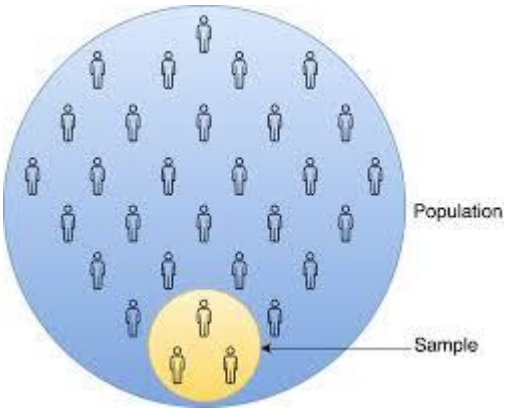
2.1 Discuss statistical concepts to include discrete and continuous variables, sampling, variance, skewness

Sampling

In statistics, quality assurance, and survey methodology, sampling is the selection of a subset (a statistical sample) of individuals from within a statistical population to estimate characteristics of the whole population.



Primary data

This is data collected by an individual (eg yourself, a company, a business, etc)

An advantage of using primary data is that researchers are collecting information for the specific purposes of their study. In essence, the questions the researchers ask are tailored to elicit the data that will help them with their study. Researchers collect the data themselves, using surveys, interviews and direct observations.

Secondary data

This is data collected by a third party and used by an individual. (eg CSO website, published data in magazines, online, etc)

There are several types of secondary data. They can include information from the national population census and other government information collected by Central Statistics Office (www,cso.ie). There are any number of examples: motor vehicle registrations, hospital intake and discharge records, workers' compensation claims records, and more.

Mean (arithmetic mean or average)

Example: Add all numbers and divide by the number of numbers

Mode (most common value or the value with the biggest frequency)

Example: Most common age of students in the college.

Median (this is the "middle" value in the list of numbers)

To find the median, your numbers have to be listed in numerical order from smallest to largest, so you may have to rewrite your list before you can find the median.

Example: 1, 4, 5, 7, 9 The median is 5 as it is the number in the middle.

Example: 2, 4, 6, 6, 9, 16 When there are two numbers in the middle.

The media is calcuated as follows: $\frac{6+6}{2} = 6$

Range (highest x value – lowest x value)

Example: If the oldest student in the college is 58 and the youngest is 18

Then the range is 58 – 18 = 40

Range = 40

Standard Deviation (this is a quantity expressing by how much the members of a group differ from the mean value for the group)

Example: Explain based on the age of students in the college for example.

Formula

Mean

From list: $\mu = \frac{\Sigma x}{n}$

From frequency table: $\mu = \frac{\Sigma fx}{\Sigma f}$

Standard deviation

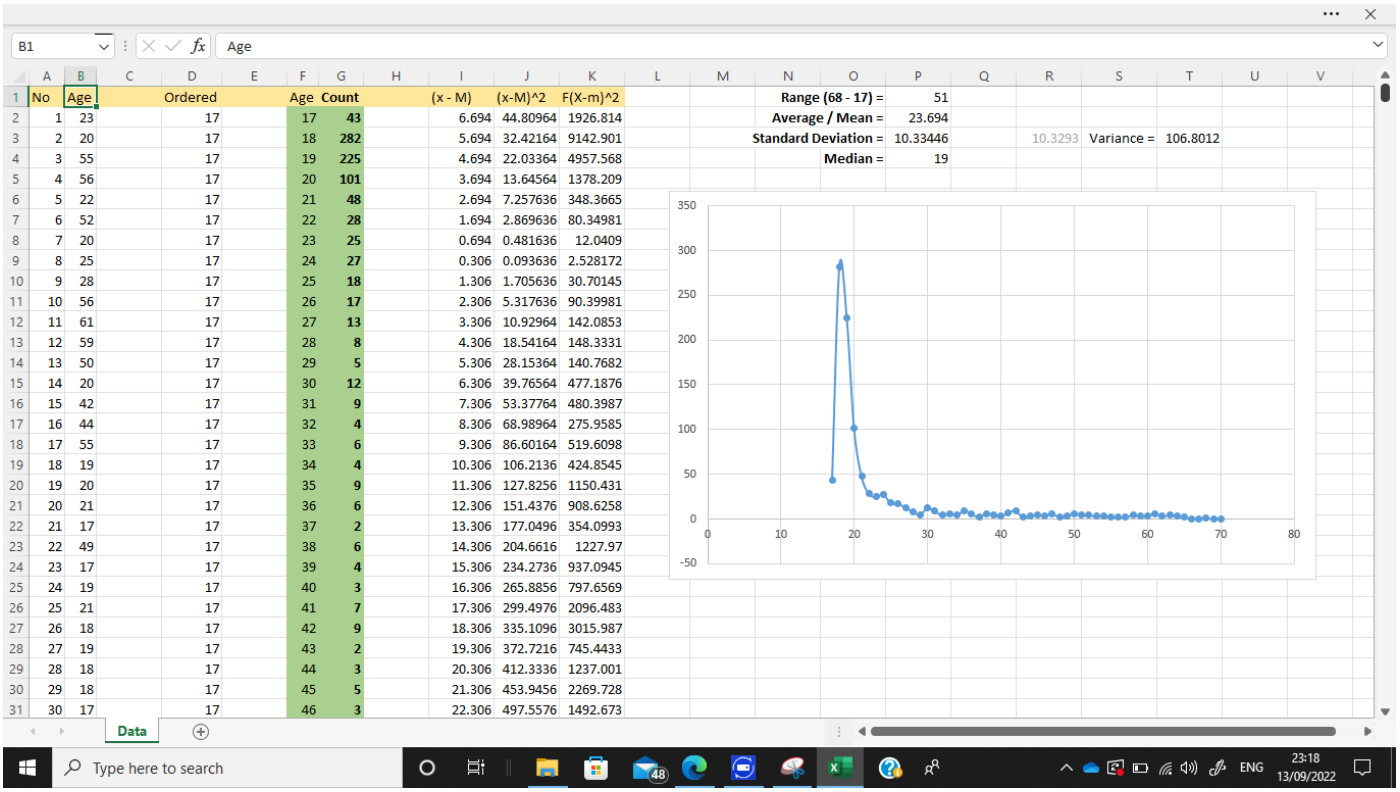
From list: $\sigma = \sqrt{\frac{\Sigma (x-\mu)^2}{n}}$

From frequency table: $\sigma = \sqrt{\frac{\Sigma f (x-\mu)^2}{\Sigma f}}$

Variance

The sample variance, (standard deviation) ^{2} , is used to calculate how varied a sample is. A sample is a select number of items taken from a population. For example, if you are measuring American people's weights, it wouldn't be feasible (from either a time or a monetary standpoint) for you to measure the weights of every person in the population. The solution is to take a sample of the population, say 1000 people, and use that sample size to estimate the actual weights of the whole population. The variance helps you to figure out how spread out your weights are.

Example: Excel file with ages of sample 1000 students in college



No	Age	Ordered	Age	Count	(x - M)	(x-M)^2	F(X-m)^2
1	23	17	17	43	6.694	44.80964	1926.814
2	20	17	18	282	5.694	32.42164	9142.901
3	55	17	19	225	4.694	22.03364	4957.568
4	56	17	20	101	3.694	13.64564	1378.209
5	22	17	21	48	2.694	7.257636	348.3665
6	52	17	22	28	1.694	2.869636	80.34981
7	20	17	23	25	0.694	0.481636	12.0409
8	25	17	24	27	0.306	0.093636	2.528172
9	28	17	25	18	1.306	1.705636	30.70145
10	56	17	26	17	2.306	5.317636	90.39981
11	61	17	27	13	3.306	10.92964	142.0853
12	59	17	28	8	4.306	18.54164	148.3331
13	50	17	29	5	5.306	28.15364	140.7682
14	20	17	30	12	6.306	39.76564	477.1876
15	42	17	31	9	7.306	53.37764	480.3987
16	44	17	32	4	8.306	68.98964	275.9585
17	55	17	33	6	9.306	86.60164	519.6098
18	19	17	34	4	10.306	106.2136	424.8545
19	20	17	35	9	11.306	127.8256	1150.431
20	21	17	36	6	12.306	151.4376	908.6258
21	17	17	37	2	13.306	177.0496	354.0993
22	49	17	38	6	14.306	204.6616	1227.97
23	17	17	39	4	15.306	234.2736	937.0945
24	19	17	40	3	16.306	265.8856	797.6569
25	21	17	41	7	17.306	299.4976	2096.483
26	18	17	42	9	18.306	335.1096	3015.987
27	19	17	43	2	19.306	372.7216	745.4433
28	18	17	44	3	20.306	412.3336	1237.001
29	18	17	45	5	21.306	453.9456	2269.728
30	17	17	46	3	22.306	497.5576	1492.673

Range (68 - 17) =	51			
Average / Mean =	23.694			
Standard Deviation =	10.33446	10.3293	Variance =	106.8012
Median =	19			

Exercise

Find the mean, median, mode, and range for the following list of values:
13, 18, 13, 14, 13, 16, 14, 21, 13

The mean is the usual average, so I'll add and then divide:
(13 + 18 + 13 + 14 + 13 + 16 + 14 + 21 + 13) ÷ 9 = 15

Note that the mean, in this case, isn't a value from the original list. This is a common result. You should not assume that your mean will be one of your original numbers.

The median is the middle value, so first I'll have to rewrite the list in numerical order:
13, 13, 13, 13, 14, 14, 16, 18, 21

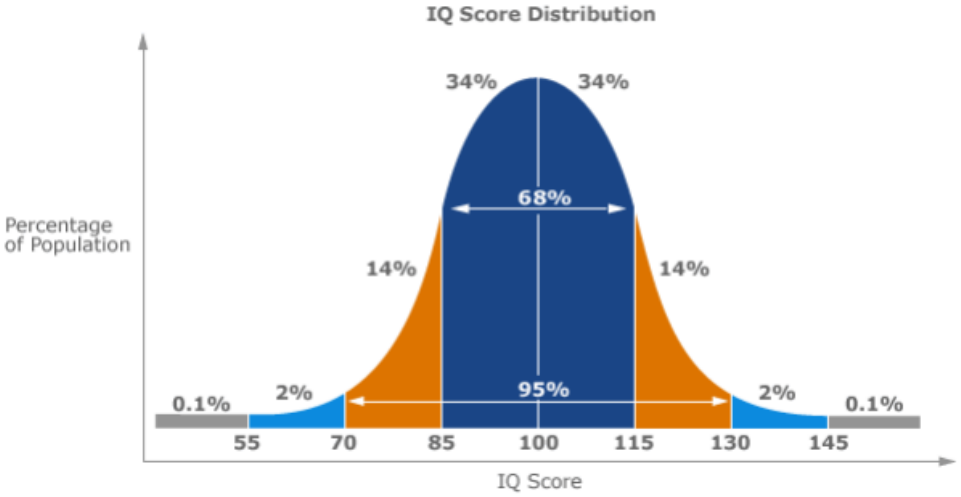
There are nine numbers in the list, so the middle one will be the (9 + 1) ÷ 2 = 10 ÷ 2 = 5th number:
13, 13, 13, 13, 14, 14, 16, 18, 21

So the median is 14.
The mode is the number that is repeated more often than any other, so 13 is the mode.
The largest value in the list is 21, and the smallest is 13, so the range is 21 – 13 = 8.

Mean: 15
Median: 14
Mode: 13
Range: 8

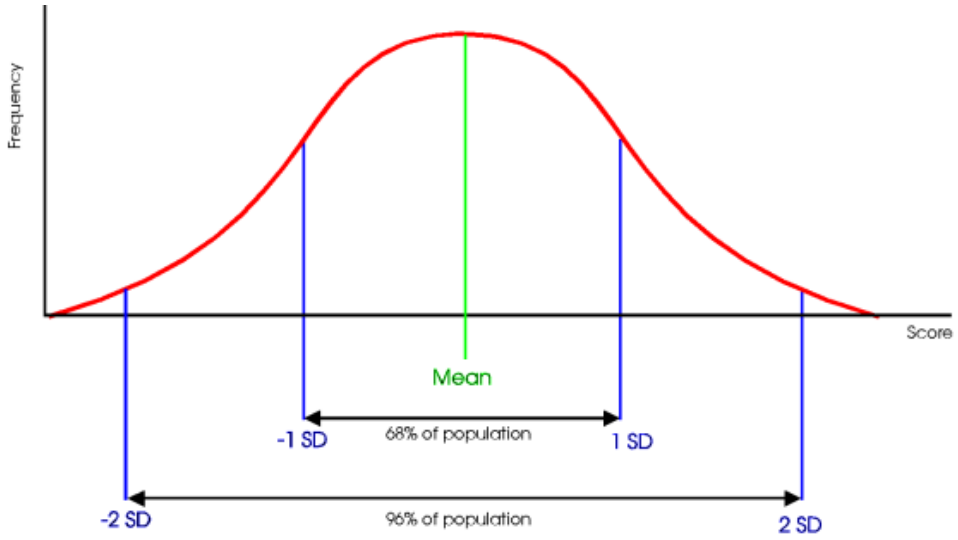
Statistics Basics

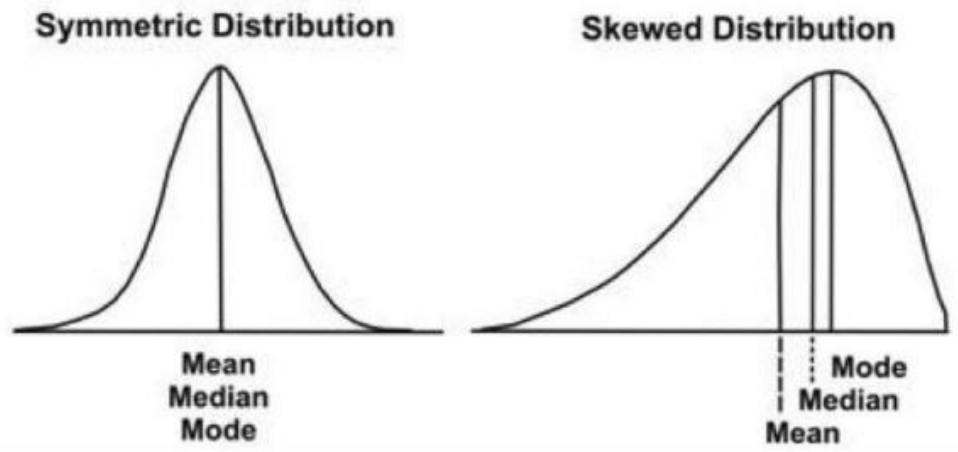
The most common basic statistics terms you'll come across are the mean, mode and median. These are all what are known as "Measures of Central Tendency." Also important in this early chapter of statistics is the shape of a distribution. This tells us something about how data is spread out around the mean or median. Perhaps the most common distribution you'll see is the normal distribution, sometimes called a bell curve. Heights, weights, and many other things found in nature tend to be shaped like this:



Example of a **Normal Distribution**

A normal distribution is an arrangement of a data set in which most values cluster in the middle of the range and the rest taper off symmetrically toward either extreme. Height is one simple example of something that follows a normal distribution pattern: Most people are of average height, the numbers of people that are taller and shorter than average are fairly equal and a very small (and still roughly equivalent) number of people are either extremely tall or extremely short.





Measure of Central Tendency

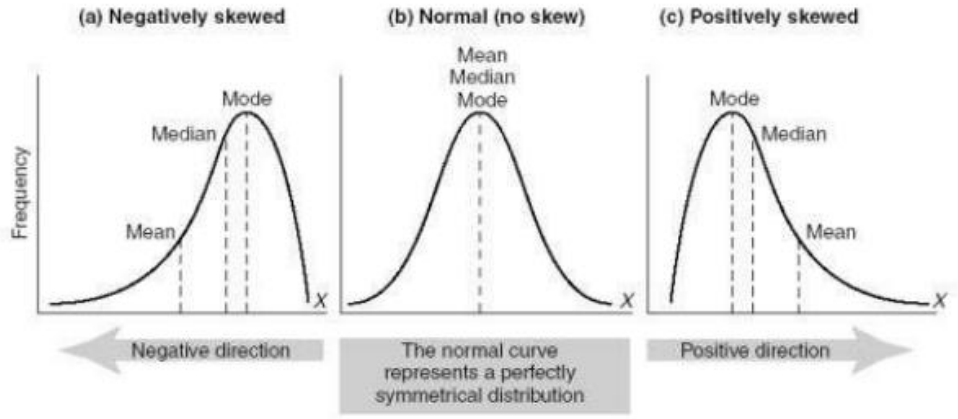
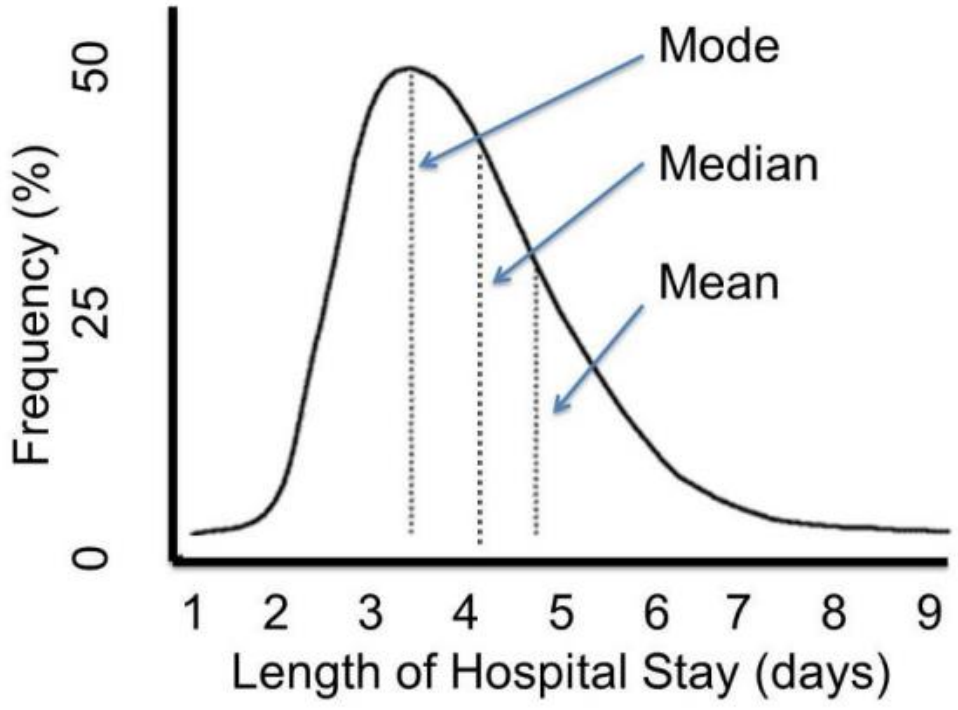


FIGURE 15.6 Examples of normal and skewed distributions



Bias in Statistics

Bias is the tendency of a statistic to overestimate or underestimate a parameter/value/data. Sampling error is the tendency for a statistic not to exactly match the population.

For example, let's say you have a population in the United States with an average height of 5 feet 9 inches. If you take a sample, even a fairly sizable sample of say, 10,000 people, it's unlikely that you'll get exactly 5 feet 9 inches. You might get very close, perhaps to within a fraction of an inch. If you repeat the experiment, you might get another very close result.

For example, in experiment 1 you might get 5 feet 8.9 inches and in experiment 2 you might get 5 feet 9.1 inches. The tendency for statistics to get very close, but not exactly right, is called sampling error.

Note: If the statistic is unbiased, the average of all statistics from all samples will average the true population parameter/value/data.