Department of Mathematics

Descriptive Statistics

Section 15.2-15.3

Dr. John Ehrke
Department of Mathematics

Fall 2012

Measures of Central Tendency

Definition

The *mode* of a set of measurements is defined to be the measurement that occurs most often (with the highest frequency).

Definition

The *median* of a set of measurements is defined to be the middle value when the measurements are arranged from lowest to highest.

- The median for an even number of measurements is the average of the two middle values when the measurements are arranged from lowest to highest.
- When there are an odd number of measurements, there are an equal number of measurements above and below the median.
- When there is frequency data involved, calculate the relative cumulative frequency. The median of the data set lies in the first interval for which the relative cumulative frequency is greater than or equal to 0.5.

Descriptive Statistics with Formulas

Sample Mean: $\bar{x} = \frac{1}{n}\sum x_i$

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

Standard Deviation: $S_x = \sqrt{s^2}$

Standard Error:[1] $SE_{\bar{x}} = \frac{s}{\sqrt{n}}$

Confidence Interval: $CI = \bar{x} \pm SE_{\bar{x}}$

Grouped Mean: $\bar{x}_f = \frac{\sum xf}{\sum f}$

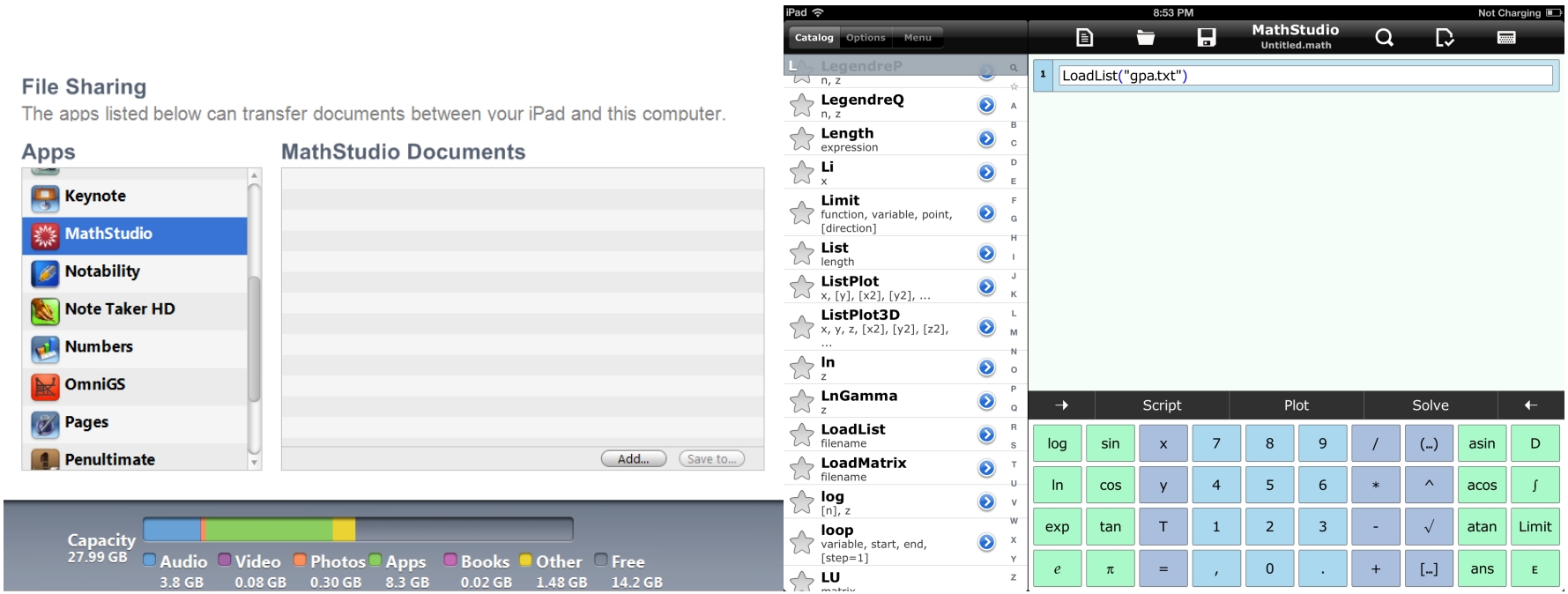
Grouped Variance: $S^2_{xf} = \frac{1}{n-1}\left(\sum x^2 f - \frac{(\sum xf)^2}{n}\right)$

Grouped Standard Deviation: $S_{xf} = \sqrt{S^2_{xf}}$

[1] If we assume the statistic is calculated on a standard normal then $s = 1$ and we call this the margin of error statistic.

Descriptive Statistics with Calculators

Even though we have several formulas on the previous slide that we will tackle at some point in this lecture a fair number of these formulas are built into most calculators. In this example we will walk through various calculations for a data set we will load into our calculators.



Example

Calculate the mean, median, range, and standard deviation of the data set “gpa.txt” located on the course blog.

Calculating Mean and Median of Grouped Data

For grouped data, or data sets involving frequencies, our approach for calculating the mean and median changes a bit. Consider the data set below that has been organized into a frequency table.

Data (x)	Frequency (f)	Cumulative Frequency	xf
1	4		
2	10		
3	9		
4	6		
5	4		
	$\sum f =$		$\sum xf =$

(a) Calculate the mean of the grouped data by filling in the rest of the table.

(b) Calculate the median of the grouped data by inspecting the cumulative frequency column of the table.

Polling Question #15

Suppose 40 students take an exam and 21 of them make the same grade, 70. Which of the following is true?

(a) The average test score is 70.

(b) The mode cannot be determined.

(c) The data set is skewed left.

(d) The median test score is 70.

Polling Question #16

A recent quiz in MATH 120 had the following scores with frequencies. Use this data to calculate the grouped mean.

Grades	Frequency
4	2
5	2
6	4
7	5
8	4
9	2
10	1

(a) $\overline{x} = 8.17$

(b) $\overline{x} = 7$

(c) $\overline{x} = 6.85$

(d) $\overline{x} = 6.9$

Measures of Variability

Data sets may have the same center (i.e. mean, median, mode) but look different because of the way the numbers *spread out* from the center. **Variability** or **dispersion** is a very important characteristic of data. The simplest measure of variation is the range.

Definition (Range)

The **range**, R, of a set of n measurements is the defined as the difference between the largest and smallest measurements.

Consider a small data set consisting of test grades:

98 95 79 89 92 96

Clearly, the range of this data set is $R = 98 - 79 = 19$. The range is easy to calculate, easy to interpret, and is an adequate measure of variability in many cases.

What are some cases in which the range is not a suitable measure of variability?

Another Measure Possible?

Definition (Sample Variance)

The **variance of a sample** of n measurements is the sum of the squared deviations of the measurements about their mean, $\bar{x}$ divided by $(n-1)$. The sample variance is denoted by s^2 and is given by the formula,

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}.$$

The sample standard deviation is calculated by taking the square root of the variance, that is, $s = \sqrt{s^2}$.

Data (x)	Deviation $(\overline{x} - x_i)$	$(\overline{x} - x_i)^2$
98		
95		
79		
89		
92		
96		
	$\sum =$	$\sum =$

Grouped Standard Deviation

A recent quiz in MATH 120 had the following scores with frequencies. Use this data to calculate the grouped standard deviation. Recall the formula for the grouped variance is given by

$$S_{xf}^2 = \frac{1}{n-1}\left(\sum x^2 f - \frac{(\sum xf)^2}{n}\right).$$

Grades	Frequency	xf	x^2f
4	2		
5	2		
6	4		
7	5		
8	4		
9	2		
10	1		
	$\sum f =$	$\sum xf =$	$\sum x^2 f =$

Polling Question #17

If the range of a data set is 20, which of the following must be true?

(a) The standard deviation of the set is equal to 20.

(b) The standard deviation of the set is less than 20.

(c) The standard deviation of the set is greater than 20.

(d) The standard deviation is greater than the mean of the data set.

Coefficient of Variation

The final measure of variability is called the *coefficient of variation*.

Definition (Coefficient of Variation)

The coefficient of variation, CV, expresses the standard deviation as a percentage of the sample mean. The coefficient of variation is given by the formula,

$$CV = \frac{s_x}{\bar{x}} \times 100\%.$$

Some interesting observations about the coefficient of variation:

- CV is useful when you are interested in the size of variation relative to the size of observation.
- CV is independent of the units of observation.

For example, the value of the standard deviation of a set of height data will be different depending on whether the data was collected in inches or in feet. The CV, however, will be the same in both cases as it does not depend on the unit of measurement. We will verify this in our next example.

Independent of Units

Example

The heights of the five starters for the ACU women's basketball team are listed as, 6-0, 5-9, 5-6, 6-3, and 5-9 recorded as feet and inches. Using this data, calculate the standard deviation and coefficient of variation for this data. Compare your results when the data is measured in inches only.

Solution:
We are comparing two data sets, one in feet and the other in inches:

$$x = [6, 5.75, 5.5, 6.25, 5.75]$$
$$y = [72, 69, 66, 75, 69]$$

If you calculate the coefficient of variation for each data set we obtain,

$$CV_x = \frac{s_x}{\overline{x}} = 4.87\% \qquad CV_y = \frac{s_y}{\overline{y}} = 4.87\%.$$

Measures of Relative Standing

A **percentile** is a measure of relative standing most often used for large data sets.

Definition (Percentile)

Suppose a set of n measurements of the variable x have been arranged from smallest to largest. The p^{th} **percentile** is the value of x that is greater than p% of the measurements and is less than the remaining (100-p)%.

Example

Suppose you have been notified that your score on the verbal portion of the SAT was 610, and that this value was good enough to place you in the 85^{th} percentile in the distribution of scores. Where does you score stand in relation to the scores of others who took the exam?

In general, the 85th percentile for the variable x is a point on the horizontal axis of the data distribution that is greater than 85% of the data measurements and less than 15% of the total measurements.

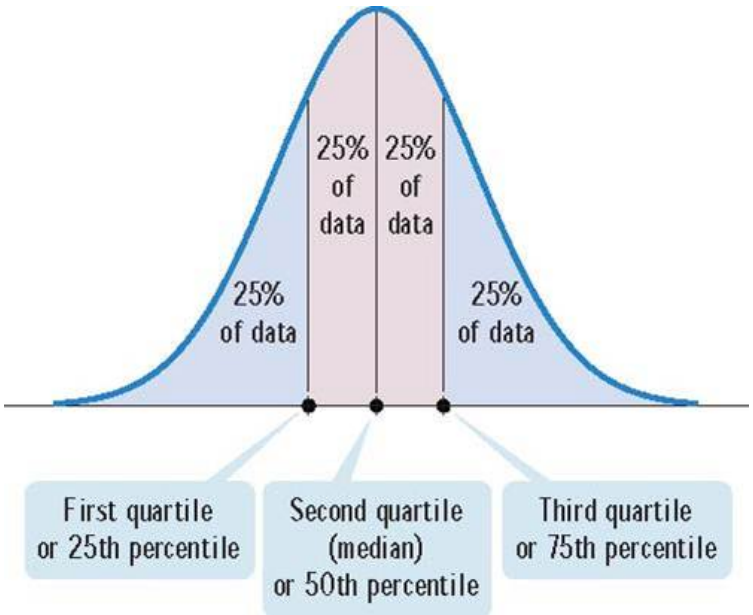
What is another name for the 50th percentile?

Quartiles

There are three particular percentiles that are worth noting:

- 25th percentile, lower quartile (Q1)
- 50th percentile (Median)
- 75th percentile, upper quartile (Q3)

Together the lower and upper quartiles, along with the median, located points in the distribution that divide the data into four sets, each containing an equal number of measurements. Twenty five percent of the measurements will be less than the first quartile, 50% less than the median, and 75% less than the upper quartiles.



Calculating Sample Quartiles

For small data sets, it is often impossible to divide the set into four groups each of which contains exactly 25% of the measurements. For example, when $n = 10$ you would need 2.5 measurements in each group which is not possible. Even when you can perform this task (i.e. $n = 12$) there are too many numbers that could be considered as quartiles. To avoid this ambiguity we use the following rule to locate sample quartiles:

- Arrange the measurements from smallest to largest.
- Calculate the median of the data set. This will divide the data set into halves.
- The lower quartile, Q_1 is the median of the first half of the data.
- The upper quartile, Q_3 is the median of the second half of the data.

The Five Number Summary

Definition (Five Number Summary)

The five number summary consists of the smallest number, the lower quartile, the median, the upper quartile, and the largest number, presented in order from smallest to largest:

Min Q_1 **Median** Q_3 **Max**

By definition, one fourth of the measurements in the data set lies between each of the four adjacent pairs of numbers.

The five number summary can be used to create a simple graph called a **box plot** to visually describe the data distribution. From the box plot, you can quickly detect any skew in the distribution, as well as clearly identify any **outliers**.

Polling Question #18

Choose the answer which represents the five number summary (min, Q1, med, Q3, max) for the set of measurements below.

$$\text{DATA} = [16, \quad 25, \quad 4, \quad 18, \quad 11, \quad 13, \quad 20, \quad 8, \quad 11, \quad 9]$$

(a) 16, 4, 12, 8, 9

(b) 4, 9, 13, 18, 25

(c) 4, 10, 12, 19, 25

(d) 4, 9, 12, 18, 25

Interquartile Range

Because the median and the quartiles divide the distribution into four parts, each containing approximately 25% of the data. Q_1 and Q_3 are the upper and lower boundaries for the middle 50% of the distribution. We can measure the range of this "middle 50%" by calculating the interquartile range.

Definition

The interquartile range (IQR) for a set of measurements is the difference between the upper and lower quartiles; that is, IQR = Q_3 - Q_1.

From our previous example,

$$IQR = Q_3 - Q_1 = 18 - 9 = 9$$

Together with the quartiles, median, min, and max values of the data set, the IQR allows us to construct another graph useful for describing data sets.

Outliers

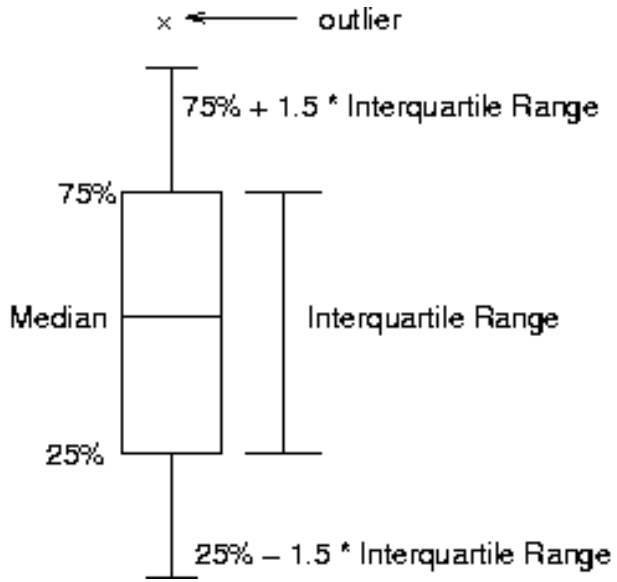
Outliers are common occurrences in statistics. An outlier may result from transposing digits during recording, from incorrectly reading an instrument panel, a defect in the experimental design, or from other problems. Even when there are no recording or observational errors, a data say may contain one or more valid measurements that, for one reason, or another, differ markedly from the others in the set. These outlier can cause a marked distortion in the values of $\bar{x}$ and s_x.

In fact, outliers may themselves contain important information not shared with the other measurements in the set. Therefore, isolating outliers, if they are present, is an important step in the preliminary assessment of a data set. The box plot is designed expressly for this purpose.

How unusual is too unusual? (i.e. When does a valid data point become an outlier?)

Constructing a Box Plot

- Calculate the median, the upper and lower quartiles, and the IQR for the data.
- Draw a horizontal line representing the scale of measurement. Form a box just above the horizontal line with left end at Q_1 and right end at Q_3.
- Mark any outliers with an asterisk (*) on the graph.
  - Lower Fence: $Q_1 - 1.5(IQR)$ (values below this number are outliers)
  - Upper Fence: $Q_3 + 1.5(IQR)$ (values above this number are outliers)
- Extend horizontal lines (called “whiskers”) from the ends of the box to the smallest and largest observations that are not outliers.



An Example With Outliers

Example

The miles per gallon (mpg) for each of 20 medium sized cars selected from a production line during the month of March are listed below:

23.1	21.3	23.6	23.7	20.2
24.4	25.3	27.0	24.7	22.7
26.2	23.2	25.9	24.7	24.4
24.2	24.9	22.2	22.9	24.6

(a) Find the mean, median, and mode for this sample.

(b) Arrange the data from smallest to largest and find the five number summary for this data.

(c) Would you consider the largest and smallest observations to be outliers? Why or why not.