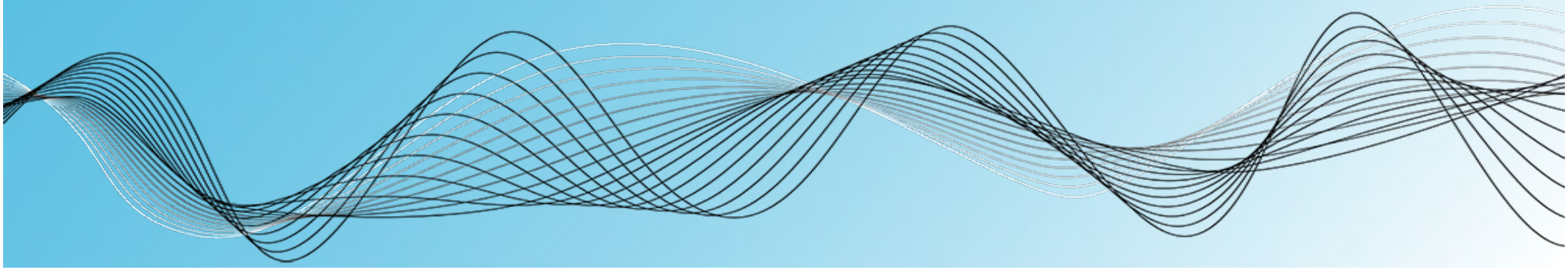
Descriptive Statistics



Population Vs Sample

Statistics

The science of collecting, organizing, presenting, analyzing, and interpreting data to assist in making more effective decisions.

Population

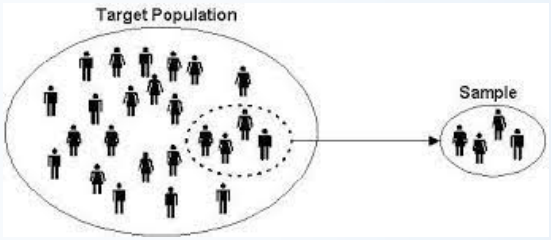
Generally, population refers to the people who live in a particular area at a specific time.

But in statistics, population refers to data on your study of interest. It can be a group of individuals, objects, events, organizations, etc.

Sample

A sample is defined as a smaller and more manageable representation of a larger group.

A subset of a larger population that contains characteristics of that population. A sample is used in statistical testing when the population size is too large for all members or observations to be included in the test.



Sampling

Simple Random Sampling

Each individual is chosen entirely by chance and each member of the population has an equal chance, or probability, of being selected

Sampling

Stratified Sampling:

The population is first divided into subgroups (or strata) who all share a similar characteristic. It is used when we might reasonably expect the measurement of interest to vary between the different subgroups, and we want to ensure representation from all the subgroups

Example—A student council surveys 100 students by getting random samples of 25 freshmen, 25 sophomores, 25 juniors, and 25 seniors.

Sampling

Cluster random sample: The population is first split into groups. The overall sample consists of every member from some of the groups. The groups are selected at random.

Example—An airline company wants to survey its customers one day, so they randomly select 5 flights that day and survey every passenger on those flights.

Sampling

Systematic random sample: Members of the population are put in some order. A starting point is selected at random, and every nth member is selected to be in the sample.

Example—A principal takes an alphabetized list of student names and picks a random starting point. Every 20th student is selected to take a survey.

Parameter vs Statistic

Parameter

Parameters are numbers that describe the properties of entire populations.

Statistic

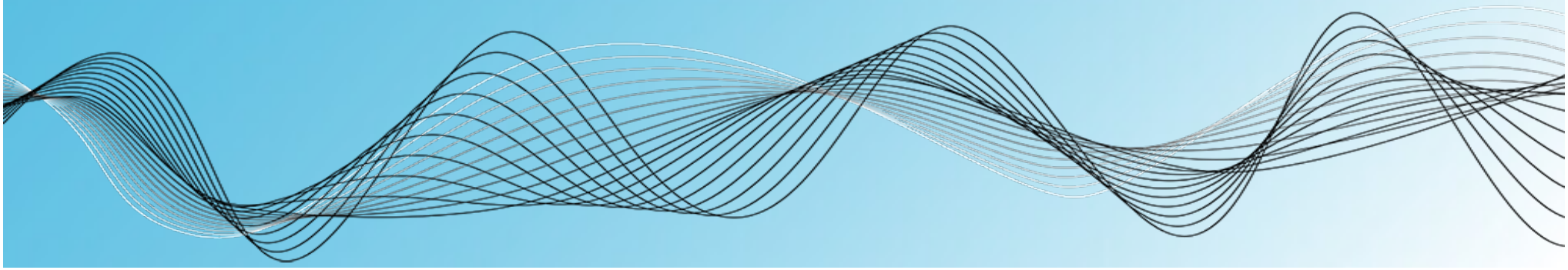
Statistics are numbers that describe the properties of samples.

Example :

The average income for the India is a population parameter.

The average income for a sample drawn from the India is a sample statistic.

Descriptive Statistics : Central Tendency



Central Tendency

Measures of Central Tendency

Central tendency is defined as “the statistical measure that identifies a single value as representative of an entire distribution. It represents the single value of the entire population or a dataset.

We will consider five Measures of Central Tendency

- the arithmetic mean
- the median
- the mode
- the weighted mean

Arithmetic Mean

Population Mean

the population mean is the sum of all the values in the population divided by the number of values in the population.
To find the population mean, we use the following formula.

$$\text{Population mean} = \frac{\text{Sum of all the values in the population}}{\text{Number of values in the population}}$$

It is denoted by μ

$$\mu = \frac{\Sigma x}{N}$$

Example

Listed below are the distances between exits (in miles).

11	4	10	4	9	3	8	10	3	14	1	10	3	5
2	2	5	6	1	2	2	3	7	1	3	7	8	10
1	4	7	5	2	2	5	1	1	3	3	1	2	1

$$\mu = \frac{\Sigma x}{N} = \frac{11 + 4 + 10 + \cdots + 1}{42} = \frac{192}{42} = 4.57$$

Median

Median

When our data contains one or two very large or very small values, the arithmetic mean may not be representative.
The center for such data is better described by a measure of location called the median.
The midpoint of the values after they have been ordered from the minimum to the maximum values.

Formula for median

n is odd,

$$\text{Median} = \left(\frac{n+1}{2}\right)^{th} \text{observation}$$

n is even,

$$\text{Median} = \frac{\left(\frac{n}{2}\right)^{th} + \left(\frac{n}{2}+1\right)^{th} \text{observation}}{2}$$

Where n is total number of data points in our sample.

Median

Example 1 (Odd numbers):

21,32,65,40,30,90,26

Ordered list : 21,26,30,32,40,65,90

n = 7

Median

$= (\frac{7+1}{2})^{th}$ Observation

$= 4^{th}$ observation = 32

Example 1 (Even numbers):

10,9,7,12,8,11

Ordered list : 7,8,9,10,11,12

n = 6

Median

$$= \frac{(\frac{6}{2})^{th}\ observation + (\frac{6}{2}+1)^{th} observation}{2}$$

$$= \frac{3^{rd}\ observation + 4^{th}\ observation}{2}$$

$$= \frac{9+10}{2}$$

= 9.5

Mode

In a dataset mode is the value of the observation that **appears most frequently.**
The value of the observation that appears most frequently**.**

Example :

11	4	10	4	9	3	8	10	3	14	1	10	3	5
2	2	5	6	1	2	2	3	7	1	3	7	8	10
1	4	7	5	2	2	5	1	1	3	3	1	2	1

The frequency table for above data is shown below

Distance in Miles between Exits	Frequency
1	8
2	7
3	7
4	3
5	4
6	1
7	3
8	2
9	1
10	4
11	1
14	1
Total	42

As we can see that value 1 is occurring most number of time (8 times) the mode is 1.

Note* : A dataset can consist of more than 1 mode which is called multimodal dataset.

Weighted mean

The weighted mean is a convenient way to compute the arithmetic mean when there are several observations of the same value.

the weighted mean of a set of numbers designated x1, x2, x3, . . . , xn with the corresponding weights w1, w2, w3, . . . , wn is computed by:

WEIGHTED MEAN

$$\bar{x}_w = \frac{w_1x_1 + w_2x_2 + w_3x_3 + \cdots + w_nx_n}{w_1 + w_2 + w_3 + \cdots + w_n}$$

EXAMPLE

The Carter Construction Company pays its hourly employees $16.50, $19.00, or $25.00 per hour. There are 26 hourly employees, 14 of whom are paid at the $16.50 rate, 10 at the $19.00 rate, and 2 at the $25.00 rate. What is the mean hourly rate paid to the 26 employees?

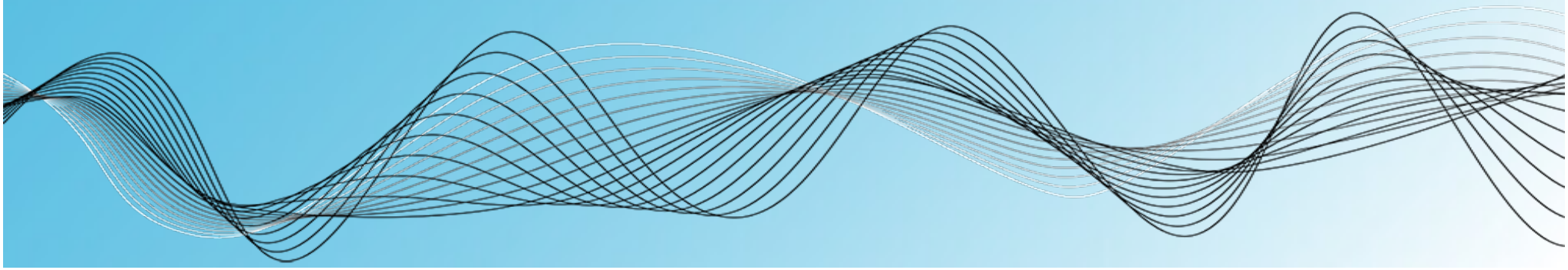
SOLUTION

To find the mean hourly rate, we multiply each of the hourly rates by the number of employees earning that rate. From formula (3–3), the mean hourly rate is:

$$\bar{x}_w = \frac{14(\$16.50) + 10(\$19.00) + 2(\$25.00)}{14 + 10 + 2} = \frac{\$471.00}{26} = \$18.1154$$

The weighted mean hourly wage is rounded to $18.12.

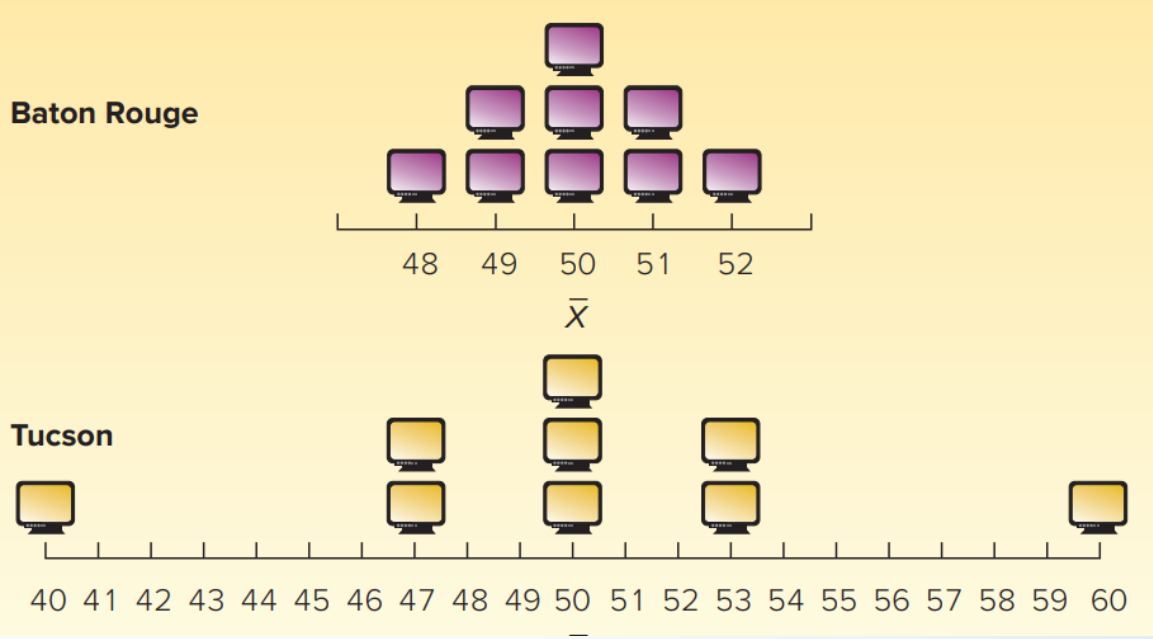
Measures of Dispersion



Measure of Dispersion

Why should we study Measure of Dispersion?

A measure of location, such as the mean, median, or mode, only describes the center of the data.
It is valuable from that standpoint, but it does not tell us anything about the spread of the data.



Range

The simplest measure of dispersion is the range.
It is the difference between the maximum and minimum values in a data set.

Range = Maximum value – Minimum value

Example :

11	4	10	4	9	3	8	10	3	14	1	10	3	5
2	2	5	6	1	2	2	3	7	1	3	7	8	10
1	4	7	5	2	2	5	1	1	3	3	1	2	1

In the above dataset
Minimum = 1
Maximum = 14
Therefore
Range = 14-1 = 13

Variance

A limitation of the range is that it is based on only two values, the maximum and the minimum; it does not take into consideration all of the values. The variance does.

It measures the mean amount by which the values in a population, or sample, vary from their mean.

VARIANCE The arithmetic mean of the squared deviations from the mean.

Formula for variance is

$$\sigma^2 = \frac{\Sigma (x - \mu)^2}{N}$$

Example :

	A	B	C
1		California Airports	
2		Orange County	Ontario
3		20	20
4		40	45
5		50	50
6		60	55
7		80	80
8			
9	Mean	50	50
10	Median	50	50
11	Range	60	60

F	G	H
Calculation of Variance for Orange County		
Number Sold	Each Value - Mean	Squared Deviation
20	20 - 50 = -30	900
40	40 - 50 = -10	100
50	50 - 50 = 0	0
60	60 - 50 = 10	100
80	80 - 50 = 30	900
	Total	2000

Source: Microsoft Excel

$$\text{Variance} = \frac{\Sigma(x - \mu)^2}{N} = \frac{(-30^2) + (-10^2) + 0^2 + 10^2 + 30^2}{5} = \frac{2{,}000}{5} = 400$$

Standard deviation

When we compute the variance, it is important to understand the unit of measure and what happens when the differences in the numerator are squared.

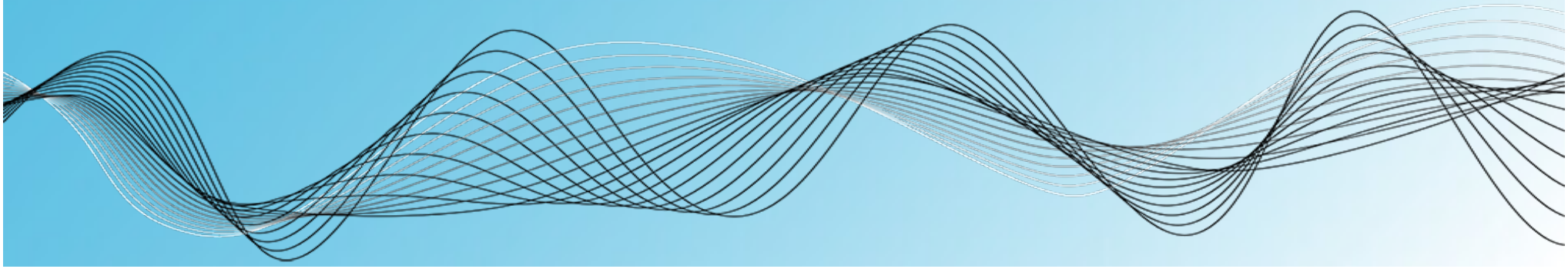
Units of standard deviation is same as units of our dataset.

Standard deviation is squared root of variance

Formula:

STANDARD DEVIATION $\sigma = \sqrt{\frac{\Sigma(x - \mu)^2}{N}}$

Normal Distribution



Normal Distribution

Normal Distribution

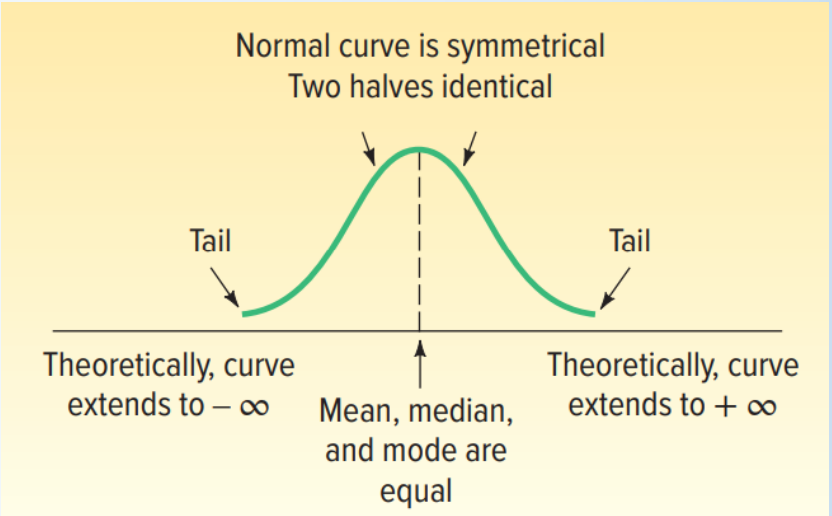
If X is a continuous random variable which follows normal distribution with parameters (mean = μ, Standard deviation = σ) then its denoted as **X ~ Normal(μ, σ)** and its pdf is given by

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

$-\infty \leq X \leq \infty$

Properties of Normal Distribution

- Bell Shaped and has single peak at the center.
- The mean, mode and median are all equal.
- The curve is symmetric at the center (i.e. around the mean, μ).
- Exactly half of the values are to the left of center and exactly half the values are to the right.
- The total area under the curve is 1.



Standard normal variable

Standard Normal Distribution

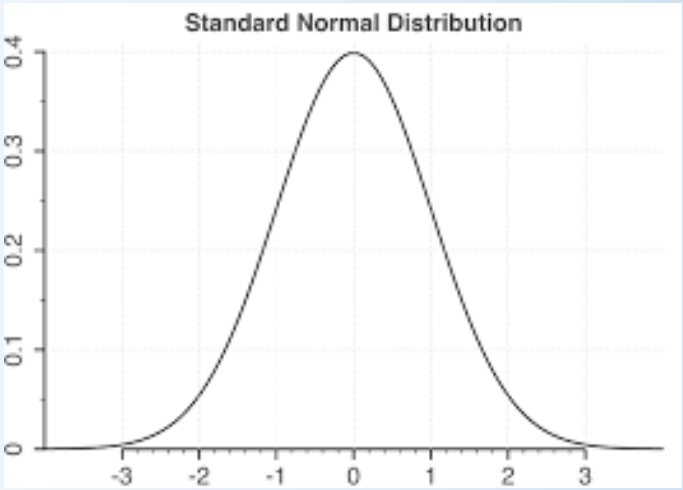
- Any normal probability distribution can be converted into a standard normal probability distribution by subtracting the mean from each observation and dividing this difference by the standard deviation.
- So, a z value is the distance from the mean, measured in units of the standard deviation. The formula for this conversion is:

$$z = \frac{x - \mu}{\sigma}$$

where:
x is the value of any particular observation or measurement.
μ is the mean of the distribution.
σ is the standard deviation of the distribution.

Properties of standard normal variable

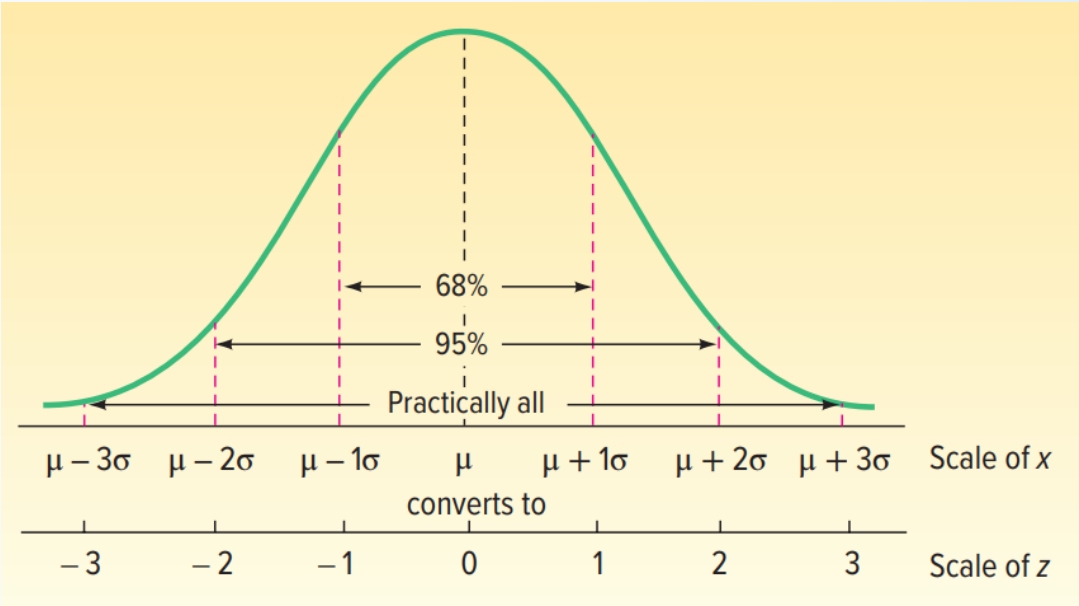
1) The graph of standard normal distribution is bell shaped.
2) The curve is symmetric about mean (μ =0)
3) Standard deviation is equal to 1



The Empirical Rule

Normal Distribution

1) Approximately 68% of the observations will lie within 1 standard deviation of the mean.
2) About 95% of the observations will lie within 2 standard deviations of the mean.
3) Practically all, or 99.7% of the observations will lie within 3 standard deviations of the mean.



Normal Distribution Example

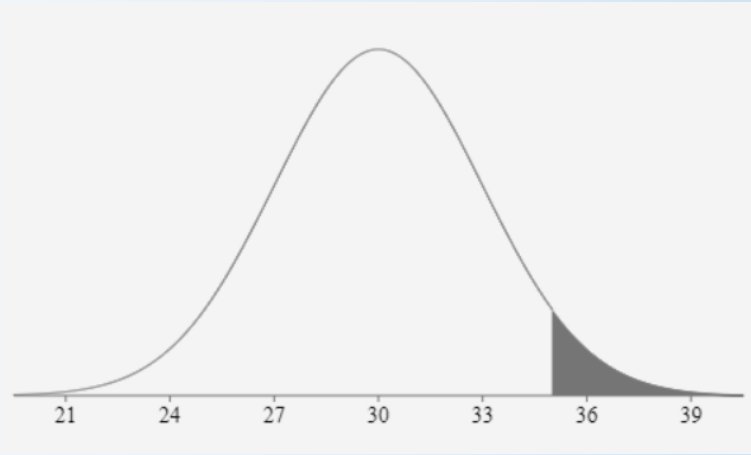
Normal Distribution

The lifetime of a bulbs in a certain region are normally distributed with a mean of 30 days and standard deviation of 3 days. What is the probability that bulbs will working even after 35?

Solution:

Let X be the random variable denoting the lifetime of bulbs.

$X \sim Normal(\mu = 30, \sigma = 3)$



Probability that bulb is working for more than 35 days

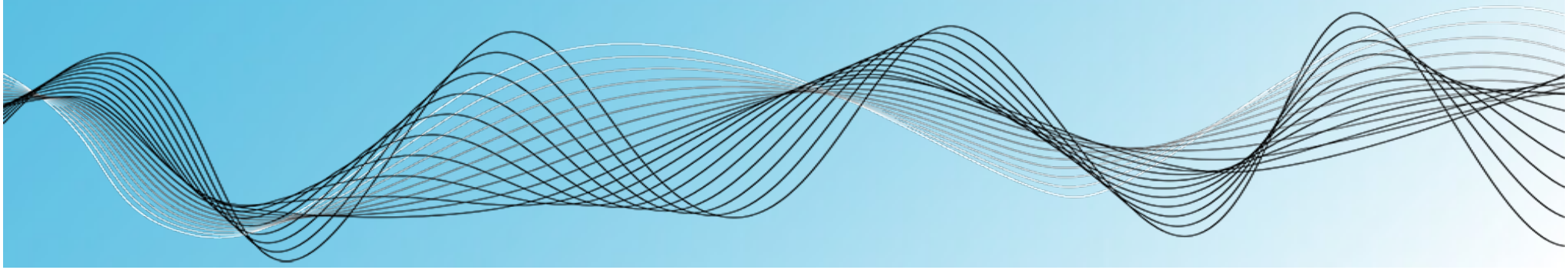
= P(X > 35)

$= P(Z > \frac{35-30}{3})$

= P(Z > 1.67)

= 0.0478

Inferential Statistics



Inferential Statistics

- Descriptive statistics describes data (for example, a chart or graph) and inferential statistics allows you to make predictions ("inferences") from that data.
- With inferential statistics, you take data from samples and make generalizations about a population.
- For example, you might stand in a mall and ask a sample of 100 people if they like shopping on weekends. You could make a bar chart of yes or no answers (that would be descriptive statistics) or you could use your research (and inferential statistics) to reason that around 75-80% of the population (all shoppers in all malls) like shopping on weekends.

There are two main areas of inferential statistics:

- **Estimating parameters :** This means taking a statistic from your sample data (for example the sample mean) and using it to say something about a population parameter (i.e. the population mean).
- **Hypothesis tests :** This is where you can use sample data to answer research questions. For example, you might be interested in knowing if a new cancer drug is effective. Or if breakfast helps children perform better in schools.

Central Limit Theorem

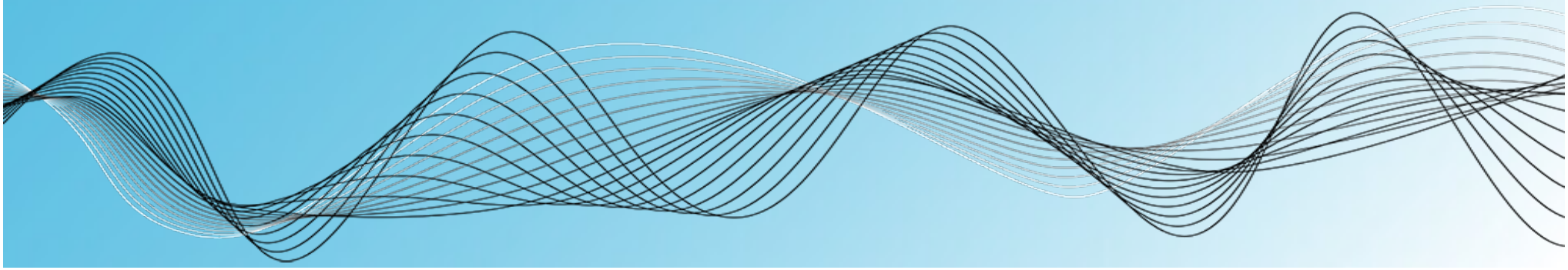
Central limit theorem

Assume that $\boldsymbol{x_1, x_2}$,…, $\boldsymbol{x_n}$ are n observations from a random sample.

1) If $\boldsymbol{x_1, x_2}$,…, $\boldsymbol{x_n}$ are from a normal distribution with mean μ and standard deviation σ then sample mean will follow normal distribution with mean = μ and standard deviation = $\frac{\sigma}{\sqrt{n}}$
2) If $\boldsymbol{x_1, x_2}$,…, $\boldsymbol{x_n}$ are not from a normal distribution but sample size is large then also central limit theorem holds.

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Point Estimate



Point Estimate

Point Estimate

A point estimate is a single statistic which is computed from sample data and is used to estimate a population parameter.

Example

Suppose Best Buy Inc. wants to estimate the mean age of buyer who purchase LCD HDTV televisions.

They select a random sample of 75 recent purchases, determine the age of each buyer, and compute the mean age of the buyers in the sample.

The **mean of this sample is a point estimate** of the population mean.

Here

Sample size : 75

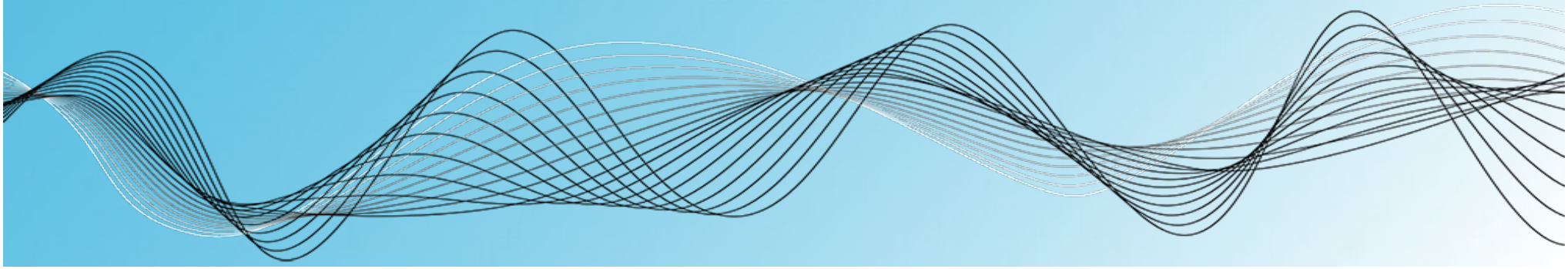
Population : All the buyer who purchase LCD HDTV televisions

Sample : All the buyer of 75 recent purchases

Parameter : mean age of all buyer who purchase LCD HDTV televisions

Statistic : mean age of all the buyer of 75 recent purchases

Standard Error



Standard Error

Standard Error

- The standard error (SE) of a statistic is the approximate standard deviation of a statistical sample population.
- The standard error is a statistical term that measures the accuracy with which a sample distribution represents a population by using standard deviation.
- In statistics, a sample mean deviates from the actual mean of a population; this deviation is the standard error of the mean.

The smaller the standard error, the more representative the sample will be of the overall population.

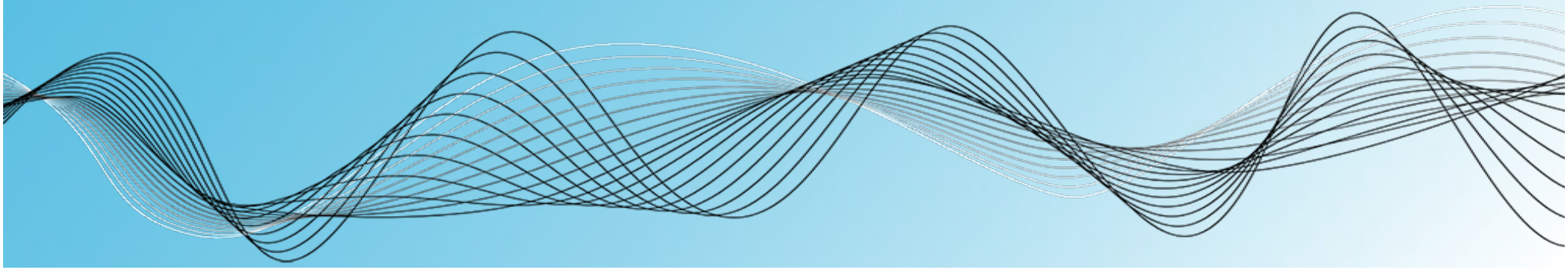
The formula for standard deviation is given by

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

← Standard deviation

← Number of samples

Confidence Interval

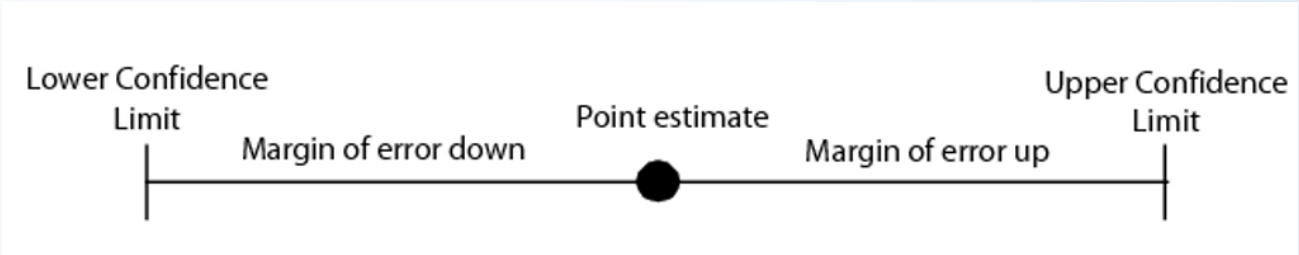


Confidence Interval

A point estimate is a single value estimate of a parameter. For instance, a sample mean is a point estimate of a population mean. An interval estimate gives you a range of values where the parameter is expected to lie.

A confidence interval is the most common type of interval estimate.

- A confidence interval displays the probability that a parameter will fall between a pair of values around the mean.
- Confidence intervals measure the degree of uncertainty or certainty in a sampling method.
- They are also used in hypothesis testing and regression analysis.
- Statisticians often use p-values in conjunction with confidence intervals to gauge statistical significance.
- They are most often constructed using confidence levels of 95% or 99%.



Margin of Error

The margin of error is defined as the range of values below and above the sample statistic in a confidence interval.

A margin of error tells you how many percentage points your results will differ from the real population value.

For example, a 95% confidence interval with a 4 percent margin of error means that your statistic will be within 4 percentage points of the real population value 95% of the time.

$$\text{Margin of error} = z^* \left(\frac{\sigma}{\sqrt{n}} \right)$$

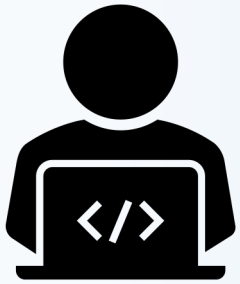
here, z^* is the critical value of the test and

$\left(\frac{\sigma}{\sqrt{n}} \right)$ is the standard deviation of the test.

Example

A poll might report that a certain candidate is going to win an election with 51 percent of the vote. Plus, the confidence level is 95 percent and the error is 4 percent. If we assume that the poll was repeated using the same techniques, then the pollsters would expect the results to be within 4 percent of the stated result (51 percent) 95 percent of the time. In other words, 95 percent of the time they would expect the results to be between:

- 51 – 4 = 47 percent and
- 51 + 4 = 55 percent



Keep Learning…… Keep Coding…… Keep going……