Chapter 2, continued. (pages 59 - 71)
Homework

- Homework due today by 4:00p
- HW 2 will be assigned by the end of the day
Diamonds Data Set

53940 obs. of 10 variables:

- carat - numeric
- cut - categorical (ordinal)
- color - categorical
- clarity - categorical (ordinal?)
- depth - numeric
- table - numeric
- price - numeric
- x - numeric
- y - numeric
- z - numeric
Boxplots as a Basis for Comparison

Boxplot of Carat versus Cut Quality

Cut Quality

Fair  Good  Very Good  Premium  Ideal
Boxplots as a Basis for Comparison

Boxplot of Price versus Cut Quality

Price

Cut Quality

Fair
Good
Very Good
Premium
Ideal

0
10000
20000
30000
40000
50000
60000
70000
80000
90000
100000
110000
120000
130000
140000
150000
Looking deeper into the Diamond Data

Bar Chart of Diamonds by Cut

Frequency

Fair | Good | Very Good | Premium | Ideal

Cut
Looking deeper into the Diamond Data

Histogram of Carat

- Frequency
- Carat

- Carat range: 0 to 4
- Frequency range: 0 to 7500

- Peaks at approximately 0.5 and 1.5 carats
Looking deeper into the Diamond Data

Scatter Plot of Price versus Carat
Suppose numeric data are given by \( y_1, \ldots, y_n \). Then

**Range**

The measure of dispersion called the range is given by

\[
\text{range} = y(n) - y(1);
\]

in other words, the range equals the maximum minus the minimum.
Let’s look at the range of the variables price and carat in the diamond data:

- price: $18,823 - 326 = 18,497$
- carat: $5.01 - .20 = 4.81$

What is a natural question upon seeing this information?
Let’s look at the range of the variables price and carat in the diamond data:

- price: $18,823 - 326 = 18,497$
- carat: $5.01 - .20 = 4.81$

What is a natural question upon seeing this information?
Inadequacy of the Range

It says nothing about the 52,000+ observations between $y(1)$ and $y(n)$:
Histogram of Carat Size of Diamonds

Carat Size
Frequency
0 1 2 3 4 5
0 5000 10000 15000
The range is not robust.

There are many cases where the range is not reliable: suppose, for example, that $y_{(n)}$ is a major outlier.
There are other measures of dispersion. Suppose numeric data are given by $y_1, \ldots, y_n$.

**Sample Variance**

The sample variance is denoted by $s^2$ and is given by the sum of the squared deviations divided by $n - 1$, or

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

Why do we emphasize “Sample”?
Definition

There are other measures of dispersion. Suppose numeric data are given by $y_1, \ldots, y_n$.

Sample Variance

The sample variance is denoted by $s^2$ and is given by the sum of the squared deviations divided by $n - 1$, or

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

Why do we emphasize “Sample”? 
More commonly,

**Sample Standard Deviation**

The sample standard deviation is denoted by $s$ and is given by

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

- $s$ is a statistic.
- $s$ is composed of deviations.
- $s$ has the same units as the $y_i$.
- Is $s$ robust?
More commonly,

**Sample Standard Deviation**

The sample standard deviation is denoted by \( s \) and is given by

\[
s = \sqrt{s^2} = \sqrt{\frac{1}{n - 1} \sum_{i=1}^{n} (y_i - \bar{y})^2}.
\]

- \( s \) is a *statistic*.
- \( s \) is composed of *deviations*.
- \( s \) has the same units as the \( y_i \).
- Is \( s \) robust?
More commonly,

**Sample Standard Deviation**

The sample standard deviation is denoted by $s$ and is given by

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

- $s$ is a *statistic*.
- $s$ is composed of *deviations*.
- $s$ has the same units as the $y_i$.
- Is $s$ robust?
Definition

More commonly,

Sample Standard Deviation

The sample standard deviation is denoted by $s$ and is given by

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

- $s$ is a statistic.
- $s$ is composed of deviations.
- $s$ has the same units as the $y_i$.
- Is $s$ robust?
Definition

More commonly,

Sample Standard Deviation

The sample standard deviation is denoted by \( s \) and is given by

\[
s = \sqrt{s^2} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2}.
\]

- \( s \) is a **statistic**.
- \( s \) is composed of **deviations**.
- \( s \) has the same units as the \( y_i \).
- Is \( s \) robust?
Why \( n - 1 \)?

Why isn’t

\[
s^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2?
\]

Why does it have \( n - 1 \) instead?

- See STAT 309/310 for more explanation
- \( s^2 \) is a function of deviations.
- Sum of deviations = \( \sum_{i=1}^{n} y_i - \bar{y} = 0. \)
- Because this is always true, \( s^2 \) is better thought of as an average of \( n - 1 \) random components instead of \( n \).
- We “lost” some randomness.
Example

Compute $s^2$ and $s$ for a toy example. Suppose the data are given by

$$1 \quad 3 \quad 4 \quad 4$$
Example

Compute $s^2$ and $s$ for a toy example. Suppose the data are given by

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

$$\bar{y} = \frac{1}{4} (1 + 3 + 4 + 4) = \frac{1}{4} (12) = 3.$$  The deviations are then

$$1 - 3 = -2, \quad 3 - 3 = 0, \quad 4 - 3 = 1, \quad 4 - 3 = 1.$$  

So that

$$s^2 = \frac{1}{4-1} ((-2)^2 + 0^2 + 1^2 + 1^2) = \frac{1}{3} 6 = 2.$$  Then

$$s = \sqrt{2} = 1.41.$$
Example

From the diamond data, we can calculate $s$ for price and for carat:

$s$:
- price: $3,989
- carat: 0.47

Recall, for range:
- price: $18,497
- carat: 4.81
From the diamond data, we can calculate $s$ for price and for carat:

**s:**
- price: $3,989
- carat: 0.47

Recall, for range:
- price: $18,497
- carat: 4.81
The Empirical Rule

If a distribution is unimodal and somewhat symmetric, then

- 68% of the observations fall within $\pm 1s$ of the mean
- 95% of the observations fall within $\pm 2s$ of the mean
- 99% of the observations fall within $\pm 3s$ of the mean

How is this possible? Is this arbitrary?
The Empirical Rule

If a distribution is unimodal and somewhat symmetric, then

- 68% of the observations fall within $\pm 1s$ of the mean
- 95% of the observations fall within $\pm 2s$ of the mean
- 99% of the observations fall within $\pm 3s$ of the mean

How is this possible? Is this arbitrary?
For diamond price, we have $\bar{y} = 3932.8$ and $s = 3989.44$. A calculator yields that

- 86% of the diamond prices fall in the interval $(\bar{y} - s, \bar{y} + s)$
- 93% of the diamond prices fall in the interval $(\bar{y} - 2s, \bar{y} + 2s)$
- 98% of the diamond prices fall in the interval $(\bar{y} - 3s, \bar{y} + 3s)$

What happened?
Assumptions

Unimodal: ok; Symmetric: not ok.
Test Empirical Rule on Student Interest in Class
For diamond price, we have $\bar{y} = 3.02$ and $s = 0.84$. A calculator yields that

- 50% of the interest levels fall in the interval $(\bar{y} - s, \bar{y} + s)$
- 93% of the interest levels fall in the interval $(\bar{y} - 2s, \bar{y} + 2s)$
- 100% of the interest levels fall in the interval $(\bar{y} - 3s, \bar{y} + 3s)$

What happened? Are we close enough? (Recall empirical rule predicts 68%, 95%, 99%.)
Does the Empirical Rule Ever Work?
Yes, and we will see cases where it does later. It is especially likely to work in biological systems with continuous numerical data.
In terms of increasing robustness:

- Measures of Center: $\bar{y} < \tilde{y}$
- Measures of Dispersion: range $< s < IQR$
Suppose that some function $f(x)$ takes the form

$$f(x) = mx + b.$$ 

We then call $f$ a linear transformation.

Examples:

- 1 USD = .77 Euro
- Degrees Fahrenheit = $\frac{9}{5}$ Degrees Celsius +32
- 1 inch = 2.54 cm
Linear Transformations

Suppose that some function $f(x)$ takes the form

$$f(x) = mx + b.$$  

We then call $f$ a linear transformation.

Examples:
- 1 USD = .77 Euro
- Degrees Fahrenheit = $\frac{9}{5}$ Degrees Celsius + 32
- 1 inch = 2.54 cm
Example

The weather forecast for several time points today is given by

\[ 6.11 \, 9.44 \, 9.44 \, 6.66 \, 7.78 \]

in Degrees Celsius. In Fahrenheit, these are

\[ 43 \, 49 \, 49 \, 44 \, 46 \]

Find the mean and standard deviation of these numbers and compare them.
Example

\[ \bar{y} = 7.89 \text{ and } s = 1.54. \]

\[ \bar{y} = 46.2 \text{ and } s = 2.78. \]

It happens that \( \frac{9}{5} \times 7.89 + 32 = 46.2 \) and \( (9/5)(1.54) = 2.78. \)
Example

\[
\begin{array}{cccccc}
6.11 & 9.44 & 9.44 & 6.66 & 7.78 \\
\end{array}
\]

\[\bar{y} = 7.89 \text{ and } s = 1.54.\]

\[
\begin{array}{cccccc}
43 & 49 & 49 & 44 & 46 \\
\end{array}
\]

\[\bar{y} = 46.2 \text{ and } s = 2.78.\]

It happens that \( \frac{9}{5} \cdot 7.89 + 32 = 46.2 \) and \( (9/5)(1.54) = 2.78. \)
Suppose the data take the form $y_1, \ldots, y_n$. Furthermore, suppose there is a linear transformation such that $x_i = my_i + b$ for every observation.

**Mean Linear Transformation**

$$\bar{x} = m\bar{y} + b$$

**Standard Deviation Linear Transformation**

$$s_x = |m|s_y$$

or

$$s_x^2 = m^2s_y^2$$