I removed #6 from the assignment for 9.1.

Just looking at the dot plots for data sets R, S and T, we can visualize how spread out the data is, and possibly estimate a 'typical' score. We will discuss more precise ways to determine this information.

The **mean (or arithmetic mean or average)** tells us where the center of the data set is. Conceptually we can talk about the mean of the set $\{7, 5, 7, 3, 8, 6\}$ by thinking of stacks of blocks.
♦ How do we determine the average height of the blocks? (move the blocks to create equal heights)

This shows us that arithmetically, $6 \cdot 6 = 7 + 5 + 7 + 3 + 8 + 6$, or equivalently $6 = \frac{7+5+7+3+8+6}{6}$.

The mean of a collection of data is $\overline{x} = \frac{S}{n}$, where S is the sum of the data, n is the number of values.

♦ Calculate $\overline{x}_R = 1771/24 \approx 73.8$, $\overline{x}_S = 1839/25 \approx 73.6$, $\overline{x}_T = 1637/23 \approx 71.2$

The means for R and S are reasonably close to our guess for the typical score, but T may seem low because the low scores really pull it down.

The **median**, $\hat{x}$, of a collection of data is the middle value in the sorted data set. If there are an odd number of values, this works out easily. If there are an even number of values, you must average the two middle values.

♦ Calculate $\hat{x}_R = \frac{76+76}{2} = 76$, $\hat{x}_S = 73$, $\hat{x}_T = 79$
♦ What is the median of our set from before ($\{7, 5, 7, 3, 8, 6\}$)? (6.5)

The **mode** of a data set is the most frequently occuring value (or values).
♦ The mode of R is 77; the modes of S are $69, 73$ and 76; the mode of T is 40.
♦ Which of these is most representative for each?

Example 9.6: All 12 players on the Uni Hi basketball team played in their 78 to 65 win over Lincoln. Jon Highpockets, Uni Hi's best player scored 23 points in the game. How many points did each of the other players average? ($\frac{78-23}{11} = 55/11 = 5$)

Example 9.7: The owner of a factory earned \$800k last year, the assistant manager earned \$55k, the three secretaries earned \$15k each, and each of the other 15 employees earned \$20k. Compute the mean, median and mode of the salaries of people working at the factory.
($S = 800 + 55 + 45 + 300 = 1200; \overline{x} = 1200/20 = 60, \hat{x} = mode = 20$)

The **range** of a data set is the difference of the highest and lowest values.
♦ Calculate the range of R, S and T? (range(R)=46, range(S)=15, range(T)=56)

In a data set sorted from lowest value to highest value, the **lower quartile**, Q_L, is the median of the lower 'half' of the data. The **upper quartile**, Q_U, is the median of the upper 'half' of the data.
♦ Find Q_L and Q_U for R. ($Q_L = \frac{71+71}{2} = 71$, $Q_U = \frac{77+79}{2} = 78$)

♦ Find $\hat{x}, Q_L$ and Q_U for each data set:

A $\{12, 7, 14, 15, 9, 11, 10, 11, 0, 8, 17\}$ $(0 - 7.5 - 10.5 - 13 - 17)$
B $\{27, 14, 13, 12, 26, 22, 24, 22, 23, 19, 10, 19, 22\}$ $(10 - 13.5 - 22 - 23.5 - 27)$
C $\{16, 22, 20, 15, 12, 14, 16, 14, 21\}$ $(12 - 14 - 16 - 20.5 - 22)$

The **innerquartile range (IQR)** of a data set is IQR$= Q_U - Q_L$.
An **outlier** is a data value which is *less than* $Q_L - (1.5 \cdot \text{IQR})$ or which is *greater than* $Q_U + (1.5 \cdot \text{IQR})$.

♦ Are there any outliers in R? (lower$= 71 - 1.5(7) = 60.5$, upper$= 78 + 1.5(7) = 88.5$, outliers: 46, 92)

The **5-number summary** of a data set is a list of the lowest value, Q_L, $\hat{x}$, Q_U and the largest value.
♦ What is the 5-number summary for R? $(46 - 71 - 76 - 78 - 92)$

A **box and whisker plot** is a pictorial representation of the 5-number summary. The 'whiskers' lead from the smallest value (represented by a tick mark) to Q_L and from Q_U to the largest value (represented by a tick mark). The box is made connecting Q_L to Q_U with a slit at the median.

♦ Draw a box and whisker plot to represent R. (See graphs)

Standard deviation is an even better measure of variability - it tells us how far the data is from the mean, on average.

$$s = \sqrt{\frac{(\overline{x} - x_1)^2 + (\overline{x} - x_2)^2 + \ldots + (\overline{x} - x_n)^2}{n}} \qquad \text{or} \qquad s = \sqrt{\frac{x_1^2 + x_2^2 + \ldots + x_n^2}{n} - \overline{x}^2}$$

♦ Calculate the standard deviation of $\{2, 4, 4, 4, 5, 5, 7, 9\}$ using both definitions. $(\overline{x} = 5, s = 2)$
♦ Compute the fraction (expressed as a percent) of data values that fall within (a) 1 std. dev. of the mean, (b) 2 std. devs. of the mean.
(a) $6/8 = 75\%$ (b) $8/8 = 100\%$

Show how much of R, S, T lie within a few std. deviations of their means.

Section 9.2 HW: #1,3,5b-e,9,13,24,33 (part of HW2)