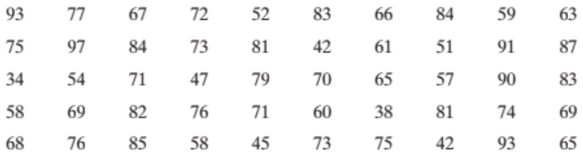
Chapter 6: Point Estimation

6.2 Exploratory Data Analysis

Stem-and-leaf display.

- A stem-and-leaf display or stem-and-leaf plot is a device for presenting quantitative data in a graphical format, similar to a histogram, to assist in visualizing the shape of a distribution.
- We can do much the same thing as a frequency table and histogram can, but keep the original values, through a stem-and-leaf display.

Example 1. Say we have the following 50 test scores on a statistics examination:



93	77	67	72	52	83	66	84	59	63
75	97	84	73	81	42	61	51	91	87
34	54	71	47	79	70	65	57	90	83
58	69	82	76	71	60	38	81	74	69
68	76	85	58	45	73	75	42	93	65

Stem-and-leaf display

Ordered stem-and-leaf display

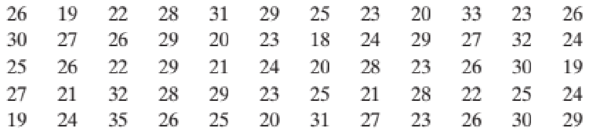
Another Modification of stem-and-leaf display

We can do this by recording leaves 0, 1, 2, 3, and 4 with a stem adjoined with an asterisk () and leaves 5, 6, 7, 8, and 9 with a stem adjoined with a dot (●).

Tukey's Modification

- Leaves are recorded as zeros and ones with a stem adjoined with an asterisk ($*$)
- twos and threes with a stem adjoined with t
- fours and fives with a stem adjoined with f
- sixes and sevens with a stem adjoined with s
- eights and nines with a stem adjoined with a dot (●).

Example 2. The following numbers represent ACT composite scores for 60 entering freshmen at a certain college:



26	19	22	28	31	29	25	23	20	33	23	26
30	27	26	29	20	23	18	24	29	27	32	24
25	26	22	29	21	24	20	28	23	26	30	19
27	21	32	28	29	23	25	21	28	22	25	24
19	24	35	26	25	20	31	27	23	26	30	29

Order Statistics: There is a reason for constructing ordered stem-and-leaf diagrams. For a sample of n observations, $x_1, x_2, ..., x_n$, when the observations are ordered from smallest to largest, the resulting ordered data are called the **order statistics** of the sample.

Note: Sometimes we give ranks to these order statistics and use the rank as the sub- script on y. Consider the example 1:

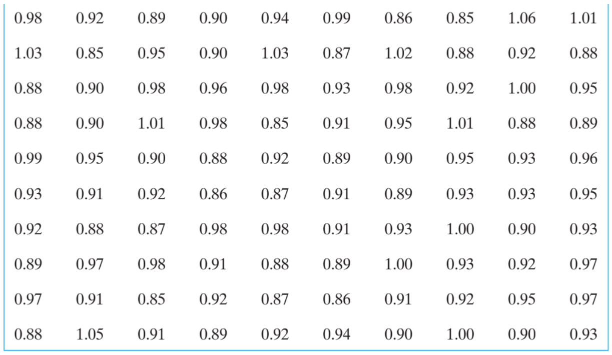
The first order statistic $y_1 = 34$ has rank 1; the second order statistic ________________ has rank _____; the third order statistic________________ has rank _____; the fourth order statistic ________________ has rank _____, . . . ; and the 50th order statistic ________________ has rank _____. It is also about as easy to determine these values from the ordered stem-and-leaf display. We see that ____________________________.

Sample Percentiles

- If $0 < p < 1$, then the $(100p)^{th}$ sample percentile has approximately np sample observations less than it and also $n(1p)$ sample observations greater than it.
- One way of achieving this is to take the $(100p)^{th}$ sample percentile as the $(n+1)p^{th}$ order statistic, provided that $(n+1)p$ is an integer.
- If $(n+1)p$ is not an integer but is equal to r plus some proper fraction—say, a/b—use a weighted average of the r^{th} and the $(r+1)^{st}$ order statistics.
- That is, define the $(100p)^th$ sample percentile as

- The 50th percentile is the ________________ of the sample.
- The 25th, 50th, and 75th percentiles are, respectively, the ________________, ________________, and ________________ ________________ of the sample.
- For notation, we let ____
- The 10th, 20th, . . . , and 90th percentiles are the ________________ of the sample,
- The 50th percentile is also the ________________, the ________________, and the ________________.

Example 3. A manufacturer of fluoride toothpaste regularly measures the concentration of fluoride in the toothpaste to make sure that it is within the specification of 0.85 to 1.10 mg/g. Following table lists 100 such measurements.



0.98	0.92	0.89	0.90	0.94	0.99	0.86	0.85	1.06	1.01
1.03	0.85	0.95	0.90	1.03	0.87	1.02	0.88	0.92	0.88
0.88	0.90	0.98	0.96	0.98	0.93	0.98	0.92	1.00	0.95
0.88	0.90	1.01	0.98	0.85	0.91	0.95	1.01	0.88	0.89
0.99	0.95	0.90	0.88	0.92	0.89	0.90	0.95	0.93	0.96
0.93	0.91	0.92	0.86	0.87	0.91	0.89	0.93	0.93	0.95
0.92	0.88	0.87	0.98	0.98	0.91	0.93	1.00	0.90	0.93
0.89	0.97	0.98	0.91	0.88	0.89	1.00	0.93	0.92	0.97
0.97	0.91	0.85	0.92	0.87	0.86	0.91	0.92	0.95	0.97
0.88	1.05	0.91	0.89	0.92	0.94	0.90	1.00	0.90	0.93

This following ordered stem-and-leaf diagram is useful for finding sample percentiles of the data.

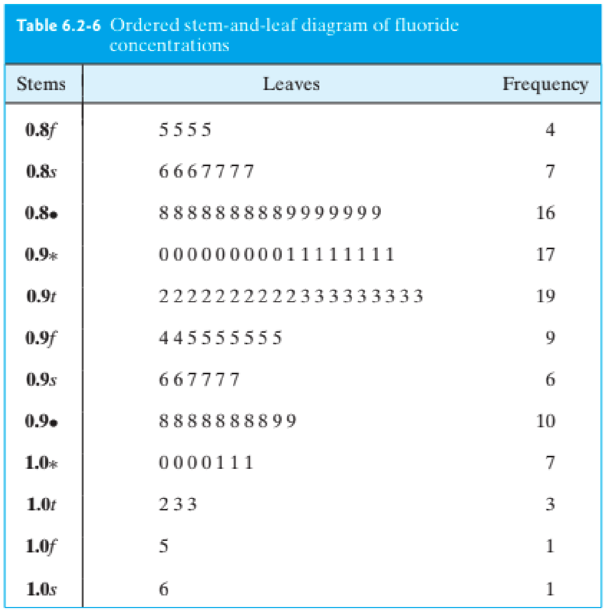


Table 6.2-6 Ordered stem-and-leaf diagram of fluoride concentrations

Stems	Leaves	Frequency
$\mathbf{0.8}f$	5 5 5 5	4
$\mathbf{0.8}s$	6 6 6 7 7 7 7	7
$\mathbf{0.8}\bullet$	8 8 8 8 8 8 8 8 8 9 9 9 9 9 9 9	16
$\mathbf{0.9}*$	0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1	17
$\mathbf{0.9}t$	2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3	19
$\mathbf{0.9}f$	4 4 5 5 5 5 5 5 5	9
$\mathbf{0.9}s$	6 6 7 7 7 7	6
$\mathbf{0.9}\bullet$	8 8 8 8 8 8 8 8 9 9	10
$\mathbf{1.0}*$	0 0 0 0 1 1 1	7
$\mathbf{1.0}t$	2 3 3	3
$\mathbf{1.0}f$	5	1
$\mathbf{1.0}s$	6	1

Five-number summary

Interquartile range (IQR)

Box-and-whisker diagram - Box plot:

Step1: Draw a horizontal axis that is scaled to the data.

Step2: Above the axis, draw a rectangular box with the left and right sides drawn at ____________ and ____________ and with a vertical line segment drawn at the median, ____________.

Step3: A ____________ is drawn as a horizontal line segment from the minimum to the midpoint of the left side of the box, and a ____________ is drawn as a horizontal line segment from the midpoint of the right side of the box to the maximum.

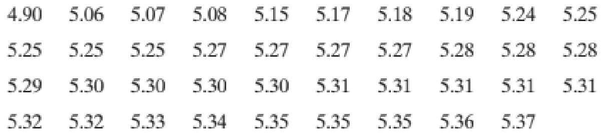
Note: The length of the box is equal to the ____________.

The left and right whiskers represent the ____________ and ____________ quarters of the data.

The two ____________ of the data are represented, respectively, by the two sections of the box, one to the left and one to the right of the median line.

Example 4. Boxplot for fluoride data:

Example 5. The following data give the ordered weights (in grams) of 39 gold coins that were produced during the reign of Verica, a pre-Roman British king:



4.90	5.06	5.07	5.08	5.15	5.17	5.18	5.19	5.24	5.25
5.25	5.25	5.25	5.27	5.27	5.27	5.27	5.28	5.28	5.28
5.29	5.30	5.30	5.30	5.30	5.31	5.31	5.31	5.31	5.31
5.32	5.32	5.33	5.34	5.35	5.35	5.35	5.36	5.37	

Draw a box plot for this data.

Outliers

- Sometimes we are interested in picking out observations that seem to be much
  larger or much smaller than most of the other observations. That is, we are looking for ____________________

- In a box-and-whisker diagram, construct ____________________ to the left and right of the box at a distance of 1.5 times the IQR.

- ____________________ are constructed in the same way at a distance of 3 times the IQR.
- Observations that lie between the inner and outer fences are called ____________________.
- Observations that lie beyond the outer fences are called ____________________.
- The observations beyond the inner fences are denoted with a circle (●).
- The whiskers are drawn only to the extreme values within or on the inner fences.

Note: When you are analyzing a set of data, suspected outliers deserve a closer look and outliers should be looked at very carefully. It does not follow that suspected outliers should be removed from the data, unless some error (such as a recording error) has been made. Moreover, it is sometimes important to determine the cause of extreme values, because outliers can often provide useful insights into the situation under consideration (such as a better way of doing things).

a) Midrange = average of the extremes
b) Range = difference of the extremes.
c) Interquartile range = difference of third and first quartiles ____________________________