PODS Lab 8: Sampling and Hypothesis Testing

Hamza Alshamy

Center for Data Science, NYU
ha2486@nyu.edu

03/14/25

1. **Inference**
2. **Sampling**
3. **Hypothesis Testing Framework**
   - Null (H_0) and Alternative Hypothesis (H_1)
   - Significance Level
   - Statistical Significance
4. **Quiz + Discussion**

Switching from prediction to inference

- **Before:** Characterizing data (central tendency and dispersion), correlation, regression (prediction)
- **Now:** Inference, sampling, hypothesis testing

Motivating Example: The Need for Sampling in Inference

- **Scenario:** You want to find the average female height in the U.S.

Motivating Example: The Need for Sampling in Inference

- **Scenario:** You want to find the average female height in the U.S.
- **Ideally,** we would literally measure every female's height in the U.S. and take the average

Motivating Example: The Need for Sampling in Inference

- **Scenario:** You want to find the average female height in the U.S.
- **Ideally,** we would literally measure every female's height in the U.S. and take the average
- **Problem:** Not possible! Cannot coordinate, too many data points, expensive, time constraint, etc

Motivating Example: The Need for Sampling in Inference

- **Scenario:** You want to find the average female height in the U.S.
- **Ideally,** we would literally measure every female's height in the U.S. and take the average
- **Problem:** Not possible! Cannot coordinate, too many data points, expensive, time constraint, etc
- **Solution:** Take a subset of the female population (e.g. 1000), measure their heights, and take the average

- We are estimating the height of the entire U.S. female population using a sample

What is Statistical Inference?

- **Inference:** The process of drawing conclusions about a population using a sample.
- **Key Idea:** We rely on a subset of the population (sample) to estimate unknown population characteristics.
- **Example:** Estimating the average height of U.S. females by measuring a subset of the population.

What is Statistical Inference?

- **Inference:** The process of drawing conclusions about a population using a sample.
- **Key Idea:** We rely on a subset of the population (sample) to estimate unknown population characteristics.
- **Example:** Estimating the average height of U.S. females by measuring a subset of the population.
- Since sampling is a core idea in inference, let's talk about it

Sampling Terminology

- **Population:** The entire group we want to study (e.g., all U.S. females).
  - **Population Parameter (θ):** A fixed, unknown value describing the population (e.g., true average height of all U.S. females, denoted as μ).

Sampling Terminology

- **Population:** The entire group we want to study (e.g., all U.S. females).
  - **Population Parameter (θ):** A fixed, unknown value describing the population (e.g., true average height of all U.S. females, denoted as μ).
- **Sample:** A subset of the population used for analysis.
  - **Sample Statistic ($\hat{\theta}$):** A measurable value computed from the sample that estimates the population parameter (e.g., sample mean $\bar{x}$).

	Parameter (θ)	**Statistic ($\hat{\theta}$)**
Mean	μ	$\bar{x}$
Standard Deviation	σ	s

Key Idea: Statistical inference estimates a population parameter θ using a sample statistic $\hat{\theta}$.

Why Does This Work? Law of Large Numbers (LLN)

- **Law of Large Numbers (LLN):** If sampling is independent, representative, and random, then as the sample size increases, the sample statistic (e.g., sample mean) converges to the population parameter (e.g., population mean).

Why Does This Work? Law of Large Numbers (LLN)

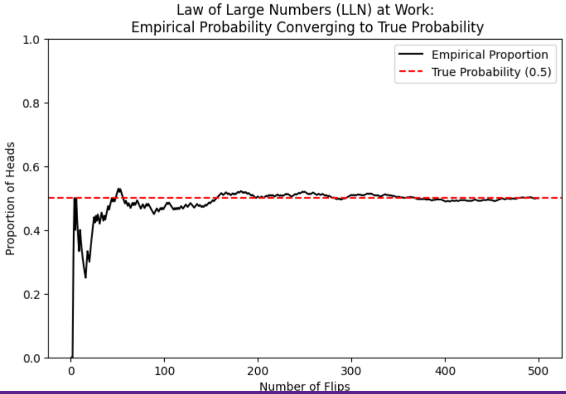
- **Law of Large Numbers (LLN):** If sampling is independent, representative, and random, then as the sample size increases, the sample statistic (e.g., sample mean) converges to the population parameter (e.g., population mean).
- **Intuition:**
  - With a small sample, randomness can cause large fluctuations.
  - As the sample grows, these fluctuations average out, bringing the estimate closer to the true value.
- $\lim_{n\to\infty} P(|\bar{x} - \mu| \geq \varepsilon) = 0$

Why Does This Work? Law of Large Numbers (LLN)

- **Law of Large Numbers (LLN):** If sampling is independent, representative, and random, then as the sample size increases, the sample statistic (e.g., sample mean) converges to the population parameter (e.g., population mean).
- **Intuition:**
  - With a small sample, randomness can cause large fluctuations.
  - As the sample grows, these fluctuations average out, bringing the estimate closer to the true value.
- $\lim_{n\to\infty} P(|\bar{x} - \mu| \geq \varepsilon) = 0$
- **Assumptions of LLN:**
  1. **Independence:** Each observation in the sample is independent of the others.
  2. **Random Sampling:** The sample is selected randomly to avoid bias.

Example: Law of Large Numbers in Coin Flipping

- **Key Idea:** As the number of coin flips increases, the **sample statistic** (empirical probability of heads) converges to the **true probability** (0.5).
- **Observation:** The fluctuations are large for small samples but stabilize as the sample size grows.



Sampling Bias

- **Question:** What do we mean by sampling bias?

Sampling Bias

- **Question:** What do we mean by sampling bias?
- **Definition:** Sampling bias occurs when certain groups in the population are systematically over or underrepresented in a sample, leading to misleading conclusions.

Sampling Bias

- **Question:** What do we mean by sampling bias?
- **Definition:** Sampling bias occurs when certain groups in the population are systematically over or underrepresented in a sample, leading to misleading conclusions.
- **Example:**
  - Survey people about their sleeping habits at 7 AM on the street.
  - You find that most respondents are early risers.
  - **This leads to a biased conclusion:** You vastly overestimate the proportion of early risers in the population.

Sampling Distribution

- **Key Idea:** Each time we draw a sample, we get a slightly different estimate.
- **Sampling Distribution:** The distribution of a sample statistic (e.g., sample mean $\bar{x}$ or sample standard deviation S) across many repeated samples from the same population.

Sampling Distribution

- **Key Idea:** Each time we draw a sample, we get a slightly different estimate.
- **Sampling Distribution:** The distribution of a sample statistic (e.g., sample mean $\bar{x}$ or sample standard deviation S) across many repeated samples from the same population.

- **Example:**
  - Draw 1,000 random samples of 50 people ($n = 50$) from a population
  - Calculate the mean height of each sample
    - $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_{1000}$
  - The distribution of those sample means would be the **sampling distribution of the mean.**

Central Limit Theorem (CLT) – Link for CLT Simulation (clickable!)

- **Central Limit Theorem (CLT):** Regardless of the population's shape, the sampling distribution of the sample mean becomes approximately normal as the sample size increases.
- **Implication:**
  - Even if the population is skewed, multimodal, or discrete, the sample mean will **follow a normal distribution** for sufficiently large *n*.

Formal Statement:

- If we take a large number of random samples of size n from any population with mean μ and standard deviation σ, then the sample means $\bar{x}$ follow:

$$\bar{x} \sim \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Standard Error of the Mean (SEM)

- **Standard Error of the Mean (SEM):** The standard deviation of the **sampling distribution** of the sample mean.

Theoretical (True Population)	**Practical (Estimate w/ Data)**
$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$	$SEM = \frac{s}{\sqrt{n}}$

- **Where:**
  - $\sigma_{\bar{x}}$: Standard deviation of sample mean (i.e., the standard error).
  - σ: Population standard deviation.
    - Practically, we use the sample standard deviation s since σ is unknown.
  - n is the sample size.
- **Note!** Decreases as a function of the square root of the sample size.

Hypothesis Testing Framework

- **What is Hypothesis Testing? (A falsification approach)**
  - A statistical method used to make inferences about a population based on sample data.
  - It helps us determine whether an observed effect is real or due to random chance.
- **Hypothesis Testing Framework:**
  1. **Formulate Null (H_0) and Alternative (H_1) Hypotheses**
     - H_0: Assumption of no effect or no difference.
     - H_1: Assumption of an effect or a difference.
  2. **Choose a Significance Level (α)**
     - Common choices: $\alpha = 0.05, 0.01$.
  3. **Determine Test Statistic**
     - Depends on data type and assumptions.
     - Examples: Z-test, T-test, KS test, Mann-Whitney U test.
  4. **Compute P-value**
     - **Reject H_0:** If $p \leq \alpha$ (statistically significant).
     - **Fail to reject H_0:** If $p > \alpha$ (not enough evidence).

Null (H_0) and Alternative (H_1) Hypothesis

1. **Null Hypothesis (H_0):** The assumption that there is no effect, no difference, or no relationship in the population.
   - Represents the status quo or baseline assumption.
   - We assume H_0 is true unless we have strong evidence against it.
2. **Alternative Hypothesis (H_1):** The hypothesis that there is an effect, a difference, or a relationship in the population.
   - Represents what we want to test for.
   - If there is enough evidence, we reject H_0 in favor of H_1.

P-Value and Interpretation

- **What is the P-value?** Probability of obtaining a test statistic as extreme as (or more extreme than) the observed one, assuming the null hypothesis H_0 is true.

- **P-Value Calculation Algorithm**
  1. Compute the test statistic.
  2. Find the probability of obtaining that test statistic using a corresponding probability distribution.

- **Interpretation:**
  - The probability of observing the data (or something more extreme) given the null hypothesis H_0 is true.

Significance Level (α) and Statistical Significance

- **Significance level (α):**
  - A threshold that is set to determine whether the results of a statistical test are significant or not
  - **Common convention:** $\alpha = 0.05$ (5%) is widely used.
  - **Why is 0.05 standard?**
    - Ronald Fisher (1925) found it convenient as a practical cutoff
- **Statistical Significance:**
  1. **Statistically significant ($p \leq \alpha$)**
     - The observed pattern of data is **unlikely due to chance alone**
     - **Decision:** Reject the null hypothesis (H_0)
  2. **NOT statistically significant ($p > \alpha$)**
     - The observed pattern of data is **plausible under random chance**.
     - **Decision:** Fail to reject the null hypothesis (H_0)

Full Example: Z-test as a Test Statistic (one-tail)

- Algorithm:
  1. Compute Z-score
  2. Look up the Z-score in a standard Z-table (Clickable link!).
  3. Determine the corresponding probability to the Z-score (P-value)

- **Example:** $\mu = 82$, $\sigma = 13$, $\bar{x} = 85$, $n = 200$, $H_0 : \mu = 82$, $H_1 : \mu > 82$

$$Z = \frac{\bar{x} - \mu}{SEM} \approx \frac{85 - 82}{0.92} \approx 3.26$$

- $P(Z \geq 3.26) = 1 - .99944 \approx 0$
- Since $p = 0.00093$ is **less than** $\alpha = 0.05$, we **reject** H_0 (statistically significant result).

Errors

- Our conclusion can be wrong!

		Reality	
		Yes	No
Significant?	Yes	Correct (1-β)	Type I Error False Positive (α)
	No	Type II Error False Negative (β)	Correct (1-α)

Table: Confusion Matrix for errors we might see

Quiz + Discussion