

Now, recompute \bar{x} without the outlier.

Def: Because the mean, \bar{x} , is so significantly affected by a few extreme observations it is called a "non resistant" measure of the center

Def: The "median" is the midpoint of a distribution. It is a number such that half the observations are smaller than M and half are larger.

Ex: Find the median of 7, 9, 4.

Ex: Find the median of 5, 10, 10, 14, 20, 25, 100

Note: the location of the median in an ordered list is $\left(\frac{n+1}{2}\right)$.

Thus in the first example of 4, 7, 9 we know that the location is $\frac{3+1}{2} = 2$.

Ex: (If n is even) Find the median of 5, 10, 10, 14, 20, 25.

The position of M is $\frac{6+1}{2} = 3.5$.

So it is the number midway between the 3rd and 4th numbers.

Ex: Compare the median of 5, 10, 15
with the median of 5, 10, 15000

Note: Unlike the mean, we see that the median is very resistant to extreme outliers.

Ex: Compare the mean of 5, 10, 15
with the mean of 5, 10, 15000

Applet: Mean and Median

Ex: Household Incomes — a skewed distribution
— slide 15 —

Notice that the extremely high incomes pull the mean higher but don't affect the median.

Thus the mean is always further in the direction of skew than the median.

Which (mean or median) would be further left in a distribution that is skewed to the left?

Measuring the Spread.

A simple way to measure spread is max-min.

However, the min and max values may be outliers.

Def: The "1st quartile" is a number which separates the first quarter of the observations from the larger ones.

Def: The "3rd quartile" is a number which separates the first three quarters from larger observations.

Ex: Find the quartiles of
5, 10, 10, 10, 10, 12, 15, 20, 20, 25, 30, 30, 40

1st quartile 10

2nd quartile 15 (Median)

3rd quartile 27.5

Ex: Find the quartiles of the above set when the last observation (40) is omitted.

$$(10, 13.5, 22.5) = (Q_1, Q_2, Q_3)$$

Def: The "five number summary" of a distribution is Min, Q1, M, Q2, Max

The "five number summary" provides a good description of the center and spread of a distribution.

A beautiful graph using the five number summary is a "box plot."

Ex: Travel Time to Work
slide 16

Compare the median and spread.

Ex: Investment types
slide 17

What can we learn from the box plots?

Advantages of box plots over histograms & stem plots:
- great for comparisons

Disadvantages:
- less information

Def: The "interquartile range" is $Q_3 - Q_1$.
This is a measure of a distribution's spread.

The interquartile range (IQR) helps us identify outliers.

We will call an observation a "potential outlier" if it is more than $1.5 * IQR$ above Q_3 or below Q_1 .

Ex: Find the 5-number summary and identify potential outliers.

5 10 10 15 15 15 15 20 20 20 25 30 30 40 40 45 60 60 65 85

min	5
Q_1	15
M	22.5
Q_3	42.5
max	85

$$IQR = 42.5 - 15 = 27.5$$

$$1.5 * IQR = 41.25$$

Potential outliers would be less than $15 - 41.25 = -26.25$
or greater than $42.5 + 41.25 = 83.75$

Thus 85 is the only potential outlier.

Why do we care about outliers?

Ex: A house in Manhattan, Kansas was mistakenly recorded as being worth 200 million rather than 60,000.

The county, school board, and city all receive revenue from property tax. Their budgets were calculated without realizing the mistake. This outlier Jacked up the total appraised value of real estate by 6.5%.

Thus they actually received much less revenue than they planned for.

It pays to check for outliers before trusting your data.

Def: The variance, s^2 , of a set of observations is a measure of spread and is calculated by

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n-1}$$

It is an average of how far observations fall from the mean.

Def: The standard deviation, s , is the most common measure of spread and is calculated as

$$s = \sqrt{\text{variance}}$$

Ex: Metabolic Rate (ie Calories burned over 24 hours)
Find the standard deviation of:

1792 1666 1362 1614 1460 1867 1439

$$\bar{x} = 1600$$

observation	Deviation	Squared deviation
1792	192	36,864
1666	66	4,356
1362	-238	56,644
		Sum = 214,870

$$\text{Thus } s^2 = \frac{214,870}{7-1} = 35,811.67$$

$$\text{and } s = \sqrt{35,811.67} = 189.24$$

Ex: Recompute the standard deviation using your calculator. This is S_x on the TI-30X IIS.

Note: The average we use for variance divides by $n-1$ (NOT by n).

The reason for this is that the deviations always sum to zero. Thus if we know $n-1$ of them, we can determine the last one

Def: Since only $n-1$ of them are freely determined, we say there are $n-1$ "degrees of freedom."

Advanced: Dividing by $n-1$ makes s^2 an unbiased estimator. (Don't Discuss This)

Properties of s : $s \geq 0$ why? When is $s=0$?

s is NOT resistant to outliers

s has the same units as the original observation

Ex: When measuring "metabolic rates (Calories per day)" \bar{x} and s are measured in Calories per day