

Notes: Describing Distributions Numerically

When describing distributions, we need to discuss shape, center, and spread. How we measure the center and spread of a distribution depends on its shape. The center of a distribution is a “typical” value. If the shape is unimodal and symmetric, a “typical” value is in the middle. If the shape is skewed, however, a “typical” value is not necessarily in the middle.

For skewed distributions, use the median to determine the center of the distribution and the interquartile range to describe the spread of the distribution.

The median:

- ❖ is the middle data value (when the data have been ordered) that divides the histogram into two equal areas
- ❖ has the same units as the data
- ❖ is resistant to outliers (extreme data values)

The range:

- ❖ is the difference between the maximum value and the minimum value
- ❖ is a number, NOT an interval
- ❖ is sensitive to outliers

The interquartile range (IQR):

- ❖ contains the middle 50% of the data
- ❖ is the difference between the lower (Q1) and upper (Q3) quartiles
- ❖ is a number, NOT an interval
- ❖ is resistant to outliers

The Five-Number Summary gives: minimum, lower quartile, median, upper quartile, maximum.

A graphical display of the five-number summary is called a boxplot.

How many hours, on average, do you spend watching TV per week? _____ Collect data from the entire class and record the values in order from smallest to largest. Calculate the five-number summary:

Construct both a histogram and a boxplot (using the same scale). Compare the displays.

Average Number of Hours per Week Spent Watching TV

For symmetric distributions, use the mean to determine the center of the distribution and the standard deviation to describe the spread of the distribution.

The mean:

- ❖ is the arithmetic average of the data values
- ❖ is the balancing point of a histogram
- ❖ has the same units as the data
- ❖ is sensitive to outliers
- ❖ is given by the formula $\bar{x} = \frac{\sum x}{n}$

The standard deviation:

- ❖ measures the “typical” distance each data value is from the mean
- ❖ Because some values are above the mean and some are below the mean, finding the sum is not useful (positives cancel out negatives); therefore we first square the deviations, then calculate an adjusted average. This is called the variance. This statistics does not have the same units as the data, since we squared the deviations. Therefore, the final step is to take the square root of the variance, which gives us the standard deviation.
- ❖ is given by the formula $s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$
- ❖ is sensitive to outliers, since its calculation involves the mean

Find the mean and standard deviation of the average number of hours spent watching TV per week for this class.