MULTIPLE CHOICE.

1. The correct answer is (a). I is correct since the mean is pulled in the direction of the large maximum value, 150 (well, large compared to the rest of the numbers in the set). II is not correct because the mode is $y$—there are three $y$s and only two 26s. III is not correct because 150 is an outlier (you can't actually compute the upper boundary for an outlier since the third quartile is $y$, but even if you use a larger value, 33, in place of $y$, 150 is still an outlier).

2. The correct answer is (e).

$z = \dfrac{70-65}{2.6} = 1.92 \rightarrow$ percentile $= 0.9726$ (see drawing below):
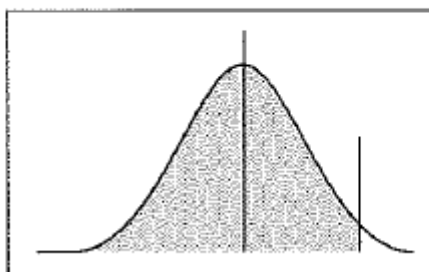


3. The correct answer is (d). The effect on the mean of a dataset of subtracting the same value is to reduce the old mean by that amount (that is, $\mu_{x-k} = \mu_x - k$). Because the original mean was 19, and 19 has been subtracted from every term, the new mean is 0. The effect on the standard deviation of a dataset of dividing each term by the same value is to divide the standard deviation by that value, that is,

$$\sigma_{x/_k} = \frac{\sigma_x}{k}.$$

Because the old standard deviation was 4, dividing every term by 4 yields a new standard deviation of 1. Note that the process of subtracting the mean from each term and dividing by the standard deviation creates a set of $z$-scores

$$z_x = \frac{x - \bar{x}}{s}$$

so that any complete set of $z$-scores has a mean of 0 and a standard deviation of 1. The shape is normal since any linear transformation of a normal distribution will still be normal.

4. The correct answer is (b). The maximum length of a "whisker" in a modified boxplot is $1.5(IQR) = 1.5(40 - 18) = 33$.

5. The correct (best) answer is (d). Using Table A, the area under a normal curve between 63 and 75 is 0.6247 ($z_{63} = -1.5 \Rightarrow A_1 = 0.0668$, $z_{75} = 0.5 \Rightarrow A_2 = 0.6915 \Rightarrow A_2 - A_1 = 0.6247$). Then $(0.6247)(5,000) = 3123.5$. Using the TI-83/84, `normal-cdf(63,75,72,6) × 5000 = 3123.3`.

6. The correct answer is (b). Since we do not know that the empirical rule applies, we must use Chebyshev's rule.

Since $72 - k(6) = 58$, we find $k = 2.333$. Hence, there are at most $\dfrac{1}{2.333^2}\% = 18.37\%$ of the scores less than 58. Since there are 5000 scores, there are at most $(0.1837)(5,000) = 919$ scores less than 58. Note that it is unlikely that there are this many scores below 58 (since some of the 919 scores could be more than 2.333 standard deviation *above* the mean)—it's just the strongest statement we can make.

7. The correct answer is (e). The graph is clearly not symmetric, bi-modal, or uniform. It is skewed to the right since that's the direction of the "tail" of the graph.

8. The correct answer is (a). The median is resistant to extreme values, and the mean is not (that is, extreme values will exert a strong influence on the numerical value of the mean but not on the median). II and III involve statistics equal to or dependent upon the mean, so neither of them is resistant.

9. The correct answer is (d). $z = 1.65 = \dfrac{43 - 40}{\sigma} \Rightarrow \sigma = \dfrac{3}{1.65} = 1.82$.

10. The correct answer is (b). A score at the 90th percentile has a z-score of 1.28. Thus, $z_x = \dfrac{x - 2.05}{1} = 1.28 \Rightarrow x = 3.33$.

FREE RESPONSE.

1. Using the calculator, we find that $\bar{x} = 29.78$, $s = 11.94$, Q1 = 21, Q3 = 37. Using the 1.5(IQR) rule, outliers are values that are less than $21 - 1.5(37 - 21) = -3$ or greater than $37 + 1.5(37 - 21) = 61$. Because no values lie outside of those boundaries, there are no outliers by this rule.

   Using the $\bar{x} \pm 2s$ rule, we have $\bar{x} \pm 2s = 29.78 \pm 2(11.94) = (5.9, 53.66)$. By this standard, the year he hit 54 home runs would be considered an outlier.

2. (a) is a property of the *standard normal distribution*, not a property of normal distributions in general. (b) is a property of the normal distribution. (c) is not a property of the normal distribution—*Almost* all of the terms are within four standard deviations of the mean but, at least in theory, there are terms at any given distance from the mean. (d) is a property of the normal distribution—the normal curve is the perfect bell-shaped curve. (e) is a property of the normal distribution and is the property that makes this curve useful as a probability density curve.

3.

| | | | |
|---|---|---|---|
| 1 | 3589 | 1 | 3 |
| 2 | 12337 | 1 | 589 |
| 3 | 01457 | 2 | 1233 |
| 4 | 02 | 2 | 7 |
| 5 | 24 | 3 | 014 |
| | | 3 | 57 |
| | | 4 | 02 |
| | | 4 | |
| | | 5 | 24 |

What shows up when done by 5 rather than 10 is the gap between 42 and 52. In 16 out of 18 years, Mantle hit 42 or fewer home runs. He hit more than 50 only twice.

4.  $\bar{x} = 76.4$ and $s = 10.17$.

$$z_{84} = \frac{84-76.4}{10.17} = 0.75. \quad z_{89} = \frac{89-76.4}{10.17} = 1.24.$$

Using the Standard Normal Probability table, a score of 84 corresponds to the 77.34th percentile, and a score of 89 corresponds to the 89.25th percentile. Both students were in the top quartile of scores after the program and performed better than all but one of the other students. We don't know that there is a cause-and-effect relationship between the pilot program and the high scores (that would require comparisons with a pretest), but it's reasonable to assume that the program had a positive impact. You might wonder how the student who got the 98 did so well!

5.
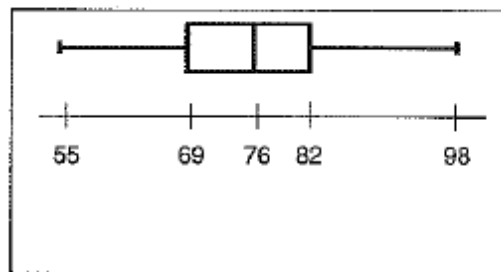


The most distinguishing feature is that the range (43) is quite large compared to the middle 50% of the scores (13). That is, we can see from the graph that the scores are packed somewhat closely about the median. The shape of a histogram of the data would be symmetric and mound shaped.

6.  $z_{0.370} = \dfrac{0.370-0.265}{0.032} = 3.28. \Rightarrow$ Area of the left of 3.28 is 0.9995.

That is, Bond's average in 2002 would have placed him in the 99.95th percentile of batters.

7. There are 18 values in the stemplot. The median is 17 (actually between the last two 7s in the row marked by the (5) in the count column of the plot —it's still 17). Because there are 9 values in each half of the stemplot, the median of the lower half of the data, Q1, is the 5th score from the top. So, Q1 = 14. Q3 = the 5th score counting from the bottom = 24. Thus, IQR = 24 − 14 = 10.

8. There are 3 values in the first bar, 6 in the second, 2 in the third, 9 in the fourth, and 5 in the fifth for a total of 25 values in the dataset. Of these, 3 + 6 + 2 = 11 are less than 3.5. There are 25 terms altogether, so the proportion of terms less than 3.5 is 11/25 = 0.44.

9. With the exception of the one outlier for Bonds, the most obvious thing about these two is just how similar the two are. The medians of the two are almost identical and the IQRs are very similar. The data do not show it, but with the exception of 2001, the year Bonds hit 73 home runs, neither batter ever hit 50 or more home runs in a season. So, for any given season, you should be overjoyed to have either on your team, but there is no good reason to choose one over the other. However, if you based your decision on who had the most home runs in a single season, you would certainly choose Bonds.

10. Let $x$ be the value in question. Because we do not want to be in the top 20%, the area to the left of $x$ is 0.8. Hence $z_x = 0.84$ (found by locating the nearest table entry to 0.8, which is 0.7995 and reading the corresponding $z$-score as 0.84). Then

$$z_x = 0.84 = \frac{x-185}{25} \Rightarrow x = 206.$$

11. $\bar{x}$ = $3.36 million, $s$ = $1.88 million, $Med$ = $3.35 million, IQR = $2.6 million. A boxplot of the data looks like this:
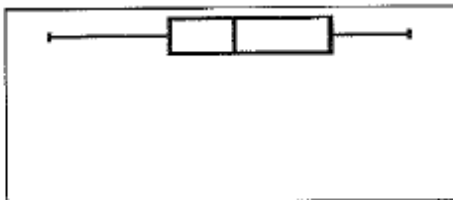


The fact that the mean and median are virtually the same, and that the boxplot shows that the data are more or less symmetric, indicates that either set of measures would be appropriate.

12. The easiest way to do this is to use the calculator. Put the age data in L1 and the frequencies in L2. Then do 1-Var Stats L1,L2 (the calculator will read the second list as frequencies for the first list).

   - The mean is 2.48 years, and the median is 2 years. This indicates that the mean is being pulled to the right—and that the distribution is skewed to the right or has outliers in the direction of the larger values.
   - The standard deviation is 2.61 years. Because one standard deviation to left would yield a negative value, this also indicates that the distribution extends further to the right than the left.

   - A histogram of the data, drawn on the TI–83/84, is drawn below. This definitely indicates that the ages of these pennies is skewed to the right.
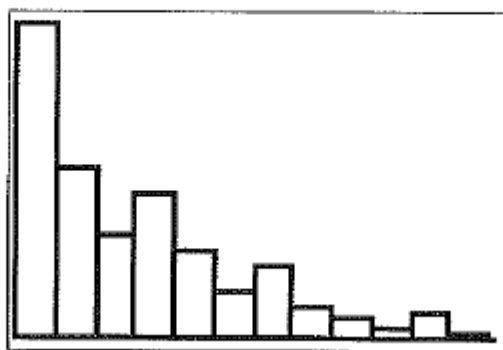


13. Since we don't know the shape of the distribution of coin values, we must use Chebyshev's rule to help us solve this problem. Let $k$ = the number of standard deviations that 170 is above the mean. Then $130 + k(15) = 170$. So, $k \approx 2.67$. Thus,

   at most $\dfrac{1}{k^2} \approx \dfrac{1}{(2.67)^2} \approx 0.14$ , or 14%, of the coins are valued at more than $170. Her

   requirement was that $\dfrac{225}{1450} \approx 0.155$ , or 15.5%, of the coins must be valued at more

   than $170. Since at most 14% can be valued that highly, she should not buy the collection.

14. The new mean is $5(35 - 10) = 125$.
   The new median is $5(33 - 10) = 115$.
   The new variance is $5^2(6^2) = 900$.
   The new standard deviation is $5(6) = 30$.
   The new IQR is $5(12) = 60$.

15. First we need to find the *proportion* of women who would be less than 62" tall:

$$z_{55} = \frac{62-65}{2.5} = -1.2 \Rightarrow \text{Area} = 0.1151.$$

So 0.1151 of the terms in the distribution would be less than 62". This means that 0.1151(300) = 34.53, so you would expect that 34 or 35 of the women would be less than 62" tall.

16. a, c, and d are properties of the standard deviation. (a) serves as a definition of the standard deviation. It is independent of the number of terms in the distribution in the sense that simply adding more terms will not necessarily increase or decrease s. (d) is another way of saying that the standard deviation is independent of the mean—it's a measure of spread, not a measure of center.

The standard deviation is *not* resistant to extreme values (b) because it is based on the mean, not the median. (e) is a statement about the interquartile range. In general, unless we know something about the curve, we don't know what proportion of terms are within 2 standard deviations of the mean.

17. For these data, Q1 = $2.3 *million,* Q3 = $4.9 *million.* To be an outlier, Erick would need to make at least 4.9 + 1.5(4.9 − 2.3) = 8.8 million. In other words, he would need a $2.6 million dollar raise in order to have his salary be an outlier.

18. You need to estimate the median and the quartiles. Note that the histogram is skewed to the left, so that the scores tend to pack to the right. This means that the median is to the right of center and that the boxplot would have a long whisker to the left. The boxplot looks like this:
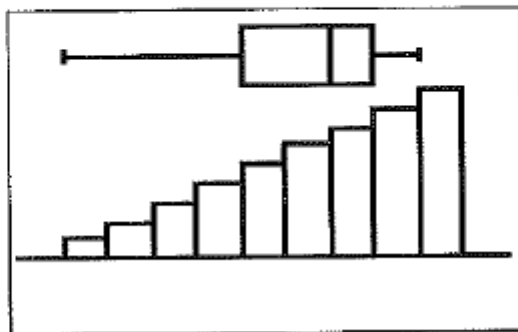


19. If you standardize both scores, you can compare them on the same scale. Accordingly,

$$z_{80} = \frac{80-72}{6} = 1.333, \quad z_{76} = \frac{76-65}{8} = 1.375.$$

Nathan did slightly, but only slightly, better on the second test.

20.

$$s = 20 = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}} = \sqrt{\frac{13600}{n-1}}$$

$$20^2 = \frac{13600}{n-1} \Rightarrow n = 35$$

$$\bar{x} = \frac{\sum x}{n} = \frac{245}{35} = 7$$