Chapter 10

Describing A Numerical Response

10.1 Pictures

So far, the response has been either a dichotomy or a count that follows the Poisson or Binomial distribution. In this chapter we extend our work to counts that are neither Poisson nor Binomial and to responses that are measurements.

Suppose the subjects are students in this class. Below are some examples of numerical responses.

- **Counting:** Number of zeroes on homework to date; number of credits this semester; number of persons living in current household.

- **Measuring:** Height; weight; age.

As often happens in life, the boundary between these options can be blurry. For example, consider annual income. Literally, annual income is determined by **counting** the number of cents earned in the year, but economists and other researchers tend to treat it as a measurement. The general guideline is that if a count variable has many many values in a population, and no one value dominates others in terms of relative frequency, it is usually mathematically more convenient to treat the variable as a measurement.

Two important words are: **precise** and **accurate**. Accurate means close to the truth. For example, if I state that my dog Casey lived for 15.5 years, that is accurate. If I state that my grandfather Wardrop lived to be 150, that is highly inaccurate.

Precise is most useful for measurements. If I state: Yesterday I ran one mile in 250.376 seconds, this is incredibly precise (to the nearest one-thousandths of a second), but ridiculously inaccurate. If I say I ran it 'In less than one hour' it is accurate, but not the least precise.

Here is a good general guideline for science: measurements should be precise enough to create variation in our population or subjects of interest, but there is no need to get carried away with it!

Precise is somewhat meaningless for counts that take on small values. For example, it is accurate to say that 2 cats live in my house. It is no more precise to say I have 2.000 cats!

Table 10.1: Sorted Speeds, in MPH, by Time, of 100 Cars.

Speeds at 6:00 pm

26	26	27	27	27	27	28	28	28	28	28	28	28	28	28	28	28	28
28	29	29	29	29	29	29	29	29	29	29	29	30	30	30	30	30	30
30	30	31	31	31	31	32	33	33	33	34	34	35	43				

Speeds at 11:00 pm

27	28	30	30	30	31	31	31	32	32	32	32	32	32	32	33	33	33
33	33	33	33	34	34	34	34	34	34	35	35	35	35	36	36	36	37
37	37	37	37	37	38	38	39	39	40	40	40	40	40				

For large counts, precision does become meaningful. For example, if forced to guess, I would say that there are 300 million people living in the US. I suspect that this is accurate, but clearly I am not being very precise.

10.1.1 Dot Plot

We begin with an example of measurement data, taken from a student project in my Statistics 301 class.

On a spring evening, a Milwaukee police officer measured the speeds of 100 automobiles. The data were collected on a street in a "warehouse district" with a speed limit of 25 MPH. Fifty cars were measured between roughly 5:45 and 6:15 pm, referred to below as 6:00 pm. The remaining 50 cars were measured between roughly 10:40 and 11:20 pm, referred to below as 11:00 pm.

Each car's speed was measured to the nearest MPH. The sorted data, by time, are in Table 10.1. What do these lists reveal? We can see the smallest and largest speeds, but not much else.

Here is a very important point: With a dichotomous response, it is easy to summarize accurately a list of data; simply count S's and F's. With a numerical response the issue of summarizing is much more complicated (interesting?).

The first idea is to draw a picture of the data. Statisticians use a variety of pictures; we begin with **dot plots**, also called **dot diagrams**. The dot plots of the speeds, by time, are given in Figure 10.1. Examine these plots briefly before reading on. What do you see?

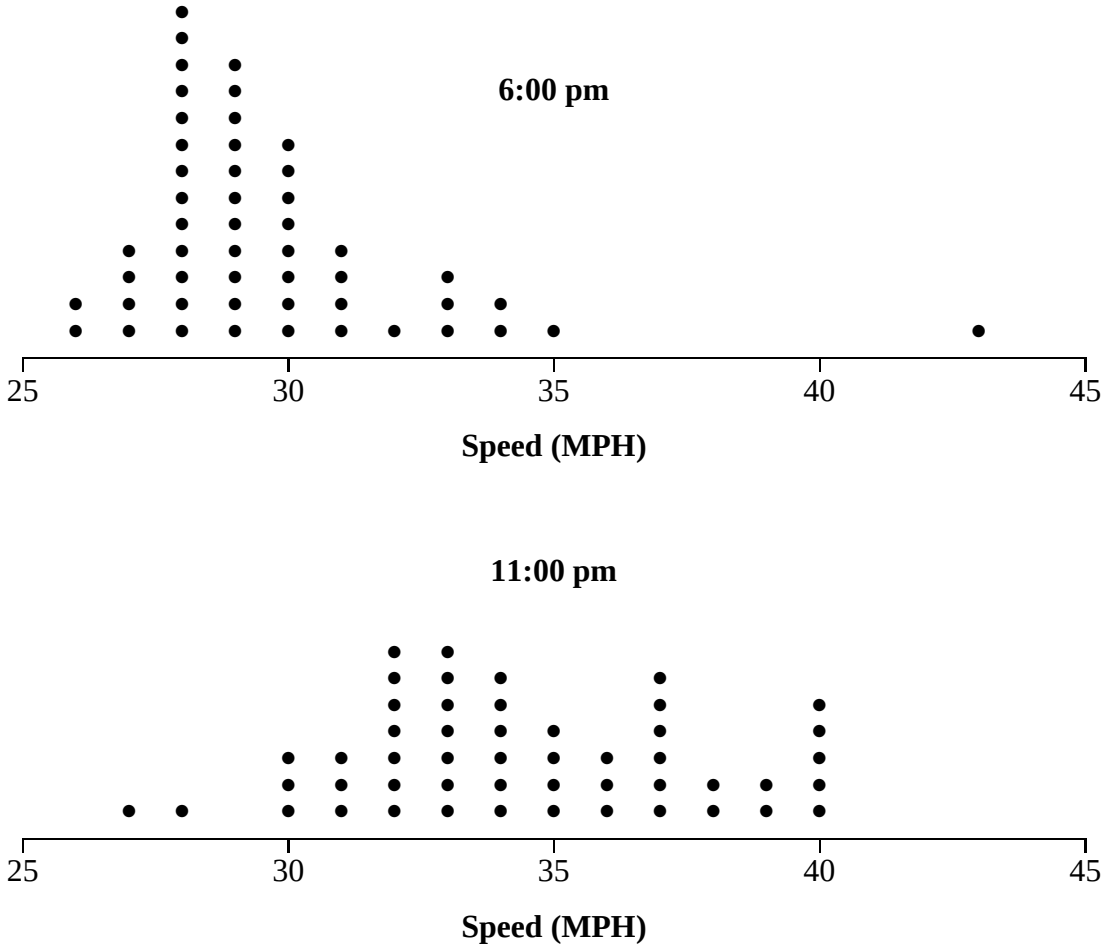
Statisticians have a number of suggestions for what to look at in an individual dot plot:

- **Outliers:** The 6:00 plot has one large outlier at 43. For 11:00, I don't see outliers; but one could label 27 and 28 as small outliers.

- **Gaps:** Of course, an outlier creates a gap as we see at 6:00 for the gap from 36 to 42. What I mean here are interior gaps, which we do not have in our plots.

- **Peaks:** There are two peaks in the 6:00 plot: at 28 and at 33. There are three peaks in the 11:00 plot: at 32–33, 37 and 40.

Figure 10.1: Dot Plots of Speeds, by Time.



- **Symmetry:** Neither dot plot is symmetric, but it very rare for a dot plot of real data to be (perfectly) symmetric. Thus, we look for approximate symmetry. In my opinion, neither of our dot plots is approximately symmetric, but 6:00 is clearly more asymmetric than 11:00. We will return to this topic below when we discuss shape.

I find it useful to note that outliers and peaks (and other summaries that we will learn about) can be **fragile** in some data sets. Fragile is not a standard term, and I don't know why b/c it is very important to note.

Fragile, the technical term, means pretty much what it does in real life, as in, "Don't touch your Great Aunt's collectibles b/c they are fragile." Here is what I mean, illustrated with our speed data.

An outlier is always fragile in the sense that if the subject/trial who gave the outlier hadn't shown up for the study, it wouldn't be there! For example, if the guy who drove 43 mph at 6:00 had taken a different route, we would have no outliers in our data sets.

Peaks are very interesting to statisticians and scientists. In many ways (as we shall see) it is easier to describe and think about data sets with one peak. Other times, however, it can be very exciting to note that a data set has more than one peak. Whenever I find a data set with more than one peak, I first decide whether any of the peaks are fragile.

Consider the 6:00 data. I consider the peak at 33 to be fragile b/c if one the persons driving 33 mph had slowed to 32 mph, then the peak would disappear. Also, and this is important, I can't think of any reason why 33 would be more popular than its neighbors for the speed of a car. B/c I view the peak at 33 as fragile, I label it unimportant and decide that the 6:00 plot has one important peak. Of course, you may reasonably disagree with me.

Now, consider the 11:00 data. I would not label any of the peaks fragile. Thus, I am resigned to there being three peaks.

Next, we will consider the **shape** of a dot plot. B/c it takes a great deal of time to draw pictures for these notes, I will show pictures of the following in lecture.

One shape for a dot plot is **rectangular**. This dot plot is symmetric with one peak, although the whole picture is the peak! It is not very important in practice.

The most important shape, by far, is **bell-shaped**. This shape is symmetric and looks like a normal curve.

If a plot has one peak and is not symmetric, then we should examine its tails. If the right tail is longer and heavier (longer is self explanatory, heavier means more data) than the left tail, we say that the dot plot is **skewed to the right**. If the left tail is longer and heavier than the right tail, we say that the dot plot is **skewed to the left**. I would describe the 6:00 plot as follows. It has one important peak at 28 mph and it is skewed to the right with a large outlier at 43.

Note that for dot plots with one peak, the labels: rectangular, bell-shaped, skewed to the right and skewed to the left, are **not exhaustive**. I often have data sets that fit none of these prototypes. That is ok. These labels are a help, not a requirement.

If a dot plot has multiple peaks I don't try to assign a shape to it, other than to say, for example, "It has multiple peaks."

Thus, I would not assign a name to the shape of the 11:00 dot plot. It has three peaks. We will discuss the meaning (if any) of these peaks in lecture.

Table 10.2: Sorted Heights, in Inches, of 50 College Men.

63.6	64.1	64.2	64.3	65.0	66.0	66.6	66.8	66.8	66.8
66.8	66.9	66.9	67.4	67.4	67.5	67.9	67.9	68.3	68.4
68.4	68.5	68.6	68.6	68.7	68.8	68.8	68.8	69.0	69.0
69.2	69.2	69.4	69.6	69.7	69.9	69.9	70.1	70.1	70.2
70.5	71.1	71.3	71.7	71.9	72.6	73.1	73.8	74.8	77.2

Table 10.3: Frequency Table of Heights, in Inches, of 50 College Men.

Class Interval*	Width (w)	Frequency (freq)	Relative Freq. rf $=$ (freq$/n$)	Density (rf$/w$)
63.0–65.0	2	4	0.08	0.04
65.0–67.0	2	9	0.18	0.09
67.0–69.0	2	15	0.30	0.15
69.0–71.0	2	13	0.26	0.13
71.0–73.0	2	5	0.10	0.05
73.0–75.0	2	3	0.06	0.03
75.0–77.0	2	0	0.00	0.00
77.0–79.0	2	1	0.02	0.01
Total	—-	$n = 50$	1.00	—-

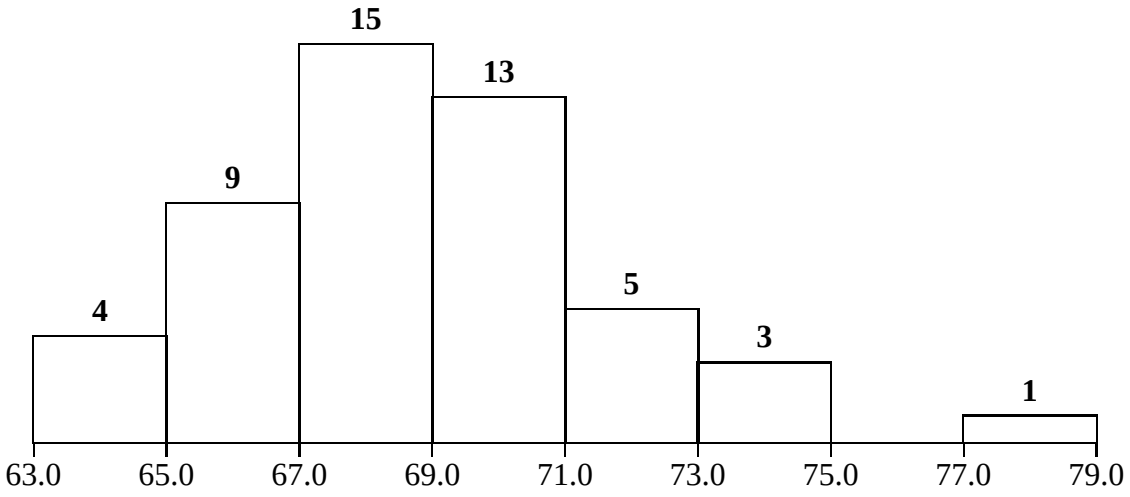
*Each class interval includes its left endpoint but not its right

10.1.2 Histograms

Table 10.2 presents the sorted heights of 50 college men, measured to the nearest tenth of an inch. If you draw a dot plot of these data you get a mess (trust me on this!). By my count (and I might be off a bit) there are nine peaks and many gaps. And I really can't say that any of the peaks are meaningful or important. And the only reason for the gaps is that we have too little data spread over too large a range. The only redeeming feature of the dot plot is that it reveals that the tallest man, at 77.2 inches, might be considered a large outlier. In this situation one should consider removing some of the detail in the data before drawing a picture. A histogram does this for us.

The first step in drawing a histogram is to create a frequency table of the data, such as the one shown in Table10.3. Note that each 'class interval' has a 'width' which is the arithmetic difference of the endpoints; i.e. it is the upper endpoint minus the lower endpoint. In this table, all widths are the same; this is not a requirement, but see below for issues that can arise if the widths vary. I hope that the rest of the column headings are self-explanatory. The endpoint convention needs to be noted. Each class interval includes its left endpoint, but not its right endpoint. Thus, the man who measured 65.0 inches is counted in the class interval 65.0 to 67.0. Figure 10.2 is the frequency histogram for these data. The directions for drawing it are quite simple:

Figure 10.2: Frequency Histogram of Heights.



1. Draw a horizontal number line and label the endpoints of the class intervals.

2. Above each class interval draw a rectangle with height equal to the frequency of the class interval.

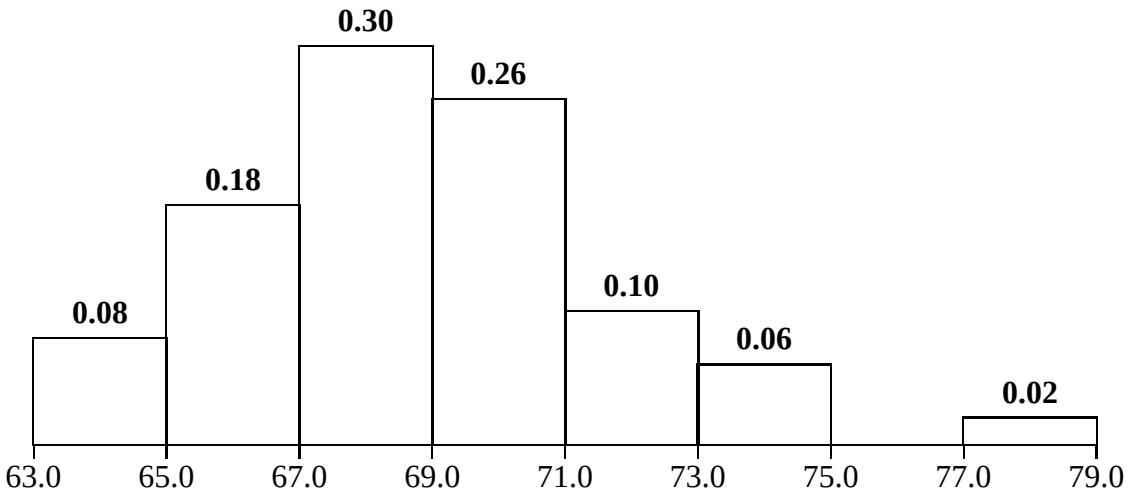
A frequency histogram is well-suited to answer the question: How many? As in, "How many men have heights between 67.0 and 69.0 inches?" (remembering our endpoint convention). The answer, from the picture, is 15.

The next type of histogram is the relative frequency histogram, pictured in Figure 10.3. To construct a relative frequency histogram, follow step 1 for a frequency histogram, but in step 2 make the height of the rectangle equal to the relative frequency of the class interval.

A relative frequency histogram is especially well-suited to answer the question, "What proportion?" as in "What proportion of men have a height between 69.0 and 71.0 inches?" The answer is the height of the rectangle above 69.0 to 71.0, which we read to be 0.26.

When I compare Figures 10.2 and 10.3, I can't help but think of the noted Rick Moranis film, *Honey, I Shrunk the Kids*, b/c, of course, if we take the picture in Figure 10.2 and shrink the height of each rectangle by a factor of $n = 50$, we get Figure 10.3. The third, and final, histogram is the density scale histogram. Horizontally, this histogram is the same as the frequency and relative frequency histograms; the difference is that now the height of each rectangle is its density. The density scale histogram for the heights is in Figure 10.4. Now it is true that this density scale histogram is a shrinkage of the relative frequency histogram, but this is true b/c $w = 2 > 1$. If $w < 1$, the rectangles in the density scale histogram will be taller than the rectangles in the relative frequency histogram and if $w < 1/n$ then the rectangles in the density scale histogram will be taller than even the rectangles in the frequency histogram. So, we need to remember the very unnecessary sequel, *Honey I Blew Up the Kid*.

Figure 10.3: Relative Frequency Histogram of Heights.



Unlike our earlier histograms, the heights in a density scale histogram do not answer any question. This is b/c the density scale histogram is about areas, not heights. To see this, note that the area of any rectangle is

$$\text{Base } \times \text{ Height } = w \times \frac{\text{rf}}{w} = \text{ rf}.$$

Thus, the density scale histogram represents relative frequencies by areas.

This begs the question: Why do we have density scale histograms? After all, they require us to calculate an area to find a number which is the height in the relative frequency histogram.

There are two reasons why density scale histograms are important.

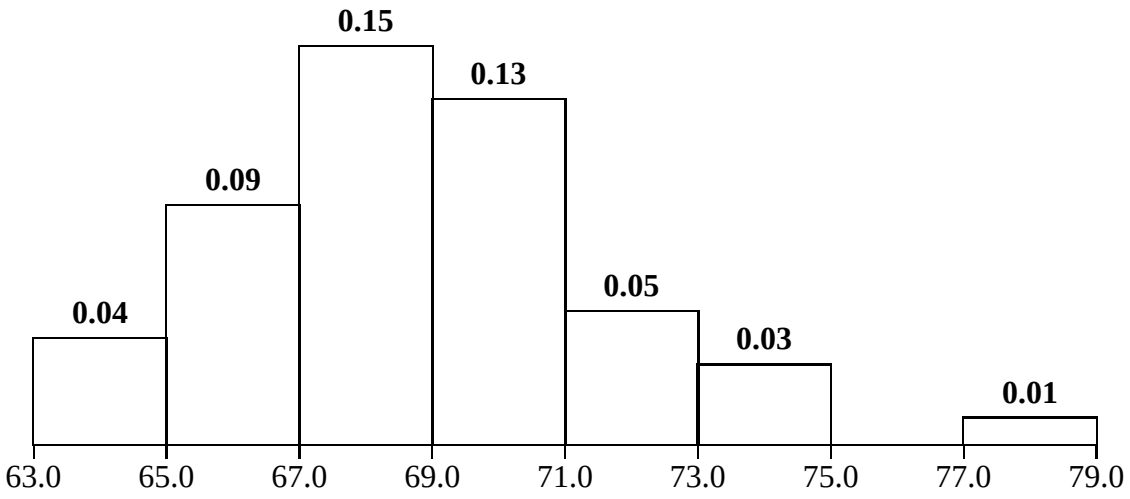
1. For the development of theory it is convenient to have a picture whose total area is one.

2. If we allow the widths to vary, we *should* use the density scale histogram. This is explained below.

For our data on heights of men, we have three histograms. There are two uses for histograms; one use is relatively unimportant, but the other is very important. The less important reason is to answer specific questions about the data set. I discussed this earlier; frequency histograms are good for *how many* questions; relative frequency histograms are good for *what proportion* questions; and density scale histograms are not particularly good for this. I say that this is relatively unimportant b/c we can also get answers to these questions just by looking at the data and counting, which is pretty easy with a computer.

The more important use is to find a shape for the data. For the heights data, b/c of the shrinking issue, all three histograms have the same shape. (We can label shapes of histograms with the same terms we used for shapes of dot plots.)

Figure 10.4: Density Scale Histogram of Heights.



Look at any one of the histograms. What shape do you see? Literally, there is one peak, in the interval 67 to 69, and the right tail is clearly longer and heavier than the left tail. Thus, the shape is skewed to the right. But I could also see the histogram as being described as approximately symmetric, with a peak at 67 to 71 inches.

It can be difficult to say definitively what a shape is. This is illustrated with the height data and Table 10.4 and its frequency histogram in Figure 10.5. There is again one peak, but now while the left tail is a bit heavier, the right tail is a bit longer. In my opinion, it does not fit any of our named shapes very well.

10.1.3 Variable Width Histograms

Thus far, our two examples have had constant-width class intervals. While it is common for scientists and statisticians to follow this *restriction*, often we can do better if we don't. In particular, constant width means that we have the same amount of detail across the range of the data. It often makes more sense to have more detail where data are plentiful and less detail where data are scarce. (This concept is followed in my road atlas which devotes two pages to New York City and one page to Alaska.) I will illustrate these ideas with some data from the 2009 Major League Baseball season.

With the help of the website cnnsi.com I found the 100 major leaguers with the most official at-bats during 2009, ranging from a high of 682 for Aaron Hill of Toronto to a low of 514 for Jason Kubel of Minnesota. For each player I copied his number of home runs; the sorted values are in Table 10.5.

Table 10.6 is a frequency table for the number of home runs and Figure 10.6 is its density scale histogram. Notice the widths of the class intervals all equal 10. Table 10.7 is a second frequency

Table 10.4: Frequency Table of Heights, in Inches, of 50 College Men.

Class Interval*	Width (w)	Frequency (freq)	Relative Freq. rf $=$ (freq$/n$)	Density (rf$/w$)
62.0–64.0	2	1	0.02	0.01
64.0–66.0	2	4	0.08	0.04
66.0–68.0	2	13	0.26	0.13
68.0–70.0	2	19	0.38	0.19
70.0–72.0	2	8	0.16	0.08
72.0–74.0	2	3	0.06	0.03
74.0–76.0	2	1	0.02	0.01
76.0–78.0	2	1	0.02	0.01
Total	—-	$n = 50$	1.00	—-

*Each class interval includes its left endpoint but not its right

Figure 10.5: Frequency Histogram of Heights.

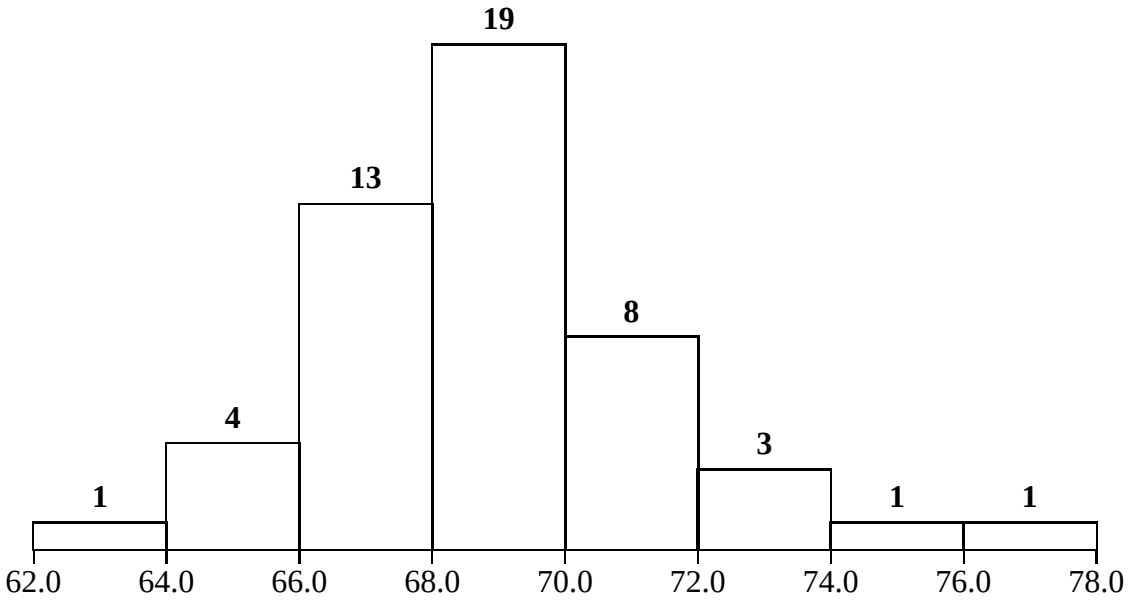


Table 10.5: Sorted Number of Home Runs of 100 Major Leaguers in 2009.

2	3	4	5	6	6	7	7	8	8
9	9	9	9	10	10	10	11	11	11
11	11	12	12	12	12	12	12	13	13
14	14	15	15	15	15	15	15	15	15
15	16	16	17	18	18	18	18	20	20
20	21	21	22	23	23	23	24	24	24
24	24	25	25	25	25	25	25	26	26
26	27	28	28	28	28	30	31	31	31
31	32	32	32	33	33	34	34	35	35
36	36	36	38	39	40	44	45	46	47

Table 10.6: Frequency Table of Number of Home Runs.

Class Interval*	Width (w)	Frequency (freq)	Relative Freq. rf = (freq/n)	Density (rf/w)
0–10	10	14	0.14	0.014
10–20	10	34	0.34	0.034
20–30	10	28	0.28	0.028
30–40	10	19	0.19	0.019
40–50	10	5	0.05	0.005
Total	—-	$n = 100$	1.00	—-

*Each class interval includes its left endpoint but not its right

Table 10.7: Frequency Table of Number of Home Runs.

Class Interval*	Width (w)	Frequency (freq)	Relative Freq. rf = (freq/n)	Density (rf/w)
0–5	5	3	0.03	0.006
5–10	5	11	0.11	0.022
10–15	5	18	0.18	0.036
15–20	5	16	0.16	0.032
20–25	5	14	0.14	0.028
25–30	5	14	0.14	0.028
30–40	10	19	0.19	0.019
40–50	10	5	0.05	0.005
Total	—-	$n = 100$	1.00	—-

*Each class interval includes its left endpoint but not its right

Figure 10.6: Density Scale Histogram of Number of Home Runs.

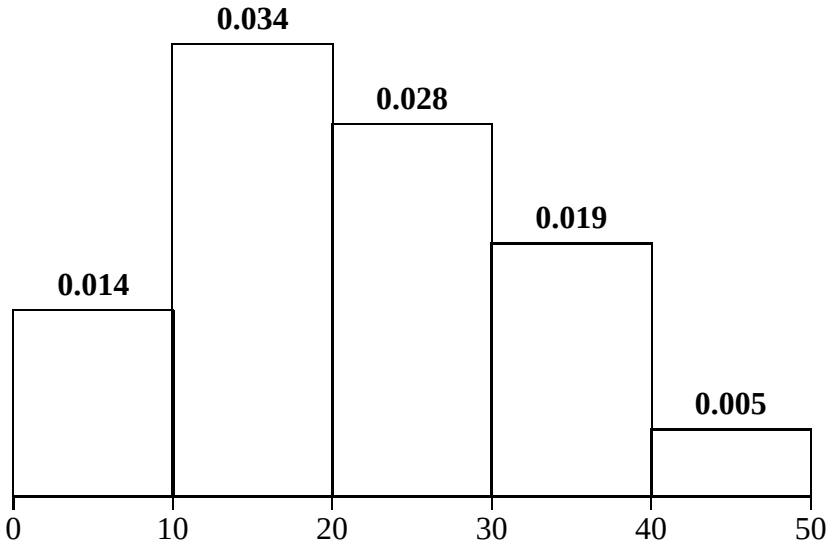


Figure 10.7: Variable Width Density Scale Histogram of Number of Home Runs.

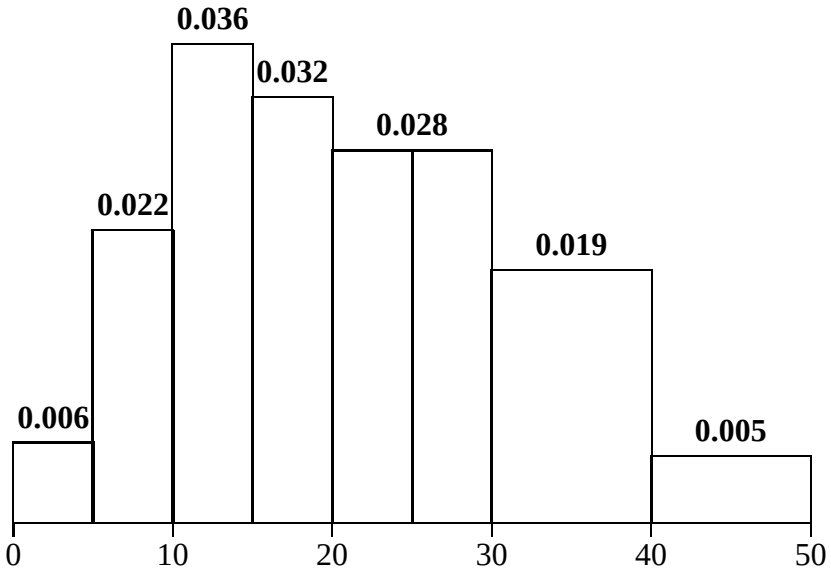


Figure 10.8: (Misleading) Variable Width Frequency Histogram of Number of Home Runs.

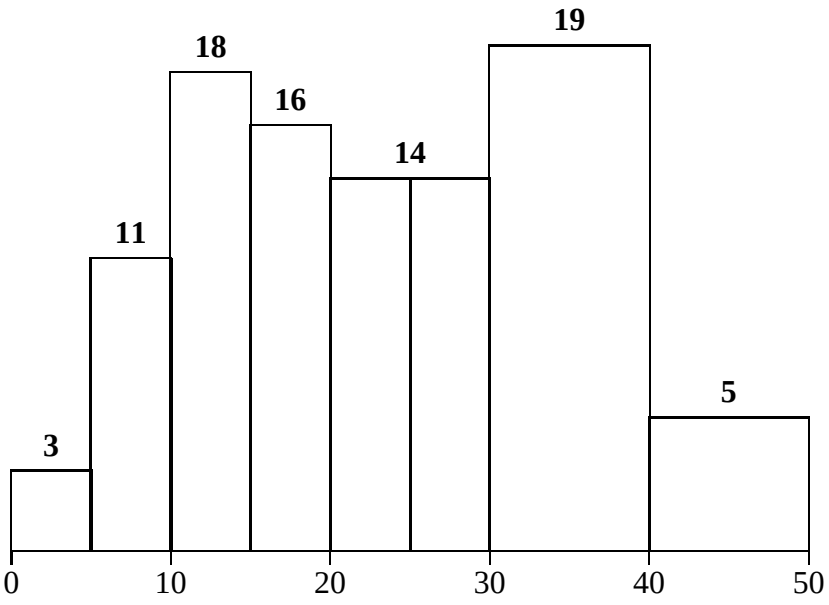


table for the number of home runs. Figures 10.7 and 10.8 are its density scale and frequency histograms. Notice the widths of the class intervals are not constant. Why do I label the frequency histogram misleading?

In view of the above, let me summarize.

- If all class intervals have the same width, the three types of histograms all give the same shape.

- If all class intervals do not have the same width, you should use the density scale histogram. The other histograms can be misleading.

10.2 Numerical Summaries

Numerical summaries fall into three broad categories:

- **Measures of Center:** The mean, median and mode.

- **Measures of Position:** The percentiles and quantiles, including quartiles.

- **Measures of Spread:** The range, interquartile range, variance and standard deviation.

10.2.1 Measures of Center

I will try to keep this brief; first, b/c I suspect you know much of this and second b/c it is pretty dull, even by the standards of this course.

The **mean** of a set of numbers is its arithmetic average. For example, the mean of 5, 1, 4 and 10 is:

$$(5 + 1 + 4 + 10)/4 = 20/4 = 5.$$

It will help us later if we introduce some notation now. Denote a collection of n numbers by:

$$x_1, x_2, x_3, \ldots x_n.$$

With this notation, the mean of the numbers is calculated as

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

We denote the mean by $\bar{x}$, read x-bar. (If the data are denoted by y's instead of x's, we call the mean y-bar; and so on.)

As you have noted above, we often like to sort our data from smallest to largest. We denote the sorted data by:

$$x_{(1)}, x_{(2)}, x_{(3)}, \ldots x_{(n)};$$

i.e. we put parentheses around the subscripts to denote sorting.

For example, suppose we have $n = 5$ numbers:

$$x_1 = 8, x_2 = 3, x_3 = 1, x_4 = 8, x_5 = 6.$$

After sorting, these numbers are:

$$x_{(1)} = 1, x_{(2)} = 3, x_{(3)} = 6, x_{(4)} = 8, x_{(5)} = 8.$$

The idea of the **median** of a set a numbers is to find the number in the center position of the sorted list. This requires some care b/c the answer depends on whether the sample size is an odd number or an even number. For example, for the five numbers above, there is a unique center position, position 3, and the number in this position, 6, is the median.

If, however, the sample size is even, we need to be more careful. For example, consider four sorted numbers: 1, 4, 5 and 10. In four positions, positions 2 and 3 have equal claim to being a center position, so the median is taken to be the arithmetic average of the numbers in positions 2 and 3; in this case the median is the arithmetic average of 4 and 5, giving us 4.5.

If we denote the data by x's, then the median is denoted by $\tilde{x}$, read x-tilde. We have a formula for calculating the median:

- If n is an odd integer, define $k = (n + 1)/2$, which will be an integer. Then,

$$\tilde{x} = x_{(k)}.$$

- If n is an even integer, define $k = n/2$, which will be an integer. Then,

$$\tilde{x} = [x_{(k)} + x_{(k+1)}]/2.$$

Refer to the above case when n is an even integer. If $x_{(k)}$ and $x_{(k+1)}$ are different numbers, then sometimes the interval $[x_{(k)}, x_{(k+1)}]$ is called the **interval of medians**.

Finally, I will define the **mode**. Suppose that we have a set of n numbers:

$$x_1, x_2, x_3, \ldots x_n.$$

If these are n different numbers, then each of the numbers is a mode and we have, obviously, n modes to the data set. If these are not n different values , then determine the frequency of each distinct values. If one of these frequencies is uniquely the largest of the frequencies, then the value associated with said frequency is the mode. If several of these frequencies tie for being the largest, then the values associated with this largest frequency are all called modes. An example might help.

Suppose we have $n = 7$ sorted values:

$$2, 3, 7, 12, 13, 15, 19.$$

This data set has seven modes. If the data are

$$2, 2, 7, 12, 13, 15, 19,$$

then we have one mode, which is equal to 2. If the data set is

$$2, 2, 2, 12, 15, 15, 15,$$

then we have two modes, which are 2 and 15.

Below is a list of important features of these three measures of center.

1. There is one exact connection between a picture of a data set and its measure of center:

   The mean of a set of data is equal to the center of gravity of its dot plot.

2. The mean is sensitive to the presence of even one wild outlier. For example, the mean of 1, 4, 5 and 10 is 5. The mean of 1, 4, 5 and 1000 is 252.5. For either set of data the median is 4.5.

3. The median can be fragile and, hence, a misleading measure of center. Consider the following two frequency distributions:

	Data Set:	
Value	A	B
0	50	51
1	26	25
2	25	25
Total	101	101
Mean	0.752	0.743
Median	1	0

These data sets are nearly identical; data set A becomes B by changing only a single observation, out of 101 observations, from 1 to 0. The fact that the data sets are nearly identical is well reflected in the means, they are almost the same. The medians, however, are deceptive b/c they suggest that the distributions are very different.

4. The mode can be fragile. Consider the sorted data:

$$1, 1, 1, 1, 3, 5, 6, 9, 12, 12, 12.$$

The mode is 1. If a 1 changes to a 2, there are two modes, 1 and 12. If another 1 changes to anything, the mode is 12.

5. For measurement data the mode depends critically on the precision of the measurement. For example, for the heights, measured to the nearest tenth of an inch, of 50 men in Table 10.2, the mean is 68.842 inches, the median is 68.75 inches and the mode is 66.8 inches.

   If we now round the heights to the nearest inch, the mean and median don't change much: the mean becomes 68.86 inches and the median becomes 69 inches. The mode, however, changes to 69 inches, which is quite different from 66.8 inches.

I have presented the mean, median and mode as alternative ways to describe a set of data. It is useful to make this whole idea a bit more rigorous.

We think of a 'loss' as a bad thing. Losing money, a job, a partner, a game, all of these hurt. Mathematicians like to formalize the notion of loss in the context of describing a set of data. We will consider three loss functions:

- **A miss is as good as a mile!**
- **Absolute error loss.**
- **Squared error loss.**

Here is the framework. Our data consist of

$$x_1, x_2, x_3, \ldots x_n.$$

We want to choose the best number c to describe these data. If we describe $x_j = x$ by c, we incur a loss denoted by $L(x, c)$. The rules for the loss function L are:

1. A perfect description incurs 0 loss; symbolically, this means $L(c, c) = 0$.
2. An imperfect description incurs a positive loss: $L(x, c) > 0$ for any $x \neq c$.

The best choice for c is the number that minimizes

$$\sum_{i=1}^{n} L(x_i, c).$$

We call this sum the total loss and denote it by $L(\text{Total})$.

We now consider three choices for our loss function L.

A miss is as good as a mile. The loss function is $L(x, c) = 1$ for any $x \neq c$. The idea is that all imperfect descriptions are equally bad. Clearly the best choice for c is to set it equal to any one of the modes. The total loss is $n - f$, where f is the frequency of any mode.

Absolute error loss. The loss function is $L(x, c) = |x - c|$. It will be shown in lecture that the best choice of c is the median of the data, $\tilde{x}$. If n is an odd integer, then the median is the unique minimizer of the total loss; if n is an even integer, then any number in the interval of medians will minimize the total loss.

Squared error loss. The loss function is $L(x, c) = (x - c)^2$. The total loss is

$$\sum_{i=1}^{n}(x_i - c)^2 = \sum_{i=1}^{n}(x_i - \bar{x} + \bar{x} - c)^2 = \sum_{i=1}^{n}[(x_i - \bar{x})^2 + 2(x_i - \bar{x})(\bar{x} - c) + (\bar{x} - c)^2] =$$

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 + 2(\bar{x} - c)\sum_{i=1}^{n}(x_i - \bar{x}) + n(\bar{x} - c)^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - c)^2,$$

this last equality is true b/c $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$. This final expression is minimized by taking $c = \bar{x}$. Thus, for squared error loss, the best description is the mean of the data.

10.2.2 Measures of Position

We learned about the median earlier in these notes. The median is also called the 50th percentile and the 0.50 quantile. Quantiles are the decimal version of a percentile. For example, the 63rd percentile is the same as the 0.63 quantile.

To make this abstract, but I hope not confusing, let π be a number strictly between 0 and 1; i.e. $0 < \pi < 1$. The π quantile is the same number as the 100π percentile. My example of the previous paragraph illustrates this notion with $\pi = 0.63$; the 0.63 quantile is the same number as the $100(0.63) = 63$rd percentile.

By convention, there are 99 percentiles—corresponding to the integers 1 thru 99—allowing π to be any of the values: 0.01, 0.02, 0.03, ... 0.99.

We begin with a new way to think about the median. As I will show you, the following definition agrees with our earlier definition of the median.

- *At least* one-half of the data are *less than or equal to* the median.
- *At least* one-half of the data are *greater than or equal to* the median.

For any odd sample size, there is a unique number that satisfies this definition, namely the number in the center position. For our example data of 4, 9 and 20, the number in the center position, 9, is the unique number that satisfies this definition, as argued below.

- Two-thirds of the data are less than or equal to 9.
- Two-thirds of the data are greater than or equal to 9.

Thus, 9 satisfies our definition. But is it unique? Yes. (For any candidate smaller than 9, the first item fails; for any candidate larger than 9, the second item fails.)

Next, consider an even sample size. This is trickier, so I will use our earlier data of 4, 7, 9 and 20. You can check that our median 8 satisfies both conditions. But it is not unique. Actually, any number between 7 and 9 inclusive will satisfy our definition of the median. Also, any number smaller than 7 or larger than 9 will fail our definition. Thus, in a strictly literal math sense, for an even sample size there can be an entire interval of numbers that satisfy the definition of median. Most statisticians and mathematicians agree that if there is an interval of medians, we call the midpoint of the interval *the median.* And we will always do this in this course.

You might well wonder what was the point of all of the above. Before you read this we all agreed on what the median was, and now I have just made it more complicated. Well, we need the above for percentiles and quantiles. I will illustrate.

First, let us be specific. Suppose we want to find the 35th percentile, which we will denote by $P_{35} = Q_{0.35}$, the 0.35 quantile. We define the 35th percentile to be any number that has the following two properties.

- *At least* 35% of the data are *less than or equal to* it.
- *At least* 65% of the data are *greater than or equal to* it.

If an entire interval of numbers satisfy both properties, we take the 35th percentile to be the midpoint of the interval.

Next, I give you the algorithm for calculating the 35th percentile.

1. Calculate $k = 0.35n$.
2. If k is an integer, then the 35th percentile equals
$$[x_{(k)} + x_{(k+1)}]/2.$$
3. If k is not an integer, round it up to the next integer and call it k'. The 35th percentile equals $x_{(k')}$.

For example, suppose $n = 100$. Then,

$$0.35n = 0.35(100) = 35$$

is an integer and the 35th percentile equals

$$[x_{(35)} + x_{(36)}]/2.$$

Let's check that this works.

First, suppose that $x_{(35)}$ and $x_{(36)}$ are different numbers. Then exactly 35% of the data are less than the percentile and exactly 65% of the data are greater than the percentile. If $x_{(35)}$ and $x_{(36)}$ are the same number then at least 36% of the data are less than or equal to the percentile and at least 66% of the data are greater than or equal to the percentile.

As another example, suppose that $n = 150$. Then

$$k = 0.35n = 0.35(150) = 52.5$$

is not an integer, so we round it up to $k' = 53$ and the 35th percentile equals $x_{(53)}$. Let's check that it works.

First, clearly at least 53 observations are less than or equal to $x_{(53)}$, and $53/150 = 0.353$ is at least 35%. Second, at least $150 - 52 = 98$ observations are greater than or equal to $x_{(53)}$, and $98/150 = 0.653$ is at least 65%.

Now that we understand the 35th percentile, we will consider any arbitrary percentile. First, the definition. We define the 100π percentile to be any number that has the following two properties.

- *At least* 100π% of the data are *less than or equal to* it.
- *At least* $100(1 - \pi)$% of the data are *greater than or equal to* it.

If an entire interval of numbers satisfies both properties, we take the 100π percentile to be the midpoint of the interval.

Next, I give you the algorithm for calculating the 100π percentile.

1. Calculate $k = \pi n$.
2. If k is an integer, then the 100π percentile equals

$$[x_{(k)} + x_{(k+1)}]/2.$$

3. If k is not an integer, round it up to the next integer and call it k'. The 100π percentile equals $x_{(k')}$.

Final comment. The 25th percentile is called the first quartile; the 50th percentile—in addition to being called the median—is called the second quartile; and the 75th percentile is called the third quartile. Note the word is quartile, not quantile. To add to the confusion, the quartiles are denoted Q_1, Q_2 and Q_3. Thus, a Q with a subscript can be a quantile or a quartile. If the subscript is 1, 2 or 3, its a quartile; if the subscript is smaller than 1, it is a quantile.

10.2.3 Measures of Spread

All measures of spread must have the following properties.

1. For any data set, the measure of spread is a number that is nonnegative.
2. For any data set, the measure of spread equals 0 if, and only if, there is no spread in the data set. (Although see the exception for the IQR.)
3. For two data sets, the data set with the larger value of the measure of spread is deemed to be the data set with more spread.

We will learn about three ways to measure spread. If you hope to become famous by inventing a new measure of spread, make sure it satisfies the conditions above; o.w. people will ignore your work and you won't become famous.

The first measure of spread is the **range**, denoted by R. Consider again the data on the number of home runs in Table 10.5. The smallest number in the data set is 2 and the largest is 47. In everyday language, I would say, "The numbers of home runs *range from* 2 *to* 47. But statisticians don't like to have words (from, to) in a summary, nor do they like to have two numbers (2, 47) in a summary. Thus, somewhat bizarrely, statisticians define the range to be:

$$R = \text{Maximum } - \text{ Minimum }.$$

Thus, $R = 47 - 2 = 45$, for the home run data.

The range is not a very popular measure of spread. It has the following bad properties.

1. Just like the mean, the range is sensitive to even one wild outlier.

2. Imagine that you are collecting data from a source, one observation at-a-time, building your data set of size n. After each new observation you recalculate the range. With each recalculation, the range can remain the same or it can increase, but—and this is the key point—it can never decrease. Added to this, and this is not at all obvious, there is no good way to adjust for sample size in the range. As a result, a large range might mean a lot of spread in a small set of data or a moderate amount of spread in a huge data set.

The next measure of spread is the interquartile range, abbreviated IQR. Recall the definition of the quartiles in the previous subsection. The IQR is computed as:

$$\text{IQR } = Q_3 - Q_1.$$

I will now explain the motivation behind the IQR. Remember that the three quartiles divide (approximately) the data set into quarters: approximately one-quarter of the data are smaller than Q_1; approximately one-quarter of the data are between Q_1 and Q_2; approximately one-quarter of the data are between Q_2 and Q_3; and approximately one-quarter of the data are larger than Q_3. Thus, approximately one-half of the data—the center half—are between Q_1 and Q_3; for this reason the IQR is interpreted as the range of the center half of the data.

View the IQR as an attempt to improve on the range. The IQR focuses on the center half of the data and ignores the values in the lower and upper quarters of the data. As a result, it is not influenced by a small number of wild outliers. Also, b/c it focuses on the center half of the data, it does adjust for sample size; i.e. it avoids the two bad problems of the range.

The IQR has two main weaknesses.

1. It really does not help us solve any scientific problems. It summarizes the data and that is it.

2. It strikes me as very odd that anyone would want to ignore half the data when measuring spread. Especially to ignore the smallest and largest quarters of the data; isn't that where we see spread?

As a result of this second weakness, we get some rather strange answers using the IQR. For example, consider the two data sets below:

- **Data Set A:** 1000, 1000, 1000, 1000 and 1000.
- **Data Set B:** 0, 1000, 1000, 1000 and 10,000.

You can verify that $Q_1 = Q_3 = 1000$ for both data sets; thus, both data sets have IQR $= 0$; but it seems strange to me that one would want to say these sets have the same spread or to say that the second set has no spread.

Many statisticians love the IQR, but I believe they are largely misguided. Many people always like anything that is new, even if it is nearly worthless.

So, is there any measure of spread that I recommend? Yes, the standard deviation—and mathematically equivalent variance—discussed below. Yes, the standard deviation is not perfect, but especially if we learn and remember its limitations, it is a very useful tool for us.

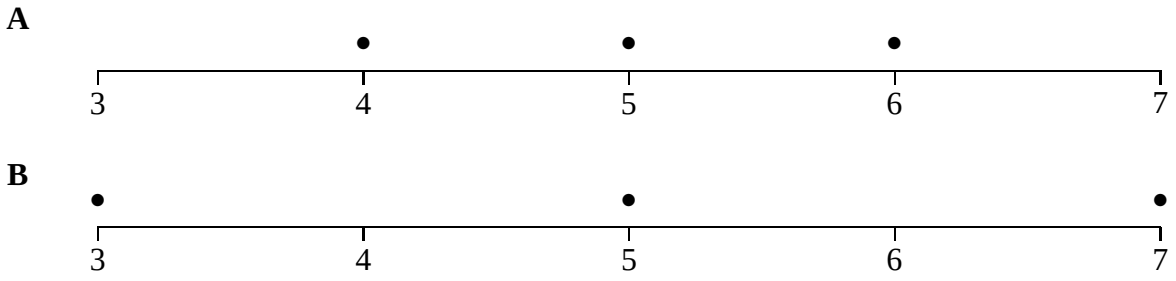
10.2.4 The Standard Deviation and Variance

Earlier in these notes we learned about the standard deviation, σ, and variance, σ^2, of a probability distribution. Also, we learned formulas for them for the Binomial and Poisson distributions. In this section we learn about the standard deviation, s, and variance, s^2, for a set of numerical data.

We have learned three measures of center: the mode, which is rarely used; the median, which focuses on position in the sorted list of data; and the mean, which is obtained by doing arithmetic (adding, then dividing) on a set of data.

Our three measures of spread are: the range, which is rarely used; the IQR, which focuses on positions in the sorted list of data; and the standard deviation, which is obtained by doing arithmetic (subtracting, squaring, adding, dividing, taking the square root) on a set of data. As a result, there is a tendency among researchers to match the median with the IQR and to match the mean with the standard deviation. You *can* mix them, but usually it makes sense to decide whether positions or arithmetic is more meaningful for your scientific problem and data.

Below are two data sets, A and B. Clearly B has more spread than A. I will use these sets to introduce the formula for the standard deviation.



Both data sets have a mean of 5. The first idea behind the standard deviation is that we measure spread relative to the center of the data set. We *compare* each observation with the center. We compare by subtracting. Thus, for each observation x, we calculate $x - \bar{x}$ which is called the deviation in x (relative to the mean). Below are the deviations for data sets A and B.

The second idea is that standard deviation is a function of the deviations; i.e. two data sets with the same deviations will have the same standard deviation. For example, if we define data set C to consist of 198, 200, 202, its deviations will be -2, 0 and $+2$. Hence, data set C will have the same standard deviation as data set B above.

It is helpful to create the following tables.

A		B	
x	$x - \bar{x}$	x	$x - \bar{x}$
4	-1	3	-2
5	0	5	0
6	$+1$	7	$+2$
Total	0	Total	0

Note that for both data sets, $\sum(x - \bar{x}) = 0$. This is, in fact, true for every data set: the total of the deviations is always 0.

We say that n deviations have $(n - 1)$ **degrees of freedom.** I will present an example of why we use this term in lecture.

We need to combine the deviations to get an overall measure of spread. Clearly, summing them does not work b/c the negative deviations will cancel the positive ones.

A deviation of 0 denotes no spread and as the deviation moves away from 0, in either direction, it reflects greater distance from the center, and, hence, greater spread.

Thus, it would seem that it would be a good idea to take the absolute value of the deviations before combining them. Sadly, whereas it 'makes sense' to take the absolute value, this operation turns out not to be useful in any way! What turns out to be very useful—as we shall see—is to square each deviation, as shown in the table below.

A			B		
x	$x - \bar{x}$	$(x - \bar{x})^2$	x	$x - \bar{x}$	$(x - \bar{x})^2$
4	-1	1	3	-2	4
5	0	0	5	0	0
6	$+1$	1	7	$+2$	4
Tot.	0	2	Tot.	0	8

Now we come to a major disagreement between mathematicians and statisticians. Both agree to measure spread by using the total of the squared deviations: and both realize that they must adjust for sample size. Mathematicians opt to calculate the mean squared deviation; i.e. they divide by n. Statisticians divide by the degrees of freedom to get the following formula for the **variance**:

$$s^2 = \frac{\sum(x - \bar{x})^2}{n - 1}.$$

For data set A, $s^2 = 2/(3-1) = 1$; and for data set B, $s^2 = 8/(3-1) = 4$.

The variance would be an acceptable way to measure spread, except for two difficulties.

- It has no interpretation other than the obvious: the variance is almost the mean of the squared deviations.
- It has the units wrong. (Will be discussed in lecture.)

Statisticians prefer the standard deviation, s, which is the (positive) square root of the variance. For data set A, $s = \sqrt{s^2} = \sqrt{1} = 1$; and for data set B, $s = \sqrt{s^2} = \sqrt{4} = 2$.

There is an interpretation for s, but it is bit strange.

Recall the mean: it is the center of gravity of the dot plot of the data. This is a direct, exact statement. The interpretation of s is indirect and only approximate. And sometimes the approximation is bad.

Imagine the data are presented as a dot plot and that you are standing at the mean. You decide,

> I want to reach out my arms in each direction (imagine you are elastic man/woman) and capture data within them.

You then ask the question:

> How far must I reach to capture 68% of the data?

The **empirical rule (ER)** says that you must reach out s units (in both directions).

I will illustrate the ER with several sets of data. Below is data set D, with $n = 100$.

−5	67	108	137	160	179	196	210	224	236	248	258	268	278	287
295	304	312	319	327	334	341	348	354	361	367	373	379	385	391
397	403	409	414	420	425	430	436	441	446	452	457	462	467	472
477	482	490	490	497	503	508	513	518	525	525	533	538	543	548
554	559	564	570	575	580	586	591	597	603	609	615	621	627	633
639	649	649	663	663	673	681	688	696	705	713	727	727	742	758
758	776	790	804	821	840	863	892	933	1005					

Here are some facts about D: the mean is 500.0, the standard deviation is 200.0 and a histogram of the data is approximately bell-shaped.

By counting, 68 of these numbers are between 300 and 700 ($\bar{x} \pm s$), a perfect agreement with the ER prediction. **In general, if the data are approximately bell-shaped, then the ER approximation is very good.**

Below is data set E, with $n = 100$.

106	108	109	111	112	112	113	117	117	122	124	125	127	129	130
133	134	138	140	141	142	145	146	156	156	158	164	170	172	173
175	175	179	180	186	190	194	195	202	202	204	206	207	207	217
221	226	228	228	233	239	245	247	250	261	266	285	286	289	300
303	306	307	310	314	324	332	337	338	341	344	354	357	366	373
384	389	403	411	415	421	421	430	439	449	472	472	523	529	567
584	610	635	637	641	651	664	793	973	1324					

Here are some facts about E: the mean is 300.0, the standard deviation is 200.0 and a histogram of the data is strongly skewed to the right.

By counting, 87 of these numbers are between 100 and 500 ($\bar{x} \pm s$), a bad agreement with the ER approximation. **In general, if the data are strongly skewed (right or left) the actual amount of data in the interval $\bar{x} \pm s$ is much larger than predicted by the ER.**

Finally, below is data set G, with $n = 100$.

63	100	120	135	147	157	166	174	181	188	194	201	207	212	218
223	227	233	237	243	247	251	256	260	265	270	275	279	284	288
292	298	302	308	312	317	323	328	334	341	347	354	361	369	378
388	400	415	420	447	459	472	477	491	504	514	523	531	538	545
551	557	564	569	574	580	584	589	594	599	604	608	613	617	621
627	631	636	640	645	649	654	659	664	669	674	679	685	691	697
704	711	718	726	735	745	757	771	792	828					

Here are some facts about G: the mean is 446.0, the standard deviation is 200.0 and its histogram has two peaks and is approximately symmetric.

By counting, 60 of these numbers are between 246 and 646 ($\bar{x} \pm s$), a somewhat bad agreement with the ER approximation. **In general, if the data have two peaks, the actual number of data points in the interval $\bar{x} \pm s$ is smaller than predicted by the ER. The more separated the peaks, the smaller the count in the interval.**

The ER can be generalized. The most common generalization is to say that approximately 95% of the data are within two standard deviations of the mean (in the interval $\bar{x} \pm 2s$).

You can check the following counts:

- In data set D, 96 observations are in the interval $[100, 900]$.
- In data set E, 97 observations are in the interval $[-100, 700]$.
- In data set G, all 100 observations are in the interval $[46, 846]$.

In this class, you will **never** be required to calculate s.

Earlier, we learned about standardizing a random variable. Recall that if X is a random variable with mean μ and standard deviation σ, then the standardized version of X is denoted by Z and they are related by the following equation:

$$Z = \frac{X - \mu}{\sigma}.$$

This relationship extends to any particular value of X, denoted by x, to give

$$z = \frac{x - \mu}{\sigma}.$$

In this last relationship, z is called the z-score of x.

A similar idea is developed for data. Given a set of data

$$x_1, x_2, x_3, \ldots x_n,$$

with mean $\bar{x}$ and standard deviation s. For any particular x value, define its **t-score** by

$$t = \frac{x - \bar{x}}{s}.$$

Note that this is a t-score, not a z-score. A z-score requires the mean and standard deviation of the population/probability distribution. A t-score requires the mean and standard deviation of a set of data.