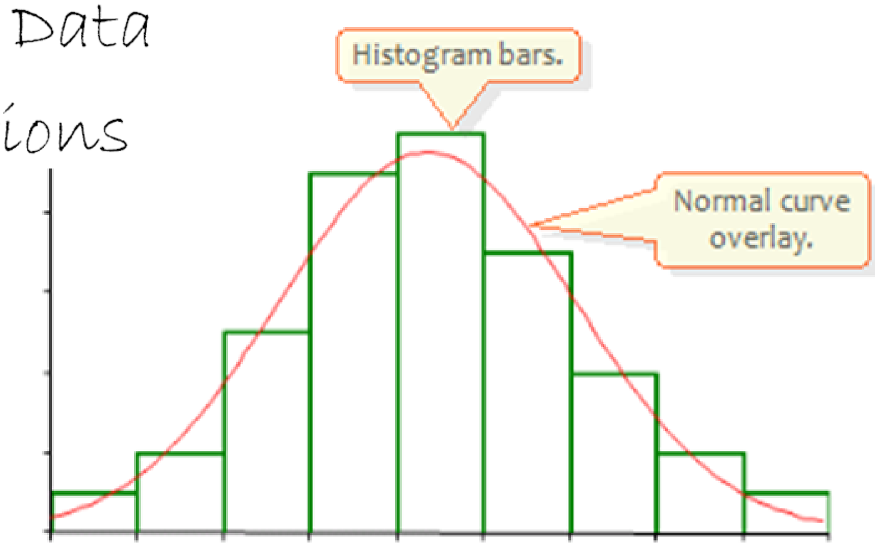
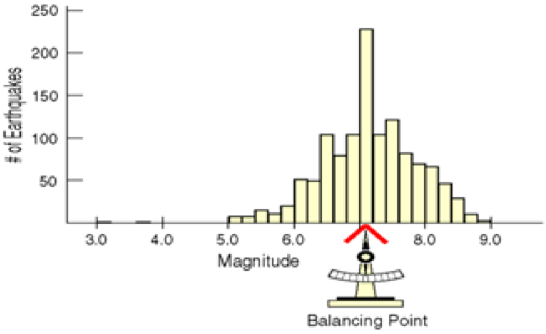
Unit 13:

Modeling Data Distributions



Mean of a distribution: average of data values that make up the distribution

-Also known as the balancing point.



Standard deviation:

-Closer the data is to the center, the smaller the standard deviation.

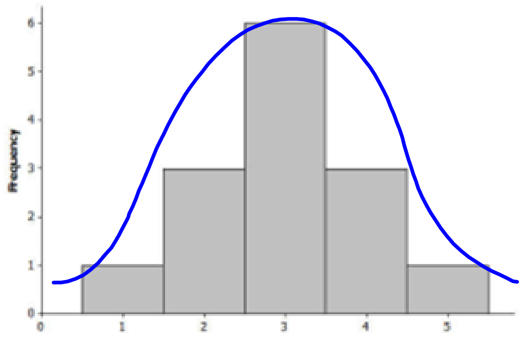
-Further the data is from the center, the larger the standard deviation.

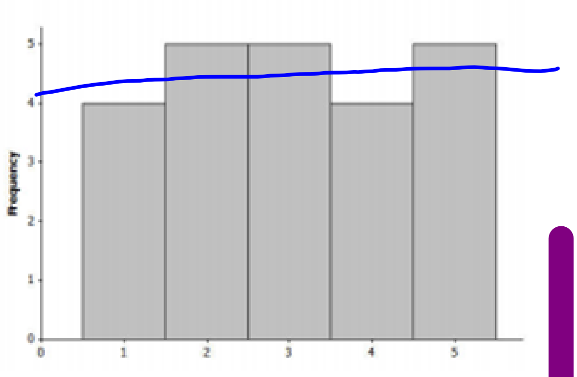
-Will be calculated with calculator in later units.

Distributions can be described by the shape (mound symmetric or skewed or uniform), the center mean or median, and the spread (variability) of the distribution.

Std. dev., Range, IQR
max - min Q3 - Q1

In each of the following, discuss the shape and center.



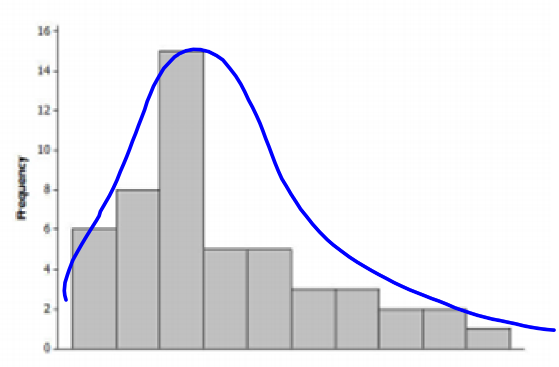


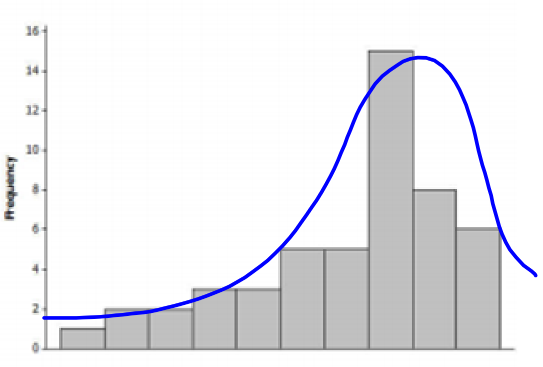
Shape: symmetric (mound shaped)

Mean: ~3

Shape: uniform

Mean: ~3





Shape: skewed right

Mean: pulled to the right of the peak

Shape: skewed left

Mean: pulled to the left of the peak

- tail on left

For skewed distributions, think of the following test grades you received: 100, 95, 98, 10 (hey it was a bad day-ok a REALLY bad day). What happens to your average after you get the 10?

average goes down - gets pulled to the left

Conclusion:

or Std. dev.

symmetric Max - min (#) - use mean to describe center

- If distribution is mound shaped: Use the Range______ to describe the spread of the distribution.
- If distribution is skewed: Use the IQR______ to describe the spread of the distribution.

Inner Quartile Range

use median to describe center.

Q3 - Q1 = IQR

IQR - Ignores skewed data and looks at the middle range

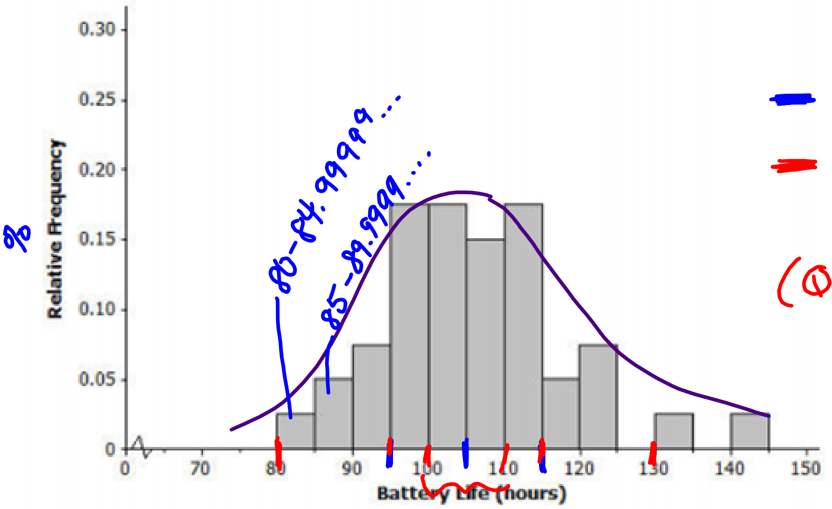
↳(on the ends)

- 50% of the data

→ ~ 2/3 of data should be within 1 Std. dev. of the mean

Relative Frequency Histogram: Heights of bars represent proportions (%) of the observations within each interval. NOT the number of observations within the interval.

Use the following histogram to answer the following questions:



Question 4

mean ?

standard deviation?
(Question 5)

1. What is the width of each bar?

   5 hours

2. What does the height of each bar represent?

   Proportion of all batteries with a life in the interval corresponding to the bar.

3. Would you describe the distribution of battery life as approximately symmetric or as skewed? Explain your answer.

   Approx. (roughly) symmetric. Right and left halves of the distribution are similar

4. Is the mean of the battery life distribution closer to 95, 105, or 115 hours? Explain your answer.

   Good habit - mark each one to see which is the best answer. Don't "look check"

   105. The data appears to be centered around 105.

5. Consider 5, 10, or 25 hours as an estimate of the standard deviation for the battery life distribution. Want majority but not all - shouldn't reach "the edge"

(~2/3 of data

a. Consider 5 hours as an estimate of the standard deviation. Is it a reasonable description of a typical distance from the mean? Explain your answer.

(100-110)

Most of the distances from the mean are greater than 5 hrs. Not a good estimate of SD.

b. Consider 10 hours as an estimate of the standard deviation. Is it a reasonable description of a typical distance from the mean? Explain your answer.

(95 to 115)

10 looks like a reasonable estimate of a typical distance from the mean. It's a reasonable est. of SD.

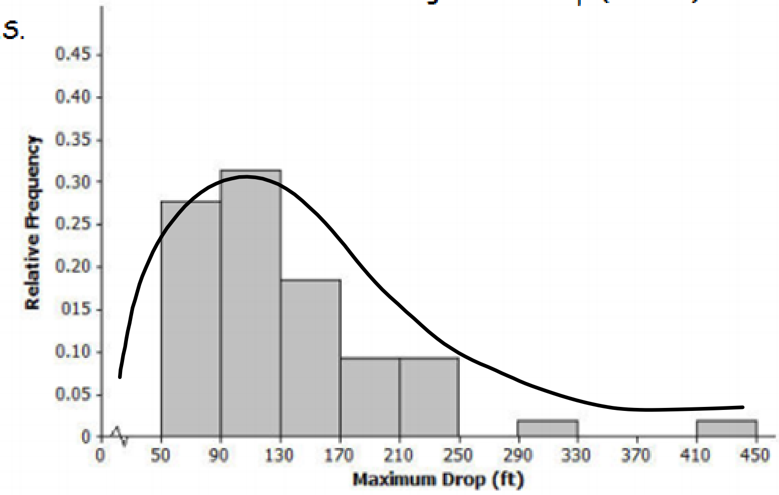
(approx. 2/3 of data is within 10 of the mean)

c. Consider 25 hours as an estimate of the standard deviation. Is it a reasonable description of a typical distance from the mean? Explain your answer.

(80 to 130)

Nearly all the values are less than 25 hrs. from the mean of 105. It's not a good est. of SD.

The histogram below shows the distribution of the greatest drop (in feet) for 55 major roller coasters in the U.S.



6. Would you describe this distribution of roller coaster maximum drop as approximately symmetric or as skewed? Explain your answer.

Skewed right because there's a long tail on the right side of the distribution

7. Is the mean of the maximum drop distribution closer to 90, 135, or 240 feet? Explain your answer.

135 because 90 is too small and 240 is too large to be considered a typical value for this data set.

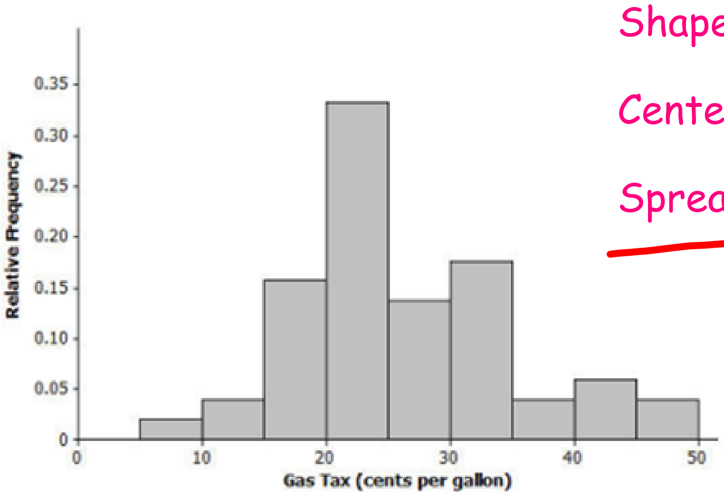
8. Would you use the range or IQR to describe the spread of the maximum drop distribution? Explain your answer.

IQR to describe the spread of max. drop distribution because it's skewed

(IQR leaves out the really high drop)

Vork with your partner on the following:

9. The histogram below shows the distribution of gasoline tax per gallon for the 50 states and the District of Columbia in 2010. Describe the shape, center, and spread (would you use the range or IQR) of this distribution.



Shape: Skewed Right

Center: ~(25 - 27)

Spread: IQR

or

Symmetric

Center - 2~~4~~5
25

Spread - Range

Consider the following histograms: Histogram 1, Histogram 2, Histogram 3, and Histogram 4. Descriptions of four distributions are also given. Match the description of a distribution with the appropriate histogram.

Histogram	Distribution
1	B
2	A
3	C
4	D

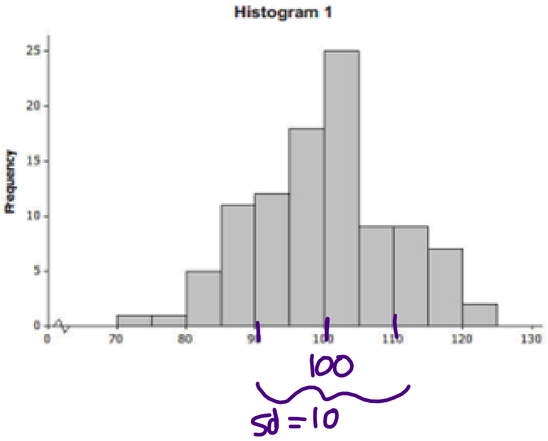
Description of distributions:

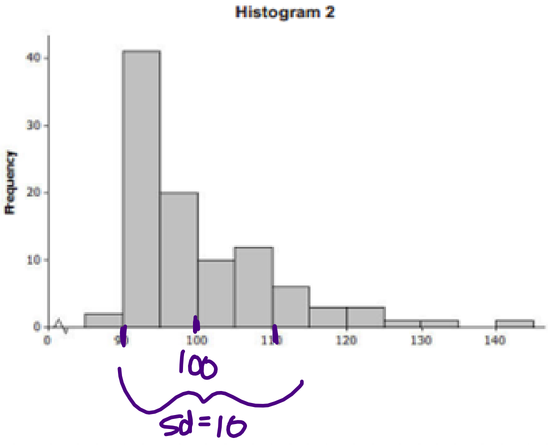
Distribution	Shape	Mean	Standard Deviation
A	Skewed to the right	100	10
B	Approximately symmetric, mound shaped	100	10
C	Approximately symmetric, mound shaped	100	40
D	Skewed to the right	100	40

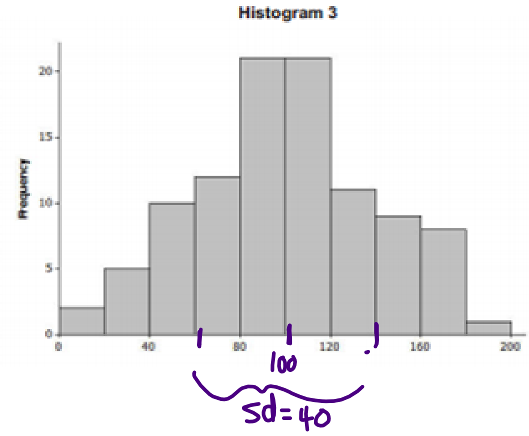
all = 100

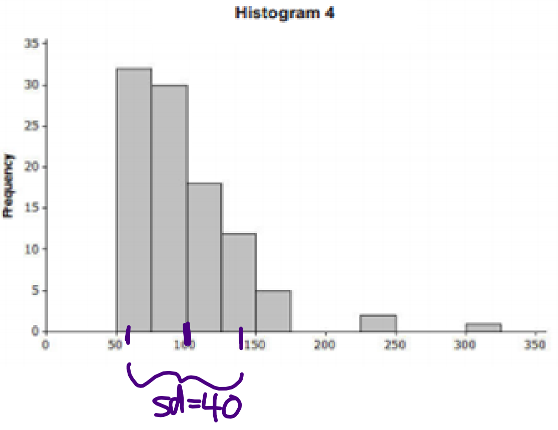
Given 2 choices for sd

Histograms:

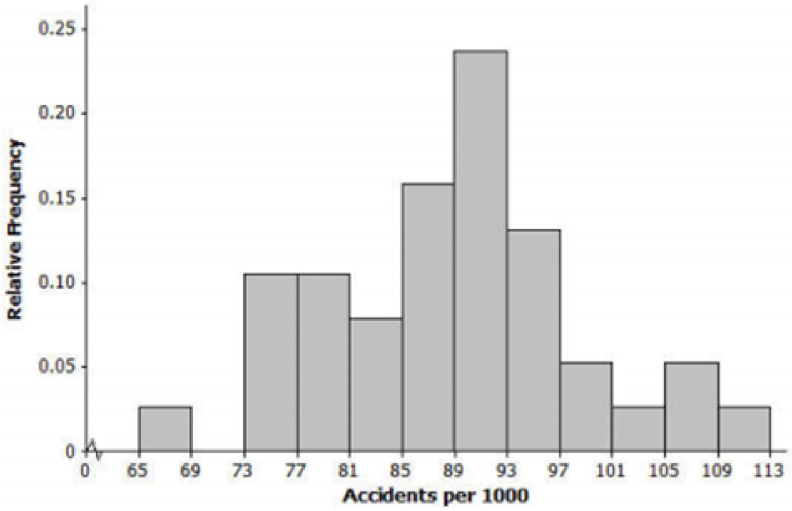








10. The histogram below shows the distribution of the number of automobile accidents per year for every 1,000 people in different occupations. Describe the shape, center, and spread (would you use the range or IQR) of this distribution.



Shape: ~ Symmetric; Mound Shaped

Center: ~89

Spread: Range (Symmetric ∴Range)