Lecture 3:
Lesson and Activity Packet

MATH 232: Introduction to Statistics

September 12, 2016

Last time, we discussed:

- Measures of Dispersion
  - Range
    - Not terribly enlightening, but easy to compute
  - Variance
    - Useful as an intermediate step to compute standard deviation
    - Is zero only when all data points are identical
  - Standard Deviation
    - Tells how far the data points generally are from their mean
    - For many data sets, the vast majority of points fall within two standard deviations of the mean—a quick way of identifying "unusual" data
- Difference between population and sample

Questions?

Last time, we discussed *measures of central tendency* and *measures of dispersion* for what is called **ungrouped** data—that is, data where you have a list of points.
Sometimes, though, we will be limited to **grouped data**—where we do not have access to the data sets themselves, but to a *frequency table* or *histogram* that shows groups.

Example 1 (*Grouped Data*)

Suppose that the Berkshire Eagle publishes the results of a study on the amount of PFOA[a] in water tested at 80 locations in Berkshire County, but instead of giving the values of the 80 tests the study conducted, the newspaper report showed only the following table:

PFOA range (ppt)	*Frequency*
5.0 – 8.9	*3*
9.0 – 12.9	*10*
13.0 – 16.9	*14*
17.0 – 20.9	*25*
21.0 – 24.9	*17*
25.0 – 28.9	*9*
29.0 – 32.9	*2*

The numbers in the right-hand column of the table are called class frequencies*, and show how many items fall into each class; the smallest and largest values permitted in any given class are called the* upper and lower class limits*.*

For example, Line 1 of the table means that 3 tests showed results between 5.0 and 8.9 parts per trillion (abbreviated 'ppt' in the table heading). The class frequency is 3, and the class limits are 5.0 and 8.9.

Notice that if any tests measure more digits of precision than are given in the class limits, their still need to be placed in one of the columns; this is done by rounding. For example, if a particular location tested at 12.94 parts per trillion, it would be put in Class 2 according to the above table; if at 12.96 parts per trillion, it would be put in Class 3. Really, then, Class 2 contains values ranging from 8.95 to 12.95. These are referred to as the class boundaries *or the* real class limits*.*

[a] PFOA, or *Perfluorooctanoic Acid*, is a toxic substance that pollutes groundwater; it causes various types of cancers and is toxic to the liver and immune system, interferes with human development, and exerts hormonal effects including alteration of thyroid hormone levels. It can result from irresponsible dumping of waste from factories that coat cookware with nonstick substances like Teflon (and was used for many other things, too, e.g., microwave popcorn bags, Gore-Tex boots, and many water repellents). Factories run for decades by the ChemFab and Saint-Gobain chemical companies have been blamed for the presence of PFOA in groundwater in excess of the states' limits in the towns of Hoosick Falls, NY and in North Bennington, VT, both just across the borders from North Adams. The Massachusetts state government currently does not require municipalities like North Adams to test their public water supply for PFOA, and so citizens of this city do not know if their water contains this dangerous chemical. So the data given in this example is completely fictional, mainly because it either does not exist, or is not being released to the public. Please see Canvas module for more information about PFOA.

Group Exercise 1

Take the data from Problem 1 of Homework 1, and make a grouped distribution from it using five classes that evenly divide the data's range.

Definition 1 (*Class Mark*)

The class mark *of a class in a frequency distribution is the mean of the upper and lower class limits. For the first class in Example 1, the class mark is computed as*

$$x_1 = \frac{5.0 + 8.9}{2} = \frac{13.9}{2} = 6.95.$$

Group Exercise 2

For the data you classified in Exercise 1, compute the class marks.

Even if we only have access to grouped data, we sometimes are interested in measures of central tendency or deviation.

Definition 2 (*Mean of Grouped Data*)

The mean of grouped data is computed as the mean for ungrouped data, assuming that each item in the class has exactly the value of the corresponding class mark.

For a distribution with k classes that have class marks x_1, x_2, ..., x_k and corresponding class frequencies f_1, f_2, ..., f_k, the mean of the distribution is given by

$$\frac{\sum_{i=1}^{k} x_i \cdot f_i}{\sum_{i=1}^{k} f_i},$$

and is called $\bar{x}$ if the data was a sample, and μ if the data was a population.

Group Exercise 3

What is the mean of the data shown in Example 1?

Definition 3 (*Class Interval*)

For a class in a table of grouped data, the class interval *is the upper class limit minus the lower class limit. It is merely the length of a class, or the range of values it can contain.*

Group Exercise 4

Find the class intervals for the data you classified in Exercise 1.

Definition 4 (*Median of Grouped Data*)

The median of grouped data *is computed assuming that the values in each class are spread evenly throughout that class. The idea is to find class in which the median value occurs by adding all class frequencies from either end until finding one class that contains the $n/2$ (or $\frac{n+1}{2}$, in case n is odd) position. Then add the fraction of the class range that is spanned by the number of data points still needed to get to $n/2$ or $\frac{n+1}{2}$.*

As a formula, this is

$$L + \frac{j}{f}c,$$

where L is the lower boundary of the class into which the medium must fall, f is that class's frequency, c is its class interval, and j is the number of items we still lack when we reach L.

An example follows.

Example 2

For the data in Example 1, we had $n = 80$. *This is an even number, so we seek the value of the data point at position* $\frac{n}{2} = \frac{80}{2} = 40$. *We must count 40 of the tests starting at either end; starting from the bottom, we find that* $3 + 10 + 14 = 27$ *of the values fall into the first three classes, and that* $3+10+14+25 = 52$ *of the values fall into the first four classes.*

Therefore, we must count $40 - 27 = 13$ *values beyond the lower limit of the fourth class. On the assumption that the 25 values in the fourth class are spread evenly throughout the data, we can do this by adding* $\frac{13}{25}$ *of the class interval (which is* $20.95 - 16.95 = 4$*) to the lower class boundary (which is 16.95, the value beyond which a data point will be rounded up into the fourth class). This gives us that the median is*

$$16.95 + \frac{13}{25} \cdot 4 \approx 19.03.$$

Group Exercise 5

Compute the median of the data from Exercise 1.

Definition 5 (*Modal Class of Grouped Data*)

The modal class of grouped data *is the class mark of the class with the maximum frequency.*

Group Exercise 6

What is the modal class of the data in Exercise 1? What is the modal class of the data in Example 1?

Definition 6 (*Range for grouped data*)

The range for grouped data *is the difference between the maximum upper class limit and the minimum lower class limit.*

Group Exercise 7

Find the range of the data in Exercise 1 and find the range of the data in Example 1.

Definition 7 (*Variance and Standard Deviation of Grouped Data*)

The variance of grouped data *is given by the formula*

$$\frac{\sum_{i=1}^{k} x_i^2 f_i - \left(\frac{\sum_{i=1}^{k} x_i f_i}{\sum_{i=1}^{k} f_i}\right)^2}{\left(\sum_{i=1}^{k} f_i\right) - 1},$$

in the case of a sample, where we assume there are k many classes; that the x_i denote the class marks and the f_i the corresponding class frequencies. For data from a population instead of a sample, we write the variance as

$$\frac{\sum_{i=1}^{k} x_i^2 f_i - \left(\frac{\sum_{i=1}^{k} x_i f_i}{\sum_{i=1}^{k} f_i}\right)^2}{\left(\sum_{i=1}^{k} f_i\right)}.$$

The standard deviation of grouped data *is the square root of the variance of the grouped data.*

Example 3

Find the mean and the standard deviation of the following distribution, giving the amounts of time that 80 college students devoted to leisure activities during a typical school week:

Hours	*Frequency*
10–14	*8*
15–19	*28*
20–24	*27*
25–29	*12*
30–34	*4*
35–39	*1*

Before using the formulas from the definitions, we make it easier on ourselves by creating a new table:

Class Mark x	x^2	*Frequency* f	$x \cdot f$	$x^2 \cdot f$
12	*144*	*8*	*96*	*1152*
17	*289*	*28*	*476*	*8092*
22	*484*	*27*	*594*	*13068*
27	*729*	*12*	*324*	*8748*
32	*1024*	*4*	*128*	*4096*
37	*1369*	*1*	*37*	*1369*

Then we substitute the necessary values into the formulas to find:

$$\textit{mean} = \frac{\sum_{i=1}^{k} x_i \cdot f_i}{\sum_{i=1}^{k} f_i} = \frac{1655}{80} \approx 20.6875 \approx 20.69 \textit{ hours},$$

and

$$\textit{variance} = \sqrt{\frac{\sum_{i=1}^{k} x_i^2 f_i - \left(\frac{\sum_{i=1}^{k} x_i f_i}{\sum_{i=1}^{k} f_i}\right)^2}{\left(\sum_{i=1}^{k} f_i\right)}} = \sqrt{\frac{36525 - \frac{1655^2}{80}}{79}} \approx 5.38 \textit{ hours}.$$

Group Exercise 8

Find the standard deviation of the data in Example 1.

Theorem 1 (*Chebyshev's Theorem*)

For any data set (population or sample) and any constant k greater than 1, the proportion of data that must lie within k standard deviations on either side of the mean is at least

$$1 - \frac{1}{k^2}.$$

Example 4

We can be sure that at least $1 - \frac{1}{2^2} = \frac{3}{4}$, or 75%, of the values in ***any*** *data set lie within two standard deviations on either side of the mean; at least $1 - \frac{1}{3^2} = \frac{8}{9}$, or 88.9 percent, of the data must lie within three standard deviations on either side of the mean; and at least $1 - \frac{1}{5^2} = \frac{24}{25}$, or 96 percent, of the data must lie within five standard deviations on either side of the mean. To compute these, we let $k = 2, 3,$ and 5.*

Group Exercise 9

If all the one-pound cans of coffee filled by a food processor have a mean weight of 16.00 ounces with a standard deviation of 0.02 ounces, at least what percentage of the cans must contain between 15.860 and 16.20 ounces of coffee?

Chebyshev's Theorem applies to **any** kind of data, but it tells us only "at least what percentage" must lie between certain limits. For nearly all sets of data, the actual percentage of data lying between the limits is much greater than that specified by Chebyshev's Theorem. For general bell-shaped distributions, the much stronger statement holds:

Theorem 2 (*Empirical Rule*)

For a data set with a bell-shaped distribution, approximately 68% of the values lie within one standard deviation of the mean; approximately 95% of the data will lie between two standard deviations of the mean; and about 99.7% of the data will lie within three standard deviations of the mean.

Group Exercise 10

If the actual mean of the data in Example 1 is 18.896, and the actual standard deviation is 5.6565, then what percentage of the data actually falls within one standard deviation of the mean? Within two standard deviations? Three?

Definition 8 (*z-Scores*)

A z-score is the number of standard deviations that a given observation, x, is below or above the mean. It tells us how the particular data point compares with others in its same set. It is computed by

$$z := \frac{x - \bar{x}}{s} \text{ or } z := \frac{x - \mu}{\sigma},$$

depending on whether the data was a sample or a population.

Example 5

Two-year-old models of a certain kind of car have been selling for a mean price of \$7,860 with a standard deviation of \$820, whereas three-year-old models of the same kind of car have been selling at a mean price of \$6,400 with a standard deviation of \$960. Leaving all other considerations aside, is a two-year-old model priced at \$6,960 a greater bargain than a three-year-old model priced at \$5,400?

Converting both prices into their z-scores (sometimes called ***standard units*** *instead), we have*

$$z_2 = \frac{6960 - 7860}{820} \approx -1.10$$

for the two-year-old car, and

$$z_3 = \frac{5400 - 6400}{960} \approx -1.04$$

for the three-year-old car. Even though the two-year-old model is priced \$900 below average and the three-year-old model is priced \$1,000 below average, the two-year-old model is priced relatively lower for cars of the same kind, and hence, it is the greater bargain.

Group Exercise 11

The mean height in a group of men is 68.34 inches, with standard deviation 3.02 inches; the mean weight is 172.55 pounds with standard deviation 26.33 pounds.

Which is more extreme: a man who is 76.2 inches tall, or a man who weighs 237.1 pounds?

Definition 9 (*Coefficient of Variation*)

The coefficient of variation *expresses the standard deviation as a percentage of what is being measured; that is, the standard deviation divided by the mean:*

$$V := \frac{s}{\bar{x}} \cdot 100\%, \text{ or } V := \frac{\sigma}{\mu} \cdot 100\%,$$

depending on a population or a sample being measured.

Example 6

As in the previous exercise, the mean height in a group of men is 68.34 inches, with standard deviation 3.02 inches; the mean weight is 172.55 pounds with standard deviation 26.33 pounds.

Computing the coefficients of variation will tell us whether the heights, or the weights, vary more.

The coefficient of variation of the heights is

$$CV_h = \frac{s}{\bar{x}} \cdot 100\% = \frac{3.02\ in}{68.34\ in} \cdot 100\% \approx 4.42\%,$$

and of the weights is

$$CV_w = \frac{s}{\bar{x}} \cdot 100\% = \frac{26.33\ lb}{172.55\ lb} \cdot 100\% \approx 15.26\%.$$

Although the standard deviation of 3.02 inches cannot be compared to the standard deviation of 26.33 pounds, we can compare the coefficients of variation, and in doing so, see that heights (CV=4.42%) have considerably less variation than weights (CV=15.26%). This should not be surprising.

Group Exercise 12

Have you ever seen a man twice as tall as another? What about twice as heavy? Does it makes sense that the variation in weight is much greater than variation in height?

Recap

- Grouped data
  - Class frequencies
  - Class limits
  - Class boundaries
  - Class mark
- Measures of Central Tendency and Dispersion for Grouped Data
  - Mean for Grouped Data
  - Median for Grouped Data
  - Modal class for Grouped Data
  - Range for Grouped Data
  - Variance for Grouped Data
  - Standard Deviation for Grouped Data
- Chebyshev's Theorem
- Empirical rule
- z-scores
- Coefficients of Variation

Homework

- Homework 1 due today at beginning of class; quiz at beginning of Wednesday's class on everything learned so far.