Homework 1

Sahana Sarangi

1 April 2024

Problem 1: Lauren and Tasneem are classmates. They had a midterm exam on Friday, but something's gone wrong with the gradebook - they can't see their scores, and the professor isn't answering email! Of course they could just wait until Monday and ask the professor in person, but they've decided to launch an investigation instead. By asking around, looking at the Canvas, and comparing their own answers, they know the following information.

- Tasneem got a higher score than Lauren.
- The class average was 77.5% (counting the missing scores).
- The class sample standard deviation was 30.82% (counting the missing scores).
- Of the seven other people in the class, three of them got 100%, two got 80%, one got 75%, and one didn't show up to the exam and got 0%.

What was Tasneem's score?

Solution: We can let Lauren's score be represented by l and Tasneem's score be represented by t. We are given the scores of everyone in the class (except Lauren's and Tasneem's), so we can set up an equation using these scores (and the variables l and t) to find the average of all the scores in the class (including Lauren's and Tasneem's). The average of a dataset is the sum of all the elements divided by the number of elements in the dataset, so the average of the scores in the class would be $(100+100+80+80+75+0+l+t)$ divided by 9. We are given that the average is 77.5, so we can write the equation

$$77.5 = \frac{100+100+100+80+80+75+0+l+t}{9}$$

Simplifying this, we have $162.5 = l+t$. Next, we can find the sample standard deviation of all the scores in the class. The sample standard deviation of a dataset is found by finding the sums of the squares of the differences of each element and the mean of the dataset, dividing the sum by one less than the length of the dataset, and then square rooting the quotient. The standard deviation of all the scores is then

$$\sqrt{\frac{22.5^2+22.5^2+22.5^2+2.5^2+2.5^2+(-2.5)^2+(-77.5)^2+(l-77.5)^2+(t-77.5)^2}{8}}$$

We are given that the standard deviation of all the scores (including Lauren and Tasneem) was 30.82. Simplifying our expression for the standard deviation and then setting it equal to 30.82, we have

$$30.82 = \sqrt{\frac{7537.5+(l-77.5)^2+(t-77.5)^2}{8}}$$

Simplifying this:

$$61.4792 = (l-77.5)^2+(t-77.5)^2$$

Now that we have two equations in terms of l and t, we can solve for l and t. Solving for l in our first equation, we have $l = 162.5 - t$. Substituting this for l in our second equation, we have $61.4792 = (162.5-t-77.5)^2+(t-77.5)^2$. Solving for t, we get

$$t = \frac{\sqrt{166771}}{100} + \frac{325}{4} \approx 85.33,\ t = -\frac{\sqrt{166771}}{100} + \frac{325}{4} \approx 77.17$$

Because we have two possible values for Tasneem's score, we need to find what Lauren's score would be in both scenarios of Tasneem's scores. If Tasneem's score was $\frac{\sqrt{166771}}{100} + \frac{325}{4}$, Lauren's score (using our first equation) would be $162.5 - \frac{\sqrt{166771}}{100} + \frac{325}{4} \approx 77.17$. If Tasneem's score was $-\frac{\sqrt{166771}}{100} + \frac{325}{4}$, Lauren's score would be $162.5 + \frac{\sqrt{166771}}{100} - \frac{325}{4} \approx 85.33$. We are given that Tasneem got a higher score than Lauren. When Tasneem scored approximately 85.33%, she earned a higher score than Lauren, who earned 77.17%. When Tasneem scored approximately 77.17%, she earned a lower score than Lauren, who earned 85.33%. Hence, Tasneem must have scored $\boxed{\text{approximately } 85.33\%, \text{ or } \left(\frac{\sqrt{166771}}{100} + \frac{325}{4}\right)\%.}$

Problem 2: Derek is curious about eye colors, so he conducts a survey in his home town. He surveys 1000 people, and notes each person's eye color; every eye was either brown, blue, or green. But he wants to know what the "average" eye color is, so he translates each eye color into a number (0 for brown, 1 for blue, 2 for green) and takes the average of this number. He finds that the average is 0.866. Based on this information, Derek concludes that the "average" eye color in his home town is mostly blue, with a little bit of brown (because 0.866 is close to 1, but leaning a little towards 0). Does Derek's reasoning make sense? Why or why not?

Solution: Derek's reasoning does not make sense. While his calculation of the average is correct, his interpretation of the average is incorrect. The numerical values that he's assigned to different eye colors is linear, but eye colors aren't. The numbers 0, 1, and 2 are all arbitrary numerical values that he has assigned, meaning that in this scenario, 1 is not necessarily the midpoint between 0 and 2. If Derek were to have 1 person with green eyes, 1 person with blue eyes, and 1 person with brown eyes, he would have a dataset with the numbers 0, 1, and 2. If he were to take the average of these numbers (which is 1), it would suggest that the average eye color among the three people is blue. This is because Derek's system assumes that blue is the intermediate eye color between brown and green and that eye color can be measured on a linear scale from brown to green. This is not a correct assumption. Hence, the average eye color in his hometown is not really blue with a little bit of brown.

Problem 3: Suppose we have a data set $a_1, a_2, \ldots, a_N$ with mean μ_a, and another data set $b_1, b_2, \ldots, b_N$ with mean μ_b.

(a) Let r be a constant. Suppose we construct another data set $c_1, c_2, \ldots, c_N$ by taking $c_i = ra_i$. What will the mean of this data set be?

(b) Suppose we construct another data set $d_1, d_2, \ldots, d_N$ by taking $d_i = a_i + b_i$. What will the mean of this data set be?

Part (a) Solution: The mean of the original array (adding all the elements and dividing by the number of elements) can be expressed as $\frac{1}{N}\sum(a_i)$. The mean of the new array could then be written as $\frac{1}{N}\sum(ra_i)$. Using the sigma rule $\sum(ca_k) = c\sum(a_k)$, the mean of the new array can be written as $\frac{r}{N}\sum(a_i)$. This expression is equal to the expression for the mean of the original array multiplied by r. Hence, if the mean of the original array is μ_a, the mean of the new array is $\boxed{r\mu_a.}$

Part (b) Solution: The mean of the new dataset can be written as $\frac{1}{N}\sum(a_i + b_i)$. Using the sigma rule $\sum(a_k + b_k) = \sum(a_k) + \sum(b_k)$, we can rewrite this as $\frac{1}{N}\sum(a_i) + \frac{1}{N}\sum(b_i)$. This is the same as the mean of the dataset with as plus the mean of the dataset with bs. Their means are μ_a and μ_b respectively, so the mean of the new dsataset is $\boxed{\mu_a + \mu_b.}$

Problem 4: Suppose we have a data set $a_1, a_2, \ldots, a_N$ with standard deviation σ_a, and another data set $b_1, b_2, \ldots, b_N$ with standard deviation σ_b.

(a) Let r be a constant. Suppose we construct another data set $c_1, c_2, \ldots, c_N$ by taking $c_i = ra_i$. What will the standard deviation of this data set be?

(b) Suppose we construct another data set $d_1, d_2, \ldots, d_N$ by taking $d_i = a_i + b_i$. Can we find the standard deviation of this data set?

Part (a) Solution: The sample standard deviation of the dataset a_i (using the standard deviation formula) is

$$\sigma = \sqrt{\frac{\Sigma(a_i - \mu)^2}{N-1}}$$

where a_i is the ith value in the dataset, μ is the mean of the dataset, and N is the total amount of elements in the dataset. If we were to create a new dataset by multiplying each value of the original dataset by r, we would substitute a_i for ra_i and μ for $r\mu$ (as we found in part (a) of Problem 3, the mean will also get multiplied by r):

$$\sigma = \sqrt{\frac{\Sigma(ra_i - r\mu)^2}{N-1}}$$

To simplify this, we can first factor out r^2 in the numerator:

$$\sigma = \sqrt{\frac{r^2\left(\Sigma(a_i - \mu)^2\right)}{N-1}} = r\sqrt{\frac{\Sigma(a_i - \mu)^2}{N-1}}$$

Our resulting expression is the same as the expression for the standard deviation of the dataset a_i, but multiplied by r.Hence, the standard deviation of the new dataset is $\boxed{r\sigma_a.}$

Part (b) Solution: To answer this question, we can consider the dataset a_i to be $[2, 6]$, which has a population standard deviation of 2. We can consider the dataset b_i to be $[1, 3]$, which has a population standard deviation of 1. If we were to construct a new dataset d_i where $d_i = a_i + b_i$, it would be $[3, 9]$, which has a population standard deviation of 3. Now, we can pick another dataset b_i with population standard deviation 1: $[3, 1]$. If we were to construct d_i using this new dataset b_i, it would be $[5, 7]$, which has a population standard deviation of 1. Because we got different standard deviations for d_i even when we kept the standard deviation of b_i the same, we know that the actual position that each value is in within b_i is what matters. Because we don't know what position each value is in, we $\boxed{\text{cannot find the standard deviation of the dataset } d_i.}$