The shock of the mean

Simon Raper recounts the history of the arithmetic mean, why scientists of the past rejected the idea, and why their concerns are still relevant in the ongoing struggle to communicate statistical concepts

In 1755 the mathematician Thomas Simpson wrote to the Earl of Macclesfield to put his weight behind a controversial new technique:

> It is well known to your Lordship that the method practised by astronomers, in order to diminish the errors arising from the imperfections of the instruments, and of the organs of sense, by taking the Mean of several observations has not been so generally received, but that some persons, of considerable note, have been of the opinion, and even publickly maintained, that one single observation, taken with due care, was as much to be relied on as the Mean of a great number.[1]

A statistical average is a thing so familiar to us now, spilling out of news bulletins, government reports, and business presentations, that it seems odd that it was ever in need of justification. But, to the scientists of the eighteenth century, the use of the arithmetical mean to summarise data was anything but obvious. As Simpson's letter shows, the prevailing method was to take the best of one's observations as the most reliable estimate, and this seemed good and right. To such men, all measurement meant measurement of an object, and accuracy was about carrying this out in the most skilful way possible. To combine the best of one's observations with inferior attempts would have seemed perverse. And if combining observations on a single object was considered radical, doing the same for measurements made on many different objects was almost unthinkable. As we shall see, it took the force of new metaphor, the idea of the *average man*, to pull this off.

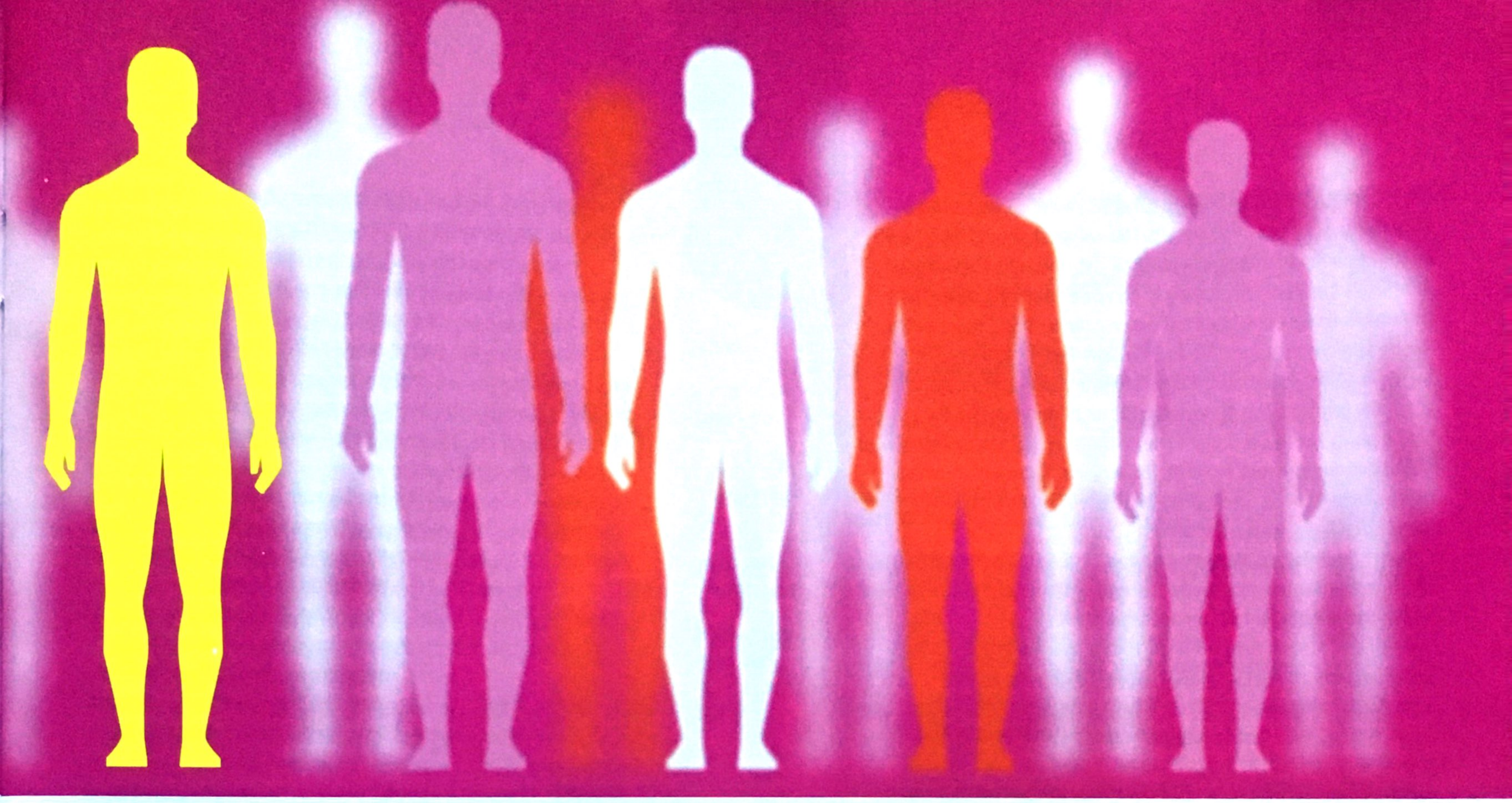
This article is about the origin and meaning of the arithmetic mean, and the struggle to justify and understand what now seems to be the simplest of statistical ideas. Why should this hold our interest? Well, on the one hand, it allows us to glimpse the world as our predecessors saw it, before the idea of a mean became commonplace. This exposes neglected arguments against its use which can shock us out of complacency. On the other hand, understanding why the arithmetic mean was originally so hard to grasp can help unravel why it is that statistical concepts are still so difficult to communicate to non-statisticians – a fact that dramatically reduces the effectiveness of our work.

Simon Raper is a statistician and the founder of Coppelia, a London-based company that uses statistical analysis, simulation and machine learning to solve business problems.

The origin of the mean

It is hard to shake the idea that the arithmetic mean must have made an appearance early on in human history. It seems too obvious and too useful an idea not to have crept in before the eighteenth century. However, in *The Seven Pillars of Statistical Wisdom* the leading historian of statistics, Stephen Stigler, goes in search of earlier references to the arithmetic mean and comes up empty.[2] Instead he uncovers plenty of evidence that the arithmetic mean was a difficult and counterintuitive idea. As Stigler points out, for all their ingenuity in the application of mathematics to practical problems, neither the Greeks nor the Romans, nor the even the Arab astronomers and scientists of the Middle Ages, thought to calculate an average from their data.

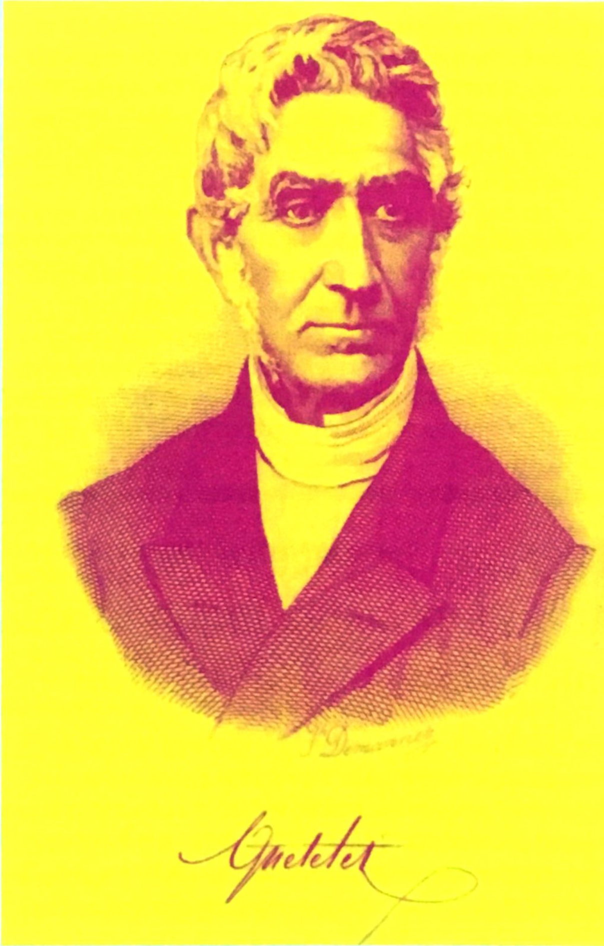
It did, however, surface early on as a purely abstract idea. Around 280 BC the Pythagoreans mention the arithmetic mean in the context of music and proportion, along with the geometric and harmonic means, but there is no suggestion of using it for summarising data. There were, though, precursors to its practical application – moments in which groups or individuals appear to stumble around the idea, half grasping it, only to be held back by the same prejudice towards the concrete that Simpson was still battling hundreds of years later. For example, in 428 BC the Greek historian Thucydides describes the process of estimating the size of an enemy's defences by counting its height in bricks. Several people made



the count and the most common value (what we would call the mode) was taken as the best estimate. The Greeks clearly understood that there were benefits to pooling data, but they clung on to the assumption that it is the best observations that count, the mode being just one way (by consensus) of deciding on the best.

We have to wait until the early sixteenth century for the first true instance of the use of an arithmetic mean for a practical purpose, although it is neither named as such nor explicitly linked to the mathematical concept. This is the attempt by the mathematician Jakob Köbel to set a standard for a unit of land measurement, the rod, which was defined as 16 feet long. The difficult matter of whose feet should be used was solved by picking 16 individuals and lining them up toe to heel to define the official length of a rod. We would say that the differences between the individuals' feet lengths were *averaged out* in the aggregate. But, as Stigler points out, the notion that there was something like an average foot length, in which the unique characteristics of any particular foot were discarded, was still a long way from being recognised. The individuals whose feet make up the rod are drawn in meticulous detail in an engraving that depicts the process. It is significant that "their identity was not discarded; it was the key to the legitimacy of the rod".

Unsurprisingly, given its reliance on multiple observations, it is in astronomy that we see the first general trend towards systematically combining data. At the end of the sixteenth century, Tycho Brahe recommends the repetition of measurements without specifying a method for combining them. We then find astronomers experimenting with a wide range of techniques for doing so, including means, mid-ranges and medians, without arriving at a consensus. (Ironically, when we do get the first recorded use of the term "arithmetic meane", in 1635, it is not used to refer to a mean at all but rather to a mid-range. The astronomer Gellibrand uses it to describe



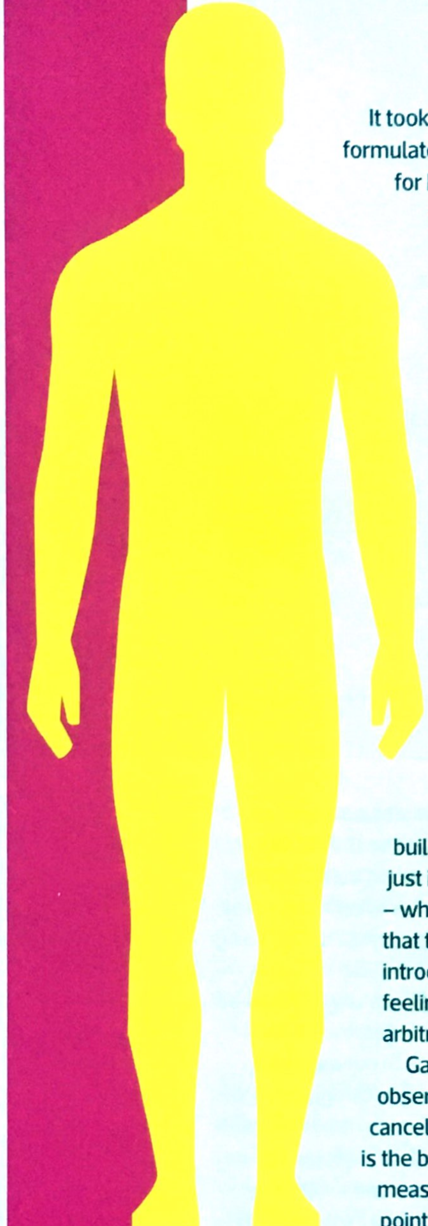
ABOVE The father of the "average man": Belgian astronomer, mathematician, statistician and sociologist Adolphe Quételet (1796–1874). Portrait by Joseph-Arnold Demannez (1825–1902). Steel engraving, recoloured.
Source: Library of Congress (cph 3b11632)

the midpoint between the highest and lowest value recorded, effectively a mean of two values.) However, it should be noted that whenever an average is used by astronomers they proceed extremely cautiously, refusing to combine anything other than observations made in identical circumstances. As Stigler notes: "Astronomers averaged measurements they considered to be equivalent, observations they felt were of equal intrinsic accuracy because the observations had been made by the same observer, at the same time, in the same place, with the same instrument, and so forth. Exceptions, instances in which measurements not considered to be of equivalent accuracy were combined, were rare before 1750."

We then enter a fascinating period, lasting from the middle of the seventeenth century to the end of the eighteenth, in which the supporters of the average gradually gain ground over its detractors. Around 1660 the eminent scientist Robert Boyle argues against repeated experiments, comparing a single experiment of high quality to a valuable oriental pearl, not to be traded for any number of cheap and inferior specimens. In 1722, in a posthumously published article that was largely ignored, Roger Cotes argues for the average by giving a justification based on the centre of gravity of a set of weights, whose positions on a horizontal bar represent their values. By the time we get to Simpson in 1755, the balance seems to be shifting sufficiently in favour of the average to warrant a fight-back by "persons of considerable note". Finally, a letter from Daniel Bernoulli in 1777 seems to imply that taking an average had become the norm.

The wrongs that make a right

The final acceptance of the arithmetic mean required another important conceptual shift: from the more intuitive idea that observational errors stack up over time (as they do, for example, when making mathematical deductions) to the less obvious thought that, provided there is no bias, they will cancel each other out. Stigler cites the medieval Trial of the Pyx, a



Final acceptance of the arithmetic mean required an important conceptual shift

test applied to the coins produced at the Royal Mint, as an early example where it was thought that errors (in this case, differences between the desired weight of a coin and its actual weight) would accumulate rather than average out. He also attributes the mathematician Euler's failure in 1749 to solve the problem of three bodies (an important astronomical problem of the time) to his belief that errors should be summed: "He distrusted the combination of equations, taking the mathematician's view that errors increase with aggregation rather than taking the statistician's view that random errors tend to cancel one another."

It took an explicit theory of errors, as formulated by Carl Gauss as the foundation for his work on the normal distribution, to fully dislodge the idea of cumulative error. Saul Stahl, in an excellent article on the history of the normal distribution,[3] shows how Gauss derived the normal distribution from only three assumptions:

1. Small errors are more likely than large errors.
2. For any real number ε, the likelihoods of errors of magnitudes ε and $-\varepsilon$ are equal.
3. In the presence of several measurements of the same quantity, the most likely value of the quantity being measured is their average.

The derivation itself is a perfect example of how mathematics can build something new and unexpected from just its own rules and a set of simple axioms – which makes it even more frustrating that this step is missed out of almost every introduction to statistics, leaving the student feeling that the normal curve is something arbitrary and mysterious.

Gauss's normal curve, as a model of observational errors, implies both that errors cancel each other out and that the average is the best estimate of the quantity being measured. But, as Stahl points out, the latter point was brought in quite brazenly as the third axiom – it was assumed, not proved. It was Laplace in 1810 who provided the final brick in the edifice by showing, in the central limit theorem, that the sample mean is itself distributed normally around the population mean, thus justifying the third axiom.

Although the significance of the central limit theorem was not fully appreciated until the late nineteenth century, it is fair to say that, by 1810, the claim that the arithmetic mean was the best method of combining many observations of the same phenomenon stood on solid foundations and was broadly accepted – and yet, still nowhere was there any inkling that an average might express a measurement across many dissimilar cases. That idea, with its implied wastefulness of simply throwing away all the information that pertained to individual cases, remained beyond consideration.

Quetelet's back door

When this all changed, in the mid-nineteenth century, it was largely due to the efforts of an underappreciated Belgian mathematician and astronomer named Adolphe Quetelet.

Quetelet (pictured on page 13) pioneered the application of statistical methods to social data and, in doing so, radically reinterpreted Gauss's error curve. The pivotal moment came while Quetelet was examining a set of data giving the chest sizes of 5738 Scottish soldiers. As he examined the distribution of the chest measurements, he did something unheard of for the time: he calculated the average chest size across all the unique and physically distinct Scottish soldiers. However, as Todd Rose points out in his fascinating book, *The End of Average*, applying the calculation to the data was not his most radical step. "After Quetelet calculated the average chest circumference of Scottish soldiers, he concluded that each individual soldier's chest size represented an instance of naturally occurring 'error'," Rose writes, "whereas the average chest size represented the 'true' soldier – a perfectly formed soldier free from any physical blemishes or disruption, as nature intended a soldier to be."[4]

This is a clear echo of Plato's philosophy of the forms, where a hidden reality is populated with the one true version of each object or idea while the realm of appearances is populated by their imperfect copies. In a way, this is no accident. Quetelet saw his work as a contribution towards an eventual "social physics" in which statistical analysis uncovers not just the true objects of the social world but also the laws that govern their interactions. As such, this was part of an enlightenment programme of scientific discovery that has its roots in Western philosophy and the search for a more real, mathematical world behind the world of appearances.

Thanks in part to his clever use of metaphor – he asks us to imagine 1000 flawed copies of a statue of a gladiator whose average would recover the original – Quetelet's idea of the "average man" caught on, and his statistical laws describing human behaviour (one, for example, shows regular patterns in suicide rates) caused a sensation as the public struggled to reconcile their private and apparently free behaviour at an individual level with the strictures of a deterministic law.

On the surface, this was a revolution. It launched the statistical study of social data which underpins all the social sciences, and Quetelet's direct influence can be traced through thinkers such as Marx, Galton and Wundt to the social sciences of the present day. However, it was an incomplete revolution and the fact that it is unfinished continues to have consequences for the understanding of statistics. For there is a continuity between Quetelet and the eighteenth-century scientists who railed against the use of the average – and this continuity, although in some ways obvious, is easily missed. Quetelet's predecessors had insisted that measurement was always measurement of something. We can think of them as holding up a tape measure to a physical thing and reading off as accurately as they can the precise dimensions of nature. By invoking the ideal average, Quetelet has by no means done away with this picture. Rather he brings these objects in again via the back door. There is still a tape measure being held against an object, only this time the object being measured is its Platonic form.

Timeline of the arithmetic mean, up to Quetelet

Date	Event	Source
428 BC	Thucydides describes the use of the mode for estimating the size of an enemy's defences	Stigler,[2] p. 30
280 BC	Pythagoreans mention the arithmetic mean in the context of music and proportion	Stigler,[2] p. 28
2nd century BC	Between 162 and 127 BC, Hipparchus uses the midrange to combine astronomical measurements	Stahl,[3] p. 98
Beginning of 16th century	Jakob Köbel makes use of an average implicitly in his definition of a 16-foot rod	Stigler,[2] p. 31
End of 16th century	Tycho Brahe uses repeated measurements in astronomy but does not tell us how he combines them	Stahl,[3] p. 98
1600	Kepler's treatment of astronomical data shows there is no agreed method for combining measurements	Stahl,[3] p. 99
1632	Galileo gives the first systematic treatment of random errors	Stigler via Stahl,[3] p. 99
1635	Gellibrand uses the term "arithmetic meane" to refer to a mid-range	Stigler,[2] p. 24
1660 (approx.)	Robert Boyle argues against combining measurements	Stahl,[3] p. 100
1722	Roger Cotes argues for the average as the best estimator and explains the average in physical terms as the centre of gravity	Stahl,[3] p. 100
1749	Euler's failure to solve the problem of three bodies is partly due to his belief that errors should be summed	Stigler,[1] p. 29
1755	Simpson's letter to the Earl of Macclesfield and his experiments with error curves	Stigler,[1] p. 91
1774	Laplace's first error curve	Stahl,[3] p. 102
1777	Daniel Bernoulli's letter shows that averaging has become the norm	Stahl,[3] p. 104
1809	Gauss uses the assumption that the average is the best estimator to derive his normal curve	Stahl,[3] p. 105
1810	Laplace publishes his central limit theorem	Stahl,[3] p. 106

The wrong sized cannon ball

What is wrong with a little Platonism? Surely this is just a harmless metaphysical overlay on an otherwise sound approach to social data? The danger lies in what Todd Rose has called, following Paul Molenaar, "the ergodic switch" (elsewhere it has been called the ecological fallacy). This is the mistake of assuming that the properties of a group can be used to make predictions about the members of that group.

(Ergodic theory is a branch of mathematics that looks at the conditions under which this inference does in fact follow through. The ergodic switch occurs when the leap from group to individual is made in error because these ergodic conditions do not hold.)

Rose gives the example of a hypothetical study of the relationship between typing speed and typing accuracy. Within a group we might see a positive correlation between the two metrics, but this is driven by the fact that the faster typists are the better typists and therefore also the most accurate. To say that I (a terrible typist) should improve my accuracy by typing faster would be to make the unjustified ergodic switch from group to individual.

As Rose correctly points out, in positing the "average man" Quetelet was explicitly pinning the properties of a group onto an individual, even if that individual was ideal rather than actual. It paved the way for many subsequent misunderstandings, which is why Rose calls it the "original sin" at the founding moment of the "Age of Average". In fact, there is more continuity in the story than this. Quetelet has the same picture as his predecessors of what measurement fundamentally is, only he is carrying out his measurement on an ideal object rather than a distant star.

In the case of the arithmetic mean, the price of falling for the ergodic switch was spotted early on. In 1878 John Venn gives the example of two spies providing conflicting reports on the calibre of cannons in a fort about to be captured. One spy says they require cannon balls of 8 inches, the other says 9 inches. The captain of the attacking vessel would like to order the right cannon balls to fire from these cannons once they have captured the fort and are defending it. Should he then order cannon balls of size 8.5 inches?

This same mistake is repeated each time a business asks for the average of a customer attribute – say, customer spend – and then acts on it, even though this attribute may have a multimodal distribution (a distribution with many peaks). There might, for example, be one group of customers who spend very little and another group who spend a lot, with no one occupying the middle ground. Antoine Augustin Cournot, a French philosopher and mathematician, argued that Quetelet's average man would be a monstrosity, pointing out that if you were to average the sides of all right-angled triangles you would not get anything like a right-angled triangle. In the same way, the average customer may end up looking like no one at all, in which case targeting them would be a serious error.

This simple, but surprisingly common, pitfall illustrates the practical implications discussed at the beginning of this article. First, it is naive to assume that any audience will treat an average, presented in this way, as the property of a group. By mentioning the "average customer", we create an almost irresistible pull towards the ergodic switch. Second, by reaching immediately for a summary statistic we have thrown away a huge quantity of individual-level information that might have alerted us to any misleading ideas suggested by the average. (As Todd Rose puts it, we need to "analyse then aggregate", not the other way around.) It is this wastefulness more than anything that would have shocked the scientists of the eighteenth century. They would have been amazed at our blasé and unthinking use of the average, much as we might be at their hostility to the idea. ■

Further reading

In a companion article, "An average understanding", I explore how the ergodic switch is related to, and perhaps explained by, the idea of the "discontinuous mind", proposed by Richard Dawkins. His theory addresses our natural aversion to grey areas (including, I argue, those caused by statistical distributions) and our favouring of discrete and familiar objects, which helps explain why we are so willing to transfer the properties of a crowd, with its uncertain, constantly shifting identity, over to the individual, which is, by contrast, satisfyingly tangible. I also discuss ways in which we might guard against making the ergodic switch. Read the article online at significancemagazine.com/571.

> Cournot argued that Quetelet's average man would be a monstrosity

References

1. Stigler, S. (1986) *The History of Statistics*. Cambridge, MA: Harvard University Press.
2. Stigler, S. (2016) *The Seven Pillars of Statistical Wisdom*. Cambridge, MA: Harvard University Press.
3. Stahl, S. (2006) The evolution of the normal distribution. *Mathematics Magazine*, **79**, 96–113.
4. Rose, T. (2015) *The End of Average*. New York: Harper Collins.