NUCLEAR CRISIS OUTCOMES: WINNING, UNCERTAINTY, AND THE NUCLEAR BALANCE

National Security Report

Kelly Rooker | James Scouras

APL
JOHNS HOPKINS
APPLIED PHYSICS LABORATORY

Image on the left: Photograph of MRBM Launch Site No. 2 in San Cristobal, Cuba, November 1, 1962, from Collection JFK-5047: Department of Defense Cuban Missile Crisis Briefing Materials, 1962–1963. Retrieved from the National Archives and Records Administration catalog, https://catalog.archives.gov/id/193933

NUCLEAR CRISIS OUTCOMES:

Winning, Uncertainty, and the Nuclear Balance

Kelly Rooker

James Scouras

APL
JOHNS HOPKINS
APPLIED PHYSICS LABORATORY

NSAD-R-18-030

Contents

Figures

Tables

Summary

The importance of the US nuclear balance with respect to its principal adversary has been the subject of intense but unresolved debate since the Soviet Union first acquired nuclear weapons some seven decades ago. One critical policy-relevant question for the United States is whether nuclear-superior states are more likely than nuclear-inferior states to achieve their objectives ("win") in nuclear crises. While scholars have studied past nuclear crises to answer this and related questions, there remains great uncertainty in estimates of the probability of winning. Our work is motivated by the importance of appropriately quantifying such uncertainty.

In particular, we show how the data on nuclear crisis outcomes can be represented using a binomial distribution. Binomial distribution-based methods are broadly applicable across many disciplines. In cases where data can be modeled using a binomial distribution, an estimate for the binomial distribution parameter is often produced. However, the uncertainty surrounding this estimate is only sometimes reported, in part because of the opacity of the various methods available. Failing to appropriately analyze uncertainty can lead to incomplete, if not erroneous, conclusions. Here, we explore both Bayesian and frequentist methods for quantifying uncertainty in the binomial distribution parameter, and discuss each method's various advantages and limitations. We use nuclear crisis outcome data as a tangible example to compare the various methods explored.

As background, we first describe Bernoulli trials and the binomial distribution in general. We then discuss the major differences between Bayesian and frequentist approaches for quantifying uncertainty in the binomial distribution parameter and the basic assumptions of each. Next, we describe selected Bayesian and frequentist approaches in more detail and illustrate applications of these methods. For the Bayesian approach, this includes comparing various possible priors, understanding the impact of the choice of prior, and conducting Bayesian point estimation. For the frequentist approach, this includes maximum likelihood estimation and Clopper–Pearson intervals. We then produce and compare results, obtained using both Bayesian and frequentist methods, for the example of nuclear crisis outcomes. Finally, we discuss the assumptions and limitations of these approaches and compare this focus on uncertainty with hypothesis testing.

We demonstrate that there are several methods to quantify uncertainty in the binomial distribution parameter and show that there is little difference between using the frequentist method of Clopper–Pearson intervals and Bayesian methods using uninformative priors. We also show how the choice of an informative prior can significantly affect results when one uses Bayesian methods. More generally, the work presented here supports the view that overreliance on the combination of hypothesis testing and regression analysis can be highly problematic. In particular, we highlight that methods of estimation—for example, those presented here—are an important supplement, or even a valid alternative, to hypothesis testing and regression analysis.

The binomial distribution is widely used because of both its simplicity and applicability to real-world phenomena. The example that motivates our study is the question of whether nuclear-superior states are more likely than nuclear-inferior states to achieve their primary goals ("win") in nuclear crises. Matthew Kroenig was the first to comprehensively examine this question empirically. Using a statistical analysis of adversarial pairs of states (termed *dyads*) in historical nuclear crises, Kroenig showed that states enjoying nuclear superiority over their opponents are more likely to achieve their primary objectives.[1]

In Kroenig's analyses, the outcomes of nuclear crises follow a binomial distribution (although Kroenig does not explicitly recognize the binomial in his work). Kroenig's dyads consist of some state A and an opposing state B, where, at the time of the crisis, one state had nuclear superiority over the other state. Nuclear crises with more than two states involved were broken up into component dyads.

Note that each dyad actually consists of two directed dyads, denoted A → B and B → A. The first state in each directed dyad, either the nuclear-superior or the nuclear-inferior state, has some principal objective it wishes to achieve in the crisis, while the second state in each directed dyad presents the main obstacle to the first state achieving this objective. Indeed, each first state of the directed dyad will have only two possible outcomes: win (achieve its primary objective) or lose (fail to achieve its primary objective). The probabilities of winning for nuclear-superior and nuclear-inferior states were each assumed to be constant between trials (i.e., in the basic framework, no other factors were considered as affecting the outcome of a nuclear crisis).

Twenty historical nuclear crises were considered, producing a total of twenty-six dyads, or fifty-two directed dyads. Note these fifty-two come from the twenty-six directed dyads where the nuclear-superior state is listed first and the twenty-six where the nuclear-inferior state is listed first. The state listed first in each directed dyad is the basis for determining a win versus a loss. Unless the primary objectives of the two states are diametrically opposed, both states are either able to both win (achieve their objectives) or both lose (not achieve their objectives). Because of this, the probability of the nuclear-superior state winning is not simply one minus the probability of the nuclear-inferior state winning (since one state winning does not always equate to the other state losing, and vice versa).

The analysis methods Kroenig uses are quite common in the social sciences: hypothesis testing and regression analysis. However, for the analysis of nuclear crisis outcomes, these methods do not specifically answer questions about the uncertainty in the respective probabilities of nuclear-superior and nuclear-inferior states winning in a nuclear crisis. Modeling the problem explicitly using a binomial distribution can help answer some of these questions, which is precisely what we do here.

[1] Matthew Kroenig, "Nuclear Superiority and the Balance of Resolve: Explaining Nuclear Crisis Outcomes," *International Organization* 67, no. 1 (2013): 141–171.

In general, there can be many problems with an exclusive reliance on hypothesis testing and regression analysis.[2] Because of this, statisticians are increasingly supplementing, or even replacing, p-value-based approaches with other approaches—those that emphasize estimation over hypothesis testing and regression analysis. The focus of this report is on two such methods, confidence and credible intervals, but other methods of estimation exist, including likelihood ratios and Bayes factors, among others.[3]

This report is directed at those in the social science community with a limited background in statistics, but researchers with a higher level of mathematical sophistication could also find this report helpful. We are not inventing new methods, but rather showing how certain methods, which have been used on myriad other problems, also have applicability here. In particular, we formulate the probability of either a nuclear-superior or a nuclear-inferior state winning in a nuclear crisis as a binomial distribution. We are interested in not just an estimate for this probability of winning, but also the level of uncertainty around the estimate. Our aim is to explain all of these methods well, demonstrate how they can be used on the problem at hand, and illustrate how they are able to answer the questions of most concern to policy makers.

All of our methods' explanations are general, so that these same methods can be applied to other questions in social science research and not just this question of nuclear superiority influencing the probability of winning in a nuclear crisis. Still, the examples used here will all be based on twenty-six trials representing the twenty-six dyads from twenty nuclear crises. For states with nuclear superiority, there were fourteen victories (i.e., achievement of goals) out of those twenty-six dyads. For states with nuclear inferiority, there were only four victories out of those twenty-six dyads. Recall that since "winning" is defined as a state achieving its primary objective, it is possible for one, both, or neither states in the dyad to "win" any nuclear crisis. See Table 1 for a summary of the nuclear crisis outcome data from Kroenig.[4]

Table 1. Summary of Nuclear Crisis Outcome Data (From Kroenig)

Relative Nuclear Arsenal Size	Nuclear Crisis Outcome		Total
	Win	Lose	
Superior	14 (54%)	12 (46%)	26
Inferior	4 (15%)	22 (85%)	26

[2] Lisa L. Harlow, Stanley A. Mulaik, and James H. Steiger, eds., *What If There Were No Significance Tests?*, 2nd ed. (New York: Routledge, 2016).

[3] Ronald L. Wasserstein and Nicole A. Lazar, "The ASA's Statement on p-Values: Context, Process, and Purpose," *The American Statistician* 70, no. 2 (2016): 129–133.

[4] Kroenig, "Nuclear Superiority."

While odds ratios may seem like a good way to analyze these data, we chose not to use odds ratios here. The odds ratio in our case would be the odds of winning in a nuclear crisis given that a state is nuclear superior, compared to the odds of winning in a nuclear crisis given that a state is nuclear inferior. Odds ratios are well used in mathematics and statistics but lack the intuitive meaning of a simple probability; odds compare two probabilities, but then odds ratios compare two odds. Because of this, we chose to focus on the probabilities of winning for nuclear-superior and nuclear-inferior states, and the uncertainties surrounding these probabilities. Doing so is both more intuitive for a nonquantitative person and makes the analysis more accessible for use by policy makers.

Problem Setup

Definitions

An event is random if its outcome is uncertain. Common examples of a random event include the flip of a coin or the roll of a die. The probability of any outcome of a random event is the chance of that outcome occurring (i.e., the proportion of times the outcome occurs given a large number of repetitions). For example, the probability of a fair coin landing heads up on a coin flip is 50%. Any two events are independent provided the probabilities of one event's outcomes are unaffected by any outcome of the other event occurring (in other words, any subsequent outcome is not influenced by any prior outcome). For example, the outcome of one coin flip will not affect the outcome of a second coin flip, meaning two regular coin flips are independent events.

A random variable is any variable that can take only values that are outcomes of a random event. Random variables can be either continuous or discrete.[5] An example of a continuous random variable is a number drawn randomly from the interval [0, 1]. An example of a discrete random variable is the number rolled on a six-sided die. Note that all possible values of a random variable need not have equal probabilities. For example, using an unfair coin where the probability of flipping heads (outcome = 1) is 0.75 and tails (outcome = 0) is 0.25, the outcome of the coin flip is still a discrete random variable.

A trial is any procedure with well-defined possible outcomes; a random trial is any trial whose outcomes are uncertain. A Bernoulli trial is any random trial with exactly two possible outcomes (often termed *success* and *failure*) and a probability of success that remains the same should that trial be repeated. A binomial experiment is any fixed number of independent Bernoulli trials, where the probability of success is the same in every trial of the experiment.

[5] Harry F. Martz and Ray A. Waller, *Bayesian Reliability Analysis* (New York: John Wiley and Sons, 1982).

A discrete (binomial) random variable can be defined by the total number of successes from a binomial experiment. This random variable will have a binomial distribution.[6]

For example, consider flipping a fair coin ten times in a row. This constitutes a binomial experiment with ten independent Bernoulli trials. Defining success as flipping heads and failure as flipping tails, the probability of success (heads) is 50%. Then the total number of coin flips resulting in heads (out of those ten trials) is a binomial random variable.

Using Kroenig's example, the fixed number of trials is the total number (twenty-six, for each of nuclear-superior and nuclear-inferior states) of directed state dyads involved in the twenty nuclear crises considered.[7] Each such trial can result in only two possible outcomes for the first country in each directed dyad: success (the country achieves its objectives) or failure (the country does not achieve its objectives). The probability of success is assumed to be the same in every trial, meaning the probabilities of winning for nuclear-superior and nuclear-inferior states are assumed to be constant between trials. Finally, the trials are assumed to be independent since the outcome of any one trial is assumed to not affect the outcome of any other trial. This means that the total numbers of victories for each of the nuclear-superior and nuclear-inferior states (out of the twenty-six total trials for each) are binomial random variables.

Notation

For this binomial experiment, let N be the total number of trials and x be the total number of successes out of those N trials. Let $0 \leq p \leq 1$ be the probability of success in each trial. In the previous example, $N = 10$, $p = 0.5$, and x is the total number of coin flips resulting in heads.

The probability of obtaining exactly x successes out of N trials, with p being the probability of success in any one trial, is given by:

$$\mathbb{P}(\text{exactly } x \text{ successes}) = \frac{N!}{x!(N-x)!} p^x (1-p)^{N-x}, \tag{1}$$

with $\mathbb{P}$ denoting probability.[8] Figure 1 depicts the range of the possible numbers of successes x (and their corresponding probabilities) with $N = 10$, $p = 0.5$.

For this report, we assume that N and x are known, while p is unknown. We are interested in obtaining more information on the uncertainty in this parameter, the probability of success (p), given a specific binomial experiment and its outcome (i.e., N and x).

[6] Oliver C. Ibe, *Fundamentals of Applied Probability and Random Processes*, 2nd ed. (New York: Elsevier, 2014).

[7] Kroenig, "Nuclear Superiority."

[8] Ibe, *Fundamentals*.

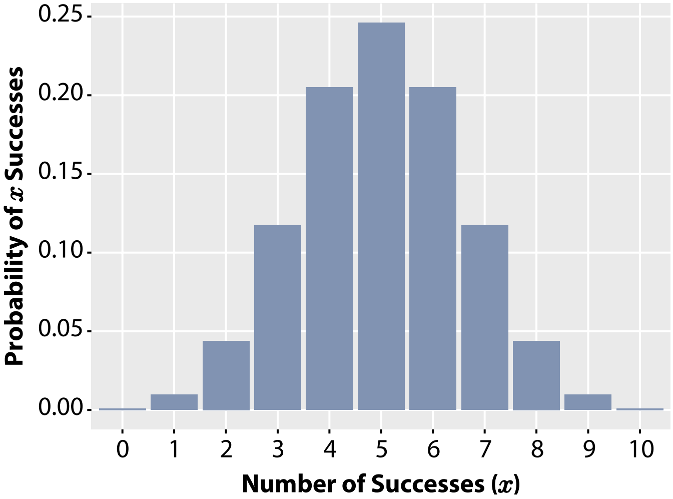


Figure 1. Binomial Distribution for $N = 10$, $p = 0.5$

Bayesian versus Frequentist Approaches

We will approach this problem from both Bayesian and frequentist perspectives. The main differences between Bayesians and frequentists lie in their underlying assumptions.[9] Both Bayesians and frequentists consider the likelihood, which is the probability distribution of the observed data, given unknown parameters.[10] Frequentists rely strongly on the data they have, so they emphasize this likelihood, including the maximum likelihood estimate. Bayesians instead rely on both the data on hand as well as any prior data or information, so they emphasize the likelihood less than frequentists do.

A hallmark of Bayesian thought, not present in frequentist thinking, is the ability to formally include prior information in analyses. Bayesians also view the parameter p as a random variable described by a probability distribution. Indeed, Bayesians assume that p has an associated probability distribution even before any analysis of new data is conducted. This initial distribution is known as the a priori probability distribution (often referred to simply as the prior). An informative prior coming from data from a previous experiment, input from experts, etc., or an uninformative prior meant to capture as little information as possible, could be used. After deciding on a prior, Bayesian inference is accomplished by collecting data and using those data to calculate the likelihood. The prior and likelihood are then used to calculate an updated probability distribution for the parameter. This new distribution, the conditional distribution of the unknown parameter p given the observed

[9] For a review of these differences, see James O. Berger, *Statistical Decision Theory and Bayesian Analysis*, 2nd ed. (New York: Springer, 1985).

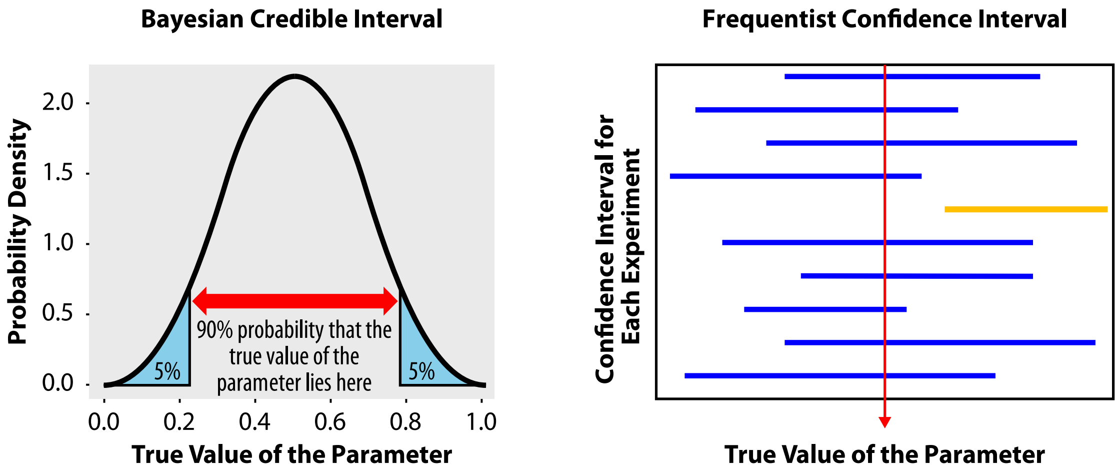
[10] Bradley P. Carlin and Thomas A. Louis, *Bayesian Methods for Data Analysis*, 3rd ed. (New York: CRC Press, 2008).

data, is called the a posteriori probability distribution, or more simply, the posterior. To a Bayesian, all statistical inferences come from the posterior, meaning that no separate theories of estimation, testing, etc. are needed.

A hallmark of the frequentist approach, not present in the Bayesian approach, is to base evaluation on imagining repeated sampling. Whenever only a sample can be taken from a population, frequentists recognize that such a sample is only one out of the many possibilities that sample could have been. In addition, frequentists assume that the parameter p is fixed (i.e., not random), but it is unknown.

These differences also affect Bayesians' and frequentists' respective definitions of credible and confidence intervals, meant to capture uncertainty in p. For Bayesians, a 90% credible interval means that there is a 90% probability that the true value of p lies within that interval.[11] Frequentists instead have confidence intervals, created by imagining repeated sampling or repeated runs of the experiment. Hence for frequentists, a 90% confidence interval is one that, given infinite sampling, captures the true value of the unknown parameter 90% of the time. These differences are illustrated graphically in Figure 2.

Bayesians criticize the frequentist assumption of unknown parameters not being random, the exclusion of prior information, and the unintuitive notion of a confidence interval.



Comparison of the definitions between the Bayesian 90% credible interval (left) and the frequentist 90% confidence interval (right). For the Bayesian, given a prior and the observed data, there is a 90% probability that the true value of the parameter will lie within the credible interval. For the frequentist, there is a 90% probability that when a confidence interval is computed from data of any sample of the population, the true value of the parameter will lie within that interval (the yellow interval depicts the one out of ten shown where it does not).

Figure 2. Comparison of a Bayesian Credible Interval and a Frequentist Confidence Interval

[11] George Casella and Roger L. Berger, *Statistical Inference*, 2nd ed. (New York: Duxbury Press, 2002).

Bayesians view the frequentist confidence interval, defined by imagining an infinite number of samples, as unhelpful since most situations do not offer multiple (much less infinite) samples with which to calculate many different confidence intervals. Instead, there is just one and the goal is to know how that one sample can be used to estimate uncertainty in p. To frequentists, the probability of the true value of the parameter lying inside any frequentist confidence interval is simply 0 or 1, which is not a very informative statement.

Frequentists challenge the Bayesian approach on other grounds. Most fundamentally: how can Bayesians justify the claim that the value of p has a probability distribution? How is the method saying anything real when the results are so dependent on the choice of a prior? And how can Bayesians defend their choice of a prior when the choice is quite subjective and even an uninformative prior provides some information?

Indeed, there is often strife between devout Bayesians and devout frequentists. It is not our goal here to support one side or the other, but rather to explain both approaches and show that there are valid criticisms of each view. We also show, for our problem, that there is much overlap between Bayesian results using an uninformative prior and frequentist results.

Bayesian Approach

This section dives deeper into the Bayesian approach. First, Bayes' theorem is discussed and, in particular, how this theorem applies to priors and posteriors. Then the prior is discussed in more depth, including the two most common uninformative priors, the uniform and the Jeffreys. Problems with each are also discussed, as well as their respective applications to the binomial distribution. Finally, informative priors are discussed, along with the impact of this choice of prior with the binomial distribution.

Connection to Bayes' Theorem

Bayes' theorem relates the probabilities of any two events A and B, where $\mathbb{P}(B) \neq 0$:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A) \cdot \mathbb{P}(A)}{\mathbb{P}(B)}. \tag{2}$$

For example, say you are interested in the probability that you have a certain disease (event A) given that the test for that disease came back positive (event B). This $\mathbb{P}(A|B)$ is an unknown value, but it can be calculated with other information: the probability that the test comes back positive given a person has the disease ($\mathbb{P}(B|A)$), the overall probability of getting a positive test ($\mathbb{P}(B)$), and the overall frequency of the disease in the population ($\mathbb{P}(A)$). Bayes' theorem allows you to use these probabilities to calculate the probability most important to you, $\mathbb{P}(A|B)$.

Bayes' theorem follows from the definition of conditional probability: for any two events A, B, where $\mathbb{P}(A), \mathbb{P}(B) \neq 0$,

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}, \tag{3}$$

and

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}. \tag{4}$$

Combining these two equations, we also know:

$$\mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B) = \mathbb{P}(B|A)\mathbb{P}(A), \tag{5}$$

meaning:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}. \tag{6}$$

This formula can also be applied to probability density functions (PDFs),[12] and in particular:

$$g(p|x) = \frac{f(x|p) \cdot \pi(p)}{m(x)}, \tag{7}$$

for parameter p, observations x (for example, the number of successes observed), posterior PDF $g(p|x)$, likelihood function $f(x|p)$, prior PDF $\pi(p)$, and marginal density function $m(x)$. Note that $m(x)$ serves as a normalization constant, meaning $m(x) = \int f(x|p)\pi(p)dp$, an integral that often needs to be computed numerically.

Hence, Bayes' theorem gives a way to describe probability based on prior, related knowledge, experience, theory, etc. The above is often written more simply as:

$$Posterior\ (g(p|x)) \propto Likelihood\ (f(x|p)) \times Prior\ (\pi(p)). \tag{8}$$

To see where this continuous version of Bayes' theorem comes from, let P be the random variable describing the distribution of parameter $0 \leq p \leq 1$, the probability of the nuclear-superior state winning in a nuclear crisis. Let A be the event that P is in an interval around p with width dp. Let B be the event that the value of the data is x (i.e., given by the marginal density function $m(x)$). Then $\mathbb{P}(A) = \pi(p)dp$, $\mathbb{P}(B) = m(x)$, and $\mathbb{P}(B|A) = f(x|p)$. Hence, using the original Bayes' theorem, we have:

$$\begin{aligned} g(p|x)dp &= \mathbb{P}(A|B) && (9) \\ &= \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)} && (10) \\ &= \frac{f(x|p)\pi(p)dp}{m(x)}, && (11) \end{aligned}$$

[12] Carlin and Louis, *Bayesian Methods*.

meaning:

$$g(p|x) = \frac{f(x|p)\pi(p)}{m(x)}. \tag{12}$$

Finally, recall the definition of marginal density in the continuous case:

$$\begin{aligned} m(x) &= m(x|\pi) \qquad (13)\\ &= \int_0^1 f(x|p)\pi(p)dp, \qquad (14) \end{aligned}$$

meaning we have the equation:

$$g(p|x) = \frac{f(x|p)\pi(p)}{\int_0^1 f(x|p)\pi(p)dp}. \tag{15}$$

Choosing a Prior

When using Bayesian methods, the choice of which prior to use is very important. In cases where one has information about a problem already (for example, data from a past experiment), using an informative prior is appropriate. On the other hand, one may not have much—if any—information about a problem beforehand and hence may want to choose a prior to express that.[13] There are many different choices for such an uninformative prior, two of which are discussed here.

The Uniform Prior

An intuitive choice for an uninformative prior is the uniform distribution. One problem with this choice is if your parameter of interest could be any number (i.e., any real number versus a number only in the interval [0, 1]), then this is not a distribution since a uniform distribution on (−∞, ∞) cannot be normalized (and hence unable to integrate to one). Such a distribution is called an improper prior. Using an improper prior will result in the posterior “distribution” not being a true distribution. Despite this, improper priors are often still used, provided that $m(x) = \int f(x|p)\pi(p)dp$ (as in Eq. 7 above) is finite.[14]

The more serious problem with using a uniform prior, in general, is parameterization. Intuitively, we would expect that order would not matter between applying a prior and changing variables. In other words, we expect that (1) applying a prior and then changing variables and (2) changing variables and then applying the same prior to each give the

[13] Carlin and Louis, *Bayesian Methods*.

[14] Luigi Pace and Alessandra Salvan, *Principles of Statistical Inference* (New Jersey: World Scientific, 1994).

same results. In general, however, this is not always the case—for example, when using a uniform prior.[15]

A truly uninformative prior is an unreachable goal since even the uniform prior contains some information (namely that every outcome has an equal probability of occurring). Hence, a minimally informative prior is the realistic aim. This means when deciding on a minimally informative prior to use, one may have to choose between using a flat prior (the uniform), which is not uninformative in its parameterization, and a prior uninformative in its parameterization, which is not flat.

The Jeffreys Prior

The Jeffreys prior is one example of a prior that directly addresses this reparameterization issue. While there are other priors that also address this issue, the Jeffreys is widely favored since it can conveniently be written as a beta distribution. The Jeffreys prior is invariant under reparameterization, meaning that (1) applying the Jeffreys prior and then changing variables and (2) changing variables and then applying the Jeffreys prior will each give the same results. This makes the Jeffreys prior minimally informative in the sense that the choice of parameterization now does not matter. Hence, despite not being a flat prior, the Jeffreys is uninformative in terms of parameterization invariance.[16] Note that this creates the trade-off mentioned earlier between having a prior that is flat and a prior that is invariant under reparameterization.

Priors for the Binomial Distribution

The beta distribution is the conjugate family for the binomial distribution. A conjugate family is any family of functions containing both the prior and posterior functions.[17] Indeed, using $Posterior \propto Likelihood \times Prior$ (from Eq. 8), the product of a binomial likelihood (since using a binomial distribution) with a beta prior will always result in a beta posterior.[18]

Conjugate families like the beta are often favored because of mathematical convenience. In addition, the beta distribution has many properties that make it a desirable choice. For example, the beta distribution is defined by only two parameters (α, β), written as $Beta(\alpha, \beta)$. The choice of these parameters can give nearly any shape of function (U-shaped, L-shaped, J-shaped, skewed, uniform, etc.). Since often there is only crude knowledge of a prior translating into some general shape, such flexibility is typically sufficient.

[15] Pace and Salvan, *Principles*.

[16] Pace and Salvan, *Principles*.

[17] Casella and Berger, *Statistical Inference*.

[18] Martz and Waller, *Bayesian Reliability Analysis*.

For a distribution $Beta(a, b)$, its PDF is given by:

$$f(p) = \frac{(a+b-1)!}{(a-1)!(b-1)!}p^{a-1}(1-p)^{b-1}, \tag{16}$$

for $0 \leq p \leq 1$ (see Martz and Waller).[19] For example, using the uniform distribution ($Beta(1, 1)$), the above becomes:

$$f(p) = 1 \tag{17}$$

(by substituting in $a = 1$, $b = 1$).

We use the notation $\sim$ to mean "is given by the distribution." For a binomial distribution with parameter p and prior $f(p) \sim Beta(\alpha, \beta)$, it follows that the posterior distribution is $f(p|x) \sim Beta(x + \alpha, N - x + \beta)$. With more data, the choice of prior becomes less important (i.e., if N, x are able to overwhelm the α, β). The uniform prior for the binomial distribution is $Beta(1, 1)$, while the Jeffreys prior for the binomial distribution is $Beta(\frac{1}{2}, \frac{1}{2})$ (see Martz and Waller).[20] Table 2 and Figure 3 depict these priors and posteriors, with Figure 3 using the data from Table 1.

Table 2. Comparison of Binomial Priors and Posteriors

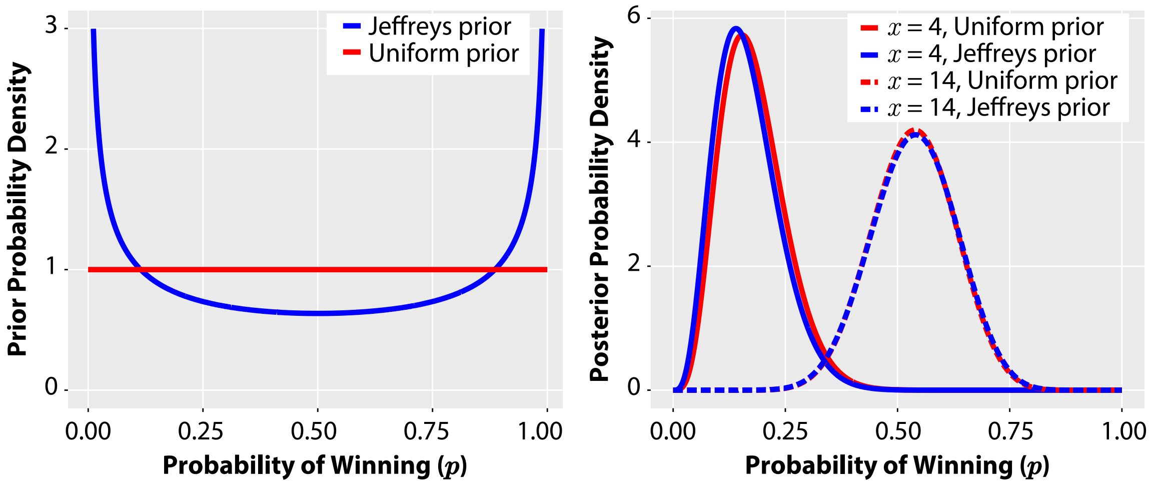
Name	Prior	Posterior
Uniform	$Beta(1, 1)$	$Beta(x+1, N-x+1)$
Jeffreys	$Beta(\frac{1}{2}, \frac{1}{2})$	$Beta(x+\frac{1}{2}, N-x+\frac{1}{2})$

We will now illustrate how the posterior is derived when using the uniform prior for the binomial distribution. The likelihood function for the binomial distribution comes from Eq. 1: $f(x|p) = \frac{N!}{x!(N-x)!}p^x(1-p)^{N-x}$. From Eq. 17 above, we know the PDF of $Beta(1, 1)$ is $f(p) = 1$. Hence the prior PDF is $\pi(p) = 1$. Then:

$$\begin{aligned}
m(x) &= \int f(x|p)\pi(p)dp \qquad (18)\\
&= \int_0^1 \left[\frac{N!}{x!(N-x)!}p^x(1-p)^{N-x}\right]dp \qquad (19)\\
&= \left(\frac{N!}{x!(N-x)!}\right)\int_0^1 \left[p^x(1-p)^{N-x}\right]dp \qquad (20)\\
&= \left(\frac{N!}{x!(N-x)!}\right)\left(\frac{x!(N-x)!}{(N+1)!}\right) \qquad (21)\\
&= \frac{1}{N+1}. \qquad (22)
\end{aligned}$$

[19] Martz and Waller, *Bayesian Reliability Analysis*.

[20] Martz and Waller, *Bayesian Reliability Analysis*.



Left, Comparison of the prior probability densities for the uniform prior $Beta(1,1)$ and the Jeffreys prior $Beta(\frac{1}{2}, \frac{1}{2})$. Right, Comparison of the posterior probability densities using a uniform and Jeffreys prior for each of $x = 4$, 14 successes out of $N = 26$ trials. In both graphs, the uniform prior is depicted in red and the Jeffreys prior in blue.

Figure 3. Comparison of Uniform and Jeffreys Priors

Then from Eq. 7, we have the posterior density:

$$g(p|x) = \frac{f(x|p) \cdot \pi(p)}{m(x)} \tag{23}$$

$$= \frac{\left(\frac{N!}{x!(N-x)!} p^x (1-p)^{N-x}\right)(1)}{\frac{1}{N+1}} \tag{24}$$

$$= \frac{(N+1)!}{x!(N-x)!} p^x (1-p)^{N-x} \tag{25}$$

$$= \frac{[(x+1)+(N-x+1)-1]!}{[(x+1)-1]![(N-x+1)-1]!} p^{(x+1)-1} (1-p)^{(N-x+1)-1}. \tag{26}$$

Note this follows the form of Eq. 16 with $a = x + 1$ and $b = N - x + 1$, meaning we in fact have the posterior distribution $Beta(x + 1, N - x + 1)$.

Impact of the Choice of Prior

Figure 3 shows that there is not much difference in the posteriors using either a uniform or Jeffreys prior. Is that still true when choosing a more skewed (or informative) prior?

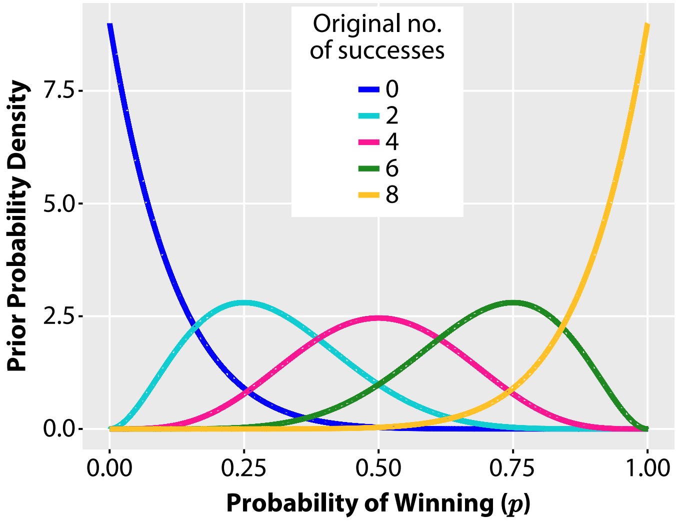
To explore this, we assume there was a previous experiment with a total of $\hat{N} = 8$ trials with $\hat{x} = 0, 2, 4, 6, 8$ successes ("original data"). How do the priors based on these varying numbers of successes affect the posterior distributions for the data $N = 26$, $x = 4$, 14 ("new data")? How do these compare to using the uniform prior as above?

Table 3. Comparison of Informative Priors and Their Respective Posteriors

Prior for Original Data	Original Data	Prior for New Data	New Data	Posterior for New Data
Beta(1, 1)	$\hat{N} = 8, \hat{x} = 0$	*Beta*(1, 9)	$N = 26, x = 4$	*Beta*(5, 31)
	$\hat{N} = 8, \hat{x} = 2$	*Beta*(3, 7)		*Beta*(7, 29)
	$\hat{N} = 8, \hat{x} = 4$	*Beta*(5, 5)		*Beta*(9, 27)
	$\hat{N} = 8, \hat{x} = 6$	*Beta*(7, 3)		*Beta*(11, 25)
	$\hat{N} = 8, \hat{x} = 8$	*Beta*(9, 1)		*Beta*(13, 23)
	$\hat{N} = 8, \hat{x} = 0$	*Beta*(1, 9)	$N = 26, x = 14$	*Beta*(15, 21)
	$\hat{N} = 8, \hat{x} = 2$	*Beta*(3, 7)		*Beta*(17, 19)
	$\hat{N} = 8, \hat{x} = 4$	*Beta*(5, 5)		*Beta*(19, 17)
	$\hat{N} = 8, \hat{x} = 6$	*Beta*(7, 3)		*Beta*(21, 15)
	$\hat{N} = 8, \hat{x} = 8$	*Beta*(9, 1)		*Beta*(23, 13)

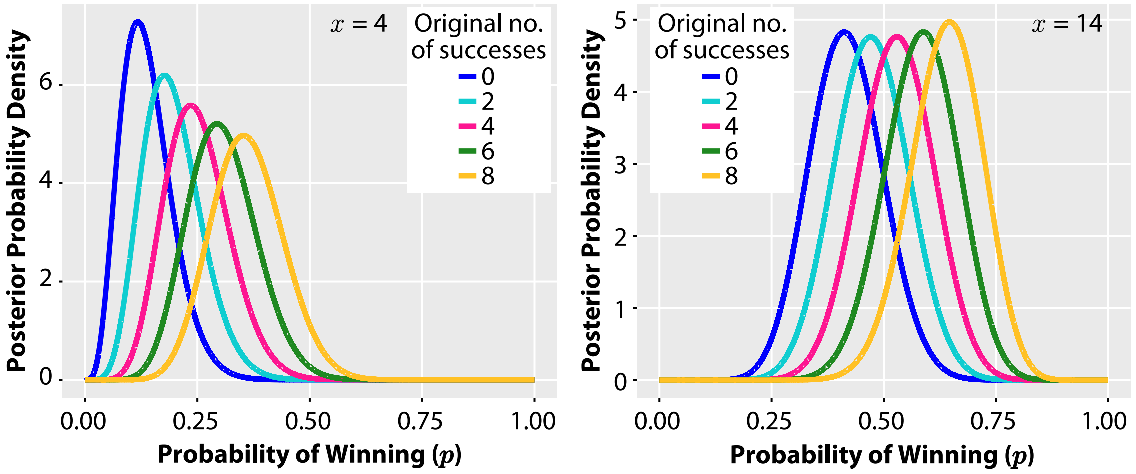
Table 3 shows what the respective priors and posteriors would be for each of these scenarios. The priors for the new data are illustrated in Figure 4. Note the varied shapes of these informative priors, all using the beta distribution.

Figure 5 shows the posteriors' PDFs. As these results show, using an informative prior can strongly alter the results of the posterior.



Comparison of the prior probability densities for the various beta distributions used as priors for the new data (column 3 in Table 3).

Figure 4. Comparison of Priors for the Beta Distribution



Comparison of the posterior distribution densities obtained using the original data of $\hat{N} = 8$ trials and $\hat{x} = 0, 2, 4, 6, 8$ successes, respectively. Left, Using the new data of $N = 26$, $x = 4$. Right, Using the new data of $N = 26$, $x = 14$.

Figure 5. Comparison of Posterior Distribution Densities Using Informative Priors

Bayesian Estimation

As discussed earlier, Bayesian methods output an entire probability distribution as the posterior. However, it is often more useful to characterize this information using descriptive statistics rather than the distribution as a whole. Several values can be used as the Bayesian point estimate, depending on the problem at hand, although all of these still come from the posterior distribution. For example, a Bayesian point estimate could be the mean, median, or mode of the posterior distribution; using the mean and variance of the posterior distribution is often favored.[21]

In addition, probability intervals can be constructed using the posterior distribution to provide a level of uncertainty for the point estimate. In particular, to construct the $100(1 - \alpha)\%$ Bayesian interval estimate, one can simply take the $\frac{\alpha}{2}$ and $1 - \frac{\alpha}{2}$ percentiles of the posterior probability distribution.[22]

Frequentist Approach

This section dives deeper into the frequentist approach. First, maximum likelihood estimation (MLE) is discussed, both in general and as it applies to the binomial distribution. Next, Clopper–Pearson intervals are discussed as a way to quantify uncertainty in p. The general Clopper–Pearson method is explained, as well as its connection to the beta distribution.

[21] Carlin and Louis, *Bayesian Methods*.

[22] Casella and Berger, *Statistical Inference*.

Maximum Likelihood Estimation

General Concept

Recall a primary difference between frequentist and Bayesian reasoning: frequentists assume a parameter like p to be fixed but unknown, while Bayesians instead assume the parameter to be a random variable. This means that Bayesians can calculate a posterior probability distribution of the parameter p if given the data and a prior. Conversely, for frequentists, there can be no such distribution on p, since p is viewed as a fixed quantity and not a random variable.

Instead, frequentists use MLE to find the value of the parameter p such that the likelihood $L(p|x)$ is maximized.[23] The value equal to this maximum is denoted $\hat{p}$ and, again, is a point estimate and not a random variable. This value $\hat{p}$ is interpreted as being the most likely value of p to result in the obtained data.[24]

MLE for the Binomial Distribution Likelihood Function

Assume we have a binomial distribution. Then letting Y be a random variable for the number of x successes out of N total trials, each with a probability p of success, we have $Y = \text{binomial}(N, p)$. Since the binomial distribution is a discrete probability distribution, the likelihood function for the parameter p given the observed number of successes x, $L(p|x)$, is simply the probability mass function considered as a function of p (see Ibe[25]):

$$\begin{aligned} L(p|x) &= \mathbb{P}(Y = x) && (27) \\ &= \frac{N!}{x!(N-x)!} p^x (1-p)^{N-x}. && (28) \end{aligned}$$

Since the log function is monotonic, finding the value of p that maximizes L is the same as finding the value of p that maximizes $\log(L)$. Hence we can write the loglikelihood function:

$$\begin{aligned} \ell(p|x) &= \log(L(p|x)) && (29) \\ &= \log\left(\frac{N!}{x!(N-x)!} p^x (1-p)^{N-x}\right) && (30) \\ &= \log\left(\frac{N!}{x!(N-x)!}\right) + \log\left(p^x\right) + \log\left((1-p)^{N-x}\right) && (31) \end{aligned}$$

[23] Ibe, *Fundamentals.*

[24] Ibe, *Fundamentals.*

[25] Ibe, *Fundamentals.*

$$= \log\left(\frac{N!}{x!(N-x)!}\right) + x\log(p) + (N-x)\log((1-p)). \tag{32}$$

Now, we want to find the maximum of this function (i.e., take the derivative, set it equal to zero, solve for p, and check that its second derivative is negative):

$$\frac{d}{dp}(\ell(p|x)) = \frac{d}{dp}\left(\log\left(\frac{N!}{x!(N-x)!}\right)\right) + \frac{d}{dp}(x\log(p)) + \frac{d}{dp}((N-x)\log((1-p))) \tag{33}$$

$$= x\frac{d}{dp}(\log(p)) + (N-x)\frac{d}{dp}(\log((1-p))) \tag{34}$$

$$= x\left(\frac{1}{p}\right) + (N-x)\left(\frac{1}{1-p}\right)(-1) \tag{35}$$

$$= \frac{x}{p} - \frac{N-x}{1-p}. \tag{36}$$

Then we have:

$$\frac{d}{dp}(\ell(p|x)) = 0 \tag{37}$$

$$\frac{x}{p} - \frac{N-x}{1-p} = 0 \tag{38}$$

$$\frac{x}{p} = \frac{N-x}{1-p} \tag{39}$$

$$x(1-p) = p(N-x) \tag{40}$$

$$x = Np \tag{41}$$

$$p = \frac{x}{N}. \tag{42}$$

Finally, to check that this value is a maximum:

$$\frac{d^2}{dp^2}(\ell(p|x)) = \frac{d}{dp}\left(\frac{x}{p} - \frac{N-x}{1-p}\right) \tag{43}$$

$$= \frac{-x}{p^2} - (N-x)\left(\frac{1}{(1-p)^2}\right) \tag{44}$$

$$= \frac{x-N}{(1-p)^2} - \frac{x}{p^2} \tag{45}$$

$$\leq 0 \tag{46}$$

(since a positive term is being subtracted from a negative term), so $p = \frac{x}{N}$ is indeed a maximum.

Interpretation

This means the value of p that is most likely to result in the obtained data of x successes out of N trials is $\frac{x}{N}$, a very intuitive result.[26] Note that this value will not necessarily align with the mode of the posterior probability distribution obtained using Bayesian methods. While MLE returns the point estimate maximizing $L(p|x)$, Bayesian methods are returning the entire distribution of p given the data. Since these posterior distributions rely on the prior, they can be drastically different (see Figure 5) and may or may not have a mode aligning with the MLE.

The mode of the posterior distribution is called the generalized maximum likelihood estimate. This is because when using a uniform prior, the mode of the posterior distribution is equal to the MLE, $\frac{x}{N}$ in this case.[27]

Clopper–Pearson Intervals

Although MLE is often a useful estimation, it does not help answer our primary question of how to quantify uncertainty in p. Confidence intervals are useful for seeing how extreme p could be while still being consistent with the observed data. There are a variety of techniques for calculating confidence intervals with the binomial distribution. However, there is no single agreed-upon best method, with all of them having various limitations.

The most common of these is the Wald interval.[28] The Wald interval uses the standard calculation for a confidence interval:

$$p \pm z \cdot \sqrt{\frac{p(1-p)}{N}},$$

with z coming from the standard normal distribution (for example, $z = 1.960$ for a two-sided 95% confidence interval). While widely used, the Wald interval is known to have poor coverage (meaning the probability that the calculated interval includes the true population value may be an underestimate). In particular, the Wald interval will become a worse estimate with smaller sample size.[29]

We have already seen (in the section on Bayesian estimation) Bayesian techniques as one example of an alternative to Wald intervals. In the frequentist realm, much work has also been done on modifying and extending Wald intervals to be better estimates.[30] However,

[26] Ibe, *Fundamentals*.

[27] Carlin and Louis, *Bayesian Methods*.

[28] Robert G. Newcombe, *Confidence Intervals for Proportions and Related Measures of Effect Size* (New York: CRC Press, 2013).

[29] Newcombe, *Confidence Intervals*.

[30] Newcombe, *Confidence Intervals*.

these extensions all inherently rely on the normal approximation, which is not always appropriate, particularly in situations (like our problem) with small sample sizes.

The Clopper–Pearson interval was developed to not rely on any normal approximation and to be a conservative estimate (or have good coverage), meaning the probability that the calculated interval includes the true population value will always be an overestimate.[31] The Clopper–Pearson method inverts two single-tailed binomial tests at the desired confidence level. In other words, the lower bound of this interval is the value of p that creates a binomial distribution with exactly $\frac{\alpha}{2}$ area in its upper tail (from x to N), while the upper bound of the interval is the value of p creating a binomial distribution with exactly $\frac{\alpha}{2}$ area in its lower tail (from 0 to x).

More specifically, the goal is to find the confidence interval for p with a confidence level of $1 - \alpha$. For example, $\alpha = 0.1$ corresponds to a $100(1 - \alpha) = 90\%$ confidence interval, or a confidence level of 0.9. For a given N, x, α, the Clopper–Pearson interval is the closed interval $[p_L, p_U]$ such that the following two equations are satisfied[32]:

$$\frac{\alpha}{2} = \sum_{i=x}^{N} \binom{N}{i} p_L^i (1 - p_L)^{N-i} \tag{47}$$

$$\frac{\alpha}{2} = \sum_{i=0}^{x} \binom{N}{i} p_U^i (1 - p_U)^{N-i}. \tag{48}$$

While the goal is to solve these equations for p_L and p_U, these equations are impossible to solve analytically in their general form, and often still impossible to solve analytically even after specific values are inputted. However, these equations can be solved numerically for p_L, p_U for any fixed values of N, x, α. A graphical illustration of these equations can be found in Figure 6, where $N = 26$, $x = 14$, $\alpha = 0.1$, and $p_L = 0.362$, $p_U = 0.708$ have been solved for.

For a fixed N and x, we can find these $[p_L, p_U]$ confidence intervals for varying values of α and then plot them on the same graph (see Figure 7). Intuitively, these intervals are larger at higher confidence levels, with [0, 1] being the confidence interval for 100% confidence. Note that Clopper–Pearson serves as a conservative estimate; even with a confidence level of 0%, the interval does not converge to a single value. Also intuitively, the confidence intervals cover smaller values of p with $x = 4$ versus $x = 14$. One would expect the probability of success p to be larger when there are fourteen observed successes out of twenty-six trials versus only four observed successes out of twenty-six trials.

Alternatively, for a fixed α, we can plot $[p_L, p_U]$ confidence intervals for varying values of x on the same graph (see Figure 8). Intuitively, we see larger intervals for $\alpha = 0.05$ (corresponding to the 95% confidence interval) than $\alpha = 0.10$ (corresponding to the 90% confidence

[31] Newcombe, *Confidence Intervals.*

[32] Newcombe, *Confidence Intervals.*

interval). We also see smaller intervals toward the extreme values of x (i.e., x closer to 0 and N), since there are fewer options for what p could likely be to obtain those extremes. Finally, similar to the above, we also see the values of p increasing as x increases (since one expects the probability of success to be larger when more successes are observed).

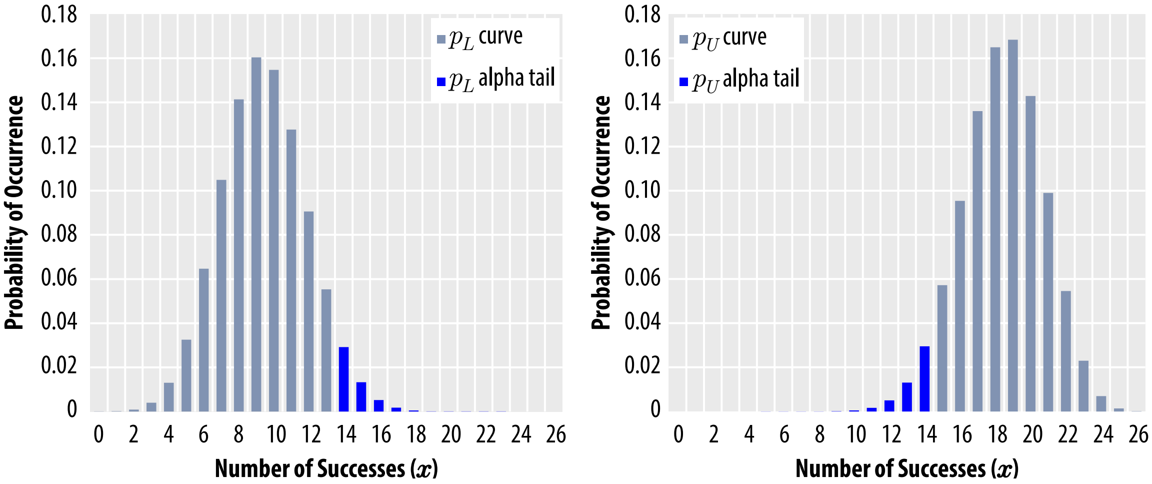
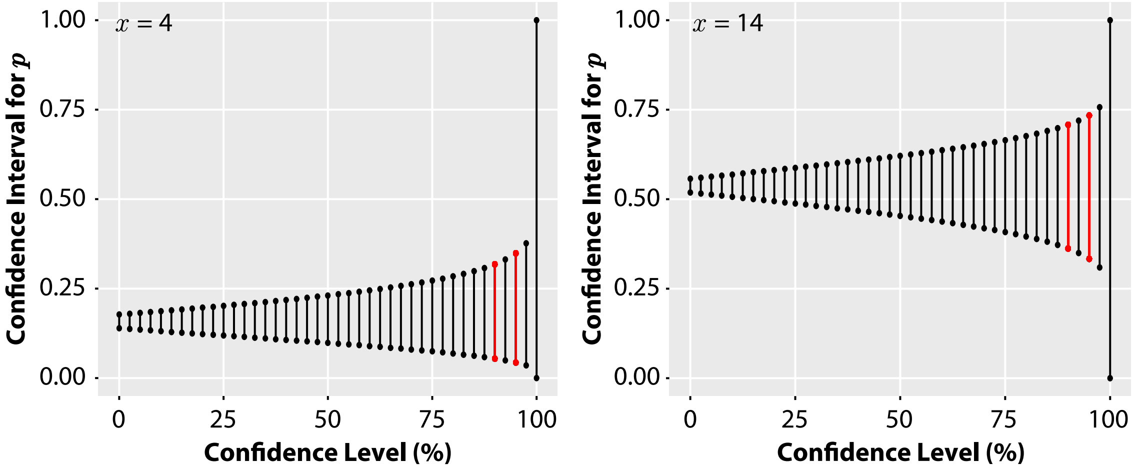


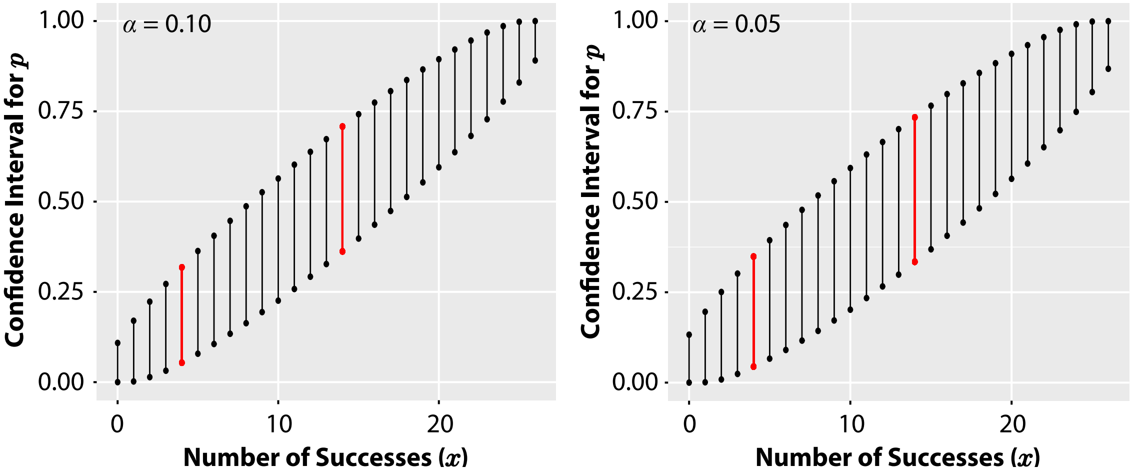
Illustration of the Clopper–Pearson equations for N = 26, x = 14, α = 0.1. Left, light blue p_L represents the binomial probabilities obtained when there is 5% of area in the upper tail (i.e., dark blue p_L alpha tail). Right, light blue p_U represents the binomial probabilities obtained when there is 5% of area in the lower tail (i.e., dark blue p_U alpha tail).

Figure 6. Illustration of the Clopper–Pearson Method



For N = 26, and x = 4 (left), x = 14 (right), forty-one different confidence intervals are plotted for varying α, with the confidence level being 100(1 – α)%. For example, the 90% and 95% confidence intervals are each highlighted in red.

Figure 7. Comparison of Clopper–Pearson Intervals for Varying Confidence Levels



For $N = 26$, and $\alpha = 0.10$ (left), $\alpha = 0.05$ (right), different confidence intervals are plotted for varying x, number of successes observed. For example, the confidence intervals for $x = 4$ and $x = 14$ are each highlighted in red.

Figure 8. Comparison of Clopper–Pearson Intervals for Varying Numbers of Successes

Clopper–Pearson Intervals and the Beta Distribution

Since the Clopper–Pearson system of equations cannot be solved analytically, it is useful to find other ways to write the Clopper–Pearson equations. One alternative is writing the Clopper–Pearson equations in terms of the beta distribution. Although this new form of Clopper–Pearson, written using the beta distribution, will still need to be solved numerically, the vast research and intuition available for the beta distribution can now be leveraged for Clopper–Pearson as well.[33]

For both derivations below, we will use the following identity, obtained via integration by parts:

$$\sum_{w=k}^{n} \binom{n}{w} p^w (1-p)^{n-w} = \int_0^p \left(\frac{n!}{(k-1)!(n-k)!} \right) z^{k-1} (1-z)^{n-k} dz. \tag{49}$$

We will use the definition for the cumulative distribution function $F(y;\, a,\, b)$ for the beta distribution with any two parameters a, b:

$$F(y; a, b) = \left(\frac{(a+b-1)!}{(a-1)!(b-1)!} \right) \int_0^y t^{a-1} (1-t)^{b-1} dt. \tag{50}$$

We also introduce notation for the inverse of this beta distribution cumulative distribution function: $F^{-1}(p;\, a,\, b)$. Note the following equations are equivalent, by definition of being inverses:

$$p = F(y; a, b) \tag{51}$$

$$y = F^{-1}(p; a, b). \tag{52}$$

[33] Newcombe, *Confidence Intervals*.

Let N and x still be given by the number of trials and number of successes, respectively. First, for the lower bound of p, p_L, we have:

$$\begin{aligned} \frac{\alpha}{2} &= \sum_{i=x}^{N} \binom{N}{i} p_L^i (1-p_L)^{N-i} && (53) \\ &= \int_0^{p_L} \left(\frac{N!}{(x-1)!(N-x)!} \right) z^{x-1}(1-z)^{N-x} dz && (54) \\ &= \left(\frac{N!}{(x-1)!(N-x)!} \right) \int_0^{p_L} z^{x-1}(1-z)^{N-x} dz && (55) \\ &= F(p_L; x, N-x+1), && (56) \end{aligned}$$

meaning the lower bound for p is given by:

$$p_L = F^{-1}\left(\frac{\alpha}{2}; x, N-x+1\right). \qquad (57)$$

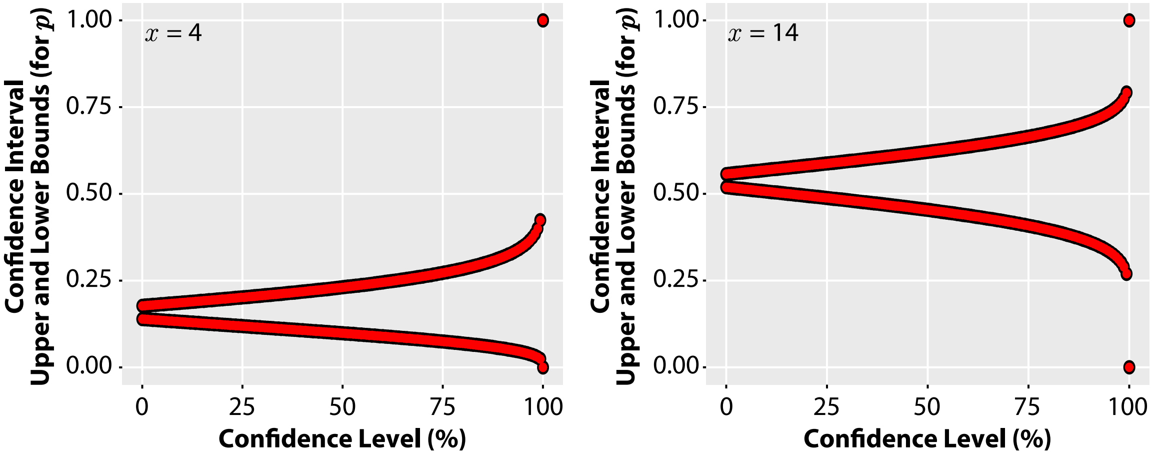
Then the upper bound for p, p_U, can similarly be found:

$$\begin{aligned} \frac{\alpha}{2} &= \sum_{i=0}^{x} \binom{N}{i} p_U^i (1-p_U)^{N-i} && (58) \\ &= 1 - \sum_{i=x+1}^{N} \binom{N}{i} p_U^i (1-p_U)^{N-i} && (59) \\ &= 1 - \int_0^{p_U} \left(\frac{N!}{x!(N-x-1)!} \right) z^{x}(1-z)^{N-x-1} dz && (60) \\ &= 1 - \left(\frac{N!}{x!(N-x-1)!} \right) \int_0^{p_U} z^{x}(1-z)^{N-x-1} dz && (61) \\ 1 - \frac{\alpha}{2} &= F(p_U; x+1, N-x), && (62) \end{aligned}$$

meaning the upper bound for p is given by:

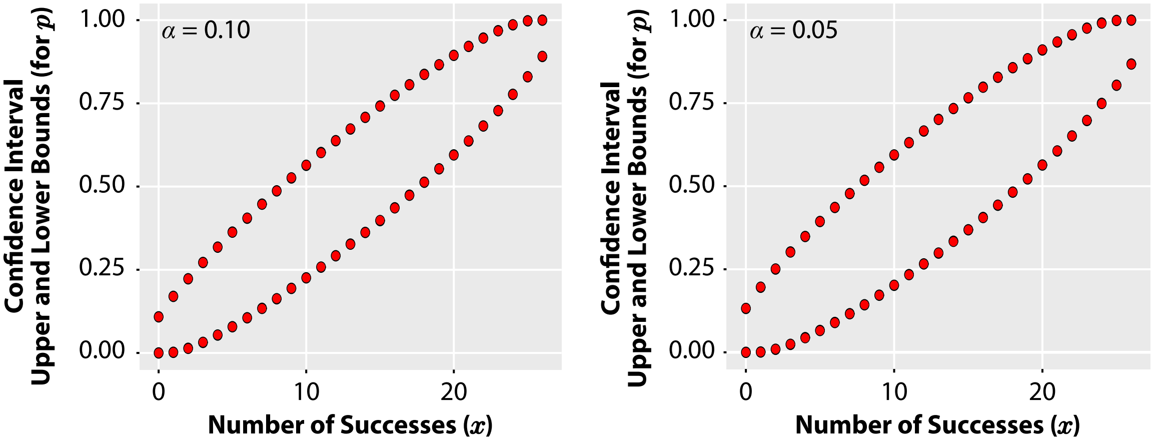
$$p_U = F^{-1}\left(1 - \frac{\alpha}{2}; x+1, N-x\right). \qquad (63)$$

To confirm that these identities between the Clopper–Pearson upper/lower bounds and the beta distribution are true, we plot these equations in Figures 9 and 10. They are similar to to Figures 7 and 8, respectively, but now with only the lower and upper bounds of the confidence interval $[p_L, p_U]$ plotted to make the graphs easier to read. The bounds obtained from the F^{-1} functions (Eqs. 57 and 63) are plotted in red over top the bounds obtained from the Clopper–Pearson interval sums (Eqs. 53 and 58) in black. (Since these are all identical, the points in black are larger so that they are still visible behind the red points.)



For $N = 26$, and $x = 4$ (left), $x = 14$ (right), 150 different confidence interval lower and upper bounds are plotted for varying α, with the confidence level being $100(1 - \alpha)\%$. Black points are obtained via Clopper–Pearson sums, and red points from the analogous beta distribution.

Figure 9. Comparison of Clopper–Pearson Intervals and Their Analogous Beta Distribution for Varying Confidence Levels



For $N = 26$, and $\alpha = 0.10$ (left), $\alpha = 0.05$ (right), different confidence interval lower and upper bounds are plotted for varying numbers of successes (x). Black points are obtained via Clopper–Pearson sums, and red points from the analogous beta distribution.

Figure 10. Comparison of Clopper–Pearson Intervals and Their Analogous Beta Distribution for Varying Numbers of Successes

Clopper–Pearson versus Bayes

Functions

Recall that the posterior probability distribution for a binomial distribution using a uniform prior ($Beta(1, 1)$) is a beta distribution with parameters $x + 1$ and $N - x + 1$. Hence if we wanted to find the 100(1 – α)% confidence interval for p using a uniform prior, we would use:

$$p_L = F^{-1}\left(\frac{\alpha}{2}; x+1, N-x+1\right) \tag{64}$$

$$p_U = F^{-1}\left(1-\frac{\alpha}{2}; x+1, N-x+1\right). \tag{65}$$

See the similarity between these and the Clopper–Pearson formulas found previously (Eqs. 57 and 63, respectively):

$$p_L = F^{-1}\left(\frac{\alpha}{2}; x, N-x+1\right) \tag{66}$$

$$p_U = F^{-1}\left(1-\frac{\alpha}{2}; x+1, N-x\right). \tag{67}$$

Finally, we can instead use the Jeffreys prior ($Beta\ (\frac{1}{2}, \frac{1}{2})$) on the binomial distribution to obtain the formulas for the 100(1 – α)% confidence interval for p:

$$p_L = F^{-1}\left(\frac{\alpha}{2}; x+\frac{1}{2}, N-x+\frac{1}{2}\right) \tag{68}$$

$$p_U = F^{-1}\left(1-\frac{\alpha}{2}; x+\frac{1}{2}, N-x+\frac{1}{2}\right). \tag{69}$$

Graphs

Because the Clopper–Pearson equations can be written in terms of the beta distribution, we can also plot their respective probability density and cumulative distribution functions. Note the above formulas for p_L and p_U using the uniform prior (Eqs. 64 and 65) use the same beta distribution. Similarly, the formulas for p_L and p_U using the Jeffreys prior (Eqs. 68 and 69) use the same beta distribution. However, the formulas for p_L and p_U obtained using Clopper–Pearson (Eqs. 66 and 67) use a different beta distribution for each of p_L and p_U. This means, when plotting these distributions, using the uniform and Jeffreys priors will result in plotting only one curve for each of the uniform and Jeffreys, while using Clopper–Pearson will result in two curves, one for each of p_L and p_U.

The formulas for Clopper–Pearson (Eqs. 66 and 67) versus using Bayesian techniques (Eqs. 64 and 65 using the uniform prior and Eqs. 68 and 69 using the Jeffreys prior) clearly differ, and in fact are using different parameters to create different beta distributions. As such, one would expect their graphs to differ as well. Indeed, that is what we find. Figure 11 plots four different curves for each of x = 4 and x = 14 in different colors:

(1) Green, the posterior probability distribution for a binomial distribution using a uniform prior ($Beta$(1, 1)): $Beta$(x + 1, N – x + 1)

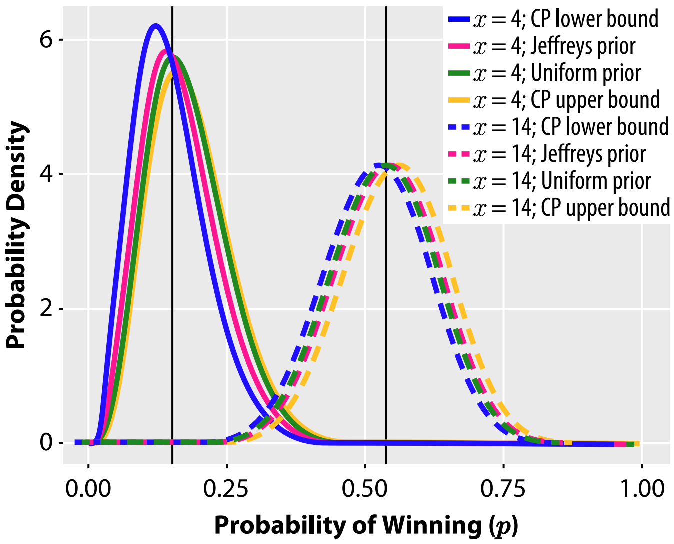
(2) Red, the posterior probability distribution for a binomial distribution using a Jeffreys prior ($Beta\,\frac{1}{2}, \frac{1}{2})$): $Beta(x + \frac{1}{2}, N - x + \frac{1}{2})$

(3) Blue, the distribution of the lower bound obtained via Clopper–Pearson: $Beta(x, N - x + 1)$

(4) Gold, the distribution of the upper bound obtained via Clopper–Pearson: $Beta(x + 1, N - x)$

Figure 11 displays these distributions via their PDFs. Note for the posterior probability distributions, the probability density of p is being plotted. On the other hand, from Clopper–Pearson, it is instead the probability densities of the upper and lower bounds of p being plotted. Also note that the mode of the posterior PDF, when using a uniform prior, aligns with the MLE ($x = 4$, 14, respectively), as discussed earlier. While the probability density curves obtained using Clopper–Pearson do not bound the curves obtained using an uninformative prior, the area under such curves will be bounded by the area under the Clopper–Pearson curves (for example, as in their cumulative distribution functions).

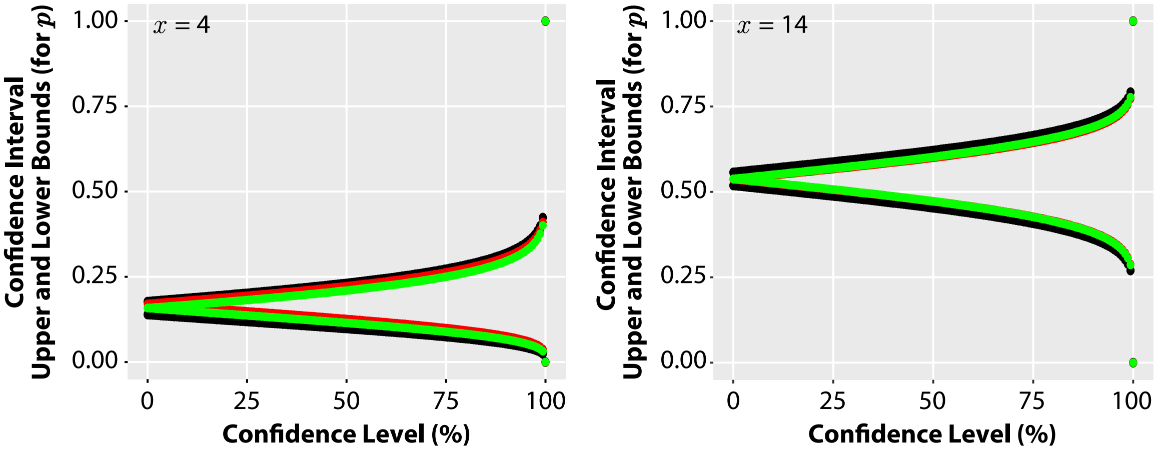
Finally, Figures 12 and 13 plot only the lower and upper bounds of the confidence (or credible, for the Bayesian approach) interval $[p_L, p_U]$. In both figures, black points are the bounds obtained via Clopper–Pearson, red points are from the beta distribution using a uniform prior, and green points are from the beta distribution using a Jeffreys prior. Again, we see



The PDFs for the distributions of the lower and upper bounds using Clopper–Pearson, and the posteriors using the uniform and Jeffreys priors, for both $x = 4$ (solid lines) and $x = 14$ (dashed lines). Vertical black lines depict the MLE for each of $x = 4$, 14 ($\frac{4}{26}$ and $\frac{14}{26}$, respectively).

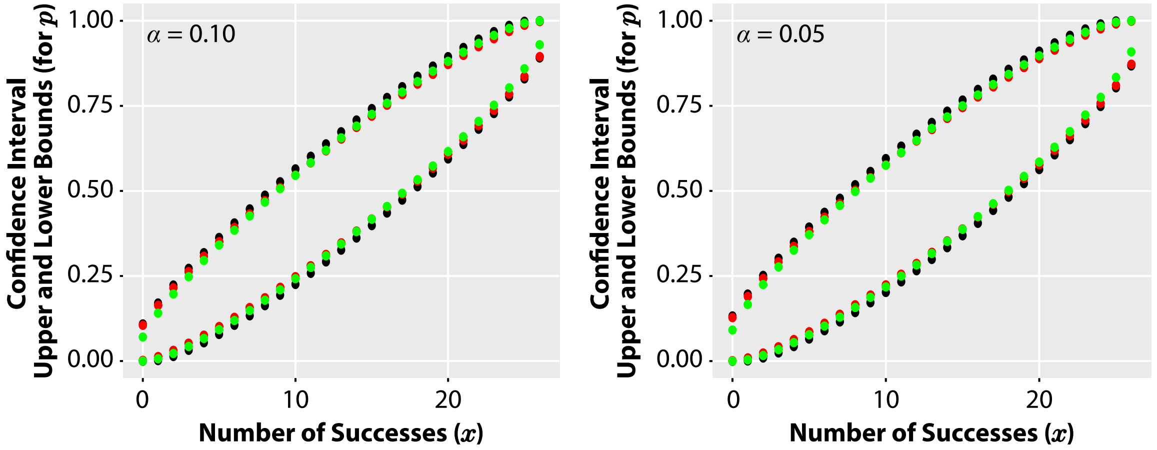
Figure 11. Comparison of Probability Densities Using Clopper–Pearson and Bayesian Methods

that Clopper–Pearson serves as a conservative estimate (i.e., actual confidence level ≥ α). This can be useful in that, as seen below, Clopper–Pearson bounds the intervals obtained via Bayesian methods using uninformative priors.



For $N = 26$, and $x = 4$ (left), $x = 14$ (right), confidence interval lower/upper bounds are plotted for varying α, with the confidence level being $100(1 - \alpha)$%. Black points are the bounds obtained via Clopper–Pearson, red points are from the beta distribution using a uniform prior, and green points are from the beta distribution using a Jeffreys prior.

Figure 12. Comparison of Confidence/Credible Intervals Using Clopper–Pearson/Bayesian Methods, Varying the Level of Confidence



For $N = 26$, and $\alpha = 0.10$ (left), $\alpha = 0.05$ (right), confidence interval lower/upper bounds are plotted for varying numbers of successes (x). Black points are the bounds obtained via Clopper–Pearson, red points are from the beta distribution using a uniform prior, and green points are from the beta distribution using a Jeffreys prior.

Figure 13. Comparison of Confidence/Credible Intervals Using Clopper–Pearson/Bayesian Methods, Varying the Number of Successes

Discussion

Assumptions and Limitations

As with any modeling work, we made assumptions that can lead to limitations of our work. In particular, all the work presented here is dependent on characterizing the data as outcomes of a binomial experiment. The binomial experiment assumes there are only two possible, mutually exclusive outcomes (i.e., success and failure). While this is often a useful simplifying assumption, real-world problems may instead consist of more than two options, or two options that could happen simultaneously. Returning to our motivating example, Kroenig combined the three separate outcomes of loss, stalemate, and compromise into the one category of lose so that he had only the two categories, win and lose.[34]

A binomial experiment also requires the assumption that outcomes are independent—that is, the outcome of a previous event does not impact the outcome of some subsequent event. Independence is not always a valid assumption when dealing with real-world events and needs to be carefully considered. For example, if the United States and the Soviet Union are a dyad involved in multiple nuclear crises, is it really true that the outcome of one crisis between these two nuclear powers had no effect on the outcome of some second nuclear crisis between them?

Similarly, while the assumption that the probability of success p (for either nuclear-superior or nuclear-inferior states) is constant for all trials is mathematically convenient, the validity of this assumption must be examined for the particular problem under study. For example, since "winning" in a nuclear crisis is defined as a state achieving its primary objective, it is possible that one, both, or neither states in the dyad will "win" the crisis. Does the probability of success change when both states have a chance of winning, compared to when the states' primary objectives are diametrically opposed (as in the Cuban missile crisis between the United States and the Soviet Union)?

In addition to the assumptions of independence between outcomes and p being constant between trials, there is also the inherent assumption that the probability of winning for a nuclear-superior state is independent of the probability of winning for a nuclear-inferior state. While stalemates and compromises are allowable outcomes of nuclear crises, in some historical nuclear crises, one country achieving its primary objective inherently implies that the other country will not achieve its primary objective. In other words, if the probability of the nuclear-superior state winning is p, then the probability of the nuclear-inferior state winning is $1 - p$. For instance, in the Cuban missile crisis during the Cold War, the principal objective of the United States was to get the Soviet Union to remove its missiles from Cuba. Conversely, the main objective of the Soviet Union was to have missiles in Cuba. By the

[34] Kroenig, "Nuclear Superiority."

United States achieving its primary objective from this crisis (a "success"), the Soviet Union was unable to achieve its primary objective (a "failure").

There are also limitations regardless of the assumption of having a binomial distribution. For example, observational data in the social sciences often have a small sample size. While this is frequently viewed as undesirable, it is also sometimes unavoidable. For example, in the case of Kroenig's nuclear crisis data, the sample size is twenty-six directed dyads for nuclear-superior states and twenty-six directed dyads for nuclear-inferior states.[35] Despite these being small sample sizes, twenty-six constitutes the total number of dyads involved in nuclear crises, by nature a small number. Indeed, in many situations, obtaining a larger sample size is costly, difficult, or even impossible. Rather than use this limitation as an excuse to forgo using statistics or the most appropriate quantitative methods, one should accept the small sample size as fact and still use the best quantitative methods available. While large uncertainty can result from simply having a small sample size, this is an important conclusion and should not be ignored.

Sample size can instead be irrelevant if one has a population. However, there is often ambiguity between a sample and population, as is the case with nuclear crises. For example, the population could be defined as all nuclear crises that have occurred (by some date in history), which would imply that the entire population is known. However, while there have been exactly twenty-six state dyads from nuclear crises occurring prior to that date, a new crisis could occur in the world at any moment, adding to a new population size. The population could instead be defined as all possible nuclear crises. In this case, the population would be infinite and the nuclear crises that have already occurred would be considered a sample of this population. However, this sample would not be random, since that would assume that every possible nuclear crisis is equally likely to happen (for example, a nuclear crisis between the United States and United Kingdom being just as likely to occur as a nuclear crisis between India and Pakistan).

Regardless of how one defines population in the case of nuclear crises, there will be ambiguity and debate surrounding the definition. Despite this debate, one could still be interested in the uncertainty of a parameter estimate (for example, the probability of success for the state with the superior nuclear arsenal). If the entire population is known, this parameter would also be known. If instead the entire population is unknown, and potentially ambiguous, interval estimates are helpful, using either Bayesian or frequentist methods.

[35] Kroenig, "Nuclear Superiority."

Bayesian versus Frequentist Methods

In general, there are many trade-offs between using Bayesian versus frequentist approaches.[36] Bayesians view parameters probabilistically, while frequentists view parameters as fixed phenomena, all of which are either known or unknown. Bayesians use prior information when approaching a problem, while frequentists assume they have no prior information. Even with no prior information (and hence using an uninformative prior), Bayesian methods can still be desired if p is viewed as a random variable, rather than a fixed quantity. Bayesian methods also tend to be more computationally intensive and have fewer analytic solutions.

Despite these and other differences, we showed here that there is not much difference between using the frequentist method of Clopper–Pearson and Bayesian methods using uninformative priors. With large sample sizes, Bayesian methods will produce results similar to those produced with frequentist approaches (because with a large sample size, the prior becomes less influential). Such accord between these approaches is not unusual and has led some to think of frequentist inference as simply a special case of Bayesian inference.[37]

However, just because uninformative priors align with Clopper–Pearson is not a good reason to use them. In particular, if previous information is known, an informative prior should be used. Such guidance can be problematic when there is disagreement about what, if any, previous information is known about a problem, which can lead to the Bayesian methods being an approximation (because the prior can only be an approximation). In fact, one of the common criticisms of Bayesian methods is the lack of justification for the chosen prior (particularly when it is chosen subjectively or only for mathematical convenience).

It is impossible to get exact answers to the questions posed here, whether using Bayesian or frequentist techniques. While Clopper–Pearson is called the "exact" confidence interval for the binomial distribution parameter p, it is only an approximation, just like all other methods. In general, one can use other frequentist methods when there is a large sample size. By contrast, sample size does not affect the choice of which Bayesian method to use.

In addition, the Bayesian approach allows for a distribution to be calculated for the unknown parameter p. While a sentence such as "There is a 95% probability that the true value of a parameter lies in this interval" is quite intuitive in its meaning, one can only accurately say it while using Bayesian methods. Rather than a probability distribution of p, the frequentist approach allows for either a point estimate (e.g., mean, median), which gives less information overall, or Clopper–Pearson intervals.

[36] See Berger, *Statistical Decision Theory* for a more complete list.

[37] Carlin and Louis, *Bayesian Methods*.

Similarly, when Bayesian techniques are found to have obtained "wrong" results, they are more likely to be wrong by only some small margin. In contrast, frequentists do not weight how wrong a result might be. For example, a 90% frequentist confidence interval means that 90% of all confidence intervals computed will contain the true value of the parameter. This means that 10% of the confidence intervals will not contain the true value of the parameter, but "not contain" could mean that the confidence interval could be 0.000001 away from the true value of the parameter or 100,000 away from the true parameter, with no distinction made between these results. Since Bayesian methods assume p to have a probability distribution, these distances get weighted by probabilities.

Comparison to Hypothesis Testing and Regression Analysis

In general, there are important advantages to evaluating uncertainty in addition to hypothesis testing. Indeed, there are many criticisms of hypothesis testing.[38] One such criticism is that hypothesis testing provides the opposite information from what is actually desired.[39] What is desired is to know how likely it is that the null hypothesis is true given the data, while hypothesis testing shows how likely it is that the data were produced given that the null hypothesis is true. In contrast, evaluating uncertainty provides an entire distribution to use for comparison to a null hypothesis.

In addition, hypothesis testing inherently relies on p-values, but the idea of making an all-or-nothing conclusion based on whether to reject versus fail to reject a null hypothesis with a p-value of 0.049 versus 0.051 is intuitively senseless[40] and often biased.[41] With little to no guidance for choosing the cutoff between rejecting or failing to reject a null hypothesis, 0.05 is often used by default. However, this choice can have considerable consequences in any conclusions reached in a statistical analysis.

P-values require random sampling in acquiring the data, an assumption that is often not acknowledged and may greatly affect the results.[42] Evaluating uncertainty, even in addition to hypothesis testing, allows conclusions to be based on an entire distribution, rather than a single p-value alone. Hypothesis testing does not measure effect size or the importance of

[38] Harlow, Mulaik, and Steiger, *What If There Were No Significance Tests?*

[39] Ronald Carver, "The Case against Statistical Significance Testing," *Harvard Educational Review* 48, no. 3 (1978): 378–399.

[40] Michael Oakes, *Statistical Inference: A Commentary for the Social and Behavioral Sciences* (New York: John Wiley and Sons, 1986).

[41] James O. Berger and Donald A. Berry, "Statistical Analysis and the Illusion of Objectivity," *American Scientist* 76, no. 2 (1988): 159–165.

[42] Norman Cliff, "Dominance Statistics: Ordinal Analyses to Answer Ordinal Questions," *Psychological Bulletin* 114, no. 3 (1993): 494–509.

a given result.[43] Smaller p-values do not necessarily imply larger effects (or even *any* effect). Similarly, larger p-values can occur even when a large effect is present. This is because sample size and measurement precision will affect the p-value; for example, a small p-value can be produced simply by the sample size or measurement precision being high enough.[44]

In general, hypothesis testing puts disproportionate weight on sample size.[45] Sample size is very important in calculating p-values, meaning hypothesis testing inherently favors one large-sample study that finds statistical significance over multiple small-sample studies that find similar results but each have sample sizes too small to achieve statistical significance. While having multiple studies come to the same conclusion is quite informative, such an experimental design is often overlooked since the setup would have insufficient power to produce statistical significance.[46] As such, hypothesis testing can influence other aspects of experimental design as well. Hypothesis testing encourages large sample sizes, even at the expense of measurement error or poor methodology. Encouraging less measurement error or better methodology may in fact be more important than having a large sample size.[47] The methods presented here for uncertainty estimation do not require certain sample sizes, meaning that uncertainty estimation can work with either a large or small sample size, eliminating this potential biasing effect of sample size on results.

Conclusion

Many real-world problems can be modeled as a binomial experiment. We showed here that one can choose several different methods to explore uncertainty in the binomial parameter p, using both Bayesian and frequentist approaches. In particular, we showed that there is not much difference between using the frequentist method of Clopper–Pearson and Bayesian methods using uninformative priors. We also showed, when prior information is available, that the choice of an informative prior is very important when using Bayesian methods.

Recall our original motivating example: are nuclear-superior powers more likely to win in nuclear crises? We showed here that both the probability of winning in a nuclear crisis as the side with the superior nuclear arsenal and the probability of winning in a nuclear crisis as the side with the inferior nuclear arsenal are highly uncertain, reflections of both the small data set and the importance of variables other than the nuclear balance. Without considering these uncertainties, the probability of a state winning a nuclear crisis is significantly

[43] Wasserstein and Lazar, "ASA's Statement."

[44] Wasserstein and Lazar, "ASA's Statement."

[45] Harlow, Mulaik, and Steiger, *What If There Were No Significance Tests?*

[46] John W. Tukey, "Analyzing Data: Sanctification or Detective Work?," *American Psychologist* 24, no. 2 (1969): 83–91.

[47] David A. Savitz, "Is Statistical Significance Testing Useful in Interpreting Data?," *Reproductive Toxicology* 7, no. 2 (1993): 95–100.

lower if the state has an inferior nuclear arsenal. These results suggest that if a nuclear state anticipates nuclear crises in its future and wishes to win, it should strive to avoid nuclear inferiority. However, even the side with the superior nuclear arsenal should not confidently expect to win in a nuclear crisis, based on both the point estimate ($\frac{14}{26} \approx 0.54$) and the large amount of uncertainty present. More generally, the uncertainty present in this probability of winning in a nuclear crisis is both important and large. Indeed, such uncertainty, whether large or small, is important information to include in any conclusions being drawn from statistical analyses.

Kroenig used hypothesis testing and regression analysis to conclude that nuclear-superior states are indeed more likely to win in nuclear crises.[48] However, his methods did not demonstrate the uncertainty associated with these conclusions. Policy makers, or other nonstatisticians, may be more receptive to interval analysis, since they already know and appreciate the uncertainty present in problems and may want to see it quantified as well. For example, an analyst may care less about whether a nuclear-superior state is more likely to win and more about the probability of the nuclear-superior state winning in a nuclear crisis and the variability surrounding that estimate. Although these types of questions are the ones that policy makers are often the most interested in, they are not the ones most easily addressed using hypothesis testing and regression analysis. In such cases so prevalent in the social sciences, parameter estimation and uncertainty quantification have proven quite useful, and we advocate for more widespread use of such techniques, either by themselves or in conjunction with hypothesis testing and regression analysis.

[48] Kroenig, "Nuclear Superiority."

Acknowledgments

We gratefully acknowledge the reviews of earlier drafts of this paper by Joseph Warfield and Isabel Slavinsky. We also thank Jane Booker and Jane Pinelis for help in thinking through this complex problem. Their technical expertise and insights helped make our work better.

About the Authors

Kelly Rooker is a member of the Senior Professional Staff at the Johns Hopkins University Applied Physics Laboratory. Dr. Rooker has served as technical lead on multiple projects spanning mathematics, statistics, and data science. Her work has supported the Departments of Defense, Homeland Security, and State. Dr. Rooker earned her PhD in mathematics from the University of Tennessee.

James Scouras is a Senior Scholar at the Johns Hopkins University Applied Physics Laboratory and the former chief scientist of the Defense Threat Reduction Agency's Advanced Systems and Concepts Office. Previously, he was program director for risk analysis at the Homeland Security Institute, and held research positions at the Institute for Defense Analyses and the RAND Corporation. Among his publications are the book *A New Nuclear Century: Strategic Stability and Arms Control* (Praeger, 2002), coauthored with Stephen Cimbala, and his forthcoming edited volume, *On the Risk of Nuclear War*. Dr. Scouras earned his PhD in physics from the University of Maryland.

APL

JOHNS HOPKINS
APPLIED PHYSICS LABORATORY