Averages

October 19, 2005

Discussion item: When we talk about an average, what exactly do we mean? When are they useful?

1 The Arithmetic Mean

When we talk about an average, we can mean different things depending on the context, but the word "average" is most commonly associated with with is called the **arithmetic mean**. The definition of the arithmetic mean is relatively simple: given a bunch of numerical data, the arithmetic mean of the data is defined to be the sum of the individual values that make up the data divided by the number of values. For example, suppose we went out on the street and asked the first ten people we saw what their age is and that we got the following answers:

$$\{17, 28, 19, 59, 73, 65, 24, 20, 33, 47\}$$

Then we can compute the arithmetic mean of the age of these ten people as follows:

$$\text{First, add up the ages: } 17 + 28 + 19 + 59 + 73 + 65 + 24 + 20 + 33 + 47 = 385$$

$$\text{Now, divide the total by the number of values: } \frac{385}{10} = 38.5.$$

Thus, the arithmetic mean (or average) of the ages of these ten people is 38.5. Note that, according the the U.S. Census, the average age of Americans is 35; what does this mean about our sample?

Of course, running out and asking ten random people their age is a somewhat silly thing to do, so the question becomes: is finding an arithmetic mean a useful thing to do in non-silly situations? Since the arithmetic mean is obviously something that someone thought was important enough to teach you, one would hope the answer is "yes".

Discussion item: What are some situations where finding the arithmetic mean might be a useful thing to do?

Example: Sports. The arithmetic mean comes up literally all the time in sports statistics. For example, in baseball a hitter's batting average is an arithmetic mean. Here's the formula for figuring out a player's batting average:

$$\text{Batting Average } = \frac{\text{number of hits the player has this season}}{\text{number of at-bats the player has}}$$

It may not be immediately obvious how this is an arithmetic mean; after all, we defined the arithmetic mean as the sum of a bunch of values divided by some number, but in this formula we

haven't added anything. How can we see that the batting average really is an arithmetic mean? (Answer: think of each at-bat as a data point. We give the at-bat the value 1 if its result was a hit and 0 otherwise. Then adding up all the 1's and 0's gives the number of hits, so we see that the number of hits is, in some sense, a sum of values, so the batting average really is an arithmetic mean).

Exercise: In 2005, Jimmy Rollins had 196 hits in 677 at-bats. What was his batting average? (Answer: $\frac{196}{677} \approx .290$)

Another example is in football: one of the ways that both teams and spectators evaluate running backs is by finding their "average yards per carry". To find the average yards per carry of a running back, add up how many yards each rush gained for and divide by the total number of carries. Note that adding up how many yards each rush gained gives the players total rushing yards, so the short formula is:

$$\text{Average Yards Per Carry } = \frac{\text{total rushing yards}}{\text{number of carries}}.$$

Exercise: In the Eagles' Sept. 18 game against the San Francisco 49ers, Brian Westbrook had rushes gaining the following yards:

$$2, -4, 8, 3, 31, 3, 6, -5, -9, 3, 9, 18, 2, 2, 20$$

What was Westbrook's average yards per carry? (Answer: $\frac{89}{15} \approx 5.9$)

Discussion Item: Can you think of scenarios where the arithmetic mean doesn't give an "average" that makes much sense? For example, are there cases where the typical value of a bunch of data is very different from the arithmetic mean of that data?

2 The Median

The last question brings us to the notion of other types of averages. There are many, many different ways to find an average; a few of the most popular are the median, the mode, the geometric mean, the harmonic mean, the generalized mean, the weighted mean, the truncated mean, and the interquartile mean. We'll focus here on the **median**.

The definition of the median is also very simple, in some ways much simpler than the definition of the arithmetic mean. To find the median of a bunch of data, simply line up the values in ascending order and the value in the middle is the median. For example, to find the median gain of Brian Westbrook's rushes from the 49ers game (listed above in chronological order), we would rearrange the rushes in ascending order:

$$-9, -5, -4, 2, 2, 2, 3, 3, 3, 6, 8, 9, 18, 20, 31$$

Now, since there are fifteen rushes, the one in the middle will be the eighth counting from either the right or the left:

$$-9, -5, -4, 2, 2, 2, 3, \boxed{3}, 3, 6, 8, 9, 18, 20, 31$$

Remember that Westbrook's average yards per carry for this game was 5.9, whereas his median yards per carry is 3. In a sense, the median gives a much better sense of how many yards Westbrook gained on a "typical" rush: 6 of his 15 rushes (40%!) were for 2 or 3 yards. However, he had several long runs as well as several rushes where he lost yardage; since he gained more on the long runs than he lost on the losses, his average yards per carry is actually **longer** than most of his carries:

only five of his carries gained more than his "average" (here I'm saying that the 6 yard carry is the "same" as the 5.9 yard average), whereas 9 (60%) gained less than his "average". On the other hand, by definition, half of Westbrook's carries gained less than his median yards per carry and half gained more.

Discussion Item: Which "average", the arithmetic mean or the median, do you think better reflects Westbrook's average run? Which is more useful in evaluating Westbrook's quality as a running back? Can you think of examples of situations where the median is a more useful "average" than the arithmetic mean? (Example: last year, there were 658 students at Sayre in grades 8-10. Figure their average age was 15. Assuming there were 50 adults working (teachers, counselors, administrators, janitors, etc.) at Sayre with an average age of 40, we get an average age for people at the school of $\frac{658\times15+50\times40}{708} = \frac{11870}{708} \approx 16.8$, even though the overwhelming majority of people at Sayre were under 16. In this case, we expect the median age of people at Sayre would probably be 15, which gives a more accurate picture)

Better Example: The U.S. Census collects a lot of data about the population of the United States; among that data, they collect extensive information on income. In 2002, the U.S. Census issued a report including the following information on household income:

Race and Hispanic origin of householder and year	Number (thousands)	Percent distribution										Median income		Mean income	
		Total	Under $5,000	$5,000 to $9,999	$10,000 to $14,999	$15,000 to $24,999	$25,000 to $34,999	$35,000 to $49,999	$50,000 to $74,999	$75,000 to $99,999	$100,000 and over	Value (dollars)	Standard error (dollars)	Value (dollars)	Standard error (dollars)
ALL RACES															
2002	111,278	100.0	3.2	5.9	7.0	13.2	12.3	15.1	18.3	11.0	14.1	42,409	139	57,852	217

Figure 1: Household income data for 2002

We see that, of the 111 million households in the U.S. in 2002, the mean household income was $57,852, while the median household income was $42,409. Why is there such a big difference? Primarily because, although only 14% of households made more than $100,000, some of those households earned **much** more than $100,000. Many CEOs, actors and professional athletes (among others) make millions of dollars. Even though there aren't very many of them, their income is so large that they skew the arithmetic mean. On the other hand, the median is unaffected by *how much* rich people make; we know, just be the definition of the median, that exactly half the households in America make more than $42,409, and exactly half make less than that amount. This is why most reports use the median as a more accurate "average" of household incomes than the arithmetic mean. (Note: because of the way the Census Bureau determines income, negative incomes [that is, situations where a household actually lost money] are probably poorly reported. If the statistics more accurately reported negative income, the mean might be closer to the median).

Discussion Item: Can you think of situations where both the arithmetic mean and the median might not be very good "averages"? (Hint: non-numerical situations)

3 The Mode

Another common way of reporting an average is called the **mode**. The definition of the mode is probably the simplest of all: the mode of a set of data is simply the most common value. Note: the mode is not necessarily unique. For example, remember Brian Westbrook's rushes against the 49ers:

$$2, -4, 8, 3, 31, 3, 6, -5, -9, 3, 9, 18, 2, 2, 20$$

The most common numbers that appear on this list are 2 and 3, which appear 3 times each. Hence, both 2 and 3 are the mode.

Discussion Item: When we're using numerical data, the mode is often not very useful, especially in comparison the the arithmetic mean and the median. Why? Examples?

However, we can't always reduce data to numbers, and in those situations, the mode may be the only useful "average" we can use.

Example: Last year at Sayre (and it's probably similar this year), 99.5% of the student body was African American and 0.5% was Latino. Now, any reasonable description of the "average" Sayre student would surely conclude that the "average" Sayre student is African American. But the arithmetic mean is worthless in this situation: what number is "African American" equal to? 1? 3? 50? 80 billion? Since we can only add numbers (what does "African American + Latino" equal?) and we have to add to find the arithmetic mean, it's silly to talk about arithmetic means in this context. The median is also non-sensical: to determine the median, we have to arrange the data points in ascending order. But which is bigger, "African American" or "Latino"? In such a simple case, it we could line the data up in whatever order we wanted to (e.g. "African American" first, then "Latino" or *vice versa*), but when there are more than two values, this becomes impossible. What if there were a hypothetical school whose students were 38% African American, 27% Latino, 19% Asian and 16% white? Then, depending on how we ordered the data, we could get any race we liked as the median.

This is a situation where the mode is useful. In both Sayre and our hypothetical school, the most common race of the students is African American, and so the mode of the data is "African American". The mode gives of a sense of the "average student".

Discussion Item: What are some other problems with the mode? (Possible Answer: If you have too many different data values, which is the most common may be pure chance and not reflective of anything. Example: In my recitations last semester, I had 76 students with 73 different last names. There were two students whose last name was Kim and two whose last name was Subramainian. Thus, "Kim" and "Subramainian" were the modes of my students' last names, even though it would be pretty strange to claim that the "average student" had Kim or Subramainian as a last name. Now, "Kim" is a pretty common last name even in the U.S., so we might be tempted to say that the fact that it was [one of] the mode[s] is indicative of the fact that it's a common last name. In some cases this holds true, but even in this context we run into problems. "Subramainian" is certainly *not* a common last name in the U.S. [I don't know about India]. Moreover, there were students with last names much more common in the U.S. than both "Kim" *and* "Subramainian": Brown, Morgan, Brooks, Lee and Rogers, to name just a few. So the mode did a poor job of picking out the most common names)

4 Distributions

Although finding averages of data (mean, median, mode) gives us some sense of what's really going on in the world the data is trying to represent, an average can never tell us the complete story. In many cases, to figure out what's going on, we need to look at the entire **distribution** of the data. For example, it's often common to represent the **frequency distribution** of data in a **histogram**. A frequency distribution is simply a count of how many times each value appears. A histogram is a graph of values vs. frequency. For example, here's a histogram showing the ancestry of U.S. residents according to the 2000 U.S. Census:

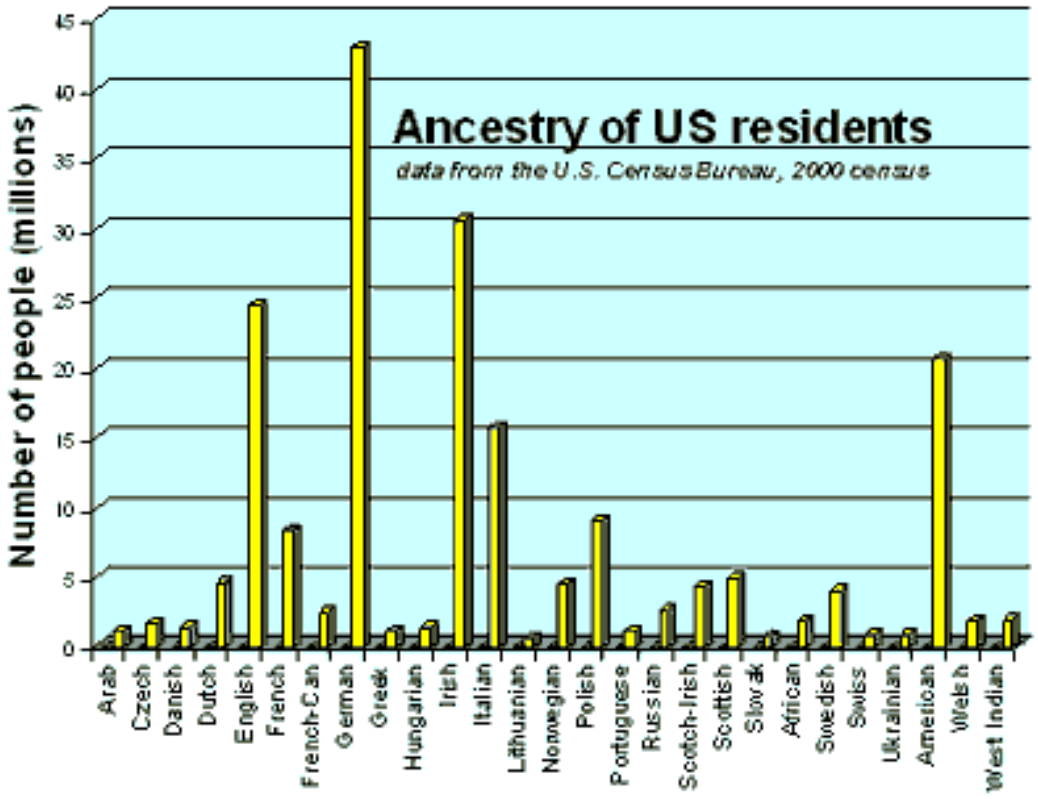

Figure 2: Ancestry of U.S. Residents

Remember the U.S. Census data on incomes we talked about earlier? Here it is again:

Race and Hispanic origin of householder and year	Number (thousands)	Percent distribution										Median income		Mean income	
		Total	Under $5,000	$5,000 to $9,999	$10,000 to $14,999	$15,000 to $24,999	$25,000 to $34,999	$35,000 to $49,999	$50,000 to $74,999	$75,000 to $99,999	$100,000 and over	Value (dollars)	Standard error (dollars)	Value (dollars)	Standard error (dollars)
ALL RACES															
2002	111,278	100.0	3.2	5.9	7.0	13.2	12.3	15.1	18.3	11.0	14.1	42,409	139	57,852	217

Figure 3: Household income data for 2002

The middle columns give a rough frequency distribution of household income. It's sort of hard to figure out what's going on just from the table, which is why histograms are so useful. Here's a histogram of this frequency distribution:

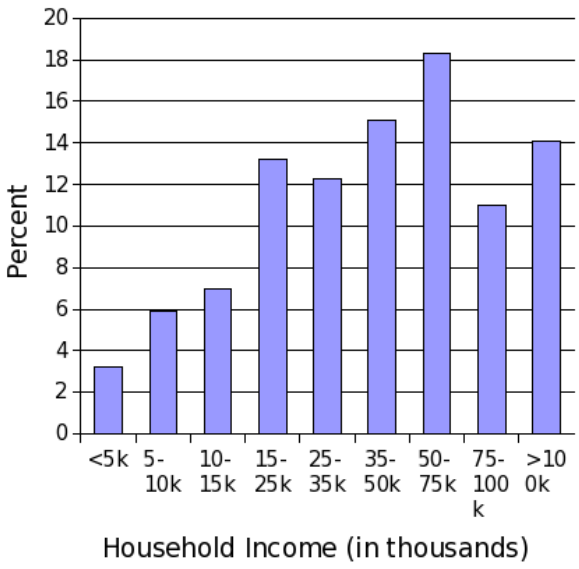

Figure 4: Household Income 2002

Discussion Item: What's misleading about this histogram? (Hint: look at the "size" of the entries; some contain a range of only $5000, like the $10,000-$15,000 entry, while others contain a much larger range, the most glaring example being the >$100,000)

Here's another histogram of the same data, where I've standardized the range of each entry and made an educated guess as to the percentage of households in each range (Note: this is only a guess! It may reflect reality **very poorly**!):

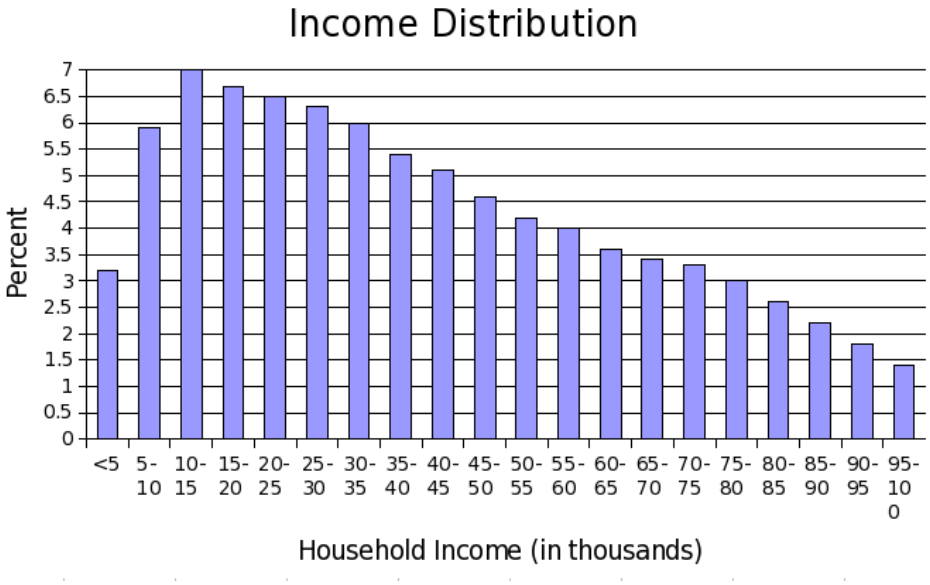

Figure 5: Household Income 2002 (Interpolated)

Here I've chopped incomes over $100,000, which will tail off a long ways to the right.

There are some distributions that come up over and over again with many different types of data. The most famous of these is called the **normal distribution**. In the normal distribution, the mean, median and mode all agree, and the further you get from this average, the less likely a value becomes. Here's a picture:

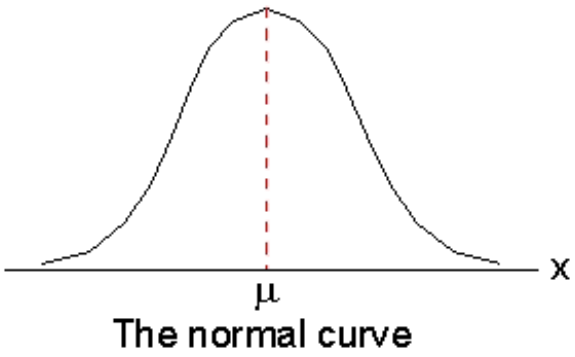

Figure 6: The normal distribution

Here, μ is both the mean and the median. You can see that the further a value is from the average, the less likely it is. The normal distribution is often called the **bell curve**, since it looks sort of like a bell.

The normal distribution tends to occur when many small effects, acting independently, cause the value being measured. For example, a person's blood pressure is determined by a very large

number of effects, any one of which contributes very little. For example, there are thousands of genes that contribute to a person's blood pressure. Also, a person's diet, exercise, medical history and various other things contribute (though it might seem like "diet" wouldn't be a "small" effect, remember that "diet" is just a catchall term than encapsulates many tiny effects; each piece of food you eat has a small effect on your blood pressure, but no one piece of food will have a significant effect). As such, it shouldn't be a surprise that, once we control for sex, blood pressure fits a normal distribution. For example, here's a histogram of the diastolic blood pressures of 1000 Australian men, with the approximate normal distribution drawn in:

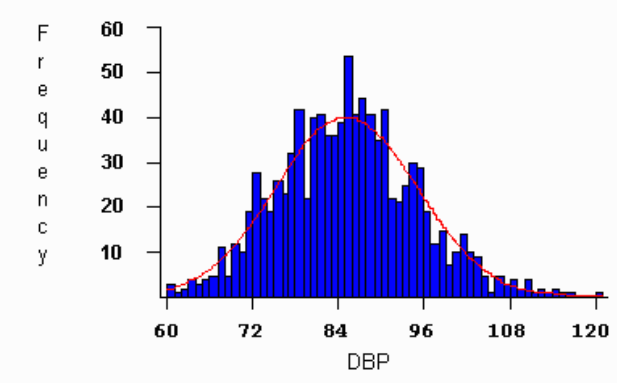

Figure 7: Diastolic Blood Pressure Distribution

Other physiological measurements also fit a normal distribution, as we would expect since they are dependent of many, many small effects. For example, height and weight fit a normal distribution (technically, they fit a lognormal distribution; that is, the natural logarithm of height and weight fit a normal distribution; this is because the factors contributing the height and weight act multiplicatively, rather than additively).

Possible Exercise: Have the class measure their heights. Calculate mean, median and mode and plot a histogram of the frequency distribution. If the class is large enough and the the variance isn't too great, this histogram should roughly look like a normal distribution. An even better exercise would be to perform the same analysis on blood pressure data collected in the course of the Medical Intake program (will work best if only male BP or only female BP is analyzed, though a combined distribution will probably still look roughly normal).

Other Examples: Of course, the normal distribution comes up in many other situations. Whenever the same quantity is measured repeatedly (for example, if you measured your height a hundred times with a very sensitive measuring device), the frequency distribution of the observed values should be a normal distribution. Also, IQ scores are normally distributed, because IQ tests are designed to give a normal distribution of scores. However, this does not mean that *intelligence* is normally distributed (which is actually a hot debate in psychology), nor that IQ tests accurately measure intelligence (whatever it may be).

Another common distribution is the **exponential distribution**. Exponential distributions are displayed graphically with *rank* along the x-axis and quantity or value along the y-axis. Here's a picture:

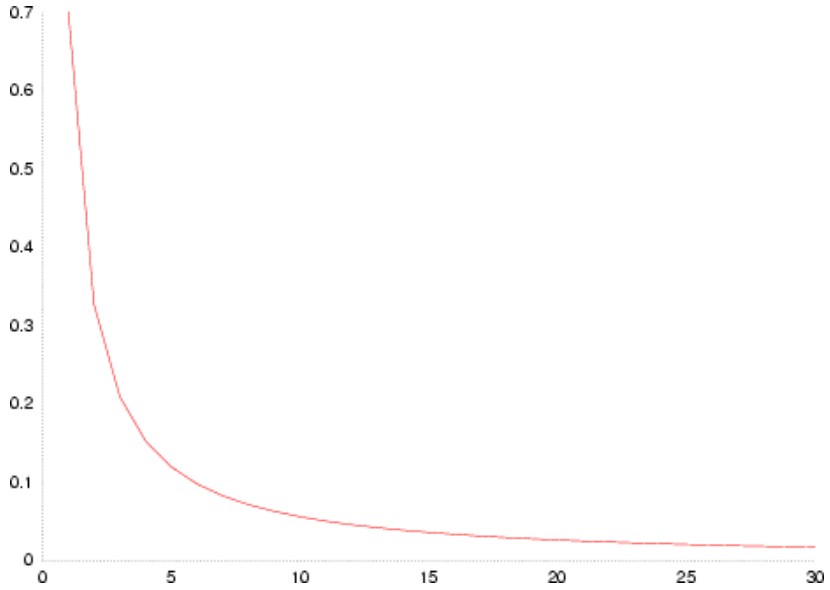

Figure 8: An exponential distribution

Exponential distributions arise in many different contexts.

Example: If we rank the words by how many times they appear in the complete works of Shakespeare, then they fit an exponential distribution. The most common words in Shakespeare are "the" and "and", each of which appears more than 25,000 times; they would be on the far left of the above graph. On the other hand, of the 23,321 words that Shakespeare used, 11,645 (almost half) appear only once or twice. (See http://ise.uvic.ca/Annex/Stats/ for the complete frequency distribution of words in Shakespeare)

Other Examples: Exponential distributions also arise in the following cases, among many, many others:

- Settlement sizes — There are a few very large cities in the world, but most settlements are small towns or villages
- Size of earthquakes — Extremely powerful earthquakes are very rare, but dozens of small earthquakes around the world each day. (See the constantly-updated USGS list of recent earthquakes around the world at: http://earthquake.usgs.gov/recenteqsww/Quakes/quakes_all.html)
- Record and book sales — Top 40 records and bestselling books sell millions of copies, but the vast majority of records and books sell very few copies. The standard rule of thumb for both is that 20% of records or books account for 80% of sales, while the remaining 80% account for only 20% of sales.

Income: Income is a special case. From our histograms earlier, we can see that income doesn't fit a normal distribution and, although it was thought for a long time that income fit an exponential distribution, it turns out that this is not the case. In fact, the distribution of income can be

viewed as two separate pieces: the top 3% of incomes **do** fit an exponential distribution; however, the bottom 97% of incomes fit a distribution curve that also, surprisingly enough, describes the spread of energies of atoms in a gas. Pictorially, the income distribution looks very similar to the exponential distribution we pictured earlier (where Bill Gates would be on the far left and normal people tail off to the right), but recent research has shown that the shape of the "tail" is slightly different. (See http://www.newscientist.com/article.ns?id=mg18524904.300).