Ursinus College

Digital Commons @ Ursinus College

Transforming Instruction in Undergraduate
Mathematics via Primary Historical Sources
(TRIUMPHS)

Fall 2017

Quantifying Certainty: the p-value

Dominic Klyve
Central Washington University, Dominic.Klyve@cwu.edu

Follow this and additional works at: https://digitalcommons.ursinus.edu/triumphs_statistics

Part of the Curriculum and Instruction Commons, Educational Methods Commons, Higher Education Commons, Science and Mathematics Education Commons, and the Statistics and Probability Commons
Click here to let us know how access to this document benefits you.

Quantifying Certainty: the p-value

Dominic Klyve[*]

July 17, 2019

1   Introduction

One of the most important ideas in an introductory class in statistics is that of the p-value. These p-values help us understand how unlikely an outcome is, given an assumption (called a null hypothesis) about how the world works. While the formal theory of p-values arose in the twentieth century [**??**], similar ideas had been around for centuries, and a study of these older ideas can give us insight and understanding into the modern theory of statistics.

This project has three main parts. We shall begin with an idea from outside of the world of statistics called "proof by contradiction," and then consider a probabilistic version of the same argument. We next examine the work of two thinkers who used the basic idea of a p-value long before it was formally defined by Ronald Fisher. Next, we shall consider the common claim used in several fields that we should reject a null hypothesis if $p < 0.05$, and ask why this value is used.

2   Proof by contradiction

Historian of statistics David Bellhouse has characterized eighteenth-century ideas about probability and decision-making as modifications of an old mathematical idea of "proof by contradiction"[1]. This idea goes back more than two thousand years, at least to the Greek philosopher Chrysippus (see **?**), and is used in mathematics today to *prove* or *disprove* a logical statement (that is, to explain using logic why the statement must be true or false). If we have two logical statements, called A and B, we can characterize the three-part structure of this argument as follows:

1. If A is true, then B is true.

2. B is not true.

3. Therefore "A" is not true.

---

[*]Department of Mathematics, Central Washington University, Ellensburg, WA 98926; `dominic.klyve@cwu.edu`.
[1]Students of logic will recognize proof by contradiction as the principle of *modus tollens*

Suppose, for example, that a friend is rolling a die with an unknown number of sides. You predict that it is a six-sided die with sides numbered 1, 2, 3, 4, 5, and 6. If your friend announced that she had just rolled an 8, you would know that your prediction was incorrect.

Task 1   Describe the die-rolling example above by defining logical statements A and B to set up a proof-by-contradiction argument.

3   Proof by the highly improbable

Bellhouse has further suggested that in the eighteenth century, mathematicians and thinkers began using a similar form of reasoning, not to prove statements, but to conclude that they are very likely true. This new kind of thinking can be written as follows:

1. If A is true, then B almost certainly is true.

2. B is not true.

3. Therefore A is almost certainly not true.

Task 2   Suppose your friend with the die above now pulls out a suspicious-looking coin, and proceeds to flip heads 20 times in a row. Would you believe that the coin is "fair"? That is, would you believe that the coin will, in the long run, come up as "heads" half of the time? Why or why not?

Task 3   Write the reasoning you used in the previous Task as a three-part argument like the one given above.

As we shall see, the idea of "proof by the highly improbable" is closely related to the modern idea of p-values studied in statistics classes today. In order to explore this connection, we first turn to the interesting work of an eighteenth-century writer seemed, on the surface, to be interested in something very different than statistics.

4   Boys and girls, births and baptisms

Our story of the early p-value begins with a doctor and satirist named John Arbuthnot. In 1710, Arbuthnot became curious about the sex ratio of births in England. That is, he wanted to know the ratio of male to female births in the country. There were no hospital records for him to use (largely because there were few hospitals, and they were almost never used for births), and the government didn't collect birth information, so he first needed to find a data source. He soon realized that there was a very similar set of information he could use.

Each parish and church that was part of the Church of England, the official church of the United Kingdom, kept a register of all babies christened (or baptized) and all of these records for the City of London had been combined by the Church in the early 1700s. The records were quite sparse, and listed only the number of boys and the number of girls baptized each year.

Task 4   How similar do you think the baptismal records that Arbuthnot collected are to the actual birth numbers? What might cause these to be different?

When Arbuthnot looked at the data gathered, he found an interesting trend. Consider the first 38 years of his data, given below.

| | Christened. | | | Christened. | |
Anno.	Males.	Females.	Anno.	Males.	Females.
1629	5218	4683	1648	3363	3181
30	4858	4457	49	3079	2746
31	4422	4102	50	2890	2722
32	4994	4590	51	3231	2840
33	5158	4839	52	3220	2908
34	5035	4820	53	3196	2959
35	5106	4928	54	3441	3179
36	4917	4605	55	3655	3349
37	4703	4457	56	3668	3382
38	5359	4952	57	3396	3289
39	5366	4784	58	3157	3013
40	5518	5332	59	3209	2781
41	5470	5200	60	3724	3247
42	5460	4910	61	4748	4107
43	4793	4617	62	5216	4803
44	4107	3997	63	5411	4881
45	4047	3919	64	6041	5681
46	3768	3536	65	5114	4858
47	3796	3536	66	4678	4319

Task 5 What do you notice about the number of boys and the number of girls born each year? Can you think of an explanation for this?

What you noticed may have matched Arbuthnot's primary observation – that the number of boys born each year was greater than the number of girls born, and indeed this was the case for all 82 years of data he was able to collect. From this observation, he came to a rather far-reaching conclusion, suggested by the title of his essay, *An Argument for Divine Providence, taken from the Constant Regularity observed in the Births of both Sexes*[2] **?**. Before discussing his conclusion, however, Arbuthnot first wanted to demonstrate just how unlikely this discrepancy was to have occurred by chance.

---

[2]Titles of eighteenth-century books and articles were usually a lot longer than those written today.

⚬⚬
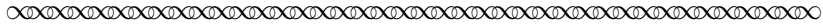
Problem. A lays against B. that every Year there shall be born more Males than Females: To find A's Lot, or the Value of his Expectation.

Let his [A's] Lot be equal to $\frac{1}{2}$ for one year. If he undertakes to do the same thing 82 times running, his Lot will be $\overline{\frac{1}{2}}|^{82}$, which will be easily found by the Table of Logarithms to be $\frac{1}{4\ 8360\ 0000\ 0000\ 0000\ 0000\ 00000}$. But if A wager with B, not only that the Number of Males shall exceed that of Females, every Year, but that this Excess shall happen in a constant Proportion, and the Difference lie within fix'd limits; and this not only for 82 Years, but for Ages of Ages, and not only at *London*, but all over the World (which it is highly probable is the Fact, and designed that every Male may have a Female of the same Country and suitable Age), then A's Chance will be near an infinitely small Quantity, at least less than any assignable fraction. From whence it flows, that it is Art, not Chance, that governs.

⚬⚬
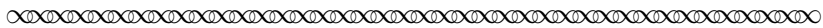
(Note that in the source above, to "lay against" means to bet against, and that the line above the fraction $\frac{1}{2}$ is the equivalent to putting parentheses around the fraction today.)

Task 6 Do you agree with the mathematics of Arbuthnot's calculation of $\left(\frac{1}{2}\right)^{82}$? If not, how might you explain the different answers?

Task 7 Arbuthnot concluded that the difference in the number of births of boys and girls could not be due to chance. Do you agree? Why or why not?

Even today, scholars debate about how to interpret statistics. Arbuthnot's own interpretation of this difference is interesting: he wanted to use the differences in the number of births by sex to make an argument for both the existence of God and for God's involvement in the world. To do this, he needed to explain why more boys being born than girls was good for humanity.

⚬⚬
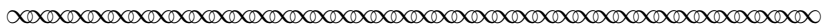
We must observe that the external Accidents to which Males are subject (who must seek their Food with danger) make a great havock of them, and that this loss exceeds far that of the other Sex occasioned by Diseases incident to it, as Experience convinces us. To repair that Loss, provident Nature, by the Disposal of its wise Creator, brings forth more Males than Females; and that in almost a constant proportion.

⚬⚬
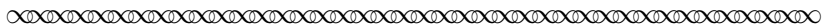
Task 8 Restate and summarize Arbuthnot's explanation using more modern terms.

Although he didn't state it directly, Arbunot seems to have used the type of reasoning we described above in the "Proof by the Highly Improbable" section.

Let's try to be restate his statement and conclusion more explicitly, using the three-step argument above.

<div style="border: 1px solid black; display: inline-block; padding: 2px;">**Task 9**</div>

- First, write Arbuthnot's main premise as an "if A, then almost-certainly B" statement.
- Next, write the contradiction (the "not B" step).
- Finally, write Arthnot's conclusion.
- Does the structure of his argument in fact match that of the "Proof by the Highly Improbable" above?

5 Stating a null hypothesis

One of the most important parts of your work in Task 9 was identifying the statement we have been calling A. Not only does this statement set up the "if-then" structure of the argument, but it is this statement that we eventually reject ("...therefore, not A"). Today we call statement A, the "null hypothesis," and good statisticians know that stating the null hypothesis carefully is a crucial step in statistical reasoning. It's the first step in an argument, a temporary claim which we may reject if we have good reason to do so.

Because it plays such an important role in statistical reasoning, it is crucial to state a null hypothesis as precisely as possible. Temporarily assuming that the same number of girls and boys are born each year, Arbuthnot's null hypothesis might be stated as follows:

Null hypothesis[3]: the probability that more boys are born than girls in any year is $\frac{1}{2}$.

5.0.1 Modifying the null hypothesis

It's worth noting that even in Arbuthnot's time, not everyone was convinced by his arguments. Nicholas Bernoulli calculated that if the probability of a male birth were just slightly higher than $\frac{1}{2}$, say $\frac{18}{35}$, then Arbuthnot's data would not be surprising. If you have learned about binomial distributions at this point in your course, it's a fun exercise to work out the math and to decide whether Bernoulli was correct. For a detailed study of Bernoulli's argument and calculations, see **?**.

6 Choosing a significance threshold

We have examined thus far a pair of eighteenth-century thinkers who used similar reasons about very different questions. Both of them, at least implicitly, formulated a hypothesis, then used data to show that the hypothesis was unlikely, and finally rejected that hypothesis. (Arbuthnot rejected the idea

---

[3]Statisticians like to abbreviate things when they can. Since "hypothesis" starts with "h", and since 0 is sometimes called null, the null hypothesis becomes simply H_0 in many books.

that the probability of a baby being born a boy was 1/2) The name given to this type of reasoning is "hypothesis testing".

The first two steps in hypothesis testing are:

1. identifying a particular claim that we want to test using ideas from probability theory (the null hypothesis), as we did above, and

2. using mathematics to calculate the probability that our data would occur if that claim is correct.

We've worked a lot with the first step already, and so far step two has used quite straightforward mathematics.

The next step is interpreting this value; for many researchers, this means choosing a particular "threshold" value in advance, and deciding that they will reject their null hypothesis (and stop believing it) if the calculated probability is below that value.

In many fields and for many years, researchers have used $p = 0.05$. Long before this standard was accepted, Geroge LeClerc, the Comte de Buffon, had a very different idea in mind. A "comte" is a "count"; King Louis XVI gave LeClerc this title of nobility near the end of his life, and it's now customary to refer to him as "Buffon". Buffon was a prolific author – he wrote an enormous 20-volume work on nature (the *Histoire Naturelle*) in which he discussed everything from the formation of the oceans to the habits of birds and foxes. At the end of one of these volumes, he included an essay on what he called "moral arithmetic"[4].

Buffon was interested in the idea of "moral certainty" (*certitude morale*), where "moral" was not meant to indicate an ethical position, but rather to indicate certainty which would be sufficient for human decision making. He contrasted this to "physical certainty" (*certitude physique*), which he defined as follows:

⊗⊗
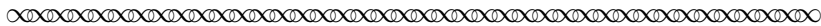
> Physical certainty, that is, the most certain of all certainties, is nevertheless only the almost infinite probability that an effect, an event that never failed to happen, will happen again; for example, because the sun has always risen, it is thenceforth physically certain that it will rise tomorrow.

⊗⊗
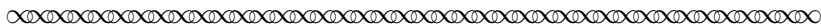
Task 10 How would you explain Buffon's "almost infinite probability" today?

Task 11 Give another example of something which is "physically certain."

---

[4]Essais d'Arithmétique Morale (Essays on Moral Arithmetic) **?**. The translations of Buffon's work are based on the translation in **?**, and have been modified by the author.

Of course, physical certainty is hard to achieve, and in practice, we need a lower threshold before we can decide that we believe something to be true.

Task 12 Suppose that you know that there was one chance in ten million that you would get in a car crash if you drove to the movie theater tonight. Would that stop you from going? What if there was one chance in ten?

Task 13 How unlikely would something have to be before you were willing, in practice, to assume that it won't happen? Come up with a specific value and explain why you chose that.

Buffon himself tried very hard to come up with a value of moral certainty which he could use in practice. He finally settled on the following:

∞∞∞
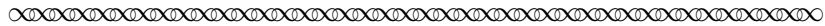
After having reflected on it, I have thought that of all the possible moral probabilities, the one that most affects man in general is the fear of death . . . . I seek therefore for what is actually the probability that a man who is doing well, and consequently has no fear of death, dies nevertheless in the twenty-four hours: consulting the Mortality Tables, I see one can deduce that there are only ten thousand one hundred eighty-nine to bet against one, that a fifty-six year old man will live more than a day. Now as any man of that age, when reason has attained its full maturity and the experience all its force, nevertheless has no fear of death in the twenty-four hours. . . ; from this I conclude that any equal or smaller probability must be regarded as zero, since any fear or any hope below ten thousand must not affect us or even occupy for a single moment the heart or the mind.

∞∞∞
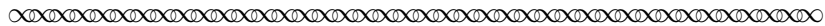
Task 14 What value did Buffon settle on as his threshold for moral certainty, and why?

Task 15 Do you think this is a reasonable value? Why or why not?

7 When to reject a belief: $p = 0.05$

During the century after Buffon wrote, not much progress was made in codifying statistical methods for whether to accept or to reject a belief or claim. Then in 1900 Karl Pearson carefully described the mathematics of a χ^2 ("chi-squared") test in an essay which would launch statistics into the 20th Century. The precise meaning of χ^2 is not important – for now it's just helpful to know that this is a way to measure how closely a set of data matches what a theory would predict.

Twenty-five years later, many of the tools of modern statistics had been developed, and statistician Ronald Fisher decided to make these technical and complex tools available to non-mathematicians. He taught a generation of scientists how to use statistics with his landmark work, *Statistical Methods for Research Workers* **?**. Among other things, this is the book in which he first defined what is now called the "p-value". His first discussion of this in the book appeared in reference to a particular value known as χ^2 .
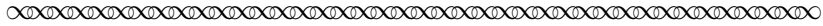
For any value of n, which must be a whole number, the form of distribution of χ^2 was established by Pearson in 1900; it is therefore possible to calculate in what proportion of cases any value of χ^2 will be exceeded. This proportion is represented by P, which is therefore "the probability that χ^2 shall exceed any specified value. To every value of χ^2 there thus corresponds a certain value of P; as χ^2 is increased from 0 to infinity, P diminishes from 1 to 0. Equally, to any value of P in this range there corresponds a certain value of χ^2.
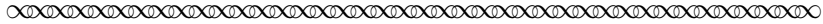
Task 16 Look up (or possibly see your course notes) for a picture of what the distribution of these χ^2 values looks like, and draw a picture to demonstrate what Fisher meant in this passage. Is it true that every value of P corresponds to one value of χ^2, and that every value of χ^2 corresponds to one value of P?

Trying to determine what value of P Fisher believed should make a researcher reject a hypothesis is trickier. Sometimes he seemed to be very clear about what he thought. Consider the following two excerpts, one taken from *Statistical Methods for Research Workers*, and the other from a paper Fisher wrote on agricultural experiments.
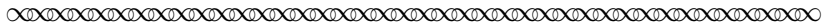
The value for which $P = 0.05$, or 1 in 20, is 1.96 or nearly 2; it is convenient to take this point as a limit in judging whether a deviation ought to be considered significant or not. Deviations exceeding twice the standard deviation are thus formally regarded as significant. Using this criterion we should be led to follow up a false indication only once in 22 trials, even if the statistics were the only guide available. Small effects will still escape notice if the data are insufficiently numerous to bring them out, but no lowering of the standard of significance would meet this difficulty.
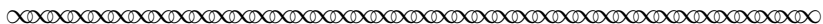
Task 17 Here Fisher seemed to be arguing for $P = 0.05$ as his threshold value for whether a deviation form what is expected should be considered "significant". Describe two of the reasons Fisher gave for choosing this value.

Task 18 Are these reasons strong enough that you believe we should always choose 0.05 as a guide to what is significant? Why or why not?

Task 19 What did Fisher mean when he wrote "Small effects will still escape notice if the data are insufficiently numerous to bring them out"? Describe a case in which a "small effect" might be missed.

Compare the reading about to another time in which Fisher discussed this threshold value **?**:

ⲟⲟⲟ
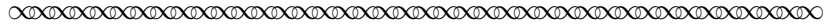
... it is convenient to draw the line at about the level at which we can say: "Either there is something in the treatment, or a coincidence has occurred such as does not occur more than once in twenty trials."... If one in twenty does not seem high enough odds, we may, if we prefer it, draw the line at one in fifty (the 2 per cent point), or one in a hundred (the 1 per cent point). Personally, the writer prefers to set a low standard of significance at the 5 per cent point, and ignore entirely all results which fail to reach this level. A scientific fact should be regarded as experimentally established only if a properly designed experiment rarely fails to give this level of significance.

ⲟⲟ
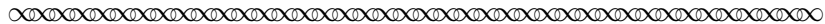
Task 20 How is this similar to (or different than) the previous quotation we read from Fisher's work?

In some ways, trying to find the value that Fisher used is doomed to fail, as he argued repeatedly throughout his life that there is no absolute value which would be appropriate to use in all cases. Gerard Dallal has explained some of the confusion around the idea of P values, writing "Part of the reason for the apparent inconsistency is the way Fisher viewed P values. When [other statisticians writing at the same time] Neyman and Pearson proposed using P values as absolute cutoffs in their style of fixed-level testing, Fisher disagreed strenuously. Fisher viewed P values more as measures of the evidence against a hypotheses, as reflected in [this] quotation from Fisher (1956, p 41-42)" [**?**, Note 31]

ⲟⲟ
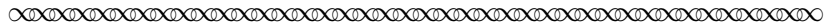
The attempts that have been made to explain the cogency of tests of significance in scientific research, by reference to hypothetical frequencies of possible statements, based on them, being right or wrong, thus seem to miss the essential nature of such tests. A man who "rejects" a hypothesis provisionally, as a matter of habitual practice, when the significance is at the 1% level or higher, will certainly be mistaken in not more than 1% of such decisions. For when the hypothesis is correct he will be mistaken in just 1% of these cases, and when it is incorrect he will never be mistaken in rejection. This inequality statement can therefore be made. However, the calculation is absurdly academic, for in fact no scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas. Further, the calculation is based solely on a hypothesis, which, in the light of the evidence, is often not believed to be true at all, so that the actual probability of erroneous decision, supposing such a phrase to have any meaning, may be much less than the frequency specifying the level of significance.

ⲟⲟ
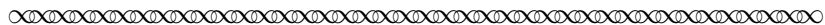
Task 21   Explain Fisher's argument that if a researcher only rejects a hypothesis if $p < 0.01$ will be "mistaken in not more than 1% of such decisions."

Task 22   If a researcher chooses a very high probability for p (say $p = 0.2$), and uses it every time to decide which hypotheses to reject, explain what the negative consequences of this would be.

Task 23   If a researcher chooses a very low probability for p (say $p = 0.001$), and uses it every time to decide which hypotheses to reject, explain what the negative consequences of this would be.

Task 24   What would you now recommend to a researcher who asks you what value of p she should choose for her own research?

Instructor Notes for "Quantifying Certainty: the p-value"
Dominic Klyve[5]
March 27, 2018

This set of notes accompanies the Primary Source Project "Quantifying Certainty: the p-value" written as part of the TRIUMPHS project (see end of notes for details).

PSP Content: Topics and Goals

This Primary Source Project (PSP) has a few goals. The first is to provide students with an intuitive introduction to hypothesis testing and p-values through reading the work of some eighteenth-century thinkers (John Arbuthnot and the Compte de Buffon) who used similar approaches to answer questions which arose naturally in other contexts. Arbuthnot used a line of reasoning very similar to to a modern hypothesis test, and as students work to make his argument explicit, they can discover for themselves the necessary components of hypothesis testing. Buffon, on the other hand, is less convincing, and made an argument that doesn't match the modern methodology as well. As the students discover the strengths and weaknesses of both arguments, they develop an intuitive understanding of modern approaches to hypothesis testing.

The second goal is to let students wrestle with what it means to reject a hypothesis, and with the tricky question of what the threshold value should be to do so. They read another argument of Buffon in which he tries, possibly for the first time in Western history, to find a threshold for "certainty", and then several excerpts from the writing of Ronald Fisher in which the statistician describes the reason that $p = 0.05$ may be a good idea, while reminding readers that the choice is a bit arbitrary.

This PSP is designed to be used in an introductory statistics class (see below for details), but much of it would likely also work well in a quantitative reasoning or "baby stats" class. It's best used at the beginning of the chapter on hypothesis testing. Rather than giving the standard lecture on the thought process behind hypothesis testing, the PSP (I hope) allows students to develop a rigorous understanding of the topic, and thus to begin to use it more easily and accurately.

Student Prerequisites

This project has almost no formal prerequisites. Indeed, as it offers an introduction to hypothesis testing beginning with an intuitive approach to the basic ideas of the field, it could be used even in a general education, "quantitative reasoning"-style class. Probably the most important requisite skill, besides a willingness to engage with original source texts, is knowledge of the "multiplicative rule" from probability theory, and comfort with calculation at the level of high school algebra.

---

[5]Department of Mathematics, Central Washington University, Ellensburg, WA 98926 `dominic.klyve@cwu.edu`.

Commentary on PSP Design and individual tasks

- **Sections 1–3**: Introduction and Proofs by contradiction / highly improbable (0 class days)

  Sections 2 and 3, though short, provide the basic reasoning behind hypothesis testing. Following an example of David Bellhouse (in an unpublished work), we first introduce proof by contradiction, and then its cousin, "proof by the highly improbable." The goal of the sections is that students will develop mental structures to think about hypothesis testing before being formally introduced to the idea.

- **Section 4**: Boys and girls, births and baptisms (1 class day)

  John Arbuthnot's work on baptismal records and the predisposition of human births to be male has been repeatedly cited as the earliest example of hypothesis testing. Starting with the observation that more boys were born (in truth, baptized) than girls each year in London for 82 consecutive years, Arbuthnot reasoned that this was so improbable that it could not be due to chance. Arbuthnot doesn't explicitly give his null hypothesis, but his calculation clearly shows that he has one in mind – namely, that the probablity that more boys are born than girls in a given year is 1/2, and students have the opportunity to turn his slightly fuzzy reasoning into something formal. Arbuthnot's conclusion that the difference in birth rates by gender must be due to the intervention of God, though amusing to many students, provides a useful opportunity to reflect on the difference between what precisely a hypothesis test tells us, and the conclusions that are often drawn from it.

- **Section 5**: Stating a Null Hypothesis

  At this point the students have used the notion of a null hypothesis at least twice, without giving it a name, or indeed without a lot of attention drawn to the notion at all. In this section we make the notion explicit, and then examine another early author (the Compte de Buffon) to use these ideas to calculate a probability of how certain he was of something (in this case, that the sun would rise tomorrow). Buffon's final conclusion is bizarre (and, I would argue, simply incorrect), and students will thus have an opportunity to wrestle with both a good and a poor example of drawing conclusions using methods of hypothesis testing from original sources.

- **Sections 6 and 7**: Significance thresholds, and rejecting null hypotheses.

  An important part of modern hypothesis testing is deciding when a p-value is low enough that we should reject a null hypothesis. In some introductory courses this question is swept under the "$p = 0.05$" rug, but the idea is worth considering closely. These sections start with (again) the work of Buffon, as he presents a clever argument for how unlikely an event would have to be before he would be "morally certain" that it wouldn't happen. (I believe that this is the first time in history that an author seeks a value for a "significance level.")

  After this, the project covers some of the work of Ronald Fisher as he proposes some suggestions for a good significance level, including a clear statement of why we now often use $p = 0.05$.

Suggestions for Classroom Implementation

I give this project to students after a discussion of sampling distributions, but before hypothesis testing. Indeed, the project assumes that students have not seen hypothesis testing formally described.

The PSP includes several open-ended discussion questions, and lends itself well to group work. I suggest assigning groups of three students (or letting students choose their own, as your classroom culture warrants). The schedule given below is based on a 50-minute class period.

Sample Implementation Schedule

- **Day 0:** I give a brief (5-minute) introduction to the project, and tell students explicitly that we'll be spending class time on this project for the next few days, to help them get a strong understanding of some concepts which have been tricky for past students to learn. Assign Sections 1–3. (Students should read the sections and complete Tasks 1–3 as homework.)

- **Day 1:** Working in groups, students work through Section 4, with the goal of finishing Tasks 4–9. Unfinished tasks should be completed as homework.

  Homework: Any unfinished tasks through Task 9. Long sections of reading are best done individually, and not in class. Students should read the first part of Section 5, and complete Tasks 10 and 11.

- **Day 2:** In groups, students discuss their answers to Tasks 10 and 11, and move on to complete Section 5 (Tasks 10–14.) Groups then start Section 6 (the reading goes quickly) and complete Tasks 15–18.

  Homework: Any unfinished tasks through Task 18. Read the rest of Section 6 and complete Tasks 19 and 20. Read the beginning of Secton 6 and complete Task 21.

- **Day 3:** Groups should complete as much of the rest of the project as possible. Some groups may finish everything. Assign the rest of the PSP as homework.

- **Days 4–n.** I have found, when using hypothesis testing throughout the rest of the course, that referring back to the arguments of Arbuthnot or Fisher help students understand some of the more theoretical ideas. Use the fact that they all now have a shared background in hypothesis testing to your advantage!

Commentary on Selected Student Tasks

- Task 5: Students (and some instructors!) often ask why more boys are born than girls. The difference is real, still true today – many sources cite 51.9% as a good estimate for the proportion of males among live births – and is hard to explain. The curious instructor may want to look at Wikipedia's "Human Sex Ratio" page for a fairly comprehensive summary of the major theories and ideas in his area. For the purpose of the PSP, any reasonable guess by students is sufficient; it's important only that they engage with the question.

- Tasks 12 and 13: Anecdotal evidence suggests that students will get quite involved in these questions – expect discussion!

Connections to other Primary Source Projects

The following additional projects based on primary sources are also freely available for use in teaching standard topics in an introductory statistics course.

- F37: Representing and Interpreting Data from Playfair, by Diana White, River Bond, Joshua Eastes, and Negar Janani
- M02: Regression to the Mean, by Dominic Klyve
- M27: Seeing and Understanding Data, by Beverly Wood and Charlotte Bolch
- M31: Playfair's Introduction of Bar Graphs and Pie Charts to Represent Data, by Diana White, River Bond, Joshua Eastes, and Negar Janani

Acknowledgments