Chapter 2. From Buccaneers to Guinea Pigs: The Genesis of Causal Inference

And yet it moves. – Attributed to Galileo Galilei (1564-1642)

For close to two centuries, one of the most enduring rituals in British science has been the Friday Evening Discourse at the Royal Institution of Great Britain, in London. This was where many discoveries of the nineteenth century were first announced to the public: Michael Faraday and the principles of photography in 1839; J. J. Thomson and the electron in 1897; James Dewar and the liquefaction of hydrogen in 1904.

Pageantry was an important part of the occasion; it was literally science as theater, and the audience, the cream of British society, were expected to be dressed to the nines (tuxedos with black tie for men). At the appointed hour, a chime would strike and the evening's speaker would be ushered into the auditorium. Traditionally he would begin the lecture immediately, without introduction or preamble. Experiments and live demonstrations were part of the spectacle.

On February 9, 1877, the evening's speaker was Francis Galton, F.R.S., first cousin of Charles Darwin, noted African explorer, inventor of fingerprinting, and the very model of a Victorian gentleman scientist. Galton's topic was "Typical Laws of Heredity." His experimental apparatus for the evening was a curious contraption that he called a quincunx, but is now often called a Galton board. A similar game has often appeared on the televised game show *The Price is Right*, where it is known as Plinko. The Galton board consists of a triangular array of pins or pegs, into which small metal balls can be inserted through an opening at the top. The balls bounce downward from one row to the next, pinball style, before settling into one of a line of slots at the bottom. (See Frontispiece.) For any individual ball, the zigs and zags to the left or right look completely random. However, if you pour a lot of balls into the Galton board, a

Frontispiece. Sir Francis Galton demonstrates his "Galton board" or "quincunx" at the Royal Institution. He saw this pinball-like apparatus as an analogy for the inheritance of genetic traits like stature. The pinballs accumulate in a bell-shaped curve that is similar to the distribution of human heights. The puzzle of why human heights don't spread out from one generation to the next, as the balls would, led him to the discovery of "regression to the mean." *(Drawing by Dakota Harr.)*

startling regularity emerges: the accumulated balls at the bottom will always form a rough approximation to a bell-shaped curve. The slots nearest the center will be stacked high with balls, and the number of the balls in each slot gradually tapers down to zero at the edges of the quincunx.

This pattern has a mathematical explanation. The path of any individual ball is like a sequence of independent coin flips. Each time a ball hits a pin, it has to bounce either to the left or right, and from a distance its choice seems completely random. The sum of the results—say, the excess of the rights over the lefts—determines which slot the ball ends up in. According to the Central Limit Theorem, proven in 1810 by Pierre-Simon Laplace, any such random process—one that amounts to a sum of a large number of coin flips—will lead to the same probability distribution, called the *normal distribution* (or bell-shaped curve). The Galton board is simply a visual demonstration of Laplace's theorem.

The Central Limit Theorem is truly one of the miracles of nineteenth-century mathematics. Think about it: Even though the path of any individual ball is unpredictable, the path of 1000 balls is extremely predictable. This fact is convenient for the producers of *The Price is Right*, because they can predict accurately how much money the contestants are going to win at Plinko over the long run. This is the same law that makes insurance companies extremely profitable, despite the uncertainties in human affairs.

What does this have to do with the laws of heredity? After all, the title of the lecture promised heredity, so surely the well-dressed audience at the Royal Institute must have wondered. To answer the question, Galton showed them some data collected in France on the heights of military recruits. These also follow a normal distribution: many men are about of average height, with a gradually diminishing number who are either extremely tall or extremely

short. In fact, it does not matter whether you are talking about 1000 military recruits or 1000 balls in the Galton board: the numbers in each slot (or height category) are almost the same.

Thus, to Galton, the quincunx was a model for the inheritance of stature or, indeed, many other genetic traits. It is a causal model. In simplest terms, Galton believed the balls "inherit" their position in the quincunx in the same way that humans inherit their stature.

But if we accept this model—provisionally—it poses a puzzle, which was Galton's chief topic for the evening. The width of the bell-shaped curve depends on the number of rows of pegs placed between the top and the bottom. Suppose that we doubled the number of rows. This would be a model for two generations of inheritance, the first half of the rows representing the first generation and the second half representing the second. You would inevitably find more variation after the second generation than the first, and in succeeding generations, the bell-shaped curve would get wider and wider still.

But this is not what happens with actual human stature! In fact, the width of the distribution of human heights stays relatively constant over time. We didn't have nine-foot humans a century ago, and we still don't. This was an enigma that Galton had been puzzling over for roughly eight years, since the publication of his book *Hereditary Genius* in 1869: What explains the stability of the population's genetic endowment?

As the title of the book suggests, Galton's true interest was not carnival games or human stature, but *human intelligence*. As the member of an extended family with a remarkable amount of scientific genius, it's natural that Galton would have liked to prove that genius ran in families. And that was exactly what he had set out to do in his book. He painstakingly compiled pedigrees of 605 "eminent" Englishmen from the preceding four centuries. But he found that the sons and fathers of these eminent men were somewhat less eminent, and the grandparents and grandchildren less eminent still.

It's easy enough for us now to call Galton's book pseudo-scientific poppycock. What, after all, is the definition of eminence? And did he not realize that people in eminent families might have been successful because of their privilege, rather than because of their talent? Even though critics of his book pointed this out, Galton remained remarkably oblivious to the possibility.

Still, Galton was onto something, which became more apparent once he started looking at features like height, which are easier to measure and more strongly linked to heredity than "eminence." Sons of tall men tend to be taller than average—but not as tall as their fathers. Sons of short men tend to be shorter than average—but not as short as their fathers. Galton first called this phenomenon "reversion," and later "regression toward mediocrity." It can be noted in many other settings. If students take two different standardized tests on the same material, the ones who scored high on the first test will usually score higher than average on the second test, but not as high as they did the first time. This phenomenon of regression to the mean is ubiquitous in all facets of life, education, and business. For instance, in baseball the Rookie of the Year (a player who does unexpectedly well in his first season) often hits a "sophomore slump," in which he does not do quite as well.

Galton didn't know all of this, and he thought he had stumbled onto a law of heredity rather than a law of statistics. He believed that *there must be some cause* of regression to the mean, and in his Royal Institution lecture he illustrated his point. He showed his audience a two-layered quincunx (Figure 1). After passing through the first array of pegs, the balls passed through sloping chutes that moved them closer to the center of the board. Then they would pass through a second array of pegs. Galton showed triumphantly that the chutes exactly compensated for the tendency of the normal distribution to spread out. This time, the bell-shaped probability distribution kept a constant width from generation to generation.

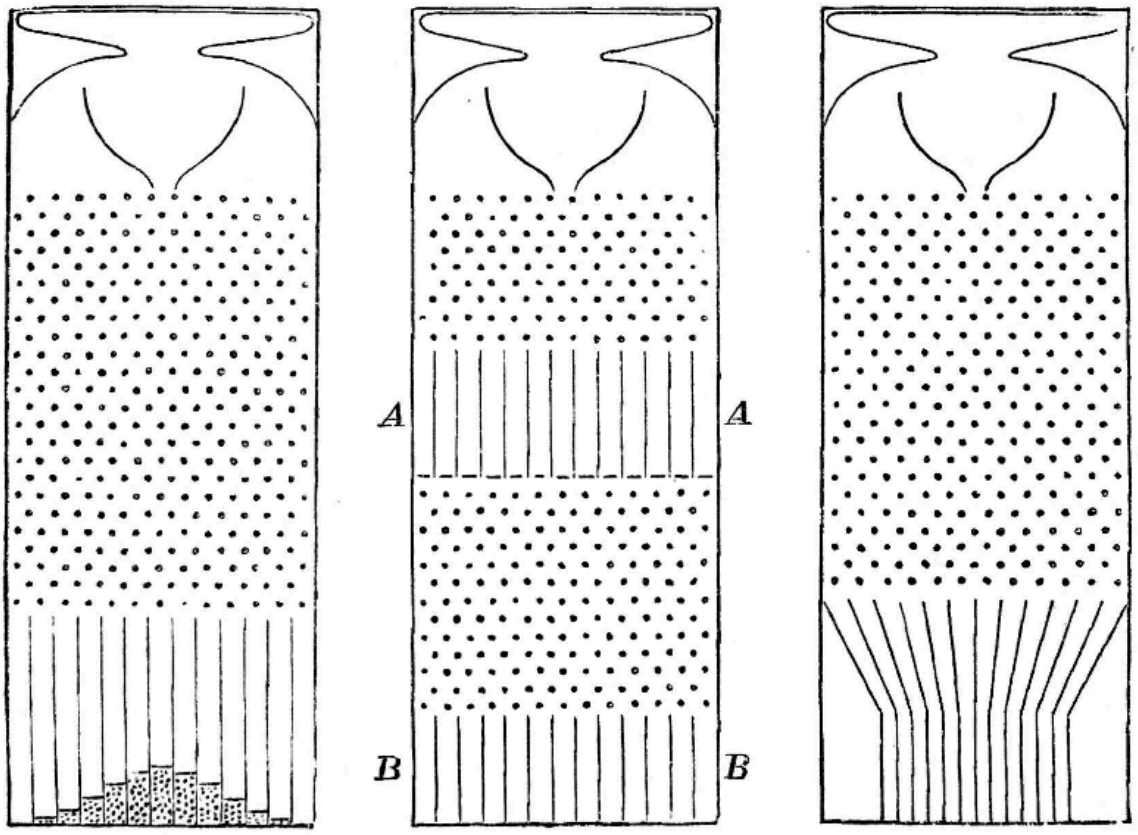

Figure 1. The Galton board, used by Francis Galton as an analogy for the inheritance of human heights. (a) When many balls are dropped through the pinball-like apparatus, their random bounces cause them to pile up in a bell-shaped curve. (b) Galton noted that on two passes A and B through the Galton board (the analogue of two generations) the bell-shaped curve would get wider. (c) To counteract this tendency, he installed chutes to move the "second generation" back closer to the center. The chutes are Galton's causal explanation for regression to the mean.

Thus, Galton conjectured, regression toward the mean was a physical process, nature's way of ensuring that the distribution of height (or intelligence) remained the same from generation to generation. "The process of reversion cooperates with the general law of deviation," Galton told his audience. He compared it to Hooke's law, the physical law that describes the tendency of a spring to return to its equilibrium length.

Keep in mind the date: in 1877, Galton was in pursuit of a causal explanation and thought that regression to the mean was a causal process, like a law of physics. He was mistaken, but he was far from alone. Many people continue to make the same mistake to this day. For example, baseball experts always look for causal explanations of a player's sophomore slump. "He's gotten overconfident," they complain, or "the other players have figured out his weaknesses." They may be right, but the sophomore slump *does not need* a causal explanation. It will happen more often than not by the laws of chance alone.

The modern statistical explanation is quite simple. As Daniel Kahneman summarizes it in his book *Thinking, Fast and Slow*: "Success = talent + luck. Great success = a little more talent + a lot of luck." A player who wins Rookie of the Year is probably more talented than average, but he also (probably) had a lot of luck. Next season, he is not likely to be so lucky, and his batting average will be lower.

By 1889, Galton had figured this out, and in the process—partly disappointed but also fascinated—he took the first huge step toward divorcing statistics from causation. His reasoning is subtle but it is worth making the effort to understand it. It is the newborn discipline of statistics uttering its first cry.

Galton had started gathering a variety of "anthropometric" statistics: height, forearm length, head length, head width, and so on. He noticed that when he plotted height against forearm length, for instance, the same phenomenon of regression to the mean took place. Tall men usually had longer than average forearms—but not as far above average as their height. Here it's clear that height is not a cause of forearm length, or vice versa. If anything, they are both caused by other factors: the man's genetic inheritance. Galton started using a new word for this kind of relationship: height and forearm length were "co-related." Eventually, he gave this up in favor of the more normal English word "correlated."

Later he realized an even more startling fact: when comparing generations, the temporal order could be reversed! That is, the fathers of sons *also* revert to the mean. The father of a son who is taller than average is likely to be shorter than his son. (See Figure 2.) Once Galton realized this, he had to give up any idea of a causal explanation for regression, because there is no way that the sons' heights could be a cause of the fathers' heights.

This realization may sound paradoxical at first. "Wait!" you're saying. "You're telling me that tall dads usually have shorter sons, and tall sons usually have shorter dads. How can both of those statements be true? How can a son be both taller and shorter than his father?"

The answer is that we are talking not about an individual father and an individual son, but two *populations*. We start with the population of 6-foot fathers. Because they are taller than average, their sons will regress towards the mean; let's say their sons average 5 feet 11 inches. However, the population of *father-son pairs with 6-foot fathers* is not the same as the population of *father-son pairs with 5-foot-11 inch sons*. Every father in the first group is by definition 6 feet tall. But the second group will have a few fathers who are taller than 6 feet and a lot of fathers who are shorter than 6 feet. Their average height will be shorter than 5-feet-11, again displaying regression to the mean.

Another way to illustrate regression is to use a diagram called a scatter plot (Figure 2). Each father-son pair is represented by one dot, with the x-coordinate being the father's height and the y-coordinate being the son's height. So a father-son pair who are both 5′9″ (or 69 inches) will be represented by a dot at (69, 69), right at the center of the scatter plot. A father who is 6′ (or 72 inches) with a son who is 5′11″ (or 71 inches) will be represented by a dot at (72, 71), in the northeast corner of the scatter plot. Notice that the scatter plot has a roughly elliptical shape—a fact that was crucial to Galton's analysis, and characteristic of normal distributions with two variables.

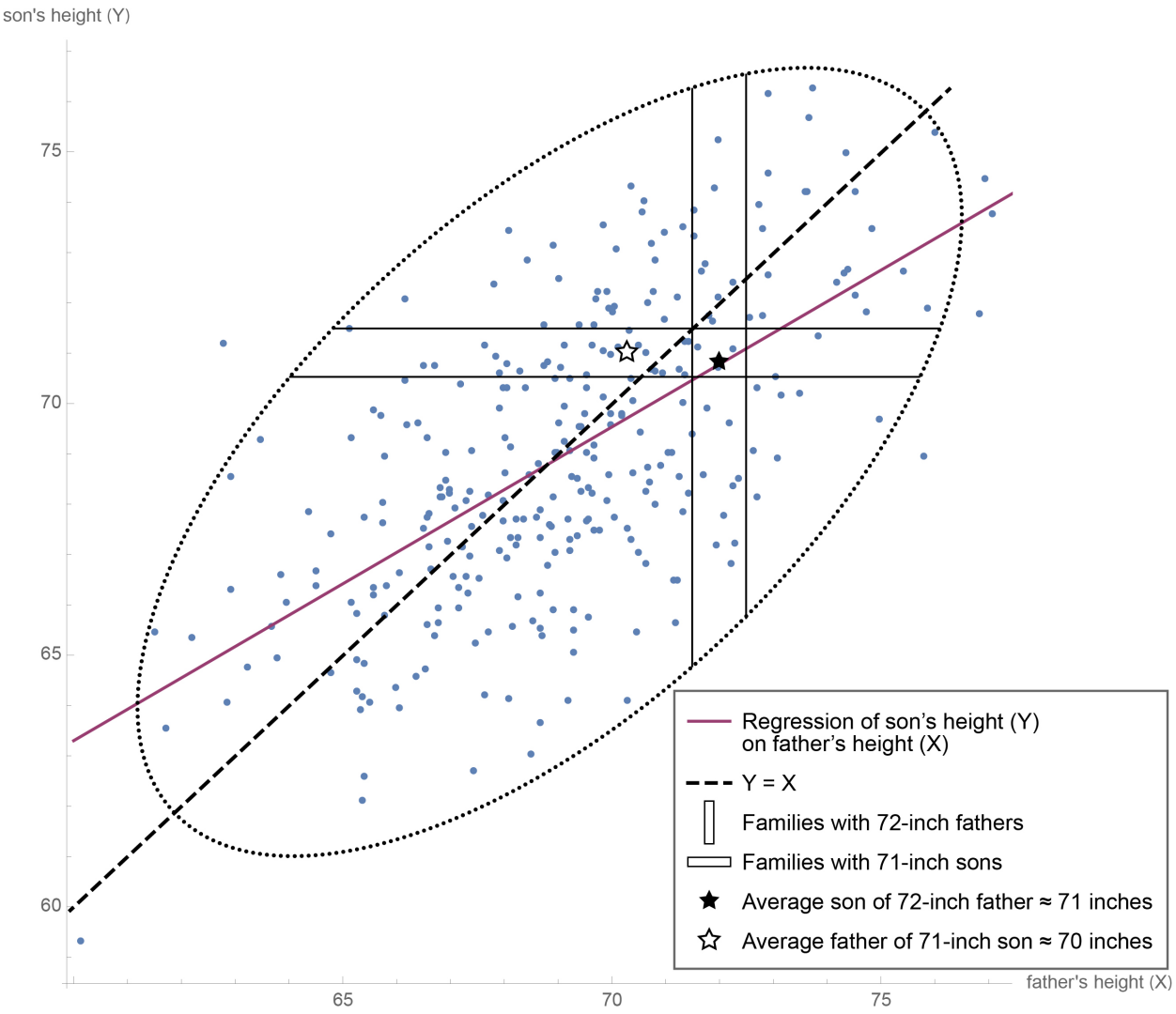

Figure 2. The scatterplot shows a dataset of heights, with each dot representing the height of a father (on the X-axis) and his son (on the Y-axis). The dashed line coincides with the major axis of the ellipse, while the solid line (called the *regression line*) connects the rightmost and leftmost points on the ellipse. The difference between them accounts for regression to the mean. For example, the black star shows that 72-inch fathers have, on the average, 71-inch sons. (That is, the average height of all the data points in the vertical strip is 71 inches.) The horizontal strip and white star show that the same loss of height occurs in the non-causal direction (backward in time). *(Figure by Maayan Harel, with a contribution from Christopher Boucher.)*

As shown in Figure 2, the father-son pairs with 72-inch fathers lie in a vertical slice centered at 72; the father-son pairs with 71-inch sons lie in a horizontal slice centered at 71. Here is visual proof that these are two different populations! If we focus only on the first population, the pairs with 72-inch fathers, we can ask, "How tall are the sons on average?" It's the same as asking where the center of that vertical slice is, and by eye you can see that the center is about 71. If we focus only on the second population with 71-inch sons, we can ask, "How tall are the fathers on average?" This is the same as asking for the center of the horizontal slice, and by eye you can see that its center is about 70.3.

We can go farther and think about doing the same procedure for *every* vertical slice. That's equivalent to asking: For fathers of height x, what is the best prediction of the son's height (y)? Alternatively, we can take each horizontal slice and ask where its center is: for sons of height y, what is the best "prediction" (or retrodiction) of the father's height?

As he thought about this question, Galton stumbled upon an important fact: *the predictions always fall on a line*, which he called the regression line, which is less steep than the major axis (or axis of symmetry) of the ellipse (Figure 3). In fact there are two such lines, depending on which variable is being predicted and which is being used as evidence. You can predict the son's height based on the father's or the father's based on the son's. The situation is completely symmetric. Once again this shows that where regression to the mean is concerned, there is no difference between cause and effect.

The slope of the regression enables you to predict the value of one variable, given that you know the other. In the context of Galton's problem, a slope of 0.5 would mean that each extra inch of height for the father would correspond, on average, to an extra half-inch for the son, and vice versa. A slope of 1 would be perfect correlation, which means every extra inch for the

father is passed deterministically to the son, who would also be an inch taller. The slope can never be greater than 1; if it were, the sons of tall fathers would be taller on average, and the sons

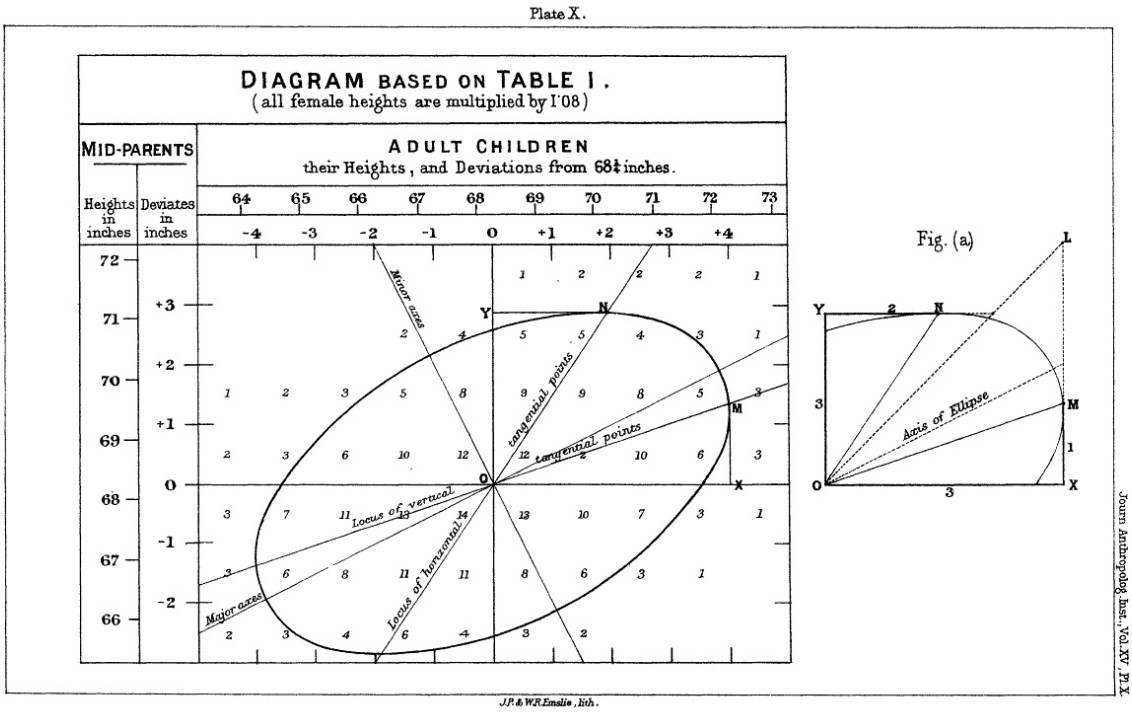

Figure 3. Galton's regression lines. Line OM gives the best prediction of a son's height if you know the height of the father; line ON gives the best prediction of a father's height if you know the height of the son. Neither is the same as the major axis (axis of symmetry) of the scatterplot.

of short fathers would be shorter—and this would force the distribution of heights to become wider over time. After a few generations we would start having nine-foot people and two-foot people, which is not what is observed in nature. So, provided the distribution of heights stays the same from one generation to the next, the slope of the regression line cannot exceed 1.

The law of regression applies even when correlating two different quantities, like height and IQ. If you plot one quantity against the other in a scatter plot, and rescale the two axes properly, then the slope of the best-fit line always enjoys the same properties. It equals 1 only

when one quantity can predict the other precisely; it is 0 whenever the prediction is no better than a random guess. The slope (after scaling) is the same no matter whether you plot *X* against *Y* or *Y* against *X*. In other words, the slope is completely agnostic as to cause and effect. One variable could cause the other, or they could both be effects of a third cause; for the purpose of prediction, it does not matter.

For the first time, Galton's idea of correlation gave an objective measure, independent of human judgment or interpretation, of how two variables are related to one another. The two variables can stand for height, intelligence, or income; they could stand in causal, neutral, or reverse-causal relation. The correlation will always reflect the degree of cross predictability between the two variables. Galton's disciple Karl Pearson later derived a formula for the slope of the (properly rescaled) regression line and called it the *correlation coefficient*. This is still the first number that statisticians all over the world compute when they want to know how strongly two different variables in a data set are related. Galton and Pearson must have been thrilled to find such a universal way of describing the relationships between random variables. For Pearson, especially, the slippery old concepts of cause and effect seemed outdated and unscientific, compared to the mathematically clear and precise concept of a correlation coefficient.

The Abandoned Search

It is an irony of history that Galton had started out in search of causation and ended up discovering correlation, a relationship that is oblivious to causation. Even so, hints of causal thinking remained in his writing. "It is easy to see that correlation [between the sizes of two organs] must be the consequence of the variations of the two organs being partly due to common causes," he wrote in 1889.

The first sacrifice on the altar of correlation was Galton's elaborate machinery to explain the stability of the population's genetic endowment. The quincunx simulated the creation of variations in height and their transmission from one generation to the next. But Galton had to invent the inclined chutes in the quincunx specifically for the purpose of reining in the ever-growing diversity in the population. Having failed to find a satisfactory biological mechanism to account for this restoring force, Galton simply abandoned the effort after eight years and turned his attention to the siren song of correlation. Historian Steven Stigler, who has written extensively about Galton, noticed this sudden shift in Galton's aims and aspirations: "What was silently missing was Darwin, the chutes, and all the 'survival of the fittest.' … In supreme irony, what had started out as an attempt to mathematize the framework of the *Origin of Species* ended with the essence of that great work being discarded as unnecessary!"

But to us, in the modern era of causal inference, the original problem remains. How do we explain the stability of the population, despite Darwinian variations that one generation bestows on the next?

Looking back on Galton's machine in the light of causal diagrams, the first thing I notice is that the machine was wrongly constructed! The ever-growing dispersion, which begged Galton for a counter-force, should never have been there in the first place. Indeed, if we trace a ball dropping from one level to the next in the quincunx, we see that the displacement at the next level inherits the sum total of variations bestowed upon it by all the pegs along the way. This stands in blatant contradiction to Kahneman's equations:

Success = talent + luck
Great success = A little more talent + a lot of luck.

According to these equations, success at generation 2 does not inherit the luck of generation 1. Luck, by its very definition, is a transitory occurrence, hence it has no impact on future

generations. But such transitory behavior is incompatible with Galton's machine!

To compare these two conceptions side by side, let us draw their associated causal diagrams. In Figure 9a (Galton's conception), success is transmitted across generations, and luck

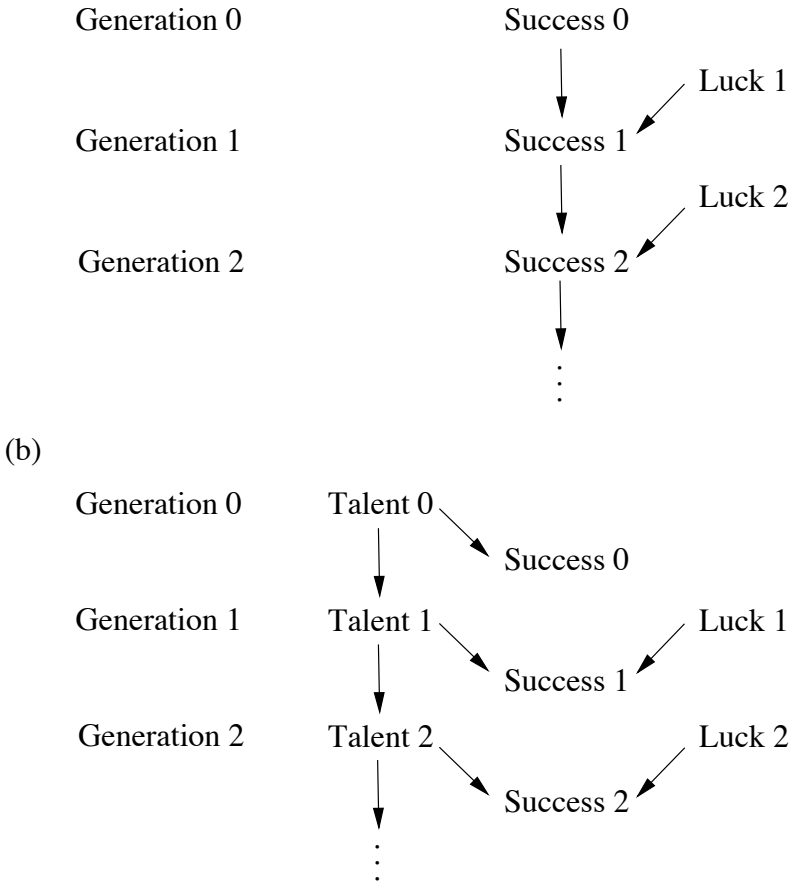

Figure 4. Two models of inheritance. (a) The Galton board model, in which luck accrues from generation to generation, leading to an ever-wider distribution of success. (b) A genetic model, in which luck does not accrue, leading to a constant distribution of success.

variations accumulate indefinitely. This is perhaps natural if "success" is equated to wealth or eminence. However, for the inheritance of physical characteristics like stature, Galton's model needs to be replaced by the model of Figure 9b. Here only the genetic component, shown here as Talent, is passed down from one generation to the next. Luck affects each generation

independently, in such a way that the chance factors in one generation have no way of affecting later generations, either directly or indirectly.

Both of these models are compatible with the bell-shaped distribution of heights. But the first model is not compatible with the stability of the distribution of heights (or Success). The second model, on the other hand, shows that to explain the stability of Success from one generation to the next, we only need to explain the stability of the genetic endowment of the population (Talent). That stability, now called the *Hardy-Weinberg equilibrium*, received a satisfactory mathematical explanation in the work of G.H. Hardy and Wilhelm Weinberg in 1908. And yes, they used yet another causal model—the Mendelian theory of inheritance.

In retrospect, Galton could not have anticipated the work of Mendel, Hardy, and Weinberg. In 1877, when Galton gave his lecture, Gregor Mendel's work of 1866 had been forgotten (it was only rediscovered in 1900); and the mathematics of Hardy and Weinberg's proofs would likely have been beyond him. But it is interesting to note how close he came to finding the right framework, and also how the causal diagram makes it easy to zero in on his mistaken assumption: the transmission of luck from one generation to the next. Unfortunately, he was led astray by his beautiful but flawed causal model, and later, having discovered correlation, he came to believe that causality was no longer needed.

As a final personal comment to Galton's story, l confess to committing a cardinal sin of history writing, one of many sins I will commit in this book. In the 1960s, it became unfashionable to write history from the viewpoint of a modern-day science, as I have done above. "Whig history" was the epithet used to mock the conventional style of writing history, which focused on the successful theories and experiments and gave little credit to failed theories and dead ends. The modern style of history writing became more democratic, treating chemists

and alchemists with equal respect, and insisting on understanding all events in the social contest of their own time.

However, when it comes to explaining the expulsion of causality from statistics, I accept the mantle of Whig historian with pride. There simply is no other way to understand how statistics became a model-blind data-reduction enterprise, except by putting on our causal lenses and re-telling the stories of Galton and Pearson in the light of the new science of cause and effect. In fact, by so doing, I rectify the distortions introduced by mainstream historians who, lacking causal vocabulary, marvel at the invention of correlation and fail to note its casualty—the death of causation.

The Zealot

It remained to Galton's disciple, Karl Pearson, to complete the task of expunging causation from statistics. Yet even he was not entirely successful…

Reading Galton's *Natural Inheritance* was one of the defining moments of Pearson's life: "I felt like a buccaneer of Drake's days—one of the order of men 'not quite pirates, but with decidedly piratical tendencies,' as the dictionary has it!" he wrote in 1934. "I interpreted … Galton to mean that there was a category broader than causation, namely correlation, of which causation was only the limit, and that this new conception of correlation brought psychology, anthropology, medicine and sociology in large part into the field of mathematical treatment. It was Galton who first freed me from the prejudice that sound mathematics could only be applied to natural phenomena under the category of causation."

In Pearson's eyes, Galton had enlarged the vocabulary of science. Causation was reduced to nothing more than a special case of correlation (namely, the case where the correlation

coefficient is 1 or -1, and the relationship between *x* and *y* is deterministic). He expresses his view of causation with great clarity in his book *The Grammar of Science* (1892):

> *That a certain sequence has occurred and reoccurred in the past is a matter of experience to which we give expression in the concept causation… Science in no case can demonstrate any inherent necessity in a sequence, nor prove with absolute certainty that it must be repeated.*

To summarize, causation for Pearson is only a matter of repetition, and causation in the deterministic sense can never be proven. As for causality in a non-deterministic world, Pearson was even more dismissive: "the ultimate scientific statement of description of the relation between two things can always be thrown back upon … a contingency table." In other words, data is all there is to science. Full stop. To Pearson, the notions of intervention and counterfactuals that we discussed in Chapter 1 do not exist, and the Ladder of Causation has only one rung.

The mental leap from Galton to Pearson is breathtaking, and indeed worthy of a buccaneer. Galton had proved only that one phenomenon—regression to the mean—did not require a causal explanation. Now Pearson was completely removing causation from science. What made him take this leap?

Here I have to turn to my UCLA colleague, historian Ted Porter, whose biography *Karl Pearson* describes how Pearson's skepticism about causation predated his reading of Galton's book. Pearson had been wrestling with the philosophical foundation of physics, and wrote (for example), "force as a cause of motion is exactly on the same footing as a tree-god as a cause of growth." More generally, Pearson belonged to a philosophical school called positivists, who believed that the universe is a product of human thought and that science is only a description of those thoughts. Thus causation, construed as an objective process that happens in the world

outside the human brain, could not have any scientific meaning. Meaningful thoughts can only reflect patterns of observations, and these can be completely described by correlations. Having decided that correlation was a more universal descriptor of human thought than causation, Pearson was prepared to discard causation completely.

Porter paints a vivid picture of Pearson throughout his life as a self-described *Schwärmer*, a German word that translates as "enthusiast," but can also be interpreted more strongly as "zealot." After graduating from Cambridge in 1879, Pearson spent a year abroad in Germany and fell so much in love with its culture that he promptly changed his name from Carl to Karl. He was a socialist long before it became popular, and he wrote to Karl Marx in 1881, offering to translate his book *Das Kapital* into English. Pearson also was arguably one of England's first feminists, who started a Men's and Women's Club in London for discussions of "the woman question." He was concerned about women's subordinate position in society and advocated for them to be paid for their work. This was a man who was extremely passionate about ideas, while at the same time very cerebral about his passions. It took him nearly half a year to persuade his future wife, Maria Sharpe, to marry him, and it seems from their letters as if she was frankly terrified of not living up to his high intellectual ideals.

When Pearson found Galton and his correlations, he at last found a focus for his passions: an idea that be believed could transform the world of science and bring mathematical rigor to sciences like biology and psychology. And he moved with a buccaneer's sense of purpose toward accomplishing this mission. His first paper on statistics was published in 1893, four years after Galton's discovery of correlation. By 1901 he had founded a journal, *Biometrika*, which still remains one of the most influential statistical journals, and one that (ironically) published my first full paper on causal diagrams in 1995. By 1903, Pearson had secured a grant from the Worshipful Company of Drapers to start a Biometrics Lab at University College London. In

1911 it officially became a department when Galton passed away and left an endowment for a professorship (with the stipulation that Pearson be its first holder). For at least two decades, Pearson's Biometrics Lab was the world center of statistics.

Once in a position of power, Pearson's zealotry came out more and more clearly. As Porter writes in his biography, "Pearson's statistical movement had aspects of a schismatic sect. He demanded the loyalty and commitment of his associates and drove dissenters from the church biometric." One of his earliest assistants, George Udny Yule, was also one of the first people to feel Pearson's wrath. Yule's obituary of Pearson, written for the Royal Society in 1936, conveys well the sting of those days, though couched in the polite language of an obituary. "The infection of his enthusiasm, it is true, was invaluable; but his dominance, even his very eagerness to help, could be a disadvantage… This desire for domination, for everything to be just as he wanted it, comes out in other ways, notably the editing of *Biometrika*—surely the most personally edited journal that was ever published… Those who left him and began to think for themselves were apt as happened painfully in more instances than one, to find that after a divergence of opinion the maintenance of friendly relations became difficult, after express criticism impossible."

Even so, there were cracks in Pearson's edifice of causality-free science, perhaps even more so among the founders than among the later disciples. For instance, it is surprising to see that Pearson himself wrote several papers about "spurious correlation," a concept that is impossible to make sense of without making some reference to causation.

Pearson noticed that it's relatively easy to find correlations that are just plain silly. For instance (an example postdating Pearson's time, but a fun one) there is a strong correlation between a nation's *per capita* chocolate consumption and its number of Nobel Prize winners.

This correlation seems silly because we cannot envision any way that eating chocolate could *cause* Nobel Prizes. A more likely explanation is that more people in wealthy, Western

countries eat chocolate and the Nobel Prize winners have also been chosen preferentially from those countries. But this is a causal explanation, which, for Pearson, is not necessary for scientific thinking. To him, causation is just a "fetish amidst the inscrutable arcana of modern science." *Correlation* is supposed to be the goal of scientific understanding. This puts him in an awkward position when he has to explain why one correlation is meaningful and another is "spurious." He explains that a genuine correlation indicates an "organic relationship" between the variables, while a spurious correlation does not. But what is an "organic relationship"? Is it not causality by another name?

Together, Pearson and Yule compiled several examples of spurious correlations. One typical case is now called confounding, and the chocolate-Nobel story is an example. (Wealth and location are confounders, or common causes of both chocolate consumption and Nobel frequency.) Another type of "nonsense correlation" often emerges in time series data. For example, Yule found that there was an incredibly high correlation (0.95) between England's mortality rate in a given year and the percent of marriages that were conducted that year in the Church of England. Was God punishing Anglicans? No! It's simply that two separate historical trends were occurring at the same time: the country's mortality rate was decreasing and the membership in the Church of England was declining. Since both were going down at the same time, there was a positive correlation between them, but no causal connection.

Possibly the most interesting kind of "spurious correlation" was discovered by Pearson as early as 1899. It arises when two heterogeneous populations are aggregated into one. Pearson, who like Galton was a fanatical collector of data on the human body, had obtained measurements of 806 male skulls and 340 female skulls from the Paris Catacombs (Figure 5). He computed the correlation between skull length and skull breadth. When the computation was done only for males or only for females, the correlations were negligible—there was no significant association

between skull length and breadth. But when the two groups were combined, the correlation was 0.197, which would ordinarily be considered significant. This makes sense, because a small skull length is now an indicator that the skull likely belonged to a female, and therefore that the

Figure 5. Karl Pearson with a skull from the Paris Catacombs. *(Drawing by Dakota Harr.)*

breadth will also be small. However, Pearson considered it a statistical artifact. The fact that the correlation was positive had no biological or "organic" meaning; it was just a result of combining two distinct populations inappropriately.

This example is a case of a more general phenomenon called Simpson's paradox. We will discuss in Chapter 6 when it is appropriate to segregate data into separate groups, and we will explain why spurious correlations can emerge from aggregation. But let's take a look at what Pearson wrote: "To those who persist in looking upon all correlations as cause and effect, the fact that correlation can be produced between two quite uncorrelated characters A and B by taking an artificial mixture of two closely allied races, must come rather as a shock." As Stephen

Stigler comments, "I cannot resist the speculation that he himself was the first one shocked." In essence, Pearson was scolding *himself* for the tendency to think causally.

Looking at the same example through the lens of causality, we can only say: what a missed opportunity! In an ideal world, such examples might have spurred a talented scientist to think about the reason for his shock, and develop a science to predict when spurious correlations appear. At the very least, he should explain when to aggregate the data and when not to. But Pearson's only guidance to his followers is that an "artificial" mixture (whatever that means) is bad. Ironically, using our causal lens, we now know that there are cases where the aggregated data, not the partitioned data, gives the correct result. The logic of causal inference can actually tell us which one to trust. I wish that Pearson were here to enjoy it!

Pearson's students did not all follow in lock-step behind him. Yule, who broke with Pearson for other reasons, broke with him over this, too. Initially he was in the hard-line camp that correlations say everything we could ever wish to understand about causation. However, he changed his mind to some extent, when he needed to explain poverty conditions in London. In 1899, he studied the question of whether "out-relief" (that is, welfare delivered to a pauper's home, versus a poor-house) increased the rate of poverty. The data showed that districts with more out-relief had a higher poverty rate, but Yule realized that it was possibly a spurious correlation: these districts might also have more elderly people, who tend to be poorer. However, he then showed that even when comparing districts with an equal proportion of elderly people, the correlation remained. This emboldened him to say that the increased poverty rate was due to out-relief. But after stepping out of line to make this assertion, he fell back into line again, writing in a footnote: "Strictly speaking, for 'due to' read 'associated with'." This set the pattern for generations of scientists after him. They would think "due to" and say "associated with."

With Pearson and his followers actively hostile toward causation, and with half-hearted dissidents such as Yule fearful of antagonizing their leader, the stage was set for another scientist from across the ocean to issue the first direct challenge to the causality-avoiding culture.

Sewall Wright, Guinea Pigs, and Path Diagrams

When Sewall Wright arrived at Harvard University in 1912, his academic background scarcely suggested the kind of lasting effect he would have on science. He had attended a small (and now defunct) college in Illinois, Lombard College, graduating in a class of only seven students. One of his teachers had been his own father, Philip Wright, an academic jack-of-all-trades who even ran the college's printing press. Sewall and his brother Quincy helped out with the press, and among other things they published the first poetry by a not-yet-famous Lombard student, Carl Sandburg.

Sewall Wright's ties with his father remained very close long after he graduated from college. Papa Philip moved to Massachusetts when Sewall did. Later, when Sewall worked in Washington DC, Philip did likewise, first at the U.S. Tariff Commission and then at the Brookings Institution as an economist. Although their academic interests diverged, they nevertheless found ways to collaborate, and Philip was the first economist to make use of his son's invention of path diagrams.

Wright came to Harvard to study genetics. At the time, genetics was one of the hottest topics in science, because Gregor Mendel's theory of dominant and recessive genes had just been rediscovered. Wright's advisor, William Castle, had identified eight different hereditary factors (or genes, as we would call them today) that affected fur color in rabbits. Castle assigned him to do the same thing for guinea pigs. After earning his doctorate in 1915, Wright got an offer for

which he was uniquely qualified: taking care of guinea pigs at the U.S. Department of Agriculture.

One wonders if the USDA knew what they were getting when they hired Wright. Perhaps they expected a diligent animal caretaker who could straighten out the chaos of twenty years of poorly kept records. Wright did all that and much, much more. Wright's guinea pigs were the springboard to his whole career and his whole theory of evolution, much like the finches on the Galapagos islands that had inspired Charles Darwin. Wright was one of the early advocates of the view that evolution was not gradual, as Darwin had said, but takes place in relatively sudden bursts.

In 1925, Wright moved on to a faculty position at the University of Chicago that was probably better suited for someone with his wide-ranging theoretical interests. Even so, he remained very devoted to his guinea pigs. An often-told anecdote says that he was once holding an unruly guinea pig under his arm while lecturing, and absent-mindedly began to erase the blackboard using the guinea pig! (Figure 6.) While his biographers agree that this story is likely apocryphal, such stories often contain more truth than dry biographies do.

It is Wright's early work at the USDA that will interest us most. The inheritance of coat color in guinea pigs stubbornly refused to play by Mendelian rules. It proved virtually impossible to breed an all-white or all-colored guinea pig, and even the most inbred families (after multiple generations of brother-sister mating) still had pronounced variation, from mostly white to mostly

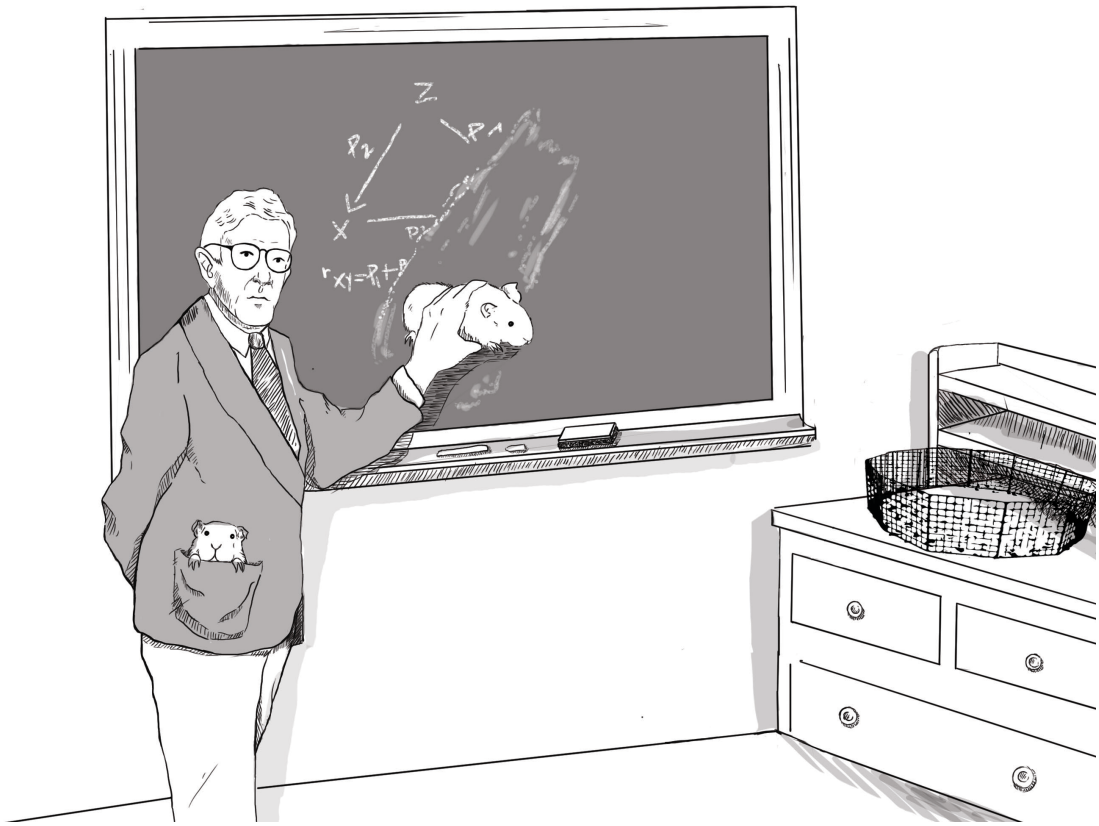

Figure 6. Sewall Wright, was the first person to develop a mathematical method for answering causal questions from data, known as path diagrams. His love for mathematics surrendered only to his passion for guinea pigs. *(Drawing by Dakota Harr.)*

colored. This contradicted the prediction of Mendelian genetics that a particular trait should become "fixed" by multiple generations of inbreeding.

Wright began to doubt that the amount of white fur was governed by genetics alone, and postulated that "developmental factors" in the womb were causing some of the variations. With hindsight, we know that he was correct. Different color genes are expressed at different places on the body, and the patterns of color depend not only on what genes the animal has inherited but where and in what combinations they happen to be expressed or suppressed.

As it often happens (at least to the ingenious!), a pressing research problem leads to new methods of analysis, which vastly transcended their origins in guinea pig genetics. Yet for Sewall Wright, it probably seemed like a college-level problem that he could have solved in his father's math class at Lombard. When you are looking for the magnitude of some unknown quantity, you first assign a symbol to that quantity, next you express what you know about this and other quantities in the form of mathematical equations, and finally, if you have enough patience and enough equations, you can solve them and find your quantity of interest.

In Wright's case, the desired and unknown quantity (shown in Figure 7) was *d*, the effect of "developmental factors" on white fur. Other causal quantities that entered into his equations included *h*, for "hereditary" factors, also unknown. Finally—and here comes Wright's ingenuity—he showed that if we knew the causal quantities in Figure 7, we could predict correlations in the data (not shown in the diagram) by a simple graphical rule. This rule sets up a bridge *from* the deep, hidden world of causation *to* the surface world of correlations. It was the first bridge ever built between causality and probability, the first crossing of the barrier between rung 2 and rung 1 on the ladder of causation. Having built this bridge, Wright could travel backward over that same bridge, *from* the correlations measured in the data (rung 1) *to* the hidden causal quantities, *d* and *h* (rung 2). He did this using the mathematics of solving algebraic

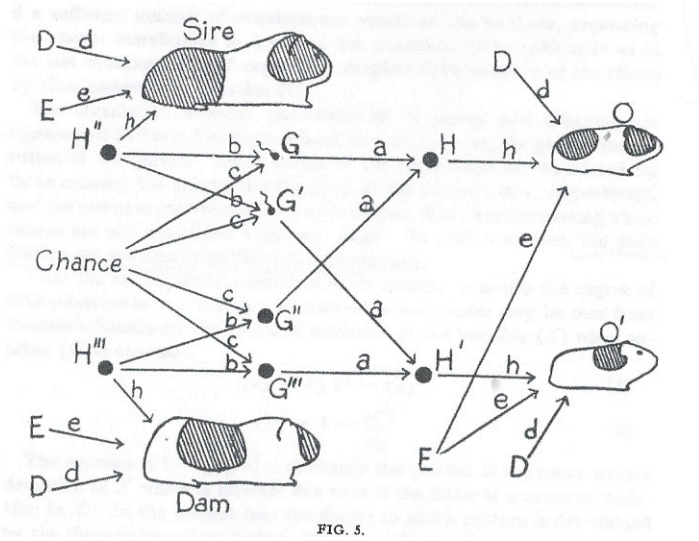

Figure 7. Sewall Wright's first path diagram, illustrating the factors leading to coat color in guinea pigs. D = developmental factors (after conception, before birth), E = environmental factors (after birth), G = genetic factors from each individual parent, H = combined hereditary factors from both parents. O, O′ = offspring. Objective of analysis was to estimate the strength of the effects of D, E, H (written as *d*, *e*, *h* in the diagram).

equations, as I mentioned in the preceding paragraph. It must have seemed simple to Wright, but it turned out to be a revolutionary idea, because it was the first proof that the mantra "correlation does not imply causation" should give way to "some correlations do imply causation!"

In the end, Wright showed that the hypothesized developmental factors were more important than heredity. In a randomly bred population of guinea pigs, 42 percent of the variation in coat pattern was due to heredity and 58 percent was developmental. By contrast, in a highly inbred family, only 3 percent of the variation in white fur coverage was due to heredity and 92 percent was developmental. In other words, the genetic variation had been all but eliminated by twenty generations of inbreeding, but the developmental factors remained.

As interesting as this result is, what is really important for our history is the way that Wright made his case. The *path diagram* in Figure 7 is the street map that tells us how to navigate over this bridge between rung 1 and rung 2. It is a scientific revolution in one picture—and it comes complete with adorable guinea pigs!

Notice that the path diagram shows every conceivable factor that could affect a baby guinea pig's pigmentation. The letters D, E, and H refer to developmental, environmental, and hereditary factors respectively. Each parent (the sire and the dam) and each child (offspring O and O′) has its own set of D, E, and H factors. The two offspring share environmental factors but have different developmental histories. The diagram incorporates the then-novel insights of Mendelian genetics: a child's heredity (H) is determined by its parents' sperm and egg cells (G and G′′), and these in turn are determined from the parents' heredity (H′′ and H′′′) via a mixing process that was not yet understood (because DNA had not been discovered yet). It was understood, though, that the mixing process included an element of randomness (labeled "Chance" in the diagram).

One thing that the diagram does not show explicitly is the difference between an inbred family and a normal family. In an inbred family there would be a strong correlation between the heredity of the sire and the dam, which Wright indicated by a two-headed arrow between H′′ and H′′′. Aside from that, every arrow in the diagram is one way and leads from a *cause* to an *effect*. For example, the arrow from G to H indicates that the sire's sperm cell may have a direct causal effect on the offspring's heredity. The absence of an arrow from G to H′ indicates that the sperm cell that gave rise to offspring O has no causal effect on the heredity of offspring O′.

When you take apart the diagram arrow by arrow in this way, I think you will find that every one of them makes perfect sense. Note also that each arrow is accompanied by a small letter (*a*, *b*, *c*, etc.). These letters are called path coefficients, and they represent the strength of

the causal effects that Wright wanted to solve for. Roughly speaking, a path coefficient represents the amount of variability in the *target* variable that is accounted for by the *source* variable. For instance, it is fairly evident that 50 percent of each child's hereditary makeup should come from each parent, so that a should be ½. (For technical reasons, Wright preferred to take the square root, so that $a = 1/\sqrt{2}$ and $a^2 = ½$.)

This interpretation of path coefficients, in terms of the amount of variation explained by a variable, was a reasonable one at the time. The modern causal interpretation is different: the path coefficients represent the results of a hypothetical intervention on the source variable. However, the notion of an intervention would have to wait until the 1940s, and could not have been anticipated by Wright when he wrote his paper in 1920. Fortunately, in the simple models he analyzed then, the two interpretations yield the same result.

I want to emphasize that the path diagram is not just a pretty picture but a powerful computational device, because the rule for computing correlations (the bridge from rung 2 to rung 1) involves tracing the paths that connect two variables to each other and multiplying the coefficients encountered along the way. Also, notice that the omitted arrows actually convey more significant assumptions than the ones that are present. An omitted arrow restricts the causal effect to zero, while a present arrow remains totally agnostic about the magnitude of the effect (unless we a priori impose some value on the path coefficient).

Wright's paper was a *tour de force*, and deserves to be considered one of the landmark results of twentieth-century biology. Certainly it is a landmark for the history of causality. Figure 7 is the first causal diagram ever published, the first step of twentieth-century science onto the second rung of the Ladder of Causation. And not a tentative step, but a bold and decisive one! The following year Wright published a much more general paper called "Correlation and Causation" that explained how path analysis worked in other settings than guinea pig breeding.

I don't know what kind of reaction the thirty-year-old scientist expected, but he surely must have been stunned by the reaction he got. It came in the form of a rebuttal published in 1921 by one Henry Niles, a student of the American statistician Raymond Pearl (no relation) who in turn was the student of Karl Pearson, the godfather of statistics.

Academia is full of genteel savagery, the wrath of which I have had the honor to experience at times in my own otherwise placid career, but even so I have seldom seen a criticism as savage as Niles'. He begins with a long series of quotes from his heroes, Karl Pearson and Francis Galton, attesting to the redundancy or even meaninglessness of the word "cause." He concludes, "To contrast 'causation' and 'correlation' is unwarranted because causation is simply perfect correlation." In this sentence he is directly echoing what Pearson wrote in *Grammar of Science*.

Niles further disparages Wright's entire methodology. He writes, "The basic fallacy of the method appears to be the assumption that it is possible to set up *a priori* a comparatively simple graphic system which will truly represent the lines of action of several variables upon each other, and upon a common result." Finally, Niles works through some examples and, bungling the computations because he has not taken the trouble to understand Wright's rules, he arrives at opposite conclusions. In summary, he declares, "We therefore conclude that philosophically the basis of the method of path coefficients is faulty, while practically the results of applying it where it can be checked prove it to be wholly unreliable."

From the scientific point of view it is perhaps not worth the time to discuss Niles' criticism in detail, but his paper is very important to us as historians of causation. First, it faithfully reflects the attitude of his generation toward causation and the total grip that his mentor, Karl Pearson, had on the scientific thinking of his time. Second, Niles' objections continue to be heard today.

For Niles, the first claim, that "causation" is nothing more than a species of correlation, was surely the most important one. But it is an outright misinterpretation of Wright. By equating causation to "perfect correlation," Niles is restricting causation to *deterministic* effects, in which a particular cause can have *only one possible* result. That is not the sense in which Wright was using the word "cause." In the modern view of causality, it is axiomatic that one variable can be a cause of another without prescribing it completely. When we say that smoking causes cancer, we do not mean that every person who smokes will get cancer. What we do mean is that smoking increases the *probability* of cancer. What Pearson and his disciples failed to realize, in a major historical blunder, was that this *probability raising* view is still a causal, not a statistical notion. In other words, the degree of probability raising cannot be determined from data alone; a causal model (e.g., a path diagram) is necessary. Wright understood this necessity, while Pearson believed that the answers are all in the data.

Niles' second argument is also fundamental. He considers it a "basic fallacy" to assert that scientific knowledge can be represented in the form of a diagram like Figure 7. How surprised he would be to find out, almost a century later, that Figure 7 is exactly how scientists *do* represent their causal knowledge! By calling diagrams a "basic fallacy," Niles was not only insulting Wright but dismissing the judgment and practice of all scientists. He was saying, in effect, that our subject-matter knowledge is worthless; that we must begin every study from scratch, putting aside our schooling, our prior experience, and most painfully our common-sense judgment.

Of course, there are times when scientists do not know the entire web of relationships between their variables. In that case, Wright argued, we can use the diagram in exploratory mode; we can postulate certain causal relationships and work out the predicted correlations between variables. If these contradict the data, then we have evidence that the relationships we

assumed were false. This way of using path diagrams was rediscovered in 1953 by the economist Herbert Simon (the 1978 Nobel laureate) and inspired much work in the social sciences.

Although we don't need to know every causal relation between the variables of interest, and we might be able to draw some conclusions with only partial information, Wright makes one point with absolute clarity: *you cannot draw causal conclusions without some causal hypotheses*. This echoes what we concluded in Chapter 1: you cannot answer a question on rung 2 of the Ladder of Causation using only data collected from rung 1.

Sometimes people ask me, "Doesn't that make causal reasoning circular? Aren't you just assuming what you want to prove?" The answer is no. By combining very mild, qualitative and obvious assumptions (e.g., coat color of the son does not influence that of the parents) with his twenty years of guinea pig data, he obtained a quantitative and by no means obvious result, which is that 42 percent of the variation in coat color is due to heredity. Extracting the non-obvious out of the obvious is not circular—it is a scientific triumph, and deserves to be hailed as such.

What makes Wright's contribution unique is that the information leading to the conclusion (of 42 percent heritability) resided in two distinct, almost incompatible mathematical languages: the language of diagrams on one side, and data on the other. This heretical idea of marrying qualitative "arrow-information" to quantitative "data-information" (two foreign languages!) was one of the miracles that first attracted me, as a computer scientist, to this enterprise.

Many people still make Niles' mistake of thinking that the goal of causal analysis is to prove that X is a cause of Y, or else to find the cause of Y from scratch. That is the problem of causal *discovery*, which was my ambitious dream when I first plunged into graphical modeling, and is still an area of vigorous research. In contrast, the focus of Wright's research, as well as

this book, is on *representing plausible causal knowledge* in some mathematical language, combining it with empirical data, and *answering causal queries* that are of practical value. Wright understood from the very beginning that causal discovery was much more difficult, and perhaps impossible. In his response to Niles, he writes: "The writer [i.e., Wright himself] has never made the preposterous claim that the theory of path coefficients provides a general formula for the deduction of causal relations. He wishes to submit that the *combination* of knowledge of correlations with knowledge of causal relations to obtain certain results, is a different thing from the *deduction* of causal relations from correlations implied by Niles' statement."

E Pur Si Muove (Yet It Moves)

If I were a professional historian, I would probably stop here. But as the "Whig historian" that I promised to be, I cannot contain myself from expressing my sheer admiration for the precision of Wright's words in the above quote, which have not gone stale in the 90 years since they were first articulated and which essentially defined the new paradigm of modern causal analysis.

My admiration for Wright's precision is second only to my admiration of his courage and determination. Imagine the situation in 1921, a self-taught mathematician facing the hegemony of the statistical establishment alone. They tell him: Your method is based on a complete misapprehension of the nature of causality in the scientific sense. And he retorts: Not so! My method generates something that is important, and that goes beyond anything that you can generate. They say: Our gurus looked into these problems already, two decades ago, and concluded that what you have done is nonsense. You have only combined correlations with correlations and gotten correlations. When you grow up, you will understand. And he continues:

I am not dismissing your gurus, but a spade is a spade. My path coefficients are *not* correlations, they are something totally different: causal effects.

Imagine that you are in kindergarten and your friends mock you for believing that 3 + 4 = 7, when everybody knows that 3 + 4 = 8. Then imagine going to your teacher for help and hearing her say, too, that 3 + 4 = 8. Would you not go home and ask yourself, perhaps there is something wrong in the way I am thinking? Even the strongest man would start to waver in his convictions. I have been in that kindergarten, and I know.

But Wright did not blink. And this was not just a matter of arithmetic, where there can be some sort of independent verification. This was a dispute that only philosophers had dared to express an opinion about. Where did Wright get this inner conviction that he was on the right track and the rest of the kindergarten was just plain wrong? Maybe it was his Midwestern upbringing, and the tiny college he went to, that encouraged his self-reliance and taught him that the surest kind of knowledge is what you construct yourself.

In one of the earliest science books I read in school, we were told how the Inquisition forced Galileo to recant his teaching that Earth revolves around the sun, and how he whispered under his breath, "And yet it moves." (*E pur si muove*.) I don't think that there is a child in the world who has read this legend without being moved by Galileo's convictions and his courage in defending them. Yet as much as we admire him for his stand, I can't help but think that he at least had his astronomical observations to fall back on. Wright had only untested conclusions, say that developmental factors account for 58 percent, not 3 percent, of variation. With nothing to lean on except his internal conviction that path coefficients tell you what correlations do not, he still declared, "And yet it moves!"

Colleagues tell me that when Bayesian networks fought against the AI establishment (see Chapter 3), I acted stubbornly, single-mindedly and uncompromisingly. Indeed, I recall being

totally convinced in my approach, with not an iota of hesitation. But I had probability theory on my side. Wright didn't have even one theorem to lean on. Causation had been abandoned by scientists, so Wright could not fall back on any theoretical framework. Nor could he rely on authorities, as Niles did, because there was no one for him to quote; the gurus had already pronounced their verdicts three decades earlier.

But one solace to Wright, and one sign that he was on the right path, must have been his understanding that he could answer questions that cannot be answered in any other way. Determining the relative importance of several factors was one such question. Another beautiful example of this can be found in his "Correlation and Causation" paper, from 1921, which asks: How much will a guinea pig's birth weight be affected if it spends one more day in the womb? I would like to examine Wright's answer in some detail, to enjoy the beauty of his method and to satisfy readers who would like to see how the mathematics of path analysis works.

Notice that we cannot answer Wright's question directly, because we can't weigh a guinea pig in the womb. What we can do though, is compare the birth weights of guinea pigs that spend (say) 66 days gestating with those that spend 67 days. Wright noted that the guinea pigs that spent a day longer in the womb averaged 5.66 grams more at birth. So, naively, it seems that a guinea pig embryo is growing at 5.66 grams per day just before it is born.

"Wrong!" says Wright. The pups who are born later are usually born later *for a reason*: they have fewer litter mates. This means that they have had a more favorable environment for growth *throughout* the pregnancy. A pup with only 2 siblings, for instance, will *already* weigh more on day 66 than a pup with 4 siblings. Thus the difference in birth weights has two causes, and we want to disentangle the two. How much of the 5.66 grams is due to spending an additional day *in utero*, and how much is due to having fewer siblings to compete with?

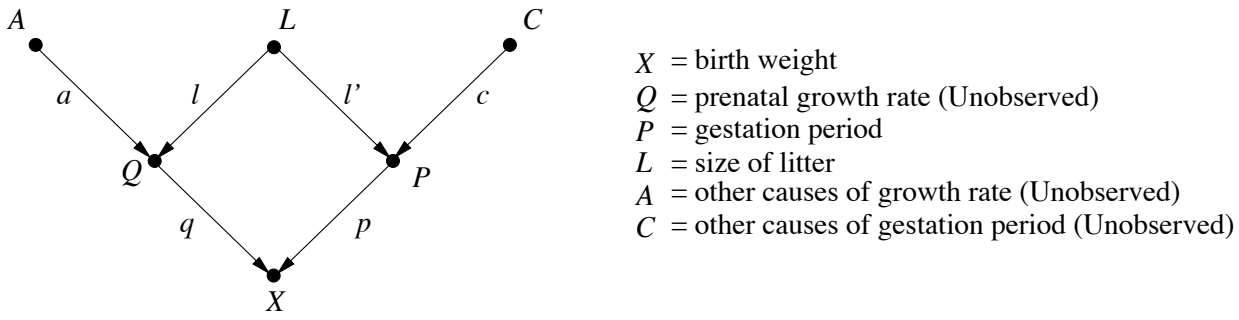

Figure 8. Causal (path) diagram for birth weight example.

Wright answered this question by setting up a path diagram (Figure 8). X represents the pup's birth weight. Q and P represent the two known causes of the birth weight: the length of gestation (P) and rate of growth *in utero* (Q). L represents litter size, which affects both P and Q (a larger litter causes the pup to grow slower and also have fewer days *in utero*). It's very important to realize that X, P, and L can be measured, for each guinea pig, but Q cannot. Finally, A and C are exogenous causes that we don't have any data about, e.g., hereditary and environmental factors that control growth rate and gestation time independently from litter size. The important assumption that these factors are independent of each other is conveyed by the absence of any arrow between them, as well as any common ancestor.

Now the question facing Wright was: what is the direct effect of the gestation period P on the birth weight X? The data (5.66 grams per day) don't tell you the direct effect; they tell you *correlation*, biased by the litter size L. To get the direct effect, we need to remove this bias.

In Figure 2, the direct effect is represented by the path coefficient *p*, corresponding to the path P → X. The bias due to litter size corresponds to the path P ← L → Q → X. And now the algebraic magic: the amount of bias is equal to the product of the path coefficients[1] along that

[1] For anyone who takes the trouble to read Wright's paper, let me warn you that he does not compute his path coefficients in grams per day. He computes them in "standard units" and then converts to grams per day at the end.

path (in other words, l times l' times q). The total correlation, then, is just the sum of the path coefficients along the two paths: algebraically, $p + (l \times l' \times q) = 5.66$ grams per day.

If we knew the path coefficients l, l', and q, then we could just work out the second term and subtract it off from 5.66, to get the desired quantity p. But we don't know them, because Q (for example) is not measured. But here's where the ingenuity of path coefficients really shines. Wright's methods tell us how to express each of the measured correlations in terms of the path coefficients. After doing it for each of the measured pairs (P, X), (L, X), and (L, P), we obtain three equations which can be solved algebraically for the unknown path coefficients, p, l', and $l \times q$. Then we are done, because the desired quantity p has been obtained.

Today we can skip the mathematics altogether and calculate p by cursory inspection of the diagram. But in 1920, it was the first time that mathematics was summoned to connect causation and correlation. And it worked! Wright calculated p to be 3.34 grams per day. In other words, had all the other variables (A, L, C, Q) been held constant and only the time of gestation increased by a day, the average increase in birth weight would be 3.34 grams per day. Note that this result is biologically meaningful. It tells us how rapidly the pups are growing a day before birth. By contrast, the number 5.66 grams per day has no biological significance, because it conflates two separate processes, one of which is not causal but anti-causal (or diagnostic) in the link P ← L. Lesson one from this example: Causal analysis allows us to quantify *processes in the real world*, not just patterns in the data. The pups are growing at 3.34 grams per day, not 5.66 grams per day. Lesson two, whether you followed the mathematics or not, is that in path analysis you draw conclusions about individual causal relationships *by examining the diagram as a whole*. The entire structure of the diagram may be needed for estimating each individual parameter.

In a world where science progresses logically, Wright's response to Niles should have produced a scientific excitement followed by an enthusiastic adoption of his methods by other scientists and statisticians. But that is not what happened. "One of the mysteries of the history of science from 1920 to 1960 is the virtual absence of any appreciable use of path analysis, except by Wright himself and by students of animal breeding," wrote one of Wright's geneticist colleagues, James Crow. "Although Wright had illustrated many diverse problems to which the method was applicable, none of these leads was followed."

Crow didn't know it, but the mystery extended to social sciences as well. In 1972, economist Arthur Goldberger lamented the "scandalous neglect" of Wright's work during that period and noted, with the enthusiasm of a convert, that it was "[Wright's] approach … which sparked the recent upsurge of causal modeling in sociology."

If only we could go back and ask Wright's contemporaries, "Why didn't you pay attention?" Crow suggests one reason: path analysis "doesn't lend itself to 'canned' programs. The user has to have a hypothesis and must devise an appropriate diagram of multiple causal sequences." Indeed, Crow put his finger on an essential point: Path analysis requires scientific thinking, as does every exercise in causal inference. Statistics in the way it is frequently practiced discourages it, and encourages "canned" procedures instead. Scientists will always prefer routine calculations on data to methods that challenge their scientific knowledge.

R.A. Fisher, the undisputed high priest of statistics in the generation after Galton and Pearson, described this difference succinctly. In 1925, he wrote: "Statistics may be regarded as… the study of methods of the reduction of data." Pay attention to the words *methods*, *reduction*, and *data*. Wright abhorred the idea of statistics as merely a collection of methods; Fisher embraced it. Causal analysis is emphatically not just about data; in causal analysis we must incorporate some understanding of the process that produces the data, and then we get something

that was not in the data to begin with. But Fisher was right about one point: once you remove causation from statistics, then reduction of data is the only thing left.

Although Crow did not mention it, Wright's biographer William Provine points out another factor that may have affected the lack of adoption of path analysis. From the mid-1930s onward, Fisher considered Wright his enemy. I previously quoted Yule on how relations with Pearson became strained if you disagreed with him, and impossible if you criticized him. Exactly the same thing could be said about Fisher. The latter carried out nasty feuds with anyone he disagreed with, including Pearson, Pearson's son Egon, Jerzy Neyman (more will be said on these two in Chapter 8), and of course Wright.

The real focus of the Fisher-Wright rivalry was not path analysis but evolutionary biology. Fisher disagreed with Wright's theory (called "genetic drift") that a species can evolve rapidly when it undergoes a population bottleneck. The details of the dispute are beyond the scope of this book, and the interested reader should consult Provine. But what is relevant here is this: from the 1920s to the 1950s, the scientific world for the most part turned to Fisher as its oracle for statistical knowledge. And you can be certain that Fisher never said one word to anyone about path analysis.

In the 1960s, things began to change. A group of social scientists, including Otis Duncan, Harold Blalock, and the economist Arthur Goldberger (mentioned earlier), rediscovered path analysis as a method of predicting the effect of social and educational policies. In yet another irony of history, Wright had actually been asked to speak to an influential group of econometricians called the Cowles Commission in 1947, but he utterly failed to communicate to them what path diagrams were about. It was only when economists arrived at similar ideas themselves that some connection was forged, albeit short-lived.

The fate of path analysis in economics and sociology followed different trajectories, each leading to a betrayal of Wright's ideas. Sociologists renamed path analysis as *structural equation modeling* (SEM), embraced diagrams and used them extensively until 1975, when a computer package called LISREL automated the calculation of path coefficients (in some cases). What followed was just what Wright would have predicted: path analysis turned into a rote method, and researchers became software users with little interest in what was going on under the hood. By the late 1980s, when publicly challenged (by statistician David Freedman) to explain the assumptions behind SEM, his challenge remained unanswered and some leading SEM experts even disavowed that they had anything to do with causality.

In economics, the algebraic part of path analysis became known as simultaneous equation models (no acronym). Economists essentially never used path diagrams and continue not to use them to this day, relying instead on numerical equations and matrix algebra. This has had two dire consequences. One is that, because algebraic equations are non-directional (that is, $x = y$ is the same as $y = x$), economists were left with no notational means to distinguish causal from regression equations and thus were unable to answer policy-related questions, even after solving the equations. As late as 1995, most economists refrained from explicitly attributing causal or counterfactual meaning to their equations. Even those who did remained incurably suspicious of diagrams, which could have saved them pages and pages of computation. Some economists continue to claim that "it's all in the data" to this very day.

For all these reasons, the promise of path diagrams remained only partially realized, at best, until the 1990s. In 1983, Wright himself was called back into the ring one more time to defend them, this time in the *American Journal of Human Genetics.*

At the time he wrote this article, Wright was past ninety years old. It is both wonderful and tragic to read his essay, written in 1983, on the very same topic he had written about in 1923.

How many times in the history of science have we had the privilege of hearing from a theory's creator *sixty years* after he first set it down on paper? It would be like Charles Darwin coming back from the grave to testify at the Scopes Trial in 1925. But it is also tragic, because in the intervening sixty years his theory should have developed, grown, and flourished; but instead it was not too far advanced from where it had been in the 1920s.

The motivation for Wright's paper was a critique of path analysis that had been published in the same journal, written by Samuel Karlin (a Stanford mathematician, recipient of the 1989 National Medal of Science, who made fundamental contributions to economics and population genetics) and two co-authors. Of interest to us are two of Karlin's arguments.

First, Karlin objects to path analysis for a reason that Niles did not raise: It assumes that all the relationships between any two variables in the path diagram are *linear*. This assumption is what allows Wright to describe the causal relationships with a single number, the path coefficient. If the equations were not linear, then the effect on Y of a 1-unit change in X might depend on what the current value of X is. What neither Karlin nor Wright realized was that a general nonlinear theory was just around the corner. (It would be developed three years later by a star student in my lab, Thomas Verma!)

But Karlin's most interesting criticism was also the one that he considered the most important: "Finally, and we think most fruitfully, one can adopt an essentially model-free approach, seeking to understand the data interactively by using a battery of displays, indices, and contrasts. This approach emphasizes the concept of robustness in interpreting results."

In this one sentence Karlin articulated how little had changed from the days of Pearson, and how much influence Pearson's ideology still had, in 1983. He is saying that the data themselves already contain all scientific wisdom; they need only be cajoled and massaged (by "displays, indices, and contrasts") into dispensing those pearls of wisdom. There is no need for

our analysis to take into account the process that generated the data. We would do just as well, if not better, with a "model-free approach." If Pearson were alive today, living in the era of Big Data, this is exactly what he would say: The answers are all in the data.

Of course, Karlin's statement violates everything we learned in Chapter 1: To speak of causality, we *must* have a mental model of the real world. A "model-free approach" may take us to the first rung of the Ladder of Causation, but no farther.

Wright, to his great credit, understood the enormous stakes, and stated in no uncertain terms: "In treating the model-free approach (3) as preferred alternative …, Karlin et. al. are urging not merely a change in method, but an abandonment of the purpose of path analysis and evaluation of the relative importance of varying causes. There can be no such analysis without a model. Their advice to anyone with an urge to make such an evaluation is to repress it and do something else."

Wright understood that he was defending the very essence of the scientific method and the interpretation of data. It is the same advice I would give today to big-data, model-free enthusiasts: of course, it is okay to tease out all the information that the data can provide, but let's ask how far this will get us. It will never get us beyond the first rung of the Ladder of Causation, and it will never answer even as simple a question as what is the relative importance of various causes. *Eppur si muove!*

From Objectivity to Subjectivity – the Bayesian Connection

One other theme in Wright's rebuttal may hint at another reason for the resistance of statisticians to causality. He repeatedly states that he did not want path analysis to become "stereotyped." This is surely intended to contrast with Fisher's approach of turning everything into a procedure. According to Wright, "The unstereotyped approach of path analysis differs

profoundly from the stereotyped modes of description designed to avoid any departures from complete objectivity."

What does he mean? First, he means that path analysis should be based on the user's personal understanding of causal processes, reflected in the causal diagram. It cannot be reduced to mechanical routines, such as those laid out in statistics manuals. For Wright, drawing a path diagram is not a *statistical* exercise, but an exercise in genetics, economics, psychology, or whatever the scientist's own field of expertise is.

Second, Wright traces the allure of "model-free" methods to their *objectivity*. This has indeed been a holy grail for statisticians since day one—or since March 15, 1834, when the Statistical Society of London was founded. Its founding charter said that *data* were to be given priority in all cases over opinions and interpretations. Data are objective; opinions are subjective. This paradigm long predates Pearson. The struggle for objectivity—the idea of reasoning exclusively from data and experiment—has been part of the way that science has defined itself ever since Galileo.

Unlike correlation and most of the other tools of mainstream statistics, causal analysis requires the user to make a subjective commitment. She must draw a causal diagram that reflects her qualitative belief, or better yet the consensus belief of researchers in her field of expertise, about the topology of the causal processes at work. She must abandon the centuries-old dogma of objectivity for objectivity's sake. Where causation is concerned, a grain of wise subjectivity tells us more about the real world than any amount of objectivity.

In the above paragraph, I said that "most of" the tools of statistics strive for complete objectivity. There is one important exception to this rule, though. A branch of statistics called *Bayesian statistics* has achieved growing popularity over the last 50 years or so. Once considered almost anathema, it has now gone completely mainstream, and you can attend an entire statistics

conference without hearing any of the great debates between "Bayesians" and "frequentists" that used to thunder in the 1960s and 1970s.

The prototype of Bayesian analysis goes like this: Prior belief + New evidence → Revised belief. For instance, suppose you toss a coin ten times and find that in nine of those tosses the coin came up heads. Your belief that the coin is fair is probably shaken, but how much shaken? An orthodox statistician would say: in the absence of any additional evidence I would believe that this coin is loaded, so I would bet 9:1 that the next toss turns up heads.

A Bayesian statistician, on the other hand, would say: Wait a minute, we also need to take into account our prior knowledge about the coin. Did I get it from the neighborhood grocery or from a shady gambler? If it's just an ordinary quarter, most of us would not let the coincidence of nine heads sway our belief so dramatically. On the other hand, if we already suspected the coin was weighted to one side, then we would be more willing to conclude that the nine heads provide serious evidence of bias.

Bayesian statistics give us an objective way of combining the observed evidence with our prior knowledge (or subjective belief) to obtain a revised belief, and hence a revised prediction of the outcome of the coin's next toss. While the exact details of this combination will be discussed in the next chapter, the point that frequentists could not abide was that Bayesians were allowing opinion, in the form of subjective probabilities, to intrude into the pristine kingdom of statistics. Mainstream statisticians were won over only grudgingly, when Bayesian analysis proved to be a superior tool for a variety of applications, such as weather prediction and tracking enemy submarines. In addition, in many cases it can be proven that the influence of prior beliefs vanishes as the size of the data increases, leaving a single objective conclusion in the end.

Unfortunately, the acceptance of Bayesian subjectivity in mainstream statistics did nothing to help the acceptance of causal subjectivity, the kind that is needed for specifying a path

diagram. Why? The answer rests on a grand linguistic barrier. Bayesian statisticians use the language of probability, the native language of Galton and Pearson. The assumptions entering causal inference, on the other hand, require a richer language (e.g., diagrams) that is foreign to Bayesians and frequentists alike. The reconciliation between Bayesians and frequentists shows that philosophical barriers can be bridged with good will and a common language. Linguistic barriers are not broken so easily.

Moreover, the subjective component in causal information does not necessarily diminish over time, even as the amount of data increases. Two people who believe in two different causal diagrams can analyze the same data and may never come to the same conclusion, regardless of how "big" the data are. This is a terrifying prospect for advocates of scientific objectivity, which explains their refusal to accept the inevitability of relying on subjective causal information.

On the positive side, causal inference *is* objective in one critically important sense: once two people agree on their assumptions, it provides a 100 percent objective way of interpreting any new evidence (or data). This is a property that it shares with Bayesian inference. So it will probably not surprise the savvy reader to find out that I arrived at the theory of causality through a circuitous route that started out with Bayesian probability and then took a huge detour through Bayesian networks. That is the story that I will tell in the next chapter.