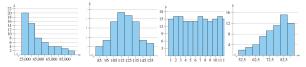**M. Singh**

Lesson 08: What is the five number summary of a distribution?

Q: *determine whether the approximate shape of the distribution in the histogram is symmetric, uniform, skewed left, skewed right, or none of these. Justify your answer.*

Match the distribution with one of the graphs above. Justify your decision.
1. The frequency distribution of 180 rolls of a dodecagon (a 12-sided die)
2. The frequency distribution of salaries at a company where a few executives make much higher salaries than the majority of employees
3. The frequency distribution of scores on a 90-point test where a few students scored much lower than the majority of students
4. The frequency distribution of weights for a sample of seventh grade boys

primethinker.com

When you want to describe a distribution...

Shape

Center

Spread

median

median

median

lower quartile

upper quartile
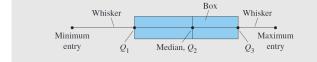
25% data below

25% data above

The **5**-number summary of a distribution reports its median, quartiles, and extremes (maximum and minimum).

The 5-number summary for the recent tsunami earthquake *Magnitudes* looks like this:

Max	9.0
Q3	7.6
Median	7.0
Q1	6.6
Min	3.7

Drawing a Box-and-Whisker Plot

1. Find the five-number summary of the data set.
2. Construct a horizontal scale that spans the range of the data.
3. Plot the five numbers above the horizontal scale.
4. Draw a box above the horizontal scale from Q_1 to Q_3 and draw a vertical line in the box at Q_2.
5. Draw whiskers from the box to the minimum and maximum entries.

Whisker Box Whisker
Minimum entry Q_1 Median, Q_2 Q_3 Maximum entry

M. Singh

Basic Computation: Five-Number Summary, Interquartile Range Consider the following ordered data:

2 5 5 6 7 7 8 9 10

(a) Find the low, Q_1, median, Q_3, and high.
(b) Find the interquartile range.
(c) Make a box-and-whisker plot.

Super Bowl points. How many points do football teams score in the Super Bowl? Here are the total numbers of points scored by both teams in each of the first 42 Super Bowl games:

45, 47, 23, 30, 29, 27, 21, 31, 22, 38, 46, 37, 66, 50, 37, 47, 44, 47, 54, 56, 59, 52, 36, 65, 39, 61, 69, 43, 75, 44, 56, 55, 53, 39, 41, 37, 69, 61, 45, 31, 46, 31

a) Find the median.
b) Find the quartiles.
c) Write a description based on the 5-number summary.

Values beyond the fences
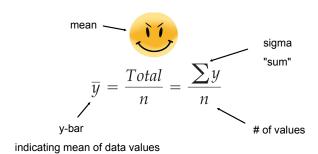of box and whisker plots

outliers

$$Upper\ fence = Q3 + 1.5\ IQR$$
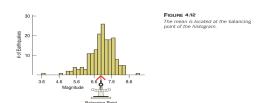$$Lower\ fence = Q1 - 1.5\ IQR$$

Center

Median is a good way to summarize the center of the distribution even when the distribution is skewed and has outliers.

But

Mean is an alternative to consider when dealing with symmetric distributions.

M. Singh

mean

$$\overline{y} = \frac{Total}{n} = \frac{\sum y}{n}$$

sigma
"sum"

y-bar

# of values

indicating mean of data values

The **mean** feels like the center because it is the point where the histogram balances

FIGURE 4.12
The mean is located at the balancing point of the histogram.

Using mean as a measure of center makes sense when the data is symmetric. Realize that an outlier can pull the mean its way.

Median is not affected by values far away.

Histogram Shape	Compare Two Measures Of Centers
If symmetric	mean and median are approximately equal
If right skewed	mean is greater than median
If left skewed	mean is less than median

Figure 4.6. Different Distributions

M. Singh

back to **spread**

1. Range: affected by outliers

2. IQR: ignores half the data

3. **Standard Deviation**: appropriate only for symmetric data.

takes into account how far each value is
from the mean

How does standard deviation work?

Suppose the batch of values is 14, 13, 20, 22, 18, 19, and 13.

(note to myself: make a list of steps)

mean $\bar{y}$

Deviation from the mean: $y - \bar{y}$

Deviation squared: $(y - \bar{y})^2$

Original Values	Deviations	Squared Deviations
14		
13		
20		
22		
18		
19		
13		

Add up the squared deviations:

Now divide by $n - 1$:

Finally, take the square root:

Notes: squaring emphasizes larger differences

M. Singh

Variance:
$$s^2 = \frac{\sum (y - \bar{y})^2}{n - 1}$$

Standard
Deviation
$$s = \sqrt{\frac{\sum (y - \bar{y})^2}{n - 1}}$$

Just Checking

6. The U.S. Census Bureau reports the median family income in its summary of census data. Why do you suppose they use the median instead of the mean? What might be the disadvantages of reporting the mean?

7. You've just bought a new car that claims to get a highway fuel efficiency of 31 miles per gallon. Of course, your mileage will "vary." If you had to guess, would you expect the IQR of gas mileage attained by all cars like yours to be 30 mpg, 3 mpg, or 0.3 mpg? Why?

8. A company selling a new MP3 player advertises that the player has a mean lifetime of 5 years. If you were in charge of quality control at the factory, would you prefer that the standard deviation of lifespans of the players you produce be 2 years or 2 months? Why?

M. Singh

Extra Credit: **up to 10 points** on any quiz

Think about why one might prefer one type of graph to another. In what ways are stem-and-leaf displays, dotplots, and histograms all similar? What information is apparent in some but not others? What are the advantages and disadvantages of using each type of graph?

Make a poster. Will be put on the class bulletin board.

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____