İST 292 STATISTICS

Sections: 05-06

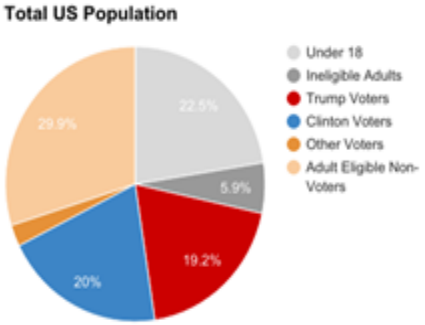
For Department of Computer Engineering

LESSON 2 DESCRIPTIVE STATISTICS

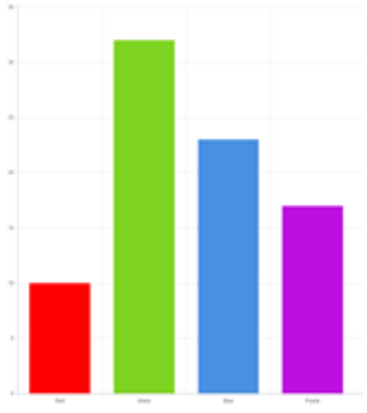
Dr. Ayten Yiğiter and Dr. Esra Polat Lecture Notes

1.DESCRIBING DATA SETS -In this section we present some common graphical and tabular ways for presenting data.

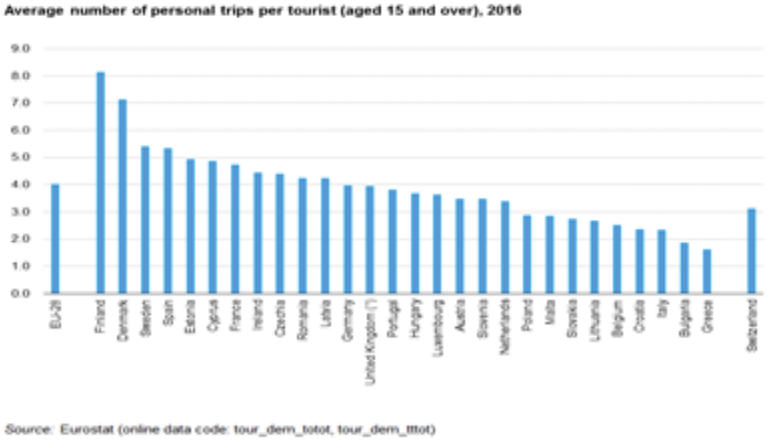
1.1. Frequency Tables and Graphs: How we display data distributions depends on the type of variable(s) or data that we are dealing with.



Categorical-pie charts:

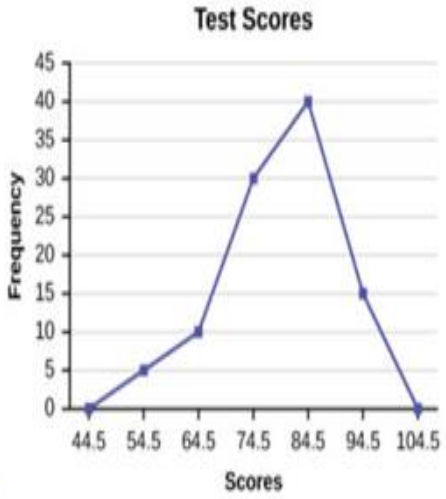


Categorical- bar charts :



Quantitative- line graphs:

Quantitative-frequency polygon:

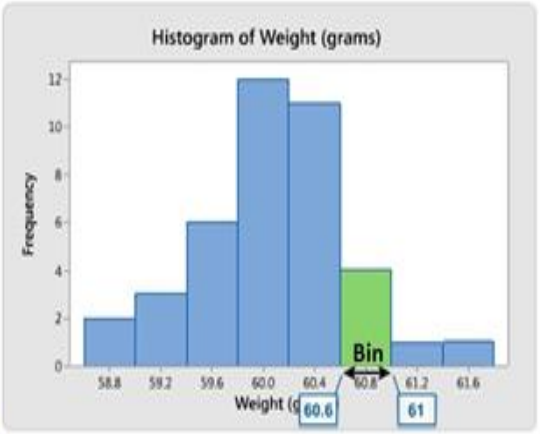


Quantitative-stemplots:

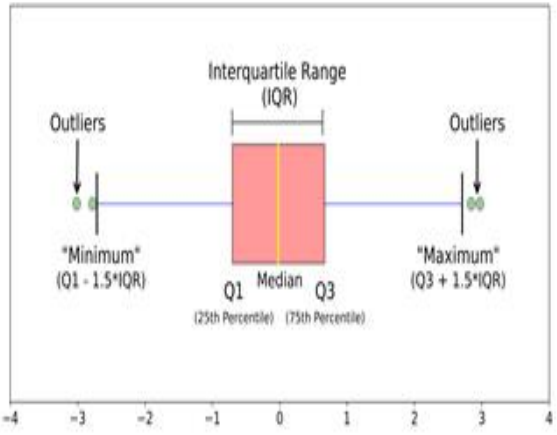
Stemplot of students weights

Stem	Leaf
4	9
5	2666
6	77
7	239
8	12779
9	1
10	3

Quantitative-histograms:



Quantitative-boxplots:



A data set having a relatively small number of distinct values can be conveniently presented in a *frequency table.* For instance, *Table1 is a frequency table* for a data set by asking 500 students how many cigarettes they smoked.

Table 1. The number of cigarettes smoked

Number of cigarettes smoked per day	frequency
0	12
5	15
8	33
10	17
12	23
15	215
20	100
25	30
30	50
40	5
Total	500

Table 1 tells us, among other things, that the *lowest number cigarettes of 0* was smoked by *12 of the students* (nonsmokers were 12), whereas the *highest number cigarettes of 40* were smoked by *5 students.* The most *common number of cigarettes were 15*, and smoked by *215 of the students.*

A)Line Graph: Data from a frequency table can be graphically represented by a line graph that plots the *distinct data values on the horizontal axis* and indicates *their frequencies by the heights of vertical lines.* A line graph of the data presented in Table 1 is shown in Figure 1.

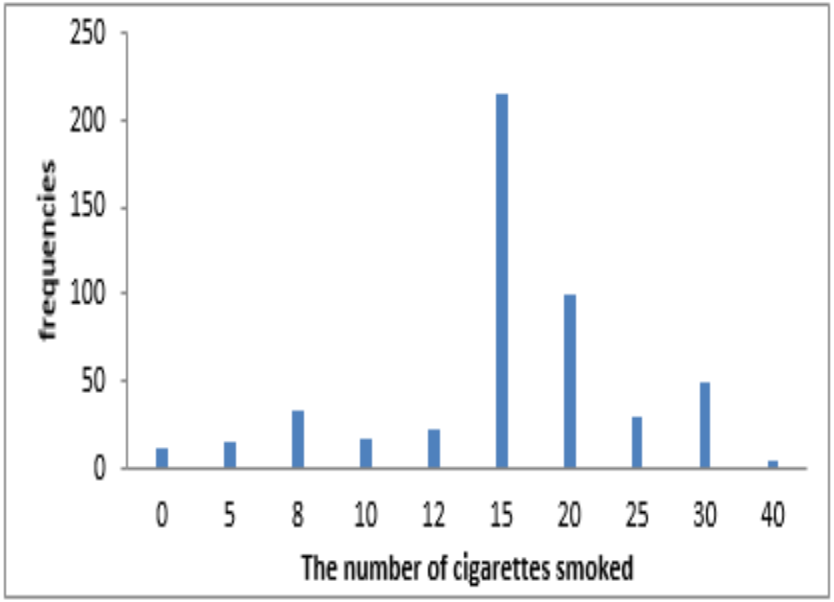


Figure 1. The number of cigarettes smoked per day.

B) Bar Graph: *When the lines in a line graph are given added thickness*, the graph is called a bar graph.

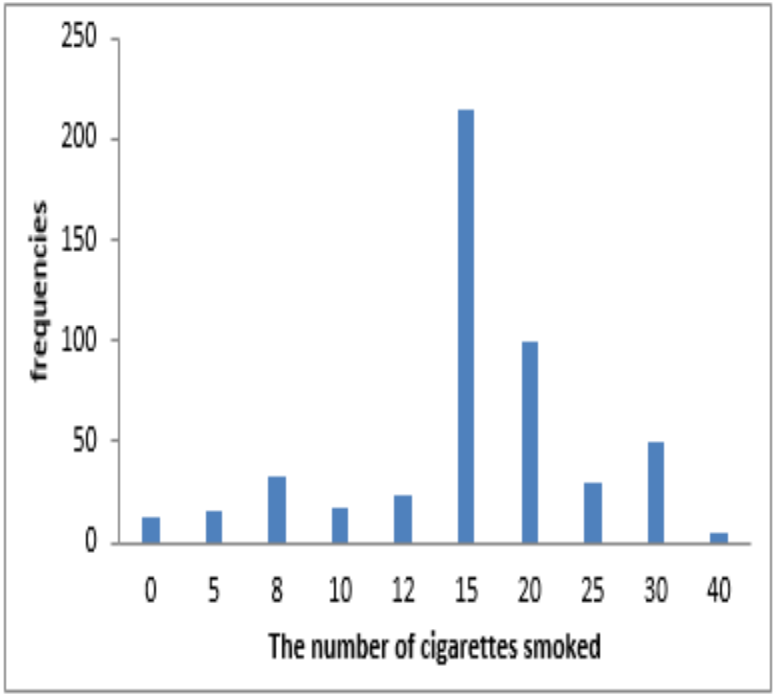


Figure 2. Bar graph for the number of cigarettes smoked per day.

Table 2 is a frequency table for a data set consisting of the starting yearly salaries (to the nearest thousand dollars) of 42 recently graduated students with B.S. degrees in electrical engineering.

Table 2. Frequency and Relative Frequency Table of Starting Yearly Salaries

Starting Salary	Frequency (f_i)	Relative Frequency $\left(p_i = \frac{f_i}{n}\right)$
47	4	4/42= 0.0952
48	1	1/42= 0.0238
49	3	3/42= 0.0714
50	5	5/42= 0.1190
51	8	8/42= 0.1905
52	10	10/42= 0.2381
53	0	0
54	5	5/42= 0.1190
56	2	2/42= 0.0476
57	3	3/42= 0.0714
60	1	1/42= 0.0238
Total	42	1

C) Frequency Polygon: It plots the *frequencies of the different data values on the vertical axis*, and then connects the plotted points with straight lines.

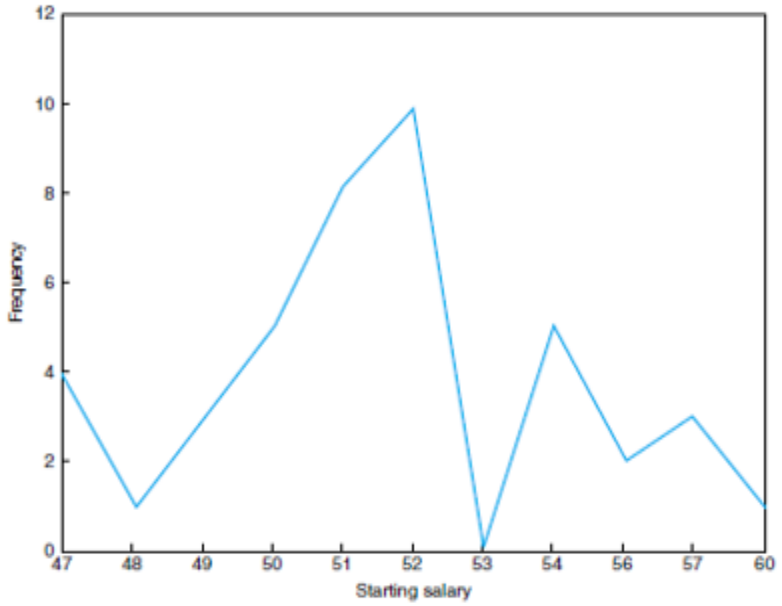


Figure 3. Frequency polygon for starting salary data.

** The *relative frequencies* can be represented graphically by *a relative frequency line* or *bar graph* or by *a relative frequency polygon.*

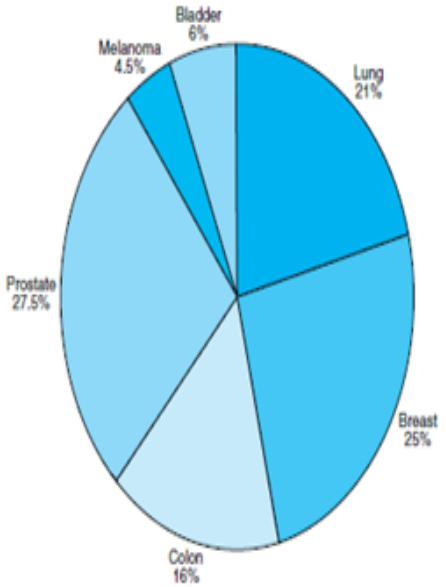
D) Pie Chart: The easiest way to portray the distribution of a categorical variable is to use a table of counts and/or percentages.

A pie chart is often used to indicate *relative frequencies when the data are not numerical in nature. A circle is constructed and then sliced into different sectors; one for each distinct type of data value.*

The relative frequency of a data value is indicated by the area of its sector (daire dilimi), this area being equal to the total area of the circle multiplied by the relative frequency of the data value. (Means that $360^0 \times p_i$)

Example: The following data relate to the different types of cancers affecting the 200 most recent patients to enroll at a clinic specializing in cancer. These data are represented in the *pie chart* presented in Figure 4. Melanoma, Bladder, Colon, Breast, Prostate. Example, 200*0.21=42 patients are Lung cancers. Example $360^0 \times 0.275 = 99^0$ is the area on pie chart for Prostate.

Type of Cancer	Number of New Cases	Relative Frequency
Lung	42	.21
Breast	50	.25
Colon	32	.16
Prostate	55	.275
Melanoma	9	.045
Bladder	12	.06



1.3.Grouped Data, Histograms, Cumulative Frequency Plot, and Stem and Leaf Plots

- For *some data sets the number of distinct values is too large*, in such cases, it is useful to divide the values into groupings, or classes, *and then plot the number of data values falling in each class.*

- Although 5 to 20 number of classes are typical, the appropriate number is a subjective choice, and of course, you can try different numbers of classes to see which of the resulting charts appears to be most revealing about the data. The number of classes is shown by *k*, if it is not given, calculated by the formula $k = 1 + 3.3 \log_{10}(n)$.

Table 4 presents the lifetimes of 200 lamps.

Table 4. Life in Hours of 200 Lamps

Item Lifetimes									
1,067	919	1,196	785	1,126	936	918	1,156	920	948
855	1,092	1,162	1,170	929	950	905	972	1,035	1,045
1,157	1,195	1,195	1,340	1,122	938	970	1,237	956	1,102
1,022	978	832	1,009	1,157	1,151	1,009	765	958	902
923	1,333	811	1,217	1,085	896	958	1,311	1,037	702
500	933	928	1,153	946	858	1,071	1,069	830	1,063
930	807	954	1,063	1,002	909	1,077	1,021	1,062	1,157
999	932	1,035	944	1,049	940	1,122	1,115	833	1,320
901	1,324	818	1,250	1,203	1,078	890	1,303	1,011	1,102
996	780	900	1,106	704	621	854	1,178	1,138	951
1,187	1,067	1,118	1,037	958	760	1,101	949	992	966
824	653	980	935	878	934	910	1,058	730	980
844	814	1,103	1,000	788	1,143	935	1,069	1,170	1,067
1,037	1,151	863	990	1,035	1,112	931	970	932	904
1,026	1,147	883	867	990	1,258	1,192	922	1,150	1,091
1,039	1,083	1,040	1,289	699	1,083	880	1,029	658	912
1,023	984	856	924	801	1,122	1,292	1,116	880	1,173
1,134	932	938	1,078	1,180	1,106	1,184	954	824	529
998	996	1,133	765	775	1,105	1,081	1,171	705	1,499
610	916	1,001	895	709	860	1,110	1,149	972	1,002

Table 5. Class Frequency Table for the data of Table 4

Group no/Class no	LL	UL	$s_i = \frac{LL_i + UL_i}{2}$	f_i	$p_i = \frac{f_i}{n}$	%
1	500	599	549.5	2	0.01	1
2	600	699	649.5	5	0.025	2.5
3	700	799	749.5	12	0.06	6
4	800	899	849.5	25	0.125	12.5
5	900	999	949.5	58	0.29	29
6	1000	1099	1049.5	41	0.205	20.5
7	1100	1199	1149.5	43	0.215	21.5
8	1200	1299	1249.5	7	0.035	3.5
9	1300	1399	1349.5	6	0.03	3
10	1400	1499	1449.5	1	0.005	0.5
				200	1	100

*The class intervals are of length 100 (c=100), with the first one starting at 500.

$$Range = Maximum - Minimum = 1499 - 500 = 999$$

$$c = \frac{999 + 1}{10} = 100$$

*Example $s_1 = \frac{LL_1 + UL_1}{2} = \frac{500 + 599}{2} = 549.5$

E) Histogram: A bar graph plot of class data, with the bars placed adjacent to each other, is called a histogram.

The vertical axis of a histogram can represent either the class frequency or the relative class frequency; in the former case the graph is called a frequency histogram and in the latter a relative frequency histogram. Figure 5 presents a frequency histogram of the data in Table 4.

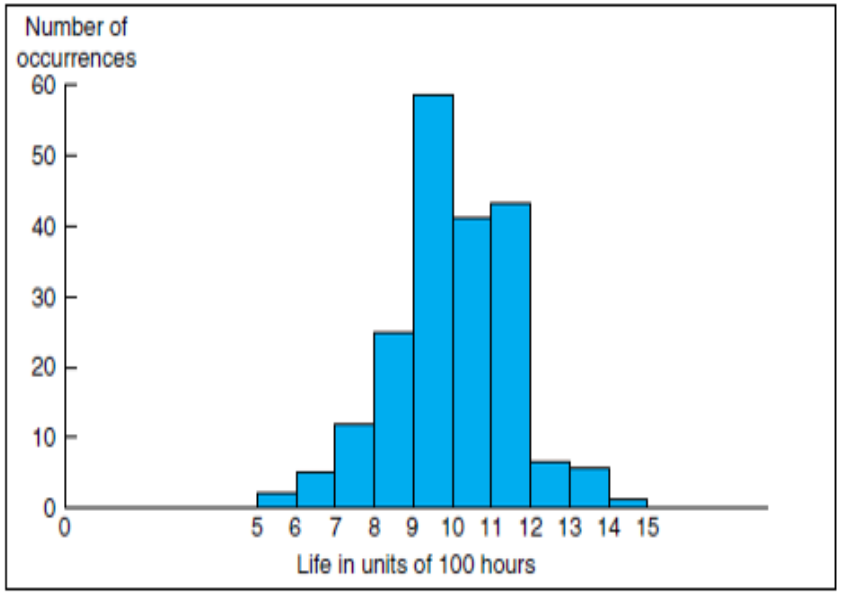


Figure 5. A frequency histogram.

F) Stem and leaf plot (Dal ve yaprak grafiği): An efficient way of organizing a small- to moderate-sized data set is to utilize a **stem and leaf plot.**

Such a plot is obtained by first dividing each data value into two parts —its stem (dal) and its leaf (yaprak). For example, if the data are all two-digit numbers, then we could let the stem part of a data value be its tens digit and let the leaf be its ones digit. Thus, for instance, the value 62 is expressed as

Stem Leaf
6 2
and the two data values 62 and 67 can be represented as
Stem Leaf
6 2,7

Example: Table 6 gives the monthly and yearly average daily minimum temperatures in 35 U.S. cities. The annual average daily minimum temperatures from Table 6 are represented in the following stem and leaf plot.

7	0.0
6	9.0
5	1.0, 1.3, 2.0, 5.5, 7.1, 7.4, 7.6, 8.5, 9.3
4	0.0, 1.0, 2.4, 3.6, 3.7, 4.8, 5.0, 5.2, 6.0, 6.7, 8.1, 9.0, 9.2
3	3.1, 4.1, 5.3, 5.8, 6.2, 9.0, 9.5, 9.5
2	9.0, 9.8

Table 6. Normal Daily Minimum Temperature— Selected Cities

[In Fahrenheit degrees. Airport data except as noted. Based on standard 30-year period, 1961 through 1990]

State	Station	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.	Annual avg.
AL	Mobile	40.0	42.7	50.1	57.1	64.4	70.7	73.2	72.9	68.7	57.3	49.1	43.1	57.4
AK	Juneau	19.0	22.7	26.7	32.1	38.9	45.0	48.1	47.3	42.9	37.2	27.2	22.6	34.1
AZ	Phoenix	41.2	44.7	48.8	55.3	63.9	72.9	81.0	79.2	72.8	60.8	48.9	41.8	59.3
AR	Little Rock	29.1	33.2	42.2	50.7	59.0	67.4	71.5	69.8	63.5	50.9	41.5	33.1	51.0
CA	Los Angeles	47.8	49.3	50.5	52.8	56.3	59.5	62.8	64.2	63.2	59.2	52.8	47.9	55.5
	Sacramento	37.7	41.4	43.2	45.5	50.3	55.3	58.1	58.0	55.7	50.4	43.4	37.8	48.1
	San Diego	48.9	50.7	52.8	55.6	59.1	61.9	65.7	67.3	65.6	60.9	53.9	48.8	57.6
	San Francisco	41.8	45.0	45.8	47.2	49.7	52.6	53.9	55.0	55.2	51.8	47.1	42.7	49.0
CO	Denver	16.1	20.2	25.8	34.5	43.6	52.4	58.6	56.9	47.6	36.4	25.4	17.4	36.2
CT	Hartford	15.8	18.6	28.1	37.5	47.6	56.9	62.2	60.4	51.8	40.7	32.8	21.3	39.5
DE	Wilmington	22.4	24.8	33.1	41.8	52.2	61.6	67.1	65.9	58.2	45.7	37.0	27.6	44.8
DC	Washington	26.8	29.1	37.7	46.4	56.6	66.5	71.4	70.0	62.5	50.3	41.1	31.7	49.2
FL	Jacksonville	40.5	43.3	49.2	54.9	62.1	69.1	71.9	71.8	69.0	59.3	50.2	43.4	57.1
	Miami	59.2	60.4	64.2	67.8	72.1	75.1	76.2	76.7	75.9	72.1	66.7	61.5	69.0
GA	Atlanta	31.5	34.5	42.5	50.2	58.7	66.2	69.5	69.0	63.5	51.9	42.8	35.0	51.3
HI	Honolulu	65.6	65.4	67.2	68.7	70.3	72.2	73.5	74.2	73.5	72.3	70.3	67.0	70.0
ID	Boise	21.6	27.5	31.9	36.7	43.9	52.1	57.7	56.8	48.2	39.0	31.1	22.5	39.1
IL	Chicago	12.9	17.2	28.5	38.6	47.7	57.5	62.6	61.6	53.9	42.2	31.6	19.1	39.5
	Peoria	13.2	17.7	29.8	40.8	50.9	60.7	65.4	63.1	55.2	43.1	32.5	19.3	41.0
IN	Indianapolis	17.2	20.9	31.9	41.5	51.7	61.0	65.2	62.8	55.6	43.5	34.1	23.2	42.4
IA	Des Moines	10.7	15.6	27.6	40.0	51.5	61.2	66.5	63.6	54.5	42.7	29.9	16.1	40.0
KS	Wichita	19.2	23.7	33.6	44.5	54.3	64.6	69.9	67.9	59.2	46.6	33.9	23.0	45.0
KY	Louisville	23.2	26.5	36.2	45.4	54.7	62.9	67.3	65.8	58.7	45.8	37.3	28.6	46.0
LA	New Orleans	41.8	44.4	51.6	58.4	65.2	70.8	73.1	72.8	69.5	58.7	51.0	44.8	58.5
ME	Portland	11.4	13.5	24.5	34.1	43.4	52.1	58.3	57.1	48.9	38.3	30.4	17.8	35.8
MD	Baltimore	23.4	25.9	34.1	42.5	52.6	61.8	66.8	65.7	58.4	45.9	37.1	28.2	45.2
MA	Boston	21.6	23.0	31.3	40.2	49.8	59.1	65.1	64.0	56.8	46.9	38.3	26.7	43.6
MI	Detroit	15.6	17.6	27.0	36.8	47.1	56.3	61.3	59.6	52.5	40.9	32.2	21.4	39.0
	Sault Ste. Marie	4.6	4.8	15.3	28.4	38.4	45.5	51.3	51.3	44.3	36.2	25.9	11.8	29.8
MN	Duluth	−2.2	2.8	15.7	28.9	39.6	48.5	55.1	53.3	44.5	35.1	21.5	4.9	29.0
	Minneapolis-St. Paul	2.8	9.2	22.7	36.2	47.6	57.6	63.1	60.3	50.3	38.8	25.2	10.2	35.3
MS	Jackson	32.7	35.7	44.1	51.9	60.0	67.1	70.5	69.7	63.7	50.3	42.3	36.1	52.0
MO	Kansas City	16.7	21.8	32.6	43.8	53.9	63.1	68.2	65.7	56.9	45.7	33.6	21.9	43.7
	St. Louis	20.8	25.1	35.5	46.4	56.0	65.7	70.4	67.9	60.5	48.3	37.7	26.0	46.7
MT	Great Falls	11.6	17.2	22.8	31.9	40.9	48.6	53.2	52.2	43.5	35.8	24.3	14.6	33.1

Source: U.S. National Oceanic and Atmospheric Administration, Climatography of the United States, No. 81.

2. SUMMARIZING DATA SETS

In this section we present some *summarizing statistics*, where a statistic is a numerical quantity whose value is determined by the data.

2.1. Measures of Location/Center (Sample Mean, Sample Median, and Sample Mode)

The most common and useful statistics are called measures of location.

Measures of location describe the central tendency of the data. Measure of location indicates where a certain part of the data is located.

2.1.1. Sample Mean

If the n observations in a sample are denoted by $x_1, x_2, \cdots, x_n$, the *sample mean* is $\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$

Example 1: Let's consider the eight observations collected from the prototype engine connectors. The eight prototype units are produced and their pull-of forces measured, resulting in the following data (in pounds):

$$x_1 = 12.6,\ x_2 = 12.9,\ x_3 = 13.4,\ x_4 = 12.3,\ x_5 = 13.6,\ x_6 = 13.5,\ x_7 = 12.6,\ x_8 = 13.1.$$

The sample mean is $\bar{x} = \dfrac{x_1 + x_2 + \cdots + x_n}{n} = \dfrac{\sum_{i=1}^{8} x_i}{8} = \dfrac{12.6 + 12.9 + \cdots + 13.1}{8} = \dfrac{104}{8} = 13.0$ pounds.

Example 2: The winning scores in the U.S. Masters golf tournament in the years from 1999 to 2008 were as follows: 280, 278, 272, 276, 281, 279, 276, 281, 289, 280. Find the sample mean of these scores: $\bar{x} = \dfrac{x_1 + x_2 + \cdots + x_n}{n} = \dfrac{\sum_{i=1}^{10} x_i}{10} = \dfrac{2792}{10} = 279.2$

The sample mean of a data set that is presented in a frequency table listing the *k* distinct values (class/group's value: average of the limits of group) $s_1,\ldots,s_k$ having corresponding frequencies $f_1,\ldots,f_k$.

Since such a data set consists of $n = \sum_{i=1}^{k} f_i$ observations, with the value s_i appearing f_i times, for each $i = 1,\ldots,k$, it follows that the sample mean of these *n* data values is $\bar{x} = \frac{\sum_{i=1}^{k} s_i f_i}{n}$

Example 3: The following is a frequency table giving the ages of members of a symphony orchestra for young adults. Find the sample mean of the ages of the 54 members of the symphony.

Age	Frequency
15	2
16	5
17	11
18	9
19	14
20	13

$$\bar{x} = \frac{\sum_{i=1}^{6} s_i f_i}{n} = \frac{15.2+16.5+17.11+18.9+19.14+20.13}{54} \cong 18.24$$

2.1.2. Sample Median

Another statistic used to indicate the center of a data set is the sample median. The *median* is the middle of the data (after data is arranged in ascending or descending order); half the observations are less than the median and half are more than the median. To get the median, we must first rearrange the data into an *ordered array*.

Generally, we order the data from the lowest value to the highest value. The median is the data value such that half of the observations are larger than it and half are smaller. *It is also the 50th percentile (we will be learning about percentiles).*

If n is odd, the median is the middle observation of the ordered array. If n is even, it is midway between the *two* central observations. Order the values of a data set of size *n* from smallest to largest. If *n* is odd, the sample median is the value in position $(n+1)/2$; if *n* is even, it is the average of the values in positions $n/2$ and $n/2+1$.

Example1: The data set is 10, 20, 30, 40, 50, 60 and n=6 and median=(30+40)/2=35

Example 2: Find the sample median for the ages of members of a symphony orchestra for young adults data. n=54 so that $i = \frac{n}{2} = \frac{54}{2} = 27$

and sample median= $\bar{x}' = \frac{x_{27} + x_{28}}{2} = \frac{18+19}{2} = 18.5$

Age	Frequency
15	2
16	5
17	11
18	9
19	14
20	13

Note that the mean and median are UNIQUE for a given set of data. That means each data only have one mean $(\bar{x})$ and median value $\bar{x}'$

ADVANTAGE: The Median is not affected by extreme values. In the previous example, if you change the 60 to 6000, the median will still be 35. The mean, on the other hand will change by a great deal. The sample mean and sample median are both useful statistics for describing the central tendency of a data set. *The sample mean makes use of all the data values and is affected by extreme values that are much larger or smaller than the others; the sample median makes use of only one or two of the middle values and is thus not affected by extreme values.*

2.1.3.Sample Mode

Another statistic that has been used to indicate the central tendency of a data set is the **sample mode,** defined to be the value that occurs with the greatest frequency. If no single value occurs most frequently, then all the values that occur at the highest frequency are called *modal values* (tepe değerleri). Mode is corresponded to the value which is the most observed in the data.

Example1: 1, 1, 1, 2, 3, 4, 5

The mode is $\hat{x} = 1$ since it occurs three times. The other values only appear once in the data set.

Example 2: 5, 5, 5, 6, 8, 10, 10, 10

The modes for this data set are 5 and 10. This is a *bi-modal (iki tepeli)* dataset.

Example 3: The following frequency table gives the values obtained in 40 rolls of a die.

Value	Frequency
1	9
2	8
3	5
4	5
5	6
6	7

Find (a) the sample mean, (b) the sample median, and (c) the sample mode.

(a) The sample mean is $\bar{x} = \frac{\sum_{i=1} s_i f_i}{n} = \frac{1.9+2.8+3.5+4.5+5.6+6.7}{40} = \frac{132}{40} = 3.3$

(b) The sample median is the average of the 20th and 21st smallest values, and is thus equal to $\bar{x}' = \frac{x_{20}+x_{21}}{2} = \frac{3+3}{2} = 3$

(c) The sample mode is $\hat{x} = 1$, the value that occurred most frequently.

Problems: The mode may not exist (for example all values are occurs one times there is no mode). The mode may not be unique (means that more than one mode could be).

2.2. Sample Quartiles and Box Plots

Quantiles are also measures of locations.

Quartiles divide the ordered set of data *into four equal parts*, the division points are called *quartiles.*

Q1 – First Quartile – 25% of the observations are smaller than Q1 and 75% of the observations are larger than Q1.

Q2 – Second Quartile – 50% of the observations are smaller than Q2 and 50% of the observations are larger than Q2. Same as the Median. It is also the 50th percentile.

Q3 – Third Quartile – 75% of the observations are smaller than Q3 and 25% of the observations are larger than Q3.

The quartiles, like the median, either take the value of one of the observations, or the value halfway between two observations.

First Quartile:

$$Q_1 = \begin{cases} x_i & i = \frac{n+1}{4},\ n \text{ is odd} \\ \frac{x_i + x_{i+1}}{2} & i = \frac{n}{4},\ n \text{ is even} \end{cases}$$

Second Quartile (Median):

$$Q_2 = \begin{cases} x_i & i = \frac{n+1}{2},\ n \text{ is odd} \\ \frac{x_i + x_{i+1}}{2} & i = \frac{n}{2},\ n \text{ is even} \end{cases}$$

Third Quartile:

$$Q_3 = \begin{cases} x_i & i = \frac{3(n+1)}{4},\ n \text{ is odd} \\ \frac{x_i + x_{i+1}}{2} & i = \frac{3n}{4},\ n \text{ is even} \end{cases}$$

Example 1: Original data: 3, 10, 2, 5, 9, 8, 7, 12, 10, 0, 4, 6
Ordered data: 0, 2, 3, 4, 5, 6, 7, 8, 9, 10, 10, 12 (To find quartiles order the data)

Mean: $\bar{x} = \frac{\sum_{i=1}^{12} x_i}{12} = \frac{76}{12} = 6.33$, **Mode=10 (because two times observed)**

n=12 (even) for Q1; $i = \frac{12}{4} = 3$ **and** $Q1 = \frac{x_3 + x_4}{2} = \frac{3+4}{2} = 3.5$ **(25% of data values smaller than 3.5; 75% of data values larger than 3.5)**

For Q2 (Median); $i = \frac{12}{2} = 6$ **and** $Q2 = \frac{x_6 + x_7}{2} = \frac{6+7}{2} = 6.5$ **(50% of data values smaller than 6.5; 50% of data values larger than 6.5)**

For Q3; $i = \frac{3 \times 12}{4} = 9$ **and** $Q3 = \frac{x_9 + x_{10}}{2} = \frac{9+10}{2} = 9.5$ **(75% of data values smaller than 9.5; 25% of data values larger than 9.5)**

Example 2: Ordered data: 210, 220, 225, 225, 225, 235, 240, 250, 270, 280

Mean: $\bar{x} = \frac{\sum_{i=1}^{12} x_i}{10} = \frac{2380}{10} = 238$, Mode=225 (because three times observed)

n=10 (even) for Q1; $i = \frac{10}{4} = 2.5$ and $Q1 = \frac{x_2 + x_3}{2} = \frac{220 + 225}{2} = 222.5$ (25% of data values smaller than 222.5; 75% of data values larger than 222.5)

For Q2 (Median); $i = \frac{10}{2} = 5$ and $Q2 = \frac{x_5 + x_6}{2} = \frac{225 + 235}{2} = 230$ (50% of data values smaller than 230; 50% of data values larger than 230)

For Q3; $i = \frac{3 \times 10}{4} = 7.5$ and $Q1 = \frac{x_7 + x_8}{2} = \frac{240 + 250}{2} = 245$ (75% of data values smaller than 245; 25% of data values larger than 245)

Example 3: Noise is measured in decibels, denoted as dB. One decibel is about the level of the weakest sound that can be heard in a quiet surrounding by someone with good hearing; a whisper measures about 30 dB; a human voice in normal conversation is about 70 dB; a loud radio is about 100 dB. Ear discomfort usually occurs at a noise level of about 120 dB. *The following data give noise levels measured at 36 different times directly outside of Grand Central Station in Manhattan. Determine the quartiles.*

82, 89, 94, 110, 74, 122, 112, 95, 100, 78, 65, 60, 90, 83, 87, 75, 114, 85
69, 94, 124, 115, 107, 88, 97, 74, 72, 68, 83, 91, 90, 102, 77, 125, 108, 65

A stem and leaf plot of the data is as follows:

6	0, 5, 5, 8, 9
7	2, 4, 4, 5, 7, 8
8	2, 3, 3, 5, 7, 8, 9
9	0, 0, 1, 4, 4, 5, 7
10	0, 2, 7, 8
11	0, 2, 4, 5
12	2, 4, 5

Becareful: You can see the data set is in an ordered form in this plot that helps you to find Q1, Q2, Q3 easily!!

n=36 (even) for Q1; $i = \frac{36}{4} = 9$ and $Q1 = \frac{x_9 + x_{10}}{2} = \frac{75 + 77}{2} = 76$

(25% of the times *noise levels* smaller than 76; 75% of times *noise levels* larger *than 76*)

For Q2 (Median); $i = \frac{36}{2} = 18$ and $Q2 = \frac{x_{18} + x_{19}}{2} = \frac{89 + 90}{2} = 89.5$

(50% of the times *noise levels* smaller than 89.5; 75% of times *noise levels* larger *than 89.5*)

For Q3; $i = \frac{3 \times 36}{4} = 27$ and $Q1 = \frac{x_{27} + x_{28}}{2} = \frac{102 + 107}{2} = 104.5$

(75% of the times *noise levels* smaller than 104.5; 25% of times *noise levels* larger *than 104.5*)

BOX PLOT

A **box plot** *is often used to plot some of the summarizing statistics of a data set.*

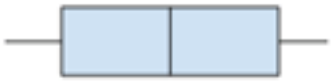
A straight line segment stretching from the smallest to the largest data value is drawn on a horizontal axis; imposed on the line is a "box," which starts at the first and continues to the third quartile, with the value of the second quartile indicated by a vertical line

Three types of boxplot giving information about shape of data, at the end of this section will be more cleared.

Median closer to 3rd quartile

Left Skew

Median in center



No Skew
"Bell Curve"

Median closer to 1st quartile

Right Skew

Example: ***Table 7 is a frequency table*** for a data set consisting of the starting yearly salaries (to the nearest thousand dollars) of 42 recently graduated students with B.S. degrees in electrical engineering.

Table 7. Starting Yearly Salaries

Starting Salary	Frequency
47	4
48	1
49	3
50	5
51	8
52	10
53	0
54	5
56	2
57	3
60	1

The 42 data values presented in Table 7 go from a low value of 47 to a high value of 60. The box plot for this data set is shown in Figure 6.

n=42 (even) for Q1; $i = \frac{42}{4} = 10.5$ and $Q1 = \frac{x_{10} + x_{11}}{2} = \frac{50 + 50}{2} = 50$

(25% of graduated students starting yearly salaries smaller than 50 thousand dollars; 75% of graduated students starting yearly salaries larger than 50 thousand dollars)

For Q2 (Median); $i = \frac{42}{2} = 21$ and $Q2 = \frac{x_{21} + x_{22}}{2} = \frac{51 + 52}{2} = 51.5$

(50% of graduated students starting yearly salaries smaller than 51.5 thousand dollars; 50% of graduated students starting yearly salaries larger than 51.5 thousand dollars)

For Q3; $i = \frac{3 \times 42}{4} = 31.5$ and $Q1 = \frac{x_{31} + x_{32}}{2} = \frac{52 + 54}{2} = 53$

(75% of graduated students starting yearly salaries smaller than 53 thousand dollars; 25% of graduated students starting yearly salaries larger than 53 thousand dollars)

Lower Fence=Q1-1.5(Q3-Q1)=50-1.5×(53-50)=45.5

Upper Fence=Q3+1.5(Q3-Q1)=53+1.5×(53-50)=57.5

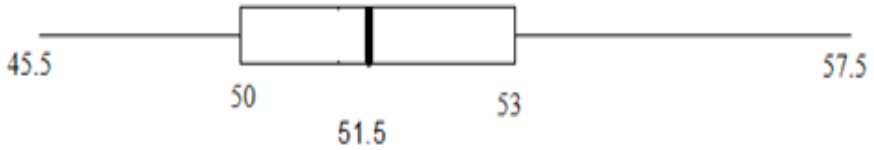


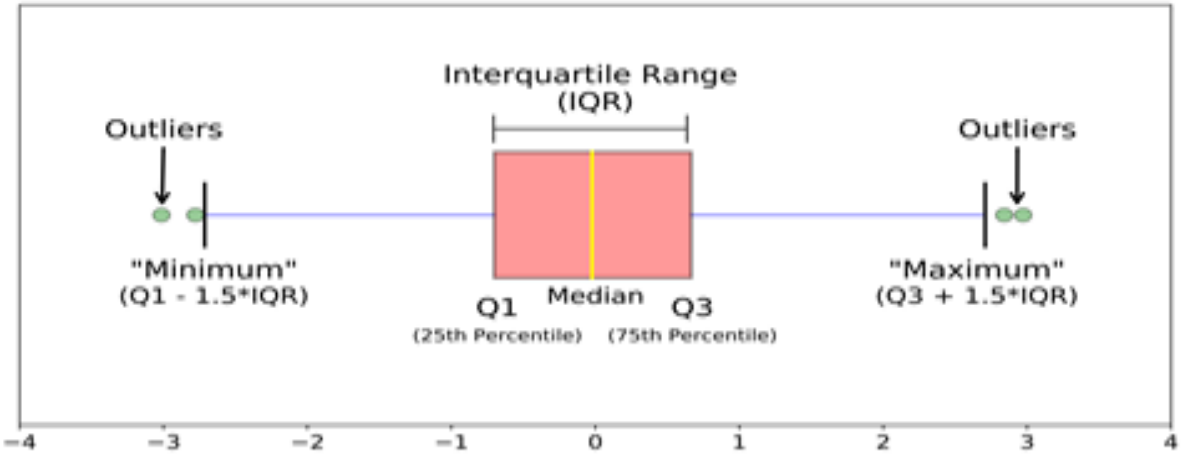
Figure 6. A box plot.

- Median in center, this is symmetric distribution.

P.C. Interquartile Range =Q3-Q1

The length of the line segment on the box plot, equal to the Upper Fence minus the Lower Fence value, is called the *range* of the data.

Also, the length of the box itself, equal to the third quartile (Q_3) minus the first quartile (Q_1), is called the *interquartile range*.



Example: Table 8 lists the populations of the 25(odd) most populous U.S. cities for the year 1994.

For this data set, find (a) the sample 25 percentile and (b) the sample 75 percentile.

Table 8. *Population of 25 Largest U.S. Cities, July 2006*

Rank	City	Population
1	New York, NY	8,250,567
2	Los Angeles, CA	3,849,378
3	Chicago, IL	2,833,321
4	Houston, TX	2,144,491
5	Phoenix, AR	1,512,986
6	Philadelphia, PA	1,448,394
7	San Antonio, TX	1,296,682
8	San Diego, CA	1,256,951
9	Dallas, TX	1,232,940
10	San Jose, CA	929,936
11	Detroit, MI	918,849
12	Jacksonville, FL	794,555
13	Indianapolis, IN	785,597
14	San Francisco, CA	744,041
15	Columbus, OH	733,203
16	Austin, TX	709,893
17	Memphis, TN	670,902
18	Fort Worth, TX	653,320
19	Baltimore, MD	640,961
20	Charlotte, NC	630,478
21	El Paso, TX	609,415
22	Milwaukee, WI	602,782
23	Boston, MA	590,763
24	Seattle, WA	582,454
25	Washington, DC	581,530

(a) Because the sample size is 25(odd) and 26(0.25) =6.5, the sample 25 percentile is average of 6th and 7th values, equal to (1448394 +1296682)/2=1372538 $\left[Q_1 = \frac{x_6 + x_7}{2}\right]$

(b) Because 26(0.75)=19.5, the sample 75 percentile is the average of the nineteenth and the twentieth values. Hence, the sample 75 percentile is (640961 +630478)/2=635719.5.

$\left[Q_3 = \frac{x_{19} + x_{20}}{2}\right]$

2.3.Measures of Dispersion / Variability (Range, Interquartile range, Sample Variance, Sample Standard Deviation, Coefficient of Variation)

- Whereas we have presented statistics that describe the central tendencies of a data set, we are also interested in ones that describe *the spread or variability of the data values.* While measures of central tendency are used to estimate "normal" values of a dataset, measures of dispersion are important for describing the spread of the data, or its variation around a central value. It shows how much the data vary from their average value.

- The measure of dispersion tells the variation of the data from one another and gives a clear idea about the distribution of the data. The measure of dispersion shows the homogeneity or the heterogeneity of the distribution of the observations.

- Two distinct samples may have the same mean or median, but completely different levels of variability, or vice versa. A proper description of a set of data should include both of these characteristics.

- **Dispersion** is the amount of spread, or variability, in a set of data. There are various methods that can be used to measure the dispersion of a dataset, each with its own set of advantages and disadvantages. There are mainly 5 major measures of dispersion:

2.3.1. Range

Range = Largest Value – Smallest Value

Example: ordered data: 1 2 3 4 8 then Range = 8 – 1 = 7
Problem: The range is influenced by extreme values at either end.

2.3.2. Interquartile Range

IQR = Q3 – Q1

It is basically the range encompassed by the central 50% of the observations in the distribution. It is less sensitive to extreme values in the sample than is the ordinary sample range.

Problem: The IQR does not take into account the variability of the *total* data (only the central 50%). We are "throwing out" half of the data.

2.3.3. Sample Variance

A statistic that could be used for this purpose would be one that measures the average value of the squares of the distances between the data values and the sample mean. This is accomplished by the sample variance, which for technical reasons divides the sum of the squares of the differences by $n-1$ rather than n, where n is the size of the data set.

The population variance is $\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$ but it is very rare that we ever take a census (tamsayım) of the population and deal with N. Normally, we work with a sample and calculate the sample measures, like the sample mean and the sample variance s^2.

Sample variance: $$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^{n}x_i^2 - \frac{\left(\sum_{i=1}^{n}x_i\right)^2}{n}}{n-1} = \frac{\sum_{i=1}^{n}x_i^2 - n\bar{x}^2}{n-1}$$

EXTRA INFORMATION: The reason we divide by n-1 instead of n is to assure that s is an unbiased estimator of σ. We have taken a shortcut: in the second formula, we are using $\bar{x}$, a statistic, instead of μ, a parameter. To correct for this – which has a tendency to understate the true standard deviation – ***we divide by n-1 which will increase s somewhat and make it an unbiased estimator of σ. Later on in the course we will refer to this as "losing one degree of freedom."***

Proof for the formula of sample variance:

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

The identity is proven as follows:

$$\begin{aligned}\sum_{i=1}^{n}(x_i-\bar{x})^2&=\sum_{i=1}^{n}\left(x_i^2-2x_i\bar{x}+\bar{x}^2\right)\\&=\sum_{i=1}^{n}x_i^2-2\bar{x}\sum_{i=1}^{n}x_i+\sum_{i=1}^{n}\bar{x}^2\\&=\sum_{i=1}^{n}x_i^2-2n\bar{x}^2+n\bar{x}^2\\&=\sum_{i=1}^{n}x_i^2-n\bar{x}^2\end{aligned}$$

Example 1: Find the sample variances of the data sets A and B given below.

A: 3, 4, 6, 7, 10 B:−20, 5, 15, 24

The sample mean for data set A: $\bar{x} = (3+4+6+7+10)/5 = 6$

its sample variance is $s^2 = \dfrac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1} = \dfrac{210-(5\times 6^2)}{5-1} = 7.5$

The sample mean for data set B is also 6; its sample variance is

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1} = \frac{1226-(4\times 6^2)}{4-1} = 360.67$$

Thus, although both data sets have the same sample mean, there is a much greater variability in the values of the B set than in the A set.

Example 2: The following data give the worldwide number of fatal airline accidents of commercially scheduled air transports in the years from 1997 to 2005.

Year	1997	1998	1999	2000	2001	2002	2003	2004	2005
Accidents	25	20	21	18	13	13	7	9	18

Source: National Safety Council.

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1} = \frac{2582-(9\times 16^2)}{9-1} = 34.75$$

2.3.4. Sample Standard Deviation

The positive square root of the sample variance is called the sample standard deviation.

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

is called the ***sample standard deviation.***

The sample standard deviation is measured in the same units as the data.

If we want to determine the sample mean of a data set that is presented in a frequency table listing the *k* distinct values (class/group's value: average of the limits of group) $s_1,\ldots,s_k$ having corresponding frequencies $f_1,\ldots,f_k$. Since such a data set consists of $n = \sum_{i=1}^{k} f_i$ observations, with the value s_i appearing f_i times, for each $i = 1,\ldots,k$, it follows that the sample variance and sample standard deviation of these *n* data values, respectively: $s^2 = \frac{\sum_{i=1}^{k} f_i s_i^2 - \frac{\left(\sum_{i=1}^{k} f_i s_i\right)^2}{n}}{n-1}$ and thus sample standart deviation $s = \sqrt{s^2}$

2.3.5. Coefficient of Variation

The problem with s^2 and s is that they are in the "original" units. This makes it difficult to compare the variability of two data sets, if they are in different units or if the magnitude of the numbers is very different.

Suppose you wish *to compare two stocks and one is in dollars and the other is in yen*; if you want to know which one is more volatile (oynak, değişken), you should use the coefficient of variation.

It is also not appropriate to compare two stocks of vastly different prices even if both are in the same units. The standard deviation for a stock that sells for around $300 is going to be very different than one where the price is around $0.25. *The coefficient of variation will be a better measure of dispersion when comparing the two stocks than the standard deviation (see example below).*

$$CV = \frac{s}{\bar{x}} \times 100\%$$

BE CAREFUL: The higher the CV, the greater the level of dispersion around the mean. It is generally expressed as a percentage. Without units, it allows for comparison between distributions of values whose scales of measurement are not comparable. (NOT ONLY COMPARING TWO STOCKS, DIFFERENT EXAMPLES COULD BE)

Example:
Which stock price is more volatile?

Closing prices over the last 8 months:

	Stock A	Stock B
JAN	$1.00	$180
FEB	1.50	175
MAR	1.90	182
APR	.60	186
MAY	3.00	188
JUN	.40	190
JUL	5.00	200
AUG	.20	210
Mean	$1.70	$188.88
s^2	2.61	128.41
s	$1.62	$11.33

The standard deviation of B is higher than for A, but A is more volatile:

$$CV_A = \frac{\$1.62}{\$1.70} \times 100\% = 95.3\%$$

$$CV_B = \frac{\$11.33}{\$188.88} \times 100\% = 6.0\%$$

2.3. Measures of Shape

A third important property of data is its shape.

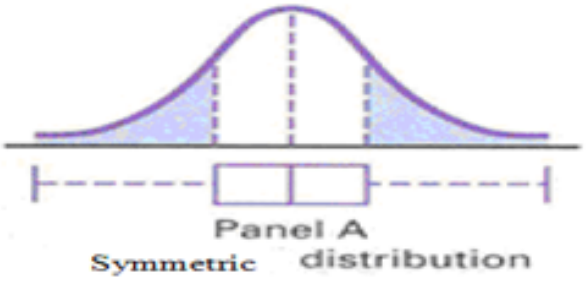
Shape is the distribution symmetric or skewed? The distribution is flat or rather sharp peak?

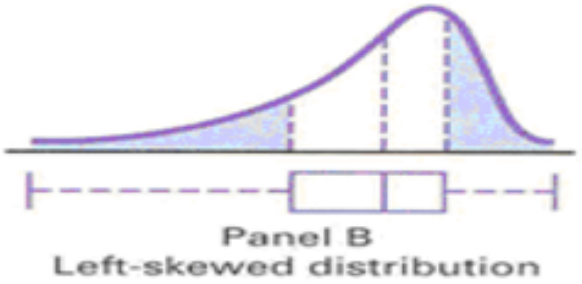
Symmetric distributions are those whose left and right-hand sides look like mirror images of one another (perfect symmetry is a rarity in real life).

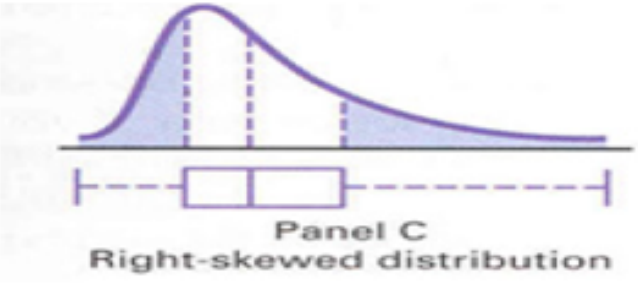
Shape can be described by degree of asymmetry (i.e., skewness) and peakedness (kurtosis) of a distribution.

If a data distribution is skewed right, the mean will be greater than the median.
(mean > median positive or right-skewness)
if a data distribution is perfectly symmetric, the median and mean will be equal.
(mean = median symmetry or zero-skewness)
if a data distribution is skewed left, the mean will be less than the median.
(mean < median negative or left-skewness)

Positive skewness arises when the mean is increased by some unusually high values. Negative skewness occurs when the mean is decreased by some unusually low values.







Example: Hours to complete a task:

2	3	8	8	9	10	10	12	15	18	22	63

$\bar{x} = \frac{180}{12} = 15$ n=12 (even) $i = \frac{n}{2} = \frac{12}{2} = 6$ $\text{median} = \bar{x}' = Q_2 = \frac{x_6 + x_7}{2} = \frac{10+10}{2} = 10$

mean > median positive or right-skewness

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1} = \frac{5568 - (12 \times 15^2)}{12-1} = 260.72$$

$$s = \sqrt{s^2} = \sqrt{260.72} = 16.15$$

$$CV = \frac{s}{\bar{x}} \times 100\% = \frac{16.15}{15} \times 100\% = 107.7$$

Kurtosis Value Range

Kurtosis *refers to a measure of the degree to which* ***a given distribution is more or less 'peaked', relative to the normal distribution.*** **The concept of kurtosis is very useful in decision-making. In this regard, we have 3 categories of distributions:**

- **Normal distribution kurtosis = 0**
- **A distribution that is more peaked and has fatter tails than normal distribution has kurtosis value greater than 0 (the higher kurtosis, the more peaked and fatter tails). Such distribution is called *leptokurtic* or *leptokurtotic.***
- **A distribution that is less peaked and has thinner tails than normal distribution has kurtosis value less than 0. Such distribution is called *platykurtic* or *platykurtotic.***