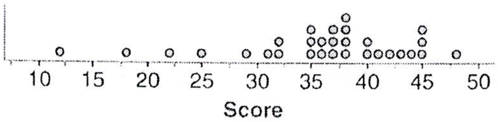
Chapter 2 Modeling Distributions of Data
Lecture Notes & Examples 2.1 Part B

Learning Targets

- Describe the effect of adding, subtracting, multiplying by, or dividing by a constant on the shape, center, and variability of a distribution of data.
- Use a density curve to model distributions of quantitative data

3. Transforming Data

Example: Below is a graph and table of summary statistics for a sample of 30 test scores. The maximum possible score on the test was 50 points.



	n	$\bar{x}$	s_x	Min	Q_1	M	Q_3	Max	IQR	Range
Score	30	35.8	8.17	12	32	37	41	48	9	36

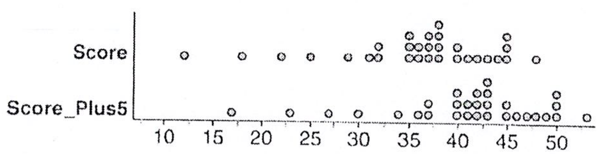
Suppose that the teacher was *nice* and added 5 points to each test score. How would this change the shape, center, and spread of the distribution?

Shape stayed the same

Center (mean + median) increased 5

Variability stayed the same (S_x, Range, IQR)

Here are the graphs and the summary statistics for the original scores and the +5 scores:



	n	$\bar{x}$	s_x	Min	Q_1	M	Q_3	Max	IQR	Range
Score	30	35.8	8.17	12	32	37	41	48	9	36
Score + 5	30	40.8	8.17	17	37	42	46	53	9	36

mean

median

Mean: 35.8 → 40.8 (+5)

Median: 37 → 42 (+5)

Min: 12 → 17 (+5)

Max: 48 → 53 (+5)

Effect of Adding (or Subtracting) a Constant

Adding the same number a (either positive, zero, or negative) to each observation:
- Adds a to measures of center and location (mean, median, quartiles, percentiles), but
- Does not change the *shape* of the distribution or measures of spread (range, IQR, standard deviation.

** Need to learn + understand **

Application: If 24 is added to every observation in a data set, the only one of the following that *is not* changed is:

adds 24 to | adds 24 to | adds 24 to | Spread/Variability stays the same | adds 24 to

(a) the mean (b) the 75th percentile (c) the median (d) the standard deviation (e) the minimum

Proof for IQR staying the same when adding "a"

① Orig IQR = $Q_3 - Q_1$

② New IQR = $(Q_3 + a) - (Q_1 + a) = Q_3 + a - Q_1 - a = Q_3 - Q_1 =$ orig IQR

Example (cont): Suppose that the teacher in the previous example wanted to convert the *original* test scores to percents. Since the test was out of 50 points, he should multiply each score by 2 to make them out of 100. Here are the graphs and summary statistics for the original scores and the doubled scores.

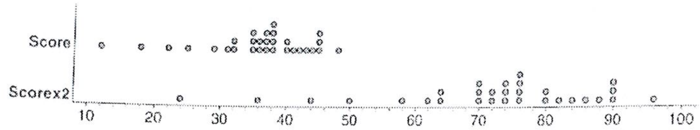
Proof IQR changing when multiplied by "6"

① Orig IQR = $Q_3 - Q_1$

② New IQR = $Q_3 \cdot 6 - Q_1 \cdot 6$

$= 6(Q_3 - Q_1)$

$= 6$ (orig IQR)



	n	$\bar{x}$	s_x	Min	Q_1	M	Q_3	Max	IQR	Range
Score	30	35.8	8.17	12	32	37	41	48	9	36
Score × 2	60	71.6	16.34	24	64	74	82	96	18	72

What happened the measures of center, location and spread? They doubled

What happened to the shape? Stays the same

Effect of Multiplying (or Dividing) by a Constant

Multiplying (or dividing) each observation by the same number b (positive, negative or 0)

- Multiplies (divides) measures of *center, location* (mean, median, quartiles, percentiles) by b,
- Multiplies (divides) measures of *spread* (range, IQR, standard deviation) by $|b|$, but
- Does not change the *shape* of the distribution.

* When Standardizing $\bar{x} = 0$ $S_x = 1$ *

4. Transformations and Z-Scores

convert to z-scores

Example (continued). Suppose we wanted to standardize the original test scores. This would mean we would subtract the mean of 35.8 from each score and then divide by the standard deviation of 8.17.

$$z_{score} = \frac{x - 35.8}{8.17}$$

	n	$\bar{x}$	s_x	Min	Q_1	M	Q_3	Max	IQR	Range
Score	30	35.8	8.17	12	32	37	41	48	9	36

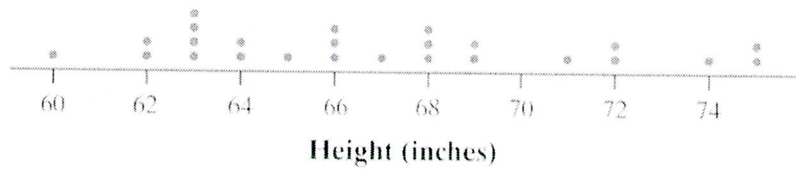
What effect would these transformations have on:

- Shape? shape stays the same
- mean * Center? Subtracting 35.8 would reduce the $\bar{x}$ by 35.8, making $\bar{x} = 0$. Dividing $\bar{x}$ by 8.17 would still have a mean ($\bar{x}$) of 0.
- Sx * Variability Spread? subtracting 35.8 would not change spread (variability) but ÷ by 8.17 would also ÷ our common measures of spread (variability) by 8.17. S_x, IQR, Range

The standard deviation would then be $\frac{8.17}{8.17} = 1$ $Sx = 1$.

CHECK YOUR UNDERSTANDING

The figure below shows a dotplot of the height distribution for Mrs. Navard's class, along with summary statistics from computer output.



Variable	n	$\bar{x}$	s_x	Min	Q_1	M	Q_3	Max
Height	25	67	4.29	60	63	66	69	75

inches $\cdot \frac{2.54 \text{ cm}}{\text{in}}$

1. Suppose that you convert the class's heights from inches to centimeters (1 inch = 2.54 cm). Describe the effect this will have on the shape, center, and spread of the distribution.

Each height will be multiplied by 2.54. The shape will not change. The center and spread will be multiplied by 2.54.

2. If Mrs. Navard had the entire class stand on a 6-inch-high platform and then had the students measure the distance from the top of their heads to the ground, how would the shape, center, and spread of this distribution compare with the original height distribution?

Each height will have 6 inches added to it. Shape and spread will not change. The center will have 6 inches added to it.

standardize $\qquad \bar{x}$ mean

3. Now suppose that you convert the class's heights to z-scores. What would be the shape, center, and spread of this distribution? Explain.

z score = $\frac{x - \text{mean}}{\text{standard deviation}}$

- The shape will not change
- The mean ($\bar{x}$) will change to 0.
- The standard deviation $S_x = 1$

$\bar{X}$ of zscore $= \frac{67-67}{4.29} = \frac{0}{4.29} = 0$

S_x of z-score $= \frac{4.29}{4.29} = 1$

need to know + understand

Page 99

5. Density Curves

Already Did in Chapter 1

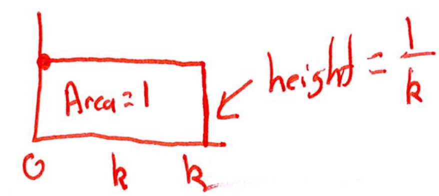
Exploring Quantitative Data
1. Always plot your data: make a graph, usually a dotplot, stemplot or a histogram.
2. Look for the overall pattern (shape, center, spread) and for striking departures such as outliers.
3. Calculate a numerical summary to briefly describe center and spread.

** New

New step:
4. Sometimes the overall pattern of a *large* number of observations is so regular that we can describe it with a *smooth curve*.

This type of *smooth curve* is called a **Density Curve**.

Density Curve



Definition: A **density curve** is a curve that

- Is always above the horizontal axis, and ← Need to know
- Has an area of exactly 1 underneath it

A density curve describes the overall pattern of a distribution. The area under the curve and above any interval of values on the horizontal axis is the proportion of all observations that fall in that interval.

Note: *no set of real data is exactly described by a density curve. The curve is an approximation that is easy to use and accurate enough for practical use.*

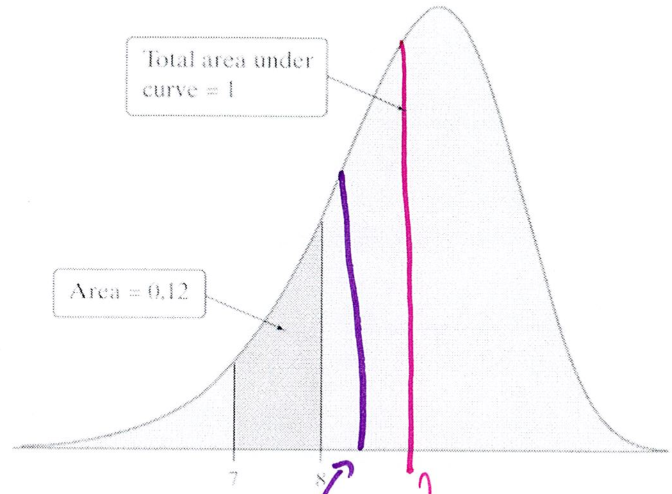
Because the density curve represents a *population* of individuals, the mean is denoted by μ (the Greek letter mu) and the standard deviation is denoted by σ (the Greek letter sigma).

Population mean = μ
S.D = σ

Distinguishing the Median and Mean of a Density Curve (Diagrams on p. 102)
- The **median** of a density curve is the *equal-areas* point, the point that divides the area under the curve in half.
- The **mean** of a density curve is the *balance point*, the point at which the curve would balance if made of solid material.
- The median and mean are the same for a perfectly symmetric density curve. They both lie at the center of the curve. The mean of a skewed curve is pulled away from the median in the direction of the long tail.

CHECK YOUR UNDERSTANDING

Use the figure shown to answer the following questions.



median

mean

1. Explain why this is a legitimate density curve.

This is a legit density curve b/c curve is positive everywhere and it has a total area of 1.

2. About what proportion of observations lie between 7 and 8?

About 12 %

3. Trace the density curve onto your paper. Mark the approximate location of the median.

Median ÷'s area under curve in half

4. Now mark the approximate location of the mean. Explain why the mean and median have the relationship that they do in this case.

The mean is less than the median in this case b/c the distribution is skewed left.

Summarize Big Ideas

Learning Target #1:

Learning Target #2:

LT #1: Given any list of #'s

- Add / Subtract a number "a" to each value...

    Shape + Variability Stay the same

    Center shifts up (down) by "a"

    Location shifts up (down) by "a"

- Multiply / Divide by a number "b"

    Shapes stays the same

    Center, Location + Variability multiply (divide) by |b|

Standardized Distribution (z scores) always mean = 0 and SD = 1

LT #2: Density Curves

- Always above x-axis
- Area = 1