Chapter 2

2.1 (a) $\overline{Y}$ = 30.563; (b) $\tilde{Y}$ (median) = 33.5; (c) $(\Sigma Y)^2$ = 239,121; (d) ΣY^2 = 16,311; (e) s = 9.543; (f) H_L = 23.5, H_U = 37.5.

2.2 (a) Call the new score *X*. Then X = 15 *x* (Y - $\overline{Y}$)/s_y + 100.
(b) Median(X) = 104.617, H_L = 88.900, and H_U = 110.905.

2.3 (a) If the mean of six scores is 47, the sum must be 6 *x* 47, or 282. However, $\Sigma Y_i = 225$. Therefore, the 6th score must be 282 - 225, or 57.

(b) The mean of the original 5 scores is 45. Adding a score equal to the mean will yield the smallest variance because the variance is the sum of squared deviations about the mean.

2.4 Outliers in box or stem-and-leaf plots and the shape of a normal probability plot suggest a heavy-tailed distribution in data set (a). Both a stem-and-leaf (or a histogram) and a normal probability plot indicate that data set (c) is skewed to the right. Data set (b) appears to be normally distributed.

(d) $\overline{X}$ = 22.70; median(X) = 16.5; $\overline{X}_{.10}$ = 20.62. The two means are considerably higher than the median, suggesting a skew to the right. This can be most clearly seen in a boxplot of the data. Note that the median lies closer to the lower than to the upper hinge, suggesting a long tail to the right.

2.5 (a) $\sum_{i=1}^{5}(X_i + Y_i) = \sum_{i=1}^{5} X_i + \sum_{i=1}^{5} Y_i = 30 + 62 = 92$. (b) $\sum_{i=1}^{5} X_i^2 = 6^2 + 5^2 + \ldots + 11^2 = 232$;

(c) $\left(\sum_{i=1}^{5} X_i\right)^2 = 30^2 = 900$; (d) $\sum_{i=1}^{5} X_i Y_i = (6)(7) + (5)(11) + \ldots + (11)(9) = 315$;

(e)

$$\sum_{i=1}^{5}(X_i + 5Y_i^2 + 27) = \sum_{i=1}^{5}(X_i) + 5\sum_{i=1}^{5}(Y_i^2) + (5)(27) = 30 + (5)(888) + (5)(27) = 4{,}605$$

2.6 (a) $\overline{Y}_{.1} = (7 + 31 + \ldots + 35)/5 = 22$; (b) $\overline{Y}_{2.} = (31 + 15 + 12)/3 = 19.333$;

(c) $\overline{Y}_{..} = (7 + 31 + \ldots + 19 + 4)/15 = 20.4$;

(d) $\sum_{i=1}^{5}\sum_{j=1}^{3} Y_{ij}^2 = 7^2 + 31^2 + \ldots + 19^2 + 4^2 = 9{,}222$;

(e) $\sum_{j=1}^{3} \overline{Y}_{.j}^2 = 22^2 + 30.6^2 + 8.6^2 = 1494.32$

2.7 (a) 286.533; (b) 245.840; (c) 2979.6.

2.8 The mean and median of the X distribution are highest, those for Y are next, and those for Z are lowest. The ranges and standard deviations are in the same order. With respect to shape, the *X* and *Y* distributions are roughly symmetric. Note that in both instances the mean and median are nearly equal to each other and the skewness value is small relative to its standard error (see the table below). In contrast, the *Z* distribution has a straggling

right tail and is clearly skewed in that direction. This impression is confirmed by the ratio of the skewness statistic to its standard error A difference between the shapes of the *X* and *Y* distributions is that the former has outlying scores in both tails. Although we might expect some outliers even when scores are sampled from a normal population, 4 (20%) of 20 scores suggests that the population, though possibly symmetric, is not normally distributed. The probability plot confirms this impression. When the expected value, assuming normality, is plotted against the observed, only the *Y* points consistently lie close to a straight line.

	X	*Y*	*Z*
N of cases	20	20	20
Minimum	10.000	15.000	9.000
Maximum	114.000	76.000	59.000
Range	104.000	61.000	50.000
Median	61.000	49.000	16.500
Mean	62.700	49.550	22.700
Std. Error	5.503	3.715	2.983
Standard Dev	24.609	16.615	13.342
Skewness(G1)	-0.042	-0.310	1.354
SE Skewness	0.512	0.512	0.512
Kurtosis(G2)	1.148	-0.449	1.614
SE Kurtosis	0.992	0.992	0.992

2.9 Standardizing each of the three sets of scores equates their means (at 0) and standard deviations (at 1). The ranges, medians, and trimmed means are not necessarily ordered as they were for the original three distributions. However, each distribution of z scores has the same shape as before the transformation; the skewness and kurtosis values, and their standard errors, as well as the normal probability plot are unchanged. The same cases are outliers as in the original data set. Standard scores are normally distributed only if the original scores are.

2.10 The following tables summarize location and variability for the Royer 3rd and 4th grade multiplication accuracy and response time scores. We also found both box and stem-and-leaf plots to be helpful in comparing gender and grade differences.

	Accuracy (Multacc)			
	Grade 3		Grade 4	
	Female	Male	Female	Male
Mean	74.39	83.18	91.45	88.32
SD	20.35	14.54	7.52	17.11
Median	72.22	84.21	92.86	93.33
H-Spread	17.68	18.26	10.05	10.53

	Response Time (Multrt)			
	Grade 3		Grade 4	
	Female	Male	Female	Male
Mean	5.05	5.10	3.58	3.76
SD	2.59	2.51	1.93	2.69
Median	4.24	4.48	3.51	2.98
H-Spread	2.65	4.50	2.49	2.57

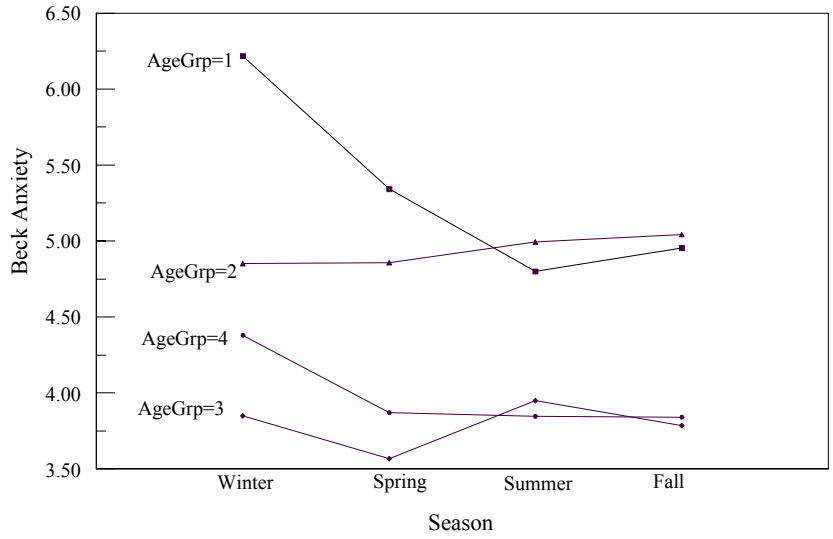
Average male accuracy scores, whether reflected in means or medians, are clearly higher in fourth grade. Standard deviations (*SD*) indicate that the girls' accuracy scores in the third grade are more variable but a consideration of density plots (box plots or stem-and-leaf plots) indicates that this is due to two very low outlying scores. The fact that the *H*-spreads are very similar confirms this impression. In the 4th grade, both distributions shift upward but the girls' average performance matches that of the boys. The girls' mean accuracy is actually somewhat higher but this is apparently due to the presence of two very low outliers in the boys data set. This is consistent with finding that the medians are quite similar and that the male SD is much larger than the female but the *H*-spreads are quite similar. With skewed distributions such as these, the median and *H*-spread, together with a knowledge of outliers, present a more useful summary of the data than do the mean and standard deviation.

In the third grade, male and female average times are quite similar. Medians are below means reflecting a skew to the right because of a few long reaction times. The *H*-spread, but not the SD, indicates that the middle 50% of male scores are more spread out. Both sexes respond more quickly in the fourth grade. The means indicate little difference between boys' and girls' averages but the male median is about a half second faster. The discrepancy between means and medians makes sense when we note two high outliers in the male distribution. This also contributes to the higher standard deviation. Box plots clearly show that the male times are generally faster.

2.11 (a) Both mean and median depression scores (Beck_D) increase noticeably as *Sayhlth* scores increase from 1 to 4 (higher *Sayhlth* scores indicate poorer self-rating of health). Although we have not performed a significance test, the size of the differences and the large numbers of scores suggest that the effect will hold for other samples from the same population.

(b) Winter depression means and medians (Beck_D1 scores) are highest in categories 1 - 3. However, individuals who rate themselves in fair health (*Sayhlth* = 4) have a somewhat higher average score in the fall season. This is particularly noticeable in the median scores.

2.12 (a) The line graph is preferable in that it more clearly reveals differences among age groups in trends over seasons.



(b) Two aspects of the graph are notable. First, the younger age groups (Agegrp = 1 and 2) have higher mean Beck anxiety scores than the older groups. Second, this is particularly pronounced in the winter season; although three of the four groups are most anxious then, this is markedly so for the youngest group.

(c) Considering the influence of outliers does not change our conclusions. Median trends over seasons within each age group show a trend similar to that for the means, though the differences among age groups are not quite as large when the median is viewed instead of the mean.

2.13 (a) Relative to the class, the student's performance declined. The *z* score for Test 1 is $z_1 = (41 - 38.6)/4.616 = .520$ whereas $z_2 = (51 - 46.84)/9.496 = .438$

(b) A score of 52 is the lowest integer value that transform the test 2 score into a *z* score exceeding .52. We arrive at this by solving $(X - 46.84)/9.496 > .52$. One point more on Test 2 would have yielded a *z* score of .543.

(c) The almost identical values of means and medians on each test suggest that the distributions are symmetric. This is confirmed by obtaining box plots. Finally, normal probability (*Q*-*Q*) plots indicate that the points lie fairly close to a straight line. A few of the upper and lower points depart slightly from a straight line but this might occur by chance in any sample drawn from a normal population.

(d) The correlation between the two sets of test scores is .543. A scatterplot shows that test 2 scores increase as test 1 scores do but there is considerable variability.

(e) Using Equations 2.10*a* and 2.10*b,* we have $b_1 = r\, x\, sy/s_x = 1.117$ and $b_0 = \bar{Y} - b_1\bar{X} = 3.737$. The regression equation is $\hat{Y} = b_0 + b_1X$; substituting for b_0 and b_1, and letting X = 40, the predicted value is 48.417.

2.14 (a) The median payroll is about $200,000 less than the mean payroll and the skew statistic is about twice its standard error. These results suggest that the data are not normally distributed and imply a straggling right tail. However, on examining the histogram and the Q-Q plot, a slightly different picture emerges. The data are approximately normally distributed except for one payroll that is about $200,000 higher than the next highest payroll.

(b) The American League average payroll is considerably higher than that of the National

League. However, the fact that the median AL payroll is higher than that of the NL indicates that the difference is not just due to one or two payrolls. There is also more variability in the AL; although, as indicated by the mean and median, salaries tend to be higher in the AL, that league also has three of the four lowest payrolls. The skew statistic is also larger for the AL; the distribution for the NL is close to the normal.

(c) A scatterplot suggests that there is an overall positive relationship between average team payrolls in 1986 and 2007. The correlation coefficient is .505, significant at the .01 level. However, when we break it down by league, the correlation is much higher for the NL (.728) than for the AL (.405). One reason for the lower AL correlation reflects the fact that 4 teams in the bottom half of payrolls in 1986 (Angels, White Sox, Mariners, and Tigers) were in the top half in 2007.

2.15 (a) The correlation is .476. and is significantly different from zero ($p < .001$). Of course, this moderately high correlation does not tell us the reason. It may reflect pitchers being more careful with players who hit more home runs, or such hitters having a better ability to distinguish between strikes and balls, or other factors.

(b)The correlation, .074,is very weak and, despite the large number of observations, is not significantly different from zero. There is no support for the existence of a relationship between offensive and defensive abilities.

(c) If you add the BAs and divide by the number of players, the Boston Red Sox mean is .268 and the NY Mets mean is .271. However, this method gives equal weight to all the individual averages even though there is a difference in the number of at bats. We should add all the hits (*H*s) for the team and divide by the total of team at bats (*AB*s). Then the Red Sox mean is .276 and the Met mean is .274

(d) The correlation is .682. Players with fewer at bats tend to have lower averages. This is why the Red Sox mean team batting average was higher (.276) when the mean was correctly calculated, taking each players ABs into consideration. When the averages are weighted by the number of at bats, the low averages have less affect on the team mean average.