CHAPTER SEVEN

# The Experimental Basis from Which to Test Hypotheses: Brownian Motion

> My major aim in this was to find facts which would guarantee as much as possible the existence of atoms of definite finite size. In the midst of this I discovered that, according to atomistic theory, there would have to be a movement of suspended microscopic particles open to observation, without knowing that observations concerning the Brownian motion were already long familiar.
>
> —Albert Einstein, *Albert Einstein: Philosopher-Scientist*, p. 47

> I have sought in this direction for crucial experiments that should provide a solid experimental basis from which to attack or defend the Kinetic Theory.
>
> —Jean Perrin, *Atoms*, p. 89

## 7.1 Brownian Motion: Some Introductory Remarks

A more full-blown example from science can now best elucidate the machinery I have assembled—the hierarchy of models in an experimental inquiry, the simple statistical test, the piecemeal check for errors by splitting off local questions, the use of canonical models of error, and the strategies for arriving at severe tests. An example often discussed by philosophers is the appraisals of hypotheses surrounding the phenomenon of Brownian motion.

Brownian motion, discovered by the botanist Robert Brown in 1827, refers to the irregular motion of small particles suspended in fluid, a motion that keeps them from sinking due to gravitation. Brown thought the particles were alive until he found that the motion occurred in inorganic as well as organic substances.[1] Attempts to explain

1. Brown discovered a piece of quartz in which a drop of water had been trapped for millions of years. Observing it under a microscope, he saw numerous particles in ceaseless, irregular motion.

this phenomenon link up with the atomic debates of the late-nineteenth and early-twentieth centuries. The atomic debate being too broad to be taken up here, I restrict my focus to the testing of the Einstein-Smoluchowski (ES) theory of Brownian motion by Jean Perrin. Perrin, who received the 1926 Nobel Prize in physics, provided the long sought after evidence in favor of the molecular-kinetic theory of gases against classical thermodynamics.

### *Brownian Motion and Paradigm Shifts*

This case lends itself not only to explicating the present program of breaking down substantive inquiries into piecemeal canonical questions but also to the exploration of the key role of experimental knowledge in larger scale theory change. The debates about the cause of Brownian motion correlate with a number of disputes between what might be regarded as rival paradigms or disciplinary matrices, as discussed in chapter 2. The disputes were between molecular and phenomenological or energeticist ontologies, mechanical and phenomenological explanation, atomic and continuous metaphysics, statistical and nonstatistical models, realism and instrumentalism, and still others. Correspondingly, the acceptance of the ES theory of Brownian motion led to the changing of the key elements that compose paradigms or disciplinary matrices. It led to a change in beliefs about fundamental entities, the existence of molecules, and the particulate nature of matter. It also led to a change in scientific methodology: a new limit to experimental accuracy due to Brownian fluctuations and "noise" in measuring systems was introduced along with corresponding canonical models of error. The entry of statistical validity into physics also conflicted with the cherished philosophical conception that physics discovers wholly exact laws. (New applications of models from Brownian motion continue up to the present time.[2])

Nevertheless, the dynamics of these changes bore no resemblance to the Kuhnian picture of holistic change, as discussed earlier. At each step of the way experimental methodology and shared criteria of reliability and severity constrained the appraisal. What we see are practitioners from ostensibly rival paradigms learning from and communicating by means of experimental arguments. We cannot make good on these claims, however, until we have properly considered Perrin's

2. Statistical models based on Brownian motion are used to understand star clustering, the evolution of ecological systems, and even the fluctuation of retail prices.

much more local work on Brownian motion. It is here that one really finds the locus of the action.

*Molecular Reality and Arguments from Coincidence*

The Perrin case often has been taken to exemplify variants of the "argument from coincidence" discussed in chapter 3. Gilbert Harman (1965) uses it to introduce inference to the best explanation, Wesley Salmon (1984) to illustrate his argument to a common cause, Nancy Cartwright (1983) to illustrate her inference to the most probable cause. Hacking (1983) ties the Perrin example to his discussion of the argument from coincidence.

They take the Perrin case to illustrate an argument from coincidence because experiments on many distinct phenomena (e.g., gases, Brownian motion, blue of the sky) gave estimates for Avogadro's number, $N$ (the mean number of molecules per gram molecule), of a similar order of magnitude.[3] Salmon argues:

> If there were no such micro-entities as atoms, molecules, and ions, then these different experiments designed to ascertain Avogadro's number would be genuinely independent experiments, and the striking numerical agreement in their results would constitute an utterly astonishing coincidence. To those who were in doubt about the existence of such micro-entities, the "remarkable agreement" constitutes strong *evidence* that these experiments are not fully independent—that they reveal the existence of such entities. (Salmon 1984, 220)

Cartwright argues along similar lines:

> We have thirteen phenomena from which we can calculate Avogadro's number. Any one of these phenomena—if we were sure enough about the details of how the atomic behaviour gives rise to it—would be good enough to convince us that Avogadro is right. Frequently we are not sure enough; we want further assurance that we are observing genuine results and not experimental artefacts. This is the case with Perrin. . . . But he can appeal to coincidence. Would it not be a coincidence if each of the observations was an artefact, and yet all agreed so closely about Avogadro's number? (Cartwright 1983, 84)

Salmon and Cartwright view these arguments as supporting realist conclusions. I think their accounts do capture the crux of the arguments that are given for the reality of atoms or to "molecular reality," as Mary Jo Nye (1972) puts it. Those arguments, however, were dis-

3. Actually, close estimates of $N$ had already been given many years before Perrin's work.

tinct from the experimental arguments Perrin and others grappled with in learning from their experiments. Of course, the arguments for the reality of molecules depended upon the successful inquiries into Brownian motion, but the success of those inquiries did not hinge on the agreement of estimates of Avogadro's number across the thirteen phenomena. For example, the possible error of experimental artifacts was put to rest by Perrin quite apart from the work on the other phenomena (blue of the sky, radiation, etc.). Otherwise he could not have arrived at a reliable estimate of Avogadro's number in the first place. The same was the case for learning the statistical nature of the second law of thermodynamics.

This is not to deny that Perrin utilized arguments from the coincidence of many distinct results—he certainly did. He repeatedly emphasized that one could only put faith in calculations arrived at in several different ways. The several different ways served two functions: to check errors in rather precarious measurements, and to arrive at standard estimates of error (needed for statistical analysis). But first and foremost, Perrin was arguing from coincidence *to obtain experimental knowledge* of the Brownian movement (of microscopic grains). Molecular reality came later. His arguments are really stunning illustrations of the development and use of canonical models of error and of experimental arguments for learning from error.

In the opening passage, Perrin declares himself searching for crucial experiments, but his idea does not fit the mold of a Popperian severe or crucial test. Brownian motion was not only known long before it was used in testing the kinetic against the classical accounts, but it was also accounted for in a number of other ways. Moreover, only after the interesting experimental work had been done could it be seen that the kinetic and classical theories give conflicting predictions. Let us turn to the interesting experimental work.

## 7.2 SOME BACKGROUND: THE FLURRY OF EXPERIMENTS ON BROWNIAN MOTION

From its initial discovery in 1827, each inquiry into the cause of Brownian motion has been a story of hundreds of experiments. The experiments are of two main classes: experiments that (arrive at and) test hypotheses that attribute Brownian motion either to the nature of the particle studied or to various factors external to the liquid medium in which the Brownian particles are suspended (e.g., temperature differences in the liquid observed, vibrations of the object glass); and experiments that test the quantitative theory of Brownian motion put forward by Einstein and (independently) by M. von Smoluchowski (the ES theory).

Each molecular-kinetic explanation of Brownian motion (first qualitatively proposed by Christian Wiener in 1863) spurred a flurry of experiments by biologists and physicists aiming to refute it. Each nonkinetic hypothesis tried, as by devising a novel way of explaining it by temperature differences, would trigger a new set of experiments to refute the challenge. The enormous variety of organic and inorganic particles studied includes sulphur, cinnabar, coal, urea, India ink, and something called gamboge. Equally numerous were the treatments to which such particles were subjected in the hope of uncovering the cause of Brownian motion—light, dark, country, city, red and blue light, magnetism, electricity, heat and cold, even freshly drawn human milk.

The scientists working on this problem (e.g., Brown, Wiener, Ramsay, Gouy, Perrin, and Smoluchowski) began by carrying out experiments to exclude all exterior causes—checking and rechecking even those suspected factors that already had been fairly well ruled out. (Even after Perrin's work and the general acceptance of molecular theory, experiments using ever-improving methods to observe hundreds of thousands of microscopic grains continued.[4]) By the end of the nineteenth century the most favored explanations in some way attributed Brownian motion to heat (e.g., the theories of Exner, Dancer, Quincke). The need for a molecular explanation began to take hold only at the start of the twentieth century. Ironically, the fact that the same Brownian particles could be used over and over again, sometimes conserved on slides for twenty years, compelled those who had been searching for *nonkinetic* explanations to admit this as strong evidence *for* the kinetic explanation. It indicated that the motion was "eternal and spontaneous," in accordance with the kinetic account. Even so, most researchers required many more quantitative experiments before abandoning nonkinetic explanations.

Only by keeping in mind that a great many causal factors were ruled out experimentally before Perrin's tests (around 1910) can his experiments be properly understood. This will become apparent only when we explicitly consider how Perrin handled the problems of experimental design and control. We can follow the central experimental arguments by means of the hierarchy of models for an experimental inquiry delineated in chapter 5. Except where noted, all references to Perrin are to Perrin [1913] 1990.

4. An excellent sourcebook detailing these modern experiments is Wax 1954.

## 7.3 MODEL OF THE PRIMARY THEORETICAL HYPOTHESIS: THE DISPLACEMENT DISTRIBUTION

The central problem with appraising the kinetic account of Brownian motion was how to formulate testable predictions. The problem was that for many years experimenters were measuring the wrong thing. What they thought had to be checked was whether the molecular effects on the *velocity* of Brownian particles accorded with that hypothesized by the kinetic theory. But this average or mean velocity had been ascertained by trying to follow the path of a Brownian particle, inevitably yielding a measured path much simpler and shorter than the actual path, which changes too fast. An important advantage of the ES theory was that it provided a testable prediction that made no reference to this unmeasurable velocity. At the same time the ES theory explained why the earlier attempts to measure it had failed.

Values obtained for the mean velocity of agitation by attempting to follow the path of a grain as nearly as possible gave the grains a kinetic energy 100,000 times too small. According to Einstein's theory, mean velocity in an interval of time $t$ is inversely proportional to the square root of $t$; it increases without limit as the time gets smaller. The meaningless results were just what the ES theory says would be expected. As Stephen Brush (1977, 369) remarks, "One can hardly find a better example in the history of science of the complete failure of experiment and observation, unguided (until 1905) by theory, to unearth the simple laws governing a phenomenon."

While Einstein's theory apparently served that guiding role, this does not mean that the tests of the ES theory depended on already accepting Einstein's theory. (I think some people mistakenly suppose that it does.) The reason attempts to measure a particle's velocity were in error was independent of the ES theory. The error turned on a (now) standard statistical point that had been articulated in other contexts but was overlooked. In particular, the point had been made in 1854 by William Thomson (Lord Kelvin) about the problem of laying the Atlantic cable.

Thomson calculated that unless the Atlantic cable is made very thick the transmission of messages between Britain and America would be very slow (Brush 1977, 369). For economic reasons, engineers were unhappy with this recommendation. In addition, attempts to measure the velocity of electricity achieved widely varying results, some appearing to contradict Thomson's prediction. Thomson defended his theoretical prediction: the diverse measurements of the velocity of electricity, he explained, were due to the time spent making the measurements. The time it takes an electric signal to cover a distance is proportional not to the distance itself, but to the square of the distance. (This is Thomson's law of squares.) The greater the length of wire used, the less the apparent velocity of electricity would seem. No wonder the values varied, measured as they were under different conditions. Brush (1977, 369–70) remarks:

> Apparently the scientists who attempted to measure the velocity of particles in Brownian movement later in the nineteenth century had not followed the dispute about Thomson's law of squares in the electric telegraph problem, and they obtained a similar collection of wildly varying results, none of them in agreement with [what would be expected according to the kinetic theory].

Perhaps if a log of canonical experimental errors had been kept, this mistake could have been instructive rather than repeated!

In any event, a key advantage of the ES theory was that it provided a testable prediction that made no reference to this unmeasurable velocity. Instead it was put in terms of the expected (or mean) *displacement* of particles.

Neglecting, therefore, the true velocity, which cannot be measured,

> Einstein and Smoluchowski chose, as the magnitude characteristic of the agitation, the rectilinear segment joining the starting and end points [of a particle]; in the mean, this line will clearly be longer the more active the agitation. (Perrin [1913] 1990, 110)

The displacement of a Brownian particle is the total distance it travels in any direction (say along the x-axis of a graph) as it weaves its zigzagged path. It is a distance that could be measured using the microscopes of the day. (See figure 7.1.) The measurement actually obtained from microscopic observations is the projection of this displacement onto a horizontal plane (e.g., the x-axis). By "observed displacement," I mean this projection onto the x-axis of the given segment.

So the question of interest concerns the quantity (abbreviated as $S_t$), the displacement (along the x-axis) after $t$ minutes of a Brownian particle from its starting point. If molecular agitation (as described by the kinetic theory of gases) causes Brownian movement, then the displacement of a Brownian particle about its mean (which by symmetry is 0) follows the Gaussian distribution, which is just the familiar Normal distribution. This distribution is given by two parameters, the mean (which is 0) and the variance. The variance is equal to $2Dt$, where $D$ is the *coefficient of diffusion* and $t$ is the time.[5] As with the familiar "bell curve," if the displacement of a Brownian particle is Normally distributed around 0, then displacements near 0 are most probable, while those further from 0 are increasingly improbable.[6] (How probable specific differences are is what the Normal distribution tells us—we need only know $D$.) As Einstein states it,

> the probable distribution of the resulting displacements in a given time $t$ is therefore the same as that of *fortuitous error,* which was to be expected. (Einstein [1926], 1956, 16; emphasis added)

So the primary theoretical hypothesis is a hypothesized statistical distribution (of displacements of suspended particles):

> The primary hypothesis $\mathcal{H}$: The displacement of a Brownian particle over time $t$, $S_t$, follows the Normal (or Gaussian) distribution with $\mu = 0$ and variance $= 2Dt$.

Having provided this hypothesized distribution by which to test the kinetic theory which entails it, Einstein concludes his 1905 paper by remarking, "It is to be hoped that some enquirer may succeed shortly in solving the problem suggested here" (Einstein [1926] 1956, 18). Perrin took up Einstein's challenge:

> It [Einstein's Theory] is well adapted to accurate experimental verification, *provided we are able to prepare spherules of measureable radius.* Consequently, ever since I became . . . acquainted with the theory, it has been my aim to apply to it the test of experiment. (Perrin, 114)

It was Perrin's dogged efforts to prepare grains with measurable and highly uniform radius that made his experimental tests so successful. As we will see, this uniformity of grains was a key assumption of the experimental testing model.

5. For the interested reader, $S_t$, the displacement after $t$ minutes of a particle in Brownian motion, follows the Normal probability (density) function $f$:

$$f_{S_t}(x) = \frac{1}{(4\pi DT)^{1/2}} e^{-x^2/4Dt}$$

where $D$ is a constant, the diffusion coefficient. $D$ depends on the absolute temperature and friction coefficient of the surrounding medium. (Strictly speaking, this assumes that $t$ is not too small.)

The mean of $S_t$, $E(S_t)$, equals 0 and the variance of $S_t$, $E[S_t^2]$, equals $2Dt$. $E$ here abbreviates the mean or expected value.

6. From our previous discussion we already know more than this. We know, for example, that it differs from its mean by more than 2 standard deviations (in either direction) less than 5 percent of the time.

## 7.4 EXPERIMENTAL AND DATA MODELS

The prediction of the ES theory (for a given type of particle) can be stated as a predicted *standard deviation*[7]—the square root of the variance $2Dt$. Since Avogadro's number, $N$, is a function of $D$, once $D$ is estimated, Avogadro's number can be calculated. The calculated value can then be compared to the value hypothesized by the kinetic theory ($N^*$).[8] So the crux of Perrin's experimental test of the ES theory is evaluating the statistical hypothesis:

> *H:* The experimental displacement distribution is from a population distributed according to Gaussian distribution $M$ with parameter value a function of $N^*$.

$N^*$ is the (probable) value for $N$ hypothesized by the kinetic theory (approximately $70 \times 10^{22}$).

I do not want to be too firm about how to break down an inquiry into different models since it can be done in many ways. The central point is that a series of models of different types (as delineated in chapter 5) is needed to link actual data with primary hypotheses from a theory.[9] Hypothesis $H$ is what the kinetic theory predicts with respect

7. The standard deviation (square root of the variance) is the displacement in the direction of the x-axis that a particle experiences on average (root mean square of displacement). The importance of this statement of variance for the experimental determination of $D$ is that it states that the mean square displacement of a Brownian particle is proportional to the time $t$. This suggests that a model for Brownian motion is provided by viewing a particle as taking a *random walk*. We can get a rough idea of how this model leads to the Normal distribution as follows. (I follow the derivation in Parzen 1960, 374–76.)

Let $X_i$ be the displacement of a Brownian particle at step $i$ (projected onto a straight line). Consider the sum $S_n$ where

$$S_n = X_1 + X_2 + \ldots + X_n.$$

$S_n$ represents the displacement of a Brownian particle from its starting point. Since it has the same chance of being displaced a given amount in the positive and the negative direction, the average value of $X_i$ equals 0. From the central limit theorem (chapter 5) we have that the sum of these $X_i$, namely, $S_n$, is approximately Normally distributed.

8. The connection is this. Estimates of the diffusion coefficient $D$ indicate the approximate rate at which a particle is moving, from which we ascertain the average number of collisions to which these Brownian particles must be subjected to have caused such diffusion. This indicates approximately how many molecules per unit area there must be, that is, $N$.

9. Each individual task in an experimental inquiry can be seen as calling for a separate primary inquiry, with its own models of experiment and data. Then the full-blown experimental argument will string together different primary experimental arguments. Alternatively, the full-blown argument can be viewed as a single primary experimental argument, but with subarguments needed at different

to a *given* experimental context $E$. It may be located one step below the primary theoretical model in our hierarchy: it is part of the experimental model.

Note that hypothesis $H$ makes two assertions: it is an assertion about the distribution $M$ and about the values of parameters. The sample data from experiment $E$ can be used to estimate or test values of $N$ only if they can be seen (i.e., modeled) as the results of observing displacements from the hypothesized Gaussian process.

Correspondingly, I suggest we designate two primary questions that needed to be asked, one about the form of the distribution, the other about parameter values. Perrin's own discussion clearly distinguished between these two tasks, which I shall call step 1 and step 2. In a nutshell, step 1 consists of checking, for each experiment $E$, whether the results of the experiment actually performed follow the given statistical distribution $M$, and step 2 involves using estimates of $D$ to estimate or test values of $N$ (Avogadro's number).

Philosophers who discuss this case tend to place the task of step 2 at the forefront. Accordingly, they locate the impressive part of Perrin's argument in his estimates of Avogadro's number, which are close to what the kinetic account predicted. In fact, what made Perrin's results so impressive (at step 2) centered on his arguments for step 1: showing that the distribution of displacements was "completely irregular." Moreover, the argument against the nonstatistical version of the second law of thermodynamics hinged on the results at this first step. (I return to this in section 7.7.) What does step 1 look like?

### *Step 1: Manipulations on Paper*

This is a good example of a case where the substantive question is identified with testing a standard type of statistical hypothesis, one that asserts that the distribution of displacements is of the "chance" (or fortuitous or nonsystematic) variety. Since this statistical hypothesis is one piece of the investigation of $H$, we had better use a different letter (lower case, to indicate it is a portion of $H$). Take $j$:

> $j$: The data from $E$ approximates a random sample from the (hypothesized) Normal process $M$.

The *denial* of $j$, denote it as $j'$, roughly asserts that

> $j'$: the sample displacements of data from $E$ are characteristic of systematic (nonchance) effects.

---

nodes of the experimental context. I choose to model the present example employing the latter way of modeling.

It is noteworthy that here the hypothesis of chance, *j*, is not rejected but rather "passes" several tests. This is done by rejecting hypothesis *j'*.

So we have split off from the full problem this one question about a low-level statistical hypothesis *j*. But appraising this low-level statistical hypothesis—far from being a preliminary side show—was the main event and feature attraction of Perrin's work.

Affirming *j*, and ruling out *j'*, corresponds to affirming the key assumption about Brownian motion. As Perrin shows, Einstein's derivation of the displacement distribution depends on "making the *single* supposition that the Brownian movement is completely irregular" (p. 112). So ruling out hypothesis *j'* was the centerpiece of Perrin's work. Asking about *j'* came down to asking whether factors outside the liquid medium might be responsible for the observed motion of Brownian particles. The general argument in ruling out possible external factors—*even without being able to list them all*—was this: if Brownian motion were the effect of such a factor, then neighboring particles would be expected to move in approximately the same direction. In fact, however, a particle's movement was found to be independent of that of its neighbors. To sustain this argument, Perrin called up experimental knowledge gleaned from several canonical cases of ("real") chance phenomena.

Consider just one of Perrin's experimental tests of *j* against *j'*. It was based on an experiment *E* that consisted of observing 500 displacements of grains of gamboge (a microscopic vegetable particle).[10] Perrin considered these particular grains to be among his most uniform grains. To get the displacements, the positions of the grains (observed with a camera lucida) are recorded every 30 seconds on paper with grids of squares.[11] The path of a single grain might look as in figure 7.1.

The actual data consist of 500 scratch marks, each measuring the displacement of gamboge grains in 4 positions. To turn this into data that can answer questions posed in the experimental model regarding hypothesis *j*, the observations must be condensed and organized. This takes us to the level of *models of the data*. The idea is to do something that will enable the 500 actual outcomes (displacements) to be seen as a single random sample from the population of possible experimental outcomes (the sample space of the experimental model).[12] If we can

10. Perrin does not say if these 500 displacements are from a larger experimental run.

11. Were the positions recorded at much shorter intervals, each single segment in the figure would be as complicated as the entire figure.

12. If the experiment consists of marking off 500 displacements of a given type of gamboge grain (say at 30-second intervals), then the sample space may be seen to refer to the different 500-fold outcomes that could have resulted. Alternatively,

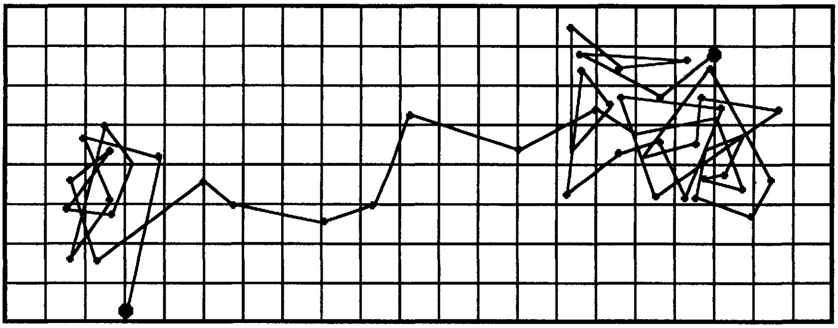

FIGURE 7.1. Tracing of horizontal projections of lines joining consecutive positions of a single grain every 30 seconds.

perform this feat, then we can ask of this *single* sample of 500 displacements whether it may be seen as a random sample from a population with the hypothesized Normal distribution.

Needed is a characteristic of the data—a statistic—such that this statistic, whose value we *can* observe, will teach us about the parent population that we cannot. Each such statistic refers to a different modeling of the data, and Perrin delineates several such models. That he does so is what makes his work such a treasure for the philosopher of experiment. Here I shall discuss the one data model that Perrin claims gives a "still more striking verification" (p. 117) than the others. It is obtained by looking at the value obtained from shifting each (horizontal) displacement to a common origin, and then counting how many are found at various distances from this common starting point. To sharpen the ability to distinguish between $j$ and $j'$, the data are condensed into 9 pigeonholes, each a different distance from the origin. As Perrin reasons,

> The extremities of the vectors obtained in this way should distribute themselves about that origin as the shots fired at a target distribute themselves about the bull's-eye. [See figure 7.2.]
> Here again we have a quantitative check upon the theory; *the laws of chance enable us to calculate how many points should occur in each successive ring.* (Perrin, 118; emphasis added)

This is a quintessential example of what I mean by "manipulations on paper." Nothing like a bull's-eye is actually observed in experiment

if the data are condensed into 9 pigeonholes, each a different size of displacement, then the sample space is the set of different numbers that could be observed in each of the 9 categories or rings.

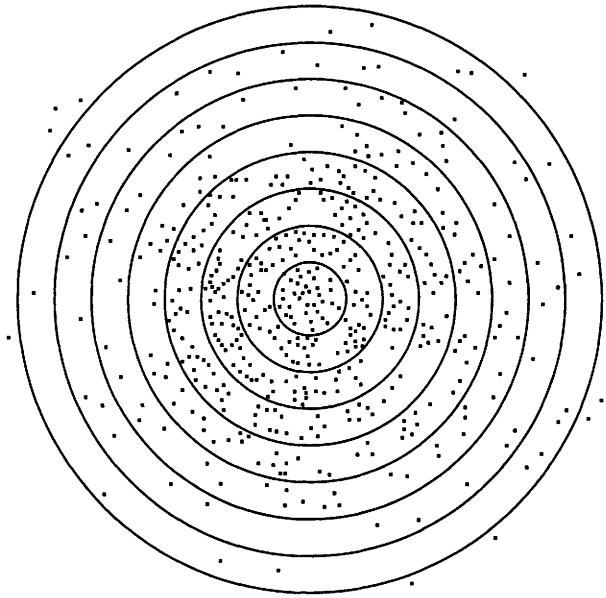

FIGURE 7.2. Manipulations on paper: displacements of 500 Gamboge grains.

*E*, but rather the displacements at 30-second intervals of several gamboge grains. The bull's-eye picture results not from physical manipulation of the grains, but from manipulations on paper. They serve much the same role in the experimental argument as physical instruments. They allow discerning patterns in the data hidden from an eye looking at 500 scratch marks. The manipulations on paper are warranted not because they represent actual experimental phenomena, but because once they are accomplished (e.g., once the data are manipulated into bull's-eye rings), "the laws of chance enable us to calculate how many points should occur in each successive ring" (Perrin, 118).

This tactic, so important yet so misunderstood, bears elaboration. The tactic is essentially this: Take a look at the handful of canonical models—in this case they are statistical distributions. In your tool kit are a bunch of random variables that have these distributions. Then think of ways of massaging and rearranging the data until you arrive at a statistic, which is a function of the data and the hypotheses of interest, and has one of the known distributions. (Your tool kit also contains standard ways of massaging and rearranging.) Nothing in front of you needs to actually have the distribution you arrive at, nor need it correspond directly to any actual event. It may simply be the distribution followed by the random variable arrived at through ma-

TABLE 7.1 (adapted from Perrin, p. 119)

| Displacement Between | Probability | $n$ calculated | $n$ found |
|---|---|---|---|
| 0 & first ring | .063 | 32 | 34 |
| 1st & 2nd rings | .167 | 83 | 78 |
| 2nd & 3rd rings | .214 | 107 | 106 |
| 3rd & 4th rings | .210 | 105 | 103 |
| 4th & 5th rings | .150 | 75 | 75 |
| 5th & 6th rings | .100 | 50 | 49 |
| 6th & 7th rings | .054 | 27 | 30 |
| 7th & 8th rings | .028 | 14 | 17 |
| 8th & 9th rings | .014 | 7 | 9 |

nipulations on paper (e.g., averaging, dividing by or adding appropriate numbers, squaring). But that is all you need to assign probabilities to various outcomes on the hypotheses being tested. Once you have this, statistical tests can be run and their error probabilities calculated. And these error probabilities (e.g., severity) *do* refer to the actual experimental test procedure.

Students of theoretical statistics are familiar with this sort of homework problem: Starting with a random variable with an unknown distribution, find a way to alter it (making use of what I call manipulations on paper) so as to arrive at a variable whose values vary in the manner of one of the known distributions.

Perrin's bull's-eye manipulations in testing hypothesis $j$ in the Perrin experiment exemplify this tactic. If the statistical hypothesis $j$ holds for the actual experiment, then we can deduce the probability that a displacement would fall in each of the 9 rings. We can deduce the number of displacements expected to fall in each ring by multiplying this probability by the number of displacements (500). This number is termed "$n$ calculated." That is, as shown in table 7.1 in the column labeled $n$ calculated, the expected frequency of displacements falling between the $i$th and the $i + 1$th ring is given by 500 multiplied by the probability of a displacement falling between the $i$th and the $i + 1$th ring.

This provides us with the ingredients for comparing the observed distribution of measurements ($n$ found) with a set of hypothesized probable measurements ($n$ calculated). The recipe comes from canonical arguments for asking: What should we make of the fit between observed and hypothesized? We want to know, especially when the differences between observed and expected are small, whether they may be merely the fluctuations typical of a sample of that size from a population of the assumed Normal distribution (as asserted in $j$), or whether a serious departure is indicated (as asserted in $j'$). We tackle this by asking what it would be like if $j$ were a correct description of the experiment. A quantitative argument is given by a statistical significance test much like the ones we have already sketched.

Here the distance measure chosen is a function of the difference between the number found and the number calculated for each of the 9 intervals:

$$\frac{n \text{ found} - n \text{ calculated}^2}{n \text{ calculated}}$$

Summing these up yields a statistic that has a known probability distribution (the chi-square distribution).[13]

What matters is being able to sustain the argument of the significance test (i.e., answer the "significance question"). This we can do because we can determine the probability that a purely random sample of measurements taken from the hypothesized model distribution $j$ would show worse agreement with the model $j$ than is shown by the actual set. If the differences observed are of the sort frequently "caused by" chance (i.e., if they are typical under $j$), then the sample data are in accord with the hypothesized experimental model. In Perrin's experiment $E$ above, it turned out that the observed differences, or one even larger, are not infrequent but rather are typical, assuming $j$. That is, the observed difference is not *statistically significant* from what is typical under $j$. Hypothesis $j$ passes.

Perrin's argument to this effect has weight only because he was able to argue further that if the model was inadequate (if $j'$ was true), we would very often get differences statistically significant from what is typical under $j$. In other words, he needed to argue that $j$ had passed a severe test. The multiple experiments for which Perrin stresses the need (e.g., Perrin, 96) are deliberately designed so that if one misses an error, another is likely to find it. The experiments are designed, to use David Hull's nifty phrase, to ensure that "errors ramify rapidly." The error of concern at this stage is that some regularity of Brownian motion has been concealed.

As is typical, Perrin's argument for severity was substantiated by reference to other tests. The overall argument goes beyond any single statistical significance test. Here is where the multitude of deliberately varied additional tests plays a particularly important role. In fact, the number of tests, checks, and rechecks Perrin performed amounted to statistical

13. The chi-square test, introduced by Karl Pearson in 1900, was actually in use before that. It was alluded to in testing assumptions in chapter 5.

overkill. He was not being overly cautious; he was well aware that others would unearth any weak point upon which to attack his arguments.

To this end, Perrin describes numerous sets of statistical analyses. Some made use of the *same* 500 measured displacements, only modeled in different ways; others involved further recorded displacements on the same gamboge preparation. Still others dealt with totally distinct gamboge experiments, where the key features were deliberately varied. (See table 7.2.) While the same question is being asked (Is $j$ adequate?), by phrasing it differently each test is designed to check if there are mistakes in the answers from other analyses.

### *Viewing Brownian Particles as Taking Random Walks*

Familiarity with standard chance mechanisms—quite independent of the ES theory—provided knowledge of experimental phenomena that are correctly described by hypothesis $j$, as well as phenomena more correctly described by $j'$. The experimental knowledge stemmed from the statistical theory of *random walk* phenomena (in one dimension). From this statistical theory about a very general type of fluctuation phenomenon, strategies for the experimental and data models in testing $j$ emerged. We see here just the sort of appeal to "real" random experiments that Neyman talked about (chapter 5).

The question in step 1, whether the displacements of Brownian particles could be characterized as the hypothesized Gaussian process $M$ (i.e., whether $j$ is adequate) is tantamount to asking, Can the displacements be modeled as a problem concerning a *simple random walk?* Such random walk problems were understood at that time. Einstein ([1926] 1956) had presented his derivation of the displacement distribution in $\mathcal{H}$ by means of a model of a random walk in one dimension[14] (see notes 7 and 22).

The displacement of a particle may be seen as the result of $k$ steps where at each step the particle has an equal chance of being displaced by a given amount in either a positive or negative direction. (This is called a *simple* random walk). Since it has the same chance of being displaced a given amount in the positive and the negative direction, on average, after $k$ steps, the displacement would be 0. Occasionally, more steps will be in one direction than the other, yielding a nonzero total displacement. That the variance is proportional to the time (the number of steps) corresponds to the fact that the more steps taken, the larger the value this nonzero total displacement can have.

14. For other clear derivations see Chandrasekhar 1954, 7 and Parzen 1960, 374–76.

A second type of canonical random experiment was used in deriving the statistical distribution in the ES theory—this one from gambling. Here one capitalizes on the fact that the displacement of Brownian particles is distributed like the winnings of a gambler who stands to win or lose a fixed amount $x$ with equal probability on each game. The more games played, the larger the gambler's loss can be.

There were yet other statistical derivations of the Normal distribution in the ES theory, all fascinating in their own right. I shall resist going into them here. My present purpose is to illuminate the strategy by which the hypothesized distribution in $\mathcal{H}$ is linked with experimental tests of $H$ by way of tests of hypothesis $j$. For each of several different experiments, abbreviated as $E_1, E_2, \ldots, E_n$, the predicted displacement distribution (in the experimental model) is given by hypothesis $j$. That is, for a given experiment $E_i$, we have

> $j$: the distribution of the $n$ displacements in $E_i$ is from a population distributed according to the Gaussian model $M$.

In other words, if displacements of Brownian particles follow the random walk model hypothesized in $\mathcal{H}$, then each experiment $E_i$ is, in the relevant respects, just like experimenting on known chance mechanisms. (This assumes of course that experimental assumptions are met, but we will deal with this separately in section 7.6.)

This probabilistic linkage between $\mathcal{H}$ and $j$ has two main functions. First, it allows deriving, for each experiment $E_i$, the experimental displacement distribution. This, recall, was the basis of the statistical significance test of $j$. Second, it is the basis for using the different experiments to cross-check and strengthen each test of $j$.

These linkages are at the heart of Perrin's argument that $j$ had passed a reliable (severe) test. The argument goes like this: if experiment $E$ was not correctly described by hypothesis $j$ (i.e., if $j'$ were true), there would *not* be an equal chance of being displaced by an amount in either direction for each particle: there would be some *dependencies*. But we know what it is like to interact with a mechanism with such dependencies. We know what would be expected in those sorts of experiments—we can even "display" it. We can actually generate the (frequency) distribution of the outcomes from experiments (on ball rollings or other chance apparatus)—where, by design, the probability of being deflected to the right is *not* equal to that of the left. (This would be an example of Neyman's "real" random experiments.) Or we can simulate these results by Monte Carlo methods or (now) by computer simulations. If we are observing such dependencies, then, we calculate, it should be fairly easy (frequent) to generate statistically significant differences in Perrin's various gamboge experiments from $j'$.

So we can argue that were $j'$, and not $j$, the case it is extremely improbable that none or even very few of the experiments $E_1, E_2, \ldots, E_n$ would have indicated this. It is very probable that a few would have shown differences statistically significant from what is expected under $j$. That is to say, the test was severe for $j$—where "the test" includes the results from several individual experiments. The pattern of arguing from error is clear. The experiments conducted by Perrin and his researchers had a very high probability of detecting a statistically significant difference from $j'$, were there dependencies in the motion, yet such differences were not detected. On the basis of such considerations, $j'$ was ruled out ($j$ was affirmed) by Perrin and others. After describing several different experiments (see table 7.2), all of which $j$ passed, Perrin remarks:

> Further verifications of the same kind might still be quoted, but to do so would serve no useful purpose. In short, the irregular nature of the movement is quantitatively rigorous. Incidentally we have in this one of the most striking applications of the laws of chance. (Perrin, 119)

## 7.5 STEP 2: ESTIMATING AND TESTING AVOGADRO'S $N$

Having found that hypothesis $j$ passed the tests, Perrin then asks, "But does it lead to these values for the molecular magnitudes that we look for?" (p. 103). Because of what was accomplished in step 1, Perrin is again in a position to split off a manageable piece from the full-blown problem in carrying out step 2. The error of concern at step 2 is that the ES theory is mistaken about the values of the magnitudes (e.g., $N$). Having chosen each experimental displacement distribution $E$ to be a function of the parameter of interest, Avogadro's number $N$, step 2 becomes a standard problem of estimating the parameter governing the displacement distribution. More precisely, step 2 involves using data models to test a hypothesis about the statistical parameter governing the probability distribution affirmed in step 1—the variance.[15]

The variance of the distribution, recall, is $2Dt$. Estimates of $D$ are experimentally obtained by calculating the mean-square displace-

15. To say that one or more parameters "govern" a statistical distribution means that if the values of the parameter(s) are given, then the probability of each possible outcome (i.e., the probability distribution) can be calculated.

ments of the numerous particles recorded.[16] Plotting the mean-square displacement against time on a graph yields a nearly straight line, whose slope is an experimental estimate of $D$. The remaining task, then, is to test a hypothesis about $D$ which is a function of $N$, Avogadro's number. (Since $D$ varies for different grains, it is preferable to work with $N$.) Perrin remarks:

> To verify Einstein's diffusion equation, it only remains to see whether the number [obtained for $N$ by substituting the estimate of $D$ into the equation $D = \frac{RT}{N}\left(\frac{1}{6\pi a\xi}\right)$] is near $70 \times 10^{22}$. (Perrin, 132)[17]

The value $70 \times 10^{22}$ is the value for $N$ that is hypothesized by the ES theory, abbreviated $N^*$. So the single task of interest has been boiled down to testing a simple statistical hypothesis $h$ expressed in the experimental model

$h$: $N = N^*$ $(70 \times 10^{22})$.

The falsity of $h$, i.e., "not-$h$," asserts that Avogadro's number is not near $N^*$. Given that Perrin wants to affirm $h$ so long as the data indicate that $N$ "is near," and not necessarily exactly equal to, $N^*$, it might be thought that $h$ should be an interval around $N^*$. For mathematical reasons it is easier to express $h$ as a "simple" or "point" hypothesis, and then take the "nearness" into account in the testing rule. That is what the standard statistical test does.

So step 2 may be broken down into an application of one or more canonical models for detecting errors in a parameter value. Although much of the modern apparatus for solving a standard estimation problem was not yet developed, Perrin's use of those notions that were available (e.g., probable errors) shows his argument to be based on standard error probability ideas. Perrin uses the data from each of the experiments $E$ to derive estimates of $N$ with known error properties (e.g., a known probable error[18]). (There is certainly no attempt to state prior probabilities and multiply them by likelihoods to yield posteriors.)

Interestingly, the estimates at step 2 come from observations on the same grains that were prepared and used in performing step 1.

16. The displacements of all the particles recorded in an experiment at some time $t$ are squared and the arithmetic average or mean is calculated. This is the mean square displacement at time $t$ (or the variance). The *root* mean square displacement is the square root of this (or the standard deviation).

17. $\xi$ is the viscosity of the fluid; $T$, its absolute temperature; $R$, the gas constant; $a$, the radius of the particles.

18. The probable error is around .7 of the standard deviation.

Only now a different question is asked of these grains and, correspondingly, different data models are formed. Here one needs to count the relative number of grains observed at different levels of the emulsion. The canonical statistical model appealed to is the exponential distribution. A complete investigation of this step would elaborate further on the variety of statistical models Perrin uses here and the shrewd manner by which he combines them to arrive at a reliable argument at step 2. Perhaps for our purposes enough has been said.

"As a matter of fact," Perrin reports, the number he obtains for $N$ "is equal to $69 \times 10^{22}$ to within $\pm$ 3 per cent" (p. 132). This is a good fit to $N^*$. Referring to the good fit of estimates from a number of experiments on Brownian motion (of emulsions) Perrin concludes:

> This remarkable agreement proves the rigorous accuracy of Einstein's formula and in a striking manner confirms the molecular theory. (P. 123)

In each case the "confirmation" of Einstein's hypothesis is based on a standard statistical argument. The difference between the estimated and hypothesized values of $N$ is the distance measure. Because "the discrepancy is well below the possible error introduced by the somewhat loose approximations . . . in making the calculations" (p. 127), hypothesis $h$ passes the test.

In describing the several sets of experiments from both steps 1 and 2, Perrin emphasizes again and again the deliberate attempt to vary several aspects of the experiment:

> I have carried out personally, or directed in others, several series of measurements, varying the experimental conditions as much as I was able, particularly the viscosity and the size of the grains. (P. 122)

He summarizes them in the following (abridged) table (table 7.2).

Nearly all the estimates of $N$ from these experiments were (statistically) insignificantly far from that predicted by the kinetic theory, $N^*$ (i.e., $70 \times 10^{22}$). Perrin declares:

> It cannot be supposed that, out of the enormous number of values *a priori* possible, values so near to the predicted number have been obtained by chance for every emulsion and under the most varied experimental conditions. (P. 105)

It should be supposed, instead, that the reason values so close to the predicted value $N^*$ can be repeatedly generated is that $N^*$ is approximately correct. In terms of experimental knowledge, this means that in a (hypothetical) population or series of experiments, the mean value

TABLE 7.2 (from Perrin, p. 123)

| $100\xi$ | Nature of the Emulsion | Radius of the grains $\mu$ | Mass $m \times 10^{15}$ | Displacements Recorded | $\frac{N}{10^{22}}$ |
|---|---|---|---|---|---|
| 1 | I. Gamboge grains | .50 | 600 | 100 | 80 |
| 1 | II. Gamboge grains | .212 | 48 | 900 | 69.5 |
| 4 to 5 | III. The same grains in sugar solution | .212 | 48 | 400 | 55 |
| 1 | IV. Mastic grains | .52 | 650 | 1,000 | 72.5 |
| 1.2 | V. Very large grains (mastic) in urea solution | 5.50 | 750,000 | 100 | 78 |
| 125 | VI. Gamboge grains in glycerine | .385 | 290 | 100 | 64 |
| 1 | VII. Gamboge grains of very uniform | .367 | 246 | 1,500 | 68.8 |
| | equality (two series) | | | 120 | 64 |

$\xi$ = the mean value of the viscosity

for Avogadro's number would be $N^*$,[19] and in actual experiments, the deviations from $N^*$ would be of the pattern ascribable to chance.

Let us summarize the argument in step 2. Step 1 affirmed that the experimental distribution is the Gaussian one hypothesized in *j*. Given step 1, we know how to design experiments that make it very difficult to generate results close to $N^*$ unless we really are sampling from a population where $N$ is approximately $N^*$, that is, unless hypothesis $h$ is true. This allows us to design an experimental test of $h$ such that if $h$ passes, then $h$ has passed a severe test.

The denial of hypothesis $h$ asserts that there is genuine discord between the hypothesized value, $N^*$, and the "true" value, where the true value refers to the mean or expected value for $N$ in a population (or long series) of experiments. If there is such discord (the extent of which can be made rigorous), then our experiment has given it a good chance (often) to manifest itself. It would manifest itself by producing an observed discrepancy statistically significantly beyond the possible experimental error introduced in estimating $N$. So each experimental test itself is a reasonably severe test of $h$. The several experiments taken together are further checks and thereby strengthen the overall severity.

The argument follows the pattern of arguing from error. If we are wrong in any single experiment that results in passing $h$, then it should

19. That is because, on the average, our estimate of $N$ equals the population mean $N^*$. That is, the average (i.e., mean) value of our estimate equals the population mean. This was discussed in chapter 5.

be very hard to reproduce results "close" to $N^*$ in experiments especially designed to display discordances (by revealing statistically significant differences). But Perrin shows we can generate at will (very frequently) estimates of $N$ near the hypothesized value $N^*$. Therefore, he can argue for the overall reliability of the argument. In terms of a probability calculation, he can say that

$P$ (such good accordance in experiments $E_1, E_2, \ldots, E_n \mid h$ is false) = very low,

meaning that $h$ passes a severe test.

The multiple experiments listed in table 7.2 rule out mistakes or "other hypotheses" in both steps 1 and 2, but the mistakes differ. In step 1 they were used to rule out systematic effects. In step 2, they improved the reliability of the estimates of $N$ and in addition helped rule out mistakes having to do with generalizability. By being deliberately varied in the experiments, any influences of the liquid, the temperature, the nature and density of the grains, and so on would affect Perrin's estimates in all directions and so cancel each other out (in the mean), leaving an extraneous effect comparable to experimental or chance error. This is *statistical control.* These multiple experiments also check errors regarding the generalizability of results. Perhaps the ES theory applies only to particular grains or experimental circumstances. Appropriate variations let us rule this out.

Now for the questions and problems that arise in checking experimental assumptions. They correspond to problems and models that would be placed "below" the data models, and our analysis would be seriously incomplete if we did not address them.

## 7.6 CHECKING EXPERIMENTAL ASSUMPTIONS: EXPERIMENTAL DESIGN AND DATA GENERATION

A key problem I have placed at the level of data models in an experimental testing context is that of checking that the various experimental assumptions are satisfied. Their violation may introduce alternative explanations for the results and may thereby vitiate experimental arguments. Checking on experimental assumptions sends us to considerations "below" the data models, to those of experimental design or data analysis. We have to look at how the experimental objects—grains in various solutions—were generated and measured to produce raw data.

In attending to the actual data-generation procedures, it becomes plain that to begin the analysis with the estimates of $N$ (Avogadro's number) as the data is too simple. Estimating $N$ calls for a full-blown inference in its own right. Each estimate of $N$ depends on being able to obtain estimates of a number of other quantities, along with their associated errors. Most notably, the experiments turn on being able to accomplish an experimental tour de force—obtaining Brownian particles each with a fairly uniform radius.

What makes Perrin's discussion so valuable for us is his careful explication of the labors required to justify experimental assumptions. He constantly stresses the need to search for and rule out errors by multiply-connected checks and tests. Here one finds ample illustration of before-trial procedures of experimental design (e.g., to ensure that the gamboge preparations are likely to be useful) and after-the-trial checks of whether assumptions are approximately met. It is impossible to appreciate the full force of Perrin's tests without delving into these details. Here are some highlights.

*Measuring Microscopic Grains of Gamboge*

After unsuccessful attempts to use the substances usually studied, Perrin hit upon gamboge:

> *Gamboge* (which is prepared from a dried vegetable latex) when rubbed with the hand under water (as if it were a piece of soap) slowly dissolves giving a splendid yellow emulsion, which the microscope resolves into a swarm of spherical grains of various sizes. (P. 94)

The force of Perrin's results, at bottom, hinged on his uncanny ability to ensure that the particles were of approximately the same size, that they could be counted and weighed, and that a host of extraneous factors could be controlled or "subtracted out"—even Einstein expressed surprise. Most impressive of all, perhaps, was the preparation of grains of uniform size. The key was the special technique Perrin developed for "fractional centrifuging."

> The emulsion having been obtained, it is subjected to an energetic centrifuging (as in the separation of the red corpuscles and serum from blood). (P. 94)

A top layer of sediment is formed and poured off, and the grains are again suspended in (distilled) water. The centrifuging and pouring off processes are repeated again and again until the emulsion is practically pure water.

> But the purified emulsion contains grains of very various sizes, whereas a *uniform* emulsion (containing grains equal in size) is required. (Pp. 94–95)

By further centrifuging, it is possible to separate out the grains according to size, the first layers of sediment containing the largest grains. Getting an adequate separation in this manner was an extremely lengthy process. In his most careful measurements, Perrin tells us, he labored over his gamboge for several *months* of treatments:

> I treated 1 kilogramme of gamboge and obtained after several months a fraction containing a few decigrammes of grains having diameters approximately equal to the diameter I wished to obtain. (P. 95)

*Reliably Measuring the Density, Volume, and Weight of the Grains*

Numerous interconnected checks and rechecks were used in scrutinizing this and other assumptions. The key to ruling out errors was a deliberate variability. Ascertaining the volume of the gamboge grains exhibits the standard pattern:

> Here again, as with the density, it is possible, on account of the smallness of the grains, to place confidence only in results obtained by several different methods. (P. 96)

The different methods desired are those that allow arguing that any error present is very likely to be detected by at least one method. When, through the several methods, Perrin obtained concordant results, he could rule out experimental artifacts. Here he is clearly arguing from coincidence along the lines sketched in Hacking's example of dense bodies (chapter 3); at this stage there is no need for a formal assessment of the degree of severity.

One of the methods used in measuring volume involved measuring the radius of the grains in the camera lucida.

> Considerable error is involved in the measurement of isolated grains (owing to the magnification by diffraction . . .). This source of error is very considerably minimised if it is possible to measure the length of a known number of grains in a row. (P. 96)

If Perrin could find a way to get his grains to line up in a row, he could appeal to a canonical technique for counting objects that had nothing to do with Brownian particles. He discovered that if he let a drop of the emulsion nearly evaporate, capillary forces made the grains run together

> and . . . collect together into groups a single grain in depth and more or less in rows, in the same way that the shot are arranged in a horizontal section through a pile of shot. (P. 97)

It was then possible to count the number of grains lying in a row.

This exemplifies another thread woven through Perrin's work. The counting procedure did double duty: it was also used to check a measurement somewhere else. Perrin continues:

> At the same time a general check upon the uniformity of the grains sorted out by the operation of centrifuging is obtained. The method gives numbers that are perhaps a little too high (the rows not being quite perfect); but owing to its being so direct it cannot be affected by large errors. (P. 97)

For Perrin to learn what he is after, he needs to count his carefully prepared grains. If he can make them arrange themselves like a pile of shot, he can not only count them but check on his centrifuging results. Perrin may be said to have invented his techniques for preparing and working with his grains, but the models for analyzing the errors were standard and did not belong to any one domain.

In some cases mistakes were made and later detected. For instance, one estimate of *N*, while close to the predicted value, was invalidated because it was noticed "during the course of some measurements on some preparations . . . that the proximity of a boundary checked the Brownian movement. (Einstein's theory presupposes an unlimited fluid.) . . . These measurements will be repeated" (p. 124). In other experiments, Perrin deliberately exploited the phenomenon of grains sticking to the walls!

I hope this suffices to get the flavor of how tactics of observation and measurement may be pieced together into an experimental argument that allows learning about primary hypotheses. The complexity of the hierarchy of models in an experimental inquiry can be grouped into two sets of arguments. One links data models with primary hypotheses (via severe tests), a second substantiates the assumptions of the data models.

In addition to the checks of measurement errors, Perrin sought ways to check himself. To avoid a type of experimenter's bias when selecting which grain to follow at steps 1 and 2, Perrin explains that

> in order not to be tempted to choose grains which happened to be slightly more visible than the rest . . . , which would raise the value of *N* a little, I followed the first grain that showed itself in the centre of the field of vision. (P. 124)

He was trying to obtain a randomly selected grain.

In this sketch of Perrin's experimental arguments two themes I have been tracing surface. First, knowledge of the ways in which one can go wrong leads to multiple procedures that allow errors to be circumvented or dealt with. At the level of data generation and measurement, in this example, the procedures refer more to actual physical manipulations (e.g., spinning the grains, lining them up, forcing some to get stuck on a barrier) than to the "manipulations on paper," which I said was the hallmark of data modeling. Second, validating experimental assumptions is much less a matter of ensuring that errors are not made than it is of knowing how much error is likely to be introduced by the various data generation procedures. Perrin's genius as an experimenter was largely a matter of his skill at catching himself and his students making mistakes, as well as knowing when they do not matter much. These two themes become even more pronounced in considering the ceteris paribus conditions.[20]

## 7.7 The Most Risky Decision of All: Ceteris Paribus Conditions

Now for the ceteris paribus conditions, at the bottom of the Suppean hierarchy, those factors not included in the systematic checks of experimental assumptions discussed above. The manifold factors, known and unknown, that are part of this soup are often thought to be the locus of a set of alternative hypotheses that cast an ever-present shadow on any primary inference. Given the way I have broken down experimental inquiries, this set of alternative hypotheses consists of threats to the experimental assumptions (or "initial conditions") of some primary inquiry. Because some such alternative ceteris paribus factors are assumed always to exist as threats to experimental assumptions, and because the ability to pinpoint what is learned hinges on these assumptions, ceteris paribus factors are thought to threaten the correct attribution of blame or credit to primary hypotheses. As Lakatos put it:

> The plight [of the methodological falsificationist] is most dramatic when he has to make a decision about *ceteris paribus* clauses, when he has to promote one of the hundreds of "anomalous phenomena" into a "crucial experiment," and decide that in such a case the experiment was "controlled." (Lakatos 1978, 27)

20. Although in delineating the framework of models in chapter 5 I combined checking experimental assumptions with checking ceteris paribus conditions, it is often useful to make a distinction between the two within this level, as I am doing here. The former refer to those experimental assumptions that are amenable to formal statistical testing—as in the case of affirming the Normal distribution *M*. The latter (ceteris paribus factors) refer to the variety of influences that are either known to be controlled, are subtracted out, or are tested by more informal, domain-specific means.

Affirming the ES hypothesis about the distribution of Brownian particles did deny a nonstatistical version of the second law of thermodynamics—something we will discuss in the next section. So the present example involves the very thing Lakatos worried most about: turning a mere anomaly into a severe test.

Lakatos, recall, gives up on *justifying* control; at best we decide—by appeal to convention—that the experiment is controlled. While I have no desire to revive "methodological falsification," I reject Lakatos and others' apprehension about experimental control. Happily, the image of experimental testing that gives these philosophers cold feet bears little resemblance to actual experimental learning. Literal control is not needed to correctly attribute experimental results (whether to affirm or deny a hypothesis). Enough experimental knowledge will do. Nor need it be assured that the various factors in the experimental context have no influence on the result in question—far from it. A more typical strategy is to learn enough about the type and extent of their influences and then estimate their likely effects in the given experiment.

How was this problem dealt with in Perrin's experiments? First remember that a host of experiments on factors suspected of influencing Brownian motion had already been conducted before Perrin's tests (around 1910). (See section 7.2.) Much was already known about the influences of light, heat, magnetism, electricity, shaking, noises of various sorts, and so on. Michael Faraday (in an 1829 lecture) and others recognized Brown's experiments as having ruled out all the causes of the motion suggested up until that time (e.g., unequal temperatures in the water, evaporation, air currents, heat flow, capillarity, motions of the observer's hands).[21] It is worth noting that these early experiments on the possible cause of Brownian motion were not testing any full-fledged theories. Indeed, it was not yet known whether Brownian motion would turn out to be a problem in chemistry, biology, physics, or something else. Nevertheless, a lot of information was turned up and put to good use by those later researchers who studied their Brownian motion experimental kits.

As I am imagining it, in one's bag of experimental tricks, along with the experimental tools and past experimental mistakes, would be a log of the extant experimental knowledge of the phenomena in question. Astutely using the kind of log I have in mind, Perrin dispelled numerous threats to experimental control. An imaginary (but not far from actual) dialogue quoting Perrin might have gone like this (the

21. See, for example, Jones 1870, 403.

names in parentheses are those of the researchers who worked extensively on the question of interest, and Perrin's lines are direct quotes):

*Questioner:* How do you know your results are not due to variations in temperature throughout the experimental emulsion? I believe there must be some temperature variations.

*Perrin:* It makes no difference whether great care is taken to ensure uniformity of temperature throughout the drop; all that is gained is the suppression of the general convection currents, which are quite easy to recognise and which have no connection whatever with the irregular agitation under observation (Wiener, Gouy). (P. 84)

*Questioner:* Might it not be something in the composition of your grains?

*Perrin:* The nature of the grains appears to exert little influence, if any at all. In the same fluid two grains are agitated to the same degree if they are of the same size, whatever the substance of which they are composed and whatever their density (Jevons, Ramsay, Gouy). (P. 85)

*Questioner:* What about vibrations of the glass containing the emulsion? Heavy vehicles passing by have made your table shake.

*Perrin:* The Brownian movement, again, is produced on a firmly fixed support, at night and in the country, just as clearly as in the daytime, in town and on a table constantly shaken by the passage of heavy vehicles (Gouy). (P. 84)

Numerous other factors are likewise shown to be either irrelevant or accounted for.

Perrin also uses his Brownian motion log to explain why, at one time, certain factors were erroneously thought to have influenced Brownian motion. For example, it had been thought that adding impurities such as acids to the emulsion influenced the motion. The error arose from the fact that the impurities caused the particles to stick to the glass vessel when they touched the sides. In actuality, "the addition of impurities . . . has no influence *whatever* on the phenomenon (Gouy, Svedberg)" (Perrin, 85, n. 1).

Each of these separate experimental inquiries had its own set of primary questions, experimental and data models, and so on. Had they not been available—as is often the case—then the separate tests would have had to be conducted as part of Perrin's experimental testing context. As it happened, by the time he performed the tests we have been discussing it was sufficient just to cite the familiar studies.

But Perrin did not do that; he repeated nearly all the tests anyway! Why? The primary reason, I suggest, is one that emerged during our discussion of "normal science": Deliberately repeating and getting good at generating *anticipated* results teaches a great deal about interacting with one's experimental objects. It is this kind of deliberate practice, not some mysterious knack, that gives one "a feeling for" one's experimental objects.

How far we have come from sighing over infinitely many alternative hypotheses that fit the data equally well. How far away, too, from the Bayesian requirement to consider all the alternatives in the catchall, plus our degrees of belief in them, to get a single inquiry going. Clearly, ruling out alternatives is not always possible, and even when it is it typically takes a lot of work and aggressive criticism. The good experimenter notices when the criticism may be lodged in terms of specific experimental questions. These questions find their place at some level and at some node of the hierarchy of models. With respect to such specific experimental questions, the infinitely many alternatives really fall into just a few categories. Experimental methods (for answering new questions) coupled with experimental knowledge (for using techniques and information already learned) enable local questions to be split off and answered. These answers in turn may be used to show that experimental assumptions are sufficiently well met for testing a primary hypothesis severely.

Table 7.3 displays the main aspects of the series of models of inquiry that we have discussed. Here I chose to place the data and experimental models side by side.

## 7.8 What Is Learned about the Sources of Experimental Effects

Our experiments teach us or indicate the correctness of any hypothesis that may be deemed to have passed a severe test. Whereas in some cases the assessment of severity comes directly from formal statistical calculations (from a test's error probabilities), in others the argument for severity is based on analogies with known canonical (statistical or other) models. In Perrin's experiments, calculations (based on significance tests) show that the two primary hypotheses in the theoretical model pass highly severe tests. Step 1 teaches Perrin that the experimental results on Brownian motion approximate a random sample from a specified Normal distribution *M*. Step 2 indicates the values of the parameters of this distribution law.

This tells us, for starters, that the Brownian motion of a variety of types of particles is satisfactorily modeled as the realization of a particular statistical process identified (in model *M*) as Normal with certain approximate parameter values—as asserted in the theoretical distribution $\mathcal{H}$. Perrin's experiments also indicate how to generate manifestations of that process. We can paraphrase that favorite passage

TABLE 7.3 Selection of Entries From Models of Inquiry: Brownian Motion

PRIMARY MODEL:
Hypotheses: $\mathcal{H}$: the displacement of a Brownian particle over time $t$, $S_t$, follows the Normal distribution with $\mu$ = 0 and variance = $2Dt$.

| MODELS OF EXPERIMENT: | MODELS OF DATA: |
|---|---|
| Hypotheses: $H$: the distribution of $n$ displacements in experiment $E_i$ follows Normal model $M$ with parameters a function of $N$ | Hypotheses: Data set is a random sample from experimental model |
| *Break down into steps:* Step 1, test of hypothesis $j$, and Step 2, test or estimation of parameters $D$ (or $N$) (using data models)<br>*Problems:* Specify the number of displacements to record, choice of experimental (test) statistics, specify adequate error probabilities | *Data Models:* $n$ displacements, observed distributions of grains, measurements of grains<br>*Problems:* How to condense data from one or more experiments to (a) arrive at suitable data models and (b) check if assumptions of experiment hold in each actual experiment $E_i$ |

CANONICAL MODELS: Normal distribution, random walks, bull's-eye model, gambling models, random selection, piles of shot

EXPERIMENTAL DESIGN AND DATA GENERATION
Fractional Centrifuging (may take several months), prepare emulsions, microscopic techniques for following displacement distributions of grains, count and weigh the grains, check for uniform radius.

*Ceteris paribus conditions*
Miscellaneous factors in experiment: uniformity of temperature, color and intensity of light, vibrations, impurities, the nature of the grains

*Problems:*
(a) How to manipulate grains and emulsions to arrive at data that satisfies experimental assumptions.
(b) How to ensure control of relevant factors, subtract out their influences, or determine that they need not be controlled. Utilize log of previous experiments on influences of background factors.

from Fisher (section 4.3): From these experiments we know how to bring it about that estimates for $D$ (or for $N$) will very rarely be significantly far from certain values. This is experimental knowledge. The currently accepted value for $N$ is still close to Perrin's values (within 19 percent).

But there is more to be said about what we learn from passing the theoretical distribution hypothesis $\mathcal{H}$. We do not learn about the distribution function of Brownian motion without also learning something about what produces it—something about molecular motions. The molecules about which the experiment teaches, however, agreeing with Nancy Cartwright, need not be the molecules of some substantive theory. The experiment teaches us about a cluster of causal or experimental properties of molecular motion. Minimally, it teaches that molecular motion reliably manifests itself as a Normal (Gaussian) process in Brownian motion experiments. Molecules in motion possess those properties that enable Perrin's experimental effects (e.g., estimates of *N*) to be reliably produced and reproduced.

Often, knowing this much (together with background knowledge) lets us go further. It may let us arrive at an understanding, if only approximate or partial, of specific properties of the underlying processes triggered in experiments. Perrin can also argue with severity, although without a precise severity assignment, that the experimental effects indicate that we are interacting with a process with certain characteristics, if only in the aggregate. We are familiar with the characteristic types of statistical processes that, when triggered in the manner of Perrin's experiments, produce data distributions of the sort he finds he can generate. More important, we know that the result of altering the underlying processes in specific ways (e.g., creating slight dependencies) would have been manifested in experiments like Perrin's.

Perrin gives the following analogy:

> Direct perception of the molecules in agitation is not possible, for the same reason that the motion of the waves is not noticed by an observer at too great a distance from them. But if a ship comes in sight, he will be able to see that it is rocking, which will enable him to infer the existence of a possibly unsuspected motion of the sea's surface. Now may we not hope, in the case of microscopic particles suspended in a fluid, that the particles may, though large enough to be followed under the microscope, nevertheless be small enough to be noticeably agitated by the molecular impacts? (Perrin, 83)

Just as the rocking of a ship indicates the motion of the sea's surface, the Brownian motion of microscopic particles indicates the motion in the liquid medium. Perrin puts it this way:

> The objective reality of the molecules . . . becomes hard to deny. At the same time, molecular movement has not been made visible. The Brownian movement is a faithful reflection of it, or, better, it is a molecular movement in itself. . . . From the point of view of agitation,

> there is no distinction between nitrogen molecules and the visible molecules realised in the grains of an emulsion. (P. 105)

The microscopic grains are small enough to be noticeably agitated by the molecular collisions yet large enough to be observed under a microscope. Of course, molecular collisions are still occurring on the macroscopic level, but they do not displace a suspended body (indeed, this encouraged the initial skepticism toward the kinetic theory). The reason is that the breadth of surface area on average counterbalances the many collisions in different directions (a consequence of the law of large numbers). With microscopic particles, in contrast, the impulses from the collisions do not generally counterbalance each other; the particles are tossed about irregularly.

From Perrin's experiments, and with the knowledge of fluctuation phenomena, we can delimit at least major aspects of the kinds of things that can produce all of this. It must be something in the liquid medium—a discrete-hit type of process approximating a random walk.[22] This is what we can give a severe argument for, at least limiting ourselves to Perrin's experiments.

In this connection it is important to note that the knowledge of the existence of Brownian motion led to a change in scientific methodology. A new limit to experimental accuracy due to Brownian fluctuations and "noise" in measuring systems was introduced. Methods of testing were revised accordingly, and updated experimental tool kits needed to reflect this. Indeed, Brownian motion was and is one of the most important sources of canonical models of types of errors and fluctuations.

A number of mistakes and ways of overcoming them, all gleaned from Perrin's inquiries, also go into our experimental kit. One not yet mentioned deserves special note. Understanding Brownian motion unearthed a general type of statistical error that many people had overlooked. The error was the basis for an important objection to the kinetic account first raised by Karl Nägeli. The objection was based on the common assumption that in order for the molecules to cause

22. The major approximation in the modeling of Brownian processes stems from the fact that it can only be seen as a random walk when the interval of time *t* is not too small. As Einstein notes:

> The movements of one and the same particle after different intervals of time must be considered as mutually independent processes, so long as we think of these intervals of time as being chosen not too small. (Einstein [1926] 1956, 12–13)

See also Chandrasekhar 1954, 89.

Brownian motion, they would have to move in a coordinated fashion. Random motion, it was objected, could not explain Brownian motion.

Gouy was the first to come close to answering this objection by citing the law of large numbers (chapter 5). Smoluchowski gave a more rigorous explanation based on a statistical argument of which Nägeli and others had been unaware. The argument is based on the canonical model of random walk phenomena discussed earlier. To use the gambler analogy, what the argument shows is how a gambler can lose a great deal of money, even with an even chance of winning or losing a fixed amount on each game, provided that he plays long enough. Analogously, unlike what Nägeli supposed, the jerks of Brownian particles do not need coordinated motion to explain them; with enough hits, the average displacement can be large, even when each hit has an equal chance of moving the particle a given amount to the right or left.

Models of fluctuation phenomena added considerably to the methods and strategies of experimental tool kits.

## 7.9 ACCEPTING A STATISTICAL VERSION OF THE SECOND LAW: A BIG SHAKE-UP TURNS ON A "SMALL" RESULT

Perrin's results on Brownian motion are sometimes considered to have provided a crucial test between the kinetic theory and classical thermodynamics taken as a whole. But his experiments themselves are not a severe test of the full kinetic account. There are many ways the full kinetic theory can be in error that Perrin's experiments did not direct themselves to uncovering.

Nevertheless, they do provide a severe and crucial test of one small though key piece of the kinetic account against one piece of the classical one. The piece turned on the severe test that took place at step 1: the test of the hypothesized Normal distribution in $j$ against $j'$. And what was learned from it, based only on local statistical tests, took on enormous importance in the debates. It provided a severe test in favor of a statistical version of the second law of thermodynamics (also referred to as Carnot's principle). Friedrich Ostwald, Ernst Mach (at times), and the mathematician Ernst Zermelo based at least part of their opposition to the molecular-kinetic theory on the fact that it would allow exceptions to the absoluteness of the second law of thermodynamics. Mach and Ostwald held that a phenomenological description, such as thermodynamics, contains sufficient information while escaping the various problems that plagued atomic theory (e.g., the use of entities deemed hypothetical).

Einstein ([1926] 1956) begins by stressing that the two theories give conflicting predictions about Brownian motion:

> If the movement discussed here can actually be observed (together with the laws relating to it that one would expect to find), then classical thermodynamics can no longer be looked upon as applicable with precision to bodies even of dimensions distinguishable in a microscope: an exact determination of actual atomic dimensions is then possible. On the other hand, had the prediction of this movement proved to be incorrect, a weighty argument would be provided against the molecular-kinetic conception of heat. (Pp. 1–2)

Perrin, as expressed in the second epigraph that opens this chapter, regarded his experiments as providing such a crucial test.

The kinetic theory, in contrast with the classical theory, views dissolved molecules as differing from suspended particles only in their size; their motion would be the same. If Brownian motion could be explained as caused by something outside the liquid medium or something within the particles themselves, it would not be in conflict with the classical theory. If, on the other hand, it could be shown that Brownian motion was caused by a molecular motion in the liquid medium, as given in the kinetic theory, it *would* be in conflict. Moreover, it would show that a statistical process was responsible and that the second law (or Carnot's principle) requires a statistical rendering.

That Brownian motion, if indeed spontaneous, would be an exception to the (nonstatistical version of the) second law was recognized even before the ES theory was formulated. For if it is true that without temperature differences in the system, a Brownian particle (denser than water) rises spontaneously, then it constitutes a case in which part of the heat of the medium is transformed into work. This recognition is explicitly discussed by Gouy around 1890. Jules-Henri Poincaré, persuaded by Gouy's arguments, declares:

> But we see under our eyes now motion transformed into heat by friction, now heat changed inversely into motion, and that without loss since the movement lasts forever. This is the contrary of the principle of Carnot. (Poincaré 1905, 610)

In step 1, Perrin showed that one can generate at will an observable process due to an agitation *not* attributable to the particles or external energy sources. Thus in carrying out step 1, Perrin demonstrates the existence of violations of the nonstatistical version of the second law. Perrin even describes his experiments as methods for generating such violations:

> Briefly, we are going to show that sufficiently careful observation reveals that at every instant, in a mass of fluid, there is an irregular spontaneous agitation which cannot be reconciled with Carnot's principle *except just on the condition of admitting that his principle has the probabilistic character suggested to us by molecular hypotheses.* (Perrin 1950, 57; emphasis added)

This is precisely what is established by the severe test of the distribution in hypothesis *j*.

The statistical version involves a standard frequentist interpretation of probability: it means that it will be violated extraordinarily rarely—so rarely that it can practically be discounted. Perrin calculates, for example, that to have a better than even chance of seeing a one-kilogram brick suspended by a rope rise to a level by virtue of its Brownian motion, one would have to wait more than $10^{10}$ billion years. "Common sense tells us, of course, that it would be foolish to rely upon the Brownian movement to raise the bricks necessary to build a house" (p. 87). So, practically speaking, the second law is unaffected. Perrin suggests that we can best understand the law by stating it as follows:

> *On the scale of magnitudes that are of practical interest to us, perpetual motion of the second kind is in general so insignificant that it would be foolish to take it into consideration.* (P. 87)

Perhaps enough has been said for our purposes, which were to illustrate the hierarchy of models in a single experimental testing context, the breakdown of larger inquiries into small pieces, and strategies for arriving at severe tests. There is ample work by others on how Perrin's results bear on theories higher in the hierarchy as well as on more global disputes arising from the atomic debates.[23] I limit myself to a few brief remarks.

### *Going Higher in the Hierarchy*

At yet a higher level in the hierarchy of models one could place more general questions about the molecular-kinetic theory as a whole. The more global molecular-kinetic theory refers not only to Brownian motion but also to theories about gases, radiation, diffusion of light, and others. Here is where the discussion of the thirteen phenomena enters. In experiments upon each of these widely different phenomena, estimates of Avogadro's number $N$ were obtained, and good agreement was found. Although Perrin takes several chapters to discuss the tests on these other phenomena, they are distinct from his

23. Examples are Brush 1977; Clark 1976; Gardner 1979; and Salmon 1984.

tests, involving primary hypotheses different from those he considered. While passing these further molecular-kinetic hypotheses adds weight to the Brownian motion tests, they largely come into play only when going beyond the Brownian motion tests that are my focus.

The good agreement among the thirteen phenomena on the molecular magnitudes effectively ruled out the worry that extrapolations from one phenomenon to another would not hold up. It was also at the heart of arguments for the reality of atoms, as Salmon and others have maintained; which is why those arguing for realism began with the argument from the thirteen phenomena. Even ardent antiatomists (probably excepting Mach) construed Perrin's experiments as telling. On the basis of such experimental evidence, even Ostwald reversed himself on the atomic-kinetic theory in 1909:

> *I have convinced myself that we have recently come into possession of experimental proof of the discrete or grainy nature of matter, for which the atomic hypothesis had vainly sought for centuries. . . .* This evidence now justifies even the most cautious scientist in speaking of the *experimental* proof of the atomistic nature of space-filling matter.[24]

As I have said, I am confining myself to what is given by experimental knowledge, and it is not clear that this does not take one as far as one would like. What did Perrin think? In some passages one hears him arguing for molecular reality (see Achinstein 1994, Nye 1972). But this is not the chief concern of his experimental work. Even on the role of the thirteen phenomena, Perrin has this to say near the end of *Atoms:*

> Yet, however strongly we may feel impelled to accept the existence of molecules and atoms, we ought always to be able to express visible reality without appealing to elements that are still invisible. And indeed it is not very difficult to do so. We have but to eliminate the constant $N$ between the 13 equations that have been used to determine it to obtain 12 equations in which only realities directly perceptible occur. (P. 216)

As an example, Perrin explains that by eliminating the molecular parameter between the equations from black body radiation and Brownian motion, we arrive at an equation that lets us predict the rate of diffusion of Brownian particles in water by measuring the intensity of the light in the radiation issuing from a furnace of molten iron:

24. This translated quotation is from Brush 1977, 381. The source is Wilhelm Ostwald 1909. The quotation is from the "Vorbericht."

> Consequently the physicist who carries out observations on furnace temperatures will be in a position to check an error in the observation of the microscopic dots in emulsions! And this without the necessity of referring to molecules. (P. 216)

The thirteen equations make fundamental connections among very different phenomena, therefore providing an effective way of using one such phenomenon to check errors regarding vastly different phenomena. This is a powerful source of progress in experimental knowledge.