Summarizing Data: Listing and Grouping

Listing Numerical Data

Listing and thus, organizing the data is usually the first task in any kind of statistical analysis.

EXAMPLE: Consider the following data, representing the lengths (in centimeters) of 60 sea trout caught by a commercial trawler in Bay Area:

19.2	19.6	17.3	19.3	19.5	20.4	23.5	19.0	19.4	18.4
19.4	21.8	20.4	21.0	21.4	19.8	19.6	21.5	20.2	20.1
20.3	19.7	19.5	22.9	20.7	20.3	20.8	19.8	19.4	19.3
19.5	19.8	18.9	20.4	20.2	21.5	19.9	21.7	19.5	20.9
18.1	20.5	18.3	19.5	18.3	19.0	18.2	21.9	17.0	19.7
20.7	21.1	20.6	16.6	19.4	18.6	22.7	18.5	20.1	18.6

The mere gathering of this information is so small task, but it should be clear that more must be done to make the numbers comprehensible.

What can be done to make this mass of information more usable? Some persons find it interesting to locate the extreme values, which are 16.6 and 23.5 for this list. Occasionally, it is useful to sort the data in an ascending or descending order. The following list gives the lengths of the trout arranged in an ascending order.

16.6	17.0	17.3	18.1	18.2	18.3	18.3	18.4	18.5	18.6
18.6	18.9	19.0	19.0	19.2	19.3	19.3	19.4	19.4	19.4
19.4	19.5	19.5	19.5	19.5	19.5	19.6	19.6	19.7	19.7
19.8	19.8	19.8	19.9	20.1	20.1	20.2	20.2	20.3	20.3
20.4	20.4	20.4	20.5	20.6	20.7	20.7	20.8	20.9	21.0
21.1	21.4	21.5	21.5	21.7	21.8	21.9	22.7	22.9	23.5

Sorting a large set of numbers in an ascending or descending order can be a surprisingly difficult task. It is simple, though, if we can use a computer or a graphing calculator.

If a set of data consists of relatively few values, many of which are repeated, we simply count how many times each value occurs and then present the results in the form of a Table or a **dot diagram**. In such a diagram we indicate by means of dots how many times each value occurs.

EXAMPLE: An audit of twenty tax returns revealed $0, 2, 0, 0, 1, 3, 0, 0, 0, 1, 0, 1, 0, 0, 2, 1, 0, 0, 1$, and 0 mistakes in arithmetic.

(a) Construct a table showing the number of tax returns with 0, 1, 2, and 3, mistakes in arithmetic.

(b) Draw a dot diagram displaying the same information.

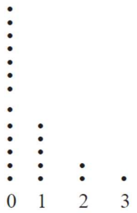
EXAMPLE: An audit of twenty tax returns revealed $0,2,0,0,1,3,0,0,0,1,0,1,0,0,2,1,0,0,1,$ and 0 mistakes in arithmetic.

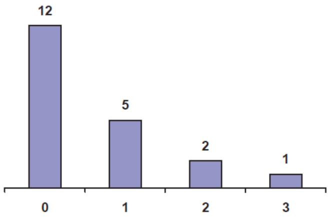
(a) Construct a table showing the number of tax returns with 0, 1, 2, and 3, mistakes in arithmetic.

(b) Draw a dot diagram displaying the same information.

Solution: Counting the number of 0's, 1's, 2's and 3's we find that they are, respectively, 12, 5, 2, and 1. This information is displayed as follows, in tabular form on the left and n graphical form on the right.

Number of mistakes	Number of the returns
0	12
1	5
2	2
3	1





Stem-And-Leaf-Display

Dot diagrams are impractical and ineffective when a set of data contains many different values or categories, or when some of the values or categories require too many dots to yield a coherent picture.

In recent years, an alternative method of listing data has been proposed for the exploration of relatively small sets of numerical data. It is called a **stem-and leaf display** and it also yields a good overall picture of the data without any appreciable loss of information.

EXAMPLE: Consider the following data on the number of rooms occupied each day in a resort hotel during a recent month of June:

55 49 37 57 46 40 64 35 73 62 61 43 72 48 54 69 45 78 46 59 40 58 56 52 49 42 62 53 46 81

The smallest and largest values are 35 and 81, so that a dot diagram would require that we allow for 47 possible values. Actually, only 25 of the values occur, but in order to avoid having to allow for that many possibilities, let us combine all the values beginning with a 3, all those beginning with a 4, all those beginning with a 5 and so on. This would yield

37 35
49 46 40 43 48 45 46 40 49 42 46
55 57 54 59 58 56 52 53
64 62 67 69 62
73 72 78
81

This arrangement is quite informative, but it is not the kind of diagram we use in actual practice. To simplify it further, we show the first digit only once for each row, on the left and separated from the other digits by means of a vertical line. This leaves us with

3	7	5									
4	9	6	0	3	8	5	6	0	9	2	6
5	5	7	4	9	8	6	2	3			
6	4	2	1	9	2						
7	3	2	8								
8	1										

And this is what we refer to as a stem-and leaf display. In this arrangement, each row is called a **stem**, each number on a stem to the left of the vertical line is called a **stem label**, and each number on a stem to the right of the vertical line is called a **leaf**. As we shall see later, there is a certain advantage to arranging the leaves on each stem according to size, and for our data this would yield

3	5	7									
4	0	0	2	3	5	6	6	6	8	9	9
5	2	3	4	5	6	7	8	9			
6	1	2	2	4	9						
7	2	3	8								
8	1										

Now suppose that in the room occupancy Example we had wanted to use more than six stems. Using each stem label twice, if necessary, once to hold the leaves from 0 to 4 and once to hold the leaves from 5 to 9, we would get

3	5	7					
4	0	0	2	3			
4	5	6	6	6	8	9	9
5	2	3	4				
5	5	6	7	8	9		
6	1	2	2	4			
6	9						
7	2						
7	8						
8	1						

Frequency Distributions

When we deal with large sets of data, and sometimes even when we deal with not so large sets of data, it can be quite a problem to get a clear picture of the information that they convey. This usually requires that we rearrange and/or display the raw (untreated) data in some special form. Traditionally, this involves a frequency distribution or one of its graphical presentations, where we group or classify the data into a number of categories or classes.

EXAMPLE: A recent study of their total billings (rounded to the nearest dollar) yielded data for a sample of 4,757 law firms. Rather than providing printouts of the 4,757 values, the information is disseminated by means of the following table:

Total billings	Number of law firms
Less than $300,000	2,405
$300,000 to $499,999	1,088
$500,000 to $749,999	271
$750,000 to $999,999	315
$1,000,000 or more	678
Total	4,757

This distribution does not show much detail, but it may well be adequate for most practical purposes.

EXAMPLE: The following table summarizes the 2,439 complaints received by an airline about comfort-related characteristics of its airplanes:

Nature of complaint	Number of complaints
Inadequate leg room	719
Uncomfortable seats	914
Narrow aisles	146
Insufficient carry-on facilities	218
Insufficient restrooms	58
Miscellaneous other complaints	384
Total	2,439

When data are grouped according to numerical size, as in the first example, the resulting table is called a **numerical or quantitative distribution**. When they are grouped into nonnumerical categories, as in the second example, the resulting table is called a **categorical or qualitative distribution.**

The construction of a frequency distribution consists essentially of three steps:

1. **Choosing the classes (intervals or categories).**
2. **Sorting or tallying the data into these classes.**
3. **Counting the number of items in each class.**

Since the second and third steps are purely mechanical, we concentrate here on the first, namely, that of choosing a suitable classification.

For numerical distributions, this consists of deciding how many classes we are going to use and from where to where each classes should go, both of these choices are essentially arbitrary, but the following rules are usually observed:

We seldom use fewer than 5 or more than 15 classes; the exact number we use in a given situation depends largely on how many measurements or observations there are.

Clearly, we would lose more than we gain if we group five observations into 12 classes with most of them empty, and we would probably discard too much information if we group a thousand measurements into three classes.

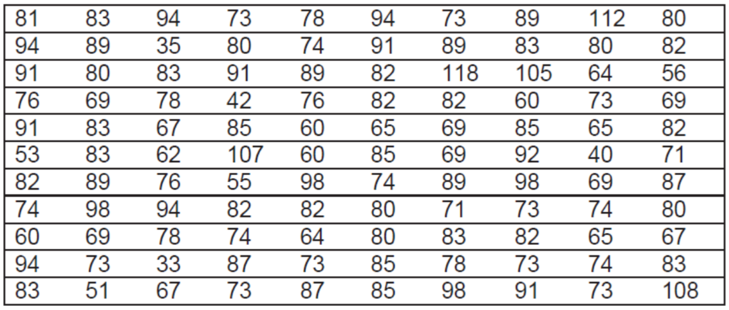
We always make sure that each item (measurement or observation) goes into one and only one class.

To this end, we must make sure that the smallest and largest values fall within the classification, that none of the values can fall into a gap between successive classes, and that the classes do not overlap, namely, that successive classes have no values in common.

Whenever possible, we make the classes cover equal ranges of values.

Also, if we can, we make these ranges multiples of numbers that are easy to work with, such as 5, 10, or 100, since this will tend to facilitate the construction and the use of a distribution.

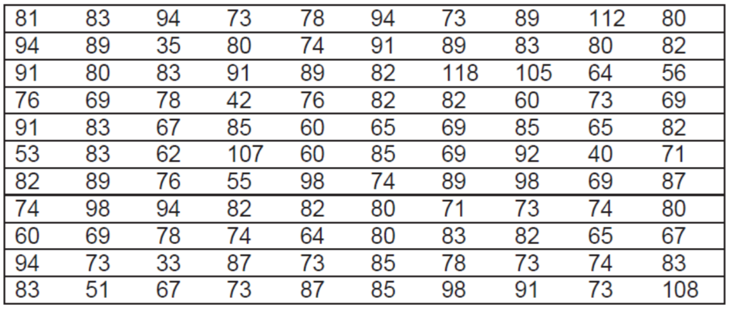
EXAMPLE: Based on 1997 figures, the following are 11.0 “waiting times” (in minutes) between eruptions of the Old Faithful Geyser m Yellowstone National Park:



81	83	94	73	78	94	73	89	112	80
94	89	35	80	74	91	89	83	80	82
91	80	83	91	89	82	118	105	64	56
76	69	78	42	76	82	82	60	73	69
91	83	67	85	60	65	69	85	65	82
53	83	62	107	60	85	69	92	40	71
82	89	76	55	98	74	89	98	69	87
74	98	94	82	82	80	71	73	74	80
60	69	78	74	64	80	83	82	65	67
94	73	33	87	73	85	78	73	74	83
83	51	67	73	87	85	98	91	73	108

Construct a frequency distribution.

EXAMPLE: Based on 1997 figures, the following are 11.0 “waiting times” (in minutes) between eruptions of the Old Faithful Geyser m Yellowstone National Park:

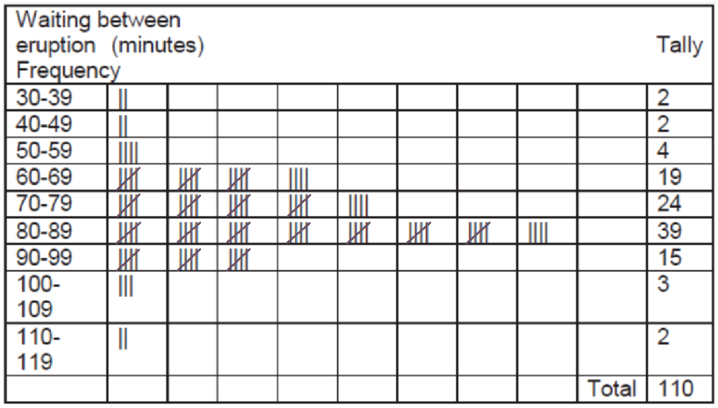


81	83	94	73	78	94	73	89	112	80
94	89	35	80	74	91	89	83	80	82
91	80	83	91	89	82	118	105	64	56
76	69	78	42	76	82	82	60	73	69
91	83	67	85	60	65	69	85	65	82
53	83	62	107	60	85	69	92	40	71
82	89	76	55	98	74	89	98	69	87
74	98	94	82	82	80	71	73	74	80
60	69	78	74	64	80	83	82	65	67
94	73	33	87	73	85	78	73	74	83
83	51	67	73	87	85	98	91	73	108

Construct a frequency distribution.

Solution: Since the smallest value is 33 and the largest value is 118, we have to cover an interval of 86 values and a convenient choice would be to use the nine classes 30 -39, 40 - 49, 50 - 59, 60 - 69, 70 - 79, 80 - 89, 90 - 99, 100 - 109, and 110-119. These classes will accommodate all of the data, they do not overlap, and they are all of the same size. There are other possibilities (for instance, 25 - 34, 35 - 44, 45 - 54, 55 - 64, 65 - 74, 75 - 84, 85 - 94, 95 - 104, 105 - 114, and 115 - 124), but it should be apparent that our first choice will facilitate the tally.

We now tally the 110 values and get the result shown in the following table:



Waiting between eruption (minutes) Frequency										Tally
30-39	||									2
40-49	||									2
50-59	||||									4
60-69	卌	卌	卌	||||						19
70-79	卌	卌	卌	卌	||||					24
80-89	卌	卌	卌	卌	卌	卌	卌	||||		39
90-99	卌	卌	卌							15
100-109	|||									3
110-119	||									2
									Total	110

The numbers given in the right-hand column of this table, which show how many values fall into each class, are called the **class frequencies**. The smallest and largest values that can go into any given class are called its **class limits**, and for the distribution of the waiting times between eruptions they are 30 and 39, 40 and 49,50 and 59, . . ., and 110 and 119. More specifically, 30, 40, 50, . . ., and 110 are called the **lower class limits**, and 39, 49, 59, . . ., and 119 are called the **upper class limits**.

Numerical distributions also have what we call class marks and classes intervals. **Class marks** are simply the midpoints of the classes, and they are found by adding the lower and upper limits of a class (or its lower and upper boundaries) and dividing by 2. A **class interval** is merely the length of a class, or the range of values it can contain, and it is given by the difference between its boundaries. If the classes of a distribution are all equal in length, their common class interval, which we call the class interval or the distribution, is also given by the difference between any two successive class marks. Thus, the class marks of the waitingtime distribution are 34.5, 44.5, 54.5, . . ., and 114.5, and the class intervals and the class interval of the distribution are all equal to 10.

There are essentially two ways in which frequency distributions can be modified to suit particular needs. One way is to convert a distribution into a percentage distribution by dividing each class frequency by the total number of items grouped, and then multiplying by 100.

EXAMPLE: Convert the waiting-time distribution of the previous Example into a percentage distribution.

Solution: The first class contains

$$\frac{2}{110} \cdot 100 = 1.82\%$$

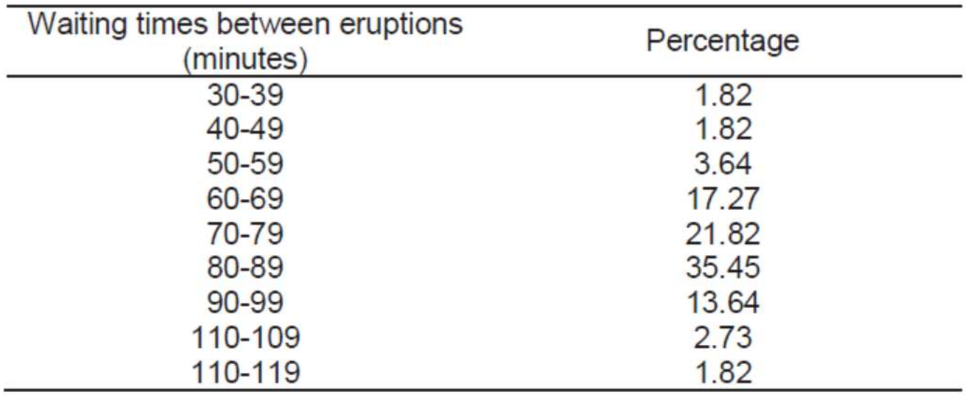
of the data (rounded to two decimals), and so does the second class. The third class contains

$$\frac{4}{110} \cdot 100 = 3.64\%$$

of the data, the fourth class contains

$$\frac{19}{110} \cdot 100 = 17.27\%$$

of the data, ..., and the bottom class again contains 1.82% of the data. These results are shown in the following table:



Waiting times between eruptions (minutes)	Percentage
30-39	1.82
40-49	1.82
50-59	3.64
60-69	17.27
70-79	21.82
80-89	35.45
90-99	13.64
110-109	2.73
110-119	1.82

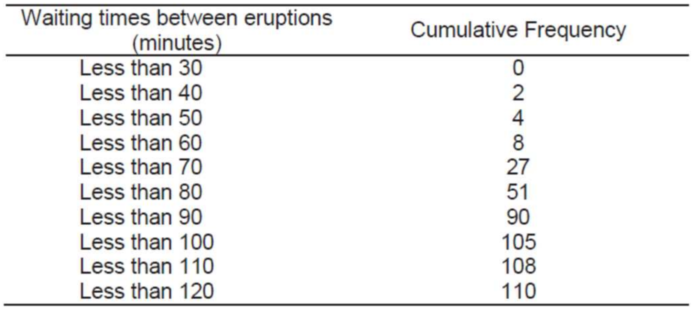
The percentages total 100.01, with the difference, of course, due to rounding.

The other way of modifying a frequency distribution is to convert it into a “less than,” “or less,” “more than,” or “or more” **cumulative distribution**. To construct a cumulative distribution, we simply add the class frequencies, starting either at the top or at the bottom of the distribution.

EXAMPLE: Convert the waiting-time distribution of the Example above into a cumulative “less than” distribution.

EXAMPLE: Convert the waiting-time distribution of the Example above into a cumulative "less than" distribution.

Solution: Since none of the values is less than 30, 2 of the values are less than 40, $2+2=4$ of the values are less than 50, $2+2+4=8$ of the values are less than 60, ..., and all 110 of the values are less than 120, we get



Waiting times between eruptions (minutes)	Cumulative Frequency
Less than 30	0
Less than 40	2
Less than 50	4
Less than 60	8
Less than 70	27
Less than 80	51
Less than 90	90
Less than 100	105
Less than 110	108
Less than 120	110

Note that instead of "less than 30" we could have written "29 or less," instead of "less than 40" we could have written "39 or less," instead of "less than 50" we could have written "49 or less," and so forth.

In the same way we can also convert a percentage distribution into a cumulative percentage distribution. We simply add the percentages instead of the frequencies, starting either at the top or at the bottom of the distribution.

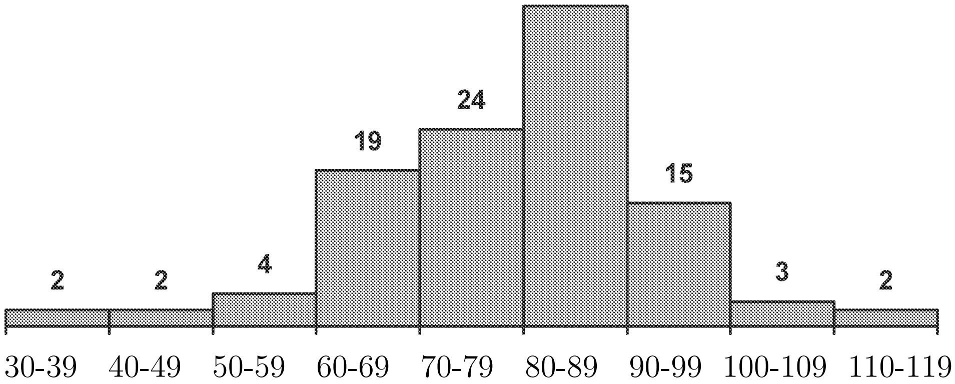
Graphical Presentation

When frequency distributions are constructed mainly to condense large sets of data and present them in an "easy to digest" form, it is usually most effective to display them graphically.

For frequency distributions, the most common form of graphical presentation is the **histogram**. Histograms are constructed by representing the measurements or observations that are grouped on a horizontal scale, the class frequencies on a vertical scale, and drawing rectangles whose bases equal the class intervals and whose heights are the corresponding class frequencies.

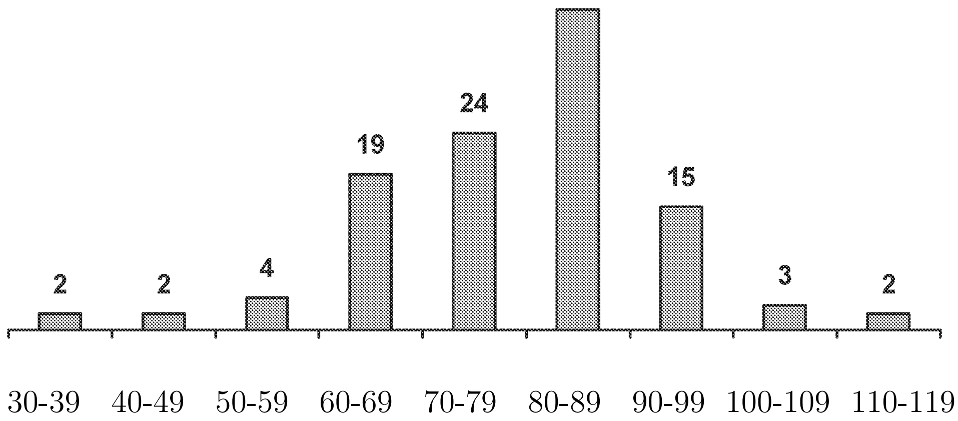
The marketing on the horizontal scale of histogram can be the class limits, the class marks, the class boundaries, or arbitrary key values. For practical reasons, it is usually preferable to show the class limits, even though the rectangles actually go from one class boundary to the next.

EXAMPLE:



Histogram of waiting times between eruptions of old faithful geyser

Also referred to at times as histograms are bar charts, such as the one shown in the Figure below. The heights of the rectangles, or bars again represent the class frequency but there is no pretense of having a continuous horizontal scale.



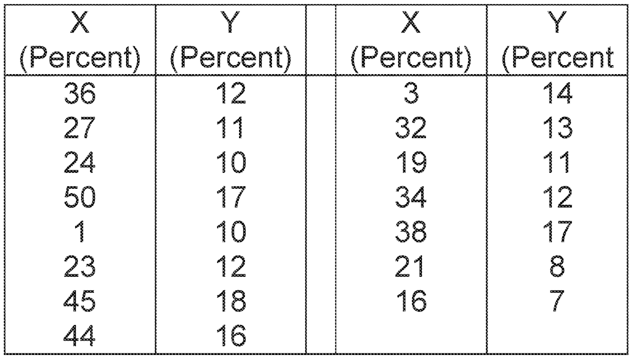
Bar Chart of distribution of waiting times between eruptions of old faithful geyser

Summarizing Two-Variable Data

So far we have dealt only with situations involving one variable – the room occupancies, the waiting times between eruptions of Old Faithful, and so on. In actual practice, many statistical methods apply to situations involving two variables x and y.

Pairs (x, y), in the same way which we denote points in the plane, with x, and y being their x- and y-coordinates. When we actually plot the points corresponding to paired values of x and y, we refer to the resulting graph as a **scatter diagram**, a **scatter plot**, or a **scatter gram**. As their name implies, such graphs are useful tools in the analysis of whatever relationship there may exist between the x's and the y's namely, judging whether there are any discernible patterns.

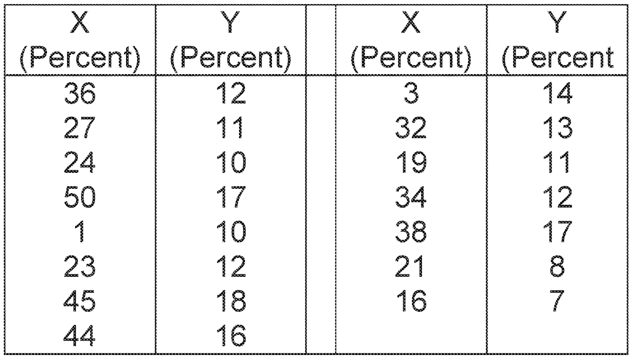
EXAMPLE: Raw materials used in the production of synthetic fiber are stored in a place that has no humidity control. Following are measurement of the relative humidity in the storage place, x, and the moisture content of a sample of the raw material, y, on 15 days



X (Percent)	Y (Percent)		X (Percent)	Y (Percent
36	12		3	14
27	11		32	13
24	10		19	11
50	17		34	12
1	10		38	17
23	12		21	8
45	18		16	7
44	16			

Construct a scatter gram.

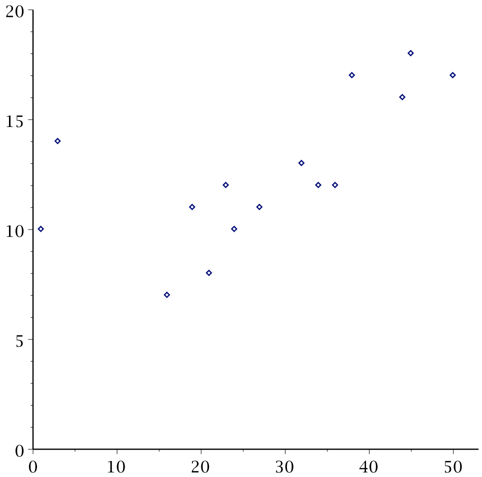
EXAMPLE: Raw materials used in the production of synthetic fiber are stored in a place that has no humidity control. Following are measurement of the relative humidity in the storage place, x, and the moisture content of a sample of the raw material, y, on 15 days



X	Y		X	Y
(Percent)	(Percent)		(Percent)	(Percent
36	12		3	14
27	11		32	13
24	10		19	11
50	17		34	12
1	10		38	17
23	12		21	8
45	18		16	7
44	16			

Construct a scatter gram.

Solution: Scatter grams are easy enough to draw, yet the work can be simplified by using appropriate computer software or a graphing calculator.


20
15
10
5
0
0
10
20
30
40
50