Probability and Statistics

Nels Grevstad

Metropolitan State University of Denver

ngrevsta@msudenver.edu

January 22, 2019

Topics

Objectives

Objectives:

- Compute and interpret the mean and median of a data set
- Know the properties of the mean and median
- Compute and interpret the variance, standard deviation, and interquartile range of a data set
- Know the properties of the variance, standard deviation, and interquartile range
- Construct a boxplot

Measures of Center (1.3)

- A **sample** of size n will be denoted

$$x_1, x_2, \ldots, x_n$$

Measures of Center (1.3)

- A **sample** of size n will be denoted

$$x_1, x_2, \ldots, x_n$$

- The ***sample mean***, denoted $\bar{x}$, is the arithmetic average of the observations in the data set:

Sample Mean:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Example

A person's metabolic rate is the rate at which the body consumes energy while at rest. The data below are the metabolic rates (in calories per day) for 7 men who took part in a study of dieting and weight loss.

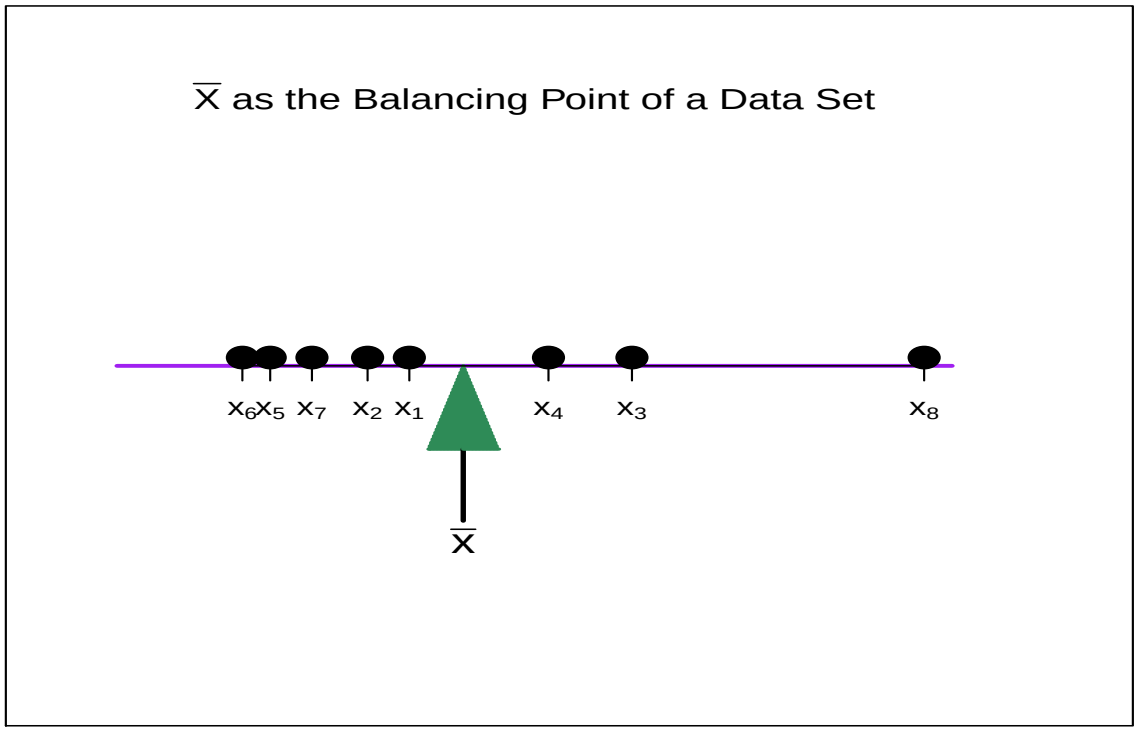
1792 1666 1362 1614 1460 1867 1439

The sample mean is

$$\begin{aligned} \bar{x} &= \frac{1792 + 1666 + 1362 + 1614 + 1460 + 1867 + 1439}{7} \\ &= 1600. \end{aligned}$$

- The mean is the "**balancing point**" of a data set – weights placed at positions $x_1, x_2, \ldots, x_n$ along a (weightless) axis would balance at $\bar{x}$.

- The mean is the "**balancing point**" of a data set – weights placed at positions $x_1, x_2, \ldots, x_n$ along a (weightless) axis would balance at $\bar{x}$.



Proposition

Suppose $x_1, x_2, \ldots, x_n$ are a sample and c is any nonzero constant.

1. If

$$y_1 = x_1 + c, \quad y_2 = x_2 + c, \quad \ldots, \quad y_n = x_n + c$$

then

$$\bar{y} = \bar{x} + c.$$

2. If

$$y_1 = cx_1, \quad y_2 = cx_2, \quad \ldots, \quad y_n = cx_n$$

then then

$$\bar{y} = c\bar{x}.$$

where $\bar{x}$ and $\bar{y}$ are the sample means of the x's and y's.

Proposition

Suppose $x_1, x_2, \ldots, x_n$ are a sample. Then

$$\sum_{i=1}^{n}(x_i - \bar{x}) \;=\; 0.$$

Example (Cont'd)

Here are the metabolic rates data again

1792 1666 1362 1614 1460 1867 1439

The deviations away from the mean (1600) sum to 0:

$$192 + 66 + (-238) + 14 + (-140) + 267 + (-161) \;=\; 0.$$

- We say the n deviations

$$x_1 - \bar{x}, \quad x_2 - \bar{x}, \quad \ldots, \quad x_n - \bar{x}$$

have only $n - 1$ ***degrees of freedom*** because any $n - 1$ of them determines the remaining one.

- For a finite population with N individuals, the ***population mean***, denoted μ, is defined as

Population Mean:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i.$$

- For a finite population with N individuals, the ***population mean***, denoted μ, is defined as

Population Mean:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i.$$

- Later we'll see a more general way to compute a population mean.

- For a finite population with N individuals, the ***population mean***, denoted μ, is defined as

Population Mean:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i.$$

- Later we'll see a more general way to compute a population mean.
- A sample mean $\bar{x}$ is used to **estimate** an unknown population mean μ.

- The ***sample median*** (or ***50th percentile***), denoted $\tilde{x}$, is the middle value in the (ordered) sample.

Sample Median:

$$\tilde{x} = \begin{cases} \text{The } (\frac{n+1}{2})\text{th ordered value if } n \text{ is odd.} \\ \text{The average of the } (\frac{n}{2})\text{ th and the } (\frac{n+2}{2})\text{th} \\ \text{ordered values if } n \text{ is even.} \end{cases}$$

Example (Cont'd)

Here are the metabolic rates **ordered** from smallest to largest.

1362 1439 1460 1614 1666 1792 1867

Because $n = 7$ is **odd**, the sample median is the middle value,

$$\tilde{x} = 1614.$$

Example

Here are data on the lifetimes (in hours) of a certain type of light bulb, **ordered** from smallest to largest.

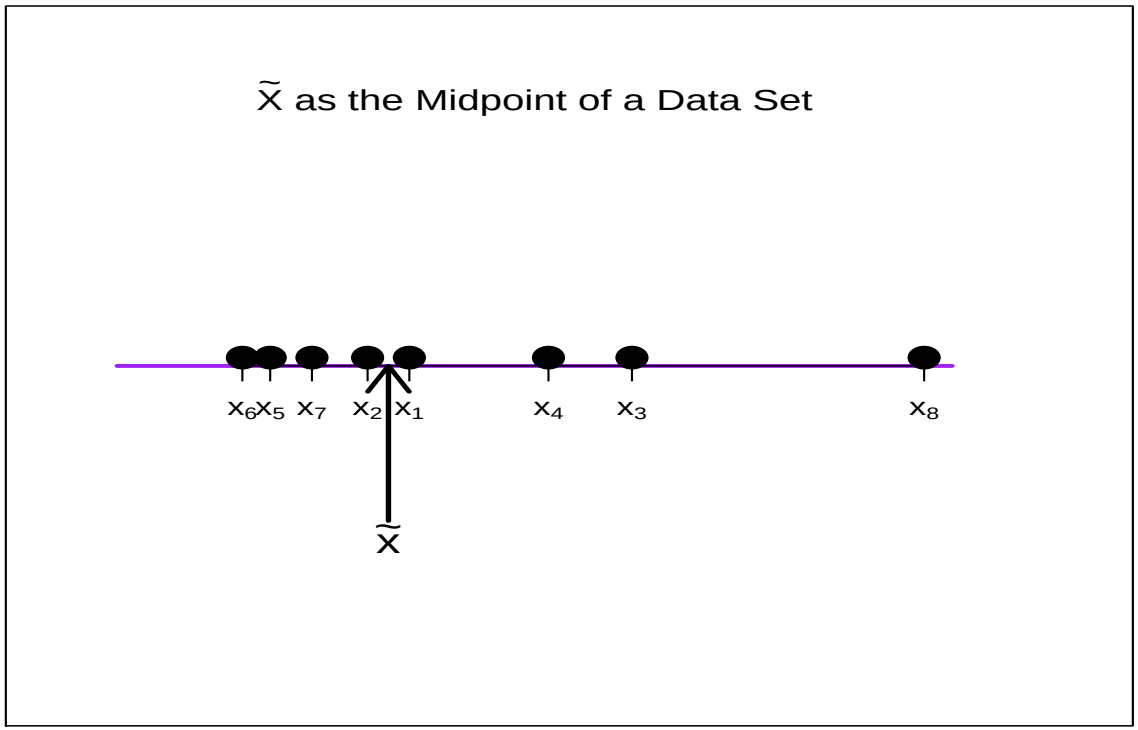
898 964 970 983 1003 1016 1022 1029 1058 1085

Because $n = 10$ is **even**, the sample median is the average of the two middle values,

$$\tilde{x} = \frac{1003 + 1016}{2} = 1009.5.$$

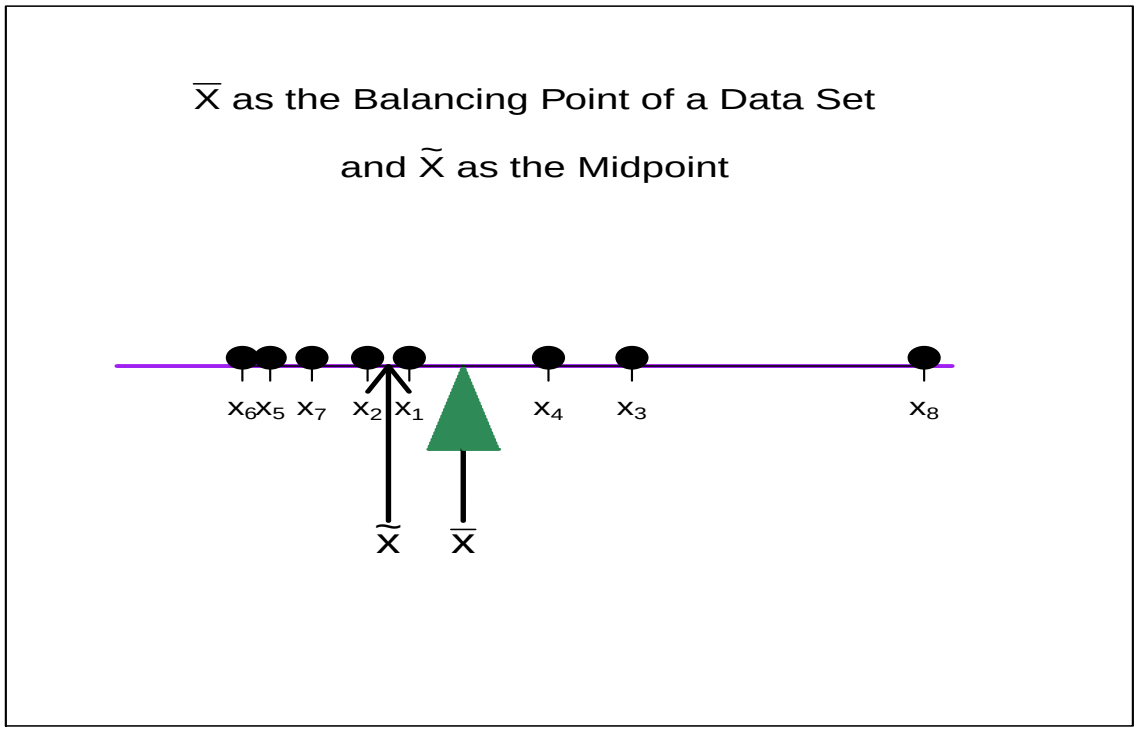
- The sample median is the **"middle point"** of a data set – it splits the data set in half.

- The sample median is the **"middle point"** of a data set – it splits the data set in half.



- Both mean and median give some idea of where the data are centered, but they're usually different.

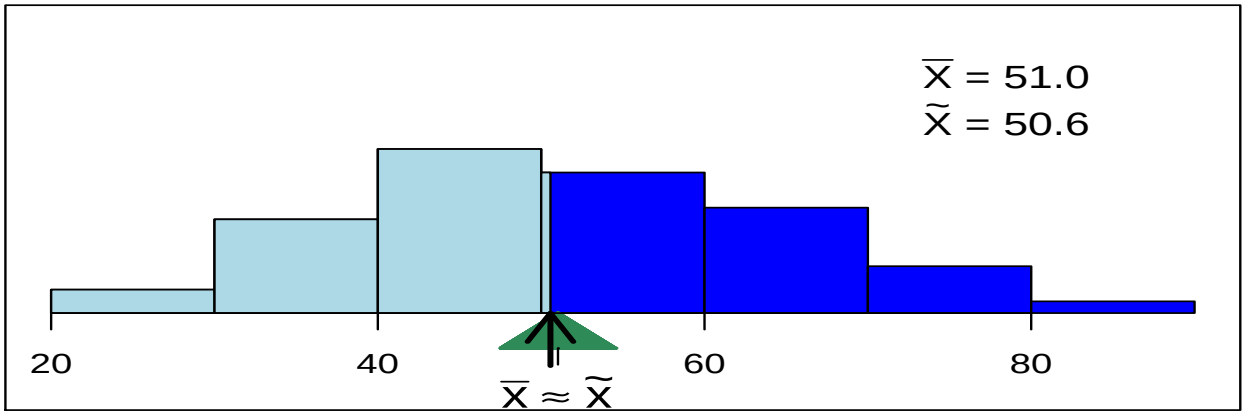
- Both mean and median give some idea of where the data are centered, but they're usually different.



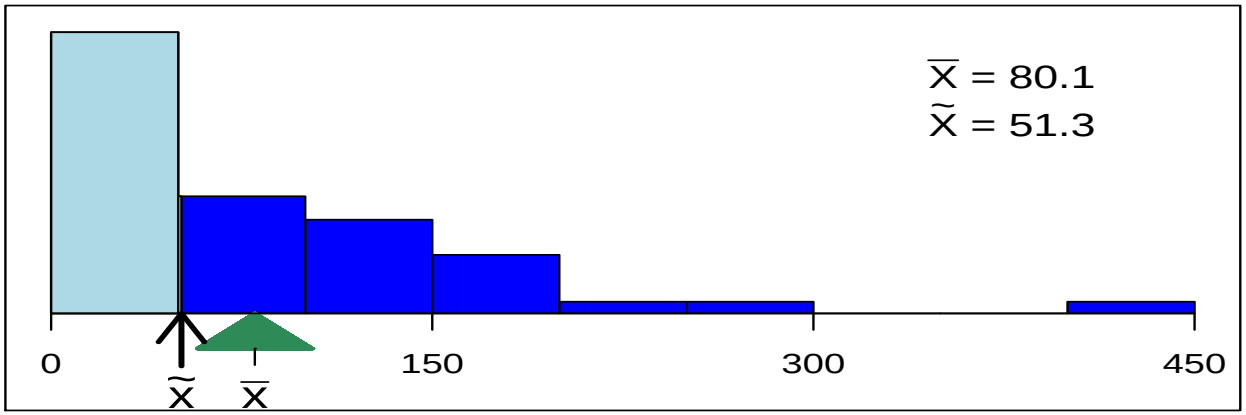
- Why might they be different?

- Why might they be different?
  - The median is not affected by a few outliers (it's ***resistant***)
  - The mean can be affected by a few outliers (it's ***not resistant***)
  - The median divides the area of the histogram in half
  - The mean is the "balancing point" of the histogram

Symmetric Distribution $\overline{X} \approx \widetilde{X}$



Right Skewed Distribution $\overline{X} > \widetilde{X}$



- If the histogram is fairly symmetric, the mean and median will be similar
- If the histogram is right skewed, the mean will be larger

- Comments:

- If the histogram is fairly symmetric, the mean and median will be similar
- If the histogram is right skewed, the mean will be larger

- Comments:
  - If the mean and median are not similar, make sure you report them both
  - The mean is most commonly used in testing and estimation (later in the course)

Measures of Variation (1.4)

- The ***sample variance***, denoted s^2, is defined as

Sample Variance:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Measures of Variation (1.4)

- The ***sample variance***, denoted s^2, is defined as

Sample Variance:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

- It represents the size of a **typical squared deviation** away from the mean.

Measures of Variation (1.4)

- The ***sample variance***, denoted s^2, is defined as

Sample Variance:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

- It represents the size of a **typical squared deviation** away from the mean.
- It's measured in the *squares* of the original units.

Measures of Variation (1.4)

- The ***sample variance***, denoted s^2, is defined as

Sample Variance:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

- It represents the size of a **typical squared deviation** away from the mean.
- It's measured in the *squares* of the original units.
- We'll see later why we use $n-1$ instead of n.

- The ***sample standard deviation***, denoted s, is the square root of the variance.

Sample Standard Deviation:

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

- It measures the size of a **typical deviation** away from the mean.
- It's is measured in the original units.

Example

Here are the metabolic rates data again.

1792 1666 1362 1614 1460 1867 1439

Recall that $\bar{x} = 1600$ and the deviations away from 1600 are

192 66 -238 14 -140 267 -161

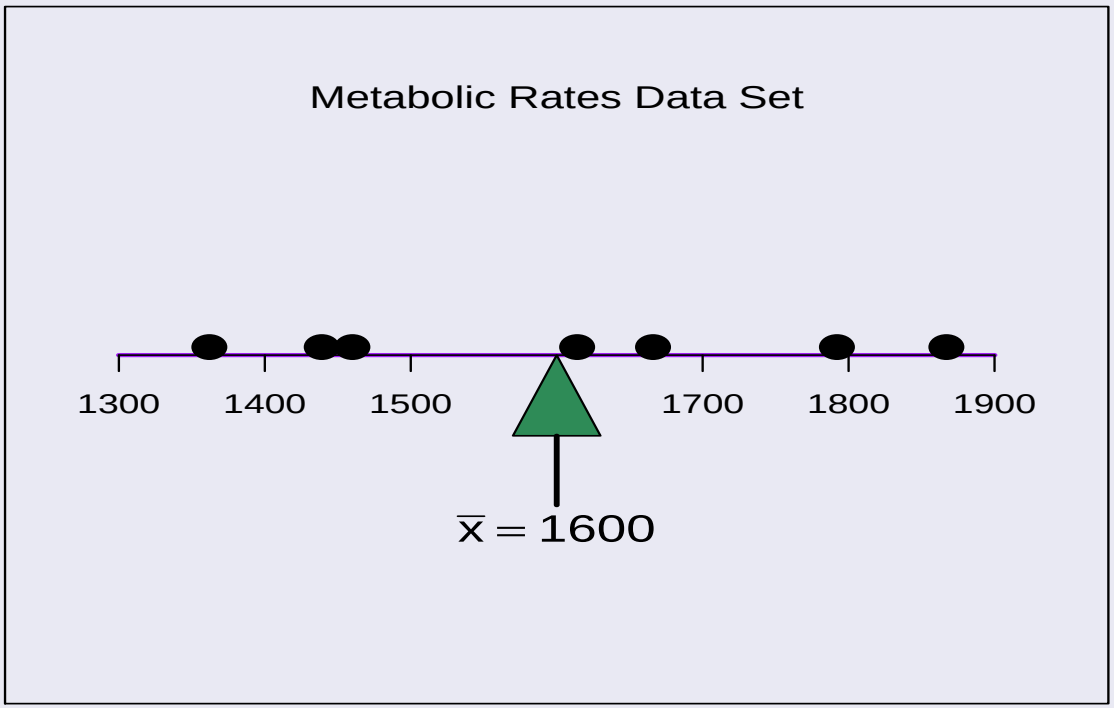
The sample variance is

$$\begin{aligned} s^2 &= \frac{1}{6}(192^2 + 66^2 + (-238)^2 + 14^2 + (-140)^2 + 267^2 + (-161)^2) \\ &= 35,811.7. \end{aligned}$$

The standard deviation is

$$s = \sqrt{35,811.7} = 189.2$$

calories per day.


Metabolic Rates Data Set
1300
1400
1500
1700
1800
1900
$\bar{x} = 1600$

- $s = 0$ if and only if $x_1, x_2, \ldots, x_n$ are all the same value.
- Otherwise $s > 0$, with larger values of s indicating more variation in the data.
- s^2 and s are **not resistant** to outliers.

Proposition

Suppose $x_1, x_2, \ldots, x_n$ are a sample and c is any nonzero constant.

1. If
$$y_1 = x_1 + c, \quad y_2 = x_2 + c, \quad \ldots, \quad y_n = x_n + c$$
then
$$s_y^2 = s_x^2 \qquad \text{and} \qquad s_y = s_x.$$
2. If
$$y_1 = cx_1, \quad y_2 = cx_2, \quad \ldots, \quad y_n = cx_n$$
then
$$s_y^2 = c^2 s_x^2 \qquad \text{and} \qquad s_y = |c| s_x.$$

where s_x^2 and s_y^2 are the sample variances of the x's and y's.

- For a finite population with N individuals, the ***population variance*** and ***population standard deviation*** , denoted $\boldsymbol{\sigma^2}$ and $\boldsymbol{\sigma}$, are

Population Variance and Standard Deviation:

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

and

$$\sigma = \sqrt{\sigma^2}.$$

- For a finite population with N individuals, the ***population variance*** and ***population standard deviation*** , denoted $\boldsymbol{\sigma^2}$ and $\boldsymbol{\sigma}$, are

Population Variance and Standard Deviation:

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

and

$$\sigma = \sqrt{\sigma^2}.$$

- Later we'll see a more general way to compute a population variance and standard deviation.

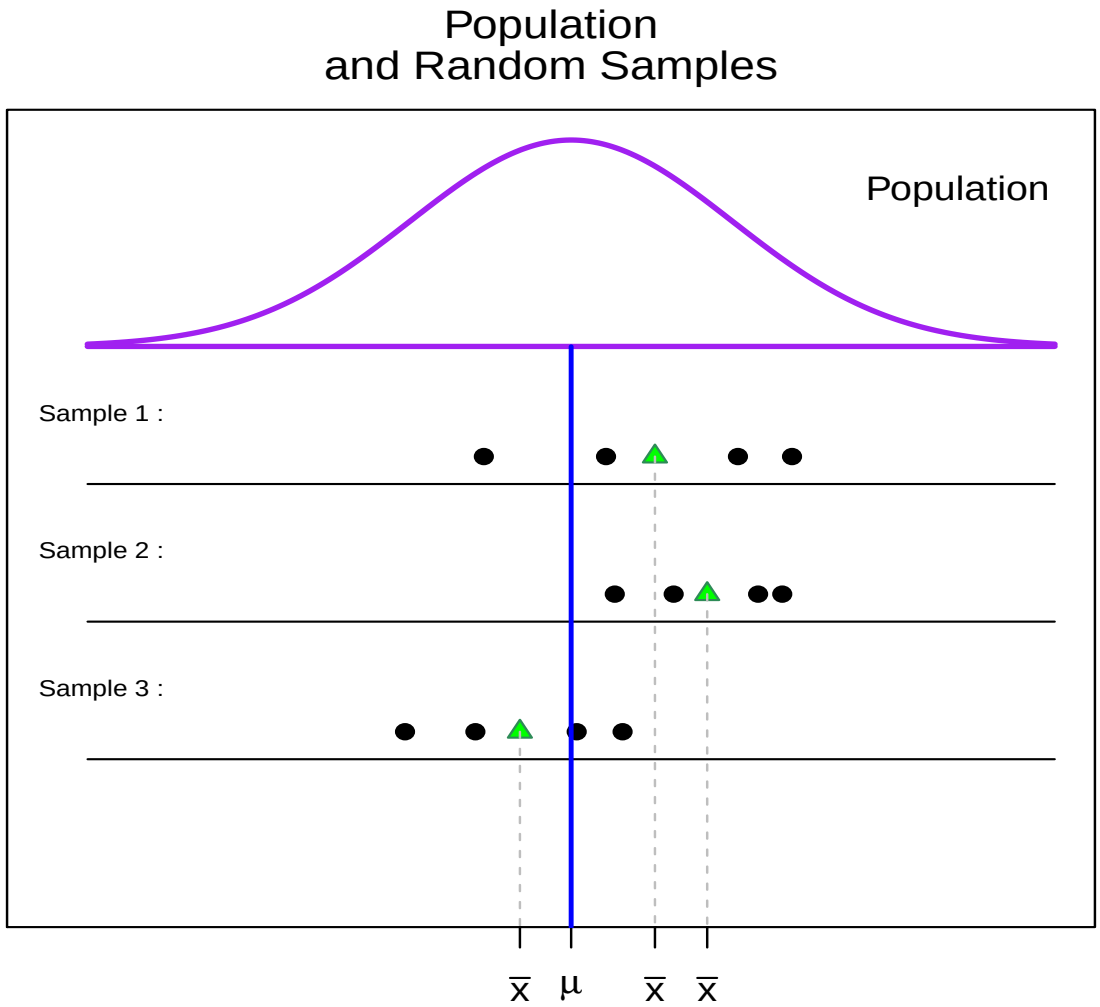
- A sample variance s^2 is used to **estimate** an unknown population variance σ^2 (and s is used to estimate σ).

- A sample variance s^2 is used to **estimate** an unknown population variance σ^2 (and s is used to estimate σ).
- We divide by $n-1$ when computing s^2 but by N when computing σ^2.

- A sample variance s^2 is used to **estimate** an unknown population variance σ^2 (and s is used to estimate σ).
- We divide by $n-1$ when computing s^2 but by N when computing σ^2.
- Why? If we divided by n, s^2 would tend to **underestimate** σ^2.

- A sample variance s^2 is used to **estimate** an unknown population variance σ^2 (and s is used to estimate σ).
- We divide by $n-1$ when computing s^2 but by N when computing σ^2.
- Why? If we divided by n, s^2 would tend to **underestimate** σ^2.
- Dividing by $n-1$ compensates for this tendency to under estimate σ^2.

- A sample variance s^2 is used to **estimate** an unknown population variance σ^2 (and s is used to estimate σ).
- We divide by $n-1$ when computing s^2 but by N when computing σ^2.
- Why? If we divided by n, s^2 would tend to **underestimate** σ^2.
- Dividing by $n-1$ compensates for this tendency to under estimate σ^2.
- Why would s^2 underestimate σ^2 if we divided by n? The deviations $x_i - \bar{x}$ used to compute s^2 tend to be **smaller** than the deviations $x_i - \mu$ used to compute σ^2.


Population
and Random Samples
Population
Sample 1 :
Sample 2 :
Sample 3 :
x̄ μ x̄ x̄

- It turns out that the value of c that makes the quantity

$$\sum_{i=1}^{n}(x_i - c)^2$$

as small as possible is $c = \bar{x}$, so

$$\sum_{i=1}^{n}(x_i - \bar{x})^2$$

will be smaller than

$$\sum_{i=1}^{n}(x_i - \mu)^2$$

- It turns out that the value of c that makes the quantity

$$f(c) = \sum_{i=1}^{n}(x_i - c)^2$$

as small as possible is $c = \bar{x}$, so

$$\sum_{i=1}^{n}(x_i - \bar{x})^2$$

will be smaller than

$$\sum_{i=1}^{n}(x_i - \mu)^2$$

- It turns out that the value of c that makes the quantity

$$f(c) = \sum_{i=1}^{n}(x_i - c)^2 \qquad f'(c) = 0$$

as small as possible is $c = \bar{x}$, so

$$\sum_{i=1}^{n}(x_i - \bar{x})^2$$

will be smaller than

$$\sum_{i=1}^{n}(x_i - \mu)^2$$

- The ***interquartile range*** (or ***fourth spread***), denoted $\boldsymbol{IQR}$, is the range of the middle 50% of the data:

Interquartile Range:

$$IQR = Q_3 - Q_1$$

where $\boldsymbol{Q_1}$ is the ***first quartile*** (or ***25th percentile***) and $\boldsymbol{Q_3}$ is the ***third quartile*** (or ***75th percentile***), defined by:

$$\begin{aligned} Q_1 &= \text{The median of the observations that are less than or equal to the overall median } \tilde{x} \\ Q_3 &= \text{The median of the observations that are greater than or equal to the overall median } \tilde{x} \end{aligned}$$

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
Tamiasciureus hud.	18

Here are the data again, but **ordered**.

11 18 18 18 21 23 45 45 150 165 360

The quartiles are

$$\begin{aligned} Q_1 &= \text{Median of } 11, 18, 18, 18, 21, 23 \\ &= 18 \\ Q_3 &= \text{Median of } 23, 45, 45, 150, 165, 360 \\ &= 97.5 \end{aligned}$$

so the interquartile range is

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

- The interquartile range is **resistant** to outliers.

- The interquartile range is **resistant** to outliers.
- The ***five number summary*** of a data set is

Five Number Summary:

Minimum, Q_1, $\tilde{x}$, Q_3, Maximum

- The interquartile range is **resistant** to outliers.
- The ***five number summary*** of a data set is

Five Number Summary:

Minimum, Q_1, $\tilde{x}$, Q_3, Maximum

- It summarizes the center, spread, *and* (we'll see) skewness of the data.

- A ***boxplot*** depicts the five number summary graphically, and consists of:

- A ***boxplot*** depicts the five number summary graphically, and consists of:
  - Vertical axis whose units are the 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 $\tilde{x}$.
  - ”Whiskers” (vertical lines) extending from the box up to the largest data value and from the bottom of the box down to the smallest (unless there are outliers).

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.

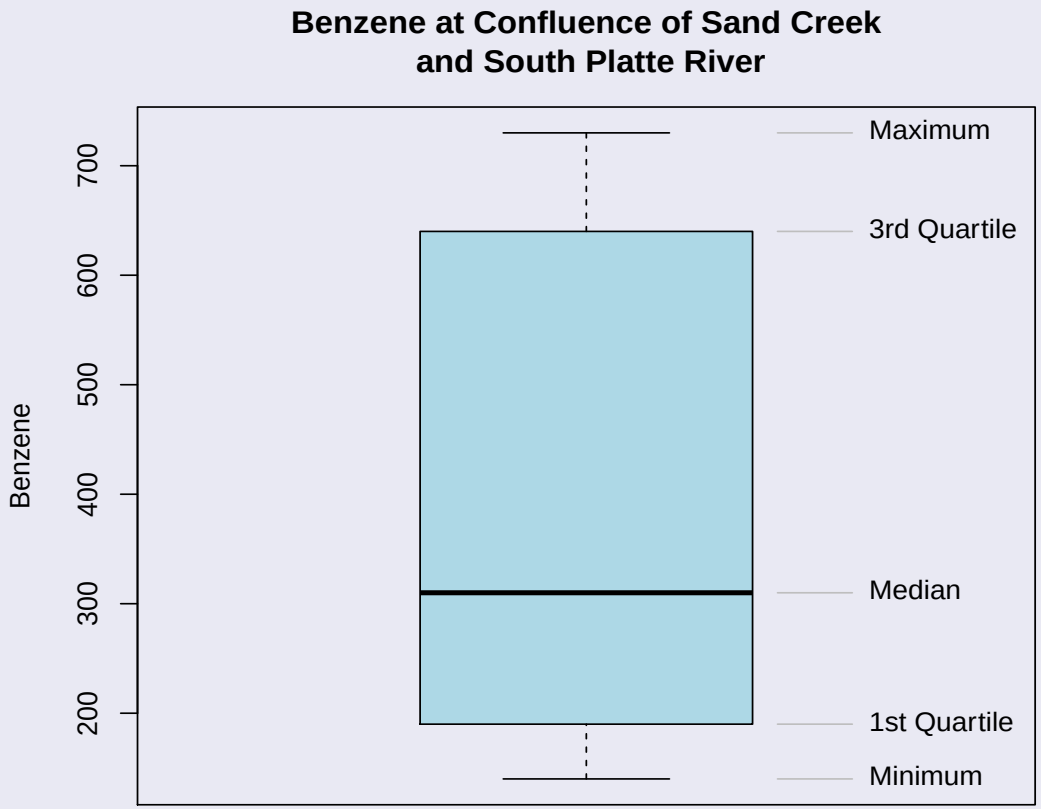
The Colorado Department of Public Health and Environment monitored the spill by measuring benzene concentrations (ppb) on 13 days at two locations, the confluence of Sand Creek and the South Platte River, and a location on the South Platte farther downstream. The data are below.

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

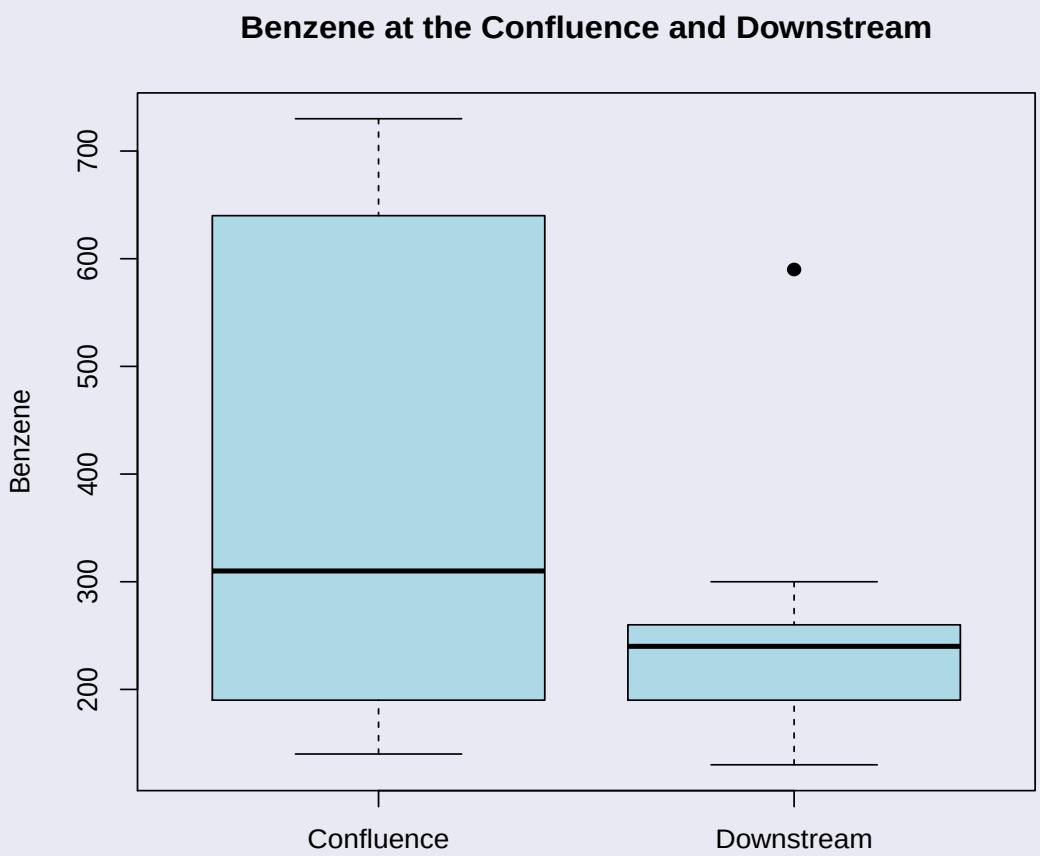
The five number summary of the data from the first location is

Min	Q_1	$\tilde{x}$	Q_3	Max
140	190	310	640	730

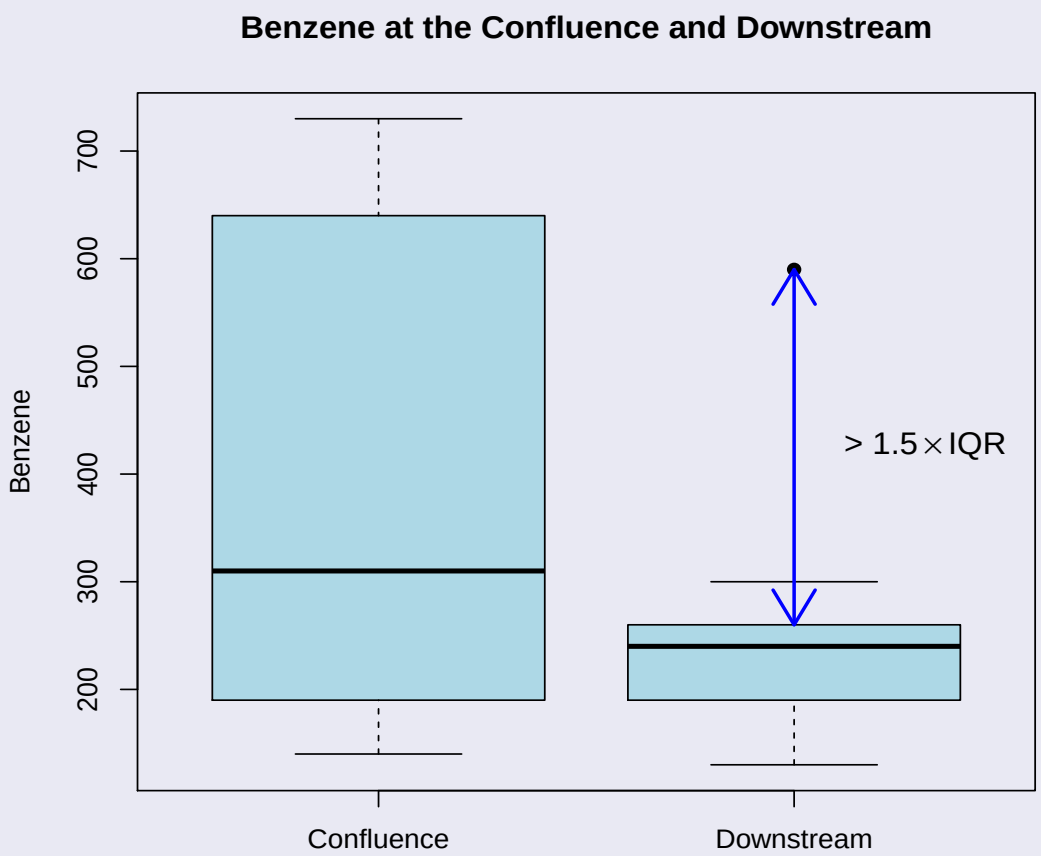
and the boxplot (with labels) is shown below.


Benzene at Confluence of Sand Creek
and South Platte River
Benzene
200
300
400
500
600
700
Maximum
3rd Quartile
Median
1st Quartile
Minimum

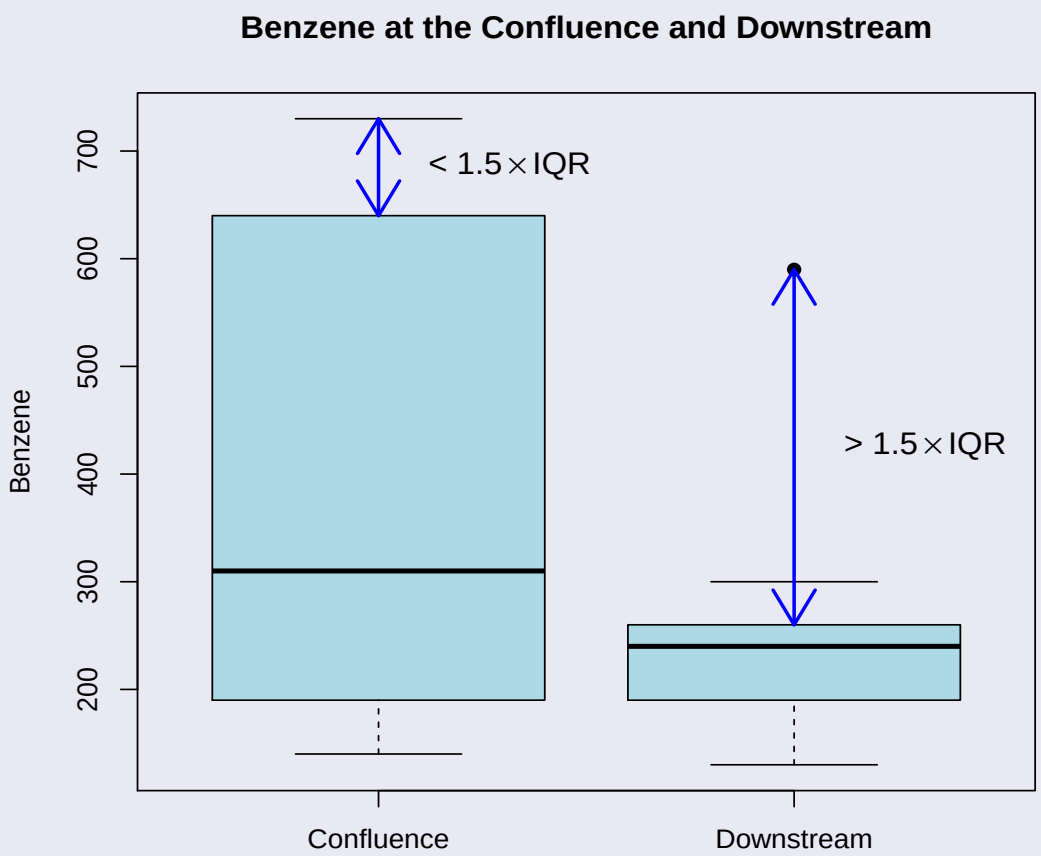
Boxplots are most often used to compare two or more samples.



Boxplots are most often used to compare two or more samples.



Boxplots are most often used to compare two or more samples.



- Right and left skewed data sets have **asymmetrical** boxplots. The plots below are in jumbled order. Can you match the boxplot with its histogram?

- Right and left skewed data sets have **asymmetrical** boxplots. The plots below are in jumbled order. Can you match the boxplot with its histogram?