Frequentist Statistical Inference without Repeated Sampling

Paul Vos*& Don Holbert†

December 9, 2021

Department of Biostatistics, East Carolina University

Abstract

Frequentist inference typically is described in terms of hypothetical repeated sampling but there are advantages to an interpretation that uses a single random sample. Contemporary examples are given that indicate probabilities for random phenomena are interpreted as classical probabilities, and this interpretation of equally likely chance outcomes is applied to statistical inference using urn models. These are used to address Bayesian criticisms of frequentist methods. Recent descriptions of p-values, confidence intervals, and power are viewed through the lens of classical probability based on a single random sample from the population.

Keywords: classical probability, equally likely outcomes, statistical ensemble, multiset, p-value, confidence interval.

*corresponding author: vosp@ecu.edu, Biostatistics, ECU, Greenville NC 27858 USA
†hobertd@ecu.edu

1 Introduction

Frequentist inference, as a subset of statistical inference, appears to require hypothetical repeated sampling. Cox (2006, page 8) describes frequentist inference as follows:

> Arguments involving probability only via its (hypothetical) long-run frequency interpretation are called *frequentist*. That is, we define procedures for assessing evidence that are calibrated by how they would perform were they used repeatedly. In that sense they do not differ from other measuring instruments.

The entry "Frequency Interpretation in Probability and Statistical Inference" in the *Encyclopedia of Statistical Sciences* (ESS) also restricts the interpretation to repeated trials.

> ... ordinary people ... [and] many professional people, both statisticians and physicists, ... will confine themselves to probabilities only in connection with hypothetically repeated trials. (Sverdrup, 2006)

Without proper context these quotes could misrepresent these authors as only concerned with long-run behavior. Cox (2006) recognizes the importance of interpreting specific data.

> We intend, of course, that this long-run behavior is some assurance that with our particular data currently under analysis sound conclusions are drawn. This raises important issues of ensuring, as far as is feasible, the relevance of the long run to the specific instance.

We contend that the probability used to describe results from a particular study should not be restricted to the interpretation of hypothetical repeated trials. Studies can be more effectively described by using probability in the classical sense of equally likely out comes.

Bayesians and philosophers see problems with probability interpretations that use hypothetical repeated trials. We describe settings where these problems are lessened, but we agree that in other cases it is better to interpret probability using equally likely outcomes. The dialog between frequentist and Bayesian statisticians will improve when frequentists recognize the validity of concerns regarding hypothetical repeated trials, and when Bayesians recognize that their criticism applies to an interpretation of a frequentist method and not to the method itself.

The classical interpretation of probability is not without criticism. The entry "Foundations of Probability" in the *Encyclopedia of Biostatistics* states

> Though influential in the early development of the subject, and still valuable in calculations, the classical view fails because it is seldom applicable. (Lindley, 2005)

When 'probability' describes epistemic uncertainty, as it does in Bayesian inference, the classical view of 'equally likely' is of limited use. However, stochastic probabilities viewed as proportions fit naturally in the context of statistical inference. Introductory texts use 'frequency' and 'relative frequency' interchangeably with 'count' and 'proportion', respectively.[1] In a population, the proportion of individuals having a certain characteristic provides the same numerical value as the probability that a single randomly chosen individual will have that characteristic.

Requiring that frequentist inference include repeated trials is unnecessary in all, or nearly all, situations. Interpreting probabilities simply as proportions will allow frequentists to better communicate p-values and other inferential concepts. In addition, more

[1] See, for example, Johnson (1996) pages 22 and 23.

substantial discussions between frequentists and Bayesians will occur when the criticism that long-run behavior is not relevant to a specific instance is addressed by a probability interpretation that does not require repeated sampling.

2 Frequentist Statistical Inference

Before discussing how probability interpretations are used in statistical inference, we give a brief description of the latter. Cox and Hinkley (2000) and Romeijn (2017) provide a more thorough description of statistical inference. Our description is incomplete but will lay the groundwork for the role probability interpretations play in frequentist inferential methods. We note two salient features of frequentist statistical inference: *randomization* is used to produce data and *probability only describes the data* for a given model – not the probability that the model is true or correct. Probability interpretations that recognize these features avoid difficulties that arise in more general settings. Section 3 provides details on the interpretations used in frequentist statistical inference.

2.1 Inference for a Deck of Cards

A standard poker deck consists of 52 cards 13 of which are hearts. There are $\binom{52}{5}$ possible five-card hands. If X is the number of hearts in a five-card hand then the proportion of hands with x hearts is given by the hypergeometric distribution

$$\text{Proportion of hands with } x \text{ hearts when deck contains 13 hearts} = \frac{\binom{13}{x}\binom{52-13}{5-x}}{\binom{52}{5}}.$$

This becomes an inference problem when we do not know the number of hearts in the

deck. Consider a simplified deck of 52 cards where the cards are identical on the back and on the front each card has a unique numeral label from 1 to 52. In addition to the label, each card may or may not have a heart. We are presented with a deck of 52 cards with no information regarding the number of hearts in the deck. The deck is thoroughly shuffled and a five-card hand is dealt, two of which contain a heart, $x = 2$. What can we infer about the number of hearts in the deck?

There are 53 possible decks: $D_0, D_1, \ldots, D_{52}$ where D_h is the deck where h of the cards are hearts. The deck from which a hand with 2 hearts was dealt, D_{pop}, cannot be D_0 or D_1 because these decks have less than 2 hearts. Likewise, D_{pop} cannot be D_{50}, D_{51}, or D_{52} since these decks have fewer than 3 blank cards. What about the other decks? If the population deck had just 2 hearts, it is possible but not very likely that these 2 hearts show up in the hand that was dealt. If the population had 3 hearts, the additional heart makes the hand with 2 hearts somewhat more likely but still improbable. The hand with 2 hearts provides evidence against the claim that the population has 3 hearts ($D_{pop} = D_3$), and more evidence against the claim of 2 hearts ($D_{pop} = D_2$). Similarly, the hand with 2 hearts provides evidence against the claim of 48 hearts and even more evidence against the claim of 49 hearts.

The frequentist statistician formalizes these ideas using hypergeometric models for each of the possible decks

$$\text{Proportion of hands with 2 hearts when deck contains } h \text{ hearts} = \frac{\binom{h}{2}\binom{52-h}{5-2}}{\binom{52}{5}},$$

where $h = 2, 3, \ldots, 49$. If the population deck has few hearts, say 2 or 3, observing 2 hearts

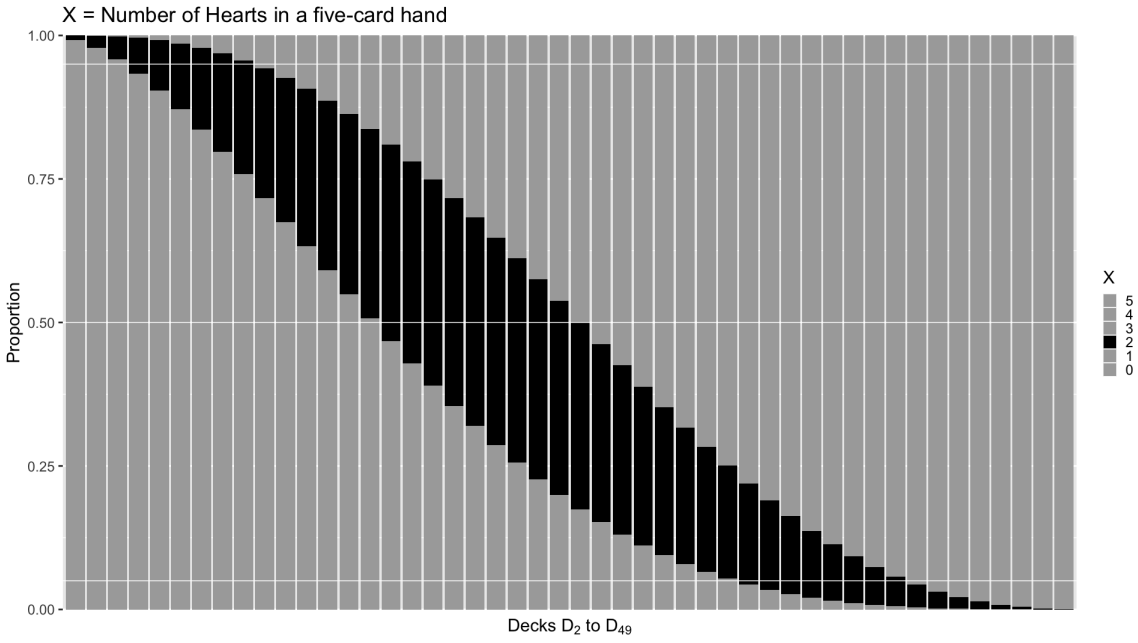

Figure 1: The left most bar presents the distribution of the number of hearts in five-card hands for D_2, the deck with 2 hearts. Nearly all the hands have 0 or 1 hearts (the height of gray bar gives the proportion of all such hands) and the remainder of the hands have 2 hearts (height of the black bar gives the proportion). Hands with 3, 4, or 5 hearts are not possible with deck D_2. The next bars present the corresponding distributions for decks D_3, D_4,..., D_{48}, and D_{49}. Each bar consists of 6 segments, one for each possible value of X. Since $x = 2$ was observed we use only two colors: black to represent the observed value and gray for values that are smaller (0 and 1) or larger (3, 4 or 5). The horizontal line at 0.50 intersects the black bars for decks D_{17} through D_{26} indicating the observed value of 2 hearts is the median and so represents no evidence against these models. The remaining two horizontal lines are at 0.05 and 0.95, the traditional values used to indicate samples sufficiently extreme as to provide significant evidence against the model.

is a surprisingly large number of hearts. On the other hand, if the population deck has many hearts, observing 2 hearts is a surprisingly small number. The observation of 2 hearts is compared to each of the models by calculating how extreme 2 hearts would be for that distribution. The evidence of 2 hearts against a particular model is measured by how far 2 hearts is in the tail of that distribution. Specifically, the evidence is measured by the p-value, the proportion of hands in the tail of the distribution. The observation of 2 hearts provides evidence against any deck for which the p-value is small.

Note that each hypergeometric model specifies a distribution of cards; it is not used to model random phenomena. If the cards in D_{pop} were not randomized, we would not be justified in treating all five-card hands as equally likely. If the cards were ordered and all the hearts came first, $D_{pop} = D_2$, and if all the hearts came last, $D_{pop} = D_{49}$[2].

Figure 1 displays the distribution of X for each of the possible models D_2 through D_{49}. For bars on the left, representing decks with a small number of hearts, the p-value is the height of the black segment together with the gray segment appearing above the black bar (indicating hands with 3, 4, or 5 hearts). For bars on the right, representing decks with a large number of hearts, the p-value is the height of the black segment together with the gray segment appearing below the black bar (indicating hands with 0 or 1 hearts). The middle bars that intersect the horizontal line where the proportion is 0.50 represent decks D_{17} through D_{26}. For these decks, observing 2 hearts is the median, the least extreme observation, and so observing a hand with 2 hearts provides no evidence against these decks.

While it would be natural to say the population deck *probably* is not the deck with only

[2]There is no way to order the decks D_0, D_1, D_{50}, D_{51}, and D_{52} to obtain a hand with 2 hearts, so that randomization is not required to eliminate these decks as being the population deck. The conclusions regarding these five decks are obtained by deduction, not statistical inference.

2 hearts, the frequentist statistician is careful to distinguish when the word 'probability' is used to describe random events compared to epistemic uncertainties. This distinction is made by using the term 'confidence'. Having observed a hand with 2 hearts, we are 95% confident that the deck contains at least 5 hearts. *Confidence* is a technical term that in this case means: either the deck has 5 or more hearts, or the hand we got is among the upper 5% of all possible hands. In Figure 1 the observation of a hand with 2 hearts is above the 95th percentile for the three bars on the left that correspond to decks with less than 5 hearts. At this point it is useful to describe the other main approach to statistical inference, Bayesian inference.

If we changed the problem to say that the population deck had been randomly selected in such a way that each of the 53 decks had an equal chance of being selected, then the frequentist statistician would use Bayes theorem to find the probability that $D_{pop} = D_h$ for $h = 0, 1, 2, \ldots, 52$. Using Bayes theorem in the original problem where we are given no information on how the population deck was chosen is an example of Bayesian inference. Bayesian inference does not require that each deck is equally likely, only that there is a probability distribution over the possible decks and this distribution need not be generated by randomization. This is the prior distribution. In some contexts, assigning equal probability to each deck when there has been no randomization, is considered part of classical probability. This is not how we use that term; classical probability only applies to equally likely sample outcomes obtained from randomization.

2.2 The P-value and Model Percentiles

Since the p-value is an important tool of frequentist inference and since it is often misunderstood, we emphasize two key elements regarding its definition and its interpretation.

One, the p-value is computed from a model for the population distribution, not the actual distribution. Two, the p-value measures how extreme the observed sample is compared to the distribution of values specified by the model and this measure does not involve randomization.

The p-value is a numerical measure relating an observed sample $x_1, x_2, \ldots, x_n$ to a specified distribution F (the model). The definition is

$$p\text{-value for } F = P(\{(y_1, y_2, \ldots, y_n) : T(y_1, y_2, \ldots, y_n) \geq T(x_1, x_2, \ldots, x_n)\})$$

where T is a real-valued function, P is the set function defined by

$$P(A) = \int_A dF^{(n)},$$

and $F^{(n)}$ is the product measure defined by the model F[3]. Because P satisfies Kolmogorov's axioms it is a probability function and so the p-value for F is a probability.

We illustrate these points with an artificial example where there is just one observation and $T(x) = x$. Suppose information is transmitted about a patient. The transmission is incomplete and contains only the person's age, 45, and their height, 5' 3". The person's sex is not known. However, if the person were male, he would be rather short. Using the population of men we can find the proportion of males that are 5' 3" or shorter. This is the p-value for the observed height for this population of men. This requires no sampling from the population as there are tables for the height percentiles of this population (these tables provide the model). Effectively, we imagine all men are ordered from shortest to tallest

[3]The sample space of the hypergeometric model used in the card example is finite and so this integral is a sum.

and the height of 5' 3" specifies a percentile for this distribution. This height is the 2nd percentile; that is, 2% of all men are 5' 3" or shorter[4]. The p-value for the corresponding population of females will be a larger value, 5' 3" is not especially short for a woman, but the interpretation is the same and does not require random sampling. It is reasonable to infer that this is a female patient since individuals this short are uncommon in the male population.

It is true that if we repeatedly sampled the male population, the proportion standing 5' 3" or less would be close to .02 and would equal this value in the limit. But this is not helpful. Our interest is in the population, not a random process, and stating that 2% of males in this population are 5' 3" or less provides a much clearer description of the population.

The previous example considered inference for an individual and so is not an example of statistical inference which is concerned with features of a collection of individuals, i.e., the population. So we now consider a sample of 50 individuals from a population for which the distribution of height is unknown. We are interested in the mean height of the population. There are a few differences from the patient example but the logic is the same. The sampling distribution of the population consists of all samples of size 50 and these are ordered from shortest mean height to tallest mean height. We observe just one of these samples, the others are possible or hypothetical samples that we could have observed. We use a collection of models, similar to the example with 53 possible decks of cards, that provide varying distributions for the mean height of samples of size 50. As in the previous example, we find how extreme the observed sample mean is in the distribution specified by the model.

[4]Percentile obtained from `https://dqydj.com/height-percentile-calculator-for-men-and-women/`.

What role does randomization play? For the patient example, Campbell and Franklin (2004) would argue that randomization is not required; all that is required is that there be no evidence that the selection of the patient was not representative of the population. Randomization plays a more important role in frequentist statistical inference because models are constructed under the assumption of randomization. The frequentist statistician will distinguish between data that were obtained using randomization and data obtained without randomization, often called observational studies. If randomization is not used, the analysis is vulnerable to the criticism that the sample was not representative. The assumption that the data in hand was obtained by random sampling addresses this concern. However, no hypothetical randomizations are required as the p-value is, by definition, the tail area associated with a particular percentile of the model.

3 Interpretations of Probability

Hájek (2019) describes many interpretations of probability but we limit our discussion to two that are applicable to frequentist inference. Before describing these interpretations we distinguish 'definition' from 'interpretation'.

3.1 Definition versus Interpretation

For our purposes, probability is defined by a mathematical model. A single probability model may be applied to two or more distinct settings. While the definition of probability is the same for each setting, the meaning and interpretation will depend on the application. Different applications will be better served by different interpretations. The distinction we

make between definition and interpretation is consistent with what is found in Hájek (2019).

This distinction can be understood by considering how different disciplines describe a 'vector'. A physicist says a vector is a quantity that has magnitude and direction. A computer scientist says a vector is a one dimensional array. A mathematician says a vector is an element of a vector space. While the mathematician's response may seem glib, it recognizes the structure that is common to what each of the scientists calls a vector and excludes any unnecessary descriptions. In particular, the concepts of magnitude and direction play no role in the definition of a vector space. By making these concepts part of the interpretation rather than the definition allows the same idea of vector to apply in computer science and many other fields. It is the interpretation that gives meaning to a vector, and it is the flexibility of multiple interpretations that allows the notion of a vector to apply to a variety of disciplines.

The same is true for probability. Instead of the two scientists considered above, we have a theoretical statistician and an applied statistician. The theoretical statistician says a probability describes a random process while the applied statistician says a probability describes a population or other distribution of values. For the mathematician, a probability is a set function that satisfies Kolmogorov's axioms. Notably, the definition of a probability space does not involve randomization.

The example of the two statisticians differs from that of the two scientists in that the latter are not likely to be in the position of interpreting the same vector. Statisticians may employ different methods and models to the same data, and this will result in p-values and other inferences that are in fact, different by definition. Using the vector space analogy, the methods may result in different vectors in the same vector space or in different vector spaces. That is not what we consider here. We have two interpretations for the same

probability model.

To insist that probability only describes random events and to define this in terms of limiting relative frequencies makes as much sense as requiring that the computer scientist define vectors as quantities that have magnitude and direction. Instead, interpreting a vector as a column of numbers is simply a more useful interpretation for computer science applications. The same is true for probability: a probability model describes the distribution of the population and a single random sample is better described in terms of a proportion (or, a percentile) rather than an infinite limit of frequencies obtained by hypothetical random samples.

3.2 Two Interpretations

We focus on two interpretations of probability, the *classical* (equally likely events) and the *limiting relative frequency* interpretations, but it is useful to recognize how these relate to other interpretations. Hájek (2019) describes three main concepts of probability:

> 1. An epistemological concept, which is meant to measure objective evidential support relations. ... 2. The concept of an agent's degree of confidence, a graded belief. ... 3. A physical concept that applies to various systems in the world, independently of what anyone thinks.

The two interpretations we consider belong to the third concept of physical probability. Statisticians are also interested in epistemology because the data provides evidence that relates to competing models for the population. The physical probabilities describe what can be learned about the population from the data, but frequentist statisticians do not assign probabilities to hypothetical models. As described above, statisticians use the word

'confidence' as a technical term and it is not a probability.

Hájek (2019) says the following about classical probability

> The guiding idea is that ... probability is shared equally among all the possible outcomes, so that the classical probability of an event is simply the fraction of the total number of possibilities in which the event occurs. It seems especially well suited to those games of chance that by their very design create such circumstances....

This describes how we use this term. Randomization used in statistical inference is the same as that used in games of chance, and so is well served by the classical interpretation. In the card example of Section 2.1 the event is the observation of 2 hearts and the classical interpretation can be used to describe this probability for each of the models considered. We do not use classical probability to address questions such as 'What is the probability that the deck contains h hearts?' A statistician using Bayesian methods would attempt to answer this question with a prior distribution.

Finite frequentism is related to the classical interpretation but differs in that only *actual* outcomes are considered. Hájek describes finite frequentism as "the probability of an attribute A in a finite reference class B is the relative frequency of actual occurrences of A within B."

The limiting relative frequency interpretation is closely related to *hypothetical frequentism* which can be described as follows

> ... we are to identify probability with a hypothetical or counterfactual limiting relative frequency. We are to imagine hypothetical infinite extensions of an *actual sequence of trials* [emphasis added]...Hájek (2019).

This description applies to how the term limiting relative frequency is applied in many situations. However, limiting relative frequency can also be used to describe a mathematical model in which case actual trials are not considered. The use of mathematical models also addresses the 'problem of the single case' that affects both finite and hypothetical frequentism[5]. Section 6.3 discusses the role of mathematical models and interpretation.

3.3 Criticism of these Interpretations

> This brings us to one of the chief points of controversy regarding the classical interpretation. Critics accuse the principle of indifference of extracting information from ignorance. Hájek (2019)

This point involves the epistemological use of probability. We use the classical interpretation only to describe physical probabilities and so avoid this controversy. The deck of 52 cards example shows that frequentist statisticians do not use classical probability for epistemic uncertainty regarding the population. We would agree with the critic that assigning equal probability to the 53 possible decks would be "extracting information from ignorance."

Some criticisms of limiting relative frequency do not apply to statistical inference. Hájek gives an example of this as the *reference class problem* illustrated by asking what is the probability that an individual lives to be 80 years old. The problem is that it is not clear to what population, or class, the individual should be considered a member. This is generally not an issue with statistical methods where inference is for a population rather than an

[5]Hájek (2019) provides this description "... a coin that is tossed exactly once yields a relative frequency of heads of either 0 or 1, whatever its bias. ... Famous enough to merit a name of its own, ... [this is an example of] the so-called 'problem of the single case'."

individual.

Hájek claims there is another problem, the *reference sequence problem* that 'probabilities must be relativized not merely to a reference class, but to a sequence within the reference class'. This could be a problem depending on the setting, but we do not see these criticisms as a reason to dismiss the limiting relative frequency interpretation outright. Hájek's statement regarding the broad interpretation of probability reflects our view of probability interpretation restricted to frequentist statistical inference.

> Each interpretation that we have canvassed seems to capture some crucial insight into a concept of it, yet falls short of doing complete justice to this concept. Perhaps the full story about probability is something of a patchwork, with partially overlapping pieces and principles about how they ought to relate.

Even within the domain of statistical inference there are overlapping pieces. In Section 5.1, we introduce the concept of scope to identify applications where the limiting relative frequency interpretation is and is not useful.

4 Common Understanding of Probability

We provide examples to show that, at least in some instances, probability is understood in terms of a proportion. These examples also serve as a platform for us to introduce the terms 'scope' and 'focus' of Section 5 to categorize applications as to which interpretation is more relevant. We make no claims regarding the prevalence of the interpretation of probability as a proportion but assert that it appears in enough settings to warrant considering it a common interpretation.

An argument can be made that the connection between probability and proportions is stronger than indicated by these examples. Each example interprets probability as a proportion. Conversely, there are examples of a proportion being interpreted as a probability[6]. Courant and Robbins (1978, p. 28) who "denote by A_n the number of primes among the integers 1,2,3,..., n" interpret the proportion A_n/n as a probability:

> The "density" of the primes among the first n integers is given by the ratio A_n/n, and may be computed empirically for fairly large values of n. [When $n = 10^9$ this ratio] may be regarded as giving the probability that an integer picked at random from among the first 10^9 integers will be prime...

No proof of this is offered. This is noteworthy because the authors emphasize the importance of proving statements that may seem obvious to the layman. That probability defined on a sample space of equally likely outcomes is a proportion is taken as an intuitive notion not requiring proof or elaboration.

4.1 *ESS* Example

The following example appears in the aforementioned *ESS* entry.

> A convict with a death sentence hanging over his head may have a chance of being pardoned. He is to make a choice between white and black and then draw a ball randomly from an urn containing 999 white balls and 1 black ball. If the color agrees with his choice he will be pardoned.

[6]Another example is Lang(2010, page 11) who motivates the proof for a theorem on the distribution of primes by writing "Roughly speaking, the idea is that the probability for a positive integer n to be prime is $1/\log n$."

Instead of using the proportion of white balls in the urn to describe a single random selection, the convict considers an unspecified number of hypothetical drawings.

> The convict replies that he will choose white because ... out of many hypothetical drawings he will in 99.9% of the trials be pardoned and in 0.1% of the trials be executed. ... the convict ... attaches 99.9% *probability to the single trial about to be performed*.

The article says the convict can attach a probability to a *single* trial because that

> probability is a very real thing to the convict and it is reliably estimated from past experiences concerning urn drawings.

It would seem we need to add the condition that the convict has sufficient experience with urn drawings.

Even if that were true, we would expect he would be open to the equally likely interpretation that clearly applies to a random draw from an urn. There is no need for a history of "past experiences concerning urn drawings" or a hypothetical future where convicts are executed repeatedly.

4.2 Gambling Examples

The broadcast of the 2018 Final Table in the 49th No-limit Hold-em main event held in Las Vegas (aired 13 July 2018 on ESPN's World Series of Poker) listed the player Cada as having a 14% chance of winning while his opponent Miles had an 86% chance. These probabilities were based on two cards held by Cada, two held by Miles, and four cards on the table. These cards were dealt after the deck was thoroughly shuffled so that each

ordering of the 52 cards was equally likely, or, at least treated as such. There is one more card to be dealt and the announcer says that Cada has 6 outs – cards that would provide him with a better hand than Miles. There are 44 cards remaining so the chance that Cada wins is 6/44 =14%.

North Carolina, like many states, has a lottery where numbers are selected by having balls jumbled with shots of air in a confined transparent space. The Pick-3 game consists of three clear boxes each with 10 balls that are labeled with the numerals 0, 1, ..., 9. These balls are jumbled for a few seconds and then one is allowed to come to the top. The jumbling is vigorous enough so that each ball is assumed to be equally likely to come up. While there may have been some players who waited for there to be sufficient history of Pick-3 drawings before placing a bet, we are confident there are many who did not require such history and still understood the probability of winning.

4.3 Clinical Trial Example

The examples above each had a known sample space of equally likely outcomes and this allowed for the calculation of the proportion that provided, under suitable randomization, the interpretation for probability. For statistical inference, simple random sampling from the population provides equally likely outcomes so that these probabilities can also be interpreted as proportions. However, unlike the previous examples, not all population values are known so that proportions cannot be calculated without specifying a model for these values.

Consider a trial of 60 participants in which 30 are assigned randomly to treatment A and the remainder to treatment B. For simplicity we take the response variable to be dichotomous with values 'favorable' and 'unfavorable'. The population is the 60 participants

and the value for each participant is the ordered pair indicating the outcome, favorable or unfavorable, under treatment A and under treatment B. Only one value of each pair is observed. Suppose the number responding favorably to A is 25 and to B is 17.

One way to compare the treatments is by testing the hypothesis that the two treatments have the same effect on each participant; that is, that the values are identical in each of the 60 outcome pairs. Under this hypothesis there would be exactly 42 favorable responses regardless of the treatment assignment. The population values consist of 42 favorable and 18 unfavorable outcomes. By chance 25 of the 42 favorable outcomes were assigned to treatment A. Each possible assignment of 30 outcomes to A can be enumerated and the proportion where 25 or more are favorable can be calculated. This proportion is 0.0235. Likewise, the proportion of 25 or more favorable responses in group B is also 0.0235. The interpretation is as follows: 4.7% of all possible treatment assignments have a discrepancy between groups as great or greater than the observed discrepancy of 25 versus 17. Because the actual assignment was done in a manner such that each possible assignment was equally likely, this *proportion* is the *probability* of an observation as extreme or more extreme than 25 vs 17. That is, the p-value is 0.047 and its interpretation does not require that we consider additional hypothetical random assignments of subjects to treatments.

5 Relationship between the Interpretations

Randomization and mathematical models are the most important features that distinguish frequentist statistical inference from more general forms of inference. Each of the examples we have considered represent these two features. Depending on the particular application, there may be interest in the results from a single randomization or from many randomiza-

tions. Scope is used to distinguish settings where interest is in a single instance and those where repeated randomizations are more relevant[7]. While frequentist statistical methods use mathematical models to make inference regarding the population, whether the emphasis is on the population or a model will depend on the particular method. Focus is used to make this distinction.

5.1 Scope - Specific or Generic

The importance of repeatability for an interpretation will depend on specific features of the application that, for the examples we consider, are closely tied to the intended audience. In the poker example, if the audience is Cada, the player holding a specific hand, probability is more usefully described as was done on the broadcast, as a proportion of equally likely cards. More generally, for casino gambling, if the audience is the house then probability is usefully described as a limiting relative frequency that describes an unspecified, but very large, number of hands.

The Lottery example did not include an interpretation of probability. However, if the audience is a ticket holder, then clearly there is interest in a specific drawing and the probability is naturally described as a proportion. On the other hand, the Lottery Commission is more concerned with on-going drawings and so long-run frequencies, obtained by repeated randomizations, are natural for this audience.

In the *ESS* example, where the audience is the convict, the proportion of white balls and the notion of equally likely provide a simpler description than hypothetical repeated drawings that involve this or other convicts. After a single draw, the convict is not likely

[7]Hájek (2019) describes *unrepeatable* events as a problem for the finite frequency interpretation. We use scope to describe the role of repeatability in the limited context of statistical inference.

to be interested in repeated drawings, especially if the ball indicates his execution. The collection of future repeated draws and consequent executions would be relevant to the state.

For the investigators of the clinical trial or anyone interested in the particular outcome of the study, the proportion of assignments to treatments resulting in a discrepancy as great as 25 and 17 provides a simple interpretation for the p-value. For statisticians interested in calibrating how inference procedures such as Fisher's exact test "would perform were they used repeatedly" then significance levels would be specified and probabilities would be described in terms of limiting relative frequencies of repeated randomizations[8].

The common factor in comparing the potential audience in each of these examples is the scope, either specific or generic, to which the probability extends. For a specific outcome, be it a hand of cards that could determine whether a player continues in the tournament, a lottery draw for a ticket holder, a convict whose life depends on a single draw from an urn, or a physician wanting to assess the evidence from a single study for the merits of a specific treatment, a proportion provides the natural interpretation for the probability related to a single randomization.

The scope is generic when the application is described in terms of a collection of outcomes.For statisticians who are concerned with how their methods perform in general, it is natural for the scope to be generic. However, results from a specific study will be communicated more effectively when statisticians recognize that the scope is specific for their audience.

Scope is related to Cox's distinction between "long-run behavior" and a "specific instance" but differs in that the collection of outcomes when the scope is generic need not be

[8]The quoted material is from the D.R. Cox displayed quote that appears in Section 1

constructed in the long-run. An interpretation for the confidence interval having generic scope that does not require repeated sampling is given in Section 7.2.

5.2 Focus - Population or Model

Scope applies to the interpretation of random phenomena whether or not these are used for inference. Focus is meaningful only in the context of statistical inference where we are concerned with an *unknown* distribution of numerical values. We call this distribution, whether it be measurements on individuals in a population or values obtained from random phenomena, the population distribution, or simply the population when the context makes it clear that we are considering a distribution of numerical values rather than a collection of individuals.

Statistical inference proceeds by positing that a known distribution, the model, is the same as, or an approximation to, the unknown population distribution. While statistical inference is always concerned with the population distribution, some inference procedures address the population directly and others indirectly using one or more models for the population. That is, the focus of an inference procedure can be on the population or a model.

The probability calculated for the clinical trial is a p-value and the calculation of any p-value requires the specification of a model (determined by the null hypothesis along with other assumptions). Unless the population is the same as the model, it is difficult to interpret the p-value as directly describing the population.

On the other hand, probability used to describe confidence intervals can have as its focus either the population or a family of models for the population. For the former, the interpretation of a 95% confidence interval for the mean, say 0.03 to 41.83, is that this

interval was the result of an interval generating procedure applied to the population that has the property that 95% of the intervals from this procedure contain the population mean. Since 95% describes the procedure and not the specific interval, the scope of this interpretation is generic and the focus is the population.

Fisher (1949, pages 190-191) provides the following interpretation.

> An alternative view of the matter is to consider that variation of the unknown parameter, μ, generates a continuum of hypotheses each of which might be regarded as a null hypothesis, which the experiment is capable of testing. In this case the data of the experiment, and the test of significance based upon them, have divided this continuum into two portions. One, a region in which μ lies between the limits 0.03 and 41.83, is accepted by the test of significance, in the sense that the values of μ within this region are not contradicted by the data, at the level of significance chosen. The remainder of the continuum, including all values of μ outside these limits, is rejected by the test of significance.

Here the focus is on a collection of models. The scope is specific because each model is assessed in terms of how extreme the specific data would be for that model.

6 Urn Models

Urn models are a conceptual construction that provide a convenient tool for describing inferential results in terms of classical probability. One should conceive of a bowl filled with N balls that are indistinguishable in regard to their possible selection but completely distinguishable in terms of at least one feature. This distinguishable feature is needed to count the balls. The urn model is an example of a multiset which is like a set except

multiplicities are allowed. For sets, $\{1,2\} \cup \{2,3\} = \{1,2,3\}$ while for urns, $\lfloor 1,2 \rfloor \cup \lfloor 2,3 \rfloor = \lfloor 1,2,2,3 \rfloor$. Unions and other basic set operations used below also hold for multisets.

6.1 Population Urn

A population can be described using the conceptional construction of an urn model. This model may be thought of as a bowl that contains one ball for each member in the population. For a variable of interest X, the population urn $\lfloor X \rfloor_{pop}$ is the bowl where the numerical value for each member is written on the corresponding ball. In most cases the values on the balls and the number of balls N are unknown. From the population urn we construct another urn $\lfloor X \rfloor_{pop}^{n}$ containing $\binom{N}{n}$ balls. Each ball in $\lfloor X \rfloor_{pop}^{n}$ represents a unique sample of size n from the $\binom{N}{n}$ possible samples from $\lfloor X \rfloor_{pop}$; this ball is labeled with an n-tuple of values obtained from the balls of the corresponding sample from $\lfloor X \rfloor_{pop}$. The only restriction on n is that it is a positive integer not greater than N. Notationally, this conceptual construction is

$$\lfloor X \rfloor_{pop} \xrightarrow{C_n} \lfloor X \rfloor_{pop}^{n} \tag{1}$$

where the arrow indicates an enumeration of all possible samples of n balls so that the observed sample corresponds to a ball $(x)^{obs}$ in $\lfloor X \rfloor_{pop}^{n}$.[9]

6.2 Model Urns

For inference regarding the population, a model is posited for $\lfloor X \rfloor_{pop}$ and the urn for the model is written $\lfloor X \rfloor_{\theta}$ because often there will be a set of models indexed by a parameter $\theta \in \Theta$. To assess how well $\lfloor X \rfloor_{\theta}$ approximates $\lfloor X \rfloor_{pop}$, the observed sample $(x)^{obs}$ from

[9] Sampling plans other than SRS would require a different enumeration.

$\lfloor X \rfloor_{pop}^{n}$ is compared to the possible samples in the model, $\lfloor X \rfloor_{\theta}^{n}$, where

$$\lfloor X \rfloor_{\theta} \xrightarrow{C_n} \lfloor X \rfloor_{\theta}^{n}. \tag{2}$$

Unlike $\lfloor X \rfloor_{pop}^{n}$, the n-tuples on all balls in $\lfloor X \rfloor_{\theta}^{n}$ are known.[10]

The samples in $\lfloor X \rfloor_{\theta}^{n}$ are compared to the observed sample using a test statistic T_θ, a real valued function on $\mathbb{R}^n$. Simple test statistics such as the sample mean will not be a function of the parameter so we write $T = T_\theta$ for notational simplicity. The value of the observed test statistic is $t^{obs} = T\,(x)^{obs}$. The plausibility of a specific model $\lfloor X \rfloor_{\theta_o}$ as an approximation to $\lfloor X \rfloor_{pop}$ is assessed by comparing $(x)^{obs}$ to the samples in $\lfloor X \rfloor_{\theta_o}^{n}$. Specifically, by finding the *proportion* of balls whose test statistic value is greater than or equal to t^{obs}. This proportion is written as

$$\Pr\lfloor T \geq t^{obs} \rfloor_{\theta_o}^{n} \tag{3}$$

where

$$\Pr\lfloor T \geq t \rfloor_{\theta}^{n} = \frac{|\,\{b \in \lfloor X \rfloor_{\theta}^{n} : T(b) \geq t\}\,|}{|\lfloor X \rfloor_{\theta}^{n}|}. \tag{4}$$

No randomizations were used to construct the model urn $\lfloor X \rfloor_{\theta_o}^{n}$. However, for the proportion in (3) to be meaningful as a probability, the observed sample must have been obtained using a simple random sample (SRS) from the population. Given this randomization, the proportion in (3) is the p-value for testing $H_o : \lfloor X \rfloor_{pop} = \lfloor X \rfloor_{\theta_o}$ using the test statistic T.

[10]The number of balls in model urn $\lfloor X \rfloor_{\theta}$ need not equal the number in the population urn. The relevant features are proportions rather than counts.

The $(1-\alpha)100\%$ confidence interval[11] for θ obtained from $(x)^{obs}$ is found by allowing θ_o in (3) to range over all possible values for θ,

$$C^{\alpha}_{(x)^{obs}} = \left\{\theta : \Pr\lfloor T \geq t^{obs} \rfloor^n_{\theta} \geq \alpha\right\}. \tag{5}$$

The interval in (5) represents all the models, indexed by θ, for which the observed data would not be in the most extreme $\alpha 100\%$ observations as measured by the ordering of the test statistic T. Even though the confidence interval $C^{\alpha}_{(x)^{obs}}$ involves many models there is still only one randomization that is required – the randomization used to obtain the data from the population.

The procedural interpretation of the confidence interval can be described using an urn of confidence intervals

$$\lfloor X \rfloor^n_{pop} \longleftrightarrow \lfloor C^{\alpha} \rfloor^n_{pop} \tag{6}$$

where the urn on the right is obtained by letting $(x)^{obs}$ in (5) range over all possible samples of size n from $\lfloor X \rfloor_{pop}$.

6.3 Compared to Repeated Sampling

The sampling urns for the population and for models are constructed using enumeration[12]. In contrast, the limiting relative frequency interpretation involves the conceptual construction of an infinite sequence where each term in the sequence is obtained by a hypothetical

[11] This notation and interpretation allow generalizing to a confidence region.

[12] This enumeration is conceptual to describe the relationship between a model and the sampling distribution obtained from the model. Both the model and the sampling distribution are mathematical objects that simply exist and do not require enumeration or construction.

random sample. Notationally,

$$\lfloor X \rfloor_{pop} \overset{SRS_n}{\longrightarrow} (x)_1, (x)_2, \ldots \tag{7}$$

where $(x)_i$ is the n-tuple obtained from the ith hypothetical sample. Because these are random samples, another sequence

$$\lfloor X \rfloor_{pop} \overset{SRS_n}{\longrightarrow} (x)'_1, (x)'_2, \ldots \tag{8}$$

could be used. The sequences in (7) and (8) are different but have the same limiting relative frequency.

It is important to remember that $\lfloor X \rfloor_{pop}$ represents the collection of values obtained from the population, so that (7) and (8) represent repeated random samples from a collection of numbers, not from the actual population. Imagining sampling from the actual population is hypothetical frequentism where frequencies are counterfactuals. Our discussion concerns mathematics not metaphysics.

While the sampling of (7) and (8) describe mathematics, the mathematics that is involved goes beyond measure theory. The notion of an infinite random sequence is required and there is no single accepted mathematical model for randomness. Chapter 2 of Khrennikov (2016) discusses three distinct models for classical randomness and a separate chapter comparing these to quantum randomness.

In addition to putting the discussion on shaky foundational ground, there is little payoff in terms of intuition or understanding. When we conceptualize a very large number of random samples of size n from a population of size N, the limiting relative frequency of

each of the possible samples is just $1/\binom{N}{n}$. This is precisely the frequency distribution of $\lfloor X \rfloor^n_{pop}$. So why cloud this message with the language of hypothetical repeated samples?

7 Confidence Intervals

The Fisher interpretation for the observed interval is naturally described without repeated sampling using $C^{\alpha}_{(x)^{obs}}$. The interpretation of a confidence interval as having been produced by a procedure is typically described using repeated sampling. Section 7.1 shows that, in fact, a single random sample can be used for the procedural interpretation. Sections 7.2 and 7.3 compare the single random sample interpretations of $C^{\alpha}_{(x)^{obs}}$ and $\lfloor C^{\alpha} \rfloor^n_{pop}$.

7.1 $\lfloor C^{\alpha} \rfloor^n_{pop}$

Greenland et al. (2016) provide the following interpretation for the 95% confidence interval,

> ... the 95% refers only to how often 95% confidence intervals computed from very many studies would contain the true effect if all the assumptions used to compute the intervals were correct.

It seems the word "only" is used to discourage other procedural interpretations since earlier in their paper the observed confidence interval is described in terms of testing which we understand to be Fisher's interpretation.

Even if the word "only" applies just to the procedural interpretation, this statement is too strong. As the urn models show, this interpretation need not be described in terms of limiting relative frequency. When the family of models contains the true model, $\lfloor X \rfloor_{pop} = \lfloor X \rfloor_{\theta^*}$ for some θ^*, then the urn $\lfloor C^{.05} \rfloor^n_{pop}$ defined by (6) has the property that 95% of

these intervals contain the true parameter value, θ^*. The proportion 0.95 is a probability when each interval in $\lfloor C^{.05} \rfloor^n_{pop}$ is given an equally likely chance of being selected; i.e., the observed data were obtained by an SRS from the population. The procedural interpretation for the confidence interval does not require the procedure to be repeated many times, just as understanding Cada's probability of winning did not require repeatedly shuffling the remaining poker cards.

7.2 Comparing the Interpretations

In terms of scope and focus the interpretations represented by $C^{\alpha}_{(x)^{obs}}$ and $\lfloor C^{\alpha} \rfloor^n_{pop}$ are very different. The interval $C^{\alpha}_{(x)^{obs}}$ is specific to the data that was observed, $(x)^{obs}$, and the focus is on a collection of models. Figure 1 shows these models for the card example and Figure 2 in the next section shows these models for the clinical trial example. The collection of intervals $\lfloor C^{\alpha} \rfloor^n_{pop}$ can be represented by a table of all possible confidence intervals with proportions obtained from the population. Table 1 shows this for the clinical trial example. This table is generic, confidence intervals for all possible observations are considered, and the focus is on the population when the additional assumption is made that there is a model with parameter θ^* such that $\lfloor X \rfloor_{\theta^*}$ is a close approximation to $\lfloor X \rfloor_{pop}$. This assumption is not required for the interpretation represented by $C^{\alpha}_{(x)^{obs}}$.

Coverage probability and expected length apply to $\lfloor C^{\alpha} \rfloor^n_{pop}$ but not to $C^{\alpha}_{(x)^{obs}}$. When intervals are defined with these two criteria in mind but without inverting a test, there is great flexibility in how individual intervals are chosen. As a result, observed intervals can have poor properties when interpreted in terms of testing.[13] To maintain fidelity to

[13]This issue arises when the sample space is discrete and the intervals are considered too conservative in terms of coverage probability. See, for example, Vos and Hudson (2008).

the Fisher interpretation, Vos and Hudson (2005) introduce the criteria p-confidence and p-bias that apply to $C^{\alpha}_{(x)^{obs}}$.

7.3 Clinical Trial Example Revisited

We return to the clinical trial considered above in which 30 of 60 patients were randomly assigned to treatment A and the remaining patients received treatment B. The study resulted in 25 and 17 favorable responses for treatments A and B, respectively. The p-value 0.047 calculated above is the probability of a discrepancy as large as 25 and 17 when a total of 42 favorable responses were observed *if* the probability of a favorable response is the same for the two treatments, i.e., $p_A = p_B$. The assumption $p_A = p_B$ provides a mathematical model for the probability for each of the 19 possible results given that there were a total of 42 favorable responses: $(12, 30), (13, 29), \cdots, (29, 13), (30, 12)$. Because the total number of favorable responses is fixed, it is enough to record X, the number of favorable responses for treatment A.

Confidence intervals describe models where the two success probabilities differ. For all possible values of p_A and p_B the distribution of X depends only on the odds ratio

$$\theta = \frac{p_A/(1-p_A)}{p_B/(1-p_B)}.$$

The 95% confidence interval for the odds ratio when $X = 25$ is all the values for θ from 1.01 to 16.0. We compare the interpretations $C^{\alpha}_{(x)^{obs}}$ and $\lfloor C^{\alpha} \rfloor^{n}_{pop}$ applied to this 95% confidence interval: $1.01 < \theta < 16.0$.

Figure 2 illustrates the interpretation $C^{\alpha}_{(x)^{obs}}$ where $(x)^{obs} = 25$ and $\alpha = 0.05$. In this interpretation each value for the odds ratio specifies a distribution for X and the values for

θ between 1.01 and 16.0 are those for which observing $X = 25$ is *not* among the 5% most extreme (2.5% for one tail). Distributions for $\theta = 1.01$ and $\theta = 16.0$ appear as vertical white lines in Figure 2. Vertical bars were used in Figure 1 to show the distribution for each of the models. Bars are replaced with lines because there are infinitely many models for the clinical trial example. While the number of favorable responses under Treatment A could have been a value other than 25, these are not considered in this interpretation. The scope is specific to the data actually observed and is useful for anyone interested in this particular study.

Table 1 illustrates the interpretation $\lfloor C^{\alpha} \rfloor^{n}_{pop}$ where $\alpha = 0.05$. The fact that the value observed for X was 25 is not required. This interpretation is useful for anyone interested in methodology, in this case, that of confidence intervals for the odds ratio. The scope of this interpretation is generic – it applies to any randomized trial with two treatment groups each of size 30 in which there are 42 total favorable responses. Table 1 lists the 95% confidence interval for each value for X. While there are only 19 different intervals, the proportion of each is a function of θ since the distribution of X is a function of θ. Table 1 provides the distribution for $\lfloor C^{\alpha} \rfloor^{n}_{\theta}$ where θ can be any value larger than 0[14]. These intervals have the property that the proportion of intervals that contain θ is at least 95% for every $\theta > 0$. With the added assumption that one of the models indexed by θ is the population model, that is, there is a value θ_{pop} such that $\lfloor C^{.05} \rfloor^{n}_{\theta_{pop}} = \lfloor C^{.05} \rfloor_{pop}$, the focus can be changed from an infinite collection of models to the population. We don't need to know the population distribution to know that at least 95% of the intervals will contain

[14]For $\lfloor C^{\alpha} \rfloor^{n}_{\theta}$ to be a finite urn requires θ to be rational. This is not a limitation since any real number can be approximated to arbitrary precision by a rational number. Also, the important point here is that the model provides a *distribution* for a collection, possibly infinite, of numerical values rather than a model for random phenomena.

x	95% CI for θ	Proportion $P(X = x; \theta)$
12	0 to 0.1147	$k\binom{30}{12}\binom{30}{30}\theta^{12}$
13	0.0006 to 0.2141	$k\binom{30}{13}\binom{30}{29}\theta^{13}$
14	0.0065 to 0.3386	$k\binom{30}{14}\binom{30}{28}\theta^{14}$
$\vdots$	$\vdots$	$\vdots$
25	1.0146 to 15.9881	$k\binom{30}{25}\binom{30}{17}\theta^{25}$
$\vdots$	$\vdots$	$\vdots$
29	4.6706 to 1622.2	$k\binom{30}{29}\binom{30}{13}\theta^{29}$
30	8.7200 to ∞	$k\binom{30}{30}\binom{30}{12}\theta^{30}$

Table 1: $\lfloor C^{0.05} \rfloor^n_{pop} = \lfloor C^{0.05} \rfloor_{\theta_{pop}}$. The Proportion column specifies the distribution of X, the number of favorable responses under treatment A, for the model where the odds ratio is θ; $k = k(\theta)$ is the proportionality constant chosen so that the sum of this column is 1. For every possible value of θ, the proportion of intervals that contain θ is at least 95%. For example, if $\theta = 0.0005$ then only the first interval contains θ which means $k\binom{30}{12}\binom{30}{30}0.0005^{12} \geq 0.95$. If $\theta = 2$, it can be checked that only intervals corresponding to $x = 20$ to $x = 26$ contain the value 2 which means that the sum of the proportions for these intervals is at least 0.95: $k\sum_{x=20}^{26}\binom{30}{x}\binom{30}{42-x}2^x \geq 0.95$. Since this property holds for all possible values for θ, if the population is described by one of these values then this property holds for the population.

θ_{pop}.

We compare these interpretations with the standard frequentist interpretation: the observed 95% confidence interval $1.01 < \theta < 16.0$ either does or does not contain θ_{pop}, 95% refers to the procedure that generated this interval and this procedure produces intervals that cover the population parameter value θ_{pop} at least 95% of the time. Difficulties arise when "95% of the time" is explained in terms of repeatedly applying the procedure. When applied to a single study, these repeated applications are subject to the criticism of hypothetical frequentism. In particular, for the 60 patients in this study, what would it mean to repeatedly randomize to treatment groups? Having received a treatment, a patient is

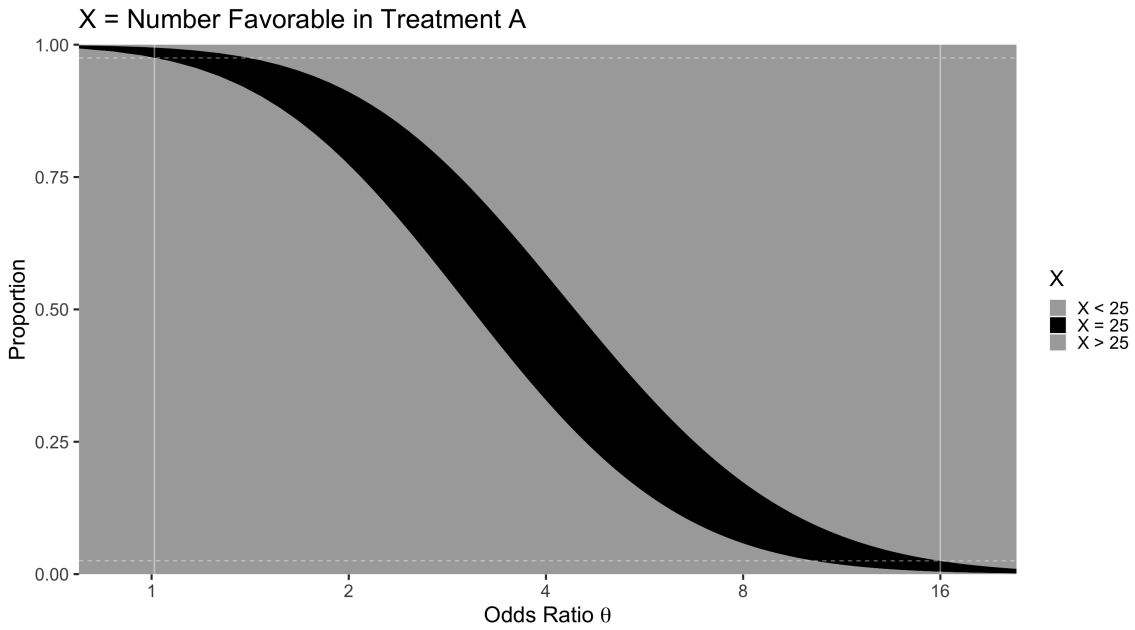

Figure 2: $C^{0.05}_{(x)^{obs}}$ consists of the models with odds ratio between 1.01 and 16.0. Horizontal dashed white lines are at 0.025 and 0.975. The vertical white line at $\theta = 1.01$ shows the distribution of X for the model where the odds of success with Treatment A are 1.01 times that of Treatment B and the total number of successes observed is 42. For this model, the observation of 25 favorable responses for Treatment A is in the upper 2.5% of possible values for X making the observation significantly larger than would be expected. Because patients were randomly assigned to treatments, the distribution of X given by the model with odds ratio $\theta = 1.01$ provides the probability of observing 25 or more favorable responses for treatment A. Vertical lines where $\theta > 1.01$ show this probability increases for these models. The vertical white line at $\theta = 16.0$ shows the distribution of X for the model where the odds of success with Treatment A are 16.0 times that of Treatment B. For this model, the observation of 25 favorable responses for Treatment A is in the lower 2.5% of possible values for X making the observation significantly smaller than would be expected – the probability of observing 25 or fewer favorable responses for treatment A is 0.025. Vertical lines where $\theta < 16.0$ show this probability increases for these models. The 95% confidence interval for θ consists of value between 1.01 and 16.0.

no longer the same – the population has changed. This explanation works better from the perspective of the theoretical statistician: the procedure is applied to many studies but only once to each study[15]. The scope for the theoretical statistician is generic, this interpretation is problematic when the scope is specific, when interest is in only one particular study.

The interpretation $\lfloor C^{.05} \rfloor_{pop}$ differs from the standard interpretation in that there is no need to use the notion of "95% of the time" which is not essential to interpreting probability. The urn model $\lfloor C^{.05} \rfloor_{pop}$ consists of balls labelled with confidence intervals such that at least 95% of these contain θ_{pop}, the results of the randomized trial is equivalent to a single random selection from this urn. The distinction between these interpretations can be understood with the simple example of the probability of rolling a '6' with a fair die. The probability is 1/6 because if you roll the die repeatedly the proportion of times that the face with '6' comes up will be come very close to 1/6. Or, the probability is 1/6 because it is equivalent to a random selection from an urn where exactly one of 6 balls is labelled with '6'. The distinction in this simple example is less useful since repeatedly rolling a die is less problematic than repeatedly conducting the same randomized trial.

Bayesian criticism that frequentist inference is hypothetical frequentism is valid for the frequentist *interpretation* that uses repeated sampling. $\lfloor C^{\alpha} \rfloor_{pop}$ which does not use repeated sampling shows that this criticism does not apply to frequentist *methods*. The conversation between Bayesians and frequentists will be improved when the distinction between a model and its interpretation is recognized, and when frequentists provide better explanations for their methods, especially when applied to the results of a specific study.

[15]Procedures are constructed that apply to numbers of patients, n_A and n_B, and success totals, n_T, other than $n_A = 30$, $n_B = 30$, and $n_T = 42$ so these apply more generally.

Another Bayesian criticism of frequentist inference is that there is no reference to the data actually observed. This is a valid criticism of the interpretation $\lfloor C^{\alpha} \rfloor_{pop}$ which is ill-suited when the scope is specific, but not of frequentist inference as the interpretation $C^{\alpha}_{(x)^{obs}}$ uses the observed value $(x)^{obs}$ to conduct inference for each model labelled by the parameter. As Figure 2 shows, this interpretation is more than a simple dichotomization of models as having parameter values either in or outside the confidence interval. The observation $X = 25$ is farther in the tails of the distribution for models with parameters near the endpoints of the interval than for those having parameter values near the center of the interval.

Frequentist methods can be described without hypothetical frequentism and with direct reference to the observed data. These observations will not resolve the disagreements between Bayesians and frequentists but will move the discussion to more productive areas. Ideally, the adjectives 'Bayesian' and 'frequentist' could be moved from 'statistician' to 'methods' so that statisticians would discuss the role of Bayesian and frequentist methods in specific applications.

8 P-values

Confidence intervals allow for an interpretation that is population focused. Interpreting p-values in terms of population focus can lead to problems associated with hypothetical frequentism. As an example we consider the issue of potential comparisons raised by Gelman (2016) who claims

> ... to compute a valid p-value you need to know what analyses *would have been done* had the data been different. Even if the researchers only did a single

> analysis of the data at hand, they well could've done other analyses had the data been different.

Gelman considers repeated sampling from the population but the p-value is a probability that describes a model – generally, a model thought to be a poor candidate for the unknown population distribution[16]. Comments by Fisher (1959, page 44) apply here

> In general tests of significance are based on hypothetical probabilities calculated from the null hypotheses. They do not generally lead to any probability statements about the real world, but to a rational and well-defined measure of reluctance to the acceptance of the hypotheses they test.

Certainly p-values can be misused but Gelman's statement is too strong because it makes p-values invalid even when there has been no *actual* misuse. A potential misuse of a p-value, or any inference procedure, does not invalidate a single instance of proper use. Consider the following example from Texas Hold'em Poker. A gambler calculates the probability of making a specific hand based on the proportion of unseen cards. This calculation is done under the following conditions: he is well rested, sober, and knows the dealer, and he has no reason to suspect cheating. The result of this calculation is a valid probability. The gambler's wife might say that if he were to play too much poker, then he would become sleepy, drink too much, and gamble at shady establishments. Regarding the long run outcome of his gambling, these are legitimate concerns that bring the validity (utility) of future probability calculations into questions. However, these potentialities do not affect the gambler's specific calculation made under the actual conditions. The scope for the gambler is specific while for his wife it is generic.

[16]If Gelman is describing the actual population rather than the unknown distribution of population values, then this is hypothetical frequentism.

The reader might find differences between our example and the discussion of potential comparisons. Our hope is that we could agree that hypothetical long run sampling is problematic when used to address a specific instance, and our point is that repeated sampling is not required to interpret inference for the data actually observed.

9 Power

We have seen that confidence intervals and p-values can be interpreted using a single random sample. Power calculations are done before data have been collected and do not require any randomization or hypothetical repetitions. This is in contrast to how power is often discussed. For example, Greenland et al. (2016) describe power as a probability "defined over repetitions of the same study design and so is a frequency probability."

Power can be described using the population of men's heights and the population of women's heights from Section 2.2 An inference regarding the sex of an individual will be made using the following procedure: select an individual at random from one of the two populations (we do not know which) and obtain their height, if the height is less than 5' 4"[17], infer that the individual is a woman, otherwise infer that the individual is a man. This procedure is described by two probabilities: the probability of correctly identifying an individual as a man, p_1, and the probability of correctly identifying an individual as a woman, p_2. In the language of hypothesis testing, if the null hypothesis is the claim that the individual is a man, then the significance level of the test is $1 - p_1$ and the power is p_2.

While randomization is part of the procedure, power calculations are done *before* a study is conducted; that is, before any randomization. Using the distribution of heights, p_1

[17]The procedure could be defined using any height. This particular height was chosen so that the probability is 0.05 of incorrectly identifying the individual as a woman.

is the proportion of men who are taller than 5'4" and p_2 is the proportion of women who are shorter than 5'4". Using data collected from the 2015-16 National Health and Nutrition Examination Survey as models for these two populations, $p_1 = 0.95$ and $p_2 = .55$[18]. The single randomization specified by the procedure makes these proportions meaningful as probabilities. Repeated randomizations are not required.

Urn models can be used to extend the height example to a general setting. Power calculations are done by comparing the model specified by a null hypothesis to a competing model. The urn $\lfloor X \rfloor_o^n$ of the null model is compared to the urn $\lfloor X \rfloor_1^n$ of the competing model in terms of a test statistic T_o. Specifically, the significance level α for a test T_o where large values are evidence against the null model defines a value t^* such that

$$\Pr\lfloor T_o \geq t^* \rfloor_o^n = \alpha$$

and the power β is given by

$$\Pr\lfloor T_o \geq t^* \rfloor_1^n = \beta.$$

Both α and β are proportions. The power is the proportion of all samples of size n from the competing model (posited as an approximation to the population) that are more extreme than t^*. These proportions are meaningful as probabilities and useful for inference regarding the population when the observed data is obtained by an actual randomization from the population. Hypothetical repetitions from the population or one of the models are not required.

[18] 2015-16 NHANES data is used for the website `https://dqydj.com/height-percentile-calculator-for-men-and-women/`

10 Discussion

Describing the observed confidence interval as having been obtained from a procedure is often the only interpretation that is considered, but there are authors who recognize Fisher's interpretation. Examples include, Kempthorne and Folks (1971) who call Fisher's interpretation a *consonance interval* and Mayo (2018) who describes inference in terms of severe testing that appears to be very close to Fisher's interpretation.

Other authors also see pitfalls with the introduction of the concept of infinity. For example, Hacking (1976, p. 7) "However much they have been a help, I shall argue that hypothetical infinite populations only hinder full understanding of the very property von Mises and Fisher did so much to elucidate."

We have restricted urns to be finite for simplicity. Allowing an urn to have an infinite number of balls results in a *statistical ensemble.* According to the Wikipedia entry (2021)

> ... an **ensemble** (also **statistical ensemble**) is an idealization consisting of a large number of virtual copies (sometimes infinitely many) of a system, considered all at once, each of which represents a possible state that the real system might be in.

A single simple random sample of n individuals from a population creates a statistical ensemble where the possible states consist exactly of the possible samples of size n from the population.

The conceptualization of a statistical ensemble differs from repeated sampling in that a large number is considered *all at once* and this idea avoids several pitfalls associated with repeated sampling. Repeated sampling and terms such as "long run" introduce the notion of time even though time is not included in the definition of probability. Adding to the

confusion is that when the scope is generic, such as a statistician defining procedures in terms of "how they would perform were they used repeatedly", time fits naturally in that particular interpretation. Furthermore, repetition generates a sequence and the order of this sequence has nothing to do with the structure of the collection so the idea of independence is needed to appropriately describe a random sequence. By considering the collection all at once, whether it is balls in an urn or states of an ensemble, these complications are avoided. A statistical ensemble can be applied when the scope is generic or specific but is especially useful in the latter case.

Recognizing that the focus can be either the population or the model sheds light on the role of randomization in statistical inference. Using a model for inference is justified by using a single actual random sample from the population. The model specifies a distribution of possible samples to which the observed sample is compared and the relationship between these is expressed in terms of a tail proportion or percentile, neither of which involves randomization. Hypothetical repeated randomizations may be introduced as a means to interpret the percentile, but these hypothetical randomizations, and the consequent confusion with the required randomization from the population, can be avoided by using classical probability described by urn models.

References

Campbell, S. and J. Franklin (2004, January). Randomness and the justification of induction. *Synthese 138*(1), 79–99.

Courant, R. and H. Robbins (1978). *What is mathematics? an elementary approach to ideas and methods*. Oxford: Oxford Univ. Press. OCLC: 256570914.

Cox, D. R. and D. V. Hinkley (2000). *Theoretical statistics*. Boca Raton: Chapman & Hall/CRC.

Cox, P. D. R. (2006). *Principles of Statistical Inference*. Cambridge University Press.

Fisher, R. (1949). *The design of experiments. 1949* (5th ed.). New York: Hafner Publishing Company Inc.

Fisher, R. (1959). *Statistical methods and scientific inference* (2nd ed.). Hopetoun Street, University of Edinburgh: T and A Constable Ltd.

Gelman, A. (2016). The problems with p-values are not just with p-values. *The American Statistician 70*(2), online.

Greenland, S., S. J. Senn, K. J. Rothman, J. B. Carlin, C. Poole, S. N. Goodman, and D. G. Altman (2016). Statistical tests, p values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology 31*(4), 337–350.

Hacking, I. (1976). *Logic of statistical inference*. Cambridge England, New York: University Press.

Hájek, A. (2019). Interpretations of Probability. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2019 ed.). Metaphysics Research Lab, Stanford University.

Johnson, R. (1996). *Statistics : principles and methods*. New York: Wiley.

Kempthorne, O. and L. Folks (1971). *Probability, Statistics, and data analysis*. Iowa State University Press.

Khrennikov, A. (2016). *Probability and randomness: quantum versus classical*. Covent Garden, London: Imperial College Press.

Lang, S. (2010). *Undergraduate algebra: Serge Lang*. New York: Springer. OCLC: 878836542.

Lindley, D. (2005). Foundations of probability. In T. Armitage, Peter & Colton (Ed.), *Encyclopedia of biostatistics* (2 ed.), Volume 3, pp. 1993–2001. Hoboken, N.J: John Wiley & Sons.

Mayo, D. G. (2018). *Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars*. Cambridge University Press.

Romeijn, J.-W. (2017). Philosophy of Statistics. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2017 ed.). Metaphysics Research Lab, Stanford University.

Sverdrup, E. (2006). Frequency interpretation in probability and statitical inference. In S. Kotz (Ed.), *Encyclopedia of statistical sciences* (2 ed.), Volume 4, pp. 2530–2536. Hoboken, N.J: Wiley-Interscience.

Vos, P. W. and S. Hudson (2005). Evaluation criteria for discrete confidence intervals: Beyond coverage and length. *The American Statistician 59*(2), 137–142.

Vos, P. W. and S. Hudson (2008). Problems with binomial two-sided tests and the associated confidence intervals. *Australian & New Zealand Journal of Statistics 50*(1), 81–89.

Wikipedia contributors (2021). Statistical ensemble (mathematical physics) — Wikipedia, the free encyclopedia. [Online; accessed 181-May-2021].

11 Acknowledgements

We are grateful to two anonymous reviewers for their constructive comments and insightful questions.