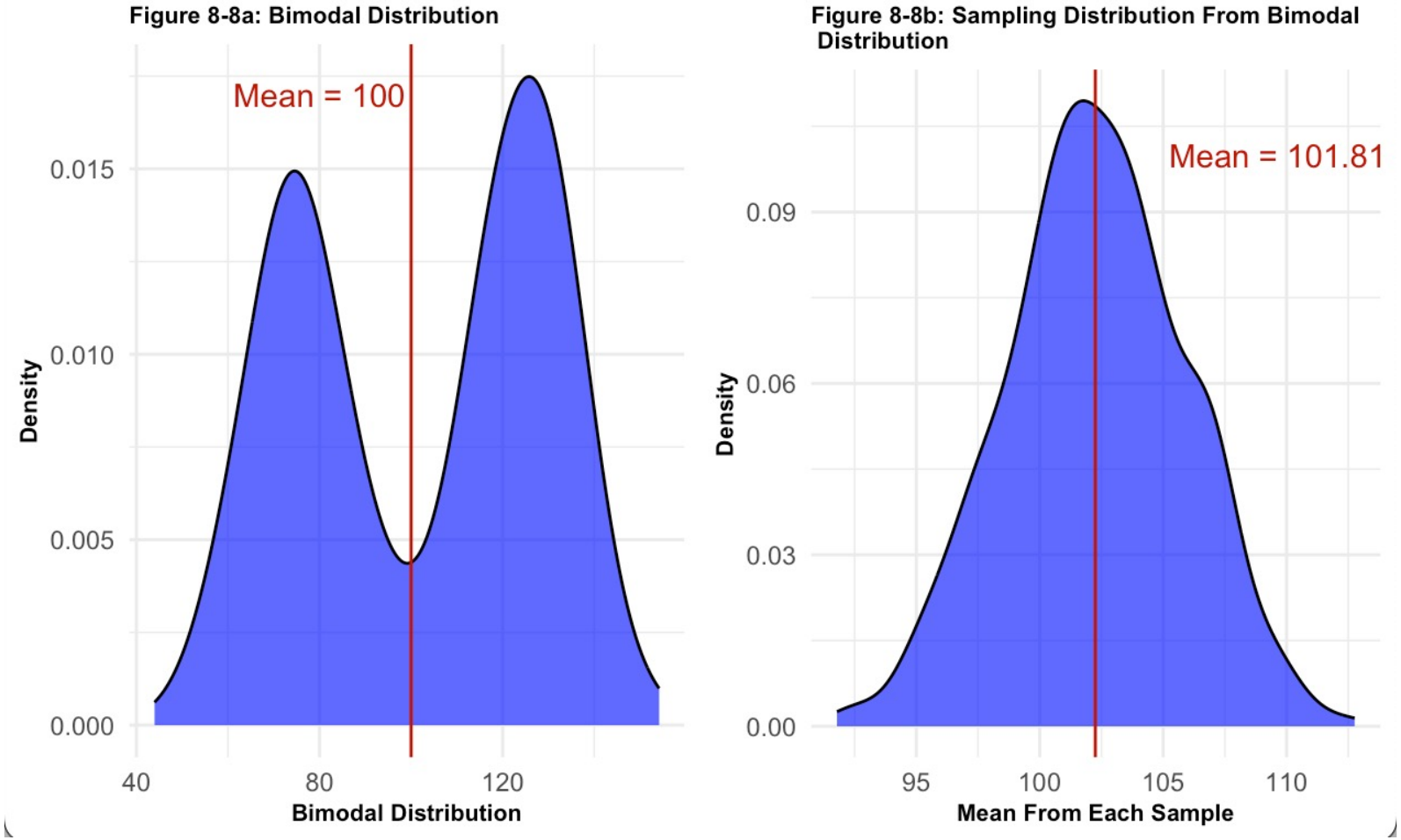



# Handout: Sampling from different distributions



# Handout: Sampling from different distributions



# Compare LLN and CLT

- LLN: If we increase the number of samples (5, 10, 50 samples, each n = 10), then the mean of the sampling distribution will approach the population parameter (the true population mean).

- CLT: If we increase the number of observations in each sample (n = 10, 50, 100), then the shape of the sampling distribution will converge to a normal distribution.

# Questions?

- I will be walking around while you work through the worksheet