To be effective as infection preventionists, we must understand how to collect data, analyze them, and present them in ways that best support decisions about interventions, resource allocation, patient safety, and more. Statistics are how we turn that mere data into useful information. Understanding those processes also enhances our statistical literacy so we can make sense of the work being done by others, whether that's research articles, conference posters, or projects in our own facilities. This series is designed to help you develop that understanding by digging into statistical concepts and linking them to infection prevention.

# Making use of frequency tables and distributions

BY DANIEL BRONSON-LOWE, PHD, CIC AND
CHRISTINA BRONSON-LOWE, MS, CCC-SLP, CLD

**AN INFECTION PREVENTIONIST (IP)** is asked by her director if they are doing a good enough job with healthcare worker (HCW) influenza vaccination compared to other facilities in the region or if they need to put more resources into the vaccination campaign next flu season. The IP discovers that the mean vaccination rate in her state last winter was 86 percent. The rate at her facility was 95 percent, so it looks like they are doing comparatively well, but she is also aware that means can be misleading at times *(see the Fall 2016 From Data to Decisions article for a discussion of mean, median, and mode).* Curious about how her facility really compares to the other facilities in her state, the IP downloads the full HCW influenza vaccination data set and begins examining it.[1]

The data set is a table with 150 rows, one row for each of the facilities included. Such a table makes it difficult to understand patterns in the data. One solution is to generate summary measures such as measures of central tendency: mean, median, and mode. The IP discovers that

the median is 89.5 percent, somewhat higher than the mean. This reveals that more than 50 percent of the values are above the mean and lends credence to her concern that the mean is not an accurate representation of how most facilities perform. To get a better idea of how the data are actually distributed, the IP generates a ***frequency table***. This table lists all possible values of a measurement and how often each occurs in the data.

It is possible to create a frequency table for either quantitative or categorical data (see Figures 1 and 2). The tables in those figures are relatively simple, but the frequency table for a variable with a wide variety of responses can end up almost as unwieldy as the original data set. Graphing the data can make patterns clearer and can suggest next steps in analysis. Graphs of frequency data are called ***frequency distributions***. The most common type used with quantitative variables is the ***histogram***, a graph with the possible measurement values along the horizontal axis (the x-axis) and the number of times each value occurs on the vertical axis (the y-axis).

As an example, HCW influenza vaccination rates can range from 0-100 percent, resulting in a frequency table with 100 rows. That is too large to interpret easily—even though it is a step better than our original data set with its 150 rows—but when converted to a histogram the data become clearer (see Figure 3). It is then easier to identify the ***range*** of the data as well as any ***outliers*** (for glossary, see page 34). In this case, the range is from 22 percent to 100 percent. The facilities with 22-percent and 30-percent vaccination could be considered outliers. The presence of these outliers explains why the mean is lower than the median. The mode can also be identified by looking for the highest bar in the graph. In Figure 3, there are three equally high bars, indicating three modes, all well above the mean. This is more evidence that the mean does not represent this data set well.

Another characteristic that can be assessed using a histogram is the symmetry of the data with respect to the mean. Figure 4

**Figure 1: Quantitative Data and Frequency Table**
Quantitative data set = 1, 3, 1, 3

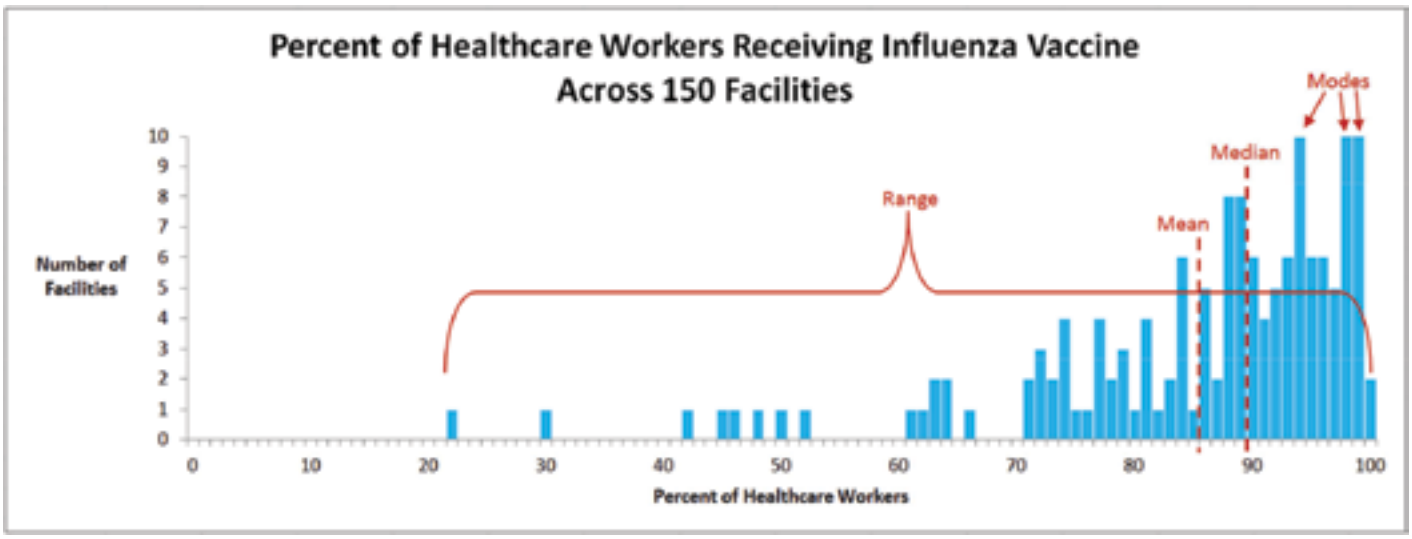
| Possible Values of the Variable | Occurrence of Frequency |
|---|---|
| 1 | 2 |
| 2 | 0 |
| 3 | 2 |

**Figure 2: Categorical Data and Frequency Table**
Categorical data set = *E. coli, E. coli,* MRSA, *E. coli*

| Possible Values of the Variable | Frequency of Occurrence |
|---|---|
| *E. coli* | 3 |
| MRSA | 1 |
| *K. pneumoniae* | 0 |

"Graphs of frequency data are called *frequency distributions*. The most common type used with quantitative variables is the *histogram*, a graph with the possible measurement values along the horizontal axis (the x-axis) and the number of times each value occurs on the vertical axis (the y-axis)."

**Figure 3: Histogram with 100 Bins**



## GLOSSARY

- **Frequency table:** A table listing the frequency (number) or relative frequency (fraction or percentage) of observations.
- **Frequency distribution:** Graphs of frequency tables.
- **Histogram:** A graph with the possible measurement values along the horizontal axis (the x-axis) and the number of times each value occurs on the vertical axis (the y-axis).
- **Outlier:** a value that is distinctly separate from, or "lies outside" of, the other values in a data set.
- **Range:** the difference between the largest and smallest values in a data set. In statistics, this is reported out as the difference (e.g., a range of 88). When speaking epidemiologically, this is often reported in terms of the minimum and maximum (e.g., a range from 22 to 100).
- **Bins:** In plotting a histogram, one starts by dividing the range of values into a set of non-overlapping intervals, called bins or class intervals, in such a way that every data point is contained in these value groupings.
- **Skewed:** An asymmetric distribution of the data. Data sets can be either positively or negatively skewed.

contains three examples of frequency distributions made using symmetric data sets. In each instance, the distributions of the frequencies on both sides of the mean are equal.[2] The vaccination data shown in Figure 3, on the other hand, are decidedly not symmetric. An asymmetric pattern like this is referred to as a ***skewed*** distribution. The skew can be positive or negative; it depends on which direction the extreme asymmetric values fall with respect to the mean. If the extreme values create a tail that extends out to the left, as is seen in Figure 3, this is called being skewed to the left or a ***negative skew***. If the extreme values create a tail that extends out to the right, this is called being skewed to the right or a ***positive skew***.

As useful as Figure 3 is, there may be times when a histogram with a bar for each possible value is too busy or complex. To simplify it, let each bar represent a range of values rather than a single value. These groupings of values are referred to as ***bins***. Figure 5 contains the same vaccination rate data as Figure 3 but displays it using only 10 bins instead of 100, making it more compact. This simplification comes at the cost of losing the resolution necessary to spot the mode(s) or the exact minimum or maximum, but it may provide a better understanding of the overall pattern. The trick is to avoid consolidating the data to the point that the histogram no longer provides useful information (see Figure 6).

When it comes to deciding just how many bins to have in your histogram, there is no best answer. Here are some guidelines to consider:

- Make the bins the same size. If one bin contains values 1-5, the next bin should not contain values 6-20.
- Select bin sizes that are easier to understand. For example, when working with percentages, bins that cover five or 10 percentage points (e.g., 11-20 percent, 21-30 percent) will be easier to understand than bins with unusual ranges (e.g., 7-13 percent, 42-48 percent).
- Consider bins that make use of goal values or clinically important thresholds. This ties in with the idea that your graph should be designed to emphasize the message or pattern you want to relay to the audience. If the real interest is in how the data are distributed with respect to the value of 88

## CONCEPT QUIZ

You are an infection preventionist at a healthcare facility with 20 nursing units. Hand hygiene compliance has been an ongoing problem requiring extra attention. You report to your superiors that last month the median hand hygiene compliance percentage for the units was at a record high: 86 percent. In light of this news, your superiors suggest it is time to move resources away from improving hand hygiene compliance and to focus on other issues. You decide to look more closely at the data before agreeing to that change.

Last month's hand hygiene compliance percentages for the 20 units were:
75, 99, 34, 36, 99, 86, 27, 96, 100, 89, 39, 44, 97, 99, 99, 81, 86, 38, 48, 88.

**Activity 1**
Create a frequency table that combines the hand hygiene compliance percentages into 10 groups: 0-10 percent, 11-20 percent, 21-30 percent, and so on.

**Activity 2**
Create a histogram based on the data in your frequency table.

**Activity 3**
What information from the frequency distribution might you use to emphasize the importance of continuing to support hand hygiene compliance improvement?

*Answers on page 37.*

**Figure 4: Histograms of Symmetric Frequency Distributions**

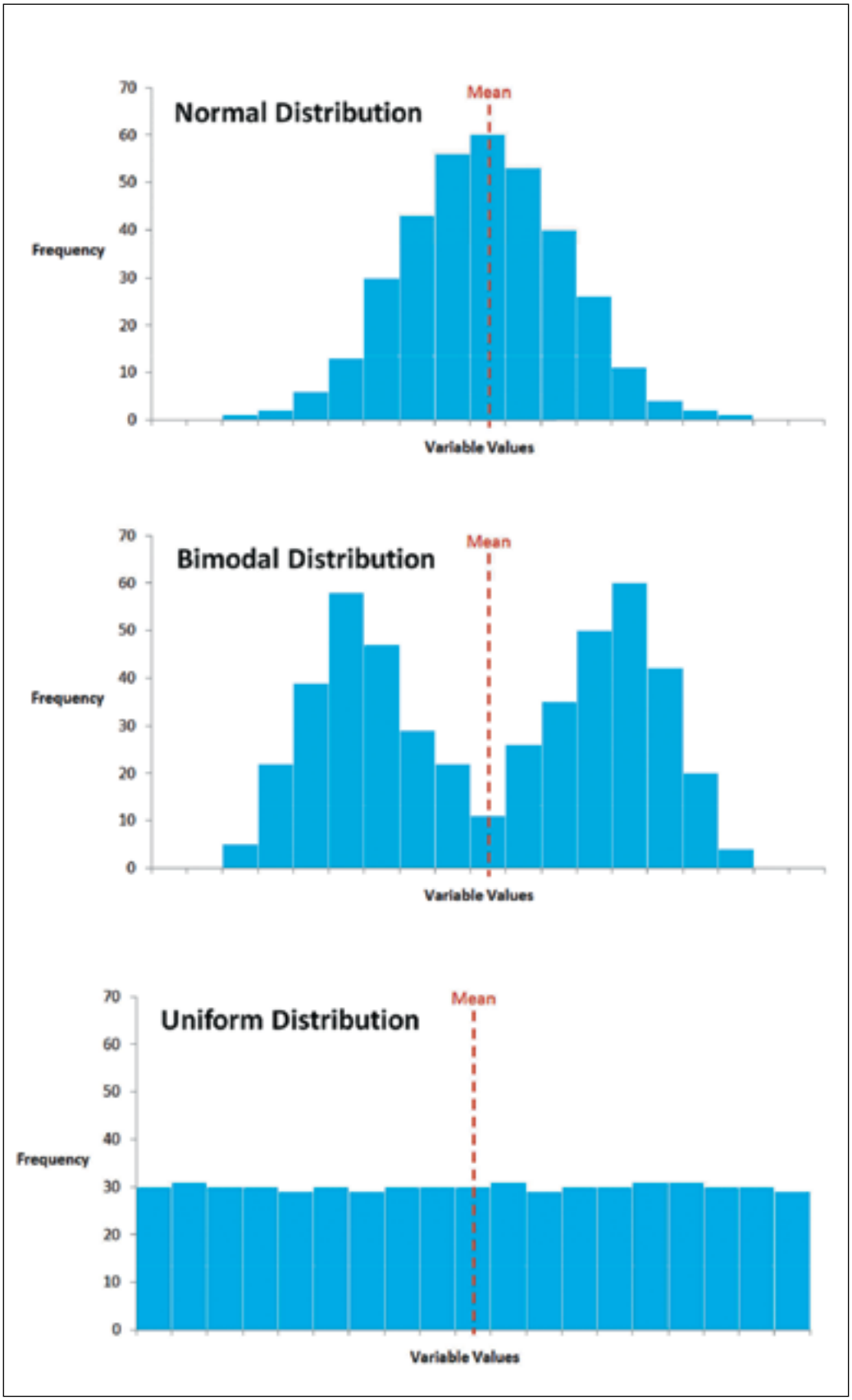


**Figure 5: Histogram with 10 Bins**

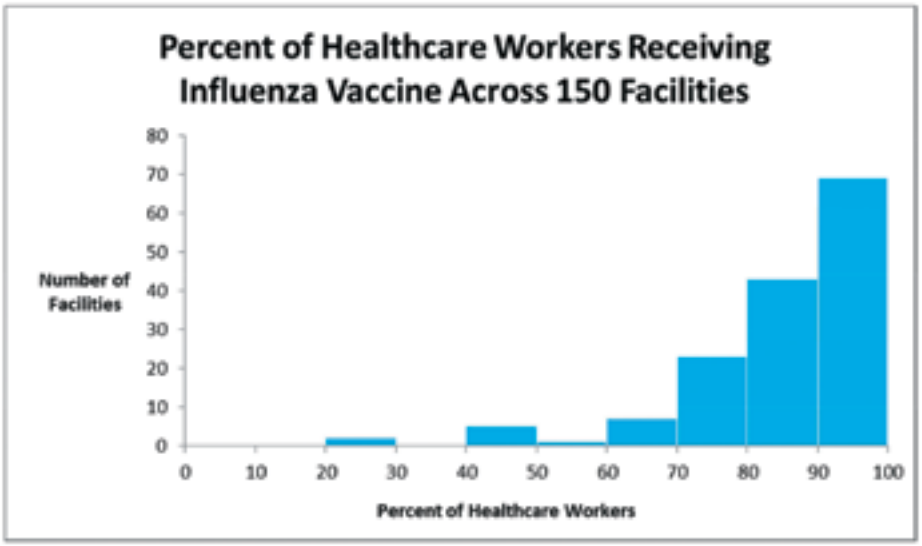
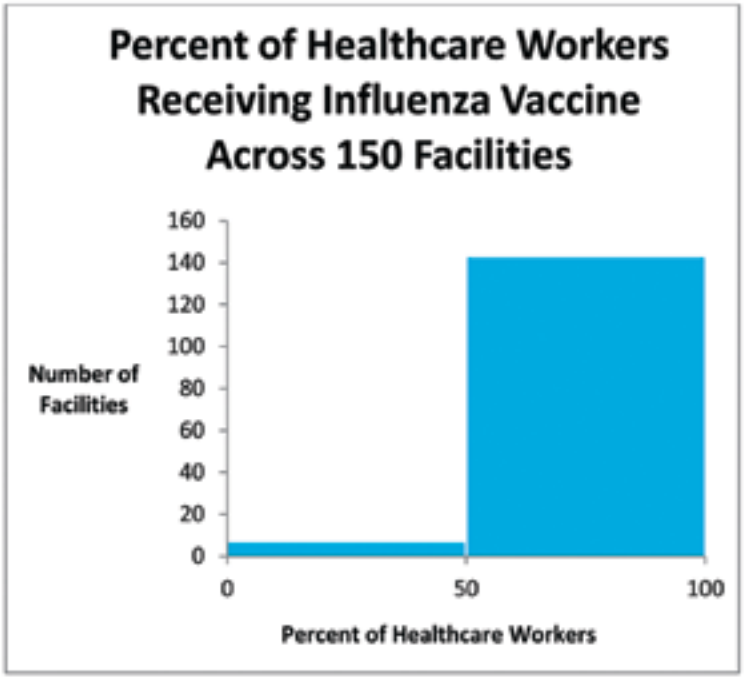



**Figure 6: Histogram with 2 Bins**




"When in doubt, aim for somewhere between five to 15 groups. Different numbers of bins can reveal different aspects of the data. Don't be afraid to experiment."

percent, ensure that 88 percent is at one end of the values contained within the applicable bin (e.g., 80-88 percent).

- When in doubt, aim for somewhere between five to 15 groups. Different numbers of bins can reveal different aspects of the data. Don't be afraid to experiment.

## NEXT STEPS

Armed with her examination of the frequency distribution thus far, the IP in our scenario has verified that the mean is not the best option for comparison given the skewness of the vaccination data. Since her facility has a better vaccination rate than the state median, she knows they are in the top 50 percent reporting facilities, but she lacks a more specific figure. To obtain that information, she needs to make use of quantiles, the topic of our next article about distributions.

If you have any questions or comments, you can contact us at IPandEpi@gmail.com. PS

*Daniel Bronson-Lowe, PhD, CIC, has been an infection preventionist, an infectious disease epidemiologist, and a statistics lecturer. He is now the instructor for APIC's "Basic Statistics for Infection Preventionists" Virtual Learning Lab and a senior clinical manager with Baxter Healthcare Corporation.*

*Christina Bronson-Lowe, MS, CCC-SLP, CLD, is a speech-language pathologist and PhD candidate who has worked in hospitals, inpatient and outpatient rehabilitation, skilled nursing facilities, and home health care.*

### Notes

1. The HCW influenza vaccination data used in this article were pulled from the Hospital Compare data sets made available by the Centers for Medicare & Medicaid Services (https://data.medicare.gov/). We examined the 2014-2015 data for the 150 reporting Illinois facilities.
2. Or at least as close to equal as you will ever see in real life. Actual data sets never produce frequency distributions that are perfectly symmetrical.

### Additional Resources

Potts, A. Chapter 13: Use of Statistics in Infection Prevention. In: Patti Grota, et al., editors. APIC Text Online. APIC; 2014.

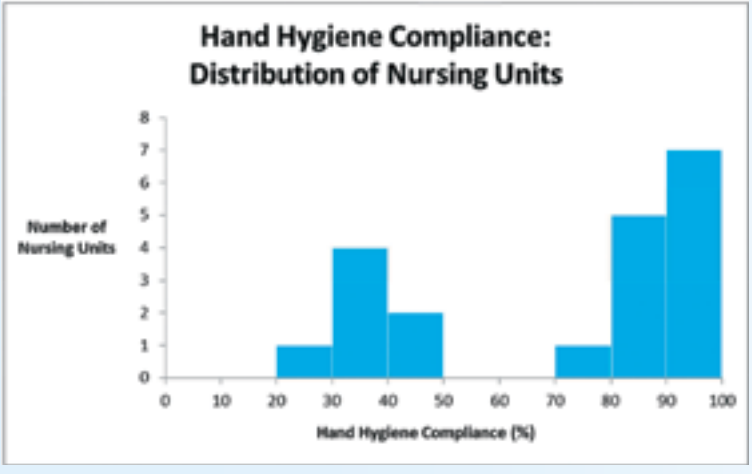
Centers for Disease Control and Prevention. Principles of Epidemiology in Public Health Practice: An Introduction to Applied Epidemiology and Biostatistics, 3rd Edition. 2012. (http://www.cdc.gov/ophss/csels/dsepd/SS1978/SS1978.pdf)

## ANSWERS

**1**

| Hand Hygiene Compliance | Number of Nursing Units |
|---|---|
| 0-10% | 0 |
| 11-20% | 0 |
| 21-30% | 1 |
| 31-40% | 4 |
| 41-50% | 2 |
| 51-60% | 0 |
| 61-70% | 0 |
| 71-80% | 1 |
| 81-90% | 5 |
| 91-100% | 7 |

**2**




**3**

Enough units are doing well that the median looks good. However, there continues to be a cluster of units with less than 50-percent compliance. These units still need extra attention.