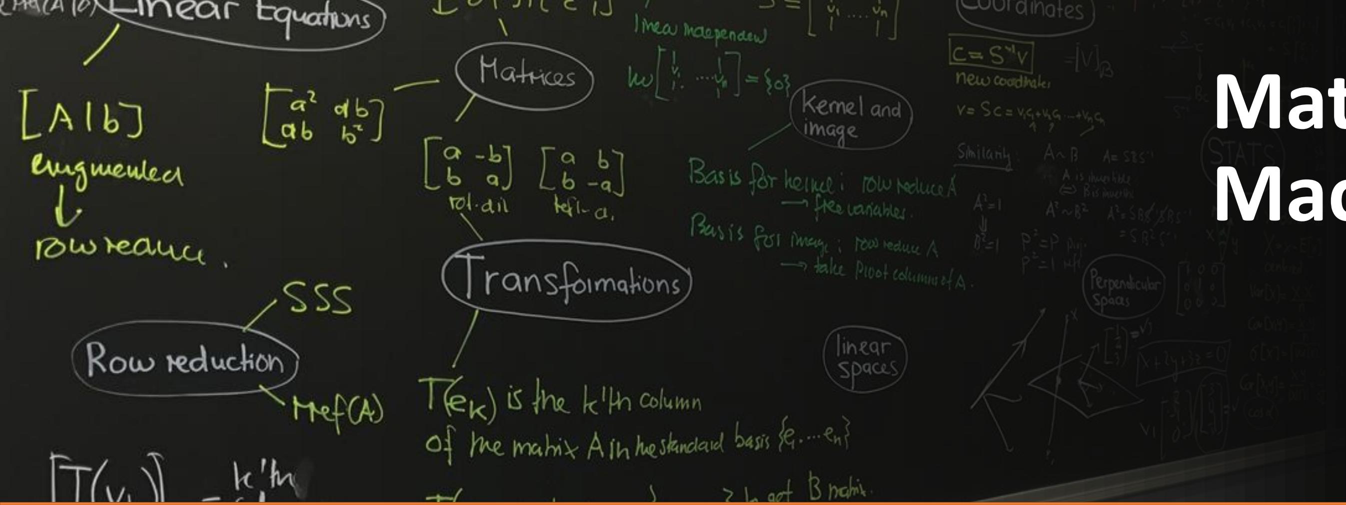


Mathematical Foundations for Machine Learning and Data Science

Statistical Inference

Dr. Zubair Khalid

Department of Electrical Engineering
School of Science and Engineering
Lahore University of Management Sciences

https://www.zubairkhalid.org/ee212_2021.html

LUMS
A Not-for-Profit University

Outline

- What is the statistical inference?
- Understanding Statistical Inference Process
- Hypothesis Testing – One sample z-Test

Statistical Inference

Overview:

Statistical Inference refers to drawing conclusions about a population from sample data.

Idea: Use a random sample to learn something about a larger population.

Types:

- Hypothesis testing
- Estimation of parameters
  - Point estimates
  - Interval estimates

Statistical Inference

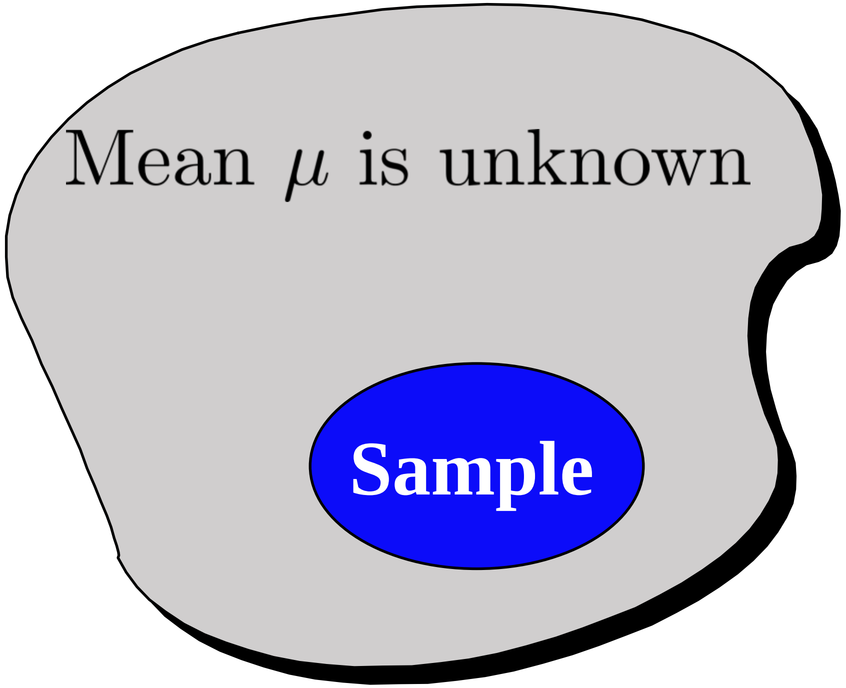
Why? Role in Data Science - Examples:

- *Assess whether all sales across all of the stores of the chain at the same level over one month*
- *Find the average weight or height of men population in a country*
- *Understand whether Pfizer (drug producer) made the same profit after COVID-19*
- *Prove a particular claim about a population wrong*

Statistical Inference

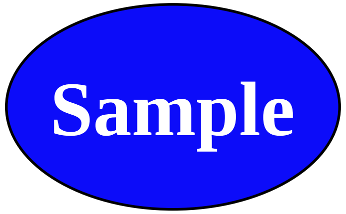
Estimation of Parameters - Example:

Population



Point estimate

Sample mean $\bar{x}$



Interval estimate

I am 95% confident that μ is between 40 and 60.

Statistical Inference

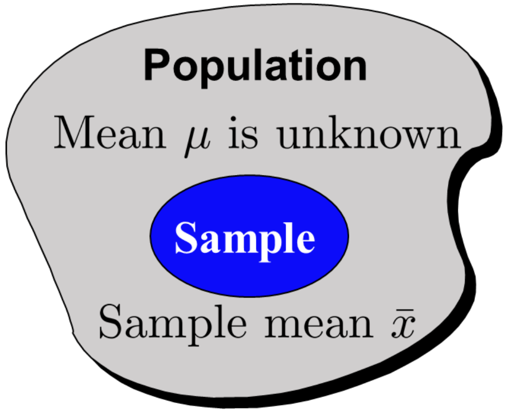
Estimation of Parameters - Example:

Claim (before collecting data):

Body weight of population of men has ***mean μ*** *= 170 pounds and standard deviation σ = 40 pounds.*

Objective:

Collect a sample and test a claim about population mean.



Alternative Claim:
- *Is mean body weight of a particular population of men higher than 170 pounds?*
- *Is mean body weight of a particular population of men not equal to 170 pounds?*

- Sample mean $\bar{x}$ can serve as a good measure of popultion mean under some asumptions.

Statistical Inference allows us to answer the following questions:

How confident are we? How can we reject this claim?

Is there a way to calculate this claim so that we can be sure? Can we prove it via quantifiable measures?

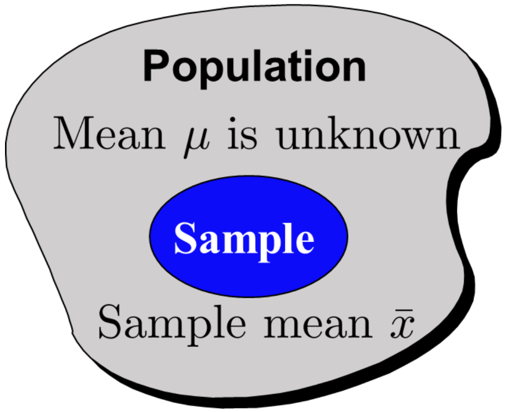
Statistical Inference

Estimation of Parameters:

The process of **test statistics** can be used to help to test the ***hypothesis*** about the parameters of the population from its sample

Idea:

A sample can be thought of as a random variable having its own probability distribution, patterns, and trends. We can collect a number of samples and workout their means, standard deviation, and variances to gain better insight into the data.



Statistical Inference

One sample z-Test:

Hypothesis Testing:

Null Hypothesis (H_0)— what is known as the accepted truth and what we want to prove wrong

Alternate Hypothesis — what we need to accept if the Null Hypothesis is not true. This is what we believe is true.

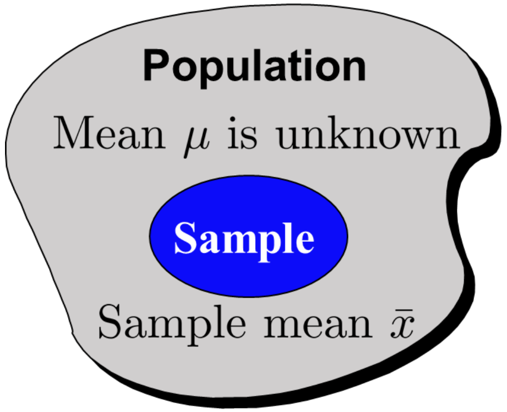
We introduce hypothesis testing **concepts** with the **most basic** testing procedure: the **one-sample z** test

Objective:

Test a claim about population mean.

Assumptions:

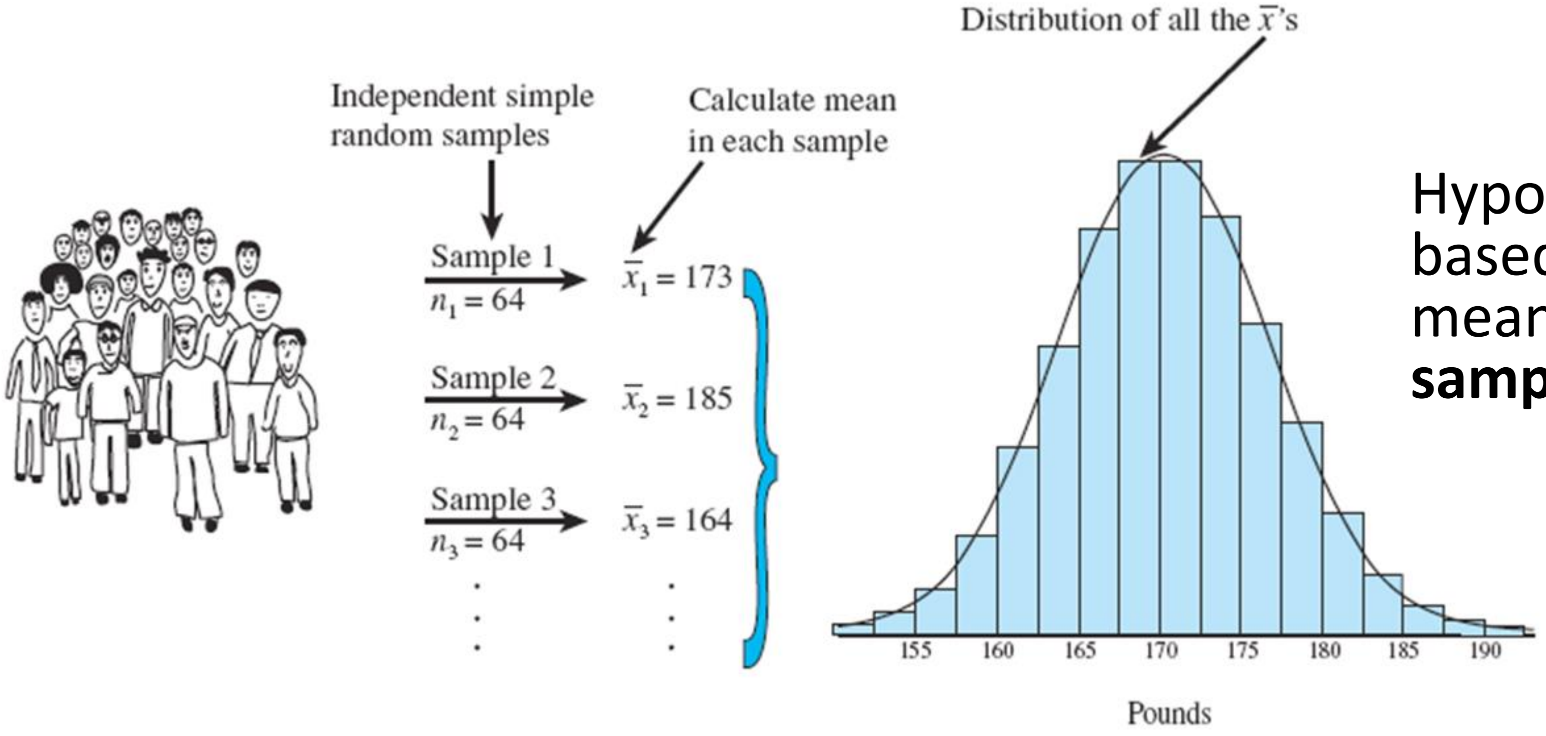
- **S**imple **R**andom **S**ample (SRS)
- Population **Normal** *or* **sample large**
- The value of **σ is known**
- The value of **μ is NOT known**



Statistical Inference

Sampling Distribution of Mean:

- What is the mean weight **μ** of a population of men?
- Sample *n* = 64 and calculate sample mean “x-bar”
- If we sampled again, we get a different x-bar



Repeated samples from the same population yield different sample means

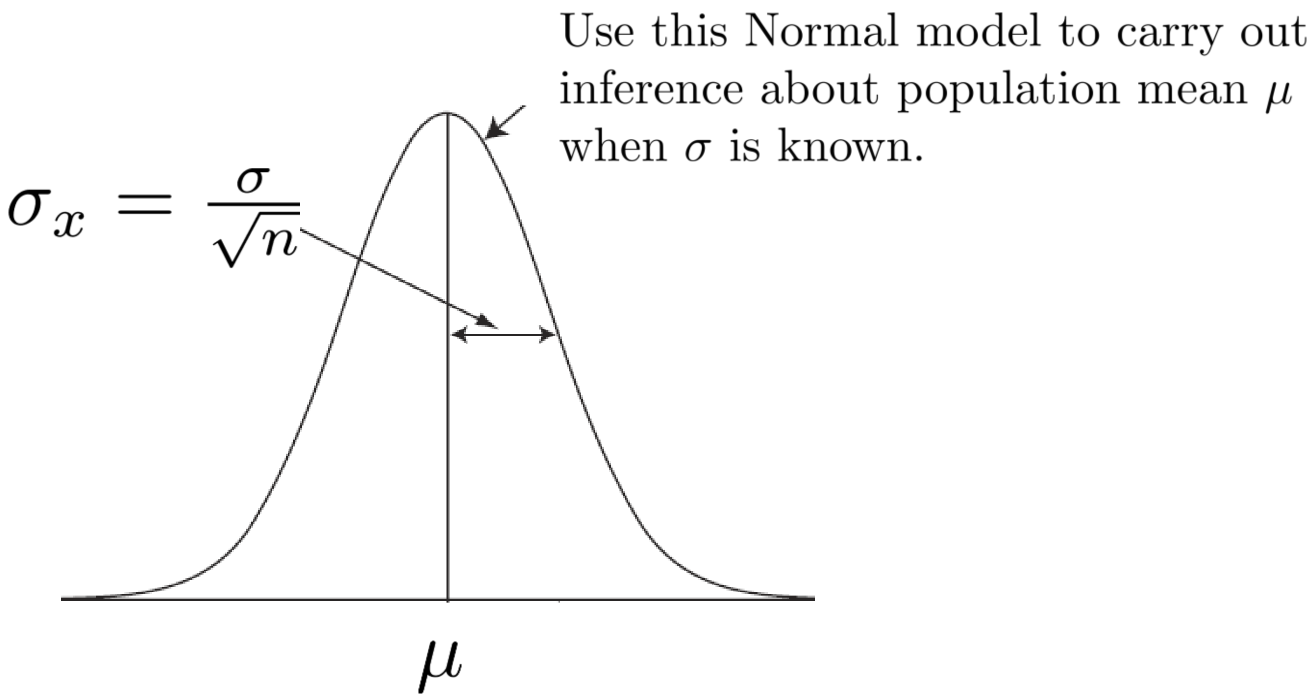
Hypothetical **probability model** based on the differing sample means. This distribution is called the **sampling distribution of the mean.**

Statistical Inference

Sampling Distribution of Mean – Model and Characteristics:

- Tend to be Normal
- Centered on population mean μ
- Standard deviation

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



Statistical Inference

One sample z-Test- Procedure:

1. Hypothesis statements:

Start by stating the widely believed claim about population mean which is known as the **Null hypothesis** H_0. Formulate an alternative hypothesis.

Null hypothesis H_0: $\mu = \mu_o$

Alternative hypothesis: H_a: $\mu < \mu_o$ **(one-sided)** OR
H_a: $\mu > \mu_o$ **(one-sided)** OR
H_a: $\mu \neq \mu_o$ **(two-sided)**

2. Test statistic

Calculate our sample results mean and standard deviation. Calculate the test statistics (z-statistic) assuming null hypothesis is true as

$$z_{\text{stat}} = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}$$

- μ Population mean claimed in null hypothesis
- $\bar{x}$ sample mean
- $\sigma_{\bar{x}}$ is the standard deviation of $\bar{x}$, that is related to population standard deviation σ as $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$.

Statistical Inference

One sample z-Test- Procedure:

3. Convert z statistics to a P-value *(the probability of having a sample "more extreme" than the ones observed in the given data assuming null hypothesis is true):*

- For H_a: $\mu > \mu_0$
  P-value = $\Pr(Z > z_{stat})$ = right-tail beyond z_{stat}

- For H_a: $\mu < \mu_0$
  P-value = $\Pr(Z < z_{stat})$ = left tail beyond z_{stat}

- For H_a: $\mu \neq \mu_0$
  P-value = 2 × one-tailed P-value

4. P-value interpretation before we can reject the claim.

Statistical Inference

One sample z-Test - Example:

Claim (before collecting data):

*Body weight of population of men has **mean μ** = 170 pounds and standard deviation σ = 40 pounds.*

1. Hypothesis Statements to be tested:

Null hypothesis H_0: $\mu = 170$

Alternative hypothesis: H_a: $\mu > 170$ **(one-sided)** *OR*
H_a: $\mu \neq 170$ **(two-sided)**

2. 2. Test statistic

Take an SRS of n = 64 and calculate a sample mean equal to 173.

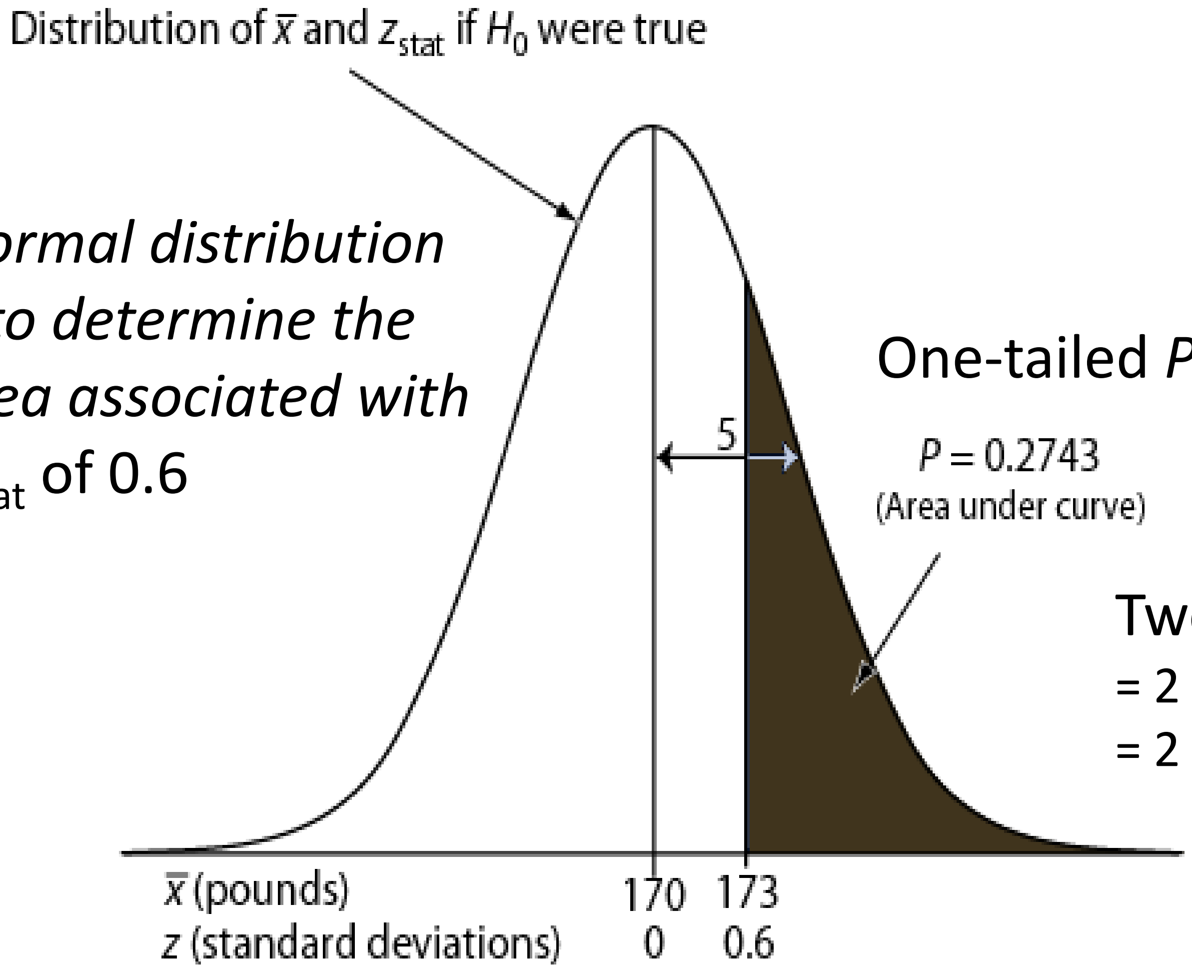
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = 5$$

$$z_{\text{stat}} = \frac{\bar{x}-\mu}{\sigma_{\bar{x}}} = 0.6$$

Statistical Inference

One sample z-Test - Example:

3. Convert *z* statistics to a *P*-value **Alternative hypothesis**: H_a: μ ≠ 170 **(two-sided)**



Use normal distribution table to determine the tail area associated with the z_{stat} of 0.6

One-tailed P = .2743

Two-tailed P
= 2 × one-tailed P
= 2 × .2743 = .5486

Statistical Inference

One sample z-Test - Example:

P-value interpretation:

- *P*-value is the the probability of the observed test statistic **when H_0 *is true***?

- Smaller and smaller *P*-values provide stronger and stronger evidence against H_0

Conventions:

$P > 0.10 \Rightarrow$ poor evidence against H_0

$0.05 < P \leq 0.10 \Rightarrow$ marginally evidence against H_0

$0.01 < P \leq 0.05 \Rightarrow$ good evidence against H_0

$P \leq 0.01 \Rightarrow$ very good evidence against H_0

Statistical Inference

One sample z-Test - Summary:

Draw an SRS of size n from a Normal population that has unknown mean μ and known standard deviation σ. To test the null hypothesis that μ has a specified value,

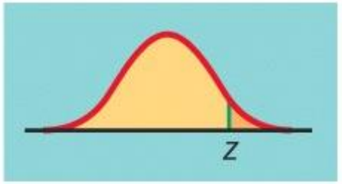
$$H_0: \mu = \mu_0$$

calculate the **one-sample z statistic**

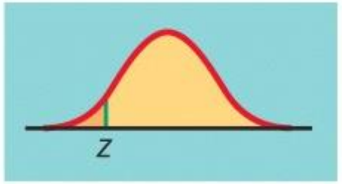
$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

In terms of a variable Z having the standard Normal distribution, the P-value for a test of H_0 against

$H_a: \mu > \mu_0$ is $P(Z \geq z)$



$H_a: \mu < \mu_0$ is $P(Z \leq z)$



$H_a: \mu \neq \mu_0$ is $2P(Z \geq |z|)$