Was regression to the mean really the solution to Darwin's problem with heredity?

Essay Review of Stigler, Stephen M. 2016. *The Seven Pillars of Statistical Wisdom*. Cambridge, Massachusetts: Harvard University Press.

Adam Krashniak and Ehud Lamm(*)

The Cohn Institute for the History and Philosophy of Science and Ideas, Tel Aviv University.
Ramat Aviv, Tel Aviv 6997801, Israel

(*) Corresponding author

Abstract

Statistical reasoning is an integral part of modern scientific practice. In *The Seven Pillars of Statistical Wisdom* Stephen Stigler presents seven core ideas, or pillars, of statistical thinking and the historical developments of each of these pillars, many of which were concurrent with developments in biology. Here we focus on Stigler's fifth pillar, regression, and his discussion of how regression to the mean came to be thought of as a solution to a challenge for the theory of natural selection. Stigler argues that the purely mathematical phenomenon of regression to the mean provides a resolution to a problem for Darwin's evolutionary theory. Thus, he argues that the resolution to the problem for Darwin's theory is purely mathematical, rather than causal. We show why this argument is problematic.

Keywords Francis Galton - Regression to the mean - Mathematical explanations - Biology and statistics - Evolution and inheritance

Introduction

Modern scientific reasoning is largely based on statistical tools and methods. In *The Seven Pillars of Statistical Wisdom* (Harvard University Press, 2016) Stephen Stigler sets forth to present the seven core ideas, or pillars, that are at the base of statistical thinking, in an attempt to give an answer to the question "what exactly is statistics?". Both a professor of statistics and a distinguished historian of the field, Stigler does not only describe each of the seven pillars, but also provides an account of their historical development, with each chapter in the book dedicated to one of the pillars. Philosophers of science and biology will greatly benefit from the book, which clearly explains complicated statistical ideas and their development; ideas that are essential for a good understanding of scientific methodology. Many of the episodes discussed in the book concern the interaction between the development of statistics and the development of biology. In what follows, we briefly present each of the chapters of the book and then discuss at greater length Stigler's description in chapter 5 of the concept of regression to the mean, as well as his arguments regarding the explanatory value of this concept in evolutionary biology. The arguments presented in this chapter are highly relevant to philosophers of biology and relate to the philosophical debate about mathematical explanations in science. After presenting Stigler's analysis we explain where we find it wanting.

The first chapter in Stigler's book deals with the discovery of the notion of data aggregation, that is, the idea that a statistical summary of a collection of data (such as an average) can provide us with more information than can be obtained by simply attending to individual data items. Stigler clearly demonstrates how radical this idea was prior to its acceptance, as it involves the counter-intuitive argument that in order to gain information we should discard parts of the data we gathered (the individual items, the order in which they were gathered, etc.). The second chapter, which deals with the concept of information, presents the development of another counter-intuitive conclusion, namely that the accuracy of our inferences does not increase proportionally with the amount of data we gather. The more data we gather, the less each new piece of data contributes to the accuracy of the investigation. This leads, for example, to the non-intuitive realization that the second 20 observations we make are not as valuable to our inferences as the first 20 observations we made, even though they would have been more valuable had they been the first 20 observations to be gathered (which would render the other 20 observations less valuable).

The third chapter deals with the concept of likelihood, which is necessary when we try to understand whether some data supports or contradicts a belief or a hypothesis we have. To do that, we need to be able to determine whether data that were gathered support or go against a hypothesis. Philosophers will be aware of the discussion of the law of likelihood by Ian Hacking (2016 [1965]) and, more recently, by Deborah Mayo (1996) and Elliott Sober (2008). The most common method in use to determine whether our data supports or contradicts a hypothesis is Null Hypothesis Significance Testing, in which the data is tested against a null hypothesis, and the hypothesis is accepted if the probability of the data given this null hypothesis (its p-value) is low enough (usually less than 0.05). The uncritical acceptance and use of significance tests in many scientific fields has long been known, and the historical development of the concept of likelihood and significance testing provided by Stigler is an accessible way to understand the foundational debates through which these ideas developed. It is particularly enlightening given the current "Replication Crisis" to recall the historical debate between Fisher and Neyman and Egon Pearson.

The fourth chapter deals with intercomparison, the idea that some inferences can be made solely on the basis of an analysis of the interior variation in the data. That is, some inferences can be made without any exterior reference. For example, the p-value of an empirical result can be inferred based on the standard deviation in our sample, without any reference to the exact standard deviation in the entire population. As usual key innovations were made by Francis Galton, Karl Pearson, and Ronald Fisher. Galton used intercomparison in his infamous 1869 book, *Hereditary Genius*, allowing him to use biographical dictionaries to compare talent in populations without having to explicitly define talent.

The fifth chapter deals with the concept of regression. We elaborate on this chapter in the next section. The sixth chapter deals with the design of experiments, and how the planned statistical analysis of results should guide this design. It is common wisdom that well-designed experiments are necessary for a thorough statistical analysis to be possible, and that a statistical analysis cannot fix a study that was not designed well. Stigler's discussion goes further to show how combining the right design and statistical methods yielded new insights that were not available before. Stigler highlights a lovely quote from Fisher who wrote that "Nature… will best respond to a logical and carefully thought out questionnaire; indeed, if we ask her a single question, she will often refuse to answer until some other topic has been discussed." (p. 153). This flies against the older tradition, still very much alive today, which considers good experiments to be ones that ask a single, often dichotomous, question. For example, Stigler demonstrates how additive models helped to recognize the causal effects of factors such as seeds and fertilizers in agriculture by comparing several multi-factor treatments simultaneously (different seeds, different fertilizers) and measuring the variation between all the treatments. These effects would not have been recognized had only treatment for one factor been made, while ignoring the variation in outcome that is due to other factors. Fisher developed these techniques while doing agricultural research at Rothamsted Experimental Station. Another interesting topic Stigler discusses in this chapter is the introduction of randomization to experimental design, which helps making inferences without the need to make assumptions about normality which would otherwise be required. Stigler's discussion of the role of randomization in making statistical inferences valid and even at times in establishing what he calls the "objects of inference" is highly interesting and thought-provoking.

The seventh chapter deals with the concept of residual, the idea that complicated phenomena can be accounted for by first subtracting the effects of known causes and then referring to the remaining effects as those that require further explanation. Stigler shows how the logic of this idea led to the development of statistical methods of comparison of complex models in scientific practice.

While each and every one of the pillars raises interesting philosophical questions, and has an interesting history often related to biology, we find Stigler's discussion of the fifth pillar, regression, to be of special interest to philosophers of biology. We turn now to the issues we have with Stigler's discussion of Francis Galton's work on inheritance and what its implications were for Natural Selection.

Darwin's problem and regression to the mean

The fifth chapter in Stigler's book describes Francis Galton's work on inheritance and his discovery of regression to the mean. We find this chapter of great interest to philosophers of biology for several reasons. First, it describes a problem with Darwin's theory of evolution by natural selection that is much less often recognized than the famous problem articulated by Fleeming Jenkin (i.e., the problem posed to the theory by blending inheritance). Second, Stigler provides a highly informative description of how Galton worked towards a solution to this problem using a physical model to represent and understand the inherited variation of characters across generations. Thus, the chapter can be of interest to those philosophers concerned with modelling. Third, Stigler argues that the solution to the problem is mathematical rather than causal, but we are not quite sure that this is correct (more on this in a minute). This connects directly with the literature on mathematical explanations and their differences and relations to causal explanations. [1]

Stigler starts his discussion of Galton's work with a description of a problem in Darwin's theory of natural selection that only Galton seemed to have recognized. Darwin's theory is built on the core assumption that each parent produces offspring that are not identical to it, and thus creates what Stigler refers to as intergenerational variability. In other words, parents create additional variation in the population when they reproduce. This pattern of inheritance seems to imply that the overall variation in the population will increase with every generation, since each reproduction event adds to the variation in the population. However, this is not what we observe in nature: in most species the distribution of character traits tends to stay the same across generations (even with no apparent selection).

[1] An important debate concerning the relation between mathematical and causal explanations in the philosophy of biology literature is the debate over whether explanations of population change that refer to natural selection and random drift are causal or mathematical explanations (e.g. Walsh et al. 2002; Shapiro & Sober 2007). We will not discuss how our analysis of Stigler's views relates to this debate.

So, while Darwin's assumption about intergenerational variability leads to the conclusion that the variation in characters should increase with each generation, in effect we see that the variation remains quite the same. This of course casts doubt on Darwin's assumption that reproduction events contribute to the variation in the population, without which his theory cannot work, and thus creates a problem for the structure of the theory. Hence, the problem Galton identified (which Stigler refers to as "Darwin's problem" (pp. 130, 131)) is that intergenerational variability and stable variation across generations seem to be in conflict, and need to be reconciled in order for Darwin's theory to work. In other words, to overcome the problem an explanation is needed for the fact that offspring traits are distributed around parental traits, yet the overall population distribution does not change from generation to generation. This problem is related to a more general concern, namely that for natural selection to operate populations need to harbor sufficient variation and such variation must therefore be maintained. Darwin's problem, as Stigler here identifies it, is a special case of this more general concern, since sufficient variation need not be the result of having stable intergenerational variation. Galton tried to find an explanation for intergenerational stability in variation, while offspring traits are distributed around parental ones, by looking for a force that "counteracted the increased variability yet also conformed with heritable intergenerational variation" (p. 115). In other words, he looked for a causal explanation. [2]

Stigler provides a clear and illuminating description of how Galton worked towards a solution to his problem, which we will follow in our discussion. Galton invented and used a device called the quincunx, in which lead balls fall through rows of pins from the top of the device to one of several compartments at the bottom. At each row of

[2] An alternative interpretation of this chapter is that Stigler describes two distinct problems: "Darwin's problem" being the maintenance of sufficient variation in natural populations and "Galton's version of Darwin's problem" being the problem of the apparent conflict between intergenerational variability and stable variation. We thank an anonymous referee for this suggestion. We prefer our reading of the chapter and note that in the section "The Solution to Darwin's Problem", Stigler describes only one problem, which is the problem we focus on here. The alternative interpretation does not fundamentally affect our argument, which is concerned with Galton's solution for the problem of the apparent conflict between intergenerational variability and stable variation in natural populations.

pins, after the ball hits one of the pins it has the same chance of falling to the right or to the left of the pin. This implies that each ball is most likely to reach the bottom compartment which is directly below the location where it was dropped. The balls are less likely to reach the compartments that are further to the left or to the right, and thus after a series of balls are dropped from the center we get a bell-shaped distribution of balls, with the majority of the balls laying at the center compartment and fewer laying at the edges.

The quincunx was used by Galton to represent the variation within a population and to demonstrate how this variation can contribute to the variation in the next generation. We can refer to the compartments at the bottom of the quincunx as representing the values of a given trait, say height, with middle compartments representing medium height, the compartments to the left representing lower heights, and the ones to the right representing higher heights. The balls in each compartment represent the number of individuals that possess each value of the trait, and thus the normal distribution of balls across the compartments after a series of drops from the center of the device represents the normal distribution of height in the population. Furthermore, if we imagine that we release the balls that are now distributed across the bottom compartments for another "round" in the quincunx, we can infer the amount of variation in the following generation. The balls will fall from their current locations and end in new compartments at the bottom, and this represents the distribution of the trait in the following generation.

Now, if we release a population of balls that are distributed normally for a second round in the quincunx, what we get is a larger range of variation at the end of the second round (i.e., in the third generation). The balls at the edges are now quite likely to end even further to the right or to the left (even though they are most likely to end at the same location from which they were dropped). This, of course, is a simple demonstration of what was said earlier about Darwin's assumption, namely that intergenerational variation leads to the expectation of ever increasing variation in the overall population. Galton used the quincunx to try and explain how such intergenerational variation can occur while the overall variation in the population remains the same, and in 1877 he came up with a possible explanation. If we let the balls that are distributed across the compartments fall only from the compartments at the center of the quincunx, the distribution will be compressed before it will be subjected to further variation, and if we compress it enough, the variation will increase only to the same range of variation that exists in the parent generation. He named the factors that are responsible for this compression 'inclined chutes'. However, he did not have a good explanation of what accounted for the occurrence of such a compression in natural populations.

Galton kept searching for evidence for the compression of intergenerational variation in natural populations represented by the inclined chutes. He gathered a large amount of data about heights of parents and children and noticed that the average height of children was not the average parental height, but rather a value that is closer to the average value of the population. In other words, Galton noted that tall fathers have sons who are taller than average, but to a lesser extent than the fathers and hence closer to the population mean. This seemed like evidence for the operation of the inclined chutes, but Galton also noticed that the same pattern is observed when he averaged groups of children, with the parents of each group of children showing an average height that is closer to the population average (p. 123). Furthermore, he discovered that the pattern is observed yet again in data gathered from pairs of brothers, such that for a group of individuals with a given average height, grouping their siblings does not yield the same average height, but once again a height which is closer to the average in the overall population.

Galton concluded that the phenomenon he observed was not a biological phenomenon at all but rather a statistical one, which we now refer to as regression to the mean. Put formally, for a given value of X the predicted value of Y using ordinary least squared regression, is fewer standard deviations from its own mean than X is from its mean.

Stigler goes beyond noting that regression to the mean is a mathematical phenomenon. He argues in addition that regression to the mean solves Darwin's problem presented above. [3] Stigler does not provide an account of what makes an explanation a mathematical explanation, as opposed to a causal one, and in what sense is regression to the mean a mathematical phenomenon[4]. But recent years have seen a growing literature on the nature of mathematical explanations, which can shed light on this question. There is somewhat of a consensus among different accounts of mathematical explanations that explanations using regression to the mean are indeed mathematical. André Ariew, Collin Rice and Yasha Rohwer argue that an explanation is mathematical if the explanandum can be deduced as a consequence of some mathematical facts without citing any specific causes (Ariew et al. 2015).[5] Thus, these authors argue that Galton's explanation using regression to the mean is mathematical because it shows how the existence of stable variation and intergenerational variability can be deduced from mathematical parameters which determine a normal distribution (Ariew et al. 2015, p. 645). Similar arguments regarding the mathematical nature of Galton's explanation using regression to the mean were made by Sober (1980) and Hacking (1990), whom these authors cite as well. Marc Lange provides a different account of mathematical explanations, according to which an explanation is mathematical if the explanandum is shown to be the result of mathematical facts, *and* these facts seem to account for the explanandum to a stronger degree than any causal facts could account for it (Lange 2013a). Lange also takes explanations using regression to the mean to be mathematical (he claims they belong to a sub-type of mathematical explanations he calls "Really Statistical" explanations), because these explanations show that the explanandum is simply "a statistical fact of life" (Lange 2013b, p. 173).

[3]It is important to note that this is Stigler's argument, not necessarily Galton's. We are not providing an interpretation of Galton's work in this paper, but rather analyze Stigler's interpretation in his analysis of Galton's work.

[4] From this point on, whenever we speak of mathematical explanations we mean pure mathematical explanations. That is, explanations that are mathematical rather than causal.

[5] These authors take explanations for population changes that appeal to natural selection or drift to be mathematical and not causal because these changes "can be explained by referring to the deductive consequences of statistical models, independent of considerations of causation" (Ariew et al. 2015, p. 636). They note that to be explanatory, it is not sufficient that the explanandum will simply be deduced from the mathematical facts. The mathematical facts need to also provide counterfactual information, by telling us "how things would have been different in various counterfactual situations" (Ariew et al. 2015, p. 655).

While Ariew et al, Sober, and Hacking describe the mathematical nature of Galton's explanation using regression to the mean, none of them discusses whether this explanation resolves Darwin's problem. However, Stigler in the book goes further, to argue that regression to the mean is the resolution of Darwin's problem, and thus that the resolution to this problem is purely mathematical. According to Stigler, the apparent conflict between stable variation across generations and Darwin's requirement of intergenerational variability was resolved once Galton discovered that the two can coexist due to regression to the mean. In a section titled "The Solution to Darwin's Problem" Stigler writes: "The problem Galton had *identified* was not a problem after all, but was instead due to a statistical effect that no one had identified before. Population equilibrium [i.e. stable variation] and intergenerational variability were not in conflict" (p. 130, our italics). Stigler further argues, in regards to Darwin's problem, that "…[Galton] showed that, properly understood, there was no problem" (p. 131). In other words, Stigler argues that once regression to the mean was discovered, Darwin's problem turned out to be a pseudo problem, since evidence for both stable variation and intergenerational variability in a population were not contradictory.[6]

However, we find it hard to see how regression to the mean by itself can fully explain the coexistence of intergenerational variability and stable variation and resolve Darwin's problem. As the quotes in the paragraph above indicate, Stigler argues that the problem was solved because the discovery of regression to the mean showed that stable variation and intergenerational variability were not in conflict. But what was really shown with the discovery of regression to the mean, and we believe Stigler will agree on that, is that stable variation and intergenerational variability were not *necessarily* in conflict. That is, regression to the mean is a phenomenon that includes the co-occurrence of stable variation and intergenerational variability, implying that the two can co-occur. But this means that Galton's discovery of regression to the mean only suggested that an explanation for Darwin's problem is *possible*, it did not solve it. In order to be a full resolution to Darwin's problem, it

[6] The argument made here by Stigler, namely that regression to the mean solved Darwin's problem, did not appear in his earlier works on Galton and Darwin's problems (e.g. Stigler 2010). We thank an anonymous referee for pointing this out.

must be clarified why an explanation using regression to the mean in fact applies to natural populations.

While it is a general phenomenon, deducing regression to the mean requires making statistical assumptions, in particular about the distributions involved. It is easy to imagine a situation in which there is intergenerational variation between parents and offspring in such a way that there is no stable variation and regression to the mean does not occur. This is the case in the ever-expanding range of variation case Stigler begins his discussion with, as well as in blending inheritance, probably the objection that most worried Darwin. Moreover, stable variation entails regression to the mean, but regression to the mean does not entail stable variation. Once you have stable variation, that is the parental and offspring generations have the same distribution of traits, the central assumption for mathematically deducing regression to the mean is satisfied. Not the other way around.

Thus, in order to convincingly conclude that regression to the mean is the explanation why stable variation and intergenerational variability which are observed in natural populations co-occur it is not enough to understand the phenomenon of regression to the mean. One must also show why the preconditions for the occurrence of regression to the mean apply to the populations involved. In the case of the natural populations that are relevant to Darwin's problem, some causal account of how traits are inherited must be given since, as noted above, traits may be inherited in ways that would not make it possible to explain away Darwin's problem by appeal to regression to the mean. Thus, Darwin's problem was resolved and received a satisfactory explanation once the Mendelian account of genetic inheritance explained intergenerational variability and stable variation. Hence, we argue, the resolution to this problem was not purely mathematical.

It seems more enlightening to understand Galton's work as elucidating the mathematical conditions of correlation between parents and offspring and regression to the mean; conditions that shed light on properties the causal inheritance systems should possess. But this of course does not contradict the fact that only a good understanding of these causal details of inheritance could provide a full explanation of the coexistence of intergenerational variability and stable variation and resolve Darwin's problem. Thus, Galton's mathematical explanation should not be thought of as replacing the causal explanation focusing on the inheritance system. The two explanations are not independent of one another; each sheds light and constrains the other.[7]

How does this interpretation sit with the correct and well known observation that regression to the mean is a statistical phenomenon not a causal one? The thing to note is that the question of predicting traits of offspring from those of parents, which is what Galton was studying, is indeed a statistical question. Galton discovered that regression to the mean does not require an independent causal explanation, being a manifestation of how the statistical expectation is derived. Darwin's problem as defined above is different. It deals with actual multi-generational variation. What Darwin's problem amounts to is how genetic inheritance produces intergenerational variability and stable variation. This is a causal question. Galton's problem, in contrast, involved the interpretation of evidence. What seemed like contradictory conclusions from collected data turned out not to be contradictory. This conundrum had a mathematical explanation.

Concluding remarks

Stigler's book provides a valuable account of the core ideas in statistics and their historical development. Anyone interested in statistical reasoning and more generally in modern scientific practice will find the themes in the book highly interesting. Furthermore, the discussion of regression to the mean and Darwin's problem in chapter five is highly relevant to the philosophical discussions on mathematical explanations. We objected to Stigler's argument that regression to the mean provides a full resolution to Darwin's problem, which implies that this resolution is mathematical and not causal. We tried to show why a full resolution to Darwin's problem must take into account the causal details of inheritance. Stigler's lucid presentation helps see the issues at stake more clearly.

[7] This line of thought is compatible with recent work by Andersen (forthcoming) who suggests that mathematical and causal descriptions should be thought of as complementary explanations of phenomena.

Whether the resolution to Darwin's problem is purely mathematical or not, and the little we can say here surely does not exhaust this debate, what is beyond dispute, and is beautifully presented by Stigler, is that a biological problem that concerned Galton led to an important discovery in the field of statistics. This, like many other examples in Stigler's book, reminds us of the tightly linked history of the fields of biology and statistics.

Stigler reminds us how little use Darwin himself had for mathematics. In an 1855 letter to William Darwin Fox he proclaimed "*I have no faith in anything short of actual measurement and the Rule of Three,*" the Rule of Three being an arithmetic rule taught to elementary school kids: if *a* is to *b*, as *c* is to *d*, *a* can be deduced if *b,c*, and *d* are known. But, as Stigler explains, even the faith in the Rule of Three was misplaced. The Rule of Three fails whenever there is variation and correlation. Reflecting on Stigler's history seems to us to suggest that Darwin the naturalist, with his concentrated focus on variation and ecological interactions, had good reason to be suspicious of mathematics. The mathematics of his day was not up to the task of supporting the scientific program Darwin played a key role in. Variation and interactions were not as well understood mathematically as they are today.

Another source suspicion, in the generations following Darwin, came from the limited biological insight offered by various idealized models, as recounted in Evelyn Fox Keller's 2003 *Making sense of life: Explaining biological development with models, metaphors, and machines*. Reading Keller's book together with Stigler's book is a good way to encourage students to reflect on the relations between biology and mathematics and physics (see Lamm, 2013 for further discussion of these relations). Mathematics, most prominently in the form of Statistics, developed in the years following Darwin's work, at least in part with the influence of biological questions. Understanding this rich history is helped significantly by reading this enjoyable and enlightening book.

References

Andersen, H. (forthcoming). Complements, not competitors: causal and mathematical explanations. *The British Journal for the Philosophy of Science*.

Ariew, A., Rice, C., & Rohwer, Y. (2015). Autonomous-statistical explanations and natural selection. *The British Journal for the Philosophy of Science*, *66*(3), 635-658.

Hacking, I. (1990). *The taming of chance* (Vol. 17). Cambridge University Press, New York.

Hacking, I. (2016). *Logic of statistical inference*. Cambridge University Press, New York.

Keller, E. F. (2003). *Making sense of life: Explaining biological development with models, metaphors, and machines*. Harvard University Press, Cambridge.

Lamm, E. (2013). Theoreticians as Professional Outsiders: The Modeling Strategies of John von Neumann and Norbert Wiener. In *Oren Harman and Michael R. Dietrich (eds.), Outsider Scientists: Routes to Innovation in Biology.* University of Chicago Press, Chicago, pp 291-312.

Lange, M. (2013a). What makes a scientific explanation distinctively mathematical?. *The British Journal for the Philosophy of Science*, *64*(3), 485-511.

Lange, M. (2013b). Really statistical explanations and genetic drift. *Philosophy of Science*, *80*(2), 169-188.

Mayo, D. G. (1996). *Error and the growth of experimental knowledge*. University of Chicago Press, Chicago.

Shapiro, L., & Sober, E. (2007). Epiphenomenalism--the Do's and the Don'ts. In Machamer, P. K., & Wolters, G. (eds.), *Thinking about causes: from greek philosophy to modern physics*. University of Pittsburgh Press, Pittsburgh.

Sober, E. (1980). Evolution, population thinking, and essentialism. *Philosophy of Science*, *47*(3), 350-383.

Sober, E. (2008). *Evidence and evolution: The logic behind the science*. Cambridge University Press, New York.

Stigler, S. M. (2010). Darwin, Galton and the statistical enlightenment. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *173*(3), 469-482.

Walsh, D. M., Lewens, T., & Ariew, A. (2002). The trials of life: Natural selection and random drift. *Philosophy of Science*, *69*(3), 429-446.