Lecture 6

What's Unusual? The Empirical Rule and z-Scores

§3.2

1. What's Unusual? The Empirical Rule

The standard deviation measures the typical distance of observations from the mean. When a distribution is approximately symmetric and unimodal, we can use the **Empirical Rule** to describe how much of the data lies near the mean.

Definition.
The **Empirical Rule** is a guideline describing how observations are distributed around the mean for a **unimodal and symmetric** distribution.
Approximately:

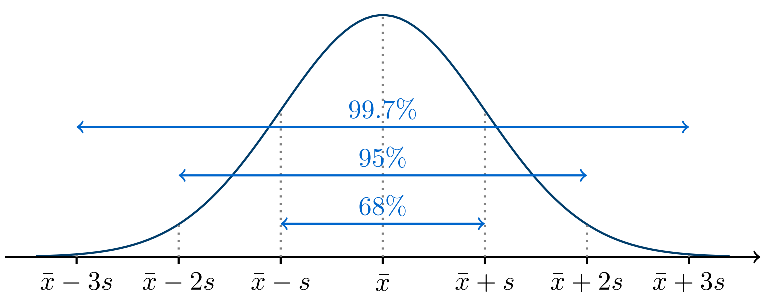
- 68% of the observations are within one standard deviation of the mean;
- 95% of the observations are within two standard deviations of the mean;
- 99.7% of the observations are within three standard deviations of the mean.

Equivalently,

$$\boxed{\begin{array}{c} 68\% \text{ within } \bar{x} \pm s \\ 95\% \text{ within } \bar{x} \pm 2s \\ 99.7\% \text{ within } \bar{x} \pm 3s \end{array}}$$

1.1 The 68–95–99.7 Rule

The Empirical Rule can be visualized using the following diagram.



Thus, the farther we move from the mean, the smaller the proportion of observations we expect to find.

$$\boxed{68\% - 95\% - 99.7\%}$$

For a unimodal, symmetric distribution:

$$\boxed{\begin{array}{cl} \bar{x}-s \text{ to } \bar{x}+s & \longrightarrow 68\% \\ \bar{x}-2s \text{ to } \bar{x}+2s & \longrightarrow 95\% \\ \bar{x}-3s \text{ to } \bar{x}+3s & \longrightarrow 99.7\% \end{array}}$$

How to use the Empirical Rule?

The Empirical Rule can be used whenever we know the mean and standard deviation and the distribution is approximately unimodal and symmetric.

The procedure is simple:

1. Start with the mean.
2. Add and subtract one, two, or three standard deviations.
3. Interpret the resulting interval using the Empirical Rule.

Example: Gas Prices

Recall the data set of gas prices from the previous lecture:

\$2.19	\$2.19	\$2.39	\$2.19
\$2.24	\$2.39	\$2.27	\$2.29
\$2.17	\$2.29	\$2.30	\$2.29

The mean gas price is

$$\bar{x} \approx \$2.2667$$

and the standard deviation is approximately

$$s \approx \$0.074.$$

If this sample is representative of a larger data set and the distribution is approximately symmetric, we can use the Empirical Rule.

Gas Price Summary

If the gas-price distribution is approximately unimodal and symmetric:

68%	:	\$2.19 to \$2.34
95%	:	\$2.12 to \$2.41
99.7%	:	\$2.04 to \$2.49

The percentages describe the expected proportion of observations in each interval.

Example: Daily High Temperatures in San Francisco

Suppose the mean daily high temperature in San Francisco is

$$\bar{x} = 65°\text{F}$$

with standard deviation

$$s = 8°\text{F}.$$

We want to find the temperature ranges containing approximately 68%, 95%, and 99.7% of the data.

68% of the Data

Within one standard deviation:

$$65 - 8 = 57$$

and

$$65 + 8 = 73.$$

Therefore,

$$\boxed{57°\text{F to } 73°\text{F}}$$

contains approximately 68% of the daily high temperatures.

95% of the Data

Within two standard deviations:

$$65 - 2(8) = 49$$

and

$$65 + 2(8) = 81.$$

Therefore,

$$\boxed{49°\text{F to } 81°\text{F}}$$

contains approximately 95% of the daily high temperatures.

99.7% of the Data

Within three standard deviations:

$$65 - 3(8) = 41$$

and

$$65 + 3(8) = 89.$$

Therefore,

$$\boxed{41°\text{F to } 89°\text{F}}$$

contains approximately 99.7% of the daily high temperatures.

68%	:	57°F to 73°F
95%	:	49°F to 81°F
99.7%	:	41°F to 89°F

What Is Considered Unusual?

By the Empirical Rule, approximately 95% of observations are within two standard deviations of the mean.

Therefore, observations that are more than two standard deviations away from the mean are relatively unusual.

Rule of Thumb:
An observation more than 2 standard deviations away from the mean is considered **unusual**. In symbols,

$$\boxed{|x - \bar{x}| > 2s \quad \Longrightarrow \quad \text{unusual}}$$

For the San Francisco temperatures,

$$\bar{x} = 65, \qquad s = 8.$$

Consider a day with a maximum temperature colder than

$$49°\text{F}.$$

Since

$$49 = 65 - 2(8),$$

a temperature below 49°F is more than two standard deviations below the mean.
Therefore,

$$\boxed{\text{A maximum temperature below } 49°\text{F would be unusual.}}$$

2. Working Backward with the Empirical Rule

The Empirical Rule can also be used when the mean and standard deviation are not given directly. If we know the interval containing approximately 68% of the data, we know that the endpoints are one standard deviation below and above the mean.

Example: Finding the Mean and Standard Deviation

Suppose the Empirical Rule tells us that approximately 68% of the data lies between

$$6.5 \qquad \text{and} \qquad 14.78.$$

Because this interval contains approximately 68% of the data,

$$\boxed{\bar{x} - s = 6.5}$$

and

$$\boxed{\bar{x} + s = 14.78.}$$

We can use these two equations to find the mean and standard deviation.

Step 1: Find the Mean

The mean is the midpoint of the two endpoints:

$$\bar{x} = \frac{6.5 + 14.78}{2}.$$

Therefore,

$$\bar{x} = \frac{21.28}{2} = \boxed{10.64}.$$

Step 2: Find the Standard Deviation

The standard deviation is the distance from the mean to either endpoint:

$$s = 10.64 - 6.5$$

so

$$\boxed{s = 4.14}.$$

We can check using the upper endpoint:

$$10.64 + 4.14 = 14.78.$$

The calculation is consistent.

Step 3: Find the 95% Interval

Approximately 95% of the data lies within two standard deviations of the mean.
Thus,

$$\bar{x} - 2s = 10.64 - 2(4.14) = 2.36$$

and

$$\bar{x} + 2s = 10.64 + 2(4.14) = 18.92.$$

Therefore,

$$\boxed{2.36 \text{ to } 18.92}$$

contains approximately 95% of the data.

Answer:

$$\boxed{\bar{x} = 10.64}$$

$$\boxed{s = 4.14}$$

and approximately 95% of the data lies between

$$\boxed{2.36 \text{ and } 18.92}.$$

3. z-Scores

The Empirical Rule tells us whether an observation is close to or far from the mean in terms of standard deviations.
A ***z*-score** gives us a precise numerical measure of this distance.

Definition.
A z**-score** measures how many standard deviations an observed value x is from the mean. The formula is

$$\boxed{z = \frac{x - \bar{x}}{s}}$$

where

- x is the observed data value;
- $\bar{x}$ is the sample mean;
- s is the sample standard deviation.

3.1 Interpreting a z-Score

The sign of the z-score tells us whether the observation is above or below the mean.

- If $z > 0$, the observation is above the mean.
- If $z < 0$, the observation is below the mean.
- If $z = 0$, the observation is exactly equal to the mean.

The absolute value tells us how far the observation is from the mean.
For example,

$$z = 2$$

means that the observation is two standard deviations above the mean.
Similarly,

$$z = -2$$

means that the observation is two standard deviations below the mean.

$$\boxed{z = 2 \quad \Longrightarrow \quad 2 \text{ standard deviations above the mean}}$$

$$\boxed{z = -2 \quad \Longrightarrow \quad 2 \text{ standard deviations below the mean}}$$

Example: Heights of Men

Suppose a data set contains the heights, in inches, of 247 men.
The mean height is

$$\bar{x} = 70$$

inches, and the standard deviation is

$$s = 3$$

inches.
We can use z-scores to determine how unusual different heights are.

A Man Taller Than Two Standard Deviations Above the Mean

We want to determine which men have

$$z > 2.$$

Using the z-score formula,

$$z = \frac{x - 70}{3}.$$

Thus,

$$\frac{x - 70}{3} > 2.$$

Multiplying by 3,

$$x - 70 > 6.$$

Therefore,

$$x > 76.$$

So men taller than 76 inches have z-scores greater than 2.

$$\boxed{z > 2 \quad \iff \quad x > 76 \text{ inches}}$$

The exact number of men is determined by counting the observations above 76 inches on the dotplot.

A Man More Than Two Standard Deviations Below the Mean

We want to determine which men have

$$z < -2.$$

Using the formula,

$$\frac{x - 70}{3} < -2.$$

Multiplying by 3,

$$x - 70 < -6.$$

Therefore,

$$x < 64.$$

Thus,

$$\boxed{z < -2 \quad \iff \quad x < 64 \text{ inches}}$$

The exact number of men is determined by counting the observations below 64 inches on the dotplot.

A Man Who Is 75 Inches Tall

For a man who is 75 inches tall,

$$x = 75.$$

Therefore,

$$z = \frac{75 - 70}{3}.$$

Thus,

$$z = \frac{5}{3} \approx \boxed{1.67}.$$

A man who is 75 inches tall has a z-score of approximately

$$\boxed{z = 1.67}.$$

Interpretation: his height is approximately 1.67 standard deviations above the mean height.

4. Using z-Scores to Compare Observations

A major advantage of z-scores is that they allow us to compare observations measured on different scales.

The original scores may have different means and standard deviations, but z-scores put the observations on the same scale.

When comparing two observations, the observation with the **larger z-score** performed better relative to its group.

$$\boxed{\text{Larger } z \quad \Longrightarrow \quad \text{better relative performance}}$$

Example: Maria's Statistics Exams

On which exam did Maria perform better when compared to the whole class?

The class statistics for the two exams were:

	Maria's Score	Mean	Standard Deviation
Exam 1	80	70	10
Exam 2	85	80	5

Although Maria received a higher raw score on Exam 2, we need to account for the different means and standard deviations.

Step 1: Calculate Maria's z-Score on Exam 1

For Exam 1,

$$x = 80, \qquad \bar{x} = 70, \qquad s = 10.$$

Therefore,

$$z_1 = \frac{80 - 70}{10}.$$

Thus,

$$z_1 = \frac{10}{10} = \boxed{1}.$$

Maria scored one standard deviation above the class mean on Exam 1.

Step 2: Calculate Maria's z-Score on Exam 2

For Exam 2,

$$x = 85, \qquad \bar{x} = 80, \qquad s = 5.$$

Therefore,

$$z_2 = \frac{85 - 80}{5}.$$

Thus,

$$z_2 = \frac{5}{5} = \boxed{1}.$$

Maria scored one standard deviation above the class mean on Exam 2.

Step 3: Compare the z-Scores

We found

$$z_1 = 1 \qquad \text{and} \qquad z_2 = 1.$$

Therefore, Maria performed equally well relative to her classmates on the two exams.

Answer:
Maria performed **equally well relative to the class** on the two exams.

$$\boxed{z_1 = z_2 = 1}$$

On both exams, Maria scored exactly one standard deviation above the class mean.
Although her raw score was higher on Exam 2, her relative performance was the same on both exams.