

Statistical and Mathematical Methods for Data Analysis

Dr. Syed Faisal Bukhari

Associate Professor

Department of Data Science

Faculty of Computing and Information Technology

University of the Punjab

Textbooks

- ❑ **Probability & Statistics for Engineers & Scientists**, Ninth Edition, Ronald E. Walpole, Raymond H. Myer
- ❑ **Elementary Statistics: Picturing the World**, 6th Edition, Ron Larson and Betsy Farber
- ❑ **Elementary Statistics**, 13th Edition, Mario F. Triola

Textbooks

- ❑ **Probability & Statistics for Engineers & Scientists**, Ninth Edition, Ronald E. Walpole, Raymond H. Myer
- ❑ **Elementary Statistics: Picturing the World**, 6th Edition, Ron Larson and Betsy Farber
- ❑ **Elementary Statistics**, 13th Edition, Mario F. Triola

Reference books

- ❑ **Probability and Statistical Inference, Ninth Edition,** Robert V. Hogg, Elliot A. Tanis, Dale L. Zimmerman
- ❑ **Probability Demystified,** Allan G. Bluman
- ❑ **Practical Statistics for Data Scientists: 50 Essential Concepts,** Peter Bruce and Andrew Bruce
- ❑ **Schaum's Outline of Probability,** Second Edition, Seymour Lipschutz, Marc Lipson
- ❑ **Python for Probability, Statistics, and Machine Learning,** José Unpingco

References

Readings for these lecture notes:

- ❑ **Schaum's Outline of Probability, Second Edition (Schaum's Outlines)** by by Seymour Lipschutz, Marc Lipson
- ❑ **Probability & Statistics for Engineers & Scientists**, Ninth Edition, Ronald E. Walpole, Raymond H. Myer
- ❑ <https://wordwatchtowers.wordpress.com/2009/12/21/underestimate-or-overestimate/>
- ❑ Elementary Statistics, Tenth Edition, Mario F. Triola
- ❑ <http://www.sjsu.edu/faculty/gerstman/>

These notes contain material from the above resources.

Populations and Samples

- ❑ The totality of observations with which we are concerned, whether their number be finite or infinite, constitutes what we call a **population**.
- ❑ A **population** consists of the totality of the observations with which we are concerned.
- ❑ A **sample** is a subset of a population.

Bias

Any **sampling procedure** that produces inferences that consistently **overestimate** or consistently **underestimate** some characteristic of the population is said to be **biased**.

To eliminate any **possibility of bias** in the sampling procedure, it is desirable to choose a **random sample** in the sense that the observations are made **independently** and at **random**.

Overestimate vs. Underestimate

- ❑ **Overestimate** means 'to form too high an estimate of'
- ❑ **Underestimate** means to estimate that something is smaller or less important than it actually is

Parameter vs. Statistic

- ❑ **Statistical inference** involves drawing conclusions about **characteristics of populations**.
- ❑ Among these characteristics are constants which are called **population parameters**. Two important parameters are the **population mean** and the **population variance**.
- ❑ Any function of the random variables constituting a **random sample** is called a **statistic**.

Sampling Distribution [1]

- ❑ The probability distribution of a statistic is called a **sampling distribution**.
- ❑ The field of **statistical inference** is basically concerned with **generalizations** and **predictions**.
- ❑ For example, we might claim, based on the opinions of several people interviewed on the street, that in a forthcoming election **60% of the eligible voters** in the city of Detroit favor a certain candidate. In this case, **we are dealing with a random sample of opinions from a very large finite population**.

Sampling Distribution [2]

- ❑ As a second illustration we might state that the **average cost** to build a residence in Charleston, South Carolina, is between **\$330,000 and \$335,000**, based on the **estimates of 3 contractors** selected at random from the 30 now building in this city. The population being sampled here is **again finite** but **very small**.

Sampling Distribution [3]

□ Finally, let us consider a **soft-drink machine** designed to dispense, on average, **240 milliliters** per drink. A company official who computes the mean of **40** drinks obtains $\bar{x} = 236$ milliliters and, on the basis of this value, decides that the machine is still dispensing drinks with an average content of $\mu = 240$ milliliters. The **40** drinks represent a sample from the **infinite population** of possible drinks that will be dispensed by this machine.

Inference about the Population from Sample Information [1]

- ❑ In each of the examples above, we computed a **statistic** from a **sample selected** from the **population**, and from this **statistic** we made various statements concerning the values of **population parameters** that **may or may not be true**.
- ❑ The company official made the decision that the soft-drink machine dispenses drinks with an average content of **240 milliliters**, even though the sample mean was **236 milliliters**, because he knows from sampling theory that, if $\mu = 240$ milliliters, such a sample value could easily occur.

Inference about the Population from Sample Information [2]

- ❑ In fact, if he ran similar tests, say every hour, he would expect the values of the statistic \bar{x} to fluctuate above and below $\mu = 240$ milliliters. Only when the value of \bar{x} is substantially different from 240 milliliters will the company official initiate action to adjust the machine.
- ❑ Since a statistic is a **random variable** that depends only on the observed sample, it must have a **probability distribution**.
- ❑ The probability distribution of a **statistic** is called a **sampling distribution**.

Sampling Distribution of a Statistic

The **sampling distribution of a statistic** (such as a **sample proportion** or **sample mean**) is the distribution of all values of the statistic when **all possible samples** of the **same size n** are taken from the same population.

Sampling Distribution of the Mean

The **sampling distribution of the mean** is the distribution of **sample means**, with **all samples** having the **same sample size n** taken from the **same population**.

Sampling Distribution of the Proportion

- The **sampling distribution of the proportion** is the distribution of **sample proportions**, with all samples having the **same sample size n** taken from the **same population**.

The Central Limit Theorem

Central Limit Theorem: If \bar{X} is the mean of a random sample of size n taken from a **population** with **mean μ** and **finite variance σ^2** , then the limiting form of

the distribution of $Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$,

as $n \rightarrow \infty$, is the **standard normal distribution $n(z; 0, 1)$** .

The Central Limit Theorem

- ❑ The normal approximation for \bar{X} will generally be good if $n \geq 30$, provided the population distribution is not **terribly skewed**.
- ❑ If $n < 30$, the approximation is good only if the population is **not too different** from a **normal distribution**.
- ❑ As stated above, if the **population is known to be normal**, the **sampling distribution of \bar{X}** will follow a **normal distribution** exactly, no matter how **small the size of the samples**.

The Central Limit Theorem

- ❑ The sample size $n = 30$ is a guideline to use for the **Central Limit Theorem**.
- ❑ However, as the statement of the theorem implies, the presumption of normality on the distribution **of \bar{X}** becomes more accurate **as n grows larger**.

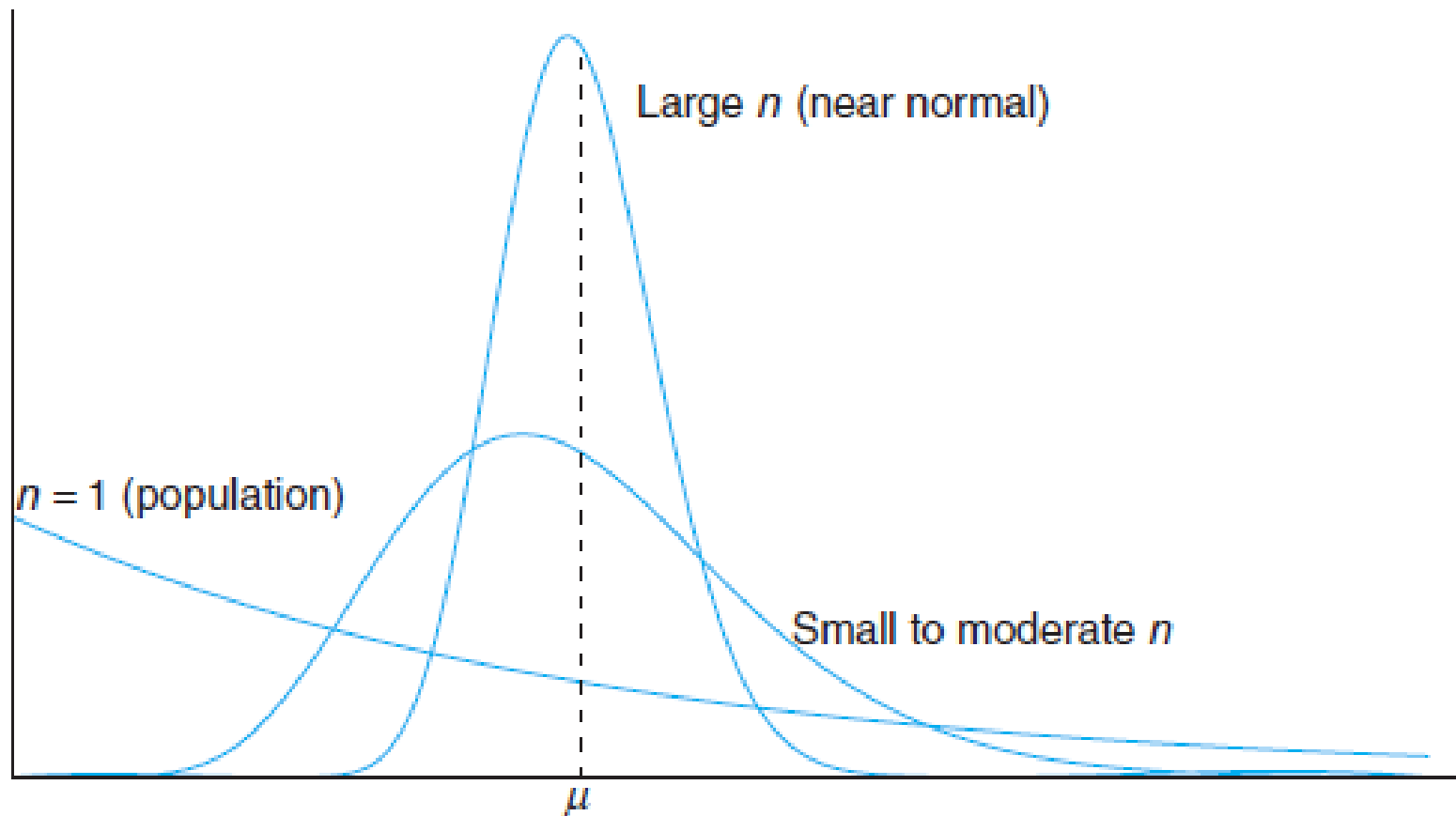


Illustration of the Central Limit Theorem
(distribution of \bar{X} for $n = 1$, moderate n , and large n).

The Central Limit Theorem

When selecting a **simple random sample** from a **population** with **mean** and **standard deviation**, it is essential to know these principles:

1. If **$n > 30$** , then the **sample means** have a distribution that can be **approximated by a normal distribution** with **mean μ** and **standard deviation σ / \sqrt{n}** (This guideline is commonly used, regardless of the distribution of the original population.)
2. If **$n \leq 30$** and the **original population** has a **normal distribution**, then the **sample means** have a normal distribution with mean **μ** and standard deviation **σ / \sqrt{n}**

The Central Limit Theorem

3. If $n \leq 30$ but the **original population** does not have a **normal distribution**, then the methods of this section do not apply.

□ Try to keep this big picture in mind: As we sample from a population, we want to know the behavior of the sample means.

□ The ***central limit theorem*** tells us that if the **sample size is large enough**, the distribution of **sample means** can be **approximated by a normal distribution**, even if the **original population is not** normally distributed.

The Central Limit Theorem and the Sampling Distribution of \bar{x}

Given:

1. The **random variable x** has a **distribution (which may or may not be normal)** with mean μ and standard deviation σ .
2. **Simple random samples** all of the **same size n** are **selected from the population**. (The samples are selected so that all possible samples of size n have the same chance of being selected.)

The Central Limit Theorem and the Sampling Distribution of \bar{x}

Conclusions:

1. The distribution of **sample means \bar{x}** will, as the sample size increases, approach a ***normal distribution***.
2. The **mean of all sample means** is the population mean (That is, the normal distribution from Conclusion 1 has mean μ).
3. The **standard deviation** of all **sample means** is **σ / \sqrt{n}** (That is, the normal distribution from Conclusion 1 has standard deviation **σ / \sqrt{n}**)

The Central Limit Theorem and the Sampling Distribution of \bar{x}

Practical Rules Commonly Used

If the original population is not itself normally distributed, here is a common guideline:

1. For samples of **size n greater than 30**, the distribution of the **sample means** can be **approximated reasonably well** by a **normal distribution**. (There are **exceptions**, such as **populations** with very **nonnormal** distributions requiring sample **sizes larger than 30**, but such exceptions **are relatively rare**.) **The approximation gets better as the sample size n becomes larger.**

The Central Limit Theorem and the Sampling Distribution of \bar{x}

Practical Rules Commonly Used

2. If the **original population** is itself **normally distributed**, then the **sample means** will be **normally distributed** for **any** sample size n (not just the values of n larger than 30).

Notation for Sampling Distribution of \bar{x}

The **central limit theorem** involves **two different distributions**: the distribution of the **original population** and the distribution of the **sample means**.

□ We use the symbols μ and σ to denote the mean and standard deviation of the original population, but we use the following **new notation** for the **mean** and **standard deviation** of the **distribution of sample means**.

$$\mu_{\bar{X}} = \mu \text{ and}$$

$$\sigma_{\bar{X}} = \sigma / \sqrt{n}$$

Example Water Taxi Safety

We noted that some passengers died when a water taxi sank in Baltimore's Inner Harbor. **Men are typically heavier than women and children, so when loading a water taxi, let's assume a worst-case scenario in which all passengers are men.** Based on data from the National Health and Nutrition Examination Survey, assume that weights of men are normally distributed with a **mean of 172 lb** and a **standard deviation of 29 lb**.

Example Water Taxi Safety cont.

- a. Find the probability that if an individual man is randomly selected, his weight will be **greater than 175 lb**.
- b. Find the probability that **20** randomly selected men will have a **mean that is greater than 175 lb** (so that their total weight exceeds the safe capacity of 3500 lb).

$\mu = 172$ and $\sigma = 29$

a) $Z = \frac{x - \mu}{\sigma} = Z = \frac{175 - 172}{29} = \mathbf{0.10}$

$$P(X > 175) = P(Z > 0.10) = 1 - P(Z < 0.10) = 1 - 0.5398$$

$P(X > 175) = 0.4602$ ans

b) $Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} = \frac{175 - 172}{29 / \sqrt{20}} = \frac{3}{6.4846} = \mathbf{0.46}$

$$P(\bar{x} > 175) = P(z > 0.46) = 1 - P(z \leq 0.46) \\ = 1 - 0.6772$$

$P(\bar{x} > 175) = 0.3228$ ans

Basics of inference[1]

- ❑ **Statistical inference** is the act of **generalizing** from a **sample to a population** with calculated degree of certainty. The two forms of statistical inference are **estimation** and **hypothesis testing**.
- ❑ A statistical **population** represents the set of all possible values for a variable. In practice, we do not study the entire population.

Basics of inference[2]

- ❑ Instead, we use data in a **sample** to shed light on the wider population.
- ❑ The term **parameter** is used to refer to a numerical **characteristic** of a **population**. Examples of parameters include the **population mean (μ)** and the population **standard deviation (σ)**.

Basics of inference[3]

- ❑ A numerical characteristic of the **sample** is a statistic.
- ❑ We introduce a particular type of **statistic** called an **estimate**. The **sample mean \bar{x}** is the natural estimator of **population mean μ** . Sample standard **deviation s** is the natural estimator of population **standard deviation σ** .
- ❑ The parameter is a fixed **constant**. In contrast, the **estimator varies from sample to sample**.

Basics of inference[4]

	Parameter	Estimators
Source	Population	Sample
Value known?	No	Yes (calculate)
Notation	Greek (μ)	Roman (\bar{x})
Vary from sample to sample	No	Yes
Error-prone	No	Yes

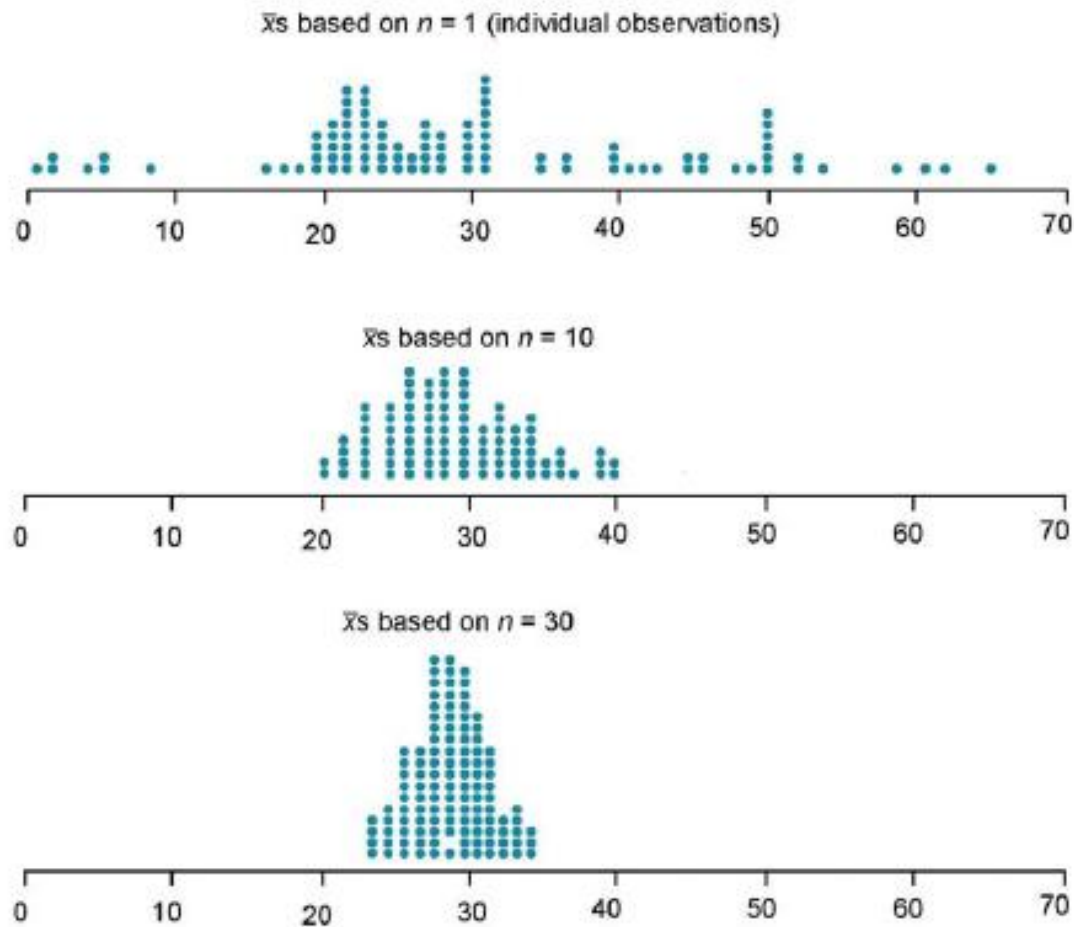
Sampling distribution of a mean (SDM)

- ❑ If we had the opportunity to take repeated samples from the same population, samples means (\bar{x} s) would vary from sample to sample and form a **sampling distribution means (SDM)**.
- ❑ Let's run a simulation experiment. Our simulation will be based on sampling a population of **$N = 600$** age values. The population mean **age $\mu = 29.5$** . The population standard deviation **$\sigma = 13.6$**

Sampling distribution of a mean (SDM)

- Imagine taking repeated samples, each of $n = 10$. Do this **100 times**.
- In one experiment, it just so happened that the first \bar{x} was **36.4**, the second \bar{x} was **30.2**, and the third \bar{x} was **24.6**

Sampling distribution of \bar{x}



Statistical Inference

- ❑ **Statistical inference** consists of those methods by which one makes **inferences or generalizations** about a **population**.
- ❑ The trend today is to distinguish between the **classical method** of estimating a **population parameter**, whereby inferences are based strictly on information obtained from a **random sample** selected from the **population**.
- ❑ Statistical inference may be divided into two major areas: **estimation** and **tests of hypotheses**.

Estimation

A candidate for public office may wish to **estimate** the true **proportion** of voters favoring him by obtaining opinions from a **random sample of 100** eligible voters.

The fraction of voters in the sample favoring the candidate could be used as an estimate of the **true proportion in the population** of voters.

A knowledge of the **sampling distribution** of a **proportion** enables one to **establish the degree of accuracy** of such an **estimate**. This problem falls in the **area of estimation**.

Tests of Hypotheses

A floor wax is more scuff-resistant than brand B floor wax. He or she might hypothesize that **brand A is better than brand B** and, after proper testing, accept or reject this hypothesis.

In this example, we do not attempt to **estimate a parameter**, but instead we try to arrive at a correct decision about a **prestated hypothesis**.

Once again we are dependent on **sampling theory** and the use of data to provide us with some measure of accuracy for our decision.

Point Estimate [1]

A **point estimate** of some population parameter θ is a single value of a statistic $\hat{\theta}$.

For example, the value \bar{x} of the statistic \bar{X} , computed from a sample of size n , is a point estimate of the population parameter μ . Similarly, $\hat{p} = x/n$ is **a point estimate** of the **true proportion** p for a binomial experiment.

An estimator is not expected to estimate the population parameter **without error**. We do not expect \bar{X} to estimate μ exactly, but we certainly hope that **it is not far off**.

Point Estimate[2]

For a particular sample, it is possible to obtain a closer estimate of μ by using the sample median \tilde{X} as an estimator. Consider, for instance, a sample consisting of the values **2, 5, and 11** from a population whose **mean is 4 ($\mu = 4$)** but is supposedly unknown.

We would estimate μ to **be $\bar{x} = 6$** , using the sample mean as our estimate, or **$\tilde{x} = 5$** , using the sample median as our estimate. In this case, the estimator \tilde{X} produces an estimate closer to the true parameter than does the estimator \bar{X} .

Point Estimate[3]

On the other hand, if our random sample contains the values **2, 6, and 7**, then $\tilde{\bar{x}} = 6$ and $\bar{x} = 5$, so \bar{x} is the better estimator.

Not knowing the true value of μ , we must decide in advance whether to use \bar{x} or $\tilde{\bar{x}}$ as our estimator.

Unbiased Estimator

A statistic $\hat{\theta}$ is said to be an **unbiased estimator** of the parameter θ if $\mu_{\hat{\theta}} = E(\hat{\theta}) = \theta$

Unbiased Estimator

Example 1: $E(\bar{X}) = \mu$, so \bar{X} is an **unbiased estimator** of μ

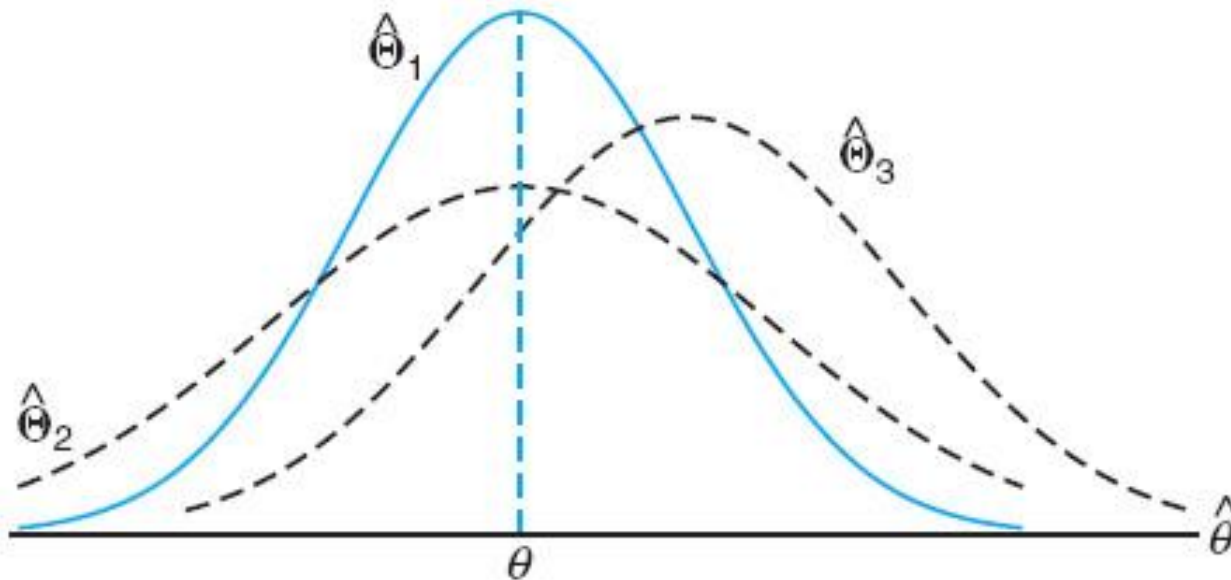
Example 2: $E(s^2) = \sigma^2$, so s^2 is an **unbiased estimator** of σ^2

Variance of a Point Estimator [1]

- If $\hat{\theta}_1$ and $\hat{\theta}_2$ are two unbiased estimators of the same population parameter θ , we want to choose the **estimator whose sampling distribution has the smaller variance.**
- Hence, if $\sigma^2_{\hat{\theta}_1} < \sigma^2_{\hat{\theta}_2}$, we say that $\hat{\theta}_1$ is a **more efficient estimator** of θ than $\hat{\theta}_2$.

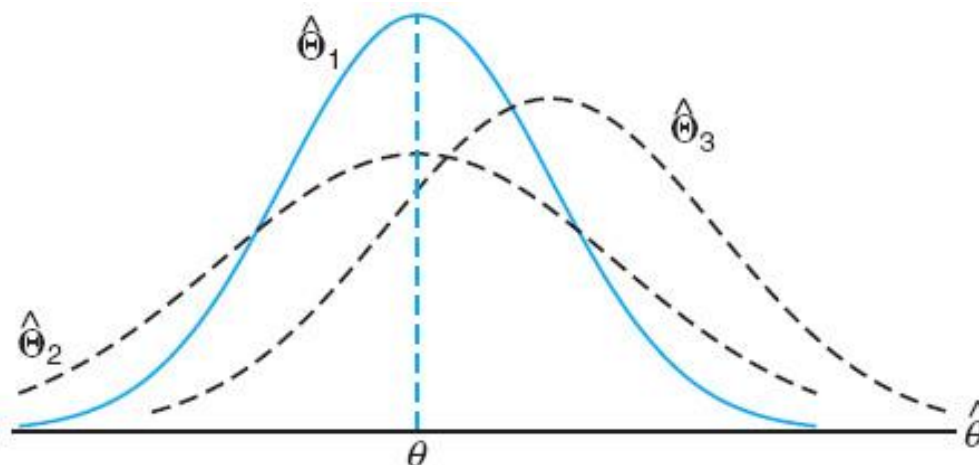
Variance of a Point Estimator [2]

□ If we consider all possible unbiased estimators of some parameter θ , the one with the smallest variance is called the **most efficient estimator** of θ .



Variance of a Point Estimator [3]

The figure illustrates the sampling distributions of three different estimators, $\hat{\theta}_1$, $\hat{\theta}_2$, and $\hat{\theta}_3$, all estimating θ . It is clear that only $\hat{\theta}_1$ and $\hat{\theta}_2$ are **unbiased**, since their distributions are centered at θ . The estimator $\hat{\theta}_1$ has a smaller variance than $\hat{\theta}_2$ and is therefore more efficient. Hence, our choice for an estimator of θ , among the three considered, would be $\hat{\theta}_1$.



Variance of a Point Estimator [2]

For normal populations, one can show that both \bar{X} and \tilde{X} are **unbiased estimators** of the population mean μ , but the variance of \bar{X} is smaller than the variance of \tilde{X} .

Thus, both estimates \bar{x} and \tilde{x} will, on average, equal the population mean μ , but \bar{x} is likely to be closer to μ for a given sample, and thus \bar{X} is **more efficient** than \tilde{X} .