Quantifying Certainty: the p-value

Dominic Klyve*

August 19, 2017

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 [Pearson, 1900, Fisher, 1925], 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 accept 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 Lodder [2013]), 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 not 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 who himself seemed to have no interest in statistics at all.

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 he collected? 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?b

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] Arbuthnot [1700]. Before discussing his conclusion, however, Arbuthnot first wanted to demonstrate just how unlikely this discrepancy was to have occurred by chance.

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.

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

> 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 $\frac{1}{4\ 8360\ 0000\ 00000\ 00000\ 00000\ 0000}$. 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 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.

Task 9

- 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 hypotehsis

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 Shoesmith [1985].

5.1 Buffon and sunrise

Another early thinker who was interested in using observations to estimate how likely a statement (a null-hypothesis) may be was Geroge LeClerc, the Comte de Buffon. 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].

One of the questions Buffon tackled in this essay is something you may never have wondered about – the probability that the sun will rise tomorrow. Buffon used a popular idea in the philosophy of his day: a person who knew nothing and who had no pre-conceived ideas, who appeared fully-formed in the world one day. Buffon wrote:

[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.

[4]Essais d'Arithmetique Morale (Essays on Moral Arithmetic) GLC [1777]. The translations of Buffon's work are based on the translation in Hey et al. [2010], and have been modified by the author.

Imagine him struck for the first time by the appearance of the sun; he sees it shine from high in the skies, then go down and finally disappear; what can he conclude? Nothing, except that he saw the sun, that he saw it follow a certain route, and that he no longer sees it. But this star would reappear and disappear again on the next day. This second sight is a first experience, which must produce in him the hope to see the sun again, and he begins to believe that it could return – nevertheless he is very much in doubt. the sun would reappear again; this third sight is a second experience which reduces his doubt as much as it increases the probability of a third return. A third experience increases the probability to the point that he no longer doubts that the sun will return a fourth time; and finally when he will have seen this star of light appear and disappear regularly ten, twenty, a hundred times, he will believe to be certain that he will see it always appear and disappear and to move the same way.

The more similar observations he will have, the greater will be the certainty to see the sun rise the next day; every observation, that is, every day, produces a probability, and the sum of these probabilities together, as it is very great, gives the physical certainty; one will therefore always be able to express this certainty by numbers, dating back to the origin of the time of our experience and it will be the same for all the others effects of Nature; for example, if one wants to reduce here the age of the world and of our experience to six thousand years, the sun has risen for us only 2 million 190 thousand times, and as to date back to the second day that it rose, the probabilities of rising the next day increase as the sequence $1, 2, 4, 8, 16, 32, 64 \ldots$ or 2^{n-1} . One will have (where the natural sequence of the numbers, n is equal to 2,190,000), I say, $2^{n-1} = 2^{2,189,999}$; this already is such a prodigious number that we ourselves cannot form an idea, and it is by this reason that one must look at the physical certainty as composed from an immensity of probabilities; since by moving back the creation date by only two thousand years, this immensity of probabilities becomes $2^{2,000}$ times more than $2^{2,189,999}$.

Task 10 Write one question and one comment you have about this passage.

Task 11 Why do you think Buffon wanted to imagine reducing the age of the world to only 6000 years?

Buffon isn't very explicit about the mathematics that he used. Let's see if we can find it more explicitly. His primary mathematical claim seems to be that the probability of the sun not rising after n days is $1/2^{n-1}$.

Task 12 Why did Buffon use $1/2^{n-1}$ and not $1/2^n$ as the probability of the sun not rising again after seeing it for n days?

Task 13 State the null hypothesis that Buffon seems to have used. You might start it this way: "Let p be the probability that the sun will rise on any given day. Then $p = \ldots$."

Task 14 Assuming your null hypothesis, what is the probability of the sun rising 11 consecutive days? Using our modern idea of hypothesis testing, does it follow that the probability that the sun will not rise the next day is $1/2^n$?

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 that the probability of a baby being born a boy was 1/2, and Buffon similarly rejected that idea that the probability that the sun will rise tomorrow is 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, Buffon had a very different idea in mind. 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 15 How would you explain Buffon's "almost infinite probability" today?

Task 16 Give another example of something which is "physically certain."

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

Task 17 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 18 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, and I felt from that time that any fear or any hope, whose probability would be equal to the one that produces the fear of death, can morally be taken as the unit to which one must relate the measure of the other fears; and I relate to the same even the one of hopes, since there is no difference between hope and fear, other than from positive to negative; and the probabilities of both must be measured in the same way. 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, although there is only ten thousand one hundred eightynine to bet against one that he will die in this short interval of time; 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 19 What value did Buffon settle on as his threshold for moral certainty, and why?

Task 20 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* Fisher [1925]. 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 21 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 22 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 23 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 24 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 Fisher and Wishart [1930]:

> ... 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.

Task 25 How is this similar to (or different than) the first 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)" [Dallal et al., 1999, Note 31]

> 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.

Task 26 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 27 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 28 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 29 What would you now recommend to a researcher who asks you what value of p she should choose for her own research?

References

J Arbuthnot. An argument for divine providence taken from the constant regularity observed in the births of both sexes. *Philosophical Transactions of the Royal Society of London*, 27, 1700.

Gerard V Dallal et al. *The little handbook of statistical practice*. Gerard V. Dallal, 1999.

Ronald Aylmer Fisher. *Statistical methods for research workers*. Genesis Publishing Pvt Ltd, 1925.

Ronald Aylmer Fisher and John Wishart. *The arrangement of field experiments and the statistical reduction of the results*. Number 10. HM Stationery Office, 1930.

Buffon GLC. Essai darithmétique morale. *Oeuvres Complètes de Buffon*, 3:338–405, 1777.

John D Hey, Tibor M Neugebauer, and Carmen M Pasca. Georges-louis leclerc de buffons essays on moral arithmetic. In *The Selten School of Behavioral Economics*, pages 245–282. Springer, 2010.

Jerry Lodder. Deduction through the ages: A history of truth. *MAA Convergence*, 2013. URL `https://www.maa.org/press/periodicals/convergence/deduction-through-the-ages-a-history-of-truth`.

Karl Pearson. 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 supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50 (302):157–175, 1900.

Eddie Shoesmith. Nicholas bernoulli and the argument for divine providence. *International Statistical Review/Revue Internationale de Statistique*, pages 255–259, 1985.

Instructor Notes

Goals

Background

Prerequisite knowledge

Acknowledgments

The development of this student project has been partially supported by the Transforming Instruction in Undergraduate Mathematics via Primary Historical Sources (TRIUMPHS) Program with funding from the National Science Foundation's Improving Undergraduate STEM Education Program under grant number 1523494. Any opinions, findings, and conclusions or recommendations expressed in this project are those of the author and do not necessarily represent the views of the National Science Foundation. For more information about TRIUMPHS, visit `http://webpages.ursinus.edu/nscoville/TRIUMPHS.html`.
.