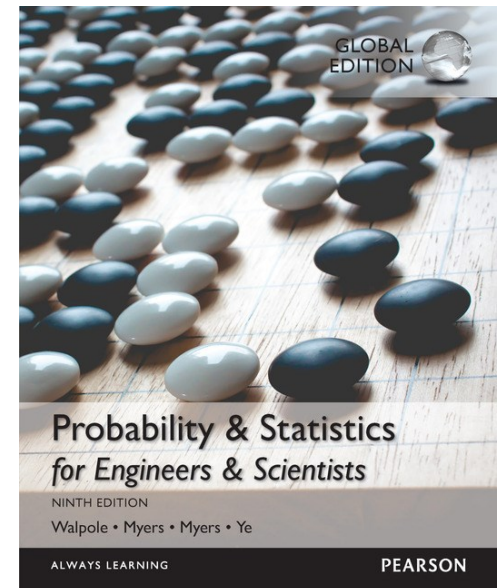


Chapter 9

One- and Two- Sample Estimation (1/2)



- The theory of statistical inference consists of some methods by which one **makes inferences or generalizations about a population**.
- **Classical Method**
 - Estimating a population parameter and making inferences are based strictly on information obtained from a **random sample** selected from the population.
- **Bayesian Method**
 - Utilizing **prior subjective knowledge** about the probability distribution of the unknown parameters in conjunction with the information provided by the **sample data**.

Statistical Inference

- **Estimation** (Covering this chapter, some in last chapter)
 - Taking a random sample from the distribution to **elicit some information about the unknown parameter θ** .
 - Example
 - A candidate for public office may wish to **estimate** the true proportion of voters favoring him by obtaining the opinions from a random sample of 100 eligible voters
- **Hypothesis Testing** (Next chapter)
 - We do not attempt to estimate a parameter, but instead we try to arrive at a **correct decision about a pre-stated hypothesis**.
 - Example
 - One is interested in finding out whether brand 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.

Classical Methods of Estimation

- **Point Estimate**

- A point estimate is single value for a population parameter.
 - *Examples of parameters : mean, median, variance, ...

- **Interval Estimate**

- An interval estimate is an interval, or range of values, used to estimate a population parameter.

- Point Estimate

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

- Example 1

The value of \bar{x} of the statistic \bar{X} , computed from a sample of size n , is a point estimate of the population parameter μ .

- **An estimator** is **not** expected to estimate the population parameter without error, but **we certainly hope that it is not far off**.

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

Good estimator ?

- What are the desirable properties of a “good” decision function that would influence us to choose one estimator rather than another?

Definition 9.1

- **Unbiased Estimator**

A statistic $\hat{\Theta}$ is said to be an **unbiased estimator** of the parameter θ if

$$\mu_{\hat{\Theta}} = E(\hat{\Theta}) = \theta.$$

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

Example 9.1:

Point Estimate

Interval Estimate

- Show that \mathbf{S}^2 is an unbiased estimator of the parameter σ^2 .

Recall from "Sampling_p2" Distribution of the Statistic $(n-1)S^2/\sigma^2$

$$\begin{aligned}\sum_{i=1}^n (X_i - \mu)^2 &= \sum_{i=1}^n [(X_i - \bar{X}) + (\bar{X} - \mu)]^2 \\&= \sum_{i=1}^n (X_i - \bar{X})^2 + \sum_{i=1}^n (\bar{X} - \mu)^2 + 2(\bar{X} - \mu) \sum_{i=1}^n (X_i - \bar{X}) \\&= \sum_{i=1}^n (X_i - \bar{X})^2 + n(\bar{X} - \mu)^2.\end{aligned}$$

Thus,

$$\sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{i=1}^n (X_i - \mu)^2 - n(\bar{X} - \mu)^2.$$

Cont.

$$E(S^2) = E \left[\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 \right]$$

$$\sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{i=1}^n (X_i - \mu)^2 - n(\bar{X} - \mu)^2.$$

$$= \frac{1}{n-1} \left[\sum_{i=1}^n E(X_i - \mu)^2 - nE(\bar{X} - \mu)^2 \right] = \frac{1}{n-1} \left(\sum_{i=1}^n \sigma_{X_i}^2 - n\sigma_{\bar{X}}^2 \right).$$

However,

$$\sigma_{X_i}^2 = \sigma^2, \text{ for } i = 1, 2, \dots, n, \text{ and } \sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}.$$

Therefore,

$$E(S^2) = \frac{1}{n-1} \left(n\sigma^2 - n\frac{\sigma^2}{n} \right) = \sigma^2.$$

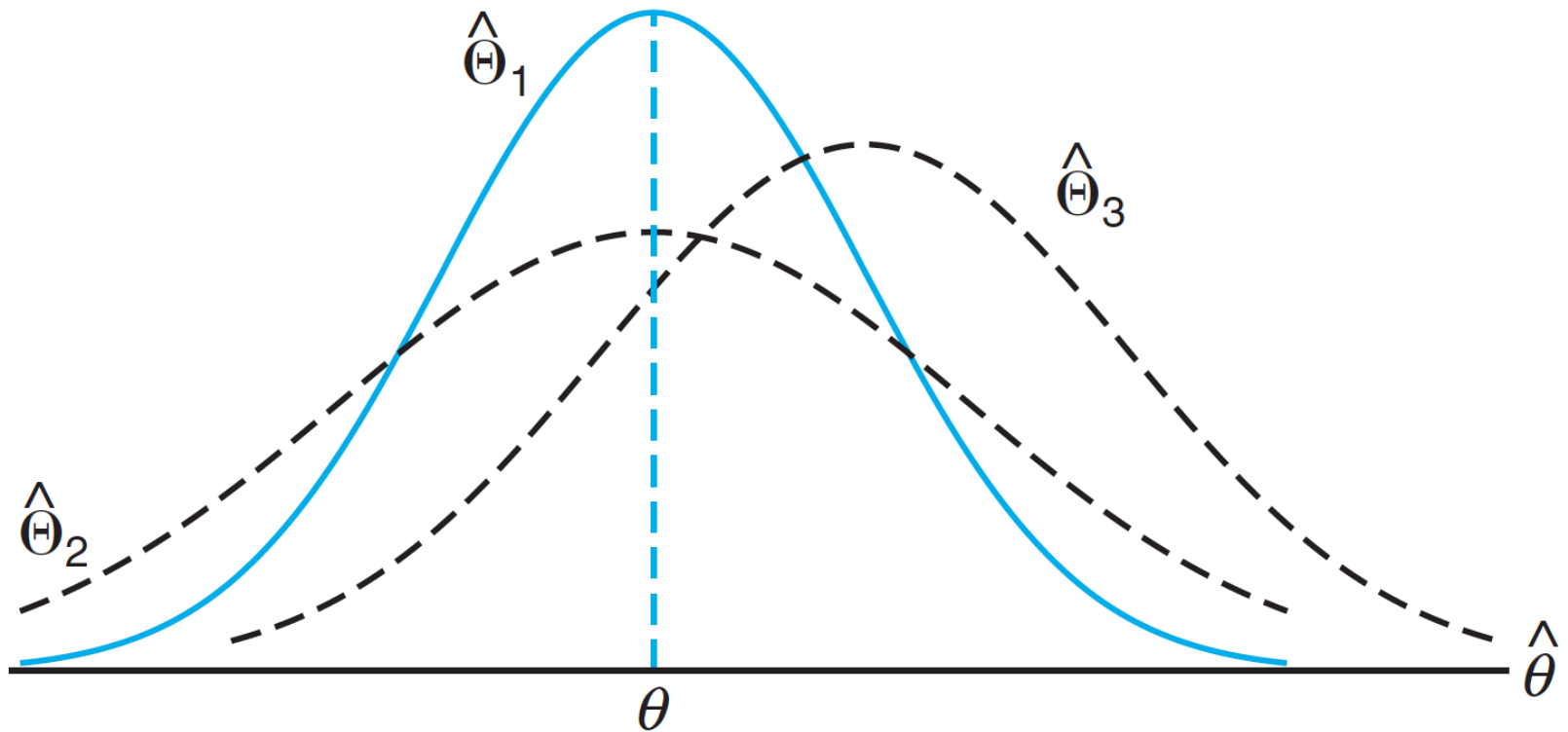
This example illustrates **why we divide by $n - 1$** rather than n when the variance is estimated!

Variance of a Point Estimator?

Most Efficient Estimator

- 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 .

If $\hat{\Theta}_1$ and $\hat{\Theta}_2$ are two unbiased estimators of the same population parameter θ and $\sigma_{\hat{\Theta}_1}^2 < \sigma_{\hat{\Theta}_2}^2$, we say that $\hat{\Theta}_1$ is **more efficient estimator** of θ than $\hat{\Theta}_2$.



Sampling distributions of different estimators of θ

- Question

- A Mayor wants to know the average zinc (아연) concentration in a river.
- How can we explain the result to the mayor ?

- Measurement

- The average zinc (아연) concentration recovered from a sample of measurements taken in 36 different locations in a river is found to be 2.6 grams per milliliter.

- One Way

- $P(\text{value1} < \mu < \text{value2}) = \alpha$

Interval Estimation

- Even the most efficient unbiased estimator is unlikely to estimate the population parameter exactly. It is true that **our accuracy increase with large samples**, but there is still no reason why we should expect a **point estimate** from a given sample **to be exactly equal to the population parameter** it is supposed to estimate.
- There are many situations in which it is preferable to determine an interval within which we would expect to find the value of the parameter. Such an interval is call **interval estimate**.

Interval Estimation

- Example

A random sample of SAT verbal scores for the students of entering freshman class might produce an interval from 530 to 550 within which we expect to find the true average of all SAT verbal scores for the freshman class. The values of the endpoints, 530 and 550, will depend on the computed sample mean \bar{x} and the sampling distribution of \bar{X} . As the sample size increases, we know that $\sigma_{\bar{X}}^2 = \sigma^2/n$ decreases, and consequently our estimate is likely to be closer to the parameter σ , by its length, the accuracy of the point estimate.

Thus, the interval estimate indicates, by its length (interval), the accuracy of the point estimate

Interpretation of Interval Estimation

Since different samples will generally yield different values of $\hat{\Theta}$ and, therefore, different values for $\hat{\theta}_L$ and $\hat{\theta}_U$, these endpoints of the interval are values of corresponding random variables $\hat{\Theta}_L$ and $\hat{\Theta}_U$. From the sampling distribution of $\hat{\Theta}$ we shall be able to determine $\hat{\Theta}_L$ and $\hat{\Theta}_U$ such that $P(\hat{\Theta}_L < \theta < \hat{\Theta}_U)$ is equal to any positive fractional value we care to specify. If, for instance, we find $\hat{\Theta}_L$ and $\hat{\Theta}_U$ such that

$$P(\hat{\Theta}_L < \theta < \hat{\Theta}_U) = 1 - \alpha,$$

for $0 < \alpha < 1$, then we have a probability of $1 - \alpha$ of selecting a random sample that will produce an interval containing θ .

Confidence Interval

$$P(\hat{\Theta}_L < \theta < \hat{\Theta}_U) = 1 - \alpha,$$

- The interval $\hat{\theta}_L < \theta < \hat{\theta}_U$, computed from the selected sample, is called a **100(1 - α)% confidence interval**, the fraction $1 - \alpha$ is called the **confidence coefficient** or the degree of confidence, and the endpoints, $\hat{\theta}_L$ and $\hat{\theta}_U$, are called the **lower and upper confidence limits**.

The wider the confidence interval is, the more confident we can be that the given interval contains the unknown parameter

Example

- It is better to be 95% confident that the average life of a certain television transistor is between 6 and 7 years than to be 99% confident that it is between 3 and 10 years.
- Ideally, we prefer a

Short
or
Long

 interval with a

Low
or
high

 degree of confidence. Sometimes, restrictions on the size of our sample prevent us from achieving short intervals without sacrificing some of our degree of confidence.

Single Sample: Estimating the Mean

Sample Mean is a good estimator

- Statistic \bar{X} is an unbiased estimator
- Its variance is smaller than that of any other estimators of μ

$\sigma_{\bar{X}}^2 = \sigma^2/n$, so a large sample will yield a value of \bar{X} that comes from a sampling distribution with a small variance. Hence, \bar{x} is likely to be a very accurate estimate of μ when n is large.

If n is sufficiently large, according to the **central limit theorem**, we can establish a confidence interval for μ by considering the sampling distribution \bar{X} . We expect the sampling distribution of \bar{X} to be approximately normally distributed with mean $\mu_{\bar{X}} = \mu$ and standard deviation $\sigma_{\bar{X}} = \sigma^2/\sqrt{n}$. Then, we have

$$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha,$$

where

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}.$$

Hence

$$P\left(-z_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) = 1 - \alpha$$

$$P\left(\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

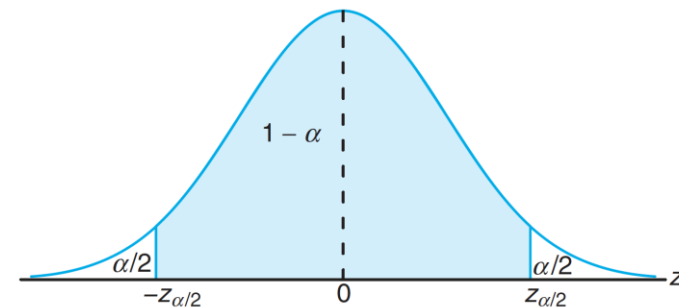


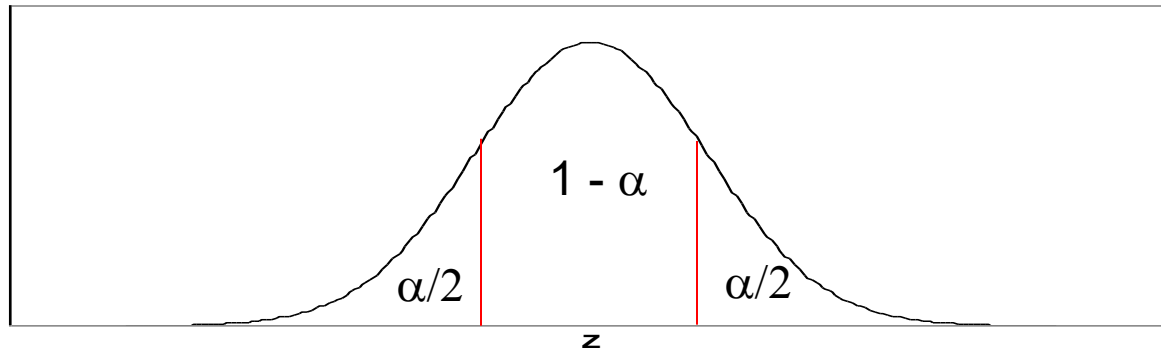
Figure 9.2: $P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$.

Confidence Interval of μ (if σ is known)

- Given:
 - σ is known and \bar{X} is the mean of a random sample of size n ,
- Then,
 - the $(1 - \alpha)100\%$ *confidence interval* for μ is

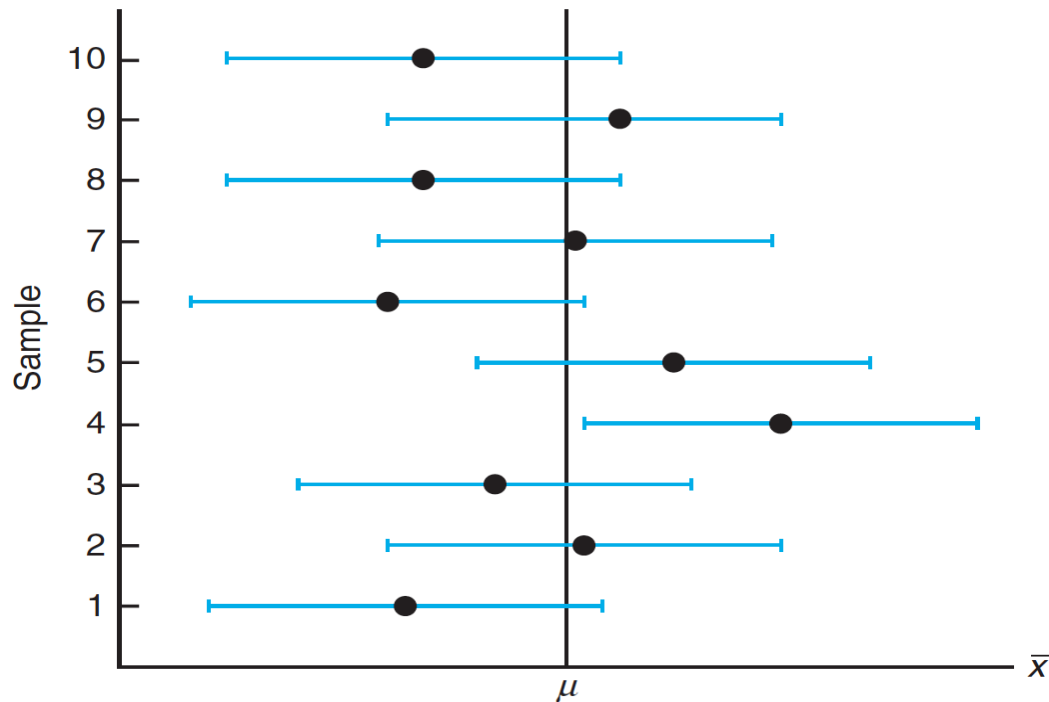
$$\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

where $z_{\alpha/2}$ is the z -value leaving an area of $\alpha/2$ to the right.



Note

Different samples will yield different values of \bar{x} and therefore produce different interval estimates of the parameter μ as shown in the figure below. Most of the intervals are seen to contain μ , but not in every case. Note that all of these intervals are of the same width, since their widths depend only on the choice of $\alpha/2$ once \bar{x} is determined.



Guidelines to Construct a Confidence Interval for μ

- 1. Identify the sample statistics, n and \bar{x} , and known parameter σ^2
- 2. Identify α .
- 3. Find the critical value $z_{\alpha/2}$.
- 4. Find the left and right endpoints and form the confidence interval.

• Example 9.2

- The average zinc (아연) concentration recovered from a sample of measurements taken in 36 different locations in a river is found to be 2.6 grams per milliliter. Find the 95% and 99% confidence intervals for the mean zinc concentration in the river.
- Assume that the population standard deviation is 0.3 gram per milliliter.

The point estimate of μ is $\bar{x} = 2.6$. The z -value leaving an area of 0.025 to the right, and therefore an area of 0.975 to the left, is $z_{0.025} = 1.96$ (Table A.3). Hence, the 95% confidence interval is

$$2.6 - (1.96) \left(\frac{0.3}{\sqrt{36}} \right) < \mu < 2.6 + (1.96) \left(\frac{0.3}{\sqrt{36}} \right),$$

which reduces to $2.50 < \mu < 2.70$.

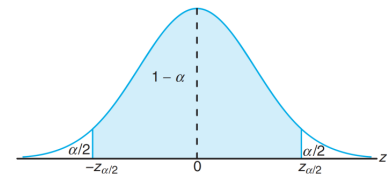


Figure 9.2: $P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$.

To find a 99% confidence interval, we find the z -value leaving an area of 0.005 to the right and 0.995 to the left. From Table A.3 again, $z_{0.005} = 2.575$, and the 99% confidence interval is

$$2.6 - (2.575) \left(\frac{0.3}{\sqrt{36}} \right) < \mu < 2.6 + (2.575) \left(\frac{0.3}{\sqrt{36}} \right),$$

or simply

$$2.47 < \mu < 2.73.$$

We now see that a longer interval is required to estimate μ with a higher degree of confidence. └

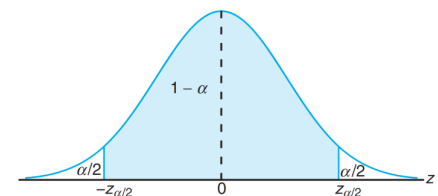


Figure 9.2: $P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$.

• Example

- A professor wants to estimate the hours per week students use computers at home. If a random sample of 100 students, the mean length of time a computer was used at home was 40 hours.
- From past studies, the professor assumes $\sigma = 5$ hours. Assume the length of time a computer was used are approximately normally distributed. Use this information to construct the 90% and 95% confidence intervals for the population mean length of time.

$$\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

where $z_{\alpha/2}$ is the z -value leaving an area of $\alpha/2$ to the right.

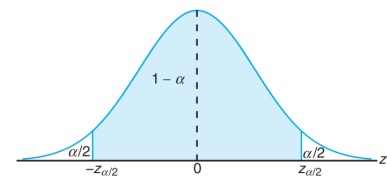
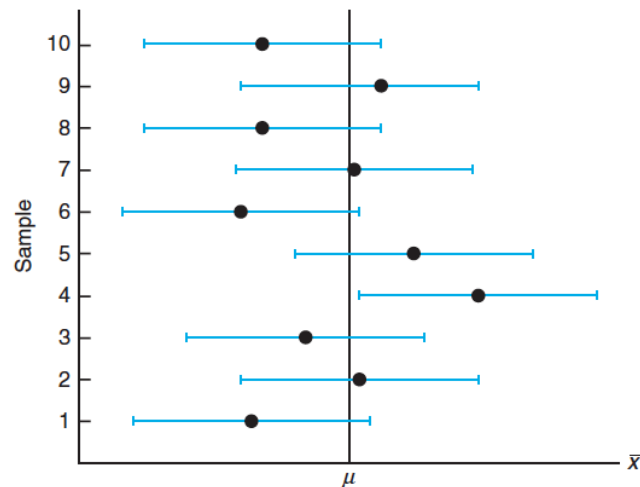


Figure 9.2: $P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$.



μ is actually the center value of the interval, but \bar{x} may estimate μ with error.

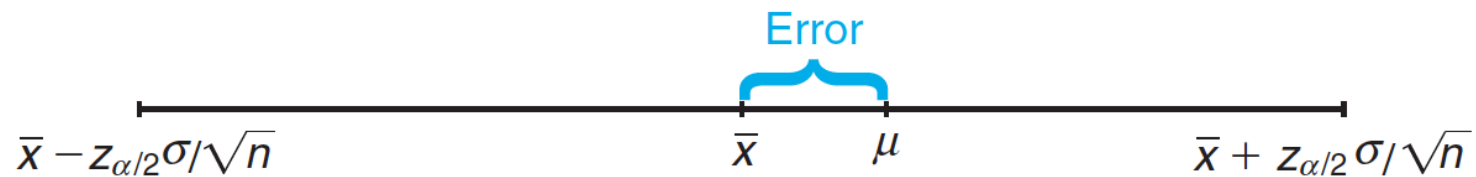


Figure 9.4: Error in estimating μ by \bar{x} .

Most of time \bar{x} is not be exactly equal to μ and the point estimate is in error. The size of this error will be the absolute value of difference between μ and \bar{x} , and we can be $100(1-\alpha)\%$ confident that this difference will not exceed $z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$.

- Theorem 9.1

If \bar{x} is used as an estimate of μ , we can be $100(1 - \alpha)\%$ confident that the error will not exceed $z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$.

- In Example 9.2, we are 95% confident that the sample mean $\bar{x} = 2.6$ differs from the true mean μ by an amount less than $(1.96)(0.3)/\sqrt{36} = 0.1$ and 99% confident that the difference is less than $(2.575)(0.3)/\sqrt{36} = 0.13$.

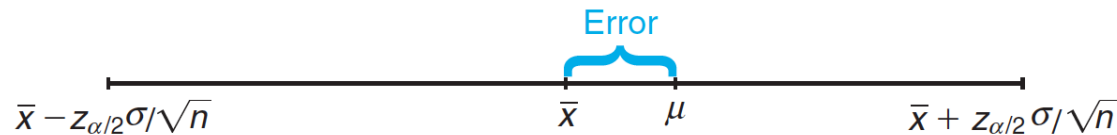


Figure 9.4: Error in estimating μ by \bar{x} .

- Frequently, we wish know how large a sample is necessary to ensure that the error in estimating μ will be less than a specified amount e .

- Theorem 9.2

If \bar{x} is used as an estimate of μ , we can be $100(1 - \alpha)\%$ confident that the error will not exceed a specified amount e when the sample size is

$$n = \left(\frac{z_{\alpha/2}\sigma}{e} \right)^2 .$$


- Example

- How large a sample is required if we want to be 95% confident that our estimate of μ in Example 9.2 is off by less than 0.05?

- Solution:

The population standard deviation is $\sigma = 0.3$. Then, by Theorem 9.2,

$$n = \left[\frac{(1.96)(0.3)}{0.05} \right]^2 = 138.3.$$

Therefore, we can be 95% confident that a random sample of size 139 will provide an estimate \bar{x} differing from μ by an amount less than 0.05. 

One-Sided Confidence Bounds (σ^2 is known)

- One-sided confidence interval are developed in the same fashion as two-sided intervals.

By the **central limit theorem**

Lower one-sided bound

$$P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_\alpha\right) = 1 - \alpha \Rightarrow P\left(\mu > \bar{X} - z_\alpha \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

Upper one-sided bound

$$P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} > -z_\alpha\right) = 1 - \alpha \Rightarrow P\left(\mu < \bar{X} + z_\alpha \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

One-Sided Confidence Bounds (σ^2 is known)

- If \bar{X} is the mean of a random sample of size n from a population with variance σ^2 , the one-sided $100(1 - \alpha)\%$ confidence bounds for μ are given by

upper one-sided bound:	$\bar{x} + z_{\alpha}\sigma/\sqrt{n};$
lower one-sided bound:	$\bar{x} - z_{\alpha}\sigma/\sqrt{n}.$

- Example 9.4

- In a psychological testing experiment, 25 subjects are selected randomly and their reaction time, in seconds, to a particular stimulus is measured. Past experience suggests that the variance in reaction times to these types of stimuli is 4 sec² and that the distribution of reaction times is approximately normal. The average time for the subjects is 6.2 seconds. Give an upper 95% bound for the mean reaction time.

The upper 95% bound is given by

$$\begin{aligned}\bar{x} + z_{\alpha}\sigma/\sqrt{n} &= 6.2 + (1.645)\sqrt{4/25} = 6.2 + 0.658 \\ &= 6.858 \text{ seconds.}\end{aligned}$$

Hence, we are 95% confident that the mean reaction time is less than 6.858 seconds.





<https://www.avepoint.com/blog/microsoft-teams/microsoft-teams-qa/>