STATISTIC

Statistics is the science of data and the specialized aspect of organizing and analyzing data for the purpose of decision-making based on analysis.

THE MAIN ASPECT IN STATISTICS:

- Data

They are facts collected in a statistical case.

- Variable

Any phenomenon in which there are differences between its vocabulary (values) is denoted by x or y or z a symbol.

- Population:

Set of all possible observations

- Sub set of population or set of observation or data chosen in some way in the community (part of community) so that they can be used for the purpose of studying the community as a whole.
- Statistical symbols:

$$x_i^n = X_1, X_2, X_3 \ldots\ldots\ldots\ldots X_n$$

Where i is sequence and n is the total number of data

1. $\sum_{i=1}^{n} x_i = x_1 + x_2 + x_3 + \ldots\ldots\ldots\ldots\ldots x_n = \sum x_i$

2. $\sum_{i=1}^{n} (x_i)^{**}2 = = (X_1)^2 , (X_2)^2 , (X_3)^2 \ldots\ldots\ldots\ldots (X_n)^2$

3. $(\sum x_i)^{**}2 = (x_1 + x_2 + x_3 + \ldots\ldots\ldots\ldots x_n)^2$

4. $\sum_{i=1}^{n} x_i\, y_i = x_1 y_1 + x_2 y_2 + x_3 y_3 \ldots\ldots\ldots x_n y_n$

5. $\sum x_i \,.\sum y_i = \{x_1 + x_2 + \ldots x_n\}.\{y_1 + y_2 + \cdots y_n\}$

6. $\sum_{i=1}^{n} (x_i \pm \mathrm{k}) = \sum x_i \mp nk$

$$\sum_{i=1}^{n} k = nk \qquad\qquad k = constant$$

اسلوب جمع البيانات

1. collection method Data and we will not deal with it now.
2. The style of samples اسلوب العينات

وهو الاسلوب المتبع في دراسة علم الاحصاء ويكون على نوعين:

- عينات معينه
- عينات عشوائيه Random samples

وفي هذا الكورس سنركز على العينات العشوائيه

The random samples is

A: Systematle random sampling

حيث يقسم المجتمع الى مجاميع وكل مجموعه:

K= N/n where,

K = عدد المجاميع

N= عدد مفردات المجتمع

n= حجم العينه

Example:

We have 24 students arrange them According to their score in descending order The required sample is 6 students.

Solution:

K= N/n

K= 24/6 =4

SO , FOUR GROUP ARE CREATED AS FOLLOWS:

G1: 1 2 3 4 5 6

G2: 7 8 9 10 11 12

G3: 13 14 15 16 17 18

G4: 19 20 21 22 23 24

THEN IF WE START WITH 3 THEN:

3 , 7 , 11, 15 , 19 , 23

Stratified Random Sampling: Definition

Stratified random sampling is a type of probability sampling using which a research organization can branch off the entire population into multiple non-overlapping, homogeneous groups (strata) and randomly choose final members from the various strata for research which reduces cost and improves efficiency. Members in each of these groups should be distinct so that every member of all groups get equal opportunity to be selected using simple probability. This sampling method is also called “random quota sampling”.

In this approach, each stratum sample size is directly proportional to the population size of the entire population of strata. That means each strata sample has the same sampling fraction.

Proportionate Stratified Random Sampling Formula: $n_h = (N_h / N) * n$

$\mathbf{n_h}$= Sample size for h^{th} stratum

$\mathbf{N_h}$= Population size for h^{th} stratum

N = Size of entire population

n = Size of entire sample

If you have 4 strata with 500, 1000, 1500, 2000 respective sizes and the research organization selects ½ as sampling fraction. A researcher has to then select 250, 500, 750, 1000 members from the respective stratum.

Stratum	A	B	C	D
Population Size	500	1000	1500	2000
Sampling Fraction	1/2	1/2	1/2	1/2
Final Sampling Size Results	250	500	750	1000

Irrespective of the sample size of the population, the sampling fraction will remain uniform across all the strata.

EXAMPLE 2:

Stratum	A	B	C	D
Population Size	500	1000	1500	2000
Sampling Fraction	1/2	1/3	1/4	1/5
Final Sampling Size Results	250	333	375	400

Frequency histogram

A histogram is a graphical representation of a frequency distribution, in which vertical rectangular blocks are drawn so that:

(a) the centre of the base indicates the central value of the class and

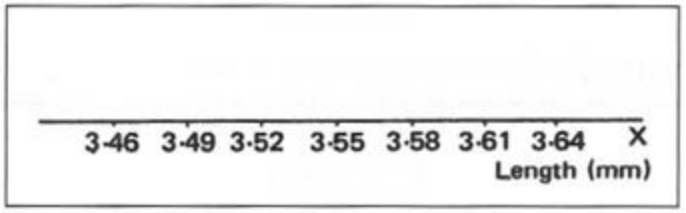
(b) the area of the rectangle represents the class frequency.

If the class intervals are regular, the frequency is then denoted by the height of the rectangle.

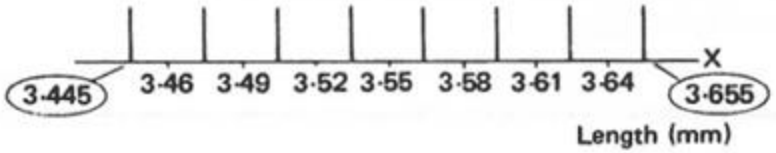
For example, measurement of the lengths of 50 brass rods gave the following frequency distribution:

Length (mm) (x)	Lower class boundary	Upper class boundary	Central value	Frequency (f)
3·45–3·47	3·445	3·475	3·460	2
3·48–3·50	3·475	3·505	3·490	6
3·51–3·53	3·505	3·535	3·520	12
3·54–3·56	3·535	3·565	3·550	14
3·57–3·59	3·565	3·595	3·580	10
3·60–3·62	3·595	3·625	3·610	5
3·63–3·65	3·625	3·655	3·640	1

First, we draw a base line and on it mark a scale of x on which we can indicate the central values of the classes. Do that for a start.

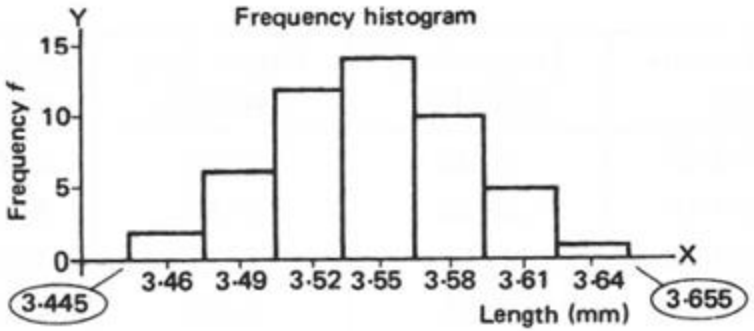


Since the classes are of regular class interval, the class boundaries will coincide with the points mid-way between the central values, thus:



Note that the lower boundary of the first class extends to 3·445 and that the upper boundary of the seventh class extends to 3·655. Because the class intervals are regular, we can now erect a vertical scale to represent class frequencies and rectangles can be drawn to the appropriate height. Complete the work and give it a title.

Relative frequency histogram



The same diagram can be made to represent the relative frequency distribution by replacing the scale of frequency along the *y*-axis with a scale of relative frequencies (percentages).

Considerable information can be gleaned from a histogram on sight. For instance, we see that the class having the highest frequency is the ..... class.

fourth

and the class with the lowest frequency is the ..... class.

---

seventh

Most of the whole range of values is clustered within the middle classes and knowledge of the centre region of the histogram is important. We can put a numerical value on this by determining a *measure of central tendency* which we shall now deal with.

There are three common measures of central tendency, the *mean*, *mode* and *median*, of a set of observations and we shall discuss each of them in turn.

Measure of central tendency:

Mean

The arithmetic mean $\bar{x}$ of a set of n observations x is simply their average,

$$\text{i.e. mean} = \frac{\text{sum of the observations}}{\text{number of observations}} \quad \therefore \ \bar{x} = \frac{\sum x}{n}$$

When calculating the mean from a frequency distribution, this becomes

$$\text{mean} = \bar{x} = \frac{\sum xf}{n} = \frac{\sum xf}{\sum f}$$

For example, for the following frequency distribution, we need to add a third column showing the values of the product $x \times f$, after which the mean can be found:

Variable (x)	Frequency (f)	Product (xf)
15	1	15
16	4	64
17	9	153
18	10	180
19	6	114
20	2	40

$$n = \sum f = 32 \text{ and } \sum xf = 566 \qquad \therefore\ \bar{x} = \frac{\sum xf}{n} = \frac{566}{32} = 17{\cdot}69$$

When calculating the mean from a frequency distribution with grouped data, the central value, x_m, of the class is taken as the x-value in forming the product xf. So, for the frequency distribution

Variable (x)	12–14	15–17	18–20	21–23	24–26	27–29
Frequency (f)	2	6	9	8	4	1

$$\bar{x} = 19{\cdot}9$$

Because 13x2+16x6+19x9+22x8+25x4+28=597

$$n = \sum f = 30 \text{ and } \sum x_m f = 597 \qquad \therefore\ \bar{x} = \frac{\sum x_m f}{n} = \frac{597}{30} = 19{\cdot}9$$

Here is one more.

Measurement in millimetres of 60 bolts gave the following frequency distribution:

Length x (mm)	30·2	30·4	30·6	30·8	31·0	31·2	31·4
Frequency f	3	7	12	17	11	8	2

The mean $\bar{x} =$ .....

$$\boxed{\bar{x} = 30{\cdot}79}$$

Length (mm) (x)	Frequency (f)	Product (xf)
30·2	3	90·6
30·4	7	212·8
30·6	12	367·2
30·8	17	523·6
31·0	11	341·0
31·2	8	249·6
31·4	2	62·8

$n = \sum f = 60$ $\quad\downarrow$

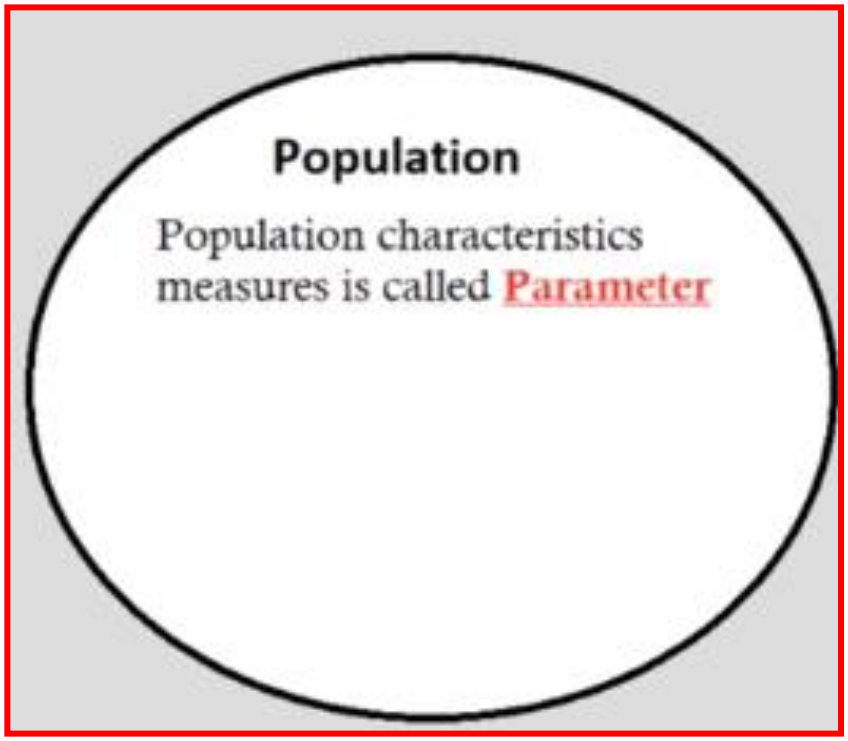
$\sum xf = 1847{\cdot}6$

$$\therefore\ \bar{x} = \frac{\sum xf}{n} = \frac{1847{\cdot}6}{60}$$

$$= 30{\cdot}79$$

$$\therefore\ \bar{x} = 30{\cdot}79$$

Measures of variation (dispersions):



- Now; our characteristic is "**Variability**".
- Measures of Variability:

1. Range. **2. Variance.**

3. Standard Deviation. **4.Coefficient of Variation.**

1. Range

Range

Range = Largest observation − Smallest observation

- The advantage of the range is its simplicity.
- The disadvantage is also its simplicity.

→Because the range is calculated from only two observations, it tells us nothing about the other observations.

→Consider the following two sets of data:

set #1]: 4 4 4 4 4 50

set#2]: 4 8 15 24 39 50

Range = 50 − 4 = 46

Range = 50 − 4 = 46

→ The range of both sets is 46. The two sets of data are completely different, yet their ranges are the same.

→To measure variability, we need other statistics that incorporate all the data and not just two observations.

2. Variance

- The variance and its related measure, the standard deviation, are arguably the most important statistics.

→They are used to measure variability

→They play a vital role in almost all statistical inference procedures.

Variance

$$\text{Population variance:} \quad \sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

$$\text{Sample variance:}^* \quad s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n - 1}$$

The population variance is represented by σ^2(Greek letter *sigma* squared).

Shortcut for Sample Variance

$$s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right]$$

Interpreting the Variance

- Variance provides us with only a **rough idea** about the amount of variation in the data.
- However, this statistic is useful when comparing two or more sets of data of the same type of variable.

→If the variance of one data set is larger than that of a second data set, we interpret that to mean that the observations in the first set display more variation than the observations in the second set.

- The **problem of interpretation** is caused by the way the variance is computed.

→ Because we squared the deviations from the mean, the **unit attached to the variance is the square of the unit attached to the original observations**.

3. Standard Deviation

Standard Deviation

Population standard deviation: $\sigma = \sqrt{\sigma^2}$

Sample standard deviation: $s = \sqrt{s^2}$

- The **unit associated with the standard deviation is the unit of the original data set**.

→standard deviation solved the problem of interpretation of variance.

4. Coefficient of Variance (CV)

- Why Coefficient of variance exist?
- Is a standard deviation of 10 a large number indicating great variability or a small number indicating little variability?

→The answer depends somewhat on the magnitude of the observations in the data set.

→ If the observations are in the millions, then a standard deviation of 10 will probably be considered a small number.

→ On the other hand, if the observations are less than 50, then the standard deviation of 10 would be seen as a large number.

→ This logic lies behind yet another measure of variability, the coefficient of variation.

Population coefficient of variation: $CV = \frac{\sigma}{\mu}$

Sample coefficient of variation: $cv = \frac{s}{\bar{x}}$

EXAMPLE 1:

calculate the variance of the following data:

9	3	7	4	1	7	5	4

Solution:

x_i	9 + 3 + 7 + 4 + 1 + 7 + 5 + 4	$\bar{x} = \frac{40}{8} = 5$
x_i^2	81 + 9 + 49 + 16 + 1 + 49 + 25 + 16	$\Sigma x_i^2 = 246$
$(x_i - \bar{x})^2$	16 + 4 + 4 + 1 + 16 + 4 + 0 + 1	$\Sigma (x_i - \bar{x})^2 = 46$

$\Sigma x_i = 40$ $\quad n = 8$

method 1:

$$S^2 = \frac{\Sigma (x_i - \bar{x})^2}{n-1} = \frac{46}{7} = 6.57$$

method 2:

$$S^2 = \frac{\Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n}}{n-1} = \frac{246 - \frac{1600}{8}}{7} = \frac{246 - 200}{7} = 6.57$$

Example 2 -Home work:

Calculate the variance of the following data: follow the same way of example 1

4 5 3 6 5 6 5 6

==========

Example 3: Determine the variance and standard deviation of the following sample.

12 6 22 31 23 13 15 17 21

- Solution

$\Sigma x_i^2 = 3278$ $\Sigma x_i = 160$ $n = 9$

$$s^2 = \frac{3278 - \frac{25600}{9}}{8} = \boxed{54.194}$$

$$s = \sqrt{54.194} = \boxed{7.362}$$

Example 4 : Find the variance and standard deviation of the following sample.

0 -5 -3 6 4 -4 1 -5 0 3

HOME WORK

==========

Example 5: Create a sample of five numbers whose mean is 6 and whose standard deviation is 0.

6, 6 ,6 ,6 ,6

==========

Standard deviation for grouped data:

If we consider groups of student as per the following table Whom degree declare as per group . The standard deviation could be calculated as follows:

Grade	f	m	f·m	$\bar{x}$	$m-\bar{x}$	$(m-\bar{x})^2$	$f(m-\bar{x})^2$
50-59	3	54.5	163.5	79.2	-24.7	610.09	1830.27
60-69	5	64.5	322.5	79.2	-14.7	216.09	1080.45
70-79	9	74.5	670.5	79.2	-4.7	22.09	198.81
80-89	12	84.5	1014	79.2	5.3	28.09	337.08
90-100	8	95	760	79.2	15.8	249.64	1997.12
	37		2930.5				5443.73

$$s = \sqrt{\frac{\Sigma f(m-\bar{x})^2}{n-1}} = \sqrt{\frac{5443.73}{37-1}} = \sqrt{151.2147\ldots} \approx 12.3$$

Standard Error:

1. 5 students in a college were selected at random and their ages were found to be 18, 21, 19, 20, and 26. (a) Calculate the standard deviation of the ages in the sample. (b) Calculate the standard error.

$$\bar{x} = \frac{18+21+19+20+26}{5} = \frac{104}{5} = 20.8$$

$$s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$$

$$s = \sqrt{\frac{(18-20.8)^2+(21-20.8)^2+(19-20.8)^2+(20-20.8)^2+(26-20.8)^2}{5-1}}$$

$$s = \sqrt{\frac{(-2.8)^2+(0.2)^2+(-1.8)^2+(-0.8)^2+(5.2)^2}{4}}$$

$$s = 3.114$$

$$SE = \frac{s}{\sqrt{n}} = \frac{3.114}{\sqrt{5}}$$

$$SE = 1.393$$

Frequency Distribution Table – Data Collection

In our day to day life, recording information is very crucial. A piece of information or representation of facts or ideas which can be further processed is known as data. The weather forecast, maintenance of records, dates, time, and everything is related to data collection.

The collection, presentation, analysis, organization and interpretation of observations or data is known as statistics. We can make predictions about the nature of data based on the previous data using statistics. Statistics are helpful when a large amount of data is to be studied and observed.

The collected statistical data can be represented by various methods such as tables, bar graphs, pie charts, histograms, frequency polygons, etc.

What is Frequency Distribution Table in Statistics?

In statistics, a frequency distribution table is a comprehensive way of representing the organisation of raw data of a quantitative variable. This table shows how various values of a variable are distributed and their corresponding frequencies. However, we can make two frequency distribution tables:

(i) Discrete frequency distribution

(ii) Continuous frequency distribution (Grouped frequency distribution)

How to Make a Frequency distribution table?

Frequency distribution tables can be made using tally marks for both discrete and continuous data values. The way of preparing discrete frequency tables and continuous frequency distribution tables are different from each other.

In this section, you will learn how to make a discrete frequency distribution table with the help of examples.

An example: In a quiz, the marks obtained by 20 students out of 30 are given as:

12,15,15,29,30,21,30,30,15,17,19,15,20,20,16,21,23,24,23,21
This data can be represented in tabular form as follows:

Table 1: Frequency Distribution Table (Ungrouped)

Marks obtained in quiz	Number of students(Frequency)
12	1
15	4
16	1
17	1
19	1
20	2
21	3
23	2
24	1
29	1
30	3
Total	20

The number of times data occurs in a data set is known as the frequency of data. In the above example, frequency is the number of students who scored various marks as tabulated. This type of tabular data collection is known as an ungrouped frequency table.

What happens if, instead of 20 students, 200 students took the same test. Would it have been easy to represent such data in the format of an ungrouped frequency distribution table? Well, obviously no. To represent a vast amount of information, the data is subdivided into groups of similar sizes known as class or class intervals, and the size of each class is known as class width or class size.

Frequency Distribution table for Grouped data

The frequency distribution table for grouped data is also known as the continuous frequency distribution table. This is also known as the grouped frequency distribution table. Here, we need to make the frequency distribution table by dividing the data values into a suitable number of classes and with the appropriate class height. Let's understand this with the help of the solved example given below:

Question:

The heights of 50 students, measured to the nearest centimetres, have been found to be as follows:

161, 150, 154, 165, 168, 161, 154, 162, 150, 151, 162, 164, 171, 165, 158, 154, 156, 172, 160, 170, 153, 159, 161, 170, 162, 165, 166, 168, 165, 164, 154, 152, 153, 156, 158, 162, 160, 161, 173, 166, 161, 159, 162, 167, 168, 159, 158, 153, 154, 159

(i) Represent the data given above by a grouped frequency distribution table, taking the class intervals as 160 – 165, 165 – 170, etc.

(ii) What can you conclude about their heights from the table?

Solution:

(i) Let us make the grouped frequency distribution table with classes:

150 – 155, 155 – 160, 160 – 165, 165 – 170, 170 – 175

Class intervals and the corresponding frequencies are tabulated as:

FREQUENCY DISTRIBUTION TABLE STATISTICS

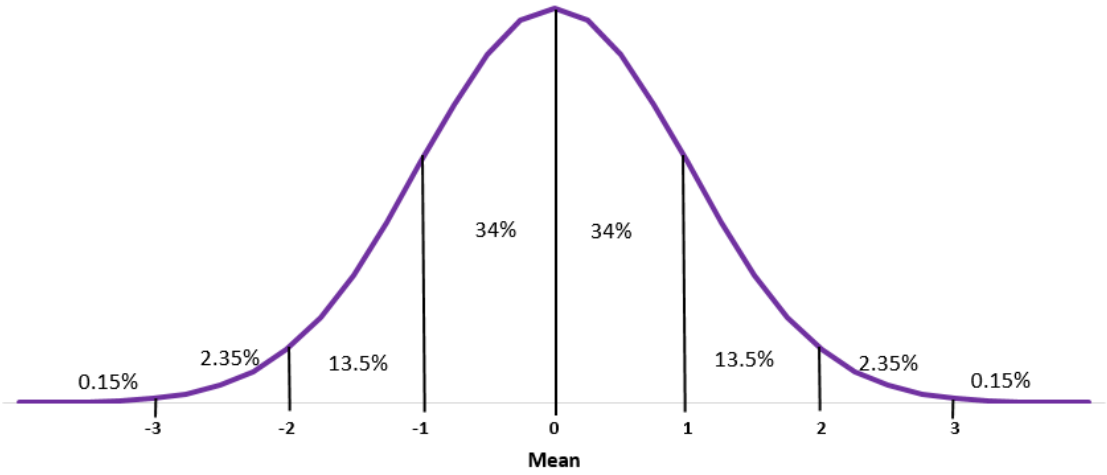
BYJU'S The Learning App

Class intervals	Frequency	Corresponding data values
150 – 155	12	150, 150, 151, 152, 153, 153, 153, 154, 154, 154, 154, 154
155 – 160	9	156, 156, 158, 158, 158, 159, 159, 159, 159
160 – 165	14	160, 160, 161, 161, 161, 161, 161, 162, 162, 162, 162, 162, 164, 164
165 – 170	10	165, 165, 165, 165, 166, 166, 167, 168, 168, 168
170 – 175	5	170, 170, 171, 172, 173
Total	50	

(ii) From the given data and above table, we can observe that 35 students, i.e. more than 50% of the total students, are shorter than 165 cm.

The Normal Distribution

The **normal distribution** is the most common probability distribution in statistics.



Normal distributions have the following features:

- Bell shape
- Symmetrical
- Mean and median are equal; both are located at the center of the distribution
- About 68% of data falls within one standard deviation of the mean
- About 95% of data falls within two standard deviations of the mean
- About 99.7% of data falls within three standard deviations of the mean

The last three bullet points are known as the **Empirical**

Rule, sometimes called the **68-95-99.7 rule**.

The last three bullet points are known as the **Empirical Rule**, sometimes called the **68-95-99.7 rule**.

How to Draw a Normal Curve

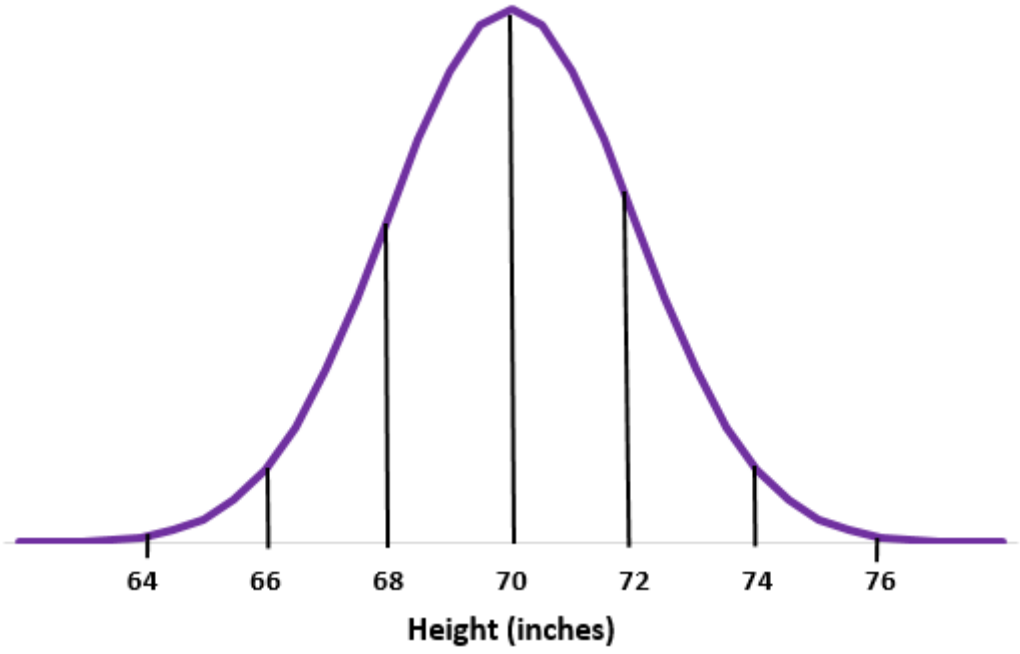
To draw a normal curve, we need to know the mean and the standard deviation.

Example 1: ***Suppose the height of males at a certain school is normally distributed with mean of μ=70 inches and a standard deviation of σ = 2 inches. Sketch the normal curve.***

Step 1: Sketch a normal curve.

Step 2: The mean of 70 inches goes in the middle.

Step 3: Each standard deviation is a distance of 2 inches.



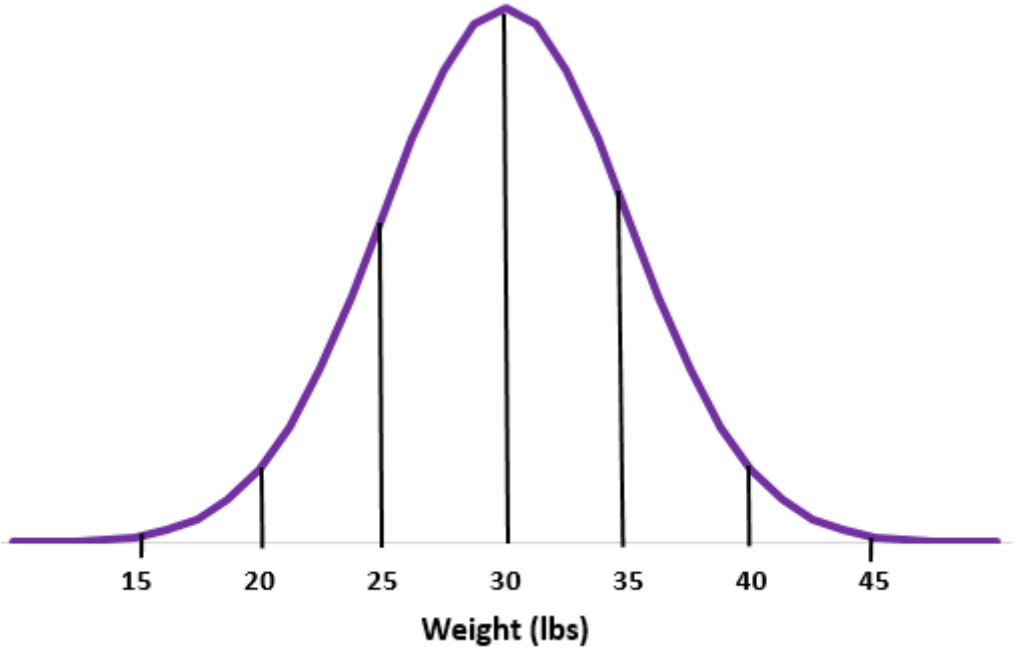
Example 2: ***Suppose the weight of a certain species of otters is normally distributed with mean of μ=30 lbs and a***

standard deviation of σ = 5 lbs. Sketch the normal curve.

Step 1: Sketch a normal curve.

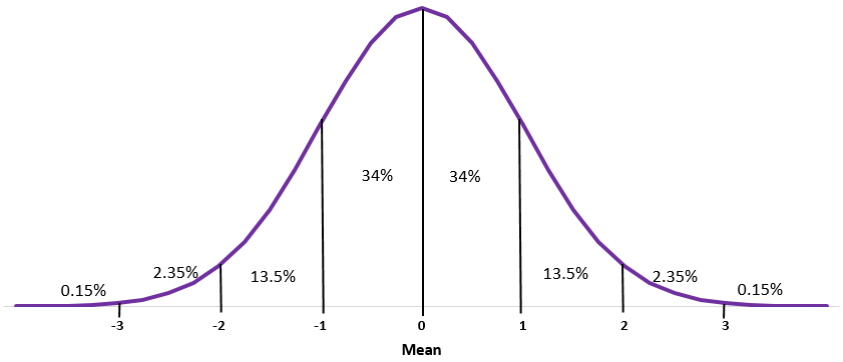
Step 2: The mean of 30 lbs goes in the middle.

Step 3: Each standard deviation is a distance of 5 lbs



How to Find Percentages Using the Normal Distribution

The **empirical rule**, sometimes called the **68-95-99.7 rule**, says that for a random variable that is normally distributed, 68% of data falls within one standard deviation of the mean, 95% falls within two standard deviations of the mean, and 99.7% falls within three standard deviations of the mean.



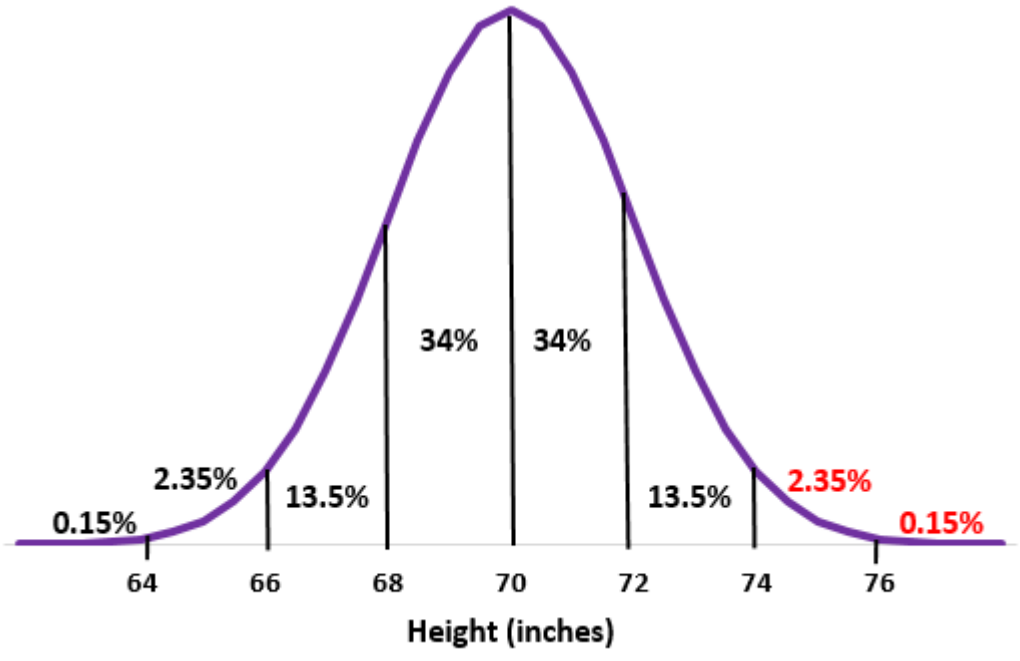
Using this rule, we can answer questions about percentages.

Approximately what percentages of males at this school are taller than 74 inches?

Solution:

Step 1: Sketch a normal distribution with a mean of μ=70 inches and a standard deviation of σ = 2 inches.

Step 2: A height of 74 inches is two standard deviations above the mean. Add the percentages above that point in the normal distribution.



2.35% + 0.15% = **2.5%**

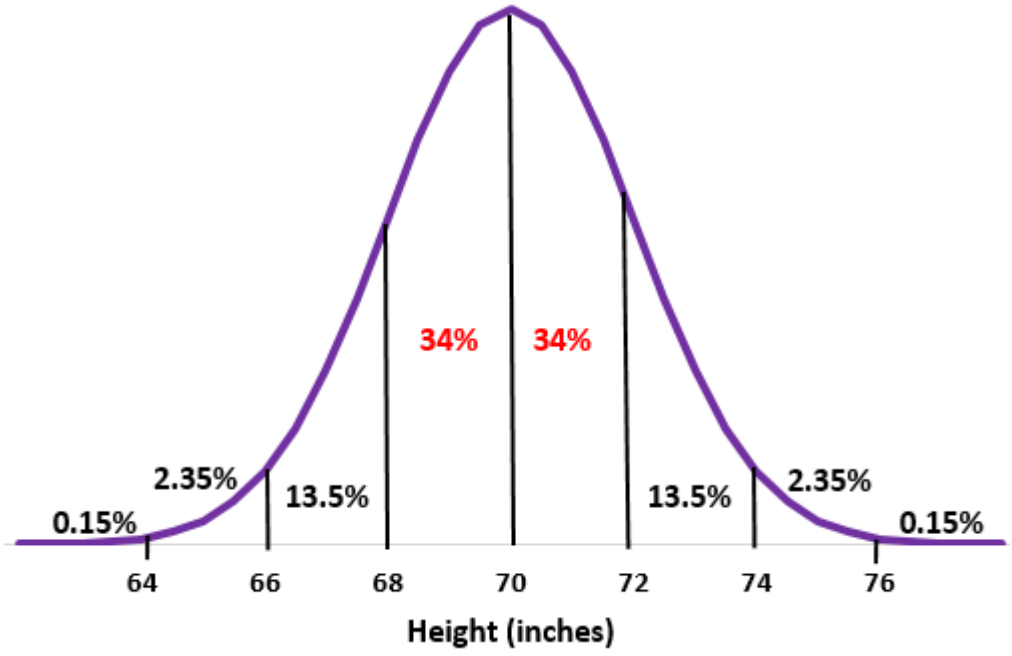
Approximately **2.5%** of males at this school are taller than 74 inches.

Approximately what percentage of males at this school are between 68 inches and 72 inches tall?

Solution:

Step 1: Sketch a normal distribution with a mean of μ=70 inches and a standard deviation of σ = 2 inches.

Step 2: A height of 68 inches and 72 inches is one standard deviation below and above the mean, respectively. Simply add the percentages between these two points in the normal distribution.



34% + 34% = **68%**

Approximately **68%** of males at this school are between 68 inches and 72 inches tall.

Q1: 95% of students at school are between 1.1m and 1.7m tall.

Assuming this data is normally distributed.

calculate the mean and standard deviation?

SOLUTION:

The mean is halfway between 1.1m and 1.7m:

Mean = (1.1m + 1.7m) / 2 = 1.4m

95% is two standard deviation i.e. total 4 standard deviation

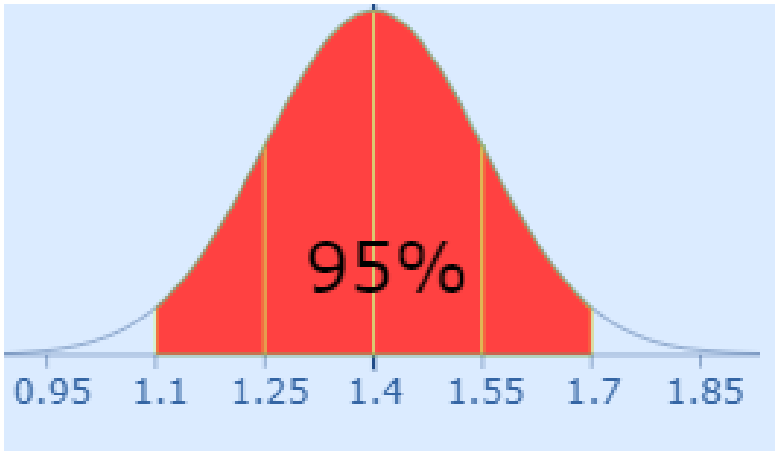
95% is 2 standard deviations either side of the mean (a total of 4 standard deviations) so:

standard deviation = (1.7m-1.1m) / 4
= 0.6m / 4
= 0.15m

OR

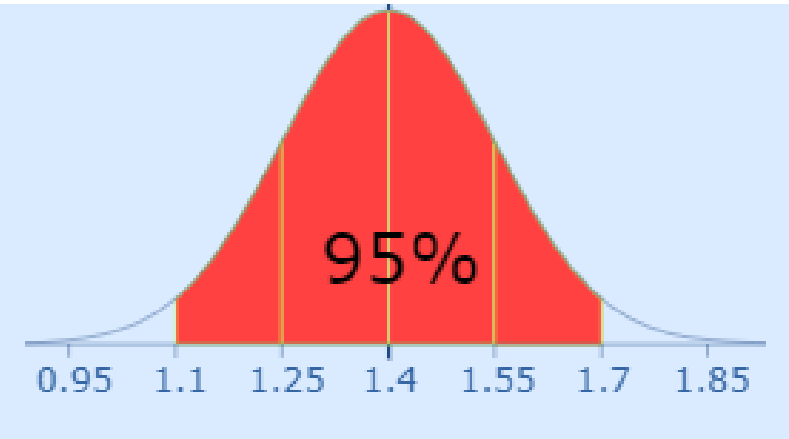
(1.7-1.4)/2

=0.3/2 =0.15



Q2: In the above same school one of student is 1.85m tall

You can see on the bell curve that 1.85m is **3 standard deviations** from the mean of 1.4, so:



It is also possible to **calculate** how many standard deviations 1.85 is from the mean

How far is 1.85 from the mean?

It is 1.85 - 1.4 = **0.45m from the mean**

How many standard deviations is that? The standard deviation is 0.15m, so:

0.45m / 0.15m = **3 standard deviations**

Q3: 95% of students at school weigh between 62 kg and 90 kg.
Assuming this data is normally distributed, what are the mean and standard deviation?

Q4: A machine produces electrical components.
Centered on the mean, 99.7% of the components have lengths between 1.176 cm and 1.224 cm. Assuming this data is normally distributed, what are the mean and standard deviation?

Q5: You work for a small company of 1,000 people and want to find out how they are saving for retirement. Use stratified random sampling to obtain your sample.

Age	Total Number of People in Strata
20-29	160
30-39	220
40-49	240
50-59	200
60+	180

The sample size is 50

*Sample size of the strata = size of entire sample / population size * layer size*

Use the stratified sample formula (Sample size of the strata = size of entire sample / population size * layer size) to calculate the proportion of people from each group:

Age	Number of People in Strata	Number of People in Sample
20-29	160	50/1000 * 160 = 8
30-39	220	50/1000 * 220 = 11
40-49	240	50/1000 * 240 = 12
50-59	200	50/1000 * 200 = 10
60+	180	50/1000 * 180 = 9

So the general formula of stratified sample is as follow

N = N1/p * n

Where,

N= sample size of the strata

N1= sample size

P= population size

n= layer size

Note that all of the individual results from the stratum add up to your sample size of 50: 8 + 11 + 12 + 10 + 9 = 50

Q6: A survey was conducted for the students of the ALmustaqbal College about medical services and the required sample 110 for the following departments: computers 600 students, pharmacy 500 students, engineering 800 students, and dental 1000 students. How many students are selected for each department to carry out the survey?

p= w as in your example note = 2900

N1= sample for each department

n = 110

computer=(600/2900) *110 = 23 =N

pharmacy= (500/2900) *110= 19=N

engineering = (800/2900)*110 = 30 =N

dental= (1000/2900)*110 = 38 = N

23+19+30+38=110 which is equal to the required sample