On the Units of the *Sample Arithmetic Mean*

Neil J. Hatfield
School of Mathematical Sciences,
University of Northern Colorado

Revised May 2019

This reading centers on one of the best known statistics: the *Sample Arithmetic Mean*. While this statistic is one of the most widely used, there are many instances where people struggle to reason with the *Sample Arithmetic Mean* properly. Before you read any further, I recommend that you read S. Raper's "The Shock of the Mean" in the Dec. 2017 volume of *Significance*[1].

What follows is what I hope you'll take as a prompt for critical thinking about the *Sample Arithmetic Mean* and how quantitative reasoning plays an overlooked but critical role in using this statistic to make decisions.

1 A Brief History of the *Sample Arithmetic Mean*

The *Sample Arithmetic Mean* has a long history in both Mathematics and the Sciences. One of the earliest usages of the *Sample Arithmetic Mean* stems from astronomy and navigation: to account for measurements being made in different conditions and to mitigate the impact of measurement errors. Astronomers wanted to know the "true" position of stars and other astronomical bodies but were faced with variation in their measures. In these early days, the astronomers accounted for this variation through heuristics. For example, they would ignore measurements not made under the same weather conditions, toss out the observations they made when they had a cold, or use a person's [perceived] reputation to give more/less credence to a particular observation.

These early researchers became increasingly concerned with the use of (questionable) heuristics and started exploring/developing new methods. One of the more productive methods eventually lead to what we call "The Method of Least Squares" (which you might have heard of in regression). The power behind the Least Squares method comes from these early scientists and statisticians realizing that they could combine measurements, even observations that they suspected of having measurement error, together. They reasoned that the measurement errors would not necessarily compound each other, but rather, the errors might balance each other out. These scientists and statisticians believed that by using the Least Squares method, they could distill the "true value" from a collection of "error-prone" measurements. When you are dealing with a single attribute, the Least Squares method reduces to what

[1] A PDF of this article is available in Canvas.

we know as the *Sample Arithmetic Mean*.

The idea that the *Sample Arithmetic Mean* allows for the distillation of the "true value" for some attribute based upon a collection of observations echoes throughout Statistics. One of the main parameters we need to specify for many distributions is that of the location parameter. The *Sample Arithmetic Mean* is one of the best estimators for the location parameter, especially when the location parameter is the Expected Value. This makes the *Sample Arithmetic Mean* a popular statistic.

2 Units on the Values of the *Sample Arithmetic Mean*

In this section, we will discuss the two viewpoints on what the units of the *Sample Arithmetic Mean* should be. As a matter of convention (and without loss of generality), we will suppose that we have n widgets in a collection we'll call $\mathcal{X}$ and we've weighed each widget's mass in grams. Let X represent widget mass (g).

2.1 The Quantitative Approach

Quantitative Reasoning focuses on thinking about what we're measuring, how we measure, and what a measurement means when we get one.

Imagine that we have a collection of widgets and that our collection has the goal of being the most massive. We know that we can combine each widget's mass with every other widget's mass to find the [total] mass of the collection; this would be the value of the GP^+ statistic, which tells us how well a collection does towards some accumulation goal. However, if we want to be able to compare our collection's performance against *any* other widget collection's performance, we need to account for the sizes of each collection. Since we are not in the habit of removing members of a collection (or making up data), we cannot make two collections of different sizes the same size. Thus, we need to find each collection's value of group performance–adjusted for the size of the collection (a.k.a. adjusted group performance).

Since mass is additive, we can make a comparison between the [total] mass of the collection (i.e., group performance) and the size of the collection. The comparison we want to make is to compare the magnitude of the collection's group performance to the magnitude of the collection's size; we end up with the relative magnitude of the two measures of these quantities. We get relative magnitudes through the mathematical operation of division. This lets us establish the *Sample Arithmetic Mean* as

$$SAM(\mathcal{X}) = \frac{GP^+(\mathcal{X})\ [\text{grams}]}{Count(\mathcal{X})\ [\text{widgets}]} = \frac{\sum_{i=1}^{n} x_i\ [\text{grams}]}{\sum_{i=1}^{n} 1\ [\text{widgets}]} = \frac{\sum_{i=1}^{n} x_i\ [\text{grams}]}{n\ [\text{widgets}]}$$

In the above definition, I've included the units for each of the two quantities involved in the comparison; grams in the numerator and widgets in the denominator. Remem-

ber, we're looking at how many times the measure of the collection's performance (an amount of mass) is as large as the measure of the collection's size (a number of widgets). In looking at the relative magnitude, we are doing exactly the same thing as when we study an average-rate-of-change[2]. An average-rate-of-change of 34 mph means that the net change in the distance we travel (in miles) is 34 times as large as the corresponding net change in the time spent traveling that distance (in hours).

As we imagine going from an empty collection (0 widgets) to our full collection with n widgets, the net change in the collection's mass is going to be $SAM(\mathcal{X})$ times as large as the change in the number of widgets (i.e., n). If we know that the $SAM(\mathcal{X}) = 14.2$, then we know that the collection's mass is 14.2 times as large as the n widgets. The rate given by the *Sample Arithmetic Mean* allows us to compare our collection of widgets with any other collection of widgets. This establishes the unit of the *Sample Arithmetic Mean* as grams per widget for our present context. More generally, the unit on the *Sample Arithmetic Mean* will reflect the fact that the output of this statistic is a rate-of-change and will have the general form of _____ per _____.

2.2 The Traditional Approach

The Traditional Approach focuses on finding the "true value".

In the traditional approach, the idea is that we're out to find the "true mass" of a widget. While we have a collection of widgets, we must view their measured values of mass as being error-prone. To account of the errors and hone in on the "true value" we need a decently sized sample (many believe that the larger, the better) and to begin combining the values together. This leads to the following:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i \text{ [grams]}}{\sum_{i=1}^{n} 1} = \frac{\sum_{i=1}^{n} x_i \text{ [grams]}}{n}$$

You'll notice in this definition that there are no units listed in the denominator; just in the numerator. In this approach, the only quantity that exists is that of mass for individual widgets. The collection is not an entity that has attributes; in fact, the collection is essentially forgotten about. While we still count the number of widgets we have, we are not to think of this value as being the measurement of a quantity. Thus, the denominator is a unitless entity (a scalar) that we don't need to seriously consider. Since there is only one measurement unit in the traditional approach to the *Sample Arithmetic Mean*, then there can only be one possible unit for the output of the *Sample Arithmetic Mean*; grams. More generally, the unit of the *Sample Arithmetic Mean* in this traditional approach is identical to the unit used to measured each of our observations.

[2]The phrase "average-rate-of-change" is one of two acceptable usages of the word "average".

2.3 Moving Between the Quantitative and Traditional Approaches

Suppose that we start from the Quantitative Reasoning approach and believe that the unit on values of the *Sample Arithmetic Mean* should be grams/widget. If we were to imagine a new collection that has the same performance and same size as our actual collection, but only contains identical widgets, then the mass of each widget must then be equal to $\left(SAM(\mathcal{X})\frac{\text{grams}}{\text{widget}} \cdot 1 \text{ widget}\right)$ grams. Notice that we needed to think of a brand new collection that did not just contain one widget, but n widgets; not just any widgets, but imaginary widgets that were perfectly identical. Thus, when we begin with the quantitative reasoning approach we can arrive at the same construction as in the Traditional Approach, but we're aware that we've made some critical leaps in our thinking.

Suppose that we start with the Traditional approach and believe that the unit on values of the *Sample Arithmetic Mean* should be grams. There is not a clear path to get to the other viewpoint. Since the focus is on this single, fictitious widget, our thinking about the collection has disappeared.

3 The More Productive Viewpoint

So, which unit is the "right" unit? The unit on values of the *Sample Arithmetic Mean* cannot simultaneously be both grams/widget and grams. We are also not dealing with a case of measurement systems like metric versus imperial (cm vs. in). I freely admit that I am in the minority when I stress that the appropriate unit on values of the *Sample Arithmetic Mean* should reflect that we are dealing with a rate (e.g., grams/widget). Being in the minority does not mean that a person is incorrect (i.e., a person cannot be "wrong" just because he/she does not agree with the majority). Nor does being in the majority make a person "right". The issue is that the "grams/widget" people and the "grams" people are speaking of ***two different things***.

3.1 A Tale of Two Objects

Take a moment and re-examine the first two subsections of Section 2 and ask yourself the following question: What is the ***thing*** whose attribute is getting measured by the *Sample Arithmetic Mean*?

For the Quantitative Reasoning case, the focus is on measuring something about the ***collection*** of widgets, not any one widget. For the Traditional Approach, the focus is on measuring something about a ***fictitious/imaginary widget***. While the same arithmetic operations were used, what is ultimately being imagined is different. This difference in the imagined object lies at the heart of what's going on.

In the traditional approach, the goal is to find the "true value" of whatever is being measured. Historically, the thing that we were trying to measure was a planetary

body like the Moon, Jupiter, or Saturn. We were attempting to determine the "true position" of these celestial bodies. The Belgian scientist Quételet capitalized on the approaches in Astronomy through the introduction of a powerful construct: *l'homme moyen*. The *l'homme moyen* or "average man"[3] represents a fictitious being that a researcher would mentally create and imbue with particular features. Quételet first developed this construct as he studied the relationships and differences between different groups of people. This enabled Quételet to talk about the propensity of the French *l'homme moyen* to commit a crime as opposed to the Spanish *l'homme moyen*. Quételet's creation was generally well-received by researchers, scientists, and statisticians of the time (circa 1835).

3.2 Quételet's Folly

While *l'homme moyen* played an important role in the expansion of Statistics into the Social Sciences, Quételet's construct quickly grew wild. While on a certain level Quételet fully acknowledged that *l'homme moyen* as imaginary, he operated as if there actually was an actual man (or woman).

In Astronomy, the scientists used the *Sample Arithmetic Mean* to combine multiple measurements taken on the ***same object*** to state something about ***that object***. Quételet used the *Sample Arithmetic Mean* to combine single measurements taken on ***multiple people*** to state something about an idealized person who ***never existed***. The heart of Quételet's approach is based on the fallacy that each person is an imperfect copy of this idealized person. He often explained his approach to the *Sample Arithmetic Mean* through the "Statue of the Gladiator" metaphor:

> Imagine a marble statue of a Roman gladiator. Suppose that we bring in 1000 sculptors and provide each with the tools and raw materials to copy the statue. The resulting 1000 sculptures will always have imperfections; none will be a perfect copy of the original statue. However, if you take all 1000 sculptures and combine them together, you could recreate the original statue.

Quételet's folly of treating single measures from multiple people as if they were multiple measures of a single object has haunted us ever since. A more recent example comes from the late 1940s, early 1950s. The US Air Force began noticing a disturbing trend in their new jet planes: the pilots would lose control and crash—one day saw 17 separate crashes. Upon investigation, the Air Force discovered that mechanical issues with the new planes were not the root cause, nor were the pilots. Engineers decided that they needed to revisit the design specifications of the cockpits. Back in 1926, the Air Force took measures on various attributes from hundreds of pilots and then took the values of the *Sample Arithmetic Mean* as being <u>*the*</u> values to base the entire cockpit design off of. In the '40s, the Air Force worried that the pilots were bigger. Thus, they proceeded to take new measurements on 140 attributes for over 4000 pilots. One of the junior researchers tasked with taking the measurements did not believe in the existence of *l'homme moyen*. Instead Gilbert Daniels did his duty, took all the measurements, and then undertook an interesting study: how many pilots

[3] Discussing *l'homme moyen* is the second and last acceptable usage of "average".

are "average"? Daniels found the value of the *Sample Arithmetic Mean* of the 4000 pilots along the ten attributes believed to be the most important. To be generous, he then added and subtracted 3/10ths of the *Sample Arithmetic Standard Deviation* to each value of the *Sample Arithmetic Mean*, constructing an interval for each of the ten attributes. Daniels then looked at all of the pilots and checked to see how many of them fit within the ten intervals. To his (and everyone's) surprise there were ***no*** pilots whose measurements fell in all 10 intervals. If he went down to just three of those intervals, less than 3.5% of the pilots would have measures within those intervals. Prior to this discovery, most of the individuals involved would have stated that most of the pilots would fall in Daniels's intervals. Today, most researchers would have expected nearly a quarter of the pilots to fall within the intervals. According to the so-called "Empirical Rule", which is steeped in Quételet's folly, approximately 23.58% of observations should be within 3/10ths of the *Sample Arithmetic Standard Deviation* of the value of the *Sample Arithmetic Mean*. Daniels's research convinced the Air Force of the *l'homme moyen* fallacy and they proceeded to make a radical change to the design specifications for the cockpits: the cockpits needed to fit all pilots whose measurements fell between the fifth and ninety-fifth percentiles. This new demand spawned the invention of adjustable seats, helmet straps, and foot pedals (to name a few); things we take for granted each time we get into a vehicle.

3.3 From "True Value" to Comparing Groups

While in the 17th and 18th centuries there was a focus on finding the "true value", this emphasis began to give way to new lines of scientific inquiry such as "how do these groups differ?" Even within the Least Squares Methods, the focus shifted from finding a "true value" to one of exploring how various collections differed from each other (often thought of as exploring how groups "out-perform" one another). To be certain, Quételet's *l'homme moyen* contributed to Statistics's development, but as the notion of group comparison grew and shaped Inferential Statistics, his construct faded into history. When you look at the methods of statistical inference that arose after Quételet's time, there is an underlying focus on groups. Even the way that we write statistical research questions reflect group comparison focus: "Do patients who get Drug A have shorter recovery times than patients who get Drug B?".

The *Sample Arithmetic Mean* measures adjusted group performance for any collection. To take an attribute of the collection, such as adjusted group performance or size, and apply that same attribute to an individual object/living being is what Raper referred to as the "Ergodic Switch" (a.k.a. the "Ergodic Fallacy"). Just as we would not expect a piece of marble to have the attribute "favorite color" or speak of a human's value of hardness on Moh's scale, we should guard against trying to apply attributes of groups/collections to single objects. A classic example of doing this is something that I'm sure that you have done in every class after you've taken an exam: you've compared your score to the value of the *Sample Arithmetic Mean*. Each time you've made such a comparison, you are engaging in the ergodic switch; the fallacy of Quételet. **There is no *l'étudiant moyen* ("average student").** To use a second analogy, when you compare your grade to the value of the *Sample Arithmetic Mean*, you are essentially comparing an amount of distance you've travelled to a speed. Just

as we do not make direct comparisons between the value "3.1 miles" and the speed "65 mph", we should avoid doing so with percentages (your grade) and percentage per student (the value of the *Sample Arithmetic Mean*). We can compare how far two individuals have run on a certain day to each other and we can compare the speeds for two different roads (or even two different sections of the same road) in meaningful ways; comparing a runner to a road makes little sense. The use of the *Sample Arithmetic Mean* is for making comparisons between collections/groups/classes, not between individual students.

I've held off on answering the question that I opened the section with. I will propose the following tweak to the question: Which unit is the *more productive* unit? In my opinion as a statistician and as a Statistics Education researcher, viewing the unit of the *Sample Arithmetic Mean* as a rate (e.g., grams/widget) is the more productive way to think about this statistic. This Quantitative Reasoning Approach more fully embraces critical thinking skills necessary for statisticians (i.e., quantitative reasoning), avoids flawed thinking (i.e., the ergodic switch), and supports a way of thinking that underpins statistical inference (i.e., the comparison of groups). The Traditional Approach does not provide such opportunities; this approach essentially encourages making logical fallacies and doing "bad" science. When we encounter someone who believes that the "correct" unit is the same as the data (e.g., grams), we must share what we understand about the *Sample Arithmetic Mean* and discuss the difference in view points.

Unfortunately the Traditional Approach is rather ingrained in today's society as Quételet's fallacy holds unprecedented sway. Sadly, the way that most individuals learned/developed this view is far removed from the historical development. If you've had a Statistics class/unit before, did you ever talk about *l'homme moyen*? In the 40+ textbooks I have, none mention or highlight that their treatment of the *Sample Arithmetic Mean* hinges on imagining a fictitious object or acknowledges the ergodic switch. Rather the emphasis is on the procedure of "adding things up and dividing by the number of things you added". This procedural meaning is divorced from the quantitative reasoning that not only birthed the *Sample Arithmetic Mean* but helps you to understand why we would even want to use this statistic in the first place.

Dear Student, remember that I want to support you in developing productive meanings for every concept in Statistics. This means that you are expected to develop your understanding of the Sample Arithmetic Mean to view the value as a rate that tells us something about the collection, not of a single fictitious object isolated from a collection.