Definitions and Terminologies

We are going to study statistical methods to help us describe a data set, display a data set, and make inference from a data set.

We may work with a data set that is already available or may collect data to understand the behavior of a variable.

Data are of two types:

Categorical: Such data can not be used to do computations with. Example of variables that can be termed categorical or qualitative are gender, ethnicity, blood group etc

Numerical: We can do computations with numerical data or in simple words, such a variable takes on numerical values. Examples of such variables are age, height, cholesterol level etc.

Later we shall subdivide the above categories to discuss variables that can be ordinal, nominal, interval and ratio.

Organization and Display of Data

The following data were obtained to study the response to the drug Captopril on the patients with hypertension. The systolic and diastolic blood pressures of 15 patients were recorded immediately before and after taking the drug. The units are mmHg. The data are from (1: MacGregor, G.A., Markandu, N.D., Roulston, J. E. , and Jones, J.C. (1979) Essential Hypertension: Effect of an oral inhibitor of angiotensin-converting enzyme. *British Medical Journal* , **2**, 1106-1109.)

Variable	Description
Patient Number	The number assigned to a patient
SB	Systolic Blood Pressure before taking Captopril
SA	Systolic Blood Pressure 2-hr after taking Captopril
DB	Diastolic Blood Pressure before taking Captopril
DA	Diastolic Blood Pressure 2-hr after taking Captopril
RS	Reduction in systolic blood pressure
RD	Reduction in diastolic blood pressure

Patient Number	SB mmHg	SA mmHg	DB mmHg	DA mmHg
1	210	201	130	125
2	169	165	122	121
3	187	166	124	121
4	160	157	104	106
5	167	147	112	101
6	176	145	101	85
7	185	168	121	98
8	206	180	124	105
9	173	147	115	103
10	146	136	102	98
11	174	151	98	90
12	201	168	119	98
13	198	179	106	110
14	148	129	107	103
15	154	131	100	82

Now that we have this data, we would like to communicate with the data. First, note these data are numerical or quantitative. We are interested in studying in the reduction in both the readings on each of the subjects before and after.

Look at the reduction in each patient for both the systolic and the diastolic blood pressure

Patient Number	RS mmHg	RD mmHg
1	9	5
2	4	1
3	21	3
4	3	-2
5	20	11
6	31	16
7	17	23
8	26	19
9	26	12
10	10	4
11	23	8

```
  12              33      21
  13              19      -4
  14              19       4
  15              23      18
```

First, let us start by taking at univariate study in which we study only one variable. We shall take the variable RS, which is reduction systolic blood pressure. The values of the variable RS are listed below again.

```
RS in mmHg
      9      4     21      3     20     31     17     26     26     10
     23     33     19     19     23
```

We can get a better sense of these values if order the values

```
RS in mmHg  (sorted)


      3      4      9     10     17     19     19     20     21     23
     23     26     26     31     33
```

The display shown below is called an ordered Stem-Leaf plot of this data.

Stem	Leaf	
3*	1 3	Shows the values 31,33
2*	0 1 3 3 6 6	Shows the values 20,21,23,23,26,26
1*	0 7 9	Shows the values 10,17,19
0*	3 4 9	Shows the values 3,4,9

If a branch becomes too large, we can split the stems. In the above example, we may split each stem in two equal parts as shown below

3* 1 3
2* 6 6
2* 0 1 3 3
1* 7 9
1* 0
0* 9
0* 3 4

For a small data set, stem plot has the following properties that make it useful

1. It takes each entry into account
2. Gives an idea of the shape and the locations
3. Helps to identify outliers

When analyzing data later, we shall take up interpretations of the stem plots as well as numerical methods to identify the outliers.

But for a large data set, it is not very convenient to look at a stem plot. For a numerical data set, it is convenient to organize the data in a frequency table and then display the table by using a histogram.

The next data that we are going to consider is the data file "laheart" that is available at the University of Massachusetts website at http://www.umass.edu/statdata/statdata/data/index.html

These data are a subset of the data from an epidemiological heart disease study on Los Angeles County employees.

Let us look at the variable "Age_50" that ranges over the ages of 200 employees in the data. The ages are the ages in years recorded in the year 1950.

The raw data (sorted) is shown below

Age_50 Sorted

20	22	23	24	25	25	25	26	26	26
26	28	28	29	29	29	30	30	30	30
30	30	30	32	32	33	33	33	34	34
34	34	34	34	34	34	35	35	35	35
36	36	36	36	36	37	37	37	38	38
38	39	39	39	40	40	40	41	41	41
42	42	42	42	42	42	42	42	43	43
43	43	43	43	44	44	44	44	44	44
45	45	45	45	45	45	45	45	46	46
46	46	46	46	46	46	47	47	47	47
47	47	47	47	47	48	48	48	48	48
48	48	48	48	49	49	49	49	49	49
49	49	50	50	50	50	50	50	51	51
51	51	51	51	51	52	52	52	52	52
53	53	53	53	53	53	53	53	53	54
54	54	54	55	55	55	55	55	56	56
56	57	57	57	57	57	57	57	57	57
57	58	58	58	59	59	59	60	60	60
60	61	61	61	61	62	62	62	63	63
63	63	64	64	64	64	65	65	68	69

Note that the ages range from 20 years to 69 years. We are going to

organize this data in 14-different classes of width of 5-each. The classes must be disjoint and must cover the entire data set. If we are going to create a histogram, the classes must be equally wide.

If we count the number of entries (the data values) in each class,

note that

there are four entries 20,22,23 ,24 in the range 20≤age<25

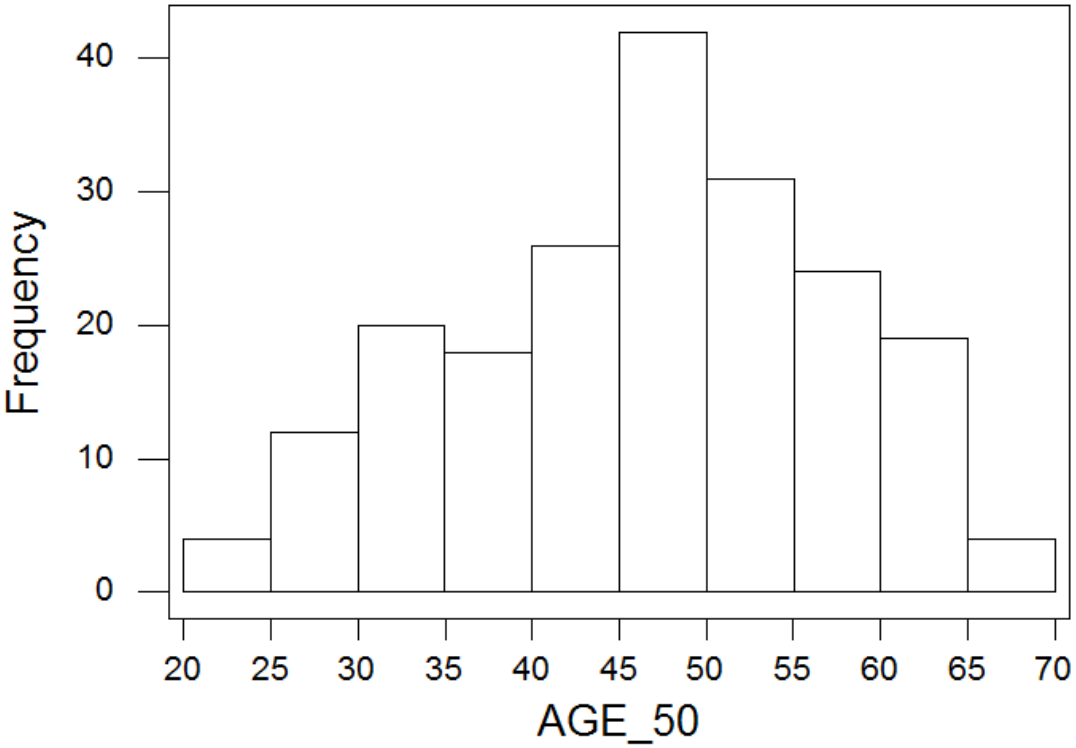
there are twelve entries 25 ,25, 25,26,26,26,26,28,28,29,29,29 in the range 25≤age<30

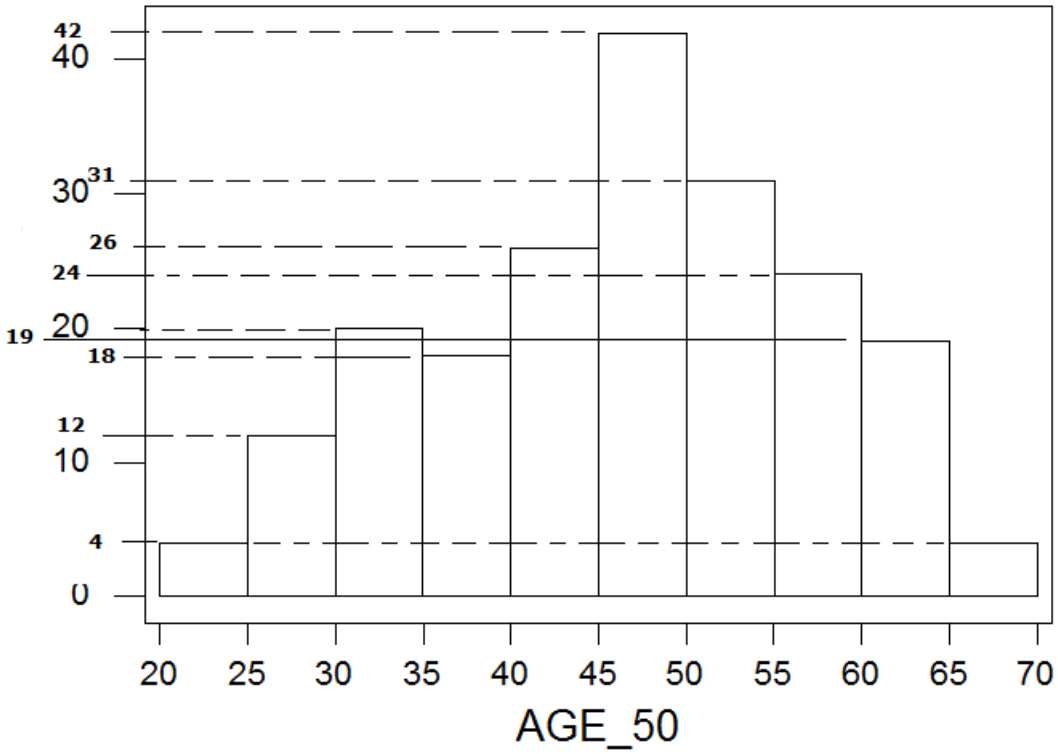
continuing in this manner

we obtain the table

Class	frequency
20≤age<25	4
25≤age<30	12
30≤age<35	20
35≤age<40	18
40≤age<45	26
45≤age<50	42
50≤age<55	31
55≤age<60	24
60≤age<65	19
65≤age<70	4
	Total=200

To make a histogram to display the above table, we can indicate the classes on the horizontal scale and the frequencies on the vertical scale and draw rectangles with altitude as the frequency of the class in a manner shown below. Note that we do not leave gaps between the bars of a histogram unless there is a class with zero frequency.


Frequency
0
10
20
30
40
20 25 30 35 40 45 50 55 60 65 70
AGE_50


42
40
31
30
26
24
19
20
18
12
10
4
0
20 25 30 35 40 45 50 55 60 65 70
AGE_50

Of course, in practice, we shall use a computer package for making a histogram and usually the program will pick a suitable number of equally wide intervals. Please refer to the link **Instructions to use JMP7** to construct a histogram with JMP7.

A histogram gives us an idea about the shape of the distribution. We may indicate approximate values of the percentiles by using a histogram of a large data set.

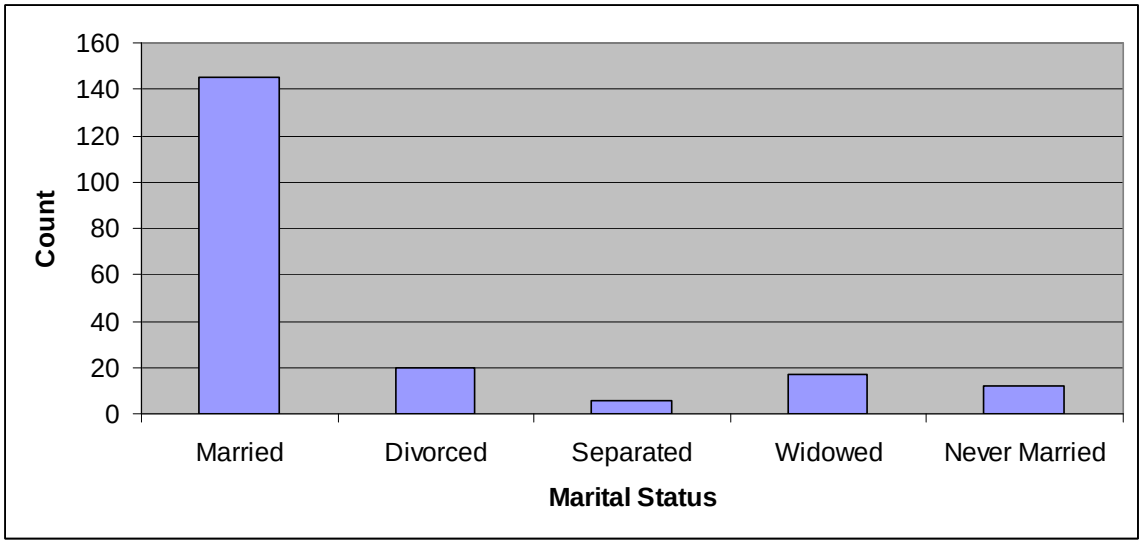
Displaying Categorical data

Let us look the variable, Marital Status of 200 subjects in the Benign Breast Disease data from http://www.umass.edu/statdata/statdata/data/index.html

Tallying the data gives us the following for the Marital status

Marital Status	Count
Married	145
Divorced	20
Separated	6
Widowed	17
Never Married	12

Such a table may be represented by a bar graph (generally with bars separated) as shown below



Or look at the data regarding AIDS Clinical Trials Group Study 320 Data (actg320.dat) from the same website http://www.umass.edu/statdata/statdata/data/index.html

Another way to display a categorical data is by using a pie chart.

Let the following table show the expenses of a certain research lab

Category	Percentage of the budget
Equipment	29
Salaries	37
Supplies	16
Building	15
Others	3

When we have "others" as a category and also for a display of a variable showing the shares of various variables, a pie chart will communicate better.