EDP 613 Fall 2020

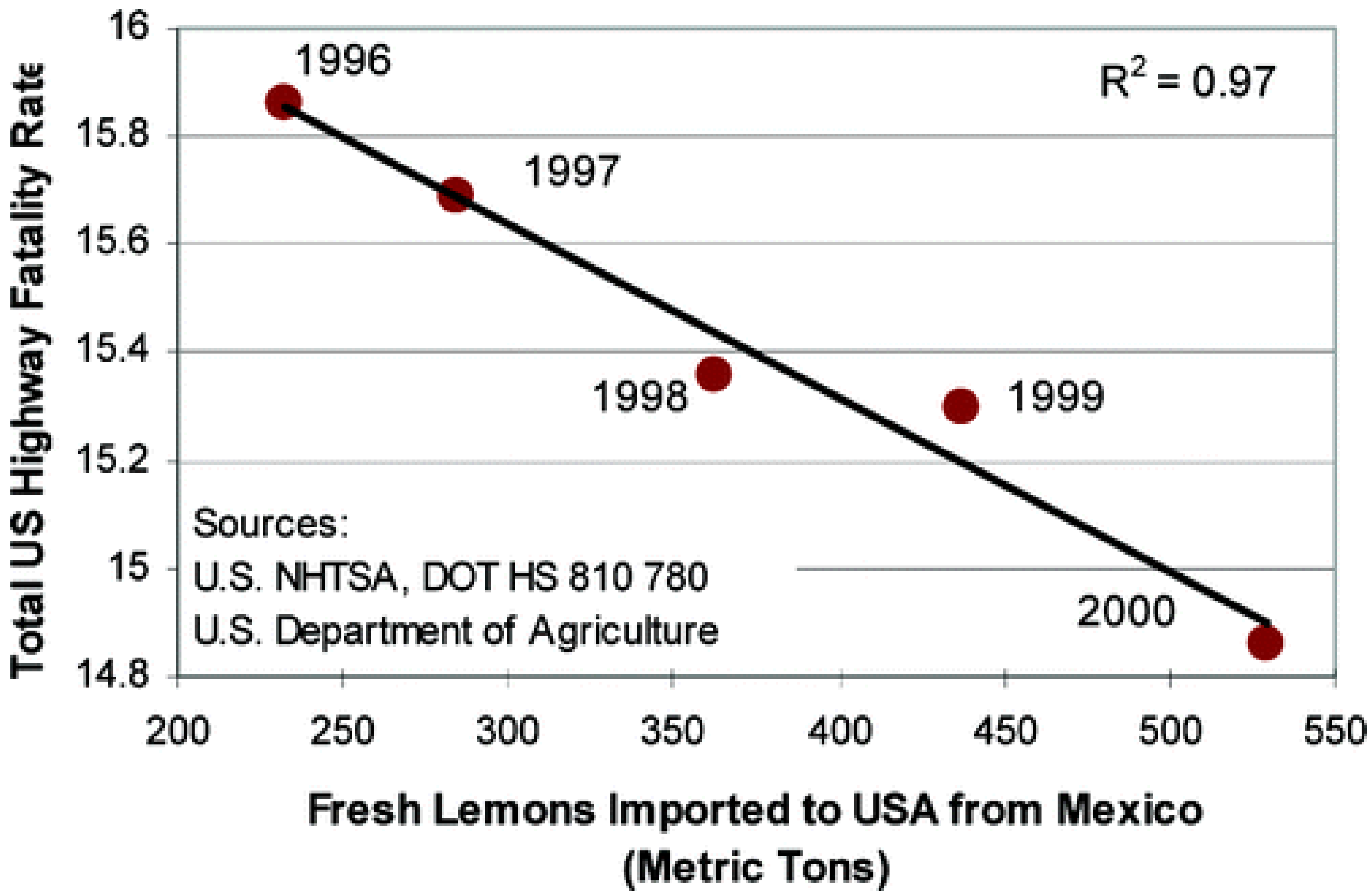
Chapter 4: Measures of Variability

Abhik Roy

Abhik.Roy@mail.wvu.edu

West Virginia University

Always Remember!



CORRELATION DOES NOT PROVE CAUSATION!

Definition

Variability is just how spread out a data set is.

Measure of Variability: **The Range**

The **range** of a data set is the difference between the largest value and the smallest value:

$$\text{Range} = \text{Largest value} - \text{Smallest value}$$

Compute the range for the following sample:

$$4 \quad 1 \quad 1 \quad 3 \quad 4 \quad 7$$

The largest value is 7 while the smallest value is 1. Therefore:

$$\text{Range} = 7 - 1 = 6$$

Compute the range for the following sample:

$3.61 $3.84 $3.79 $3.61 $4.09 $3.96

The largest value is \$4.09 while the smallest value is \$3.61. Therefore:

$$\text{Range} = \$4.09 - \$3.61 = \$0.48$$

Measure of Variability: **Quartiles**

Every data set has three quartiles:

- The **first quartile**, denoted Q_1, is the 25^{th} percentile. Q_1 separates the lowest 25% of the data from the highest 75%.

Every data set has three quartiles:

- The **first quartile**, denoted Q_1, is the 25^{th} percentile. Q_1 separates the lowest 25% of the data from the highest 75%.
- The **second quartile**, denoted Q_2, is the 50^{th} percentile. Q_2 separates the lowest 50% of the data from the highest 50%. (NOTE: $Q_2 =$ median)

Every data set has three quartiles:

- The **first quartile**, denoted Q_1, is the 25^{th} percentile. Q_1 separates the lowest 25% of the data from the highest 75%.
- The **second quartile**, denoted Q_2, is the 50^{th} percentile. Q_2 separates the lowest 50% of the data from the highest 50%. (NOTE: $Q_2 =$ median)
- The **third quartile**, denoted Q_3, is the 25^{th} percentile. Q_3 separates the lowest 75% of the data from the highest 25%.

The **five-number summary** of a data set consists of the following quantities:

- Minimum
- First Quartile
- Second Quartile (Median)
- Third Quartile
- Maximum

An **outlier** is a value that is considerably larger or smaller than most of the values in a data set.

The **interquartile range (IQR)** is found by subtracting the first quartile from the third quartile

$$\text{IQR} = Q_3 - Q_1$$

The IRQ Method for Finding Outliers:

1. Find Q_1 and Q_3.
2. Compute the IQR.
3. Compute the cutoff points for determining outliers, or **outlier boundaries**,

$$\begin{aligned} \text{Lower Outlier Boundary} &= Q_1 - 1.5 \cdot \text{IQR} \\ \text{Upper Outlier Boundary} &= Q_3 + 1.5 \cdot \text{IQR} \end{aligned}$$

4. Any number $<$ Lower Outlier Boundary or $>$ Upper Outlier Boundary is an outlier.

Jamie drives to work every weekday morning and keeps track of her time (in minutes) for 35 days. Her measurements are displayed below:

15	17	17	17	17	18	19
19	19	19	19	19	20	20
20	20	20	21	21	21	21
21	21	21	21	21	22	23
23	24	26	31	36	38	39

Construct a box plot for the data

1. We have a minimum value of 15 minutes and a maximum value of 39 minutes

2. Computing Q_1

$$
\begin{aligned}
L_{\text{first}} &= \frac{25}{100} \cdot 35 \\
&= 0.25 \cdot 35 \\
&= 8.75 \\
&\approx 9
\end{aligned}
$$

In position 9, the data value is 19 minutes.

3. Computing Q_3

$$\begin{aligned} L_{\text{third}} &= \frac{75}{100} \cdot 35 \\ &= 0.75 \cdot 35 \\ &= 26.25 \\ &\approx 27 \end{aligned}$$

In position 27, the data value is 22 minutes.

4. IQR:

$$\begin{aligned} \text{IQR} &= 22 - 19 \\ &= 3 \end{aligned}$$

so...

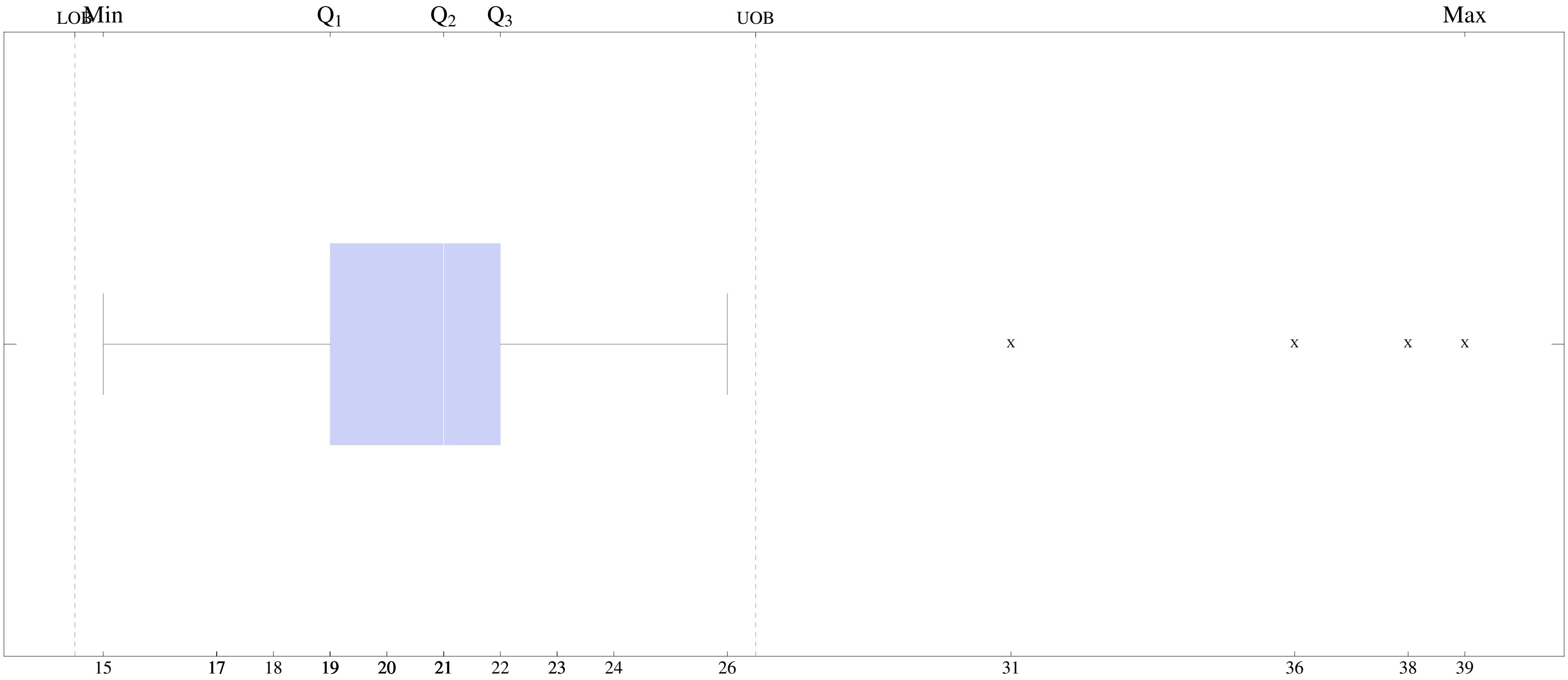
LOB & UOB:

$$\begin{aligned}\text{Lower Outlier Boundary} &= 19 - 1.5 \cdot 3 \\ &= 19 - 4.5 \\ &= 14.5 \text{ minutes}\end{aligned}$$

and

$$\begin{aligned}\text{Upper Outlier Boundary} &= 22 + 1.5 \cdot 3 \\ &= 22 + 3 \\ &= 26.5 \text{ minutes}\end{aligned}$$

Boxplot



LOB
Min
Q1
Q2
Q3
UOB
Max
x
x
x
x
15
17
18
19
20
21
22
23
24
26
31
36
38
39

To determine the shape of a box plot, use the following

- If the median is closer to the first quartile then to the third OR the upper whisker is longer than the lower whisker, the data are skewed to the right.
- If the median is closer to the third quartile then to the third OR the lower whisker is longer than the upper whisker, the data are skewed to the left.
- If the median is approximately halfway between the first and third quartiles AND the upper whisker is similar in length to the lower whisker, the data is normal or approximately normal.

OYO: Following are the number of grams of carbohydrates in 12-ounce espresso beverages offered at Starbucks

14	43	38	44	31	27	39	59	9	10	54
14	25	26	9	46	30	24	41	26	27	14

Construct a box plot for the data

1. We have a minimum value of 59 minutes and a maximum value of 9 grams.

2. Computing Q_1

$$
\begin{aligned}
L_{\text{first}} &= \frac{25}{100} \cdot 22 \\
&= 0.25 \cdot 22 \\
&= 5.50 \\
&\approx 6
\end{aligned}
$$

In position 6, the data value is 14 grams.

3. Computing Q_3

$$
\begin{aligned}
L_{\text{third}} &= \frac{75}{100} \cdot 22 \\
&= 0.75 \cdot 22 \\
&= 16.5 \\
&\approx 17
\end{aligned}
$$

In position 17, the data value is 41 grams.

4. IQR:

$$\begin{aligned} \text{IQR} &= 41 - 14 \\ &= 27 \end{aligned}$$

so...

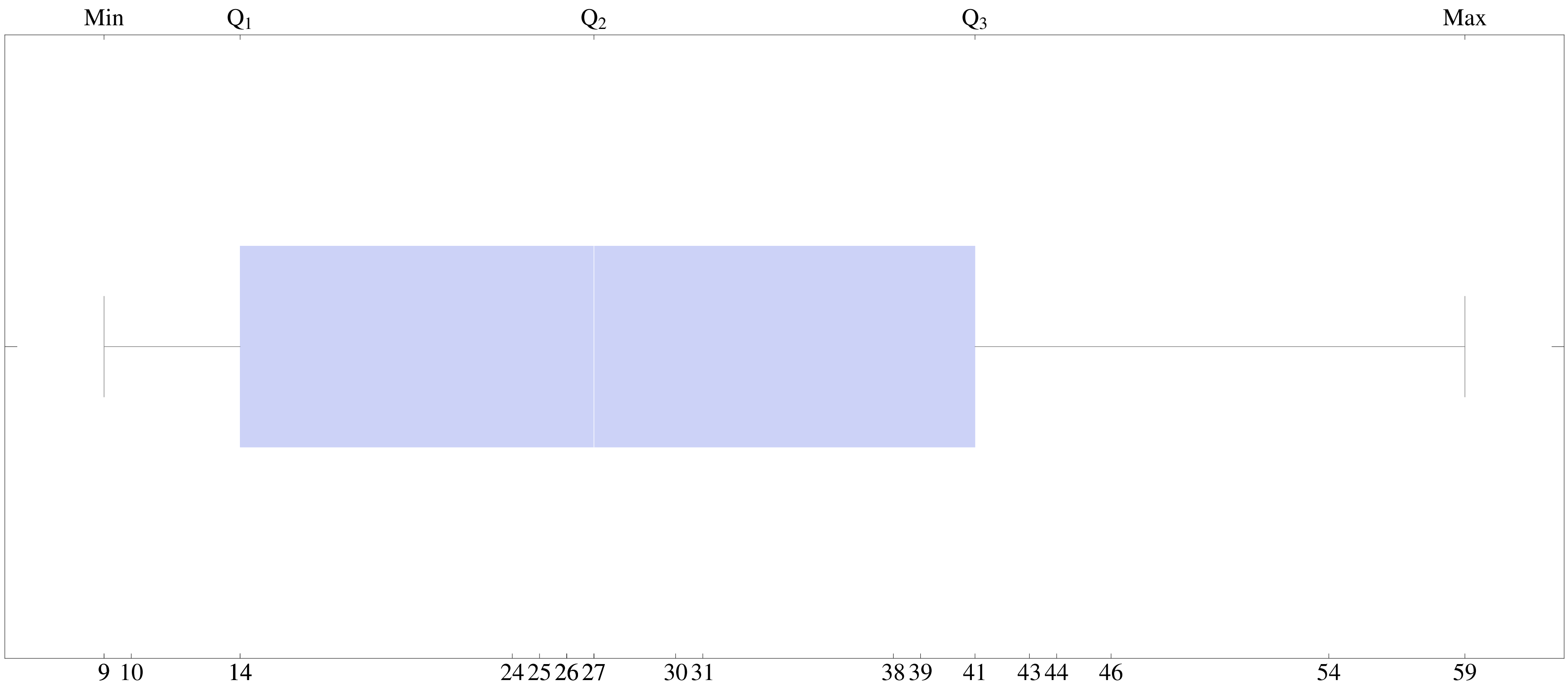
LOB & UOB:

$$\begin{aligned}\text{Lower Outlier Boundary} &= 14 - 1.5 \cdot 27 \\ &= 14 - 40.5 \\ &= -26.5 \text{ grams}\end{aligned}$$

and

$$\begin{aligned}\text{Upper Outlier Boundary} &= 41 + 1.5 \cdot 27 \\ &= 41 + 40.5 \\ &= 81.5 \text{ grams}\end{aligned}$$

Boxplot



Measure of Variability: **The Standard Deviation**

Nutsehell:

- A **standard** is the roughly considered to be the average.
- The **standard deviation** can be thought of as roughly the average distance of all of the data points from the mean.

Formally:

- The **standard deviation** measures how much - on average - individual scores of a given group vary (or deviate) from the mean score for this same group.

Formula:

- The **(sample) standard deviation** s is the square root of the sample variance:

$$s = \sqrt{\frac{\sum (Y - \overline{Y})^2}{N - 1}}$$

Compute the population standard deviation for the table below that lists the names of the professors along with the number of years they have worked in a department and their rank:

Name	Tenure in years	Rank
Erik	2	Assistant Professor
Ray	15	Full Professor
Kerrie	8	Associate Professor
Sam	7	Associate Professor
Karah	2	Assistant Professor

C

ompute the mean:

$$
\begin{aligned}
\overline{Y} &= \frac{\sum Y}{N} \\
&= \frac{2+15+8+7+2}{5} \\
&= \frac{34}{5} = 6.8
\end{aligned}
$$

Compute the deviations and square them:

Y	$Y - \overline{Y}$	$(Y - \overline{Y})^2$
2	$2 - 6.8 = -4.8$	$(-4.8)^2 = 23.04$
15	$15 - 6.8 = 8.2$	$(8.2)^2 = 67.24$
8	$8 - 6.8 = 1.2$	$(1.2)^2 = 1.44$
7	$7 - 6.8 = 0.2$	$(0.2)^2 = 0.04$
2	$2 - 6.8 = -4.8$	$(-4.8)^2 = 23.04$

Sum the squared deviations:

$$\sum (Y - \overline{Y})^2 = 23.04 + 67.24 + 1.44 + 0.04 + 23.04$$
$$= 114.80$$

Divide by $N-1$:

$$\begin{aligned}\frac{\sum(Y-\overline{Y})^2}{N-1} &= \frac{114.80}{4} \\ &= 22.70\end{aligned}$$

Take the square root:

$$\begin{aligned} s &= \sqrt{22.70} \\ &\approx 4.76 \end{aligned}$$

So the people are about 4.76 years away from the mean.

Measure of Variability: **The Variance**

Nutsehell:

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

Formally:

- The **variance** is a measure of how far the values in a data set are from the mean, on average that is always positive.

Formula:

- The **(sample) variance** s is:

$$s^2 = \frac{\sum (Y - \overline{Y})^2}{N - 1}$$

In the previous example, the variance was just

$$s^2 = 22.70$$

.

On your own:
A random sample of 10 American college students reported sleeping 7, 6, 8, 4, 2, 7, 6, 7, 6, 5 hours, respectively. What is the (sample) standard deviation and variance?