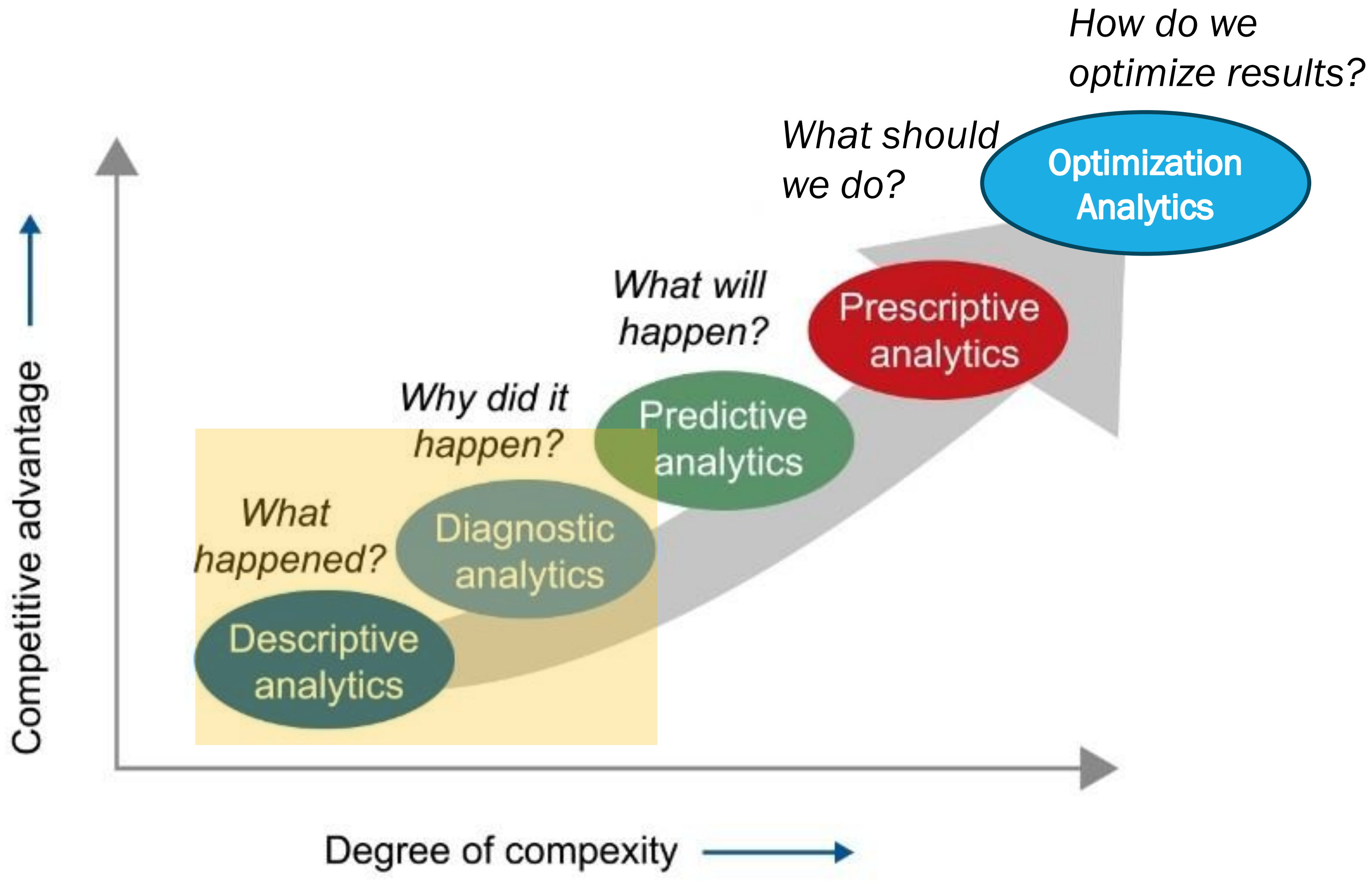
DATA SAMPLING

HIERARCHY OF DATA ANALYSIS TYPES



BYU

POPULATION VS SAMPLE

- Ideally, analyze all data for insights.
- Examples:
  - A mobile company wants to assess all potential customers.
  - A government must consider all citizens' needs for a new service.
- Full data collection is often impractical due to:
  - High costs of data acquisition
  - Time constraints
  - Computational and storage limitations
  - Increased complexity in processing large datasets
- Solution: Select a representative, make inference

DATA SAMPLING

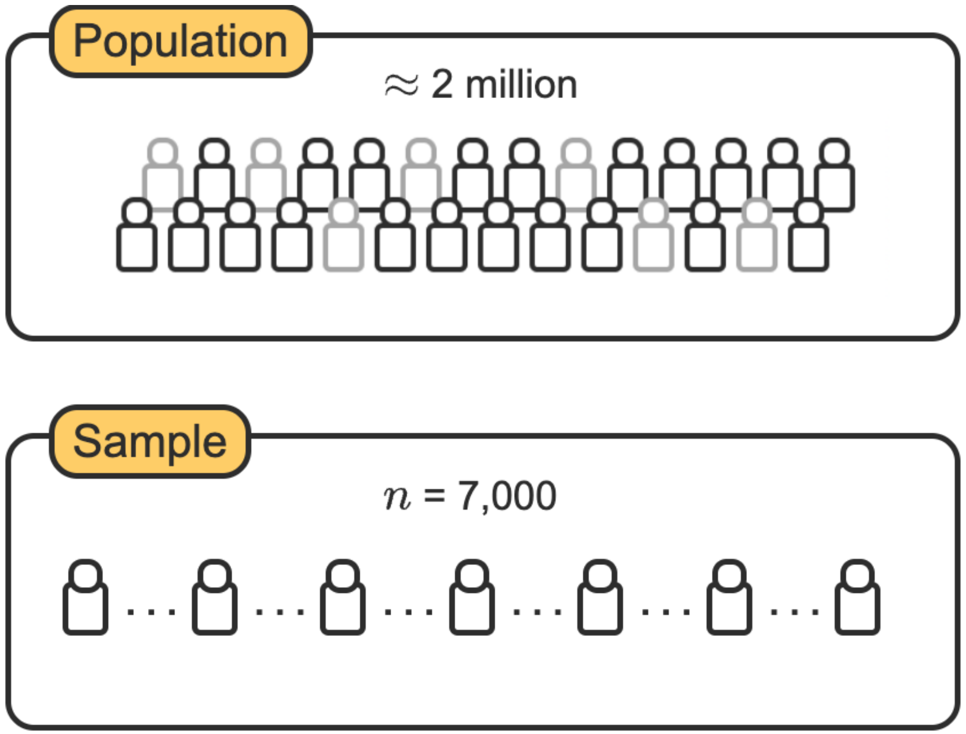
A ***Sampling Method*** is a process to select a subset of observations from the entire population. Ideally, the observations in the sample are representative of the population. Common methods include:

- Random Sampling: Each subset of n units is equally likely to be chosen.
- Stratified Sampling: The population is divided into meaningful groups (strata), and samples are drawn from each.
- Cluster Sampling: The population is divided into clusters (unrelated to key study features), and some clusters are randomly selected.
- Systematic Sampling: Every kth observation is selected from a random starting point, where $k \approx$ (population size) / n.
- Convenience Sampling: Easily accessible observations are selected (non-random).

BYU

SAMPLING SCENARIO

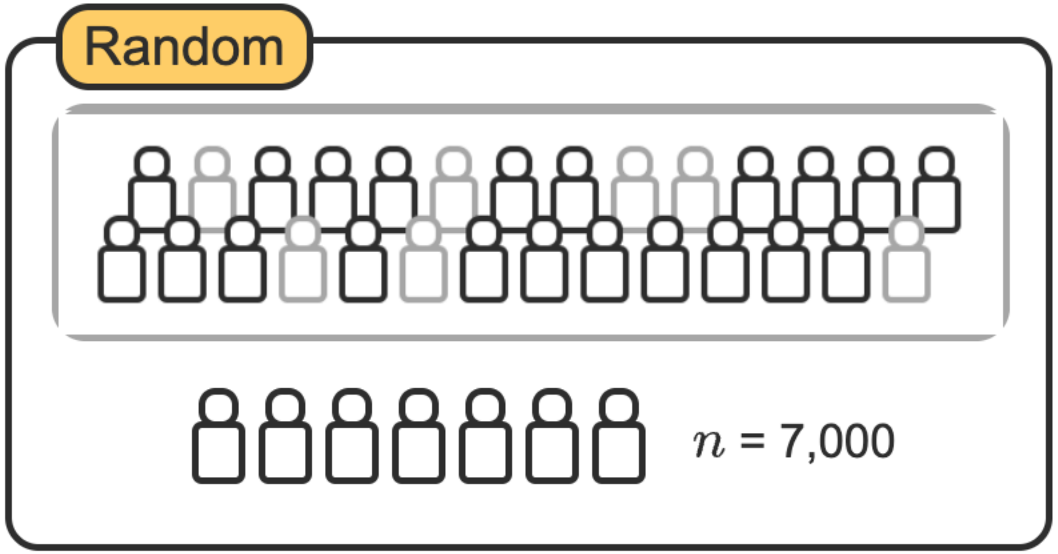
- A ***population*** is the entire set of all individuals, items, or events of interest.
- An ***observational unit (aka observation)*** is an individual, item, or event of the population where data is recorded.
- A ***sample*** is a subset of observations from the population used for analysis.
- Example: Transportation satisfaction survey across 5 cities.



BYU

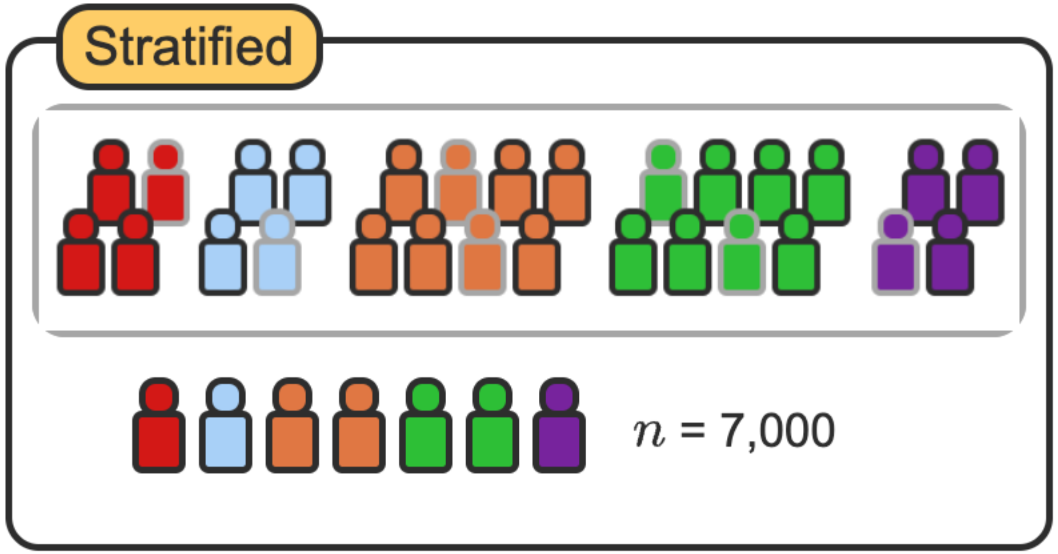
RANDOM SAMPLING

- In random sampling, passengers are selected at random from a list of all passengers in the five cities.
- Random sampling reduces the potential for sampling bias.
- But this could result in missing important events that occur less frequently.



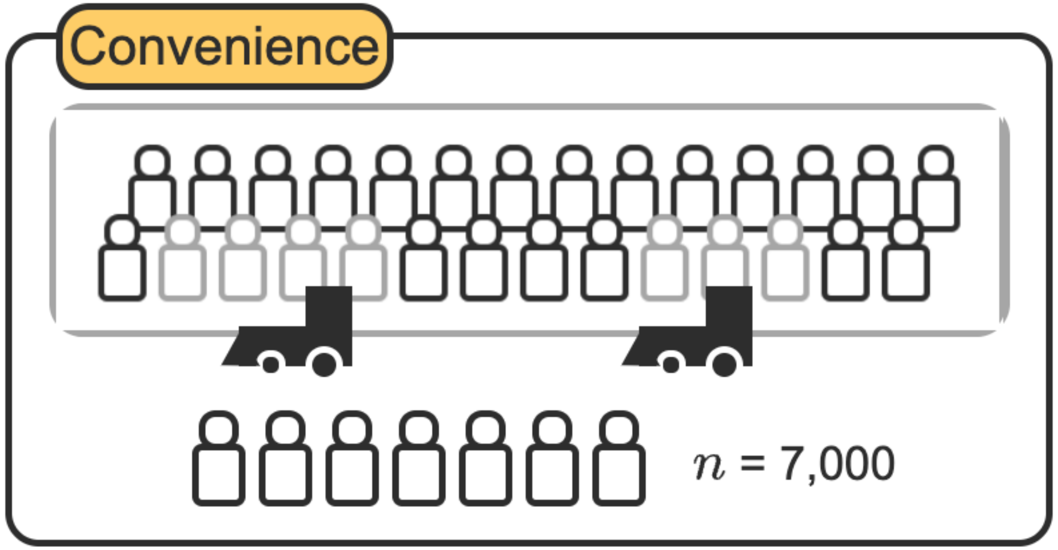
STRATIFIED SAMPLING

- Passengers are first divided into groups based on city.
- Then from each group, passengers are selected at random.
- Unlike pure random sampling, stratified sampling ensures adequate representation from each city.
- This is especially important when working with data that includes events that are relatively rare (e.g., customer churn, network intrusion, cancer cell detection, etc.)



CONVENIENCE SAMPLING

- Select passengers waiting in the train stations uses convenience sampling.
- This method is easy and quick, but the sample is not likely representative of all train passengers.

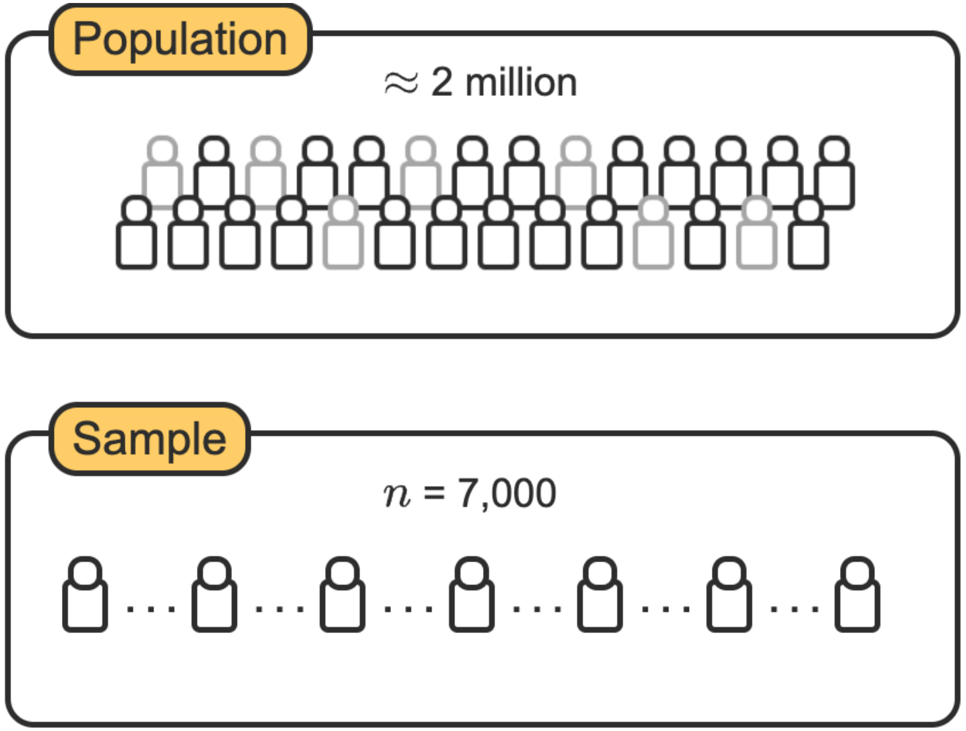


BYU

SYSTEMATIC SAMPLING

- Every 286th passenger from a list of all 2 million potential passengers is selected for the sample.
- Population / sample size = selection criteria
- Depending on ordering of the list, this could be close to random, or highly biased.

Select every 286th person



SAMPLING DISTRIBUTION – NOT ALL SAMPLES EQUAL

A **sampling distribution** is like taking many small samples from a big jar of jellybeans and calculating the proportion of red beans in each sample. If you repeat this process over and over, you'll get a bunch of different averages.

Now, imagine plotting those averages on a graph—you'll notice that most of them cluster around the true average of the whole jar. This pattern of sample proportions is the **sampling distribution.** It helps us understand how much the results can vary and lets us make better guesses about the whole jar without checking every single jellybean!

BYU

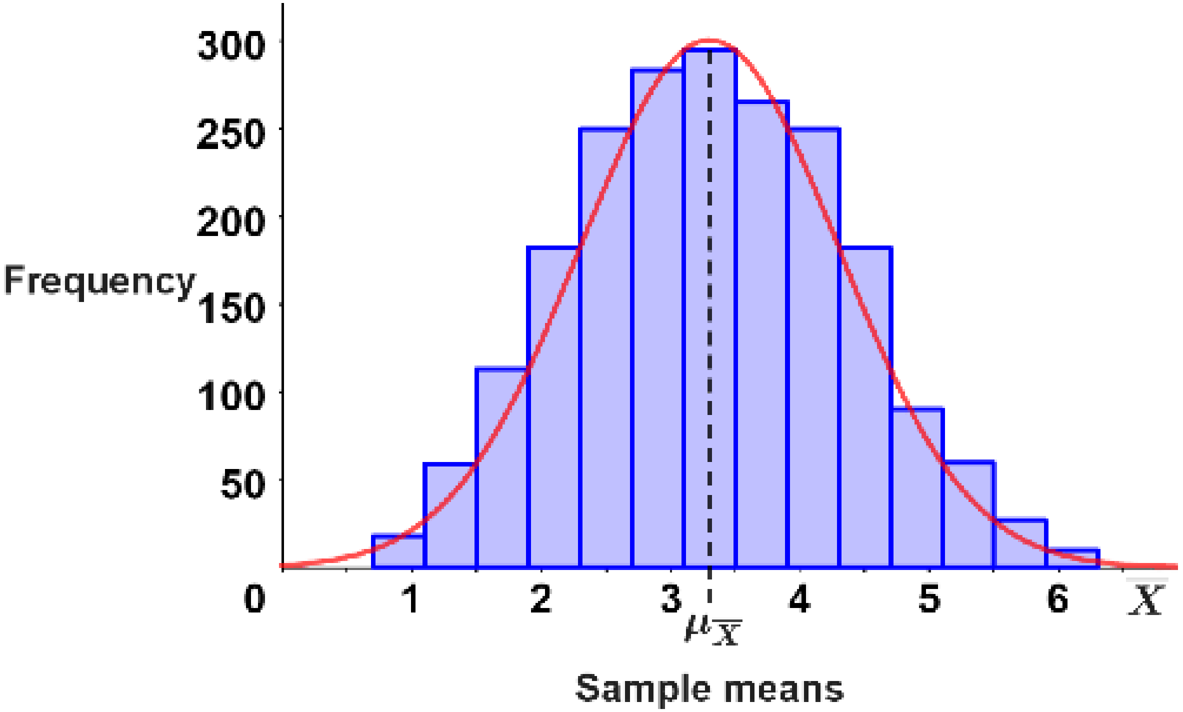
SAMPLING ACTIVITY: COIN TOSS

- Go to the coin flip simulator here: https://flipsimu.com/
- Flip 10 coin, record the number of heads.
- Repeat 5 times.
- Calculate the average number of heads.
- Does this average seem to well represent the expected number of heads, assuming the coin is fair?
- Record your average here: Coin Tosses

BYU

THE CENTRAL LIMIT THEOREM

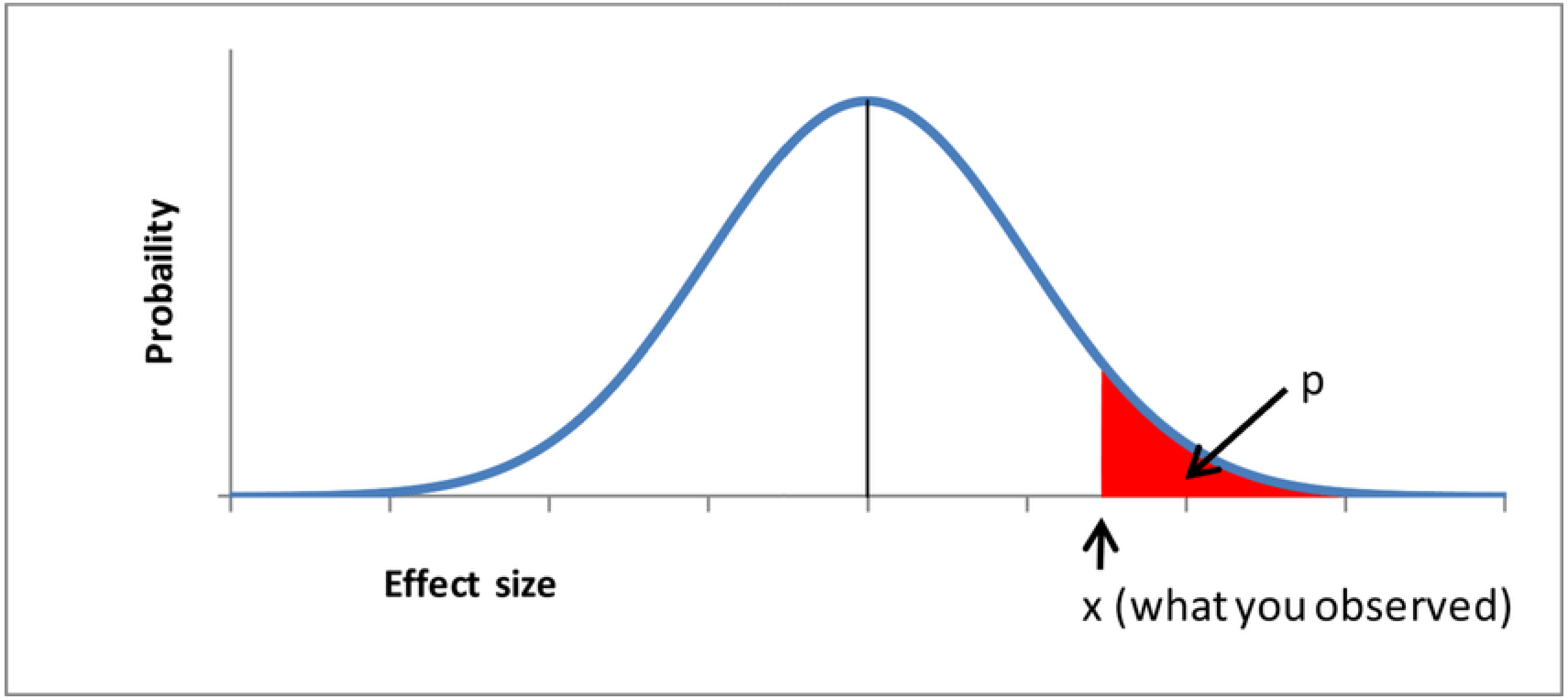
- **The sampling distribution of the sample mean approaches a normal distribution as sample size increases, regardless of the population's original distribution** (assuming random sampling).
- Works well for large n (>30).



BYU

IS OUR SAMPLE SIGNIFICANTLY DIFFERENT THAN THE POPULATION?

Proportion of heads …



ONE-SAMPLE T-TEST

In the case of a **one-sample t-test** (where you are comparing the sample mean against a known population mean), the equation becomes:

$$t = \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}}$$

Where:

- $\overline{X}$ is the sample mean,
- μ is the population mean,
- S is the sample standard deviation,
- n is the sample size.

This formula tests whether the sample mean $\overline{X}$ significantly differs from the population mean μ. The denominator represents the **standard error** of the mean.