Lecture 1:
Lesson and Activity Packet

MATH 232: Introduction to Statistics

September 7, 2016

How This Works: Each class, we will try to work through a packet like this (and we'll keep some local paper companies in business, I suspect; please recycle!). There are some notes, and some exercises. Individual exercises are in green boxes, and group exercises are in yellow boxes. There are also some worked examples and definitions as part of the notes. If there is ever a packet we don't complete, I'd ask you to work through the exercises and read the notes as best you can before the subsequent class; you can always ask Svetlana (the SI tutor) or me (Dr. Kiley) for help if you're having trouble understanding something. We will work through the packets together as a class, so please don't read too far ahead, unless you just can't help yourself...

When trying to think about data sets, we might want a convenient single value that somehow represents all of the data in the set at once.

For example, from MCLA's web site, we learn that the "average class size" is 18 (we will find out shortly how to skew this figure up or down for the particular example of class size). There are many, many classes offered at MCLA every semester, and the "average class size" is a way of quickly giving one value that is representative of all of the class sizes.

Definition 1 (*Measure of Central Tendency*)
A measure of central tendency *is a "typical" numerical value that helps describe a data set.*

An important note! Mathematicians and statisticians *avoid* using the word "average" to describe a measure of central tendency, because it can actually mean several things.

We will discuss four popular measures of central tendency today:

- Mean
- Median
- Mode
- Midrange

The *mean* is what many people think of, when they think of "averages". For example, the term "grade point average" typically denotes the mean of a student's course grades. The mean is defined as follows:

Definition 2 (*Mean and Sample Size*)

For the data set[a]

$$\{x_1, x_2, \ldots, x_n\},$$

the mean value *is denoted* $\bar{x}$, *and has value*

$$\bar{x} := \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}.$$

Here, the number n *is referred to as the* sample size*—because there are* n *many points in the sample.*

[a]Data sets are always written as lists of numbers enclosed by "curly braces" { }. For example, the set of ages of people in my knitting group is $\{30, 22, 60, 83, 37, 42, 19\}$.

Notes about this definition:

- The definition relies on a *mathematical abstraction*; that is, the data set that is given as an example is not a set of actual numbers, but each of the x_i stands for a number in a possible data set. In terms of this abstraction, the "knitting group" data set has $n = 7$, $x_1 = 30$, $x_2 = 22$, $x_3 = 60$, and so forth.
- The definition uses *summation notation*. If this notation is not familiar to you, please check Canvas for a link that explains it.
- The equals sign in the expression $\frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1+x_2+x_3+\cdots+x_n}{n}$ and the equals sign in the expression $\bar{x} := \frac{1}{n}\sum_{i=1}^{n} x_i$ look different (the latter one has a colon in front of it). Some authors like to distinguish the assertion "A is equal to B" from "define A as an object equal to B". If you have taken a computer science course, then you have probably already learned this distinction; for those who have not and who are curious, you will find a link on Canvas to further explanations. A great deal of mathematics has to do with fine distinctions like these.

Group Exercise 1

How would you describe the above formula in words?

The mean is the sum of the values of the data points, divided by the total number of data points.

Example 1 (*Computing the Mean*)

I said earlier that the set of ages of people in my knitting group is $\{30, 22, 60, 83, 37, 42, 19\}$. *The mean age is computed as follows:*

$$\text{Mean Age} = \frac{30 + 22 + 60 + 83 + 37 + 42 + 19}{7} = \frac{293}{7} \approx 41.857.$$

If I were reporting this mean in an official document where I would not want to mislead readers, I would round this number to the nearest whole, and report that the mean age among the members of my group is 42[a].

[a]I would round because if I reported the mean age as 41.857 years, it might lead readers to believe that I had more precise knowledge of the ages to begin with—that I knew more *significant digits* (decimal places) than I actually did. My data set consists of just whole, two-digit numbers with no decimals, so any measure of central tendency of that set should not contain more than two significant figures either. See Canvas for a link that explains significant figures.

Individual Exercise 2

Suppose that the total number of people who entered Village Pizza on Eagle Street on Saturday was 67; on Sunday, it was 33; on Monday, it was 54; on Tuesday, it was 20; and on Wednesday, it was 26. Write this data as a set. What is the sample size? What is the mean number of people who entered the pizzeria each day?

Data set is $\{67, 33, 54, 20, 26\}$.

Mean is $\frac{67+33+54+20+26}{5} = \frac{200}{5} = 40$ people

units important!

In some cases (where all the data points are non-negative numbers), the mean actually gives us a cap on the maximum value among the data. Look at the formula again:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}.$$

Multiply both sides by n:

$$n\bar{x} = \sum_{i=1}^{n} x_i = x_1 + x_2 + x_3 + \cdots + x_n.$$

If all data points (that is, all of the x_i) are non-negative, then in the extremal case, one of them is nonzero, and the rest are zero; in that case, the only nonzero value is equal to $n\bar{x}$, the sample size times the mean. The maximum possible value for any of the data points, then, is the sample size times the mean. In math words, that is, for all data points x_i in the set of sample size n (that is, for whole-number i values where $1 \leq i \leq n$), it is true that $x_i \leq n\bar{x}$.

Group Exercise 3

If the mean annual salary paid to the three top administrators of a private university is \$156,000, can one of them receive an annual salary of \$500,000?

We know the sample size is $n = 3$, and the mean is $\bar{x} = \$156,000$.

The maximum possible value in the data set is therefore $n \cdot \bar{x} = 3(\$156,000)$

$= \$468,000.$

Since \$500,000 is greater than the \$468,000 that we found to be the maximum possible salary, we know that no an administrator cannot receive that much.

What if there is one extreme value? Say that in addition to the seven people already in my knitting group originally, a 97-year-old joins us next Wednesday. Let's see what that does to the mean age:

$$\text{Mean Age} = \frac{30 + 22 + 60 + 83 + 37 + 42 + 19 + 97}{8} = \frac{390}{8} = 48.75.$$

Because one 97-year-old joined, this bumped the mean age up to 49, instead of 42 as it was before. But the others all remained the same age. Since the mean cannot resist substantial changes caused by extreme values, we say that the mean is **not** a *resistant* measure of central tendency.

Example 2 (*Class Size Paradox*)

Let's first compute the mean class size at a university by recording the size of each class, and letting our data set be the list of class sizes (so that the sample size is equal to the number of classes offered). We then take the mean of that data set. For example, suppose that our college offers five classes with sizes in the set $\{36, 124, 15, 7, 150\}$. Then the mean would be:

$$\textit{Mean Class Size} = \frac{36 + 124 + 15 + 7 + 150}{5} = \frac{332}{5} = 66.4.$$

Now, let's compute the mean class size experienced by students—which is a different concept than the mean class size, because there are more students in larger classes. For each student, the size of each class he or she is taking is put into a large data set. For example, if Jehan is taking classes of size 36, 15, and 7, and Maria is taking classes of size 36, 15, and 124, then together, they contribute those six points to the data set. In this case, the sample size is the sum of the class sizes (because each student in each class contributes that class's size once to the data set). We then compute the mean of that data set. For our example from above, there are 36 students contributing 36 to the data set, 124 students contributing 124 to the data set, 15 students contributing 15 to the data set, and so on, so that the data set is

$$\{\underbrace{36, 36, \ldots, 36}_{36\ \textit{times}}, \underbrace{124, 124, \ldots, 124}_{124\ \textit{times}}, \ldots, \underbrace{150, 150, \ldots, 150}_{150\ \textit{times}}\}$$

The mean class size experienced by students is therefore:

$$\textit{Mean} = \frac{36 \cdot 36 + 124 \cdot 124 + 15 \cdot 15 + 7 \cdot 7}{36 + 124 + 15 + 7 + 150} = \frac{39,446}{332} \approx 118.8.$$

As you can see, these two methods produce wildly different results. The notion that "the average student is in a larger-than-average class" is referred to as the class size paradox[a]. *Most of the time in their literature, universities do not specify how they compute the mean class sizes they report, which a statistician would consider to be bad scientific practice.*

[a] A scientific paper was published on this; see Canvas for the link.

The *median* of a set of data is another measure of central tendency that is important to know, because it helps avoid the possibility of being misled by very small or very large values in the data set.

When you list the data in increasing order, the median represents a "center" or "middle" value, where the number of data values less than the median is the same as the number of data values greater than the median. To make this definition mathematically rigorous, we need to consider sample sizes that are both even and odd. When the sample size is odd, the median is exactly the middle value. When the sample size is even, the median is the mean of the two middle values.

Definition 3 (*Median*)

If the set $\{x_1, x_2, \ldots, x_n\}$ has values listed in increasing order (that is, $x_1 \le x_2 \le \cdots \le x_n$), then the median *of the set is $x_{((n+1)/2)}$ if n is odd, and is the mean of $\{x_{(n/2)}, x_{(n/2+1)}\}$ if n is even.*

Example 3 (*Computing the Median (odd)*)

In a recent month, our state's Department of Conservation of Resources reported 53, 31, 67, 53, and 36 hunting and fishing violations in five different counties. The median number of violations is found first by listing the data in increasing order:

31, 36, 53, 53, 67.

Because the sample size is odd, the median is the middle value: for $n = 5$, the index of the middle is $\frac{5+1}{2} = 3$, and so the median is the third value in our list. That is, the median is 53.

Example 4 (*Computing the Median (even)*)

Repeat the previous example, including a sixth county where there were 42 violations. We again list the data in increasing order:

31, 36, 42, 53, 53, 67.

Because our sample size is now even, the median is the mean of the middle two items. For $n = 6$, the middle two items are $\frac{6}{2} = 3$ and $\frac{6}{2} + 1 = 4$, and so the median is the mean of the third and fourth items in the list. That is, the median is

$$\frac{42 + 53}{2} = \frac{95}{2} = 47.5.$$

Group Exercise 4

On the third hole of a certain golf course, eight golfers scored 4, 3, 4, 5, 4, 3, 4, and 3. Find the median.

List in inc. order: 3, 3, 3, 4, 4, 4, 5.

Sample size is $n = 8$, so median is the mean of the 4^{th} & 5^{th} values in the list – i.e., $\frac{4+4}{2} = \frac{2(4)}{2} = 4$.

Group Exercise 5

If a data set contains $n = 45$ values, and if these values are listed in increasing order, find the position of the median value.

Pos'n of median for odd n is $\frac{n+1}{2} = \frac{45+1}{2} = \frac{46}{2} = 23$

So the 23^{rd} point in the ordered list is the median

Group Exercise 6

If a data set contains $n = 28$ values, and if these values are listed in increasing order, the median is the mean of the values in which two positions?

For even n, median is the mean of the $\left(\frac{n}{2}\right)^{th}$ and $\left(\frac{n}{2}+1\right)^{th}$ points in the list. For $n = 28$, that is the 14^{th} & 15^{th}.

Yet another measure of central tendency is the *mode*. This is the value in a data set that occurs with the highest frequency (and more than once). One of its main advantages is that it requires no calculation or sorting (only counting), making it less computationally expensive for extremely large data sets.

Example 5 (*Mode*)

20 of the talks at last year's Undergraduate Research Conference were attended by 26, 25, 28, 23, 25, 24, 24, 21, 23, 26, 28, 26, 24, 32, 25, 27, 24, 23, 24, and 22 of its members. Find the mode.

Among the twenty numbers, 21, 22, 27, and 32 each occurs once; 28 occurs twice; 23, 25, and 26 each occurs three times; and 24 occurs five times. Thus, 24 is the modal attendance.

Group Exercise 7

Does there always exist a mode?[a] *[Hint: Is there a mode for the set* $\{0, 1, 9, 3, 6\}$*?] What about means and medians—do those always exist?*

[a] "Existence" [of solutions/winners/objects/etc.] is a very important notion in mathematics!

NO — some data sets (like $\{0, 1, 9, 3, 6\}$) have no mode, because they have no repeated values. Means and medians do always exist.

For the previous example, the set $\{0, 1, 9, 3, 6\}$ is a *counterexample* to the assertion that a mode always exists—that is, that particular example proves the assertion wrong. If someone had told me that all horses were brown, I could show a white horse as a counterexample to prove that person wrong.

Group Exercise 8

Can there ever be more than one mode? In other words, is a mode **unique**?[a] *Provide an example.*

[a] "Uniqueness" is also a very important notion in mathematics!

Yes, there can be more than one mode.

Consider $\{1, 1, 2, 2\}$. Two modes: 1 and 2.

If there are two modes for a set of data, the set is called *bimodal*; if there are three modes, it is called *trimodal*; the Latin prefixes continue *ad nauseam*, and at some point, if there are many modes, the data set is called *multimodal*.

The final measure of central tendency to discuss is the *midrange*. It is exceedingly simple to compute.

Definition 4 (*Midrange*)

For a set S of data, the midrange *is computed as*

$$\text{Midrange} = \frac{\max(S) + \min(S)}{2}.$$

Note that we give our general set the name S in this definition. We would write this as $S := \{x_1, x_2, \ldots, x_n\}$.

Example 6

With the URC attendance example discussed above, the midrange is $\frac{21+32}{2} = \frac{53}{2} = 26.5$.

Group Exercise 9

Tell your group members your age in years (lie if you like). As a group, find the mean, median, mode, and midrange of this data.

For example, say my groupmates have ages in the set $\{19, 17, 18, 18, 18, 23\}$.

Mean is $\frac{19+17+18+18+18+23}{6} = \frac{113}{6} \approx 19$ years.

(Notice I rounded the decimal — significant digits!)

Median is the avg. of the $(\frac{6}{2}+1)^{\text{st}}$ & $(\frac{6}{2})^{\text{th}}$ positions,

i.e., the avg. of the 3rd & 4th positions in the ordered list. Ord. list is 17, 18, 18, 18, 19, 23,

So avg. of 3rd & 4th vals is $\frac{18+18}{2} = \frac{2(18)}{2} = 18$ years

Mode is 18 (the only repeated value).

Midrange is $\frac{\min(S)+\max(S)}{2} = \frac{17+23}{2} = \frac{40}{2} = 20$ years.

Units (years) important!

Recap

We learned about the following **measures of central tendency**:

- Mean
  - Not a *resistant* measure of central tendency
  - Tells us about the maximum value, in case of non-negative data
- Median
  - Computed differently for even and odd sample sizes
  - As many values in the set are below the median as are above it
- Mode
  - Not always guaranteed to exist
  - When a mode does exist, it is not guaranteed to be unique
- Midrange
  - Easy to compute
  - Not resistant to extreme values

We also learned the mathematical notions of:

- **Existence** [of means, medians, modes, etc.]
- **Uniqueness**
- **Counterexample** to disprove an assertion

We will see these notions again and again!

Homework

- Please find the syllabus module on Canvas, and complete the quiz there. This quiz will not contribute to your grade, but it is required before you complete Homework 1.
- Homework 1 has been posted to Canvas, and will be due September 12.