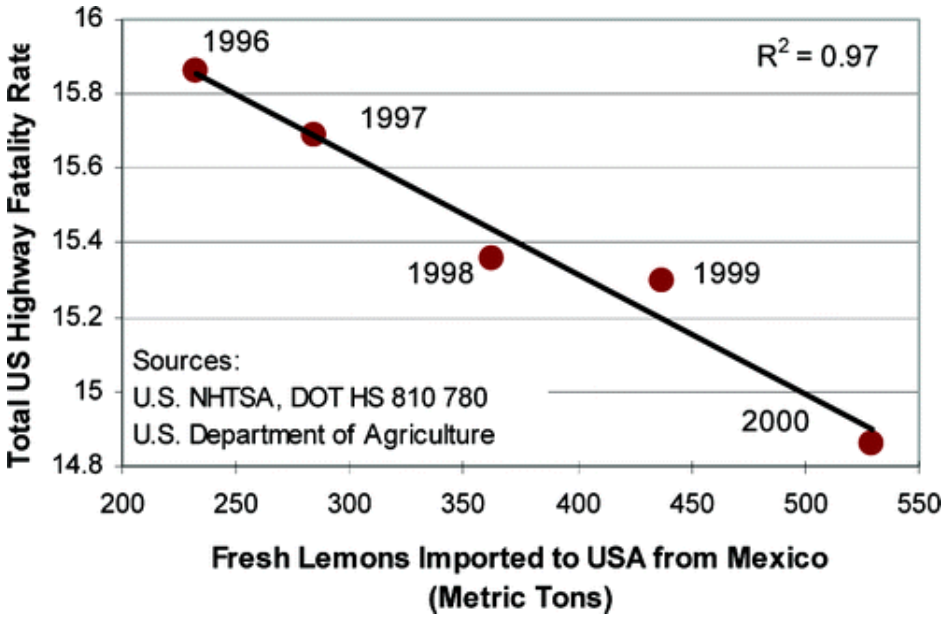
EDP 613 Chapter 4 Notes



CORRELATION DOES NOT PROVE CAUSATION!

(Johnson, 2007)

Measures of Variability

Definition: Variability is just how spread out a data set is.

The Range

Definition:

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

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

Example: Compute the range for the following sample:

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

Solution:

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

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

On Your Own: Compute the range for the following sample of gas prices (in dollars):

$$\$3.61 \quad \$3.84 \quad \$3.79 \quad \$3.61 \quad \$4.09 \quad \$3.96$$

Solution:

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

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

Quartiles

Definition:

- 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

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

Minimum First Quartile Second Quartile (Median) Third Quartile Maximum

Outliers

Definition:

- An **outlier** is a value that is considerably larger or smaller than most of the values in a data set. (NOTE: While deletion of an outlier is possible in certain circumstances, typically this is not performed unless that data point is an error.)

Definition:

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

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

NOTE: 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 $x <$ Lower Outlier Boundary or $x >$ Upper Outlier Boundary is an outlier.

Boxplots

Definition:

- A **boxplot** is a graph that presents the five-number summary along with some additional information about the data set. NOTE: There are many types, but the one presented here is called a modified boxplot.

NOTE: Procedure for Constructing a Boxplot:

1. Compute the first, second, and third quartiles and draw vertical lines representing them.
2. Draw horizontal lines between the first and third quartiles.
3. Compute the lower and upper outlier boundaries.
4. Use the largest data value less than the upper outlier boundary and a horizontal line from the third quartile to this value (called a **whisker**).
5. Use the smallest data value greater than the lower outlier boundary and a horizontal line from the first quartile to this value (also called a **whisker**).
6. If outliers exist, plot them separately.

Example: 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.

Solution:

First we have a minimum value of 15 minutes and a maximum value of 39 minutes. Now

1. For the first quartile, we have

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

In position 9, the data value is 19 minutes.

For the first quartile, we have

$$\begin{aligned} L_{\text{second}} &= \frac{50}{100} \cdot 22 \\ &= 0.50 \cdot 35 \\ &= 17.5 \\ &\approx 18 \end{aligned}$$

In position 18 the data value is 21 minutes.

For the third quartile, we have

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

In position 27, the data value is 22 minutes.

2. (Next page)

3. The IQR is given by:

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

so

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

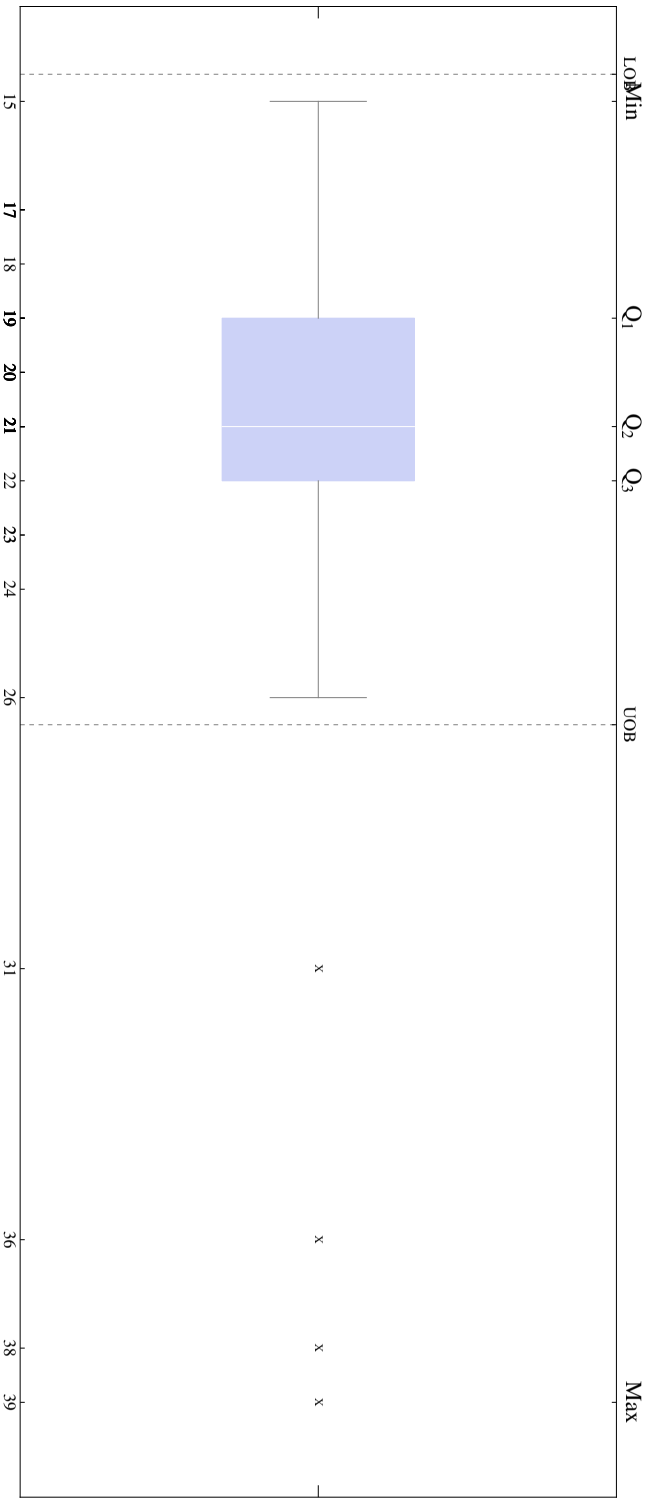
and

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

Thus the outliers are 31, 36, 38, 39 minutes respectively.

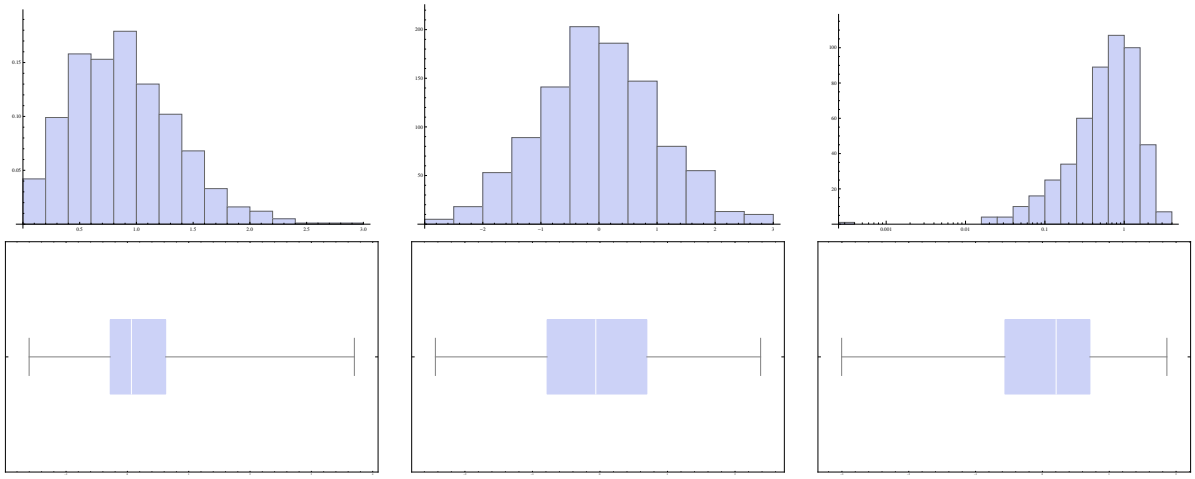
4. (Next page)

5. (Next page)


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

NOTE: 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.



On Your Own: 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

Source: www.starbucks.com

(a) Find the first, second and third quartiles of these data.

(b) Find the upper and lower boundaries.

(c) The beverage with the most carbohydrates is a Peppermint White Chocolate Mocha, with 59 grams. Is this an outlier?

(d) Construct a box plot for these data.

(e) Describe the shape of this distribution.

(f) There are 38 grams of carbohydrates in an Iced Dark Cherry Mocha. What percentile is this?

Solution:

First we reorder the data:

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

(a) For the first quartile, we have

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

In position 6, the data value is 14 grams.

For the first quartile, we have

$$\begin{aligned} L_{\text{second}} &= \frac{50}{100} \cdot 22 \\ &= 0.50 \cdot 22 \\ &= 11 \end{aligned}$$

In position 11, the data value is 27 grams.

For the third quartile, we have

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

In position 17, the data value is 41 grams.

(b) The IQR is given by:

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

So

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

and

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

1. No since 59 grams is between -26.5 grams and 81.5 grams.
2. (Next page)

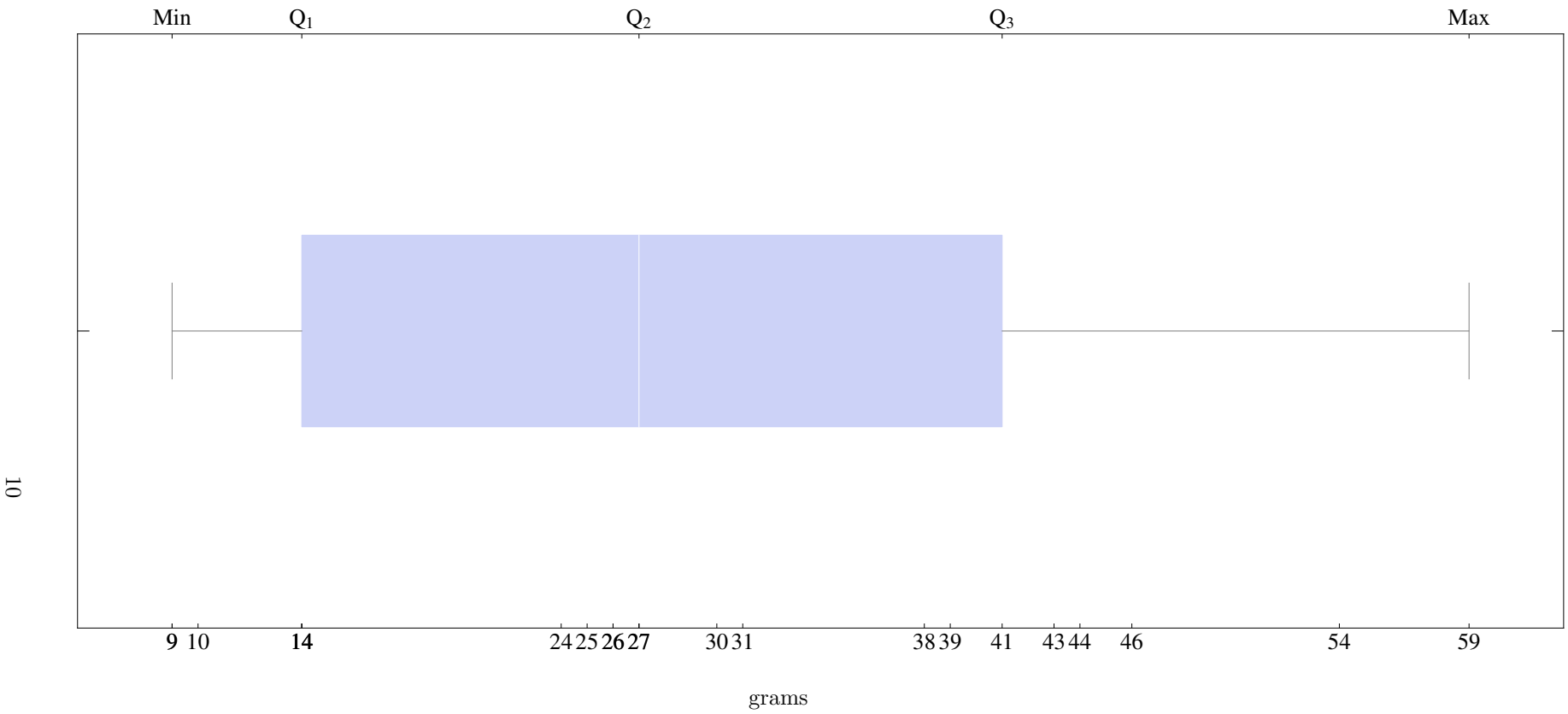
(e) The shape is skewed left since the lower whisker is significantly shorter than the upper whisker.

(f) We have

$$\begin{aligned} \text{Percentile} &= 100 \cdot \frac{14 + 0.5}{22} \\ &= 100 \cdot \frac{14.5}{22} \\ &\approx 100 \cdot 0.6591 \\ &= 65.91 \\ &\approx 66 \end{aligned}$$

Thus the Iced Dark Cherry Mocha is in the 66^{th} percentile according to the table.

(d)


Min
Q1
Q2
Q3
Max
10
9 10 14 24 25 26 27 30 31 38 39 41 43 44 46 54 59
grams

Deviation

Definition: A **deviation** is the distance from a data point to its mean.

The Standard Deviation

Nutshell:

- 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.

Formal:

- 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

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

Example: 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

Solution:

Using previous results, we have $\sigma^2 = 22.96$. Thus $\sigma = \sqrt{22.96} \approx 4.79$.

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?

Solution:

Using previous results, we have $s^2 = 3.07$. Thus $s = \sqrt{3.07} \approx 1.75$ hours away from the mean number of hours sleeping (5.8 hours).

Note: The greater the standard deviation, the greater the probability that any given measurement will have a value noticeably different from the mean.

The Variance

Nutshell:

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

Formal:

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

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

Formula:

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

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

Steps to compute the Population or Sample Variance:

1. Compute the mean μ or $\overline{x}$.
2. For each value x, compute the deviation ($x - \mu$ or $x - \overline{x}$).
3. Square the deviations yielding $(x - \mu)^2$ or $(x - \overline{x})^2$.
4. Sum the squared deviations yielding $\sum(x - \mu)^2$ or $\sum(x - \overline{x})^2$.
5. Divide the sum obtained in step 4 by N (for population) or $n - 1$ (for sample) to obtain the variance σ^2 or s^2.

Example: Compute the population variance 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

Solution:

1. $\mu = \frac{\sum x}{N} = \frac{2 + 15 + 8 + 7 + 2}{5} = \frac{34}{5} = 6.8$
2. (Included in Step 3)

3.

x	$x-\mu$	$(x-\mu)^2$
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	$2-6.8=-4.8$	$(-4.8)^2=23.04$

4. $\sum(x-\mu)^2 = 23.04+67.24+1.44+0.04+23.04 = 114.80$

5. $\sigma^2 = \dfrac{\sum(x-\mu)^2}{N} = \dfrac{114.80}{5} = 22.96.$

So the years of tenure are, on average, a distance of 22.96 squared years away from the mean number of tenure years (6.8 years).

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 variance?

Solution:

1. $\overline{x} = \dfrac{\sum x}{N} = \dfrac{7+6+8+4+2+7+6+7+6+5}{10} = \dfrac{58}{10} = 5.8$

2. (Skip to Step 3)

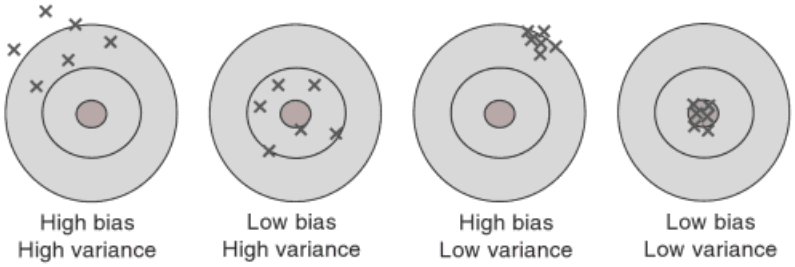
3.

x	$x-\overline{x}$	$(x-\overline{x})^2$
7	$7-5.8=1.2$	$(1.2)^2=1.44$
6	$6-5.8=0.2$	$(0.2)^2=0.04$
8	$8-5.8=2.2$	$(2.2)^2=4.84$
4	$4-5.8=-1.8$	$(-1.8)^2=3.24$
2	$2-5.8=-3.8$	$(-3.8)^2=14.44$
7	$7-5.8=1.2$	$(1.2)^2=1.44$
6	$6-5.8=0.2$	$(0.2)^2=0.04$
7	$7-5.8=1,2$	$(1.2)^2=1.44$
6	$6-5.8=0.2$	$(0.2)^2=0.04$
5	$5-5.8=-0.8$	$(-0.8)^2=0.64$

4. $\sum(x-\overline{x})^2 = 1.44+0.04+4.84+3.24+14.44+1.44+0.04+1.44+0.04+0.64 = 27.60$

5. $s^2 = \dfrac{\sum(x-\overline{x})^2}{n-1} = \dfrac{27.60}{10-1} = \dfrac{27.60}{9} = 3.07.$

So the hours of sleeping are, on average, a distance of 3.07 squared hours away from the mean number of hours sleeping (5.8 hours).



Bias Variance Decomposition. Figure 1. The bias-variance decomposition is like trying to hit the bullseye on a dart-board. Each dart is thrown after training our "dart-throwing" model in a slightly different manner. If the darts vary wildly, the learner is *high variance*. If they are far from the bullseye, the learner is *high bias*. The ideal is clearly to have both low bias and low variance; however this is often difficult, giving an alternative terminology as the bias-variance "dilemma" (*Dartboard analogy*, Moore & McCabe (2002))

Note:

Table II

Comparison of Variance and Standard Deviation

Criteria	Variance	Standard Deviation
Provides a measure of the spread of repeated measurements either side of the mean.	✓	✓
Always positive.	✓	
Bias is easily corrected.	✓	
Units are the same as those of the measurement.		✓
Allows you to determine how many significant figures are appropriate when reporting a mean value.		✓
Resistant		

References

Johnson, S. R. (2007). The trouble with QSAR (or how I learned to stop worrying and embrace fallacy). *Journal of Chemical Information and Modeling, 48*(1), 25-26.

Moore, D.S., McCabe, G.P., & Craig, B. (2010). *Introduction to the practice of statistics.* New York, NY: Freeman.