Why Randomization Based Inference Works

Jake Caldwell & Riley Coburn

Denison University *

December 18, 2021

1 Introduction

In mathematical statistics, a problem that often comes up with basic tests (t-tests, Analysis of Variance, etc) is dealing with small sample sizes. When looking specifically at the two-sample t-test, the condition that has to be met and is taught in all introductory statistics courses is normality of both samples. However, in the case where this conditions cannot be met, we hope to have a sample size greater than 30. But this magic number 30 can be misleading. Real world data often contains more than 30 observations. Creating models such as linear regression models is difficult because real world data often fails to meet critical model conditions such as normality of residuals and homoescedasticity. Because of this, we cannot blindly trust parameters like a $\hat{\beta}_i$'s in linear regression models and residual deviance in logistic regression models. There are also times when real world data is too small and then we can't even use the Central Limit Theorem to assume our statistics come from an approximation of a normal distribution. However, there is a solution to both of these problems: randomization based inference. Another common solution is bootstrapping, however, that is most useful when we want to develop robust estimates for statistics and standard errors as well as counteract sampling error. We are focusing on randomization methods, which focus on randomization of units within our sample. Forgetting all model distribution assumptions, randomization-based inference only cares whether the sample that you have is typical of the population. In this paper, we will explain what randomization based inference is in detail while walking through a short example pointing out each step in the context of our paper, explain why this process works, elaborate on some of the limitations that come with this test, and explore Monte Carlo methods for re-randomization.

*Dr. David White

2 What is Randomization Based Inference

2.1 Definition

The general explanation goes something like this: If the data we are analyzing falls to meet our model conditions but passes the randomness assumption, we can randomly "shuffle" or permute our response variable (or treatment groups, or our multiple explanatory variables, etc.) n times to estimate a distribution of test statistics. From this distribution, we calculate the number of test statistics that are as extreme or more extreme than our original test statistic, coming from the original model we created. This proportion is then our p-value. We can trust this p-value and use it to make a judgement on whatever to reject or fail to reject our null hypothesis.

Algorithmically, suppose we have some randomly sampled dataset, A, with test statistic, S_0. We can make J-many permutations of the column of responses in A where these permuted matrices are A_j^* with test statistic S_j.

$$A = \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & y_1 \\ x_{2,1} & x_{2,2} & \cdots & y_2 \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & y_n \end{pmatrix} \longrightarrow A_j^* = \begin{pmatrix} x_{1,1} & x_{2,2} & \cdots & y_i \\ x_{2,1} & x_{2,2} & \cdots & y_i \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & y_i \end{pmatrix}$$

Where our p-value is calculated as

$$\hat{p} = \frac{\sum_{j=1}^{J} I(|S_j| \geq |S_0|)}{J}$$

2.2 In Practice

There is often more to this picture than just the above definition. Different kind of data require different kinds of permutation methods to perform statistical inference. Take a two-sample t-test for example. We start with the sampling process. Let us imagine that we have two treatments, A and B with n_1 and n_2 randomly assigned patients, respectively.

$$A = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_{n_1} \end{pmatrix} \qquad B = \begin{pmatrix} x_{n_1+1} \\ x_{n_1+2} \\ \vdots \\ x_{n_1+n_2} \end{pmatrix}$$

If from here we randomly sample n_1 patients into A without replacement and put the rest into B, we will have ensured that our new data is random as the treatments were randomly assigned in our initial sample. That is, under the assumption of the null hypothesis, if there was no difference in treatments A and B, we could assign each data point from the whole study to each treatment randomly.

$$A^* = \begin{pmatrix} x_k \\ x_k \\ \vdots \\ x_k \end{pmatrix} \qquad B^* = \begin{pmatrix} x_k \\ x_k \\ \vdots \\ x_k \end{pmatrix}$$

From here, we don't have to check our model conditions anymore because we have committed to moving forward with randomization based inference. The next step in the process will be to calculate our test statistic from this permutation of the data. For this example with two treatments, we could calculate a difference in sample means, generating a t-statistic. Following this, we repeat the initial sampling procedure and extract a difference in means t-statistic N-many times.

$$A_n^* = \begin{pmatrix} x_k \\ x_k \\ \vdots \\ x_k \end{pmatrix} \qquad B_n^* = \begin{pmatrix} x_k \\ x_k \\ \vdots \\ x_k \end{pmatrix}$$

For $k \in \{1, 2, 3 \ldots, n_1 + n_2\}$ for N-many permutations.

When this process is done, we take our t-statistics and plot them on a dotplot. The result we get is called a randomization or reference distribution. Consequently, that distribution is centered on the null hypothesis of the given test we are conducting meaning that the reference distribution we created is also a null distribution. Using this distribution, we locate where our original test statistic calculated using A and B would fall on this distribution. We then count the number of new test

statistics that are as or more extreme than the original one and divide that by the number of total test statistics calculated. This proportion that is yielded is our p-value. The error rate of this p-value has been maintained at the original significance level we started with as well. We can then take this p-value and use it to either reject or fail to reject the null hypothesis of the test we conducted.

3 Why Does RBI Work?

3.1 Layman's Terms

Let's assume we're conducting a simple linear regression. Suppose that our model failed the the homoscedasticity and normality of residuals conditions, our data is random, and we have n > 30 observations in our dataset. By the central limit theorem, we can then assume our slope coefficient, β_1 is approximately normally distributed. However, we can't wholly trust the p-value associated with that slope. This is because for statistical inference, we need certain conditions to be met depending on the test. In this example of linear regression, we need a independence, a linear relationship between our response and predictor, and homoscedasticity of residuals. If it's the case where these conditions cannot be met but we do have a sample representative of the population, we can create our randomization model. This randomization model is the distribution of n test statistics that we get from scrambling the data n times and calculating the given test statistic. Our randomization model also assumes the null hypothesis of the test statistic being zero or no statistically significant effect being present, creating what is called a null distribution. This is because the model has no parameters and that each time the data is shuffled, the observations become independent of their outcomes. Our original model, and typically most models, are tested at a significance level of $\alpha = .05$. Because our randomization model assumes observations are independent of their outcomes, a constant significance level, or type 1 error rate, is held constant for the randomization model. Furthermore, by creating this randomization model, our randomization distribution is approximately the null distribution of whatever statistic we're calculating. Now, we calculate the p-value from the proportion of test statistics calculated from our randomly permuted dataset that are as or more extreme than our original test statistic. We can trust this p-value because the p-value comes from a now uniform distribution.

3.2 Sharp Null Hypothesis & Asymptotic Normality and p-value

As with many concepts in statistics and probability, the basis for randomization-based inference lies in the Central Limit Theorem. Let's stick with our simple linear regression example from above. As we saw, when we permute our data in simple linear regression, we are basically saying that there is

independence between x and y. This is equivalent to the null hypothesis, $H_0 : \hat{\beta}_1 = 0$. In fact, for many permutations, this is equivalent to Fischer's sharp null hypothesis which says that there is no effect for each pairwise set of x and y values. Traditionally Fischer's exact test is defined as

$$H_0 : Y_i(1) = Y_i(0) \quad \forall i = 1, 2, \ldots, N \text{ for } N \text{ many permutations.}$$

As our example is dealing with linear regression we can't have this specific null hypothesis, but instead we can interpret it as

$$H_0 : \hat{\beta}_{1,i} = 0 \quad \forall i = 1, 2, \ldots, N.$$

Now, let's recall the Central Limit Theorem. The CLT says that given a population with mean μ and standard deviation σ and we take sufficiently large random samples from the population with replacement, then our sampling distribution will be approximately normally distributed about μ. What does this mean for our case of simple linear regression? Well, we can think about the population of least squares estimators as every $\hat{\beta}_1$ we would get if were to permute the response column. A critical part of this is that our data from which we are calculating the least squares estimator is a representative random sample from the population. If so, then we can validate this claim that our distribution of least squares estimators is representative of the population as required by the CLT. This is crucial if we are to conduct statistical inference as randomization-based inference spells out. This is the reason we are able to generalize and make inferences about the population, even when model conditions aren't met. Now, with this in mind, each time we take a permutation of our data, we are taking a random sample from the population of $\hat{\beta}_1$'s. As we saw above, the permutations we make lead to the assumption of the null hypothesis, $H_0 : \hat{\beta}_1 = 0$. Therefore, by the central limit theorem, our reference distribution will be centered around zero because under the null hypothesis we have $E(\beta_1) = 0$ and for a normal distribution we have $E(X) = \mu$. If we take all $n!$ permutations of an n-length response vector we would have an approximately normal distribution. As we take more permutations and include each newly calculated test statistic in the randomization distribution, we will only further approach normality to the point where more permutations with neither help nor hurt the normality of our reference distribution. This type of distribution is called an asymptotic distribution, thus, we have asymptotic normality in our reference distribution.

Since our distribution in asymptotically normal, our p-value is not only valid, but it will be approximately exact, depending on the number of permutations that we run. If we reach the point where we have approached asymptotic normality, we would not just be approximating a p-value through statistical theory, but actually calculating it on an asymptotic normal distribution. By the same virtue

as we showed asymptotic normality, we have an asymptotic p-value since we are calculating a p-value from this asymptotic distribution. We cannot take any more permutations that will influence the p-value since the distribution we are calculating the p-value on is nearly fixed for large numbers of permutations.

3.3 In General

This process that we took is true for many test statistics that we can choose. If we know that our sample is truly an approximation of the population, we are able to construct a roughly normal randomization centered around the null hypothesis of our test. Examples of other test statistics that we can use this approach for are means, proportions, difference in means, difference in proportions, and correlations. The same is true for χ^2- and F-distributions, however there is a bit more legwork that needs to happen before we can see this clearly. The formula for a χ^2 test statistic is

$$\sum \frac{(O_i - E_i)}{E_i}$$

where O_i is each observed data point and E_i is the expected value for that group. Assuming that our data has been randomly sampled from the population, one of the assumptions that we hold for randomization based inference, then O_i will be normal random variable meaning the difference $O_i - E_i$ is also a normal random variable. Assuming the null hypothesis of no relationship (as we do when we shuffle the data), $E(O_i - E_i) = 0$. Thus, the χ^2-distribution will be a sum of squared normal random variables taken from a standard normal distribution. As we shuffle the data and sample more normal random variables from the standard normal distribution, the sum of their squares will only further approximate a χ^2-distribution. From this randomization distribution we can calculate a p-value by taking the proportion of χ^2 statistics that are greater than what we observed in our original sample over the total number of permutations.

Similarly, we can show that our randomized F-distribution will in fact be F-distributed using the same reasoning as we did previously to show that random permutations of our data will lead to a χ^2-distribution. Again, we're assuming the null in this case when we randomize our data. For an F-test, the null hypothesis is that both models or groups are the same. When we calculate our proportion that we do for an F-statistic, we will be taking a proportion of χ^2 random variables (multiplied by a factor of the inverse of each χ^2 random variable's degrees of freedom). By definition, this will be an F random variable. Thus with many permutation we will have the null distribution of F-statistics which is centered at 1. Since we've satisfied that we can approximate our necessary distribution, we

can calculate a p-value just as above.

4 Limitations

Although what we have discussed in this paper so far is robust, there are some chinks in Randomization Based Inference's armor.

1. **Random Data Condition**

   The first and most glaring limitation of RBI is that it is not truly conditionless. The one condition that it does have is that the initial data has to be random. Some basic cases where this is not met is when you have access to the whole population, in which case you do not need randomization based inference, or when collecting data you failed to use a randomization procedure. The latter is the situation we are more focused on in this paper. In general, when a sample fails to be random, it is no longer assumed to be representative of the population. We can connect this to randomization based inference in that we are trying to find an asymptotically correct p-value. In order to do this, we have to randomly shuffle the data and extract our test statistic and build the randomization distribution and then calculate the p-value from that. This process becomes invalid when the original data fails to be random because of two key issues. We can start with the simpler one: The original sample contains high levels of bias. Let's assume that we are trying to figure out if a bill on an upcoming ballot does not have equal support in a community. Instead of using your city hall and a random number generator to sample people, you decide to ask everyone who walks into the local Hobby Lobby. In your mind, this seems reasonable, considering that you don't know who is going to walk in and what they are thinking about the bill. However, you have failed to consider the population of people who shop at Hobby Lobby. This can introduce bias into your sample. It could be response bias or under-coverage bias (specific groups of people left out). The fatal flaw that both the former and the latter present is that they alter the null hypothesis of your test. If you were to continue with your experiment and use randomization based inference, you won't be able to trust any of your permuted test statistics and your final p-value due to the bias present.

   The second and more complex issue of not obtaining a random sample focuses on the obvious connection between a random sample and randomization based inference. The main point to be made here is that if a sample is not random, we cannot assume it to be representative of the population our sample comes from. In the Hobby Lobby story, the sample taken could be representative of the population. Here, even if some randomized sampling procedure is taken, the

sample could still be not random. Typically when we code in R, we use tests like Bartel's test for randomness to determine if data has been sampled randomly. If the test fails to reject the null hypothesis that the data is random, then we are fair to move forward with randomization-based inference. When the null hypothesis is rejected, randomization-based inference will not work. This is due to the fact that we are typically using a test to make an inference about the population. We use the term correlation, because our tests can't conclude causation. Regardless of the prior statement, the main problem here is that our random variables, no matter the type of test, are not approximately normally distributed. This is because the Central Limit Theorem needs randomly distributed variables for it to apply. When a sample is not random, the variables we are observing are also not random. This makes the random variable we're observing to be not approximately normally distributed. No matter the test statistic calculated we will not be allowed us to trust our test statistic, causing a breakdown of the entire randomization process.

In summation, in order for our randomization based inference to be of value, we need a sample free of bias and that is random. This ensures that our variables are random variables and we can trust the test statistic created from them.

2. **Small Sample Sizes**

   For most tests, there is typically a condition that requires the initial sample to be of an acceptable size. For example, when trying to determine if two different treatment groups have a difference in means that is not equal to zero, we require that for the Central Limit Theorem to work, we need at least 30 data points, assuming our sample is random. Although randomization based inference can generate many permutations of a given sample, the one thing it can't do is increase the sample size. The randomization based inference does not deal with the counterfactual statement that "if the sample size was larger, we could be more sure of our conclusion". Instead, the point here is that our randomization distribution using a small original sample size will be less approximately normal than a satisfactory sample size. This affects the robustness of our randomization based inference. Sure, it created something that could pass as normal to any given statistician, but the level of asymptotic certainty we have is in jeopardy.

3. **Population Parameter Estimation**

   One of the important aspects of model creation in statistics is the idea of parameter estimation and statistical accuracy. We often fit one model at the beginning of our experiment and then go through rigorous testing and fitting procedures to develop the best model. This best model

is not only accurate at its forecasting and initial predictions, but also does not over-fit to the data so that it can be generalized to new data. The limitation we are getting at here is that randomization based inference is not something we can use to develop better population parameter estimates.

In general, we use randomization based inference to help us figure out if the perceived effects we see in our original model are trustworthy. However, if we are doing linear regression, our beta estimates are only as good as the original sample we have. In randomization based inference, we resample data without replacement. This means we assume our null hypothesis is true. However, if we wanted to make sure our sample statistics were better at estimating the true population parameters given our test, we would have to take a different approach. Instead of just randomly shuffling our response variable, we could take all of our observations and pool them back up, then choose one response value at random and assign it to a given observation. We then could put the response value back into the pool of responses and repeat the process. This is sampling with replacement. Instead of building a randomization distribution which looks to estimate the distribution of the given test statistic, this estimates the sampling distribution. This estimated sampling distribution is centered around our original sampling mean and has a standard deviation from our original sample as well. However, with this distribution, we can make confidence intervals that give us more precise estimates of the population parameter. This process is referred to as bootstrapping. In most cases, we should really pair these models so that we can have better model accuracy as well as more clarity around the effect we are testing for. However, the computational acumen required for both tests is quite high. Lastly, the level of interpretability when using both methods of resampling is murky due to the different procedures used. Depending on what your goal is, picking one of these methods will be more useful than including both.

5 Monte Carlo Simulations

5.1 Approximation of π

We will use this toy example as a brief introduction to Monte Carlo Simulations. Lets assume that we are trying to find an approximation for π. We can still do this using the same randomization-based process as before. We will use the statistical and programming software R for our process. This is a particularly interesting example, as there are very little conditions to actually be satisfied. The only condition that needs to be met as has been said time again is that the sample is representative of the

population. In our case, the population is the points $x, y \in R : x, y \sim U(-0.5, 0.5)$.

```r
runs <- 100000
x <- runif(runs,min=-0.5,max=0.5)
y <- runif(runs,min=-0.5,max=0.5)
in.circle <- x^2 + y^2 <= 0.5^2
mc.pi <- (sum(in.circle)/runs)*4
```

We've made sure that that x and y are representative of each population by selecting a random value from $U(-0.5, 0.5)$ using the **runif()** command which stands for "random uniform". Rather than a least squares estimator, in this case, our test statistic will be the Bernoulli random variable representing whether a point is contained within the circle which inscribes our population distribution. Now that we have our test statistics, we can approximate π. Empirically, $A_{square} = (2r)^2 = 4r^2$. Suppose we don't know a value for π, but instead know that $A_{circle} \propto r^2$. That is, $A_{circle} = cr^2$ where c is a positive constant. We can now approximate π where c is our approximation. This is as simple as taking the proportion of A_{circle} and A_{square}.

$$\frac{A_{circle}}{A_{square}} = \frac{cr^2}{4r^2} = \frac{c}{4} \longrightarrow \frac{4 \cdot A_{circle}}{A_{square}}$$

The simulated A_{circle} that we have is just the total number of points inside of the circle while A_{square} will be the total number of points that we sampled. The approximation for π in this example is 3.14252, not too far off from the true value of 3.14159.

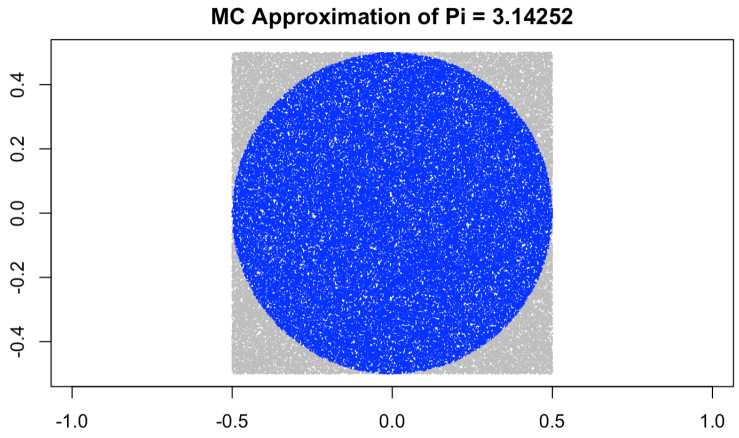

Figure 1: Plot of sampling distribution and inscribed circle

5.2 Monte Carlo Simulation and RBI

The Monte Carlo simulation is a do-it-all kind of method which has many useful applications in statistics and mathematics in general. For randomization based inference, Monte Carlo simulations manifest themselves as a streamlined way to get the asymptotically correct p-value we want. The process is applied as follows: First, instead of taking $n!$ permutations of our data, we take a random sample of permutations usually between 5 to 100. We then calculate the proportion of those permutations that yield a test statistics that are as or more extreme than our original test statistic. Now we have our random variable P (the aforementioned proportion) with which we create a 95% confidence interval using the largest possible binomial standard deviation, $\sqrt{n * (0.5)(1 - 0.5)}$. Thus, we create a sufficiently wide confidence interval that, if created over and over again, will contain the true p-value for our effect 95% of the time.

6 Conclusion

Throughout this paper, we've discussed the importance of randomization-based inference in statistics, the situations that this method is best attuned to, and most importantly, why this method works for so many statistical hypothesis tests. We've also briefly touched on Monte Carlo Simulations and the importance of these methods in performing accurate randomization-based inference. Randomization-based inference is a powerful non-parametric tool which specifies that we only check one condition: randomness. This creates an extremely versatile process which can be implemented in situations where other model conditions cannot be met in some cases. As with many topics in statistics, we learned that the utility for randomization-based inference derived from the Central Limit Theorem.

References

Chihara, L., & Hesterberg T. 2019. “Mathematical statistics with resampling and R.”.

Ding, Peng. 2017. “A Paradox from Randomization-Based Causal Inference.” *Statistical Science* 32(3):331 – 345.

Hesterberg, Tim. 2014. “What Teachers Should Know About the Bootstrap: Resampling in the Undergraduate Statistics Curriculum.” *The American Statistician* 69.

Janssen, Arnold and Thorsten Pauls. 2003. “How Do Bootstrap and Permutation Tests Work?” *The Annals of Statistics* 31(3):768–806.

Kruschke, John K. 2015. “Doing Bayesian Data Analysis.”.

Mammen, E. 1992. “Bootstrap, wild bootstrap, and asymptotic normality.” *Th. Rel. Fields* (93):439–455.

Owens, Art. 2015. “The Simple Monte Carlo.”.

Raychaudhuri, Samik. 2008. “Introduction to Monte Carlo simulation.” pp. 91–100.

Wang, Y., Rosenberger W.F. & Uschner D. 2019. “Randomization-based inference and the choice of randomization procedures.” pp. 395–404.