12-2009

Nonparametric Inference Using Randomization and Permutation Reference Distribution and Their Monte-Carlo Approximation

Minh D. Nguyen
Portland State University

Nonparametric Inference using Randomization and Permutation Reference Distribution and their Monte-Carlo Approximation

Minh D. Nguyen
Master of Science in Statistics
Portland State University, Portland, OR

Statistic 501 Project
Advisor: **Dr. Mara Tableman**
Second Reader: Dr. Yunming Mu

December 7, 2009

Abstract The purpose of this paper is to explain the concepts and replicate some results in the article *Permutation Methods: A Basis for Exact Inference* by Michael D. Ernst (2004). In the paper, we discuss how to perform nonparametric inferences using randomization and permutation reference distributions. We clearly explain the differences and similarities between the two distributions and show how to use Monte-Carlo sampling to approximate them. In the process, we introduced and proved two theorems that are the basis for some of the inferences. Specifically, one of the theorems describes the exact p-value of a two-sample test as a left-continuous decreasing step function of the treatment effect Δ, which enables us to construct a confidence interval for Δ. This description appears to be new and the proof is independently derived. The second theorem is a basis for constructing a confidence interval for this exact p-value based on binomial distribution which results from the Monte-Carlo approximation. This theorem was given as an exercise in a book by Casella and Berger (C-B, 2002, Exercise 9.21, page 451). In addition, we introduce two self-written R functions, `pval` and `cint`, and demonstrate how to use them to replicate all the results in the article. Readers will see that, in some situations, the two functions are more versitile than standard packages such as Resampling Stat (Resampling Stats, Inc., 2003) used by Higgins (2004) in his book on the topic of nonparametric method and StatXact (Cytel Software Corporation, 2003) used by Ernst in his article. The functions will soon be put in a standard R package for free distribution to be used in classroom, study, and research.

Key words and phrases Nonparametric inference, permutation method, distribution free, randomization reference distribution, permutation reference distribution, Monte-Carlo sampling

1 Introduction

A common problem in many parametric tests is that they rely on the assumptions made on the distributions of the populations where the samples are drawn from. It is always assumed that the forms of these distributions can be exactly specified except on a finite number of parameters. For example, in a two-sample test, the parametric approach is to use either the two-sample Z statistic (if the population variances are assumed known), the two-sample pooled T statistic (if the population variances are assumed unknown but equal), or the two-sample general T statistic (if the population variances are assumed unknown and not equal.) In any case, it is always assumed that the populations where the samples are drawn from are normally distributed. If not, the chosen statistics do not have their corresponding reference distribution. As a result, Type I error can not be controlled and the test result could be erroneous.

In light of this problem, alternative approaches that do not rely on the assumptions made on the distributions of the populations where the samples are drawn from have been devised, and they are the foundations for today's nonparametric statistics. For example, in a two-sample test, the basic idea of the nonparametric approach is to generate many permutations of the data into the two samples. For each permutation, the value of a chosen statistic is computed, and the set of all the values computed for all the permutations generate a reference distribution for the chosen statistic. Under the null hypothesis, this reference distribution can be exactly specified, and most importantly, it does not reply on any assumptions made on the distributions of the populations where the two samples are drawn from. Under the randomization model, this reference distribution is called randomization reference distribution, and under the permutation model, it is called permutation reference distribution. We will discuss the two models in details in the paper. In any case, this method of obtaining a reference distribution based on the many permutations of the data is called permutation method.

Permutation method, basically the method of permuting data, dates back to Fisher (1936) who wrote that "the statistician does not carry out this very simple and very tedious process, but his conclusions have no justification beyond the fact that they agree with those which could have been arrived at by this elementary method." What Fisher wrote not only confirms the validation of the method but also reflects its limitation back in the old days, which is the reason for its unpopularity until now. To get an idea of of this "simple and very tedious" process that many statisticians did not carry out, let's look at an example. Suppose we have a sample of 6 data points and we want to permute them to two samples of 3 each, then we have a total of ${}_6C_3 = 20$ unique permutations which can be written down nicely. However, if we have a sample of 30 data points and we want to permute them to two samples of 15 each, then we have a total of ${}_{30}C_{15} = 155,117,520$ permutations, which is impossible for any statistician to write down. Thus back in the old days, "the methods were thought of as quick, hand-calculation methods, suitable only for simple designs and small data sets," wrote Higgins (2004). However, with today's powerful computing capacity, we have no reason not to use the method. Our self-written R functions, `pval` and `cint`, are examples of this. We discuss in detail how to use them to generate the permutations of the data, and based on these permutations, genererate the appropriate reference distribution, which are used in all of the inferences.

We begin the paper by introducing the randomization model in a two-sample experiment. We explain what it is and discuss the basis for generating exact randomization reference distributions of some chosen statistics. We then show how to estimate the treatment effect in a two-sample experiment. This is where we introduce the lemma and theorem that is used to construct a treatment effect that is much

more accurate and efficient than the standard approach. We then introduce the permutation model in a two-sample experiment. We show that the permutation model is different than the randomization model in theory but the mechanics to make inferences in the two models are the same. Thus, we could use the same method in the randomzation model to make inferences in the population model. Next, we discuss the method of Monte-Carlo sampling, which is a fast and highly accurate way of generating an approximation of the reference distribution and obtaining estimates of the results in previous sections. We then show that Monte-Carlo is very efficient to use when the data are complicated such as in the case of multivaritate data. This is where we show our self-written functions are more versatile than standard packages because it can compute the exact reference distribution whereas other packages can't. Finally, we conclude the paper by expanding the method to a one-way layout or k-sample experiment.

2 Two-Sample Experiments

A two-sample experiment is one in which there are two treatments or samples. Depending on how the data was obtained, the model under a two-sample experiment is either a randomization model or a permutation model. We discuss the randomization model first in the following subsection.

2.1 Randomization Model

A randomization model is a model in which available subjects are randomly assigned to different treatments. For example, consider an experiment in which a researcher wants to find out if a new treatment improves the recovery time (in days) of a postsurgical patient compared to the standard treatment. Since the new treatment has not been carried out, we do not have a population of patients who were assigned the new treatment to draw a sample from. Thus, instead, we choose some available subjects and randomly assign each subject to either the new or the standard treatment. This is a randomization model. The benefit of this model is that it is easy to obtain a sample and randomly assign each subject to the treatments. The drawback is that any inferences are limited to subjects in the study only.

Suppose there are N subjects available in the study, of which n are randomly assigned to receive the new treatment while the remaining m receive the standard treatment. Below is the recovery time data (in days) of 7 patients, of which 4 were assigned to receive the new treatment and 3 received the standard treatment.

Table 1: *Recovery time (in days) of 7 patients*

New treatment	Standard treatment
19, 22, 25 26	23, 33, 40

We want to test if the new treatment decreases the recovery time, thus our hypotheses are

H_0: There is no difference between the treatments
H_1: The new treatment decreases receovery time

Table 2: *All possible permutations of $N = 7$ recovery times to two treatment groups of size 4 and 3 along with the randomization reference distributions of 4 chosen statistics: Diference In Means, Sum of New, Difference in Medians, Wilcoxin Rank Sum*

	Randomization									
	New				Standard			Difference	Sum Of	Difference	Wilcoxin
No.	Treatment				Treatment			In Means	New	In Medians	Rank Sum
1	19	22	25	26	23	33	40	-9.00*	92	-9.5	12
2	19	22	25	23	26	33	40	-10.75	89	-10.5	10
3	19	22	25	33	26	23	40	-4.92	99	-2.5	13
4	19	22	25	40	26	23	33	-0.83	106	-2.5	14
5	19	22	26	23	33	40	25	-10.17	90	-10.5	11
6	19	22	26	33	23	40	25	-4.33	100	-1.0	14
7	19	22	26	40	23	33	25	-0.25	107	-1.0	15
8	19	22	23	33	40	25	26	-6.08	97	-3.5	12
9	19	22	23	40	33	25	26	-2.00	104	-3.5	13
10	19	22	33	40	25	26	23	3.83	114	2.5	16
11	19	25	26	23	33	40	22	-8.42	93	-9.0	13
12	19	25	26	33	23	40	22	-2.58	103	2.5	16
13	19	25	26	40	23	33	22	1.50	110	2.5	17
14	19	25	23	33	40	26	22	-4.33	100	-2.0	14
15	19	25	23	40	33	26	22	-0.25	107	-2.0	15
16	19	25	33	40	26	23	22	5.58	117	6.0	18
17	19	26	23	33	40	22	25	-3.75	101	-0.5	15
18	19	26	23	40	33	22	25	0.33	108	-0.5	16
19	19	26	33	40	23	22	25	6.17	118	6.5	19
20	19	23	33	40	22	25	26	4.42	115	3.0	17
21	22	25	26	23	33	40	19	-6.67	96	-9.0	14
22	22	25	26	33	23	40	19	-0.83	106	2.5	17
23	22	25	26	40	23	33	19	3.25	113	2.5	18
24	22	25	23	33	40	26	19	-2.58	103	-2.0	15
25	22	25	23	40	33	26	19	1.50	110	-2.0	16
26	22	25	33	40	26	23	19	7.33	120	6.0	19
27	22	26	23	33	40	25	19	-2.00	104	-0.5	16
28	22	26	23	40	33	25	19	2.08	111	-0.5	17
29	22	26	33	40	23	25	19	7.92	121	6.5	20
30	22	23	33	40	25	26	19	6.17	118	3.0	18
31	25	26	23	33	40	19	22	-0.25	107	3.5	18
32	25	26	23	40	33	19	22	3.83	114	3.5	19
33	25	26	33	40	23	19	22	9.67	124	7.5	22
34	25	23	33	40	26	19	22	7.92	121	7.0	20
35	26	23	33	40	19	22	25	8.50	122	7.5	21

Let $Y_1, Y_2, \ldots, Y_n$ denote the response values in the new treatment and $X_1, X_2, \ldots, X_m$ denote the response values in the standard treatment. If the null hypothesis is true, then the recovery time of each subject will be the same regardless of the treatment he/she received. For example, if a subject's recovery time under the new treatment is 19 days, then his/her recovery time under the standard treatments is also 19 days. This implies that *under the null hypothesis, our sample is one out of* ${}_NC_n = {}_7C_4 = 35$ *equally likely randomizations of the data into the two treatments*, so if we choose $T = \bar{Y} - \bar{X}$ as the test statistic and compute its value for all of these equally likely randomizations, we obtain a reference distribution under the null hypothesis for T, which as we called earlier the randomization reference distribution of T. We then use this randomization reference distribution to compute the p-value for our test by computing the probability that T gets a value that is at least as extreme as the observed value t^* in the direction of the alternative, which is left-tailed in this case. Because of the equally likeliness of all the permutations, this probably is just the ratio of the number of statistic values t_i's $(i = 1, 2, \ldots, {}_N C_n)$ that is less than or equal to t^*. That is

$$\text{p-value} = P(T \leq t^* | H_0) = \frac{\sum_{i=1}^{{}_NC_n} I(t_i \leq t^*)}{{}_NC_n}.$$

Table 2 on page 4 shows all possible randomizations of the data into the two treatments along with the reference distributions of four chosen statistics: Difference In Means, Sum of New (sum of response values in the new treatment), Difference In Medians, Wilcoxin Rank Sum (sum of ranks of the response values in the new treatment). The table was created using a combination of self-written R functions and Latex. Specifically, the function `mcomb1` generates all the randomizations and the function `pval` generates all the reference distributions. Details are below.

```
> RT <- c(19,22,25,26,23,33,40) # Recovery Times data
> pmatrix <- mcomb1(RT,4) # The matrix that contains all possible permutions
> rt <- pval(RT,4,"l","a",vect = TRUE)
> data.frame(pmatrix,rt$mean.diff.vec,rt$sum.vec,rt$median.diff.vec,rt$wilcox.vec)
```

From Table 2, we see that there are three values of the T (the underlined ones) that are less than or equal to our observed value $t^* = -9$, thus the p-value for the left-tailed test based on the randomization reference distribution of the difference in means statistic is $3/35$, which is approximately 0.0857. This p-value implies that if we choose a level of significance $\alpha = 0.05$, then the test is not significant and we conclude that the new treatment does not decrease the recovery time significantly. However, if we are more flexible and choose a level of significance $\alpha = 0.10$, that is the chance of getting a Type I error to be 10%, then the test is significant and we conclude that the new treatment does decrease the recovery time significantly.

Note that as the sample size increases and the number of randomizations increases, it becomes a "tedious" and impossible task to write down all possible randomizations of the data, which are used to generate a reference distribution of a statistic, and to compute the p-value based on that. Fortunately, the R function `pval` that we introduced earlier computes this p-value as well as the p-values for the left-tailed test based on the reference distributions of other test statistics that we mentioned earlier. Details are below.

```
> pval(RT,4,"l","m") # p-value based on the difference in mean statistic
```

```
[1] 0.08571429
> pval(RT,4,"l","s") # p-value based on the sum of new [treatment] statistic
[1] 0.08571429
> pval(RT,4,"l","d") # p-value based on the difference in medians statistic
[1] 0.08571429
> pval(RT,4,"l","w") # p-value based on the wilcoxin rank sum statistic
[1] 0.1142857
```

The results show that the p-values for the test using different statistics are very close to each other, which means we could use any of these statistics to perform our analysis. Which one is better than the other and how are they compared to their parametric counterparts is the topic of another debate. See page 64 in Higgins (2004) for more details on this topic.

The `pval` function also has the capability to generate the graph of the reference distribution of a chosen statistic. Simply set the plot argument to TRUE and the type argument to the type of plot ("l" for a line plot and "h" for a histogram) you want and it will display the graph for you. Details are below. Figure 1 displays the reference distribution of the difference in mean statistic and Figure 2 displays the reference distributions of all the statistics used earlier.

```
> pval(RT,4,"l","m",plot = TRUE, type = "l") #Plot of the ramdomization reference
distribution of the difference in means
> pval(RT,4,"l","a",plot = TRUE, type = "l") #Plot of the ramdomization reference
distribution of the all four statistics
```

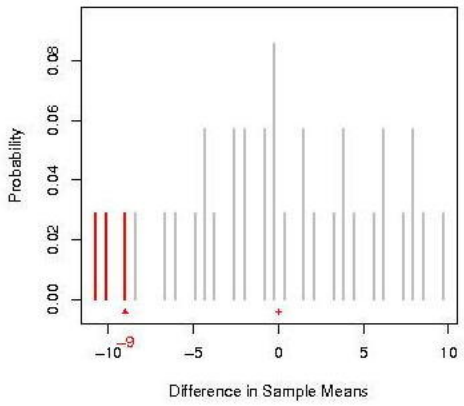

Figure 1: *The randomization reference distribution of the difference in means statistic from Table 2.*

A quick look at two graphs suggest that the randomization reference distributions of the difference in means statistic and that of the first treatment sum appear to be the same. This is not a coincidence because there is an increasing monotone relationship between the difference in means statistic and the first treatment sum statistic. That is $\bar{Y} - \bar{X} = \frac{m+n}{mn} \sum_{j=1}^{n} Y_j - \frac{1}{m} \left(\sum_{j=1}^{n} Y_j + \sum_{i=1}^{m} X_i \right)$.This is an important observation because "the sum of the responses from one treatment group is often used rather than the t statistic or the difference in means since it is computationally more efficient" (Ernst.)

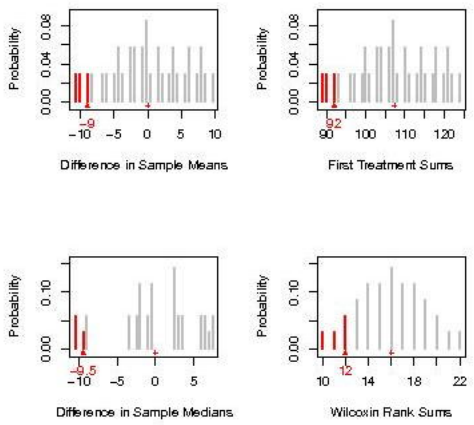

Figure 2: *The randomization reference distribution of the all four statistics from Table 2.*

2.2 Estimation Of The Treatment Effect

In a two-sample experiment, if we assume that the first treatment (the new treatment in this case) has a constant additive treatment effect Δ above the second treatment (the standard treatment in this case), then the response values in the new treatment, after subtracted Δ, will have similar magnitude to those in the standard treatment. This means the test for no treatment effect based on new treatment's shifted response values and the standard treatment's original response values will not be significant. We use this to construct a confidence interval for the treatment effect for the randomization model. We first discuss the standard procedure used to construct the confidence interval. Then we discuss another method that is independently derived to obtain the confidence interval faster and more accurate. We also introduce and prove the lemma and thoerem that is the basis for this new method.

Again, let $Y_1, Y_2, \ldots, Y_n$ denote the response values in the new treatment, $X_1, X_2, \ldots, X_m$ denote the response values in the standard treatment. If Δ is the constant additive effect of the new treatment above the standard treatment, then the new set of response values in the new treatment, $Y_{1\Delta} = Y_1 - \Delta, Y_{2\Delta} = Y_2 - \Delta, \ldots, Y_{n\Delta} = Y_n - \Delta$, and the original set of response values in the standard treatment, $X_1, X_2, \ldots, X_m$ will be similar in magitude, and the test for H_0 no treatment effect based on these values will be not significant. Let T_Δ be the shifted difference in mean statistic, that is $T_\Delta = \bar{Y}_\Delta - \bar{X}$, and let t^*_Δ be its observed value. Let $p_1(\Delta)$ and $p_2(\Delta)$ be the left-tailed and right-tailed areas of the randomization distribution of T_Δ, respectively. Then $p_1(\Delta) = P(T_\Delta \leq t^*_\Delta)$ and $p_2(\Delta) = P(T_\Delta \geq t^*_\Delta)$. To find a $1-\alpha$ left-tailed confidence interval for Δ, we invert the left-tailed test, which means finding the set of all possible values of Δ such as $p_1(\Delta) > \alpha$. We show that this set is equivalent to $(-\infty, \Delta_U)$ where $\Delta_U = \max\{\Delta | p_1(\Delta) > \alpha\}$. Using the same approach, we obtain a $1-\alpha$ right-tailed confidence interval for Δ to be $(\Delta_L, \infty,)$ where $\Delta_L = \min\{\Delta | p_2(\Delta) > \alpha\}$. As a result, a $1-2\alpha$ two-tailed confidence interval for Δ is (Δ_L, Δ_U).

The standard procedure to compute those end points Δ_U and Δ_L "involves a tedious and laborious search that requires the recomputation of the randomization distribution of T_Δ for each value of Δ that is tried," wrote Ernst. Ernst also wrote that "Garthwaite (1996) described an efficient method

for constructing confidence intervals from randomization tests, but this method is not implemented in any commercial software" and most importantly, as we read the paper, we noticed that this efficient method is still based on the recomputation of the randomization distribution of T_Δ. Either way, Ernst used an R program borrowed from Cliff Luneborge to construct a $(1-2/35)100\% \approx 94.29\%$ confidence interval for Δ, which results in $(-\infty, 2]$.

Originally, we attempted to replicate this result by writing an R function called `cint`. We succeeded in replicating the result. We were also able to obtain a graph of $p_1(\Delta)$ versus Δ, as displayed in Figure 3. Details are below.

```
> cint(RT,4,"l","m",1-2/35,plot = TRUE)
$LB
[1] "-inf"
$UB
[1] 2
```

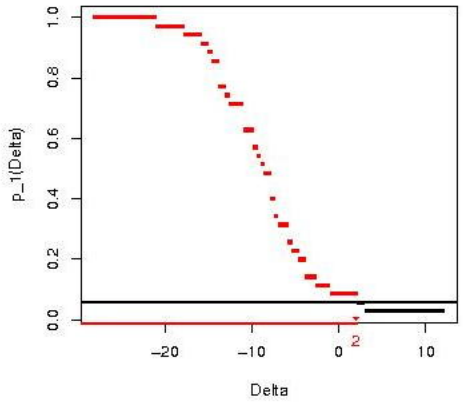

Figure 3: *The 94.29% left-tailed confidence interval for the effect of the new treatment along with graph of* $p_1(\Delta)$ *as a descreasing step function of* Δ.

The graph of $p_1(\Delta)$ versus Δ indicates that it is a decreasing step function, which prompted us to investigate this behavior further. After further investigation, we discovered the exact behavior of $p_1(\Delta)$ as a function of Δ, which we write as a theorem. In the process, we develope and state a lemma which we use to prove the theorem. Details are in the following paragraph. We show that the theorem not only provides a proof for the standard recomputation approach to obtain a confidence interval for the treatment effect Δ of a two-sample experiment as shown above; it also provides a basis for another method that does not require the recomputation of the reference distribution of T_Δ for each Δ that is tried.

Again, let denote $Y_1, Y_2, \ldots, Y_n$ the response values in the new treatment group before subtracting the treatment effect Δ, and $X_1, X_2, \ldots, X_m$ denote the response values in the standard treatment group. We assume that $n \geq m$ without loss of generosity. A permutation of the data into the two treatment

groups is just a switch of k response values in the standard treatment group with k response values in the new treatment group where $k = 0, 1, \ldots, m$. For a fixed k value, there are ${}_nC_k \times {}_mC_k$ ways of doing the switching. Consider the first of these switches, that is switching the first k response values in the standard treatment group with the first k response values in the new treatment effect, then this randomization results in a sample that contains $X_1, \ldots, X_k, Y_{k+1}, \ldots, Y_n$ in the new treatment group and $Y_1, \ldots, Y_k, X_{k+1}, \ldots, X_m$ in the standard treatment group. Let $t_{k,1}$ be the value of the difference in means statistic in this permutation, where k denotes the number of response values that were switched and 1 denotes the first of these switches, then

$$t_{k,1} = \frac{\sum_{i=1}^{k} X_i + \sum_{j=k+1}^{n} Y_j}{n} - \frac{\sum_{j=1}^{k} Y_j + \sum_{i=k+1}^{m} X_i}{m}.$$

After subtracting the treatment effect Δ from the response values in the new treatment group, we have $Y_{1\Delta} = Y_1 - \Delta, Y_{2\Delta} = Y_2 - \Delta, \ldots, Y_{n\Delta} = Y_n - \Delta$, as response values in the new treatment group and the response values in the standard treatment group are still the same, which are $X_1, X_2, \ldots, X_m$. Now the permutations are done on these shifted, $Y_{1\Delta}, Y_{2\Delta}, \ldots, Y_{n\Delta}$, and original, $X_1, X_2, \ldots, X_m$, response values. Again, a permutation is just a switch of k response values in the standard treatment group with k response values in the new treatment group. Consider the same switch as the one done above, that is switching the first k response values in the standard treatment groups with the first k respone values in the standard treatment effect, then this permutation resuls in a sample that contains $X_1, \ldots, X_k, Y_{k+1,\Delta}, \ldots, Y_{n,\Delta}$ in the new treatment group and $Y_{1,\Delta}, \ldots, Y_{k,\Delta}, X_{k+1}, \ldots, X_m$ in the standard treatment group. Let $t_{k,1,\Delta}$ be the value of the difference in means statistic in this permutation, where k denotes the number of response values that were switched, 1 denotes the first of these switches, and Δ is the shifted amount of the response values in the new treatment group, then

$$\begin{aligned}
t_{k,1,\Delta} &= \frac{\sum_{i=1}^{k} X_i + \sum_{j=k+1}^{n} Y_{j,\Delta}}{n} - \frac{\sum_{j=1}^{k} Y_{j,\Delta} + \sum_{i=k+1}^{m} X_i}{m} \\
&= \frac{\sum_{i=1}^{k} X_i + \sum_{j=k+1}^{n} (Y_j - \Delta)}{n} - \frac{\sum_{j=1}^{k} (Y_j - \Delta) + \sum_{i=k+1}^{m} X_i}{m} \\
&= \frac{\sum_{i=1}^{k} X_i + \sum_{j=k+1}^{n} Y_j - (n-k)\Delta}{n} - \frac{\sum_{j=1}^{k} Y_j - k\Delta + \sum_{i=k+1}^{m} X_i}{m} \\
&= \frac{\sum_{i=1}^{k} X_i + \sum_{j=k+1}^{n} Y_j}{n} - \frac{\sum_{j=1}^{k} Y_j + \sum_{i=k+1}^{m} X_i}{m} - \frac{(n-k)\Delta}{n} + \frac{k\Delta}{m} \\
&= t_{k,1} - \frac{(n-k)\Delta}{n} + \frac{k\Delta}{m} \\
&= t_{k,1} - \Delta + k\left(\frac{1}{n} + \frac{1}{m}\right)\Delta.
\end{aligned}$$

Note that, from the above derivation, similar relationship also holds for $t_{k,d,\Delta}$ and $t_{k,d}$, where k is the number of response values that were switched, d is the d^{th} of these switches (recall that for a fixed k, there are ${}_nC_k \times {}_mC_k$ ways of doing the switching, so $d = 1, 2, \ldots, {}_nC_k \times {}_mC_k$), and Δ is the shifted amount of response values in the new treatment group, which is the treatment effect of interest. Thus we have proven the following lemma:

Lemma 1. Let $t_{k,d}$ be the value of the [difference in means] statistic in one of the randomizations of the original data obtained by switching k response values between the two treatment groups and d

is the d^{th} of these switches. Let $t_{k,d,\Delta}$ be the value of the statistic in the same randomization of the shifted data where Δ is the shifted amount of the respone values in the new treatment group, then

$$t_{k,d,\Delta} = t_{k,d} - \Delta + k\left(\frac{1}{n} + \frac{1}{m}\right)\Delta.$$

We now state and prove the theorem that describes the exact behavior of $p_1(\Delta)$.

Theorem 1. Let q be the number of distinct values of $w_{k,d} = \frac{t_0 - t_{k,d}}{k\left(\frac{1}{n}+\frac{1}{m}\right)}$ for $\forall k \neq 0, d$. Let $w_{(i)}$ be the i^{th} distinct ordered value of $w_{k,d}$ where $i = 1, 2, \ldots, q$ and p_i be the number of ties at $w_{(i)}$. Also, let $w_{(0)} = -\infty$ and $w_{(q+1)} = \infty$. Then, for $w_{(i-1)} < \Delta \leq w_{(i)}$ where $i = 1, 2, \ldots, q+1$,

$$p_1(\Delta) = \frac{\sum_{j=i}^{q} p_j + 1}{{}_N C_n}.$$

Proof. We have

$$\begin{aligned}
p_1(\Delta) &= P(T_\Delta \leq t^*_\Delta) \\
&= \frac{\sum_{k,d} I(t_{k,d,\Delta} \leq t_{0,\Delta})}{{}_N C_n} \\
&= \frac{\sum_{k,d} I(t_{k,d} - \Delta + k\left(\frac{1}{n}+\frac{1}{m}\right)\Delta \leq t_0 - \Delta)}{{}_N C_n} && \text{(using equation in Lemma 1)} \\
&= \frac{\sum_{k,d} I(k\left(\frac{1}{n}+\frac{1}{m}\right)\Delta \leq t_0 - t_{k,d})}{{}_N C_n} \\
&= \frac{\sum_{k\neq 0,d} I(k\left(\frac{1}{n}+\frac{1}{m}\right)\Delta \leq t_0 - t_{k,d})}{{}_N C_n} + \frac{1}{{}_N C_n} \\
&= \frac{\sum_{k\neq 0,d} I\left(\Delta \leq \frac{t_0 - t_{k,d}}{k\left(\frac{1}{n}+\frac{1}{m}\right)}\right)}{{}_N C_n} + \frac{1}{{}_N C_n} \\
&= \frac{\sum_{k\neq 0,d} I(\Delta \leq w_{k,d})}{{}_N C_n} + \frac{1}{{}_N C_n} \\
&= \frac{\sum_{i=1}^{q} I(\Delta \leq w_{(i)})}{{}_N C_n} + \frac{1}{{}_N C_n} \\
&= \frac{p_i + p_{i+1} + \ldots + p_q}{{}_N C_n} + \frac{1}{{}_N C_n} && \text{(for } w_{(i-1)} < \Delta \leq w_{(i)}\text{)} \\
&= \frac{\sum_{j=i}^{q} p_j + 1}{{}_N C_n}.
\end{aligned}$$

□

The theorem shows that for $\Delta \in (w_{(i-1)}, w_{(i)}]$, $p_1(\Delta)$ is the proportion of the number of $w_{k,d}$ that are greater than or equal to $w_{(i)}$ plus 1 to the total number of randomizations. Thus, we have the following Corallary:

Corollary 1. $p_1(\Delta)$ is a left-continuous decreasing step function of Δ with a maximum value of 1, minimum value of $\frac{1}{{}_N C_n}$, and the step down points are the distinct ordered values of $w_{k,d}$.

Corallary 1 justifies that if $\Delta_U = \max\{\Delta | p_1(\Delta) > \alpha\}$, then for $\forall \Delta \in (-\infty, \Delta_U]$, $p_1(\Delta) > \alpha$. We now state and prove the following Corollary

Corollary 2. $(-\infty, \Delta_U]$ is a $1-\alpha$ left-tailed confidence interval for the treatment effect Δ.

Proof. We have $P(\Delta \in (-\infty, \Delta_U]) = P(p_1(\Delta) > \alpha)$. But $p_1(\Delta) = P(T_\Delta \leq t^*_\Delta)$ is a discrete cdf which is stochastically greater than a Uniform(0,1) (C-B, Exercise 2.10, page 77). As a result, $P(p_1(\Delta) > \alpha) \geq 1-\alpha$. Thus, $(-\infty, \Delta_U]$ is a $1-\alpha$ left-tailed confidence interval for the treatment effect Δ. □

Corollary 2 proves that $(-\infty, \Delta_U]$ is a $1-\alpha$ left-tailed confidence interval for the treatment effect Δ, which Ernst stated but did not prove.

On a more important note, the theorem not only provides a proof for the standard recomputation approach to obtain a confidence interval for the treatment effect Δ of a two-sample experiment; it also provides a basis for another method that does not require the recomputations of the reference distribution for each Δ that is tried. Let us clarify the statement by looking at the result of Theorem 1 again. The theorem shows that for $\Delta \in (w_{(i-1)}, w_{(i)}]$, $p_1(\Delta)$ is the proportion of the numbers of $w_{k,d}$ that are greater than or equal to $w_{(i)}$ plus 1 to the total number of randomizations. As a result, this proportion decreases as Δ crosses through each $w_{(i)}$. We want this proportion to be greater than α, so we find the maximum $w_{(i)}$ such that the number of $w_{k,d}$ that are greater than or equal to $w_{(i)}$ is greater than $\alpha \times_N C_n - 1$. To do that, we order $w_{k,d}$ and start counting the number of $w_{k,d}$ from the largest down until this number greater than $\alpha \times_N C_n - 1$, then the value of $w_{k,d}$ where the count stops is Δ_U.

We apply this method in the recovery time example. Below is Table 3 which is Table 2 excluding the three columns for the three statistics: Sum of First, Difference In Medians, Wilcoxin Rank Sum, but including three extra columns, one for the values of k, one for the values of $w_{k,d}$ and one for the ordered values of $w_{k,d}$. For $\alpha = 2/35, \alpha \times_N C_n - 1 = 2/35 \times 35 - 1 = 1$, so we count from the largest $w_{k,d}$ down and stop at the 2^{nd} value. From Table 3, the ordered value of $w_{k,d}$ where the count stops is 2, thus a $1 - 2/35$ left-tailed confidence interval for Δ is $(-\infty, 2]$ This result matches the one resulted from the standard approach, but it only requires the computation of the reference distribution once, which is a lot faster than the standard approach! On a side note, a program to implement this new method automatically is in the process of being developed and will be available soon.

2.3 Permutation Model

A permutation model is a model in which subjects are randomly drawn from different populations. This is different from a randomization model where subjects are fixed and randomly assigned to different treatments. The benefit of this model is that inferences can be generalized to the populations and not limited to the subjects in the study only. The drawback is that truly random samples sometimes are not easy to obtain. However, despite the difference, we show below that the mechanics to perform tests or construct confidence intervals in a permutation model are the same as in a randomization model.

Let's start with the study quoted in Ernst' article. Two biologists discovered two spieces of small flies, call them S1 and S2 for simplicity purpose, that belong to an insect family termed "biting midges" discovered in the jungles of Central and South Ameria. In an attempt to distinguish between the two species, the biologists obtained a random sample of 9 flies from S1 and 6 flies from S2 and measured their wing lengths (WL) and antenna length (AL) respectively. The data are shown in Table 4. We

Table 3: *Update of Table 2 that excludes the three columns for the three statistics: Sum of First, Difference In Medians, Wilcoxin Rank Sum, but includes three extra columns, one for the values of k, one for the values of $w_{k,d}$ and one for the ordered values of $w_{k,d}$.*

Randomization											
	New				Standard			Difference			
No.	Treatment				Treatment			In Means	k	$w_{k,d}$	$w_{(i)}$
1	19	22	25	26	23	33	40	-9.00*	0		
2	19	22	25	23	26	33	40	-10.75	1	3.0	-21.0
3	19	22	25	33	26	23	40	-4.92	1	-7.0	-18.0
4	19	22	25	40	26	23	33	-0.83	1	-14.0	-16.0
5	19	22	26	23	33	40	25	-10.17	1	2	-15.0
6	19	22	26	33	23	40	25	-4.33	1	-8.0	-14.5
7	19	22	26	40	23	33	25	-0.25	1	-15.0	-14.0
8	19	22	23	33	40	25	26	-6.08	2	-2.5	-14.0
9	19	22	23	40	33	25	26	-2.00	2	-6.0	-14.0
10	19	22	33	40	25	26	23	3.83	2	-11.0	-13.0
11	19	25	26	23	33	40	22	-8.42	1	-1.0	-12.5
12	19	25	26	33	23	40	22	-2.58	1	-11.0	-11.0
13	19	25	26	40	23	33	22	1.50	1	-18.0	-11.0
14	19	25	23	33	40	26	22	-4.33	2	-4.0	-11.0
15	19	25	23	40	33	26	22	-0.25	2	-7.5	-10.0
16	19	25	33	40	26	23	22	5.58	2	-12.5	-9.67
17	19	26	23	33	40	22	25	-3.75	2	-4.5	-9.5
18	19	26	23	40	33	22	25	0.33	2	-8.0	-9.0
19	19	26	33	40	23	22	25	6.17	2	-13.0	-8.67
20	19	23	33	40	22	25	26	4.42	3	-7.67	-8.0
21	22	25	26	23	33	40	19	-6.67	1	-4.0	-8.0
22	22	25	26	33	23	40	19	-0.83	1	-14.0	-7.67
23	22	25	26	40	23	33	19	3.25	1	-21.0	-7.5
24	22	25	23	33	40	26	19	-2.58	2	-5.5	-7.5
25	22	25	23	40	33	26	19	1.50	2	-9.0	-7.0
26	22	25	33	40	26	23	19	7.33	2	-14.0	-6.0
27	22	26	23	33	40	25	19	-2.00	2	-6.0	-6.0
28	22	26	23	40	33	25	19	2.08	2	-9.5	-5.5
29	22	26	33	40	23	25	19	7.92	2	-14.5	-4.5
30	22	23	33	40	25	26	19	6.17	3	-8.67	-4.0
31	25	26	23	33	40	19	22	-0.25	2	-7.5	-4.0
32	25	26	23	40	33	19	22	3.83	2	-11	-2.5
33	25	26	33	40	23	19	22	9.67	2	-16	-1.0
34	25	23	33	40	26	19	22	7.92	3	-9.67	2.0
35	26	23	33	40	19	22	25	8.50	3	-10.00	3.0

want to test if the two populations of flies based on the wing lengths are the same. Thus our hypotheses are:

Table 4: *The wing lengths (WL) and antennae lengths (AL) of 15 flies in two samples from population 1 (S1) and population 2 (S2).*

S1		S2	
WL	AL	WL	AL
1.72	1.24	1.78	1.14
1.64	1.38	1.86	1.20
1.74	1.36	1.96	1.30
1.70	1.40	2.00	1.26
1.82	1.38	2.00	1.28
1.82	1.48	1.96	1.18
1.90	1.38		
1.82	1.54		
2.08	1.56		

H_0: The two populations of flies based on the wing lengths are the same
H_1: The two populations of flies based on the wing lengths are different

Let $Y_1, Y_2, \ldots, Y_n$ denote the wing lengths of the sample flies in S1 and $X_1, X_2, \ldots, X_m$ denote the wing lengths of the sample flies in S2 where $n = 9$, $m = 6$, and $N = 15$ total number of flies in this example. Because the samples of flies are random, $Y_1, Y_2, \ldots, Y_n$ and $X_1, X_2, \ldots, X_m$ are random. Given the set of observed wing lengths, there are ${}_NC_n = {}_{15}C_6 = 5005$ divisions of these observed values into the two samples. If the null hypothesis is true, then the populations of flies based on the wing lengths are the same. This implies $Y_1, Y_2, \ldots, Y_n$ and $X_1, X_2, \ldots, X_m$ come from the same population. Thus, *under the null hypothesis, conditioned on the observed values of the wing lengths, the divisions of these observed values into the two samples are equaly likely.* If we choose $T = \bar{Y} - \bar{X}$ as the test statistic and compute its value for all of these equally likely permutations of the observed values, we obtain a reference distribution under the null hypothesis for T, which as we called earlier the permutation reference distribution of T. Note that this permutation reference distribution is obtained based on equally likely assignments, much like the way a randomization reference distribution is obtained. As a result, we can use the same mechanics that we used to perform a randomization test or to construct a construct a confidence interval for the treatment effect Δ of the two treatments to perform a permutation test or to construct a confidence interval for the shift effect between the two populations.

More specifically, in this example, because we have a two-tailed test, our p-value is the probably that T gets a value that is at least as extreme as the sample value t^* in both directions, and thus if $\bar{t}$ is the mean of the reference distribution

$$p - value = P(|T - \bar{t}| \geq |t^* - \bar{t}| \mid H_0).$$

Using `pval`, we obtain the p-values for our permutation test based on the wing lengths and antenna

length. We also turn on the plot parameter to obtain the graphs for the permutation reference distribution of the wing length and antennae length, as displayed in Figure 4 and Figure 5.

```
WL <- c(1.72,1.64,1.74,1.70,1.82,1.82,1.90,1.82,2.08,1.78,1.86,1.96,2.00,2.00,1.96)
AL <- c(1.24,1.38,1.36,1.40,1.38,1.48,1.38,1.54,1.56,1.14,1.20,1.30,1.26,1.28,1.18)
> pval(WL,9,"b","s",plot = TRUE,type = "l") # p-values using wing lengths
[1] 0.07172827
> pval(AL,9,"b","s",plot = TRUE,type = "l") # p-values using antennae lengths
[1] 0.002197802
```

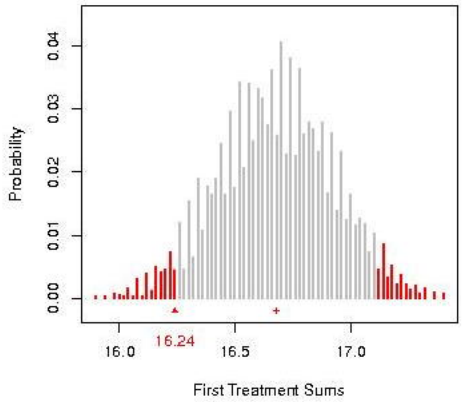

Figure 4: *The permutation reference distribution of the sum of wing lengths of S1 species where + marks* $\bar{t}$ *and 16.24* $= t^*$

The p-values suggest that the two species differ significantly in antennae length but not in wing length. We also obtain the confidence interval for the shift effect between the two populations in the same manner that we use to obtain the confidence interval for the effect between the two treatments in previous subsection. Specifically, we'll use the function `cint` again to obtain a 94.90% confidence interval for the shift effect using the wing lengths and a 94.94% confidence interval for the shift effect using the antennae lengths along with their graphs, as displayed in Figure 6 and Figure 7.

We note here that `cint` computes the upper and lower limits, Δ_U and Δ_L, based on $p_1(\Delta)$ and $p_2(\Delta)$ as defined on page 7. However, in Figure 6 and Figure 7, `cint` plots $p(\Delta)$ defined as

$$p(\Delta) = P(|T_\Delta - \bar{t}_\Delta| \geq |t^*_\Delta - \bar{t}_\Delta|).$$

We observe that the interval $[\Delta_L, \Delta_U] \subset \{\Delta | p(\Delta) > \alpha\}$, which provides us with another confidence interval for Δ. Although this interval is wider and hence less efficient, it is a companion to the two-sided test.

```
>cint(WL,9,"b","m",.9490,plot = TRUE)
$LB
```

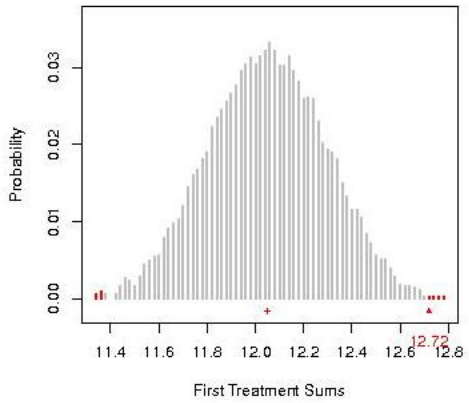

Figure 5: *The permutation reference distribution of the sum of antennae lengths of S1 species where + marks* $\bar{t}$ *and 12.72 =* t^*

```
[1] -0.246
$UB
[2] 0.01
>cint(AL,9,"b","m",.9490,plot = TRUE)
$LB
[1] 0.087
$UB
[2] 0.285
```

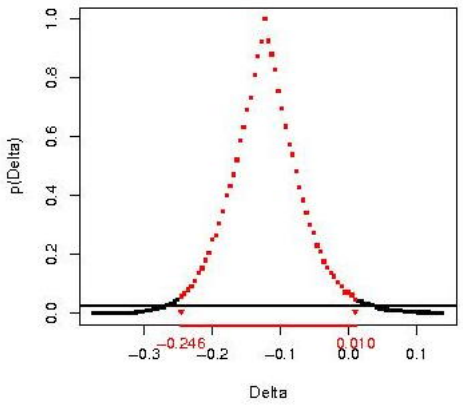

Figure 6: *The 94.90% two-tailed confidence interval for the shift effect of the two populations using wing lengths is given by (-0.246,0.010).*

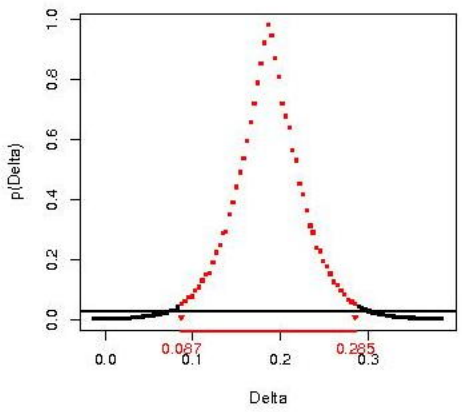

Figure 7: *The 94.90% two-tailed confidence interval for the shift effect of the two populations using antennae lengths is given by (0.087,0.285).*

The two confidence intervals reported by `cint` are very closed within ones in Ernst's article. He reported that the 94.90% confidence interval for the mean difference in wing length is (-0.250, 0.010) and a 94.94% confidence interval for the mean difference in anennae length is (0.087, 0.286). The conclusions derived from these intervals also match the results of the tests.

2.4 Monte-Carlo Sampling

From previous subsections, it was shown that both randomization and permutation reference distribution are caculated based on ${}_NC_n$ unique permutations of the data into two samples of size n and m. This poses two computational problems: (1) As the sample size or the number of samples increases, the number of permutations becomes very large. For example, there are ${}_{30}C_{15} \approx 155$ million ways of permuting 30 data points into two groups 15 each, and there are ${}_{30}C_{10,10} \approx 5.5$ trillion ways of permuting 30 data points into three groups 10 each; and (2) The permutation has to be unique, which is not easily done as the number of permutations increase. Both of these problems can be handled by handled by using fast computers and efficient algorithms, but even when that's the case, as the number of permutations increases, the amount of time it takes to process all the permutations or to repeat this process many times, e.g. to compute the limit of the confidence interval for the treatment effect in a previous subsection, increases signficantly. This problem, however, can be solved with the method of Monte-Carlo sampling.

The method of Monte-Carlo sampling is not new. It has been used in many other fields. However, the ideas are basically the same. First, we define a domain of all possible inputs of interest. Then we randomly generate a set of inputs from the domain using some specified probability distribution. And finally, we use this set of randomly generated inputs to study the domain of interest. In this paper, we are interested in studying the randomization or permutation reference distribution of a chosen statistics, which is the domain of interest. We can randomly generate a set of values from this distribution simply by repeated (let say M times) and randomly divide our sample of size N into two samples of size n and m and recalcuating the value of our chosen statistic for each division. Note that

this procedure is not the same as the procedure we use to obtain our randomization or permutation reference distribution in two ways: (1) The divisions are not necessarily unique; and (2) The number of divisions is pre-specified. Because of these differences, it's a lot easier to obtain the M divisions computationally than obtaining the ${}_NC_n$ unique permutations. From the way it was obtained, each of the M divisions is obviously just one out of ${}_NC_n$ permutations of the data set into two groups of size n and m. Under the null hypothesis, the ${}_NC_n$ permutations are equally likely, and because of that, each of the M divisions is indeed a random permutation chosen from the ${}_NC_n$ permutations. Therefore, the set of M statistic values obtained from the M divisions is a random sample chosen from the randomization or permutation reference distribution!

2.4.1 P-value and Effect Estimation

By Monte-Carlo sampling, we easily obtain a random sample from the randomization or permutation reference distribution we study. We use this random sample, which is an approximation of the randomization and reference distribution, to study the distribution themself. For example, let's say we want to compute the p-value for left-tail randomization test in Subsection 2.1. We can either generate the randomization reference distribution of the difference in means statistic and compute the proportion of statistic values that are less than or equal to the observed value like we did in that section, or since the p-value is the proportion of the statistic values that are less than or equal to the observed value in the population, it is a population proportion, and we can estimate it by obtaining a random sample from the reference distribution of the difference in means statistic and calculate the sample proportion. To be consistent with the fact that the population proportion contains the observed value in its caculation, we'll include it in our sample as well. Thus if $t_i, i = 1, \ldots, M$, are the Monte-Carlo values, then the sample proportion, which is the estimated p-value for a left-tailed test, is

$$\hat{p} = \frac{1 + \sum_{i=1}^{M} I(t_i \leq t^*)}{M + 1}.$$

To computationally generate a random sample from the randomization reference distribution and calculate the estimated p-value based on it, we use the R function `pval` with the optional argument `monte` (TRUE to turn it on) and the accompaning argument `B` (to specify the number of Monte-Carlo samplings). We'll use M = 999 and thus `B` = 999. For convenience purpose, we include the code to compute a Monte-Carlo estimated p-value for the two tailed permutation test in subsection 2.3 using wing length and antannae length.

```
> pval(RT,4,"l","m",monte = TRUE,B = 999)
[1] 0.079
> pval(WL,9,"b","s",monte = TRUE,B = 999)
[1] 0.062
> pval(AL,9,"b","s",monte = TRUE,B = 999)
[1] 0.003
```

Compared to the exact p-values for these tests, which are 0.0857, 0.0717, and 0.002, we conclude that Monte-Carlo estimated p-values are very close to the actual p-values. Another benefit of Monte-Carlo sampling is it can also be used to estimate the confidence interval for the treatment effect and the shift effect nicely. Recall that these confidence intervals are computed by the recomputation of the

p-values for each value of Δ until we get to one that is just above α. This process takes time. Indeed, the calculation of the confidence intervals for the shift effect based on the wing length and antennae length took us 5 minutes each. The calculation times are shortened to 4 minutes when we used a faster computer, but still it takes time considering our sample in still relatively small. We use the same function `cint` with similar optional argument monte and B to obtain estimated values for these intervals for Monte-Carlo sampling.

```
> date()
[1] "10:24:49 2009"
> cint(RT,4,"l","m",1-2/35,monte = TRUE,B = 999)
$LB
[1] "-inf"
$UB
[1] 2.937
> date()
[1] "10:26:51 2009"
> cint(WL,9,"b","m",.9490,monte = TRUE,B = 999)
$LB
[1] -0.248
$UB
[1] 0.01
> date()
[1] "10:28:13 2009"
> cint(AL,9,"b","m",.9494,monte = TRUE,B = 999)
$LB
[1] 0.086
$UB
[1] 0.285
> date()
[1] "10:29:14 2009"
```

These estimated values of the confidence intervals, compared to the exact confidence intervals, which are $(-\infty, 2), (-0.246, 0.01), (0.087, 0.285)$, show that Monte-Carlo sampling yields very good estimates of the confidence intervals for the treatment effect and shift effect. In addition, it only takes about a minute to complete the calculation, compared with the 4, 5 minutes it takes to compute the exact confidence intervals. This is a reason we choose Monte-Carlo sampling whenever possible where small errors are allowed and time is constrained.

Another question arising from the estimation of the p-value is how to obtain a confidence interval for it. We show that the method of Monte-Carlo sampling can be used to compute a confidence interval for the true p-value. The details are developed in the following section.

2.4.2 Confidence Interval for the True P-Value

As explained earlier, the p-value for our test is the population proportion of the population values that are less than or equal to the observed value (in a left-tailed test). Thus a confidence interval for it can be computed using the Central Limit Theorem, which says for large sample sizes, a 1

- α confidence interval for the true population proportion in random sampling with replacement is $[\hat{p} - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\hat{q}}{n}}, \hat{p} + z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\hat{q}}{n}}$. This approach is familar with most readers already, so we won't go into details. We look at another another way to construct a confidence interval by pivoting the discrete cdf of a binomial distribution (C-B, page 434). We implement that method to construct the confidence interval for the true p-value. First, we state an exercise in C-B (Exercise 9.21, page 454) and prove it as a theorem, which is then used for the construction of the confidence interval.

Theorem 2 (Exercise 9.21)**.** If $X \sim binomial(n, p)$, then a $1 - \alpha$ confidence interval for p is

$$\frac{1}{1 + \frac{n-x+1}{x} F_{2(n-x+1),2x,\alpha/2}} \leq p \leq \frac{\frac{x+1}{n-x} F_{2(x+1),2(n-x),\alpha/2}}{1 + \frac{x+1}{n-x} F_{2(x+1),2(n-x),\alpha/2}}.$$

Proof. To prove this thoerem, we use the pivoting a discrete cdf theorem in C-B stated below. The proof for the theorem was outlined in their book, so it is omitted.

Theorem 9.2.14 (Pivoting a discrete cdf). Let X be a discrete statistic with cdf $F_X(x|\theta) = P(X \leq x|\theta)$. Let $\alpha_1 + \alpha_2 = \alpha$ with $0 < \alpha < 1$ be fixed values. Suppose that for each x $\in \mathcal{S}_X$, define $\theta_L(x)$ and $\theta_U(x)$ as follows.

i. If $F_X(x|\theta)$ is a decreasing function of θ for each x, define $\theta_L(x)$ and $\theta_U(x)$ by

$$\bar{F}(x|\theta_L(x)) = P(X \geq x|\theta_L(x)) = \alpha_2, \qquad F(x|\theta_U(x)) = P(X \leq x|\theta_U(x)) = \alpha_1.$$

i. If $F_X(x|\theta)$ is a increasing function of θ for each x, define $\theta_L(x)$ and $\theta_U(x)$ by

$$F(x|\theta_L(x)) = P(X \leq x|\theta_L(x)) = \alpha_2, \qquad \bar{F}(x|\theta_U(x)) = P(X \geq x|\theta_U(x)) = \alpha_1.$$

Then the random interval $[\theta_L(X), \theta_U(X)]$ is a 1 - α confidence interval for α.

We use the relationship between the binomial and the beta distribution as outlined in Exercise 2.40 (C-B, page 82) and the relationship between the beta and the F distributions as outlined in Theorem 5.3.8 (C-B, page 225). These are fundamental mathematical statistics relationships so we state them without proofs. We omit the parts that are not needed.

Exercise 2.40 (binomal-beta relationship). If $X \sim binomial(n, p)$ and $Y \sim beta(x, n - x + 1)$, then

$$P(X \geq x) = P(Y \leq p) \text{ for } x = 1, 2, \ldots, n.$$

Theorem 5.3.8 (beta-F relationship). We have

a. If $F \sim F_{p,q}$, then $1/F \sim F_{q,p}$.
b. If $F \sim F_{p,q}$, then $\frac{\frac{p}{q}F}{1+\frac{p}{q}F} \sim beta(\frac{p}{2}, \frac{q}{2})$.

Before we turn to the proof of Theorem 2, we state the following definition and lemmas. We provide the proofs for these lemmas for the readers' reference.

Definition 1 (Definition of MLR)**.** A family of pmfs or pdfs $\{f(x|\theta)\}$ for a univariate random variable X with a real-valued parameter θ has a *monotone likelihood ratio* (MLR) w.r.t θ if, for every $\theta_2 > \theta_1$, $\frac{f(x|\theta_2)}{f(x|\theta_1)}$ is monotone on $\{x : f(x|\theta_1) > 0 \text{ and } f(x|\theta_2) > 0\}$. If $\frac{f(x|\theta_2)}{f(x|\theta_1)}$ is monotone increasing, then the family has an increasing MLR. Otherwise, it has a decreasing MLR.

Lemma 1 (MLR property)**.** If a family of pmfs or pdfs $\{f(x|\theta)\}$ for a univariate random variable X with a real-valued parameter θ has an *increasing* MLR w.r.t θ, then its cmf or cdf $F(x|\theta)$ is a decreasing function of θ.

Proof. We prove the discrete case first and outline the proof for the continuos case.
Suppose $\{f(x|\theta)\}$ is a discrete family with $\mathcal{S}_X = \{x_1, x_2, \ldots\}$ where $x_i < x_j$ for $i < j$.
Suppose $\theta_2 > \theta_1$. We show $F(x_k|\theta_2) < F(x_k|\theta_1)$ for $\forall x_k \in \mathcal{S}_X$ except on the end point (if exists) where $F(x|\theta) = 1$.
Because $\{f(x|\theta)\}$ has an increasing MLR and $\theta_2 > \theta_1$, then $\frac{f(x_1|\theta_2)}{f(x_1|\theta_1)} < \frac{f(x_2|\theta_2)}{f(x_2|\theta_1)} < \ldots < \frac{f(x_k|\theta_2)}{f(x_k|\theta_1)} < \ldots$
Case 1: If $\frac{f(x_k|\theta_2)}{f(x_k|\theta_1)} < 1$, then $\frac{f(x_1|\theta_2)}{f(x_1|\theta_1)} < \frac{f(x_2|\theta_2)}{f(x_2|\theta_1)} < \ldots < \frac{f(x_k|\theta_2)}{f(x_k|\theta_1)} < 1$,
then $\frac{f(x_i|\theta_2)}{f(x_i|\theta_1)} < 1$ for $\forall i \leq k$, then $f(x_i|\theta_2) < f(x_i|\theta_1)$ for $\forall i \leq k$,
then $\sum_{i=1}^{k} f(x_i|\theta_2) < \sum_{i=1}^{k} f(x_i|\theta_1)$, then $F(x_k|\theta_2) < F(x_k|\theta_1)$.
Case 2: If $\frac{f(x_k|\theta_2)}{f(x_k|\theta_1)} \geq 1$, then $1 \leq \frac{f(x_k|\theta_2)}{f(x_k|\theta_1)} < \frac{f(x_{k+1}|\theta_2)}{f(x_{k+1}|\theta_1)} < \ldots$,
then $\frac{f(x_i|\theta_2)}{f(x_i|\theta_1)} > 1$ for $\forall i > k$, then $f(x_i|\theta_2) > f(x_i|\theta_1)$ for $\forall i > k$,
then $\sum_{i=k+1} f(x_i|\theta_2) > \sum_{i=k+1} f(x_i|\theta_1)$, then $1 - \sum_{i=1}^{k} f(x_i|\theta_2) > 1 - \sum_{i=1}^{k} f(x_i|\theta_1)$
then $\sum_{i=1}^{k} f(x_i|\theta_2) < \sum_{i=1}^{k} f(x_i|\theta_1)$, then $F(x_k|\theta_2) < F(x_k|\theta_1)$.

To prove for the continous case, use the same approach. Suppose $\theta_2 > \theta_1$, we need to show $F(x|\theta_2) < F(x|\theta_1)$. Look at the ratio of the pdfs at x and compares it with 1. Use integrals instead of sums. Details of the proof are left for the reader. □

Lemma 2 (MLR in regular exponential family)**.** If $f(x|\theta)$ belongs to a regular exponential family, i.e. $f(x|\theta) = h(x)c(\theta)e^{w(\theta)x}$ and $w(\theta)$ is an increasing function of θ, then $\{f(x|\theta)\}$ has an increasing MLR

Proof. Suppose $\theta_2 > \theta_1$, then $w(\theta_2) > w(\theta_1)$.
We need to show that $\frac{f(x|\theta_2)}{f(x|\theta_1)}$ is increasing in θ.
We have $\frac{f(x|\theta_2)}{f(x|\theta_1)} = \frac{h(x)c(\theta_2)e^{w(\theta_2)x}}{h(x)c(\theta_2)e^{w(\theta_1)x}} = \frac{c(\theta_2)}{c(\theta_1)}e^{(w(\theta_2)-w(\theta_1))x}$.
This is an increasing function of x because $c(\theta_2) > 0, c(\theta_1) > 0$, and $w(\theta_2) - w(\theta_1) > 0$.
Thus, $\{f(x|\theta)\}$ has an increasing MLR. □

Lemma 3 (binomial as regular exponential famly)**.** If $X \sim binomial(n, p)$, then $f(x|p)$ belongs to a regular exponential family; i.e., $f(x|p) = h(x)c(p)e^{w(p)x}$ and $w(p)$ is an increasing function of p.

Proof. This is easily shown and for formality purpose, we provide the proof in here.
As $X \sim binomial(n, p)$, then $f(x|p) =_n C_x p^x (1-p)^{n-x} =_n C_x e^{xlnp} e^{(n-x)ln(1-p)}$
$=_n C_x e^{nln(1-p)} e^{xln\frac{p}{1-p}} = h(x)c(p)e^{w(p)x}$ where $w(p) = ln\frac{p}{1-p}$ is an increasing function of p. □

Now that we've established all the known definition, theorems, exercises, and proved lemmas, we put them together to prove Theorem 2.

First, based on the three lemmas, we conclude that if $X \sim binomial(n, p)$ then $F(x|p)$ is a decreasing function of p. The according Theorem 9.2.14 in C-B, for $\alpha_1 = \alpha_2 = \frac{\alpha}{2}$ such that $\alpha_1 + \alpha_2 = \alpha$, we define $p_L(x)$ and $p_U(x)$ as follows such that

$$\bar{F}(x|p_L(x)) = P(X \geq x|p_L(x)) = \frac{\alpha}{2}, \text{and } F(x|p_U(x)) = P(X \leq x|p_U(x)) = \frac{\alpha}{2}.$$

For $p_L(x)$, we have

$$\begin{aligned}
\bar{F}(x|p_L(x)) &= P(X \geq x|p = p_L(x)) \\
&= P(Y \leq p_L(x)) \qquad (Y \sim beta(x, n-x+1). \text{ See binomial}-\text{beta relationship}) \\
&= P\left(\frac{1}{1+\frac{n-x+1}{x}F} \leq p_L(x)\right) \qquad (F \sim F(2(n-x+1), 2x). \text{ See beta}-\text{F relationship}) \\
&= P\left(F \geq \frac{1-p_L(x)}{\frac{n-x+1}{x}p_L(x)}\right) = \frac{\alpha}{2}.
\end{aligned}$$

This implies $\frac{1-p_L(x)}{\frac{n-x+1}{x}p_L(x)} = F_{2(n-x+1),2x,\alpha/2}$, then $p_L(x) = \frac{1}{1+\frac{n-x+1}{x}F_{2(n-x+1),2x,\alpha/2}}$.

For $p_U(x)$, we have

$$\begin{aligned}
F(x|p_U(x)) &= P(X \leq x|p = p_U(x)) \\
&= 1 - P(X \geq x+1|p = p_U(x)) \\
&= 1 - \bar{F}(x+1|p = p_U(x)) \\
&= \frac{\alpha}{2}, \text{ then } \bar{F}(x+1|p = p_U(x)) = 1 - \frac{\alpha}{2}.
\end{aligned}$$

This implies

$$\begin{aligned}
p_U(x) &= \frac{1}{1+\frac{n-(x+1)+1}{x+1}F_{2(n-(x+1)+1),2(x+1),1-\alpha/2}} \\
&= \frac{1}{1+\frac{n-x}{x+1}F_{2(n-x),2(x+1),1-\alpha/2}} \\
&= \frac{1}{1+\frac{n-x}{x+1}\frac{1}{F_{2(x+1),2(n-x),\alpha/2}}} \qquad (\text{See } F - 1/F \text{ relationship}) \\
&= \frac{\frac{x+1}{n-x}F_{2(x+1),2(n-x),\alpha/2}}{1+\frac{x+1}{n-x}F_{2(x+1),2(n-x),\alpha/2}}.
\end{aligned}$$

Thus a $1-\alpha$ confidence interval for p is

$$p_L(x) \leq p \leq p_U(x),$$

which is

$$\frac{1}{1+\frac{n-x+1}{x}F_{2(n-x+1),2x,\alpha/2}} \leq p \leq \frac{\frac{x+1}{n-x}F_{2(x+1),2(n-x),\alpha/2}}{1+\frac{x+1}{n-x}F_{2(x+1),2(n-x),\alpha/2}}.$$

□

Now that we've established a proven way to construct a $1-\alpha$ confidence interval for the p parameter of a binomial experiment, let's go back to the Recovery Time example and see how it can be used to construct a confidence interval for the p-values. We know that the exact p-value is the proportion of the statistic values in the randomization reference distribution that are less than or equal to the observed value. We used Monte-Carlo sampling and obtained a sample of size M from this reference

distribution. Let X be the number of sample values that are less than or equal to the observed value. Then we know $X \sim binomial(M+1, p)$ (we include the observed value in the sample as explained earlier) where p is the true proportion of the values that are less than or equal to the observed value in the randomization reference distribution. This true proportion is indeed the p-value of the test, and thus obtaining a confidence interval for the p-value is the same as obtaining a confidence interval for p and we use the formula given in Theorem 2.

We now use this to compute a confidence interval for the p-value for our example using Monte-Carlo sampling. In the Monte-Carlo sample, x is the actual number of sample values that are less than or equal to the observed value, and n = M + 1. Plug these numbers into the formula given in Theorem 2, we obtain a $1-\alpha$ confidence interval for p. This process, however, can be automated using our function `pval`. There are three new optional arguments to accomplish the job, which are conf.int (to turn on the confidence interval estimation), conf.level (specifies the confidence coefficient), and use ("c" for using Central Limit Theorem method, and "b" for using binomal distribution method). Note that when using these arguments, the monte argument must be turned on because otherwise, we are not estimating the p-value.

```
> pval(RT,4,"l","m",monte = TRUE,B = 999,conf.int = TRUE,conf.level = .999, use
= "b")
$mean.diff.pval.conf.int
[1] 0.07137552 0.13477500
> pval(WL,9,"b","s",plot = TRUE,type = "l",monte = TRUE,B = 999,conf.int
= TRUE,conf.level = .999, use = "b")
$sum.pval.conf.int
[1] 0.0478817 0.1027194
$sum.pval.conf.int
[1] 3.197650e-05 1.199101e-02
```

These confidence intervals do contain the exact p-values for our previous examples, which again are 0.0857, 0.0717, and 0.002. This feature is especially useful when the sample becomes too large which makes the computation based on the exact permutation distribution time-consuming. We see this more in the next two sections when we study the multivariate case and the one-way layout experiment.

2.5 Multivariate Case

Consider the same study on the two species. This times, instead of studying the wing length and antennae length data individually, we combine data and treat it as a bivariate data. We want to test if the means of these bivariate vectors differ signifiantly. Under the null hypothesis, our sample is again one of ${}_{15}C_9$ equally likely division of the observed bivariate vectors into the two samples of 9 and 6 subjects. Thus, the construction of our reference distribution is the same. We use the popular two-sample Hotelling T^2 statistic, which is defined as

$$T^2 = (\bar{X} - \bar{Y})' \left[\frac{S_X}{n} + \frac{S_Y}{m} \right]^{-1} (\bar{X} - \bar{Y})$$

where $\bar{X}$ and $\bar{Y}$ are the sample mean vectors and S_X and S_Y are the sample covariance matrices. The reference distribution of T^2 is again generated by recaculating the value of T^2 for each permutation of the data.

The task to generate this reference distribution, however, is not easy and currently has not been implemented in any commercial software. Thus, the exact p-value for our test can not be computed using any commerical software. Ernst obtained a Monte-Carlo sample approximation of this reference distribution, and based on it, he obtained an estimated p-value of 0.001 and a 99.9% confidence interval for the exact p-value (using binomial distribution) which is (0.0000005, 0.0099538). This upper limit of this confidence interval is less than a chosen level of significance $\alpha = 0.01$, thus if the test allows a 1% chance of getting a Type I error is signficant. We reject the null hypothesis and support the alternative that the means of these bivariate vectors dffer significantly. This means at least one of the univariate means differs signficantly. This is consistent with our univariate tests in which, if $\alpha = 0.01$, the means of antennae length differ significantly among the two populations of flies.

We replicate this result using our `pval` function and show that `pval` has the ability to generate and compute the exact p-value, which makes it more versatile than any commerical software in the multivariate case. Below is the exact p-value computed using `pval` along with the graph of the exact reference distribution of the Hotelling T^2 statistic, as displayed in Figure 8.

```
WA <- t(matrix(c(WL,AL),,2)) #WL, AL are the wing and antenna length vector
> pval(WA,9,"r","h",plot = TRUE,type = "h")
[1] 0.0001998002
```

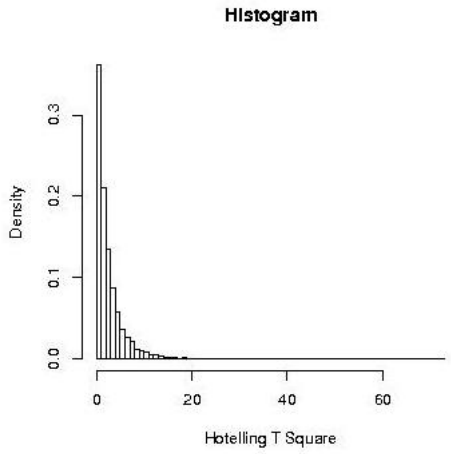

Figure 8: *The histogram of the randomization reference distribution of the Hotelling T^2 statistic.*

The exact p-value (0.001998) indicates that our exact test is significant for any popular level of significance. Note that this value is contained in the confidence interval for p-value using Monte-Carlo sampling and binomial distribution that was shown earlier. Note that the `pval` function also implements the Monte-Carlo method, which provides a sample approximation of the exact reference distribution of the Hotelling T^2 statistic, as displayed in Figure 9, and an estimated p-value along with the confidence interval for the exact p-value. Code follows.

```
> pval(WA,9,"r","h",plot = TRUE,type = "h",monte = TRUE,B = 999,conf.int = TRUE
```

```
,conf.level = .999, use = "b")
$hotelling.pval
[1] 0.001
$hotelling.pval.conf.int
[1] 5.001249e-07 9.953814e-03
```

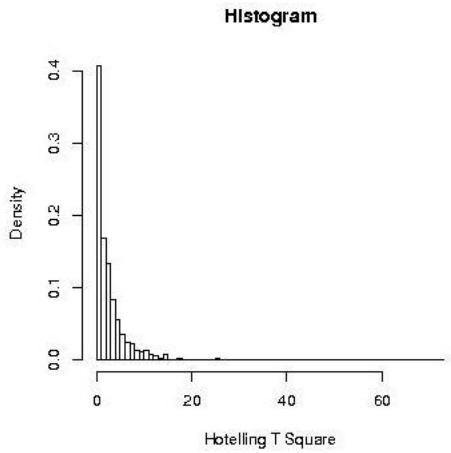

Figure 9: *A histogram of a Monte-Carlo random sample of $M = 9999$ values of the Hotelling T^2 statistic.*

3 One-Way Layout Experiment

Let's extend our study to caess when there are three or more treatments or populations. They are called one-way layout or k-sample experiments. Consider the data in Table 5. In this experiment, 15 subjects were randomly assigned to one of three different typeface styles and their reading speeds were recorde. One subject was unable to complete the exercise for reasons unrelated to the experiment, so one has four subjects. We want to test if there is a difference between these subjects, so our hypotheses are:

H_0: There is no difference between the treatments
H_1: At least one treatment is different from others

We suspect that the typeface styles do make a difference in reading speeds. Again, if the null hypothesis is true, then our sample is one of ${}_{14}C_{5,4} = 252,252$ equally likely randomization of the 14 subjects into 3 each treatments of 5, 4, and 5 subjects. Thus, the construction of our reference distribution in a one-way layout experiment is the same as that in a two-sample experiment. We could use either the one-way F statistic, which is defined as $F = MSTR/MSE = (SSTR/(k-1))/(SSE/(n-k))$ where $SSTR = \sum_{j=1}^{k} n_j(\bar{y}_{.j} - \bar{y}_{..})^2$ and $SSE = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij} - \bar{y}_{.j})^2$ or the k-sample T statistic, which is defined as $T = \sum_{j=1}^{k} n_j \bar{y}_{.j}^2$. We show below that there is an increasing monotone relationship between

Table 5: *Reading speeds for 15 subjects randomly assigned to three typeface styles.*

	Typeface Style		
j	1	2	3
	135	175	105
	91	130	147
	111	514	159
	87	283	107
	122		194
$\bar{Y}_{.j}$	109.2	275.5	142

the two statistics. We have

$$F = \frac{MSTR}{MSE} = \frac{SSTR/(k-1)}{SSE/(n-k)} = \frac{(SSTO - SSE)/(k-1)}{SSE/(n-k)} = \frac{n-k}{k-1}\left(\frac{\sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij} - \bar{y}_{..})^2}{SSE} - 1\right)$$

and

$$SSE = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij} - \bar{y}_{.j})^2 = \sum_{j=1}^{k}\sum_{i=1}^{n_j} y_{ij}^2 - \sum_{j=1}^{k} n_j \bar{y}_{.j}^2 = \sum_{j=1}^{k}\sum_{i=1}^{n_j} y_{ij}^2 - T.$$

Then

$$F = \frac{n-k}{k-1}\left(\frac{\sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij} - \bar{y}_{..})^2}{\sum_{j=1}^{k}\sum_{i=1}^{n_j} y_{ij}^2 - T} - 1\right).$$

This implies F is an increasing monotone function of T. It's easier to compute T than F, so we generate the reference distribution of T instead of F. Again, this reference distribution is generated by recaculating the value of T for each randomization of the data into the three treatments. The p-value for the test based on this reference distribution is the proportion of the test values that are greater than or equal to the observed value of T. Again, we use the `pval` function to generate a graph of the exact reference distribution of T, displayed in Figure 10, and compute the exact p-value based on this reference distribution. This is a unique feature of the function `pval` that is not available in commercial software. Details are below.

```
TS <- c(135,91,111,87,122,175,130,514,283,105,147,159,107,194) # Data vector
[1] "Thu Dec 10 09:49:51 2009"
> pval(TS,c(5,4),"r","t",plot = TRUE,type = "h")
[1] 0.01090180
> date()
[1] "Thu Dec 10 10:18:56 2009"
```

The exact p-value is 0.01 implies that if we allow a Type I error of less than 5% chance, then the test is significant. Put another way, the typeface styles make a difference in reading speeds. The task, however, is time-consuming. As seen above, it took about 30 minutes to generate the exact reference distribution of T and to compute the p-value based on this reference distribution. To shorten the computing time, we again use Monte-Carlo sampling to generate an approximation of the exact

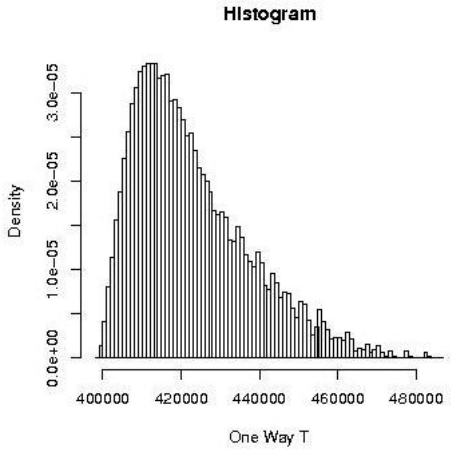

Figure 10: *The histogram of the randomization reference distribution of the test statistic* $T = \sum_{i=1}^{k} n_i \bar{Y}_i^2$ *for the data in Table 5.*

reference distribution, displayed in Figure 11 and compute an estimated p-value and its confidence interval based on this approximated reference distribution. Details are below.

```
> pval(TS,c(5,4),"r","t",plot = TRUE,type = "h",monte = TRUE,B = 9999,conf.int
= TRUE, conf.level = .99,use = "b")
$t.pval
[1] 0.0118
$t.pval.conf.int
[1] 0.009201403 0.014874214
```

The approximation reference distribution and the estimated p-value are very close to the exact results and the confidence interval for p-value contains the exact p-value. And this process only took a couple minutes. This again confirms that Monte-Carlo sampling is a fast and highly accurate way of generating an approximation of the exact reference distribution as well as caculating an estimated p-value.

Pairwise Comparisons

As concluded earlier, there is a difference between the treatments. To find out which pairs of treatments are different, we consider all pairwise comparisons and call it pair-wise comparison test. Two treatments are declared significantly different if their treatment means differ by an amount greater than or equal to a critical value C_α. The probability of commiting a Type I error is then the probability that at least one pair differ sigficantly. Put another way,

$$P(\text{Type I error}) = P(max_{1\le i<j\le k}|\bar{y}_{.i} - \bar{y}_{.j}| \ge C_\alpha | H_0) = \alpha.$$

This implies C_α is the $1-\alpha$ quantile of the reference distribution of the maximum absolute pairwise mean differences statistic $d = max_{1\le i<j\le k}|\bar{y}_{.i} - \bar{y}_{.j}|$. Note that this C_α does not depend on the

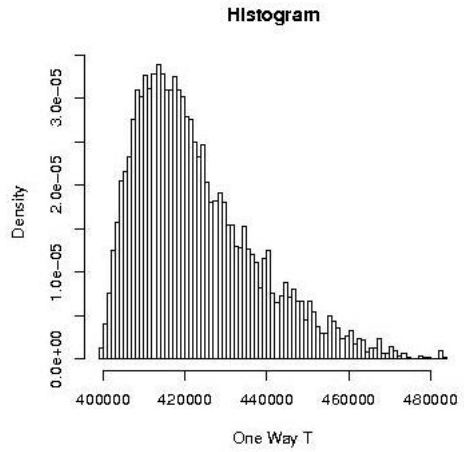

Figure 11: *A histogram of a Monte-Carlo random sample of $M = 9999$ test statistic $T = \sum_{i=1}^{k} n_i \bar{Y}_i^2$ for the data in Table 5.*

number of pairwise comparisons, which is similar to Tukey's approach (See Montgomery, 1997, page 106). However, ours is distribution-free whereas Tukey's procedure assumes normality for validity. Under the null hypothesis, our sample is one out of ${}_{14}C_{5,4} = 252,252$ equally likely randomizations of the 14 subjects into 3 each treatments of 5, 4, and 5 subjects. Thus, a reference distribution of d is constructed the same way as the reference distributions of other statistics. We use the function `pval` to construction this reference distribution, displayed in Figure 12, and compute its $1 - \alpha$ quantile as below.

```
> pval(TS,c(5,4),"r","t",sig.pairs = TRUE,sig.level = .05)
$c.alpha
   95%
142.05
```

The exact value of C_α is 142.05. Among three pairs of treatment, treatment 1 and 2 have their absolute mean difference of 166.3 that is more than 142.05. Thus, treatment 1 and 2 are different. Put another way, the reading speed differ significantly for typeface style 1 and 2. Since typeface 2 has higher mean reading speed, this implies the reading speed for typeface style 2 is significantly higher than that for typeface style 1.

The pairwse comparison test based on the exact reference distribution of the maximum absolute mean differences statistic, however, is also time-consuming. Thus, we use Monte-Carlo to approximate this reference distribution and based on it, compute an estimate C_α. It is shown in Serfling (1980, Theorem stated on page 75) that a sample quantile is a strongly consistent estimator of a population quantile. Thus $\hat{C}_\alpha$ is a strongly consistent estimator of C_α. We generate a Monte-Carlo reference distribution of d, displayed in Figure 13, and compute $\hat{C}_\alpha$ based on this approximated reference distribution as below.

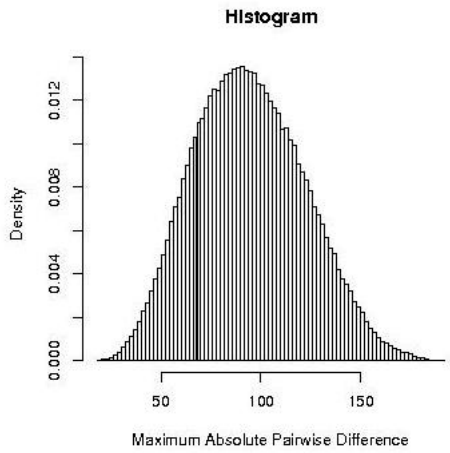

Figure 12: *The histogram of the randomization reference distribution of the maximum absolute pair-wise difference statistic for the data in Table 5*

```
> pval(TS,c(5,4),"r","t",monte = TRUE,B = 9999,conf.int = TRUE,conf.level = .99,use
= "b",sig.pairs = TRUE,sig.level = .05)
$c.alpha
  95%
142.6
```

It is once again shown that Monte-Carlo sampling gives us an approximated result that is very close to the exact result and more importantly, it is a lot faster than exact computation. This advantage has become clearer as the computation gets complicated as in the multivariate case or the number of treatment groups and the group sizes increase as in the one-way layout experiment.

4 Conclusion

In this paper, we've discussed how to perform nonparametric inferences using randomization and permutation reference distributions. We've clearly explained the differences and similarities between the two distributions and shown that Monte-Carlo sampling is an accurate, easy, and fast way to obtain an approximation of these distributions, which make inferences based on Monte-Carlo sampling valid and fast compared to those based on the exact reference distributions. We've stated and proved Theorem 1 which not only provides a proof for the standard way of constructing a confidence interval for the treatment effects Δ of a two-sample test based on recomputations of the reference distributions but also provides a basis for a new method of constructing this confidence interval based on the ordered values of $w_{k,d}$. This method does not require the recomputations of the reference distributions as the standard methods, thus it is a lot faster. We also stated and proved a theorem that was given as an exercise in C-B. This theorem is a basis for constructing a confidence interval for this exact p-value based on binomial distribution which results from the Monte-Carlo approximation. In addition, we introduced two self-written R functions, `pval` and `cint`, and demonstrate how to use them to replicate

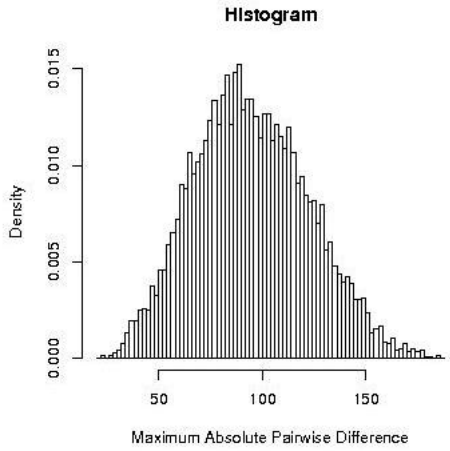

Figure 13: *A histogram of a Monte-Carlo random sample of $M = 9999$ maximum absolute pair-wise difference statistic for the data in Table 5.*

all the results in the article. In some situations, the two functions are more versitile than standard packages such as Resampling Stat or StatXact. However, the two functions are not without limitations. First, the algorithm we used only works to generate an exact reference distribution when the sample size is 24 or less. We can increase this limit to 27 using a 64-bit machine, but the limitation is still there. And secondly, the program does not generate $w_{k,d}$ automatically. Thus, for future works, we are rewriting the functions so that the limitations are overcome and $w_{k,d}$ and its ordered values $w_{(i)}$ are automatically calculated. We could then use the revised functions to execute the new method to obtain a confidence interval for Δ computationally.

Acknowledgment

I show my deepest appreciation to Dr. Mara Tableman for her support throughout this paper, especially the lively discussions that helped me to formulate and prove Theorem 1 and its corollaries and her reading recommendations that lead me to Theorem 2.

References

[1] CASELLA, G. and BERGER, R. L. (2002). *Statistical Inference,* 2nd Ed. Duxbury, Pacific Grove, CA

[2] ERNST, M. D. (2004). Permutation methods: a basis for exact inference. *Statistical Science* **19** 676-685

[3] GARTHWAITE, P. H. (1996). Confidence intervals from randomization tests. *Biometrics* **52** 1387-1393

[4] HIGGINS, J. J. (2004). *An Introduction to Modern Nonparametric Statistics.* Brooks/Cole, Pacific Grove, CA.

[5] MONTGOMERY, D. C. (1997). *Design and Analysis of Experiments,* 4th Ed. John Wiley & Sons, N.Y.

[6] SERFLING, R. J. (1980). *Approximation Theorems of Mathematical Statistics.* John Wiley & Sons, N.Y.