Sections 6A & 6B

Characterizing Data

&

Measures of Variation

The **distribution** of a variable (or data set) describes the values taken on by the variable and the frequency (or relative frequency) of these values.

The graphs we talked about earlier all show us how the data were distributed over various categories.

What is Average?

The **mean** is the arithmetic average and is given by:

$$\text{mean} = \frac{\text{sum of all values}}{\text{total number of values}}$$

The **median** is the middle value of the sorted data set.

To find the median: 1. Arrange the data in order.

2. odd # of values – median is the middle value.
3. even # of values – median is halfway between the middle two values

The **mode** is the most common value (or group of values).

Find the mean, median, and mode of the following exam scores:
63, 71, 74, 77, 79, 79, 80, 81, 81, 82, 82, 85, 86, 87, 91, 93

$$\begin{aligned}\text{mean} &= \frac{(63+71+\cdots+91+93)}{16} \\ &= \frac{1291}{16} \\ &= 80.6875\end{aligned}$$

The **mean is 80.6875**.

There are an even # of data values. The middle two values are 81 & 81, so the **median is 81**.

79, 81, and 82 all occur twice, which is the most that any value occurs. So the **modes are 79, 81, and 82**.

An **outlier** is a data value that is much higher or much lower than almost all other values.

Example: Consider the following salaries: $2500, $2500, $5000, $6000, $50000

$$\text{mean} = \frac{2500+2500+5000+6000+50000}{5} = \frac{66000}{5} = 13{,}200$$

The **mean is $13,200**.

The **median is $5,000**.

The **mode is $2,500**.

Note that the outlier drastically affects the mean, but does not affect the median or the mode.

Shapes of Distributions

Number of Peaks

We are primarily interested in the *general* shapes of distributions.

unimodal – single-peaked

bimodal – double-peaked

uniform distributions – no peaks

Note that the peak represents the mode of the distribution.

Shapes of Distributions

Symmetry or Skewness

A distribution is **symmetric** is its left half is a mirror image of its right half. (see top of page 376)

A distribution is **skewed** if it is not symmetric.

left-skewed – the values are more spread out on the left side

right-skewed – the values are more spread out on the right side

(See pictures on the bottom of page 376.)

Shapes of Distributions

Variation describes how widely data values are spread out about the center of a distribution.

low variation – most data values are clustered together

high variation – data values are spread apart

It is possible for two *different* data sets to have the same mean and median, but to have completely different shapes.

Measures of Variation

The **range** of a data set is the difference between its highest and lowest data values:

range = highest value – lowest value

The **lower quartile (first quartile)** – Q_1 – the median of the *lower half* of a data set

The **middle quartile (second quartile)** – the overall median

The **upper quartile (third quartile)** – Q_3 – the median of the *upper half* of a data set

Find the range and the quartiles for the exam scores:
63, 71, 74, 77, 79, 79, 80, 81, 81, 82, 82, 85, 86, 87, 91, 93

range = highest value – lowest value

= 93 – 63

= 30

The median (second quartile) is 81.

The lower half of the data is: 63, 71,74, 77, 79, 79, 80, 81

The lower (first) quartile, Q_1, is 78.

The upper half of the data is: 81, 82, 82, 85, 86, 87, 91, 93

The upper (third) quartile, Q_3, is 85.5

Five-Number Summary & Boxplot

The **five-number summary** consists of the following five numbers: low value, Q_1, median, Q_3, high value

A **boxplot** shows the five-number summary visually.

To draw a boxplot:

1. Draw and appropriately label a number line.
2. Draw a rectangular box that extends from Q_1 to Q_3.
3. Draw a line inside the box to indicate the median.
4. Draw “whiskers” extending to the low & high values.

Give the five-number summary and draw a boxplot for the exam scores.

Five-number summary

low = 63

Q_1 = 78

median = 81

Q_3 = 85.5

high = 93

The **standard deviation** is a measure of how far data values are spread around the mean of a data set.

$$\text{standard deviation} = \sqrt{\frac{\text{sum of (deviations from mean)}^2}{\text{total \# of data values} - 1}}$$

Find the standard deviation of the following set of data:
2.03, 0.27, 0.92, 1.07, 2.38

The mean is 1.334

Value	Deviation (value – mean)	(Deviation)^2
2.03	2.03 – 1.334 = 0.696	$(0.696)^2 = 0.484416$
0.27	0.27 – 1.334 = - 1.064	$(-1.064)^2 = 1.132096$
0.92	0.92 – 1.334 = - 0.414	$(-0.414)^2 = 0.171396$
1.07	1.07 – 1.334 = - 0.264	$(-0.264)^2 = 0.069696$
2.38	2.38 – 1.334 = 1.046	$(1.046)^2 = 1.094116$
		Sum = 2.95172

$$\text{The standard deviation is } \sqrt{\frac{2.95172}{5-1}} = \sqrt{\frac{2.95172}{4}}$$

$$= 0.8590285211$$

Range Rule of Thumb

The standard deviation is *approximately* related to the range of a set of data by the **range rule of thumb**:

$$\text{standard deviation} \approx \frac{\text{range}}{4}$$

The high and low values can be estimated by:

$$\text{low value} \approx \text{mean} - 2 \times \text{standard deviation}$$

$$\text{high value} \approx \text{mean} + 2 \times \text{standard deviation}$$

Use the range rule of thumb to estimate the standard deviation for the following set of data:
2.03, 0.27, 0.92, 1.07, 2.38

Range = high value – low value

= 2.38 – 0.27

= 2.11

$$\text{standard deviation} \approx \frac{\text{range}}{4} = \frac{2.11}{4}$$

The standard deviation is *approximately* 0.5275