Permutation Methods

Applied Statistics

Rui M. Castro

Permutation Methods

A simple but powerful non-parametric idea originating with Fisher and Pitman (1937):

• Before computers became pervasive these methods were viewed as **computationally cumbersome** - and were brushed aside for a long time.

• Provide a unified framework for the study of rank-based tests

Bradley (1968, p. 85):
Eminent statisticians have stated that the randomization test is the truly correct one and that the corresponding parametric test is valid only to the extend that it results in the same statistical decision.

Permutation vs. Randomization Tests

Typical Setting: We have observations from two different groups, or under two different conditions.
Question: Is there a difference between the two groups or conditions?

- **Randomization Model:** available subjects are randomly assigned to treatments. The resulting inference pertains only the observed individuals (there is no concept of a population)
  - **Randomization Tests**

- **Population Model:** subjects are randomly sampled from different subpopulations
  - **Permutation Tests**

Advantages: conceptually very easy, and results in exact inference procedures.
Disadvantages: Cannot be used for every testing problem and is computationally intensive

The Randomization Model - Example

Basis: subjects are randomly assigned to different treatments (usual practice in medicine)

- The only random aspect of the model is the assignment of treatments
- Inference is limited to subjects under study (there is no population)

Typical Setting:

N patients are randomly assigned to treatment/control group. The effect to treatment/control is recorded:

A treatment to speed-up post-surgical recovery: n patients are randomly assigned a specific treatment, and $m = N - n$ are assigned to a control group and given a placebo treatment. The corresponding recovery times are recorded:

$$\underbrace{x_1, x_2, \ldots, x_m}_{\text{control}} \qquad \underbrace{y_1, y_2, \ldots, y_n}_{\text{treatment}}$$

The Randomization Model - Example

A treatment to speed-up post-surgical recovery: n patients are randomly assigned a specific treatment, and $m = N - n$ are assigned to a control group and given a placebo treatment. The corresponding recovery times are recorded:

$$\underbrace{x_1, x_2, \ldots, x_m}_{\text{control}} \qquad \underbrace{y_1, y_2, \ldots, y_n}_{\text{treatment}}$$

More specifically let $N = 7$, $n = 4$ and $m = 3$ and

$$(x_1, x_2, x_3) = (23, 33, 40) \qquad \textbf{(control)}$$

$$(y_1, y_2, y_3, y_4) = (19, 22, 25, 26) \qquad \textbf{(treatment)}$$

A sensible statistic: $t = \bar{y} - \bar{x}$

In this case $t = -9$. It seems sensible to state the treatment makes a difference if t is small, but "how small is small"?

The Randomization Model - Example

Notation: Identify the patients with numbers $1, \ldots, N$ and let $z_1, \ldots, z_N$ denote the recovery times of the N patients. Let d be the set of patients receiving treatment

$$d = \{i \in \{1, \ldots, N\} : \text{patient } i \text{ received treatment}\} \ .$$

Clearly

$$\bar{x} = \frac{1}{m} \sum_{i=1}^{m} z_i \mathbf{1}\{i \notin d\} \quad , \quad \bar{y} = \frac{1}{n} \sum_{i=1}^{n} z_i \mathbf{1}\{i \in d\}$$

Two hypothesis:

H_0 : no difference between treatment/control

H_1 : treatment reduces recovery time

Test statistic: $t_d = \bar{y} - \bar{x}$

The Randomization Model - Example

The only random aspect of this setting is the random assignment of treatment. So we can check if randomly chosen assignements would result in very different outcomes...

Let D be a subset of $\{1, \ldots, N\}$ with cardinality n, chosen uniformly at random over all the possible $\binom{N}{n}$ possible sets.

Define

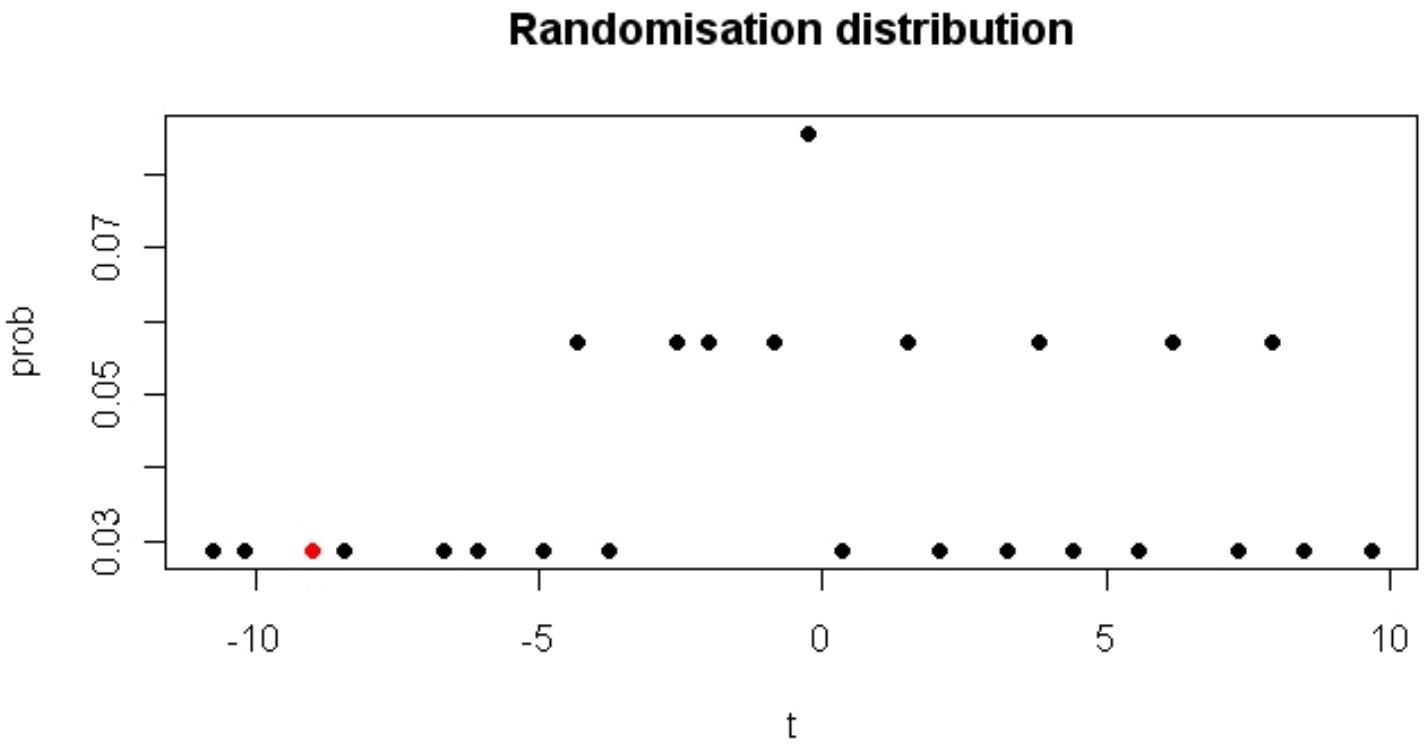
$$t_D = \frac{1}{n} \sum_{i \in D} z_i - \frac{1}{m} \sum_{i \notin D} z_i \ .$$

This is our test statistic, and we observed the value t_d.

Under the null hypothesis the observed value t_d should be a sample from the distribution t_D. We can easily compute this distribution by trying out all possible assignments!!!

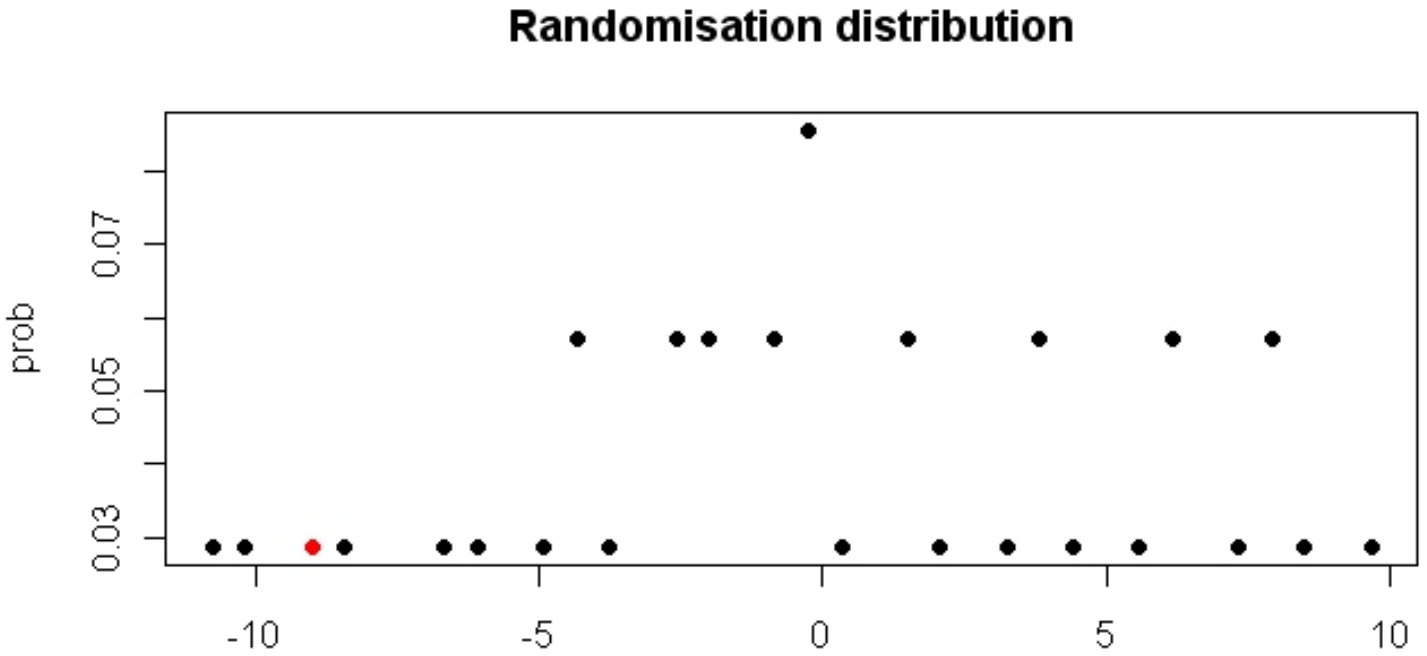
The Randomization Model - Example

No.	X_1	X_2	X_3	X_4	Y_1	Y_2	Y_3	t_D
1	19	22	25	26	23	33	40	-9.00
2	22	23	25	26	19	33	40	-6.67
3	22	33	25	26	19	23	40	-0.83
4	22	25	26	40	19	23	33	3.25



The observed value t_d is quite extreme.

The Randomization Model - Example



The p-value of the randomization test is simply given by

$$\mathbb{P}_{H_0}(t_D \leq t_d) = \frac{1}{\binom{N}{n}} \sum_{d' \in \mathcal{D}} \mathbf{1}\{t'_d \leq t_d\} \ ,$$

where $\mathcal{D}$ is the set of ALL possible assignments.

In the example $p = \frac{3}{35} = 0.0857$.

The Randomization Model

These tests are "exact", in the sense that we can control exactly the probability of type I error. However, we need to be a bit careful as the possible values of the test statistic are discrete, and therefore the test cannot be exact for an arbitrary significance level.

Lemma:

If the significance level is $\alpha = \frac{k}{\binom{N}{n}}$ and we take as critical region $(-\infty, t_{(k)}]$ the randomization test is exact, that is

$$\mathbb{P}_{H_0}(\text{reject } H_0) = \alpha \ .$$

In the above, $t_{(k)}$ is the kth smallest value in $\{t'_d : d' \in \mathcal{D}\}$.

Remarks: we can obvioulsly use other test statistics, e.g. difference in the group medians, or the sum of responses in one of the groups:

$$s_D = \sum_{i=1}^{n+m} z_i \mathbf{1}\{i \in D\} \ .$$

Exercise: Show that this test statistic is equivalent to t_d.

The Randomization Model

More Remarks: Asymptotically the randomization test results in the same inference procedure as the 2-sample t-test (although the basis of the inference is completely different).

If we apply these ideas for ranks we recover the famous Wilcoxon-rank-sum test.

Bradley (1968, p. 85):
Eminent statisticians have stated that the randomization test is the truly correct one and that the corresponding parametric test is valid only to the extend that it results in the same statistical decision.

The Population Model

Basis:

- Data is assumed to correspond to **random** samples from different populations.
- This is a stronger assumption than in the previous setting but...
- Conclusions can be generalized to populations.

$$X_1, \ldots, X_m \overset{i.i.d}{\sim} F \quad \text{and} \quad Y_1, \ldots, Y_n \overset{i.i.d}{\sim} G \ .$$

$$H_0 : F = G \qquad \text{versus} \qquad H_1 : F \neq G$$

Remarks: The mechanics is the same as before, but the reasoning is different:

Under the null hypothesis all samples come from the same distribution. Therefore, conditionally on the values of $\{X_1, \ldots, X_m, Y_1, \ldots, Y_n\}$ any possible assignment into two different populations is equally likely.

The Population Model

$$X_1, \ldots, X_m \overset{i.i.d}{\sim} F \quad \text{and} \quad Y_1, \ldots, Y_n \overset{i.i.d}{\sim} G \ .$$

$$H_0 : F = G \qquad \text{versus} \qquad H_1 : F \neq G$$

Let $N = n + m$ and define

$$Z_1 = X_1, \ldots, Z_m = X_m, Z_{m+1} = Y_1, \ldots, Z_N = Y_N \ ,$$

and denote by $z_1, \ldots, z_N$ the observed values.

Define the event

$$\begin{aligned} E &= \left\{(Z_1, \ldots, Z_N) = (z_{p(1)}, \ldots, z_{p(N)}) \text{ for some permutation } p\right\} \\ &= \left\{(Z_{(1)}, \ldots, Z_{(N)}) = (z_{(1)}, \ldots, z_{(N)})\right\} \ . \end{aligned}$$

The idea is that we are going to condition on the event that the order statistics are the ones we observed (event E)

The Population Model

Define the event

$$E = \left\{ (Z_1, \ldots, Z_N) = (z_{p(1)}, \ldots, z_{p(N)}) \text{ for some permutation } p \right\} .$$

Clearly

$$\mathbb{P}_{H_0}(Z_1 = z_1, \ldots, Z_N = z_N | E) = \frac{1}{N!} ,$$

so we have

$$\mathbb{P}_{H_0}(\text{observed division over groups} | E) = \frac{1}{\binom{N}{n}} .$$

As before, all we have to do is to construct a test statistic T_d, and compare it with a (conditional) reference distribution T_D under the null.

The Population Model

For instance let

$$T_d = \left|\bar{X} - \bar{Y}\right| \ ,$$

so we will test if the means of the two populations are significantly different.

The conditional p-value of this test is simply given by

$$p = \mathbb{P}_{H_0}(T_D \geq T_d | E) = \frac{1}{\binom{N}{n}} \sum_{d' \in \mathcal{D}} \mathbf{1}\{T'_d \geq T_d\} \ .$$

Lemma:

If the significance level is $\alpha = \frac{k}{\binom{N}{n}}$ and we take as critical value T' the permutation test is exact, that is

$$\mathbb{P}_{H_0}(\text{reject } H_0 | E) = \mathbb{P}_{H_0}(T_D \geq T' | E) = \alpha \ ,$$

where T' is the kth largest value in $\{T_{d'} : d' \in \mathcal{D}\}$.

The Population Model

Lemma:

If the significance level is $\alpha = \frac{k}{\binom{N}{n}}$ and we take as critical value T' the permutation test is exact, that is

$$\mathbb{P}_{H_0}(\text{reject } H_0|E) = \mathbb{P}_{H_0}(T_D \geq T'|E) = \alpha \ ,$$

where T' is the kth largest value in $\{T_{d'} : d' \in \mathcal{D}\}$.

Note that, because the test controls the type I error regardless of event E, this test is also **unconditionally exact**. This is why we can issue meaningful statistical statements about the populations.

Let's see an example...

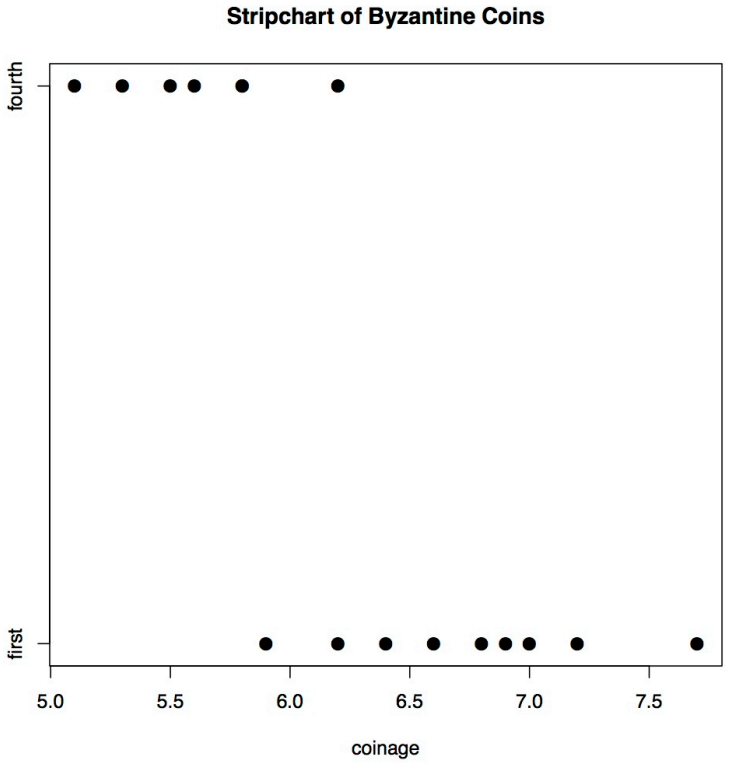
The Population Model - Example

Byzantine coins: We wish to investigate the silver content (%Ag) of a number of Byzantine coins discovered in Cyprus. The coins are from the first and fourth coinage in the reign of King Manuel I, Commenus (1143-1180). First we visualize the data, using a stripplot.

```
c1<-c(5.9, 6.8, 6.4, 7.0, 6.6, 7.7, 7.2, 6.9, 6.2)
c4<-c(5.3, 5.6, 5.5, 5.1, 6.2, 5.8, 5.8)
    coindata<-data.frame(coinage=c(c1,c4),
        type=c(rep("first",9),rep("fourth",7)))
    stripchart(coinage~type,data=coindata,cex=1.5,pch=16,
        main='Stripchart of Byzantine Coins')
```

In this case let X's denote the data relative to the first coinage and Y's the data pertaining the fourth coinage, therefore $m = 9$, $n = 7$.

The Population Model - Example



Suppose we want to test the null hypothesis that there is no difference among the silver content in different coinages versus an alternative where there is a difference.

A reasonable test statistic is

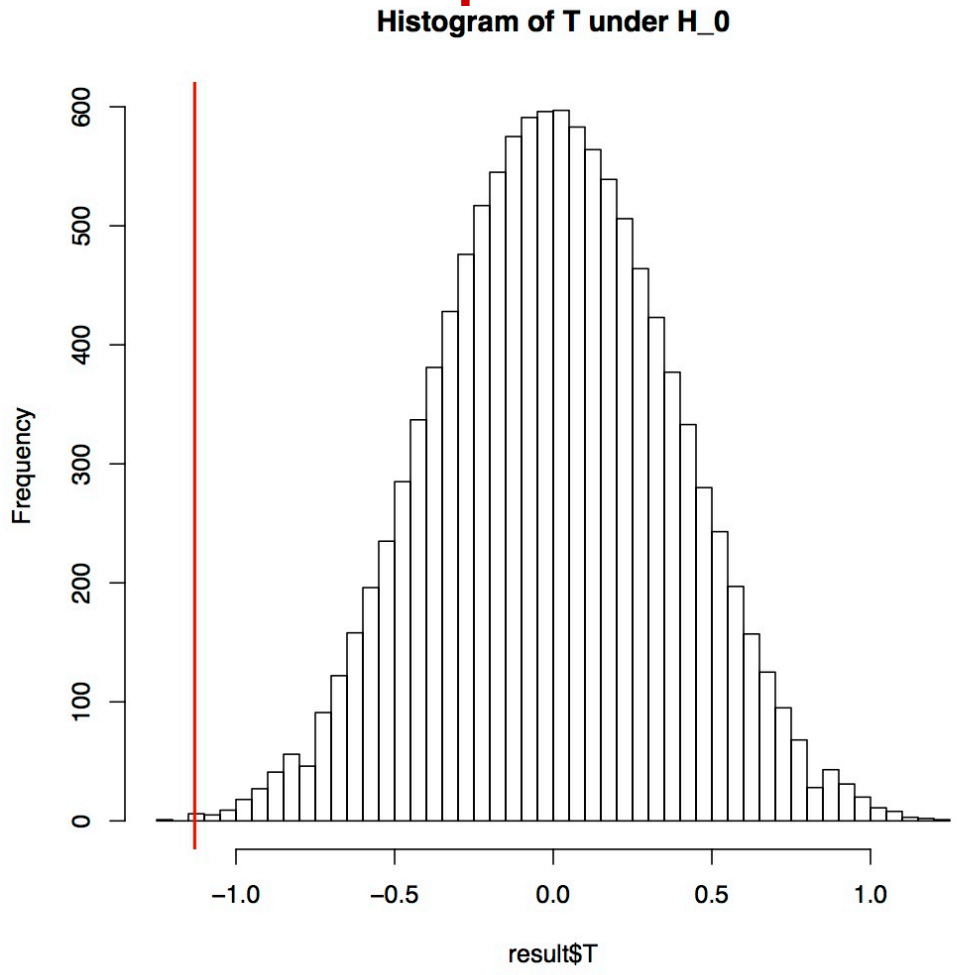
$$T = \bar{Y} - \bar{X} \ ,$$

where X_i correspond to the data relative to the first coinage and Y_i corresponds to the data of the fourth coinage.

Since the sample is relatively small there are only $\binom{16}{9} = 11440$ permutations.

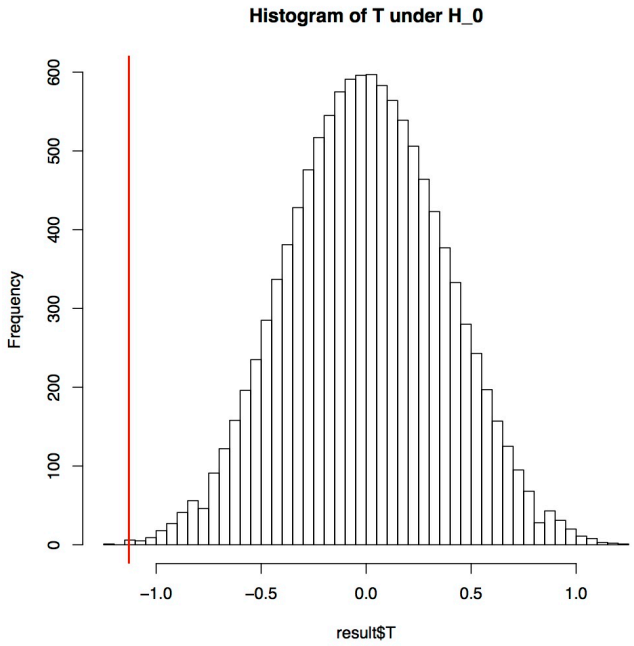
The Population Model - Example

```
twosample.perm.test<-function(x,y)
{
   z<-c(x,y)
   m<-length(x)
   n<-length(y)
   N<-n+m
   mat<-combn(N,m)

   poss<-ncol(mat)
   t<-teststat(x,y)
   T<-numeric(poss)
   for ( i in 1:poss)
   {
      x<-z[mat[,i]]
      y<-z[-mat[,i]]
      T[i]<-teststat(x,y)
   }
   return(list(T=T,t=t))
}

teststat<-function(x,y)      mean(y)-mean(x)

result<-twosample.perm.test(c1,c4)

hist(result$T,50, main='Histogram of T under H_0');abline(v=result$t, col='red', lwd=2);

p_value<-mean(abs(result$T-mean(result$T))>=abs(result$t-mean(result$T)))
cat('P-value for the test ',p_value,'\n')
```



There are better ways to code this!!!

The Population Model - Example



Observed value is quite extreme. We can safely reject the null hypothesis with high confidence.

The p-value can be computed as described in the article by Ernst, and we get $p = 0.000699$, so we can quite confidently reject the null hypothesis.

The conditional inference (`coin`) package can also do the calculations (and much faster) using the following code.

```
oneway_test(coinage~type,data=coindata,distribution="exact")
```

This package offers much more functionality for computing p-values for permutation tests. However, it takes some study to understand what's going on and what user input is necessary.

Monte-Carlo Computations

Computation of all the possible permutations is seldomly possible for most problems. For instants if $N = 30$ and $n = 15$ we have

$$\binom{N}{n} \approx 155 \times 10^6 \ .$$

Often one can enumerate permutations efficiently, and are able to compute distributions involving millions of permutations.

Nevertheless, it is quite easy to sample from the permutation distribution!!! Therefore, we can estimate p-values easily using Monte-Carlo methods.

Monte-Carlo Computations

Suppose the test statistic is t_d and we want to compute the p-value for a one-sided test rejecting the null if $t_d \geq \tau$. We can generate M i.i.d. samples from the test statistic under the null and estimate the p-value as

$$\widehat{p} = \frac{1 + \sum_{i=1}^{M} \mathbf{1}\{t_i \geq t_d\}}{M+1} \ .$$

Since we compute $\widehat{p}$ from a binomial random variable we can actually get a CI for p, and therefore we can easily get a conservative bound for the p-value.

Final Remarks

When applicable, permutation methods are a sensible way to perform statistical inference. Furthermore, given the computational power available to us these are also extremelly practical.

Not all statistics can be used with permutation methods. Under the null the distribution of the test statistic must be invariant under permutations. This might not always be the case, for instance:

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{S_X^2/m + S_Y^2/n}}$$

This is the t-statistic for comparing means when sampling from two independent normal populations with possibly different variances. If the variances are different then the distribution of T is not invariant under permutations, even when the means are identical.