1. INTRODUCTION

Statistics Definition

The practice or science of collecting and analyzing numerical data in large quantities, especially for the purpose of inferring proportions in a whole from those in a representative sample.

Probability Definition

The extent to which something is likely to happen or be the case.

Why Study Statistics?

Answers is very simple: for making better decisions and choices of actions

Statistics and Engineering

The impact of the recent growth of statistics has been felt strongly in engineering and industrial management. It enables engineers to understand phenomena subject to variation and to effectively predict or control them.

A Case Study: Visually Inspecting Data to Improve Product Quality

This study illustrates the important advantages gained by appropriately plotting and then monitoring manufacturing data. Brick width was measured on three different parts selected from production every half hour during the first shift from 6 a.m. to 3 p.m. The data in the given table were obtained on a Friday. The sample mean, or average, for the first sample of 214, 211, and 218 (MMs) is

$$\frac{214+211+218}{3}=\frac{643}{3}=214.3$$

This value is the first entry in row marked $\bar{x}$**.** Let us plot graphical representation called ***X*-bar** chart, consists of plotting the sample averages versus time order.

Table 1.1 Slot depth (thousandths of an inch)

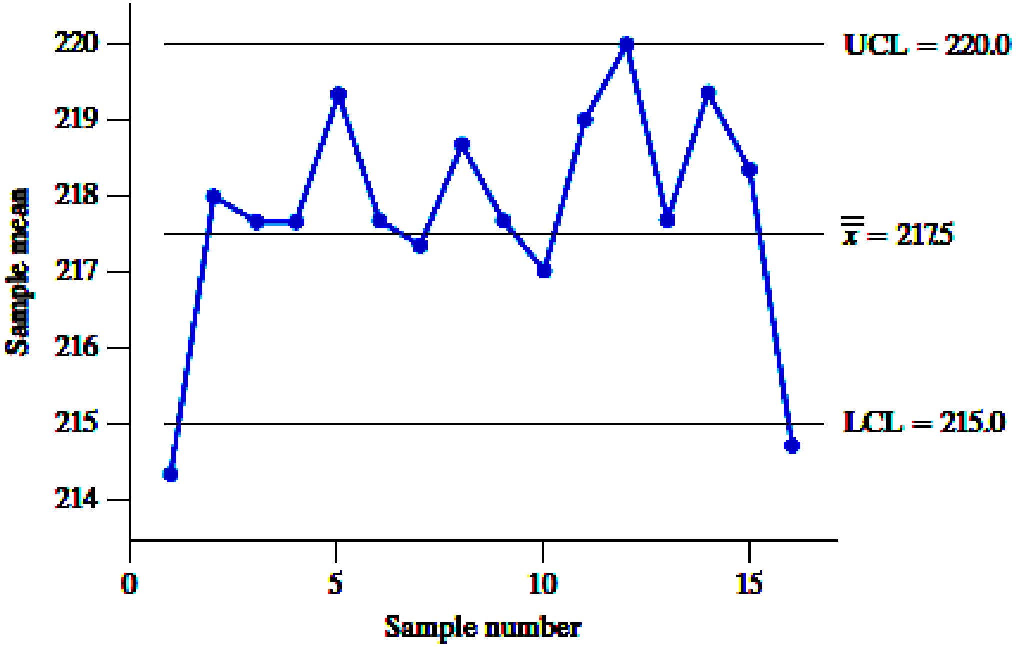
Time	6:30	7:00	7:30	8:00	8:30	9:00	9:30	10:00
1	214	218	218	216	217	218	218	219
2	211	217	218	218	220	219	217	219
3	218	219	217	219	221	216	217	218
Sum	643	654	653	653	658	653	652	656
$\bar{x}$	214.3	218.0	217.7	217.7	219.3	217.7	217.3	218.7
Time	10:30	11:00	11:30	12:30	1:00	1:30	2:00	2:30
1	216	216	218	219	217	219	217	215
2	219	218	219	220	220	219	220	215
3	218	217	220	221	216	220	218	214
Sum	653	651	657	660	653	658	655	644
$\bar{x}$	217.7	217.0	219.0	220.0	217.7	219.3	218.3	214.7

Suppose the process was stable and that it varied about a value of 217.5 mm. This value will be taken as the central line of the *X*-bar chart in the following figure . It was also assumed that the average slot dimension for a sample remained between certain control limits.

Lower control limit: LCL = 215.0

Upper control limit: UCL = 220.0

What does the chart tell us?



The first sample, taken at approximately 6:30 a.m., is outside the lower control limit. Further, a measure of the variation in this sample range

= largest − smallest = 218 − 211 = 7

2. ORGANIZATION AND DESCRIPTION OF DATA

Statistical data, obtained from surveys, experiments, or any series of measurements, are often so numerous that they are virtually useless unless they are condensed, or reduced into a more suitable form.

2.1 Pareto Diagrams and Dot Diagrams

Pareto Diagrams

This display, which orders each type of failure or defect according to its frequency, can help engineers identify important defects and their causes.`

When a company identifies a process for improvement, the first step is to collect data on the frequency of each type of failure. For example, the performance of pre-concrete slabs manufacturing is shown the following causes of faults and their frequencies:

power fluctuations	6
controller not stable	22
operator error	13
worn tool not replaced	2
other	5

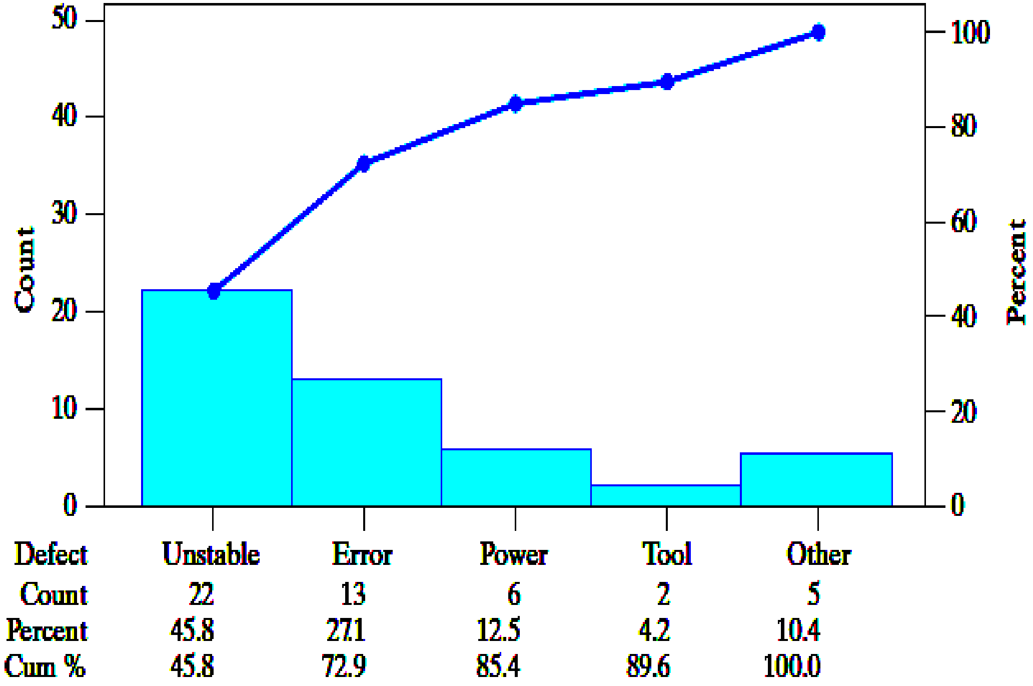
These data are presented as a special chart called a **Pareto diagram** as shown the following figure.

This diagram graphically depicts **Pareto’s empirical law** that any variety of events consists of a few major and many minor elements. Typically, two or three elements will account for more than half of the total frequency.

controller not stable	22	(22/48)X100 = 45.8 %
operator error	13	{(22+13)/48}X100 =72.9%
power fluctuations	6	{(22+13+6)/48}X100= 85.4%
worn tool not replaced	2	{(22+13+6+2)/48}X100= 89.6%
other	5	{(22+13+6+2+5)/48}X100= 100%
Sum	**48**	

The cumulative percentages are shown in the figure as a line graph whose scale is on the right-hand side of the Pareto diagram.

In the sake of quality improvement, this graph visually emphasizes the importance of reducing the frequency of controller defects.



Defect	Unstable	Error	Power	Tool	Other
Count	22	13	6	2	5
Percent	45.8	27.1	12.5	4.2	10.4
Cum %	45.8	72.9	85.4	89.6	100.0

Dot diagrams

A dot chart or dot plot or dot diagram is a statistical chart consisting of data points plotted on a fairly simple scale, typically using filled in circles. It is a type of simple. A dot plot is similar to a bar graph because the height of each “bar” of dots is equal to the number of items. To draw a dot plot, count the number of data points and draw a stack of dots that number.

Example:

A major food processor regularly monitors bacteria along production lines that include a stuffing process for meat products. An industrial engineer records the maximum amount of bacteria present along the production line, in the units Aerobic Plate Count per square inch (APC⁄in2), for n = 7 days.

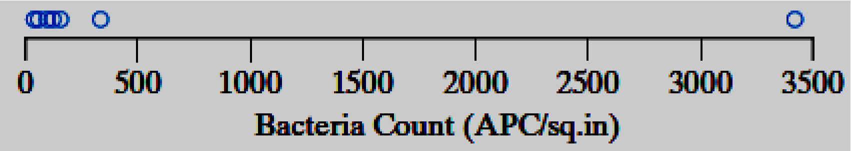
96.3 155.6 3408.0 333.3 122.2 38.9 58.0

Solution: First, ordered the data

38.9 58.0 96.3 122.2 155.6 333.3 3408.0

Second, by using open circles, the dot diagram shown in the figure can help to differentiate the crowded smaller values. The one very large bacteria count is the prominent feature. It indicates a possible health concern. Statisticians call such an unusual observation an ***outlier*** **نشاذ أو قيمة متطرفة**. Usually, outliers merit further attention.

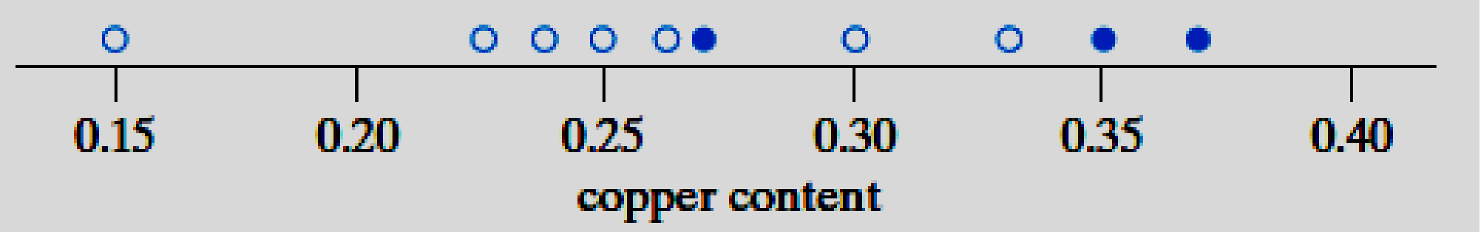
Maximum bacteria counts on seven days.



A dot diagram for multiple samples reveals differences

Example: The vessels الاوعية that contain the reactions at some nuclear power plants consist of two hemispherical components welded together. Copper in the welds could cause them to become brittle after years of service. Samples of welding material from one production run used in one plant had the copper contents 0.27, 0.35, 0.37. Samples from the next heat had values 0.23, 0.15, 0.25, 0.24, 0.30, 0.33, 0.26. Draw a dot diagram that highlights possible differences in the two production runs of welding material. If the copper contents for the two runs are different, they should not be combined to form a single estimate.

Solution: We plot the first group as solid circles and the second as open circles. It seems unlikely that the two production runs are alike because the top two values are from the first run. The two runs should be treated separately. The copper content of the welding material used at the power plant is directly related to the determination of safe operating life. Combining the sample would lead to an unrealistically غير واقعي low estimate of copper content and too long an estimate of safe life.



Dot diagram of copper content

2.2 Frequency Distributions

A **frequency distribution** is a table that divides a set of data into a suitable number of classes (categories), showing also the number of items belonging to each class. Instead of knowing the exact value of each item, we only know that it belongs to a certain class. On the other hand, grouping often brings out important features of the data, and the gain in “legibility” usually more than compensates for the loss of information.

To illustrate the construction of a frequency distribution, consider the following heights of 50 Nano-pillars were measured in nanometers (nm), or $10^{-9}\times$ meters during the fabricating a new transmission type electron multiplier of a flat silicon membrane.

245	333	296	304	276	336	289	234	253	292
366	323	309	284	310	338	297	314	305	330
266	**391**	315	305	290	300	292	311	272	312
315	355	346	337	303	265	278	276	373	271
308	276	364	390	298	290	308	**221**	274	343

The following steps should be followed:

1- The **maximum** number of classes may be determined by formula:

Number of Classes = **C**=1+3.3 * log **n** or **C**= $\sqrt{\mathbf{n}}$; where **n** is the total number of observations in the data. As it can be seen the number of classes depends on the number of observations, but it is seldom to use fewer than 5 or more than 15. The exception to the upper limit is when the size of the data set is several hundred or even a few thousand.

The following steps should be followed:

1- The **maximum** number of classes may be determined by formula:

Number of Classes = **C**=1+3.3 * log **n** or **C**= $\sqrt{\mathbf{n}}$; where **n** is the total number of observations in the data. As it can be seen the number of classes depends on the number of observations, but it is seldom to use fewer than 5 or more than 15. The exception to the upper limit is when the size of the data set is several hundred or even a few thousand.

2- Calculate the range of the data **(Range = Max – Min)** by finding minimum and maximum data value. Range will be used to determine the class interval or class width.

3- Decide about the **approximate** width of the class denote by **h** and obtained by:

h = Range/Number of Classes

Generally the class width is the same for all classes. The classes must cover from the lowest value (minimum) in the data set up to the highest (maximum) value. Also note that class intervals together must avoid a large number of empty, or almost empty classes. Starting point of the first class is arbitrary, and should be less than or equal to the minimum value. Usually it is started before the minimum value in such a way that the midpoint (the average of lower and upper class limits of the first class) is properly placed.

From the previous given measurements the largest observation is **391** and the smallest is **221**, accordingly the **Range** can be calculated as **391−221 = 170**. Also , the maximum number of classes is $\boldsymbol{\sqrt{50} \approx 7}$. However, consider The actual taken number of classes to be **5**, and the width of the each class to be **39**. Therefore, the class intervals can be: 206-245, 246-285, 286-325, 326-365, 366-405.

The number of observations in each class is counted to obtain the frequency distribution:

Limits of Classes	**Frequency**	
206–245	3	221,234,245
246–285	11	253,265,266,271,272,274,276,276,276,278,284
286–325	23	289,290,290,292,292,296,297,298,300,303,304,305,305, 308,308,309.310,311,312,314,315,315,323
326–365	9	330,333,336,337,338,343,346,355,364
366–405	4	366,373,390,391
Total	**50**	

In the preceding example, the data on heights of measurements may be thought of as values of a continuous variable which, conceivably, can be any value in an interval. But if we use classes such as 205–245, 245–285, 285–325, 325–365, 365–405, there exists the possibility of confusion, 245 could go into the first class or the second. To avoid this difficulty, we take an alternative approach. We make an **endpoint convention**. For the measurements height data, we can take (205, 245] as the first class, (245, 285] as the second, and so on through (365, 405]. That is, for this data set, we adopt the convention that the right-hand endpoint is included but the left-hand endpoint is not. For other data sets we may prefer to reverse the endpoint convention so the left-hand endpoint is included but the right-hand endpoint is not. Whichever endpoint convention is adopted, it should appear in the description of the frequency distribution.

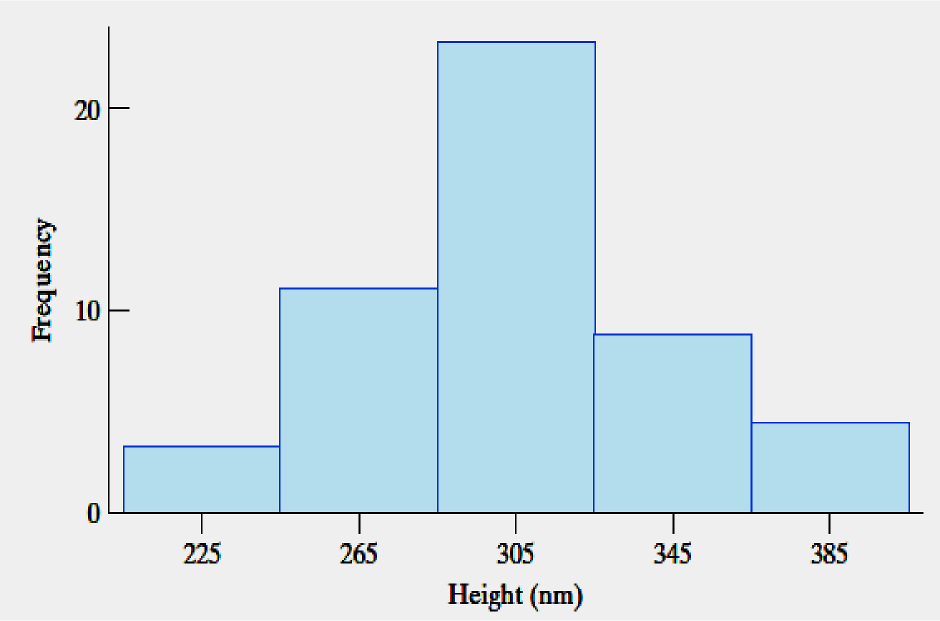
Under the convention that **the right-hand endpoint** is included, the frequency distribution of the Nano-pillar data is:

Class (nm)	Frequency
(205, 245]	3
(245, 285]	11
(285, 325]	23
(325, 365]	9
(365, 405]	4
Total	50

2.3 Graphs of Frequency Distributions

The most common form of graphical presentation of a frequency distribution is the **histogram**. The histogram of a frequency distribution is constructed of adjacent rectangles. Provided that the **class intervals** are equal, the heights of the rectangles represent the class frequencies and the bases of the rectangles extend between successive **class boundaries**. **A histogram** of the heights of Nano-pillars data is shown in The given figure.

Using right-hand endpoint convention, the interval (205, 245] that defines the first class has frequency 3, so the rectangle has height 3, the second rectangle, over the interval (245, 285], has height 9, and so on. The highest rectangle is over the interval (285,325] and has height 23. The histogram has a single peak and is reasonably symmetric. Almost half of the area, representing half of the observations, is over the interval 285 to 325 nanometers.



A density histogram

When a histogram is constructed from a frequency table having classes of unequal lengths, the height of each rectangle must be changed to:

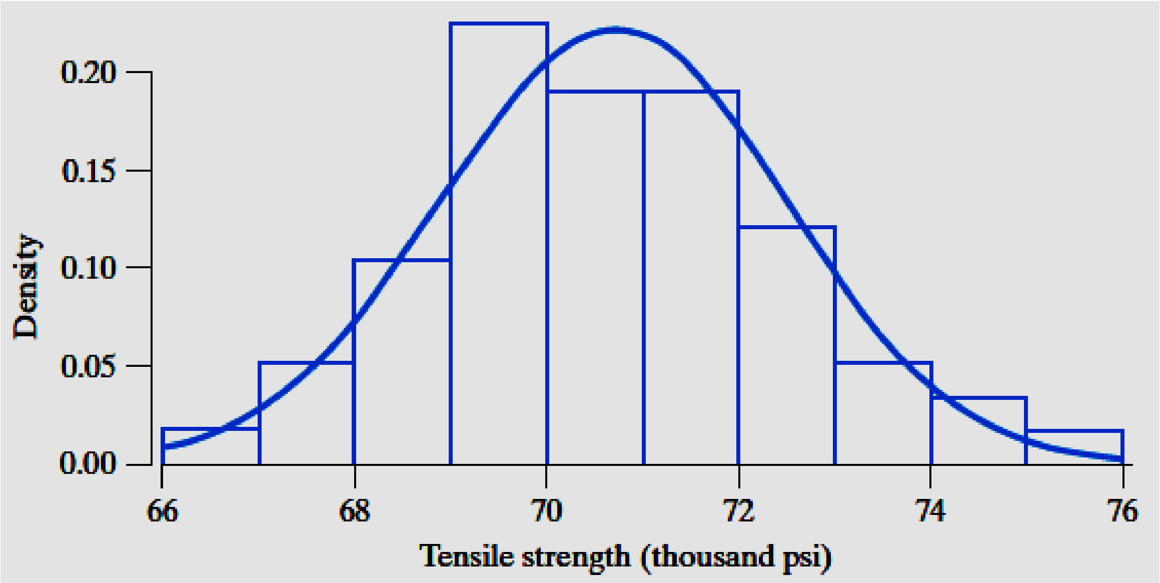
$$\text{height} = \frac{\text{relative frequency}}{\text{width}}$$

Example: Compressive strength was measured on 58 specimens of a new aluminum alloy undergoing development as a material for the next generation of aircraft.

66.4	67.7	68.0	68.0	68.3	68.4	68.6	68.8	68.9	69.0
69.2	69.3	69.3	69.5	69.5	69.6	69.7	69.8	69.8	69.9
70.0	70.1	70.2	70.3	70.3	70.4	70.5	70.6	70.6	70.8
71.0	71.1	71.2	71.3	71.3	71.5	71.6	71.6	71.7	71.8
71.9	72.1	72.2	72.3	72.4	72.6	72.7	72.9	73.1	73.3
69.1	70.0	70.9	71.8	73.5	74.2	74.5	75.3		

Class	Frequency	Density
(66,67]	1	0.017241
(67,68]	3	0.051724
(68,69]	6	0.103448
(69,70]	13	0.224138
(70,71]	11	0.189655
(71,72]	11	0.189655
(72,73]	7	0.12069
(73,74]	3	0.051724
(74,75]	2	0.034483
(75,76]	1	0.017241
Total	58	

Solution : The height of each rectangle equal to ***relative frequency / width***, so that its area equals the relative frequency. The resulting histogram, constructed has a nearly symmetric shape.



Histogram of aluminum alloy tensile strength

Height (nm)	Frequency	Density
(66,67]	1	0.017241
(67,68]	3	0.051724
(68,69]	6	0.103448
(69,70]	13	0.224138
(70,71]	11	0.189655
(71,72]	11	0.189655
(72,73]	7	0.12069
(73,74]	3	0.051724
(74,75]	2	0.034483
(75,76]	1	0.017241
Total	58	

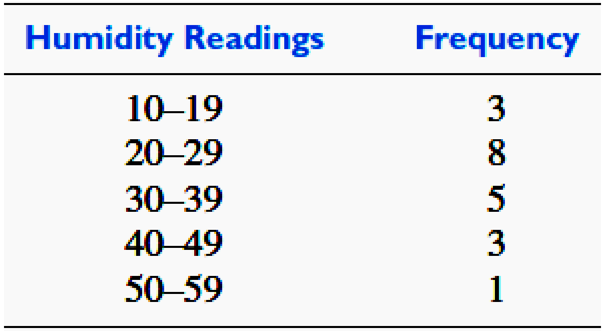
This example suggests that histograms, for observations that come from a continuous scale, can be approximated by smooth curves.

2.4 Stem-and-Leaf Displays

Generally, a stem and leaf plot, or stem plot, is a technique used to classify either discrete or continuous variables. A stem and leaf plot is used to organize data as they are collected. To illustrate, consider the following humidity readings rounded to the nearest percent:

29	**44**	**12**	**53**	**21**	**34**	**39**	**25**	**48**	**23**
17	**24**	**27**	**32**	**34**	**15**	**42**	**21**	**28**	**37**

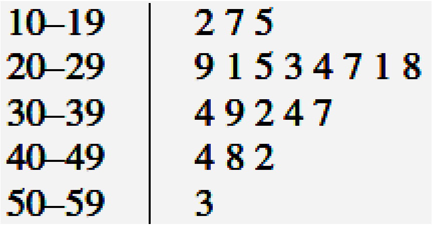
Proceeding as in **frequency distribution**, these data might be grouped into the following distribution:



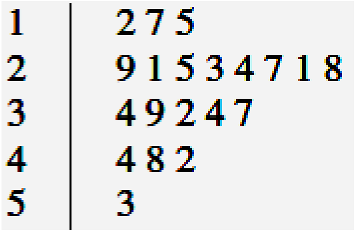
Humidity Readings	Frequency
10–19	3
20–29	8
30–39	5
40–49	3
50–59	1

If we wanted to avoid the loss of information in the table, it could keep track of the last digits of the readings within each class as the following:

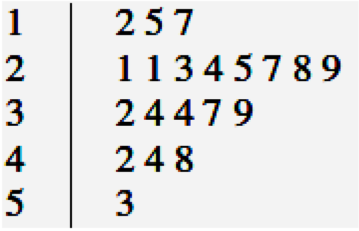
A stem and leaf plot looks something like a bar graph. Each number in the data is broken down into a stem and a leaf. The stem of the number includes all but the last digit. The leaf of the number will always be a single digit.



10–19	2 7 5
20–29	9 1 5 3 4 7 1 8
30–39	4 9 2 4 7
40–49	4 8 2
50–59	3



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



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

2.5 Descriptive Measures

Histograms, dot diagrams, and stem-and-leaf diagrams summarize a data set pictorially so it can be visually determined the overall pattern of variation. However, numerical measures can increase visual displays when describing a data set.

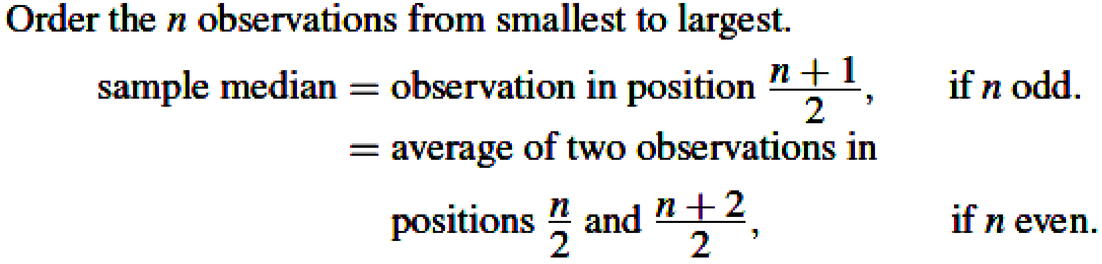
Mean and Median

Given a set of n measurements or observations, $X_1, X_2, \ldots, X_n$, there are many ways in which we can describe their center (middle, or central location). Most popular among these are the **arithmetic mean** and the **median.** The arithmetic mean (or more briefly the **mean**) is defined as the sum of the observations divided by sample size.

$$\textbf{Sample mean} = \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

The notation $\bar{x}$, read x **bar**, represents the mean of the x_i. To stress that it is based on the observations in a data set, it is often referred to $\bar{x}$ as the sample mean.

Sometimes it is preferable to use the **sample median** as a descriptive measure of the center, or location, of a set of data. More precisely, if the observations are arranged according to size and n is an **odd number**, the **median** is the value of the observation numbered **(n + 1)/ 2** ; if n is an **even number**, the **median** is defined as the mean (average) of the observations numbered $n/2$ and **(n + 2)/2**. To calculate **sample median** the following should be followed:



Order the n observations from smallest to largest.

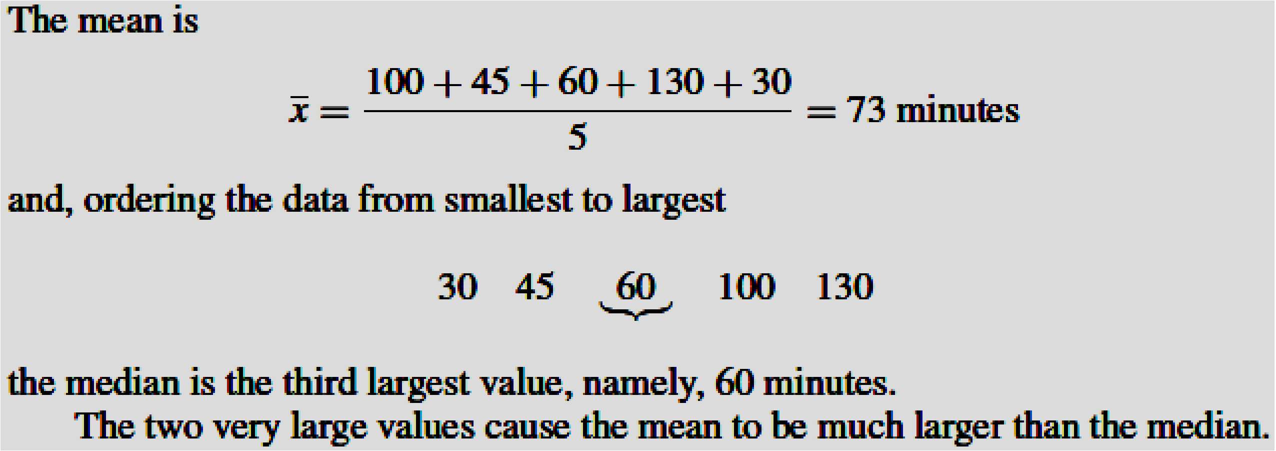
$$\begin{aligned} \text{sample median} &= \text{observation in position } \frac{n+1}{2}, && \text{if } n \text{ odd.} \\ &= \text{average of two observations in} \\ &\quad \text{positions } \frac{n}{2} \text{ and } \frac{n+2}{2}, && \text{if } n \text{ even.} \end{aligned}$$

Example: A sample of five university students responded to the question “How much time, in minutes, did you spend on the social network site yesterday?”

100 45 60 130 30

Find the mean and the median.

Solution:



The mean is

$$\bar{x} = \frac{100 + 45 + 60 + 130 + 30}{5} = 73 \text{ minutes}$$

and, ordering the data from smallest to largest

$$30 \quad 45 \quad \underbrace{60} \quad 100 \quad 130$$

the median is the third largest value, namely, 60 minutes.

The two very large values cause the mean to be much larger than the median.

Sample Variance

Mostly, it is always danger when summarizing a set of data in terms of a single number. In fact, the mean and median describe one important aspect of a set of data but they tell us nothing about the extent of variation.

One of the most important characteristics of almost any set of data is the vary among themselves. It is of basic importance in statistics. Therefore, it would seem reasonable to measure the variation of a set of data in terms of the amounts by which the values deviate from their mean. Consider the observations 11, 9, 17, 19, 4, 15, where $\boldsymbol{\bar{x}}$ = 12.5 is the balance point. The six deviations are −1.5, −3.5, 4.5, 6.5, −8.5, and 2.5. The sum of positive deviations: 4.5 + 6.5 + 2.5 = 13.5 exactly cancels the sum of the negative deviations: −1.5 − 3.5 − 8.5 = −13.5 so the sum of all the deviations is 0. In fact, he sum of the deviations is always zero. That is:

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0$$

Because the deviations sum is always zero, their signs need to be removed. It is convenient and agreed to use the most common measure of variation by square each deviation. The **sample variance**, $\boldsymbol{s^2}$, is essentially the average of the squared deviations from the mean, $\boldsymbol{\bar{x}}$, and is defined by the following formula.

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

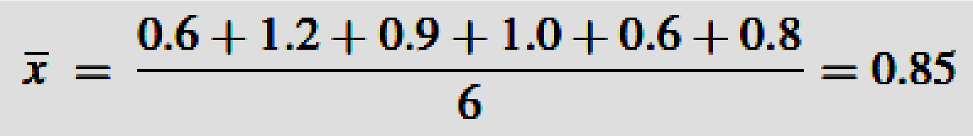
Note that there are only ***n*** **−1 (**instead of ***n*** is)?!

Example:

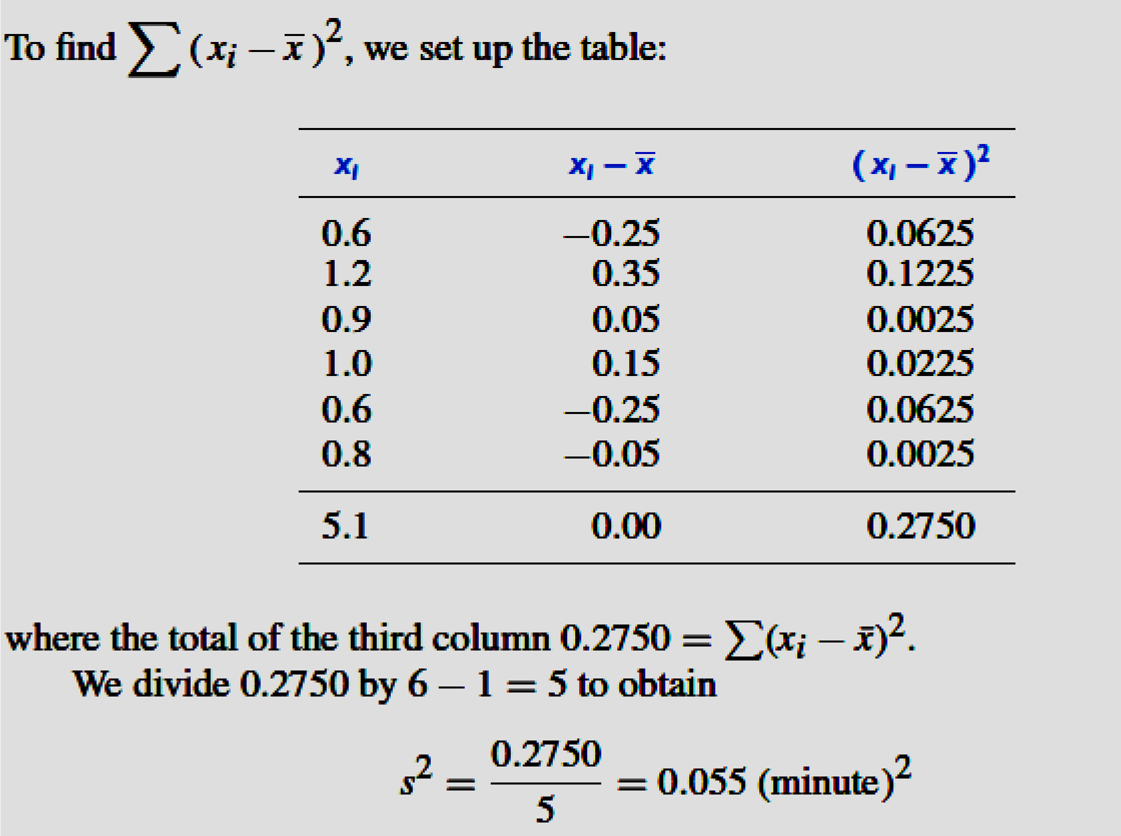
The delay times (handling, setting, and positioning the tools) for cutting 6 parts on an engine lathe are 0.6, 1.2, 0.9, 1.0, 0.6, and 0.8 minutes. Calculate s^2.

Solution:

First we calculate the mean



$$\bar{x} = \frac{0.6 + 1.2 + 0.9 + 1.0 + 0.6 + 0.8}{6} = 0.85$$



To find $\sum (x_i - \bar{x})^2$, we set up the table:

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
0.6	–0.25	0.0625
1.2	0.35	0.1225
0.9	0.05	0.0025
1.0	0.15	0.0225
0.6	–0.25	0.0625
0.8	–0.05	0.0025
5.1	0.00	0.2750

where the total of the third column $0.2750 = \sum (x_i - \bar{x})^2$.

We divide 0.2750 by $6 - 1 = 5$ to obtain

$$s^2 = \frac{0.2750}{5} = 0.055 \text{ (minute)}^2$$

Standard Deviation

Notice in the previous example that the units of s^2 are not those of the original observations. The data are delay times in minutes, but s^2 has the unit **(minute)**^{2}. Consequently, we define the **standard deviation** of n observations $x_1, x_2, . . . , x_n$ as the square root of their **variance**, namely: **Sample standard deviation:**

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

The standard deviation is by far the most generally useful measure of variation. Its advantage over the variance is that it is expressed in the same units as the observations.

Example: With reference to the previous example, calculate s.

Solution: From the previous example, $\mathbf{s}^2 = 0.055$. Take the square root and get

$$s = \sqrt{0.055} = 0.23 \text{ minute}$$

The **standard deviation** and the **variance** are measures of **absolute variation**; that is, they measure the actual amount of variation in a set of data, and they depend on the scale of measurement. To compare the variation in several sets of data, it is generally desirable to use a measure of **relative variation**, for instance, the **coefficient of variation**, which gives the standard deviation as a percentage of the mean:

Coefficient of variation : $$V = \frac{s}{\bar{x}} \cdot 100\%$$

Example:

Measurements made with one micrometer of the diameter of a ball bearing have a mean of 3.92 mm and a standard deviation of 0.0152 mm, whereas measurements made with another micrometer of the unstretched length of a spring have a mean of 1.54 inches and a standard deviation of 0.0086 inch. Which of these two measuring instruments is relatively more precise?

Solution: For the first micrometer the coefficient of variation is

$$V = \frac{0.0152}{3.92} \cdot 100 = 0.39\%$$

and for the second micrometer the coefficient of variation is

$$V = \frac{0.0086}{1.54} \cdot 100 = 0.56\%$$

Thus, the measurements made with the first micrometer are relatively more precise.

2.7 The Calculation of $\bar{x}$ and s

There are different methods for calculating $\bar{x}$ and s from especially data that are already grouped into intervals. An alternative formula for s^2 forms the basis of the grouped data formula for variance. It was originally introduced to simplify hand calculations.

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

Example: Find the mean and the standard deviation of the following miles per gallon (mpg) obtained in 20 test runs performed on urban roads with an intermediate-size car:

19.7	21.5	22.5	22.2	22.6
21.9	20.5	19.3	19.9	21.7
22.8	23.2	21.4	20.8	19.4
22.0	23.0	21.1	20.9	21.3

Solution: The sum of the data is 427.7 and that the sum of their squares is 9,173.19. Consequently,

$$\bar{x} = \frac{427.7}{20} = 21.39 \text{ mpg}$$

and

$$s^2 = \frac{9{,}173.19 - (427.7)^2/20}{19} = 1.412$$

Accordingly, the standard deviation is **S** = 1.19 mpg

To calculate $\overline{x}$ and **s** from grouped data, it is assumed something about the distribution of the values within each class. Each value is represented within a class by the corresponding class mark. Then the sum of the **x'**s and the sum of their squares can be written:

$$\sum_{i=1}^{k} x_i f_i \quad \text{and} \quad \sum_{i=1}^{k} x_i^2 f_i$$

where X_i is the class mark of the ith class, f_i is the corresponding class frequency, and k is the number of classes in the distribution. Substituting these sums into the formula for x and the computing formula for s^2, we get:

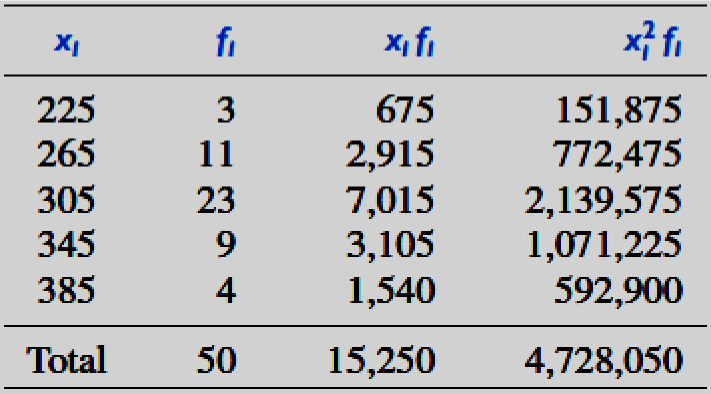
$$\bar{x} = \frac{\sum\limits_{i=1}^{k} x_i f_i}{n}$$

$$s^2 = \frac{\sum\limits_{i=1}^{k} x_i^2 f_i - \left(\sum\limits_{i=1}^{k} x_i f_i \right)^2 / n}{n - 1}$$

Example: Use the distribution obtained previously to calculate the mean, variance, and standard deviation of the Nano-pillar heights data:

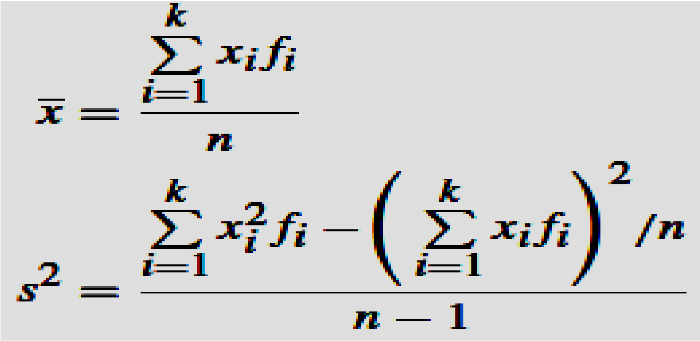
Height (nm)	Frequency
(205, 245]	3
(245, 285]	11
(285, 325]	23
(325, 365]	9
(365, 405]	4
Total	50

Solution: Recording the class marks and the class frequencies in the first two columns and the products $\boldsymbol{x_i\ f_i}$ and $\boldsymbol{x_i^2\ \ f_i}$ in the third and fourth columns,



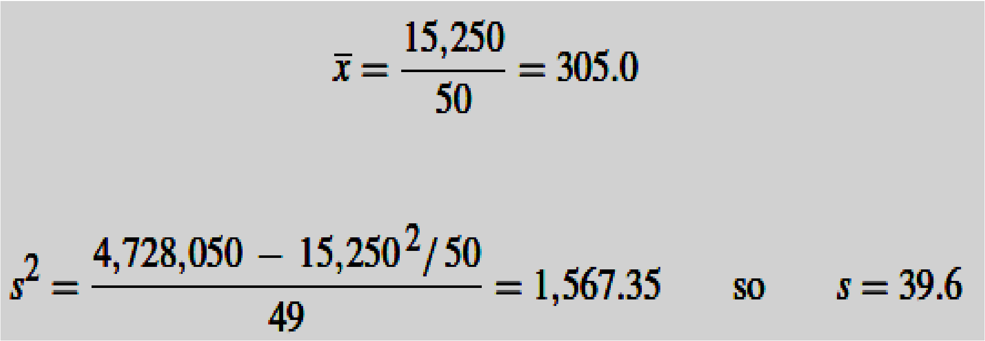
x_i	f_i	$x_i f_i$	$x_i^2 f_i$
225	3	675	151,875
265	11	2,915	772,475
305	23	7,015	2,139,575
345	9	3,105	1,071,225
385	4	1,540	592,900
Total	50	15,250	4,728,050

Then, substitution into the formula yields:



$$\bar{x} = \frac{\sum_{i=1}^{k} x_i f_i}{n}$$

$$s^2 = \frac{\sum_{i=1}^{k} x_i^2 f_i - \left(\sum_{i=1}^{k} x_i f_i\right)^2 / n}{n-1}$$



$$\bar{x} = \frac{15{,}250}{50} = 305.0$$

$$s^2 = \frac{4{,}728{,}050 - 15{,}250^2/50}{49} = 1{,}567.35 \quad \text{so} \quad s = 39.6$$