Drew University

College of Liberal Arts

Two Approaches to Statistical Research:

Frequentist and Bayesian

A Thesis in Mathematics

by

Jamie Lynn Connors

Submitted in Partial Fulfillment

of the Requirements

for the Degree of

Bachelor in Arts

With Specialized Honors in Mathematics

May 2023

Abstract

This study investigates current controversies surrounding the use of *p*-values in statistics and related fields. The definition of statistical terms, such as statistical significance, are investigated through their contributions to *p*-hacking, publication bias, and misconceptions in statistics education. Past datasets are utilized to analyze the effectiveness and correctness of *p*-values and other statistical analysis methods. Further, a series of statistical studies are conducted to conclude that *p*-values have a few limitations when compared to alternative statistical measures, such as Bayesian statistics, through the use of statistical modeling. These studies prompt the discussion that the understanding around *p*-values requires clarification and modification for some. Thus, I clarify specific cautions on the use of *p*-values and discuss alternate methods of analysis.

Introduced as an alternative means of analysis, Bayesian statistics involves a distinct school of thought that is not solely based on the data itself. For this reason, I pursue both classic non-Bayesian and Bayesian methods further to gauge their respective strengths and weaknesses through the implementation of a simple statistical task of estimating the probability of a potentially biased coin. Modeling and simulation results yield Bayesian statistics as the more adaptable analysis method due to its probabilistic reasoning and incorporation of prior knowledge. The advantage of the Bayesian model over the usage of *p*-values is further discussed using modern day applications with severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2).

Keywords: Bayesian, publication bias, *p*-hacking, *p*-value, significance

Table of contents

1. Introduction

The current controversies surrounding the use of *p*-values is a prevalent topic in both mathematics and fields that require statistical research methods, as some claim the probability value is insufficient for analysis (Halsey, 2019). Notably, *p*-value is a commonly used statistic in many analyses, often incorporated to determine statistical significance of a hypothesis. While this is an element taught in many courses and widely employed by statisticians, its use is a debated topic (Lytsy, 2022). Fuel for the debate, however, lies in occasional misuses of the statistic, rather than its weaknesses (Lakens, 2021). Specifically, I support three main facts regarding *p*-values to enhance their use and promote more valid conclusions: *p*-values should be augmented with an understanding of power, *p*-values should be paired with effect size, and *p*-values should not be dichotomized into statistical significance and non-statistical significance. To substantiate these aforementioned ideas, I first explore the definition of *p*-values and their associated statistics, then examine each argument further. Moreover, I discuss the use of Bayesian models as an alternative framework to make statistical inference. This framework is compared to the traditional use of *p*-values, and I argue that both methods should be implemented with their respective advantages and limitations in mind.

1.1. Overview of the Thesis

Throughout this thesis, I seek to clarify the use of and promote understanding about *p*-values alongside specific methodology. The rest of the thesis is organized as follows. I first provide some brief introduction and historical remarks of both the classic, non-Bayesian hypothesis testing and the alternative Bayesian method (sections 1.2 and 1.3). I then investigate

the specific act of dichotomization and its contribution to publication bias and *p*-hacking with respect to the former head of the Cornell Food and Brand Lab, Dr. Brian Wansink (section 2.1).

To both connect and contrast modes of analysis, I discuss the specific statistical task of estimating the probability of heads for a potentially biased coin (section 3). Specifically, hypothesis testing based on *p*-values are compared to a Bayesian beta-binomial model via a series of simulations. Lastly, I survey modern-day real world applications of the beta-binomial model to further demonstrate the limitations of *p*-values and the strength of the Bayesian method (section 4).

1.2. Historical remarks on *p*-values

In introductory statistics courses and research, *p*-values are an integral factor (Cobb, 2015). Defined as the probability of selecting a sample with a statistic that is at least as extreme as the observed statistic in the direction of the alternative hypothesis, assuming the null hypothesis is true (Carucci, n.d.); it is often used to determine statistical significance (Nahm, 2017). The format for a statistical analysis involves stating the null (H_0) and alternative (H_a) hypotheses, then formulating a conclusion based on the obtained *p*-value. For example, a common one-sample *z*-test of proportions is used for testing the following hypotheses:

H_0: the proportion of the population (p) = *a number* (p_0);

H_a: the proportion of the population (p) ≠ *a number* (p_0).

Similarly, a common one-sample *t*-test of means is used for testing:

H_0: the mean of the population (μ) = *a number* (μ_0);

H_a: the mean of the population (μ) $\neq$ *a number* (μ_0).

For *p*-values less than 0.05, one is advised to conclude the results are statistically significant and reject the null hypothesis in favor of the alternative; for *p*-values greater than or equal to 0.05, one is advised to conclude the results are not statistically significant and fail to reject the null hypothesis. These steps are the main elements of an NHST.

Hypothesis tests are built from Thomas Bayes's essay, "An Essay towards solving a Problem in the Doctrine of Chances" (1763). To primarily analyze the chances of an event occurring, he highlights, "After having observed for some time the course of events it would be found that the operations of nature are in general regular, and that the powers and laws which prevail in it are stable and permanent" (p. 410). Significantly, Bayes formulated a theorem which influenced a matter of forming and accepting hypotheses regarding these events.

Extending Bayes's theorems, Ronald Fisher introduces *p*-values in "On the Mathematical Foundations of Theoretical Statistics." He cites the later-coined NHST's involve a "hypothetical infinite population" with a limited number of parameters, representative of the data to determine the "*probability* of a certain object fulfilling a certain condition [and] all such objects to be divided into two classes, according as they do or do not fulfil the condition" (1922). Shortly after Fisher identified the importance of utilizing the population measurement as a parameter, and formulating null and alternative hypotheses, two other statisticians sought further development. In the 1933 article, "On the problem of the most efficient tests of statistical hypotheses," Jerzy Neyman and Egon Pearson warn, "We may accept or we may reject a hypothesis with varying degrees of confidence; or we may decide to remain in doubt…. From the point of view of mathematical theory all that we can do is to show how the risk of the errors may be controlled

and minimised." That being said, the concept of an analysis is clarified to be an estimate, rather than a true reflection of a population. Neyman and Pearson further extend Fisher's ideas to include power analysis and discussions of Type I and Type II errors. The three aforementioned statisticians are credited with building the foundations of today's frequentist hypothesis testing theory, thus inspiring the whole of frequentist, or classic, statistics.

1.3. Introduction to Bayesian Analysis

Graduating from the three aforementioned concepts to better augment *p*-values, I seek to introduce another approach in conducting analyses. Bayesian statistics, a vast field under statistics and statistical theory, is built upon Bayes's Theorem and conditional probability. The latter defines the probability of one event, A, occurring, given that another event, B, has already occurred, defined as $P(A|B) = \frac{P(A \cap B)}{P(B)}$, where $P(A \cap B)$ is the probability of both events A and B occurring, and $P(B)$ is the probability of only B occurring. Bayes's Theorem is employed when the probability of B conditioned on A is either simpler than the probability of A conditioned on B, or provided. The probability of A conditioned on B can be written mathematically as $P(A|B) = \frac{P(A)*P(B|A)}{P(B)}$. Connecting Bayes's theorem and hypothesis testing, we can identify the hypothesis as an event (H), and data as an event (D). Plugging into Bayes's formula: $P(H|D) = \frac{P(H)*P(D|H)}{P(D)}$. P(H|D), also known as the posterior probability, is the probability of the hypothesis occurring, with influence of the data; P(H), also known as the prior probability, is the probability of the hypothesis occurring without influence of the data; P(D|H), also known as the likelihood, is the probability of the data based on the hypothesis; P(D) is the

probability of the data with respect to any hypothesis. More generally, to make inference on a parameter p, the goal in Bayesian analysis is to derive the posterior distribution of p, conditioning on data.

2. Cautions on the use of *p*-values

In modern day, statisticians are seemingly divided: some claim frequentist statistics and its methods, such as the NHST, are outdated (Szucs & Ioannidis, 2017); others believe the faults of classic methods lie majorly in misinterpretations (Lakens, 2021). I seek to provide guidelines to promote better understanding and augmentation of *p*-values and hypothesis testing, while comparing other methods of statistical inference, such as Bayesian statistics. In doing so, I caution against the dichotomization of *p*-values in terms of significance and employing them without consideration of power and effect size.

While a mere cluster of statisticians have misconceptions regarding the definition of a *p*-value (Lytsy, 2022), this should be cautioned. To counter this association, Stanford University professor John Ioannidis suggests there to be further analysis and less reliance on the *p*-value itself: his definition of the statistic leads to the association of a significant or nonsignificant *p*-value with more analysis or observation (Dusheck, 2016). Furthering this approach, Ronald Wasserstein urges statisticians to rely more on statistical thinking and less on statistical significance: “While it concerns an observation regarding statistical significance, many believe it answers the question regarding practical importance” (Gelman & Stern, 2006). That being said, it is integral to understand statistical significance with respect to the statistic within the context of the problem. Thus, statisticians are urged to move away from focusing exclusively on a *p*-value’s statistical significance, and instead pursue additional information (Wasserstein, 2019). Overall, it must be remembered that practical and statistical significance are not equivalent; statistical and non-statistical significance are not to be fixed concepts with an unchangeable threshold.

Equally important, it is essential to understand *p*-value in the context of effect and sample size. When dealing with real-world samples and populations, it is important to focus on analyzing a *p*-value thoroughly. As mentioned before, computing and reporting a *p*-value alone does not guarantee significance or lack thereof, it instead motivates further analysis and calculation. One reason for this extension of analysis is the existence of Type I and Type II errors in a study: the probability of rejecting a true null hypothesis (α) or failing to reject a false null hypothesis (β), respectively. To overcome the likelihood of these errors, one can utilize statistical power, which is defined as the probability of not making a Type II error (Weiss, 2012, p. 416). Mathematically, power is equivalent to (1- β). Although not universally employed, it is a necessary add-on to null hypothesis testing to correctly and further analyze *p*-value in connection to the hypotheses and in context of the sample (Weiss, 2012, p. 417).

Notably, statistical power takes into consideration sample size, effect size, number of tails, and the alpha level of the test. When analyzing *p*-values it is important to note the sample and effect sizes as they can alter results: "Smaller *p*-values do not necessarily imply the presence of larger or more important effects, and larger *p*-values do not imply a lack of importance or even lack of effect" (Wasserstein, 2016). Furthermore, according to Gail Sullivan and Richard Feinn, effect size should always be reported with the *p*-value for readers to understand its context in the study (2012). Conforming with the aforementioned statisticians, statistical power helps correctly gauge *p*-values with respect to sample sizes and effect sizes. Therefore, incorporating power simulations with null hypothesis tests can be a more suitable approach in many applications.

To demonstrate statistical power in the context of statistical analyses, I performed a series of simulations[1], focusing specifically on independent samples *t*-tests. *See an example of power analysis in Tables 1, 2 and Figure 1 of the Appendix.* Upon randomizing scores based on a mean difference of 5 and standard deviation of 10 among two groups, I conduct *t*-tests for sample sizes of 30 and 100 in SPSS, an IBM statistical tool. With the randomization command, I obtained a variety of effect sizes in terms of Cohen's *d*, and failed to notice any odd combinations, such as the pairing of $p < 0.05$ and $d < 0.2$. I then conducted the same randomization among two groups with sample sizes exclusively of 100, a mean difference of 2, and standard deviation of 10, to decrease the value of Cohen's *d*. Three simulations were carried out under this setup, resulting in different conclusions. The first simulation led to a *p*-value greater than 0.05 and strong effect size; the second simulation led to a *p*-value less than 0.05 and weak effect size. Both of these conclusions align with the idea that greater sample sizes lead to a correlation between *p*-values and Cohen's *d*: typically, significant *p*-values tend to be associated with greater effect sizes, and vice versa. However, the third simulation produced a very significant *p*-value and weak Cohen's *d*, disrupting the predicted trend. This idea further complicates *p*-value itself by invalidating the idea of dichotomized significance. In this case, with a sample size of 100, the test does not have enough power to consistently detect a mean difference of 2, and conclusions drawn by *p*-values alone can be misleading. Results motivate a need to tackle these complications through a thorough analysis of power-influenced simulations for The Euro Problem.

2.1. Cautionary Examples

[1]Appendix: Tables 1, 2; Figure 1

Notably, the publication crisis, the idea that if results are not statistically significant, the study will not get published in prominent journals (Dickerson and Min, 1993), enforces biased and corrupt practices such as *p*-hacking, the process of manipulating data to obtain a misleading *p*-value. This leads to a lack of diversity in published results and prevents studies associated with less practical significance from being shared with the public frequently (ASA, 2021). Some observe that due to a moderate reliance on statistical significance in journals, researchers can be motivated to take part in data manipulation (Franco et al., 2014), exhibited through Wansink. The controversial researcher was found to have committed misconduct in several of his studies, concluding several inaccuracies (Servick, 2018). Specifically, he conducted a study claiming that advertising popular cartoons on apples would influence children to select fruit instead of cookies at lunchtime (Wansink, 2012). Despite obtaining a *p*-value greater than 0.05, Wansink urged a student to falsely manipulate the data to obtain significant results (Lee, 2018)[2]. *See evidence of this in Figure 2 of the Appendix.* Despite knowingly contributing to the spread of misinformation, the reproducibility crisis also plays a hand in influencing Wansink's mentality associated with the heavy reliance on statistical significance.

Another tale of *p*-hacking is the infamous study delineating that chocolate promotes weight loss. Conducted with the intention of exhibiting the dangers of falsified or misleading studies surrounding diet culture in a documentary film, Dr. Johannes Bohannon and his team ran a clinical trial concluding accelerated weight loss in a group of individuals who had consumed a daily bar of dark chocolate. What news outlets, such as *Shape*, *Men's Health*, *Shape*, and *Times*

[2]Appendix: Figure 2

of India, failed to realize was the crucially small sample size. Bohannon, a science journalist, purposely planned the experiment to output a significant result:

> Here's a dirty little science secret: If you measure a large number of things about a small number of people, you are almost guaranteed to get a "statistically significant" result. Our study included 18 different measurements—weight, cholesterol, sodium, blood protein levels, sleep quality, well-being, etc.—from 15 people. (One subject was dropped.) That study design is a recipe for false positives (2015).

The intentional fraudulent study exhibited both the ease of data dredging and simplicity of spreading misinformation. Although several publications accepted the results, Bohannon compared an analysis of their results to the concept of reading tea leaves to determine the future.

While manipulators are to blame for the proliferation of the situation, it is also essential to combat the crisis by graduating from the dichotomization of *p*-values. To further demonstrate its impact, C. Glenn Begley and Lee M. Ellis urged publications to include nonsignificant studies to increase credibility, public awareness, and create more successful outreach. Regarding preclinical cancer studies, "There must be more opportunities to present negative data….Funding agencies, reviewers and journal editors must agree that negative data can be just as informative as positive data" (2012). Claiming trials and experiments are not reaching the radars of drug companies and other field scientists, Begley and Ellis propose more flexible guidelines for publication, specifically regarding replicable, valid data.

Acknowledging that *p*-hacking is a focal point of the publication crisis, data manipulation, however, extends to other disciplines and spans many transgressions. For one, financial motivations are a leading contributor to the spread of misinformation (Colomina,

Sanchez, & Youngs, 2021, p. 8), often occurring when a researcher is paid-off to ensure the data tells a one-sided story. Andrew Wakefield, a former physician with considerable influence on the public, published a study falsely evidencing the measles, mumps, and rubella (MMR) vaccine's connection with Autism Spectrum Disorder. The authors claimed to find a link between gastrointestinal disease and developmental regression in "a group of previously normal children," whose parents associated the MMR vaccine with behavioral symptoms indicative of autism (Wakefield, et al., 1998). Despite the authors releasing a statement defending the scope of the study and consequently retracting the article (Murch, et al., 2004), Wakefield and his cohort of authors were charged with ethical violations, scientific misrepresentation, fraud (Rao & Andrade, 2011), and guilty of financial deception. The study, which promoted vaccine skepticism (Motta & Stecula, 2021) and falsified data, was motivated by more than a £400,000 pay-out by lawyers seeking to show the vaccine was detrimental to society (Deer, 2006).

3. The Euro Problem

When conducting probability experiments with a coin, a typical hypothesis claims the coin is fair, and it is of interest to test if the probability of landing on heads is equivalent to that of tails. If the coin were not assumed to be fair, however, it is advisable to then make inference on the value of p, defined as the probability of success (or obtaining heads for a coin). Take, for example, The Euro Problem, concisely introduced by David MacKay in "Think Bayes 2e:"

> A statistical statement appeared in *The Guardian* on Friday January 4, 2002: When spun on edge 250 times, a Belgian one-euro coin came up heads 140 times and tails 110. 'It looks very suspicious to me', said Barry Blight, a statistics lecturer at the London School of Economics. 'If the coin were unbiased the chance of getting a result as extreme as that would be less than 7%'. But do these data give evidence that the coin is biased rather than fair? (2003)

In the following sections, we seek to answer the euro coin question from a non-Bayesian perspective, utilizing a standard normal distribution hypothesis test, and from a Bayesian perspective, utilizing a Beta-binomial model. In this case, data is the number of heads out of n = 250 tosses and the parameter is p, the probability of heads. For both approaches, we model the number of heads with a binomial distribution, which describes the probability of getting k successes out of n trials with the probability mass function, given by $P(k) = \binom{n}{k} p^k (1 - p)^{n-k}$. For a numeric example, if we take the success probability to be 0.5, assuming it is a fair coin, we obtain $\binom{250}{k} * 0.5^k (1 - 0.5)^{250-k}$. This function can be utilized to calculate the probability of, for example, obtaining exactly 150 heads out of 250 tosses, by substituting k = 150. There are

two assumptions to the binomial model: trials must be independent and the probability of success p must be constant from trial to trial. Since one coin flip does not influence another, we can assume each toss is independent from each other; since we are using the same coin, each toss has the same probability of getting heads, so probability of success is constant. Thus, the euro problem is robust to any violations of the two assumptions.

3.1. Non-Bayesian Approach

In The Euro Problem, the goal is to estimate p based on the provided data. A natural point estimate for p, denoted $\hat{p}$, is the sample proportion of heads (i.e., $\hat{p} = \frac{x}{n}$). *See an example distribution of $\hat{p}$ in Figure 1 of the Appendix.* In the case of the euro coin problem, $\hat{p}$ is 0.56, calculated by $\hat{p} = \frac{140}{250}$. To evaluate if the coin is biased, we conduct a hypothesis test on the null hypothesis (the true probability of heads is equal to 0.5), against a two-sided alternative (the true probability of heads is not equal to 0.5). Conforming to the notation style, H_0: $p = 0.5$ and H_1: $p \neq 0.5$. The test is based on the point estimate $\hat{p}$, with the test statistic z follows the formula: $z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$, such that p is the null hypothesis value (0.5 in this case). In the context of the euro problem, the test statistic is equivalent to 1.897, calculated by $z = \frac{(0.56) - (0.5)}{\sqrt{\frac{0.5*(1-0.5)}{250}}}$. Under the null hypothesis, the test statistic follows a standard normal distribution, called a reference distribution, as a result of the Central Limit Theorem.

The *p*-value is then obtained by doubling the value obtained from the corresponding value derived from a *z*-score table (Frost, 2022). In this case, *p*-value is 0.067 (with continuity

correction). Thus, we conclude, at the 0.05 level, that we do not have enough evidence to reject the null hypothesis. A notable drawback of the *p*-value is that it is not intuitive: a common misconception is to make a seemingly equivalent conclusion that we have "accepted" the null hypothesis and have "proved" it is a fair coin, but this is simply not true. The purpose of a hypothesis test for a non-statistically significant result is to conclude a lack of enough evidence for the alternative hypothesis; one cannot prove, with certainty, which hypothesis is true based on the *p*-value.

In addition to the hypothesis test, we can provide a confidence interval, expressed as a proportion, for p. In a similar manner, the confidence interval is based on the normal distribution of the point estimate p̂, and is given by $\widehat{p} \pm \left(Z_{(\frac{\alpha}{2})} \sqrt{\frac{\widehat{p}*(1-\widehat{p})}{n}} \right)$, where $Z_{(\frac{\alpha}{2})}$ is the critical value of a standard normal distribution such that the probability to the right of $Z_{(\frac{\alpha}{2})}$ is $\frac{\alpha}{2}$ ("One Sample Proportion," n.d.).

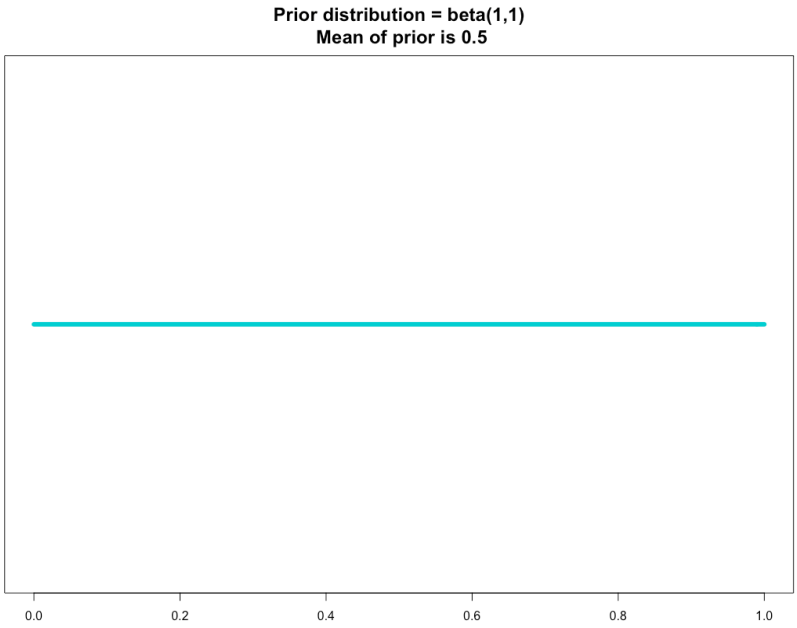
The confidence interval provides a range of values to accept if we were to conduct a hypothesis test. In this case, the 95% confidence interval is $(0.498, 0.622) = 0.56 \pm \left(1.960 \sqrt{\frac{0.56*0.44}{250}} \right)$. Relating to the hypothesis test, since a value of 0.5 is in the confidence interval, we do not have enough evidence to reject this value. Thus, the approach of a NHST provides a concise, clear-cut means of analysis in determining if the coin is fair.

3.2. Bayesian Approach

In the Euro problem, the key difference of a Bayesian approach is to regard the parameter p as a random variable, rather than a fixed, unknown constant. The first step is then to assign a probability distribution that describes the randomness in p, identified as the prior distribution in Bayesian analysis. *See an example of prior distributions in Figure 2 of the Appendix.* We then make inference by updating the prior distribution according to the collected data, which leads to the posterior distribution. Frequently, the beta distribution is used to represent the prior distribution of p, and has a range of (0,1). The shape of the distribution is determined by its two parameters, α and β ($\alpha > 0$, $\beta > 0$). Its expected value is given by $\mu = \frac{\alpha}{\alpha+\beta}$, with variance given by $\sigma^2 = \frac{\alpha*\beta}{(\alpha+\beta)^2*(\alpha+\beta+1)}$. A special case of the beta distribution is the uniform distribution (**Figure 1**), in which $\alpha = \beta = 1$. *See further examples of beta distributions in Figures 3,4, and 5 of the Appendix.*

Figure 1: Examples of Beta distributions

The uniform distribution indicates the probability of heads can be any value between 0 and 1, with equal probability.

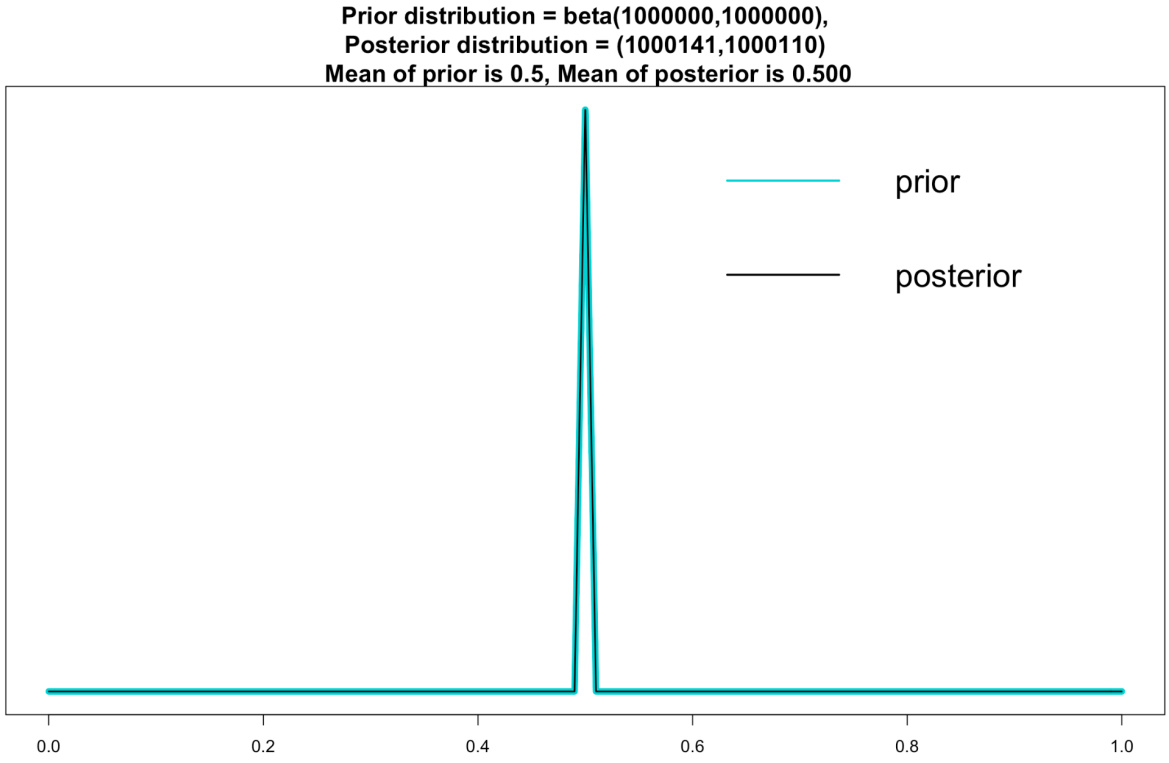


There are three main benefits of using a beta prior for the parameter p. Primarily, the shape of the distribution is flexible, depending entirely on the two aforementioned parameters; the range of positive values (0,1) is suitable to model probabilities. Second, when a beta prior is used, the posterior is also a beta distribution, and its derivation is mathematically simple: if the prior is beta(α, β), then the posterior is beta($\alpha + x$, $\beta + (n-x)$) (recall that x is the number of successes, or heads, and n-x is the number of failures, or tails). Lastly, it incorporates prior bias concerns. Simplistically, the choice of the prior distribution represents a prior belief on the skewness of the coin, chosen subjectively. It is typical to select smaller integers as parameters of the prior beta distribution, as larger parameter values correspond to strong prior beliefs that tend to not be affected much by data.

A prior distribution of beta(1000000, 1000000) is displayed in **Figure 2**, indicating a strong belief that the value of p is concentrated around 0.5. The posterior distribution is beta(1000140, 1000110), and does not differ much from the prior distribution. Due to the prior distribution's parameters being so large, it has a dominance over data, influencing the posterior mean to not largely differ from 0.5. We note that, when conducting Bayesian analysis, it is usually important to conduct sensitivity analysis to study the effect on different prior choices.

Figure 2: Updated posterior distribution with large prior parameters

The prior distribution centers around 0.5, indicating it is not a biased coin.



3.3. Simulations

To further contrast NHST's and Bayesian analyses, I conducted a series of simulations for The Euro Problem, utilizing two sets of data values: the first set with $x = 140$ and $n = 250$ (from the Euro coin example), and the second with $x = 150$ and $n = 250$. *P*-values and confidence intervals are provided in **Table 1** and **Table 2**. We note the second set of data only increases the number of heads by 10, but causes a drastic difference in the *p*-value and resulting conclusions. This data was chosen to exhibit another limitation of *p*-values: minor alterations of data can have a significant effect on the hypothesis testing conclusions because the *p*-value is typically compared to a strict, artificial cutoff. We compare the *p*-value conclusions to the results of Bayesian analysis based on a series of different prior choices. For each choice of the prior distribution, we calculate 95% credible intervals and determine whether the results of the Bayesian method indicated a fair coin: the conclusion is based on whether 0.5 was an element of the 95% credible interval. Results are shown in Table 1 and Table 2.

Table 1

Bayesian analysis of the euro problem with x = *140 and* n = *250*

Table 1:	Test results indicate that we do not have enough evidence to conclude it is not a fair coin, $t(249) = 1.897$, $p = 0.067$, 95% CI = [0.496,0.622].						
Row	**Prior**	**P(p>0.5) prior**	**Posterior**	**P(p>0.5) posterior**	**95% Cred.**	**Fair coin?**	**Consistent?**
1	beta(4,6)	0.254	beta(144,116)	0.959	(0.493, 0.614)	Y	Y
2	**beta(4,2)**	**0.813**	**beta(144,112)**	**0.978**	**(0.501, 0.623)**	**N**	**N**
3	beta(4,3)	0.656	beta(144,113)	0.974	(0.499, 0.620)	Y	Y
4	beta(4,4)	0.500	beta(144,114)	0.969	(0.497, 0.618)	Y	Y
5	beta(4,5)	0.363	beta(144,115)	0.965	(0.495, 0.616)	Y	Y
6	beta(10,12)	0.332	beta(150,122)	0.956	(0.492, 0.610)	Y	Y
7	beta(12,10)	0.668	beta(152,120)	0.974	(0.500, 0.617)	Y	Y
8	beta(200,2)	1.000	beta(340,112)	1.000	(0.711, 0.791)	N	N
9	beta(2,200)	0.000	beta(142,310)	0.000	(0.272, 0.358)	N	N
10	beta(200,100)	1.000	beta(340,210)	1.000	(0.577, 0.658)	N	N
11	beta(100,200)	0.000	beta(240,310)	0.001	(0.395, 0.478)	N	N
12	beta(1000,800)	1.000	beta(1140,910)	1.000	(0.535, 0.578)	N	N
13	beta(10000,12000)	0.000	beta(10140,12110)	0.000	(0.449, 0.462)	N	N
14	beta(100000,99999)	0.501	beta(100140,100109)	0.528	(0.498, 0.502)	Y	Y
15	beta(100000,70000)	1.000	beta(100140,70110)	1.000	(0.586, 0.591)	N	N
16	beta(12,11)	0.584	beta(152,121)	0.970	(0.498, 0.615)	Y	Y
17	beta(25,24)	0.557	beta(165,134)	0.964	(0.495, 0.608)	Y	Y
18	beta(25,23)	0.615	beta(165,133)	0.968	(0.497, 0.610)	Y	Y
19	beta(25,26)	0.444	beta(165,136)	0.953	(0.492, 0.604)	Y	Y

Table 2

Bayesian analysis of the modified euro problem with x = *150 and* n = *250*

Table 2: Tests results indicate we reject the null hypothesis and conclude the coin is not fair, $t(149) = 3.16227766$, $p = 0.002$, 95% CI = [0.536, 0.661].

Row	Prior	P(p>0.5) prior	Posterior	P(p>0.5) posterior	95% Cred.	Fair coin?	Consistent?
1	beta(4,6)	0.254	beta(154,106)	0.999	(0.532, 0.651)	N	Y
2	**beta(4,2)**	**0.813**	**beta(154,102)**	**0.999**	**(0.541, 0.661)**	**N**	**Y**
3	beta(4,3)	0.656	beta(154,103)	0.999	(0.539, 0.658)	N	Y
4	beta(4,4)	0.500	beta(154,104)	0.999	(0.539, 0.658)	N	Y
5	beta(4,5)	0.363	beta(154,105)	0.999	(0.534, 0.654)	N	Y
6	beta(10,12)	0.332	beta(160,112)	0.998	(0.529, 0.646)	N	Y
7	beta(12,10)	0.668	beta(162,110)	0.999	(0.537, 0.653)	N	Y
8	beta(200,2)	1.000	beta(350,102)	1.000	(0.735, 0.812)	N	Y
9	beta(2,200)	0.000	beta(152,300)	0.000	(0.293, 0.380)	N	Y
10	beta(200,100)	1.000	beta(350,200)	1.000	(0.596, 0.676)	N	Y
11	beta(100,200)	0.000	beta(250,300)	0.016	(0.413, 0.496)	N	Y
12	beta(1000,800)	1.000	beta(1150,900)	1.000	(0.539, 0.582)	N	Y
13	beta(10000,12000)	0.000	beta(10150,12100)	0.000	(0.450, 0.463)	N	Y
14	beta(100000,99999)	0.501	beta(100150,100099)	0.545	(0.498, 0.502)	Y	N
15	beta(100000,70000)	1.000	beta(100150,70100)	1.000	(0.586, 0.591)	N	Y
16	beta(12,11)	0.584	beta(162,111)	0.999	(0.535, 0.651)	N	Y
17	beta(25,24)	0.557	beta(175,124)	0.998	(0.535, 0.651)	N	Y
18	beta(25,23)	0.615	beta(175,123)	0.999	(0.531, 0.642)	N	Y
19	beta(25,26)	0.444	beta(175,126)	0.998	(0.525, 0.636)	N	Y

The results in Table 1 show limited discrepancies between Bayesian analysis and the NHST. The latter method indicates a lack of evidence to reject the null hypothesis: $t(249) = 1.897$, $p = 0.067$, 95% CI = [0.496,0.622]. In Row 2, the prior distribution beta(4, 2) leads to a posterior probability P(p>0.5) = 0.978, with a 95% credible interval of (0.501, 0.623). Since the result of the credible interval did not contain p = 0.5, there is enough reason to conclude the coin is heads-biased, directly contrasting the conclusion obtained from the *p*-value. It is worth noting that rows that are not consistent with the results of the NHST correspond to biased prior distributions.

The results of Table 2 convey a different message: most of the results of the Bayesian analysis are consistent with that of the NHST. The original hypothesis test indicates enough evidence to reject the null hypothesis and conclude the coin is not fair ($t(149) = 1.897$, $p =$

0.002). In Row 4, a weak prior distribution of beta(4, 4), which reflects a symmetric and unimodal distribution, results in a 95% posterior credible interval of (0.5387454, 0.6582353), thus concluding the coin is likely to be head-biased. In both trials, we obtain the posterior probability of the coin being Heads-biased (P(p>0.5)). In Table 1, the second row outputs a probability of the true sample proportion being greater than 0.5 as 0.978, which is a relatively high chance. However, as concluded based on the *p*-value, we do not have enough evidence to conclude it is not a fair coin. This highlights a discrepancy between the *p*-value and posterior probability. In the second row of Table 2, we see the probability of p being greater than 0.5 is 0.999. Significantly, the credible intervals and posterior probabilities in both cases suggest the coin is likely to be biased. This conclusion highlights that, unlike NHST's, Bayesian analysis neither relies on fixed cut-offs nor provides unambiguous answers regarding the fairness of the coin.

In summary, in the Bayesian setting, researchers do not seek to make a dichotomous decision (i.e., reject or fail to reject the null hypothesis); rather, the Bayesian analysis views p as a continuous random variable and, thus, not constant. This is associated with a new way of thinking that can provide some uncertainty assessment in the distribution of p. While Bayesian analysis can associate probabilities for specific statements, such as the probability that p is greater than 0.5, Hypothesis testing and *p*-values, conversely, do not follow the same methodology. While *p*-values are commonly mistaken for an associated probability, they instead only analyze the strength of a conclusion. In addition, when comparing the results of the NHST's in Tables 1 and 2, we see that increasing the number of heads caused a dramatic effect on the *p*-value. While it is a seemingly insignificant change of data, it drastically alters the significance

of the NHST. The lack of prior distribution in NHST implies that its conclusion is solely based on the data. In comparison, the Bayesian method adds another layer of information in the form of the prior distribution.

4. Application

While the euro coin problem is a simple application of the beta binomial model, it can be implemented to a plethora of real-world topics, such as COVID-19 analyses. In "A primer on Bayesian estimation of prevalence of COVID-19 patient outcomes" (Gao and Dong, 2020), the parameter p is the probability of COVID-19 death in Iceland (instead of the probability of heads in the coin example); the number of deaths (y) replaced the number of heads (x); the total number of infected (N) replaced the total number of rolls. A uniform prior was used and the following conclusions were obtained: there is a 95% probability that the true infection fatality for those younger than 70 years lie within (0.04%–0.29%), given the evidence provided by the observed data, the true infection fatality for those between 70 and 80 years lie within (0.85%–6.65%), and the true estimate of infection fatality for those older than 80 years lie within (4.30%–24.22%). In another application, the parameter p is the percentage of the asymptomatic children in the United States and data is the the number of asymptomatic children out of the total amount of infected children. Similarly, a uniform prior was used and 95% posterior credible intervals were reported to be (0.66%-0.94%) for the Western region, (0.56%-1.04%) for the Midwest region, (0.44-0.78) for the Southern region, and (0.73%-1.33%) for the Northeast region.

Gao and Dong (2020) also examined two COVID-19 studies that utilized Bayesian analysis on two previously published studies to provide a different perspective on the data. The first cited article, "Humoral Immune Response to SARS-CoV-2 in Iceland" (Gudbjartsson et al., 2020) focuses on the risk of infection fatality measured through antibodies in individuals diagnosed with, exposed to, and not exposed to SARS-CoV-2. Generalizing to the population of

Iceland, research sought to assess SARS-CoV-2 seroprevalence — the proportion of individuals with antibodies against SARS-CoV-2 in a given population — within four months of infection. Notably, the sample seems to be representative, as samples are weighted by specific demographics and have adequate sample sizes. The second cited article, "Prevalence of SARS-CoV-2 Infection in Children Without Symptoms of Coronavirus Disease 2019" (Sola et al., 2020) analyzes the rate of positive cases for asymptomatic children treated in hospitals for conditions other than SARS-CoV-2. Deriving the data from Johns Hopkins University's database, the researchers obtained the amount of asymptomatic individuals under 18 who tested positive at hospitals, and calculated prevalence proportions based on individual hospitals and their corresponding regions. Results indicated that 0.7% of asymptomatic children were infected with SARS-CoV-2 during the time period, with low prevalence. The total pooled prevalence percentage was 0.65%, associated with a 95% confidence interval (0.47%, 0.83%).

Throughout the article, Gao and Dong caution on the use of point estimates, uncertainty, and confidence intervals in non-Bayesian methods. Point estimation is warned to increase the uncertainty of specific inferred values. Instead, authors suggest employing a probability distribution to look at a series of probabilities, rather than one value. Their next point references the misconception that *p*-values and confidence intervals are associated with the probability of a specific event occurring, which is incorrect. As mentioned earlier, one cannot prove, with certainty, which hypothesis is true. Lastly, the authors discuss the importance of prior knowledge to influence more honest estimations. Overall, they "advocate the use of Bayesian methods for researchers who work in this important field for COVID-19 research, as it enables them to overcome the above limitations by deriving a probability for every possible value of the

unknown parameter of interest." Thus, while *p*-values and other non-Bayesian methods used in biostatistical analysis are not incorrect, there are advantages in incorporating probabilistic Bayesian inference.

The use of Bayesian methods extend beyond the aforementioned studies to many other applications. In "Multiplicity Eludes Peer Review: The Case of COVID-19 Research," Oliver Gutiérrez-Hernández and Luis Ventura García (2021) conducted an investigation into popular COVID-19 observational studies to analyze different statistical analyses and their associated prevalence. The main focus of the study was to target multiplicity, and the dangers of false-discoveries. The former is the increased risk of a type I error, previously defined as the probability of rejecting a true null hypothesis (α), due to a large amount of simultaneous inferences or tests. To reduce the spread of misinformation and risks of multiplicity, the authors suggest amendments to investigated studies utilizing *p*-values without the acknowledgment of potential errors or limitations. From their sample size of 100 peer-reviewed articles, the authors discovered that 99 articles citing multiple *p*-values in the main text failed to conduct any method of controlling for potential type I errors, concluding a potential prevalence of apocryphal findings.

5. Conclusion

This research intends to capture the misunderstandings surrounding NHST's and *p*-values, and compare these methods to the alternative Bayesian methods. Inspired by the debate between subjectivism and frequentism (Benhamini et al., 2021), I seek to detail the benefits of Bayesian analysis with the limitations and cautions regarding *p*-values in mind.

On one hand, Bayesian analysis provides an arguably more intuitive and probabilistic analysis of the hypothesis, promoting an alternative way to measure the strength of data. Conversely, *p*-values rely on a fixed threshold to make a conclusion, usually without offering more information. Moreover, lack of a prior distribution in NHST's indicates the conclusion is solely based on the data and can be very sensitive to minor changes in the data. *P*-values and confidence intervals provide valid reasonings for tests of proportions — the statistic provides a preliminary analysis for significance for one- and two-tailed tests, and confidence intervals yield a range of values to accept upon conducting a hypothesis test (du Prel et al., 2009). In comparison, Bayesian analysis takes a different approach with the use of probabilistic inference, holding influence on uncertainty, and based on probability theory (Chater & Oaksford, 2012). Bayesian statistics treats the parameter as random and studies its posterior distribution by considering both the prior belief and the data.

Despite justifying my inclination to practice Bayesian methods, I argue the choice of approach lies in the researcher, especially when conducting both analyses adequately. This is achieved, when using *p*-values, by avoiding dichotomization and implementing an understanding of power and effect size, and, when using Bayesian methods, by discussing the sensitivity of prior choices. A. Blasco, a professor in the Department of Animal Science at Valencia

Polytechnic University, discusses the choice of technique both in theoretical and practical approaches. Concluding his argument, he writes:

> Both Bayesian and frequentist schools of inference are well established…. To choose one school or the other should be related to whether there are solutions in one school that the other does not offer, to how easily the problems are solved, and to how comfortable the scientist feels with the particular way of expressing the results. Both schools present formal problems and paradoxes, although problems and paradoxes with methods of greater familiarity are often better tolerated (2001).

Interpreting Blasco's thoughts in the context of modern day statistics colloquies, the choice of method relies on one's comfort level. Both approaches have their benefits and drawbacks, and whichever method is more understood by the researcher and exhaustive in its analysis, is suitable. Through my examples, simulations, and applications, I have deemed Bayesian statistics to be more beneficial. Nevertheless, much like the branch of statistics, the incorporation of one's opinion is both integral and welcomed.

6. References

American Statistical Association. (2021, August 1). *ASA President's Task Force Statement on Statistical Significance and Replicability.* AMSTAT News. https://magazine.amstat.org/blog/2021/08/01/task-force-statement-*p*-value/

Blasco A. (2001). The Bayesian controversy in animal breeding. J*ournal of animal science, 79*(8), 2023–2046. https://doi.org/10.2527/2001.7982023x

Bohannon, J. *I Fooled Millions Into Thinking Chocolate Helps Weight Loss. Here's How.* Gizmodo, https://gizmodo.com/i-fooled-millions-into-thinking-chocolate-helps-weight-1707251800

Chater, N., & Oaksford, M. (2008). *The Probabilistic Mind: Prospects for Bayesian cognitive science.* Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199216093.001.0001

Cobb, G. W. (2015), "Mere Renovation is Too Little Too Late: We Need to Re- think Our Undergraduate Curriculum from the Ground Up," *The American Statistician, 69*(4), 1-33. https://doi.org/10.1080/00031305.2015

Colomina, C., Sánchez Margalef, H., & Youngs, R. (2021). *The Impact of Disinformation on Democratic Processes and Human Rights in the World*. European Parliament. https://www.europarl.europa.eu/RegData/etudes/STUD/2021/653635/EXPO_STU(2021)653635_EN.pdf.

Deer, B. (2006, December 31). MMR doctor given legal aid thousands. *The Sunday Times*. Retrieved from https://briandeer.com/mmr/st-dec-2006.htm

Dickersin, K., & Min, Y. I. (1993). Publication bias: the problem that won't go away. *Annals of the New York Academy of Sciences, 703*, 135–148. https://doi.org/10.1111/j.1749-6632.1993.tb26343.x

du Prel, J. B., Hommel, G., Röhrig, B., & Blettner, M. (2009). Confidence Interval or *P*-Value?: Part 4 of a Series on Evaluation of Scientific Publications. *Deutsches Arzteblatt International, 106*(19), 335–339. https://doi.org/10.3238/arztebl.2009.0335

Dusheck, J. (2016, March 15). *Misleading* p-*values showing up more often in biomedical journal articles*. Stanford School of Medicine. https://med.stanford.edu/news/all-news/2016/03/misleading-*p*-values-showing-up-more-often-in-journals.html

Fisher, R. A. & Russell, E. J. (1922). *On the Mathematical Foundations of Theoretical Statistics*. Springer New York. DOI: doi: 10.1098/rsta.1922.0009

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science, 345*(6203), 1502-1505. https://doi.org/10.1126/science.1255484

Frost, J. (2022, November 10). *Z-table.* Statistics By Jim, https://statisticsbyjim.com/hypothesis-testing/z-table/

Gelman, A. & Stern, H. (2006, November). The Difference Between “Significant” and “Not Significant” is not Itself Statistically Significant. *The American Statistician 60*(4). https://doi.org/10.1198/000313006X152649

Gudbjartsson, D. F., Norddahl, G. L., Melsted, P., Gunnarsdottir, K., Holm, H., Eythorsson, E., Arnthorsson, A. O., Helgason, D., Bjarnadottir, K., Ingvarsson, R. F., Thorsteinsdottir, B.,

Kristjansdottir, S., Birgisdottir, K., Kristinsdottir, A. M., Sigurdsson, M. I., Arnadottir, G. A., Ivarsdottir, E. V., Andresdottir, M., Jonsson, F., Agustsdottir, A. B., … Stefansson, K. (2020). Humoral immune response to SARS-COV-2 in Iceland. *New England Journal of Medicine, 383*(18), 1724-1734. https://doi.org/10.1056/NEJMoa2026116

Gutiérrez-Hernández, O., & García, L. V. (2021). Multiplicity Eludes Peer Review: The Case of COVID-19 Research. *International journal of environmental research and public health, 18*(17). https://doi.org/10.3390/ijerph18179304

Halsey, L. G. (2019). The reign of the *p*-value is over: What alternative analyses could we employ to fill the power vacuum?. *Biology letters, 15*(5). https://doi.org/10.1098/rsbl.2019.0174

Lakens, D. (2021). The Practical Alternative to the *p* Value Is the Correctly Used *p* Value. *Perspectives on Psychological Science, 16*(3), 639–648. https://doi.org/10.1177/1745691620958012

Lee, S. M. (2018, September 20). *Cornell Just Found Brian Wansink Guilty Of Scientific Misconduct And He Has Resigned*. Buzzfeed News. https://www.buzzfeednews.com/article/stephaniemlee/brian-wansink-retired-cornell

Lytsy, P., Hartman, M., & Pingel, R. (2022). Misinterpretations of *P*-values and statistical tests persist among researchers and professionals working with statistics and epidemiology. *Upsala Journal of Medical Sciences, 127*(1). https://doi.org/10.48101/ujms.v127.8760

MacKay, D. J. C. (2003). *Information theory, inference and learning algorithms.* Cambridge University Press.

Motta, M. & Stecula, D. (2021) Quantifying the effect of Wakefield et al. (1998) on skepticism about MMR vaccine safety in the U.S. *PLoS ONE, 16*(8): e0256395. https://doi.org/10.1371/journal.pone.0256395

Murch, S. H., Anthony, A., Casson, D. H., Malik, M., Berelowitz, M., Dhillon, A. P., Thomson, M. A. Valentine, A., Davies, S. E., & Walker-Smith, J. A. (2004). Retraction of an interpretation. *The Lancet 363*(9411), https://doi.org/10.1016/S0140-6736(04)15715-2

Nahm F. S. (2017). What the *P* values really tell us. *The Korean Journal of Pain, 30*(4), 241–242. https://doi.org/10.3344/kjp.2017.30.4.241

Neyman, J. & Pearson, E. S. (1933). IX. On the Problem of the most Efficient Tests of Statistical Hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, *231*, 694-706. https://doi.org/10.1098/rsta.1933.0009

Rao, T. S., & Andrade, C. (2011). The MMR vaccine and autism: Sensation, refutation, retraction, and fraud. *Indian journal of psychiatry*, *53*(2), 95–96. https://doi.org/10.4103/0019-5545.82529

Servick, K. (2018, September 21). *Cornell Nutrition Scientist Resigns after Retractions and Research Misconduct Finding.* Science, https://www.science.org/content/article/cornell-nutrition-scientist-resigns-after-retractions-and-research-misconduct-finding

Sola, A. M., David, A. P., Rosbe, K. W., Baba, A., Ramirez-Avila, L., & Chan, D. K. (2021). Prevalence of SARS-CoV-2 Infection in Children Without Symptoms of Coronavirus

Disease 2019. *JAMA pediatrics, 175*(2), 198–201. https://doi.org/10.1001/jamapediatrics.2020.4095

Szucs, D., & Ioannidis, J. P. A. (2017). When Null Hypothesis Significance Testing Is Unsuitable for Research: A Reassessment. *Frontiers in Human Neuroscience, 11*(390). https://doi.org/10.3389/fnhum.2017.00390

Wakefield, A. J., Murch, S. H., Anthony, A., Linnell, J., Casson, D. M., Malik, M., Berelowitz, M., Dhillon, A. P., Thomson, M. A., Harvey, P., Valentine, A., Davies, S. E., & Walker-Smith, J. A. (1998). Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children. *Lancet (London, England)*, *351*(9103), 637–641. https://doi.org/10.1016/s0140-6736(97)11096-0 (Retraction published Lancet. 2010 Feb 6;375(9713):445)

Wansink B., Just D. R., & Payne C. R. (2012). Can Branding Improve School Lunches? *Archives of Pediatric Adolescent Medicine, 166*(10), 967–968. https://doi.org/10.1001/archpediatrics.2012.999

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA statement on *p*-values: Context, process, and purpose. *The American Statistician, 70*(2), 129–133. https://doi.org/10.1080/00031305.2016.1154108

Wasserstein, R. L., Schirn, A. L., & Lazar, N. A. (2019, March 20). Moving to a World Beyond "$p < 0.05$." *The American Statistician, 73*(1), 1-19. https://doi.org/10.1080/00031305.2019.1583913

Weiss, N.A. (2012) *Introductory Statistics*. Boston: Addison-Wesley Pearson Inc.

Weinberg, S., & Abramowitz, S. (2002). *Data analysis for the behavioral sciences using SPSS*. Cambridge University Press.

Xiang G., & Qunfeng D., (2020). A primer on Bayesian estimation of prevalence of COVID-19 patient outcomes, *JAMIA Open, 3*(4), 628–631, https://doi.org/doi.org/10.1093/jamiaopen/ooaa062

8.1 - One Sample Proportion. (n.d.) PennState: Eberly College of Science. Retrieved March, 2023, from https://online.stat.psu.edu/stat200/lesson/8/8.1

7. Appendix

Table 1

Distribution of p-*value and Cohen's* d *with mean difference of 5*

Sample size	Mean difference	Standard Deviation	P-value	Cohen's d
30	5	10	0.03	-0.497
30	5	10	0.013	-0.594
30	5	10	0.001	-0.81
30	5	10	0.01	-0.616
30	5	10	0.051	-0.429
100	5	10	<0.001	-0.528
100	5	10	<0.001	-0.463
100	5	10	<0.001	-0.617
100	5	10	<0.001	-0.513
100	5	10	0.015	-0.307
100	5	10	0.015	-0.309

Table 2

Distribution of p-*value and Cohen's* d *with mean difference of 2*

Sample size	Mean difference	Standard Deviation	P-value	Cohen's d
100	2	10	<0.001	-0.829
100	2	10	0.463	-0.013
100	2	10	0.018	-0.337

Figure 1

Power Analysis demonstrating there is only a 76.686% chance that p *is less than 0.05, assuming the alternative hypothesis is true.*

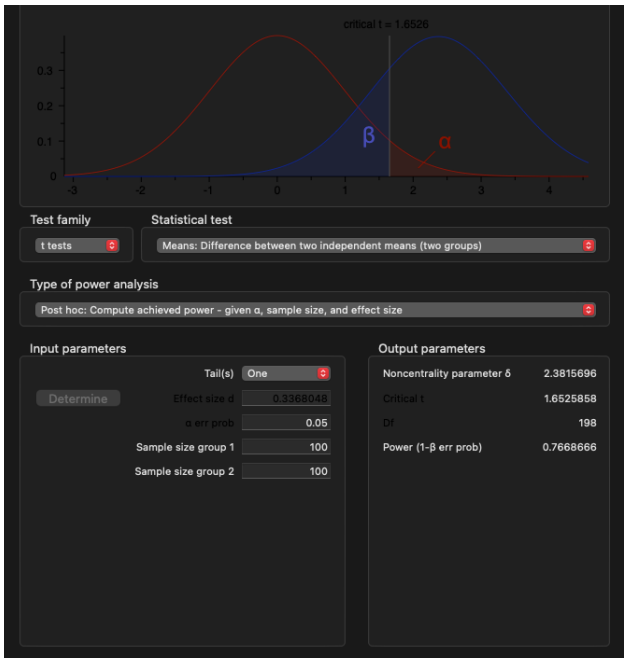


Figure 2

Email from Brian Wansink, provided by Buzzfeed News

Hi David,

Here's the Elmo study we are going to spin off and submit.
I think we start with the AJPH as a Brief (80 word abstract and 800 word paper), and go from there. I'll give Sandra a list of the journals and the priority order we should consider. Let's consider these two first:

Brief -- American Journal of Public Health

Research Letter – Archives of Pediatric and Adolescent Medicine

One sticking point is that although the stickers increase apple selection by 71%, for some reason this is a p value of .06. It seems to me it should be lower. Do you want to take a look at it and see what you think. If you can get the data, and it needs some tweeking, it would be good to get that one value below .05.

Best,

Brian

Figure 1

Probability distribution for $\hat{p}$

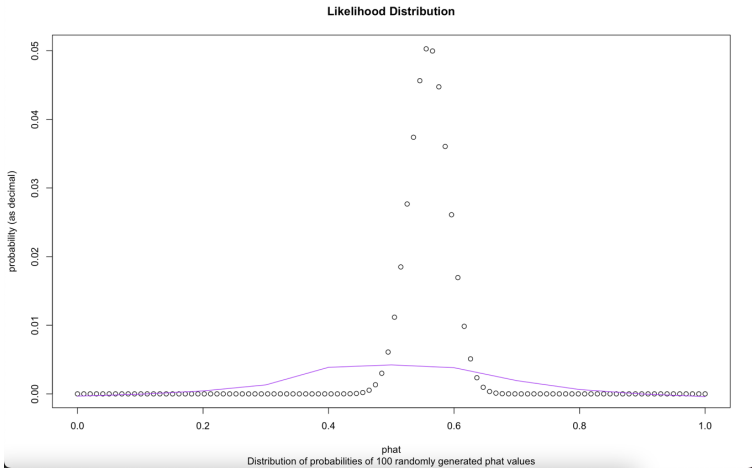


Figure 2

Prior distributions simulated from alpha and beta values with ranges [1,4]

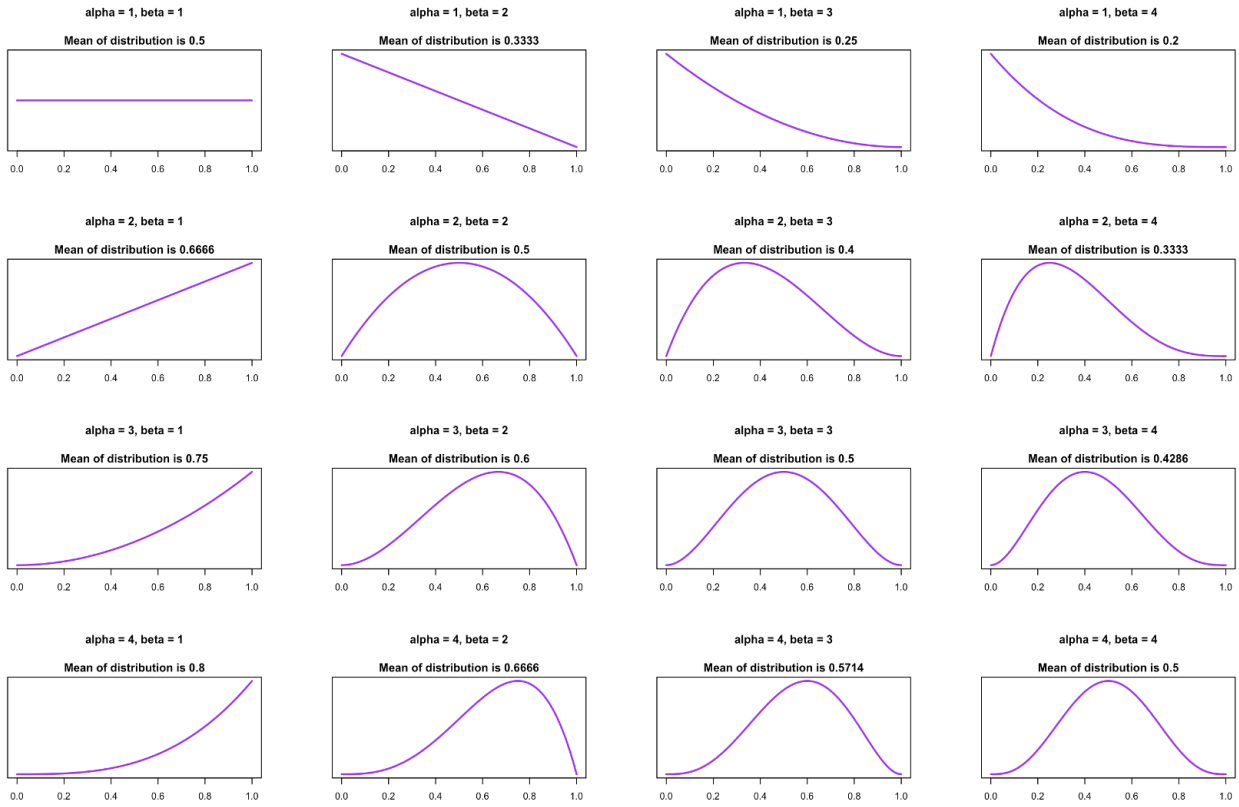


Figure 3

Uniform prior distribution and its corresponding posterior distribution

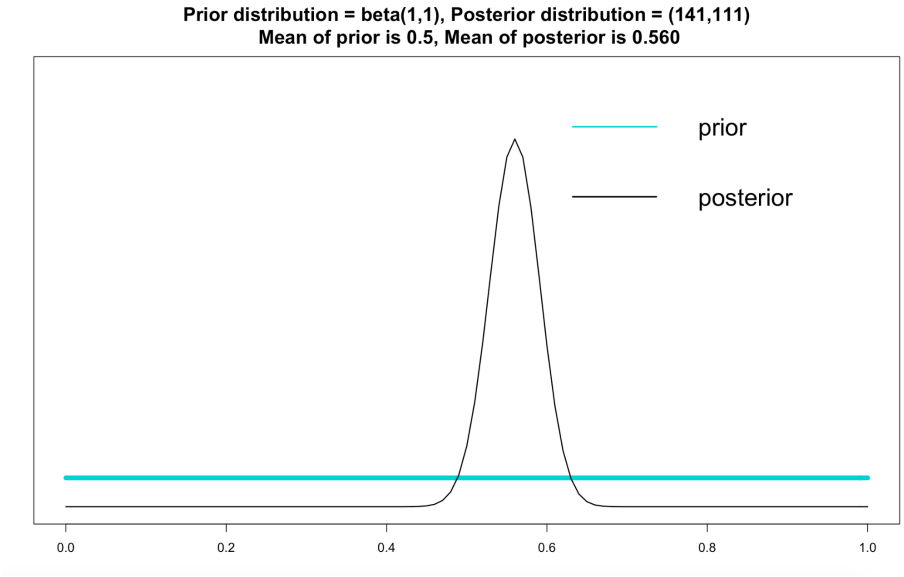


Figure 4

Prior distribution of beta(4, 2) and its corresponding posterior distribution

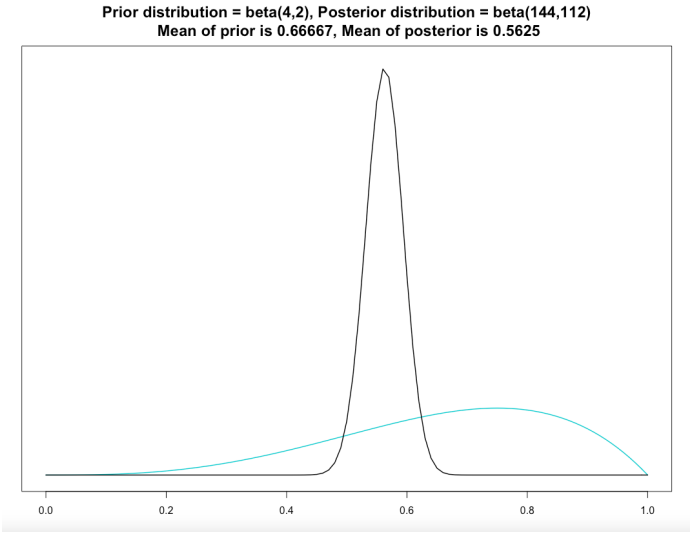


Figure 5

Prior distribution of beta(1000000, 1000000) and its corresponding posterior distribution