

---

---

---

---

---

---

---

---

## Introduction to Statistics

Nels Grevstad

Metropolitan State University of Denver  
ngrevsta@msudenver.edu

August 28, 2019

Nels Grevstad

---

---

---

---

---

---

---

---

## Topics

- 1 Summarizing Quantitative Data: Measures of Variation
- 2 The Five Number Summary and Boxplots
- 3 Using Statistics to Estimate Population Parameters

Nels Grevstad

---

---

---

---

---

---

---

---

## Objectives

### Objectives:

- Compute and interpret the range of a data set.
- Compute and interpret the quartiles and the IQR of a data set.
- Compute and interpret the sample standard deviation of a data set.
- Know which statistic (range, IQR, or standard deviation) is resistant to outliers.
- Produce and interpret a boxplot of a set of data.

Nels Grevstad

---

---

---

---

---

---

---

---

## Summarizing Quantitative Data: Measures of Variation

(3.2, 3.4)

### Three Ways to Measure Variation

- We'll look at three statistics that measure the **variation** (or **spread**) in a set of data:
  1. The range
  2. The sample standard deviation
  3. The interquartile range

Nels Grevstad

### The Sample Range

- The **sample range** is the difference between the maximum (largest) and minimum (smallest) observations:

#### Sample Range:

$$\text{Sample Range} = \text{Max} - \text{Min}$$

where **Max** is the maximum observation in the data set and **Min** is the minimum observation.

### Exercise

Here are the metabolic rates (in calories per 24 hrs) of 7 men who took part in a study of dieting and weight loss.

1792 1666 1362 1614 1460 1867 1439

Find the **range** of these data.

- The **sample range is not resistant to outliers.**

### The Sample Standard Deviation (and Sample Variance)

- To compute the *sample standard deviation*, we first need to calculate the **sample variance**, which is the *square* of the standard deviation and is denoted  $s^2$ :

#### Sample Variance:

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n - 1}$$

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

- The **variance is interpreted as an "average" (dividing by  $n - 1$  instead of  $n$ ) of the squared deviations  $(x_i - \bar{x})^2$  of the  $x_i$ 's away from their mean  $\bar{x}$ .**
- It's measured in the units of the original data **squared** (e.g.  $\text{cm}^2$  if the data are in cm).

- The **sample standard deviation**, denoted  $s$ , is obtained by taking the square root of the sample variance:

**Sample Standard Deviation:**

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

- The sample standard deviation is measured in the **same** units as the original data (e.g. cm if the data are in cm).

• **Step-by-Step Instructions for Computing the Sample Standard Deviation:**

1. Calculate  $\bar{x}$ .
2. Calculate each of the deviations  $x_i - \bar{x}$  of the observations away from  $\bar{x}$ .
3. Square each of the deviations.
4. "Average" the squared deviations by adding them up and dividing by  $n - 1$ . This gives the *sample variance*  $s^2$ .
5. Take the square root of the sample variance. This gives the *sample standard deviation*  $s$ .

• **Properties and Interpretation of the Standard Deviation:**

- The **standard deviation**  $s$  is interpreted as a **typical deviation** of an individual  $x_i$  value away from the mean  $\bar{x}$ .
- $s = 0$  only when there is no variation in the data. This happens **when all observations in the data set have the same value**.
- If the observations don't all have the same value, then  $s > 0$ .
- The **more variation** there is in the data, the **larger**  $s$  will be.

**Exercise**

Here are the metabolic rates from data from an earlier exercise.

1792 1666 1362 1614 1460 1867 1439

- a) Calculate **standard deviation** of this data set. **Hint:** The mean is  $\bar{x} = 1600$ .
- b) Interpret the standard deviation as a typical deviation away from the mean using the dot plot on the next slide.

Notes

---

---

---

---

---

---

---

---

---

---

Notes

---

---

---

---

---

---

---

---

---

---

Notes

---

---

---

---

---

---

---

---

---

---

Notes

---

---

---

---

---

---

---

---

---

---

---

---

---

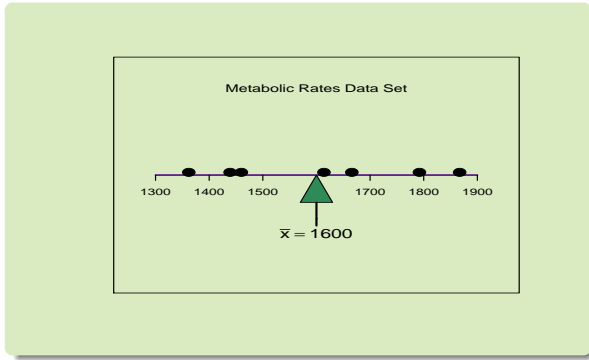
---

---

---

---

---



Nels Grevstad

- The **sample standard deviation  $s$  is not resistant to outliers**. The presence of one or more outliers will inflate the value of  $s$ .

---

---

---

---

---

---

---

---

Nels Grevstad

### Computing the Standard Deviation from Grouped Data

- The standard deviation of a set of data can be computed even when the data are **grouped** (as in a frequency distribution table). See the book for more info.

#### Sample Standard Deviation (Grouped Data Formula):

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

where each  $x_i$  is one of the **unique** values in the data set and  $f_i$  is the **frequency** of that value.

---

---

---

---

---

---

---

---

Nels Grevstad

### The Population Standard Deviation (and Population Variance)

- To compute a *population standard deviation*, first compute the **population variance**, which is denoted  $\sigma^2$ :

#### Population Variance:

$$\sigma^2 = \frac{1}{N} \sum (x_i - \mu)^2$$

where  $N$  is the population size and  $\mu$  is the population mean.

---

---

---

---

---

---

---

---

Nels Grevstad

- The **population standard deviation**, denoted  $\sigma$ , is the square root of the population variance:

**Population Standard Deviation:**

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{N} \sum (x_i - \mu)^2}$$

**The Interquartile Range**

- The **interquartile range** of a data set, denoted **IQR**, is the range of the middle 50% of the data:

**Interquartile Range:**

$$IQR = Q_3 - Q_1$$

where  $Q_1$  and  $Q_3$  are the *first* and *third quartiles*, defined below.

- The **first quartile**  $Q_1$  separates the smallest 25% of the observations from the rest of the data set.

The **third quartile**  $Q_3$  separates the smallest 75% of the observations from the rest.

- Formally, the **quartiles** are defined by:

**First and Third Quartiles:** After *sorting* the data from smallest to largest,

$Q_1$  = The median of the lower half of the sample

$Q_3$  = The median of the upper half of the sample

If  $n$  is odd, the median of the entire sample is included in *both* halves.

Notes

---

---

---

---

---

---

---

---

Notes

---

---

---

---

---

---

---

---

Notes

---

---

---

---

---

---

---

---

Notes

---

---

---

---

---

---

---

---

- The first quartile, overall median, and third quartile split the data set into fourths.
- **Interpretation of the IQR:** The *IQR* indicates how spread out the *middle 50%* of the data set is.

Nels Grevstad

### Example

Here are typical onset times (in days) for locomotion in  $n = 11$  different mammal species.

Species	Locomotion Begins (days)
Homo sapiens	360
Gorilla gorilla	165
Felis catus	21
Canis familiaris	23
Rattus norvegicus	11
Turdus merula	18
Macaca mulatta	18
Pan troglodytes	150
Saimiri sciurens	45
Cercocebus alb.	45
Tamiasciurus hud.	18

Nels Grevstad

Here are the data again, but **sorted** from smallest to largest.

11 18 18 18 21 23 45 45 150 165 360

The **quartiles** are

$$Q_1 = \text{Median of } 11, 18, 18, 18, 21, 23 \\ = 18$$

$$Q_3 = \text{Median of } 23, 45, 45, 150, 165, 360 \\ = 97.5$$

(Note that the overall median, 23, is included in *both halves* of the data set when  $n$  is odd.)

Nels Grevstad

The **interquartile range** is

$$IQR = 97.5 - 18 = 79.5.$$

This statistic measures variation in the locomotion onset times, and is interpreted as the spread in the middle 50% of the data.

Nels Grevstad

Notes

---

---

---

---

---

---

---

---

Notes

---

---

---

---

---

---

---

---

Notes

---

---

---

---

---

---

---

---

Notes

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

Exercise

Here are the metabolic rates from data from an earlier exercise, sorted from smallest to largest.

1362 1439 1460 1614 1666 1792 1867

Calculate the **interquartile range** of the data and **interpret** the result.

---

---

---

---

---

---

---

---

- The *IQR* is resistant to outliers.

---

---

---

---

---

---

---

---

**Guidelines for Choosing a Measure of Variation**

- Here are suggestions for choosing between the standard deviation and the *IQR* for reporting how much variation is in a data set (the range is seldom used):
  - The standard deviation is preferable whenever the mean is used as the measure of center, i.e. for data that don't have any strong outliers and whose distribution is not too skewed.
  - The *IQR* is preferable whenever the median is used as the measure of center, i.e. if the data set has strong outliers or its distribution is very skewed (in either direction).

---

---

---

---

---

---

---

---

**The Five Number Summary and Boxplots (3.4)**

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

**Five Number Summary:**

1. Minimum (smallest) value
2. First quartile  $Q_1$
3. Median
4. Third quartile  $Q_3$
5. Maximum (largest) value

### Constructing a Boxplot

- A **boxplot** depicts the *five number summary* graphically. It consists of:
  - Vertical axis whose units are the same as those of the data values.
  - A box whose top is level with  $Q_3$  and whose bottom is level with  $Q_1$ .
  - A horizontal line through the box level with the median.
  - "Whiskers" extending from the box up to the largest data value and down to the smallest (unless there are outliers – more about this later).

Nels Grevstad

#### Example

On November 28, 2011 a spill of toxic materials from the Suncor Energy oil refinery north of Denver was discovered seeping into Sand Creek.

The Colorado Department of Public Health and Environment monitored the spill by measuring benzene concentrations (ppb) on 13 days at two locations: the confluence of Sand Creek and the South Platte River, and a location farther downstream on the South Platte.

The data are on the next slide.

Nels Grevstad

Date	Benzene at Confluence with Sand Creek	Benzene Downstream from the Confluence
Dec. 27	640	190
Dec. 28	240	300
Dec. 29	140	130
Dec. 30	190	130
Dec. 31	170	160
Jan. 2	300	240
Jan. 3	730	250
Jan. 4	630	240
Jan. 5	650	240
Jan. 6	190	590
Jan. 7	310	260
Jan. 8	400	260
Jan. 9	720	240

Nels Grevstad

The data from the **confluence** with Sand Creek, in sorted order, are

140 170 190 190 240 300 310 400 630 640 650 720 730

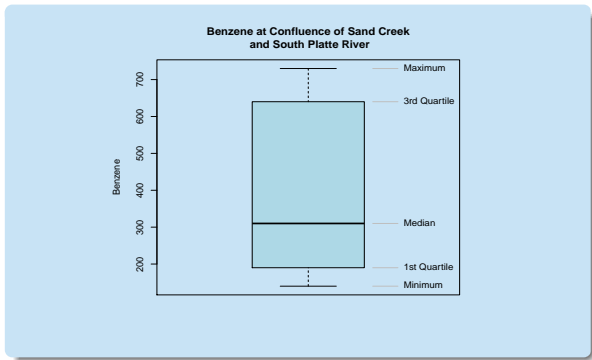
The **five number summary** is

Min	$Q_1$	Med	$Q_3$	Max
140	190	310	640	730

and the **boxplot** (with five number summary labels) is on the next slide.

Nels Grevstad





Nels Grevstad

Notes

---

---

---

---

---

---

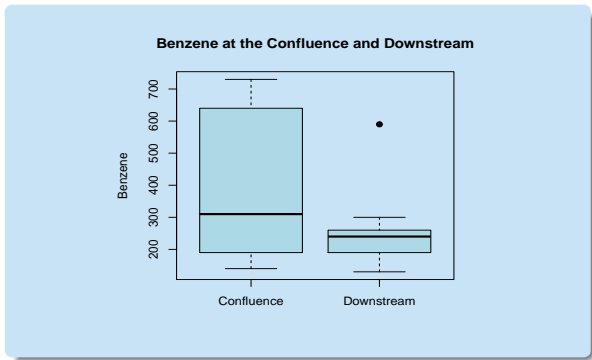
---

---

Boxplots are most often used to **compare** two or more samples. The next slide shows **side-by-side boxplots** of the two benzene samples.

The isolated point in the right boxplot is an **outlier**.

Nels Grevstad



Nels Grevstad

Notes

---

---

---

---

---

---

---

---

Notes

---

---

---

---

---

---

---

---

We see that the benzene concentrations tend to be lower farther downstream from the chemical spill. Also, there's more variation in the concentrations at the confluence (which was nearer to the spill).

Nels Grevstad

Notes

---

---

---

---

---

---

---

---

### Exercise

The table below lists 20 popular breakfast cereals, classified as either intended for **adults** or for **children**, and the amount of **sugar** per serving for each cereal.

Adults' Cereals		Children's Cereals	
Cereal	Sugar (mg)	Cereal	Sugar (mg)
Corn Flakes	2,000	Cheerios	1,000
Crispix	3,000	Wheaties	3,000
Grape Nuts	3,000	Honey Nut Cheerios	10,000
Special K	3,000	Frosted Flakes	11,000
All Bran	5,000	Honeycomb	11,000
Life	6,000	Capt Crunch	12,000
Frosted MW	7,000	Cinnamon Toast	13,000
Oatmeal Raisin Crisp	10,000	Fruit Loops	13,000
Crackling Oat Bran	10,000	Apple Jacks	14,000
Raisin Bran	12,000	Sugar Smacks	15,000

Nels Grevstad

Notes

---

---

---

---

---

---

---

---

Side-by-side boxplots of the sugar amounts for adults' and children's cereals are on the next slide.

Draw any conclusions that seem appropriate.

Nels Grevstad

Notes

---

---

---

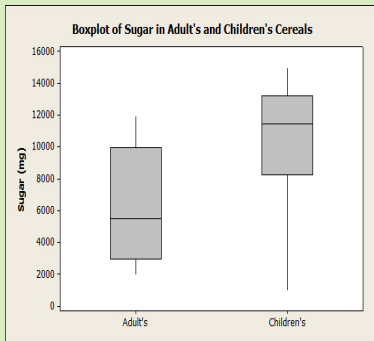
---

---

---

---

---



Nels Grevstad

Notes

---

---

---

---

---

---

---

---

### Constructing Boxplots That Show Outliers

- **Outliers in Boxplots:** Outliers are shown as isolated points in boxplots, and whiskers extend only as far as the largest and smallest observations that *aren't* outliers.

Nels Grevstad

Notes

---

---

---

---

---

---

---

---

- To decide whether an observation's an **outlier**, use the following criterion.

**Identifying Outliers:** An observation is considered to be an **outlier** if it lies farther than 1.5 *IQR*'s away from the nearest quartile. Thus an observation is an outlier if it lies outside the "fences":

$$\text{Lower Fence} = Q_1 - 1.5 \times IQR$$

$$\text{Upper Fence} = Q_3 + 1.5 \times IQR$$

**Example**

For the toxic spill data collected **downstream** of the confluence with Sand Creek, the five number summary is

Min	$Q_1$	Med	$Q_3$	Max
130	190	240	260	590

and so the *IQR* is  $260 - 190 = 70$ .

We'll deem as **outliers** any observations less than

$$\text{Lower Fence} = Q_1 - 1.5 \times IQR = 190 - 1.5(70) = 85$$

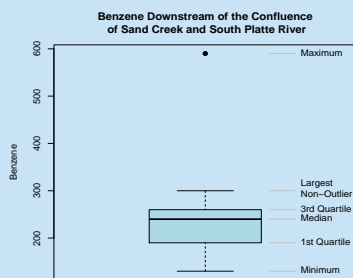
or greater than

$$\text{Upper Fence} = Q_3 + 1.5 \times IQR = 260 + 1.5(70) = 365$$

The largest benzene concentration, **590 ppb**, is therefore an **outlier**, but it's the only one.

The boxplot is shown (again) on the next slide.

Notice that the upper whisker extends only as far as 300 ppb, the largest observation that's *not* an outlier.



Notes

---

---

---

---

---

---

---

---

Notes

---

---

---

---

---

---

---

---

Notes

---

---

---

---

---

---

---

---

Notes

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

---

Exercise

The following data represent the amount of time spent studying each night (in minutes) by a sample of  $n = 30$  female college students.

180 120 180 360 240 120 180 120 240 170  
150 120 180 180 150 200 150 180 150 180  
120 60 120 180 180 90 240 180 115 120

---

---

---

---

---

---

---

---

The five number summary is:

Min	$Q_1$	Med	$Q_3$	Max
60	120	175	180	360

- a) Determine which observations (if any) are **outliers**.
- b) Make a boxplot of the data with any outliers shown as isolated points.

---

---

---

---

---

---

---

---

Identifying Skewness from a Boxplot

- The **five number summary** and **boxplot** conveys information about the **center, spread, and skewness** of the distribution of the data.
- Boxplots of right and left **skewed** data sets are **asymmetrical** ("lopsided").  
The long "tail" of the distribution shows up as a long, stretched-out whisker in the boxplot.

---

---

---

---

---

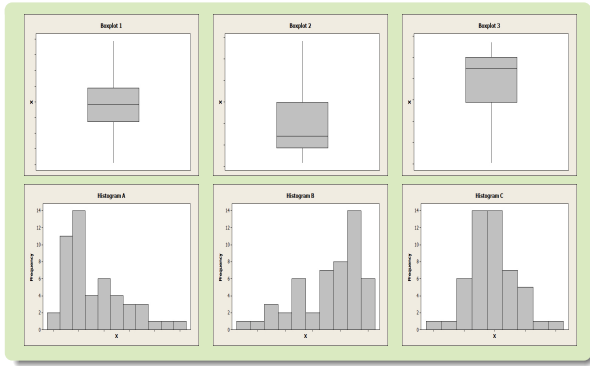
---

---

---

Exercise

For each of three sets of data, both a histogram and boxplot were made. They're shown on the next slide in jumbled order. Can you match the boxplot with its histogram?



Nels Grevstad

## Using Statistics to Estimate Population Parameters (3.5)

- We use **statistics** to **estimate** population parameters. In particular:
  - The sample mean  $\bar{x}$  is used as an **estimate** of the population mean  $\mu$ .
  - The sample standard deviation  $s$  is used as an **estimate** of the population standard deviation  $\sigma$ .

Nels Grevstad

- **Comment:** Dividing by  $n - 1$  when computing  $s$  makes  $s$  a **more accurate estimator** of  $\sigma$ .  
If we divided by  $n$ , the resulting statistic would tend to underestimate  $\sigma$ .

Nels Grevstad

Notes

---

---

---

---

---

---

---

---

Notes

---

---

---

---

---

---

---

---

Notes

---

---

---

---

---

---

---

---

Notes

---

---

---

---

---

---

---

---