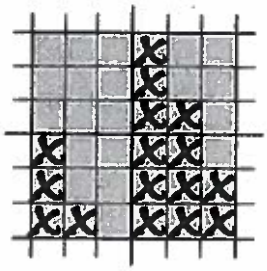


Data Terms: A Glossary

This section provides information about the statistical terms and data representations that come up in this seminar. The information is intended primarily as background for you, although you may find it useful to share parts of it with participants from time to time.

Types of data

Throughout this seminar, we refer to two types of data which we designate *numerical* and *categorical*.

Data that are *categorical* have values such as *yes* and *no*, *red* and *blue*, or *dog* and *cat*. These values cannot be compared or ordered as quantities. The order in which they are placed on a graph or table is arbitrary. While data such as these are sometimes designated by numbers—for example 0 for *men* and 1 for *women*—these numbers are only codes used to make recording easier. They cannot be added, subtracted, or averaged.

Data that are *numerical* have values that represent numerical quantities that can be ordered and compared mathematically. A respected statistics textbook by David Moore and George McCabe (1993) sums up the difference between what they term *quantitative* and *categorical* values, terms that correspond to our use of *numerical* and *categorical*. Their definition reads:

> A quantitative variable takes numerical values for which arithmetic operations such as differences and averages make sense. A categorical variable simply records into which of several categories a person or thing falls. (p. 2)

Quantitative data is a useful alternative term for *numerical data*; you may want to introduce it to your participants.

Within quantitative data, further distinctions can be made. Not all numerical values behave the same way, and your participants may come across instances of data that they are not sure how to classify. For example, consider two data sets that appear to be numerical: family size and reading level. Values for family size can be ordered and compared in all the ways we might expect for numerical data: a family that has 6 members is twice as big as a family that has 3 members; a family with 4 members has one less person than a family with 5 members. We can find an average family size in such a set of data.

Reading levels may not behave in the same way. They can be ordered and have differences that can be calculated. A student with a reading level of 6.2 as measured by some reading scale is said to be reading 2.5 grade levels above a student with a reading level of 3.7. However, it may not be reasonable to say that a reading

level of 6.2 is twice as high as a reading level of 3.1. It could be argued that we can't quantify "reading" in this way; it does not make sense to say "she is twice as good a reader as he is" if we mean that there is some quantity that is twice as great as another.

There are traditional classifications of numerical data that make finer distinctions among types of numerical data and provide rules about which statistical measures can be used in each case. You may have encountered some of these terms for data types—for example, *ordinal*, *interval*, or *ratio* data. However, this view of data types has changed in the past 30 years. Critics of the traditional classification of data types point out that while making these distinctions still has value, the classification has been applied too rigidly. What is most important is that we choose tools for analysis that make sense for the question being asked, not according to an *a priori* classification of the data.

As an example, let's do a thought experiment about grade level. Suppose there is a school in which there is one classroom of every grade from grade 1 through grade 6. Every Monday, a number from 1 to 6 is pulled out of a hat to see which grade will be in charge of making the morning announcements for that week. A graph is kept showing each grade level and how many times each has been chosen. In this case, grade level is a categorical value, a name for a particular group. We could just as well use the room number or the teacher's name.

At first, it doesn't seem that finding an average for these data would make any sense. But suppose some of the primary-grade students complain to their teachers that the fifth and sixth graders seem to get chosen a lot more than the other grades. We could simply count and compare the frequencies for each grade, just as we would do with any categorical data. Or we might instead order the data by grade and find the median grade level. If it turned out to be grade 5, we would know that fifth and sixth graders were in fact being chosen at least half the time, while the other four grade levels were sharing the remaining weeks. Because our question has changed, we can treat the data values differently.

Yet another use of grade level might be in a study focused on high school dropouts in our town. We might collect information about the highest grade level achieved by town residents in their early twenties. Now we are using grade level values to indicate years of education and could reasonably calculate the mean grade achieved by people in the study.

In this seminar, when participants want to know if they have chosen data that are "numerical," they need to look not only at the type of data they are collecting, but also at their questions. Will their questions lead them to treat their data as quantities for which (as Moore and McCabe say) "arithmetic operations such as differences and averages make sense," or as categorical values that can only be counted and compared to see which data values occur with more or less frequency? As Velleman and Wilkinson's (1993) paper on this issue states, how we can operate on a data set "depends upon the questions we intend to ask of the data."

Value and frequency One of the issues that arises in the cases, especially in chapter 4, is that students need to sort out what the different numbers in their data mean. In Isabelle's case, "How Many People, How Many Teeth?" some of the second graders are figuring out how to represent the number of teeth lost by each person. Other stu-

dents are trying to represent both the number of teeth lost (the values of the data) and *how many people* lost each number of teeth (the frequency of each value's occurrence in the data set).

Every data set has both values and frequencies. First, each piece of data has a value. This value might be a quantity such as 6 or a categorical value such as *yes*. In Isabelle's case, the values are quantities—for example, 0 or 1 or 5—representing the number of teeth lost. In Beverly's case, "Do You Like to Eat Soup?" the values of the responses are *yes* and *no*. In both cases, there are also frequencies—how many pieces of data have a particular value. There might be five students who have lost 3 teeth: the value of 3 occurs with a frequency of 5. There may be 10 students who responded *yes* to the question *Do you like to eat soup?*: the value of *yes* occurs with a frequency of 10. A graph that shows the frequencies of an ordered set of numerical values is called a *frequency distribution*. In Denise's case 14, Cara's graph is a frequency distribution, while Kenny's graph is not. (This is further discussed under "Types of Representations," p. 115.)

Terms used to describe and summarize data

Average An average is a value that describes the center of a data set. In everyday usage, people usually assume that the word *average* refers to the arithmetic mean. However, *average* is also an inclusive term for any measure that is used to summarize the center of the data. Statistical references differ in the term they use to describe mode, median, and mean. They are all sometimes referred to as "averages," sometimes as "measures of center," sometimes as "measures of location." *Mean*, *median*, and *mode* are further described in separate entries below.

Measures of variability An average or center of a data set gives an incomplete picture of the data. We also need to know how the data are spread around the center, or how they vary. Two measures of variability are the standard deviation, used to show spread around the mean, and the interquartile range, used to show spread around the median. In this seminar, the interquartile range (the middle 50 percent of the data) is introduced informally when participants learn how to construct a box plot. Whether or not specific measures of variability are being used, it is critical that participants describe the shape of the data—how the data vary as well as where they are centered.

Mean The arithmetic mean is a commonly used statistical average. It can be thought of as a "balance point" or as an evening out of the data. It takes into account all of the values in the data set.

One way to think of the mean is as the value that would result if all the values in the data set were evened out. This evening out is actually what you are doing when you use the algorithm for finding the mean. Finding the sum of all the values, then dividing by the number of values, results in the average. You might picture this as having a set of sticks of varying lengths. If we cut pieces off the taller ones and added them to the shorter ones until all the sticks were the same length, that length would be the mean value. It is as if we taped all the sticks together, end to end (added their values), then chopped this into the same number of even lengths as the original number

of sticks (divided by the number of values).

Another way to think of the mean is like the fulcrum of a seesaw. If all the pieces of data were represented on a line plot, the mean is the point at which the number line would balance. In this model, the sum of the distances of all the pieces of data on one side of the mean from the mean value equals the sum of the distances of all the pieces of data on the other side of the mean from the mean value. Both the evening out model and the balancing model are explored by participants as they use cubes and a line plot with a small set of data (bags of peanuts) in Session 7.

Since the value of the mean is affected by the actual values of each piece of data, the mean is more sensitive than the median to unusually small or unusually large values in the data set. The mean is a less stable indicator of center than the median. However, in essentially symmetrical data sets, the median and mean will be close together.

Median The median is the value of the middle piece of data (or, in the case of an even number of data points, the midpoint between the two middle pieces of data). One way to think of the median is the value that would result if you listed all the pieces of data in order of their values and found the middle of the ordered list. The median cuts the data set in half. Half of the values in the data set are either equal to or greater than the median value, while half are either equal to or less than the median value.

The median takes into account all the pieces of data in the data set, but it is not subject to much change by unusually small or large values because the median depends only on the *order* of values, not on the actual values. The median is a stable indicator of the center of a data set and is often used for data, such as housing prices, for which a middle value (rather than an "evening out") is most useful.

Mode The mode of a data set is the value at which more data occur than at any other value. A data set might have one or several modes. Modes are not generally used in statistics for the analysis of numerical data as they do not take into account all the values in the data and so may not communicate anything useful about the data set as a whole. In many data sets, the mode does not indicate anything of importance about the shape of the data. Mode is often used to describe categorical data, where the most frequent value may have more meaning ("the most popular TV program is . . . ").

Outlier An outlier is a piece of data that is well removed from the rest of the data. There are a variety of statistical procedures for determining whether a piece of data is small enough or large enough to be considered an outlier. In this seminar, we don't worry about the formal statistical definitions of outliers, but develop the notion that an outlier is a value to which we need to attend. It may be an error in the data or an unusual value of interest that should be investigated further.

Range The range of the data is the difference between the minimum and maximum values in the data set. These minimum and maximum values are called the *extreme values*. If the minimum value in a data set is 3 and the maximum value is 10, the range is 7. The word *range* is often used more informally to indicate the extent of values from minimum to maximum, so we say, for example, "the data range from 3 to 10."

Types of representations

The field of graphic representation is a lively and inventive one. For a glimpse into the many and varied ways of representing data, you might want to take a look at *The Visual Display of Quantitative Information* by Edward Tufte (1983). We encounter only a small sample of basic types of graphs in this course.

Of the common representations used by students in the elementary grades for numerical data, we see two major types, *frequency distributions* and what we have termed *case value plots*. Case value plots—such as Kenny's in Denise's case 14—show a bar or a line of counters for each piece of data. That is, each case (piece of data) is shown separately. The length of the line or bar indicates the value of that piece of data. A taller bar shows a higher value. A frequency distribution shows the number of cases at a particular value. The length of a line or bar indicates the frequency with which that value occurred. A taller bar shows more data.

On entering this course, participants may be most familiar with representations that are commonly used in the media, including bar graphs, histograms, and line graphs.

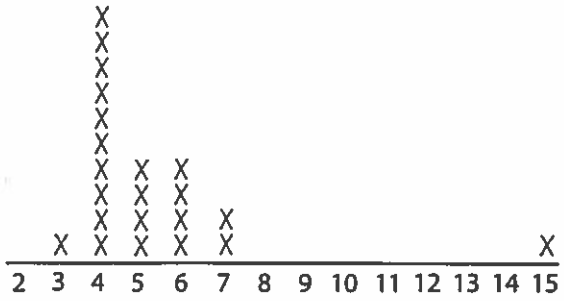
Bar graph A *bar graph* is usually defined as a graph in which each bar shows a single value for one category (10 *yes*, 15 *no*; or 40 percent *yes*, 60 percent *no*). This type of graph is frequently used in the media to compare a few values.

Histogram A *histogram* is often used to show numerical data in which there are many different values in the data set, making it impractical to show all the individual values. Each bar of the histogram shows a subset of the values. For example, the first bar might include values from 0 to 5 (equal to or greater than 0 and less than 5), the next bar from 5 to 10 (equal to or greater than 5 and less than 10), and so forth. The vertical scale indicates frequency.

Line graph A *line graph* (as distinct from a *line plot*) represents a type of data that we do not treat in this course—data that reflect a relationship between two variables, showing how one variable changes as the other changes. One common example is a graph of temperature over time. Line graphs are used to show continuous change—change in which a variable moves through all possible values as it increases or decreases (e.g., as the temperature changes from 20 to 40 degrees, it moves through all the intermediate values).

In this course we emphasize several of the plots developed by statistician John W. Tukey and elaborated by Tukey and his colleagues from the early 1970s until the present. These plots—the line plot, the stem-and-leaf plot, and the box plot—are used for what Tukey termed *exploratory data analysis*—looking at a set of data to see its shape and patterns, to consider what it might reveal, and to generate questions for further study.

Line plot A *line plot* is a frequency distribution, showing the values of the data along the horizontal axis and the frequencies along the vertical axis. Each piece of data is represented by an X. The line plot is a straightforward way to represent every piece of data in the set in a way that reveals the shape of the data—where data are concentrated, how the data are spread out, where there are gaps in the data, as in this graph of family size in Maura's case 24, "How Many People in Our Families?"



Stem-and-leaf plot A *stem-and-leaf plot*, sometimes called a *stem plot*, is constructed by ordering the data and arranging them in groups. Instead of X's or some other symbol, the actual values of the data are used.

A stem-and-leaf plot is particularly useful when the range of values is so large that it is not practical to use a scale that shows each value, as a line plot does. A stem plot is also useful when the data set is large; some guidelines suggest that up to about 250 values can be shown in a stem-and-leaf plot without becoming visually confusing.

In the classic stem-and-leaf display, the "stem" is the tens digit and the "leaves" are the ones digits, as in the plot of a group of teachers' commute times from home to school.

Teachers' commute times

0	5 5 9
1	1 5 6 6 7
2	0 0 0
3	4 5 5 5 5
4	5 5
5	0
6	
7	5

3 | 5 is 35 minutes

For these commute data, grouping by tens works well to show the shape of the data, but for some data sets, the tens might be further subdivided, giving intervals of 5 or 2, as in the example of student heights.

Student heights

5	0 0 2 4 4
•	5 6 6 7 7 7 7 8 8 8 8 9
6	0 0 2 4

5 | 2 is 52 inches

For data with much larger or much smaller values, the stem value might be hundredths, tenths, hundreds, thousands, or millions. In this case, the digit for the stem would be the next lowest place, and the other places would not be shown (the number is typically truncated). For example, following is a stem-and-leaf plot showing population of the 50 states, according to the 2000 U.S. census. The stem digit is millions and the leaves are hundred thousands. The rest of each number has been truncated. The key is used to indicate the values of the digits in the display.

Population of the 50 states (2000 census)

0	4 6 6 6 7 7 9
1	0 2 2 2 2 7 8 8 9
2	2 6 6 8 9
3	4 4 4
4	0 0 3 4 4 9
5	1 2 3 5 6 8
6	0 3
7	0
8	0 1 4
9	9
10	
11	3
12	2 4
13	
14	
15	9
16	
17	
18	9
19	
20	8
21	
22	
23	
24	
25	
26	
27	
28	
29	
30	
31	
32	
33	8

9 | 9 means 9,900,000
(numbers are truncated)

Back-to-back stem-and-leaf plots can be used to compare two sets of data, as for example to compare the heights of students in two classrooms.

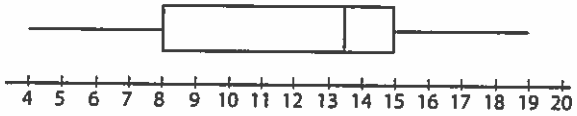
Box plot A *box plot* (or *box-and-whiskers*) does not show each individual value in the data set. Rather, it provides a summary of the center and spread of the data. The basis of the box plot is the five number summary, which consists of these values: the minimum value (or lower extreme), the lower or first quartile, the median, the upper or third quartile, and the maximum value (or upper extreme). These values divide the data into four groups, with close to the same amount of data in each group. To find these five numbers, put the data in order, as shown in the example below. Find the median, which in this case is 13.5.

4 5 5 6 8 9 12 12 13 | 14 15 15 15 15 16 18 18 19

Then find the median of all the values below the median, which is 8—this is the value of the first quartile. Do the same for the values above the median to find the value of the third quartile, which is 15. The five number summary for this sample of data, then, is as follows:

lower extreme: 4
lower quartile: 8
median: 13.5
upper quartile: 15
upper extreme: 19

The left and right edges of the "box" are the lower and upper quartiles. The vertical line within the box is the median. From the box, the "whiskers" extend to the lower and upper extremes:



The *interquartile range*—a measure of spread around the median—is the distance between the lower and upper quartile values, which here is 7. So, the middle 50 per-

cent of the data are spread within a range of 7 around the median 13.5. The interquartile range is also used to calculate outliers. One way to calculate outliers is as follows: multiply the interquartile range by 1.5. If a data value is more than this amount above the upper quartile or below the lower quartile, it is considered an outlier. Outliers in a box plot can be shown as separate points, rather than as part of the "whisker."

References

Moore, D. S., & McCabe, G. P. (1993). *Introduction to the practice of statistics.* Second edition. New York: W. H. Freeman.

Tufte, E. R. (1983). *The visual display of quantitative information.* Cheshire, CT: Graphics Press.

Velleman, P. F., & Wilkinson, L. (1993). Nominal, ordinal, interval and ratio typologies are misleading. *American Statistician 47,* 65–72.