BUDT 730
Data, Models and Decisions

Lecture 5

Central Limit Theorem & Confidence Interval

Prof. Sujin Kim

Introduction to Statistical Inference

- In a typical statistical inference problem, you want to discover one or more characteristics of a given population
- Topics that we will cover over the next few classes
  - Ch7 Sampling and Estimation
  - Ch8 Confidence Intervals
  - Ch9 Hypothesis Testing

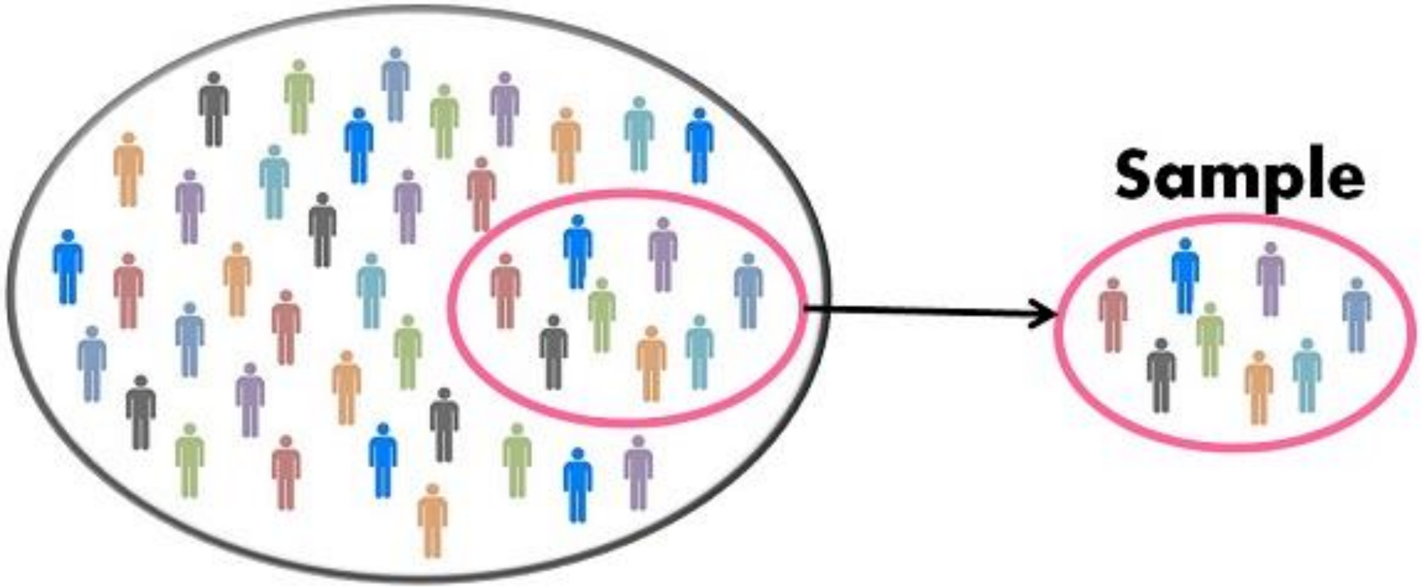
Agenda

- Overview of statistical inference
- Ch7 Sampling and Point Estimation:
  - Understand the sampling distribution of a sample mean
  - Understand the ‘Central Limit Theorem (CLT)’
  - Calculate the probabilities for a sampling distribution
- Ch8: Confidence Intervals
  - Understand the concept of a confidence interval.
  - Calculate and interpret a confidence interval for a population mean

Population vs. Sample

Population

The set of all members about which a study intends to make *inferences*



Sample

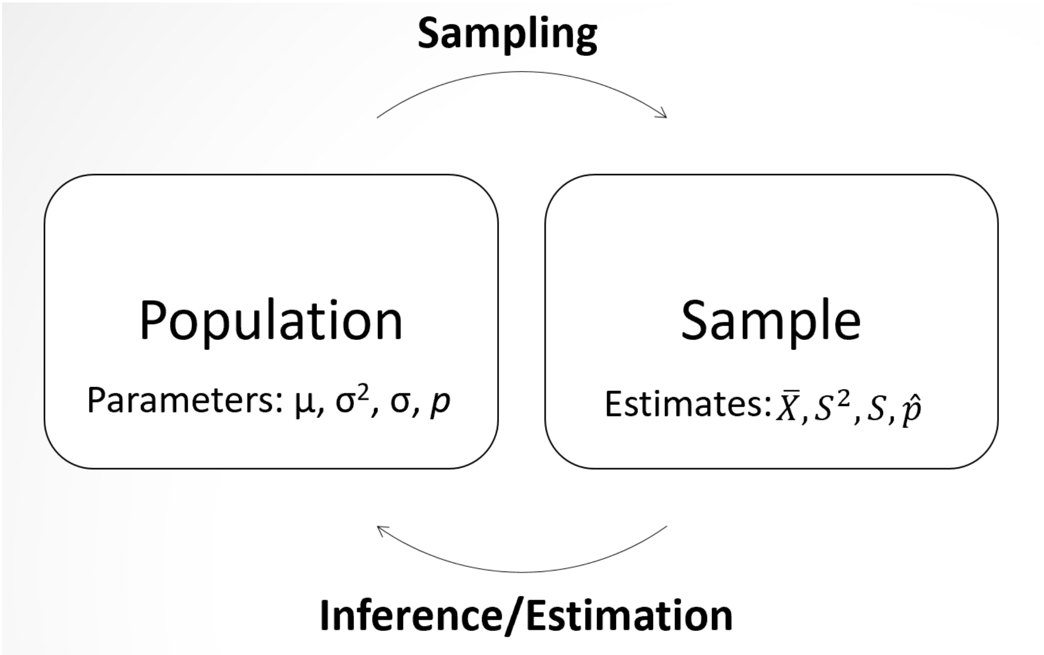
A subset of a population whose properties are studied to gain information about the population as a whole

We typically do not have access to the entire population, so we need to sample a subset, then infer the characteristics of the population based on this sample

Statistical Inference

- **Statistical inference** is the process of using sample data to infer properties of the underlying population

- It is the foundation for data analysis and divided into two major areas: parameter estimation and hypothesis testing (on parameters)



Parameter Estimation (Ch7-8)

- Two parameter estimates:
  - A **point estimate** is a single numeric value, a “best guess” of a population parameter, based on the data in a random sample.

    ex: sample mean, sample variance, sample proportion
  - A **confidence interval (CI)** is an interval around the point estimate, calculated from the sample data, that is very likely to contain the true value of the population parameter

Ch 7
Sampling and Sampling Distributions

Sampling

- **Sampling** is the act, process, or technique of selecting a suitable sample, or a representative part of a population for the purpose of determining parameters or characteristics of the whole population

- There are two basic types of samples
  - Random (Probability) sample: members are chosen according to a random mechanism
  - Judgmental sample: members are chosen according to a sampler’s judgment.

- **Random sampling** is commonly used in practice, and we focus exclusively on probability samples here on.

Types of Random Sampling

- There are many different types of random sampling Techniques, including:
  - **Simple random sampling**
  - Systematic sampling
  - Stratified sampling
  - Cluster sampling

- The choice depends on the situation
- we will focus on **simple random samples**, where the mathematical details are relatively straightforward – We cannot directly apply the standard statistical analysis to other sampling methods.

Simple Random Sampling

Simple Random Sampling

- Default sampling in statistical analysis: can generate i.i.d. samples
- The simple random sample selects each member of the population with equal probability: population size = n => the probability of being chosen =1/n
- Simple random samples have some challenges
  - How do we randomly sample people? How do we get it so that everybody is equally as likely?
  - It can be expensive (e.g. have to cover vast geographical regions, east coast to west coast, north to south)
  - It can result in under- and overrepresentation of population segments (e.g. minorities may not be represented appropriately)

Central Limit Theorem
& Sampling Distribution

Introduction to Estimation

- The purpose of any sample is to estimate properties of a population from the data observed in the sample
- The mathematical procedures for performing this **estimation depend on which population characteristic is of interest.**
- We will study the estimation of a <u>population mean</u> and a <u>population proportion</u> in this course.

Example: Meal Service

- A government contractor provided services to the military in a troubled region. The contractor claims that
  - Average of 10,000 daily meals provided, and
  - Standard deviation is 1643.17.
- The operations lasted 300 days:
  - Cost: $10/meal
  - Total charged: $30 million

- The government believes that the charges of the contractor are too high.
- The government obtains a random sample of 30 days
  - Average number of meals for 30 days: 8,983 meals served

Meal Service: What is your conclusion?

Based on the auditor's sample, what is your conclusion?

a) The contractor's charges are accurate.
b) The contractor's charges are too high.
c) It is impossible to say – the sample is way too small.

Can we estimate the charges with 100% certainty?

What can we do instead?

Properties of the Sample Mean

- Draw samples from X via the simple random sampling: $X_1, X_2, \dots, X_n$
- Then, $X_1, X_2, \dots, X_n$ are i.i.d.
- The sample mean is

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

- Then, the sample mean $\bar{X}$ itself is random!
- Since the sample mean is a random variable, we can associate with it:
  - An expected value,
  - A variance, and
  - A distribution.
- Our goal is to understand what types of statements we can make based on our sample

Expected Value of the Sample Mean

- Let's assume that the original random variable X has the mean μ and the standard deviation σ
- Expected value of $\bar{X}$:

$$E[\bar{X}] = \mu$$

- Standard deviation of $\bar{X}$:

$$stdev[\bar{X}] = \frac{\sigma}{\sqrt{n}}$$

  - Called "**standard error** (SE)"
  - The standard error decreases as the sample size *n*.

Sampling Distribution - The Central Limit Theorem

- The population distribution (distribution of X) is usually unknown.
- However, the sampling distribution (distribution of $\bar{X}$) can be estimated by the central limit theorem (CLT).
  - The CLT is the single most important result in statistics!

The Central Limit Theorem (CLT): $\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$

If the sample size (n) is sufficiently large, then the sample mean $\bar{X}$ is **normally distributed**
(no matter what the distribution of X is!)

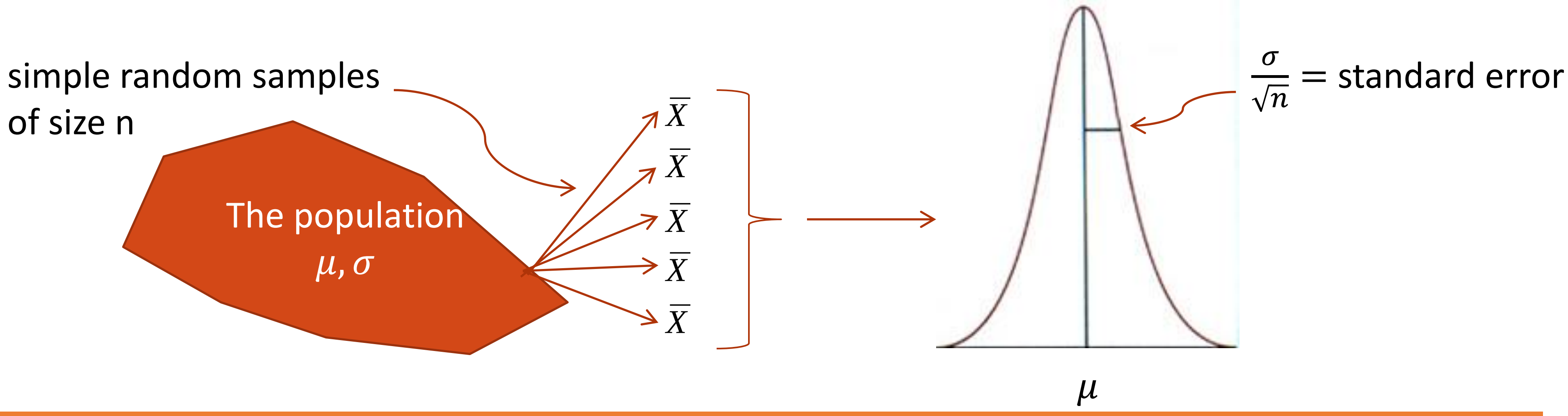
- How large does n have to be to apply the central limit theorem?
  - Typically, the normal approximation is good for **n >= 30**

Sampling Distribution - The Central Limit Theorem

The Central Limit Theorem (CLT): $\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$

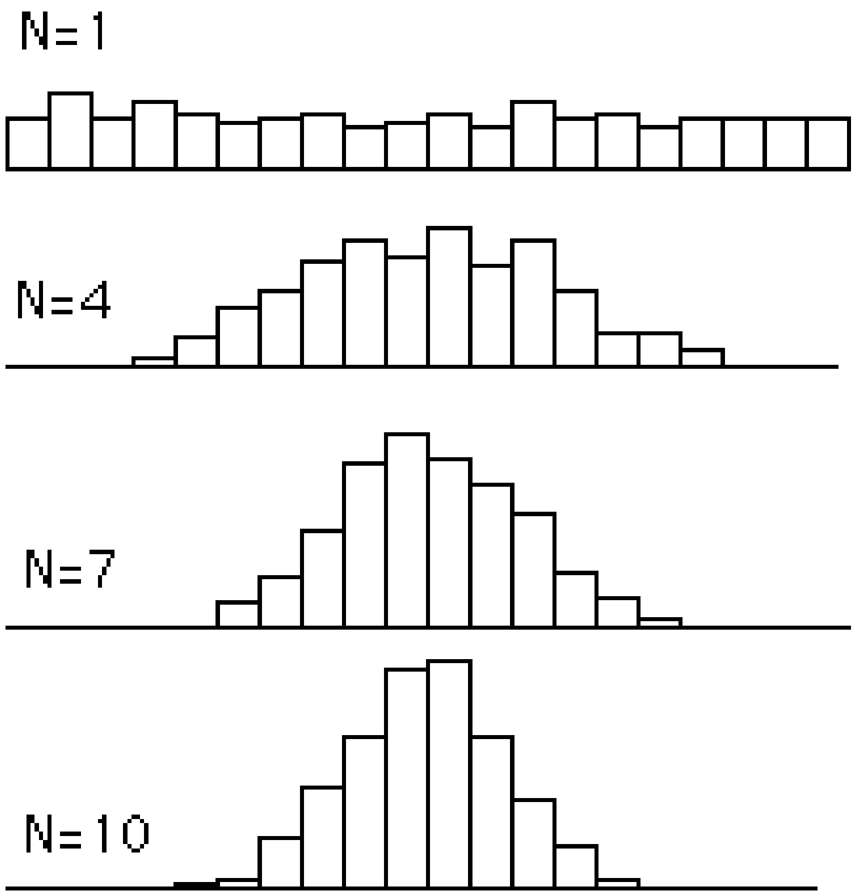
If the sample size (n) is sufficiently large, then the sample mean $\bar{X}$ is **normally distributed**

- CLT allows us to measure the probabilistic accuracy of an estimator



The Central Limit Theorem

- X is uniformly distributed
- As *n* increases, the sampling distribution more closely represents a normal distribution
- In addition, the variance of the sample means gets smaller!



Example: Meal Service

What is the distribution of the average of a random sample of 30 days, **if we believe the contractor's claim**:

- The distribution: The normal distribution
- The mean of sample mean $(E[\bar{X}] = \mu)$: 10,000
- The standard deviation of sample mean ($stdev[\bar{X}]$ = standard error $= \frac{\sigma}{\sqrt{n}}$):

1643.17 /sqrt(30) = 300.00

$$\bar{X} \sim N(\$10{,}000, \$300)$$

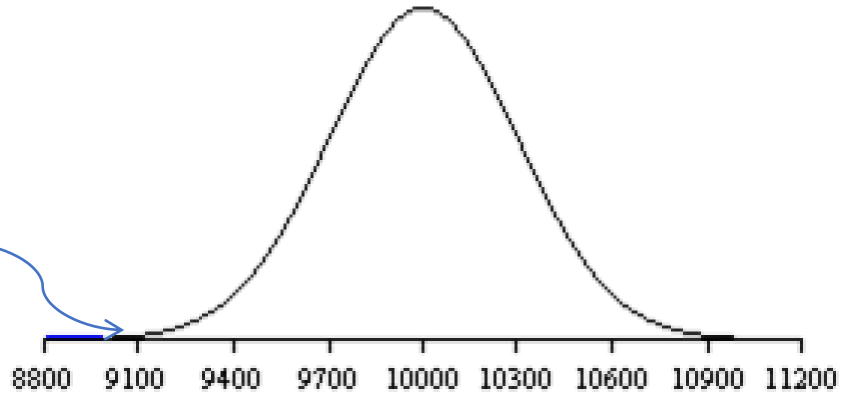
Cost of Service and the CLT

- The sample mean of the auditor (8,983) is far less then the average 10,000 daily meals the contractor claims
- How likely is it to obtain a number as small as 8,983 if the contractor’s claim was true?

Cost of Service and the CLT

Answer: Invoking the CLT, $\bar{X}$ has a normal distribution, with mean 10,000 and a standard deviation of 300.

The number 8,983 is more than three standard deviations away from the mean



$P(\bar{X} \leq 8{,}983) =$ NORM.DIST(*8983*,10000,300,1)=0.00035

Cost of Service and CLT

Based on the previous result, you would…

- believe that the contractor's claim of 10,000 daily meals served.
- not believe that the contractor's claim of 10,000 meals served.
- not be able to come to a conclusion because of small sample size and other missing information.

Recall: Sum of i.i.d. Random Variables (9/27)

- Consider the sum of i.i.d. random variables

$$Y = X_1 + X_2 + \cdots + X_n,$$

where $E(X_i) = \mu$ and $Std(X_i) = \sigma$.

- Note that by the CLT

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{Y}{n} \sim \mathrm{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right).$$

Therefore,

$$Y \approx N(n\mu, \sqrt{n}\sigma).$$

Point Estimate
& Sampling Distribution

Point Estimate of a Population Mean

- A **point estimate** is a single numeric value, a "best guess" of a population parameter, based on the data in a random sample.
  - The point estimate of the population mean is the **sample mean**, the average of the observations in the sample.
  - Denoted by $\bar{X}$.

- The **sampling error** (or **estimation error**) is the difference between the point estimate and the true value of the population parameter being estimated.
  - If $\hat{\theta}$ is a point estimate of θ, the sampling error is $\hat{\theta} - \theta$.
  - It measures how much the point estimate misses the population parameter.
  - Sampling error of sample mean = $\bar{X} - \mu$.

Sampling Distribution of a Sample Mean - Bias

- A **bias** is the difference between the mean of the point estimate and the true value of the population parameter being estimated.
  - If $\hat{\theta}$ is a point estimate of θ, the bias is $\mathrm{E}[\hat{\theta}] - \theta$.

- An **unbiased estimate** is a point estimate such that the mean is equal to the true value of the population parameter being estimated.
- The bias of sample mean is $\mathbf{E}(\bar{X}) - \mu$ **= 0.**
- **Sample mean is an unbiased** estimate of the population mean.

Sampling Distribution of a Sample Mean - SE

- The **standard error (SE) of an estimate** is the standard deviation of the sampling distribution of the estimate.
- It measures how much estimates vary from sample to sample
- The **accuracy of the point estimate** is measured by its standard error

- For sample mean,

$$\mathbf{SE}(\bar{\boldsymbol{X}}) = stdev(\bar{X}) = \frac{\boldsymbol{\sigma}}{\sqrt{\boldsymbol{n}}}$$

- The standard error decreases as the sample size *n*.

Sampling Distribution of a Sample Mean - CLT

- By the CLT, the **sampling distribution** of any point estimate is

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

where σ is the standard deviation of the population and n is the sample size.

- Note: The sampling error $(\bar{\boldsymbol{X}} - \boldsymbol{\mu})$ can be reduced by increasing the sample size *n*:

$$(\bar{\boldsymbol{X}} - \boldsymbol{\mu}) \sim \boldsymbol{N}\left(\boldsymbol{0}, \frac{\boldsymbol{\sigma}}{\sqrt{\boldsymbol{n}}}\right)$$

Example: Meal Service

- The contractors claim that on average 10,000 daily meals were provided. The standard deviation is 1643.17
- The government obtains a random sample of 30 days
  - Average number of meals for 30 days: 8,983 Meals Served
  - 8,983 is the outcome of the point estimate $\bar{X}$

- If we believe the contractors claim:

$$\bar{X} \sim \left(\boldsymbol{\mu}, \frac{\boldsymbol{\sigma}}{\sqrt{\boldsymbol{n}}}\right) = \boldsymbol{N}(\mathbf{10,000,300})$$

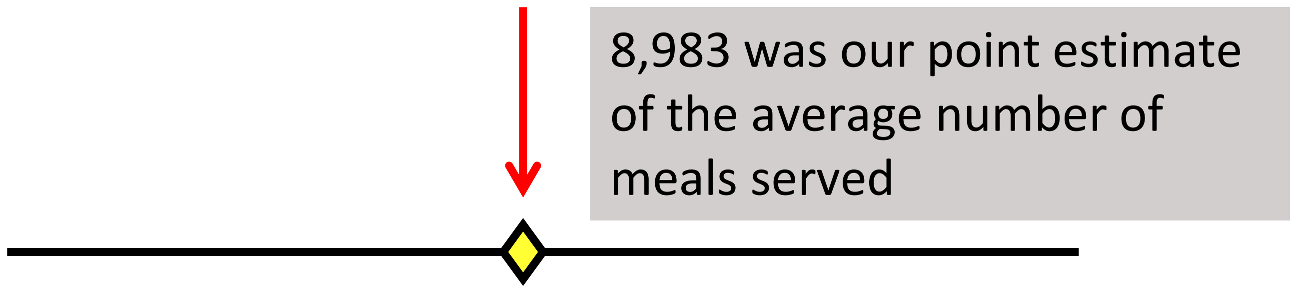
  - Bias = 0
  - Standard error = 300
  - Sampling error = 8,983 – 10,000 = -1,017

CH8
Confidence Interval Estimation

Introduction

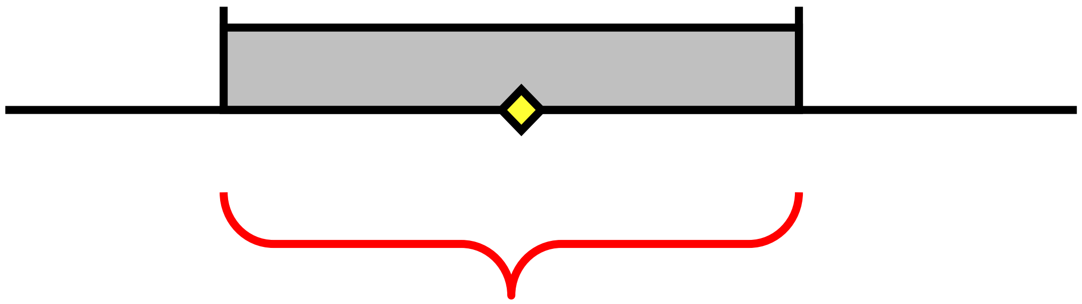
Point Estimator

- A point estimator draws inferences about a population by estimating the value of an unknown parameter using a single value or point
- Sample mean is a pointe estimate of the population mean
    - Example: 8,983 was our point estimate of the average number of meals served
- Disadvantage: Don’t know how good this estimate is



Confidence Interval Estimator

- An interval estimator draws inferences about a population by estimating the value of an unknown parameter using an interval
- A **confidence interval (CI)** is an interval estimator with an attached measure of confidence.



- We say with some ___% certainty that the population parameter of interest is between some lower and upper bounds.

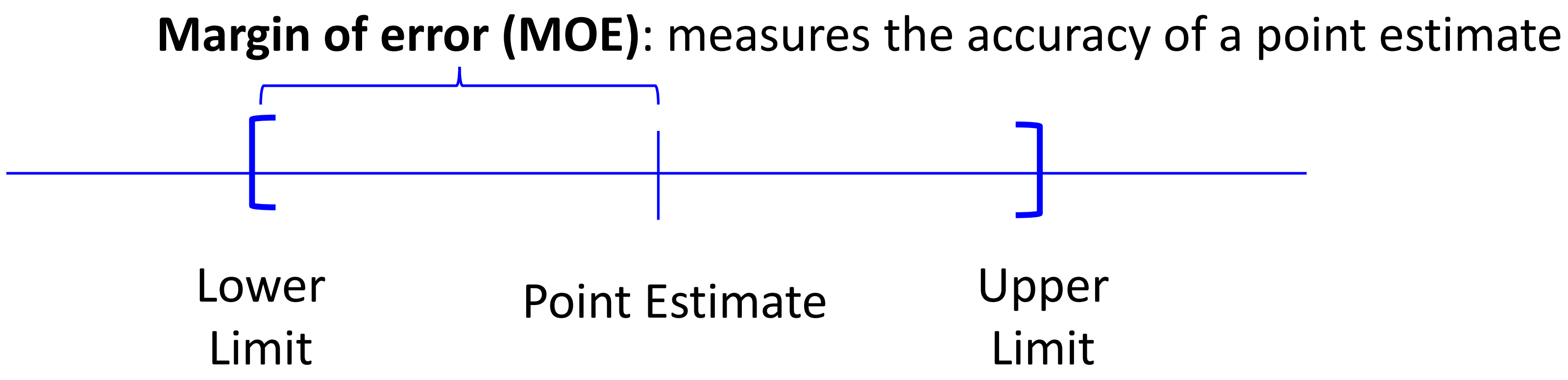
▪ The **confidence level** is usually 90%, 95%, or 99%.

Types of Confidence Intervals

- Given a random sample, we can compute confidence intervals for many population parameters
  - Mean: μ
  - Proportion: p
  - Standard deviation: σ
  - Total: T
  - Difference between means: $\mu_1 - \mu_2$ (Later in Ch 9)
  - Difference between proportions: $p_1 - p_2$
- The process for calculating each type is very similar

Confidence Interval

- In general, a confidence interval is of the form:



$$\boxed{\mathbf{CI = Point\ Estimator \pm Margin\ of\ Error\ (MOE)}}$$

Confidence Interval

- The confidence interval is more commonly written as:

$$\textbf{Point Estimator} \pm \underbrace{(\textbf{Multiple})*(\textbf{Standard Error of Point Estimator})}_{\text{MOE}}$$

- The “**multiple**” Depends on:
  - The distribution of the point estimator
  - The desired confidence level
    - The greater the desired confidence level, the <u>larger</u> the multiple
    - 95% CI is **wider** than 90% CI
- The **standard error** of the point estimator depends on the sample size
  - In general, as *n* increases, the standard error of the point estimator <u>decreases.</u>

CH8
Confidence Interval Estimation

For a Population Mean

Confidence Interval for **Mean** with **known $\boldsymbol{\sigma}$**

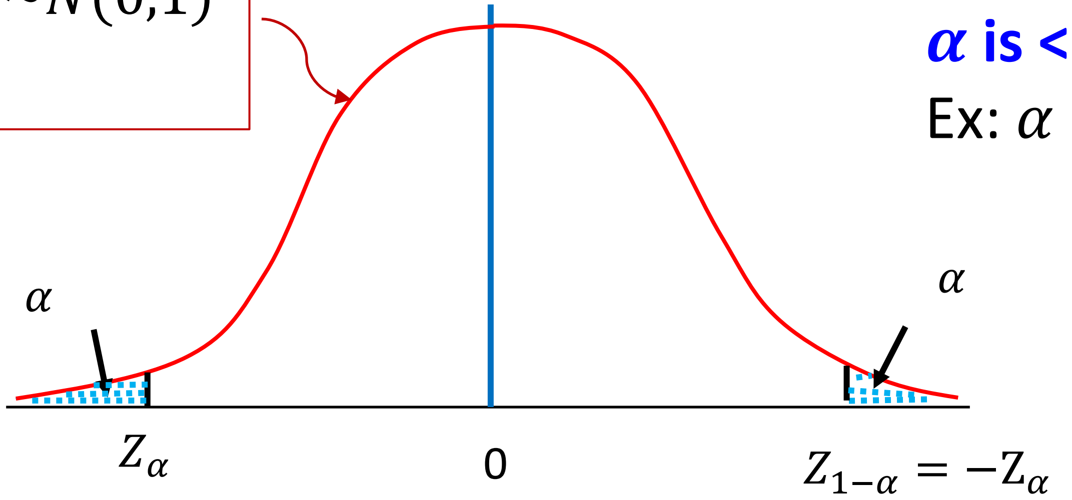
- This is the equation for determining of the confidence interval for a sample mean.

$$\boxed{\bar{X} \pm (Z - multiple) \times \frac{\sigma}{\sqrt{n}}}$$

  - $\bar{X}$: Sample mean, the center of the confidence interval.
  - $Z - multiple$: The z-value is determined by the confidence level.
  - $\frac{\sigma}{\sqrt{n}}$: The standard error of the sample mean estimator, where σ is the standard deviation of the population

Z – multiple (Z_α)

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$



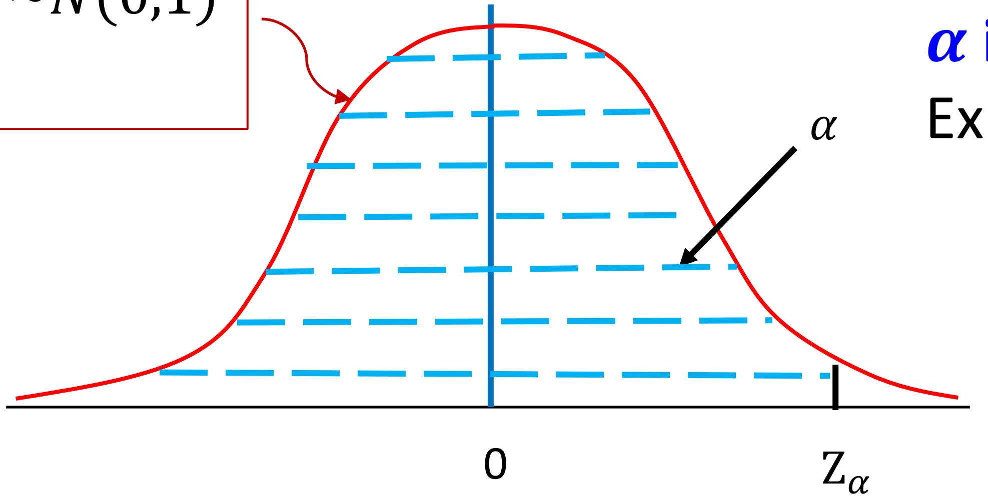
α is < 0.5

Ex: $\alpha = 0.05\ (= 5\%)$

- α is a probability: $0 \le \alpha \le 1$
- Z_α is the $\alpha * 100^{\text{th}}$ percentile, that is,
$$P(Z \le Z_\alpha) = \alpha$$
- In Excel, $Z_\alpha = NORM.S.INV(\alpha)$

Z – multiple (Z_α)

$$\boxed{Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)}$$



***α* is > 0.5**

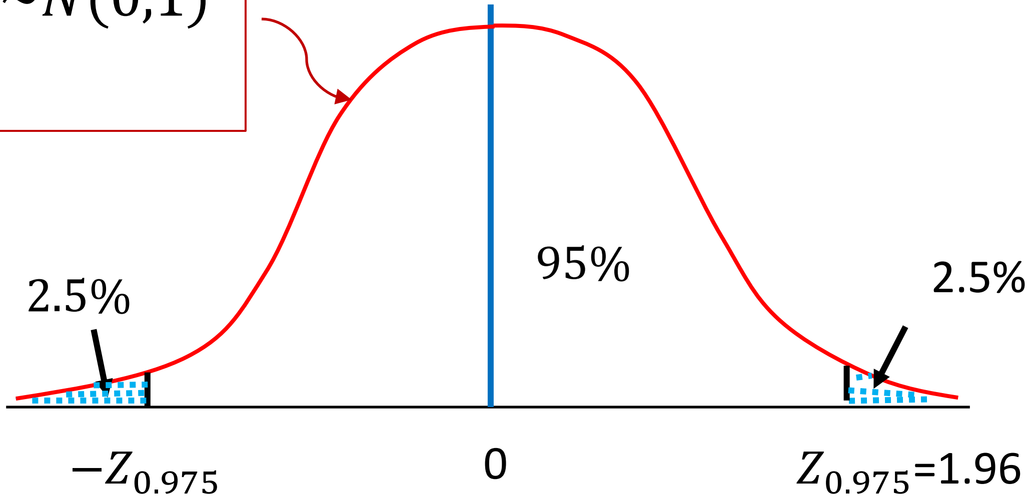
Ex: $\alpha = 0.95$ $(= 95\%)$

- α is a probability: $0 \leq \alpha \leq 1$
- Z_α is the $\alpha * 100^{\text{th}}$ percentile, that is,
$$P(Z \leq Z_\alpha) = \alpha$$
- In Excel, $Z_\alpha = NORM.S.INV(\alpha)$

Z - multiple

Confidence level	**90%**	**95%**	**99%**
$Z - multiple$	$Z_{0.95}$ =1.645	$Z_{0.975}$ =1.96	$Z_{0.995}$ =2.576

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$



Example: 95% CI

$$P(-1.96 \le Z \le 1.96) = 0.95$$

$$\Rightarrow P\left(-1.96 \le \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \le 1.96\right) = 0.95$$

$$\Rightarrow \boldsymbol{P(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + 1.96\ \frac{\sigma}{\sqrt{n}})\ = 0.95}$$

Thus, if σ is known, 95% Confidence Interval $= \bar{X} \pm 1.96\frac{\sigma}{\sqrt{n}}$

Example: Meal Service

- What would be a plausible range for the contractor's average daily meal servings, given our evidence?
- Building a 95% confidence Interval
  - The point estimator: 8,983
  - The standard error is 300
  - The Z-multiple is 1.96
- Plugging into the formula we have

  $8{,}983 \pm 1.96 * 300 = 8{,}983 \pm 588$

- Or written another way on the format [lower, upper]: [8,395, 9,571]
- Interpretation:

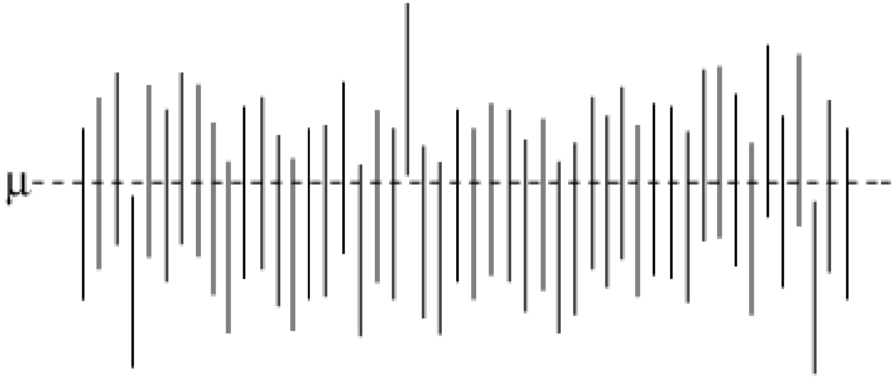
  "We are **95% confident** that the **mean number of meals served** is between 8,395 and 9,571".

Interpretation of Confidence Interval for a Mean

The interpretation of $95\%\ CI\ (\bar{X} \pm 1.96\frac{\sigma}{\sqrt{n}})$ is that

"$\boldsymbol{CI}$ covers the true mean $\boldsymbol{\mu}$ with 95% of chance"

Note that CI is random. If we generate *100* confidence intervals, we would expect 95 (95%)-CIs to contain μ



Interpretation of Confidence Interval for a Mean

- Suppose that a random sample of 100 observations is given and its 95% CI for the mean μ is (0.1, 0.9).
- What is the point estimate of μ?
- What is MOE?

- Can we say that (0.1, 0.9) includes μ with 95% probability (chance)?
- In mathematical expression, this means

$$P \text{ (} \mu \text{ is in } (0.1,0.9) \text{)} = 0.95 \text{ ?}$$

- The answer is ….

Interpretation of Confidence Interval for a Mean

- Then, what is the meaning of $(0.1, 0.9)$?
- $(0.1, 0.9)$ is just an **estimate** of the interval that covers the true mean with 95% probability (chance).
- We say that

" **we are 95% confident that the true mean is in** $\mathbf{(0.1, 0.9)}$".

- We can have an information about **the precision of the point estimate** (=0.5) via the **MOE** (=0.4) **of the CI**.

Interpretation of Confidence Interval for a Mean

- Then, can we say that "we expect 95 observations out of 100 fall within $(0.1, 0.9)$"?
  - The 95% confidence interval for the mean (or any other population parameter) DOES NOT mean that 95% of random samples will fall within the interval

- "95% confidence" that the true value is within a range has no corresponding probability to consider. The population is not repeated, and it is just one outcome.
- Rather, this 95% confidence characterizes our personal feeling of uncertainty. For the argument to work, however, that confidence needs to be a probability, even if it cannot be defined through a probability. It can be considered as a "subjective probability".

from "data analysis for business, economics, and policy" by Bekes and Kezdi

Next …

- CI for population mean with unknown standard deviation
- Other point estimates and CIs