An Introduction to Statistics for Librarians (Part Two): Frequency Distributions and Measures of Central Tendency

Caitlin J. Bakker, MLIS, AHIP[a]

[a]Discovery Technologies Librarian, Dr. John Archer Library & Archives, University of Regina, Regina, Saskatchewan, Canada, https://www.orcid.org/0000-0003-4154-8382, Caitlin.Bakker@uregina.ca

In Part One of this column, the different types of data were discussed.[1] Understanding the type of data is essential to interpreting them. If the type of data isn't correctly identified, it's not possible to answer some fundamental questions accurately. One of these fundamental questions is "what's the average value?" This is often the building block for more advanced statistical tests. In statistical terms, this question is asking us for the central tendency of the data. The central tendency is a single value that represents the midpoint of the data set. It tells us what is "average" or "normal" in the data set. There are three different ways to measure central tendency: mode, median, and mean. The measure chosen will depend on the type of data and the distribution of that data.

Frequency Distributions

Frequency distributions describe how often values occur in a dataset. These may be presented in tables, but are also often visualized as histograms. A distribution can be normal, or it can be skewed, either positively or negatively. A normal distribution is called a bell curve because it is symmetrical. Although there may be outliers, meaning extreme positive or negative values, these are relatively comparable and the majority of data clusters evenly around a center point. When a distribution is skewed, it means there are some particularly high or low values that make the distribution asymmetrical.

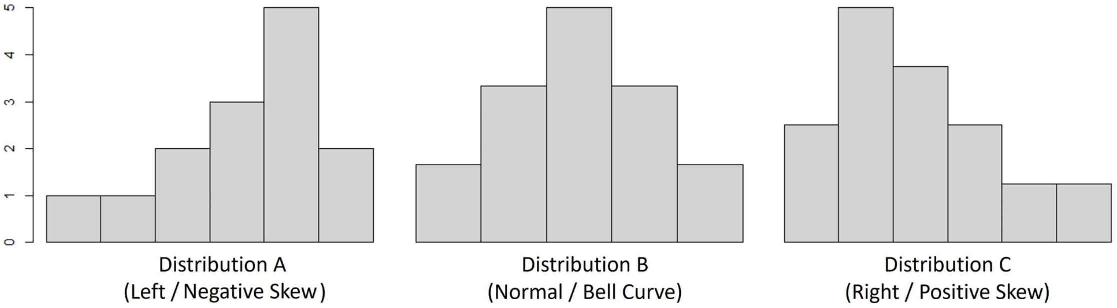

Figure 1: Skewed and normal distributions

In Figure 1, Distribution B is a bell curve or normal distribution, whereas Distribution A is left or negatively skewed, and Distribution C is right or positively skewed. Left skewness or negative skewness refers to the fact that some extreme low outliers, or negative values, are present and those negative values are what is causing the asymmetry. The inverse is true of right or positive skewness, where some particularly high outliers are causing asymmetry. Distribution matters because the measure of central tendency–that is, the way the midpoint is calculated–depends upon the distribution of the data. If we use the wrong measure of central tendency, the result will show the midpoint of the data to be either higher or lower than what is accurate.

Measures of Central Tendency

Table 1 describes the three measures of central tendency.

Table 1. Measures of central tendency and data types

	Mode	Median	Mean
	The most frequently occurring value	The middle value	The average of all values
Nominal	X		
Ordinal	X	X	
Interval	X	X	X
Ratio	X	X	X

Mode

The mode refers to the most frequently occurring value in the dataset. The benefit of the mode is that we can use this with any type of data, including nominal data. For example, if we are conducting a review of a library's chat reference service, the chat service may begin by prompting users to identify their patron type (e.g., student, staff, faculty). This would be nominal data. The mode here would be whichever value occurs most frequently. If the chat service is used by 40 students, 30 staff, and 20 faculty, the mode would be "student." If we review a library's chat reference service and look at how long each chat interaction was during one day, we would be looking at ratio data. We might have the following data: 1.5 minutes, 3 minutes, 3 minutes, 5 minutes, 7.5 minutes, 9 minutes, 10 minutes. The mode would be 3 minutes because this is the value that occurs most frequently. However, the most frequent value may not be truly representative of the data and its midpoint, as is the case in these examples. In the first case, while "student" may be the most common patron type, it still represents less than 45% of all patrons, so concluding that the "average patron" is a student overlooks the majority of chat reference patrons. In the case of chat reference time, an "average time" of 3 minutes is quite low when we look at it in the context of all of the data, given that only one of seven chat transactions was less than three minutes.

Median

The median is the middle value when all of the values in the data set are arranged in ascending or descending order. Using the above example of chat reference times, the median length of a chat reference interaction was 5 minutes, because 5 is the fourth of seven values. You can think of the median as the point at which 50% of the data will be above that point, and 50% of the data will fall below that point, as shown in Table 2.

Table 2. Example calculation of the median

1 minute	3 minutes	3 minutes	**5 minutes**	7.5 minutes	9 minutes	10 minutes
50% of values below the median			**Median value**	50% of values above the median		

The median can be calculated for most types of data. However, because the median requires that data be ranked or placed in order, we cannot use this as a measure of nominal data because nominal data cannot be ranked. The median is less impacted by outliers. This makes it a helpful measure when the data are not normally distributed, meaning that the distribution is skewed either positively or negatively rather than being symmetrical.

Mean

The mean is what people often think of when describing the average of a set of numbers. A mean involves adding all of the values in a dataset together, and then dividing it by the number of values. Using the chat reference example, we would calculate a total of 39 minutes spent conducting chat reference, and divide that by the seven chat reference interactions to have a mean value of 5.6 minutes, as seen in Figure 2.

$$\frac{(1.5 + 3 + 3 + 5 + 7.5 + 9 + 10)}{7} = \frac{\text{39 minutes}}{\text{7 interactions}} = \text{5.6 minutes per interaction}$$

Figure 2: Example calculation of the mean

To have an accurate mean, the data should be normally distributed. If the data skew right, meaning that there are some larger values as outliers, the mean will be higher than the median and will overestimate the midpoint. Where the data skew left, meaning that there are some smaller values as outliers, the mean will be lower than the median and will underestimate the midpoint.

Alongside the mean, researchers will generally also present the standard deviation. The standard deviation is a measure of variance. The smaller the standard deviation, the more the data tends to be concentrated around the mean, or has less variation. A larger standard deviation means that the data are more spread out from the mean, or it has more variation. To learn more about standard deviations and their meaning, consult Illowsky and Dean.[2]

Why Does This Matter?

Imagine the data in Figure 3 represent responses from patrons at two different libraries. The patrons are asked to rate their satisfaction with library services on a scale from 1 to 6, where 1 means very unsatisfied and 6 means very satisfied.

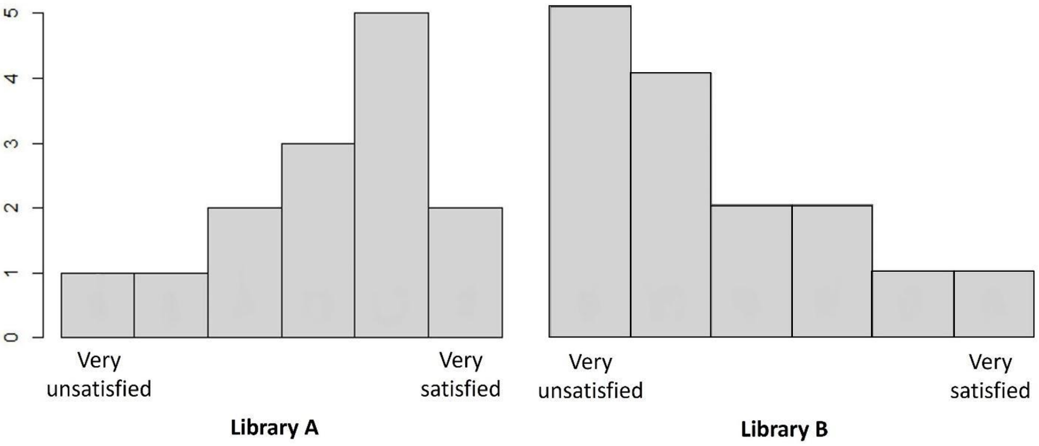

Figure 3: Patron satisfaction at two libraries

To compare these two groups, we would compare a measure of central tendency. However, the measure we select would give us quite different results. In Table 3, we see the various midpoints calculated based on the data in the visualization above.

Table 3. Patron satisfaction at two libraries

	Library A	Library B
Mean	4.1	2.5
Median	4.5	2.0
Mode	5	1

For Library A, we have a mean of 4.1, a median of 4.5, and a mode of 5. When we try to express the "average" level of patron satisfaction with library services, if we define the "average" using the mean, we would say patron satisfaction was 4.1/6. However if we defined "average" as the median, we see that patron satisfaction is 4.5/6. Because the mean doesn't account for the fact that the data are left (negatively) skewed, using the mean would lead us to conclude that patrons are less satisfied with library services than may actually be the case.

For Library B, we have a mean of 2.5, a median of 2.0, and a mode of 1. If we were to describe patron satisfaction with library services using the mean, we would conclude that the average rating was 2.5/6. If we were to use the median, we would conclude that the average rating was 2.0. As opposed to Library A, the data are right (positively) skewed in this distribution, which means that using the mean would lead us to conclude that patrons are more satisfied with library services than they may actually be.

These decisions about what is "average" matter both for individual groups, but also when we compare different groups. If we were to compare Library A and Library B using the means, we would conclude that the difference between the two groups was smaller than if we were to compare them using the medians. Using the mode would make the difference look quite dramatic. The mode, the median, and the mean all represent the "average" value, but those values can be very different.

These differences become more important as more advanced statistical tests, such as t-tests or ANOVAs, are involved, because these tests compare different midpoints between and within groups and have certain assumptions about how the data are distributed. Choosing a test that compares means when you have a skewed distribution might lead to inaccurate results and false conclusions, which can subsequently lead to poor decisions. In the next column in this series, more advanced statistical tests will be discussed, building on some of this foundational knowledge.

Further Reading

Shafer D, Zhang Z. Measures of central location: three kinds of averages. In: Introductory statistics [Internet]. Saylor Foundation; 2012 [cited 2022 Dec 17]. Available from: https://stats.libretexts.org/Bookshelves/Introductory_Statistics/Book%3A_Introductory_Statistics_(Shafer_and_Zhang)

References

1. Bakker CJ. An introduction to statistics for librarians (part one): types of data. Hypothesis Res J Health Inf Prof [Internet]. 2022 Aug 24 [cited 2022 Dec 17];34(1). Available from: https://journals.iupui.edu/index.php/hypothesis/article/view/26428
2. Illowsky B, Dean S. Measures of the spread of the data. In: Introductory statistics [Internet]. Houston, TX: OpenStax; 2013 [cited 2022 Dec 17]. Available from: https://openstax.org/details/books/introductory-statistics