



CHENNAI INSTITUTE OF TECHNOLOGY

(Affiliated to Anna University, Approved by AICTE, Accredited by NAAC & NBA)
Sarathy Nagar, Kundrathur, Chennai – 600069, India.

Lecture Notes Unit II

DEPARTMENT OF INFORMATION TECHNOLOGY

Subject: CS 3353-Foundations of Data Science

Dr.A.R.Kavitha

II Year IT/III SEMESTER

CS3353 FOUNDATIONS OF DATA SCIENCE

UNIT II DESCRIBING DATA

Types of Data – Types of Variables -Describing Data with Tables and Graphs –Describing Data with Averages – Describing Variability – Normal Distributions and Standard (z) Scores

- **What are the types of data?** 2
- **How to handling data in a Tables** 6

Frequency Distribution and Data: Types, Tables, and Graphs

Frequency distribution in statistics provides the information of the number of occurrences (frequency) of distinct values distributed within a given period of time or interval, in a list, table, or graphical representation.

Types of Frequency Distribution:

There are two types of Frequency Distribution.

- Grouped
- Ungrouped

There are two types Data is a **collection of numbers or values**

Data: Any bit of information that is expressed in a **value or numerical number** is data. Data is basically a collection of information, measurements or observations.

For example

- The marks you scored in your Math exam is data
- The number of cars that pass through a bridge in a day.

Raw data :

Raw data is an initial collection of information. This information has not yet been organized. After the very first step of data collection, you will get raw data. For example,

A group of five friends their favourite colour. The answers are Blue, Green, Blue, Red, and Red. This collection of information is the raw data.

Discrete data :***Discrete data*** is that which is recorded in whole numbers, like the number of children in a school or number of tigers in a zoo. It cannot be in decimals or fractions.

Continuous data :***Continuous data*** need not be in whole numbers, it can be in decimals. Examples are the temperature in a city for a week, your percentage of marks for the last exam etc.

Example of Data Handling:

- Pictographs
- Bar Graphs
- Histogram and Pie-Charts
- Chance and Probability
- Arithmetic Mean and Median and Mode

Frequency

The frequency of any value is the number of times that value appears in a data set. So from the above examples of colours, we can say two children like the colour blue, so its frequency is two. So to make meaning of the raw data, we must organize. And finding out the frequency of the data values is how this organisation is done.

Frequency Distribution

Many times it is not easy or feasible to find the frequency of data from a very large dataset. So to make sense of the data we make a frequency table and graphs. Let us take the example of the heights of ten students in cms.

Frequency Distribution Table

139, 145, 150, 145, 136, 150, 152, 144, 138, 138

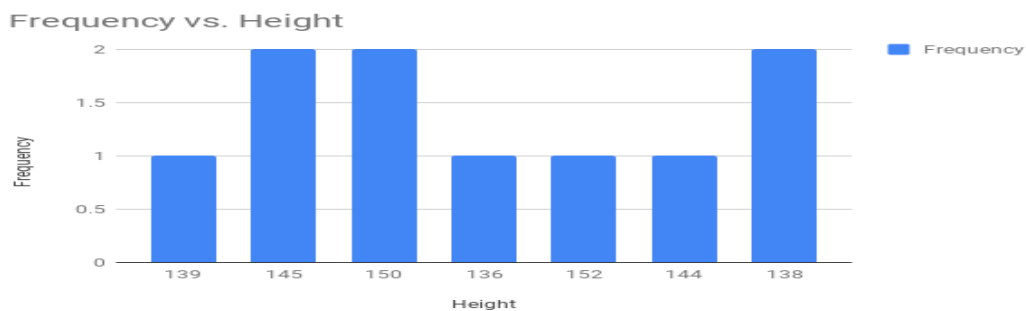
Height	Frequency
139	1
145	2
150	2
136	1
152	1
144	1
138	2

This frequency table will help us make better sense of the data given. Also when the data set is too big (say if we were dealing with 100 students) we use tally marks for counting. It makes the task more organized and easy. Below is an example of how we use tally marks.

1	I	6	I
2	II	7	II
3	III	8	III
4	IIII	9	IIII
5		10	

Frequency Distribution Graph

Using the same above example we can make the following graph:



Learn more about [Bar Graphs and Histogram here](#).

Types of Frequency Distribution

- Grouped frequency distribution.
- Ungrouped frequency distribution.
- Cumulative frequency distribution.
- Relative frequency distribution.
- Relative cumulative frequency distribution.

Grouped Data

At certain times to ensure that we are making correct and relevant observations from the data set, we may need to group the data into class intervals. This ensures that the frequency distribution best represents the data. example :the height of students.

Class Interval	Frequency
130-140	4
140-150	3
150-160	3

From the above table, you can see that the value of 150 is put in the class interval of 150-160 and not 140-150. This is the convention we must follow.

- The table gives the number of snacks ordered and the number of days as a tally. Find the frequency of snacks ordered. 2

snacks	Tally
2-4	
4-6	
6-8	
8-10	
10-12	

Answer: From the frequency table the number of snacks ordered ranging between

- 2-4 is 4 days
- 4 to 6 is 3 days
- 6 to 8 is 9 days
- 8 to 10 is 9 days
- 10 to 12 is 7 days.

So the frequencies for all snacks ordered are 4, 3, 9, 9, 7

- **How to find frequency distribution?** 2

Answer: We can find frequency distribution by the following steps:

- First of all, calculate the range of the data set.
- Next, divide the range by the number of the group you want your data in and then round up.
- After that, use class width to create groups
- Finally, find the frequency for each group.

- **Define frequency distribution in statistics?** 2

Answer: In an overview, the frequency distribution of all distinct values in some variables and the number of times they occur. Meaning that it tells how frequencies are distributed over values in a frequency distribution. However, mostly we use frequency distributions to summarize categorical variables.

- **Why are frequency distributions important?** 2

Answer: It has great importance in statistics. Also, a well-structured frequency distribution makes possible a detailed analysis of the structure of the population with respect to given characteristics. Therefore, the groups into which the population break down can be determined.

- **State the components of frequency distribution?** 2

Answer: The various components of the frequency distribution are: Class interval, types of class interval, class boundaries, midpoint or class mark, width or size o class interval, class frequency,

frequency density = class frequency/ class width,

relative frequency = class frequency/ total frequency, etc.

Descriptive Statistics and the Normal Distribution

Statistics has become the universal language of the sciences, and data analysis can lead to powerful results.

As scientists, researchers, and managers working in the natural resources sector, we all rely on statistical analysis to help us answer the questions that arise in the populations we manage. For example:

- Has there been a significant change in the mean saw timber volume in the red pine stands?
 - Has there been an increase in the number of invasive species found in the Great Lakes?
 - What proportion of white tail deer in New Hampshire have weights below the limit considered healthy?
 - Did fertilizer A, B, or C have an effect on the corn yield?
- These are typical questions that require statistical analysis for the answers.
 - In order to answer these questions, a good random sample must be collected from the population of interests.
 - We then use descriptive statistics to organize and summarize our sample data. The next step is inferential statistics, which allows us to use our sample statistics and extend the results to the population, while measuring the reliability of the result.

- But before we begin exploring different types of statistical methods, a brief review of descriptive statistics is needed.

Statistics

Statistics is the science of collecting, organizing, summarizing, analyzing, and interpreting information.

Good statistics come from good samples, and are used to draw conclusions or answer questions about a population. We use sample statistics to estimate population parameters (the truth).

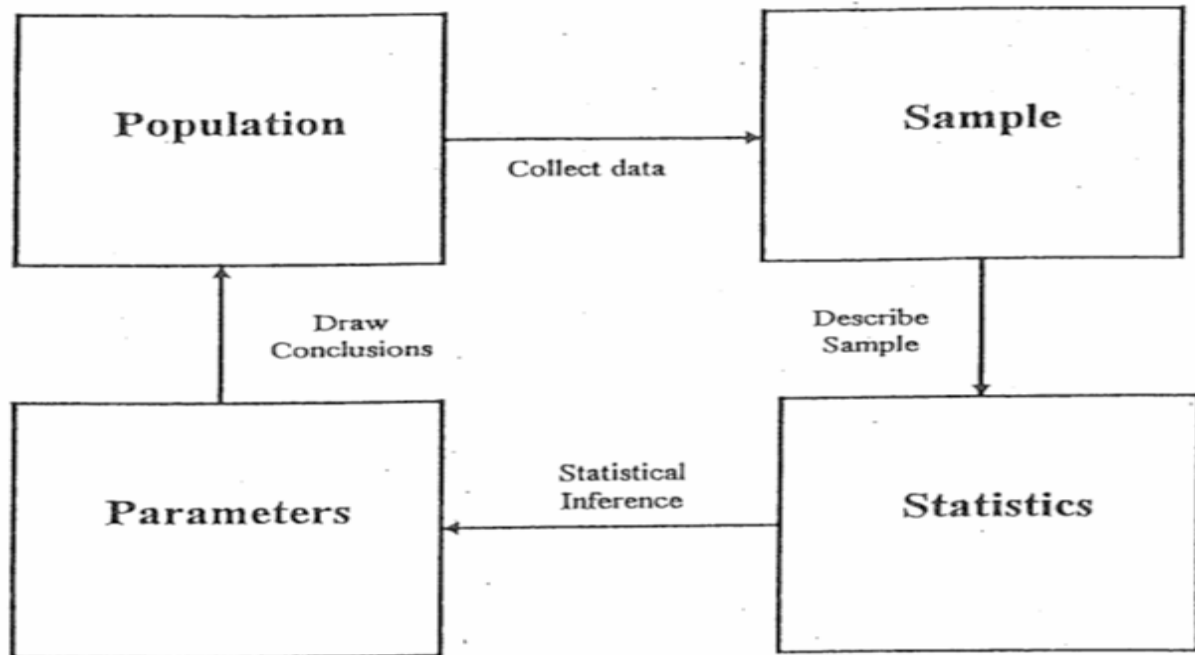


Figure 1. Using sample statistics to estimate population parameters.

Section 1: Descriptive Statistics

A population is the group to be studied, and population data is a collection of all elements in the population. For example:

- All the fish in Long Lake.
- All the lakes in the Adirondack Park.
- All the grizzly bears in Yellowstone National Park.

A sample is a subset of data drawn from the population of interest. For example:

- 100 fish randomly sampled from Long Lake.
- 25 lakes randomly selected from the Adirondack Park.
- 60 grizzly bears with a home range in Yellowstone National Park.

Populations are characterized by descriptive measures called parameters. Inferences about parameters are based on sample statistics.

For example,

The population mean (μ) is estimated by the sample mean (\bar{x}). The population variance (σ^2) is estimated by the sample variance (s^2).

Variables are the characteristics we are interested in.

For example:

- The length of fish in Long Lake.
- The pH of lakes in the Adirondack Park.
- The weight of grizzly bears in Yellowstone National Park.

Variables are divided into two major groups: **Qualitative And Quantitative**.

1. Qualitative variables

- Qualitative variables have values that are attributes or categories.
- Mathematical operations cannot be applied to qualitative variables.
- Examples of qualitative variables are gender, race, and petal color.
- Quantitative variables have values that are typically numeric, such as measurements.
- Mathematical operations can be applied to these data. Examples of quantitative variables are age, height, and length.

2. Quantitative variables

- Quantitative variables can be broken down further into two more categories: discrete and continuous variables.
- **Discrete variables** have a finite or countable number of possible values. Think of discrete variables as “hens.” Hens can lay 1 egg, or 2 eggs, or 13 eggs... There are a limited, definable number of values that the variable could take on.

- **Continuous variables** have an infinite number of possible values. Think of continuous variables as “cows.” Cows can give 4.6713245 gallons of milk, or 7.0918754 gallons of milk, or 13.272698 gallons of milk ... There are an almost infinite number of values that a continuous variable could take on.

Examples

Is the variable qualitative or quantitative?

Species	Weight	Diameter	Zip Code
(qualitative	quantitative,	quantitative,	qualitative)

Descriptive Measures

Descriptive measures of populations are called parameters and are typically written using Greek letters. The population mean is μ (mu). The population variance is σ^2 (sigma squared) and population standard deviation is σ (sigma).

Descriptive measures of samples are called statistics and are typically written using Roman letters. The sample mean is \bar{x} (x-bar). The sample variance is s^2 and the sample standard deviation is s . Sample statistics are used to estimate unknown population parameters.

These descriptive statistics help us to identify the center and spread of the data.

Measures of Center

- **Explain to find Mean, Median and Mode**

13

Mean

The arithmetic mean of a variable, often called the average, is computed by adding up all the values and dividing by the total number of values.

The population mean is represented by the Greek letter μ (mu).

The sample mean is represented by \bar{x} (x-bar).

The sample mean is usually the best, unbiased estimate of the population mean. However, the mean is influenced by extreme values (outliers) and may not be the best measure of center with strongly skewed data.

The following equations compute the population mean and sample mean.

$$\mu = \frac{\sum x_i}{N} \quad \bar{x} = \frac{\sum x_i}{n}$$

where x_i is an element in the data set, N is the number of elements in the population, and n is the number of elements in the sample data set.

Example 2

Find the mean for the following sample data set: 6.4, 5.2, 7.9, 3.4

$$\bar{x} = \frac{6.4 + 5.2 + 7.9 + 3.4}{4} = 5.725$$

Median

- The median of a variable is the middle value of the data set when the data are sorted in order from least to greatest.
- It splits the data into two equal halves with 50% of the data below the median and 50% above the median.
- The median is resistant to the influence of outliers, and may be a better measure of center with strongly skewed data.



The calculation of the median depends on the number of observations in the data set.

To calculate the median with an odd number of values (n is odd), first sort the data from smallest to largest.

Example 3

23, 27, 29, 31, 35, 39, 40, 42, 44, 47, 51

The median is 39. It is the middle value that separates the lower 50% of the data from the upper 50% of the data.

To calculate the median with an even number of values (n is even), first sort the data from smallest to largest and take the average of the two middle values.

Example 4

23, 27, 29, 31, 35, 39, 40, 42, 44, 47

$$M = \frac{35 + 39}{2} = 37$$

Mode

- The mode is the most frequently occurring value and is commonly used with qualitative data as the values are categorical.
- Categorical data cannot be added, subtracted, multiplied or divided, so the mean and median cannot be computed.
- The mode is less commonly used with quantitative data as a measure of center. Sometimes each value occurs only once and the mode will not be meaningful.
- Understanding the relationship between the mean and median is important.
- It gives us insight into the distribution of the variable.
- For example, if the distribution is skewed right (positively skewed), the mean will increase to account for the few larger observations that pull the distribution to the right.
- The median will be less affected by these extreme large values, so in this situation, the mean will be larger than the median.

- In a symmetric distribution, the mean, median, and mode will all be similar in value. If the distribution is skewed left (negatively skewed), the mean will decrease to account for the few smaller observations that pull the distribution to the left.
- Again, the median will be less affected by these extreme small observations, and in this situation, the mean will be less than the median.

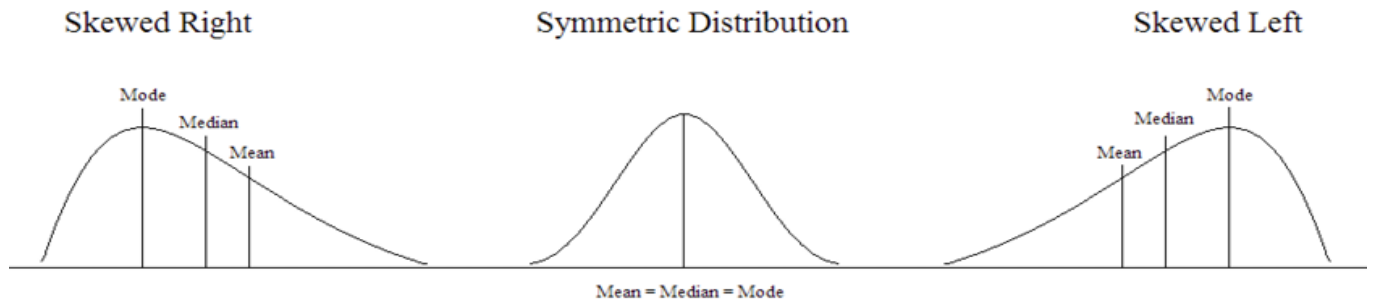


Figure 2. Illustration of skewed and symmetric distributions.

Measures of Dispersion

- Measures of center look at the average or middle values of a data set.
- Measures of dispersion look at the spread or variation of the data.
- Variation refers to the amount that the values vary among themselves.
- Values in a data set that are relatively close to each other have lower measures of variation. Values that are spread farther apart have higher measures of variation.

Examine the two histograms below. Both groups have the same mean weight, but the values of Group A are more spread out compared to the values in Group B. Both groups have an average weight of 267 lb. but the weights of Group A are more variable.

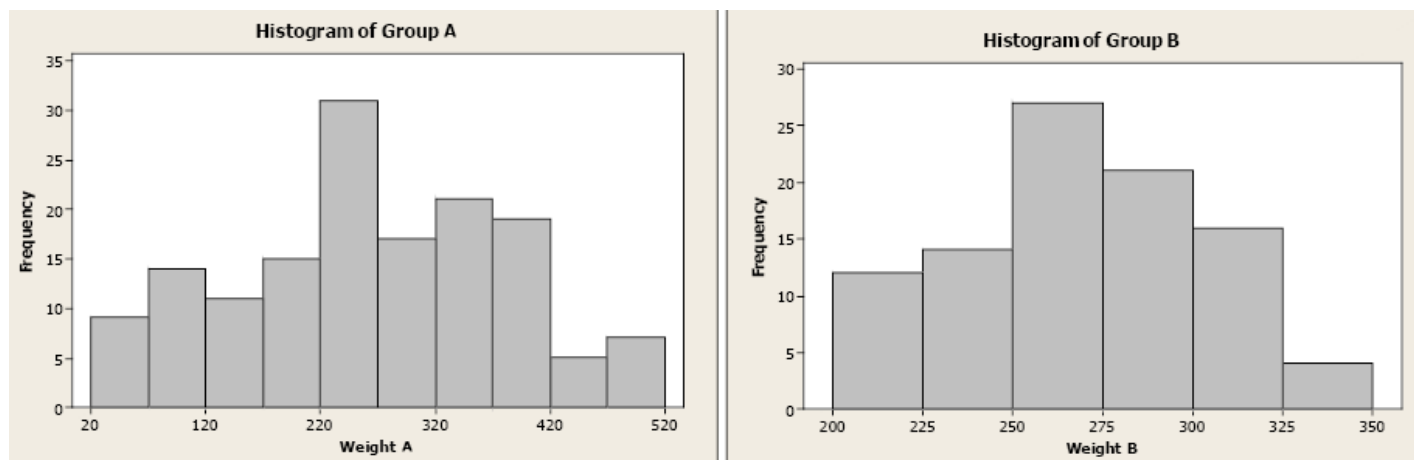


Figure 3. Histograms of Group A and Group B.

There are five measures of dispersion: range, variance, standard deviation, standard error, and coefficient of variation.

- **What is meant by Range** 2
- **Difference between range and variance** 2

Range

The range of a variable is the largest value minus the smallest value. It is the simplest measure and uses only these two values in a quantitative data set.

Example 5

Find the range for the given data set.

12, 29, 32, 34, 38, 49, 57

$$\text{Range} = 57 - 12 = 45$$

Variance

- The variance uses the difference between each value and its arithmetic mean.
- The differences are squared to deal with positive and negative differences.
- The sample variance (s^2) is an unbiased estimator of the population variance (σ^2), with $n-1$ degrees of freedom.

Degrees of freedom: In general, the degrees of freedom for an estimate is equal to the number of values minus the number of parameters estimated en route to the estimate in question.

- The sample variance is unbiased due to the difference in the denominator.
- If we used “ n ” in the denominator instead of “ $n - 1$ ”, we would consistently underestimate the true population variance.
- To correct this bias, the denominator is modified to “ $n - 1$ ”.

Population variance

Sample variance

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N} \quad s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} = \frac{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}{n-1}$$

Example 6

Compute the variance of the sample data: 3, 5, 7. The sample mean is 5.

$$s^2 = \frac{(3-5)^2 + (5-5)^2 + (7-5)^2}{3-1} = 4$$

Standard Deviation

- The standard deviation is the square root of the variance (both population and sample).
- While the sample variance is the positive, unbiased estimator for the population variance, the units for the variance are squared.
- The standard deviation is a common method for numerically describing the distribution of a variable. The population standard deviation is σ (sigma) and sample standard deviation is s .

Population standard deviation

Sample standard deviation

$$\sigma = \sqrt{\sigma^2}$$

$$s = \sqrt{s^2}$$

Example 7

Compute the standard deviation of the sample data: 3, 5, 7 with a sample mean of 5.

$$s = \sqrt{\frac{(3-5)^2 + (5-5)^2 + (7-5)^2}{3-1}} = \sqrt{4} = 2$$

Standard Error of the Means

Commonly, we use the sample mean \bar{x} to estimate the population mean μ .

For example, if we want to estimate the heights of eighty-year-old cherry trees,

- Randomly select 100 trees
- Compute the sample mean of the 100 heights
- Use that as our estimate
 - We want to use this sample mean to estimate the true but unknown population mean.
 - But our sample of 100 trees is just one of many possible samples (of the same size) that could have been randomly selected.
 - Imagine if we take a series of different random samples from the same population and all the same size:
- Sample 1—we compute sample mean \bar{x}
- Sample 2—we compute sample mean \bar{x}
- Sample 3—we compute sample mean \bar{x} Etc.

In this above example, we may get a different result as we are using a different subset of data to compute the sample mean. This shows us that the sample mean is a random variable!

The sample mean (\bar{x}) is a random variable with its own probability distribution called the sampling distribution of the sample mean. The distribution of the sample mean will have a mean equal to μ and a standard deviation equal to $\frac{s}{\sqrt{n}}$.

The standard error $\frac{s}{\sqrt{n}}$ is the standard deviation of all possible sample means.

The standard error is the standard deviation of the sample means and can be expressed in different ways.

$$s_{\bar{x}} = \sqrt{\frac{s^2}{n}} = \frac{s}{\sqrt{n}}$$

Note: s^2 is the sample variance and s is the sample standard deviation

Example 8

Describe the distribution of the sample mean.

A population of fish has weights that are normally distributed with $\mu = 8$ lb. and $s = 2.6$ lb. If you take a sample of size $n=6$, the sample mean will have a normal distribution with a mean of 8 and a standard deviation (standard error) of $\frac{2.6}{\sqrt{6}} = 1.061$ lb.

If you increase the sample size to 10, the sample mean will be normally distributed with a mean of 8 lb. and a standard deviation (standard error) of $\frac{2.6}{\sqrt{10}} = 0.822$ lb.

Notice how the standard error decreases as the sample size increases.

- The Central Limit Theorem (CLT) states that the sampling distribution of the sample means will approach a normal distribution as the sample size increases.
- If we do not have a normal distribution, or know nothing about our distribution of our random variable, the CLT tells us that the distribution of the \bar{x} 's will become normal as n increases.
- How large does n have to be? A general rule of thumb tells us that $n \geq 30$.

The Central Limit Theorem tells us that regardless of the shape of our population, the sampling distribution of the sample mean will be normal as the sample size increases.

Coefficient of Variation

To compare standard deviations between different populations or samples is difficult because the standard deviation depends on units of measure.

The coefficient of variation expresses the standard deviation as a percentage of the sample or population mean. It is a unit less measure.

Population data

Sample data

$$CV = \frac{\sigma}{\mu} * 100$$

$$CV = \frac{s}{\bar{x}} * 100$$

Example 9

Fisheries biologists were studying the length and weight of Pacific salmon. They took a random sample and computed the mean and standard deviation for length and weight (given below). While the standard deviations are similar, the differences in units between lengths and weights make it difficult to compare the variability. Computing the coefficient of variation for each variable allows the biologists to determine which variable has the greater standard deviation.

	Sample mean	Sample standard deviation
Length	63 cm	19.97 cm
Weight	37.6 kg	19.39 kg

$$CV_L = \frac{19.97}{63.0} * 100 = 31.7\% \quad CV_W = \frac{19.39}{37.6} * 100 = 51.6\%$$

There is greater variability in Pacific salmon weight compared to length.

Variability

- Variability is described in many different ways.
- Standard deviation measures point to point variability within a sample, i.e., variation among individual sampling units.
- Coefficient of variation also measures point to point variability but on a relative basis (relative to the mean), and is not influenced by measurement units.
- Standard error measures the sample to sample variability, i.e. variation among repeated samples in the sampling process.

- Typically, we only have one sample and standard error allows us to quantify the uncertainty in our sampling process.

Basic Statistics Example using Excel and Minitab Software

Consider the following tally from 11 sample plots on Heiburg Forest, where X_i is the number of downed logs per acre. Compute basic statistics for the sample plots.

ID	X_i	X_i^2	$(X_i - \bar{X})$	$(X_i - \bar{X})^2$	Order
1	25	625	-7.27	52.8529	4
2	35	1225	2.73	7.4529	6
3	55	3025	22.73	516.6529	10
4	15	225	-17.25	298.2529	2
5	40	1600	7.73	59.7529	8
6	25	625	-7.27	52.8529	5
7	55	3025	22.73	516.6529	11
8	35	1225	2.73	7.4529	7
9	45	2025	12.73	162.0529	9
10	5	25	-27.27	743.6529	1
11	20	400	-12.27	150.1819	3
Sum	355	14025	0.0	2568.1519	
	$\sum_{i=1}^n X_i$	$\sum_{i=1}^n X_i^2$	$\sum_{i=1}^n (X_i - \bar{X})$	$\sum_{i=1}^n (X_i - \bar{X})^2$	

Table 1. Sample data on number of downed logs per acre from Heiburg Forest.

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n} = \frac{355}{11} = 32.27$$

(1) Sample mean:

(2) Median = 35

(3) Variance:

$$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1} = \frac{2568.1519}{11-1} = 256.82$$

$$= \frac{\sum_{i=1}^n X_i^2 - \frac{\left(\sum_{i=1}^n X_i\right)^2}{n}}{n-1} = \frac{14025 - \frac{(355)^2}{11}}{11-1} = 256.82$$

(4) Standard deviation: $S = \sqrt{S^2} = \sqrt{256.82} = 16.0256$

(5) Range: $55 - 5 = 50$

(6) Coefficient of variation:

$$CV = \frac{S}{\bar{X}} \cdot 100 = \frac{16.0256}{32.27} \cdot 100 = 49.66\%$$

(7) Standard error of the mean:

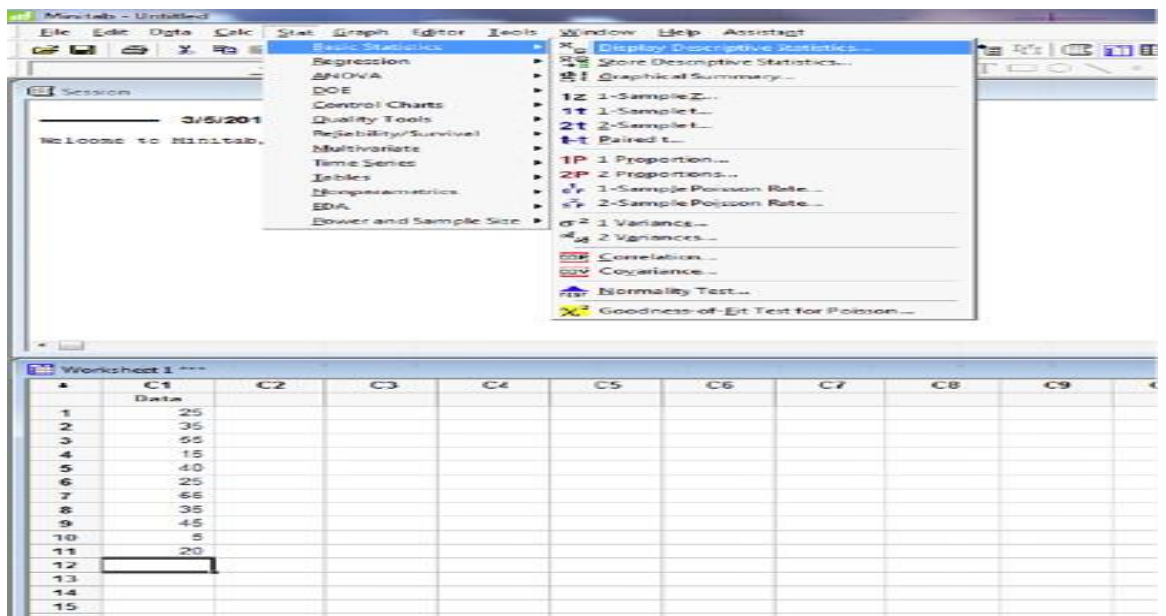
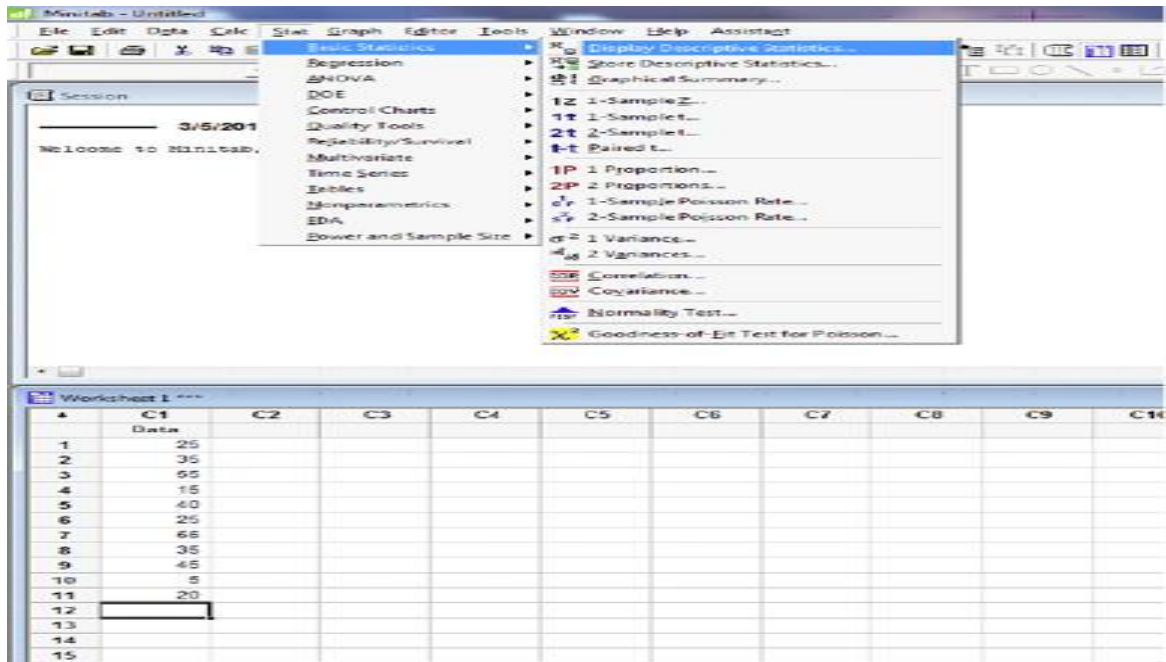
$$S_{\bar{X}} = \sqrt{\frac{S^2}{n}} = \sqrt{\frac{256.82}{11}} = 4.8319$$

$$= \frac{S}{\sqrt{n}} = \frac{16.0256}{\sqrt{11}} = 4.8319$$

Software Solutions

Minitab

Open Minitab and enter data in the spreadsheet. Select STAT>Descriptive stats and check all statistics required.



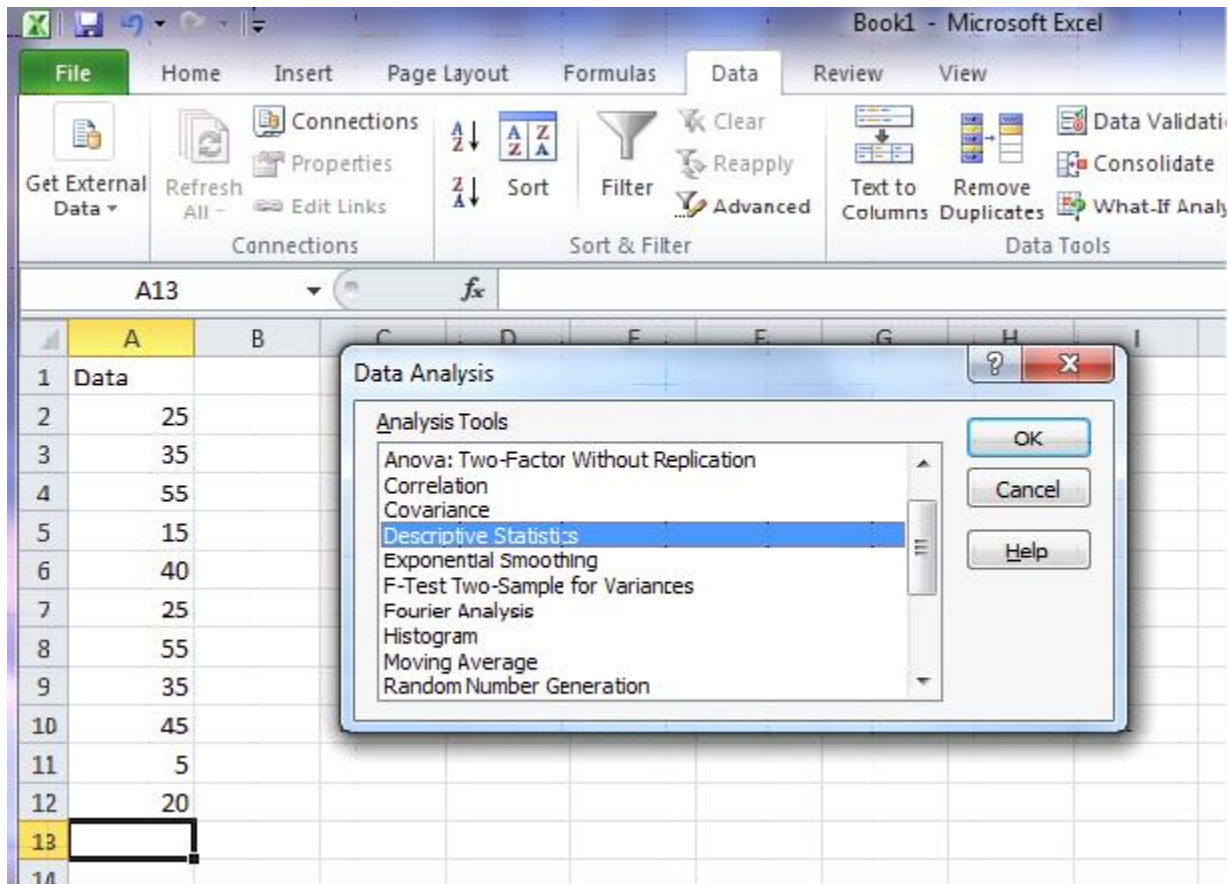
Descriptive Statistics: Data

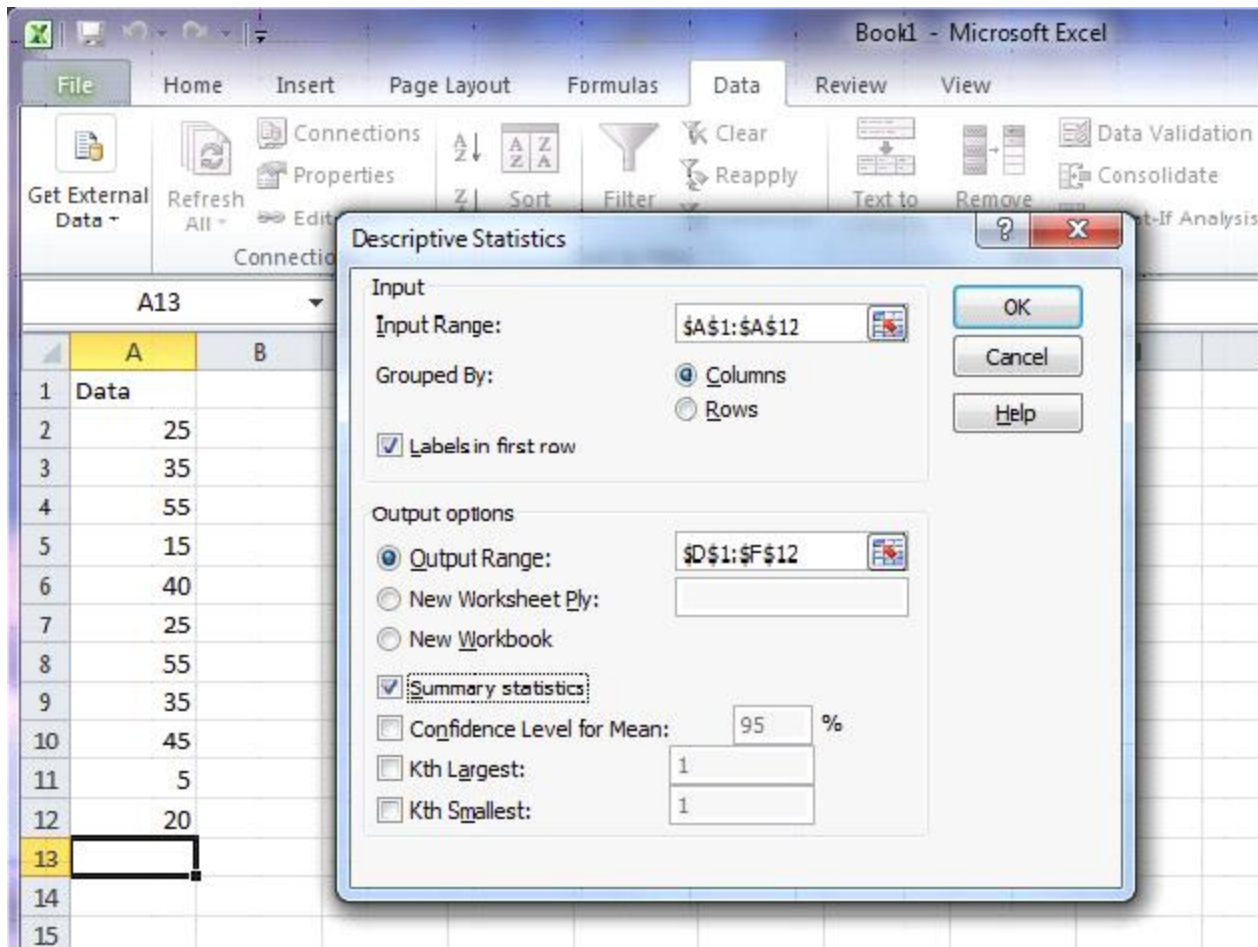
Variable	N	N*	Mean	SE Mean	StDev	Variance	CoefVar	Minimum	Q1
Data	11	0	32.27	4.83	16.03	256.82	49.66	5.00	20.00

Variable	Median	Q3	Maximum	IQR
Data	35.00	45.00	55.00	25.00

Excel

Open up Excel and enter the data in the first column of the spreadsheet. Select DATA>Data Analysis>Descriptive Statistics. For the Input Range, select data in column A. Check “Labels in First Row” and “Summary Statistics”. Also check “Output Range” and select location for output.





Data

Mean	32.27273
Standard Error	4.831884
Median	35
Mode	25
Standard Deviation	16.02555
Sample Variance	256.8182
Kurtosis	-0.73643
Skewness	-0.05982
Range	50
Minimum	5

Maximum	55
Sum	355
Count	11

Graphical Representation

Data organization and summarization can be done graphically, as well as numerically. Tables and graphs allow for a quick overview of the information collected and support the presentation of the data used in the project. While there are a multitude of available graphics, this chapter will focus on a specific few commonly used tools.

Pie Charts

Pie charts are a good visual tool allowing the reader to quickly see the relationship between categories. It is important to clearly label each category, and adding the frequency or relative frequency is often helpful. However, too many categories can be confusing. Be careful of putting too much information in a pie chart. The first pie chart gives a clear idea of the representation of fish types relative to the whole sample. The second pie chart is more difficult to interpret, with too many categories. It is important to select the best graphic when presenting the information to the reader.

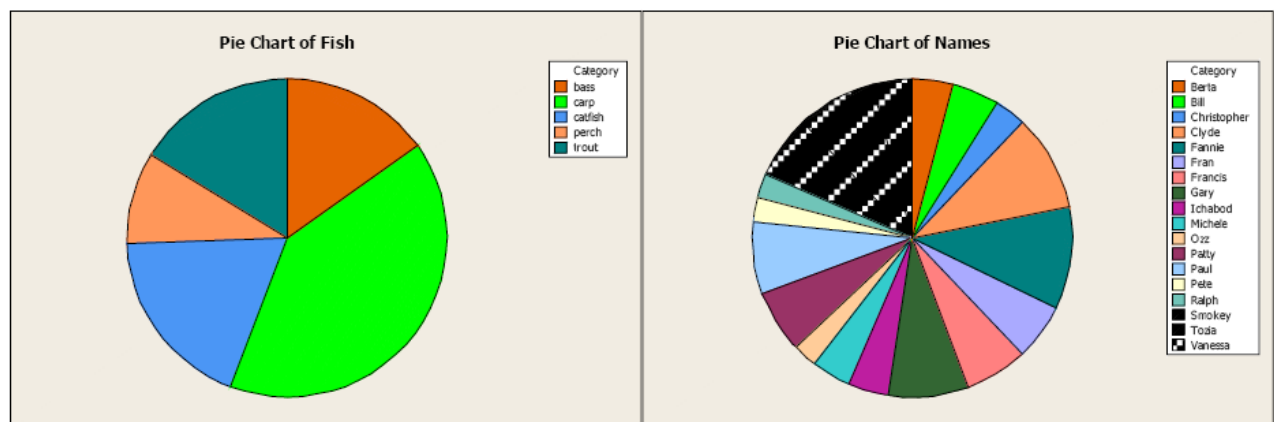


Figure 4. Comparison of pie charts.

What is a Pie Chart?

The “**pie chart**” is also known as a “circle chart”, dividing the circular statistical graphic into sectors or sections to illustrate the numerical problems. Each sector denotes a proportionate part of the whole. To find out the composition of something, Pie-chart works the best at that time. In

most cases, pie charts replace other graphs like the bar graph, line plots, histograms, etc.

Formula

The pie chart is an important type of data representation. It contains different segments and sectors in which each segment and sector of a pie chart forms a specific portion of the total(percentage). The sum of all the data is equal to 360° .

The total value of the pie is always 100%.

To work out with the percentage for a pie chart, follow the steps given below:

- Categorize the data
- Calculate the total
- Divide the categories
- Convert into percentages
- Finally, calculate the degrees

Therefore, the pie chart formula is given as

$$(\text{Given Data/Total value of Data}) \times 360^\circ$$

Note: It is not mandatory to convert the given data into percentages until it is specified. We can directly calculate the degrees for given data values and draw the pie chart accordingly.

How to Create a Pie Chart?

Imagine a teacher surveys her class on the basis of favourite Sports of students:

Football	Hockey	Cricket	Basketball	Badminton
10	5	5	10	10

The data above can be represented by a pie chart as following and by using the circle graph formula, i.e. the pie chart formula given below. It makes the size of the portion easy to understand.

Step 1: First, Enter the data into the table.

Football	Hockey	Cricket	Basketball	Badminton
10	5	5	10	10

Step 2: Add all the values in the table to get the total.

I.e. Total students are 40 in this case.

Step 3: Next, divide each value by the total and multiply by 100 to get a per cent:

Football	Hockey	Cricket	Basketball	Badminton
$(10/40) \times 100$ =25%	$(5/40) \times 100$ =12.5%	$(5/40) \times 100$ =12.5%	$(10/40) \times 100$ =25%	$(10/40) \times 100$ =25%

Step 4: Next to know how many degrees for each “pie sector” we need, we will take a full circle of 360° and follow the calculations below:

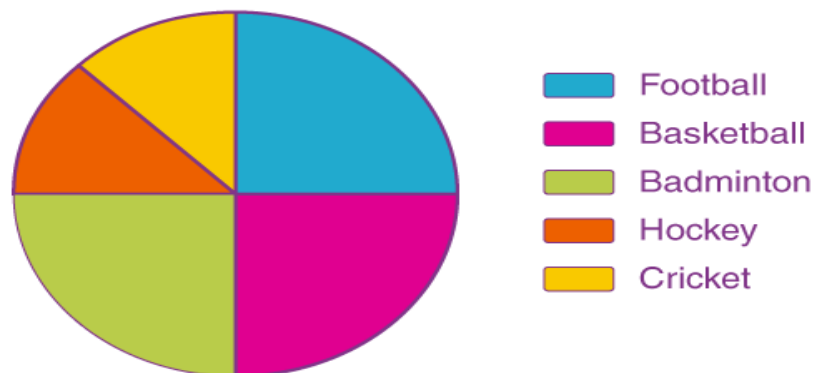
The central angle of each component = (Value of each component/sum of values of all the components) $\times 360^\circ$

Football	Hockey	Cricket	Basketball	Badminton
$(10/40) \times 360^\circ$ =90°	$(5/40) \times 360^\circ$ =45°	$(5/40) \times 360^\circ$ =45°	$(10/40) \times 360^\circ$ =90°	$(10/40) \times 360^\circ$ =90°

Now you can draw a pie chart.

Step 5: Draw a circle and use the protractor to measure the degree of each sector.

Favourite Sports Percentage



Let us take an example for a pie chart with an explanation here to understand the concept in a better way.

Question: The percentages of various crops cultivated in a village of particular district are given in the following table.

Items	Wheat	Pulses	Jowar	Groundnuts	Vegetables	Total
Percentage of crops	125/3	125/6	25/2	50/3	25/3	100

Represent this information using a pie-chart.

Solution:

The central angle = $(\text{component value}/100) \times 360^\circ$

The central angle for each category is calculated as follows

Items	Percentage of crops	Central angle
Wheat	125/3	$[(125/3)/100] \times 360^\circ = 150^\circ$
Pulses	125/6	$[(125/6)/100] \times 360^\circ = 75^\circ$
Jowar	25/2	$[(25/2)/100] \times 360^\circ = 45^\circ$
Groundnuts	50/3	$[(50/3)/100] \times 360^\circ = 60^\circ$
Vegetables	25/3	$[(25/3)/100] \times 360^\circ = 30^\circ$
Total	100	360°

Now, the pie-chart can be constructed by using the given data.

Steps to construct:

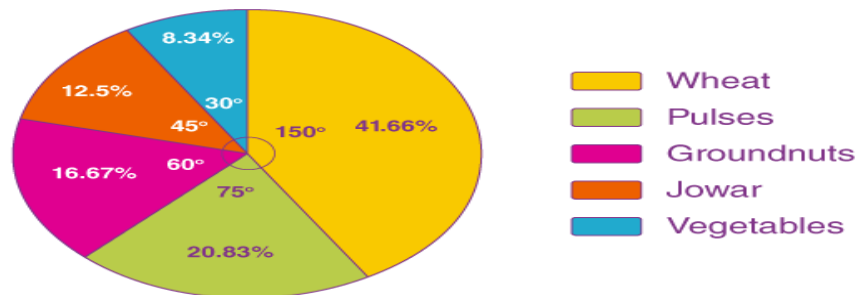
Step 1: Draw the circle of an appropriate radius.

Step 2: Draw a vertical radius anywhere inside the circle.

Step 3: Choose the largest central angle. Construct a sector of a central angle, whose one radius coincides with the radius drawn in step 2, and the other radius is in the clockwise direction to the vertical radius.

Step 4: Construct other sectors representing other values in the clockwise direction in descending order of magnitudes of their central angles.

Step 5: Shade the sectors obtained by different colours and label them as shown in the figure below.

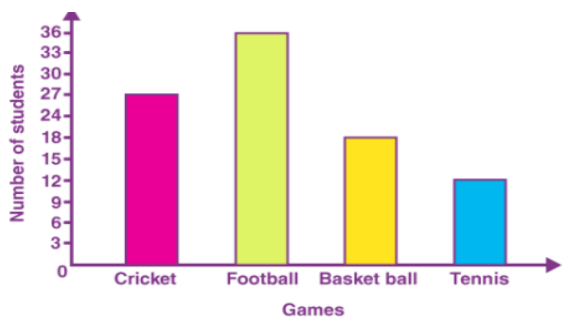
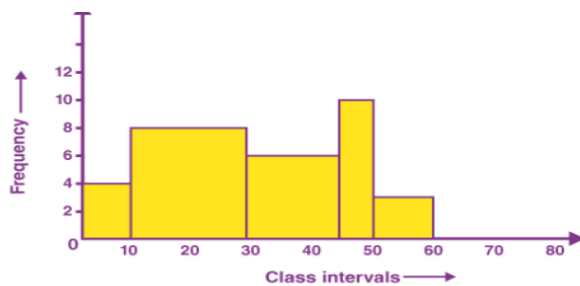


Bar Charts and Histograms

Bar charts graphically describe the distribution of a qualitative variable (fish type) while histograms describe the distribution of a quantitative variable discrete or continuous variables (bear weight).

The major **difference between Bar Chart and Histogram** is the bars of bar chart are not just next to each other. In histogram, the bars are adjacent to each other. In statistics, bar chart and histogram are important for expressing huge or big number of data. The similarity between bar chart and histogram is both are the pictorial representation of grouped data. Here, we will learn histogram vs bar graph with examples.

What is the difference between bar chart and histogram?

Bar graph	Histogram
Bar graph is the graphical representation of categorical data	Bar graph is the graphical representation of grouped data in continuous manner
There is equal space between each pair of consecutive bars	There is no space between the consecutive bars
The height of the bars shows the frequency and the width gap is zero	The frequency of the data is shown by the area of rectangular bars
	

Bar Graph V/s Histogram

A bar graph is a pictorial representation using vertical and horizontal bars in a graph. The length of bars are proportional to the measure of data. It is also called bar chart.

A histogram is also a pictorial representation of data using rectangular bars, that are adjacent to each other. It is used to represent grouped frequency distribution with continuous classes.

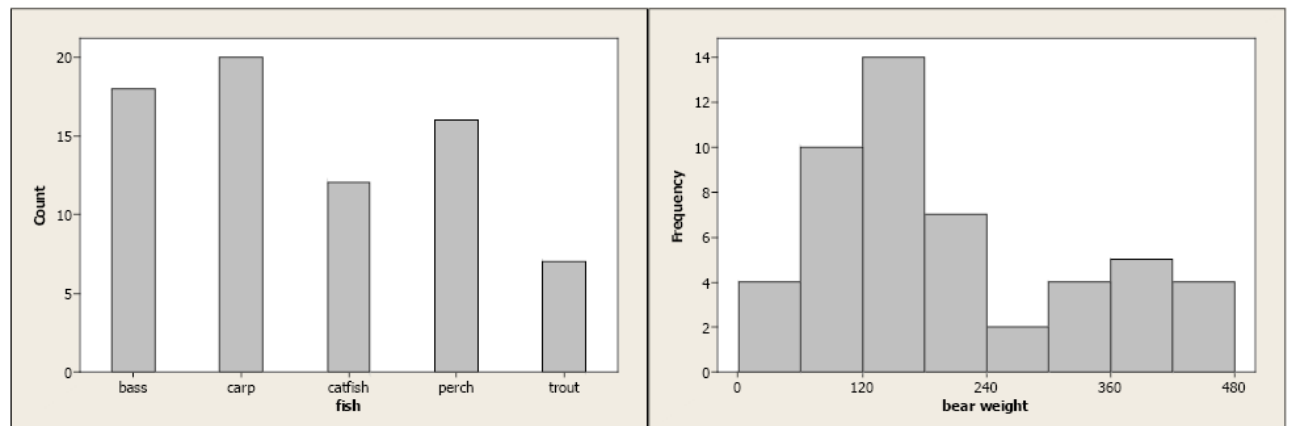


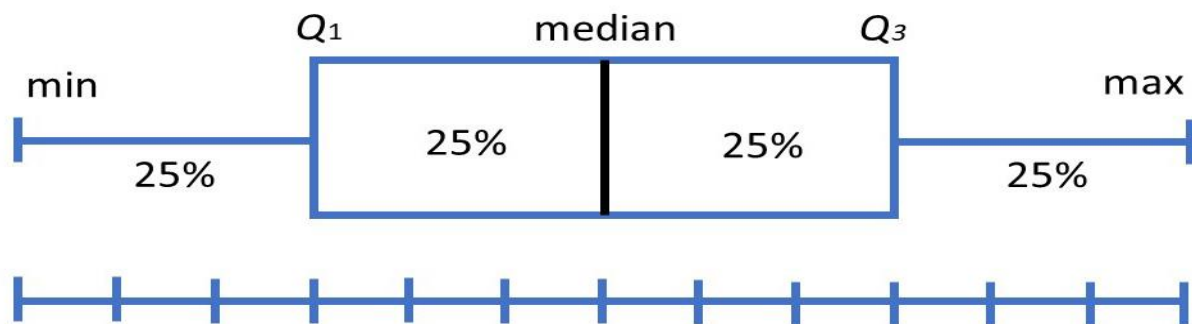
Figure 5. Comparison of a bar chart for qualitative data and a histogram for quantitative data.

In both cases, the bars' equal width and the y-axis are clearly defined. With qualitative data, each category is represented by a specific bar. With continuous data, lower and upper class limits must be defined with equal class widths. There should be no gaps between classes and each observation should fall into one, and only one, class.

Boxplots

- Boxplots use the 5-number summary (minimum and maximum values with the three quartiles) to illustrate the center, spread, and distribution of your data.
- When paired with histograms, they give an excellent description, both numerically and graphically, of the data.
- With symmetric data, the distribution is bell-shaped and somewhat symmetric.
- In the boxplot, we see that Q1 and Q3 are approximately equidistant from the median, as are the minimum and maximum values. Also, both whiskers (lines extending from the boxes) are approximately equal in length.

Box plots divide the data into sections that each contain approximately 25% of the data in that set.



Box plots are useful as they provide a visual summary of the data enabling researchers to quickly identify mean values, the dispersion of the data set, and signs of skewness.

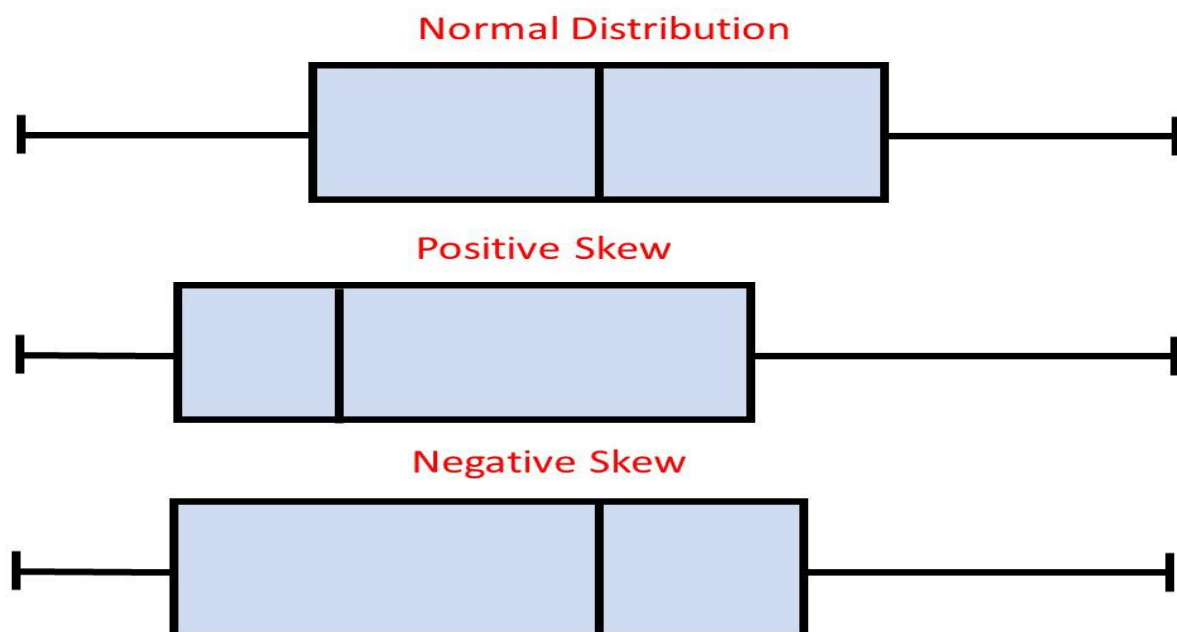
Note the image above represents data which is a perfect [normal distribution](#) and most box plots will not conform to this symmetry (where each quartile is the same length).

Box plots are useful as they show the average score of a data set.

The median is the average value from a set of data and is shown by the line that divides the box into two parts. Half the scores are greater than or equal to this value and half are less.

Box plots are useful as they show the skewness of a data set

The box plot shape will show if a statistical data set is normally distributed or skewed.



When the median is in the middle of the box, and the whiskers are about the same on both sides of the box, then the distribution is symmetric.

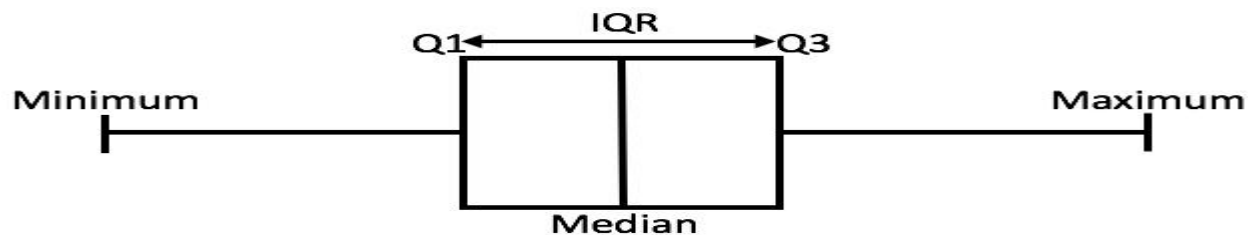
When the median is closer to the bottom of the box, and if the whisker is shorter on the lower end of the box, then the distribution is positively skewed (skewed right).

When the median is closer to the top of the box, and if the whisker is shorter on the upper end of the box, then the distribution is negatively skewed (skewed left).

Box plots are useful as they show the dispersion of a data set.

In statistics, dispersion (also called variability, scatter, or spread) is the extent to which a distribution is stretched or squeezed.

The smallest value and largest value are found at the end of the 'whiskers' and are useful for providing a visual indicator regarding the spread of scores (e.g. the range).



The interquartile range (IQR) is the box plot showing the middle 50% of scores and can be calculated by subtracting the lower quartile from the upper quartile (e.g. $Q3 - Q1$).

Box plots are useful as they show outliers within a data set.

An outlier is an observation that is numerically distant from the rest of the data.

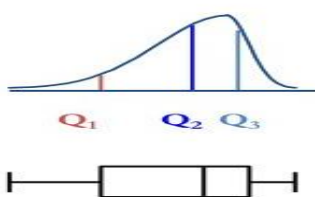
When reviewing a box plot, an outlier is defined as a data point that is located outside the whiskers of the box plot.

For example, outside 1.5 times the interquartile range above the upper quartile and below the lower quartile ($Q1 - 1.5 * IQR$ or $Q3 + 1.5 * IQR$).

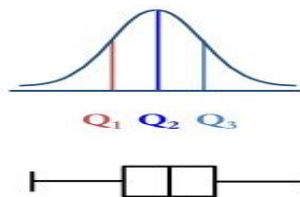
signs of skewness

If the data do not appear to be symmetric, does each sample show the same kind of asymmetry?

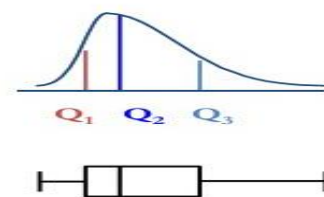
Left-Skewed



Symmetric



Right-Skewed



The median (Q_2) divides the data set into two parts, the upper set and the lower set. The **lower quartile** (Q_1) is the median of the lower half, and the **upper quartile** (Q_3) is the median of the upper half.

Example:

Find Q_1 , Q_2 , and Q_3 for the following data set, and draw a box-and-whisker plot.

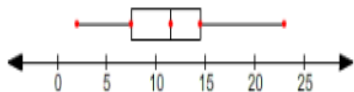
$\{2, 6, 7, 8, 8, 11, 12, 13, 14, 15, 22, 23\}$

There are 12 data points. The middle two are 11 and 12. So the median, Q_2 , is 11.5.

The "lower half" of the data set is the set $\{2, 6, 7, 8, 8, 11\}$. The median here is 7.5. So $Q_1 = 7.5$.

The "upper half" of the data set is the set $\{12, 13, 14, 15, 22, 23\}$. The median here is 14.5. So $Q_3 = 14.5$.

A box-and-whisker plot displays the values Q_1 , Q_2 , and Q_3 , along with the extreme values of the data set (2 and 23, in this case):



A box & whisker plot shows a "box" with left edge at Q_1 , right edge at Q_3 , the "middle" of the box at Q_2 (the median) and the maximum and minimum as "whiskers".

Note that the plot divides the data into 4 equal parts. The left whisker represents the bottom 25% of the data, the left half of the box represents the second 25%, the right half of the box represents the third 25%, and the right whisker represents the top 25%.

Outliers

If a data value is very far away from the quartiles (either much less than Q_1 or much greater than Q_3), it is sometimes designated an **outlier**. Instead of being shown using the whiskers of the box-and-whisker plot, outliers are usually shown as separately plotted points.

The standard definition for an outlier is a number which is less than Q_1 or greater than Q_3 by more than 1.5 times the **interquartile range** ($IQR = Q_3 - Q_1$). That is, an outlier is any number less than $Q_1 - (1.5 \times IQR)$ or greater than $Q_3 + (1.5 \times IQR)$.

Example:

Find Q_1 , Q_2 , and Q_3 for the following data set. Identify any outliers, and draw a box-and-whisker plot.

$\{5, 40, 42, 46, 48, 49, 50, 50, 52, 53, 55, 56, 58, 75, 102\}$

There are 15 values, arranged in increasing order. So, Q_2 is the 8th data point, 50.

Q_1 is the 4th data point, 46, and Q_3 is the 12th data point, 56.

The interquartile range IQR is $Q_3 - Q_1$ or $56 - 46 = 10$.

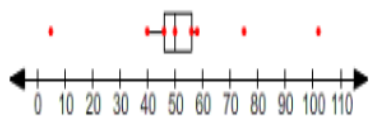
Now we need to find whether there are values less than $Q_1 - (1.5 \times \text{IQR})$ or greater than $Q_3 + (1.5 \times \text{IQR})$.

$$Q_1 - (1.5 \times \text{IQR}) = 46 - 15 = 31$$

$$Q_3 + (1.5 \times \text{IQR}) = 56 + 15 = 71$$

Since 5 is less than 31 and 75 and 102 are greater than 71, there are 3 outliers.

The box-and-whisker plot is as shown. Note that 40 and 58 are shown as the ends of the whiskers, with the outliers plotted separately.



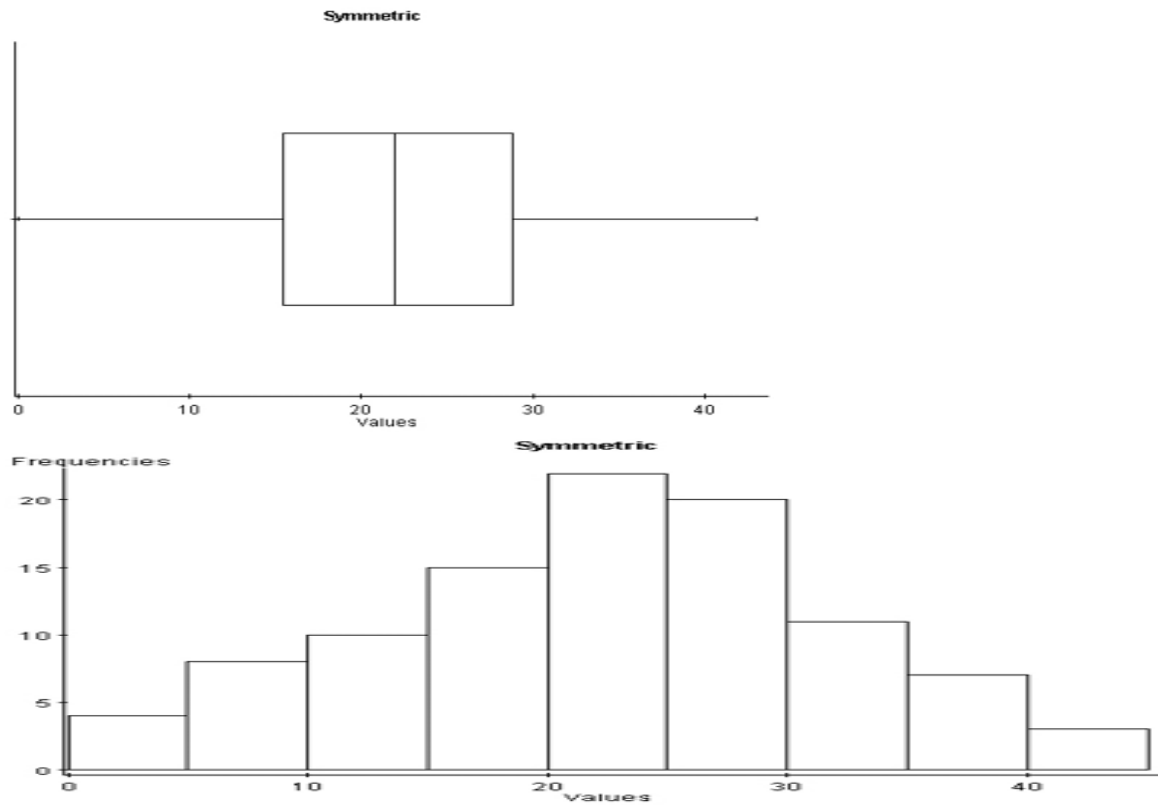


Figure 6. A histogram and boxplot of a normal distribution.

With skewed left distributions, we see that the histogram looks “pulled” to the left. In the boxplot, Q1 is farther away from the median as are the minimum values, and the left whisker is longer than the right whisker.

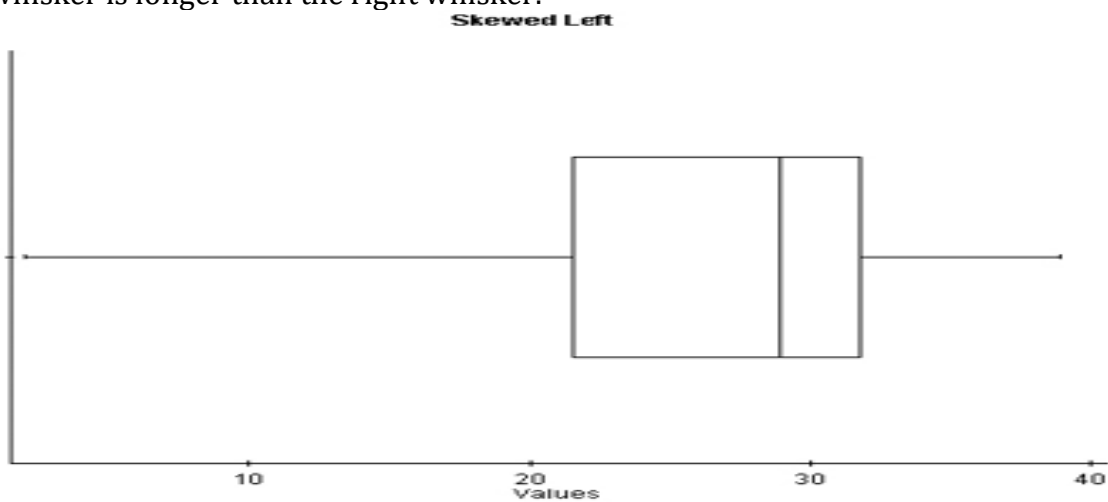
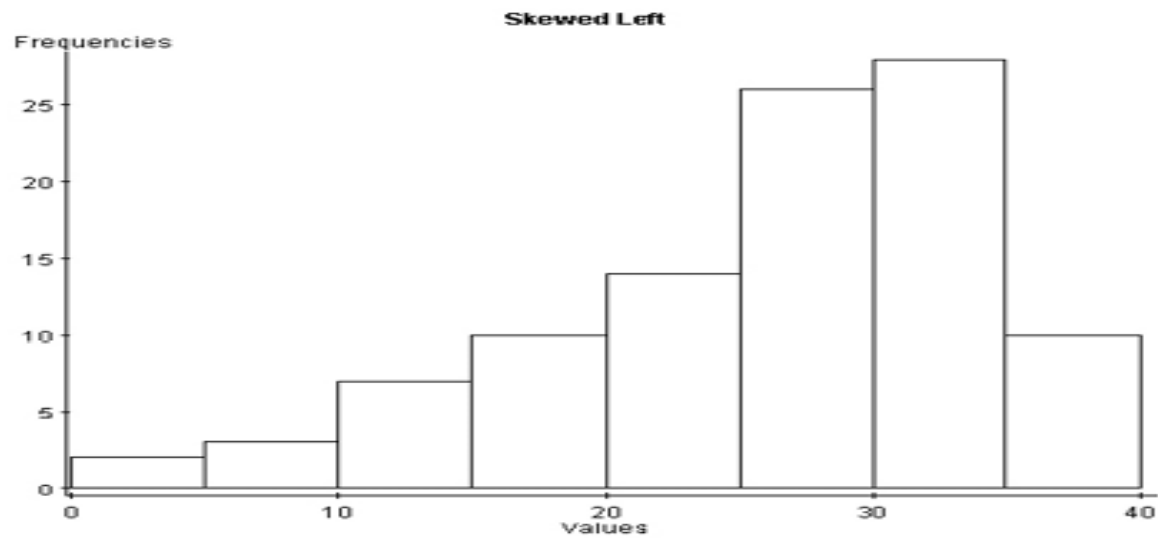


Figure 7. A histogram and boxplot of a skewed left distribution.



With skewed right distributions, we see that the histogram looks “pulled” to the right. In the boxplot, Q3 is farther away from the median, as is the maximum value, and the right whisker is longer than the left whisker.

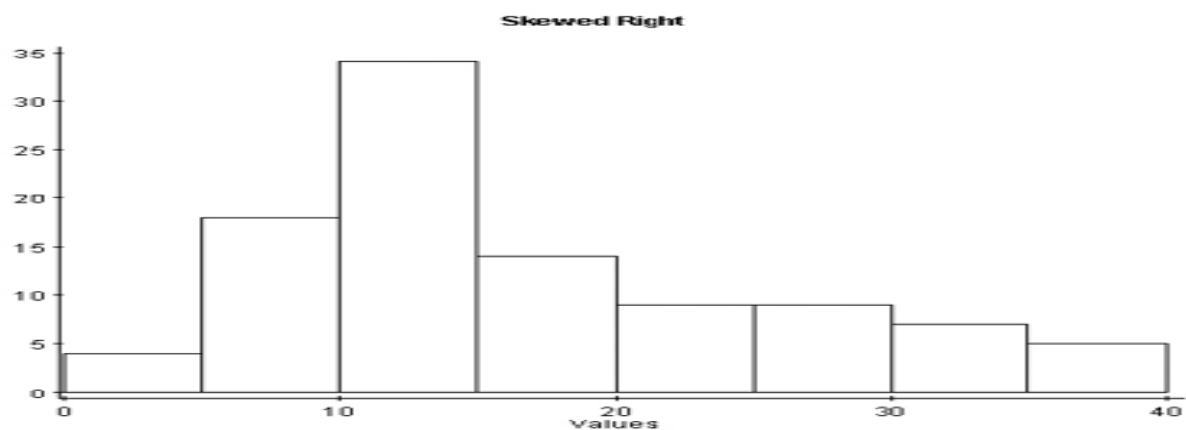
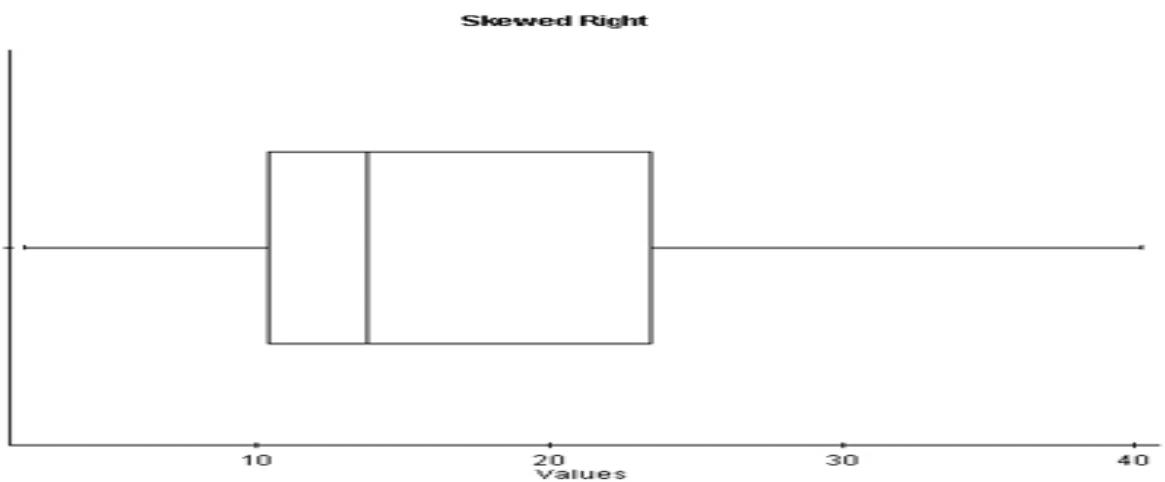


Figure 8. A histogram and boxplot of a skewed right distribution.

Probability Distribution

Once we have organized and summarized your sample data, the next step is to identify the underlying distribution of our random variable.

Computing probabilities for continuous random variables are complicated by the fact that there are an infinite number of possible values that our random variable can take on, so the probability of observing a particular value for a random variable is zero.

Therefore, to find the probabilities associated with a continuous random variable, we use a probability density function (PDF).

A PDF is an equation used to find probabilities for continuous random variables. The PDF must satisfy the following two rules:

1. The area under the curve must equal one (over all possible values of the random variable).
2. The probabilities must be equal to or greater than zero for all possible values of the random variable.

The area under the curve of the probability density function over some interval represents the probability of observing those values of the random variable in that interval.

The Normal Distribution

Many continuous random variables have a bell-shaped or somewhat symmetric distribution.

This is a normal distribution. In other words, the probability distribution of its relative frequency histogram follows a normal curve.

The curve is bell-shaped, symmetric about the mean, and defined by μ and σ (the mean and standard deviation).

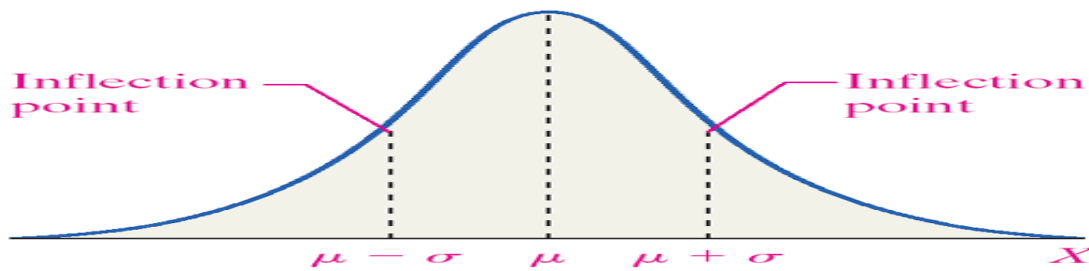


Figure 9. A normal distribution.

There are normal curves for every combination of μ and σ .

- The mean (μ) shifts the curve to the left or right.
- The standard deviation (σ) alters the spread of the curve.
- The first pair of curves have different means but the same standard deviation.
- The second pair of curves share the same mean (μ) but have different standard deviations.
- The pink curve has a smaller standard deviation. It is narrower and taller, and the probability is spread over a smaller range of values.
- The blue curve has a larger standard deviation. The curve is flatter and the tails are thicker. The probability is spread over a larger range of values.

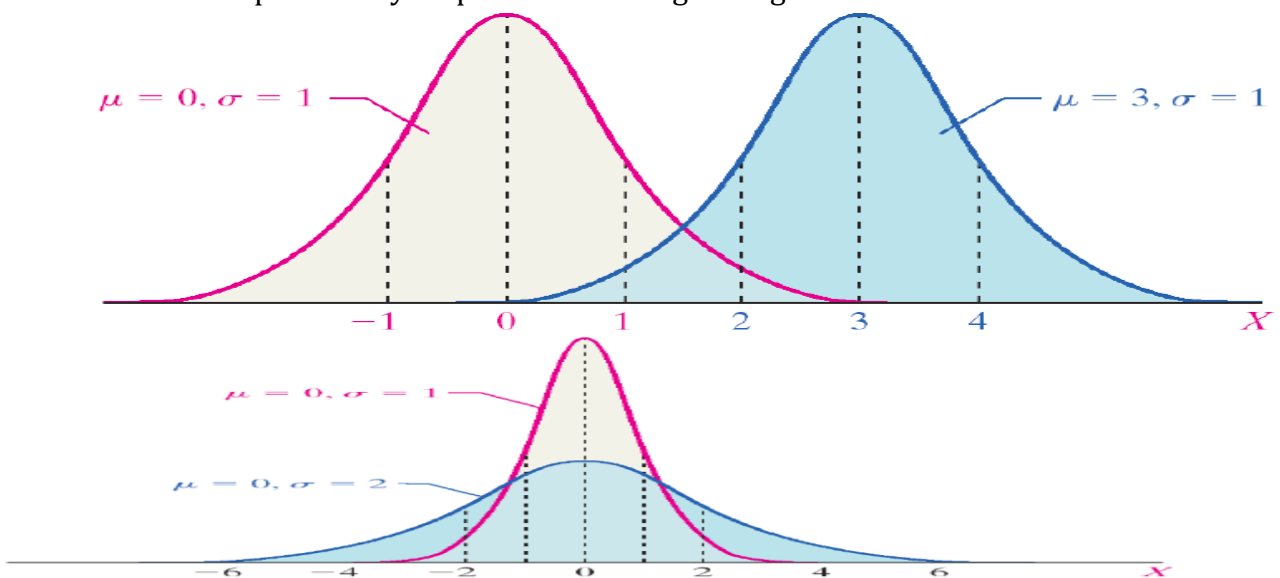


Figure 10. A comparison of normal curves.

Properties of the normal curve:

- The mean is the center of this distribution and the highest point.
- The curve is symmetric about the mean. (The area to the left of the mean equals the area to the right of the mean.)
- The total area under the curve is equal to one.
- As x increases and decreases, the curve goes to zero but never touches.

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

- The PDF of a normal curve is $y = \frac{1}{\sqrt{2\pi} \sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$.
- A normal curve can be used to estimate probabilities.
- A normal curve can be used to estimate proportions of a population that have certain x -values.

The Standard Normal Distribution

The standard normal distribution, also called the z -distribution, is a **special normal distribution where the mean is 0 and the standard deviation is 1**. Any normal distribution can be standardized by converting its values into z -scores. There are millions of possible combinations of means and standard deviations for continuous random variables.

Finding probabilities associated with these variables would require us to integrate the PDF over the range of values we are interested in.

To avoid this, we can rely on the standard normal distribution. T

he standard normal distribution is a special normal distribution with a $\mu = 0$ and $\sigma = 1$. We can use the Z -score to standardize any normal random variable, converting the x -values to Z -scores, thus allowing us to use probabilities from the standard normal table. So how do we find area under the curve associated with a Z -score?

Standard Normal Table

- The standard normal table gives probabilities associated with specific Z -scores.
- The table we use is cumulative from the left.
- The negative side is for all Z -scores less than zero (all values less than the mean).
- The positive side is for all Z -scores greater than zero (all values greater than the mean).
- Not all standard normal tables work the same way.

Example 10

What is the area associated with the Z-score 1.62?

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633

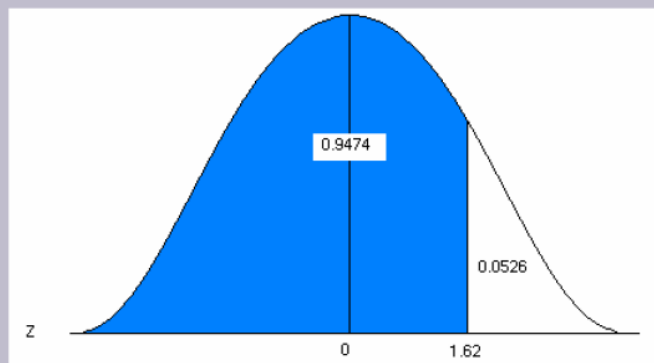


Figure 11. The standard normal table and associated area for $z = 1.62$.

Reading the Standard Normal Table

- Read down the Z-column to get the first part of the Z-score (1.6).
- Read across the top row to get the second decimal place in the Z-score (0.02).
- The intersection of this row and column gives the area under the curve to the left of the Z-score.

Finding Z-scores for a Given Area

- What if we have an area and we want to find the Z-score associated with that area?
- Instead of Z-score \rightarrow area, we want area \rightarrow Z-score.
- We can use the standard normal table to find the area in the body of values and read backwards to find the associated Z-score.
- Using the table, search the probabilities to find an area that is closest to the probability you are interested in.

Example 11

To find a Z-score for which the area to the right is 5%:

Since the table is cumulative from the left, you must use the complement of 5%.

$$1.000 - 0.05 = 0.9500$$

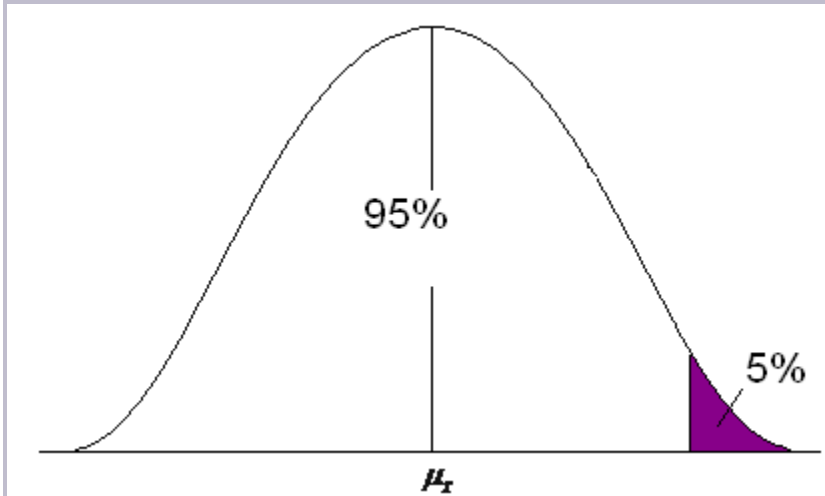


Figure 12. The upper 5% of the area under a normal curve.

- Find the Z-score for the area of 0.9500.
- Look at the probabilities and find a value as close to 0.9500 as possible.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633

Figure

13. The standard normal table.

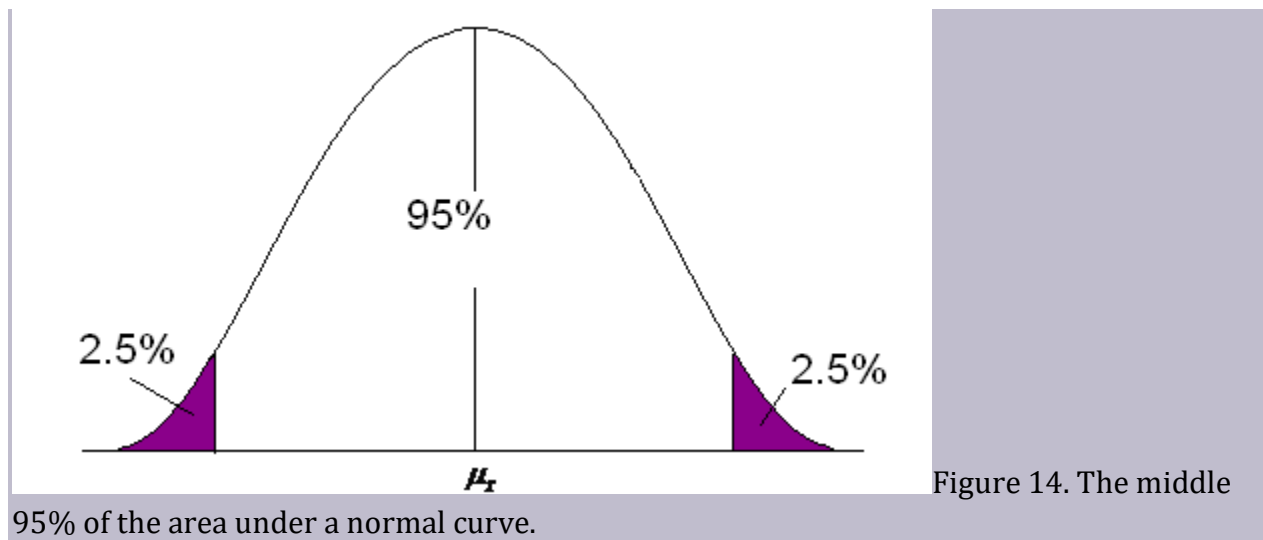
The Z-score for the 95th percentile is 1.64.

Area in between Two Z-scores

Example 12

To find Z-scores that limit the middle 95%:

- The middle 95% has 2.5% on the right and 2.5% on the left.
- Use the symmetry of the curve.



- Look at your standard normal table. Since the table is cumulative from the left, it is easier to find the area to the left first.
- Find the area of 0.025 on the negative side of the table.
- The Z-score for the area to the left is -1.96.
- Since the curve is symmetric, the Z-score for the area to the right is 1.96.

Common Z-scores

There are many commonly used Z-scores:

- $Z_{.05} = 1.645$ and the area between -1.645 and 1.645 is 90%
- $Z_{.025} = 1.96$ and the area between -1.96 and 1.96 is 95%
- $Z_{.005} = 2.575$ and the area between -2.575 and 2.575 is 99%

Applications of the Normal Distribution

Typically, our normally distributed data do not have $\mu = 0$ and $\sigma = 1$, but we can relate any normal distribution to the standard normal distributions using the Z-score. We can transform values of x to values of z .

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

For example, if a normally distributed random variable has a $\mu = 6$ and $\sigma = 2$, then a value of $x = 7$ corresponds to a Z-score of 0.5.

$$Z = \frac{7-6}{2} = 0.5$$

This tells you that 7 is one-half a standard deviation above its mean. We can use this relationship to find probabilities for any normal random variable.

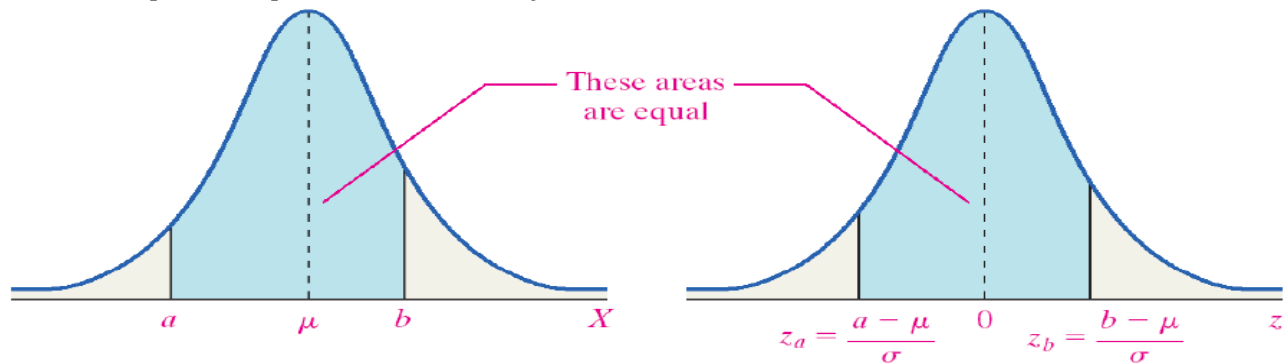


Figure 15. A normal and standard normal curve.

To find the area for values of X , a normal random variable, draw a picture of the area of interest, convert the x -values to Z -scores using the Z -score and then use the standard normal table to find areas to the left, to the right, or in between.

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

Example 13

Adult deer population weights are normally distributed with $\mu = 110$ lb. and $\sigma = 29.7$ lb. As a biologist you determine that a weight less than 82 lb. is unhealthy and you want to know what proportion of your population is unhealthy.

$$P(x < 82)$$

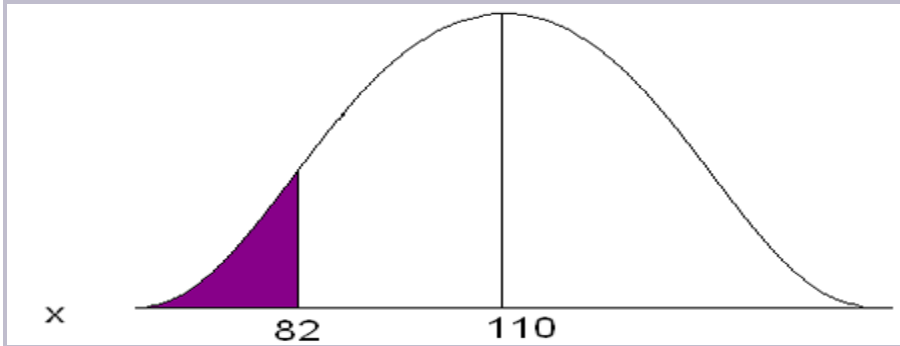


Figure 16. The area

under a normal curve for $P(x < 82)$.

$$z = \frac{82 - 110}{29.7} = -0.94$$

Convert 82 to a Z-score

The x value of 82 is 0.94 standard deviations below the mean.

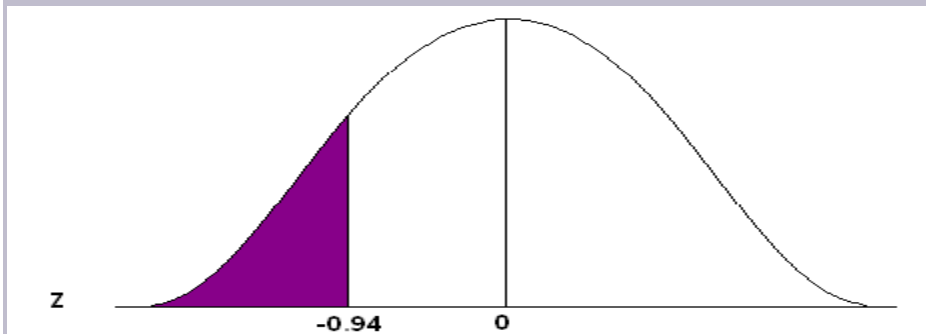


Figure 17. Area under

a standard normal curve for $P(z < -0.94)$.

Go to the standard normal table (negative side) and find the area associated with a Z-score of -0.94.

This is an “area to the left” problem so you can read directly from the table to get the probability.

$$P(x < 82) = 0.1736$$

Approximately 17.36% of the population of adult deer is underweight, OR one deer chosen at random will have a 17.36% chance of weighing less than 82 lb.

Example 14

Statistics from the Midwest Regional Climate Center indicate that Jones City, which has a large wildlife refuge, gets an average of 36.7 in. of rain each year with a standard deviation of 5.1 in. The amount of rain is normally distributed. During what percent of the years does Jones City get more than 40 in. of rain?

$$P(x > 40)$$

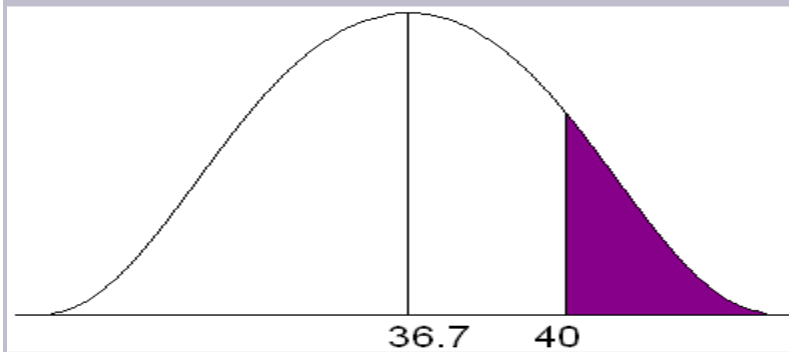


Figure 18. Area under a normal

curve for $P(x > 40)$.

$$z = \frac{40 - 36.7}{5.1} = 0.65$$

$$P(x > 40) = (1 - 0.7422) = 0.2578$$

For approximately 25.78% of the years, Jones City will get more than 40 in. of rain.

Assessing Normality

- If the distribution is unknown and the sample size is not greater than 30 (Central Limit Theorem), we have to assess the assumption of normality.
- Our primary method is the normal probability plot. This plot graphs the observed data, ranked in ascending order, against the “expected” Z-score of that rank.
- If the sample data were taken from a normally distributed random variable, then the plot would be approximately linear.
- Examine the following probability plot.
- The center line is the relationship we would expect to see if the data were drawn from a perfectly normal distribution.
- Notice how the observed data (red dots) loosely follow this linear relationship. Minitab also computes an Anderson-Darling test to assess normality.

- The null hypothesis for this test is that the sample data have been drawn from a normally distributed population. A p-value greater than 0.05 supports the assumption of normality.

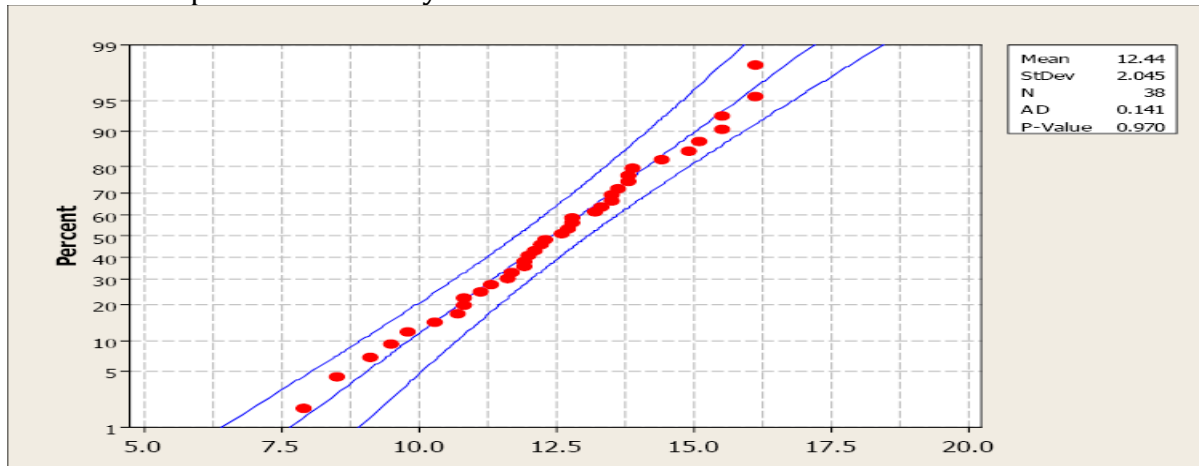


Figure 19. A normal probability plot generated using Minitab 16.

Compare the histogram and the normal probability plot in this next example. The histogram indicates a skewed right distribution.

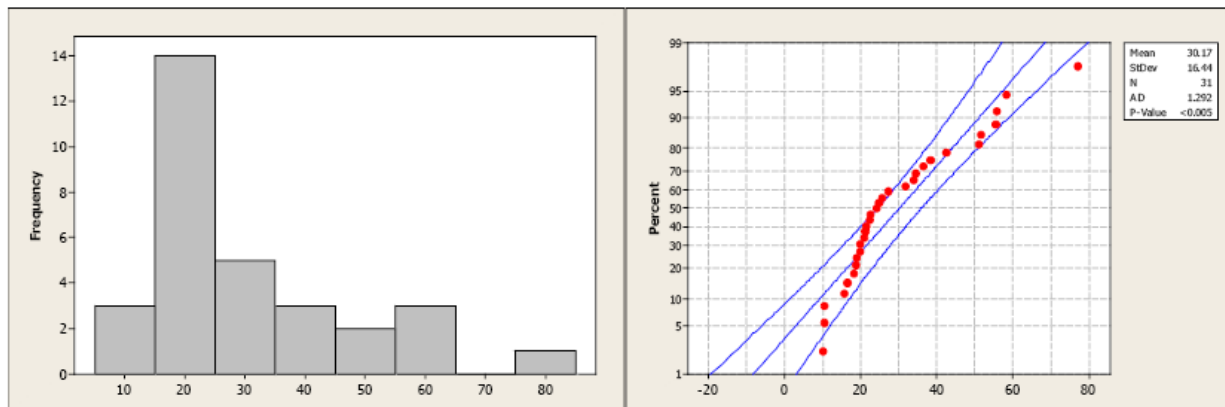


Figure 20. Histogram and normal probability plot for skewed right data.

The observed data do not follow a linear pattern and the p-value for the A-D test is less than 0.005 indicating a non-normal population distribution.

Normality cannot be assumed. You must always verify this assumption. Remember, the probabilities we are finding come from the standard NORMAL table. If our data are NOT normally distributed, then these probabilities DO NOT APPLY.

IMPORTANT QUESTIONS

1. What are the types of data.
2. List out the types of variables.
3. How to describing data with table and graph.
4. Define variability
5. What is meant by normal distribution.
6. Define standard (z) scores.
7. What is standard errors.
8. Difference between bar chart and histogram.
9. Write the properties of the normal curve
10. To find a z-scores for a given area which the area to the right at 5% and 95% of area under normal.
11. Application of the normal distribution.
12. Explain mean ,median and mode with proper example.
13. Shorts notes on range and variance
14. How to find the standard deviation of a given data 3,5,7 with a sample mean is 5.
15. Write short notes on Normal distribution.