Mean and standard deviation

135. Both mean and standard deviation are statistics used for a single quantitative variable.

136. Thus, we can take the mean and standard deviation of a set of quantitative observations like IQ, shoe size, height, weight, etc., but not a set of categorical observations like gender, party affiliation, etc.

137. Test scores are quantitative. Let's say we have the following test scores (and everyone did really well): 90, 92, 94, 96, 98, 100. We want to find the mean and the standard deviation.

138. Most students are familiar with the mean:

$$\bar{x} = \frac{90+92+94+96+98+100}{6} = 95.$$

139. Now we give each data point a number, called an index:

$$\begin{aligned} x_1 &= 90 \\ x_2 &= 92 \\ &\vdots \\ x_6 &= 100. \end{aligned}$$

140. If we want to refer to an *arbitrary* data point we use the letter i. In other words x_i is the i^{th} data point. Here i stands for a number, either 1, 2, 3, 4, 5, or 6. The subscript i is called the *index*.

141. Finally, if we want to refer to the *total* number of data points (in this case 6) we use the letter n. This use of n is common in statistics.

142. We use the sigma notation to write the formula for the mean:

$$\bar{x} = \frac{1}{n}\sum x_i.$$

143. The symbol Σ is the capital form of the Greek letter *sigma*. It stands for *sum*.

144. Other branches of mathematics require *limits* on the sum, such as

$$\sum_{i=1}^{6} x_i$$

This notation means to sum the data points x_i for values of the index i ranging from 1 to 6.

145. Statisticians often leave the limits off the sum. In this case, it is implied to sum over all of the data: sum from i ranging from 1 to n, which is the same thing as above.

146. Finally the coefficient $\frac{1}{n}$ in front of the Σ tells us to divide the sum by the total number of data points, n, in this case 6, as above.

147. The mean is a measure of the center of the distribution.

148. The standard deviation is a measure of the spread of the distribution—in other words, how close, or how far, do the data tend to fall from the mean?

149. I'll start with a formula, explained below. Confusing: there are actually two formulas for standard deviation, and many calculators give you a choice.

$$s_n = \sqrt{\frac{1}{n}\sum(x_i - \bar{x})^2}$$

$$s_{n-1} = \sqrt{\frac{1}{n-1}\sum(x_i - \bar{x})^2}$$

150. In the context of *sampling* (discussed below), the second formula is correct. Indeed, if there is only one choice given in a book or a calculator, it is usually the second one. I will call the second formula the *more common formula*, and the first formula the *less common formula*.

151. Which formula should you use? Short answer: always use the more common formula. Use of the less common formula should be noted and justified, and I would say just don't bother!

152. Why are we discussing the less common formula? Because most students have a hard time understanding the more common formula, and think the less common formula makes more sense. It *does* make more sense in certain contexts. And I think it is important to discuss these contexts to better understand the more common formula.

153. Let's unpack the formulas.

154. The quantity $(x_i - \bar{x})$ is called the deviations, or deviations from the mean.

155. Deviations are important because we are trying to estimate how far the data points fall from the mean.

156. Each data point has a corresponding deviation from the mean.

157. Consider the same test score example. The first observation, 90, is 5 points *below* the mean (which was 95). So its deviation from the mean is -5.

158. Can you guess what the other deviations from the mean are?

159. The other deviations from the mean are: -5, -3, -1, 1, 3, 5.

160. Because the data points are equally spaced, the deviations have a nice pattern to them. This pattern will not be there in most data sets.

161. However, it will always be the case that the deviations add to zero.

162. Unless all deviations are actually zero, some will be positive and others will be negative, in such a way that they will balance out, adding to zero—this property results from the fact that the deviations are from the mean and the mean is the center of the distribution.

163. Because it is useless to average the deviations (the average will always be zero), we first square the deviations: $(x_i - \bar{x})^2$. For our data set the squared deviations are: 25, 9, 1, 1, 9, 25.

164. Unless a deviation is zero, its square is positive.

165. The next step is to "average" the squares of the deviations. This number will be positive unless all of the deviations are zero.

166. The less common formula for s_n uses the mean of the deviation as the average: $\frac{70}{6} = 11.6667$.

167. The more common formula for s_{n-1} uses an adjusted mean—adjusted for the so-called number of degrees of freedom or $n - 1$: $\frac{70}{5} = 14$. The adjusted mean is what is used as the average.

168. The last step, in both formulas, is to take the square root of the result: either $\sqrt{\frac{70}{6}} = 3.4157$ or $\sqrt{\frac{70}{5}} = 3.7416$. Because we square the deviations in a previous step, we take the square root, so that the result can more easily be compared with the mean (without taking the square-root the units change).

169. The more common formula for s_{n-1} always gives a larger value than the less common formula for s_n.

170. The larger the value of n, the less difference there is between the results given by the two formulas. The difference between dividing by 6 or dividing by 5 is much greater than the difference between dividing by 1000 or dividing by 999.

171. If you skip the square root step you are left with a quantity that is also important in statistics. It is called the *variance*. The variance has different units than the data.

172. As mentioned above, in the context of sampling, we should use the more common formula for s_{n-1}. In this context, $\bar{x}$ is called the sample mean, and s_{n-1}, also written as s, is called the sample standard deviation.

173. The more common formula arises in the context of sampling.

174. What is sampling? Let's proceed with our example. Let's say you have many students in the class—a large lecture with 1000 students. But as above, you randomly pick just 6 students to assess the performance of the whole class. The 6 students comprise your *sample*.

175. A good assessment of the class average requires that the sample be *random*. Later, we will talk about what this stipulation means, why it leads to good estimates, and precisely, how good we can expect our estimates to be (they improve with the size of your sample; 6 is actually pretty small).

176. Because you don't have time to work with a thousand students, you make due with 6: you calculate the sample mean, as above, 95.

177. The whole class is called the *population* and the mean test score for the whole class is called the *population mean*.

178. The population mean is unequivocally the correct answer to the question: what is the mean test score for the class? But it requires a calculation with all 1000 test scores to figure it out.

179. Because you don't want to, or can't for some reason, deal with all 1000 students, so use the *sample mean*, as your best answer to the same question, as your estimate of the elusive population mean.

180. Using the sample of the six students shown above, your estimate of the population mean is 95.

181. You know that your estimate could be too high or too low depending on the sample you chose. Without examining all 1000 test scores, you don't know the right answer—you are stuck with uncertainty.

182. Now, in addition to estimating the population mean to assess center of the distribution, you may want to estimate the population standard deviation to assess the spread in the distribution—things get complicated.

183. The unequivocal right answer to the population standard deviation uses the *less common formula!*, summing over all 1000 students and using the unadjusted average and substituting the population mean for $\bar{x}$.

184. The question is: what is an appropriate estimate of the population standard deviation using our sample of 6 students, rather than all 1000?

185. Which formula you should use depends on what you use for $\bar{x}$.

186. If you use the population mean for $\bar{x}$, as above, you would use the less common formula, employing the usual unadjusted average. This is almost never done for lack of access to the population mean.

187. If you use the sample mean for $\bar{x}$ (after all, you want to avoid dealing with all 1000 students) you will get an answer with is prone to being too small, unless you correct it by changing the notion of average.

188. To fix this problem, you use the adjusted mean, which appropriately increases your estimate, so that on average, you get a result which is neither prone to being too high, nor too low.

189. The question is: why does the less common formula lead to an estimate which is prone to being too low?

190. Consider this fact: the correct result involves an average of squared-deviations from the population mean.

191. But now consider this fact: We want to take an average of squared deviations from the *sample mean*, not population mean.

192. The problem is, deviations from the sample mean are prone to being smaller than deviations from the population mean.

193. Why? Consider this example: What if the population mean test score, instead of being 95, was in fact 70. Our sample wouldn't be representative of the population, but that can happen, some times.

194. The deviations from the population mean would be between 20 and 30 whereas the deviations from the sample mean would be between 1 and 5, as shown above. The sample mean deviations would be too small.

195. The example above is extreme, but any time the sample mean is different from the population mean, we have a problem, because the sample mean is a better estimate for just the sample (it was derived from the sample) than for the whole population involving all the data.

196. The sample is closer to the sample mean than the population mean, but the population mean gives the right answer. The sample mean's result is too small, but we can correct this by adjusting our notion of average.

197. So why divide by $n - 1$ instead of something else, like $n - 2$. I am not sure if anyone has a satisfying non-mathematical answer to this, although it is clear from a mathematical calculation.

198. It should be pointed out that $n - 1$ is the number of "degrees of freedom" in the deviations.

199. There is one less degree of freedom in the deviations than the total number of deviations because they are constrained to add to zero as mentioned above, so you are really averaging $n - 1$ independent quantities instead of n.

200. Some books justify the more common formula with this argument concerning the degrees of freedom in the deviations, but for me the explanation falls flat and doesn't tell the whole story.