1 Lesson 6: Measure of Variation

1.1 The range

As we have seen, there are several viable contenders for the best measure of the central tendency of data. The mean, the mode and the median each have certain advantages and certain disadvantages. In any specific situation, anyone of these could provide the best intuitive value for the center. Once a center has been established, the next question is, how much does the data vary from this center? As it turns out, there are very few alternatives in mathematics for this measure.

The first measure of the variation in a data set is the range. The range of numerical data set is simply the difference between the highest value and the lowest. Let us consider our familiar example of student grades:

Name	Test 1	Test 2	Test 3
April	55	71	64
Barry	63	67	63
Cindy	88	90	91
David	97	92	87
Eileen	58	55	75
Frank	90	89	96
Gena	88	100	85
Harry	71	70	71
Ivy	65	75	85
Jacob	77	70	65
Keri	75	88	85
Larry	88	92	92
Mary	95	95	100
Norm	86	82	80

The range of the first test comes from subtracting April's score of 55 from David's 97. The range is 42. On test 2 the range is 45, and on test 3, it is 37.

The range is a rather crude measure of variability of data, but it is nevertheless rather an important one when looking for a graphical representation of the data. We have see how the interaction between the scale used in a chart and the actual range of the data in the chart can change the visual implications of a chart. Chart scales that are close to the range tend to emphasize differences in the data, while larger scales have the opposite effect.

1.2 The Variance

The next possible measure of variability in data begins with a failure of sorts. A reasonable first guess might be to find the average distance between data points

and the center, say as measured by the mean. For the first test in our class, the ,mean was 79.29. Using this, we would compute the various distances from that mean.

Name	Test 1	Distance
April	55	−23.29
Barry	63	−15.29
Cindy	88	9.71
David	97	18.71
Eileen	58	−20.29
Frank	90	11.71
Gena	88	9.71
Harry	71	−7.29
Ivy	65	−13.29
Jacob	77	−1.29
Keri	75	−3.29
Larry	88	9.71
Mary	95	16.71
Norm	86	7.71
Average	79.29	0.00

However, that is exactly what we were expecting. We have already seen that the best property that the mean has going for it is that the average distance from the average will always be 0.

One idea for fixing this is to exaggerate the distance from the center. We could try doubling it, but that will not work because it exaggerates all the distances uniformly. We need to penalize data for being further away from the center. We do this by squaring the distance. That way a distance of 1 is left alone, but a distance of 2 gets boosted to 4. And a distance of 5 gets counted as a whopping 25. The average square distance is called the variance. In our example,

Name	Test 1	Distance	Distance^2
April	55	-23.29	542.22
Barry	63	-15.29	233.65
Cindy	88	9.71	94.37
David	97	18.71	350.22
Eileen	58	-20.29	411.51
Frank	90	11.71	137.22
Gena	88	9.71	94.37
Harry	71	-7.29	53.08
Ivy	65	-13.29	176.51
Jacob	77	-1.29	1.65
Keri	75	-3.29	10.8
Larry	88	9.71	94.37
Mary	95	16.71	279.37
Norm	86	7.71	59.51
Average	79.29	0.00	181.35

The variance is a bit strange, but it is a good measure of variation. It has wonderful mathematical properties that allow mathematicians and statisticians to study it in great detail. Still it does seem a bit odd. One reason for this is the units. The distances from the mean in the example above are measured in points. When we square these digits, the units are squared as well. That means that the variance is

$$181.35 \text{ points}^2.$$

Squared points is not the most natural unit of anything except variance. For now, we will try to live with it; later it will become far less of a problem.

So what exactly is the variance of a set of data? Numerically we have said it is the average squared distance from the mean. Algebraically this is just as easy to see, although perhaps a little frightening. Suppose our data is

$$d_1, d_2, d_3, ...d_{n-1}, d_n.$$

The average of this is

$$a = \frac{d_1 + d_2 + d_3 + ...d_{n-1} + d_n}{n}.$$

The distances from the mean are

$$(d_1 - a), (d_2 - a), (d_3 - a), ... (d_{n-1} - a), (d_n. - a).$$

The square distances are

$$(d_1 - a)^2, (d_2 - a)^2, (d_3 - a)^2, ... (d_{n-1} - a)^2, (d_n. - a)^2.$$

So the variance must be

$$v = \frac{(d_1 - a)^2 + (d_2 - a)^2 + (d_3 - a)^2 + \ldots + (d_n. - a)^2}{n}.$$

However, we can take this a bit further.

$$\begin{aligned} v &= \frac{(d_1-a)^2+(d_2-a)^2+(d_3-a)^2+\ldots+(d_n.-a)^2}{n} \\ &= \frac{\left(d_1^2-2d_1a+a^2\right)+\left(d_2^2-2d_2a+a^2\right)+\ldots+\left(d_n^2-2d_na+a^2\right)}{n} \\ &= \frac{\left(d_1^2+d_2^2,+\ldots d_n^2\right)-(2d_1a+2d_2a+\ldots+2d_na)+\left(a^2+a^2+\ldots a^2\right)}{n} \\ &= \frac{\left(d_1^2+d_2^2,+\ldots d_n^2\right)-2a\left(d_1+d_2+\ldots+d_n\right)+n\cdot a^2}{n} \\ &= \frac{\left(d_1^2+d_2^2,+\ldots d_n^2\right)}{n}-\frac{2a\left(d_1+d_2+\ldots+d_n\right)}{n}+\frac{na^2}{n} \\ &= \frac{\left(d_1^2+d_2^2,+\ldots d_n^2\right)}{n}-2a\frac{(d_1+d_2+\ldots+d_n)}{n}+a^2. \end{aligned}$$

But notice that

$$\frac{d_1+d_2+...+d_n}{n}$$

appears in this last statement, and it is just the average. So we have

$$\begin{aligned} v &= \frac{\left(d_1^2+d_2^2,+\ldots d_n^2\right)}{n}-2a\frac{(d_1+d_2+\ldots+d_n)}{n} \\ &= \frac{\left(d_1^2+d_2^2,+\ldots d_n^2\right)}{n}-2a\cdot a+a^2 \\ &= \frac{\left(d_1^2+d_2^2,+\ldots d_n^2\right)}{n}-a^2 \\ &= \frac{\left(d_1^2+d_2^2,+\ldots d_n^2\right)}{n}-\left(\frac{d_1+d_2+\ldots+d_n}{n}\right)^2. \end{aligned}$$

The only reason we did this algebra is that, very often, this is the definition of variance one finds in math books or computer programs. It looks a lot different than "the average square distance from the mean," but that is just what it is. Notice the two parts of this formula:

$$\frac{\left(d_1^2+d_2^2,+\ldots d_n^2\right)}{n}$$

is the average of the squares of the data. Now

$$\left(\frac{d_1+d_2+\ldots+d_n}{n}\right)^2$$

is the square of the average of the data. Thus we have two algebraically equivalent ways of describing the variance:

- The variance is the average squared distance from the mean.

- The variance is the mean of the squares minus the square of the mean.

The first description illustrates the reason it measures variation from the center in squared points. The second description gives a formula that makes the variance easier to compute.

1.3 The Standard Deviation

The biggest problem with the variance, until you get used to it, is that it is measured in square units. In our test data, the variance in on first test is 181.35 points^2. If we want to bring these units back to normal, we can take the square root. In this case

$$\sqrt{181.35 \text{ points}^2} \simeq 13.47 \text{ points}.$$

The square root of the variance is the standard deviation.

Thus on test 1 of our example, the standard deviation is 13.47 points. On test 2 the variance is 160.27 points^2; so that makes its standard deviation 12.66 points. On test 3 the variance is 135.37 points^2; so that makes its standard deviation 11.63 points. It looks like the class grades are becoming less varied through the three tests.

The standard deviation is the most common measure of variation in data. The variance has better mathematical properties than the standard deviation, but they are so closely related that it hardly matters. What makes the standard deviation preferred is that it is measured in the natural units of the data.

As the name suggests, the standard deviation is also used as a measure in its own right. The standard deviation works as a good unit of measure when comparing the relative position of a datum within a set.

Consider the grades on test 1 above, and distances of those grades from the mean:

Name	Test 1	Distance
April	55	−23.29
Barry	63	−15.29
Cindy	88	9.71
David	97	18.71
Eileen	58	−20.29
Frank	90	11.71
Gena	88	9.71
Harry	71	−7.29
Ivy	65	−13.29
Jacob	77	−1.29
Keri	75	−3.29
Larry	88	9.71
Mary	95	16.71
Norm	86	7.71

Frank had a score of 90% . If the purpose of the test was to measure Frank's knowledge of the material covered out of a theoretical 100%, then Frank's grade was quite good. Learning 90% of the material is quite an accomplishment. Frank's performance should be judged solely on the fact that he got 90% out of 100%. If the only point is to learn the material, Frank has a good claim to have done that.

But Frank had another accomplishment of which he can be proud. Frank's 90% was the third highest grade in the class. In a competition between students, this is the important thing. If the point is to learn the material, all that matters is the grade. If the point is to outscore as many people in the class as possible, the ranking of your score is important:

Name	Test 1	Ranking	Distance
April	55	14	−23.29
Barry	63	12	−15.29
Cindy	88	4 tie	9.71
David	97	1	18.71
Eileen	58	13	−20.29
Frank	90	3	11.71
Gena	88	4 tie	9.71
Harry	71	11	−7.29
Ivy	65	10	−13.29
Jacob	77	8	−1.29
Keri	75	9	−3.29
Larry	88	4 tie	9.71
Mary	95	2	16.71
Norm	86	7	7.71

Another way to compare Frank to the rest of the class is to notice that he scored almost 12 points above the class average. That means that, in a race to the highest total score at the end of the course, he has a 12 point lead over a lot of students in the class. If the point is to establish a lead over as many people in the class as possible, the distance from the mean is the important measure.

But has Frank's achievement really distinguished him as better than the rest of the class; is a 90% an extraordinary score on this test relative to the results in the class. Here is where using a measure of standard deviations can be very useful. Frank scored 11.71 points above the mean on a test with a standard deviation of 13.47. Measured in a different unit, this is $\frac{11.71}{13.47} = 0.86934$ standard deviations above the mean. Notice that we are using "standard deviations" as a unit of standard measure. We are comparing Frank to the rest of the class using a more objective measure than the number of points. In general, a distance of 1 standard deviation or less is not consider particularly special. So Frank still did quite well, but so far, nothing of extra note compared to others in the class. If the point is to see how remarkable a test score is objectively, the distance from the mean in standard deviations is the important measure.

Look at April. Clearly April did poorly. If the purpose is to learn the

material, then April has a way to go. She had the lowest grade in the class, and so is far from the top in that competition. If she hopes to catch up, her distance from the mean of -23.29 is quite telling. However, how bad was her performance on this test? After 55% is more than half. In standard deviations, April's score was $\frac{23.29}{13.47} = 1.729$ below the mean. This is almost 2 standard deviations below the average. Two standard deviations is definitely quite a bit off, and a teacher who understands this way of measuring a student's place relative to the rest of the class will definitely be alarmed. April is definitely not learning the material as well as the other students. Certainly the fact that she is 23 points below the average shows this. The importance of the value 23, however, depends on the test, the way it was graded, the scale used, and even the number of students in the class. However in a more objective measure, she is 1.7 standard deviations below the mean. In any class of any size and under any grading scheme, this is very low.

We can measure the standings of all the students in the class in standard deviations:

Name	Test 1	Ranking	Pts Distance	S.D. Distance
April	55	14	-23.29	-1.73
Barry	63	12	-15.29	-1.14
Cindy	88	4 tie	9.71	0.72
David	97	1	18.71	1.39
Eileen	58	13	-20.29	-1.51
Frank	90	3	11.71	0.87
Gena	88	4 tie	9.71	0.72
Harry	71	11	-7.29	-0.54
Ivy	65	10	-13.29	-0.99
Jacob	77	8	-1.29	-0.1
Keri	75	9	-3.29	-0.24
Larry	88	4 tie	9.71	0.72
Mary	95	2	16.71	1.24
Norm	86	7	7.71	0.57

We always have a choice between measuring distance from the mean in original units or in standard deviations. In general, keeping the original units is best when making comparisons within the data set; while using standard deviations works best when comparing different data sets. We will say more about this later.

So while standard deviation is, on the one hand, a single measure of the variation of a collection of data, it can also be used as a unit to measure the position of an individual datum within the data set.

1.4 Quartiles

Now the variance and the standard deviation are measures of variation that treat the mean as the center of the data. This means that they are good

measures of variation when the mean is a good measure of the center. We have seen, however, that this is not always the case. There are data sets where the median is a better measure of the center. In these cases there are alternate measures of the variation as well.

The median is the half way point in the data of the set. To compute the median of a data set, we need to rank the data in order. Using our familiar test 1:

Name	Test 1	Ranking
April	55	12
Barry	63	10
Cindy	88	4 tie
David	97	1
Eileen	58	11
Frank	90	3
Gena	88	4 tie
Harry	71	9
Ivy	65	8
Jacob	77	6
Keri	75	7
Larry	88	4 tie
Mary	95	2
Norm	86	5

It would be best to rearrange this data in the order of rank:

Name	Test 1	Ranking
David	97	1
Mary	95	2
Frank	90	3
Cindy	88	4 tie
Gena	88	4 tie
Larry	88	4 tie
Norm	86	7
Jacob	77	8
Keri	75	9
Ivy	65	10
Harry	71	11
Barry	63	12
Eileen	58	13
April	55	14

There are $12 = 2 \circ 6$ scores, so the median is the average of the 6-th and 7-th scores: $\frac{88+86}{2} = 87$.

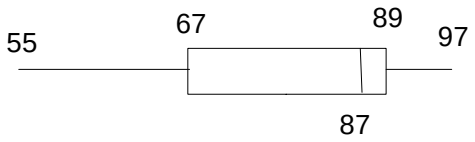
The median is the score where half the class is above that score and half the class is below it. The median divides the class in equal halves. If we divide each of those halves into their own equal halves , we get quartiles. There are

$6 = 2 \circ 3$ scores in each half, so the quartile breaks half way between the 3-rd and 4-th grade in each half. That puts the top quartile at $\frac{90+88}{2} = 89$, and the bottom quartile is put at $\frac{71+63}{2} = 67$.

The results of test one can be summarized by the following statistics:

- Minimum: 55
- Bottom Quartile: 67
- Median: 87
- Top Quartile: 89
- Maximum: 97.

There is a nice diagram that can be used to display this summary called a "Box and Whiskers" graph. The diagram is drawn on a number line. The box in the diagram is a rectangle where the left side is labeled with the bottom quartile and the right side labeled with top quartile. The median is located in the box properly placed in the interval. Some people just note its location with a *, but a line parallel to the sides is more common. Protruding from the two sides of the box are lines extending to the minimum and the maximum of the data; So our test would be summarized as



Notice how the box and whisker diagram clearly shows that most of the class did quite well. Even though the low grades did not really standout as outliers, the median grade of 87 is a better indicator of the class performance than the mean 79. The box and whisker diagram gives a pretty good idea of how the grades came out.

1.5 Percentile

The final way to measure variation we will consider is percentiles. This is not a single measure of the variation like variance or standard deviation. Rather it is a measure of the variation individual data. Computing the percentiles of each data point begins by placing the data in ranked order, with the "best" results

placed highest. In our test

Name	Test 1	Ranking
David	97	1
Mary	95	2
Frank	90	3
Cindy	88	4 tie
Gena	88	4 tie
Larry	88	4 tie
Norm	86	7
Jacob	77	8
Keri	75	9
Ivy	65	10
Harry	71	11
Barry	63	12
Eileen	58	13
April	55	14

The percentile of each score is the percentage of the rank in the total number of students. For David, this is $\frac{1}{14} \cdot 100 = 7\%$. We would say that David scored in the 7-th percentile of the class. Cindy, Gena and Larry tied for 4-th; so they would all have the same percentile: $\frac{4}{14} \cdot 100 = 29\%$. That puts them all in the 29-th percentile.

Actually this is not a good example of how percentile ratings are used. Typically percentiles are used in very large ranked data sets. Standardized test often report grades as both raw scores and percentiles. If you know the scale of a test, then a raw score of 129 gives you information. However, knowing that that score puts a student in the 85-th percentile, gives you a better idea of the meaning of the score without any additional information about the test. The student with that score is in the top quartile and

Thus percentile measurement is the approach used to measure the distance from the center as measured be the median, which is the 50-th percentile. This corresponds to measuring distance of a datum from the mean in standard deviations.

Prepared by: Daniel Madden and Alyssa Keri: May 2009