Week 11: Samples, Statistics & Sampling Distributions

Data Analysis for Psychology in R 1

Alex Doumas

Department of Psychology
The University of Edinburgh

Course Overview

Exploratory Data Analysis	Research design and data
	Describing categorical data
	Describing continuous data
	Describing relationships
	Functions
Probability	Probability theory
	Probability rules
	Random variables (discrete)
	Random variables (continuous)
	Sampling

Foundations of inference	Confidence intervals
	Hypothesis testing (p-values)
	Hypothesis testing (critical values)
	Hypothesis testing and confidence intervals
	Errors, power, effect size, assumptions
Common hypothesis tests	One sample t-test
	Independent samples t-test
	Paired samples t-test
	Chi-square tests
	Correlation

This Week's Learning Objectives

1. Understand the difference between a population parameter and a sample statistic
2. Understand the concept and construction of sampling distributions
3. Understand the effect of sample size on the sampling distribution
4. Understand how to quantify the variability of a sample statistic and sampling distribution (standard error)

Concepts to carry forward

- Data can be of different types
- We can assign probabilities to outcomes of random experiments
- We can define a probability distribution that describes the probability of all possible events
- Dependent on type (continuous vs. discrete), we can visualise and describe the distribution of data differently

Why are these concepts relevant to psych stats?

- In psychology, we design a study, measure variables, and use these measurements to calculate a value that carries some meaning
  - E.g. the difference in reaction times between groups

- Given it has meaning based on the study design, we want to know something about the value:
  - Is it unusual or not?
  - This is the focus throughout the next semester

- **Today:**
  - We will talk about populations, samples, and sampling
  - Basic concepts of sampling may seem simple and intuitive
  - These concepts will be very useful when we start talking about *statistical inference*, or how we make decisions about data

Populations vs Samples

- In statistics, we often refer to populations and samples
  - **Population:** The entire group of people about whom you'd like to make inferences
  - **Sample:** The subset of the population from whom you will collect data to make these inferences

- To get the most accurate measure of our variable, it would be ideal to collect data from the entire population; however, this is not feasible

- In almost all cases, researchers need to collect data from samples and use these results to make inferences about the population

  - The population value of the variable of interest is known as a **population parameter**
  - The sample value of the variable of interest is known as a **sample statistic**, or **point-estimate**

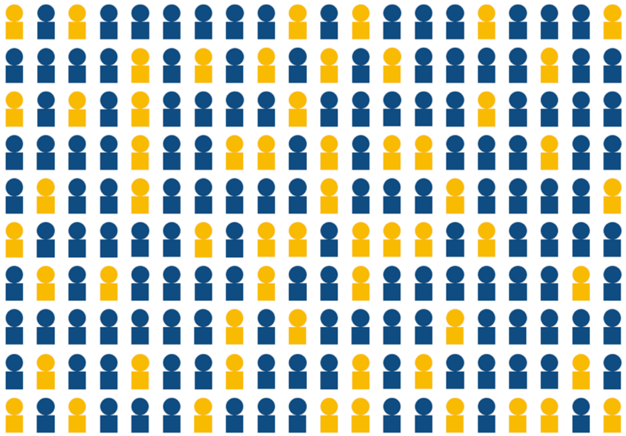
Populations vs Samples - Notation

- It's important to know that although you may have seen these different types of notation used interchangeably in the past, they are actually slightly different when one is referring to a *population* versus a *sample*:

Population	Parameter	Sample
μ	Mean	$\bar{x}$
$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$	Standard Deviation	$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$
N	Size	n

Populations vs Samples - Example

- Suppose I wanted to know the proportion of UG students at the University of Edinburgh who read sci-fi novels



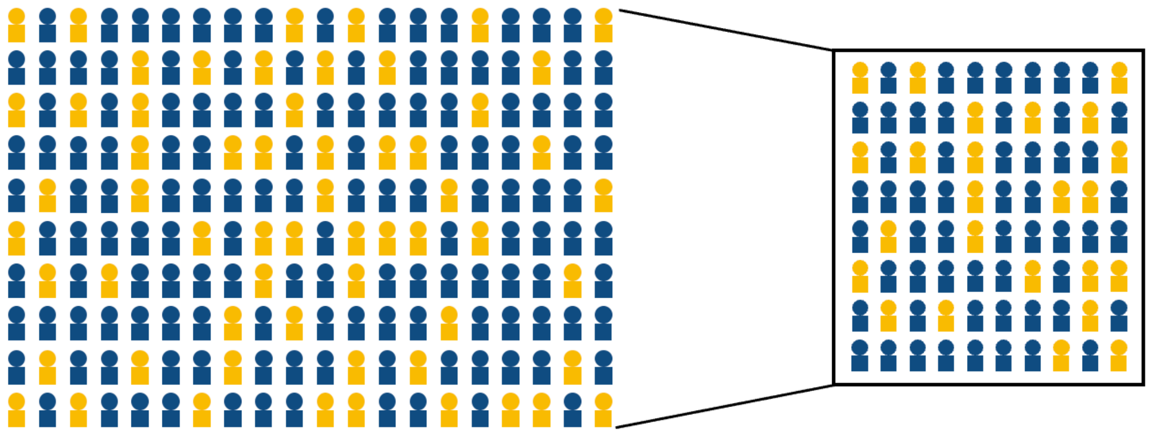
> **Test your Understanding:** What is the population in this example?
>
> What is the variable of interest?
>
> What is the parameter?

Populations vs Samples - Example

- How can we collect this information?
    - We could send out an email requesting all students to tell us if they read sci-fi...but it's not likely that all students will respond
    - We could ask instructors to collect this data from students in their classes, but not every student will attend each class, and not every instructor will comply
- Even with this relatively small, accessible population, it's unlikely we can collect information from every single member

Populations vs Samples - Example

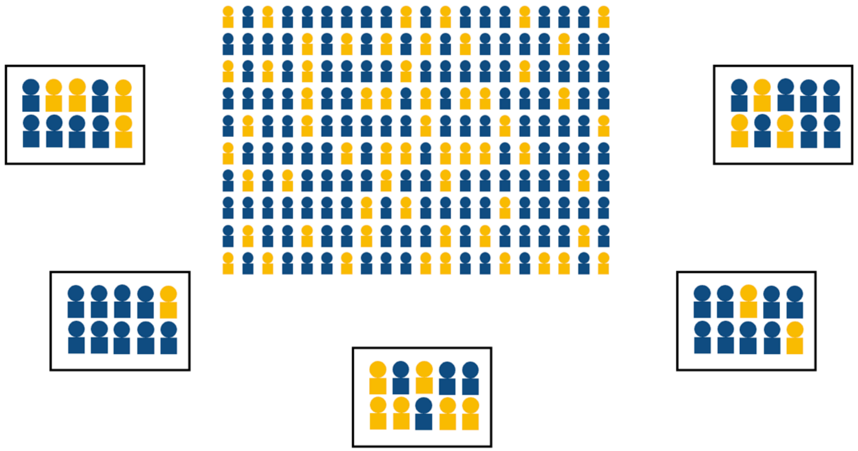
- Instead, we have to use the data from students who *do* respond to make inferences about the overall student population



> **Test your Understanding:** What is the sample?
>
> What is the statistic, or the point-estimate?

Parameters, point-estimates, and sampling distributions

- It is the population parameter (proportion of UoE students who read sci-fi novels) we are interested in: The *true* value of the world
- We can draw a sample, and calculate this proportion in the sample
    - The point-estimate from the sample is our best guess at the population parameter
- If we draw multiple samples, we can produce a **sampling distribution**, which is a probability distribution of some statistic obtained from repeatedly sampling the population



2021/22 actual proportion

- Let's use data from 21/22 to demonstrate this concept
  - These data represent the entire student body from 21/22
- Using these data, we can:

  1) Simulate gathering multiple samples of UoE students

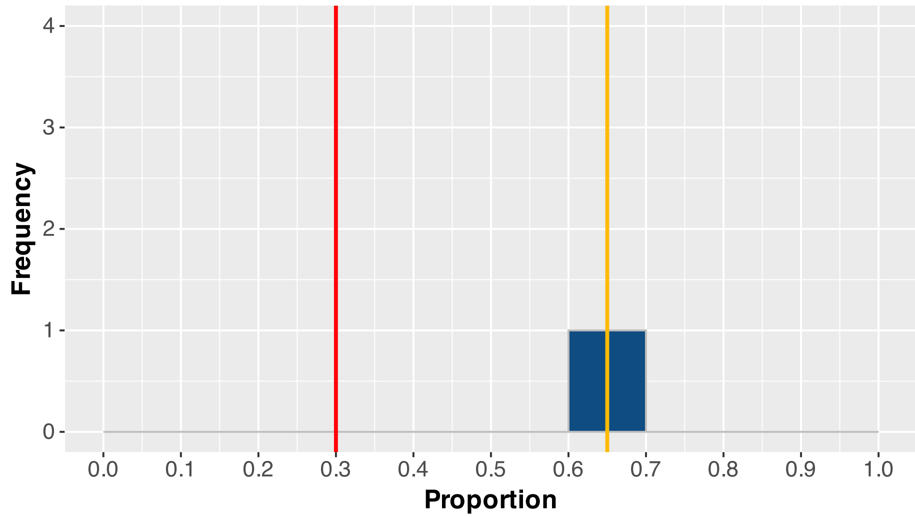
  2) Calculate the proportion students that read sci-fi in each sample

  3) Produce a frequency distribution of each sample's results

Scottish	n	Freq
No	20090	0.7
Yes	8665	0.3

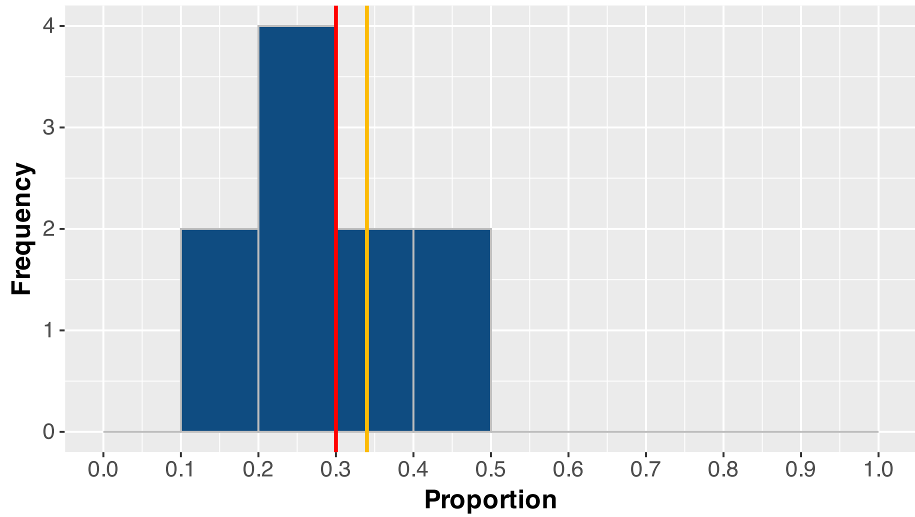
Visualising sampling distributions

- Imagine we took a single sample of 10 students
- This action demonstrates how a statistic from a single small sample may or may not capture the population parameter



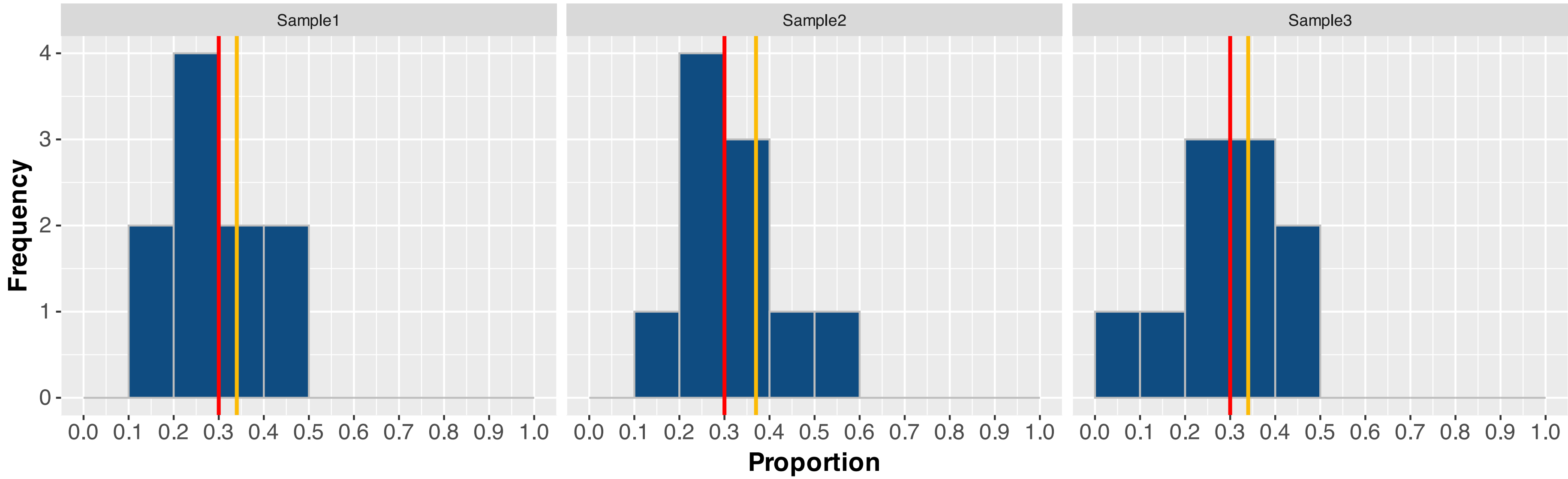
Visualising sampling distributions

- Imagine that we instead took 10 samples of 10 students each
- What happens to the difference between the mean sampling statistic and the population parameter?



Visualising sampling distributions

- If we were to repeat this process 2 more times, we can create three sampling distributions, each of which look different.

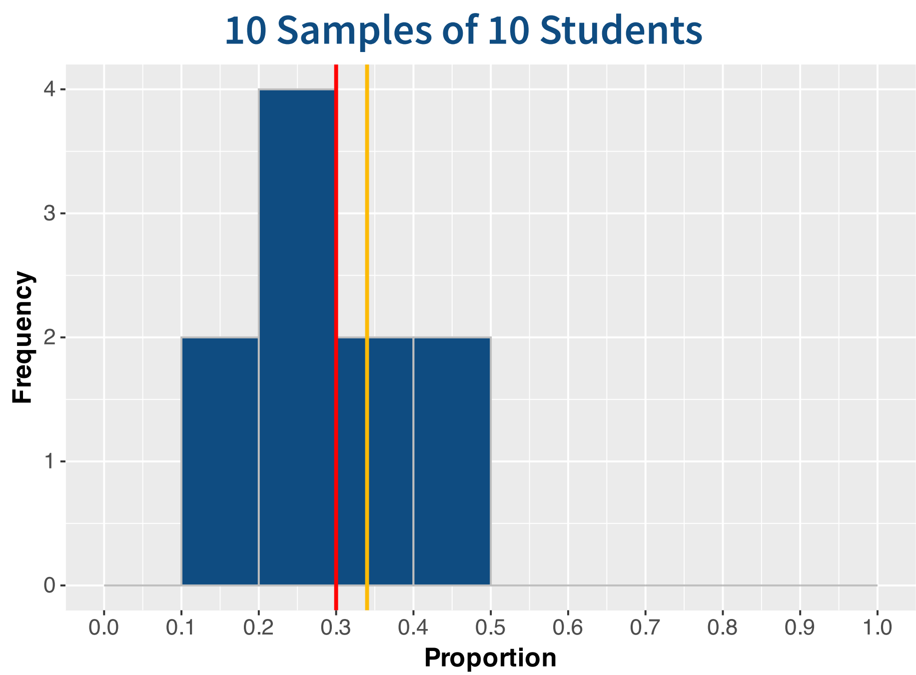


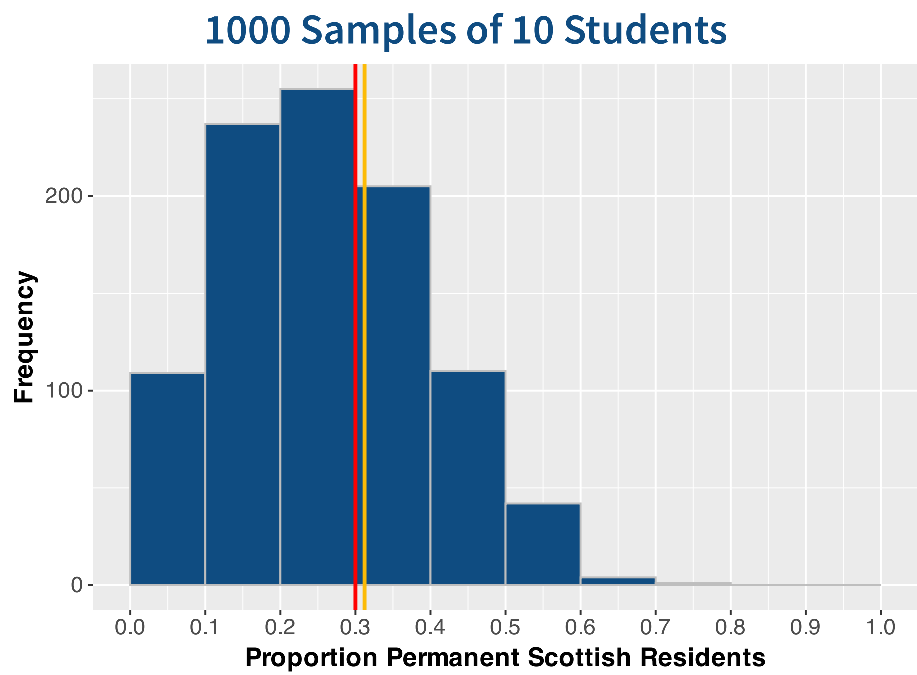
- Each sampling distribution is characterising the *sampling variability* in our estimate of the parameter of interest

- **Do samples with values close to the population value tend to be more or less likely?**

Taking more samples

- So far we have taken 10 samples... what if we took more?
- Let's imagine we sampled 10 students 100 times

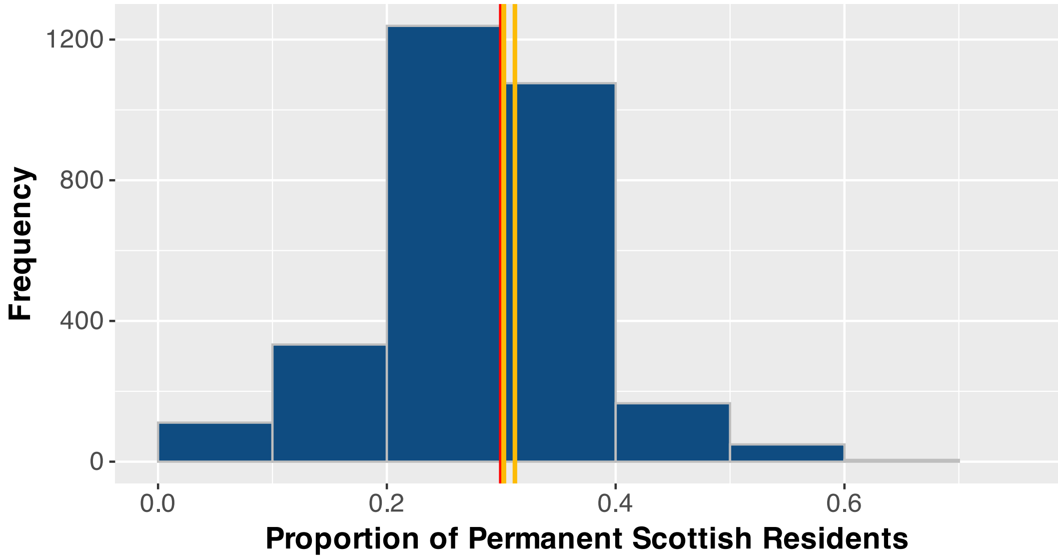




- **What differences do you notice between these two sampling distributions?**

Bigger samples

- We've been taking samples of 10 students. Let's see what happens when we increase our sample size to $n = 50$, and then $n = 100$.



- **What changes as we increase sample size?**

Properties of sampling distributions

- Sampling distributions are characterising the variability in sample estimates
    - Variability can be thought of as the spread in data/plots
- So as we increase n, we get less variable samples (the distribution of sample statistics is more tightly clustered around the population parameter
    - Harder to get an unrepresentative sample as your n increases
- Let's put this phenomenon in the language of probability:
    - As sample n increases, the probability of observing a point-estimate that is a long way from the population parameter (here 0.30) decreases (becomes less probable)
- So when we have large samples, the point-estimates from those samples are likely to be closer to the population value

Standard error

- We can formally calculate the variability of a sampling distribution, or the **standard error**

$$SE = \frac{\sigma}{\sqrt{n}}$$

- This is essentially calculating the standard deviation of the sampling distribution, with a key difference:
  - The standard deviation describes the variability *within* one sample
  - The standard error describes variability *across* multiple samples

- With continuous data, the standard error gives you a sense of how different $\bar{x}$ is likely to be from μ

- In this example, we're working with binomial data (Scottish Residency = Yes or No), so the standard error indicates how greatly a particular sample proportion is likely to differ from the proportion in the population

Properties of sampling distributions

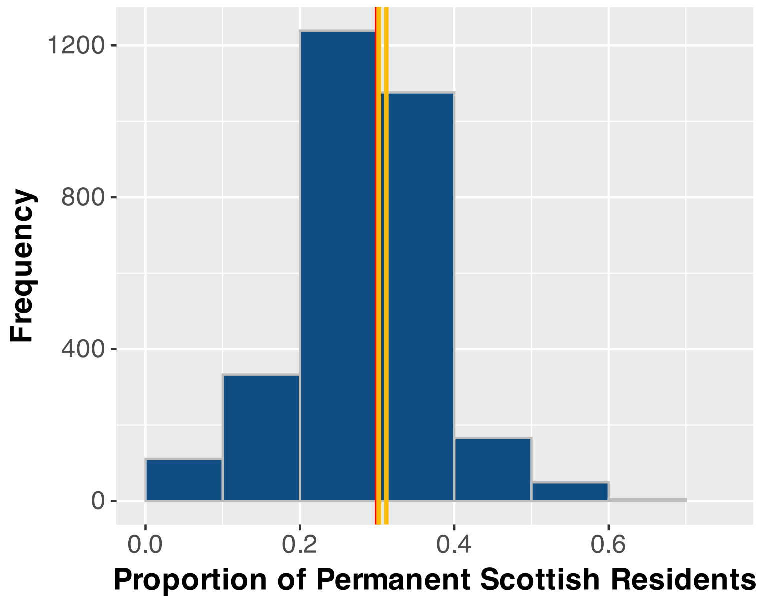
- Mean of the sampling distribution is close to μ, even with a small number of samples

- As the number of samples increases:

    - The sampling distribution approaches a normal distribution
    - Sample $\bar{x}$s pile up around μ

- As n per sample increases, the SE of the sampling distribution decreases (becomes narrower)

    - With large n, all our point-estimates are closer to the population parameter

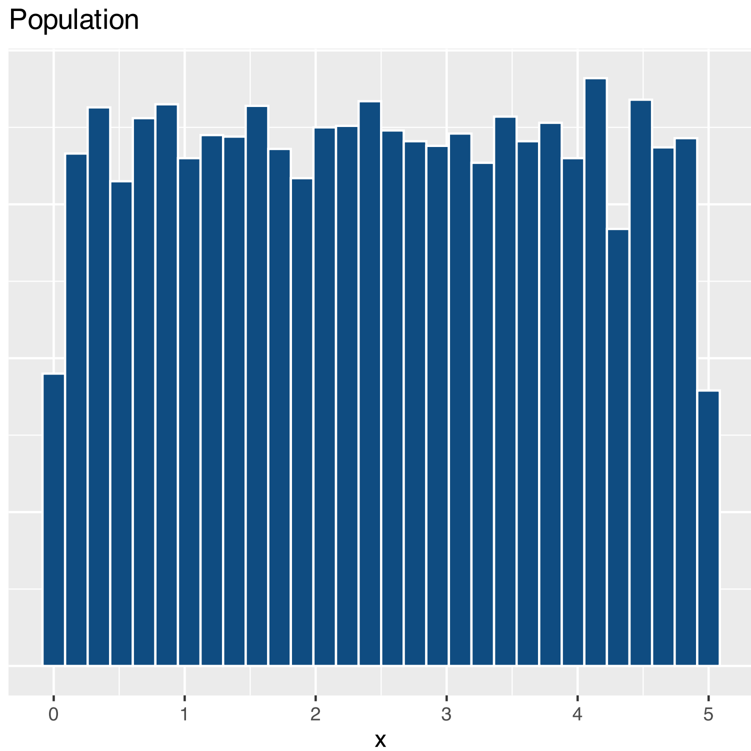


Two Related Concepts

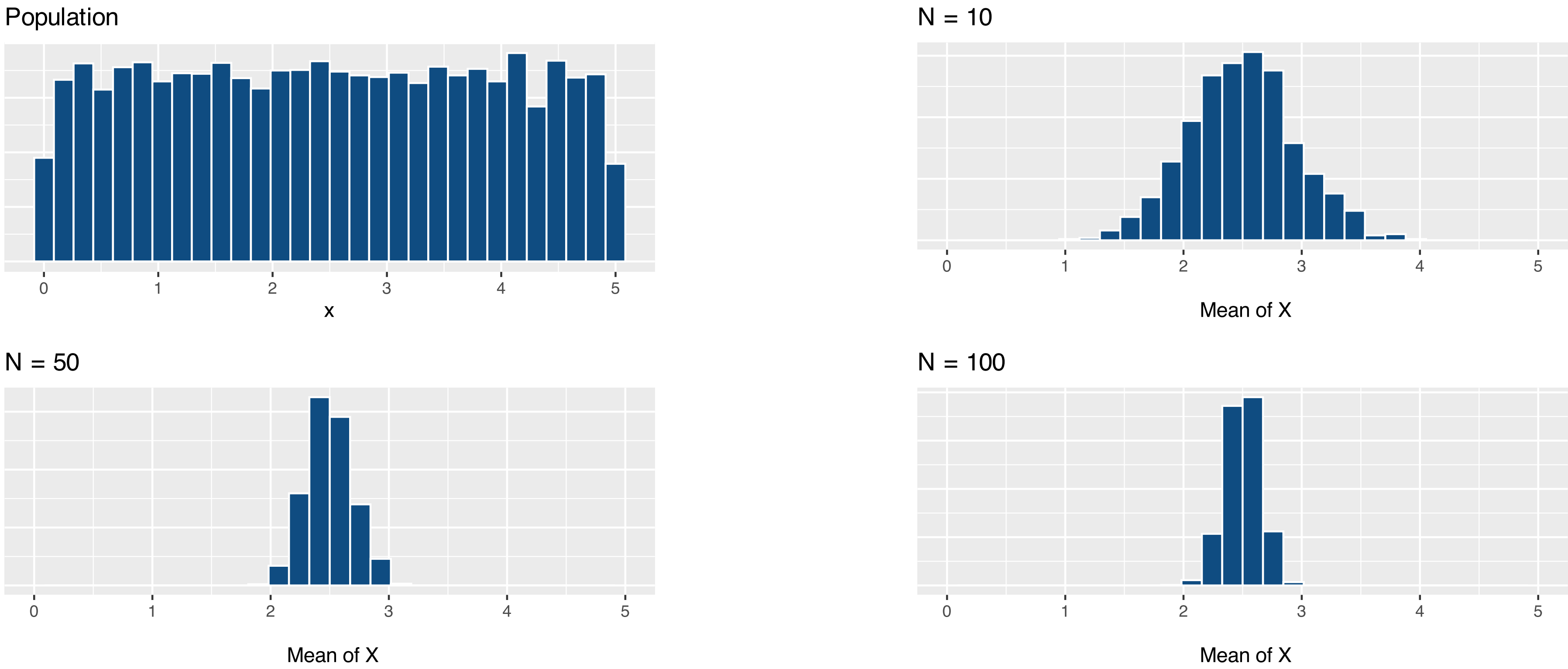
- These properties illustrate two important concepts:
- **The Law of Large Numbers:** As n increases, $\bar{x}$ approaches μ
- **Central Limit Theorem:** When estimates of $\bar{x}$ are based on increasingly large samples (n), the sampling distribution of $\bar{x}$ becomes more normal (symmetric), and narrower (quantified by the standard error)
- These concepts hold regardless of the underlying shape of the distribution
- To demonstrate this, let's explore some different distributions

Uniform distribution

- Continuous probability distribution
- There is an equal probability for all values within a given range

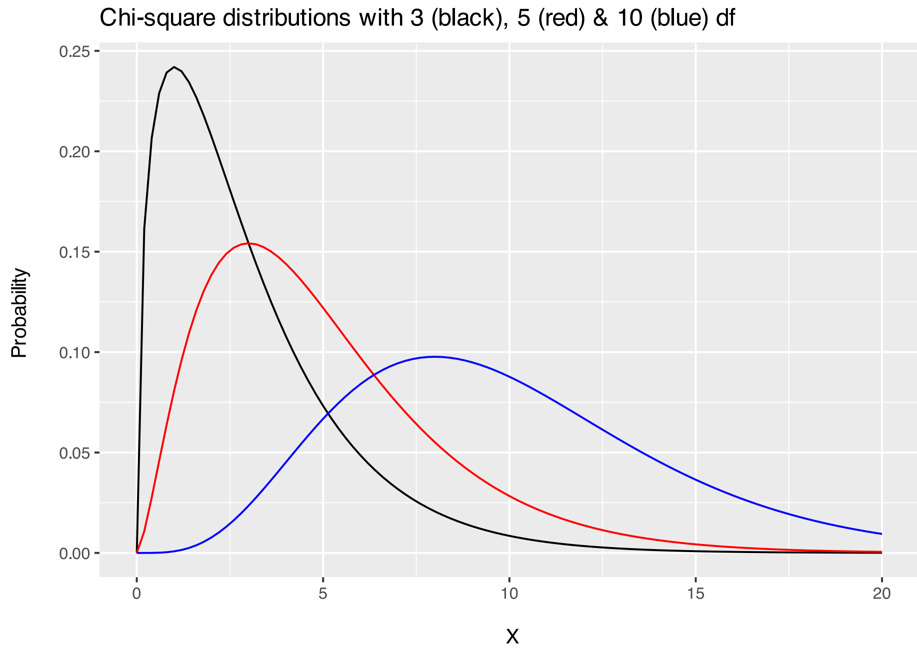


Uniform distribution

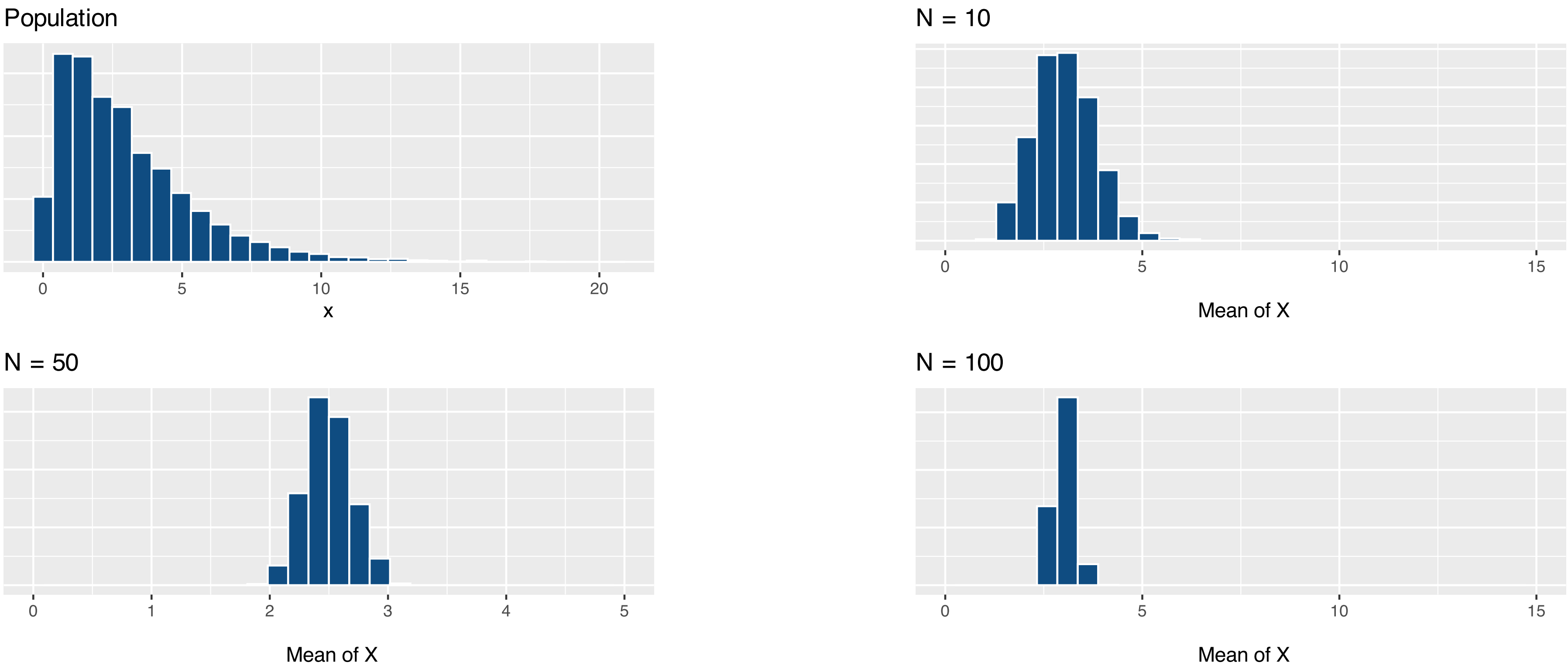


χ-square distribution

- Continuous probability distribution
- Non-symmetric

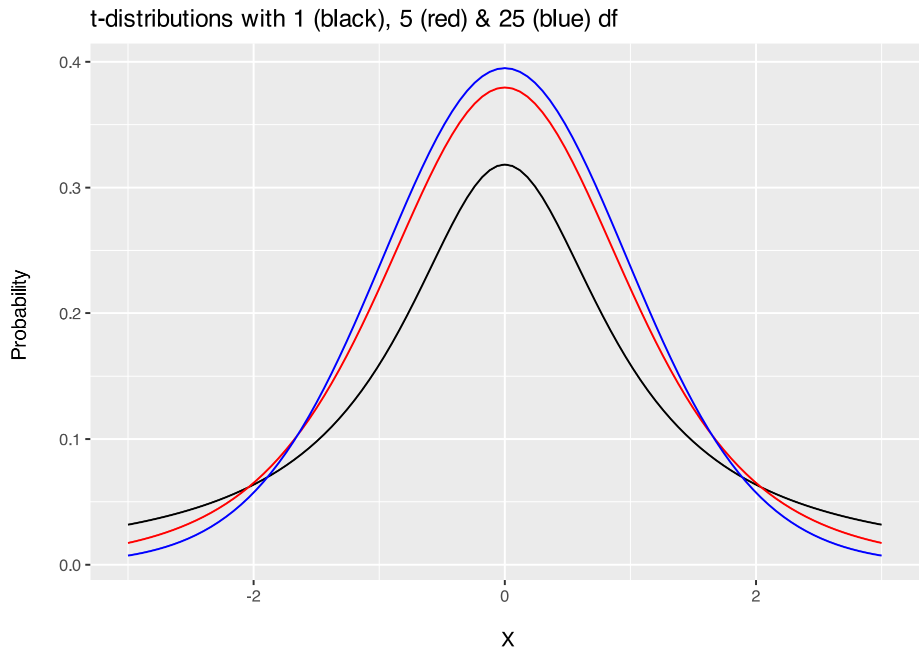


χ -square distribution

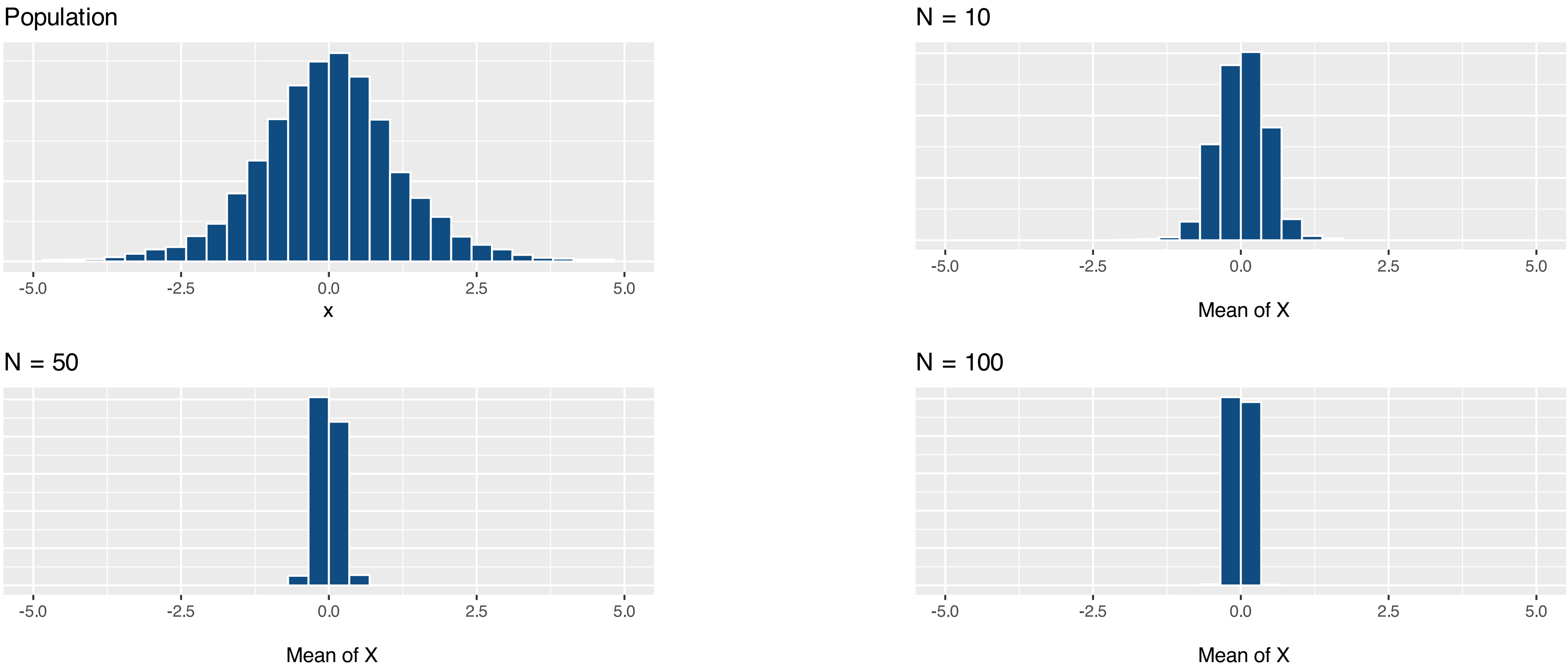


t -distribution

- Continuous probability distribution
- Symmetric and uni-modal (similar to the normal distribution)
  - "Heavier/fatter tails" = greater chance of observing a value further from the mean



t-distribution



Central Limit Theorem

- These examples all demonstrate the Central Limit Theorem
- When n is large enough, $\bar{x}$'s approximate a normal distribution around μ, regardless of the underlying population distribution

Features of samples

- Is our sample...
  - Biased?
  - Representative?
  - Random?
- If a sample of n is drawn at random, it is likely to be unbiased and representative of N
- Our point estimates from such samples will be good guesses at the population parameter

Summary of today

- Samples are used to estimate the population
- Samples provide point estimates of population parameters
- Properties of samples and sampling distributions
- Properties of good samples

This week

Tasks

- Attend both lectures
- Attend your lab and work together on the lab tasks
- Complete the weekly quiz
  - Opened Monday at 9am
  - Closes Sunday at 5pm
- Submit Formative Report B by 12 noon on Friday the 29th of November 2024

Support

- **Office hours**: for one-to-one support on course materials or assessments
  (see LEARN > Course information > Course contacts)
  - Note: No office hours between 2 Dec and 10 Jan
- **Piazza**: help each other on this peer-to-peer discussion forum
- **Student Adviser**: for general support while you are at university
  (find your student adviser on MyEd/Euclid)