Types of Average and Sampling: "Household Words" to Dwell On

Lawrence Mark Lesser
The University of Texas at El Paso
Lesser@utep.edu

Published: August 2013

Overview of Lesson
This lesson is designed to give students more insight into both mathematical and real-world assumptions that can be involved even in what appears to be a simple task: to find the "average household size" for students at their school. The context of this lesson has real-world relevance: it is not only part of the Census at School questionnaire, but is also part of what the Census Bureau regularly estimates. The lesson can efficiently set the tone for better habits of mind and questioning and more precise usage of vocabulary and more explicit awareness of assumptions and possible biases in a broad range of future student work in statistics.

GAISE Components
This investigation follows the four components of statistical problem solving put forth in the *Guidelines for Assessment and Instruction in Statistics Education (GAISE) Report*. The four components are: formulate a question, design and implement a plan to collect data, analyze the data by measures and graphs, and interpret the results in the context of the original question. This is a GAISE Level C activity because of, for example, its nuanced exploration of sampling design.

Common Core State Standards for Mathematical Practice
(with standards 1,2,6,7 especially well hit):
1. Make sense of problems and persevere in solving them.
2. Reason abstractly and quantitatively.
3. Construct viable arguments and critique the reasoning of others.
4. Model with mathematics.
5. Use appropriate tools strategically.
6. Attend to precision.
7. Look for and make use of structure.
8. Look for and express regularity in repeated reasoning.

Common Core State Standards Grade Level Content (High School)
S-ID. 1. Represent data with plots on the real number line (dot plots, histograms, and box plots).
S-IC. 1. Understand statistics as a process for making inferences about population parameters based on a random sample from that population.
S-IC. 3. Recognize the purposes of and differences among sample surveys, experiments, and observational studies; explain how randomization relates to each.
S-IC. 4. Use data from a sample survey to estimate a population mean or proportion; develop a margin of error through the use of simulation models for random sampling.

NCTM Principles and Standards for School Mathematics
Data Analysis and Probability Standards for Grades 9-12

Formulate questions that can be addressed with data and collect, organize, and display relevant data to answer them:

- understand the differences among various kinds of studies and which types of inferences can legitimately be drawn from each;
- compute basic statistics and understand the distinction between a statistic and a parameter.

Select and use appropriate statistical methods to analyze data:

- for univariate measurement data, be able to display the distribution, describe its shape, and select and calculate summary statistics.

Develop and evaluate inferences and predictions that are based on data:

- use simulations to explore the variability of sample statistics from a known population and to construct sampling distributions;
- understand how sample statistics reflect the values of population parameters and use sampling distributions as the basis for informal inference;
- evaluate published reports that are based on data by examining the design of the study, the appropriateness of the data analysis, and the validity of conclusions.

Understand and apply basic concepts of probability:

- use simulations to construct empirical probability distributions;
- understand the concepts of conditional probability and independent events.

Prerequisites

Prior to completing this activity students should have familiarity with (1) the mean, median and mode (which the Common Core Standards indicates certainly should have happened by the start of 9th grade), (2) measures of numerical and graphical descriptive statistics (being able to know enough inferential statistics to do a one-sample test of a mean would be nice, too), and (3) concepts of sampling and types of probability-based sampling such as simple random sampling, stratified random sampling, systematic random sampling, and cluster random sampling. The beauty of this activity is that there are multiple ways for students with basic knowledge to participate and contribute. Creative insights about the idea of "size-based sampling" (even if students use different words to describe this) may be no more likely to come from the students who have greater formal mathematical knowledge.

Learning Targets

Students will be able to classify a real-life "average", and become aware of a connection between type of sampling basis (i.e., unit of analysis) and type of average.

Time Required

One class period for GAISE Steps I and II, and one class period for GAISE Steps III and IV. Teachers can make the most of class time by assigning some tasks to students as homework outside of class such as one or more of these: data collection, simulation, doing some of the reflection questions, doing one of the Activity Sheets, and looking up Census Bureau information online.

Materials Required

- The instructor will be prepared to provide a copy of the Activity Sheet (page 11).
- Students need scratch paper and access to technology such as calculators (e.g., TI-84).
- The class will need access to the Internet to look up Census Bureau information.

Instructional Lesson Plan

The GAISE Statistical Problem-Solving Procedure

I. Formulate Question(s)

Start the lesson by displaying to students question #28 from the Census at School web site questionnaire (http://www.amstat.org/censusatschool/pdfs/C@SQuestionnaire.pdf), which asks "How many people usually live in your home, including yourself?" Ask students why this question is a variable of real-world interest, bringing in connections to such domains as equity, sociology, and the economy, or trends in population growth and housing needs. Continue the lesson by analyzing the wording of the question, asking students to discuss why they think the question made it a point to say "including yourself." The question did not use the word "family" or "household", so ask students to which of the two following questions the wording of question #28 seems closer: "what is the size of your family" or "what is the size of your household". By not giving an option that those phrasings are equivalent, students are forced to reflect more deeply about possible differences between family and household. Since the focus of the Census at School question is simply who is physically living in the home, not who might be related, the discussion should yield a consensus toward the word "household."

For validation, the class can do a straightforward Google search on this question, which turns up Lofquist, Lugaila, O'Connell and Feliz (2012, p. 4) defining a household as "all of the people who occupy a housing unit", thus potentially including not only relatives (by birth, marriage or adoption), but also friends, roommates, boarders, an unmarried partner, etc. Ask students whether they, as minors (i.e., people under 18), count towards household size. Since there is no reference to minimum age – only whether the person lives in the housing unit – it should be clear to the students that they "count" and therefore, that this question has direct personal relevance to them. Students might discuss why a focus on household makes more sense (than focusing on family) for the Census, in light of how the Census is conducted by mailing a form to (and sometimes following up by visiting) a physical address. For students of same-sex orientation, a focus on households may feel more inclusive, because it does not get into current debates about whether a household headed by a same-sex couple (who may be unmarried simply because their state does not offer that option) should be classified as a family household, and these matters are likely to be handled differently by the time of the next Census). For yet more ways to give this topic real-world relevance, ask students to reflect upon the decline of household size over time (in the USA, mean household size has dropped by 1.1 people since 1930) and how changes in household size might be correlated with the economy, etc. And how might this trend relate to trends in population growth and housing needs? When students make a connection to a real-life context, they have more motivation to explore the mathematics and more of a basis to interpret and assess the reasonableness of their answers and reflect on the problem's meaning, constraints, goals, variables and conditions.

Ask students if they have any ideas for a question they could formulate based on data collected on question #28. It is very likely one of the suggestions will be "what is the average household size for students at our school?" or "what is the average answer students at our school will give to question #28?". Such a suggestion and curiosity about the result can be further motivated by the introductory paragraph of Lofquist, Lugaila, O'Connell and Feliz (2012, p. 1), which indicates that "in 2010, 300.8 million people [this excludes the 8 million people living in group-quarters arrangements such as dorms, barracks, or nursing homes] lived in 116.7 million households for an average of 2.58 people per household." Ask students to discuss what is meant by "average" (a word that is sometimes used as median, or a type of mean, or something else) in that sentence. Students should logically deduce that 2.58 cannot be a median (the median for a "whole number" variable like household size might have a value of 2.5, but not 2.58) or mode (which would have to be a whole number in this case), but can reasonably be identified as the arithmetic mean because students can notice and verify that 2.58 = 300.8/116.7. In Franklin et al. (2007), we see the importance of students at Level C (generally associated with high school students) being able to "formulate questions and determine how data can be collected and analyzed to provide an answer" (p. 61) and to recognize when questions need to be made more specific to be useful (p. 64).

II. Design and Implement a Plan to Collect the Data

Have students design a survey (our focus in this STEW lesson is on the Census at School question #28, but students may decide to add additional questions as well) and discuss the size and type of sample. Use this as an opportunity for students to recall and discuss the features and benefits of some type of random sampling procedure, rather than a "haphazard sample" of just arbitrarily stopping various students passing by in the hallway or cafeteria. Ask students to assess whether their school or school office has a "frame" available of all students in the school such as a telephone directory from which a sample could be obtained. Have students discuss what variables might be important to take into account to ensure the sample is representative of the entire school, not just your students' current classroom and grade level. Teachers might reflect in advance on how to respond if a student proposes using ethnicity or religion or immigrant status as a stratifying variable under the belief that certain groups of those types have larger household sizes than others. Have students discuss the tradeoffs of a random sample from the entire school versus some type of cluster sample where intact rooms of students (e.g., Ms. Johnson's homeroom, Mr. Wilson's first period geometry class, etc.) might be the clusters. Since virtually all students are required to take mathematics, have students discuss if using mathematics classrooms could be a good way to have everyone potentially eligible to be sampled (as well as being most likely to get the cooperation of those rooms' teachers for the needed 1-2 minutes of class time). Students should also discuss whether this sample should be done with or without replacement and which of the many technological options (e.g., the TI-84 randInt command, the Microsoft Excel command RANDBETWEEN, or an online applet) they will utilize to generate random numbers to determine the sample. Having worked out the type of sample, students can then move on to discuss an appropriate sample size, thus providing a vehicle to discuss related statistical issues such as precision, margin of error, and sampling costs or time involved.

If students have discussed inferential statistics, ask students to identify one or more hypotheses to test. An example null hypothesis might be that the mean of the collected household size values equals the national mean (2.58). If students feel a more appropriate benchmark of comparison is not the national value but their state's value, they can look this up from Table 4 in http://www.census.gov/prod/cen2010/briefs/c2010br-14.pdf (Lofquist et al., 2012, p. 10), where mean household size ranges from a low of 2.30 (North Dakota) to a high of 3.10 (Utah). Finally, ask students to discuss whether they think the alternative hypothesis should be left-tailed, right-tailed, or two-tailed.

If time is especially short, another option to obtain real data is to download data (or a random sample from this data) for question #28 from the cumulative data set classes from various states from the Census at School web site (http://www.amstat.org/censusatschool/RandomSampleForm.cfm). Students should be prepared to discuss what assumptions they may be making about this pooled data and whether there is enough of it for analysis purposes. Alternatively, your school may encourage wide participation in which a large number of students all fill out the entire questionnaire online and then you as the teacher can go in and obtain the results for your school in comma-separated values format which can be imported into packages such as Excel, Fathom, or Minitab (see http://www.amstat.org/censusatschool/participantinstructions.cfm for more information).

III. Analyze the Data

First, have students go through the data they collected during Step II of the GAISE Statistical Investigative Process (e.g., from the survey they designed) and make sure there are no obvious errors (e.g., an answer of a huge number such as 500 almost surely represents a misunderstanding or joke). Ask students to make a dotplot and see what they notice about the distribution of answers as well – ask students if they are able to say if the distribution is left-skewed, right-skewed or symmetric. Ask students if there are any gaps or outliers in the distribution (perhaps there will be at least one sampled student who lives in an unusually large household) and what criterion they may have used to assert that. Have students discuss if the number of different values in the distribution is too small for it to be useful to make a graphic such as a histogram that does some grouping or collapsing of values.

Have students (aided by any appropriate use of technology) obtain numerical descriptive statistics such as mean, median, mode, quartiles, minimum, maximum. A main focus for this lesson is the mean household size of sampled students, which is obtained by summing the answers to question #28 and then dividing by the number of students sampled. The hypothesis test mentioned at the end of GAISE Step II can be tested using the test statistic $\frac{\overline{x} - \mu_0}{\frac{s}{\sqrt{n}}}$.

Likewise, a 95% confidence interval estimate of the mean household size can be constructed using the formula $\overline{x} \pm (\text{z or t}) \times \frac{s}{\sqrt{n}}$.

As an example, suppose we obtain this data set of 18 household sizes:
{4, 2, 3, 5, 4, 5, 3, 4, 5, 4, 2, 3, 4, 2, 3, 6, 5, 3}

Students could type it into L1 in their TI-84 calculator, and then use the command EDIT→SortA(L1) to sort the data.

From the sorted data, students can readily provide this frequency table:

household size	1	2	3	4	5	6
# of households	0	3	5	5	4	1

Then, students could calculate descriptive statistics by hand or using the calculator command sequence STAT→CALC→1-Var Stats L1 to obtain a sample mean of 3.72, a sample standard deviation of 1.18, and a 5-number summary of (2, 3, 4, 5, 6). Whether done with the formula $\frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$ or using the calculator commands STAT→TESTS→T-Test, the two-tailed test of the null hypothesis that the mean is 2.58 results in a *t*-value that exceeds 4 and a *p*-value that is well below .01. And so we would conclude that the mean household size in our sample significantly exceeds the national mean. This conclusion is consistent with the 95% confidence interval estimate of the mean household size, since the hypothesized mean value of 2.58 lies outside the interval (3.14, 4.31), whether obtained from a formula or from the calculator command sequence STAT→TESTS→TInterval→Data, L1, 1, .95.

IV. Interpret the Results

The main focus of interpretation is the mean because (for reasons that will soon be discussed) the answer the students obtain will almost surely be *noticeably larger* (and perhaps triggering a rejection of the null hypothesis tested in GAISE Step III) than the Census Bureau's value of 2.58 discussed previously. Comparing results to real-world benchmarks can also reinforce this dynamic of a sample providing a larger number than expected. For example, 8th-grade teacher Dean (2013, p. 164) cites the U.S. average household size and then said "averages don't necessarily reflect conditions on the ground. According to a quick in-class hand-raising poll, only about a fourth of my students lived in households with three or fewer people, with the majority living in households of four to seven. For a bit of fun, you as the teacher can dramatically open a sealed envelope to reveal a "magic prediction" you made to this effect *before* students collected data.

Ask students now to conjecture a reason why the obtained mean was larger than the Census Bureau reference point. Let students reflect on the question as individuals first for a few minutes, jotting down their "notices" and "wonders" and any assumptions they are making. Then let students form small groups and "share and compare" their ideas within small groups. Here are three possible responses to consider (that you can facilitate discovery, exploration and discussion of if the students do not readily generate them):

- People who are K-12 students come from larger than average households.
- The data collection sampled *students*, not households.
- Your school may serve a neighborhood population whose, for example, socioeconomic demographics contribute to a higher-than-average household size.

Here is some more discussion on the first two bullets:

1) People who are K-12 students come from larger than average households. Finding information from the Census Bureau's website is a good skill for students to have. Help students to find Table S2501 Occupancy Characteristics (2007-2011 American Community Survey 5-Year Estimates; http://factfinder2.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_11_5YR_S2501&prodType=table), which tells us that 27.3% of occupied housing units had a household of only one person. This one person would almost surely *not* be a K-12 student and thus this nontrivial group of one-person households is underrepresented in a sample obtained by random selection of students at a K-12 school. That same table also says that households of sizes 2, 3 and 4+ have percentages of 33.5%, 15.9% and 23.3%, respectively. Since a K-12 student lives in a household of size at least two (since he/she is surely living with at least one other person in his/her household), this means that every student at the school lives in a household whose size is greater than or equal to 61% of the households nationwide. Note that by treating 4+ as 4, students can do the expected value calculation 1*.273 + 2*.335 + 3*.159 + 4*.233 and obtain 2.35. Students can use algebra to show that if 4+ is replaced by 5 (instead of rounded down to 4), the expected value calculation yields 2.585, which happens to yield the overall mean virtually exactly. For a hands-on simulation activity, the proportions .273, .335, .159, and .233 could be approximated by dice rolls as indicated in the middle column of the table below. Or with the TI-84+ a sample of, say, 50 students can be selected using the sequence: MATH→PRB→randInt(0,999, 50)→STO→L1. These 50 numbers are now in L1, where they can be sorted(STAT→SortA(L1)), and students can scroll through and note how many fell within each range in the right-hand column of the table below:

Household size	**Sum of 2 dice**	**TI number range**
1	2, 3, 4, or 5	0-272
2	8, 9, or 10	273-607
3	7	608-766
5	6, 11, 12	767-999

2) The data collection sampled *students*, not households. (And the Census Bureau number is averaged over households, not over individuals.) Almost every school of typical sized enrollment has multiple instances of students (e.g., siblings) who live in the same household and it is not uncommon in a multiple-child family for the spacing between two of the children to be within the 3-4 years spanned by a typical high school. And so larger households have a greater chance of being sampled and are therefore overrepresented in the students' data collection. (This is called "**selection bias**".) To be more specific, a household with 2 children attending the school is twice as likely to be sampled by the students' survey as a household with 1 child attending the same school. To explore more concretely how this makes a difference, have students work through Activity Sheet 1. To remove the issue of whether the sampled person is a student, work through Activity Sheet 2.

   In effect, sampling students produces a weighted mean because students from larger households are more likely to be chosen. Students may more readily be able to visualize weighted averages when the number of students is small enough to actually write them all

out (as in the Activity Sheet). Now ask students to reflect on the importance of identifying assumptions, considering that most of them did not initially realize that they assumed implicitly a basis for sampling (and therefore for the mean). To help them make this connection explicit, give them Activity Sheet 2 (adapted from Lesser, 2010) and have them complete it in their small groups. Students will then realize their initial answers were on a "per-student basis" instead of a "per-household basis" and this makes a noticeable difference in the answers, and allow students to notice that the "per-student basis" generally yields larger values. As a nice application of the high school algebra students have likely recently studied ("look for and make use of structure" is Common Core Standard #7), have students explore this conjecture for the simple case of finding the mean household size for a village with 2 households (of sizes a and b). As outlined in Lesser (2009, pp. 377-378), the per-household mean is $(a+b)/2$ while the per-person household mean is $(a^2+b^2)/(a+b)$, an expression which Lesser(2009) shows to be always at least as large as $(a+b)/2$.

Assessment

Additional computational practice can certainly be done with a different set of data provided or collected. For example, have students find the mean number of people in a household in a community having equal numbers of households with 1, 2, 3, 4, or 5 people. The per-household mean household size is clearly 3, but the per-person mean household size (i.e., by asking each person their answer and then averaging over all the individual people) will be greater than that because there are more people in larger households. If the community has n households of each of the five sizes, there will be n people answering "1" as the size of their household, $2n$ people answering "2", $3n$ people answering "3", $4n$ people answering "4", and $5n$ people answering "5". The mean of these $n + 2n + 3n+4n+5n$ responses is

$(1n*1 + 2n*2 + 3n*3 + 4n*4 + 5n*5)/(n + 2n + 3n + 4n + 5n) =$

$(1^2 + 2^2 + 3^2 + 4^2 + 5^2)/(1 + 2 + 3 + 4 + 5) = 55/15 = 3.67$, which is noticeably larger than 3. Since the number n ends up cancelling out anyway, there is no loss of generality to have students with weaker algebra background just assume a specific number of households (such as one of each of the five types) for this problem, and their answer will be the same. Another way to practice the concepts is to apply them to the related, but different, variable of "family size" (which also is tabulated by the Census Bureau) instead of "household size".

The biggest goal of assessment for this lesson is probably more conceptual. Therefore, ask students a question that forces them to reflect on the underlying assumptions and big picture ideas. For example, ask students to explore (and make up datasets to support their conjectures) whether a statement such as "Most households have no more than 2 people, but most people live in households with more than 2 people" must be, could be, or cannot be true. (A simple example of how it could be true is a 3-household village with household sizes of 1, 2, and 4.) Or ask students to describe in detail how they would design a study to estimate the "average class size" at your school or in your school district.

Possible Extensions

The issue of sampling households versus sampling students (i.e., computing a mean household size on a per-household basis versus on a per-student basis) as described as the third bullet point in the Interpretation step of the GAISE process appears in other contexts as well. Close at hand to the education context is the matter of average class size. It turns out the mean class size on a per-student

basis is always at least as large as the mean class size on a per-class basis, and students can explore much of the intuition and algebra, as in Lesser (2009) or Lesser and Kephart (2011).
While for household size, we generally want a per-household basis (to align with the Census Bureau), for class size, the basis (i.e., unit of analysis) we want depends on whether we are wanting a school (or teacher) view versus a student view. Feld (1991, p. 1476) describes this tension as: "faculty members experience the actual average class size, while their students disproportionately experience the larger classes; as a result, even though faculty and students have similar preferences for smaller classes, students have an interest in minimizing variation in class size, while faculty have an interest in maximizing that variation."

The underlying mathematics that shows why others come from larger households (on average) than your students do can also be applied to additional contexts of weighted averages, including: why most people have fewer friends than their friends have (people with more friends show up in more people's list of friends; this applies to Facebook friends too!), mean waiting time for a bus (people disproportionately experience the longest bus waits), mean speed of cars passing by a stationary hidden radar gun (faster cars will be sampled more frequently), and mean family size of randomly selected individuals (larger families will be sampled more frequently), people experience crowdedness (of a beach, restaurant, etc.) as worse than it usually is because people disproportionately experience the most crowded times (Feld, 1991; Hemenway, 1982; Stein & Dattero, 1985).

Another extension easily carried out in a classroom is offered by Reinhardt (1981, p. 107):

> "Ask the students in the class to give the number of boys and girls in their families. Tabulate the results separately for the boys and girls in the class and calculate the average number of boys per family and the average number of girls per family….The claim is that families tend to 'run' to children of one sex as the data superficially indicate. The difficulty is that the families of the boys in the class are not a random sample of families, nor are the families of the girls."

Further connections to the Common Core can be made by having students discuss this lesson's insights using the language of conditional probability (e.g., have students distinguish P(student selected | student lives in a "large household") from P(student lives in a "large household" | student is selected)) or conducting simulation or resampling from a set of household values.

References

1. *Portions of this lesson are adapted from what the author wrote for:*

Lesser, L. M. & Kephart, K. (2011). Setting the Tone: A Discursive Case Study of Problem-based Inquiry Learning to Start a Graduate Statistics Course for In-service Teachers. *Journal of Statistics Education 19*(3), 1-29.
http://www.amstat.org/publications/jse/v19n3/lesser.pdf
[related webinar at https://www.causeweb.org/webinar/activity/2012-01/]

Lesser, L. (2009). Sizing up Class Size: A Deeper Classroom Investigation of Central Tendency. *Mathematics Teacher, 103*(5), 376-380.

Lesser, L. (2010). Sizing up class Size: Additional insights. *Mathematics Teacher, 104*(2), 86-87.

2. *Standards documents cited are:*

Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck, R., Perry, M., & Scheaffer, R. (2007). *Guidelines for Assessment and Instruction in Statistics Education (GAISE) Report: A Pre-K-12 Curriculum Framework.* Washington, DC: American Statistical Association.

National Council of Teachers of Mathematics (2000). *Principles and Standards for School Mathematics.* Reston, VA: NCTM.

National Council of Teachers of Mathematics (2009). *Focus in High School Mathematics: Reasoning and Sense Making.* Reston, VA: NCTM.

National Governors Association and Council of Chief State School Officers. *Common Core State Standards for Mathematics. Common Core State Standards (College- and Career-Readiness Standards and K–12 Standards in English Language Arts and Math).* Washington, D.C.: National Governors Association Center for Best Practices and the Council of Chief State School Officers, 2010. http://www.corestandards.org.

3. *Other references cited:*

Dean, J. (2013). The square root of a fair share: School desks and dream homes. In E. Gutstein and B. Peterson (Eds.), *Rethinking Mathematics: Teaching Social Justice by the Numbers* (2nd ed.), pp. 161-168. Milwaukee, WI: Rethinking Schools, Ltd.

Hemenway, D. (1982). Why your classes are larger than 'average'. *Mathematics Magazine, 55*(3), 162-164.

Lofquist, D., Lugaila, T., O'Connell, M. & Feliz, S. (2012). Households and families: 2010 [Census briefs]. http://www.census.gov/prod/cen2010/briefs/c2010br-14.pdf. United States Census Bureau.

Feld, S. L. (1991). Why your friends have more friends than you do. *American Journal of Sociology, 96*(6), 1464-1477.

Reinhardt, H. E. (1981). Some statistical paradoxes. In A. P. Shulte (Ed.), *Teaching Statistics and Probability* (pp. 100-108). Reston, VA: National Council of Teachers of Mathematics.

Stein, W. E. & Dattero, R. (1985). Sampling bias and the inspection paradox. *Mathematics Magazine, 58*(2), 96-97.

Types of Average and Sampling: "Household Words" to Dwell On Activity Sheet 1

Below is a roster for a 10-student school. The first seven kids come from different households that each have X people, while the last three kids come from the same household (which has a size Y). Assume Y > X.

1.) Suppose X = 2 and Y = 4. If all 10 students report the size of the household they are in, the data consists of {2, 2, 2, 2, 2, 2, 2, 4, 4, 4}. Find the mean household size by taking the mean of these 10 numbers.

2.) Suppose X = 2 and Y = 4. If each of the eight households represented by these 10 students reports its size, the data consists of {2, 2, 2, 2, 2, 2, 2, 4}. Find the mean household size by taking the mean of these 8 numbers.

3.) Compare the answers in the previous two parts and discuss how this relates to whether you are sampling individuals or sampling households.

4.) For more generality, repeat parts 1 and 2, leaving expressions in terms of X and Y. Show that you obtain (7X + 3Y)/10 and (7X + Y)/8, respectively. Now use algebra to show that (7X + 3Y)/10 is indeed always a greater value than (7X + Y)/8.

First name	Al	Bob	Carl	Dee	Ed	Flo	Gil	Hal	Ivy	Jo
Household size	X	X	X	X	X	X	X	Y	Y	Y

ANSWER KEY: Activity Sheet 1

1.) (2+2+2+2+2+2+2+4+4+4)/10 = 26/10 = 2.6

2.) (2+2+2+2+2+2+2+4)/8 = 18/8 = 2.25

3.) In question #1, we are sampling individuals and all 10 individuals have an equal chance of getting selected (which means the four-person household consisting of Hal, Ivy, Jo and their single parent contains a larger chance of selection than any other one household). In question #2, each of the eight households is given the same selection probability or weight upon averaging, so we are sampling households now.
(X +X+X+X+X+X+Y+Y+Y)/10 = (7X+3Y)/10 while (X+X+X+X+X+X+X+Y)/8 = (7X+Y)/8. We want to show that (7X+3Y)/10 > (7X+Y)/8. By writing the fractions as decimals, this is equivalent to showing that .7X + .3Y > .875X + .125Y. Combining like terms makes this inequality equivalent to .175Y > .175X, which is true if and only if Y > X. But we assumed Y > X from the start, so we are done!

Types of Average and Sampling: "Household Words" to Dwell On Activity Sheet 2

Below is a roster for a 20-person village with four households (adapted from Lesser, 2010):
1.) What are the sizes of the four households? Find the mean of those four numbers.

2.) Now fill in column 3 and then find the mean of that column's 20 numbers.

3.) Describe how answers in #1 and #2 differ and how one might be "better" for a particular purpose.

First name of village resident	Household (unique surname)	Number of people in that person's household
Al	Jones	
Bob	Jones	
Carl	Jones	
Dee	Jones	
Ed	Jones	
Flo	Jones	
Gil	Jones	
Hal	Jones	
Ivy	Jones	
Jo	Jones	
Kay	Park	
Lia	Park	
Mo	Park	
Ned	Chan	
Olga	Chan	
Pat	Chan	
Quinn	Chan	
Ray	Gomez	
Sue	Gomez	
Ted	Gomez	

ANSWER KEY: Activity Sheet 2

1.) The Jones, Park, Chan and Gomez households have sizes 10, 3, 4, and 3, respectively. The mean of these four numbers is (10+3+4+3)/4 = 5.

2.) Column 3 consists of these 20 values (in order, from Al to Ted): 10,10,10,10,10,10,10,10,10,10,3,3,3,4,4,4,4,3,3,3. The mean of these 20 values is (10+10+10+10+10+10+10+10+10+10+3+3+3+4+4+4+4+3+3+3)/20 = 134/20 = 6.7. Notice that by writing the mean this way, you can see why some refer to it as a "self-weighting mean": (10*10 + 3*3 + 4*4 + 3*3)/20

3.) In question #1, the mean was computed on a per-household basis, which yields a smaller value (it never yields a larger value) than on the per-individual basis of question #2. Intuitively, we see that the outlier value of 10 is only one of four values in question #1, but is half of the 20 values in question #2. The approach of question #1 makes more sense when the natural unit of analysis is the household (e.g., as part of calculations for housing, taxation, etc.), while the approach of question #2 makes more sense when the experience of each individual needs to be ascertained and accounted for.