Frequentist Principles in the Context of Bayesian Learning:

The Merits of Using P-Values to Update Beliefs about Causal Hypotheses

Tom Leavitt

March 3, 2017

Columbia University
tl2624@columbia.edu

Overview

Frequentism and Bayesianism

Sampling Theory

Sampling Theory and Likelihood Theory

Bayesian Confirmation Theory

Hypothetical Experiment

Exact P-Values

Worked Out Example

Conclusion

Further Questions

Frequentism and Bayesianism

Frequentism and Bayesianism

Perhaps nothing is more antithetical to Bayesianism than p-values:

- P-values are defined by the frequentist logic of repeated randomizations.
- P-values connote a binary decision theory that either rejects or fails to reject a single hypothesis on the basis of its p-value.
- Bayesianism does not reject or fail to reject a single hypothesis on the basis of its p-value.
- Bayesianism defines likelihood functions that ascribe probability to data, which researchers use to update prior beliefs that are defined over *all* possible hypotheses.

Sampling Theory

Sampling Theory

- Let's imagine that we are taking a sample of size n from a population of $y_1, \ldots, y_N$. We can denote this random sample by $Y_1, \ldots, Y_n$.
- $Y_1, \ldots, Y_n$ are random variables because the value of Y_i we observe could be any value in the population of $y_1, \ldots, y_N$ so long as each $y_1, \ldots, y_N$ has a sample inclusion probability defined on the open interval from 0 to 1.
- We have n parameters we want to estimate, which is the mean, μ_i, of the population from which Y_i was drawn for all $Y_i = Y_1, \ldots, Y_n$ in the sample.
- If each unit in the population from which we are sampling has the same sample inclusion probability (i.e., sample inclusion probabilities are uniformly distributed over all $1, \ldots, N$ units in the population), then $\mathbb{E}[Y_1] = \mu_y$.
- If all $Y_1, \ldots, Y_n$ random variables are *independent* and *identically* distributed (i.i.d), then $\mathbb{E}[Y_1] = \cdots = \mathbb{E}[Y_n] = \mu_y$.

Sampling Theory Continued

Consider the population of $y_1, \ldots, y_N$ units and each unit's respective sample inclusion probability:

y	1	0	0	0	1	1	1	1
Pr	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125

Table 1: y Values in the Population and Uniform Sample Inclusion Probabilities

Then, $Y_i \sim$ Bernouli:

$$\Pr(Y_i = y_i) = \begin{cases} \pi^{y_i}(1-\pi)^{1-y_i} & \text{if } y_i = 0, 1 \\ 0 & \text{otherwise.} \end{cases}$$

With sample inclusion probabilities uniformly distributed over all $y_i, \ldots, y_N$ units in the population, $\pi = 0.625 = \overline{y} = \frac{5}{8}$.

The set of events in Y_i is identical to the unique values of $y_1, \ldots, y_N$. And if we have uniform sample inclusion probabilities, the probability associated with each $Y_i = y_i$ is equivalent to the relative frequency of that value of y in the population.

Sampling Theory Continued

y	1	0	0	0	1	1	1	1
Pr	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.3

Table 2: y Values in the Population and Nonuniform Sample Inclusion Probabilities

$Y_i \sim$ Bernouli:

$$\Pr(Y_i = y_i) = \begin{cases} \pi^{y_i}(1-\pi)^{1-y_i} & \text{if } y_i = 0, 1 \\ 0 & \text{otherwise.} \end{cases}$$

With sample inclusion probabilities *non*uniformly distributed over all $y_i, \ldots, y_N$ units in the population, $\pi = 0.7 \neq \overline{y} = \frac{5}{8}$.

Sampling Theory and Likelihood Theory

y	1	0	0	0	1	1	1	1
Pr	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125

Table 3: y Values in the Population and Uniform Sample Inclusion Probabilities

The likelihood function for the observed value y_i is:

$$\mathcal{L}_i\left(y_i \mid \pi\right) = \pi^{y_i}\left(1-\pi\right)^{1-y_i}$$

If sample inclusion probabilities for Y_i are uniformly distributed over all $y_1, \ldots, y_N$ units in the population, then a hypothesis about the parameter π ensures that π is also a hypothesis about the population from which you are sampling.

If random variables $Y_1, \ldots, Y_n$ are independent and identically distributed, then the joint likelihood function for all units in the sample is:

$$\mathcal{L}\left(y_1, \ldots, y_n \mid \pi\right) = \prod_{i=1}^{n} \pi^{y_i}\left(1-\pi\right)^{1-y_i}$$

Bayesian Confirmation Theory

Bayesian Confirmation Theory

- We have a set of hypotheses that pertain to values of π and a subjective credence (probability) function defined over all values of π in the parameter space.
- We have a likelihood function through which each hypothesis ascribes a probability to the sample data.
- We update our credences about π via Bayesian conditionalization:

$$C\left(H \mid e\right) = \frac{\Pr_H\left(e\right) C\left(H\right)}{C\left(e\right)},$$

where $C\left(e\right) = \Pr_{h_1}\left(e\right) C\left(h_1\right) + \cdots + \Pr_{h_n}\left(e\right) C\left(h_n\right)$.

Hypothetical Experiment

Hypothetical Experiment I

Table 4: Experimental Outcomes

z	y_c	y_t
1	?	1
0	0	?
0	0	?
1	?	1
1	?	0
0	1	?
1	?	1
1	?	1
0	0	?
0	0	?

Hypothetical Experiment I Continued

- $H_{y_c} = \{0, \frac{1}{10}, \frac{2}{10}, \frac{3}{10}, \frac{4}{10}, \frac{5}{10}, \frac{6}{10}, \frac{7}{10}, \frac{8}{10}, \frac{9}{10}, 1\}$
- $H_{y_t} = \{0, \frac{1}{10}, \frac{2}{10}, \frac{3}{10}, \frac{4}{10}, \frac{5}{10}, \frac{6}{10}, \frac{7}{10}, \frac{8}{10}, \frac{9}{10}, 1\}$
- E.g., $H_{y_c} = \frac{5}{10}$:

$$\Pr(e \mid H) = \binom{n}{k} \pi^k (1-\pi)^{n-k},$$

where $n \in \mathbb{N}$ is the number of sample observations, $k \in \{0, \ldots, n\}$ is the number of 1s, $n - k \in \{0, \ldots, n\}$ is the number of 0s, and $\pi \in [0, 1]$ is the proportion of 1s in the population of experimental subjects (as specified by any given hypothesis).

$$\begin{aligned}
\Pr(e \mid H) &= \binom{5}{1} \left(\frac{1}{2}\right)^1 \left(1 - \frac{1}{2}\right)^{5-1} \\
&= (5) \left(\frac{1}{2}\right) \left(\frac{1}{16}\right) \\
&= 0.15625.
\end{aligned}$$

Hypothetical Experiment II

- But what about when the support of y_c, y_t is not $\{0, 1\}$?
- If sample inclusion probabilities for Y_i are uniformly distributed across all $y_1, \ldots, y_N$ units in the population, then $\mathbb{E}[Y_i] = \mu_y$. But how can a hypothesis about the value of μ_y ascribe probability to sample data without imposing a probability model on the data generating process?

Table 5: Experimental Population of y_c and y_t Values

y_c	y_t
18	10
20	12
38	30
48	40
11	3
58	50
43	35
48	40
30	22
36	28

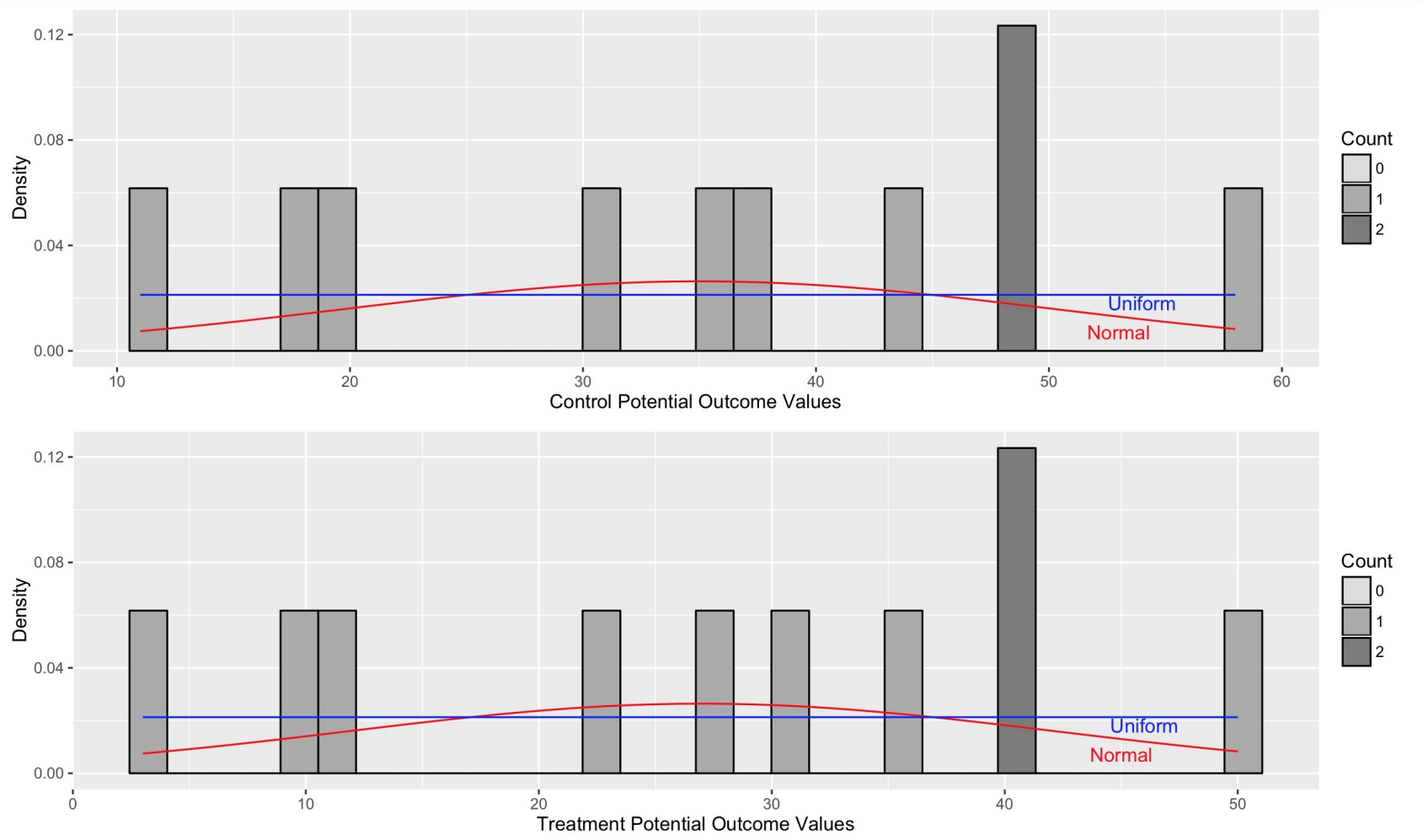

Figure 1: Distribution of Control and Treatment Potential Outcomes in Experimental Population

How can we define prior beliefs over hypotheses and update those beliefs after observing data without invoking probability models for the data generating process?

Exact P-Values

Exact P-Values

- In the hypothetical experiment above, there are $\binom{10}{5} = 252$ permutations in which 5 out of 10 units could be assigned to treatment while the other 5 out of 10 units are assigned to control.

Table 6: Treatment Assignment Permutations

$\mathbf{z}_1$	$\mathbf{z}_2$	...	$\mathbf{z}_{251}$	$\mathbf{z}_{252}$
1	1	...	0	0
1	1	...	0	0
1	1	...	0	0
1	1	...	0	0
1	0	...	1	0
0	1	...	0	1
0	0	...	1	1
0	0	...	1	1
0	0	...	1	1
0	0	...	1	1

Exact P-Values

$$\Pr\{t(\mathbf{Z}; \mathbf{y_c}) \geq T\} \equiv \sum_{\mathbf{z} \in \Omega} [t(\mathbf{Z}; \mathbf{y_c}) \geq T] \, Pr(\mathbf{Z} = \mathbf{z}),$$

where

$$[\text{event}] = \begin{cases} 1 & \text{if event occurs,} \\ 0 & \text{otherwise.} \end{cases}$$

(Rosenbaum, 2002)

Exact P-Values

There are $\binom{10}{5} = 252$ treatment assignment vectors $\mathbf{z} \in \Omega$. But what is the probability associated with each $\mathbf{z} \in \Omega$?

Proposition

In the context of complete random assignment, the probability associated with each $\left(\mathbf{Z} = \mathbf{z} \in \Omega \,\middle|\, n_t\right)$ *is:*

$$\Pr\left(\mathbf{Z} = \mathbf{z} \in \Omega \,\middle|\, n_t\right) \equiv \frac{\prod_{i=1}^{n} \pi_i^{z_i} (1-\pi_i)^{(1-z_i)}}{\sum_{i:\omega_i \in \Omega}^{|\Omega|} \left\{ \left(\prod_{i:z_i \in \omega}^{|\omega|} \prod_{i=1}^{n_t} \pi_i \right) \left(\prod_{j:z_j \in \omega'}^{|\omega'|} \prod_{j=1}^{n_c} \left(1-\pi_j\right) \right) \right\}}.$$

Worked Out Example

Worked Out Example

To illustrate the type of analysis for which I advocate, consider the aforementioned hypothetical experiment:

Table 7: Experimental Outcomes

z	y_c	y_t	y
1	?	10	10
0	20	?	20
0	38	?	38
1	?	40	40
1	?	3	3
0	58	?	58
1	?	35	35
1	?	40	40
0	30	?	30
0	36	?	36

Worked Out Example

The observed difference-in-means test-statistic is

$$
\begin{aligned}
(\overline{y} \mid z = 1) - (\overline{y} \mid z = 0) & \\
&= \frac{(10 + 40 + 3 + 35 + 40)}{5} - \frac{(20 + 38 + 58 + 30 + 36)}{5} \\
&= 25.6 - 36.4 \\
&= -10.8,
\end{aligned}
$$

where y_i, the observed outcome for each of the $i = 1, \ldots, 10$ units, is a realization of the random variable $Y_i = Z_i y_{it} + (1 - Z_i)\, y_{ic}$.

Worked Out Example

Let's say that the researcher wants to assess the probability of the observed test statistic, -10.8, or greater (in terms of absolute value) if a given hypothesis were true.

For the purposes of exposition, assume that there are only 4 hypotheses.

- H_1: The unit level causal effect is 0 for all 10 experimental subjects;
- H_2: The unit level causal effect is -5 for all experimental subjects;
- H_3: The unit level causal effect is -8 for all subjects;
- H_4: The unit level causal effect is -10 for all subjects.

Assign an equal prior probability of $\frac{1}{4}$ to each of the four causal hypotheses.

Worked Out Example

In order to assess the evidential support for hypotheses H_1, H_2, H_3 and H_4, the first step is to fill in the full schedule of potential outcomes as if each of the respective hypothesis were true:

z	y_c	y_t	$y_t - y_c$
1	10	10	0
0	20	20	0
0	38	38	0
1	40	40	0
1	3	3	0
0	58	58	0
1	35	35	0
1	40	40	0
0	30	30	0
0	36	36	0

z	y_c	y_t	$y_t - y_c$
1	15	10	-5
0	20	15	-5
0	38	33	-5
1	45	40	-5
1	8	3	-5
0	58	53	-5
1	40	35	-5
1	45	40	-5
0	30	25	-5
0	36	31	-5

Table 8: Potential Outcomes Under (from Left to Right) $H_1 : 0$, $H_2 : -5$

Worked Out Example

z	y_c	y_t	$y_t - y_c$
1	18	10	-8
0	20	12	-8
0	38	30	-8
1	48	40	-8
1	11	3	-8
0	58	50	-8
1	43	35	-8
1	48	40	-8
0	30	22	-8
0	36	28	-8

z	y_c	y_t	$y_t - y_c$
1	20	10	-10
0	20	10	-10
0	38	28	-10
1	50	40	-10
1	13	3	-10
0	58	48	-10
1	45	35	-10
1	50	40	-10
0	30	20	-10
0	36	26	-10

Table 9: Potential Outcomes Under (from Left to Right) $H_3 : -8$ and $H_4 : -10$

Worked Out Example

- For all null hypotheses, the expected value of the difference-in-mean y_c values between treatment and control units over all 252 treatment assignment permutations is always 0.
- That is, by considering only treatment and control units' y_c values, the difference in mean y_c values between units assigned to treatment and units assigned to control is, by construction, equal to 0 in expectation.

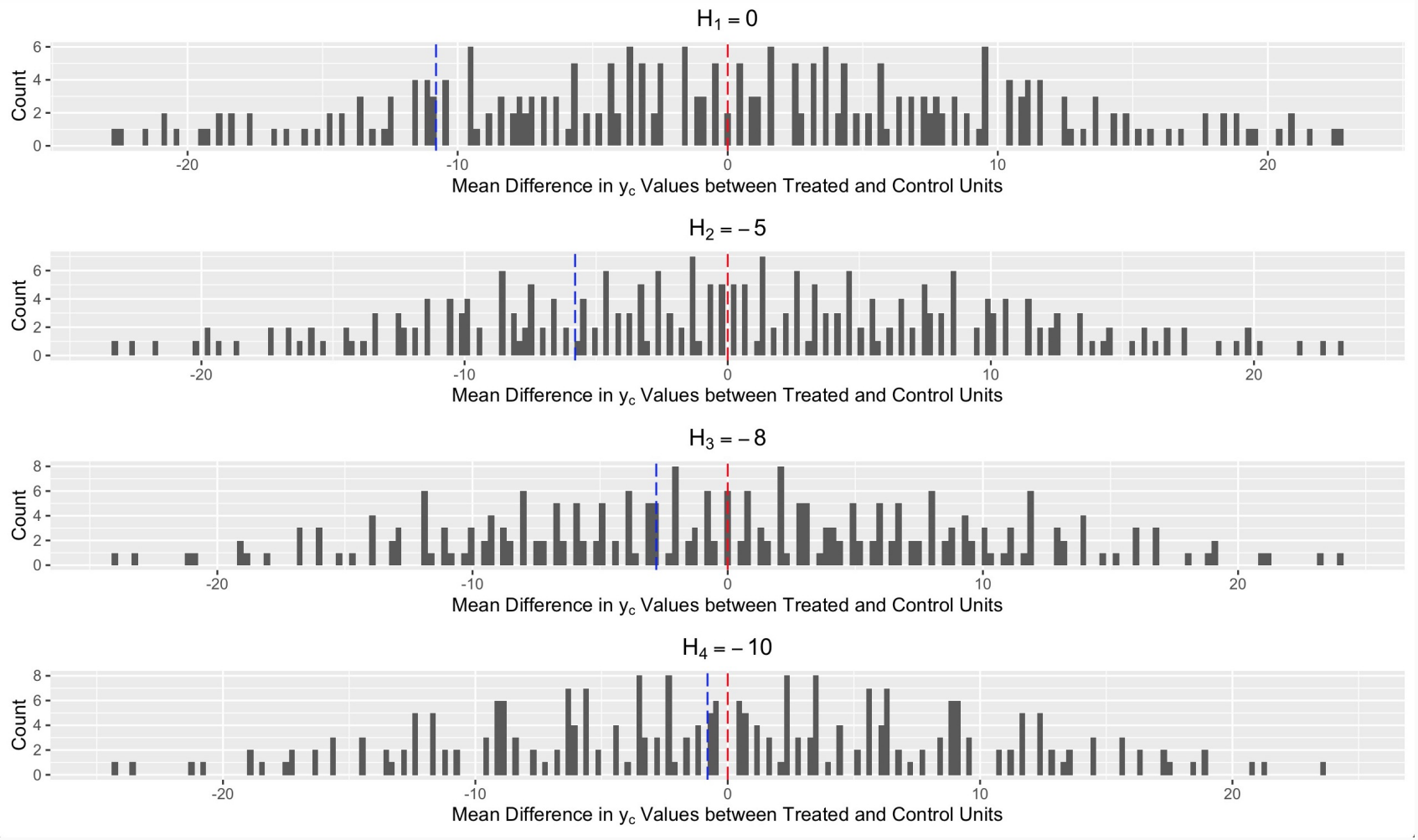

Figure 2: Randomization Distributions Under Each of the Null Hypotheses

Worked Out Example: P-Values

1. $\frac{78}{252} = \frac{13}{42} \approx 0.3095238$ for H_1
2. $\frac{146}{252} = \frac{73}{126} \approx 0.5793651$ for H_2
3. $\frac{200}{252} = \frac{50}{63} \approx 0.7936508$ for H_3
4. $\frac{232}{252} = \frac{58}{63} \approx 0.9206349$ for H_4.

Posterior Credences

One can then calculate the posterior credences in each of the four hypotheses as follows:

1. $C(H_1) = \frac{P_{H_1}(e)}{C(e)} C(H_1) = \left(\frac{\left(\frac{13}{42}\right)}{\left(\frac{41}{63}\right)} \right) \left(\frac{1}{4}\right) \approx \left(\frac{0.3095238}{2.603175}\right) \approx 0.1189024.$
2. $C(H_2) = \frac{P_{H_2}(e)}{C(e)} C(H_2) = \left(\frac{\left(\frac{73}{126}\right)}{\left(\frac{41}{63}\right)} \right) \left(\frac{1}{4}\right) \approx \left(\frac{0.5793651}{2.603175}\right) \approx 0.2225609.$
3. $C(H_3) = \frac{P_{H_3}(e)}{C(e)} C(H_3) = \left(\frac{\left(\frac{50}{63}\right)}{\left(\frac{41}{63}\right)} \right) \left(\frac{1}{4}\right) \approx \left(\frac{0.7936508}{2.603175}\right) \approx 0.304878.$
4. $C(H_4) = \frac{P_{H_4}(e)}{C(e)} C(H_4) = \left(\frac{\left(\frac{58}{63}\right)}{\left(\frac{41}{63}\right)} \right) \left(\frac{1}{4}\right) \approx \left(\frac{0.9206349}{2.603175}\right) \approx 0.3536585.$

Worked Out Example: Discussion

- The prior credences of H_3 (the true hypothesis) and H_4 both increase after observing experimental evidence, but H_4 increases by more than H_3.
- The true hypothesis, H_3, does not have the highest p-value among the four hypotheses. H_4 has the highest p-value of $\frac{58}{63} \approx 0.9206349$.
- In general, the null hypothesis that is exactly equal to the mean difference in observed y_t and y_c values—in this case, -10.8—will always have the highest two-sided p-value, which will always be 1.
- In expectation, the observed test statistic will be -8, and whenever the observed test statistic is -8, the p-value of the hypothesis that the mean unit-level causal effect is -8 will always be 1.

Worked Out Example: Discussion

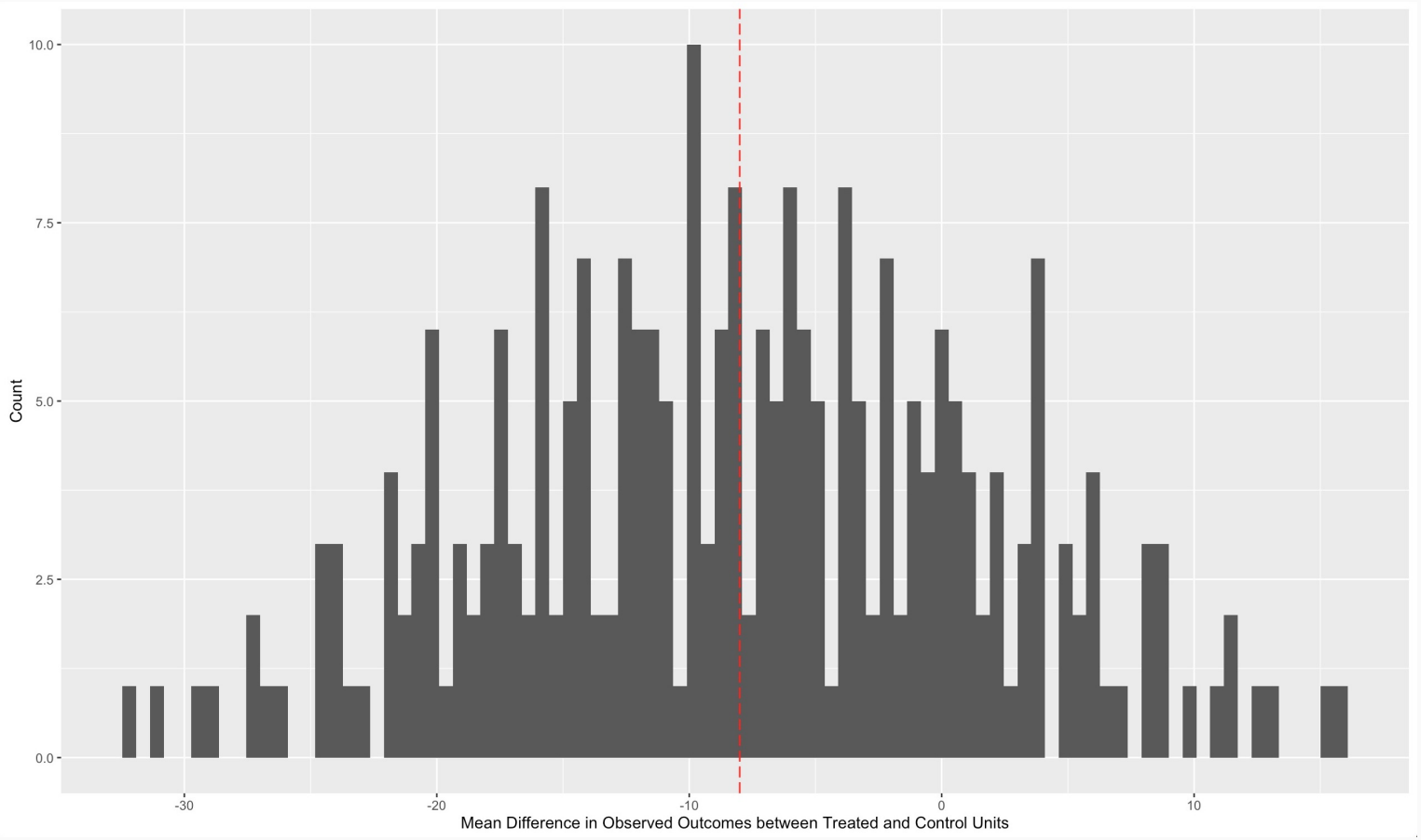

Figure 3: Randomization Distribution of Difference-in-Means Test-Statistic

Worked Out Example: Discussion

- The p-value of the hypothesis equal to the expected value is 1, but the expected p-value of the true hypothesis, $H : -8$, is *not* 1.
- The expected difference-in-means test statistic is -8, but that value occurs in only 6 out of the 252 permutations.
- A difference-in-means test statistic of -10, by contrast, occurs in 8 out of the 252 permutations even though -8 is the true average causal effect and -10 is not.
- By the finite central limit theorem, the randomization distribution converges to a normal distribution as the size of the experiment's finite population approaches infinity, in which case the expected test statistic (which is equal to the truth) is also the modal (most frequent) test statistic.
- Insofar as the randomization distribution is approximately normal, then the modal observed difference-in-means test statistic is -8, which implies that the true hypothesis will have the highest p-value more times than will any false hypothesis.

Conclusion

Conclusion

- The random assignment of experimental units to treatment and control yield *exact* p-values based only on that random assignment process and the outcome the researcher seeks to measure, not probability models of the DGP.
- Exact p-values substitute for only the likelihood function in which hypotheses ascribe probabilities to individual data (as opposed to p-values that reflect the probabilities hypotheses ascribe to the observed test statistic or greater under repeated randomizations).
- Bayesianism's incorporation of researchers' prior beliefs over *all* hypotheses remains in the approach offered herein.
- Researchers are thus able to update prior beliefs about all hypotheses via Bayesian conditionalization instead of relying on the conventional frequentist decision calculus in which only a single hypothesis is either rejected or not rejected.

Further Questions

Conclusion

- Are there any problems with using p-values to define the probability a hypothesis ascribes to empirical data?
- What about external validity?

References

Rosenbaum, P. R. (2002). *Observational Studies* (Second ed.). New York, NY: Springer.