Chapters 2. Graphics and Simple Numerical Techniques (A)

January 17, 2011

Two approaches to describe the main features of the data (population or sample)

- **Graphical Techniques** (1) Pareto diagram; (2) dot diagram; (3) histogram (3) ...
- **Numerical Techniques** (1) central tendency; (2) dispersion; (3) association

1 Graphics and their applications

Graphical statistical methods have several objectives:

- to find structure in data (including the the distribution);
- to check assumptions in statistical models;
- to communicate the results of an analysis.

For categorical variable (data): Pareto diagram

For numerical variable (data): dot diagram, histogram, scatter plot, ...

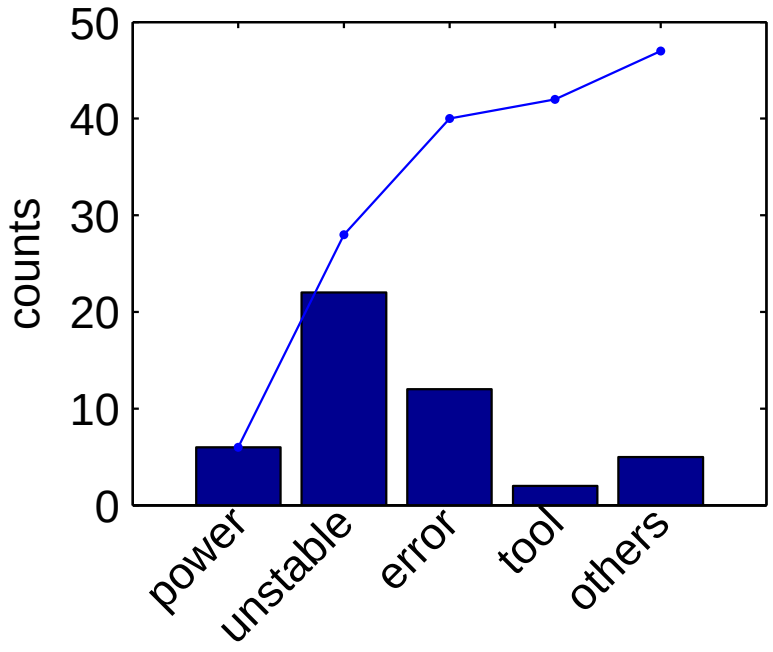
1.1 Pareto diagram (for categorical variable/data)

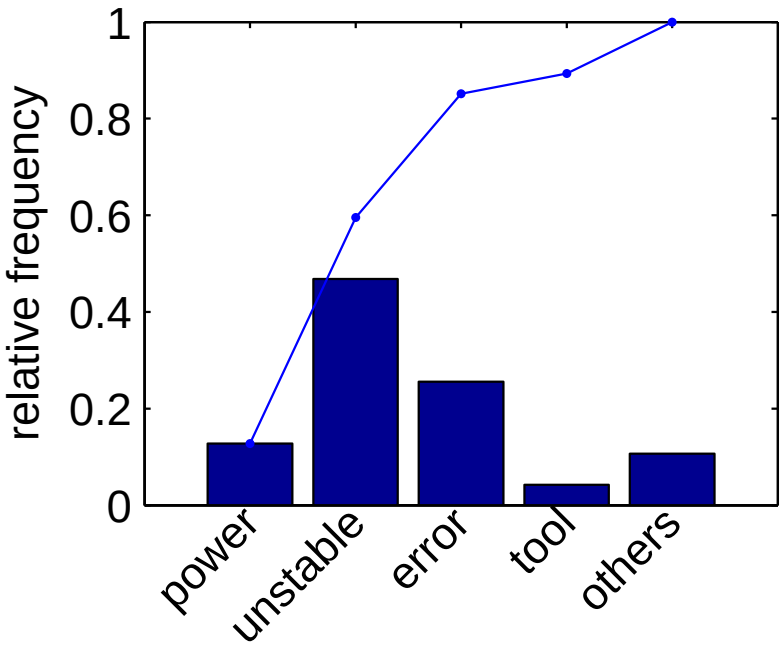
For categorical variable, the only allowable calculation is to count the frequency in each category.

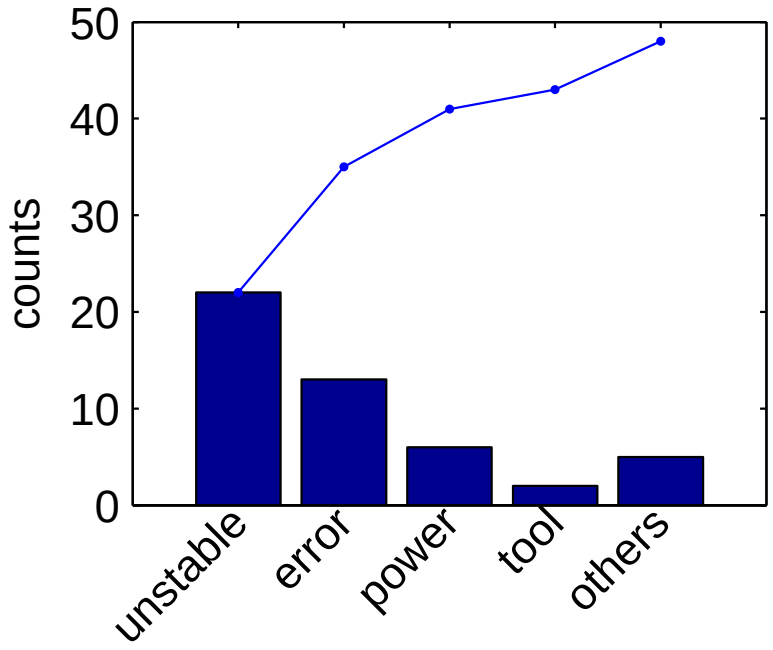
- each vertical bar represents a category with height being the frequency.
- If we arrange the bars in a descending order of the frequency and with cumulative line on the same diagram, we call the graph a Pareto diagram.
- Pareto diagram can separate the "critical few" from "trivia many".

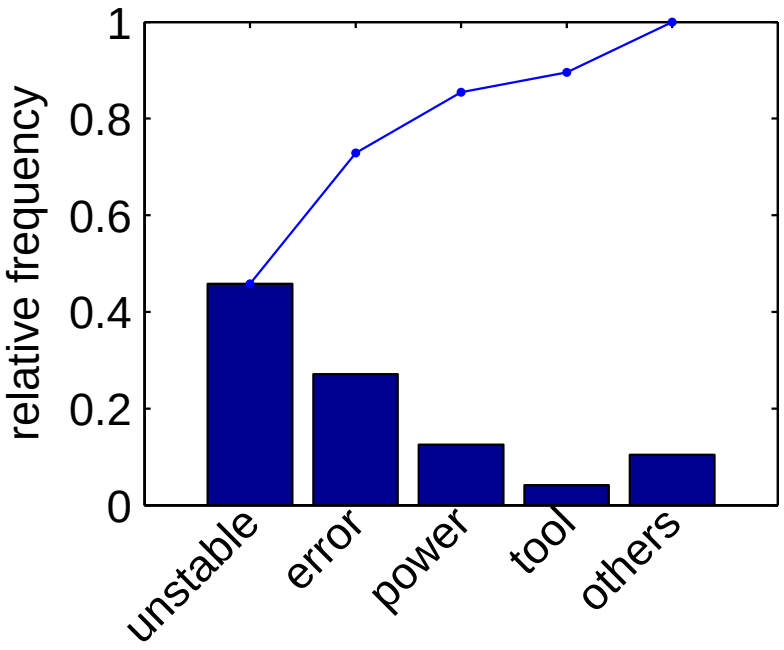
Example When a company identifies a process as a candidate for improvement, the first step is to collect data on the frequency of each type of failure.

types of failure	frequency	cumulative frequency	relative frequency	cumulative relative frequency
power fluctuations	6	6	0.1250	0.1250
controller not stable	22	28	0.4583	0.5833
operator error	13	41	0.2708	0.8542
worn tool not replaced	2	43	0.0417	0.8958
other	5	48	0.1042	1.0000


50
40
30
20
10
0
counts
power
unstable
error
tool
others


1
0.8
0.6
0.4
0.2
0
relative frequency
power
unstable
error
tool
others


50
40
30
20
10
0
counts
unstable
error
power
tool
others


1
0.8
0.6
0.4
0.2
0
relative frequency
unstable
error
power
tool
others

In R;

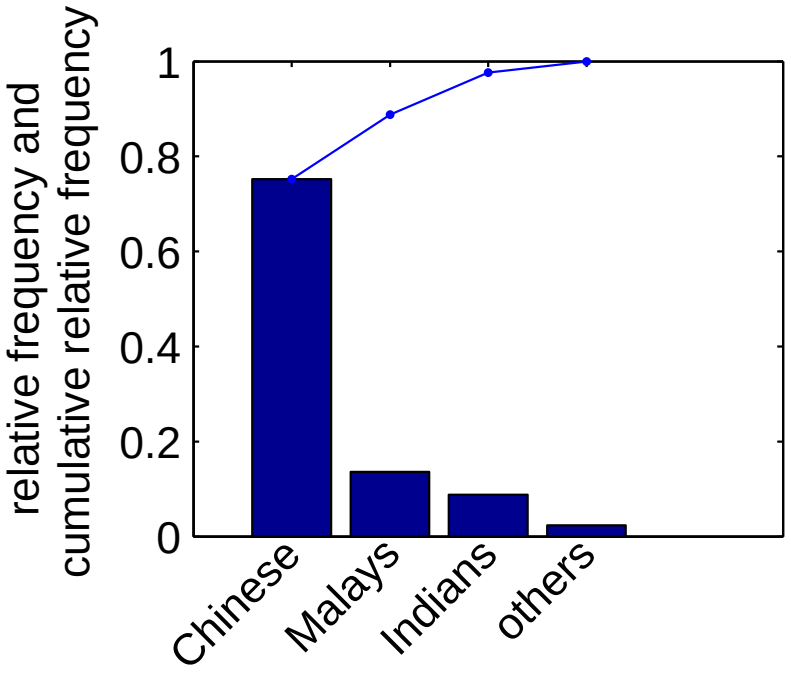
```
barplot(c(power = 6, unstable = 22, error = 13, tool=2, others
= 5), space=0)
```

or

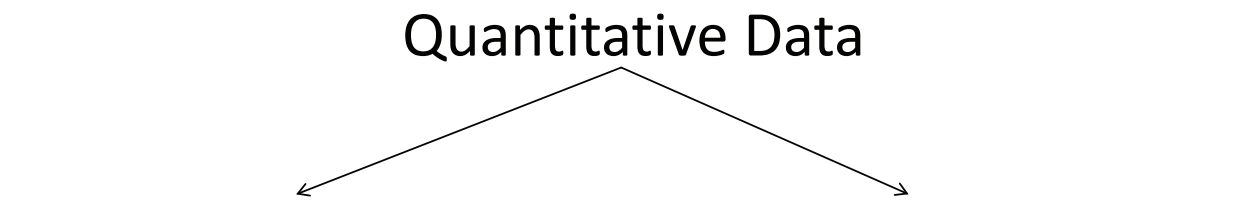
```
barplot(c(power = 6, unstable = 22, error = 13, tool=2, others
= 5), ylim=c(0, 50), space=0)


lines(c(power = 6, unstable = 28, error = 41, tool=43, others
= 48))
```

Example B (continued)

Ethnic	Relative Frequency	cumulative frelative Frequency
Chinese	75.2%	75.2%
Malays	13.6%	88.8%
Indians	8.8%	97.6%
Eurasians and others	2.4%	100%



Graphs for Quantitative Variable/Data



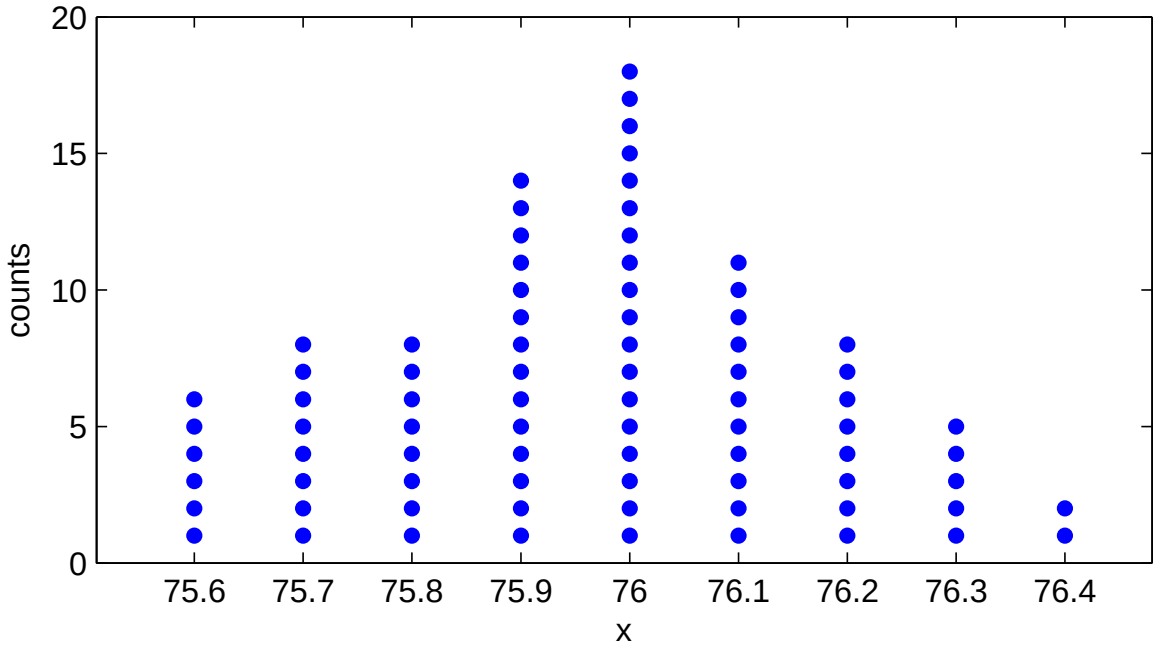
Ungrouped data (raw): **Dot-diagram, Scatter plot.**

Grouped data with frequency: **Histogram, Cumulative frequency plot.**

1.2 Dot diagrams

Dot diagram represents each individual by a dot and pile the individuals with the same value. It is a good summary of information when data is not large.

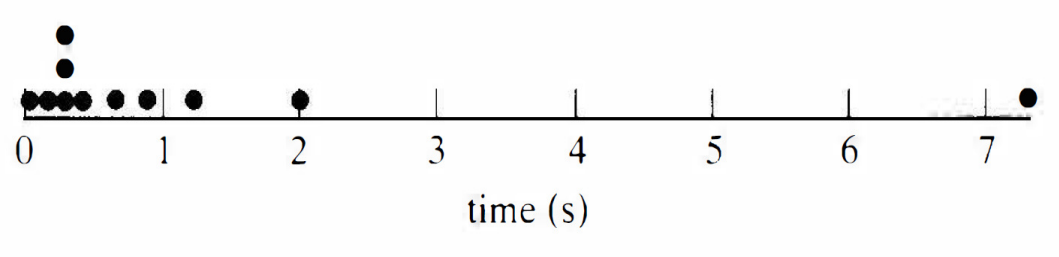
Example A (continued) (see also the figure above) The dot-diagram is



Dot-diagrams expose outliers—— observations that are numerically distant from the rest of the data.

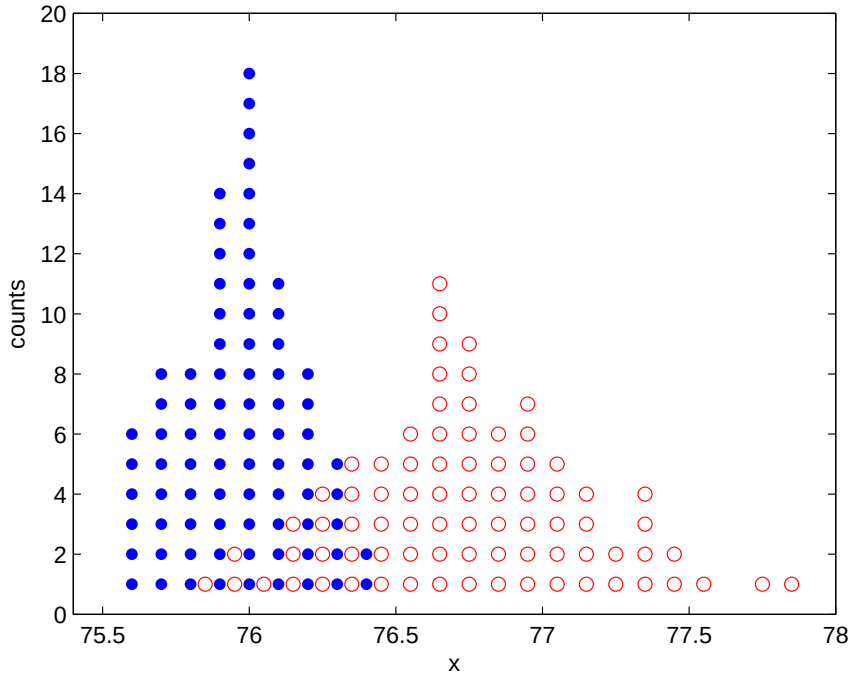
The following times (in seconds) between 2 neutrinos occurred outside of our solar system were recorded:

0.107, 0.196, 0.021, 0.283, 0.179, 0.854, 0.580, 0.19, 7.3, 1.18, 2.0

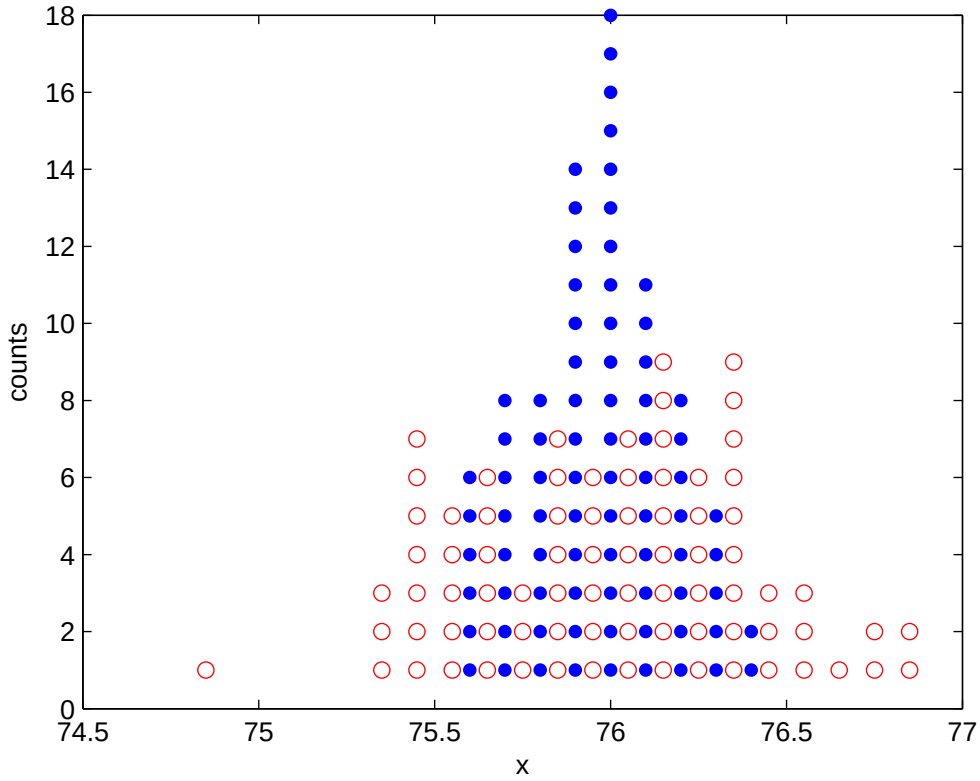


observation 7.3 is possibly an outlier, which means something unusual happened.

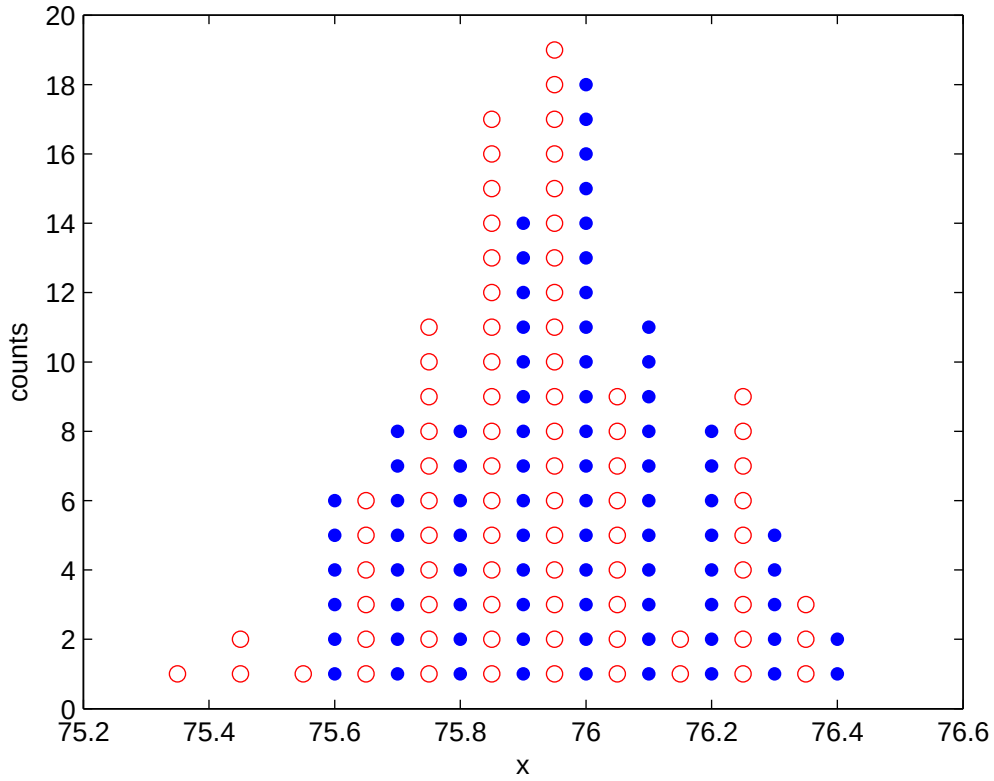
A dot-diagram for multiple samples reveals differences Each panel below has two data, denoted by "·" and "o" respectively. tell their difference.



We conclude that data marked by "o" is (possibly) statistically bigger than the blue data marked by "·".



We conclude that red data has (possibly) wider spread than the blue data statistically.



We conclude that both data have similar distributions.

1.3 Histogram and distribution

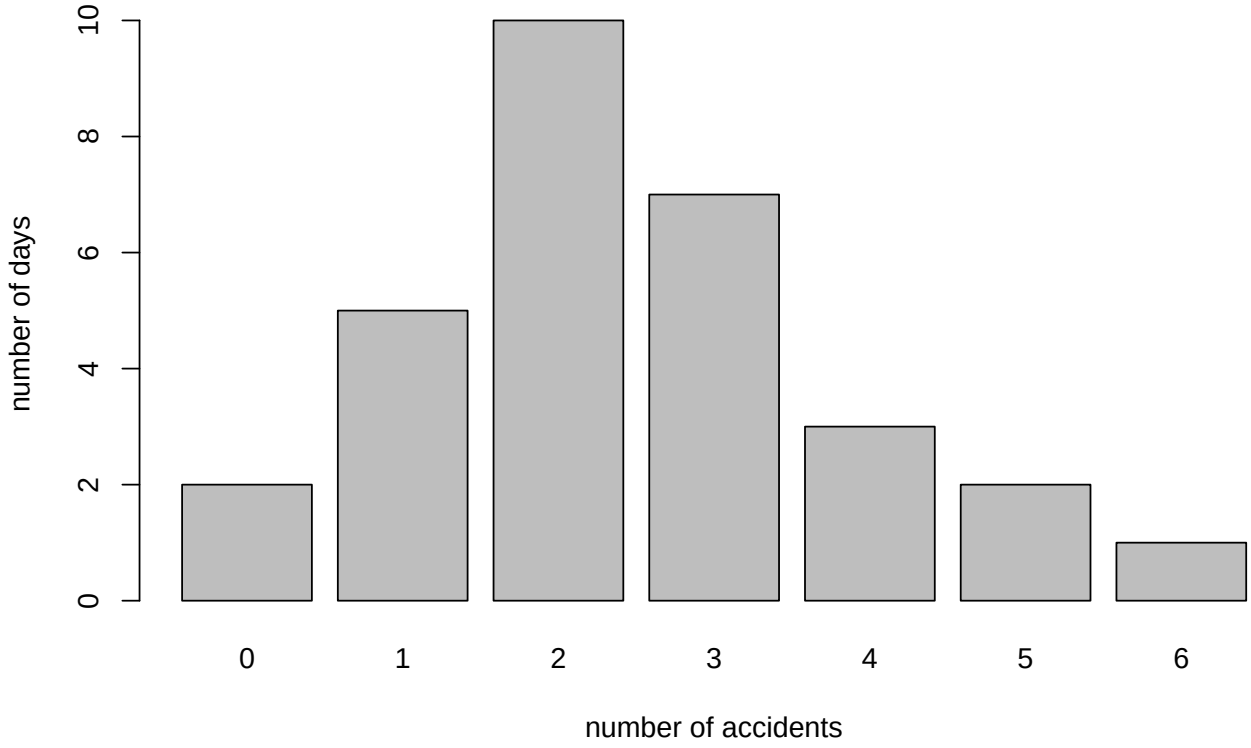
The histogram is the most common form of graphical presentation of a frequency distribution of a data (population or sample).

The histogram is a bar plot, each bar represents a group/set of data (in an interval) with its height equaling the frequency of the set.

If the variable is discrete, we can do it the same way as the Pareto diagram.

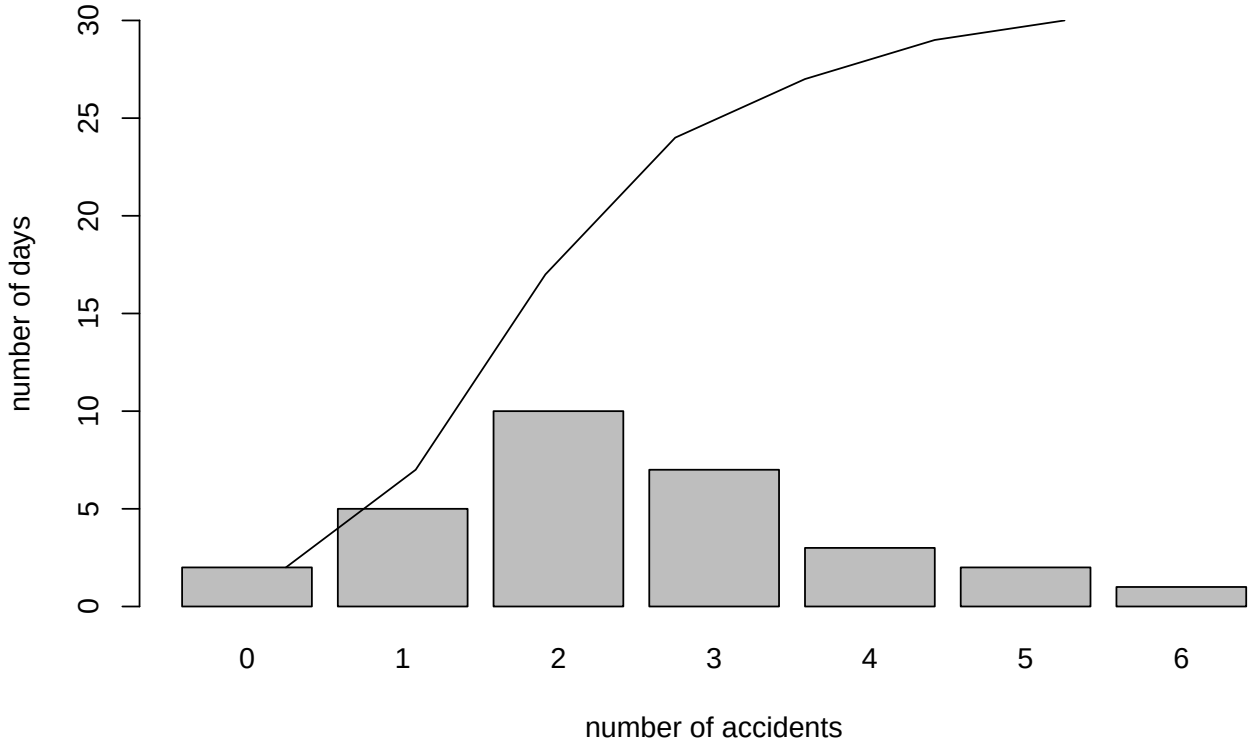
Example Daily numbers of traffic accidents and their distributions in a month.

numer of accidents	days	cumulative frelative Frequency
0	2	2
1	5	7
2	10	17
3	7	24
4	3	27
5	2	29
6	1	30



R code:

```
accidents = c(2, 5, 10, 7, 3, 2, 1)
barplot(accidents, names.arg=c("0", "1", "2", "3", "4", "5","6"), ylab="number of days", xlab="number of accidents")
```



R code:

```
accidents = c(2, 5, 10, 7, 3, 2, 1)
barplot(accidents, names.arg=c("0", "1", "2", "3", "4", "5","6"), ylab="number of days", xlab="number of accidents",
ylim=c(0, 30))
lines(cumsum(accidents))
```

Building a Histogram with continuous variable by grouping them

1. Collect the Data ?

2. Create a frequency distribution for the data?

   (a) Determine the number of classes/groups to use. Usually, the number of classes is the integer part of

   $$\sqrt{n} \text{ (or } \sqrt{N})$$

   (b) Determine how large to make each class.

      i. Look at the range of the data, that is,

      $$\text{Range} = \text{Largest Observation - Smallest Observation}$$

ii. Then each class width becomes: width$_i$ = Range / (number of classes)

(c) establish class limits (class boundaries)

(d) Place the data into each class and count the frequency in each class

3. Draw Histogram (bars with heights being one of the following 3 cases).

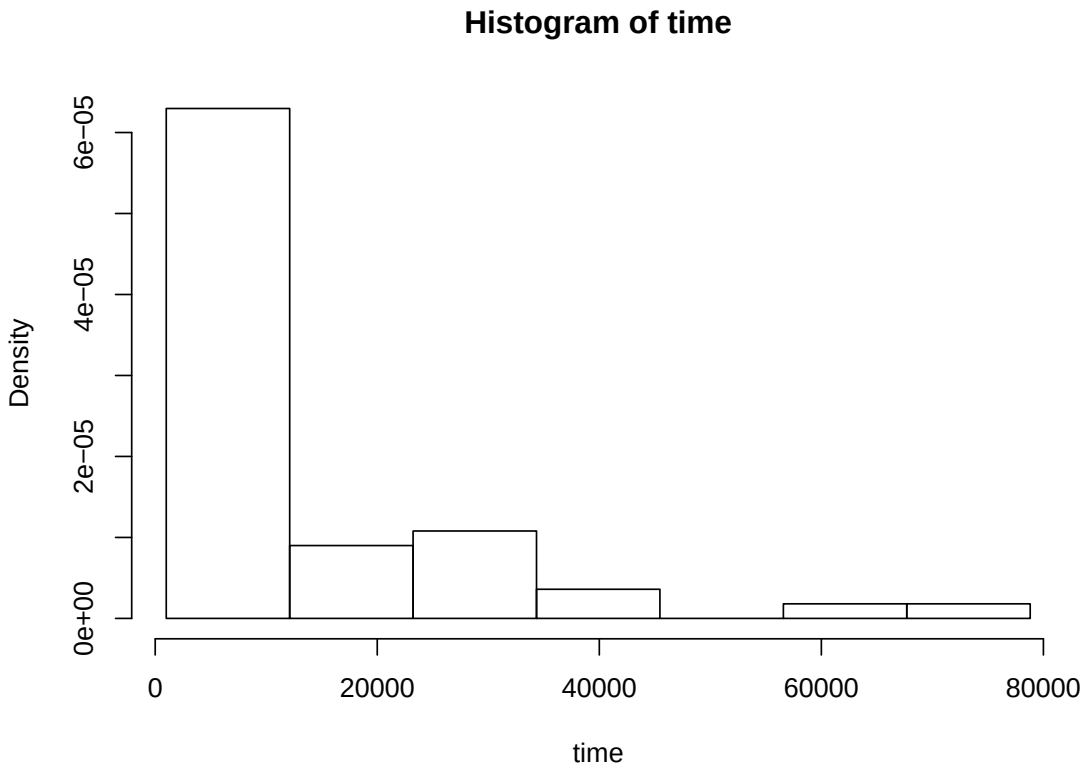
- the height of i'th bar is the frequency f_i or N_i
- the height of i'th bar is the relative frequency p_i (sometimes called density histogram), in this case the sum of all heights is 1
- the height of i'th bar is $p_i/($width$_i$ of the bar) (also called density histogram). In this case the sum of the areas of all bars is 1 because the area of i'th bar is p_i

Example E A computer scientist, trying to optimize system performance, collected data on the time, in microseconds, between requests for a particular process service, 2808, 4201, 3848, 9112, 2082, 5913, 1620, 67192, 1657, 3072, 2949, 11768, 4731, 14211, 1583, 9853, 78811, 6655, 1803, 7012, 1892, 4227, 6583, 15147, 4740, 8528, 10563, 43003, 16723, 2613, 26463, 34867, 4191, 4030, 2472, 28840, 24487, 14001, 15241, 1643, 5732, 5419, 28608, 2487, 995, 3116, 29508, 11440, 28336, 3440. Draw the histogram.

Solution: number of classes is 7 because $\sqrt{50} = 7.07$. The class width is $(78811 - 995)/7 = 11116.57$. Thus we have approximately the classes and their frequencies below

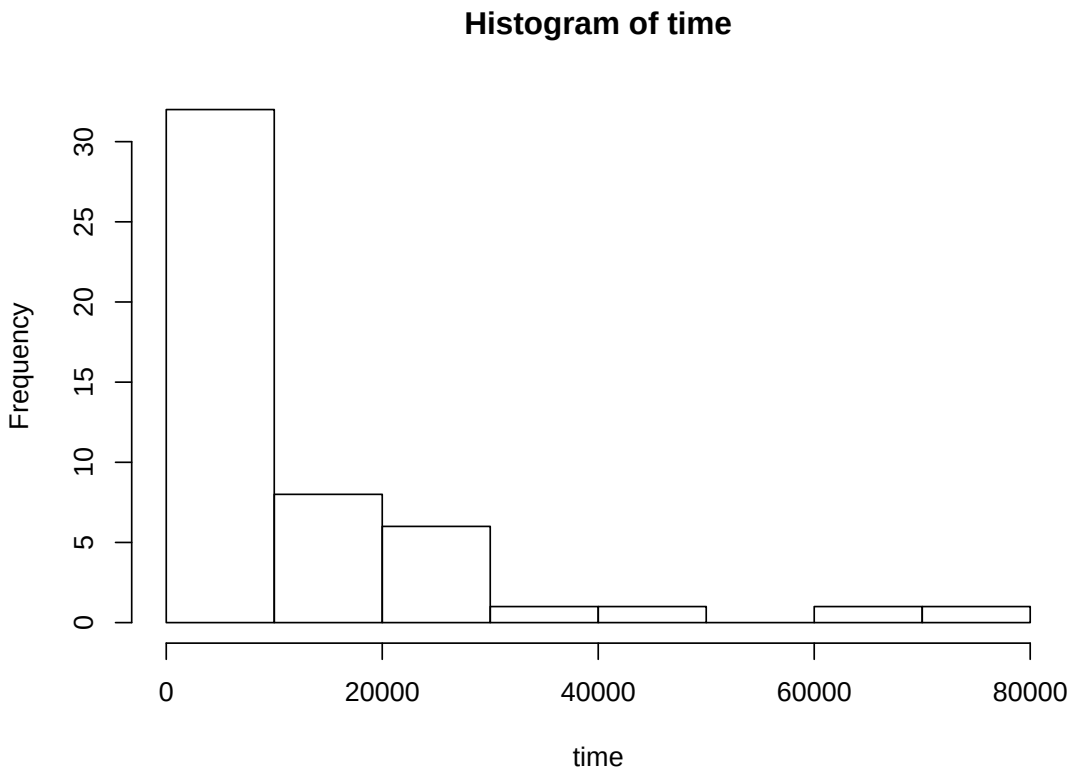
group/class	frequency	relative frequency
[995 12112]	35	0.7
(12112 23228]	5	0.1
(23228 34345]	6	0.12
(34345 45461]	2	0.04
(45461 56578]	0	0.0
(56578 67694]	1	0.02
(67694 78811]	1	0.02

please note the endpoint of the intervals!




```
time = c(2808, 4201, 3848, 9112, 2082, 5913, 1620, 67192, 1657, 3072, 2949, 11768, 4731, 14211, 1583, 9853, 78811,
6655, 1803, 7012, 1892, 4227, 6583, 15147, 4740, 8528, 10563, 43003, 16723, 2613, 26463, 34867, 4191, 4030, 2472,
28840, 24487, 14001, 15241, 1643, 5732, 5419, 28608, 2487, 995, 3116, 29508, 11440, 28336, 3440)

hist(time, breaks=c(995, 12112, 23228, 34345, 45461, 56578, 67694, 78811))
```

Or simply




```
time = c(2808, 4201, 3848, 9112, 2082, 5913, 1620, 67192, 1657, 3072, 2949, 11768, 4731, 14211, 1583, 9853, 78811,
6655, 1803, 7012, 1892, 4227, 6583, 15147, 4740, 8528, 10563, 43003, 16723, 2613, 26463, 34867, 4191, 4030, 2472,
28840, 24487, 14001, 15241, 1643, 5732, 5419, 28608, 2487, 995, 3116, 29508, 11440, 28336, 3440)
hist(time, nclass=7)
```

Example F 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 69.1 69.2 69.3 69.3 69.5 69.5 69.6 69.7 69.8 69.8 69.9 70.0 70.0 70.1 70.2 70.3 70.3 70.4 70.5 70.6 70.6 70.8 70.9 71.0 71.1 71.2 71.3 71.3 71.5 71.6 71.6 71.7 71.8 71.8 71.9 72.1 72.2 72.3 72.4 72.6 72.7 72.9 73.1 73.3 73.5 74.2 74.5 75.3. Draw a density histogram with 8 classes.

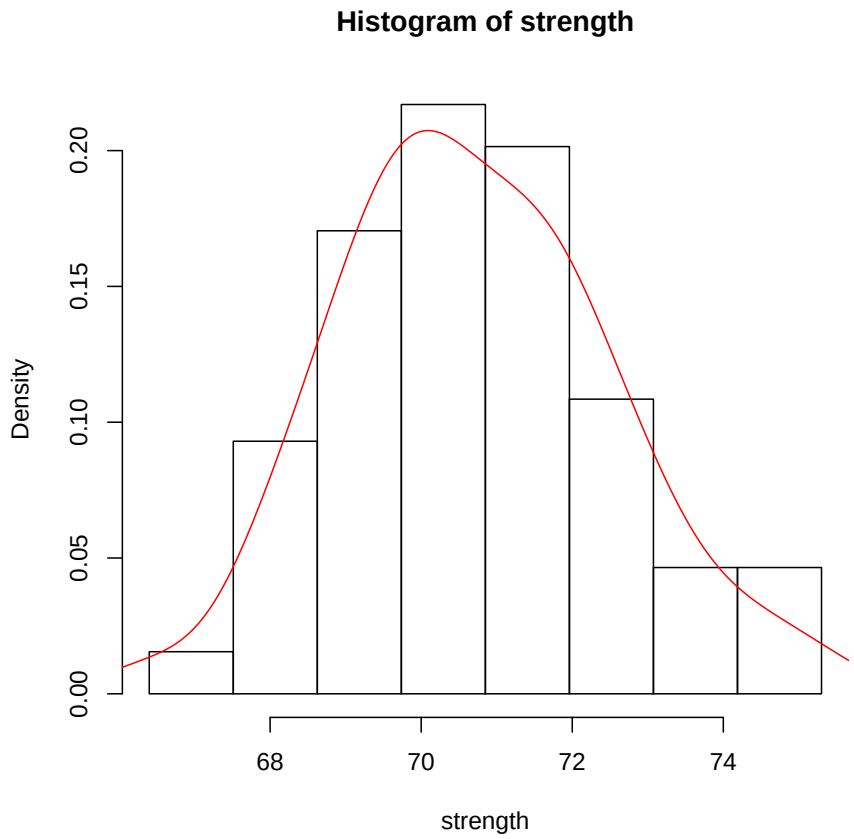
Solution:

width of each bar $= (max - min)/8 = (75.3 - 66.4)/8 = 1.2714$

in the classes, the frequencies are

group/class	frequency	relative frequency
[66.4000 67.5125]	1	0.0172
(67.5125 68.6250]	6	0.1034
(68.6250 69.7375]	11	0.1897
(69.7375 70.8500]	14	0.2414
(70.8500 71.9625]	13	0.2241
(71.9625 73.0750]	7	0.1207
(73.0750 74.1875]	3	0.0517
(74.1875 75.3000]	3	0.0517

Sometimes, people prefer to have a smooth histogram/density function (see the red line, which will be discussed later)



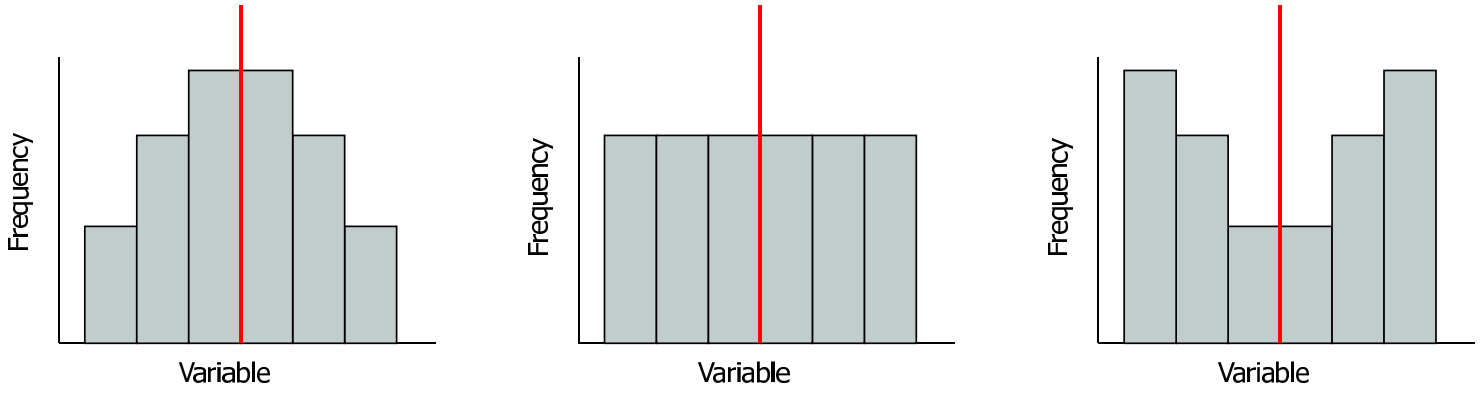
R code:

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

width = (max(strength)-min(strength))/8
cc = (0:8)*width + min(strength);
hist(strength, breaks = cc, freq=FALSE)
  # 'freq=FALSE' to draw relative frequency/density histogram


lines(density(strength), col="red") # to get a smooth version of density histogram
```

1.4 Shapes of Histogram

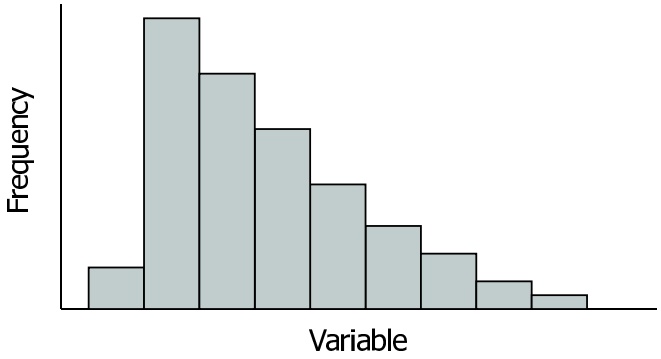
Symmetry

A histogram is said to be ***symmetric*** if, when we draw a **vertical line** down the center of the histogram, the two sides are identical in shape and size:

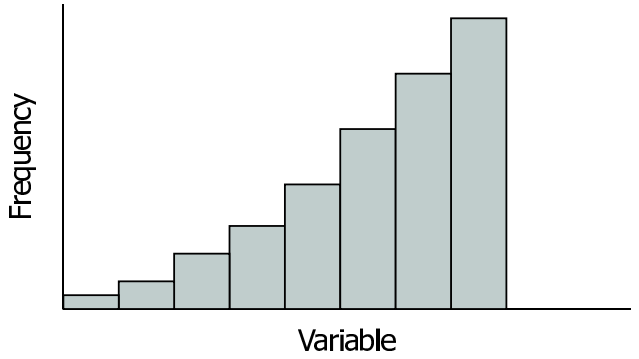


Skewness

A skewed histogram is one with a long tail extending to either the right or the left:



Positively Skewed

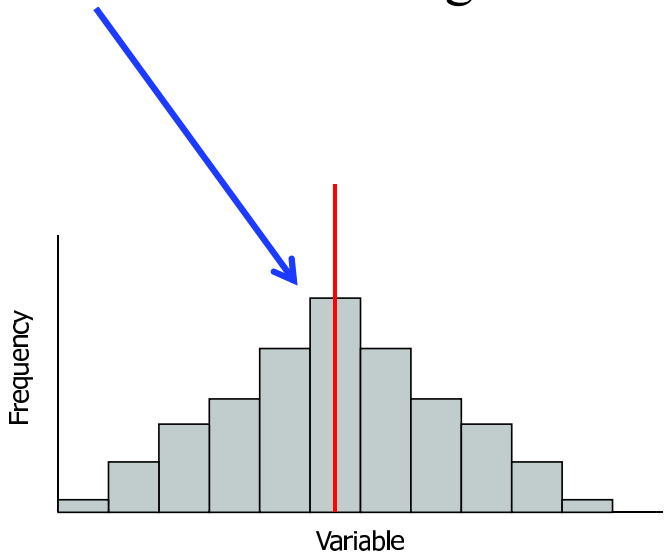


Negatively Skewed

Negative skew/(skewed to the left) if left tail is longer; positive skew/(skewed to the right) if right tail is longer

Bell Shape

A special type of ***symmetric unimodal*** histogram is one that is bell shaped:



Bell Shaped

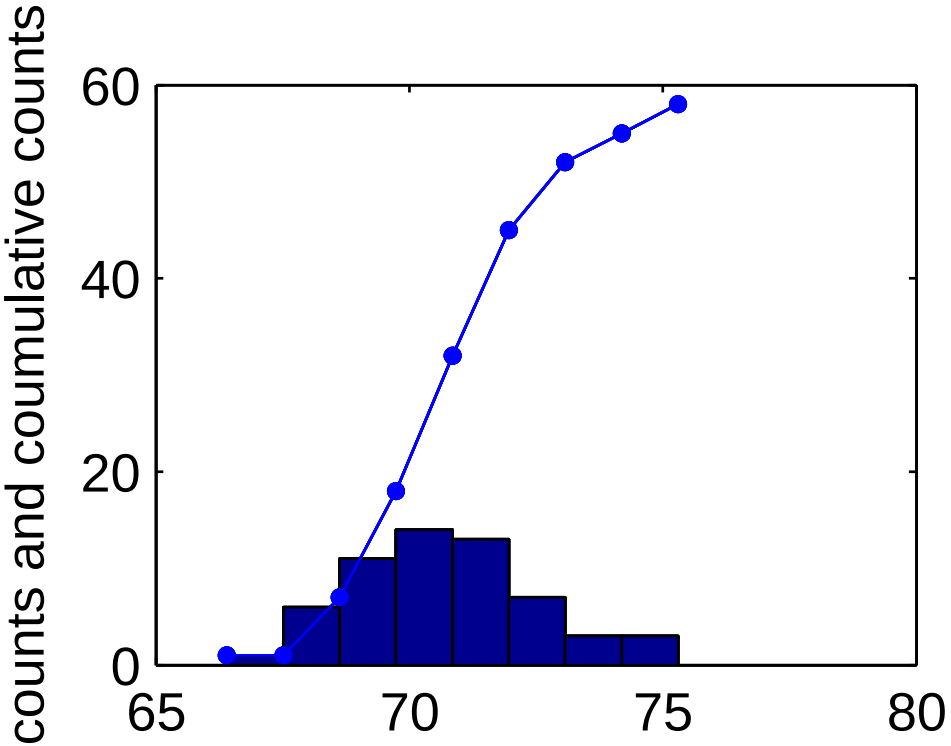
Many statistical techniques require that the population be bell shaped.

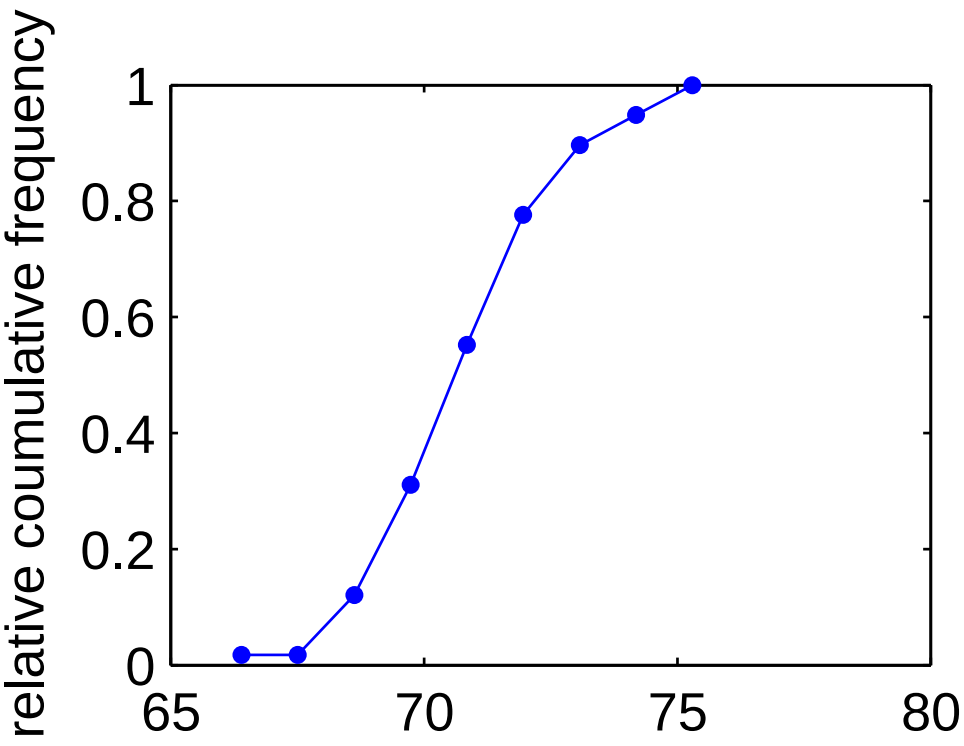
Drawing the histogram helps verify the shape of the population in question.

Cumulative frequency Cumulative distributions are usually presented graphically in the form of ogives, where we plot the cumulative frequencies at the class boundaries.

Example F (continued) We calculation the cumulative

group/class	frequency	cumulative frequency	relative frequency	cumulative relative frequency
$\leq$ 66.400	1	1	0.0172	0.0172
(66.4000 67.5125]	0	1	0	0.0172
(67.5125 68.6250]	6	7	0.1034	0.1207
(68.6250 69.7375]	11	18	0.1897	0.3103
(69.7375 70.8500]	14	32	0.2414	0.5517
(70.8500 71.9625]	13	45	0.2241	0.7759
(71.9625 73.0750]	7	52	0.1207	0.8966
(73.0750 74.1875]	3	55	0.0517	0.9483
(74.1875 75.3000]	3	58	0.0517	1.0000

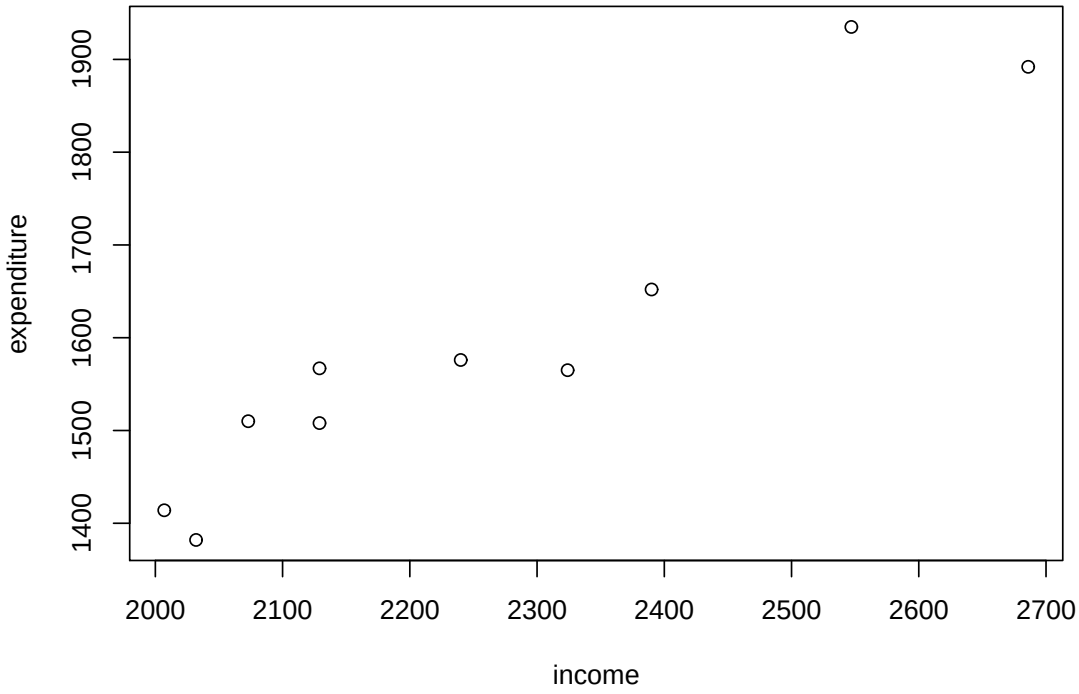

counts and coumulative counts
60
40
20
0
65
70
75
80


relative coumulative frequency
1
0.8
0.6
0.4
0.2
0
65
70
75
80

1.5 Scatter plot

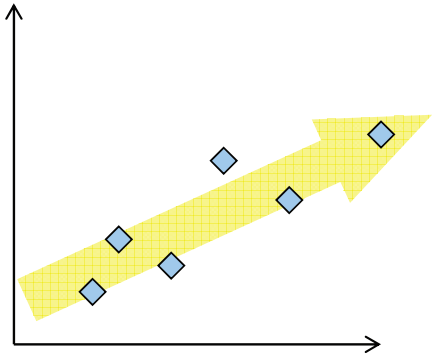
If each individual has two numerical variables, with value (x_i, y_i), we can represent the individual in the 2-dimensional plan by a dot (or any other symbol).

Example incomes and expenditures of 10 people: (2007, 1414), (2240, 1576), (2324, 1565), (2032, 1382), (2129, 1508), (2073, 1510), (2547, 1935), (2686, 1892), (2129, 1567), (2390, 1652)

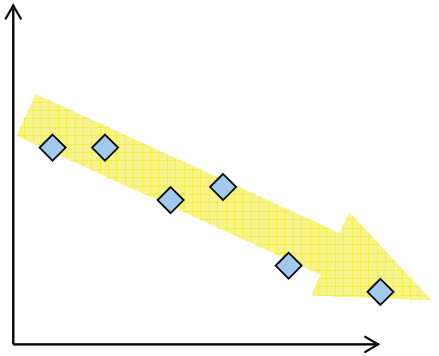



```
R code
income = c(2007, 2240, 2324, 2032, 2129, 2073, 2547, 2686, 2129, 2390)
expenditure = c(1414, 1576, 1565, 1382, 1508, 1510, 1935, 1892, 1567, 1652)
plot(income, expenditure)
```

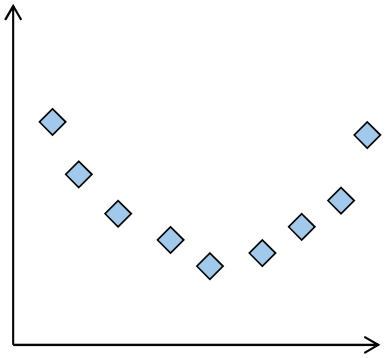
Statistical dependences are our interest



Positive Linear Relationship



Negative Linear Relationship



Weak or Non-Linear Relationship