Definition – Statistical Inference

In statistics, **inference** is the process of using data from a **sample** to draw conclusions or make predictions about a **population**.

There are two main types of statistical inference:

- **Estimation** – estimating population parameters (e.g., proportion) using sample data.
- **Hypothesis testing** – evaluating a claim about population parameters (e.g., proportion) based on sample data.

As you learned in the previous statistics class, all statistical inference about a **population proportion**—including estimation and hypothesis testing—is based on the **sample proportion**.

In this lecture, you will learn

- how sample proportions vary from one sample to another and
- how these sample proportions are distributed.

Contents

Simulation Experiment for Sample Proportions

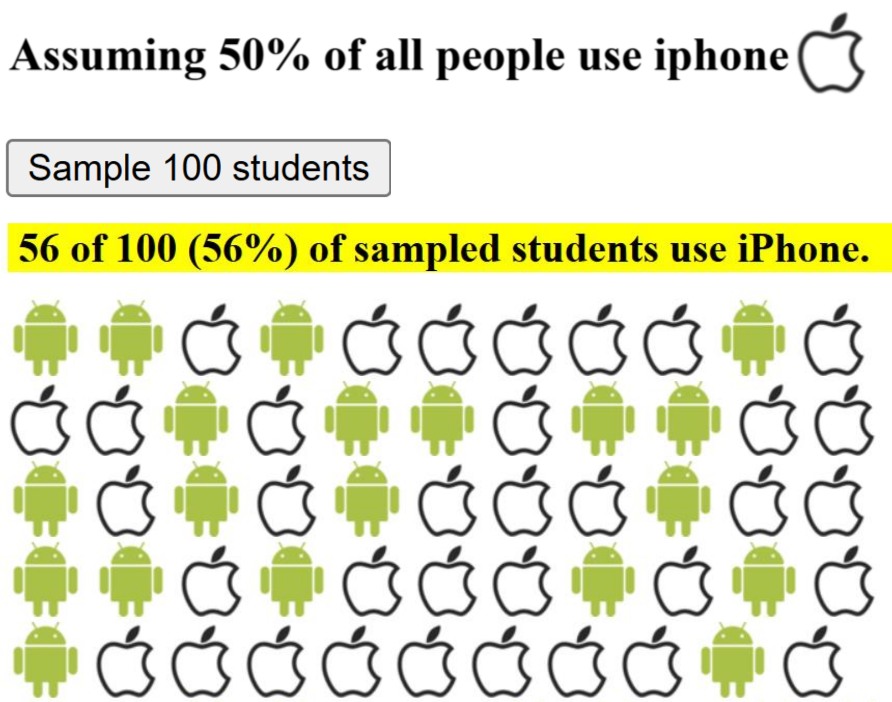
Example:
Suppose that in a population, **50% of all people** use an iPhone.

Now, let's take many random samples of 100 people from this population and examine the possible sample proportions we might obtain. To do this, we will use the following app:

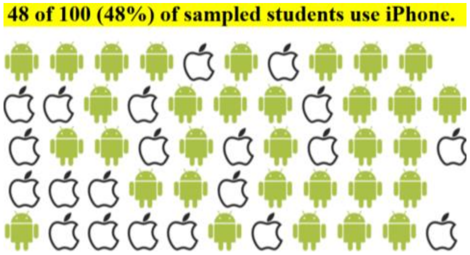
http://mylinux.langara.bc.ca/~sli/iphone/

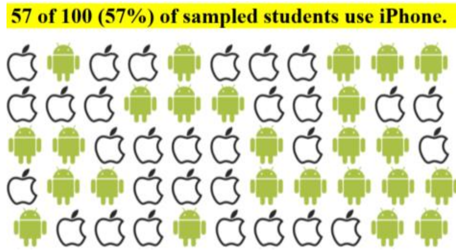
When you click the **"Sample 100 students"** button, the app will:

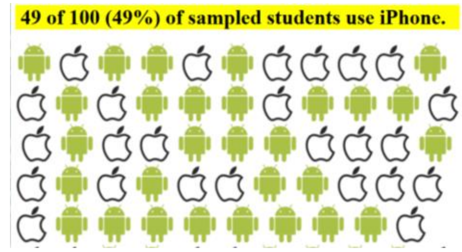
- randomly sample 100 students from the population,
- record which students use an iPhone and which do not, and
- calculate the percentage of iPhone users in the sample.

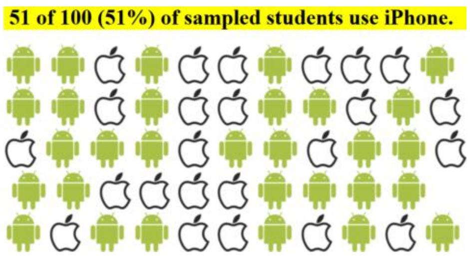


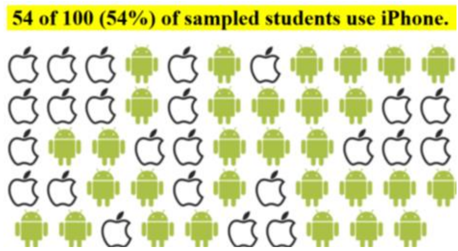
Next, we use the app to simulate 1,000 random samples and record the proportion of iPhone users in each sample.

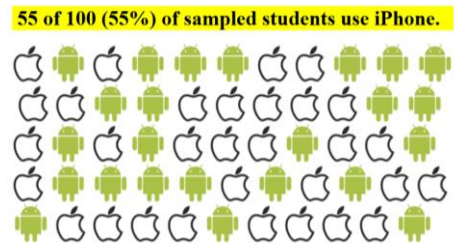










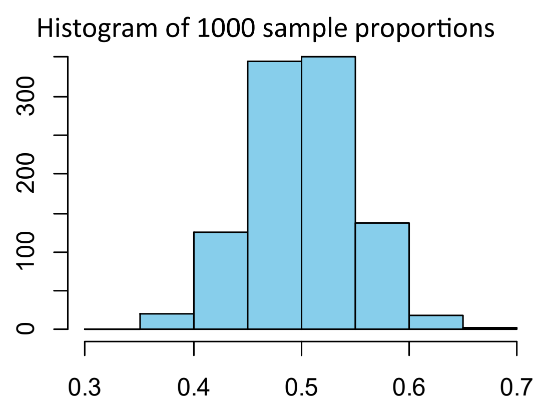


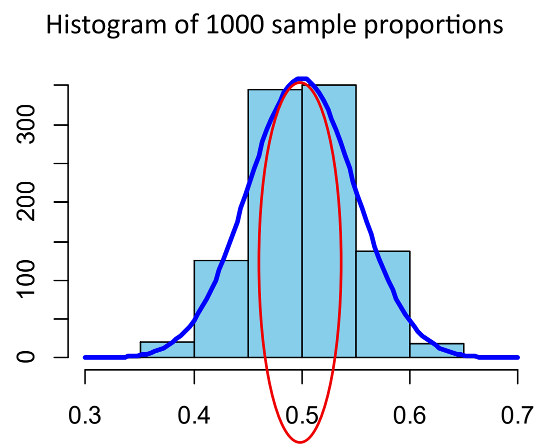
⋮ ⋮ ⋮

More results *More results* *More results*

Then we collect all the sample proportions (denoted by $\hat{p}$) and construct a histogram of these proportions.

Sample Proportion ($\hat{\boldsymbol{p}}$)
48
57
49
51
54
55
⋮
More results





Let's describe the distribution of the sample proportions.

1. The sample proportions follow an approximately **Normal distribution**.
2. **Most of the sample proportions are centered around 0.5**, which is exactly the true proportion of iPhone users in the population.

This is clearly illustrated by the simulation results.

Theory for the Distribution of the Sample Proportion

In the population, P X 100% of individuals fall in the category of interest

(e.g. **50%** of all people use iPhone)

Suppose many random samples with size n are selected from the population.

and the corresponding **sample proportions** ($\hat{p}$) are recorded.

Under certain conditions (which we will discuss later),

the distribution of the **sample proportions**:

- follows an approximately **Normal distribution**,
- has a **mean** (expected value) equal to the **population proportion** p, and
- has a **standard deviation** (also called the **standard error**) equal to

$$SD[\hat{p}] = \sqrt{\frac{p(1-p)}{n}}$$

Returning to our example, the proportion of iPhone users in the population is $p = 0.5$ (50%).

According to the theory:

- the sample proportion of iPhone users follows an approximately **Normal distribution**,
- the **mean** (expected value) of the **sample proportions** is equal to ______, and
- the standard deviation (standard error) of the **sample proportion** is

$$SD[\hat{p}] = \sqrt{\frac{p(1-p)}{n}} =$$

This means:

When we select many random samples of 100 people from the population, we expect the follow:

1. The sample proportions of people using iPhone ______________________________.
2. Each sample proportion is somewhat "**close**" to ______.
3. How **close**? On average, ____

   ____.

Conditions for the Sample Proportion to Follow a Normal Distribution

Question: How can we ensure that a sample proportion follows a Normal distribution?

The main requirement is that the sample is **large enough**. To check this, we verify that:

- The expected number of individuals in the category of interest is **at least 5**, and
- The expected number of individuals **not** in the category of interest is also **at least 5**.

Example:
In the population, 50% of all people use an iPhone.

For a random sample of 100 people, we expect:

- 50% of 100 people use iPhone.
- The other 50% (= 100% - 50%) people do not use iPhone

	Expected Number
Using iPhone	
Not using iPhone	
Total	100

Since both expected numbers are **at least 5**, the sample size is large enough for the sample proportion to be approximately Normal.

(Optional) Proof of the Mean and Standard Deviation of the Sample Proportion

The sample proportion $(\hat{p})$ of individuals in the category of interest is given by:

$$\hat{p} = \frac{\textbf{\textit{Number}} \text{ of individuals in the Category of Interest}}{n}$$

Hope you still remember the Binomial Distribution you learned in previous statistics class.

Recall the **Binomial distribution** from your previous statistics class.

The ***Number*** of individuals in the category of interest follows a **Binomial** distribution:

- Mean of Binomial Distribution $= np$ and
- SD of Binomial Distribution $= \sqrt{np(1-p)}$

Therefore,

$$\textbf{\textit{Mean or Expected value}} \text{ of } \hat{p} = \frac{\textbf{\textit{Mean or Expected Number}} \text{ of individuals in the Category of Interest}}{n}$$

$$= \frac{np}{n}$$

$$=$$

$$\textbf{SD} \text{ of } \hat{p} = \frac{\textbf{SD} \text{ of the Number of individuals in the Category of Interest}}{n}$$

$$= \frac{\sqrt{np(1-p)}}{n}$$

$$=$$

In the next lecture, we will apply the theory above to **estimation** and **hypothesis testing**.