2.2 Variability

Dr. Lauren Perry

Goals

1. Calculate and interpret measures of variation/spread.
   - standard deviation and interquartile range
   - determine which measure of variation to use for a given dataset
2. Find and interpret measures of position.
   - minimum and maximum
   - quartiles
3. Summarize data using box plots.

Measures of Variability

How much do the data vary?

Should we care? Yes! The more variable the data, the harder it is to be confident in our measures of center!

- If you live in a place with extremely variable weather, it is going to be much harder to be confident in how to dress for tomorrow's weather.
- If you live in a place where the weather is always the same, it's much easier to be confident in what you plan to wear.

Range

One easy way to think about variability is the **range** of the data:

$$\text{range} = \text{maximum} - \text{minimum}$$

- This is quick and convenient, but it is *extremely* sensitive to extreme values!
- It also takes into account only two of the observations.
  - We would prefer a measure of variability that takes into account *all* the observations.

Deviation

Deviation is the distance of an observation from the mean:

$$x - \bar{x}$$

- We want to think about how far - on average - a typical observation is from the center.
- Might think to take the average deviance. . . but it turns out that summing up the deviances will *always* result in 0!
  - Conceptually, this is because the stuff below the mean (negative numbers) and the stuff above the mean (positive numbers) end up canceling each other out until we end up at 0.

One way to deal with this is to make all of the numbers positive, which we accomplish by *squaring* the deviance.

	Deviance	Squared Deviance
x	$x - \bar{x}$	$(x - \bar{x})^2$
2	-1.2	1.44
5	1.8	3.24
3	-0.2	0.04
4	0.8	0.64
2	-1.2	1.44
$\bar{x} = 3.2$	Total $= 0$	Total $= 6.8$

Variance

Variance (denoted s^2) is the "average" squared distance from the mean:

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n - 1}$$

where n is the sample size.

- Notice that we divide by $n - 1$ and NOT by n.

Standard Deviation

Finally, we come to **standard deviation** (denoted s).

- The standard deviation is the square root of the variance.

$$s = \sqrt{s^2}$$

- We will use a computer to calculate the variance and standard deviation.

Interpretation: A "typical" observation is within about one standard deviation of the mean (between $\bar{x} - s$ and $\bar{x} + s$).

Interquartile Range

The **interquartile range** (**IQR**) represents the middle 50% of the data.

- This is another measure of variability!

Percentiles

- Recall that the *median* cut the data in half: 50% of the data is below and 50% is above the median.
- This is also called the **50th percentile**.
- The ***p*th percentile** is the value for which p% of the data is below it.

To get the middle 50%, we will split the data into four parts:

1	2	3	4
25%	25%	25%	25%

The 25th and 75th percentiles, along with the median, divide the data into four parts.

Quartiles

We call these three measurements the **quartiles**:

- **Q1**, the first quartile, is the median of the lower 50% of the data.
- **Q2**, the second quartile, is the median.
- **Q3**, the third quartile, is the median of the upper 50% of the data.

Example

Consider $\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$

- Cutting the data in half: $\{1, 2, 3, 4, 5 \mid 6, 7, 8, 9, 10\}$
- So the median (Q2) is $\frac{5+6}{2} = 5.5$.
- Q1 is the median of $\{1, 2, 3, 4, 5\}$, or 3
- Q3 is the median of $\{6, 7, 8, 9, 10\}$, or 8

Note: this is a "quick and dirty" way of finding quartiles. A computer will give a more exact result.

Interquartile Range

Then the interquartile range is

$$\text{IQR} = \text{Q3} - \text{Q1}$$

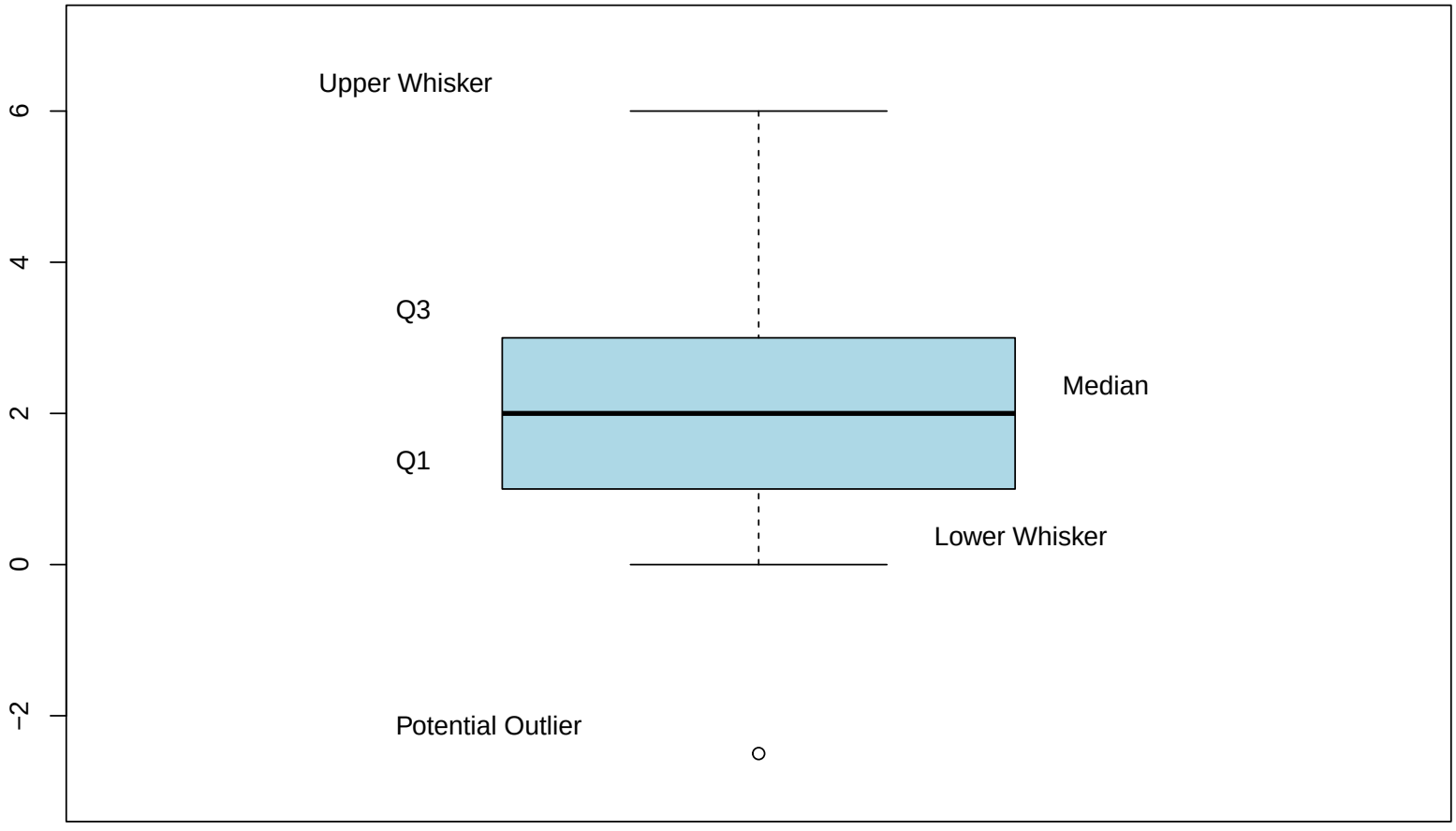
- The IQR is resistant to extreme values.

Which measure to use?

- mean and standard deviation when the data are symmetric
- median and IQR when the data are skewed

Box Plots

Box plots summarize the data with 5 statistics plus extreme observations:



Drawing a box plot:

1. Draw the vertical axis to include all possible values in the data.
2. Draw a horizontal line at the median, at Q1, and at Q3. Use these to form a box.
3. Draw the **whiskers**. The whiskers' upper limit is $\text{Q3} + 1.5 \times \text{IQR}$ and the lower limit is $\text{Q1} - 1.5 \times \text{IQR}$. The actual whiskers are then drawn *at the next closest data points within the limits*.
4. Any points outside the whisker limits are included as individual points. These are **potential outliers**.

We won't draw box plots by hand, but understanding how they are drawn will help us understand how to interpret them!

Outliers

(Potential) outliers can help us. . .

- examine skew (outliers in the negative direction suggest left skew; outliers in the positive direction suggest right skew).
- identify issues with data collection or entry, especially if the value of the outliers doesn't make sense.

Descriptive Measures for Populations

- We've thought about calculating various descriptive statistics from a sample.
- Our long-term goal is to estimate descriptive information about a population.
- At the population level, these values are called **parameters**.

- When we find a measure of center, spread, or position, we use a sample to calculate a single value.
- These single values are called **point estimates**.
  - They are used to *estimate* the corresponding population parameter.

Point Estimate	Parameter
sample mean: $\bar{x}$	population mean: μ
sample standard deviation: s	population standard deviation: σ