Sampling

148. What is sampling? Let's proceed with our example from the previous chapter: we have the test scores of 6 students: 90, 92, 94, 96, 98, 100. The mean of these test scores is 95.

149. Sampling would be indicated if, in addition to these 6 students, there were many more in the class and we wanted to use the 6 students to study the properties of the whole class.

150. Exactly what properties? Later, we will work with other statistics, but for now our focus rests squarely on the mean: we want to assess the *mean* test score of the whole class by just looking at the 6 students.

151. My example is not really best, because, in practice, the teacher would probably have all grades for the all students (even if there were 1000 students). Often the grades are in BlackBoard or in a spreadsheet and it takes one command to calculate the class mean.

152. Sampling only gives an approximation to the right answer. In the case of grades, the right answer is readily available, so sampling is not necessary or even advisable.

153. However, in many situations it is impractical to collect data on all imaginable cases.

154. Example: if you are doing a survey, you can't feasibly ask every person in the world. But it is still possible to study the world's population by sampling using a much smaller group.

155. Let's proceed with our grades example ignoring that it is often inappropriate for this application.

156. For our grades example, our sample size was 6. Six is an unusually small sample size. The larger the sample, the better.

157. Let's say the 6 students are among 1000 students in the whole class.

158. The 6 students comprise the *sample*.

159. The 1000 students (which must include the 6) comprise the so-called *population*.

160. We want to study the whole class mean: if we had access to all the grades we could find this number exactly (without sampling) by adding all 1000 grades and dividing by 1000: the usual mean.

161. But let's restrict ourselves to only the 6 students. What could we do?

162. A reasonable approach is to calculate the *mean* of the sample, or sample mean. As shown above, the sample mean is 95. That's our *estimate* of the population (i.e. whole class) mean.

163. The sample size is customarily written with the lower case letter n. In our example, $n = 6$.

164. If the variable in question is x, the sample mean is customarily written with the notation $\bar{x}$:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

165. The sample mean is an example of a *statistic*.

166. A statistic is a number that describes a sample.

167. The population size is customarily written with the upper case letter: N. In our example, $N = 1000$.

168. The population mean is written with the greek letter mu: μ. Alternatively, if the variable in question is x, this is sometimes indicated as μ_x:

$$\mu_x = \frac{1}{N}\sum_{i=1}^{N} x_i$$

169. The population mean is an example of a *parameter*.

170. A parameter is a number that describes a population.

171. Mnemonic: *S*ample and *s*tatistic go together: they both start with s. *P*opulation and *p*arameter also go together: they both start with p.

172. You use $\bar{x}$ as an estimate of μ_x, but your answer depends on your sample.

173. Specifically, if your 6 students all happen to be above average, your estimate will clearly be too high. This scenario is clearly possible.

174. If your 6 students all happen to be below average, your estimate will clearly be too low. This scenario is also clearly possible.

175. If some of your students are above average and others are below, then your estimate might be too high or it might be too low, but it is unlikely (though possible) that it will be *exactly* right.

176. Therefore sampling rarely gives the right answer.

177. But in this case, it gives an *unbiased* estimate, a concept that will be made more precise below.

178. An *unbiased* estimate is one that is neither prone to being too high nor prone to being too low.

179. What does that mean? With 1000 students we can pick our 6-person sample in many ways. In fact, there are exactly 1,368,173,298,991,500 ways of picking a six member sample from a population of 1000. (Huh, you say? I'll show you how to count samples, in the next chapter.)

180. Each sample leads to a different sample mean (although some values may repeat). Some of these values are too high, and some of these values are too low, and maybe a few values are just right.

181. Thus, there are approximately 1.37 quadrillion possible sample means (including repeats).

182. Each sample mean is a mean of 6 values (but different values for each sample).

183. What would happen if I added all 1.37 quadrillion sample means then divided by 1.37 quadrillion?

184. I would get the mean of the sample means!

185. The sense in which the sample mean is an unbiased estimator of the population mean: the mean of all possible sample means equals the population mean!

186. In other words, the mean of all possible estimates is the quantity you are trying to estimate.

187. In this sense, an unbiased estimator is neither prone to being too high, nor prone to being too low.

188. An unbiased estimator is exactly correct on average. Individual estimates will likely be too high or too low, but those errors cancel out when the average is taken.

189. Caution: our result depends on the condition that the sample be chosen at *random*.

190. For example, if the high values are more likely to be chosen than low values, then clearly the estimator will be prone to estimates that are too high.

191. A *simple random sample* is one in which all of the possible samples have an equal chance of being chosen.

192. Our grades example employs a simple random sample only if each of the 1.37 quadrillion samples had an equal (1 out of 1.37 quadrillion) chance of being chosen as the sample that we used.

193. There are other strategies for sampling, which will be discussed in time.

194. However if the professor were to select her favorite 6 students as her sample, she should not expect an accurate assessment of the whole class mean.

195. Let's explore simple random samples:

196. Suppose we want to draw a sample of 2 people from the following population of 4 people: Amy, Betty, Carl, and Dennis, each denoted by his or her initial: A, B, C, and D.

197. There are 6 possible samples: AB, AC, AD, BC, BD, and CD.

198. Each person appears in exactly half the samples. Thus each person has an equal chance of being in the sample.

199. To draw a sample as a simple random sample, we could assign six-sided die face to each of the 6 possible samples, then roll the die to make the selection. Each person would have the same probability of landing in the sample: $1/2$.

200. Some people mistakenly believe that a simple random sample means that each person has a equal probability of being in the sample.

201. Let's explore this scenario. Let's suppose we don't have a die—we only have a coin and we get lazy. We assign AB to heads and CD to tails. Then again each person in the population has the same probability, $1/2$, of being in the sample, but not every sample can be chosen: there are no coed samples possible!

202. Characteristics (such height differences among members of the sample) for which single-sex samples do not fully represent the population would not be well-studied with this sampling scheme.

203. For a simple random sample, in this example, we must make our selection among 6 samples, not 2.

Counting samples

204. How many ways are there to choose a sample of n individuals out of a population of N individuals.

205. This number has a name. It is called, appropriately enough, "N choose n".

206. The following is a mathematical notation for this number:

$$\binom{N}{n}.$$

207. What is the number $\binom{N}{n}$?

208. Calculating this number is based on counting the number of ways of arranging the N individuals in the population into an order.

209. First, how many ways are there of arranging the letters $ABCD$? There are 4 choices for the first letter, 3 for the second, 2 for the third, and 1 for the fourth: $4 \times 3 \times 2 \times 1$.

210. This number is better denoted $4!$, read "four factorial." Basic arithmetic will tell you that $4! = 24$. Likewise there are $N!$ (N factorial) ways of ordering the N individuals in the population.

211. Having enumerated all the ways of ordering the N individuals in the population, how do we pick a sample from the ordering?

212. It doesn't really matter how we pick the sample, so let's just pick one way and be consistent: from each ordering, pick the first n individuals from the ordering as the sample.

213. We have found a way of counting orderings of the population, and picking samples from ordering. Now we try to count the ways of sampling n individuals from the population of N.

214. Unfortunately, counting orderings of the population will over count the number of samples, because we can change the ordering of the population without changing the sample. Indeed, if we just reorder the first n individuals, the sample, as we have picked it, doesn't change.

215. In our example, with sample size 2, every sample of the correct size has two reorderings: for example, we can reorder AB as AB or BA. Note that we count the original ordering AB as one its possible reorderings.

216. So we should divide the number of orderings by at least 2 to get the number of samples—but we are not quite done, yet, because there is a second way of reordering the population without changing the population. In general, every sample of size n will have $n!$ possible orderings, so we should divide $N!$ by at least $n!$ to get the number of samples—but we are not quite done yet.

217. We are not quite done, yet, because, as mentioned, there are actually two ways of reordering of the N individuals in the population without changing the sample: we can reorder the first n chosen as the sample, as done above, *or* we can reorder the last $(N - n)$ left out of the sample.

218. The following 4 orderings all give the same sample AB: $ABCD$, $BACD$, $ABDC$, and $BADC$.

219. Indeed, there are 4 orderings for each of the 6 of the possible samples; so we need to divide 24 by 4 (or divide 24 by 2 twice), which gives 6, as expected. In general, we need to divide $N!$ by $n!$ and then divide again by $(N - n)!$, which gives the following result:

$$\binom{N}{n} = \frac{N!}{n!(N-n)!}$$

220. Another way of calculating $\binom{N}{n}$ is with Pascal's Triangle. Here is Pascal's Triangle:

$N = 0$: 1

$N = 1$: 1 1

$N = 2$: 1 2 1

$N = 3$: 1 3 3 1

$N = 4$: 1 4 6 4 1

$N = 5$: 1 5 10 10 5 1

$N = 6$: 1 6 15 20 15 6 1

$N = 7$: 1 7 21 35 35 21 7 1

$N = 8$: 1 8 28 56 70 56 28 8 1

$N = 9$: 1 9 36 84 126 126 84 36 9 1

$N = 10$: 1 10 45 120 210 252 210 120 45 10 1

⋮

221. The rows are labeled by the population size, N, top to bottom, $N = 0, 1, 2, 3, \ldots$,

222. There is no limit to the number of possible rows in Pascal's triangle, but the first row, the top row, the apex of the triangle, corresponds to $N = 0$, the smallest population size.

223. It is not much of a population, if it has no individuals, but the row for $N = 0$ is there for completeness.

224. The Nth row has $(N + 1)$ entries.

225. These entries correspond to the sample size, n; left-to-right as $n = 0, 1, 2, 3, \ldots N$.

226. The smallest possible sample size is zero: $n = 0$; and the largest possible sample size is the size of the whole population: $n = N$.

227. For both smallest and largest samples, there is only one possible way to draw the sample (respectively, no one in the sample, or everyone in the sample).

228. The entry corresponding to N and n equals $\binom{N}{n}$, the number we want to calculate. Note the entry for $N = 4$ and $n = 2$ is 6 as we expected. (Remember to count n from 0, not 1.)

229. The first and last entries in a row are always 1, for the smallest and largest samples. After that, do you see the pattern? To get any other entry, add the two entries above it: the one to the left, and one to the right.

Problem 1: A deli gives patrons the option of 3 different breads (rye, pumpernickel, and white), 2 different meats (chicken and roast beef) and 8 different toppings (lettuce, tomato, banana peppers, avocado, grated cheese, relish, black olives and garlic). How many ways can you make a sandwich with exactly 1 bread, exactly 1 meat, and exactly 4 different toppings? (One possible sandwich that meets the criteria is roast beef on rye with lettuce, tomato, avocado, and black olives.)

Problem 2: What is the row of Pascal's triangle corresponding to $m = 11$?

Solution 1:

$$3 \text{ breads } \times 2 \text{ meats } \times \binom{8}{4} \text{ toppings}$$

This simplifies to 420 sandwiches.

Solution 2: The $m = 11$ row of Pascal's Triangle is

$m = 11$: 1 11 55 165 330 462 462 330 165 55 11 1