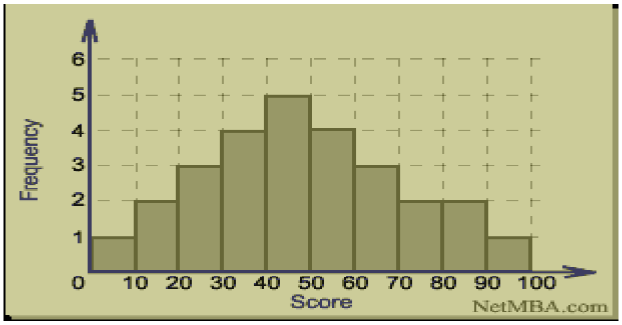
CHAPTER 1: SAMPLING AND DESCRIPTIVE STATISTICS

Graphical Summaries (Section 1.3, page 25)

Stem-and-Leaf Plots

Most of us already know what a histogram looks like:



Although a histogram tells us how many observations fall in a particular class, we lose information about the different values within a class.

A different method for displaying data which overcomes some of these difficulties, and which is easy to do, even by hand, is the Stem and Leaf Plot.

Recall that the position of an individual digit in a number tells us the value the digit represents. For example, consider the number 962.78. For this number 9 is 100's digit, 6 is 10's digit, 2 is 1's digit, 7 is .1's digit and 8 is .01's digit.

The stem and leaf plot uses selected digits (called stems) to group the sample into classes. The individual observations are then represented by the stem and the next significant digit. This is called the "truncation" method.

EXAMPLE: Consider the following sample data of English Scholastic Aptitude Test (SAT) scores.

638 574 627 621 705 690 522 612 594 581 640 653 638 760 491

The data here ranges from 491 to 760. Thus it is reasonable to group the sample using the 100's digit.

The number "638" is plotted as 6 | 3, (8 is truncated).

"6" is called a stem. The stem unit is SU=100 since the stem digit is the 100's digit.

"3" is called a Leaf. The leaf unit is LU=10, since the leaf digit is 10's digit.

6 | 3 = 630 approximates the actual value of 638.

Note: The leaf always consists of a single digit [anything more is truncated].

To illustrate these ideas, let's plot with stem unit 100:

638 574 627 621 705 690 522 612 594 581 640 653 638 760 491

Initial Plot: Stems Leaves

Final Plot: Stems Leaves

LU= _____________ , SU = ______________.

This plot is called one leaf category per stem plot (LCPS). The number of LCPS merely gives the number of lines for which the stem value is the same.

Example: Two leaf category per stem plot:

Stems Leaves

LU= ______________, SU= _________________.

Note: The increment of a stem and leaf plot is the distance from one line to the next.

INCR= SU/#LCPS

For the above example:

Histograms

A set of raw data gives us very little information as to how the observations are distributed. For example, consider the following sample of grades from a large mathematics class (the grades have been ordered from lowest to highest for convenience).

55	55	55	56	56	57	57	57	58	59	59	60	60	60
61	65	65	66	66	66	67	67	67	67	67	67	68	68
68	69	69	69	69	69	69	70	70	70	71	71	72	72
72	72	73	73	73	73	73	73	73	74	74	74	74	76
76	76	76	76	76	76	77	77	77	77	78	78	79	79
79	80	80	80	80	81	82	82	82	83	83	83	83	84
84	85	85	87	87	88	88	89	92	92	94			

What is the distribution of these grades? Are there more B's than C's? Do the grades clump around a certain grade?

Divide the sample of grades into five classes, say , [50, 60), [60, 70), [70, 80), [80, 90), and [90, 100), and find the frequency and relative frequency for each class.

<u>Frequency and Relative Frequency Distribution (or Table)</u>

Class Interval	Class Frequency (f_i)	Class Relative Frequency (f_i/n)
[50, 60)	11	11/95 = .1158
[60, 70)		
[70, 80)		
[80, 90)		
[90, 100)		

Note: The **CLASS MID-POINT** is the average of the lower and upper class boundaries.

Note: The **CLASS WIDTH (CW)** is usually (but not always) the same size within a histogram.

Frequency and Relative Frequency Histograms

A frequency (or relative frequency) histogram of a sample is simply a picture of the frequency (or relative frequency) distribution of that sample.

Example (FREQUENCY HISTOGRAM): To draw a frequency histogram of our sample of 95 grades from a large mathematics class we proceed as follows:

(a) Form an (x,y) coordinate system with the class intervals on the horizontal axis and the class frequencies marked on the vertical axis.

(b) Above each class draw a rectangle, whose height is equal to the class frequency.

Recall:

Class Interval	Class Midpoint (m_i)	Class Frequency (f_i)	Class Relative Frequency (f_i/n)
[50, 60)	55	11	11/95 = .1158
[60, 70)	65	24	24/95 = .2526
[70, 80)	75	36	36/95 = .3789
[80, 90)	85	21	21/95 = .2211
[90, 100)	95	3	3/95 = .0316

To draw a relative frequency histogram, mark the relative frequencies on the vertical axis and above each class draw a rectangle whose height is equal to the class relative frequency.

Note: The class width determines the number of classes. This choice will affect the "look" of the histogram. For example, what would the frequency histogram of the class grades look like if we used one class [50, 100).

Histograms for Categorical Data

For a categorical data, the categories themselves may be used as classes.

Example: The table below gives the final grades to a class of 100 students in an elementary statistics course. Create a categorical histogram.

Grades	A	B	C	D	F
# of students	15	25	30	20	10

Some Typical Sample Shapes

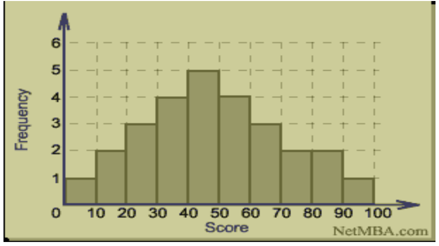
(1)A **LEFT SKEWED** sample is one whose histogram (or stem and leaf plot) has a long left tail; the sample values tend to cluster at the right end of the scale and taper off at the lower end.

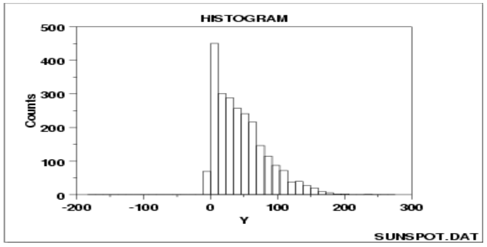
(2)A **RIGHT SKEWED** sample is one whose histogram (or stem and leaf plot) has a long right tail; the sample values tend to cluster at the left end of the scale and taper off at the higher end.

(3)A **SYMMETRIC** sample is one whose histogram (or stem and leaf plot) is distributed approximately the same on each side of some central value.

(4) A particular type of symmetric sample is a **BELL- SHAPED**; all bell shaped samples are symmetric, but not all symmetric samples are bell shaped.

Example: For the following histograms, describe the skew:





Boxplots

A boxplot is a visual display of data based on the following five-number summary:

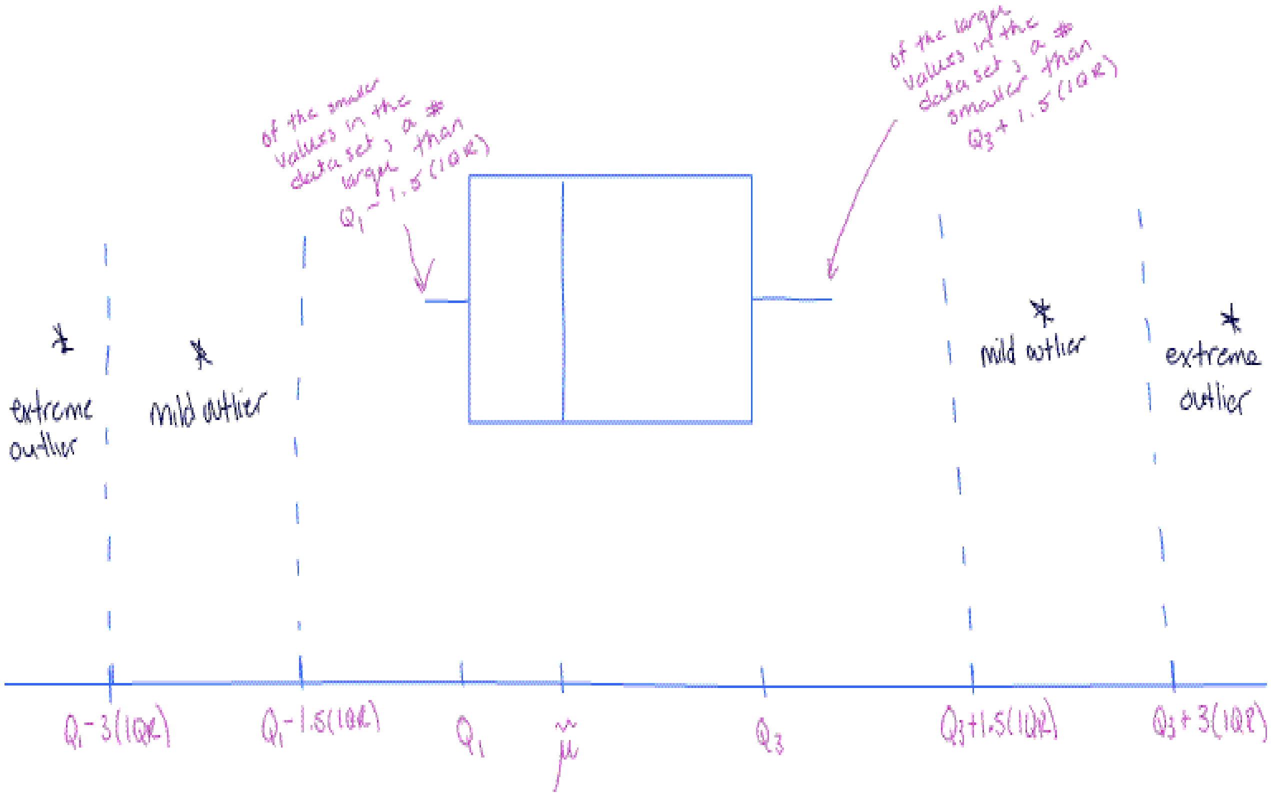
smallest x_i lower quartile median upper quartile largest x_i

What is the purpose of a boxplot? To visually see where the data points gather.

Definition

Any observation farther than 1.5 IQR from the closest quartile is an outlier. An outlier is extreme if it is more than 3 IQR from the nearest quartile, and it is mild otherwise.

The following is a display of a sample data boxplot:


of the smaller values in the data set, a # larger than Q1 − 1.5 (IQR)
of the larger values in the data set, a # smaller than Q3 + 1.5 (IQR)
extreme outlier
mild outlier
mild outlier
extreme outlier
Q1 − 3(IQR)
Q1 − 1.5(IQR)
Q1
μ̃
Q3
Q3+1.5(IQR)
Q3 + 3(IQR)

We now know how to find these measurements, so let's apply them to a boxplot!

Example: Given the following sample data, sketch the boxplot:

146 165 171 179 181 184 190 191 192 192 192 193 195 196 196 197 198 199 200 200 201 203 204 205 206 213 215 221 232 247

Note: we need to find the upper and lower quartiles, the median, and the IQR, then we can find the outliers.

Median:

Since our sample size is even (n = 30), the first ordered 15 numbers go in the lower sample and the last 15 numbers go in the upper sample:

Lower sample: 146 165 171 179 181 184 190 191 192 192 192 193 195 196 196

Upper sample: 197 198 199 200 200 201 203 204 205 206 213 215 221 232 247

mild outliers:

extreme outliers:

Boxplot:

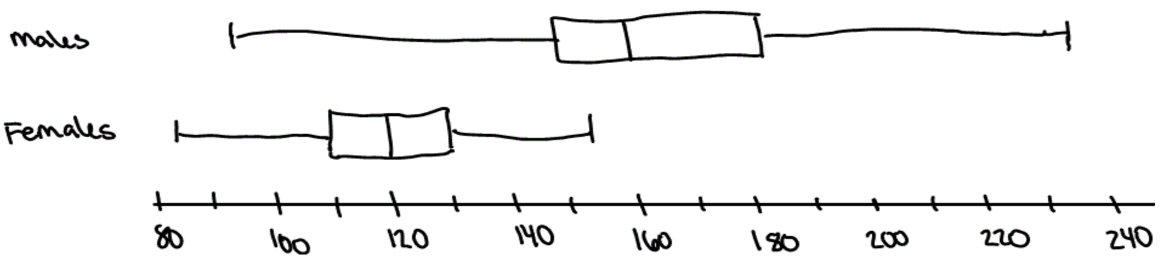
Example: Based on the data and summary statistics below, how many outliers are there? How many mild? How many extreme?

Data: 200, 215, 257, 260, 375, 384, 424, 461, 486, 488, 500, 513, 522, 528, 557, 789, 810, 1018, 1265, 1350, 1499, 2030, 2099

Summary Stats: $\tilde{x} = 513$, $Q_1 = 404$, $Q_3 = 914$

Test your knowledge:

1.) The weights of the male and female students in a class are summarized in the following boxplots:



Which of the following is NOT correct?

(a) About 50% of the male students have weights between 150 and 185 lbs.

(b) About 25% of female students have weights more than 130 lbs.

(c) The median weight of male students is about 162 lbs.

(d) The mean weight of female students is about 120 because of symmetry.

(e) The male students have less variability than the female students.

2.) Rainwater was collected in water collectors at thirty different sites near an industrial basin and the amount of acidity (pH level) was measured. The following stem-and- leaf diagram shows the pH values that ranged from 2.6 to 6.3.

Stems	Leaves
2	679
3	237789
4	1222446899
5	0556788
6	0233

The median acidity is:

(a) 4.2

(b) 4.4

(c) 4.5

(d) 4.6

(e) Average of 15 and 16.