Data 8
Spring 2016

Mean, Center, and Spread

Discussion 6

In lecture, we have seen the important uses of estimation, center, and spread. Today's discussion worksheet is focused primarily on a small recap of them all.

Estimation

Estimation is the idea of trying to estimate an unknown value using data. In lecture, we saw the example of estimating the number of warplanes based on a (possibly small) sample. In many situations, we can estimate the same quantity in multiple ways. For example, we could take the average of our sample and multiply it by 2 as one estimate for the total number of warplanes, but we could also use the maximum element in a sample.

Often, we're interested in seeing how an estimate varies with different random samples. If we have access to the full data, we can simulate many random samples, take an estimate from each, and examine a histogram of those estimates. We sometimes call this a "sampling distribution."

In terms of coding, we often accomplish this using *iteration*, using a `for` loop in order to make many estimates and create a distribution.

Center

Center, average, or mean all refer to essentially the same value. One way to calculate it is to take the sum of all elements, and then divide by the amount of elements present, a method you might be familiar with. Another way to calculate our mean is to multiply each value by its proportion in the set, and then summing each of these values up. This calculation might be helpful particularly in the case of histograms.

For example, if we had a list $[3,3,4,4,4,4,5,6]$. We could add up all of the elements in the list and divide by 8, the length of the list, but it might be easier to see that we have a 0.25 proportion of 3s, 0.5 proportion of 4s, .125 proportion of 5s, and a .125 proportions of 6s. Hence, our center can be calculated by: $3*.25+4*.5+5*.125+6*.125$, which is 4.125. Notice, also, that this element is not in our original list.

In this way, the mean is the weighted average for the distinct values, where the weights are proportions. So, the mean is the balance point of a histogram.

So, what is the relationship between the mean and the median? The median is the 50 percent point of the data, which is not necessarily equal to the mean. If the mean is larger than the median, then we will say our data is right-skewed, as there exists a tail, causing the mean to be pulled to the right.

Luckily, we can use the function `np.mean` in order to take the mean of an array or of a column of a table easily for us.

Spread

Now, we have seen why the center of a distribution is important, but it's also important to note how far away, on average, elements are from the center. So, we first look at the deviations of each of the elements from the average. This causes some deviations to be negative, so we should square these, and taking the average of these gets us the *variance*, the average squared deviation from the mean. But, depending on what units we are working with, we now have a units squared, so after all of that we need to take the square root. This calculation gets us the *standard deviation*, an important tool in exploring the spread of a distribution. Hence, the SD is the root mean square of deviations from average.

In code, `np.std` does this long computation for us on arrays.

So, why do we use SDs? Because, a bulk of our entries are guaranteed to be within a positive or negative distance of a few SDs from our mean. What's a few? Usually, it is no more than 5. This leads us to an im-

portant equation called Chebyschev's inequality, which says that for some number z, the amount of entries that within z * SD of the mean is **at least** $1 - \frac{1}{z^2}$. Notice the phrase "at least", which says that the proportion can be more, but never less.

Another useful tool is standard units, which are often denoted z. In general, a standard unit is $\frac{\text{value} - \text{mean}}{\text{standard deviation}}$, with a standard unit of 0 meaning that our value is equal to our mean. The farther away we get from the mean, the farther our z gets from 0, in either the positive of negative direction.

Practice Problems

1. Here is a list of numbers: [0.7, 1.6, 9.8, 3.2, 5.4, 0.8, 7.7, 6.3, 2.2, 4.1, 8.1, 6.5, 3.7, 0.6, 9.9, 8.8, 3.1, 5.7, 9.1]. Using it, answer the following questions

   (a) Without doing any math, guess whether the average is around 1, 5, or 10.

   (b) Without doing any math, guess whether the SD is around 1, 3, or 6.

   Answers:

   (a) We should expect the average to be around 5. This is because there are minimal values less than 1, or even close to 1, and there are no values bigger than 10.

   (b) We should expect the SD to be around 3, as 1 doesn't cover much of the data in each step, and implies that it's all very close to the center, while 6 implies a very large spread, so much so that 1 standard deviation encompasses the whole list.

2. Suppose we have two lists, one is [5, 4, 6, 3, 7, 2, 7], and the other one is [5, 4, 6, 3, 7, 2, 7, 5, 5, 5, 5]. Which of the lists has a smaller standard deviation? Explain your answer, but do not do any calculations.
   Answers:
   We should expect the second one to have a smaller Standard Deviation. This is because the average of both lists is around 5, but in the second list, we have many more 5s, so that the spread becomes smaller. Hence, the standard deviation is smaller.

3. Use the three histograms below for the following questions.

   (a) Match each histogram with a unique possible average: 40, 50, 60.

   (b) Match each histogram with a description: The median is greater than the average, The median is less than the average, The median is equal to the average

   (c) Is the standard deviation of histogram 3 closer to 5, 15, or 50?

   (d) True or false: Histogram 1 has a standard deviation much smaller than that of histogram 3. Explain.

   Answers:

   (a) i: 60, ii: 50, iii: 40. We can tell that the mass of the histogram is on the right for i, so we can expect that the average should also be near the right side, bigger than 50. Similar logic for iii, as the left side, as it's clumped on the left side of 40.

(b) The first histogram is left skewed, meaning that there is a tail on the left side of the histogram pulling the mean away from the median, so that the median is larger than the mean. Similarly, for 3, the histogram is right skewed so the mean is pulled towards the tail, causing the median to be less than the mean. For the last one, there is no skew, so the median and mean are relatively equal.

(c) We need to look at the relative spread of histogram 3. Let's assume the SD is close to 50, but then, in one standard deviation, we have covered the whole graph and more, which is too much. So, 50 is too big. Let's try out 5. Then, it would take around 12 standard deviations in order to finally encompass the whole data, which is way too much. So, 15 sounds like a more reasonable SD, as that is about 3 or 4 standard deviations to cover everything.

(d) False. Notice that the histograms are mirror images of each other. Hence, their spread/distribution is the same, the only difference being the center and if the deviations are positive or negative. But, remember, when calculating standard deviation, we square the deviations so they're roughly equal in the end.

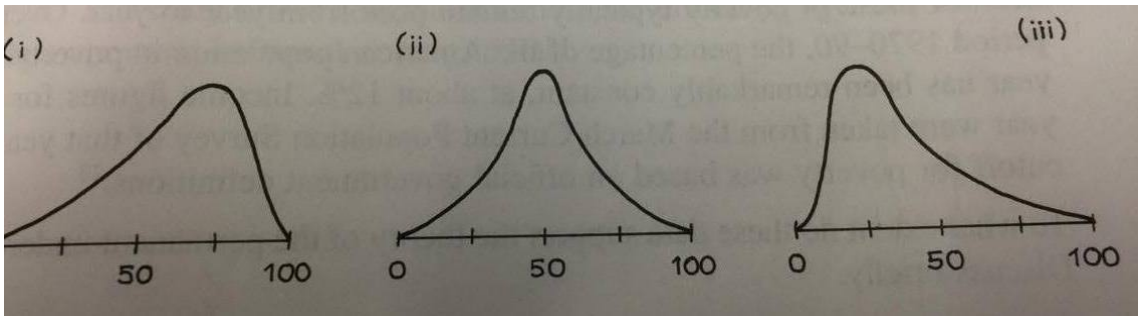

Figure 1: Histograms for question 3

4. A study on college students found that men had an average weight of 66 kg and a standard deviation of 9 kg, while the women had an average weight of about 55 kg and a standard deviation of 9 kg. If you took the men and women together, would the standard deviation of their weights be smaller, equal to, or bigger than 9 kg? Why?
   Answers:
   Bigger than. Notice that the two averages are different, so we should expect that when we add the two datasets together, they will compete for the average and the spread will increase as we have more data that are separate from each other. Hence, the SD should get bigger.

5. Assume we are given the following table:

Color	Shape	Amount	Price
Red	Round	4	1.30
Green	Rectangular	6	1.20
Blue	Rectangular	12	2.0
Red	Round	7	1.75
Green	Rectangular	9	1.40
Green	Round	2	1.0

   Use the table, named `marbles`, from discussion 4 to compare the distribution of prices for round marble packages versus rectangular marble packages. Use code in order to get the necessary aspects of our distribution.

Answers:
Any answeres are acceptable, but some important functions to use are `np.mean`, `np.std`, and `marbles.where(marbles.column('Shape') == 'Round').column('Price')`. Your analysis should include something like this somewhere in it, along with variations in orer to get the Rectangular packages.