MATH1019 Linear Algebra and Statistics for Engineers

Lecture 3: Sampling Distribution & Estimation

Overview: In this lecture, we consider the sampling distribution of the mean and use it to estimate the population mean with an interval at a particular “confidence” level.

Motivation: Whenever we estimate a population parameter such as the population mean, the estimate is based on a single random sample taken from the population of interest. This estimate will be different for different samples and it is, therefore, important to specify an uncertainty associated with the estimate. We do this using a range of values which is expected, with some quantifiable degree of confidence, to contain the value of an unknown value of interest. Some applications are quality control and modelling and simulation as a quantitative method of validation.

Learning outcomes

In today’s lecture we will learn how to:

- Check data for normality
- Obtain the sampling distribution of the mean
- Use a confidence interval to estimate a population parameter

Checks for Normality

Often we wish to know whether we can assume that our data is approximately normally distributed. This can be checked in several ways.

1. Shape of stem-and-leaf plot (or histogram).
2. Shape of box plot
3. Using the 68 - 95 - 99.7 rule
4. Normal quantile plot.

Example: Rainfall Data

- Let us test the following rainfall data for normality.

1.88	2.23	2.58	2.07	2.94	2.29	3.14
2.15	1.95	2.51	2.86	1.48	1.12	2.76
3.10	2.05	2.23	1.70	1.57	2.81	1.24
3.29	1.87	1.50	2.99	3.48	2.12	4.69
2.29	2.12					

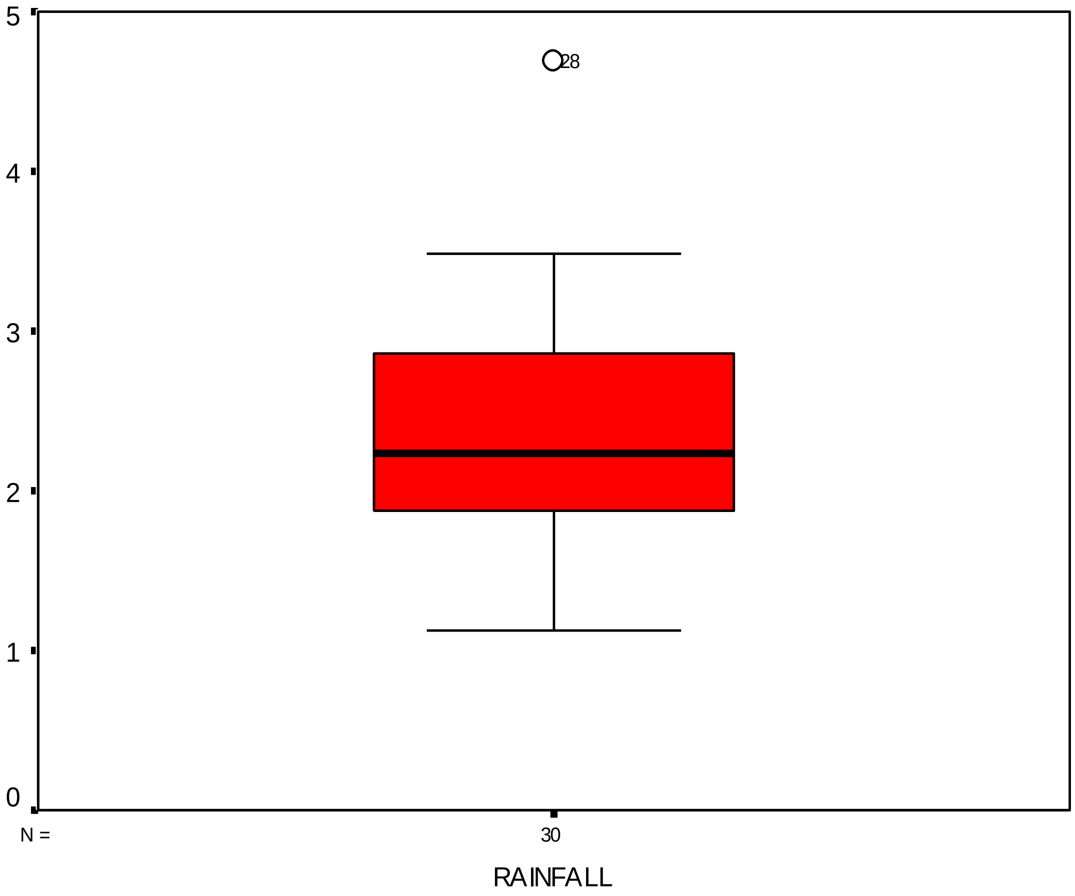
```
RainFall Stem-and-Leaf Plot

 Frequency    Stem &  Leaf

     3.00        1 .  124
     6.00        1 .  557889
     9.00        2 .  001112222
     7.00        2 .  5578899
     4.00        3 .  1124
     1.00 Extremes    (>=4.7)

 Stem width:      1.00
 Each leaf:       1 case(s)
```

Interpretation of stem plot

- Note: The data has been truncated.

- Apart from the observation of 4.69 the data appears to have a typical bell shape.

Boxplot of rainfall data



2. Boxplot

- A normal distribution would have the median in the middle of the box and the whiskers will be of approximately same length (why?).
- The whiskers should be slightly longer than half the length of the box

3. 68-95-99.7 rule

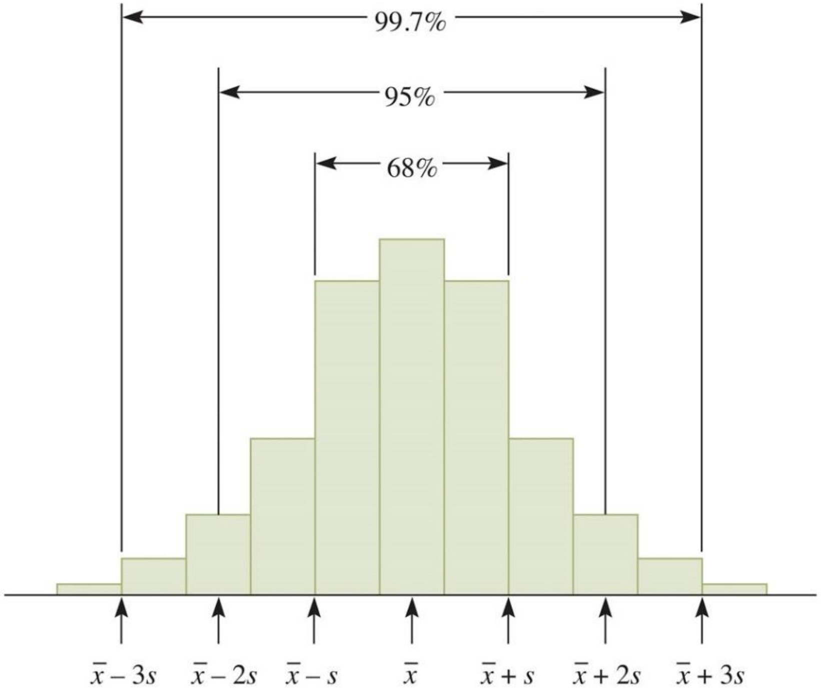
This is an empirical rule.

For a distribution that is symmetrical and bell-shaped (in, particular for a normal distribution):

Approximately 68% of the data values will lie within 1 standard deviation on each side of the mean.

Approximately 95% of the data values will lie within 2 standard deviations on each side of the mean.

Approximately 99.7% of the data values will lie within 3 standard deviation3 on each side of the mean.



An Approximate Test for Normality

- From the numerical summary
- Mean = 2.367
- Standard deviation = 0.754

- Calculate 2.367 ± 0.754, then count how many observations there are in this interval. Convert the number to a percentage. This should be close to 68% if the data is approximately normally distributed.
- Now calculate 2.367 ± 2(0.754) and 2.367 ± 3(0.754) and check the percentage of observations in these intervals. Compare with 95% and 99.7%.

4. Normal plot

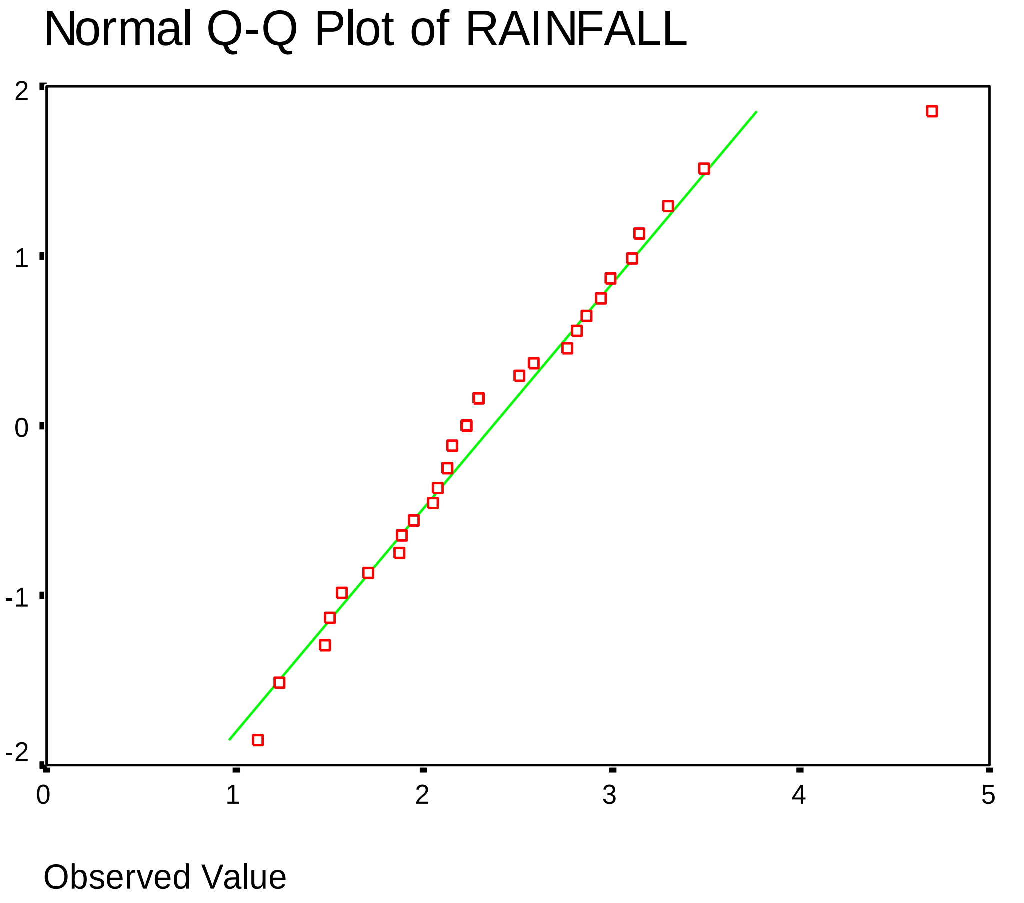
- If you have access to a software package then a normal quantile plot is the best check. If the data is normally distributed then this plot will produce a straight line. Each observation, *x* is plotted against the corresponding quantile of the standard normal distribution. If the distribution of X is normal then the plot will produce a straight line as there is a linear relationship between the two.

Test of Normality

	Kolmogorov-Smirnov			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
RAINFALL	.141	30	.134	.952	30	.287

- In general use the Shapiro-Wilks significance level

Normal quantile plot



Why Normality?

- Certain statistical methods require our data to be normally distributed.

- Luckily most methods are robust to departure from normality

- You need approximate normality only

Population and Sample

- The entire group of about which information is sought is called the **population**.

- A **sample** is a subset of the population that is examined for the purpose of gathering information.

Simple Random Sample

- A **simple random sample** of size *n* consists of *n* units from the population chosen in such a way that every possible group of *n* units has the same probability of being chosen as the sample.

Parameter

- A **parameter** is a number describing the population.
- Population mean and population standard deviation are examples of parameters.

Statistic and Estimator

- A **statistic** is a number that can be computed from the data.

- When a parameter is unknown we use a statistic to estimate it.

- A statistic used to estimate a parameter is called an **estimator** of that parameter.

Sampling Distribution

- The distribution of a statistic is called **sampling distribution**.

- For example, the sampling distribution of the mean of a sample taken from a normal population will again be normal.

- We will discuss its mean and standard deviation later.

Sampling Distribution of the Sample Mean

- Suppose a population consists of a very large and equal number of three possible outcomes.

X	1	2	3
P(X=x)	1/3	1/3	1/3

- This is a symmetric distribution with

mean = 2 variance = 2/3

- Suppose we take ALL possible samples of size *n* and write down the probability distribution of the sample mean. (note that this is an immensely complex task even for *n* not too large.) This gives us the sampling distribution of the sample mean. The sampling distribution will change with the sample size. What would its expectation and variance be?

Samples of size 2

- Let us take all possible samples of size 2 from this population and work out the mean of each sample. The distribution of these means is the sampling distribution of the mean when the sample size is 2.

$\bar{X}$	1	3/2	2	5/2	3
P($\bar{X}$)	1/9	2/9	3/9	2/9	1/9

Mean and Variance

- If the mean and variance of the above distribution are calculated we will get
  $\mathrm{E}(\bar{X}) = 2$ and $\mathrm{Var}(\bar{X}) = 1/3$

- The expectation here is the same as that of the original distribution. Variance is half of the original variance.

- Similarly, if the distribution of the sample mean for the case of $n = 3$ is constructed, we will find that the expectation is still the same and the variance will be one-third of the original variance.

Samples of size 3

- The sampling distribution of the sample mean when $n = 3$ is given below:

$\bar{X}$	1	4/3	5/3	2	7/3	8/3	3
P($\bar{X}$)	1/27	3/27	6/27	7/27	6/27	3/27	1/27

- Here the mean is 2 and the variance is 2/9.

Mean and standard deviation of a sample mean

- If $\bar{X}$ is the mean of a SRS of size *n* from a population having mean μ and standard deviation σ, then

$$\mu_{\overline{X}} = \mu, \qquad \sigma^2_{\overline{X}} = \frac{\sigma^2}{n},$$

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$$

Distribution of $\bar{X}$

- If a population has N(μ, σ) distribution, then the sample mean of *n* independent observations has distribution

$$N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

- This is because any linear combination of independent normal random variables is normally distributed.

- Therefore

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

- When the original population is not normal, as in the case of the 3-valued random variable considered above, the above results do not tell us what the distribution of the sample mean will look like.

- All we know is its mean and standard deviation.

Central Limit Theorem

- The **Central Limit Theorem** says that the distribution of the sample mean approaches a normal distribution as the sample size increases regardless of the distribution of the underlying population, provided it has a finite mean and variance.

- This is a very important result with far reaching consequences.

Example

A bottling company uses a filling machine to fill plastic bottles with a popular cola. The bottles are supposed to contain 300ml. In fact the contents vary according to a normal distribution with a mean of 298ml and standard deviation 3ml.

(a) What is the probability that an individual bottle contains less than 295ml?

(b) What is the probability that the mean contents of the bottles in a six-pack is less than 295ml?

Solution

a) $P(X < 295) = P\left[Z < \frac{295-298}{3}\right]$

$= P(Z < -1) = .1587$

b) Here we are dealing with $\bar{X}$ where $n = 6$.

$$P(\bar{X} < 295) = P\left[Z < \frac{295-298}{3/\sqrt{6}}\right]$$

$$= P(Z < -2.45)$$

$$= .0071$$

Example continued

- Here, even though there is about 16% chance for an individual bottle to have less than 295 ml, the chances of the average in a six-pack being less than 295 is only about 0.7%.

- The average for a twelve-pack will have even lower probability of going below 295.

Change in variability as n increases

- As the sample size increases the variability of the sample mean decreases and the values get closer and closer to the population mean.
- As a result the probability that the average is so far below the mean will decrease further and further.

Example

A laboratory weighs filters from a coal mine to measure the amount of dust in the mine atmosphere. Repeated measurements of the weight of the dust on the same filter vary normally with standard deviation of 0.08mg because the weighing is not perfectly precise. The dust on a particular filter actually weighs 123mg. Repeated weightings will then have the normal distribution with mean 123mg and standard deviation 0.08mg.

(a) The laboratory reports the mean of 3 weightings. What is the distribution of this mean?

(b) What is the probability that the lab reports a weight of 124mg or higher for this filter?

Solution

a) $$N\left(123, \frac{.08}{\sqrt{3}}\right)$$

b) $$P(\bar{x} > 124) = P\left[Z > \frac{124-123}{.08/\sqrt{3}}\right]$$

$$= P(Z > 21.7) = 0$$

How do you select a distribution for given problem?

- Direct theoretic arguments
- Draw histogram to explore shape
- When data is given you can statistically test for the goodness of the distribution.-goodness of fit test—more on this later
- Why do you need to study the distributions?

Inference

- The purpose of statistical inference is to draw conclusions from data.
- The methods are based on sampling distributions
  - We have a normal population
  - Parameters: mean μ, standard deviation σ
  - Unknown μ is parameter of interest
  - Assume that σ is known

ESTIMATION: AN INTRODUCTION

Definition

- The assignment of value(s) to a population parameter based on a value of the corresponding sample statistic is called ***estimation***.

- The value(s) assigned to a population parameter based on the value of a sample statistic is called an ***estimate***.

- The sample statistic used to estimate a population parameter is called an ***estimator***.

ESTIMATION: AN INTRODUCTION cont.

The estimation procedure involves the following steps:

1. Select a sample.
2. Collect the required information from the members of the sample.
3. Calculate the value of the sample statistic.
4. Assign value(s) to the corresponding population parameter.

POINT AND INTERVAL ESTIMATES

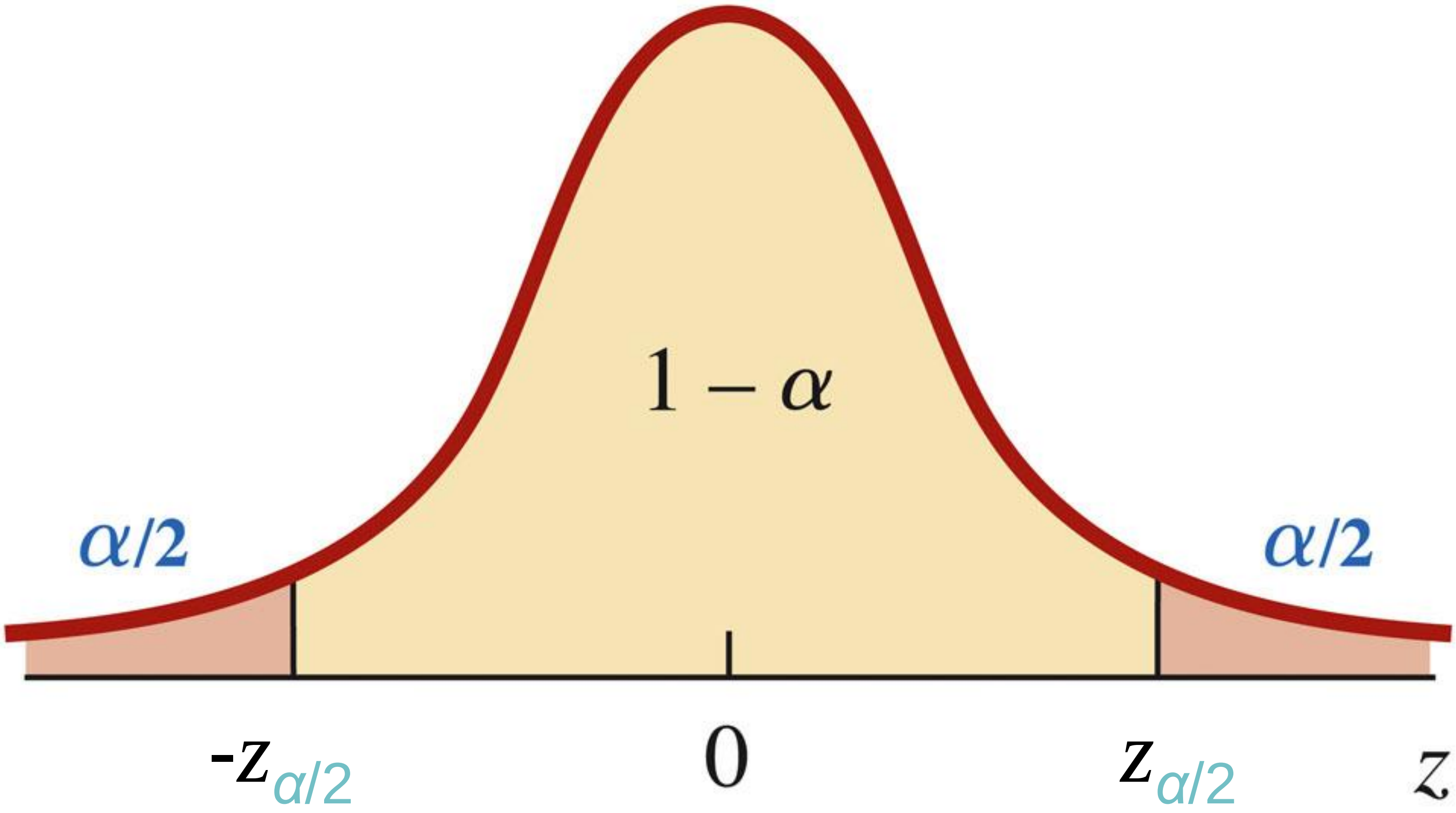
- The value of a sample statistic that is used to estimate a population parameter is called a ***point estimate***.

- In ***interval estimation***, an interval is constructed around the point estimate, and it is stated that this interval is likely to contain the corresponding population parameter.

Confidence Interval

- We know that we estimate μ by the sample mean.
- But for many situations we will need an interval where μ will belong with certain pre-specified probability 1-α.
- The interval is called a 100(1-α)% *confidence interval.*
- For example, when α=0.05, we have a 95% confidence interval.

Confidence Coefficient $z_{\alpha/2}$



Distribution of $\bar{X}$

- Recall that the sampling distribution of the mean is $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$

- So by 68-95-99.7 rule:

- 95% of the time $\bar{X}$ will be within $2\frac{\sigma}{\sqrt{n}}$ of μ, or more precisely $1.96\frac{\sigma}{\sqrt{n}}$

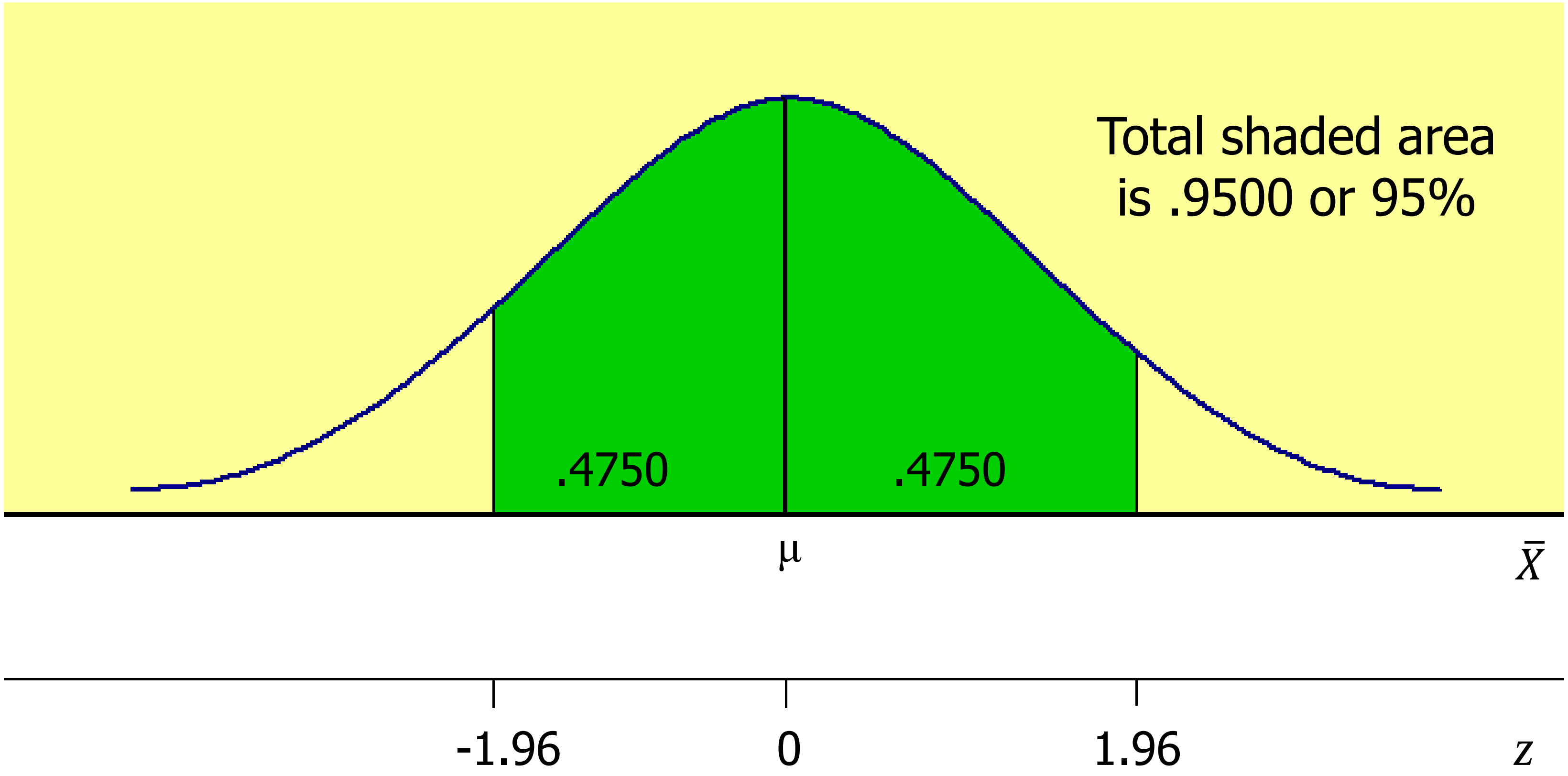
95% confidence interval for μ

- From the sample calculate $\bar{X}$ and construct an interval of $\pm 1.96\frac{\sigma}{\sqrt{n}}$ on either side of $\bar{X}$.

- This is the 95% confidence interval for μ.

- This interval has the property that 95% of the time it will contain the population mean.

Finding *z* for a 95% confidence level.



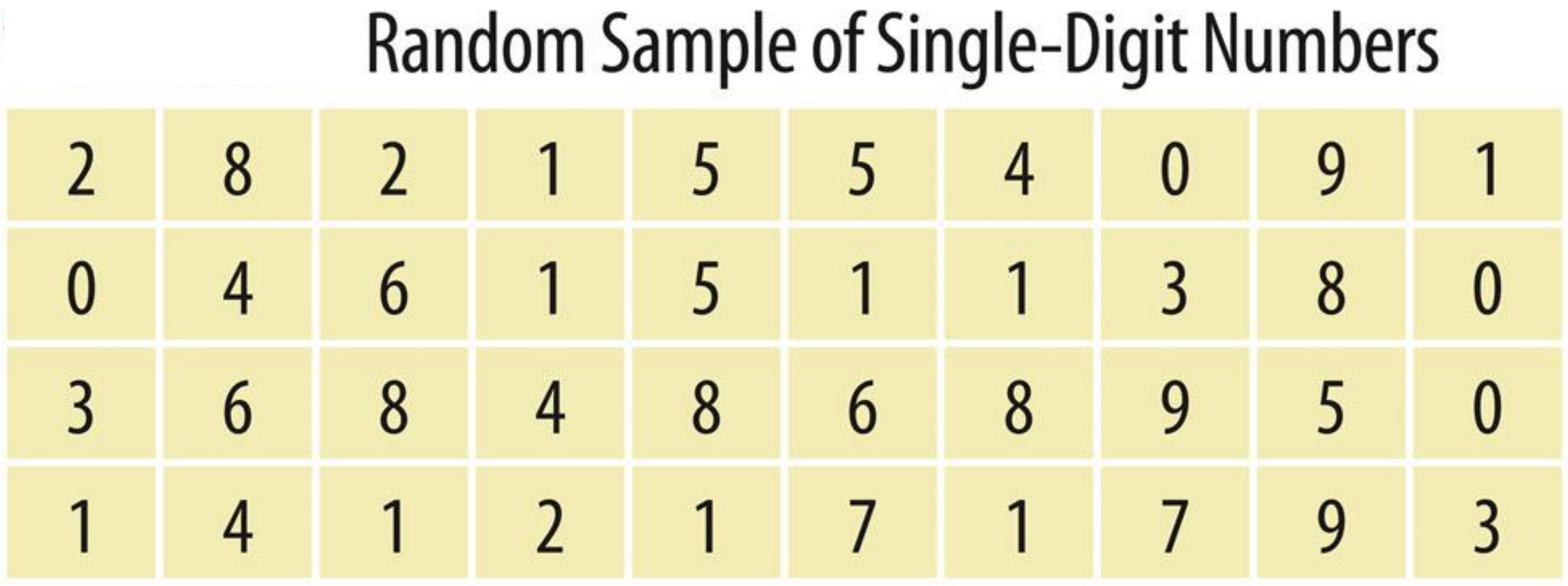
General confidence interval for μ

- The 100(1-α)% confidence interval for μ when σ is known is given by $\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$

- The value of $z_{\alpha/2}$ depends on the confidence level at which you wish to give your results
- The interval is in the form:

estimate ± margin of error.

Example



Random Sample of Single-Digit Numbers

2	8	2	1	5	5	4	0	9	1
0	4	6	1	5	1	1	3	8	0
3	6	8	4	8	6	8	9	5	0
1	4	1	2	1	7	1	7	9	3

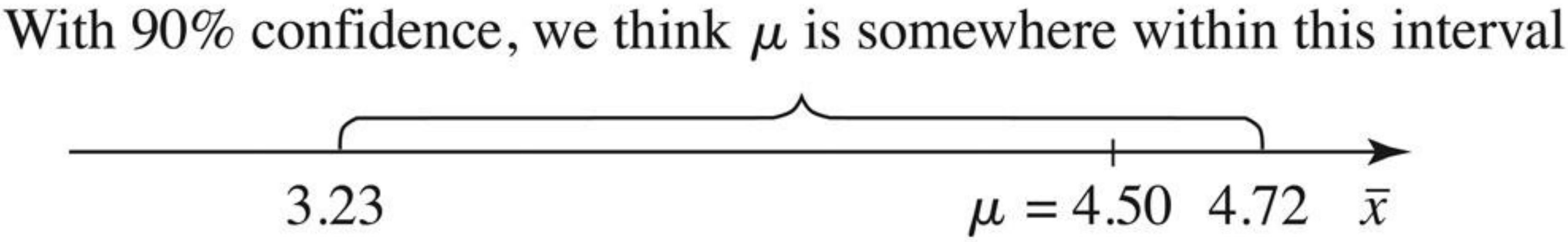
The sample statistics are $n = 40$, $\sum x = 159$, and $\bar{x} = 3.975$. Here is the resulting 90% confidence interval:

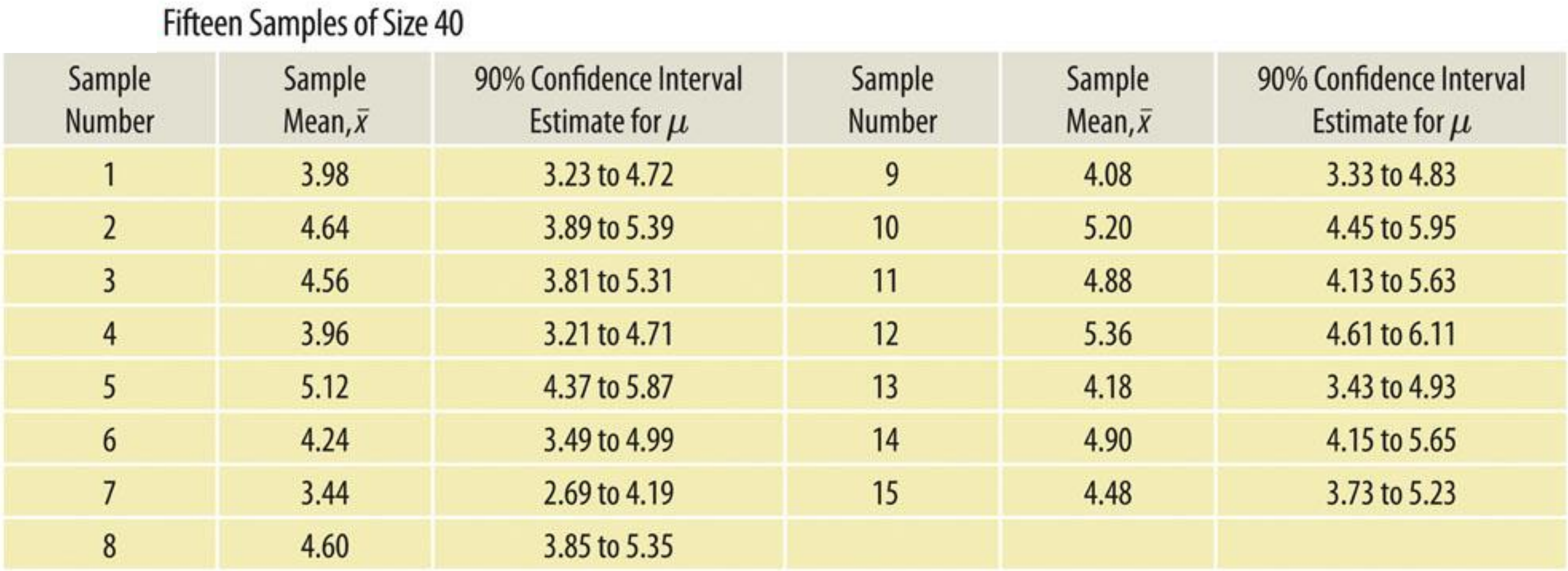
$\bar{x} \pm z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$: $\quad 3.975 \pm 1.65\left(\frac{2.87}{\sqrt{40}}\right)$

$$3.975 \pm 0.749$$

i.e. the 90% confidence interval for μ is (3.23, 4.72)

The 90% Confidence Interval

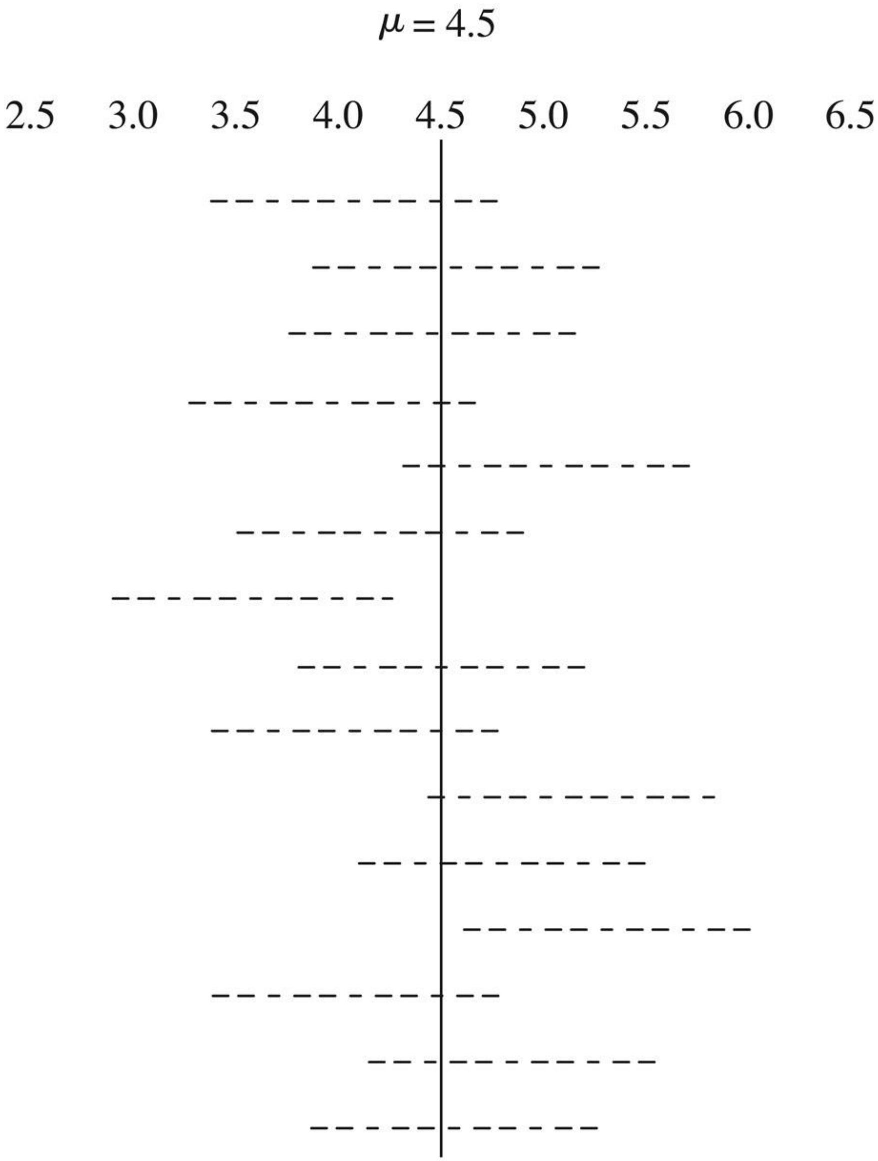

With 90% confidence, we think μ is somewhere within this interval
3.23
$\mu = 4.50$
4.72
$\bar{x}$



Fifteen Samples of Size 40

Sample Number	Sample Mean, $\bar{x}$	90% Confidence Interval Estimate for μ	Sample Number	Sample Mean, $\bar{x}$	90% Confidence Interval Estimate for μ
1	3.98	3.23 to 4.72	9	4.08	3.33 to 4.83
2	4.64	3.89 to 5.39	10	5.20	4.45 to 5.95
3	4.56	3.81 to 5.31	11	4.88	4.13 to 5.63
4	3.96	3.21 to 4.71	12	5.36	4.61 to 6.11
5	5.12	4.37 to 5.87	13	4.18	3.43 to 4.93
6	4.24	3.49 to 4.99	14	4.90	4.15 to 5.65
7	3.44	2.69 to 4.19	15	4.48	3.73 to 5.23
8	4.60	3.85 to 5.35			

Confidence Intervals from Table



Lecture Summary

- ***CLT:*** the distribution of the sample means will be approximately normally distributed. The mean of the sampling distribution will be equal to the mean of the population; and standard deviation will equal to the standard deviation of the population divided by the square root of the number of items in each sample.

- The 100(1-α)% confidence interval for μ when σ is known is given by

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$