1.3.7 Measuring Spread: The Standard Deviation

The five-number summary is not the most common numerical description of a distribution. That distinction belongs to the combination of the **mean to measure center** and the **standard deviation to measure spread**.

The ***standard deviation*** s_x measures the *average* distance of the observations from their mean.

Procedure: The *standard deviation* is calculated by finding the average of the squared distances and then taking the square root. The average *squared difference* is called the ***variance***.

Symbols/Formula:

$$s_x = \sqrt{\frac{1}{n-1}\sum(x_i - \bar{x})^2}$$

Example: These are the foot lengths (in cm) for a random sample of seven 14-year-olds from the United Kingdom:
25 22 20 25 24 24 28
The mean foot length is 24 cm. $\bar{x} = 24\text{cm}$

x	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
25	$25-24=1$	$(1)^2=1$
22	$22-24=-2$	$(-2)^2=4$
20	$20-24=-4$	$(-4)^2=16$
25	$25-24=1$	$(1)^2=1$
24	$24-24=0$	$(0)^2=0$
24	$24-24=0$	$(0)^2=0$
28	$28-24=4$	$(4)^2=16$

$$\sum(x_i-\bar{x})^2 = 1+4+16+1+0+0+16 = 38$$

$$\frac{1}{n-1}\sum(x_i-\bar{x})^2 = \frac{1}{7-1}(38) = \frac{38}{6}$$

$$s_x = \sqrt{38/6} = \sqrt{6.\bar{3}} \approx 2.52\text{ cm}$$

2.52 cm is roughly the average distance each foot length is from the mean of 24 ft.

Many calculators report **two standard deviations**, giving you a choice of **dividing by n or by n – 1**. The former is usually labeled σ_x, the symbol for the **standard deviation of a population**. If your data set consists of the entire population, then it's appropriate to use σ_x. **More often, the data we're examining come from a sample. In that case, we should use s_x.** More important than the details of calculating s_x are the properties that determine the usefulness of the standard deviation:

Properties of the Standard Deviation

- s_x measures spread about the mean and should be used only when the mean is chosen as the measure of center.
- **s_x is always greater than or equal to 0. $s_x = 0$ only when there is no variability.** This happens only when all observations have the same value. Otherwise, $s_x > 0$. **As the observations become more spread out about their mean, s_x gets larger.**
- **s_x has the same units of measurement as the original observations.** For example, if you measure metabolic rates in calories, both the mean X and the standard deviation s_x are also in calories. This is one reason to prefer s_x to the variance, which is in squared calories.
- Like the mean X, s_x is not resistant. A few outliers can make s_x very large.

The use of squared deviations makes s_x even more sensitive than X to a few extreme observations.

Check your understanding, p. 64.

The heights (in inches) of the five starters on a basketball team are 67, 72, 76,76, and 84.

1. Find and interpret the mean.

$\bar{x} = 75$ $\left(\frac{67+72+76+76+84}{5}\right) = 75$ If the total of all the heights was the same, they would each be 75 inches tall.

2. Make a table that shows, for each value, its deviation from the mean and its squared deviation from the mean.

x	$x-\bar{x}$ (Deviation)	$(x-\bar{x})^2$ Squared Deviation
67	$67-75 = -8$	$(-8)^2 = 64$
72	$72-75 = -3$	$(-3)^2 = 9$
76	$76-75 = 1$	$(1)^2 = 1$
76	$76-75 = 1$	$(1)^2 = 1$
84	$84-75$ 9	$(9)^2 = 81$
Total	0	156

3. Show how to calculate the variance and standard deviation from the values in your table.

Variance $= \frac{\Sigma(x-\bar{x})^2}{n-1} = \frac{(64+9+1+1+81)}{5-1} = \frac{156}{4} = 39 \text{ in}^2$

units are squared

Standard Deviation $= \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}} = \sqrt{39} \approx 6.24 \text{ in}$

(square root of variance)

units are the same as original data

4. Interpret the meaning of the standard deviation in this setting.

On average, the players' heights vary about 6.24 inches from the mean of 75in.

1.3.9 Choosing Measure of Center and Spread

We now have a choice between **two descriptions of the center and spread of a distribution**:

1) the median and IQR,

or

2) $\bar{X}$ and s_x.

↑ standard deviation

Because $\bar{X}$ and s_x are sensitive to extreme observations, they can be misleading when a distribution is strongly skewed or has outliers. In these cases, the median and IQR, which are both resistant to extreme values, provide a better summary. We'll see in the next chapter that the mean and standard deviation are the natural measures of center and spread for a very important class of symmetric distributions, the Normal distributions.

Choosing Measures of Center and Spread

- Skewed Distributions: median and IQR are better
- Distributions with strong outliers: Median and IQR are better
- Reasonably symmetric distributions: Mean and Standard Deviation

*****Resistance*****

- Median: More resistant than mean
- IQR: More resistant than range
- IQR: More Resistant than standard deviation

Analyzing Data Sets

From this point on, whenever you are analyzing data sets, in the "Do" step you should:

- Plot the distribution
- Create a numerical summary which includes:
  - Mean
  - Standard deviation
  - 5-Number Summary (min, Q1, median, Q3, max)

Always

Example - Who Texts More—Males or Females? Pulling it all together

For their final project, a group of AP Statistics students investigated their belief that females text more than males. They asked a random sample of students from their school to record the number of text messages sent and received over a two-day period. Here are their data:

Males:	127	44	28	83	0	6	78	6	5	213	73	20	214	28	11
Females:	112	203	102	54	379	305	179	24	127	65	41	27	298	6	130	0

What conclusion should the students draw? Give appropriate evidence to support your answer. (State, Plan, Do, Conclude)

State: Do the data give convincing evidence that females text more than males?

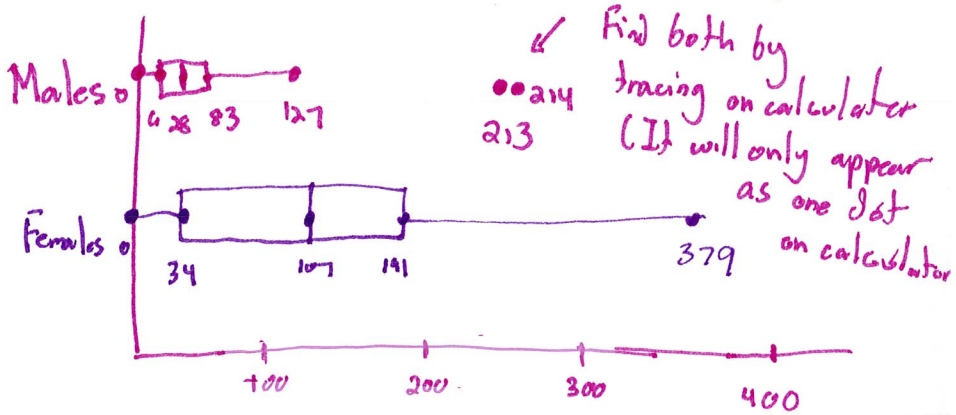
Plan: Make side by side box plots. Calculate one-variable statistics.

Compare shape, center, spread, and outliers for the two distributions.

Do:

Numerical Summary

Males ($n=15$)	Females ($n=16$)
$\bar{x}=62.4$	$\bar{x}=128.3$
$S_x=71.4$	$S_x=116.0$
min $x=0$	min $=0$
$Q_1=6$	$Q_1=34$
med $=28$	med $=107$
$Q_3=83$	$Q_3=191$
max $=214$	max $=379$
IQR $=83-6=77$	IQR $=191-34=157$



Find both by tracing on calculator (It will only appear as one dot on calculator)

* Due to strong skewness + outliers, use median + IQR to discuss center + spread **

Shape: Both distributions are heavily skewed right.

Center: On average females text more than males. (Median for females is 107 compared to males of median of 28). In fact, the median for the females is above the 3rd quartile for the males. This indicates that over 75% of the males texted less than typical (median) female.

Spread: There is much variability in texting among the females than males. The IQR for females (157) is about twice the IQR for males (77).

Outliers: There are 2 outliers in the male distribution, students who reported 213 + 214 text messages. The female distribution has no outliers.