P's in a Pod: some recipes for cooking Mendel's data

Teddy Seidenfeld
Departments of Philosophy and Statistics
Carnegie Mellon University
May-98

ABSTRACT

In 1936 R.A.Fisher asked the pointed question, "Has Mendel's Work Been Rediscovered?" The query was intended to open for discussion whether someone altered the data in Gregor Mendel's classic 1866 research report on the garden pea, "Experiments in Plant-Hybridization." Fisher concluded, reluctantly, that the statistical counts in Mendel's paper were doctored in order to create a better intuitive fit between Mendelian expected values and observed frequencies. That verdict remains the received view at least among statisticians, so I believe.

Fisher's analysis is a tour de force of so-called "Goodness of Fit" statistical tests using χ^2 to calculate significance levels, i.e., P-values. In this paper I attempt a defense of Mendel's report, based on several themes.

(1) Mendel's experiments include some important sequential design features that Fisher and (to my knowledge) others ignore. This is relevant to Fisher's charge that, for one crucial experiment, Mendel's protocol had approximately a 5% error of misclassifying hybrids as pure dominants.

(2) Fisher uses particular statistical techniques of Meta-analysis for pooling outcomes from different experiments. These methods are subject to critical debate.

and (3) I speculate on a small modification to Mendelian theory for the model of self-fertilization of garden pea that offers some relief from Fisher's harsh conclusion that, overall, Mendel's data are too good to be true. This last idea is yet another try to find an empirically plausible reason for thinking that Mendel's pea data should reflect sub-binomial variability.

P's in a Pod: Some recipes for cooking Mendel's data*

Teddy Seidenfeld

Philosophy & Statistics
Carnegie Mellon University

1. Introduction — The history, first anonymity and subsequent fame of Gregor Mendel's 1866 research report, "Experiments in Plant-Hybridization," is a familiar story.[1] The tale of neglect until 1900 and then great excitement upon the simultaneous rediscoveries by Carl Corens (in Germany), by Hugo DeVries (in Holland), and by Erich von Tschermak (in Austria), is Genetics folklore. Yet, in 1936 R.A.Fisher asked the question, "Has Mendel's Work Been Rediscovered?" What is the point of Fisher's query?

Of course, Fisher was not disputing the authenticity of the document which, in 1900, so stirred the biometrics community.[2] Concerning Fisher's own judgment, there can be no serious question about the importance he vested in Mendelian theory, and the respect he had for Mendel's genius. As a third year undergraduate at Cambridge, in 1911, Fisher saw the advantages in reconciling Biometricians (such as K.Pearson and W.Weldon) aligned with Darwinian theory, and the opposing Mendelians (such as W. Bateson). Both camps thought their theories were contrary arguing, e.g., that Mendelian laws are inconsistent with the "continuous evolution" of Darwinism.[3] Fisher's important [1918] paper gives the mathematical details (with due credit to

---

[1]. Mendel presented his work to the Natural History Society of Brünn (now Brno, Czechoslovakia) the year before, in 1865.

[2]. W.B.Provine [1971, chapter 3] gives an excellent account of the impact of the rediscovery on then ongoing disputes about the continuity of evolution. Particularly noteworthy were the 1902-1903 debates between W.F.R. Weldon and W.Bateson, whether Darwinian continuous evolution and the biometric blending of heritable traits were consistent with the Mendelian laws of dominance and segregation. (See, too, Karl Pearson's [1908] overview of Weldon's theory.)

[3]. In a preface to that undergraduate presentation Fisher concludes,
> "I have almost entirely devoted myself to the two lines of modern research which are of particular interest in Eugenics, that is to Biometrics and Mendelism; and perhaps experts and professionals will forgive the absence of more complicated details in both branches, if I explain that my object has been to give a fair view of

U.Yule) of how the two schools can be joined.[4] In his 1930 book, *The Genetical Theory of Natural Selection*, he cemented the bond. Last, in Fisher's [1955] introduction to (a reprinting of Bateson's 1909 version of an English translation of) Mendel's 1866 paper, he praises Mendel for keen methodological insights on central issues facing plant geneticists, without even so much as a hint of the sensational claims raised in his 1936 paper. (Did Fisher grow tired of the controversy he had stirred up twenty years earlier?) Fisher did not write "Has Mendel's Work Been Rediscovered?" either to question Mendel's integrity or to challenge his rightful place among those at the center of modern Genetics.

Fisher intended his provocative 1936 paper to chide the 20th century scientific and statistical researcher community for its inability to read with adequate scrutiny.

> "Mendel's contemporaries may be blamed for failing to recognize his discovery, perhaps through resting too great a confidence on comprehensive compilations. It is equally clear, however, that since 1900, in spite of the immense publicity it has received, his work has not often been examined with sufficient care to prevent its many extraordinary features being overlooked, and the opinions of its author being misrepresented. Each generation, perhaps found in Mendel's paper only what it expected to find: ... . Only a succession of publications, the progressive building up of a corpus of scientific work, and the continuous iteration of all new opinions seem sufficient to bring a new discovery into general recognition." (1936, p. 137)

The distortions of Mendel's opinions, as Fisher found them, were illustrated by, e.g., Bateson's attempts both to use Mendelian theory against the Darwinians and to represent Mendel as inclined

---

> the merits of the two methods, whose advocates have shown so little appreciation of the other school." [Bennett, 1983, p. 51]

[4]. The central theme in Fisher's 1918 work is to recover, e.g., a (normal) distribution of observed magnitudes and the observed correlations among related individuals using a large number of (independently) segregating Mendelian factors. Fisher's argument is a direct rebuttal to K.Pearson's 1909 Royal Society paper. In fact, Mendel anticipated the use of multiple factors in his speculation, derived from his abbreviated experiments, on seed and flower colors in *Phaselous* (Mendel, p.349). Ironically, K.Pearson's and R.Punnett's non-supportive referee reports of Fisher's 1916 submission to Biometrika (of his 1918 paper) focused on the question whether Mendelian theory could responsibly use more than 2 or 3 factors per observable trait (Norton and E.S.Pearson, 1976, pp, 153-155). Put in current terms, they questioned the empirical basis for evaluating Mendelian hypotheses with so many unmeasured parameters.

that way too.[5] But what are the "extraordinary" features of Mendel's publication that Fisher uncovers? The fourth section in Fisher's paper, titled "The Nature of Mendel's Discovery," begins with these words,

> "The reconstruction has been undertaken in order to test the plausibility of the view that Mendel's statements as to the course and procedure of his experimentation are to be taken as an entirely literal account, or whether, on the other hand, there is evidence that data have been assembled from various sources, or the same data rediscussed from different standpoints in different sections of his account. There can, I believe, now be no doubt whatever that his report is to be taken entirely literally, and that his experiments were carried out in just the way and much in the order that they are recounted. The detailed reconstruction of his programme on this assumption leads to no discrepancy whatever. A serious and almost inexplicable discrepancy has, however, appeared, in that in one series of results the numbers observed agree excellently with the two to one ratio, which Mendel himself expected, but differ significantly from what should have been expected had his theory been corrected to allow for the small size of his test progenies. To suppose that Mendel recognized this theoretical complication, and adjusted the frequencies supposedly observed to allow for it, would be to contravene the weight of the evidence supplied in detail by his paper as a whole. Although no explanation can be expected to be satisfactory, it remains a possibility among others that Mendel was deceived by some assistant who knew too well what was expected. This possibility is supported by independent evidence that the data of most, if not all, of the experiments have been falsified so as to agree closely with Mendel's expectations." (p. 132)

What is the basis for Fisher's startling proposal that "most, if not all of" Mendel's data were cooked?[6] "Has Mendel's Work Been Rediscovered?" is a tour de force of statistical analysis using

---

[5]. See Bateson's remark, from his 1909 biographical notice on Mendel, quoted by Fisher (1936, p.117).

[6]. Fisher's concern, even suspicion, about the close fit between Mendelian theory and Mendel's data dates, at least, from his 1911 (undergraduate!) presentation to the Cambridge University Eugenics Society. There, at age 21, he writes (without reference),

> "It is interesting that Mendel's original results all fell within the limits of probable error; if his experiments were repeated the odds against getting such good results is about 16 to one. It may have been just luck; or it may be that the worthy German abbot, in his ignorance of probable error, unconsciously placed doubtful plants on the side which favored his hypothesis." (Bennett, 1983, p.57)

However it was Weldon (1901, p.233), not Fisher, who first published a detailed analysis of the fit of Mendel's data to Mendelian theory, including the 16:1 probable error odds that such a good fit would not be duplicated. Regarding the data for the 3:1 law, he writes,

> "These results then accord so remarkably with Mendel's summary of them that if they were repeated a second time, under similar conditions and on a similar scale, the chance that the agreement between observation and hypothesis would be worse than that actually observed is about 16 to 1."

χ^2 Goodness-of-Fit tests. According to Fisher, the two claims pertaining to "sophistication" of data are these:

(**i**) Mendel conducted tests to separate hybrid from pure-bred plants. Concerning five plant traits (not to be confused with seed traits), plants were to be classified as either pure-dominant or hybrid in a single trait for which they showed the dominant feature. Mendel tested each such plant by looking at 10 of its offspring to see whether the recessive trait was displayed in any. A plant was declared to be pure-bred in case all its offspring showed the dominant trait. This procedure yields about a 5.6% ($\approx .75^{10}$) misclassification of hybrids as pure-breds, against which Mendel's

---

What is, I find, much more remarkable than the young Fisher's loose scholarship in a 1911 unpublished undergraduate paper, is his total neglect of Weldon's analysis, 25 years later, in 1936. For, not only did Weldon scrutinize Mendel's data by analyzing how many times results exceeded their probable error, thereby providing the 16:1 odds quoted above. Also, Weldon used the then new method of χ^2 Goodness-of-Fit tests to analyze some of Mendel's more complicated experiments. (Parenthetically, Pearson published his famous paper introducing χ^2 in 1900, but Elderton's tables appeared only in 1901, in the same inaugural issue of Biometrika as Weldon's paper. Thus, Weldon's may well be the first application of χ^2 as a "Goodness of Fit" test; ironically, used to argue that the fit is suspiciously too good.)

Regarding the 3-factor experiments, for testing the simultaneous distribution of three factor inheritance (with 3 categories - two for homozygotes and one for hybrids) Weldon (1901, p. 235) notes that, using Pearson's χ^2,

> "Applying the method of Pearson (No. 25) the chance that a system will exhibit deviations as great as or greater than these from the result indicated by Mendel's hypothesis is about 0.95 (see Elderton, this Journal, *ante*, p. 161), or if the experiment were repeated a hundred times, we should expect to get a worse result about 95 times, or the odds against a result as good as this or better are 20:1.

Fisher's (1936, p. 131) Table V χ^2 analysis of Mendel's trifactorial experiment on 26 ($= 3^3 - 1$) df leads to the same conclusion. Thus, Weldon anticipated Fisher not only in his finding that Mendel's counts showed surprisingly little departure from their expected values, but also even in his early application of small values of χ^2 for analyzing the extreme goodness of this fit.

Weldon follows his remark, above, with a repetition of the well known fact (and one reported by Mendel himself) that the Mendelian laws fail to apply to all the observable traits in plants, even for peas. Then Weldon concludes his critical discussion of the scope of Mendel's laws by saying,

> "In trying to summarize the evidence on which my opinion rests, I have no wish to belittle the importance of Mendel's achievements. I wish simply to call attention to a series of facts which seem to me to suggest fruitful lines of enquiry."

The reader is urged to contrast this with Fisher's parallel conclusion in 1936 (p. 137), quoted in the main text, above. Just how much of Weldon's 1901 essay did Fisher unconsciously internalize?

reported observed frequencies are statistically significant. Moreover, in these data, the departures of the observed frequencies in Mendel's data from their "corrected" expectations are all (suspiciously) in the direction of the "uncorrected" Mendelian theory, where a 2:1 hybrid-to-dominant ratio is expected.

(**ii**) Apart from the misclassification of hybrids, a thorough examination of the entirety of Mendel's published results show an exceptional agreement between observed and expected frequencies which cannot be attributed to luck. The P-value from the composite χ^2 is in excess of .999. In short, Mendel's data are too good to be true! (The χ^2 analysis for these two assertions are conveniently summarized by Fisher's Tables III and V, reproduced here in the Appendix.)

Since 1936, there have been numerous attempts at explaining away the anomalous features in Mendel's evidence that were spotlighted so intensely by Fisher. The well-known statistician A.W.F Edwards (1986) provides a thorough review of much of the literature. He gives careful reconsideration to Fisher's analysis. Edwards concludes, after adding his own analysis of the distribution of the individual (1 df) χ^2 values, in agreement with Fisher's two central findings (above), that outcomes discrepant with Mendelian expectations have been trimmed. Even current scientifically informed popular literature accepts Fisher's judgment on the point.[7] On the other side of the debate, the botanist F.Weiling (1989) continues his rebuttal (spanning 25 years) to Fisher's criticisms, to the effect that Mendel's data are not cooked. And research botanists continue to cite Mendel's data in prominent publications without hesitation, e.g., Gauch (1993).

My purpose here is several fold. There are statistical questions to be asked about the logic of "too good to be true" analysis. I want, also, to re-examine some of Fisher's criticisms — to ascertain

[7] . For example, even S.Mawer (1997, pp. 237-239) concedes that "Mendel cheated." I am not sympathetic, however, with Mawer's excuses for Mendel that both Burt and Lysenko also cheated with their data but, contrary to Mendel, created false (and dangerous) theories, nor that Darwin was simply wrong in his speculations about heritability.

whether a useful reply to the charge that the data were "sophisticated" can be founded on scientific (genetic) and statistical considerations. After all, it is in the spirit of Fisher's 1936 charge to the reader to reexamine Mendel's original work. As he wrote about it a generation later, "A first-hand study is always instructive, and often, as in this case, full of surprises." (1955, p.6). I will be satisfied if this essay prompts some to enjoy the pleasure of re-reading both Mendel and Fisher.

2. A brief outline of Mendel's paper. Before discussing Fisher's objections, allow me to sketch the experimental plan of the peas studies. I follow Mendel's account of his attempt to ascertain laws for hybrids.[8]

Mendel's paper is arranged in 11 unnumbered sections. The different trials discussed in these sections are not dated, though some temporal sequences are explicit. In general, it seems that not all trials reported in given section of the paper were conducted in the same year. For example, even when the same Mendelian law is studied for each of the seven pea characteristics, the reader should not assume these were simultaneous trials.[9]

Foremost in the selection of experimental plants were, according to Mendel, two general concerns. As stated in section 2 of his paper:

(1) The plants should possess constant differentiating forms.

(2) There should be little risk of accidental fertilization by foreign pollen.

The second point is obvious. A reason Mendel chose the garden pea for study, then, is because

> "a disturbance through foreign pollen cannot easily occur, since the fertilizing organs are closely packed inside the keel and the anther bursts within the bud, so that the stigma becomes covered with pollen even before the flower opens." (p.319)

---

[8]. Throughout, I shall refer to the translation appearing in Bateson's 1909 volume. This gives the same pagination used by Fisher in his 1936 paper.

[9]. Fisher (1936, Table VI) proposes a chronology of the experiments. Based on a subsequent review of the historical record, he revised Table VI by reducing by one each year indicated. (See Bennett, 1965, p. 59.)

and "Among more than 10,000 plants which were carefully examined there were only a very few cases where an indubitable false impregnation had occurred." (p.324)

Regarding the first point, Mendel notes as well that,

> "In order to discover the relations in which the hybrid forms stand toward each other and also toward their progenitors it appears to be necessary that all members of the series developed in each successive generation be, *without exception*, subjected to observation." (p.319)

The significance of this methodological claim is hard to overemphasize, as we shall shortly see!

Beginning with some 34 varieties of peas (*Pisum sativum*), after two-years' trial, Mendel selected 22 for cultivation throughout the years of his experiments, which ran approximately from 1856 through 1864. According to the presentation in the third section of his paper, he used the following seven pea-plant characteristics for intensive study: (The dominant form appears first in each pair.)

1. whether ripe seeds are round or wrinkled;
2. whether the ripe seed cotyledon is yellow-orange or green;
3. whether the seed-coat is grey-brown or white — associated with corresponding blossom colors (violet-red or white);
4. whether the ripe pod form is simply inflated or deeply constricted between the seeds;
5. whether the unripe pods are green or bright yellow — with corresponding coloration of the leaf stems;
6. whether flowers distribute along the stem or are terminally bunched at the top;

and 7. whether main stem is (about 6 feet) tall or (about 1 foot) short.

These seven pairs have the feature that the dominant form is almost entirely so in hybrids, as opposed to other characteristics noted by Mendel, e.g., where the hybrid flowering times stand midway between the seed and pollen parents flowering times (p. 337).

Of the seven traits, the first two belong to the seed, i.e. are of the next generation, whereas features #3-#7 belong to the (maternal) seed-plant. This division is important for understanding Mendel's

experimental design, since his garden averaged about 30 seeds/plant. Thus, for seed shape and seed color, the first two traits, Mendel had sample sizes many times larger than what he could create for the remaining five characteristics, and he had those data in-hand one growing season sooner! It hardly needs saying that, therefore, Mendel had well grounded expectations for his experiments on single (and even double) factor trials involving the 5 plant characteristics, since he had seen the parallel results (at much large sample sizes) for the 2 pea characteristics a year earlier.

The pea experiments focused on three different aspects of what, today, is called Mendelian theory:

(**a**) First, Mendel discusses experiments designed to show laws for the heritability of single-factor traits through successive generations bred from hybrids. Each of the seven characteristics (above) was separately studied through several generations. The results appear in Mendel's sections 4-7 (pp. 324-332). The grand conclusion of these trials is Mendel's "1:2:1" law for hybrids: their offspring are produced (independently) with probabilities of 1/4, 1/2, and 1/4 (respectively) for recessive, hybrid, and pure-dominant form.

(**b**) Section 8 (pp. 332-338) addresses two- and three-factor distributions. The two-factor trials examine laws for the joint heritability of the two seed-characteristics (shape and color). The three-factor trials looked at the simultaneous distribution of seed shape, seed color, and seed-coat color. These experiments led to the conclusion that the traits are distributed independently, each following the "1:2:1" law.[10]

(**c**) Last among the trials dealing with peas, Mendel's ninth section offers evidence that the genetics for the offspring of hybrids do not depend upon which gamete originates with the pollen and which with the seed parent plant. The laws for offspring are the same when the genetic role of

---

[10]. There are seven chromosomes for the garden pea. S.Blixt (1974, pp. 187-188) indicates that the 7 characteristics studied by Mendel are governed by alleles located on 4 different chromosomes. Using Mendel's ordering, as reported above, we find these 7 characteristics are associated (respectively) with chromosomes #7, #1, #1, #4, #5, #4, and #4. The reason the three-factor trial supports "independent" heritability, despite the coincidence of seed color and seed coat genes on chromosome #1, is because these alleles are at remote sites. Likewise for the comparison of flower position with either pod shape or plant height. According to Lamprecht (1968a), as reported by Blixt, only pod shape and plant height might show linkage in trials the size Mendel had; but it is not known that Mendel investigated joint frequencies for this pair.

pollen and seed are exchanged. Thus, the ninth section justifies the composite data reported in sections 3-8, where Mendel intentionally grouped data from offspring of hybrids, regardless whether the hybrid was the result of fertilizing a "recessive" seed-plant with "dominant" pollen or the other way around. (See, especially, his remarks at the top of p. 325.)

Separately, section 10 of "Experiments in Plant Hybridization" notes some preliminary research on other species, and includes the fertile speculation that some observed plant features are the consequence of numerous "Mendelian" factors (p. 349). Last, there is a very carefully worded concluding section 11, in which Mendel contrasts his findings with some (then) recent work of others investigating hybrids.

3. On the supposed error of misclassification in Mendel's test of the "2:1" law.

The first of Fisher's two objections rests on the penetrating analysis that using 10 offspring to separate pure-bred dominant plants from hybrids leads to a misclassification of about 5.6% of the hybrids. That is, the probability that all 10 offspring of a hybrid will show the dominant trait is $.75^{10}$, or slightly more than .056. Of course, all the pure breds are correctly classified this way. Thus, if the ratio of hybrid to pure-dominant plants in a given collection is 2:1, Mendel's procedure has an expectation of separating them in the ratio 1.8874 : 1.1126, rather than 2 : 1. Where does this affect Mendel's results? According to Fisher, the problem occurs in two settings.

The first occurrence of the misclassification problem is in section 6 (pp. 329-331), where Mendel conducts trials to complete his single factor studies of the "1:2:1" law. For each of the seven characteristics, hybrid plants (the "F_1" generation) were created by artificial fertilization of two pure-breeds (as described in section 3), a cross of a pure-dominant with a recessive. The F_1 hybrids produced offspring (the "F_2" generation) by self-fertilization. These F_2 children of the F_1 hybrids displayed the "3:1" law (of dominant to recessive observed types), as described in section 4 of Mendel's paper. The further division of the "3"-group into a "2:1" ratio of hybrids to pure-

dominant plants is the locus of Fisher's concern about misclassification.

χ^2s for Mendel's summary of these data are given in a Table appearing on page 36. Regarding the two seed characteristics, form and color (experiments #1 and #2), there is no issue of misclassification since we may assume that each F_2 plant produced about 30 seeds. Hence, for these two trials, with a total of about 1000 F_2 plants to be classified, the (expected) error is less than one plant. That is, the probability of mistaking a hybrid for a pure-bred ($p = .75^{30} < 2 \times 10^{-4}$) is negligible.

Concerning each of the remaining five plant characteristics, Mendel selected 100 F_2 plants to be sorted and cultivated 10 seeds from each.[11] Presumably, an F_2 plant was declared pure-dominant (rather than hybrid) if all its F_3 offspring showed the dominant trait for the characteristic tested. One plant feature (pod color) was tested twice. Thus, there was a total of 600 plants to be sorted between the two kinds: pure-dominant and hybrid (with a total of about 6000 offspring to observe).[12] Of the 600 F_2 plants, 399 were classified as hybrid and 201 were classified as pure-dominant. The χ^2 (1 degree of freedom) for these observed frequencies against the Mendelian "2:1" law (with expectations of 400 hybrid and 200 dominant) is 7.5×10^{-3}; or a P-value of about .93. However, if we follow Fisher's lead and alter the expectations to reflect the 5.6% rate of

---

[11]. There is some discussion in the literature whether Mendel *planted* 10 seeds, or *cultivated* 10 plants. I will here not cite the various authors who take opposite sides on this matter. The point, however, is rather simple. If only 10 seeds were planted, of which about 8-9 would be expected to grow, then the misclassification problem is worsened. If Mendel planted more than 10 seeds, in order to be confident of having at least 10 mature plants to observe, then the misclassification problem is diminished but also his writing is misleading for not indicating how many test plants were observed. Fisher considers the latter alternative (p. 129) and dismisses it.

[12]. Experiment #5a (for pod color) yielded a 60:40 division of 100 plants. Mendel retested this trait with another batch of 100 (called experiment #5b in the Appendix). The retest showed a 65:35 division. If the retest was in a subsequent year, as Fisher suggests (1936, p. 125), where did the second hundred plants come from? In this special case, did Mendel manage to grow a second crop in the same year? (The pea-plant requires only about 90 days to reach maturity.) Or, as I suspect, did he simply combine the results of two year's experimentation on the plant characteristics in this one grouping of his data? In that case, it further endorses my speculation (below) that Mendel reported on this part of his study using the benefit of hindsight, after all the (subsequent years') data were in.

misclassifying hybrids, the "corrected" χ^2 (1 df) is about 3.3, or a P-value of about .07. That is, with Fisher's correction, the probability of an F_2 plant being classified hybrid is only about .629 rather than .667, yet the data show a frequency of .668.

A similar problem arises in the trifactorial study where plants were to be distinguished as hybrid or pure-dominant for the plant-characteristic of seed-coat color.[13] The "uncorrected" and "corrected" expectations are given on the second row of Fisher's Table III. Combining the two sets of outcomes from the "2:1" and trifactorial experiments, Fisher notes that the "corrected" χ^2 (1 df) is about 8.05; or a P-value of about 4.5×10^{-3}. (The "uncorrected" χ^2 (1 df) is about .093; or a P-value of about .76.) If we calculate the probability of a departure from "corrected" expectations in the direction actually observed (towards the "2:1" law), the P-value is halved. As Fisher writes,

> "A total deviation of the magnitude observed, and in the right direction, is only to be expected once in 444 trials; there is therefore here a serious discrepancy." (p.129)

What are the possible replies to this analysis. Fisher examines several.

(1) Might it be that the test plants were not a random sample? In selecting the 600 (= 6×100) F_2 plants from a population of about 4600, was Mendel disposed to favor hybrids?[14] For example, in the "2:1" monofactorial study, there were more than 900 F_2 plants to choose from in experiment #3 (on seed-coat color), almost 1200 F_2 plants were available from experiment #4, 580 F_2 plants available for experiment #5, about 850 in experiment #6, and over 1050 in experiment #7. A selection bias in favor of hybrids might have increased the proportion of hybrids among the 100 F_2 test-plants chosen for each experiment.

Unfortunately, this rebuttal will not do. Fisher indicates three reasons why, two of which are cogent:

---

[13]. Recall, the misclassification issue does not affect the sorting of plants with regard the two seed-traits, for which each test plant yields perhaps 30 seeds, as explained in the main text. The bifactorial trials, therefore, were not involved, nor was there a problem for two of the three traits in the trifactorial study.

[14]. In a footnote, p.122, Fisher quotes the biologist, J.Rasmussen, that in picking from batches of whole, dry plants, a selection bias in favor of hybrids is quite plausible.

(i) It does not apply to the trifactorial study, where all plants were classified.

(ii) It is implausible that the bias was equally effective for all five characteristics—at best, selection was the result of some gross plant features typical of hybrids, e.g., plant size.

(iii) And (with circular reasoning), if the data were altered, the observed difference between "uncorrected" and "corrected" expectations agrees with the expected number of misclassifications using precisely 10, rather than 9 or 11 F_3 children to sort the F_2 parents.[15]

Like Fisher, I find the coincidences of perfectly offsetting selection biases (the second rebuttal point) more difficult to believe even than the alternative that the data were "cooked"!

(2) Fisher considers, also, the proposal that Mendel might have grown, say, 15 progeny (rather than 10) per F_2 test plant. This would eliminate the bulk of the misclassification, but it involves a denial of Mendel's own account of his method for the monofactorial study.

(3) Last, since Mendel was aware that low probability events might not appear in limited samples (see, e.g., his remarks on p. 359, concerning an outcome with 1/8 chance that might fail to be observed in a small sample), why didn't he create the 10 F_3 plants from a backcross; crossing each test F_2 plant against a plant recessive for the trait in question? That would have reduced the misclassification rate for hybrids to less than 2×10^{-4}. The answer is that, unfortunately, the effort needed for 600 artificial fertilizations makes the design infeasible.[16]

---

[15] . Fisher's third rebuttal point is without merit that I can see. The "adjustment" had to agree with the "correction" based on 10, rather than some other number of F_3 test plants. That is because Fisher calculated the "correction" (of 5.6% unclassified hybrids) based on Mendel's announced number (10) of test plants. Whatever number N of test plants Mendel had announced, that would have served as the basis for the "correction." So, of course there is no surprise that the "adjustment" Fisher calls for agrees with N = 10.

[16] . Mendel writes, p. 319, "Artificial fertilization is certainly a somewhat tedious process, ..." I estimate that all of Mendel's (reported) data on peas required fewer than 500 artificial fertilizations.

What is left to say in Mendel's defense?[17] I propose the following. Mendel did not base his classification of F_2 plants solely on the observations of 10 F_3 children. He carried his experiments to subsequent generations bred from the same hybrids. For example, as Mendel's data support, suppose about 90% of seeds germinate. Assume 10 F_3-seeds per F_2-test-plant were sown. (This is a conservative interpretation of Mendel words in section 6, p.330.[18]) Consider the design that whenever all (that is, on average about 9) F_3 offspring showed the dominant trait then, in the following year, an additional 3 F_4 plants were grown from the seeds of each of the phenotypic dominant F_3 plants. In this sequential procedure, the probability of misclassifying an F_2 hybrid is *less* than 2×10^{-3}. If only 2 F_4 plants were grown, the error in classifying an F_2 hybrid is small, about 5×10^{-3}. Even with only one F_4 observation per F_3 plant, if all 10 F_3 plants were successfully cultivated, the probability of misclassifying an F_2 hybrid still is *less* than .01. Thus, one way around Fisher's first objection is to hypothesize that Mendel used an elementary

---

[17] . F.Weiling (1971, p.76) suggests that only 8 of 10 seeds might germinate. I do not see how this explains away the misclassification problem. If only 8 F_3 plants grow, then the resulting misclassification worsens. In another paper (1989, p. 136), in response to Edwards (1986), Weiling appeals a hypergeometric distribution — which has lower variance than the binomial. After conditioning on some expected values (instead of calculating a distribution), he manages to reduce the misclassification rate to about .038. (Recall, Fisher gave a misclassification rate for hybrids of .056).

Among several difficulties I have with Weiling's statistics, I do not understand the basis for his use of the hypergeometric distribution. It is true, as he writes, that the process of choosing 10 of 30 particular seeds from a plant (as Mendel is posited to have done to make the 10 F_3 offspring per F_2-parent) follows a hypergeometric distribution, with smaller variance than the i.i.d. Binomial distribution. However, under Mendelian theory, these 30 seeds follow the i.i.d. Binomial distribution. Hence, the net (marginal) distribution for the 10 seeds, chosen from the 30, is again i.i.d. Binomial, not hypergeometric, contrary to what Weiling asserts. The challenge, taken up below, is to justify the claim that the 30 seeds are not an i.i.d. sample from the Binomial distribution.

[18]. It is interesting to note that Mendel uses the same word for all the test plants used in the 5 experiments for classifying the F_2 generation. Regarding the 100 F_2 in experiment #3, for example, he writes, "The offspring of 36 plants yielded exclusively grey-brown seed-coats, while of the offspring of 64 plants, some had grey-brown and some had white."

The German word for 'offspring' Mendel uses here is 'Nachkommen' which (according to my colleague, Wilfried Sieg) is exactly the right term to refer to future generations as well the immediate progeny. However, Mendel uses 'Nachkommen' also at the beginning of the same section (§6) in a context where it is evident he intends to refer to the next (F_3) generation only. Thus, I do not see how to settle the issue, whether Mendel employed a sequential design, merely by this choice of word in describing the protocol.

sequential design.

What is the basis for this speculation? In section 7, titled "The subsequent generations [bred] from the hybrids," coming only one page after describing the data for the "2:1" law, Mendel begins,

> "The proportions in which the descendants of the hybrids develop and split up in the first and second generations presumably hold good for all subsequent progeny. Experiments 1 and 2 have already been carried through six generations, 3 and 7 through five, and 4, 5, and 6 through four, these experiments being continued from the third generation with a small number of plants, and no departure from the rule has been perceptible. The offspring of the hybrids separated in each generation in ratio of 2:1:1 into hybrids and constant forms." (p.331)

Unfortunately, it is unclear what Mendel means here by a "small number of plants." Might it be that he carried on in the F_4 generation only with a few of the F_3 descendants of the (then) already confirmed F_2-hybrids? Then Fisher's objection is left intact. Or, did Mendel mean that he carried on with some of both kinds, hybrids and pure-breds? Then there is no misclassification bias to correct for.

In any event, misclassifications of the F_2 plants can easily be avoided using as few as 10 F_4 observations whenever all (10) F_3 children of a test plant show the dominant character. That is, by growing ten plants in the F_4 generation, one per suspect F_3 plant, the misclassification rate for F_2 hybrids drops from .056 (reported by Fisher) to less than .01. Thus, according to Fisher's correction to Mendel's design — using 10 F_3 plants per F_2 test plant, or 1000 F_3 plants per experiment — among each 100 F_2's (with probability more than .975) fewer than 47 F_2 plants/experiment would remain unclassified after examining the F_3 generation. That would entail an addition (to whatever else Mendel planned) of no more than 470 plants/experiment in the F_4 generation, or about half the space allocated for the F_3 plants. This sequential design was feasible given the size of Mendel's garden.[19] Mendel is explicit in his paper that he persisted several

[19] . Fisher (p.129, footnote) attributes to Rasmussen an estimate of space for 4000-5000 plants.

generations with these experiments. So, a sequential plan to eliminate misclassification of the F_2 hybrids seems to me to be a better reply than Fisher's alternative involving wholesale "sophistication" of the data.

There is another issue which affects Fisher's concern about misclassification, having to do with the assumption that each grouping of 10 F_3 plants from a hybrid F_2 parent constitute an *independent* sample of 10 (with binomial parameter .75) of displaying the phenotypic dominant trait. This question surfaces, also, in connection with Fisher's assertion that Mendel's data, generally, are too good to be true — to which I turn next.

4. Data too good to be true.

4.1 General remarks. It is a daunting task to address all aspects of Fisher's second challenge — the charge that the balance of Mendel's data conform too well to ("uncorrected") Mendelian expectations. One difficulty I find in responding to Fisher's argument, in-kind, is the absence of an adequate theory of (Fisherian) significance testing. Particularly troubling is the absence of Fisherian theory regulating the alternatives to the null hypothesis. Nonetheless, the central idea behind Fisher's second objection is simple enough to state.

If, for example, one were to flip a supposedly "fair" coin 500 times and observe 250 heads, the coincidence of getting exactly the expected numbers of heads and tails might be noted and then attributed to "luck." However, if in repeated experiments the observed relative frequency of heads distributed about .5 with *much* smaller variance than expected in binomial sampling, that would *seem to be* a good reason to reject the (i.i.d.) Binomial model.

Fisher uses χ^2 tests to quantify this informal reasoning. But even the naive reasoning in the simple case has obvious difficulties. For example, if the data are n flips of the supposedly "fair" coin, how many different "experiments" do they comprise? Does each partition of n constitute a

sequence of "experiments" for assessing the Binomial variance?[20]

One point of view allows arbitrary redescriptions of a data set as part of a freewheeling exploration for "significant" outcomes. H.Cramér (1946, §30.2) constructs three tests of seed-data from ten plants that Mendel reports in an illustration of sample variance in his Experiment #2. (Recall, in experiment #2, 258 plant hybrids self-fertilize. The proportions of yellow to green seeds they produce reflect the "3:1" distribution of dominant to recessive phenotypes in the next generation.) Cramér tests the Mendelian "3:1" hypothesis by constructing an array of three χ^2's as follows:

(1) From the ten plants, combined, 355 of 478 seeds were yellow. On the "null" hypothesis of "3:1," that is a χ^2 (1 df) of .137, and a corresponding P-value of about .78.

(2) Take each plant as providing an autonomous experiment, as captured in Cramér's Table 30.21. The sum of the ten χ^2's (each on 1 df) is a χ^2 (10 df) of 7.191, with a corresponding P-value of about .71.

(3) Also, as a test whether the 10 χ^2-values are distributed as i.i.d. χ^2 (1 df) variates, Cramér forms a trinomial partition of χ^2-values, with cell-probabilities of .3, .4, and .3 given the null hypothesis. The 10 χ^2-values of Table 30.2.1 yield cell counts of 2, 6, and 2, respectively. Cramér states the agreement of these data with the trinomial distribution "must be regarded as good."

Given Fisher's (1936) analysis, what I find surprising here is not Cramér's support for Mendel's conclusions, but Cramér's methodology.[21] He writes,

---

20. Recall, the number of ways of partitioning n elements into m non-empty subsets, $\mathbf{S}^{(m)}_n$), is given by Sterling numbers of the second kind:

$$\mathbf{S}^{(m)}_n = \frac{1}{m!}\sum_{k=0}^{m}(-1)^{m-k}\binom{m}{k}k^n \quad .$$

Thus, there are $\sum_{m=1}^{n}\mathbf{S}^{(m)}_n$ alternative representations of the data as a sequence of "experiments."

21. The parallel three tests relating to the 10 plants Mendel selected from Experiment #1, concerning the "3:1" ratio of round to wrinkled seeds, lead to the same conclusion: (1) Ten plants yielded 437 seeds of which 336 showed the dominant trait, for a χ^2 (1 df) of .831, and a corresponding P-value of about .36. (2) The sum of the ten, 1-df χ^2 is 5.299 with a

> Thus all our tests imply that the data of Table 30.2.1 are consistent with the 3:1 hypothesis. If either test had disclosed a significant deviation, we should have had to reject the hypothesis, at least until further experience had made it plausible that the deviation was due to random fluctuation. (p. 423)

However, repartitioning the evidence to hunt with various χ^2 tests for significant P-values makes little sense without critical control over the space of alternatives to the null hypothesis. All that a P-value provides in this context is an index of rarity, one measure of the data's so-called "discrepancy" with the null hypothesis. But, each outcome of an experiment can be made "rare" when suitably redescribed using one of many (non-equivalent) test-statistics.

Here is a completely trivial example: Index the discrepancy of an outcome inversely to its probability of occurrence under the hypothesis.[22] On n independent flips of a fair coin, the exact sequence of heads and tails observed has a probability of 2^{-n}. Unless the same discrepancy index (the improbability of the sequence) establishes that data are *not* "rare" under a serious rival to the null hypothesis, what is the point of the small P-value for sequence observed, and the large P value (of 1) for the set of outcomes at least as discrepant as the one observed?

Consider, for example, Fisher's use of χ^2-tests to evaluate Mendel's data about the "2:1" law — the ratio of heterozygous to homozygous dominant offspring among the "3-group" in the "3:1" law, discussed in the previous section. Fisher's Table III (p.128) reports the combined data, pooled from six (of the eight) experiments Mendel conducted on this subject. In these six experiments (of 100 trials each) 399 out of a total of 600 plants had the (relevant) dominant trait ($\chi^2 = .0075$, P = .93 on 1 df). And in Table V (p. 131), Fisher uses the sum of these six 1 df

---

corresponding P-value of .87. (3) Last, regarding Cramér's test that these ten values are trinomial with cell probabilities .3, .4, and .3, the respective cell-counts are 3, 6, and 1 for a χ^2 (2-df) of 2.33 and P-value of .33. None of these three is "significant." Nor do these analyses change when the combined (20) data are similarly analyzed in three tests.

[22] Fisher often refers to such a discrepancy index as an "exact" test, e.g. Fisher, 1925, §21.02. Of course, he does so when the outcomes are suitably described, e.g., through a minimal sufficient statistic with respect to an unstated, larger model that contains the null as a special case. For example, cell frequencies are the right test statistic for an exact test of independence in contingency tables, where Multinomial sampling serving as the unstated background model.

χ^2's ($\Sigma\chi^2$ = 4.58, P ≈ .60, on 6 df) and the sum of the remaining two 1 df χ^2's ($\Sigma\chi^2$ = .5983, P ≈ .74, on 2 df) as part of an analysis of all of Mendel's data. He offers the P-value only for the sum of these eight 1 df χ^2's ($\Sigma\chi^2$ = 5.1733, P = .74, on 8 df). In his Table V grouping of Mendel's data, by summing χ^2 ($\Sigma\chi^2$ = 41.6056, 84 df), Fisher arrives at a truly exceptional P-value, P > .9999, data clustering too closely about their expected Mendelian values to be believed!

However, in the spirit of Cramér's analysis, using iterations by examining distributions of distributions of χ^2 or transformations of P-values with χ^2 having other degrees of freedom[23], there exist unlimited varieties of possible meta-analyses of Mendel's data. We are threatened making P-soup! The different meta-analyses offer very different perspectives on the data, as we see next.

The two figures in the Appendix, Figures N1 and N2 (Nobile, 1992) graph the asymptotic joint distribution ($n \to \infty$), under a Binomial ($\theta = 1/3$) null hypothesis of P-values for χ^2's with six samples of size 100n, calculated either by:

(1) summing six independent (1 df) χ^2's (P_U), to yield a χ^2 (6 df)

or by (2) pooling the data first to create a single (1 df) χ^2 (P_Z).

This corresponds to the case, discussed before, of Mendel's six experiments for plant characteristic data in the "2:1" law. Observe, especially, how much of the joint distribution is off the diagonal line, where $P_U = P_Z$, especially near the large ("too good to be true") P-values.[24]

Evidently, the unstated alternative model to the null hypothesis that is used with a traditional test of the simple null hypothesis (i.e., reject $\theta = 1/3$ when the P-value is low) is not the same in the one

---

[23]. See, for example, Fisher's advice in §21.1 of (1925) where, for example, he recommends meta-analysis by converting P-values back into χ^2 values on 2 df before summing. This rule he gives because -2ln(P) is the corresponding χ^2 values on 2 df. Supplied with a table of logarithms, anyone can do meta-analysis!

[24]. Nobile (1992, Table 5) recomputes an analogue of Fisher's Table V for the whole of Mendel's data, aggregating the results of Mendel's six experiments wherever these appear by summing the corresponding χ^2 statistics as in P_Z. The effect is a χ^2 of 19.0 on 40 df, with a P-value of .998.

kind of meta-analysis as in the other. For example, the simple binomial model (varying θ) works fine with the "pooled" 1-df χ^2 test, whereas the sum of χ^2-test (6 df.) is made sense of by considering a background model in which each of the six experiments might have its own separate binomial value. That is, in this case the alternative models are nested.

In order to respond to Fisher's argument, that the data are too good, we have to provide a family of alternative hypotheses, rivals to the "null" hypothesis that Mendel's data follow the Mendelian laws, to which we may turn when the "null" hypothesis is deemed to be overly discrepant with the observed data. Also, we are obliged to make precise the standards by which "discrepancy" is judged, to insure that there exist some alternatives which are not discrepant with the data whenever the "null" hypothesis is. The cogency of Fisher's argument rests, then, on the plausibility of these tacit alternative hypotheses, so introduced.

In the informal example above, where a sequence of observed relative frequencies agrees "too well" with Binomial expectations for independent experiments, consider a rival statistical model:

(**a**) with the same first moment as in the Binomial "null" distribution,

but (**b**) with a smaller second moment.

Then, whenever the "null" hypothesis is suspect because the data are "too good to be true," such an alternative hypothesis fits the first two moments of the sequence of relative frequencies better than the "null" does.

The χ^2 distribution (on k degrees of freedom) is the distribution for the sum

$$\chi^2 = \sum_{i=1}^{k} x_i^2$$

where the x_i are i.i.d. standardized, normal $N(0,1)$ variates. More generally, if the x_i are i.i.d. normal $N(\mu,\sigma^2)$ variates, then

$$\sum_{i=1}^{k} \frac{(x_i-\mu)^2}{\sigma^2} \qquad (*)$$

is distributed as χ^2 on k degrees of freedom. Suppose the "null" hypothesis h_0 is that binary variables, $y_i \in \{cell_1, cell_2\}$, (i = 1, ..., n) form an i.i.d. Binomial sequence, with parameter θ_0. Then the cumulative $cell_1$ count, n_1, is approximately normally distributed $\boldsymbol{N}[n\theta_0, n(1-\theta_0)\theta_0]$. Let n_2 be the cumulative $cell_2$ count; so $n_1 + n_2 = n$. Hence, the familiar chi-square Goodness of Fit test-statistic,

$$\chi^2 = \frac{(n_1 - n\theta_0)^2}{n\theta_0} + \frac{(n_2 - n(1-\theta_0))^2}{n(1-\theta_0)}$$

$$= \frac{(n_1 - n\theta_0)^2}{n(1-\theta_0)\theta_0}$$

is distributed, approximately, as χ^2 (on 1 df), in accord with (*).

If, as Fisher alleges, exceptionally small values of χ^2 tell against the null-hypothesis just as exceptionally large values do, then we may ask which alternatives to h_0 have reduced discrepancy with the data, when discrepancy is indicated by such extreme (small or large) χ^2's. To repeat, unless there is some alternative hypothesis which is not discrepant, the so-called "discrepancy" cannot be the basis for discrediting just the "null."

With k-cells ($k \geq 2$), for large values of the χ^2 test-statistic, shifting to an alternative hypothesis within the Multinomial family, e.g., using the m.l.e. or the minimum χ^2 estimator, reduces the χ^2 test-statistic from the unacceptably large magnitude. As is very well known, provided one of the Multinomial hypothesis is correct, asymptotically (with increasing sample size), such estimators are consistent and even efficient. Hence, using the magnitude of χ^2 as the index of discrepancy, the usual test, satisfies the condition that some alternative hypothesis exists with low discrepancy.

However, Fisher asserts quite generally that (very) small values of the χ^2 test-statistic also indicate discrepancy with an hypothesis.

The term Goodness of Fit has caused some to fall into the fallacy

> of believing that the higher the value of P the more satisfactory is the hypothesis verified. Values over .999 have sometimes been reported which, if the hypothesis were true, would only occur once in a thousand trials. ... In these cases the hypothesis considered is as definitely disproved as if P had been .001. (1925, §20, pp. 80-81)

Then we face a dilemma, at least with 1 df tests. If hypotheses with small values of χ^2 (1 df) are discrepant too, the Binomial minimum χ^2 or the m.l.e. will *not* survive as alternatives to a Binomial "null" hypothesis when large values of χ^2 are observed. An instance occurs in Fisher's analysis of Mendel's data for the "2:1" law; the data discussed in the previous section.

Regarding the 600 F_2 plants which Mendel sorted as either pure-bred dominants or hybrids, his counts were 399 hybrids and 201 dominant types. This corresponds to a suspiciously low χ^2 of 7.5×10^{-3} on the "null" hypothesis that the Binomial ratio is 2:1. (That is a P-value of .93 on 1 degree of freedom for the "uncorrected" Mendelian "2:1" law.) However, also Fisher cites the relatively large χ^2 value 3.3 (a P-value of about .07) for the "corrected" Mendelian hypothesis — corresponding to the 1.89 : 1.11 ratio which incorporates the 5.6% misclassification rate for hybrids — as another reason to believe that the data were "cooked." The problem here is that a hypothesis corresponding to the Binomial m.l.e. (or one sufficiently close to it) *must* yield small χ^2 values on 1 df; whereas, with increasing sample sizes, Binomial hypotheses with cell expectations different from the (limiting) sample frequencies *must* yield large χ^2 values. Then the Binomial model is damned because its expectations fit too well or if not well enough!

Thus, we shouldn't use Fisher's two-tailed reasoning with χ^2 values on 1 df. Of course, Fisher's criticism of Mendel uses a χ^2 value on 84 df, so he may (consistently) work both tails of the distribution: When χ^2 is too high appeal to Binomial alternatives with different first moments (but retain the Binomial second moment), and when χ^2 is too small, appeal to alternatives with the same first moment as the "null," but with a reduced second moment. As we see next, Fisher's argument using χ^2 tests makes explicit use of one variant of such an alternative hypothesis.

Specifically, a recipe for an alternative model that conforms to the two statistical requirements (above) is ready to hand. Propose that the data are "cooked," i.e., that observed relative frequencies which depart from Binomial expectations by more than some critical amount have been censored, or they have been altered to fall within the allowed range. Using small values of χ^2 to ascertain whether the observed frequencies in Mendel's data cluster too tightly about their Mendelian expectations, Fisher reports having

> "had the shocking experience lately of coming to the conclusion that the data given in Mendel's paper must be practically all faked."[25] (Bennett, 1983, p.199)

Consider one group of experiments in Mendel's study, from §5 of his paper, titled [F_2] "The First Generation [Bred] from the Hybrids," to illustrate this point. These trials constitute the first phase, testing the "3:1" ratio, in the investigation leading up to the "1:2:1" law for offspring of hybrids. Mendel allowed the F_1 hybrids to self-fertilize and tabulated the ratio of phenotypic dominant to recessives in the next generation. The data for the "3:1" law are displayed in Table A.

TABLE A: Mendel's 7 experiments testing the "3:1" law in the F_2 generation

	Dominant form	Recessive form	Total
Experiment 1 (seed shape)	5,474 round	1,850 wrinkled	7,324 seeds
Experiment 2 (seed color)	6,022 yellow	2,001 green	8,023 seeds
Experiment 3 (seed-coat color)	705 grey-brown	224 white	929 plants
Experiment 4 (pod shape)	882 inflated	299 constricted	1,181 plants
Experiment 5 (pod color)	428 green	152 yellow	580 plants
Experiment 6 (flower position)	651 axial	207 terminal	858 plants
Experiment 7 (stem length)	787 long	277 short	1,064 plants.

[25] . This was written by Fisher to E.B.Ford, in a letter dated 2 January 1936. Fisher confesses that his suspicions about Mendel's data were aroused by the problem associated with the mis-classification of hybrids, discussed (above) in section 3.

Fisher's Table V reports a χ^2 for these data of 2.1389, P = .95 (on 7 degrees of freedom). That value is obtained by summing the seven separate χ^2's (each an "experiment" with 1 degree of freedom) and using the statistical fact the sum of χ^2's is again χ^2 (on the sum of degrees of freedom). Thus, considering these seven experiments under the "3:1" law, the odds are 19:1 that, in a repetition of these trials, the new χ^2 (on 7 degrees of freedom) will exceed the magnitude (2.1389) observed in Mendel's data.

Fisher's Table V offers a cumulative χ^2 analysis for **all** of Mendel's data, constituting 84 degrees of freedom.[26] The upshot is a P-value in excess of .9999. That is, given Mendelian expectations,

---

[26] . As has been noted by others, e.g., Edwards (1986, p.299) there are two, statistically minor points concerning Table V.

(1) The last row of Fisher's table, "Illustrations of plant variation," constitute 20 degrees of freedom, includes a double-counting of 20 plants out of 511 plants (or a doubling counting of 915 seeds out of 15,347 seeds) from the first two of Mendel's seven experiments on the "3:1" law. When the data from these 511 plants are separated from the other results of the first two experiments, the χ^2's change as follows: In experiment 1, the new χ^2 (1 df) is .57504 (to replace .26288). In experiment 2, the new χ^2 (1 df) is .04811 (to replace .01500). The net change for the sum is a new χ^2 (2 df) of .62315, to replace .27788. (This corresponds to a new P value of about .73, to replace a P value of .87.) Concerning Fisher's Table V summary of the "3:1" ratios, the new χ^2 (7 df) is 2.48413, or a new P-value of about .93 instead of Fisher's .95. Since this correction adds only .3452 to the overall χ^2 (84 df), no doubt Fisher would persist in his claim that the overall P-value, .999, remains too good to be true.

(2) Also, Fisher's Table V includes all the data pertaining to the supposed misclassified hybrids used in testing the "2:1" law— data both from the five monofactorial and from the single trifactorial studies. These constitute another 15 degrees of freedom. However, Fisher uses the "uncorrected," Mendelian "2:1" expectations in calculating the cumulative χ^2 on 84 degrees of freedom. If we accept Fisher's objection that there was a 5.6% misclassification rate for hybrids, then Fisher incorrectly uses the uncorrected "2:1" expectations to calculate χ^2. It seems to me that, to be fair to Mendel, Fisher should have calculated in one of two ways instead:

Either (i) Fisher might have taken his first objection to heart and used the "corrected" expectations regarding the plant characteristic data for the "2:1" law. Considering the six, 1 df trials (n = 100 each). With a "corrected" expectation, the cumulative χ^2 for these increase from 4.58 to 7.36. Regarding the data for the relevant nine 1 df cells in the trifactorial study, the χ^2 for these increase from 6.29 to 10.68. The net gain of 7.17 between these two raises Fisher's Table V χ^2 on 84 df to 48.78, with a corresponding P of about .9975. [Of course, I reject this defense of Mendel, as I argue that there is no need to used a "corrected" expectation for his "2:1" law data.]

Or (ii) Fisher should have excluded these 15 degrees of freedom altogether and calculated a cumulative χ^2 on 69 degrees of freedom, just as Edwards does (1986, p.308).

in a repetition of all of Mendel's experiments, the odds are better than 10,000:1 that a new χ^2 will exceed the value achieved by Mendel's data. Next, I try to make more precise when small values of χ^2 might plausibly indicate data are "cooked."

4.2 An alternative model? Apart from manipulating the data, are there sensible, rival accounts that lead to a reduced second moment for Mendel's data? The Mendelian model posits that distinct plants are the result of probabilistically independent fertilizations. A random pollen grain fertilizes a random egg cell. Regarding egg cells, biology supports this hypothesis. However, it is plausible that pollen cells form on the anther in a restricted pattern. Specifically, during meoisis, a hybrid germ cell becomes a tetrad of 2 dominant and 2 recessive pollen cells. It is plausible to suppose that these move to the surface of the anther roughly as a tetrad, maintaining their proximity. If this is correct, the dominant and recessive pollen form something approximating a checkerboard pattern on the anther.

When the anther bursts in an open flower there is mixing. The tetrad configuration appears irrelevant to the Mendelian model. However, in the garden pea, self-fertilization occurs when the keel is very tightly packed. Recall, Mendel selected the pea plant for that feature -- to minimize foreign pollen. Also, Mendel's garden averaged only about 30 peas/plant (over 511 plants) in experiments 1 and 2, compared to the 106.5 peas/plant in Bateson-Kilby's (1905) study (over 283 plants) of the same 3:1 law, or the 217.6 peas/plant in Darbishire's (1909) 1908 large-scale experiment (over 482 plants) for testing the 3:1 law. Evidently, Mendel's garden was not Eden for pea plants. Given their yields, the stems were not likely to have been subject to high water pressure from their roots, which (I understand) determines when and how vigorously the anthers burst inside the keel .

Two questions are obvious: First, is there evidence to confirm or to refute the speculative genetics that Mendel's peas are not independently distributed within self-fertilizing pods? Second, does it matter to Fisher's analysis if the model of pea genetics is not quite Mendelian but, instead, reflects

this alternative distribution of pollen cells? How much of Fisher's .9999 P-value can be explained away with some subtle correlation among the pollen?

To the best of my knowledge, the issue whether peas in a pod are probabilistically independent has not been rigorously tested by field trials, at least not for the varieties of peas Mendel grew and under similar circumstances.[27] There are some intriguing numbers in Mendel's data, however, to suggest that not all is i.i.d., as we see next.

---

[27]. Edwards (1986, pp. 300-301) argues that peas are excellent randomizers, contrary to the view offered by Weiling (1966) that they carry sub-binomial variance, for example. Edwards' reasoned position rests on a reexamination of the large data sets in Bateson & Kilby's (1905) and Darbishire's (1908, 1909) studies - all experiments that (having among other goals) re-test the Mendelian "3:1" law in self-fertilizing pea hybrids. Edwards (p. 300) reports Weiling combining these data in the fashion of Fisher's meta-analysis. That yields a total χ^2 = 1008.8 on 1,062 df, corresponding to about 95% (= 1,008.8/1,062.0 %) of the variance under ordinary (Mendelian) i.i.d. Binomial sampling. So Edwards finds no basis here for Weiling's claim that sub-binomial variability explains away Fisher's analysis with an extremely low χ^2 = 41.6 on 84 df, corresponding to less than 50% of the Mendelian variance, and I agree.

The one-sided P-value even for this meta-analysis is about .87, however. In that sense, these data are not quite as "unremarkable" as Edwards makes out, though (of course) they do not even remotely approximate the "too good to be true" χ^2 values Fisher gives for Mendel's data.

Another first-rate investigation of these two studies is found in Stephens' (1994) work. He, too, concludes (§6.4) that these large-scale studies (each) fit the traditional Mendelian model, but not so with Mendel's own data.

My analysis of these studies lead to the following χ^2 values, taking each plant as its own "experiment," i.e., with 1 df/plant. I have ignored those plants (as indicated) where either the expected number of recessives fell below 6 peas/plant, i.e. where fewer than 24 peas grew on a single plant, or where the authors report that a plant had shed "many" peas before counting. By deleting plants with fewer than 24 peas, in effect, I have increased the average χ^2 value (per d.f.), as the binomial distribution is truncated at the origin. For example, with 23 peas, even if all 23 show the dominant trait, the 1 df χ^2 = 5.99 with a P-value of .014.

Bateson & Kilby (1905) χ^2 = 289.17 on 283 df (with 38 plants not counted);
Darbishire (1909) 1906 data: χ^2 = 7.44 on 13 df;
Darbishire 1907 data χ^2 = 59.35 on 67 df (with 20 plants not counted);
and Darbishire 1908 data χ^2 = 440.48 on 482 df (with 4 plants not counted).

This gives a total χ^2 of 796.44 on 845 df, corresponding to about 94% of the Mendelian variance, with a one-sided P-value of about .88. Again, my ("conservative") analysis carries fewer degrees of freedom where I have not counted plants, as indicated above.

Suppose that Mendel tried to waste as few plants as possible. When counting peas, he counted all the peas in a pod. Concerning experiments with the first generation [F_2] bred from the hybrids, Mendel writes,

> Expt.1. Form of seed.— From 253 hybrids 7,324 seeds were obtained in the second trial year. Among them were 5,474 round or roundish ones and 1,850 angular wrinkled ones. Therefrom the ratio 2.96 to 1 deduced.
>
> Expt.2. Color of albumen.— 258 plants yielded 8,023 seeds, 6,022 yellow, and 2,001 green; their ratio, therefore, is as 3.01 to 1.
>
> In these two experiments each pod yielded usually both kinds of seed. In well-developed pods which contained on the average six to nine seeds, it often happened that all the seeds were round (Expt. 1) or all yellow (Expt. 2); on the other hand there were never observed more than five wrinkled or five green ones in one pod. It appears to make no difference whether the pods are developed early or later in the hybrid or whether they spring from the main axis or from a lateral one. In some few plants only a few seeds developed in the first formed pods, and these possessed exclusively one of the two characters, but in the subsequently developed pods the normal proportions were maintained nevertheless.

So, Mendel reports a total of 15,347 seeds taken from 511 plants. Is it "significant" that there were never more than 5 recessives in a single pod?

The answer depends upon the distribution of peas in a pod. For example, with an average of 6.65 peas/pod distributed as:

20.% pods; 22.% pods; 38.% pods; 12.5% pods; and 7.5% pods

at 5 peas; 6 peas; 7 peas; 8 peas; 9 peas

on the Mendelian model, the chance of not seeing either 6 or 7 recessives in a single pod (among about 2300 pods) is approximately .015. If we increase the average peas/pod to 6.925 with

15% pods; 20% pods; 30% pods; 27.5% pods; and 7.5% pods

at 5 peas; 6 peas; 7 peas; 8 peas; 9 peas

the chance of observing no more than 5 recessives in any of the (about 2215) pods is approximately .0055. Thus, on the simple Mendelian model, gauged by "significance," his data

are at least mildly surprising for their absence of large numbers of recessives in any pod.[28]

Reconsider the second question: Does it matter to Fisher's analysis if the model of pea genetics is not quite Mendelian but, instead, reflects this alternative distribution of pollen cells? In response, re-examine Fisher's Table V. We see that if we want to reduce the overall expected 1-sided P-level for such a table to about .98, then we require a total χ^2 (84 df) of about 59.07. That corresponds to a model with about 70% of the Mendelian (Binomial) variance, but with the same first moment. Then we would expect an observed χ^2 (summed on 84 df) of about 59. Likewise, if we want to reduce the expected overall 1-sided (84 df) P-level to the more extreme .99 level, then we require a total χ^2 (84 df) of about 56.15, i.e., a model with about 2/3rd's of the Mendelian variance. Even a rival model with only 75% of the Mendelian one carries an expected χ^2 (84 df) with a corresponding one-sided P-value in excess of .95. However, there is more to explain about the distribution the component χ^2 from Mendel's data than just the 84 df sum, as we shall discover. What follows next is a simplistic model of self-fertilization for peas, based on an 80% seed survival, that has the same (Binomial) first moment and 74.1% of the variance of the Mendelian model, corresponding to a one-sided χ^2 P-value of about .96 on (Fisher's) 84 df. (For comparison, at only 70% seed survival, this model has about 77.8% the variance of the Mendelian model, which corresponds to an 84 df one-sided χ^2 P-value of about .93) The alternative model uses a 'hypergeometric' distribution for selection of the surviving pollen, following-up on the speculative account of how self-fertilizing hybrid peas depart from the i.i.d. "fair" (binomial) chance that either kind of pollen (dominant versus recessive) fertilizes eggs within the same pod. Here, of course, it is a departure from "independence, i.e.. it is the first 'i.' of the i.i.d. Mendelian model that is removed. Call it the "Correlated Pollen" [CP] model. (Perhaps this is what Weiling intends with his appeal to a hypergeometric model of pea selection?)

[28] Weldon (1901, p.230) arrives at the opposite conclusion. Weldon considers data from Mendel's second experiment, on seed color, and then he calculates the odds that a pod (with 6, 7, 8, or 9 seeds) contains all recessive (only green) peas. This doubly dilutes the force of Mendel's observation of no more than five recessive peas, either for shape or color, in a single pod.

The Correlated Pollen Model. Suppose that within the pea-flower for hybrids, 10 egg cells form according an i.i.d. "fair" (binomial) distribution. However, approximating the speculated, checkerboard pattern that pollen have on the anther, suppose that exactly 5 of every 10 pollen cells arriving at the egg cells are dominant. Last, assume that, with equal probability, 2 of these 10 zygotes spontaneously abort, leaving 8 peas/pod. The result is a model where pollen cells are negatively correlated within a pod.

4.3 The "3:1" law. Table B displays the two sets of probabilities for the number N (out of 8 peas in a pod) of phenotypic dominants in the "3:1" law for self-fertilizing hybrids. The top row gives the ordinary Mendelian probabilities, whereas the bottom row gives the probabilities for Correlated Pollen model, with 74.1% of the Mendelian variance but the same first moment.

Table B
A comparison between two models of the "3:1" law for 8 peas in a pod
N is the number (out of 8) of dominant peas

	N = 0	N = 1	N = 2	N = 3	N = 4	N = 5	N = 6	N =7	N = 8
MENDELIAN	1.50×10^{-5}	3.66×10^{-4}	3.85×10^{-3}	2.31×10^{-2}	8.65×10^{-2}	2.08×10^{-1}	3.11×10^{-1}	2.67×10^{-1}	1.00×10^{-1}
CORRELATED	0.00	0.00	0.00	6.94×10^{-3}	6.94×10^{-2}	2.36×10^{-1}	3.61×10^{-1}	2.57×10^{-1}	6.94×10^{-2}

A more useful comparison is the distribution of the number N of dominant peas out of 32, corresponding to 4 pods/plant. That is, with 4 pods of 8 peas/plant, the probabilities for outcomes of N phenotypic dominants out of 32 is given in Table C.

TABLE C

χ^2, P-value, and Probabilities for N (out of 32) dominant peas in the 3:1 law

χ^2	P-value	N	Mendelian Model	Correlated Pollen Model
10.667	.001	32	1.00×10^{-4}	2.30×10^{-5}
8.167	.004	31	1.07×10^{-3}	3.44×10^{-4}
6.000	.014	30	5.54×10^{-3}	2.39×10^{-3}

4.167	**.041**	**29**	**1.85×10^{-2}**	**1.04×10^{-2}**
2.667	**.105**	**28**	**4.46×10^{-2}**	**3.16×10^{-2}**
1.500	**.221**	**27**	**8.32×10^{-2}**	**7.14×10^{-2}**
0.667	**.410**	**26**	**1.25×10^{-1}**	**1.24×10^{-1}**
0.167	**.670**	**25***	**1.55×10^{-1}**	**1.70×10^{-1}**
0.000	**1.000**	**24***	**1.61×10^{-1}**	**1.87×10^{-1}**
0.167	**.670**	**23***	**1.43×10^{-1}**	**1.65×10^{-1}**
0.667	**.410**	**22***	**1.10×10^{-1}**	**1.19×10^{-1}**
1.500	**.221**	**21**	**7.32×10^{-2}**	**6.89×10^{-2}**
2.667	**.105**	**20**	**4.27×10^{-2}**	**3.23×10^{-2}**
4.167	**.041**	**19**	**2.19×10^{-2}**	**1.16×10^{-2}**
6.000	**.014**	**18**	**9.90×10^{-3}**	**3.63×10^{-3}**
8.167	**.004**	**17**	**3.96×10^{-3}**	**9.44×10^{-4}**
10.667	**.001**	**16**	**1.40×10^{-3}**	**1.49×10^{-4}**
			$\Sigma P > 9.99 \times 10^{-1}$	**$\Sigma P > 9.99 \times 10^{-1}$**

***** Denotes those outcomes where the probability is greater under the Correlated Model than under the Mendelian Model.

The expected sum of χ^2 (1,000 df) from 1,000 plants (rounded to integer numbers of plants, above), each with 32 peas/plant, is 1,000 for the Mendelian model and only 734 for the Correlated Pollen Model, reflecting the reduction in variance for the latter. Of course, under the Correlated Pollen Model, no pod ever shows more than 5 recessives, since (by design!) there are at most 5 recessive pollen grains fertilizing each pod.

In his report of Experiments 1 and 2, Mendel illustrates his data with a sequence of ten plants for each of the two pea characteristics, shape and color. These 20 plants produced 915 seeds, an average of 45.75 seeds/plant. That is fully 50% greater than the overall average of 29.39 peas/plant (14,432 seeds) from the other 491 plants in Experiments 1 and 2. Thus, these illustrative plants were somewhat hardier than the average plant in Mendel's garden. Nonetheless, the sum of these 20 χ^2 is only about 12.49, corresponding to a one-sided P value of about .90.

Contrast the usual (i.i.d. binomial) Mendelian and the Correlated Pollen models, each with 4 pods of 8 peas/plant resulting in 9 possible values of χ^2 as illustrated in Table C. Grouped by χ^2 values, Table D gives the expected numbers of plants out of 20, together with the integer values obtained by rounding these to whole plants. Below each column is the expected χ^2 (20 df) and **P**-value that results. Again, Mendel's 20 plants have $\Sigma\chi^2 = 12.49$ on 20 df and **P** = .90.

TABLE D

χ^2	**Expected numbers and integer numbers of plants (out of 20)**			
	Mendelian Model		**Correlated Pollen Model**	
0.000	3.22	3.00	3.74	4.00
.167	5.96	6.00	6.72	7.00
.667	4.70	5.00	4.84	5.00
1.500	3.12	3.00	2.80	3.00
2.667	1.76	2.00	1.28	1.00
4.167	0.80	1.00	0.44	0.00
6.000	0.32	0.00	0.12	0.00
8.167	0.10	0.00	0.01	0.00
10.667	0.00	0.00	0.00	0.00
Expected χ^2 (20 df)	**19.57**	**18.33**	**14.67**	**11.67**
and **P**-values	**.50**	**.56**	**.79**	**.93**

Unfortunately, Mendel does not provide us with additional data at the plant-level for his experiments relating to the "3:1" law apart from his report about 4 extreme cases.[29] Instead, he

[29] Mendel (p.327) does give extreme plant counts in both directions for each of the two pea-characteristics. For Experiment 1 (pea shape) he reports that one plant yielded 43 round peas (which is dominant) against 2 that were angular, and another plant had 14 round and 15 angular peas. Regarding Experiment 2 (pea color) a plant had 32 yellow peas (dominant) and 1 green one. Fourth, a plant had 20 yellow and 19 green peas. Respectively, the 1 df χ^2s for these four cases are: 10.14 (a P-value of about 1.5×10^{-3}); 11.05 (a P-value of about 9.0×10^{-4}); 8.50 (a P-value of about 3.5×10^{-3}) and 11.70 (a P-value of about 6.3×10^{-4}). The latter being the extreme of the four.

However, taking into account the discreteness of the Binomial distribution, the exact Binomial counts for these "tails" are less extreme than the χ^2 values. Respectively, the Binomial probability of an outcome as or more extreme than Mendel reports are, respectively, only: 1.54 x

gives pooled data for each of the 7 experiments of the "3:1" law taken from the remaining 2 pea-characteristics experiments and the 5 plant characteristics experiments. Also, there are 2 tests of the "3:1" law contained in the two-factor experiment and 3 tests in the three-factor experiment. These data have χ^2 that sum to 3.98 (12 df) , with a P-value of about .98. Thus, these 32 tests of the "3:1" law yield χ^2s that sum to 16.47 (32 df), and a corresponding P-value of about .98 also.

It is mere speculation whether, in each case, these experiments of the "3:1" law are the results of Mendel cultivating all the peas from a set of pods, rather than sampling a few peas from different pods. If the former was the situation, i.e., if Mendel wasted as few peas as possible from each pod, the Correlated Pollen model might apply. That model supports an expected P-value of about .96, compared to the observed P-value of .98.

What about other studies of the "3:1" law that were conducted in early years of this century? Both of Bateson-Kilby's (1905) study (with more than 280 plants), and Darbishire's (1909) large-scale experiment (with over 450 plants) give plant-by-plant counts and show a cumulative χ^2 that is within 6% of the Mendelian model (see fn. 26). That is, these studies reflect very little variance reduction; in contrast with the Correlated Pollen model, which calls for about a 25% reduction in Binomial variance. Moreover, the scatter of P-values from these two experiments approximate the uniform U[0,1] distribution, in accord with the Mendelian model. Group the P-values from the 1-df χ^2 by deciles. The "null" hypothesis is that the P-values are uniformly distributed: that is, the "null" probability is .10 for each decile of P-values. Tested this way, Darbishire's 1908 data (482

10^{-3}; 1.90×10^{-3} ; 3.63×10^{-3} and 1.15×10^{-3}. For example, the Binomial probability that all 258 trials in Experiment 2 were less discrepant than the fourth case is about .742. The corresponding result for Experiment 1 is about .677. More importantly, the probability under the Mendelian model that all 511 trials are less discrepant outcomes than the fourth case (above) is about .554. By contrast, under the Correlated Pea model, the probability that all χ^2s from 511 trials (for plants of 32 peas each, i.e., 4 pods of 8 peas each) do not exceed 10.67 is about .916. That approximation gives the Mendelian model a factor of about 5:1 over the Correlated Pollen model, for the maximum discrepancy in 511 trials. That is, the best such outcome would be that the probability is .50 that the maximum would be exceeded again in 511 trials. But (1-.554)/(1-.916) is about 5.3. Again we face the question of how to index "discrepancy."

plants) has a χ^2 of 13.23 (9 df), for a P-value of about .15. Batetson & Kilby's data (283 plants) has a χ^2 of 4.03 (9 df), for a P-value of about .91.

The plants in these two studies have profuse yields: respectively, 7 and 3 times the average yield per plant compared with Mendel's garden. In these two studies, moreover, there is no correlation between plant yield and P-value. For the Bateson-Kilby data, a linear regression of P-value on plant yield fits with an intercept of .516, a slope of -2.74×10^{-4}, and a correlation of about .06. With Darbishire's 1908 data set, the linear regression of P-value on plant yield has an intercept of .510, a slope of 4.06×10^{-5}, and a vanishing correlation of .02. In short, there is no indication of any departure from the Mendelian model in either of these two studies. Certainly, there is no warrant for a Correlated Pollen model here, nor does consideration of just the plants with small yields in these studies offer any evidence for a Correlated Pollen model.

However, in Darbishire's 1907 trials of the "3:1" law his plant yields were a better approximation of Mendel's. From 87 plants he reports 3904 peas, or 44.87 peas/plant. Discount 20 plants, whose yields were less than 24 peas/plant.[30] The net result is a data set of 67 plants (each with at least 24 peas/plant) that produced 2703 Yellow and 974 Green peas: a total of 3677 peas for an average of nearly 59 peas/plant. This is roughly on the same order of Mendel's 20 plants illustrating the "3:1" law, whose average yield was about 45 peas/plant. Overall, these data give a χ^2 (1 df) of 4.35, with a P-value of about .04 - "significant" at the .05 level. Nonetheless, even with such a "significantly" large number of recessives, the 67 χ^2s (1 df) for Darbishire's 1907 study total only 59.86, or about 88% of the Mendelian variance: for a 1-sided P-value of about .62. The distribution of the 67 P-values is "significantly" far from uniform, as one might anticipate

[30] These total 227 peas: 166 Yellow - dominant, and 61 Green - recessive. To repeat, there are two reasons for "cleaning" the data this way. First, the rule-of-thumb for χ^2 calls for expected values of at least 6 per cell; hence, calling for plants with at least 24 peas to test the "3:1" law. The second reason is that the Correlated Pollen model yields increasing variance as fewer peas per pod are counted. In the extreme, with only 1 pea/pod counted, the Correlated Pollen model is the Mendelian model.

from the overabundance of recessives: Grouped by deciles the χ^2 = 19.42 (9 df), with a one-sided P-value of .022. But that thinking misses an unusual feature of the data.

The noticeable point in the departure from the uniform distribution of P-values in Darbishire's 1907 data-set is not due to an excessive number of plants (out of 67) that have individual P-values that are too small. Figures 1 and 2 display the histograms of P-values by deciles in Darbishire's 1907 data-set (N=67) and in Mendel's tests (N=32) of the "3:1" law. In fact, in Darbishire's data, the lower third of P-values is under-represented (15/67 = 22.4%) against a uniform distribution. (The upper third of P-values is over-represented (29/67 43.3% and the middle third shows about as expected (23/67 = 34.3%) for a uniform distribution.) The shortfall in low P-values in Darbishire's results duplicate a similar short-left-tail histogram of P-values from Mendel's (32 df) data on the 3:1 law. (More on this feature of plotted P-values later on!)

In summary, regarding the "3:1" law, Mendel reports the absence of pods with more than 5 recessives for either pea-characteristic (color or shape). This is at least mildly surprising under the Mendelian model (a P-value less than .02). An examination of the 32 separate tests of "3:1" law present in Mendel's data produces a histogram of P-values (Figure 2) with a variance reduction also in accord with the speculative, Correlated Pollen Model. On the other hand, neither of the large-scale studies by Darbishire or Bateson & Kilby give any support whatsoever to such a speculative model. But these studies were under conditions noticeably different from Mendel's, at least in terms of plant yield. Whether the Correlated Pollen model is negatively associated with, e.g., soil moisture (a factor that positively affects the number of pollen competing to fertilize the eggs, I believe) - a factor also positively associated with yield - is a question for field trials to decide. One (smaller) study, by Darbishire, that approximates Mendel's plant yields also has a reduced sum-of-χ^2, despite having a "significantly" different first moment from the Mendelian model.

4.4 The "2:1" law. Next, let us consider the evidence Mendel offers for his "2:1" law. that offspring of self-fertilizing hybrids have a 2:1 ratio of hybrid to pure dominants. We have already considered one aspect of these data in connection with Fisher's (first) objection, relating to the alleged misclassification of the hybrids as pure-bred dominants. But that is not what we are concerned with here. The question now is whether the Correlated Pollen Model is a serious rival to Fisher's alternative of "faked" data, based on the data being "too good to be true"?

Unfortunately, unlike his discussion of the "3:1" law, Mendel gives us only pooled counts, not any counts by outcomes of individual plants (let alone by pod) for his experiments on the "2:1" law. There are 15 such tests. Eight come from experiments on the phenotypic dominant plants that were used to test the "3:1" law. Each of the two pea-characteristics and four of the five plant characteristics were tested once to determine how the "3" divide up between hybrid and pure dominants. There were two trials testing pod-color. Four tests are available from the bi-factorial

trials, which yielded 499 peas (out of 529) of the relevant sorts.[31] To maintain rough parity with the sample sizes from the other 12 tests, in the tri-factorial study (involving altogether 639 test plants of which 632 are relevant to tests of the "2:1" law), I have pooled the counts using the coarse categories, of the individual characteristics, producing only 3 more tests.[32] A table of the resulting χ^2s is just below, showing the "too good to be true" totals.

---

[31] I have tested each character twice for the "2:1" by partitioning the bifactorial data on the phenotype of the other character.

[32] The question is how fine to partition the tri-factorial experiment into separate tests of the "2:1" law? At the extreme, there are 9 genotype configurations for the other two characteristics, opening the door to 27 tests (with 26 df), achieved by partitioning a test for each trait on each such genotypic configuration of the other two.

To partition by the phenotype of the other two traits still yields 12 tests, as the following table illustrates. However, such a partition leads to tests with samples that differ by an order of magnitude, e.g., N = 269 in Exp. 1 and N = 27 in Exp. 4.

Mendel's Trifactorial Data on the "2:1" law partitioned into 12 tests using the phenotypes of the other two traits.

Experiment	χ^2	P-value	N
1 pea shape	.00186	.97	269
AA vs **Aa** among [**BBCC + BbCC + BBCc + BbCc**]			
2 pea shape	.74419	.39	86
AA vs **Aa** among [**bbCC + bbCc**]			
3 pea shape	.02041	.89	98
AA vs **Aa** among [**BBcc + Bbcc**]			
4 pea shape	.66667	.41	27
AA vs **Aa** among **bbcc**			
5 pea color	.98327	.32	269
6 pea color	.68750	.41	88
7 pea color	.02041	.89	98
8 pea color	.23529	.62	34
9 seed-coat	.22491	.63	269
10 seed-coat	.42186	.52	96
11 seed-coat	.27841	.60	88
12 seed-coat	.00000	1.00	30
	$\Sigma\chi^2 =$ 4.285 (12 df)	.98	

A: Round pea; **a**: wrinkled pea
B: Yellow albumen; **b**: green albumen
C: Grey seed-coat; **c**: white seed-coat

If these 12 degrees of freedom replace the 3 entries in the text that represent the tri-factorial data, (i.e., numbers 12, 13, and 14) in the table, the upshot is a $\Sigma\chi^2 =$ 10.096 (24 df), and a P-value of .99.

Mendel's Data on the "2:1" law

Experiment	χ^2	P-value
1 (seed shape)	.17348	.67
2 (seed color)	.42486	.51
3 (seed coat)	.32083	.57
4 (pod shape)	.50005	.48
5a (pod color)	2.00022	.16
5b (pod color)	.12502	.74
6 (flower position)	.00500	.94
7 (plant height)	1.28000	.26
8 (bi-fac. pea shape)	.10631	.75
9 (bi-fac. pea color)	.08140	.78
10 (bi-fac. pea shape)	.75000	.38
11 (bi-fac. pea color)	.04412	.83
12 (tri-fac. pea color)	.00938	.92
13 (tri-fac. pea shape)	.33129	.57
14 (tri-fac. seed coat)	.30550	.58
	$\Sigma\chi^2$= 6.45746	.97 (15 df)

The Mendelian model, with independence between peas, tells a simple story about distribution of hybrids within a pod. From a self-fertilizing hybrid, the (conditional) chance that phenotypic dominant pea is hybrid is 2/3. Moreover, this chance is independent of the number of phenotypic dominants in a pod. The Correlated Pollen model provides a more complicated account.

In a pod of 8 peas, both models give the same (marginal) distribution for the number of hybrids.[33] However, because of the correlation between pollen in the C-P model, the conditional probability of a pea being hybrid, given that it is phenotypic dominant, increases with the number, n, of such

[33]. Under both models, with a self-fertilizing hybrid the chance of a hybrid pea is 1/2, and separate peas are independent for this genotype. In the Mendelian model, it is evident that the chance that n out of 8 peas are hybrid is just the Binomial chance of n "heads" out of 8 flips of a fair coin. For the C-P model reason as follows: exactly one egg-type matches with each pollen type to form a hybrid. But egg-types are i.i.d. Binomial, with chance 1/2 for each type. Hence, as with the Mendelian model, the chance that n out of 8 peas in a pod is hybrid is given by the Binomial distribution for i.i.d. data.

phenotypic dominant peas there are in a pod of 8. Let 'H' stand for the event that a pea ("randomly" chosen from the n) is hybrid rather than pure-dominant. In the C-P model, $p(H \mid n)$ = (.5, n = 3), (.5625, n = 4), (.61765; n = 5), (.66346; n = 6), (.70077; n=7), and (.73125; n = 8). Under this model, the probability distribution of pods, by the number n ($3 \le n \le 8$) of phenotypic dominant peas out of 8 in a pod, is in the ratios

16 : 160 : 544 : 832 : 592 : 160

for n = 3 4 5 6 7 8 phenotypic dominant peas.

This yields an expected "ratio" of hybrids to pure-dominants of only .65879, rather than the 2/3rds value of the Mendelian model.

However, this is not the relevant average to use for a contrast with Mendel's data. Mendel collected peas from pods to reach a number of test plants in creating his samples of the 2:1 law. That is, he collected about 500 seeds for each of his first two experiments (about 10% of the total available from the pervious year's two experiments on the 3:1 law) and he used 100 seeds for each test of the plant characteristics. I assume that he used all the phenotypic dominants in a pod and randomly sampled pods until he reached his quota of test seeds.

Among 100 randomly selected pods from the C-P distribution, we expect (to the nearest integer)

	1	7	24	36	26	7	pods
with n =	3	4	5	6	7	8	phenotypic dominant peas.

As there are n phenotypic peas/pod, this induces a distribution of peas, so that

	0.5%	4.6%	19.8%	35.7%	30.1%	9.3%	of the peas come from pods
with n =	3	4	5	6	7	8	phenotypic dominant peas.

Rounded to whole numbers of (100) peas, this gives a C-P expected ratio (Hybrids to Pure-Dominants) of .6665 - compared to the Mendel value of .6667. In the C-P model, the variance for this ratio is 3.29×10^{-2}. Under the same sampling rule, the Mendelian variance for this ratio is 4.39×10^{-2}. Hence, the C-P model has (approximately) 75% of the Mendelian variance for tests of the "2:1" law. The expected sum-of χ^2 P-values, then, is about .96, which figure matches the value obtained from Mendel's 15 tests of the "2:1" law.

4.5 Mendel's data from artificial fertilizations. The C-P model is a speculative proposal about the details of self-fertilizing peas which allows a re-analysis of Mendel's data relating to the "3:1" and "2:1" laws. However, fully 15 of Fisher's 84 degrees of freedom in his Table V reflect Mendel's experiments on gametic ratios. These come from section 9 of Mendel's paper, "The Reproductive Cells of the Hybrids." There, he shows that it is irrelevant whether hybrids are formed by fertilization with a dominant pollen and a recessive egg or vice versa.

Five experiments, each with 3 df, sum to a combined χ^2 of only 3.6730, with a P-value of .9987. These data reflect approximately 550 plants, all the results of *artificial* fertilizations. For each test there were 4 cells, which had an equal expected count under the Mendelian hypothesis. Four of the tests involved the two pea-characteristics, and used aproximately 100 test-plants each, with expectations of approximately 25 plants/cell. The fifth test used two plant characteristics (flower color and plant height), and had expectations of about 40 plants per cell. Twelve plants, total, were used to create the peas for the first four tests. Mendel writes that he made 45 fertilizations leading to the 166 plants of the fifth test.

Whatever merit there is to using the C-P model for re-examining the other data, it is of no relevance with these. They are the results of artificial fertilizations, all, where Mendel "dusted" the pollen onto each stigma for the flowers. I have no account of why these ratios cluster so closely around their expected values, each of which is 1/4 of its sample size.

In these experiments, classification of a plant was (in the case of pea-characteristics) based on whether all of a plant's peas were of one (dominant) kind or whether it was hybrid, showing both dominant and recessive traits. Similarly, in the studies on plant characteristics, entire plants were classified, not merely individual peas. I find it hard to understand how misclassifications occurred.

5. What model of "cheating"? The distribution of Mendel's Ps is the focus of A.W.F. Edward's (1986) important article. He concludes (p. 310) that the data were adjusted (rather than censored) to avoid extreme segregations in the record. I think there is rather good reason to agree with Edwards in that, at least, Mendel's "cheating" was not either the result of censoring extreme values, nor a more aggressive strategy of faking outcomes that cluster at the expected values.

Rather than reproducing Edward's chart of Fisher's (1 df) signed-χs consider, instead, a histogram of the (78) P-values from χ^2.[34] Under the Mendelian hypothesis, these are uniformly distributed on the unit interval. The sum of the 78 (1 df) $\chi^2 = 41.49$, with a P-value of about .9997, just as in Fisher's Table V. If we examine this distribution by deciles, the resulting (9 df) χ^2 is 19.18, with a P-value of about .023. The discrepancy is maximum for the first decile, reflecting the absence of a "left-tail" in Mendel's data, as we are well aware. However, the uppermost decile, rather than showing a excess over its expected value (of 7.8) has only 4 values. Even the upper-two deciles fail to exceed their expectations. In short, there is no evidence here of aggressive cheating, where outcomes at the highest decile of P-values are over-represented.

The final two histograms show the P-values from the data as Mendel organized it in his paper. Following his presentation, by my count, there are 48 experiments or separate steps in the tests that he devises.[35] I have plotted these by their P-values, which are uniformly distributed under the

---

[34] Edwards deletes the 6 "questionable" values from the (n = 100) tests of the "2:1" law that purportedly involves misclassification of hybrids.

[35]. Specifically, these 48 P-values comprise:

22 1 df tests of the 3:1 law with two pea-characteristics - I adjusted Mendel's two grand totals to avoid double-counting the twenty illustrative plants.;
5 1 df tests of the 3:1 law with plant characteristics;
8 1 df tests of the 2:1 law;
2 3 df tests in the bifactorial study;
2 1 df tests of the 2:1 law in the bifactorial study;
1 3 df test in the trifactorial study (according to the number of hybrid characters, 0 -3);
1 5 df test in the trifactorial study of those hybrid in two traits;
1 11 df test within the trifactorial study of those hybrid in one trait;
1 7 df test within the trifactorial study of those hybrid in no traits
5 3 df tests in the gametic study.

Mendelian model regardless of the number of degrees of freedom involved in each test. Compare this histogram with that for Mendel's data in the "3:1" law. The bulge near the median in the histogram is "significant" by χ^2. Contrast the middle two deciles [.4, .6] with the 8 deciles in the complementary set. Under the Mendelian hypothesis these have probability .2 and .8. But the (1 df) $\chi^2 = 5.33$ for this hypothesis, with a P-value of .021.

What model of cheating, then, can the reader propose that replaces extremely discrepant outcomes with ones clustered about the median of χ^2s?[36] I challenge the reader to try to adjust binomial data from samples sizes in Mendel's experiments, so that the following three features appear in the resulting distribution of P-values from the (1 df) χ^2s:

1. There is a significant reduction in the left-tail of the Ps.
2. There is no significant departure from uniformity in the right tail of the Ps.
3. There is a significant concentration of the Ps about their median, i.e., about .50.

To be fair to Mendel, this exercise should be attempted without the aid of χ^2 Tables, which distribution, the reader recalls, K.Pearson discovered only in 1900!

Note that under the Mendelian null-hypothesis these P-values are uniformly distributed regardless the number of degrees of freedom associated with each.

[36]. The problem iterates in the literature. Novitski's (1995) reanalyzes Mendel's 5 experiments involving the "Reproductive Cells of the Hybrid," where Mendel tests the hypothesis that hybrids are the same regardless which parental gamete carries dominant trait. Novitski introduces a new factor of approximately 1:29 to the already low 1:700 significance level as reported by Fisher for these 15 (= 5 x 3) df, to conclude that the observed significance level for these data are at the exceptionally small level of approximately 1: 20,000. The new factor of 1:29 he arrives at by numerical simulation of the conditional distribution for the variance of five 3-df χ^2s, given the sum of these five χ^2s (= 3.67) calculated from Mendel's data. That is, given the sum (3.67), under the Mendelian null hypothesis, the odds are about 30:1 for 5 χ^2s with larger variance than is found in Mendel's data. Novitski (p. 65) finds that this "places some additional weight on the conclusion of the majority of others who have looked at Mendel's results that his data, in the context in they were reported in his paper, are highly improbable." Alas, Novitski does not consider whether the added factor, of the reduced (conditional) variance in these 5 χ^2s, in fact makes it more or less plausible that Mendel's data were faked.

6. Conclusions. I have here reviewed each of Fisher's two principal objections to the data in Mendel's classic paper.

There is an easy reply to Fisher's charge that Mendel "cheated" in producing data for the "2:1" law. Data, in fact, which are significant against the hypothesis of a corrected ratio (of hybrids to pure-dominants) that reflects the 5.6% misclassification error Fisher uncovered in Mendel's protocol. The reply is that Mendel's protocol included (as he wrote) extending the testing through successive generations bred from the same hybrids. This, sequential aspect of Mendel's design, provides a simple rebuttal to Fisher's claim of misclassification.

This paper offers no easy, nor complete answer to Fisher's second objection that, overall, Mendel's data conform too well to their theoretical expectations. There is insufficient variability in Mendel's counts under the Mendelian model. That is a certainty. But how is it to be explained, if not by positing reduced variance by cheating?

Regarding the "3:1" and "2:1" laws of segregation for self-fertilizing hybrids, I propose the Correlated Pollen model, an alternative to the usual Mendelian (i.i.d.) distribution of peas in a pod. The C-P model has the same first moment as the Mendelian model with about 75% of the Mendelian variance. This speculative model is enough to recover the P-values in Mendel's data for each of the two, main Mendelian laws. Also, the CP model fits much better than the Mendelian model the values that Mendel reports for the maximum number of recessives in pods, from more than 2 thousand pods. It does slightly worse than the (i.i.d.) Mendelian model in fitting the extreme value (per plant) that Mendel reports for the "3:1" law from 511 plants.

The C-P model gets no support whatsoever from the two large-scale studies by Kilby & Bateson, and Darbishire on the "3:1" law, both conducted in the early 1900s. However, these studies do not duplicate the small yields (per plant) in Mendel's garden. I offer reason to think that the C-P

model, if it applies at all, does not fit luscious plants. One of Darbishire's experiments that duplicated Mendel's lower yields shows an anomalous, small sum-of χ^2s, as in Mendel's data.

Even with all this effort, I have no insight to offer about the extremely low sum of χ^2s (15 df) from the 5 experiments on gametic ratios. These tests were intelligently designed to show that, regarding the Mendelian laws, a hybrid is a hybrid regardless which parent carried the dominant trait. The C-P model fails to apply to these data because all the test plants were the result of artificial fertilizations.

Where do we stand, more than sixty years after Fisher's shocking allegations against the authenticity of data in Mendel's paper. The allegation of misclassification (of hybrids) admits such a straightforward reply that I no longer find merit in that aspect of Fisher's criticism. But, unless some alternative model with reduced variance, like the C-P model, can be justified, I see little hope of explaining away the Ps that are "too good to be true."

The C-P model can be subjected to simple field trials, providing that classic strains can be cultivated under circumstances similar to those found in Mendel's garden. The difference with the Mendelian model is evident for the "2:1" law and makes testing straightforward. The C-P model (but not the Mendelian model) introduces a probabilistic dependence between the ratio of pure-breds to hybrids and the number of phenotypic dominants in a pod.

Regardless the outcome, no matter how peas self-fertilize, I urge the reader to study Mendel's classic paper and Fisher's provocative article. Mendel's work is a standard of clarity and a delight for its intelligent, sequential designs. Fisher, as always, is a brilliant statistician and imposing geneticist. As with many of his other writings, coming to an understanding of how he argues is the key, regardless what reader thinks, in the end, of his conclusion.

***Acknowlegments**

I have struggled with the Fisher-Mendel dispute for several years and owe thanks to many for helping me to find whatever is of value here. During Fall term 1997, where the current version was mostly completed, I was the guest of the STICERD Center and the Philosophy of Science Center at the London School of Economics. I have benefitted considerably from reactions to talks I have given on this subject at: the Popper Seminar of LSE; the Department of Biometrics of Cornell University; the Harrisburg Chapter of the American Statistical Associations; and the Philosophy of Science Seminar of the University of Pittsburgh.

References

Bateson, W. and Kilby, H. (1905) "Experimental Studies in the Physiology of heredity: Peas," *Royal Soc. Reports to the Evolution Committee* **2**, 55-80.

Bateson, W. (1909) *Mendel's Principles of Heredity*. Cambridge: University Press; (1990) Classics of Medicine Library, Div. of Gryphon Editions, Inc.: Birmingham, Alabama.

Bennett, J.H., ed. (1983) *Natural Selection, Heredity, and Eugenics*. Oxford: Carendon Press

Blixt, S. (1974) "The Pea" in R.C.King (ed.) Handbook of Genetics, vol. 2. New York: Plenum Press, chapter 9.

Cramér, H. (1946) *Mathematical Methods of Statistics*. Princeton: Princeton Univ. Press.

Darbishire, A.D. (1908) "On the Results of Crossing Round with Wrinkled Peas, with Especial Reference to their Starch-grains," *J.Royal Soc. of London B, Proceedings* **80**, 122-135.

Darbishire, A.D. (1909) "An Experimental Estimation of the Theory of Ancestral Contributions in Heredity," *J.Royal Soc. of London B, Proceedings* **81**, 61-79.

Elderton, W. Palin (1901) "Tables for Testing the Goodness of Fit of Theory to Observation," *Biometrika* **1**, 155-163.

Edwards, A.W.F. (1986) "Are Mendel's Results Really Too Close?" *Bio.Rev*. **61**, 295-312.

Gauch, H.G., Jr. (1993) "Prediction, Parsimony and Noise," *American Scientist* **81**, 468-478.

Fisher, R.A (1911) "Mendelism and Biometry," in Bennett, J.H (1983), 52-58.

Fisher, R.A. (1918) "The Correlation between Relatives on the Supposition of Mendelian Inheritance," 399-

Fisher, R.A. (1925) *Statistical Methods for Research Workers*. (14th ed., 1973) New York: Hafner Press.

Fisher, R.A. (1936) "Has Mendel's Work Been Rediscovered?," *Annals of Science* **1**, 115-137.

Fisher, R.A. (1955) "Introductory Notes on Mendel's Paper," in J.H.Bennett, ed. (1965) Experiments in Plant Hybridization. London: Oliver & Boyd, pp.1-6.

Mawer, S. (1997) *Mendel's Dwarf*. London: Doubleday.

Mendel, G. (1865) "Experiments in Plant-Hybridization," *Verh. naturf. Ver. in Brunn, Ab.* **4**; translated by the Royal Horticultural Soc. and appearing in Bateson (1909), 318-361.

Mendel, G. (1869) "On Hieracium-Hybrids obtained by Artificial Fertilization," *Verh. naturf. Ver. in Brunn, Ab.* **8**; trans. Royal Horticultural Soc. and appearing in Bateson (1909), 362-368.

Nobile, A. (1992) "A Note on Goodness-of-fit χ^2 Statistics," Tech. Report 550, Dept. of Statistics, Carnegie Mellon Univ., Pgh. PA 15213.

Norton, B. and Pearson, E.S. (1976) "A Note on the Background to, and Refereeing of, R.A.Fisher's 1918 Paper 'On the correlation between relatives on the supposition of Mendelian inheritance," *Royal Society of London, Notes and Records* **31**, 151-162.

Novitski, C.E. (1995) "Another Look at Some of Mendel's Results," *J.Heredity* **86**, 62-66.

Orel, V. (1968) "Will the Story on 'too good' Results of Mendel's Data Continue?" BioScience 18, 776-778.

Pearson, K. (1900) "On the Criterion that a given System of Deviations from the Probable in the case of a Correlated System of Variables is such that it can be reasonably supped to have arisen from Random Sampling,"*Phil Magazine* **50**,157-175.

Pearson, K. (1908) "On a Mathematical Theory of Determinantal Inheritance, from suggestions and notes of the late W.F.R.Weldon," *Biometrika* **6**, 80-93.

Piegorsch, W.W. (1983) "The Questions of Fit in the Gregor Mendel Controversy," *Commun. Statist.-Theor. Meth*. **12**, 2289-2304.

Piegorsch, W.W. (1990) "Fisher's Contributions to Genetics and Heredity, with Special Emphasis on the Gregor Mendel Controversy."

Provine, W.B. (1971) *The Origins of Theoretical Population Genetics*. Chicago: U.Chicago Press.

Stephens, M. (1994) "The Results of Gregor Mendel: An analysis, and comparison with the results of other researchers." Thesis submitted for the Diploma in Mathematical Statistics, University of Cambridge. (A copy may be obtained from the author, who may be reached via his home page: www.stats.ox.ac.uk/~stephens/index.html. I thank Dr. Stephens for supplying me with a personal copy.)

Weiling, F. (1971) "Mendel's 'Too Good' Data in Pisum-Experiments," *F. Mendeliana* **6**, 75-77.

Weiling, F. (1986) "What about R.A.Fisher's statement of the 'too good' data of J.G.Mendel's Pisum paper?" *J.Heredity* **77**, 281-283.

Weiling, F. (1989) "Which points are incorrect in R.A.Fisher's statistical conclusion: Mendel's experimental data agree too closely with his expectations?" *Angew. Botanik* **63**, 129-143.

Weldon, W.F.R. (1901) "Mendel's Laws of Alternative Inheritance in Peas," *Biometrika* **1**, 228-253.

Weldon, W.F.R. (1902) "On the Ambiguity of Mendel's Categories," *Biometrika* **2**, 44-55.

Weldon, W.F.R. (1902) "Mr. Bateson's Revisions of Mendel's Theory of Herredity," *Biometrika* **2**, 286-298.

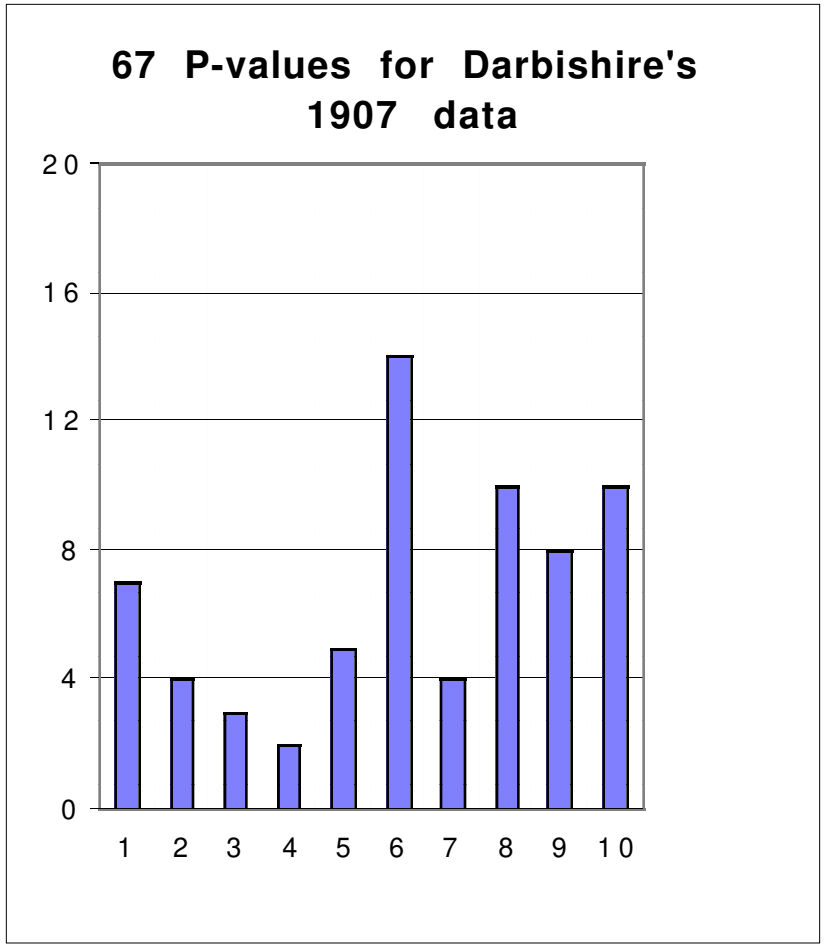


Figure 1

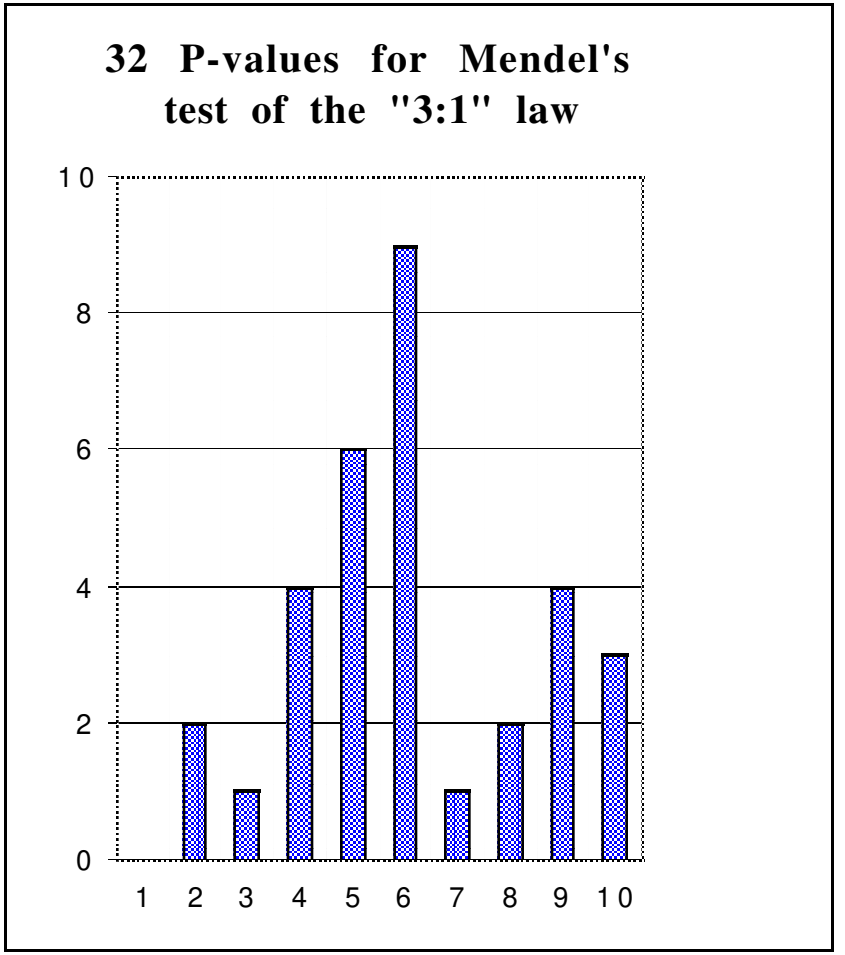
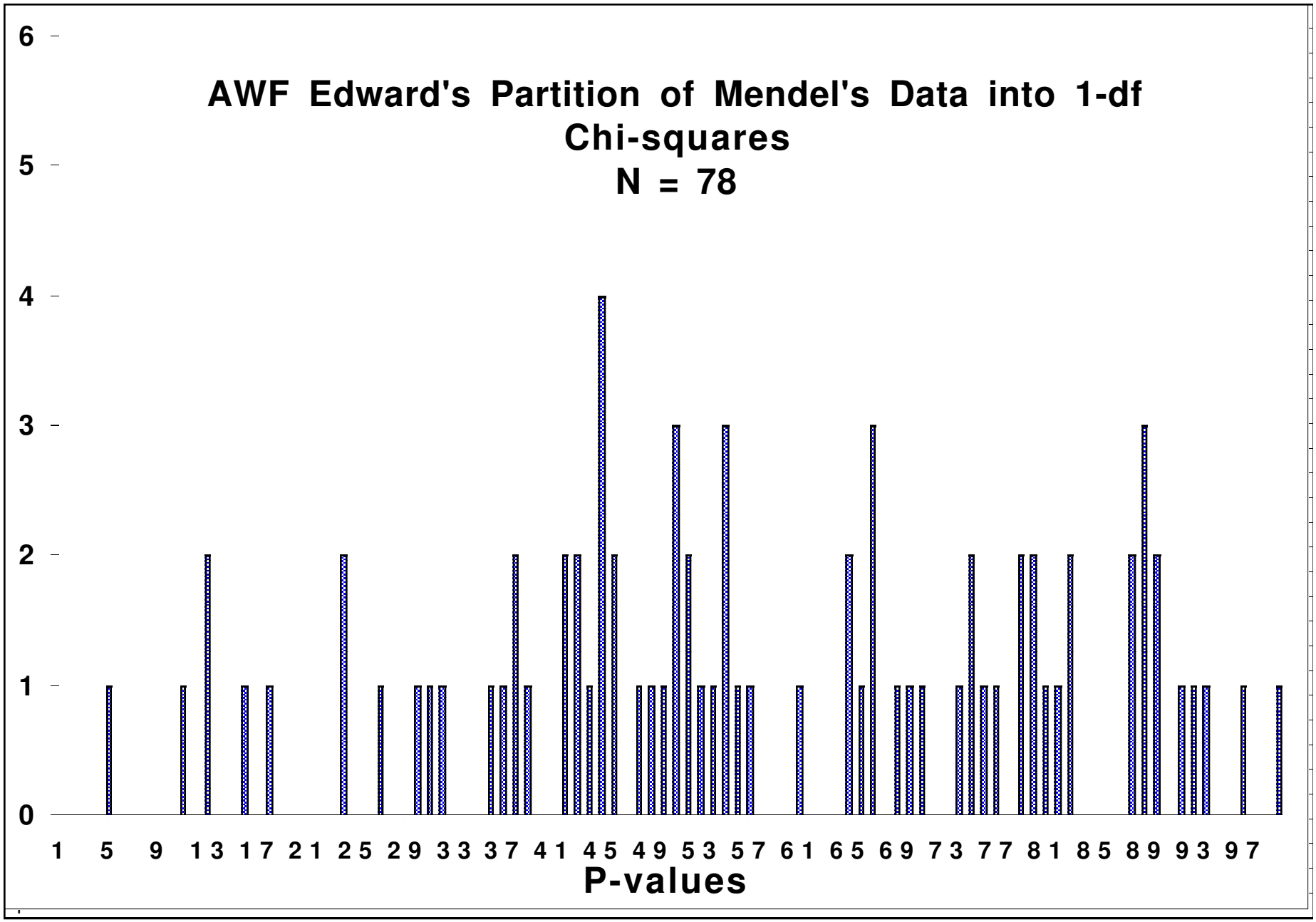
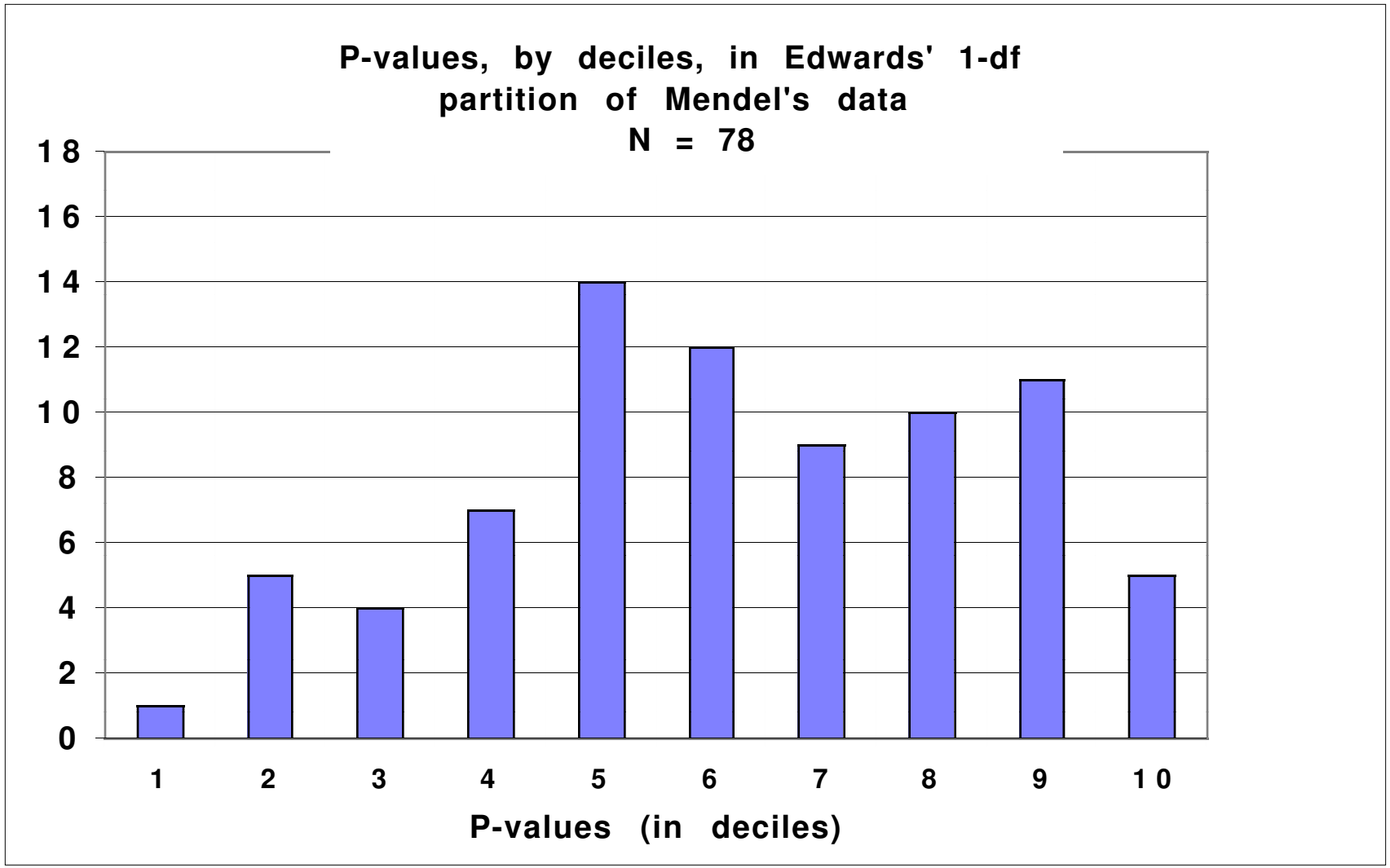
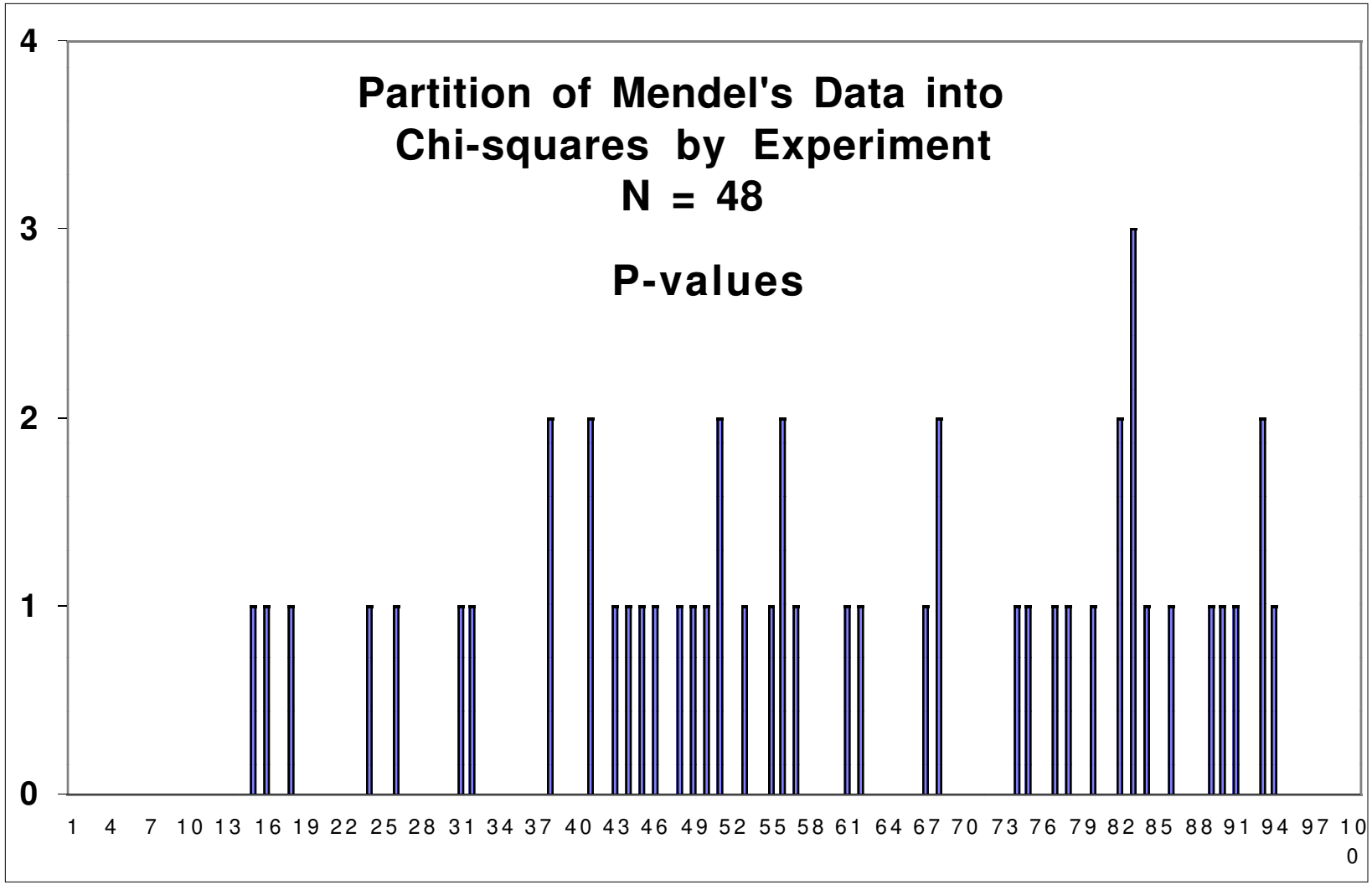
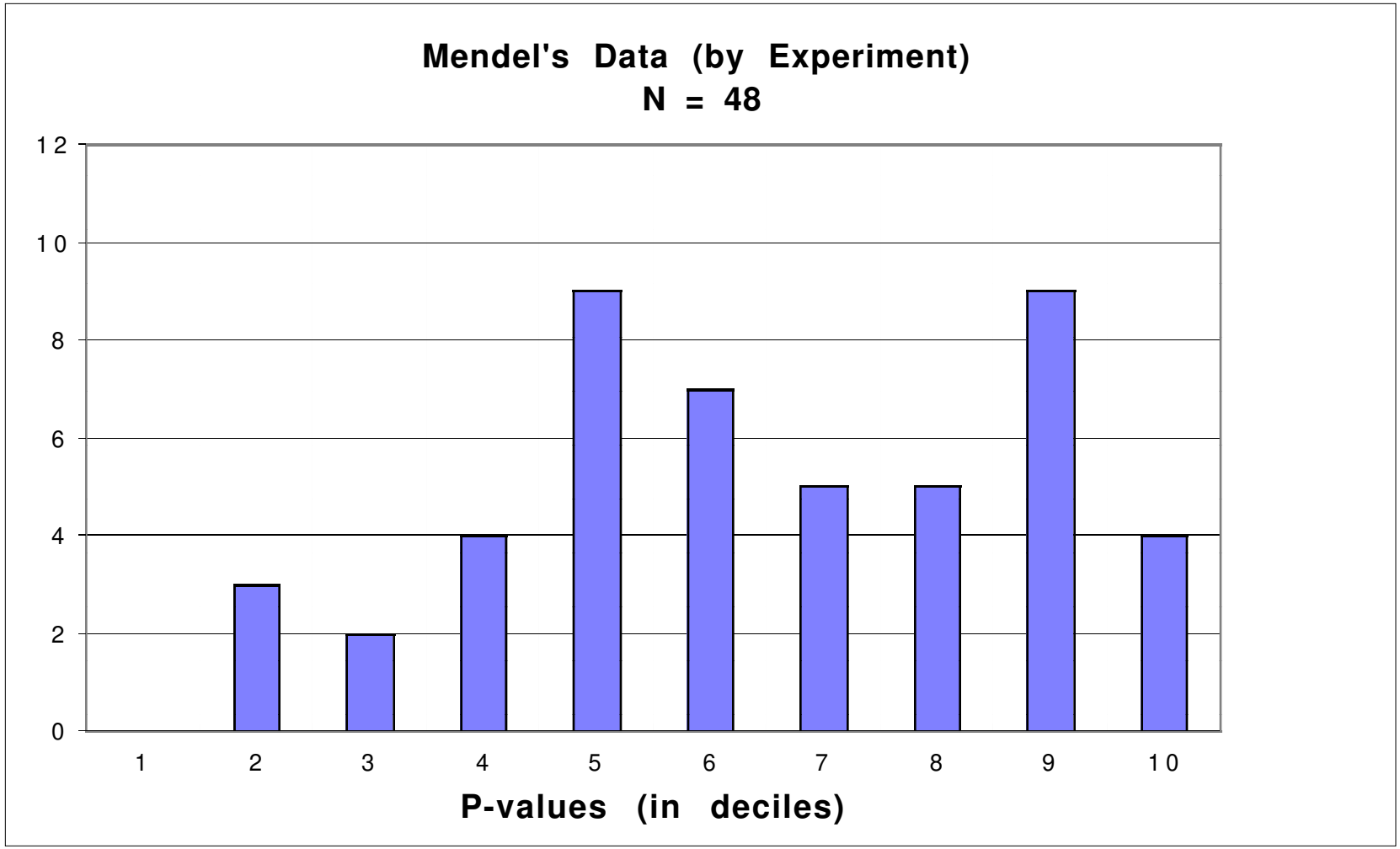


Figure 2


AWF Edward's Partition of Mendel's Data into 1-df Chi-squares
N = 78
6
5
4
3
2
1
0
1 5 9 13 17 21 25 29 33 37 41 45 49 53 57 61 65 69 73 77 81 85 89 93 97
P-values


P-values, by deciles, in Edwards' 1-df partition of Mendel's data
N = 78
18
16
14
12
10
8
6
4
2
0
1
2
3
4
5
6
7
8
9
10
P-values (in deciles)


Partition of Mendel's Data into
Chi-squares by Experiment
N = 48
P-values
4
3
2
1
0
1 4 7 10 13 16 19 22 25 28 31 34 37 40 43 46 49 52 55 58 61 64 67 70 73 76 79 82 85 88 91 94 97 10
0


Mendel's Data (by Experiment)
N = 48
12
10
8
6
4
2
0
1
2
3
4
5
6
7
8
9
10
P-values (in deciles)