

Summarizing Numerical Data: Center and Spread

- **midrange**

the number halfway between the smallest and largest data value is an estimate of the **center** of the distribution; often a poor estimate, since it is highly **sensitive** to the size of outlier values

- **median**

the middle observation in a sorted list of the data values (for an even number of values, average the *two* middle observations); a better estimate of center since it is **resistant** to the effects of outliers

- **range**
the difference between the largest and smallest data values is an estimate of the **spread** in the data; again, often a poor estimate, since it is highly sensitive to the size of outlier values
- **lower/upper quartiles** (Q_1 and Q_3)
the observations which are one quarter (Q_1) and three quarters (Q_3) of the way up the list, the median values of the half of the data located below/above the median; also, the 25th and 75th **percentiles** of the data
- **interquartile range (IQR)**
the difference $Q_3 - Q_1$ between the two quartiles; a better measure of the spread in the data since it is resistant to the presence of outliers

The **five-number summary** of a data set:

- minimum value,
- lower quartile,
- median,
- upper quartile, and
- maximum value.

[TI83: STAT CALC 1-VarStats]

- **boxplot**

graphical display of the five-number summary formed by drawing a box over a number line so that the sides of the box are located at the two quartiles, a line through the box is drawn at the location of the median, and “whiskers” are extended to the minimum and maximum data values; multiple data sets can be compared by displaying side-by-side boxplots

- **1.5 IQR rule**

standard rule of thumb for identifying outliers: any data value more than 1.5 IQR below the lower quartile or more than 1.5 IQR above the upper quartile may be tagged as an outlier, and data values more than 3 IQR away are tagged as extreme outliers

[TI83: STATPLOT, ZoomStat]

Center and Spread for Symmetric Distributions

- **mean** (\bar{y})
the arithmetical average (where the distribution “balances”)

$$\bar{y} = \frac{\sum y}{n} ;$$

in skewed distributions, the mean is pulled in the direction of the skewness (the longer tail)

- **deviation from the mean** ($y - \bar{y}$)
the difference between a data value and the mean of all the data
- **variance** (s^2)
estimates an average squared deviation from the mean

$$s^2 = \frac{\sum (y - \bar{y})^2}{n - 1}$$

- **standard deviation** (s)
measure of spread that estimates the size of a typical deviation from the mean

$$s = \sqrt{\frac{\sum (y - \bar{y})^2}{n - 1}} ;$$

like the mean, sensitive to outliers