COMPARING AND REPRESENTING DATA

Data distributions can be represented graphically with histograms and boxplots. For assistance with box plots and histograms, see the Parent Guide with Extra Practice for CPM Core Connections Course 1, 2, and 3, which is available free for download at cpm.org. For more details on describing spread and data distributions see the Math Notes boxes in Lessons 1.2.2 and 1.2.3.

Two distributions of data can be compared by comparing their center, shape, spread, and outliers.

The center, or “typical” value, of a data distribution can be described by the median. If the distribution is symmetric and has no outliers, the mean can be used to describe the center.

The spread of a distribution can be described with the interquartile range (IQR) or the standard deviation, as described in the Math Notes box in Lesson 1.2.2. Since the standard deviation is based upon the mean, it should be used only to describe the spread of distributions that are symmetric and without outliers. Since gathering data for entire populations is often impractical, most of the data sets we analyze are samples. In this course, in general, calculate the sample standard deviation.

Example 1

University professors are complaining that the English Literature classes at community colleges are not demanding enough. Specifically, the university professors claim that community college literature courses are not assigning enough novels to read. A community college statistics student collected the following data from 42 universities and community college literature courses in the state. Compare the number of novels read in the two types of colleges.

Number of novels assigned in community college literature courses:
13, 10, 15, 12, 14, 9, 11, 15, 12, 14, 9, 10, 13, 15, 12, 9, 11, 15, 12, 10, 15, 14

Checksum 270

Number of novels assigned in university literature courses:
11, 8, 14, 13, 25, 11, 7, 13, 8, 16, 11, 10, 20, 7, 8, 13, 14, 16, 18, 10

Checksum 253

Solution:

Any analysis of data distributions should begin with a graphical representation of the data. A bin width of two was chosen for the histograms that follow. So that the distributions can be compared, both graphs have the same scale on the x-axis, and the boxplots are graphed above the histograms.
The checksum is used to verify that data has been entered into the graphing technology correctly. The sum of the data set, as determined by the statistical functions of the calculator, should match the given checksum value.

When comparing the distributions, the center, shape, spread, and outliers should be considered. Since neither of the distributions is nearly symmetric, and one of the distributions has an outlier, it would not be appropriate to use the mean nor standard deviations to compare. The five number summaries are shown below each graph.

Center: Both types of colleges assign the same median of 12 novels.

Shape: The distribution for community colleges is skewed, with a low of 8 to 9 novels and increasing to a peak at 14 to 15 novels. The distribution at universities is skewed in the other direction, with a peak at 10 to 11 novels.

Spread: The variability in the number of novels assigned at the community college level is much less than the variability between courses at the university level. The IQR for community colleges is 4 novels (14 – 10 = 4), while the university IQR of 6 (15 – 9 = 6) is one-and-a-half times as wide.

Outliers: One course at a university is an outlier; 25 books are assigned in that course. Twenty-five books is far away from the bulk of the university courses. The TI-83+/TI-84+ calculator can mark an outlier or a boxplot with dots.

Conclusions: The university professors claim that their courses are more demanding because they assign more novels. However that data does not bear this claim out. Twenty-five percent of university courses assign more novels than any of the community college courses (the right “whisker,” or the top 25% of the courses, for universities is beyond the entire boxplot for community colleges). But just as dramatically, 25% of the university classes assign fewer books than any of the community colleges (the left “whisker,” or lowest 25%, for universities is below the entire boxplot for community colleges). Furthermore, the median number of novels assigned at the two universities is the same (12 books). Community colleges are more consistent from course-to-course in the number of novels they assign (IQR is 4) than are the universities (IQR is 6).
Example 2

A rabbit breeder kept track of the number of offspring from five does (female rabbits) this year. The does had: 243, 215, 184, 280, and 148 kits (baby rabbits) respectively. Show how to calculate the mean and standard deviation number of kits per doe without using the statistical functions of a calculator.

The mean is \( \frac{243 + 215 + 184 + 280 + 148}{5} = 214 \) kits.

To calculate the sample standard deviation, first calculate the distance each data point is from the mean:

\[
\begin{align*}
243 &- 214 = 29 \\
215 &- 214 = 1 \\
184 &- 214 = -30 \\
280 &- 214 = 66 \\
148 &- 214 = -66
\end{align*}
\]

Then, calculate each of the distances squared:

\[
\begin{align*}
29^2 & = 841 \\
1^2 & = 1 \\
(-30)^2 & = 900 \\
66^2 & = 4356 \\
(-66)^2 & = 4356
\end{align*}
\]

Divide by one less than the number of data points: \( \frac{841+1+900+4356+4356}{5-1} = 2613.5 \)

Take the square root: \( \sqrt{2613.5} \approx 52.122 \)

Since the precision of the original measurements was an integer, the final result should also be an integer. The mean number of kits per doe is 214 with a standard deviation of 52 kits.

Problems

1. Different types of toads tend to lay different numbers of eggs. The following data was collected from two different species. Compare the number of eggs laid by American toads to the number laid by Fowler’s toads. Is it appropriate to summarize the distributions by using mean and standard deviation? Use a bin width of 250 eggs.

   American toads: 9100, 8700, 10300, 9500, 7800, 8900, 9200, 9300, 8900, 8300, 9400, 8000, 9000, 8400, 9700, 10000, 8600, 8900, 9900, 9300  checksum 181,200

   Fowler’s toads: 9500, 9100, 9400, 8800, 9000, 8400, 9200, 9200, 8900, 9100, 8600, 9200, 8700, 9800, 9300, 8800, 9200, 9300, 9000, 9100  checksum 181,600

2. Without using the statistical functions on your calculator, calculate the sample standard deviation of the number of eggs laid by each of the first five American toads.
3. A psychologist collected the following data for the age at which children started crawling:
   low weight babies: 10, 12, 11, 11, 7, 13, 10, 12, 11, 13, 10, 11, 15, 11, 14, 10 months
   checksum 181
   average weight babies: 7, 6, 13, 9, 8, 7, 5, 7, 9, 8, 10, 8, 11, 7, 7, 10, 6, 8, 7, 6, 12, 8, 7 months
   checksum 186

   Do low birth weight babies start crawling at a later age than babies born at an average weight?

4. Compare the amount of time the flavor lasted for people chewing brand “10” chewing gum to the amount of time the flavor lasted in “Strident” chewing gum. See the graph at right. Estimate the mean for each type of gum.

Answers

1. Mean and standard deviation are appropriate statistics because both distributions are fairly symmetric with no outliers.

   Both types of toads lay a mean of between 9000 and 9100 eggs. Both distributions are single-peaked and symmetric with no apparent outliers. However there is much greater variability in the number of eggs that American toads lay. The sample standard deviation for American toads is about 654 while the sample standard deviation for Fowler’s toads is only about half as much, about 324 eggs.
2. \[ \sqrt{\frac{20^2 + (-380)^2 + 1220^2 + 420^2 + (-1280)^2}{5 - 1}} \approx 928 \text{ eggs} \]

3. 

The median and IQR will be used to compare statistics since mean and standard deviation are not appropriate—both distributions are skewed and one has an outlier.

The median age at which low-weight babies start crawling is 11 months, while the median age for average-weight babies is 8 months.

Both distributions are single-peaked and skewed. The low-weight babies appear to have an outlier at 7 months, although the calculator does not identify it as a true outlier. The average-weight babies have an outlier at 13 months.

The variability in the crawling age is roughly the same for low-weight babies (IQR is 2.5 months) as for average-weight babies (IQR is 2 months).

Low-weight babies have their development delayed by about 3 months. About 75% of low-weight babies have not even started crawling after 75% of average-weight babies are already crawling.

4. The median for both types of gum was about 18 minutes of flavor time. The times for “10” were skewed, while the times for Strident were symmetric. The lower half of the distributions for both gums was the same. But there was much more variability in the upper half of people chewing “10” than in the upper half of Strident. Indeed, more than 25% of “10” chewers reported flavor lasting longer than any of the Strident chewers. Neither gum had outliers in flavor time.

There was more variability in flavor time for “10”—the IQR was about 9 minutes \((25 - 16 = 9)\). The IQR of 4 minutes \((20 - 16 = 4)\) for Strident was less than half that of “10”. That variability is an advantage. If you chew “10,” you will probably be no worse off than chewing Strident, and you could have much longer flavor.

The mean for Strident is about the same as the median since the distribution is symmetric—about 18 minutes. But the mean for “10” is longer than 18 minutes due to the skew in the shape—maybe 22 minutes or so.